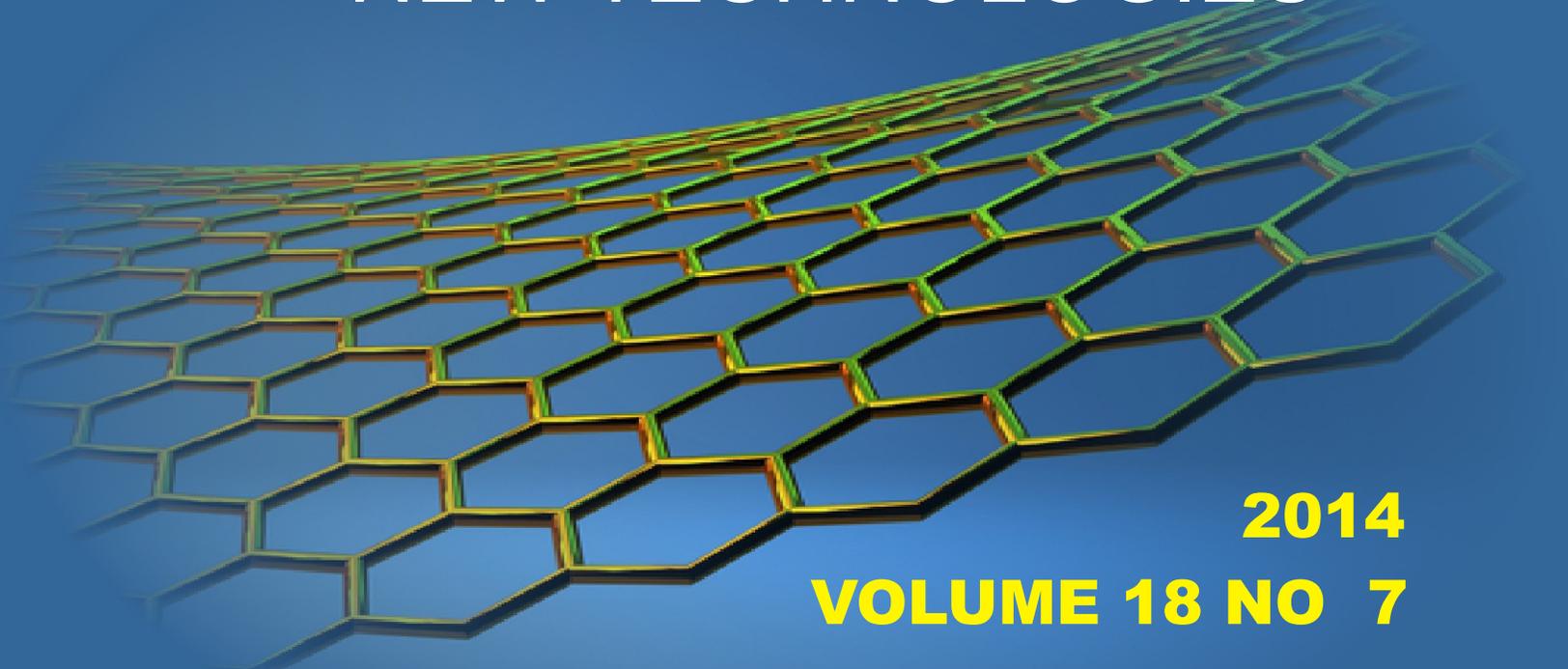


COMPUTER MODELLING
AND
NEW TECHNOLOGIES



2014
VOLUME 18 NO 7

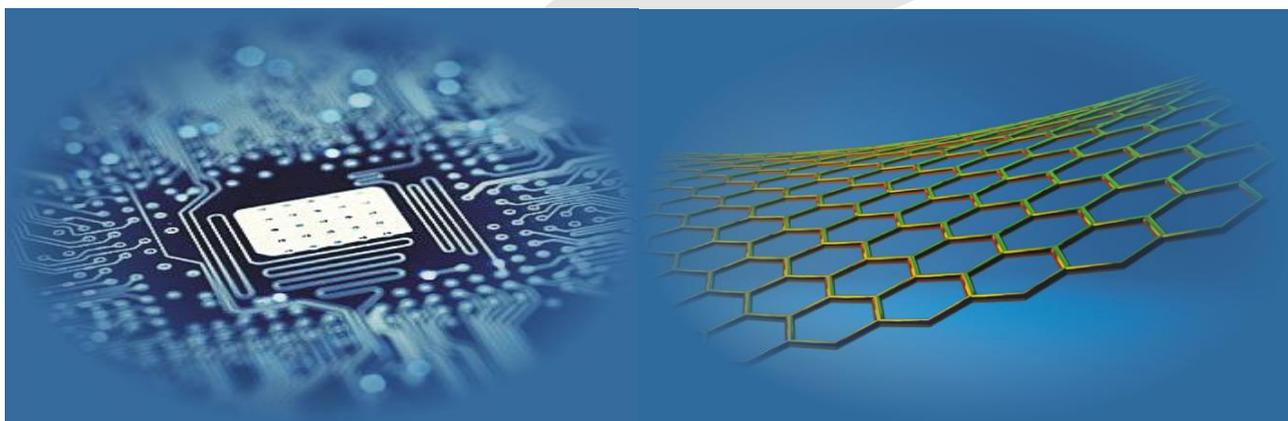
ISSN 1407-5806 ISSN 1407-5814 on-line

Transport and Telecommunication Institute
and
Latvian Transport Development and Education Association

Computer Modelling and New Technologies

2014 Volume 18 No 7

ISSN 1407-5806, ISSN 1407-5814 (*On-line: www.tsi.lv*)



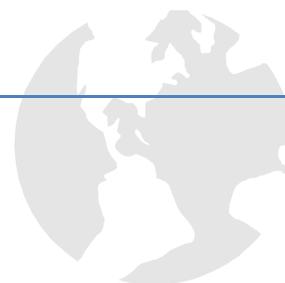
Riga – 2014

EDITORIAL BOARD

| | |
|--|--|
| Prof. Igor Kabashkin | Chairman of the Board , <i>Transport & Telecommunication Institute, Latvia</i> |
| Prof. Yuri Shunin | Editor-in-Chief , <i>Information Systems Management Institute, Latvia</i> |
| Prof. Adolfas Baublys | <i>Vilnius Gediminas Technical University, Lithuania</i> |
| Dr. Brent Bowen | <i>Embry-Riddle Aeronautical University, United States of America</i> |
| Prof. Olgierd Dumbrajs | <i>University of Latvia, Solid State Physics Institute, Latvia</i> |
| Prof. Sergey Maksimenko | <i>Institute for Nuclear Problem, Belarus State University, Belarus</i> |
| Prof. Vladimir Litovchenko | <i>V. Lashkaryov Institute of Semiconductor Physics of National Academy of Science of Ukraine, Ukraine</i> |
| Prof. Pavel D'yachkov | <i>Kurnakov Institute for General and Inorganic Chemistry, Russian Academy of Sciences, Russian Federation</i> |
| Prof. Stefano Bellucci | <i>Frascati National Laboratories – National Institute of Nuclear Physics, Italy</i> |
| Prof. Arnold Kiv | <i>Ben-Gurion University of the Negev, Israel</i> |
| Prof. Alytis Gruodis | <i>Vilnius University, Lithuania</i> |
| Prof. Michael Schenk | <i>Fraunhofer Institute for Factory Operation and Automation IFF, Germany</i> |
| Prof. Dietmar Fink | <i>University of Mexico, United Mexican States</i> |
| Prof. Ravil Muhamedyev | <i>International IT University, Kazakhstan</i> |
| Prof. Kurt Schwartz | <i>Gesellschaft für Schwerionenforschung mbH, Darmstadt, Germany</i> |
| Prof. Eva Rysiakiewicz-Pasek | <i>Institute of Physics, Wroclaw University of Technology, Poland</i> |
| | |
| Contributing Editor | Prof. Victor Gopeyenko, <i>Information Systems Management Institute, Latvia</i> |
| Literary Editor | Prof. Tamara Lobanova-Shunina, <i>Riga Technical University, Latvia</i> |
| Technical Editor , secretary of Editorial Board | MSc Comp Nataly Burluckaya, <i>Information Systems Management Institute, Latvia</i> |

| Journal topics: | Host Organization | Supporting Organizations |
|--|---|---|
| mathematical and computer modelling computer and information technologies natural and engineering sciences operation research and decision making nanoscience and nanotechnologies innovative education | Transport and Telecommunication Institute | Latvian Transport Development and Education Association Latvian Academy of Sciences Latvian Operations Research Society |
| Articles should be submitted in English . All articles are reviewed. | | |

| | |
|---|---|
| EDITORIAL CORRESPONDENCE | COMPUTER MODELLING AND NEW TECHNOLOGIES, 2014, Vol. 18, No.7 ISSN 1407-5806, ISSN 1407-5814 (on-line: www.tsi.lv) |
| Transport and Telecommunication Institute | Scientific and research journal The journal is being published since 1996 |
| 1 Lomonosova, Bld 4 , LV-1019, Riga, Latvia Phone: (+371) 67100594 Fax: (+371) 67100535 E-mail: yu_shunin@inbox.lv www.tsi.lv | The papers published in Journal 'Computer Modelling and New Technologies' are included in: INSPEC (since 2010) , www.theiet.org/resources/inspec/ VINITI (since 2011) , http://www2.viniti.ru/ CAS Database http://www.cas.org/ El Compindex |



Content

| | | |
|--|---|-----|
| Editors' Remarks | | 5 |
| Mathematical and Computer Modelling | | |
| Pengtao Jia, Jun Deng, Shuhui Liang | A fuzzy combined forecasting model of coal spontaneous combustion | 7 |
| Guoliang Cai, Shengqin Jiang Jiang, Shuiming Cai, Lixin Tian | Exponential synchronization of complex networks with non-delayed and delayed coupling via hybrid control | 12 |
| Mingming Qi, Yang Xiang | Tensor modular sparsity preserving projections for dimensionality reduction | 18 |
| Gongfa Li, Fuwei Cheng, Honghai Liu, Guozhang Jiang, Jia Liu | Coke oven production process hybrid intelligent control | 23 |
| Wei Huang | Research on output regulation for saturated systems | 30 |
| Zhang Hui, Wu Jinzhao, Tan Hongyan, Yang Hao | Approximate trace equivalence of real-time linear algebraic transition systems | 36 |
| Yijun Liu, Sheng He, Yao Wang, Xiumei Wang | A comparative study on artificial neural networks for environmental quality assessment | 41 |
| Yuan Xiangyue, Chen Zhongjia | Design of Q450 pellet molding machine and force analysis of its molding assembly based on SolidWorks | 48 |
| Yan Li, Zhe Zhang, Guihong Jiang, Xiaofeng Cui | Model driven testing distributed environment monitoring system | 54 |
| Huang Wenzhun, Zhang Shanwen | Source enumeration algorithm based on eigenvector: revisit from the perspective of information theory | 60 |
| Information and Computer Technologies | | |
| Jun Zhao, Zhong Ma, Xiangjun Wu | A method for improving real-time communication of switched Ethernet | 65 |
| Guangchun Gao, Kai Xiong, Shengying Zhao, Cui Zhang | Optimal adaptive wavelet transforms without using extra additional information | 71 |
| Jing Jiang, Shuang Xu, Guangyue Lu, Yongbin Xie, Yanxia Liang | A large-scale MIMO channel information feedback algorithm based on compressed sensing | 80 |
| Yong Li, Jiang Yu, Rong Zong, Yan Zhang, Jihong Shi, Jinsong Hu | Heterogeneous networks model for lower error using concatenated encoding | 86 |
| Zhenrong Deng, Xingxing Tang, Chuan Zhang, Xi Zhang, Wenming Huang | Improvements and implementation of the permission system based on RBAC model | 92 |
| Jin Yang, Lingxi Peng, Tang Liu | Anti-spam model based on AIS in cloud computing environments | 97 |
| Yanjun Zhao, Chunying Zhang | Research and application on set pair entity similarity model of social network | 103 |
| Jizhen Ye, Jian Wei, Yan Huang, Jingliang Peng | Comparative study of DXT1 texture encoding techniques | 110 |
| Xian Wu, Yan Huang | Real-time and interactive browsing of massive mesh models | 116 |
| Xiaofeng Li, Yanfang Yang, Limin Jia | A kernel induced energy based active contour method for image segmentation | 122 |
| Yiming Yuan, Ming Jiang, Wengen Gao | Image fusion based on MPCNN and DWT in PCB failure detection | 128 |
| Min Yang, Yaoliang Song, Qianmu Li | Research on virus transmission of online social network | 133 |
| Lin Bai, Meng Hui | SVM classification of hyperspectral images based on wavelet kernel non-negative matrix factorization | 140 |
| Operation Research and Decision Making | | |
| Wenlong Wang, Xinmei Liu, Xiaojie Zhang | The optimal promised quality defect model for service guarantees | 147 |
| Yue Yu, Yong-shi Hu, Ming-xing Xu | Research on supply chain competition advantage under repeated games | 159 |
| Yan Li, Dong Wei, Yangyang Chen | The development and evolution of bridge in Chongqing China | 166 |
| Yaoting Chen, Xiaowei Lin | A comparative study on efficiency of two different circulation modes of agricultural products based on DEA model: wholesale market and logistics distribution centre | 175 |
| Qing-Huang Huang, Ming Gao | A study on mechanism of environmental protection industry innovation under open innovation - the intermediary effect based on the enterprise network dynamic capability | 181 |
| Jiansheng Zhang, X.zhong Hu | An analysis on the growth and effect factors of TFP under the energy and environment regulation: data from China | 191 |
| Liping Fu, Juan Li | Analysis of the public satisfaction index of public cultural services based on the grey correlation AHP method | 197 |
| Yezheng Liu, Jun Liu | Asymmetric effects of exchange rate pass-through: an empirical analysis among China, the United States and Japan | 204 |
| Ning Gao, Cai-Yun Gao | Deformation forecasting with a novel high precision grey forecasting model based on genetic algorithm | 212 |
| Hua Zhang, Shunchang Liu | Effect judgment and effectiveness estimation of anti-dumping duty — an example of the case of canned mushroom | 218 |
| Zongyi Xing, Lingli Mao, Limin Jia, Yong Qin | Identification of key subsystems for urban rail vehicles based on fuzzy | 224 |

| | | |
|--|---|-----|
| | comprehensive evaluation | |
| Yu Wu | Reputation risk contagion and control of rural banks in china based on epidemic model | 229 |
| Fei Meng, Jianliang Wei | Research on the influential factor of consumer model based on online opinion leader | 236 |
| Ming-xing Xu, Yue Yu, Yong-shi Hu | Service and revenue sharing strategies in a dual-channel supply chain with fairness concerns | 244 |
| Yibo Du, Jin Zhang | Time-varying decision-making for hazardous chemical transportation in a complex transportation network | 253 |
| Hailong Lu, Yong Cen | Research and implementation on integration information platform in China tobacco industry enterprise | 259 |
| Xilong Liu, Yizeng Chen | A fuzzy clustering approach of the customers' demands, which influence the e-banking service quality | 267 |
| Yun-jun Yu, Sui Peng, Yun-tao Xue, Chao Tong, Zi-heng Xu | An autonomous decision making algorithm applied for the evaluation of power quality | 273 |
| Shanshan Shang, Jianxin You | A simulation model on the formation of knowledge-based collaborative networks | 278 |
| Rende Yu, Juan Shen | Analysis on road traffic accidents spatial distribution based on the multi-fractal theory | 283 |
| Shuiping Zhang | Research on the principal-agent problems in China's low-carbon ecological urban construction | 290 |
| Zhao-Xing Li, Li-le He | A multi objective optimization algorithm for recommender system based on PSO | 298 |
| Hoan Manh Dau, Ning Xu | The effectiveness of using methods two-stage for cross-domain sentiment classification | 304 |
| Ying Lu, Junping Xie | Multi-objective hub location problem in hub-and-spoke network | 309 |
| Ilana Ter-Saakova, Nataly Podolyakina | Analysis of necessary investments in the production and warranty service of innovative products considering the necessity of their backup | 317 |
| Xing Yu, Guohua Chen | The continuous-time optimal portfolio using a multivariate normal inverse Gaussian model | 322 |
| Peng Ma | Computer information technology and agricultural logistics management system | 325 |
| Nature Phenomena and Innovative Engineering | | |
| Junmei Zhao, Zhijie Zhang, Yifeng Ren | Research on speed regulation system for matrix converter fed induction motor | 330 |
| Xiaohui Liu, Feng Dai, Jianfeng Liu | Research on the anisotropy of the coal rock under different bedding direction | 337 |
| Yanli Feng, Dashe Li, Shue Liu | Research on the laser transmission simulation based on random phase screen in atmospheric turbulent channel | 344 |
| Zhao Han, Yuan Rao, Wentao Chen, Junkai Huang | The design of a dynamic slope compensation circuit for boost DC-DC Converter | 350 |
| Xingjie Chen, Xiaodong Chai, Xining Cao | The time-frequency analysis of the train axle box acceleration signals using empirical mode decomposition | 356 |
| Xianfeng Zheng, Zheng Fan | Research into voltage sag online detection technology based on wavelet tree | 361 |
| Chang Chen, Guojin Chen, Shaohui Su, Haiqiang Liu | Modelling and simulation of marine rudder system in a unified M&S platform | 368 |
| Zhangming Peng, Guojin Chen, Shaohui Su | Study on quantitative diagnosis method of valve clearance based on cylinder head vibration signal of diesel engine | 373 |
| Dong Jian-Gang, Zhang Feng, Zhang Yong-Heng | A water quality changing prediction model for agricultural water-saving irrigation based on PSO-LSSVR | 377 |
| Jian-Long Ding, Weifang Chen, Ji Gao | Intelligent data-collaboration mechanism under the distributed application environment | 382 |
| Yunxia Zhang, Chenglong Dai, Jifeng Cui | Lifetime forecasting for hemispherical resonator gyroscope with wavelet analysis-based GM(1,1) | 388 |
| Xiaodong Huang | Automatic license plate detection based on colour gradient map | 393 |
| Jun Li, Xiaoyu Liu, Shiping Zhao | Prediction model of recast layer thickness in die-sinking EDM process on Ti-6Al-4V machining through response surface methodology coupled with least squares support vector machine | 398 |
| Hao Wu, Dewen Seng, Xujian Fang | Construction of a computer simulation platform for optical experiments | 406 |
| Authors' Index | | 413 |
| Cumulative Index | | 414 |

Editors' Remarks

A General Summary

by Rudyard Kipling

We are very slightly changed
From the semi-apes who ranged
India's Prehistoric clay;
He that drew the longest bow
Ran his brother down, you know,
As we run men down to-day.

"Dowb," the first of all his race,
Met the Mammoth face to face
On the lake or in the cave:
Stole the steadiest canoe,
Ate the quarry others slew,
Died -- and took the finest grave.

When they scratched the reindeer-bone,
Some one made the sketch his own,
Filched it from the artist -- then,
Even in those early days,
Won a simple Viceroy's praise
Through the toil of other men.
Ere they hewed the Sphinx's visage
Favouritism governed kissage,
Even as it does in this age.

Who shall doubt "the secret hid
Under Cheops' pyramid"
Was that the contractor did
Cheops out of several millions?
Or that Joseph's sudden rise
To comptroller of Supplies
Was a fraud of monstrous size
On King Pharaoh's swart Civilians?

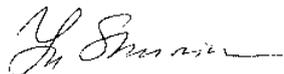
Thus, the artless songs I sing
Do not deal with anything
New or never said before.
As it was in the beginning
Is to-day official sinning,
And shall be for evermore!

Rudyard Kipling (1809-1849) *

This 18th volume No.7 presents actual papers on main topics of Journal specialization, namely, **Mathematical and Computer Modelling, Computer and Information Technologies, Operation Research and Decision Making and Nature Phenomena and Innovative Engineering.**

Our journal policy is directed on the fundamental and applied sciences researches, which are the basement of a full-scale modelling in practice. This edition is the continuation of our publishing activities. We hope our journal will be interesting for research community, and we are open for collaboration both in research and publishing. We hope that journal's contributors will consider the collaboration with the Editorial Board as useful and constructive.

EDITORS



Yuri Shunin



Igor Kabashkin

* **Joseph Rudyard Kipling** (30 December 1865 – 18 January 1936) was an English short-story writer, poet, and novelist. He is chiefly remembered for his tales and poems of British soldiers in India and his tales for children. He was born in Bombay, in the Bombay Presidency of British India, and was taken by his family to England when he was five years old. Kipling is best known for his works of fiction, including *The Jungle Book* (a collection of stories, which includes and his poems, including "Mandalay" (1890), "Gunga Din" (1890), "The Gods of the Copybook Headings" (1919), "The White Man's Burden" (1899), and "If—" (1910). He is regarded as a major "innovator in the art of the short story"; his children's books are enduring classics of children's literature; and his best works are said to exhibit "a versatile and luminous narrative gift".



A fuzzy combined forecasting model of coal spontaneous combustion

Pengtao Jia^{1*}, Jun Deng², Shuhui Liang¹

¹School of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an Shaanxi 710054, China

²School of Energy Engineering, Xi'an University of Science and Technology, Xi'an Shaanxi 710054, China

Received 25 July 2014, www.tsi.lv

Abstract

This paper focuses on the effective analysis of the coal spontaneous combustion monitoring data, so as to realize the accurate and reliable coal spontaneous combustion limit parameter prediction. Firstly, a weighted multimember fuzzy operation model was constructed. When the additive generator of the model changes, this model can generate new operation clusters. Based on it, a new combined forecasting model of coal spontaneous combustion limit parameter is proposed. The new model can use linear and nonlinear models as its single forecasting models. Its combination is variable and has good generalization ability. Then, the BP neural network model and the support vector machine were used as the single forecasting models of the new model. Finally, for realizing the optimal combination of single models, genetic algorithm and least square method were used to evaluate parameters of new model. The experimental analysis shows that the new model leads to less error and better performance than single models. It can be concluded that the new combined forecasting model is suitable for coal spontaneous combustion.

Keywords: coal spontaneous combustion, limit parameters, combined forecasting, genetic algorithm, least square method

1 Introduction

Coal spontaneous combustion is an important problem in its mining, long distance transportation, and storage, in terms of both mine safety and economics. It has large proportion and wide coverage in China mine. According to statistics, the coal spontaneous combustion fire accounted for more than 94% of the total number of coal mine fire in China [1]. The coal spontaneous combustion is a physical-chemical process, which is extremely complex, dynamic changing and automatic accelerating. Virtually, it is a slow and automatic process of oxidation, heating, and then burning. Coal spontaneous combustion requires certain external conditions, of which the limiting ones are called limit parameters. The limit parameters are considered as the strong basis for estimating the dangerous areas of coal spontaneous combustion. There are some main parameters, such as the lower oxygen concentration, the ceiling air leakage strength and the minimum float coal thickness [2]. Because influencing factors of the limit parameters are relatively complex [3], and it is meaningful to establish appropriate models to forecast the limit parameters of coal spontaneous combustion [4]. At present, there are basically two kinds of forecasting models of limit parameters as follows:

One, mathematical models based on numerical simulation [2, 5-7]. To simplify the calculation, in the process of calculation, only the effects of main factors are considered, the secondary factors are ignored or taken as constant values. Then the limit parameters can be

estimated approximately. So there often has large deviation between its calculation results and the actual ones.

Two, forecasting models based on data mining methods. In order to overcome the deficiency of mathematical models, some scholars establish the prediction model through data mining methods, such as the neural network prediction model [8, 9], rough set neural network model [10], support vector machine (SVM) model [4, 11], rough set support vector machine forecasting model [12], regression analysis [13].

As we know the generalization abilities of single forecasting models are poor. Although many scholars perform some efforts on single forecasting models for improving forecasting capability, these investigations mainly focus on exploring new forecasting algorithm and improving old forecasting algorithm, and do not overcome disadvantage of single forecasting models. One of the important directions in improvement of the forecasting ability is the integration of multiple single forecasting models. Several effective methods have been proposed to combine the results of the single forecasting models, such as product operator, mean operator, median operator, max operator, min operator and majority vote method. But these integration methods always have not good capability in different datasets. In this paper, we put forward a new fuzzy combined forecasting model to predict limit parameters of coal spontaneous combustion based on a weighted multimember operation model.

*Corresponding author e-mail: pengtao.jia@gmail.com

2 The weighted multimember fuzzy operation model

Many existing fuzzy operation models were binary operators with the same weight. But the various factors in the actual complex systems usually have more than two factors and different weights. So weighted parameters and many factors are introduced into the fuzzy operation model, we construct a weighted multimember fuzzy operation model.

Theorem 1: Assume that function $f(x)$ is an additive generator, then

$$G(x_1, x_2, \dots, x_n, \alpha_1, \alpha_2, \dots, \alpha_n) = f^{-1}(\max(f(0), \sum_{i=1}^n \alpha_i f(x_i) - 1)) \tag{1}$$

is a weighted multimember fuzzy operation model.

Where $x_i \in R^+$ ($i=1, 2, \dots, n$), α_i is the weight of $f(x)$,

$$\alpha_i \in [0, 1] \text{ and } \sum_{i=1}^n \alpha_i = 1.$$

When $f(x)$ changes, this model can generate new operation clusters. According to this variability, the model gets good generalization ability. For example, set $f(x) = x^p$, Equation (1) can generate an operator cluster as follows:

$$G(x_1, x_2, \dots, x_n, \alpha_1, \alpha_2, \dots, \alpha_n) = f^{-1}(\sum_{i=1}^n \alpha_i x_i^p - 1) = \sqrt[p]{\sum_{i=1}^n \alpha_i x_i^p - 1} \tag{2}$$

where, p is the parameter of the generator and $p \in (-\infty, 0) \cup (0, +\infty)$. Later, the combined forecast model will be constructed based on Equation (2).

3 Combined forecasting of coal spontaneous combustion limit parameters

3.1 FORECASTING MODEL BASED ON BP NEURAL NETWORK

Limit parameters of coal spontaneous combustion are influenced by many factors. There is a nonlinear relationship between influence factors and limit parameters, so the BP neural network can be used to predict the limit parameters [8]. In this paper, BP neural network with three layers is used as the single forecasting model to predict limit parameters of coal spontaneous combustion. In the input layer, five nodes are used to input the impact factors. If the impact factors need more comprehensive consideration, additional nodes should be added to the input layer. The predictive value of the limit parameter is considered as the only node in the output layer. After training and comparing the forecasting model several times, it found that the training effect is better

when only use a hidden layer with ten nodes, and select the logarithmic function *sigmod* as the excitation function. Then set the initial weights in (-1, 1) and the convergence error to 1e-6. Take the ceiling air leakage strength for instance, the structure of BP neural network forecasting model is shown in Figure 1.

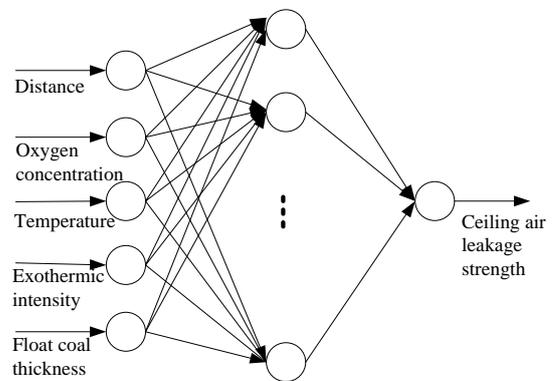


FIGURE 1 The BP forecasting model for limit parameters of coal spontaneous combustion

3.2 FORECASTING MODEL BASED ON SVM

Support Vector Machine (SVM) is a learning machine based on minimum structural risk and statistical learning theory. It has unique advantages in solving small sample, nonlinear and high dimensional pattern recognition problems. The basic idea of using the SVM algorithm to estimate the regression function is that firstly map the input space data x into a high-dimensional feature space through a nonlinear mapping, then proceed the linear regression in the high-dimensional space.

Least Squares Support Vector Machines (LS-SVM) is one kind of the support vector machine. It uses the least squares linear system instead the traditional support vector, that is, using the quadratic programming method to solve the pattern recognition problems, transforming the quadratic optimization of the original SVM algorithm into solving linear equations. So it can effectively reduce the computational complexity. In this paper, the LS-SVM is used as the single forecasting model to predict limit parameters of coal spontaneous combustion.

The modelling steps of LS-SVM forecasting model are as follows:

- 1) To describe the function fitting problem as an optimization problem;
- 2) To solve the optimization problem by using the Lagrange method, and convert it to solving the systems of linear equations. The training of the LS-SVM model is mainly to solve the systems of linear equations;
- 3) To get the LS-SVM fitting model, as follows:

$$y(x) = \sum_{k=1}^N \alpha_k K(x_k, x) + b \tag{3}$$

where, α_k is the support vector, $b \in R$ is the offset value. α_k and b can be obtained according to the training sample data. $K(x_k, x)$ is the kernel function which is a symmetric function and satisfies the Mercer condition.

Predicting by using the LS-SVM model simply needs to calculate the kernel function $K(x_k, x)$ between each training sample and the sample under test.

4) To select kernel function. Selecting the kernel function is an important part for building a model. Consider that the Gaussian radial basis kernel function (RBF) has good learning ability and wide domain of convergence, RBF is selected as the kernel function here, so:

$$K(x_k, x) = \exp(-\|x_i - x_j\|^2 / \sigma^2), \quad (4)$$

where, σ is the kernel parameter.

3.3 CONSTRUCT THE COMBINED FORECASTING MODEL

Modelling method of combined forecasting is a portfolio of predicting the same objects by using two or more predict methods. Theoretical research and practical application show that the combined forecasting has high ability to adapt to the change of the future predict environment, and it can enhance the stability of forecast, so as to achieve the purpose of improving the prediction precision [15,16]. The key of the combination forecasting model lies in its generalization ability. The model can be described as follows:

Assume that the actual observations of a certain prediction problem at time t is $y(t)$ ($t=1, 2, \dots, m$), there are n feasible forecast methods, the corresponding prediction models respectively are f_1, f_2, \dots, f_n , their predictive values respectively are:

$\hat{y}(t)$ ($t=1, 2, \dots, m; i=1, 2, \dots, n$), i.e. $\hat{y}(t) = f_i(t)$, and the weighted combination forecasting problem can be described as:

$$\hat{y}(t) = F(\hat{y}_1(t), \hat{y}_2(t), \dots, \hat{y}_n(t), \alpha_1, \alpha_2, \dots, \alpha_n) \quad (5)$$

where, the combination forecasting values are $y(t)$ ($t=1, 2, \dots, m$), F is the way of combination. Using Equation (5) aims to make the combination forecasting values better than the single prediction effects.

The combined forecasting model of this paper is based on the theory of BP neural network and support vector machine (SVM), the predictive values are $\hat{y}_{BP}(t)$ and $\hat{y}_{SVM}(t)$.

According to the characteristics and advantages of each single forecast model, different weights, such as α and $1-\alpha$ are assigned to each single one in the combined model. Consider Equation (2) as the F of Equation (5), then the combination forecasting model (CFM) can be described as:

$$\hat{y}(t) = \sqrt[p]{\alpha \hat{y}_{BP}(t)^p + (1-\alpha) \hat{y}_{SVM}(t)^p} - 1 \quad (6)$$

The parameters are estimated by combining genetic algorithm with least squares method. Due to the objectivity and inevitability of the prediction error, there are errors between the predictive values $\hat{y}(t)$ and the actual ones $y(t)$. Set:

$$E = \sqrt{(\hat{y}(t) - y(t))^2}. \quad (7)$$

Minimizing E is used as the evaluation of the objective function in genetic algorithm, and then the parameters in the combination model can be obtained.

Genetic algorithm (GA) is an adaptive global optimization search algorithm, which is formed by simulating the genetic and evolutionary process of organisms in the natural environment. Given its global optimization ability, GA is used as the parameter estimation module of CFM.

Set the following parameters of GA for parameter optimization:

- 1) The initial population is 20;
- 2) Use binary coding with eight numbers;
- 3) Select operation by using the uniform distribution random model;
- 4) Do crossover operation by the disperse cross;
- 5) Mutate operation by using gauss function.

4 Experiment and result analysis

4.1 EXPERIMENTAL DESIGN

Taking the ceiling air leakage strength prediction of coal spontaneous combustion limit parameters as an example, some main influence factors of the ceiling air leakage strength are selected as the input data, such as the exothermic intensity of coal, coal temperature, the measured oxygen concentration, distance from working face to the goaf prediction area and the float coal thickness. And the output data is the ceiling air leakage strength. The dataset from Xinzhou mine of China [8] is shown in Table 1.

Here, the data with number 1 to 20 was used as training samples, and that with number 21 to 25 was used as testing samples.

TABLE 1 Dataset of the combined forecasting model

| No. | Input data | | | | | Output data | |
|-----|-------------|-------------------------|----------------------|---|--------------------------|---|--|
| | Distance /m | Oxygen concentration /% | Coal temperature /°C | Exothermic intensity /10 ⁵ J·s ⁻¹ ·cm ⁻³ | Float coal thickness / m | Ceiling air leakage strength / cm ³ ·cm ⁻² ·s ⁻¹ | |
| 1 | 1.70 | 20.60 | 19.60 | 0.87 | 7.0 | 0.70 | |
| 2 | 2.50 | 20.04 | 20.30 | 1.04 | 6.0 | 0.83 | |
| 3 | 4.70 | 19.88 | 22.00 | 1.27 | 5.0 | 1.08 | |
| 4 | 7.60 | 19.03 | 22.50 | 1.34 | 4.0 | 1.56 | |
| 5 | 16.30 | 18.21 | 24.20 | 1.43 | 2.0 | 2.35 | |
| 6 | 20.50 | 17.99 | 25.60 | 1.51 | 3.0 | 2.58 | |
| 7 | 25.20 | 17.60 | 26.70 | 1.58 | 4.0 | 2.88 | |
| 8 | 29.10 | 17.36 | 26.80 | 1.58 | 3.0 | 3.17 | |
| 9 | 36.40 | 16.90 | 27.50 | 1.62 | 2.0 | 3.43 | |
| 10 | 43.90 | 15.74 | 28.30 | 1.67 | 3.0 | 3.87 | |
| 11 | 44.30 | 15.68 | 28.60 | 1.69 | 4.0 | 3.92 | |
| 12 | 47.00 | 14.91 | 28.10 | 1.66 | 5.0 | 4.12 | |
| 13 | 53.70 | 13.77 | 25.13 | 1.49 | 7.0 | 5.60 | |
| 14 | 56.40 | 13.09 | 24.80 | 1.47 | 6.0 | 5.77 | |
| 15 | 59.00 | 12.44 | 24.30 | 1.43 | 5.0 | 6.00 | |
| 16 | 61.20 | 11.93 | 23.60 | 1.40 | 4.0 | 6.53 | |
| 17 | 70.60 | 10.78 | 24.67 | 1.46 | 2.0 | 5.39 | |
| 18 | 74.30 | 9.81 | 26.30 | 1.55 | 2.0 | 4.18 | |
| 19 | 78.00 | 8.85 | 27.80 | 1.61 | 3.0 | 3.76 | |
| 20 | 89.20 | 7.14 | 30.40 | 1.79 | 3.0 | 2.89 | |
| 21 | 11.00 | 18.59 | 23.40 | 1.39 | 3.0 | 2.04 | |
| 22 | 39.70 | 16.50 | 27.90 | 1.64 | 2.0 | 3.54 | |
| 23 | 50.40 | 14.36 | 28.20 | 1.66 | 6.0 | 3.86 | |
| 24 | 66.80 | 11.18 | 24.20 | 1.43 | 3.0 | 6.12 | |
| 25 | 83.50 | 7.97 | 28.10 | 1.66 | 4.0 | 4.17 | |

Experimental steps are listed as follows:

1) To facilitate comparison, firstly the samples should be standardized to [0, 1]. The standardized formula is as follows:

$$norm(x_i) = \frac{x_i - \min(X)}{\max(X) - \min(X)} \tag{8}$$

2) Training the single forecasting method BP and SVM on the training sets, getting the prediction results $\hat{y}_{BP}(t)$ and $\hat{y}_{SVM}(t)$ on the test sets.

3) The CFM model is now used. The genetic algorithm is used to estimate parameters on the training sets, and then get the CFM prediction results $\hat{y}(t)$ on the test sets.

4) Normalizing all the prediction results, the normalized formula is as follows:

$$y_i = \hat{y} \times (\max(X) - \min(X)) + \min(X) \tag{9}$$

5) Using evaluation index to evaluate.

4.2 RESULTS AND ANALYSIS

The prediction results of CFM model are shown in Table 2, the parameters of the Equation (6), namely the combination forecasting model, obtained based on GA are $\alpha=0.261$ and $p=2.926$.

TABLE 2 Contrast of the test sample's expected outputs and the predicted results

| Sample number | Actual value | BP | LS-SVM | CFM |
|---------------|--------------|---------|---------|---------|
| 1 | 1.88 | 1.81977 | 1.93764 | 1.87373 |
| 2 | 3.52 | 3.66935 | 3.56403 | 3.61498 |
| 3 | 4.24 | 4.12589 | 4.37216 | 4.24232 |
| 4 | 6.01 | 6.01739 | 5.99676 | 6.00860 |
| 5 | 3.76 | 3.79867 | 3.61776 | 3.70324 |

The contrast results on the error evaluation index are shown in Table 3.

TABLE 3 Comparison of error percentage of prediction results of different models (%)

| Sample number | BP | Improved BP [8] | LS-SVM | SVM [4] | CFMSPT |
|----------------------|------|-----------------|--------|---------|--------|
| 1 | 3.20 | 8.51 | 3.07 | 3.67 | 0.33 |
| 2 | 4.24 | 0.57 | 1.25 | 1.94 | 2.70 |
| 3 | 2.69 | 8.96 | 3.12 | 3.04 | 0.05 |
| 4 | 0.12 | 1.83 | 0.22 | 0.35 | 0.02 |
| 5 | 1.03 | 10.90 | 3.783 | 2.88 | 1.51 |
| average error | 2.26 | 6.15 | 2.29 | 2.38 | 0.92 |

It can be seen from Table 3 that when forecasting the ceiling air leakage strength by using various forecasting models, the result of BP model in this paper is 2.26%, the result of the LS-SVM model is 2.29%, while the result of CFM is 0.92%, it is the lowest.

The prediction results show that BP neural network and SVM can all be used to forecast limit parameters of coal spontaneous combustion, and there is not much difference between their results. But for small samples, different number choice of hidden layer for BP neural network will lead to large difference. For example, the average relative error of improved BP [8] is 6.15%, it is very different with the BP neural network in this paper, because they have different training methods and number of hidden layer. But there is no theory evidence for selecting the number of hidden layer, they are mostly chosen based on experience, and the training time is too much. Thus, excessive fitting situation may appear easily. So, in view of the small sample data, support vector machine (SVM) is more suitable, such as the SVM in reference [4] and the LS-SVM in this paper, their results are basically identical. However, no matter BP model or

SVM model, given the limitations of single model itself, it is hard to work in all cases, so its generalization ability is weak. Gratifyingly, the combined forecasting model makes up for the inadequacy of single model, complements itself by the advantage of single model, and obtains the best prediction effect.

Any linear or nonlinear model can be used in the single forecasting model of CFM. So CFM has variability, it can always establish a combined prediction model, and CFM is most suitable for predicting the characteristics of time series data.

This method is also suitable for predicting the lower oxygen concentration and the minimum float coal thickness.

5 Conclusions

According to this study, it can be concluded as follows:

1) A combined forecasting model is proposed to predict limit parameters of coal spontaneous combustion.

References

- [1] He Q L, Wang D M 2004 Numerical simulation of spontaneous combustion process in goaf areas by fully-mechanized and caving roof coal *Journal of CUMT* **33**(1) 11-4 (in Chinese)
- [2] Xu J C, Wen H, Deng J, Zhang X H 2000 Study limit parameter of coal self-ignite *Fire Safety Science* **9**(2) 15-7 (in Chinese)
- [3] Luo D Y, Zhang G S 2010 Impact factors of survived coal spontaneous combustion in goaf *Coal Technology* **29**(11) 83-84 (in Chinese)
- [4] Meng Q, Wang H Q, Wang Y S, Zhou Y 2009 Predicting limit parameters of coal self-ignition based on support vector machine *Journal of China Coal Society* **34**(11) 1489-93 (in Chinese)
- [5] Zhang X H, Xi G, Chen X K, Deng J, Wen H 2005 Determining spontaneous combustion danger zones and predicting spontaneous combustion during mining near-neighbor coal seams *Journal of China Coal Society* **30**(6) 733-6 (in Chinese)
- [6] Wen H 2002 Dynamic numeric simulation of coal self-ignite in goaf in fully mechanized caving face *Journal of China Coal Society* **27**(1) 54-8 (in Chinese)
- [7] Zhang C, Ti Z Y, Li Z X 2012 Three-dimension heterogeneity dynamic numerical simulation of top coal spontaneous combustion in limit equilibrium zone *China Safety Science Journal* **22**(5) 37-42 (in Chinese)
- [8] Xu J C, Wang H 2002 The neural network prediction method for the limit parameters of coal ignition *Journal of China Coal Society* **27**(4) 366-70 (in Chinese)
- [9] Wang H 2011 Principal component neural network prediction model for coal self-ignition duration *Computer Engineering and Applications* **47**(26) 242-5 (in Chinese)
- [10] Hou Y B 2004 An RSNN-based prediction method for the coal mine spontaneous combustion *Information and Control* **33**(1) 93-6 (in Chinese)
- [11] Gao Y, Qin M G, Li M J 2010 Analysis on prediction of residual coal spontaneous combustion in goaf based on support vector machine *Coal Science and Technology* **38**(2) 50-4 (in Chinese)
- [12] Meng Q, Wang Y S, Zhou Y 2010 Prediction of spontaneous combustion in caving zone based on rough set and support vector machine *Journal of China Coal Society* **35**(12) 2100-4 (in Chinese)
- [13] Deng J, Xing Z, Ma L 2011 Application of multiple regression analysis in coal spontaneous combustion prediction *Journal of Xian University of Science and Technology* **31**(6) 645-8 (in Chinese)
- [14] Mizumoto M 1989 Pictorial Representation of Fuzzy Connectives, Part 1: Cases of T-norms, T-conorms and Averaging Operators *Fuzzy Sets and Systems* **31** 217-42
- [15] Zeng K S, Hu N L 2008 Model of system safety forecasting and combination forecasting *Journal of China Coal Society* **33**(10) 1123-5 (in Chinese)
- [16] Ao P, Mou L H 2011 Load combination forecasting based on power grid with environmental characteristics *Journal of China Coal Society* **36**(9) 1575-80 (in Chinese)

2) Compared with the single forecasting model, the error of the combined forecasting model is lower, and the prediction effect is better.

3) CFM model has variability and strong generalization ability. It can be trained to find the most suitable combined forecasting model for the characteristics of the dataset, thus ensuring the predicted results of CFM model at least as good as the ones of other single forecasting models.

Acknowledgments

The project is supported by the key project of national natural science foundation of China (Program No. 51134019), the natural science basic research plan in Shaanxi province of China (Program No.2012JQ8035) and the special scientific research of the department of education in Shaanxi province (Program No. 2013JK0870).

| Authors | |
|---|--|
|  | <p>Pengtao Jia, born in 1977, Xinzheng, Henan, China</p> <p>Current position, grades: Associate Professor</p> <p>University studies: M.E. degree in computer application technology at Xian University of Science and Technology, China in June 2002. Doctor's degree in computer science and technology at Northwestern Polytechnical University, China in June 2008.</p> <p>Scientific interest: Computer Science, data mining, application of artificial intelligence theory.</p> <p>Publications: more than 20.</p> |
|  | <p>Jun Deng, born in 1970, Dazhu, Sichuan, China</p> <p>Current position, grades: Professor at School of Energy Engineering, Xi'an University of Science and Technology.</p> <p>University studies: Xi'an University of Science and Technology (China).</p> <p>Scientific interest: mine safety, coal spontaneous combustion.</p> <p>Publications: more than 80.</p> |
|  | <p>Shuhui Liang, born in 1988, Shuozhou, Shanxi, China</p> <p>Current position, grades: Postgraduate at School of Computer Science and Technology, Xi'an University of Science and Technology.</p> <p>University studies: Xi'an University of Science and Technology (China)</p> <p>Scientific interest: computer science, ensemble learning.</p> <p>Publications: 3.</p> |

Exponential synchronization of complex networks with non-delayed and delayed coupling via hybrid control

Guoliang Cai, Shengqin Jiang*, Shuiming Cai, Lixin Tian

Nonlinear Scientific Research Center, Jiangsu University, Zhenjiang, Jiangsu, 212003, China

Received 1 March 2014, www.tsi.lv

Abstract

In this paper, the different structure synchronization of the two complex chaotic networks with time-varying delay and non-time-varying delay coupling is considered. Based on Lyapunov stability theory, combined with Yong inequality approach, Hybrid control including periodically intermittent control and adaptive control is designed such that the two complex chaotic networks achieve the exponential synchronization. Different numerical simulations are given to illustrate the effectiveness of the proposed method. Moreover through comparing the numerical simulations with the different functions of time delay, we can get how the time delay function impacts the complex chaotic networks synchronization in this model.

Keywords: complex chaotic networks, hybrid controller, time-varying delayed, Lyapunov stability theory

1 Introduction

In recent years, complex chaos networks have been a most important hot research area in the nonlinear science [1-4]. Based on the potential application and development foreground in physics, biology, communication, traffic, WWW and so on, the controller and synchronization have been attracted increasing attention [5, 6].

After many years of research, people have put forward a variety of effective chaotic synchronization control methods such as feedback control [7], adaptive control [8], impulse control [9] and intermittent control [10], etc. Synchronization has been applied to practical application, especially used in secure communication. In actually, the structure of drive system and response system is likely to be different. Therefore there is more practical significance for the research of synchronization of complex networks with non-identical structure.

As an important direction, many works have been done to consider the synchronization of complex networks. The synchronization of chaotic dynamic networks with unknown and mismatched parameters has been considered in [11]. In [12], Zheng et al, discussed adaptive projective synchronization in complex networks with time-varying coupling delay. In [13], Cai et al, studied the synchronization-based approach for parameters identification in delayed chaotic networks. Many novel researches were proposed in [14], which considered the synchronization of chaotic systems with time-varying delays via intermittent control. In [15], Du et al, studied function projective in complex dynamical networks with time delay via hybrid feedback control.

The above synchronization methods are based on the chaotic networks with identical structure. Sun et al,

proposed non-identical structure chaotic networks in [16]. Based on the research of [16], Cai et al, studied linear generalized synchronization between two complex networks in [17].

In light of above finding, we propose a hybrid control, consisting of an adaptive control and intermittent control, to achieve the exponential synchronization of complex networks with time delay and non-time delay. And we explore the new condition of time delay $\tau(t)$. Based on the Lyapunov stability theorem and Yong inequality, the synchronization of chaos networks has been achieved. The numerical simulations have showed the accuracy and the effectiveness of the method.

2 Description

In this paper, complex networks with time delay and non-time-delay consisting N linearly and diffusively coupled identical nodes are considered as the drive system, described as the following:

$$\begin{cases} \dot{x}_i(t) = f(x_i(t)) + \sum_{j=1}^N c_{ij} \Gamma_1 x_j(t) + \sum_{j=1}^N d_{ij} \Gamma_2 x_j(t - \tau(t)), \\ t > 0 \\ x(t) = \varphi(t), \quad -\tau \leq t \leq 0 \quad i = 1, 2, \dots, N \end{cases}, \quad (1)$$

where $x_i(t) = (x_{i1}(t), x_{i2}(t), \dots, x_{in}(t))^T \in R^n$ is the state vector of the i^{th} node, $f: R \times R^n \rightarrow R^n$ is a smooth vector-valued function, the time delay $\tau(t)$ is a constant or a bounded function, for simplicity, we

* Corresponding author's e-mail: jiangshengmeng@126.com

assume the inner connecting matrix Γ_1 and Γ_2 are diagonal matrix, where $\|\Gamma_1\| = \beta$, $\|\Gamma_2\| = \gamma$, $C = (c_{ij})_{n \times n}$, $D = (d_{ij})_{n \times n} \in R^{n \times n}$ are the weight matrices, if there is a connection from node i to node j ($i \neq j$), then $c_{ij} \neq 0$ and $d_{ij} \neq 0$, otherwise, $c_{ij} = 0$ and $d_{ij} = 0$.

In this following, we introduce a general response networks consisting of N nodes with non-time-varying and time-varying delay, regarding the Equation (1) as the drive system, described as follows:

$$\dot{y}_i(t) = B y_i(t) + g_i(t, y_i(t)) + \sum_{j=1}^N c_{ij} T_1 y_j(t) + \sum_{j=1}^N d_{ij} T_2 y_j(t - \tau_i(t)) + u_i \quad (2)$$

where $y_i(t) = (y_{i1}(t), y_{i2}(t), \dots, y_{im}(t))^T \in R^n$ is the state vector of the i^{th} node, B is an $n \times n$ constant matrix, $g_i(\cdot): R^n \rightarrow R^n$ is a nonlinear vector-valued function, which is distinct for differentiable cluster, representing the activity of an individual subsystem. u_i is a controller.

Remark 1. The coupling configuration matrix C and D is not restricted to be symmetric or irreducible.

Now we introduce some definitions, assumptions, lemmas and theorem that will be required in this paper. Definition 1: For the drive Equation (1) and response Equation (2), the following controller is called hybrid controller including adaptive control and intermittent control:

$$u_i(t) = u_{i1}(t) + u_{i2}(t), \quad (3)$$

where

$$u_{i1}(t) = f(x_i(t)) - A y_i(t) - B g(y_i(t)),$$

$$u_{i2}(t) = -h_i(t)(y_i(t) - x_i(t)).$$

$$h_i(t) = \begin{cases} k_i, & nT \leq t \leq (n + \theta)T \\ 0, & (n + \theta)T \leq t \leq (n + 1)T \end{cases}$$

$i=1, 2, \dots, N$, in which $k_i > 0$ is a constant.

Defining the synchronization error as $e(t) = y(t) - x(t)$, if the drive-response system satisfies: $\lim_{t \rightarrow \infty} \|e(t)\| = \lim_{t \rightarrow \infty} \|y(t) - x(t)\| = 0$, then the drive Equation (1) and response Equation (2) can achieve synchronization. We can derive the error dynamical networks:

$$\dot{e}_i(t) = A e_i(t) + B(g(y_i(t)) - g(x_i(t))) + \sum_{j=1}^N c_{ij} \Gamma_1 e_j(t) + \sum_{j=1}^N d_{ij} \Gamma_2 e_j(t - \tau(t)) - k e_i(t)$$

$$nT \leq t \leq (n + \theta)T,$$

$$\dot{e}_i(t) = A e_i(t) + B(g(y_i(t)) - g(x_i(t))) + \sum_{j=1}^N c_{ij} \Gamma_1 e_j(t) + \sum_{j=1}^N d_{ij} \Gamma_2 e_j(t - \tau(t))$$

$$(n + \theta)T \leq t \leq (n + 1)T.$$

Assumption 1. For any different $x_1, x_2 \in R^n$, suppose there exists a constant $L > 0$ such that $\|g(x_1) - g(x_2)\| \leq L \|x_1 - x_2\|$, $i=1, 2$. The norm $\|\cdot\|$ of a variable is defined as $\|x\| = (x^T x)^{1/2}$.

Lemma 1 [18]. For any $x(t) = (x_1(t), x_2(t), \dots, x_n(t))^T$, $y(t) = (y_1(t), y_2(t), \dots, y_n(t))^T \in R^n$, there exists a positive constants $\xi > 0$ so that the following inequality is established: $2x^T y \leq \frac{1}{\xi} x^T x + \xi y^T y$.

Lemma 2 [14]. Let $0 \leq \tau(t) \leq \tau$, $y(t)$ is a continuous and non-negative function. If $t \in [-\tau, \infty]$, and the following conditions are satisfied:

$$\begin{cases} \dot{y}(t) \leq -\gamma_1 y(t) + \gamma_2 y(t - \tau(t)) & nT \leq t < (n + \theta)T \\ x(t) = \varphi(t), & -\tau \leq t \leq 0 \\ \dot{y}(t) \leq -\gamma_3 y(t) + \gamma_2 y(t - \tau(t)) & (n + \theta)T \leq t < (n + 1)T \\ y_i(t) = \Phi(t), & -\tau \leq t \leq 0 \end{cases}$$

where $\gamma_1, \gamma_2, \gamma_3$ are constants, $n=1, 2, \dots, N$, if the conditions $\gamma_1 > \gamma_2 > 0$, $\delta = \gamma_1 + \gamma_3$ and $\eta = \lambda - \delta(1 - \theta) > 0$, so we can get $y(t) \leq \sup_{-\tau \leq s \leq 0} y(s) \exp(-\eta t)$, $t \geq 0$, in which $\lambda > 0$ is the only positive solution of function $\lambda - \gamma_1 + \gamma_2 \exp(\lambda \tau) = 0$.

3 Main results

In this section, ρ_{\min} is defined as the minimum eigenvalue of the matrix $(\Gamma_1 + \Gamma_1^T) / 2$. We assume $\rho_{\min} \neq 0$ and $\|\Gamma_1\| = \rho$. $\hat{C}^s = (\hat{C} + \hat{C}^T) / 2$, where \hat{C} is obtained through that $(\rho_{\min} / \rho) c_{ii}$ substitutes for the diagonal element c_{ii} of matrix C . Let $P = D \otimes \Gamma_2$, where \otimes stands for the Kronecker product. Now we consider how to select the appropriate $h_i(t)$ ($i=1, 2, \dots, N$), θ and T so that the drive Equation (1) and the response Equation (2) can achieve exponential synchronization.

Theorem 1. For drive Equation (1) and response Equation (2), if Assumption 1 is established, by the Definition 1, there exist a positive constants a_1, a_2, a_3

and k_i ($i=1,2,\dots,N$) such that the following conditions hold:

(i) $A + LB + \rho \hat{C}^s + (\lambda_{\max}(\frac{1}{2} PP^T) + a_1 - k)I_n \leq 0,$

(ii) $A + LB + \rho \hat{C}^s - (a_3 - a_1 - \lambda_{\max}(\frac{1}{2} PP^T))I_n \leq 0,$

(iii) $L_2 - a_1 < 0,$

(iv) $\bar{w} = \varepsilon + 2a_3(1 - \theta) > 0,$

where $\varepsilon > 0$ is the only positive solution of function $-2a_1 + \varepsilon + 2L_2 \exp\{\varepsilon\tau\} = 0$, then the trivial solution of error Equation (6) is globally asymptotically stable, which implies that drive Equation (1) and the response Equation (2) can achieve globally exponential synchronization.

Proof. Consider the following Lyapunov function:

$$V(t) = \frac{1}{2} \sum_{i=1}^N e_i^T(t) e_i(t).$$

According the Definition 1, Lemma 1 and the conditions (i)-(iii), the derivative of $V(t)$ about the trajectories of error system is given as the following: when $nT \leq t \leq (n + \theta)T$, $n=0,1,2,\dots$, we have

$$\begin{aligned} \dot{V}(t) &= \sum_{i=1}^N e_i^T(t) \dot{e}_i(t) = \sum_{i=1}^N e_i^T(t) (Ae_i(t) + B(g(y_i(t) - g(x_i(t)))) + \sum_{j=1}^N c_{ij} \Gamma_1 e_j(t) + \sum_{j=1}^N d_{ij} \Gamma_2 e_j(t - \tau(t)) - ke_i(t)) \leq \\ &\sum_{i=1}^N e_i^T(t) (A + LB - kI_n) e_i(t) + \sum_{i=1}^N \sum_{j=1}^N c_{ij} e_i^T(t) \Gamma_1 e_j(t) + \sum_{i=1}^N \sum_{j=1}^N d_{ij} e_i^T(t) \Gamma_2 e_j(t - \tau(t)) \leq \sum_{i=1}^N e_i^T(t) (A + LB - kI_n) e_i(t) + \\ &\sum_{i=1}^N c_{ii} \rho_{\min} e_i^T(t) e_i(t) + \sum_{i=1}^N \sum_{j=1, j \neq i}^N c_{ij} \rho \|e_i(t)\| \|e_j(t)\| + \\ &\frac{1}{2} e^T(t) PP^T e(t) + \frac{1}{2} e^T(t - \tau(t)) e(t - \tau(t)) \leq \\ &e^T(t) (A + LB + (\lambda_{\max}(\frac{1}{2} PP^T) - k) I_n) e(t) + \\ &\rho e^T(t) \lambda_{\max}(\hat{C} \otimes I_n) e(t) + \frac{1}{2} e^T(t - \tau(t)) e(t - \tau(t)) + \\ &(\lambda_{\max}(\frac{1}{2} PP^T) + a_1 - k) I_n e^T(t) e(t) - a_1 e^T(t) e(t) \leq \\ &-2a_1 V(t) + V(t - \tau(t)) \end{aligned}$$

when $(n + \theta)T \leq t < (n + 1)T$, $m = 0, 1, 2, \dots$

$$\begin{aligned} \dot{V}(t) &= \sum_{i=1}^N e_i^T(t) \dot{e}_i(t) = \sum_{i=1}^N e_i^T(t) (Ae_i(t) + B(g(y_i(t) - g(x_i(t)))) + \sum_{j=1}^N c_{ij} \Gamma_1 e_j(t) + \sum_{j=1}^N d_{ij} \Gamma_2 e_j(t - \tau(t))) \leq \\ &e^T(t) (A + LB + \frac{1}{2} e^T(t - \tau(t)) e(t - \tau(t)) + \\ &\lambda_{\max}(\frac{1}{2} PP^T) I_n) e(t) + \rho e^T(t) \lambda_{\max}(\hat{C} \otimes I_n) e(t) \leq \\ &e^T(t) (A + LB + \rho \lambda_{\max}(\hat{C}^s) + (a_3 - a_1) e^T(t) e(t) - \\ &(a_3 - a_1 - \lambda_{\max}(\frac{1}{2} PP^T)) I_n) e^T(t) e(t) + \frac{1}{2} e^T(t - \tau(t)) e(t - \tau(t)) \leq \\ &2(a_3 - a_1) V(t) + V(t - \tau(t)) \end{aligned}$$

From the above equation, we have:

$$\begin{cases} \dot{V}(t) \leq -2a_1 V(t) + V(t - \tau(t)), & nT \leq t \leq (n + \theta)T \\ \dot{V}(t) \leq 2(a_3 - a_1) V(t) + V(t - \tau(t)), & (n + \theta)T \leq t < (n + 1)T \end{cases}$$

From the lemma 2, one has:

$$V(t) \leq \sup_{-\tau \leq s \leq 0} V(s) \exp(-\bar{w}t), \quad t \geq 0.$$

According to the Lyapunov stability theorem and the definition of exponential synchronization, the drive Equation (1) and response Equation (2) can achieve exponential synchronization. The proof is completed.

Next, we will discuss how to select appropriate parameters, and use simple and easy data to achieve synchronization: from Theorem 1. We know:

$$m_0 = 2\lambda_{\max}(A + LB + \rho \lambda_{\max}(\hat{C}^s) + (\lambda_{\max}(\frac{1}{2} PP^T) I_n),$$

$$a_2 = \frac{1}{2}, \quad a_3 = m_0 + a_1 > 0.$$

Then the condition (ii) in Theorem 1 is established. Corollary 1. If there exists a positive $a_1 > a_2$ such that

(i) $0 < m_0 + a_1 \leq k,$

(ii) $\eta = \lambda - 2(m_0 + a_1)(1 - \theta) > 0,$

where $\varepsilon > 0$ is the only positive solution of function $-2a_1 + \varepsilon + 2a_2 \exp\{\varepsilon\tau\} = 0$, then the drive Equation (1) and response Equation (2) can achieve exponential synchronization.

Remark 2. There are many papers about complex networks based on intermittent control. The time delay and intermittent control period have strict restrict conditions, for example, $\theta = 0.5$ [20], $0 \leq \dot{\tau}(t) < 1$ [17], where the time delay $\tau(t)$ must be differentiable and bounded. In this paper, the time delay only needs bounded.

Remark 3. Adaptive synchronization is a continuous control method. Intermittent synchronization is a discrete control method. In this paper, we put the two

methods together. The model we consider is general chaos networks. And the use of intermittent controller is a pretty effective way to achieve system synchronization.

In Corollary 1, there is the only parameter a_1 . If the parameter a_1 is given as $\gamma^* (> a_2)$, we plug $a_1 = \gamma^*$ into the function $-2a_1 + \varepsilon + a_2 \exp\{\varepsilon\tau\} = 0$. So we can get $\varepsilon = \varphi(\gamma^*)$, and the following conclusion.

Corollary 2. If the parameter a_1 is given as $\gamma^* (> a_2)$, then the condition of the two systems achieve exponential synchronization is obtained in the following:

- (i) $0 < m_0 + a_1 \leq k$,
- (ii) $1 - \frac{\varphi(\gamma^*)}{m_0 + \gamma^*} < \theta < 1$.

From the corollary, we can obtain k , control rate θ . Based on the above discussion, appropriate controller can be get such that the drive system and response system can achieve synchronization.

4 Numerical simulations

In this section, we consider the Lü system:

$$\dot{x} = f(z) = \begin{bmatrix} a(z_2 - z_1) \\ cz_2 - z_1z_2 \\ -bz_3 + z_1z_2 \end{bmatrix}, \tag{5}$$

where $a = 36, b = 3, c = 20$.

Rössler system:

$$\dot{v} = Av + Bg(v) = \begin{bmatrix} -v_2 - v_3 \\ v_1 + wv_2 \\ \lambda + v_3(v_1 - \theta) \end{bmatrix}, \tag{6}$$

where $w = 0.2, \lambda = 0.2, \theta = 5.7$.

$$A = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 0.2 & 0 \\ 0 & 0 & -5.7 \end{pmatrix}, B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, g(v) = \begin{pmatrix} 0 \\ 0 \\ v_1v_2 + 0.2 \end{pmatrix}$$

in which $\alpha = \|A\| = 5.7897, \beta = \|B\| = 1$.

The five nodes are considered in drive system:

$$\dot{x}_i(t) = f(x_i(t)) + \sum_{j=1}^5 c_{ij} \Gamma_1 x_j(t) + \sum_{j=1}^5 d_{ij} \Gamma_2 x_j(t - \tau(t)),$$

Response system is described as:

$$\dot{y}_i(t) = Ay_i(t) + Bg(y_i(t)) + \sum_{j=1}^5 c_{ij} \Gamma_1 y_j(t) + \sum_{j=1}^5 d_{ij} \Gamma_2 y_j(t - \tau(t)) + u_i(t), \quad i=1, \dots, 5;$$

where the $f(x_i(t))$ as the Equation (5) shows, $g(y_i(t))$, A and B as the Equation (6) shows. $C = (c_{ij})_{n \times n}$ and $D = (d_{ij})_{n \times n} \in R^{n \times n}$ are the weight matrix.

$$C = \begin{bmatrix} -5 & 0 & 0 & 3 & 2 \\ 1 & -4 & 0 & 0 & 3 \\ 0 & 0 & -7 & 4 & 3 \\ 0 & 0 & 0 & -3 & 3 \\ 0 & 0 & 1 & 1 & -2 \end{bmatrix}, D = \begin{bmatrix} -4 & 3 & 0 & 1 & 0 \\ 2 & -3 & 1 & 0 & 0 \\ 0 & 0 & -7 & 4 & 3 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 2 & -2 \end{bmatrix}.$$

In numerical simulation, the initial values of drive-response system are chosen as $x_i(0) = (0.3 + 0.1i, 0.3 + 0.1i, 0.3 + 0.1i)^T$, $y_i(0) = (2.0 + 0.7i, 2.0 + 0.7i, 2.0 + 0.7i)^T$. Choose $L=1, \Gamma_1 = \Gamma_2 = \text{diag}(1,1,1)$, $T=1, \rho_{\min} = \rho = 1, \tau(t) = \frac{e^t}{1+e^t}$, $0 < \tau(t) < 1$. Based on the Corollary 2, we can get $\rho \lambda_{\max}(\hat{C}^s) = 1.35, \lambda_{\max}(\frac{1}{2}PP^T) = 1.3053$, so $m_0 = 36.6037$.

From the Figure 1, we can obtain $\gamma^* = 20$ and $\theta = 0.97$, which satisfy the relation of Corollary 2, such that $k \geq m_0 + a_1 = 56.6037$. Thus we choose $k = 57$. The simulation of Figure 2 shows the drive Equation (1) and response Equation (2) can achieve synchronization in a few second. A piecewise function given in the Figure 3, which means that the time delay is not differentiable, shows the correctness of the discussion in remark 2.

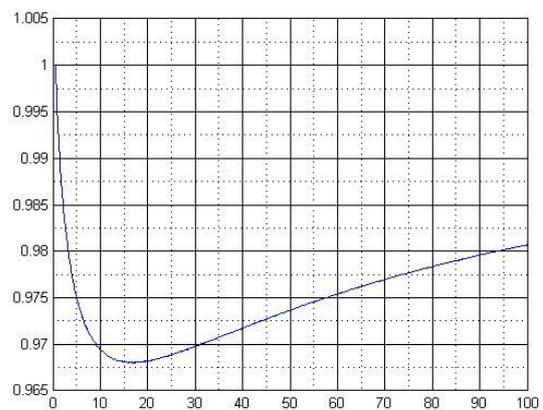
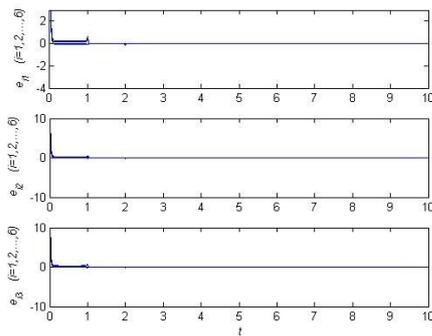
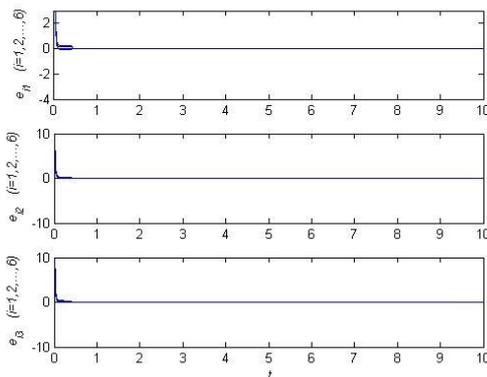


FIGURE 1 The relationship of parameter $\gamma^* - \theta$

FIGURE 2 The synchronization error when $\tau(t)$ is linear functionFIGURE 3 The synchronization error where $\tau(t) = \|t - 0.2\| - \|t - 0.6\|$

References

- [1] Ding W 2009 Synchronization of delayed fuzzy cellular neural networks with impulsive effects *Communications in Nonlinear Science Numerical Simulation* **14** 3945-52
- [2] Wu Z Y, Fu X C 2012 Cluster projective synchronization between community networks with nonidentical nodes *Physica A* **391** 6190-98
- [3] Ma Q, Lu J W 2013 Cluster synchronization for directed complex dynamical networks via pinning control *Neurocomputing* **101** 354-60
- [4] Arik S 2002 *IEEE Transactions on Circuits and Systems I-fundamental Theory and Application* **49** 1211-14
- [5] Yang Y Q, Yu X H, Zhang T P 2010 Smart variable structure control of complex network with time-varying inner-coupling matrix to its equilibrium *Control Theory Application* **27** 181-7
- [6] Lü J, Chen G A 2005 *IEEE Transactions on Automatic Control* **50** 841-6
- [7] Du H Y, Shi P, Lü N 2013 Function projective synchronization in complex dynamical networks with time delay via hybrid feedback control *Nonlinear Analysis: Real World Applications* **14** 1182-90
- [8] Zhang Z Q, Wang Y X, Du Z B 2012 Adaptive synchronization of single-degree-of-freedom oscillators with unknown parameters *Applied Mathematics and Computation* **218** 6833-40
- [9] Li X D, Rakkiyappan R 2013 Impulsive controller design for exponential synchronization of chaotic neural networks with mixed delays *Communications in Nonlinear Science Numerical Simulation* **18** 1515-23
- [10] Yu J, Hu C, Jiang H J, Teng Z D 2012 Synchronization of nonlinear systems with delays via periodically nonlinear intermittent control *Communications in Nonlinear Science Numerical Simulation* **17** 2978-89
- [11] Xiao Y Z, Xu W, Li X C 2007 Adaptive complete synchronization of chaotic dynamical network with unknown and mismatched parameters *Chaos* **17** 033118 1-8 (in Chinese)
- [12] Zheng S, Bi Q S, Cai G L 2009 Adaptive projective synchronization in complex networks with time-varying coupling delay *Physics Letters A* **373** 1553-59
- [13] Cai G L, Shao H J 2010 Synchronization-based approach for parameters identification in delayed chaotic network *Chinese Physics B* **19** 060507 1-7
- [14] Cai S M, Hao J J, He Q B, Liu Z R 2012 New results on synchronization of chaotic systems with time-varying delays via intermittent control *Nonlinear Dynamics* **67** 393-402
- [15] Du H Y, Shi P, Lu N 2013 Function projective in complex dynamical networks with time delay via hybrid feedback control *Nonlinear Analysis: Real world Applications* **14** 1182-90
- [16] Sun M, Zeng C Y, Tian L X 2010 Linear generalized synchronization between two complex networks *Communications in Nonlinear Science Numerical Simulation* **15** 2162-7
- [17] Cai G L, Yao Q, Fan X H, Ding J 2011 Linear generalized synchronization between two complex networks *In International Conference on Multimedia, Software Engineering and Computing (MSEC2011)* November 26-27 2011 Wuhan China 447-52
- [18] Yang M, Wang Y W, Wang H O, Tanaka K, Guan Z H 2008 Delay independent synchronization of complex network via hybrid control *In American Control Conference* 2008 Seattle WA 2266-71
- [19] Cai S M, Hao J B, He Q B, Liu Z R 2011 Exponential synchronization of complex delayed dynamical networks via pinning periodically intermittent control *Physics Letters A* **375** 1965-71
- [20] Li C D, Liao X F, Huang T W 2007 Exponential stabilization of chaotic systems with delay by periodically intermittent control *Chaos* **17** 013103 1-7

5 Conclusions

In this paper, based on the Lyapunov stability theorem, a hybrid controller is proposed, which concludes adaptive control and intermittent control, to achieve synchronization of time-varying and non-time-varying coupling complex networks. The time delay $\tau(t)$ only needs bound. The numerical simulations discuss two state of the time delay, which is continuous and discrete. And they also indicate the synchronization time is related to the maximum of the time delay $\tau(t)$. The examples have shown the accuracy and the effectiveness of the proposed method.

Acknowledgements

This work was supported by the National Nature Science foundation of China (Nos 51276081, 71073072), and the Students' Research Foundation of Jiangsu University (No 12A415). Especially, thanks for the support of Jiangsu University.

| Authors | |
|---|---|
|  | <p>Guoliang Cai, born on September 20, 1956, Henan, China</p> <p>Current position: Deputy director of system engineering at Nonlinear Scientific Research Center (Jiangsu university), Professor, Doctor, supervisor of postgraduate.</p> <p>University studies: Applied mathematics in Jiangsu University.</p> <p>Scientific interest: analysis and applications of nonlinear dynamical systems, applied mathematics.</p> <p>Publications: 166 articles, 8 textbooks.</p> |
|  | <p>Shengqin Jiang, born on December 19, 1989, Shangdong, China</p> <p>Current position: Graduate Student of Faculty of Science, Nonlinear Scientific Research Center Jiangsu University.</p> <p>University studies: Basic mathematics in Jiangsu University.</p> <p>Scientific interest: analysis and applications of nonlinear dynamical networks.</p> <p>Publications: 1 paper</p> |
|  | <p>Shuiming Cai, born on October 12, 1983, Fujian, China</p> <p>Current position: Lecturer, Doctor at Jiangsu University.</p> <p>University studies: Applied mathematics in Shanghai university.</p> <p>Scientific interest: analysis and applications of nonlinear dynamical systems, applied mathematics.</p> <p>Publications: 19 articles.</p> |
|  | <p>Lixin Tian, born on July 15, 1963, Jiangsu, China</p> <p>Current position: Vice-president of Jiangsu University, Professor, Doctoral Supervisor.</p> <p>University studies: applied mathematics in East China Normal University.</p> <p>Scientific interest: analysis and applications of nonlinear dynamical systems, applied mathematics.</p> <p>Publications: 150 publications, Winner of National teacher award.</p> |

Tensor modular sparsity preserving projections for dimensionality reduction

Mingming Qi^{1, 2*}, Yang Xiang¹

¹Department of Computer science and technology, Tongji University, Shanghai, 201804, China

²Yuanpei College of Shaoxing University, Shaoxing, Zhejiang 312000, China

Received 1 March 2014, www.tsi.lv

Abstract

In order to reduce the computational complexity and promote the classification performance of Modular Weighted Global Sparse Representation (MWGSR), Tensor Modular Sparsity Preserving Projections (TMSPP) for dimensionality reduction is proposed. The algorithm firstly partitions an image into several equal-sized modules and constructs these modules into a third-order tensor image; then, the algorithm makes module sparse reconstructions and some modules with less reconstruction errors are selected. These selected modules are recombined into a dataset with fewer dimensions and a new sparse reconstruction weight is gotten on the new dataset, which is denoted as the sparse reconstruction weight of original samples; finally, projection matrices are gotten with steps of tensor sparsity preserving projections on the reconstructed tensor images. The algorithm promotes the computational efficiency and the robust performance of sparse preserving projections on high-dimensional datasets. Experimental results on YaleB and AR face datasets demonstrate effectiveness of proposed algorithm.

Keywords: dimensionality reduction, modular sparsity preserving projections, sparse reconstruction, the third-order tensor

1 Introduction

The destination of dimensionality reduction is to preserve certain property as far as possible in the process of projecting data from high-dimensional data space into low-dimensional data space, reducing the complexity of disposing high-dimensional data in data. Therefore dimensionality reduction is an important step in applications of data mining. In recent years sparse representation has been widely used in classification and reduces dimensionality of machine learning [1-8] thanks to its strong representation performance. Wright et al. [12] employed the sparse representation based classification (SRC) for robust face recognition. Based on the sparse representation, Qiao et al. [3] proposed Sparsity Preserving Projections (SPP) dimensionality reduction algorithms. The main purpose of SPP is to preserve the relationship of the sparse reconstruction in high-dimensional data into low-dimensional data in the process of dimensionality reduction. However, when the number of training high-dimensional data is large, sparse reconstruction calculation of them is very large, and even harder to complete [1]. Therefore, how to improve sparse reconstruction computational efficiency of large high-dimensional data is an important issue. Some solution way is to achieve sparse reconstruction based on Principal component analysis (PCA), gabor feature and so on. However, these methods are easy to lose the realness of the original data. So Lai et al [9] proposed a Modular Weighted Global Sparse Representation

(MWGSR) method. The method firstly modularize face image and achieve sparse reconstruction of every module, and then recalculated sparse reconstruction combining modular sparse reconstruction weight with the linear weighted way. Experimental results show that the improved modular sparse learning of sparse representation improves the computational efficiency and robust performance.

In order to preserve space relationship of high-dimensional data in the process of dimensionality reduction, tensor dimensionality reduction algorithms have been introduced [10-12]. These algorithms regarded a two-dimensional face image as a second-order tensor image without transforming them into vectors. Recently, Lai et al. [13] proposed a novel modular discriminant analysis algorithm. The algorithm first modularizes uniformly face image and combined these image blocks into a three-order tensor image, and then applies Multilinear Discriminant Analysis (MDA) in built third-order tensor images.

Inspired by above analyses, a dimensionality reduction algorithm called Tensor Modular Sparsity Preserving Projections (TMSPP) is proposed in this paper. The algorithm first modularizes uniformly the two-dimensional image into several module and uses these modules to construct the corresponding three-order face tensor data; then calculate sparse reconstruction weight of each module and obtains the corresponding reconstructed image based on sparse representation; selects some module with little sparse reconstruction error and

* Corresponding author e-mail: webqmm1974@163.com

combine them into new training feature vectors; Finally, calculates sparse reconstruction weights of new training feature vectors and the projection matrix on third-order tensor data. Experimental results on real AR and YaleB face datasets show that the proposed algorithm not only improves the performance of dimensionality reduction but also promotes the efficiency of sparse learning.

The characteristics of the proposed algorithm are listed as follows:

1) Because an image is divided into a number of modular, the number of feature dimensions is greatly reduced, which improves the computing efficiency of sparse reconstruction for image modules. Therefore, the algorithm can adapt to the large-scale high-dimensional datasets.

2) Part of modules are selected to reconstruct into training feature vectors for sparse reconstruction on image with occlusion and disguise, which is more efficient in avoiding the external disturbance, so the algorithm has better robustness performance.

3) Apart from preserving sparsity reconstruction of samples, the algorithm not only preserve pixels relation in very module but also preserve modular spatial relation.

The paper is organized as follows: In Section 2 we will introduce SPP. A theoretical analysis of TMSPP is given in Section 3. The experimental results and analysis will be presented in Section 4 and conclusions are given in Section 5.

2 Sparsity preserving projections (SPP)

Sparsity reconstruction weight of Sparse representation reveals category relationships of signal. Given training sample $X = \{x_1, x_2, x_3, \dots, x_n\} \in \mathbb{R}^{d \times n}$, the destination of sparse representation is to represent $x_i \in X$ with as few other samples of x to as possible. For facilitate calculation, l_0 -norm is replaced by l_1 -norm in sparse learning as follows:

$$\begin{aligned} \min_{s_i} \|s_i\|_1 \\ \text{s.t. } x_i = Xs_i, \\ 1 = 1^T s_i \end{aligned} \quad (1)$$

where $\|s_i\|_1$ denotes the l_1 norm of s_i , $s_i = [s_{i1}, \dots, s_{ii-1}, 0, s_{ii+1}, \dots, s_{in}]^T \in \mathbb{R}^n$ denotes sparsity reconstruction weight of x_i and $1 \in \mathbb{R}^n$ denotes a vector full of 1's. s_{ij} denotes reconstruction coefficients of sample x_j reconstructing x_i , namely:

$$x_i = s_{i1}x_1 + \dots + s_{ii-1}x_{i-1} + s_{ii+1}x_{i+1} + \dots + s_{in}x_n \quad (2)$$

For each $x_i \in X$, the sparsity reconstruction matrix $S = [S_1, S_2, \dots, S_n]^T$ of the training samples can be obtained by calculating the corresponding s_i of $x_i \in X$.

Sparse preserving projection aims at preserving sparsity reconstruction relations of input data in the process of dimensionality reduction. Given the sparse preserving projection matrix T , $T^T X s_i$ denote projected points of sparsity reconstruction in high-dimensional data space, the objective function of SPP is described as follows [3]:

$$\max_T \frac{T^T X (S + S^T - S^T S) X^T T}{T^T X X^T T}. \quad (3)$$

3 Tensor modular sparsity preserving projections (TMSPP)

3.1 BASIC IDEA

In practical applications of image data mining, dimensions of image data are usually high and the size of them is large, which affect greatly the efficiency of sparse reconstruction and even fail to achieve sparse reconstruction. Furthermore, modular sparse learning is helpful for improvements on the performance because that deformity and disguise happen in part of images. When images are divided into some modules, there are two important spatial relations, namely, spatial relations of pixels in every module and partial relations of modules. Therefore, an efficient way is to construct third-tensor data with images modules and achieve third-tensor sparsity preserving projections. Figure 1 shows eight modules of a face image and a constructed third-order tensor image.



FIGURE 1 A face modules and a constructed third-order tensor image

3.2 OBJECTIVE FUNCTION

Given training samples $X = \{x_1, x_2, x_3, \dots, x_n\} \in \mathbb{R}^{d \times n}$, each image will be evenly divided into m module. Modules set $X = [X_1^T, X_2^T, X_3^T, \dots, X_m^T]$ with $X_k \in \mathbb{R}^{(d/m) \times n}$ ($1 \leq k \leq m$) are gotten.

1) Firstly, sparse learning is achieved separately in each module set X_k ($1 \leq k \leq m$) and sparsity reconstruction weight S_k of X_k is obtained. So sparsity reconstruction set of each module set, namely $Y_k = X_k S_k$ ($1 \leq k \leq m$) is obtained. Third-order images data is obtained by combining these module sets in Figure 1.

2) Then, sparsity reconstruction error $\varepsilon_k = \sum_{i=1}^n \|Y_k^i - X_k^i\|_2^2$, ($1 \leq k \leq m$) of each sample module X_k ($1 \leq k \leq m$) set is calculated and module sets in the first z minimal reconstruction error are chosen, which is denoted by

$\tilde{X} = [\tilde{X}_1^T, \tilde{X}_2^T, \tilde{X}_3^T, \dots, \tilde{X}_m^T]$ with $X_k \in R^{(z \times d/m) \times n}$, $(1 \leq k \leq m)$ are obtained. The next step is to obtain sparsity reconstruction weights \tilde{S} for \tilde{X} as follows:

$$\begin{aligned} \min_{s_i} \|s_i\|_1 \\ \text{s.t. } x_i = Xs_i \\ 1 = 1^T s_i \end{aligned} \quad (4)$$

$$\max_{U_1, U_2, U_3} \frac{Y(\tilde{S} + \tilde{S}^T - \tilde{S}^T \tilde{S})Y^T}{YY^T} = \max_{U_1, U_2, U_3} \frac{(X \times U_1 \times U_2 \times U_3)(\tilde{S} + \tilde{S}^T - \tilde{S}^T \tilde{S})(X \times U_1 \times U_2 \times U_3)^T}{(X \times U_1 \times U_2 \times U_3)(X \times U_1 \times U_2 \times U_3)^T} \quad (5)$$

3.3 ALGORITHM STEPS

Input: training sample $X = \{x_1, x_2, x_3, \dots, x_n\}$, errors ε .

Output: Projection matrix U_1, U_2 and U_3 .

Steps:

1) Initialize respectively the projection matrix U_1, U_2 and U_3 as one diagonal unit matrix.

2) Build sparse reconstruction matrix \tilde{S} of the modular remodeling image using Equation (4).

3) Loop $t = 1 \dots 7$

3.1) Loop $f = 1 \dots 3$

3.1.1) $Y_i^f = X_i \times \dots \times U_{f-1} \times U_{f+1} \times \dots \times U_3$

3.1.2) transform Equation (5) into the solution of the generalized matrix:

$$Y^f(\tilde{S} + \tilde{S}^T - \tilde{S}^T \tilde{S})Y^{fT} v_i - \lambda_i Y^f Y^{fT} v_i, 1 \leq i \leq l^f,$$

$U_f^t = [v_1, v_2, \dots, v_{l^f}]$ is obtained.

3.1.3) If $\|U_f^t - U_f^{t-1}\|_2 < \varepsilon (f = 1, 2, 3)$, jump out of loop t .

4) Get the projection matrix U_1, U_2 and U_3 .

3.4 COMPUTATIONAL COMPLEXITY ANALYSES

Sparsity reconstruction is the main process of sparse learning, analysis on the time complexity of sparse reconstruction of our proposed algorithm. Given training sample $X = \{x_1, x_2, x_3, \dots, x_n\} \in R^{d \times n}$, sparsity reconstruction of sparse learning is the problem of minimization solving based on l_1 norm, which is $O(2d^2n - 2d^3/3)$ [14]. The time complexity of TMSPP is divided into two parts:

1) TMSPP divides image X into m images modules $X = [X_1^T, X_2^T, X_3^T, \dots, X_m^T]$. The time complexity of sparsity reconstruction for each module $X_k \in R^{(d/m) \times n}, (1 \leq k \leq m)$ is $O\left(\frac{2d^2n}{m^2} - \frac{2d^3}{3m^3}\right)$, the

3) Finally, the projection matrix of third-order sparsity preserving projections on third-order images is calculated and $Y = X \times U_1 \times U_2 \times U_3$ is obtained. Combined with Equation (3), the objective optimization function is obtained:

time complexity of sparse reconstruction for all m modules is $O\left(\frac{2d^2n}{m} - \frac{2d^3}{3m^2}\right)$.

2) The time complexity of sparsity reconstruction on choosing first z modules with minimum errors for image sparse reconstruction is $O\left(\frac{2zd^2n}{m^2} - \frac{2zd^3}{3m^3}\right)$.

In short, the time complexity of TMSPP is $O\left(\frac{2(m+z)d^2n}{m^2} - \frac{2(m+z)d^3}{3m^3}\right)$.

4 Experiments and analyses

4.1 FACE DATASETS

1) AR consists of over 4000 face images of 126 individuals. For each individual, 26 pictures were taken in two sessions (separated by two weeks) and each session contains 13 images. These images include front view of faces with different expressions, illuminations and occlusions. A group of face images on AR are shown in Figure 2.

2) YaleB contains 2414 front-view face images of 38 individuals. For each individual, about 64 pictures were taken under various laboratory-controlled lighting conditions. A group of face images on YaleB are shown in Figure 3.



FIGURE 2 A group of face images on AR



FIGURE 3 A group of face images on YaleB

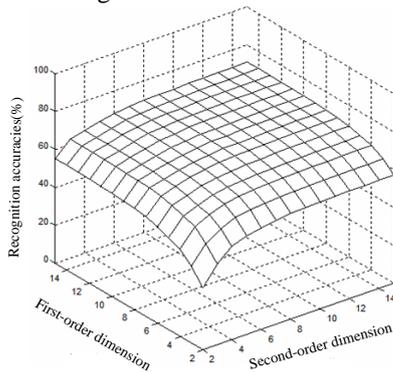
4.2 EXPERIMENTAL SETTINGS

In order to evaluate effectively the classification performance of the algorithm, MWGSR [9] and MDA [13] are selected to compare with our proposed algorithm. A certain number of images are chosen randomly as training samples from each group face images and the rest facial images as test samples. Furthermore, all the face image size is adjusted to 30×30 for computational convenience. The nearest neighbour classification algorithm is used. m is set to 8 and z is set to 3. All experiments are repeated 20 times and the average recognition accuracy is gotten as experimental results.

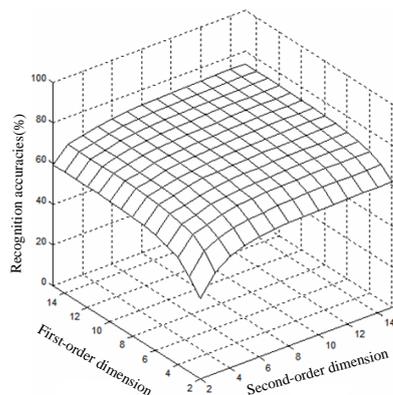
4.3 EXPERIMENTAL RESULTS AND ANALYSES

The number of modules is set to 4 and first two modules with minimum sparse reconstruction error are chosen for reconstruct feature vectors.

1) Firstly, the number of training samples in a group of samples is set to 4. With increment in feature dimension in first-order and second-order under the number of the same third-order dimension, recognition accuracies on AR are shown in Figure 4.



a) The number of the third-order dimension is 3



b) The number of the third-order dimension is 4

FIGURE 4 Recognition accuracies vs. dimensions of first-order and second-order on AR under different dimension of the third-order

From Figure 4 we can see that the recognition accuracies increase greatly with increment in dimensions of first-order and second-order and flat when the dimension exceeds certain less value. This illustrates that TMSPP can get the most classification performance in low dimension.

2) Secondly, most recognition accuracies are shown in Tables 1 2 for further verification in the performance of TMSPP under the different number of training samples.

TABLE 1 Experimental result on AR

| algorithm | The training number of a group samples | | |
|-----------|--|--------------|--------------|
| | 4 | 6 | 10 |
| MWGSR | 70.35 | 78.50 | 84.24 |
| MDA | 82.35 | 85.65 | 87.56 |
| TMSPP | 75.35 | 84.05 | 92.40 |

TABLE 3 Experimental result on YaleB

| algorithm | The training number of a group samples | | |
|-----------|--|--------------|--------------|
| | 10 | 15 | 20 |
| MWGSR | 80.50 | 85.50 | 90.50 |
| MDA | 86.53 | 90.65 | 92.45 |
| TMSPP | 81.55 | 92.40 | 95.65 |

Here bold data denote best and highest recognition accuracies under the same training sample.

The following conclusions can be drawn from Tables 1 and 2:

1) Although MWGSR inherited the feature of sparse learning, In contrast to MDA, the recognition accuracy of MDA is higher than MWGSR, which illustrates that third-tensor dimensionality reduction method has better classification performance.

2) Although MWGSR and TMSPP have taken advantage of some modules to guide sparse reconstruction. As a third-tensor dimension reduction algorithm, TMSPP not only preserve spatial relations of pixels in modules but also preserves spatial relations of modules, which is the reason that TMSPP has more obvious classification performance than MWGSR.

3) TMSPP and MDA share common characteristics of third-order tensor dimensionality reduction. When the size of samples is small, the classification performance of TMSPP is worse than that of MDA. When the size of samples exceeds the certain number, the classification performance of TMSPP is better than that of MDA. This shows that TMSPP inherits sparse learning robust performance, and is more suitable to face image with external disturb.

5 Conclusions

Despite the sparse learning has good performance of representation, but sparse reconstruction is not suitable for large-scale high-dimensional image datasets thanks to the computation complexity. Tensor Modular Sparse Preserving Projection (TMSPP) is proposed for dimensionality reduction. Apart from solving the problem of the computational efficiency, TMSPP improves the robustness performance through modularizing image. As

third-order tensor dimensionality reduction, TMSPP not only preserves pixels spatial relation in each module but also preserves module spatial relation of modules. Experimental results on AR and YaleB show that TMSPP demonstrates the efficiency of our proposed. The next work is to study how to select the optimal number of module on different face datasets. In addition, the related semi-supervised dimensionality reduction is the future research work.

References

- [1] Wright J, Yang A Y, Ganesh A, Sastry S S, Ma Y 2009 *IEEE Transaction Pattern Analysis And Machine Intelligence* **31**(2) 210-27
- [2] Yuan X T, Yan S C 2010 Visual classification with multi-task joint sparse representation *IEEE Conference Computer Vision and Pattern Recognition* Singapore June 2010 3493-3500
- [3] Qiao L S, Chen S C, Tan X Y 2010 Sparsity preserving projections with applications to face recognition *Pattern Recognition* **43**(1) 331-41
- [4] Cheng B, Yang J C, Yan S C, Fu Y, Huang T 2010 *IEEE Transactions on Image Processing* **19**(4) 85866
- [5] Gui J, Sun Z, Jia W, Hu R, Lei Y, Ji S 2012 Discriminant sparse neighborhood preserving embedding for face recognition *Pattern Recognition* **45**(2) 2884-93
- [6] Wang S, Cui L, Liu D, Huck R, Verma P, Sluss J J, Cheng S 2012 *IEEE Transactions on Intelligent Transportation Systems*, **13**(2) 955-62
- [7] Zhou T, Tao D 2012 Double Shrinking Sparse *IEEE Transactions on Image Processing* **22**(1) 244-56
- [8] Wang S J, Yang J, Sun M F, Peng X J, Sun M M, Zhou C G 2012 *IEEE Transactions on Neural Networks And Learning Systems* **(23)**6 876-88
- [9] Lai J, Jiang X D 2012 Modular Weighted Global Sparse Representation for Robust Face Recognition *Signal Processing Letters* **(19)**9 571-4
- [10] He X, Cai D, Niyogi P 2005 Tensor subspace analysis *Advances in Neural Information Processing Systems* Vancouver British Columbia, Canada, December 2005 345-56
- [11] Hua Han X-H, Chen Y-W, Ruan X 2012 *IEEE Transactions on image processing* **21**(3) 1314-26
- [12] Zhang Z, Chow W S 2012 Tensor Locally Linear Discriminative Analysis *IEEE Signal Processing Letters* **(18)**11 643-6
- [13] Jin Q Y, Huang Y Z, Wang C F 2013 Modular discriminant analysis and its applications *Artif Intell Rev* **(39)**4 285-303
- [14] Donoho D L, Tsaig Y 2008 *IEEE Transactions on Information Theory* **(54)**11 4789-812

Acknowledgments

The work is supported by NSF of China (71171148 and 61103069), National Technology Support Program of China (2012BAD35B01), Shanghai science and technology innovation plan (11dz1501703 and 11dz1210600).

Authors



Mingming Qi, born in May, 1974, Jiangxi, China

Current position, grades: Ph.D. student at the Department of Computer Science and Engineering, Tongji university, China Lecturer at Shaoxing University.

University studies: M.S. degree in computer application at Guizhou University in 2003.

Scientific interest: machine learning and image processing.

Publications: 13



Yang Xiang, born in August, 1964, Shangdong, China

Current position, grades: Professor at the Department of Computer Science and Engineering, Tongji university, China.

University studies: Ph.D. at Dalian University of Technology.

Scientific interest: machine learning, web semantic and image processing.

Publications: more than 100.

Coke oven production process hybrid intelligent control

Gongfa Li^{1, 2*}, Fuwei Cheng¹, Honghai Liu^{1, 2}, Guozhang Jiang¹, Jia Liu¹

¹College of Machinery and Automation, Wuhan University of Science and Technology, Wuhan 430081, China

²Intelligent Systems and Biomedical Robotics Group, School of Creative Technologies, University of Portsmouth, Portsmouth, PO1 2DJ, United Kingdom

Received 12 July 2014, www.tsi.lv

Abstract

Coke oven production possesses the characteristics of nonlinear, large inertia, large disturbances, and highly-coupling and so on. According to the characteristics of coke oven production process and control demand of coke oven production, intelligent control structure and models of coke oven production process was conducted. Firstly the intelligent control structure of coke oven production process was established. Then coal blending intelligent control, gas collector pressure intelligent control and combustion intelligent control of coke oven were discussed simply, while heating intelligent control, the production plan and schedule were discussed in detail. The control principle of combining the intermittent heating control with the heating gas flow adjustment was adopted, and fuzzy hybrid control was proposed to establish heating intelligent control strategy and model of coke oven, which combined feedback control, feed forward control and fuzzy intelligent control. The production plan and schedule of coke oven were optimized by utilizing the dynamic program and genetic algorithm. The practical running indicates that the system can effectively improve quality of coke and decrease energy consumption.

Keywords: Coke oven production process, Hybrid intelligent control, Heating intelligent control, Production plan and schedule

1 Introduction

Coke oven consumes energy substantially in the iron and steel enterprise, how to economize energy consumption, and improve quality and output of cokes are key questions of controlling and management of coke oven. It is significant to strengthen the competitiveness and increase economic efficiency of iron and steel enterprises to guarantee coke quality and improve output of coke steadily. The coke oven is disposed by a lot of coke-chambers and flue-chambers alternatively. The coke oven is disposed by a lot of coke-chambers and flue-chambers alternatively. The coal gas and air enter the coke-chamber to spread and burn after preheating from coke-chamber, and then heat produced is spread to coke-chamber by the stove wall. The coal material carries on the high-temperature dry distillation in the coke-chamber, and then coke is formed; waste gas produced by burning is discharged after holding retrieving the remaining heat energy via regenerator. Flow direction of coal gas, air and waste gas is exchanged per twenty minutes. According to production technology of coke oven, production process of coke oven has the following characteristics: 1) Production process of coke oven belongs to intermittent type, which is operated by single stove according to the operation plan; 2) Coke oven has characteristics of great inertia and large time-delay; 3) Mechanism of coking course is complicated, which has complexity of nonlinear and coupling parameters; 4) The variable changes violently, which results in strong interfering in the

production process; 5) The coke oven is a big hot-close system, where temperature measurement has particularity and complexity [1]. In order to effectively control and manage coke oven production process, intelligent control structure of coke oven production process has been set up at first, and then its intelligent control models are established. The production plan and schedule have described especially. The application of the system in some plants indicates that the system can stabilize production of coke oven, reach the anticipated result and is of great practical value.

2 Intelligent control structure of coke oven production process

According to the equipment state and market information, production task is made, production schedule is arranged rationally, four carts are managed in order and intelligent control of coke oven heating process is carried out in accordance with the production schedule. Therefore, coke quality is guaranteed, cost is lowered and energy consumption is economized. So production process from coal preparation to cooling cokes are controlled and managed effectively. Information and knowledge offered through the information processing of the production process and support system, production statistics, production schedules, supply-balance, production cost, equipment, quality and security are managed in real time. Meanwhile, according to the information gathered from the field, the heating and burning control of coke oven are

* *Corresponding author* e-mail: ligongfa@wust.edu.cn

optimized and controlled in real time. The production schedule of pushing coke and coal injection are adjusted in real time, the repair schedule is arranged rationally, coke oven production and stability of the whole stove temperature are guaranteed effectively. In accordance with the technological process and characteristic of production of the coke oven, the intelligent control structure of coke oven production process is set up, which is shown as Fig. 1. The intelligent control model of coke oven production process, basic automation system and actual application research are combined by using fuzzy control, mathematic analysis, linear programming, neural networks and genetic algorithm. Some models suitable for intelligent control system of coke oven production process are established. The factor relationship fit for intelligent control of coke oven production process is constructed through correlative parameters of temperature, flux, pressure, and hear-value of coke oven extracted by experiment research to validate experiment research, application and theory model. The estimation method of model is found and intelligent control of coke oven production process is realized. The DCS control system, interlocking and orientation system and PLC system in Fig. 1 are basic composition in the intelligent control of coke oven production process. Content above are not described here, because these are only the infrastructure of systematic research.

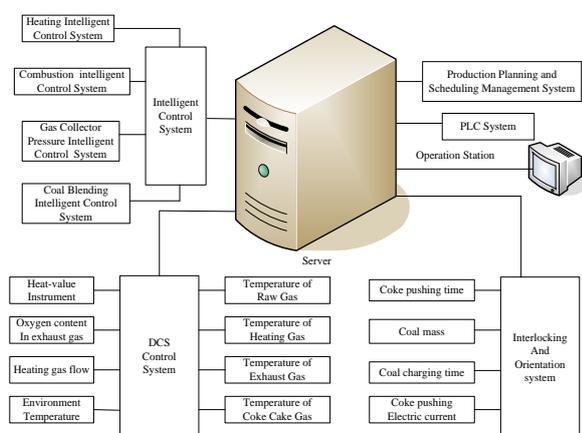


FIGURE.1 Intelligent control structure of coke oven production process

3 Intelligent control models of coke oven production process

From coke oven production technological process, intelligent control of coke oven production process involves coal blending intelligent control, gas collector pressure intelligent control, combustion intelligent control, heating intelligent control, the production plan and schedule of coke oven. Because coal blending intelligent control, gas collector pressure intelligent control, combustion intelligent control and heating intelligent control have already been conducted on the systematic research [2-5], a simple introduction is only given in this paper. The production plan and schedule of coke oven are discussed in detail.

3.1 COAL BLENDING INTELLIGENT CONTROL

In iron and steel industrial production, the blast furnace requires coke with the characteristics of low ash content, little sulphur, great intensity and high anisotropism degree. With the development of blast furnace maximization and injection technology with high pressure, requisition for coke quality is stricter and stricter. In coal blending and coking process, coal blending ratio is main factor influencing coke quality. Adopting suitable coal blending ratio to coking, coke quality can be guaranteed and coal resources can utilized rationally. But coal blending and coking process is a complicated industrial production process with a series of physics and chemical change and numerous of qualitative factors influencing coke quality. It is very difficult to describe the relationship between a kind of coal quality, coal blending ratio and coal blending quality, coke quality. The coal blending and coking process, in which much uncertainty exists, is too complicated to describe with mathematical models, and it cannot be controlled properly by traditional methods. According to the coking theory and using statistical data acquired from industry processes, a mathematical model is designed and a rule model is proposed with qualitative knowledge derived from experts' experience by utilizing statistical data acquired from industry processes. The rule model is combined with the mathematical model and qualitative and quantitative methods are integrated to establish a model for prediction of coke quality and to compute the blending ratio of coal. Furthermore, the flow of each kind of coal is controlled. The system has been running in a plant with accuracy of coal blending and prediction precision of coke quality reaches up to 97% and 95% respectively [2].

3.2 GAS COLLECTOR PRESSURE INTELLIGENT CONTROL

There is much produced raw gas which is collected from gas collectors and sent to the following work section by air delivery pipe through fan in coking process. Because the amount of raw gas changes along with the coking time, the pressure of gas collectors varies continuously, and the pressure of gas collector will have a vast fluctuation, especially in pushing coke and coal charging. When the pressure of the chamber in operation is negative, the air will enter into the chamber from the door and cover of chamber, which makes the coke burn, ash content inverse and coke quality decrease. The ingoing air also can generate chemical reactions with the constructional materials of the chamber, which leads to the denudation of the oven and shortens coke oven service life. The air also can prick up the burning of the raw gas and improve temperature of the coal gas system, which could prick up the burden of cooling system and bring needless energy consuming. When the pressure in the chamber is higher, the raw gas will burst out from the

door and the cover of the chamber, which not only leads to the discharge of soot and fire, and environment pollution, but also decreases raw gas recovery and energy waste. So the stabilization of the pressure in gas collector affects the quality of coke and coke oven service life. Therefore, it's essential to control it to a setting pressure arrange. The combination of classical control and intelligent control are utilized to control the pressure of gas collector. Intelligent pressure control of gas collector will control the press within an appropriate arrange and decrease the emanation [3].

3.3 INTELLIGENT CONTROL OF COMBUSTION

It processed production data in real-time, intelligent control the amount of air-flue ration, and controls coal gas totally burning as well as enables abolishing gas within the rational range. During actual combustion process, combustible components in flue and oxygen of air cannot carry out ideal mixing, and the air amount of supply must be more than theory needed in order to guarantee totally burning of fuel. The ration of real air amount to the theory air amount is called the air consumption coefficient, air surplus coefficient too.

$$\alpha = \frac{L_1}{L_2}, \quad (1)$$

where L1 represents real air amount, L2 represents theory air amount.

Value of α must be more than 1, and the choice of α is very important for coke oven production. If value of α is too small, the coal gas is burnt incompletely, and the flammable composition is given off with the waste gas; If value of α is too great, amount of abolished gas produced is large, and the heat of taking away increases. So that value of α is too great or small can increase amount of gas consumption. Therefore, the suitable control value of α must seek through practice. Under the general normal situation, while burning the coke oven coal gas, $\alpha = 1.2-1.3$; while burning the blast furnace gas, $\alpha = 1.15-1.2$. Aiming at the main problems existing in some coking production, with the analysis of coking process, a new control strategy is proposed, which integrates the technique of zoom chaos optimization with the method of expert control system for coke oven combustion. The application results demonstrate its feasibility and effectiveness [4].

3.4 HEATING INTELLIGENT CONTROL

Coke oven temperature mainly consists of raw gas temperature, flue temperature, cross wall temperature and so on. In order to realize automatic control of heating course in coke oven, the measurement value of various control parameters of coke oven should be got firstly, and the most key one is measurement, assessment and

prediction of different temperature of coke oven among them.

The key of furnace temperature feedback control is to establish goal flue temperature rationally and accurately, there are a lot of factors influencing goal flue temperature, in order to investigate the influence of various factors and find out quantitative relationship between them, it is necessary to carry on research on calculation model of goal flue temperature. But when establishing calculation model of goal flue temperature in fact, generally only several main factors are considered, such as the influence of coal mass, moisture of coal material, carbonization time and operating condition and so on. Analysis model of goal flue temperature is in equation (2).

$$F(j) = f(x, y, z, u, v, w, g, k), \quad (2)$$

where F is goal flue temperature; x is goal carbonization time; y is goal time; z is passing carbonization time after charging coal; u is real coal mass; v is moisture of coal; w is gas flow; g is prediction temperature of coke button in coke-chamber; j is a serial number of coke-chamber; k is revised coefficient.

Assessment and prediction of temperature in coke oven not only consider goal temperature calculation model, but also the interrelation models of various kinds of temperature of coke-chamber. For example, when analysing the interrelation model between flue temperature and temperature on top of regenerator, temperatures on top of regenerator at machine side and coke side are measured through the electric thermocouple, the average temperature on top of regenerator is changed into longitudinal temperature at machine side and coke side through the interrelation model between flue temperature and temperature and temperature on top of regenerator.

This kind of interrelation model is set up generally by adopting linear regression method, but because there is a greater error sometimes in this method, and actual physics system is non-linear system, so neural network is used to build model. When modelling, neural network of three layers is used, neural network structure is $1 \times 6 \times 1$. Input layer is one node among them, average temperature on top of regenerator is inputting value; hidden layer is 6 nodes, nodal function is linear function is linear function, flue temperature is output value. Right value and valve matrix got after neural network learning are used to construct interrelation model between flue temperature and temperature on top of regenerator.

The most important control in coke oven production is temperature control, because the coke oven temperature is the key factor of influencing coke oven quality, saving heating gas, decreasing pollution. In the same coking circle, if the coke oven temperature is too low, the coke will not mature enough, and the coke cake cannot constrict to well-balanced station, the rigidity of coke is low, and the density is high, and the pushing

electricity is high. Whereas, if the temperature is too high, the coke will mature too enough, and the coke cake constrict more than well-balanced station, the rigidity of coke is high, and the density is low, and the pushing electricity is low, which bringing too much soot in pushing course.

At present, iron and steel enterprises usually utilize the intermittent heating control method in heating control system of coke oven, which can well optimize heating control of coke oven in a situation that the heating energy of coke oven is steady and rich [5]. But when pressure in main pipe of blast furnace fluctuates violently and heating coal gas flow is insufficient, the method can't instruct attendants how to operate heat controlling. The function of control system can only be analysed and judged artificially by the attendants, and during stopping heating time, blast furnace gas and coke oven gas are stopped to be utilized at the same time and blast furnace gas is not fully utilized either. A new control principle combining the "intermittent heating control" with the heating gas flow adjustment is adopted in the control system [6]. It analyses and processes data synthetically such as temperature, flow and calorific value of gas, pushing coke, charging coal, coal mass, water content and planned carbonization time, "stopping heating time" of PLC system and the heating blast furnace gas /coke oven gas flow of DCS system are calculated and established through the model. Therefore heating of coke oven is even and stability, the whole heating level of coke oven is intelligent control, heating intelligent control of coke oven is realized. Its intelligent control model is illustrated in Fig. 2.

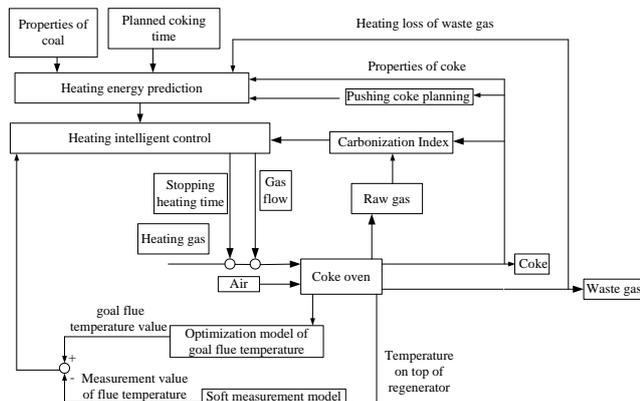


FIGURE 2 Intelligent control model of coke oven heating

Goal flue temperature of coke oven is the goal value of average temperature at machine side and coke side, is a main craft index to guarantee coke button ripe within carbonization time. There are a lot of main factors influencing goal flue temperature such as carbonization time, temperature in the centre of coke button, piling density of coal, water content, width of coke-chamber, thickness of stove wall. Because there are many variables and restrains, suitable optimization method is necessary

to adopt. Through analysing some main factors influencing goal flue temperature and course of conducting heating of coke oven, conducting heating model is set up, thus the relationship between carbonization time and width of coke-chamber, thickness of stove wall, thermal conductance rate of coal material and stove wall, thermal diffusion rate of coal material, flue temperature, temperature in the centre of coke button. Finally the optimal model of goal flue temperature is built.

Goal flue temperature value is got by optimization model of goal flue temperature, and measurement value of flue temperature is got by flue temperature soft measurement model according to temperature on top of regenerator. Deviation between goal value and measurement value, heating supplied amount of coke oven is revised real-time, thus coal gas flow and stopping heating time are adjusted.

When operation condition of production changes, the change range of temperature will often exceed (-6, +6), if the simple control method is still adopted, because of the great inertia of coke oven, big exceeding adjusting amount and two long adjustment time are caused. Aiming at above-mentioned situations, a prediction part in the controlling course has been increased (shown as Fig.3). Furthermore, the deviation of temperature is judged firstly when the range of deviation does not exceed (-6, +6), and fuzzy control is adopted. If it exceeds above-mentioned ranges, Bang-Bang intelligent control is used. So control precision and fast response of controlled target are guaranteed.

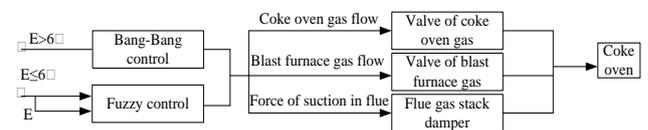


FIGURE 3 Intelligent control

A compound control system is proposed to control heating of coke oven, which combines feedback control, feedforward control and fuzzy intelligent control. Real-time data of production in coke oven are gathered by the system, such as pressure, flow, calorific value and temperature of coal gas, water content, and composition of heating gas and dynamic plans and so on. Settlement value of controlled parameters is calculated through energy prediction model, namely the feedforward. Then the value is transferred to the basic automated system to regulate. According to real-time information such as waste gas temperature, coke button temperature, flue temperature and oxygen content offered by the basic automated system at the same time, the energy balance is feedback regulated according to the fuzzy intelligent control model constantly in the course of heating. Then the settlement value is calculated again in order to enable CI to be kept within the range of control, which satisfies not only necessary temperature needed in coking, but also optimal heating control.

3.5 PRODUCTION PLAN AND SCHEFULE

Production schedule of coke oven is affected by such information as production task, technological process characteristic of coke oven, equipment state and resource situation. It's an optimal control question with multi-object and many-restrains. When the unusual situation is coming, coke oven production is influenced and the production schedule of the coke oven can be adjusted dynamically. So making a plan by men or traditional methods is more difficult to solve problems above. Automatic arrangement and control of coke pushing a plan are realized by using automatic control and computer technologies combining with linear programming and genetic algorithm in the system. Its structure is shown as Fig. 4.

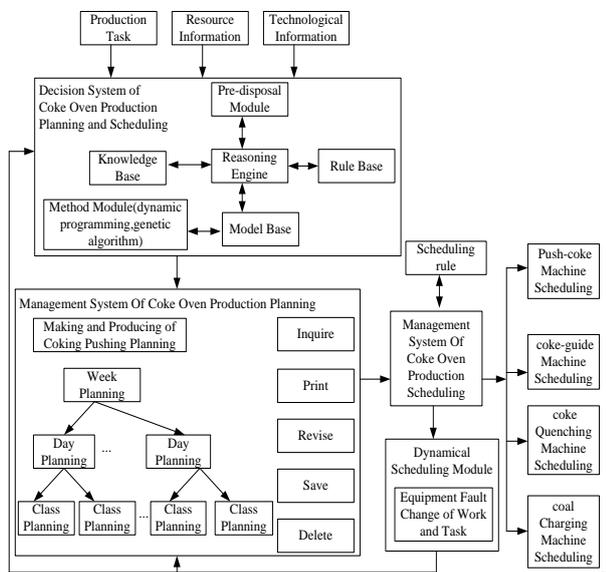


FIGURE 4 Structure of production planning and scheduling in coke oven

Coking is a process of carbonizing coal to coke in high temperature in the coke-chamber. When coke is matured, coke cake will have some constriction, and the centre temperature will reach to 950~1050°C. If the centre temperature cannot reach the temperature, the coke cake will half-cooked and does not constrict well, this will affect coke oven life consequently. If the centre temperature higher than the temperature, the coke cake is easy to broke, this will affect the quality of coke. There are 3 stages in the whole operation as following, running, pushing and setting off, which means coke pusher running to the appointed chamber to push according to the presupposed programming, and setting off the coal when larry car finished the work about the last chamber. Therefore, pushing coke should be operated according to definite sequence, and the designing model of this sequence is pushing planning model.

In large coking plants of the iron and steel companies in china, most operators estimate that the coke pushers are in the right place and alignment timely or not just by vision and experience. Sometimes the coke pushers lag or

the chamber are lead appointed, hence the master controller must operate continuously, the reducer must be restarted and stopped continuously, the hydraulic brake must be restarted continuously to reduce the security and shorten power equipment life. Coke pushing planning is still arranged by hand, and the work is trivial and stiff. Therefore, it is very essential to arrange pushing planning automatically and give coke pushing planning timely.

Automatic arrangement and control of coke pushing planning are realized by using automatic control and computer technologies combines with linear programming and genetic algorithm in the system.

Thinking about time relationship and real command, the algorithm of average subsection and proportional correction focusing on coke pushing job is adopted to arrange coke pushing planning, which means making average subsection according to commanded number of overhaul in every little circle, then getting coke pushing time of each chamber in turn and inserting overhaul time based on operation time and start pushing time of each chamber. If the pushing chamber number is less or more than total chamber number of every little circle in the arrangement, the planning will be corrected according to the scale of each section. Thinking about the flexibility and adaptability, the circle time, total chamber number, overhaul number of each little circle, starting coke pushing time, starting coke pushing chamber number and operation time of each chamber should be filled before automatic arrangement.

Coke pushing sequence can be arranged according to the algorithm of average subsection and proportional correction, then the mature of coke in each chamber should be analysed in term of the practical production of coke battery, the following pushing coke-chamber number and pushing time should be computed and transferred to the operator in coke pusher. Therefore, linear programming should be introduced and dynamic programming is used to build planning model. In the planning model, the target function is target temperature, like flue temperature or coke cake temperature and so on. The constrained conditions are the restriction of operation time and circle time, the coking time of consecutive chambers has a half discrepancy, coal charged lastly must be uniform distributed in the chamber and the distance of outputting chamber number and holding outputting chamber number must be given. Genetic algorithm can be adopted to find the solution. At first, the appropriate sample should be chosen, the target function should be mapped to adaptable function, the constrained condition should be charged properly and genetic tool box of MATLAB can be used to programming to find the solution.

Combining computer with chamber number orientation system, the real coke pushing time is recorded. Then compared with coke pushing planning, coefficient of coke pushing planning, coefficient of coke pushing execution and total coefficient of coke pushing planning are computed to estimate coke pushing

operation and realize effective operation and monitoring. In this system, the arrangement of coke pushing planning mainly use VB language, the database adopts Oracle or Access and so on. There are some additional management, such as real time pushing electricity current, history current, and coke pushing plan arrangement and job report forms and so on.

According to such real-time information as production task, resource information, craft information and so on, the production plan is made by decision system of the production plan and schedule. Production plan management system is composed of the pushing coke plan, coal injection plan, plan requirement, plan type, plan revision, plan conversation and plan delete. All plans are divided into several grades of class, day and week. The schedule management system carries out the management of four carts of coke oven in order according to schedule rules. This system coordinates production capacity online, and has inspection ability to each production process of the production line. When equipment is unusual, either operating mode or production task will vary. The system starts and adjusts the module dynamically to manage in real time, and give feedback information to the decision system to resume producing steadily and fast, and then an automatic dynamic schedule decision scheme is given to administrative staff on the spot. This system has guaranteed implement of pushing coke and coal injection on the schedule, steady production operation, and carbonization time, which improves coke quality and lengthen the furnace service life.

4 Application

The system puts into operation for more than four years in some coke oven of iron and steel company. The system

References

- [1] Gongfa Li, Jianyi Kong, Guozhang Jiang 2008 Research and Application on Compound Intelligent Control System for Coke Oven Heating *Chinese Journal of Iron and Steel* **43** 89-92
- [2] Chunhua Yang, Deyao Shen, Min Wu 2000 Synthesis of Qualitative and Quantitative Methods in a Coal Blending Expert System for Coke Oven *Chinese Journal of Acta Automatica Sinica* **26** 226-32
- [3] Guoxiong Zhou, Min Wu, Weihua Cao 2008 Variable Structure Fuzzy Control Based on Particle Swarm Optimization For Gas Collector Pressure *Chinese Journal of Information and Control* **37** 327-33
- [4] Min Jin, Deyao Shen 1999 Zoom Chaos Optimization Expert Control System for Coke Oven Combustion *Chinese Journal of Nonferrous Metals* **9** 631-5
- [5] Jukka Swanijung, Petri Palmu 1996 Development of Coke-oven Battery Process Management System at Rautaruukki Steelworks *Iron and Steel Engineer* 46-9
- [6] Guozhang Jiang, Jianyi Kong, Gongfa Li 2006 Intelligent Control System for Coke Oven Heating *Chinese Journal of Iron and Steel* **41** 73-6

runs normally and the effect is good. The system can not only regulate the size of heat-supply in time but also have stronger anti-interference ability. The system has reduced the fluctuation of the whole temperature of the coke oven, and fluctuation with furnace temperature to a great extent. It diminishes the heat of supporting that the coke oven need is also relative, so consumption of the blast furnace gas has decreased, and energy is saved by 2%-3% under the normal situation. The fluctuation of longitudinal temperature is kept straightly on 3-5 degrees centigrade. At the same time, the control system improves stable coefficient, lengthens the service life of the coke oven, and boosts coke quality.

5 Conclusions

The intelligent control of coke oven production process is carried out, the fluctuation of furnace temperature has been solved effectively, and the better control result has been achieved. The production plan and schedule of coke oven are optimized, which has realized the intelligent control of coke oven production operation effectively through utilizing the dynamic plan and genetic algorithm correctly at the same time. After the system is put into operation, it runs steadily and can stabilize stove temperature and coke oven production. The energy-conserving result is obvious, improves coke quality effectively, and has reached the anticipated effect. The system is of great practical value.

6 Acknowledgments

This research was supported in part by Hubei Provincial Department of Education (Q20141107).

Authors



Gongfa Li, born on October 7, 1979, Hubei China

Current position, grades: Associate professor. College of Machinery and Automation, Wuhan University of Science and Technology

Scientific interest: Modelling and optimal control of complex industrial process.

Publications: 100

Experience: Dr. Gongfa Li received the Ph.D. degree in mechanical design and theory from Wuhan University of Science and Technology in China. Currently, he is an associate professor at Wuhan University of Science and Technology, China. His major research interests include modelling and optimal control of complex industrial process. He has published nearly twenty papers in related journals.



Fuwei Cheng, born on June 17, 1988, born in China

Current position, grades: Currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology

Scientific interest: mechanical CAD/CAE, signal analysis and processing.

Experience: Fuwei Cheng was born in Hubei province, P.R. China, in 1988. He received B.S. degree in mechanical engineering and automation from Donghu college of Wuhan University, Wuhan, China, in 2012. He is currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology. His current research interests include mechanical CAD/CAE, signal analysis and processing.

| | |
|---|--|
|  | <p>Honghai Liu, born in 1973, China</p> <p>Current position, grades: Professor in Intelligent Systems, Head of Intelligent Systems and Biomedical Robotics, University of Portsmouth.</p> <p>Scientific interest: approximate computation, pattern recognition, multi-sensor based information fusion and analytics, human machine systems, advanced control, intelligent robotics and their practical applications.</p> <p>Publications: 300</p> <p>Experience: He previously holds research appointments at King's College London, University of Aberdeen, and project leader appointments in large-scale industrial control and system integration industry. He received a PhD in Intelligent Robotics in 2003 from Kings College, University of London, UK.</p> |
|  | <p>Guozhang Jiang, born on December 15, 1965, Tianmen, China</p> <p>Current position, grades: Professor of Industrial Engineering, and the Assistant Dean of the college of machinery and automation, Wuhan University of Science and Technology.</p> <p>University studies: He received the B.S. degree in Chang'an University, China, in 1986, and M.S. degree in Wuhan University of Technology, China, in 1992. He received the Ph.D. degree in mechanical design and theory from Wuhan University of Science and Technology, China, in 2007.</p> <p>Scientific interest: computer aided engineering, mechanical CAD/CAE and industrial engineering and management system.</p> <p>Publications: 120</p> <p>Experience: He is a Professor of Industrial Engineering, and the Assistant Dean of the college of machinery and automation, Wuhan University of Science and Technology.</p> |
|  | <p>Jia Liu, born in 1990, Hubei China</p> <p>Current position, grades: Currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology</p> <p>Scientific interest: mechanical CAD/CAE, signal analysis and processing.</p> <p>Experience: He received B.S. degree in mechanical engineering and automation from Wuchang institute of Technology, Wuhan, China, in 2012. He is currently occupied in his M.S. degree in mechanical design and theory at Wuhan University of Science and Technology.</p> |

Research on output regulation for saturated systems

Huang Wei*

School of Automation, Hangzhou Dianzi University, Hangzhou, China

Received 5 June 2014, www.tsi.lv

Abstract

In this paper, the output regulation problem is investigated, which consists of building a controller to asymptotically steer the output of a saturated linear systems to a given reference signal despite external disturbances. Particularly, for saturated systems subject to periodically time-dependent exosystem, a K -step asymptotically regulatable region was characterized by a set of all the initial states of the plant and the exosystem. Improved internal model principles were constructed on the balance between the state convergence rate and the control of all the initial state. Finally, a state feedback controller was designed to ensure exponential output regulation in the regulatable region with disturbance rejection. Simulation examples were given to illustrate the effectiveness of proposed method. The results show these systems can go into stable rapidly and periodically.

Keywords: saturation constraint, output regulation, internal model principles, feedback controller

1 Introduction

Reference signals tracking is an important subject in systems theory. Regulation theory provides a framework that allows the analysis and design of controllers capable of achieving the tracking of references, even in the presence of disturbances.

Saturation constraint is a kind of nonlinear constraint in many practical conditions. This addresses the problem of designing a feedback controller for an uncertain plant so that the closed loop system is internally stable and the output of the closed-loop system can asymptotically track a class of reference inputs in the presence of a class of disturbances. Francis and Wonham [1, 2] proposed the internal mode principle, which aims to convert the output regulation problem of a given plant into a stabilization problem of an augmented system composed of the given plant and a well defined dynamic compensator.

For the cases where the exogenous signals are constant, Francis [2] designed a linear robust regulator based on the linear approximation of the plant can solve the local structurally stable output regulation problem for the nonlinear plant. Huang and Rugh [3] made a further work and put the solution to nonlinear plant under normal disturbance. Self-Adaptive method and optimal feedback control [4-7] were used in solving the problem of globe robust output regulation for nonlinear system disturbed by uncertain exogenous signals. Disturbance suppression of a class of nonlinear systems was studied in [8-10]. However, it should be pointed that most of the studies are carried with semi-stable exosystem, the problem of output regulation for saturated systems under the action of nonlinear exosystem has received relatively less attention. The few works motivate our current research are [11-14]. In [11], robust adaptive constrained motion

tracking control methodology was derived for bounded nonlinear effects and external disturbance within the closed-loop system. Output regulation for periodic signal of constrained MIMO system subject to actuators saturated is studied in [12]. To exact output regulation for Takagi-Sugeno (T-S) fuzzy models, [13] considered the fuzzy model as a special class of linear time-varying systems, existence conditions are rigorously derived.

In the nonlinear case, the inclusion of an internal model was proved to be a necessary condition to guarantee robustness with respect to parameter variations. This internal model is obtained as an immersion of the exosystem into a dynamical system which generates all the possible steady-state inputs for any admissible parameter variation [14].

The steady-state zero-error manifold is a centre manifold, which becomes invariant by the effect of the steady-state input. Therefore, the regulation process can be understood as follows: 1) The stabilizer is responsible for taking the states of the plant toward the steady-state zero-error manifold, reducing this way the tracking error; 2) the steady-state input keeps the states of the plant on the steady-state zero-error manifold, this way achieving the exact tracking of the reference signals. Then, regulation problem consists in finding both the steady state zero-error manifold and the steady-state input [15]. See Figure 1 for the graphical representation of the nonlinear regulation problems.

In this paper we consider the regulation problem of linear system subject to actuator saturation under the action of nonlinear exosystem. Based on our earlier results mentioned in [16], a simple feedback controller was achieved by a stabilizing law for output regulation of linear system with input constrains.

* *Corresponding author* e-mail: hw@hdu.edu.cn

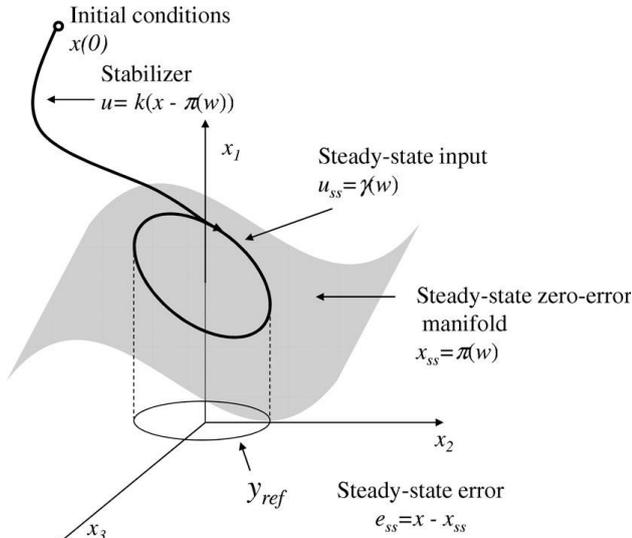


FIGURE 1 Regulation scheme for nonlinear systems

Under the action of a nonlinear exosystem action, the problem to be addressed in this paper is the following: (1) Characterize of the regulatable region. The first task of this paper is to characterize the set of initial conditions for which there exist admissible controls to keep the state bounded and to drive the tracking error to 0 asymptotically. (2) Design of constrained state feedback controller. Find a state feedback law and construct the state controller.

2 Problem statement and preliminaries

Consider the system

$$\begin{cases} x(k+1) = Ax(k) + Bu(k) + P\omega(k) \\ e(k) = Cx(k) + Q\omega(k) \\ \omega(k+1) = S\omega(k) \end{cases}, \tag{1}$$

where $A \in R^{n \times n}$, $B \in R^{n \times m}$, $P \in R^{n \times r}$, $C \in R^{p \times n}$, $Q \in R^{p \times r}$. The first plant describes a plant, with state $x \in R^n$, input $u \in R^m$ and $\|u\|_\infty \leq 1$, subject to the effect of disturbance represented by $P\omega(k)$. The error between the actual output $Cx(k)$ and a reference signal $Q\omega(k)$ is defined as $e(k)$ by the second equation. The third equation describes the exosystem with state $\omega \in R^r$ and $S \in R^{r \times r}$.

Due to the constraint input, it's well known that the initial state of the plant and exosystem can not be in the whole space. We should characterize the set of all initial states $(x_0, \omega_0) \in R^{n+r}$, on which the problem of constrained output regulation is solvable. This set is called regulatable region. If we can construct a state feedback law, $u = \phi(x, \omega)$, $\|\phi(x, \omega)\|_\infty \leq 1$ and $\phi(0, 0) = 0$, by which following conditions are satisfied:

A. Plant $x(k+1) = Ax(k) + B\phi(x, \omega)$ is asymptotically stable on the equilibrium point $x=0$.

B. For all initial states $(x_0, \omega_0) \in R^{n+r}$ in regulatable region, the close-loop system has $\lim_{k \rightarrow \infty} e(k) = 0$.

To begin with, some necessary assumptions are made:

A1. The pair (A, B) is stabilizable.

A2. S has all its eigenvalues on the unit circle and diagonalizable.

A3. $\left(\begin{bmatrix} C & Q \\ 0 & S \end{bmatrix}, \begin{bmatrix} A & P \\ 0 & S \end{bmatrix} \right)$ is measurable.

A4. There exist matrices Π and Γ solve the linear matrix equation

$$\begin{cases} \Pi S = A\Pi + B\Gamma + P \\ 0 = C\Pi + Q \end{cases}, \tag{2}$$

In this paper, we focus on two kinds of nonlinear external disturbance: the square wave and triangle wave. The square wave is discontinuous and underivable, can be described as $\omega(k+1) = S\omega(k)$, S is a unit matrix.

Let $\omega(0) = [m \ m]'$, when $k = nT/2$ ($n=0, 1, 2 \dots$), $\omega(k) = (-1)^n \omega(0)$. There are two step signals of different amplitude in one cycle, and the step signal is linear. If the period T is long enough, the action of exosystem can be viewed as tow constant disturbance that works alternatively. Review our earlier works in [16], it is possible to design an easily implementable state controller to make the close loop system stable asymptotically, simulation results are shown in section 5. Detailed study on output regulation problem is focus on the influence of periodic triangle wave.

3 The regulatable region

The triangle wave is continuous but underivable. Triangle with period T and amplitude m is described as follows, where $\omega(0) = 0$:

$$\omega(k+1) = \begin{cases} \omega(k) + a & nT \leq k < nT + T/2 \\ \omega(k) - a & nT + T/2 \leq k < (n+1)T \end{cases} \quad n=0, 1, 2, 3 \dots \tag{3}$$

at the equilibrium point, let $u(k) = \Gamma\omega(k) + Ga$, $x(k) = \Pi\omega(k)$.

By (1)

$$e(k) = Cx(k) + Q\omega(k) = C\Pi\omega(k) + Q\omega(k) = 0. \tag{4}$$

If B has full row rank, then G exists made:

$$\begin{cases} \Pi = BG & nT \leq k < nT + T/2 \\ \Pi = -BG & nT + T/2 \leq k < (n+1)T \end{cases} \quad n=0, 1, 2, 3 \dots, \tag{5}$$

$$\left\{ \begin{array}{l} \Pi\omega(k) = A\Pi\omega(k) + B\Gamma\omega(k) + P\omega(k) \\ nT \leq k < nT + T/2 \\ n = 0, 1, 2, 3 \dots \end{array} \right. \quad (6)$$

$$\left\{ \begin{array}{l} \Pi\omega(k) = A\Pi\omega(k) + B\Gamma\omega(k) + P\omega(k) \\ nT + T/2 \leq k < (n+1)T \end{array} \right.$$

Due to $\omega(k) \neq 0$, by (3), (6), the internal mode of triangle wave action is represents as (7):

$$\left\{ \begin{array}{l} \Pi = A\Pi + B\Gamma + P \\ C\Pi + Q = 0 \end{array} \right. \quad (7)$$

Consider system (1), a control signal u is said to be admissible if $\|u(k)\|_\infty \leq 1$.

Definition 3.1: For some $K > 0, (x_0, \omega_0) \in R^n \times R^r$ is said to be K -step regulatable if there exists an admissible u makes (1) satisfy $e(K) = 0$. The set of all regulatable pair (x_0, ω_0) is K -step regulatable region, denoted by $R_g(K)$.

According to classical regulation theory, there exists matrix $\Pi \in R^{n \times r}$ and matrix $\Gamma \in R^{m \times r}$ makes the equation (7) solvable, meanwhile, (7) is a zero-state equation which describes the equilibrium point as

$$u(k) = \Gamma\omega(k) + Ga, \quad x(k) = \Pi\omega(k), \quad (8)$$

where $e = 0$. Due to the restriction that $\|u(k)\|_\infty \leq 1$, $e(k)$ will go to zero asymptotically at the equilibrium point only if

$$\sup_{k \geq 0} |\Gamma\omega(k) + Ga|_\infty \leq 1. \quad (9)$$

Thus, the exosystem initial conditions corresponding to this equilibrium point are restricted in the compact set

$$W_0 = \{\omega_0 \in R^r: |\Gamma a T/2 + Ga|_\infty \leq 1, \forall k \geq 0\}. \quad (10)$$

Definition 3.2: For some $K > 0$, a state x_0 is said to be null controllable if there exists an admissible u makes the system state transforms from $x(0) = x_0$ and satisfies $\lim_{k \rightarrow \infty} x(k) = 0$. The set of all the null controllable region x_0 is null controllable region, denoted by $C(A, B)$. Specially, the set of null controllable region is called K -step null controllable region when $x(K) = 0$, denoted by $C_K(A, B)$.

By similarity transformation, we may assume

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \in R^{(n_1+n_2) \times (n_1+n_2)}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \in R^{(n_1+n_2) \times m},$$

where

A_1 has all eigenvalues inside or on the unit circle and A_2 has all eigenvalues outside the unit circle. So, the null

controllable region $C(A, B) = R^{n_1} \times C(A_2, B_2)$. We consider the condition about all the eigenvalues of A are outside the unit circle. Generally, if K is large enough (i.e. $K = 10 \sim 30$), $C_K(A, B)$ is fairly approximate to $C(A, B)$.

Correspondingly, let

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad P = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}, \quad Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}.$$

Now, we will describe the regulatable region R_g in terms of $C_K(A, B)$ and W_0 .

Lemma 1 [17]. Let $V_0 \in R^{n_2 \times r}$ be the unique solution to the linear matrix equation $V_0 S - A_2 V_0 = P_2$. Then the K -step regulatable region $R_g(K)$ is given by

$$R_g(K) = \left\{ (x_{10}, x_{20}, \omega_0) \in R^{n_1} \times R^{n_2} \times W_0 : x_{20} - V_0 \omega_0 \in C_K(A_2, B_2) \right\}. \quad (11)$$

For the first semi-cycle of triangle wave, let $T_1 = T/2$, by carrying out a similarity transformation

$$x(T_1) = A^{T_1} x_0 + \sum_{i=0}^{T_1-1} A^{T_1-i-1} B u(i) + \sum_{i=0}^{T_1-1} A^{T_1-i-1} P \omega(i) \quad (12)$$

we get

$$\begin{bmatrix} e_1(T_1) \\ e_2(T_1) \end{bmatrix} = \begin{bmatrix} Cx_1(T_1) - Q_1\omega(T_1) \\ Cx_2(T_1) - Q_2\omega(T_1) \end{bmatrix} \quad (13)$$

Since $Q_2\omega(T_1)$ is bounded for all k and $A_2^K \rightarrow \infty$ when $k \rightarrow T_1$, $\lim_{k \rightarrow T_1} e(k) = 0$ stands on

$$x_{20} + \sum_{i=0}^{T_1} A_2^{-i-1} B_2 u(i) + \sum_{i=0}^{T_1} A_2^{-i-1} P_2 i a = 0. \quad (14)$$

Denote $V_0 = -\sum_{i=0}^{T_1} A_2^{-i-1} P_2 i$, V_0 satisfies $V_0 - A_2 V_0 = (A - I)^{-1} P_2$. Let $(A - I) = D$, then $D(V_0 - A_2 V_0) = P_2$.

For the second semi-cycle of triangle wave, which can be viewed as the result of half a cycle parallel translation towards the right direction on the time axis $\omega(k+1) = \omega(k) - a$, $\omega(0) = aT_1$.

The regulator equation

$$\left\{ \begin{array}{l} \Pi = A\Pi + B\Gamma + P \\ C\Pi + Q = 0 \\ \Pi = -BG \end{array} \right. \quad (15)$$

Similarly, let $V_0 = -\sum_{i=0}^{T_1} A_2^{-i-1} P_2 (T_1 - i)$, $-(A - I) = D$, we get $D(V_0 - A_2 V_0) = P_2$.

4 State feedback controller design

In this section, we will construct a state feedback controller for above system.

Lemma 2 [18]. Let $\lambda \in (0, 1)$, for any initial condition $\tilde{x}_0 \in C_\lambda = C(\lambda^{-1}A_2, \lambda^{-1}B_2)$, there exists a state feedback law $u(k) = h[x(k)]$ such the solution of $x(k+1) = A_2x(k) + B_2u(k)$ satisfies $x(k) \in \lambda^k \rho C_\lambda(x_0) C_\lambda$ and the control signal $|u(k)|_\infty \leq \lambda^k \rho C_\lambda(x_0) \leq \lambda^k$

Lemma 2 gives a balance between the state convergence rate and the control of all the initial state in \bar{C}_λ , denoted by λ^k . The construction of this state feedback controller constructed in [16], based on which, we will construct a revised controller law for regulation problem in this paper.

Theorem 2. Assume there exists a matrix V_0 that satisfies $D(V_0 - AV_0) = P_2$, for every initial pair (x_0, ω_0) in the regulatable region, under the following controller, $u(k) = h[x(k) - \lambda^k V_2 \omega(k) - (1 - \lambda^k) \Pi_2 \omega(k)] + (1 - \lambda^k) (\Gamma \omega(k) + Ga)$ the closed-loop system satisfies $\lim_{k \rightarrow \infty} e(k) = 0$.

Proof. Corresponding to (8), we can divide system (1) in to two subsystems

$$\begin{aligned} x_1(k+1) &= A_1x(k) + B_1u(k) + P_1\omega(k) \\ x_2(k+1) &= A_2x(k) + B_2u(k) + P_2\omega(k) \end{aligned} \tag{16}$$

Denote

$$\tilde{x}_i(k) = x_i(k) - \lambda^k V_i \omega(k) - (1 - \lambda^k) \Pi_i \omega(k), \quad i = 1, 2. \tag{17}$$

By Lemma 1, for $i = 1, 2$, we get

$$\begin{aligned} \tilde{x}_i(k+1) &= A_i \tilde{x}_i(k) + B_i u(k) + (\lambda^k - 1) B_i \Gamma \omega(k) \\ &\quad - \lambda^k (I - D) P_i \omega(k) - \lambda^k V_i a - (1 - \lambda^k) \Pi_i a \end{aligned} \tag{18}$$

Based on the controller defined in Lemma2, the state feedback controller can be constructed as:

$$u(k) = h[\tilde{x}_2(k)] + (1 - \lambda^k) (\Gamma \omega(k) + Ga). \tag{19}$$

Apply it to the two subsystems

$$\begin{aligned} \tilde{x}_1(k+1) &= A_1 \tilde{x}_1(k) + B_1 h[\tilde{x}_2(k)] - \lambda^k (I - D) P_1 \omega(k) - \lambda^k V_1 a \\ \tilde{x}_2(k+1) &= A_2 \tilde{x}_2(k) + B_2 h[\tilde{x}_2(k)] - \lambda^k (I - D) P_2 \omega(k) - \lambda^k V_2 a \end{aligned} \tag{20}$$

Then we can get $\lim_{k \rightarrow T_1} \tilde{x}_2(k) = 0$, $|h[\tilde{x}_2(k)]|_{T_1} \leq \lambda^k$ by Lemma 2. Since A_1 is semi-stable and $|h[\tilde{x}_2(k)]|_{T_1} \leq \lambda^k$, $\tilde{x}_1(k)$ also convergences to the origin.

$$|u(k)|_{T_1} = |h[\tilde{x}_2(k)] + (1 - \lambda^k) (\Gamma \omega(k) + Ga)|_{T_1} \leq 1. \tag{21}$$

The closed-loop system satisfies $\lim_{k \rightarrow T_1} e(k) = 0$. Similar controller can be constructed for the second semi-cycle of a triangle cycle.

5 Numerical Examples

Example 1. A semi-stable system as follows under the action of square signal (T/2=1000)

$$\begin{aligned} x(k+1) &= \begin{bmatrix} 1.4 & 0 \\ 0.2 & 1.2 \end{bmatrix} x(k) + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} u(k) + \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \omega(k) \\ \omega(k+1) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \omega(k) \\ e(k) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x(k) - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \omega(k) \end{aligned} \tag{22}$$

With $x_0 = [-1.5 \ -0.8]^T$, $\omega(0) = [1.5 \ 1.5]^T$, the regulation equation has solutions $\Pi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$,

$$\Gamma = S - A - P = \begin{bmatrix} -0.5 & 0 \\ -0.2 & -0.3 \end{bmatrix}, \quad V = \begin{bmatrix} -0.25 & 0 \\ 0.25 & -0.5 \end{bmatrix}.$$

Applying the controller provided in [16] $u(k) = h[x(k) - 0.97^k V \omega(k) - (1 - 0.97^k) \Pi \omega(k)] + (1 - 0.97^k) \Gamma \omega(k)$

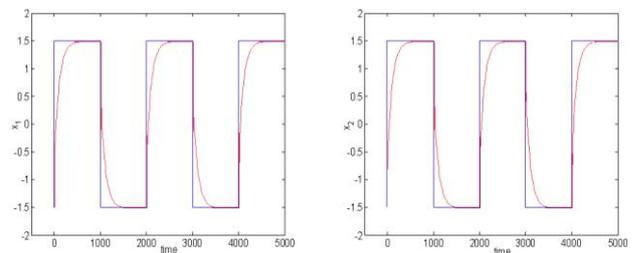


FIGURE 2 closed-loop state tracking under the square signal disturbance

The closed-loop state tracking is shown in Figure 2.

Example 2. The following system under the action of triangle signal (T=1000)

$$\begin{aligned} x(k+1) &= \begin{bmatrix} 1.4 & 0 \\ 0.2 & 1.2 \end{bmatrix} x(k) + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} u(k) + \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \omega(k) \\ e(k) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x(k) - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \omega(k) \end{aligned} \tag{23}$$

In the first semi-cycle, $x_0 = [-0.1 \ -0.01]^T$, $\omega_0 = [0 \ 0]^T$, $a = [0.003 \ 0.004]^T$. The regulation equation has solutions

$$\Pi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Gamma = S - A - P = \begin{bmatrix} -0.5 & 0 \\ -0.2 & -0.3 \end{bmatrix},$$

$$G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

$D(V-AV)=P$ has a unique solution

$$V = \begin{bmatrix} -0.625 & 0 \\ 1.875 & -2.5 \end{bmatrix}.$$

The state feedback controller $u(k)=h[x(k)-0.95^k V\omega(k)-(1-0.95^k)\Pi\omega(k)] + (1-0.95^k)(\Gamma\omega(k)+Ga)$.

The closed-loop state tracking are plotted in Figure 3.

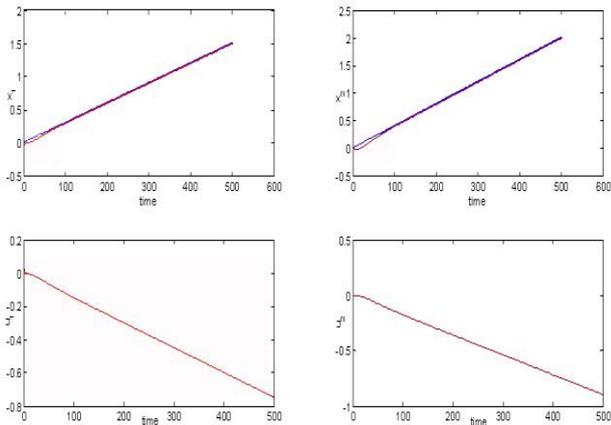


FIGURE 3 Closed-loop state tracking in first semi-cycle in Example 2

During the last semi-cycle, $x_0=[1.5 \ 2.0]^T$, $\omega_0=[1.5 \ 2.0]^T$, $a=[0.003 \ 0.004]^T$.

$$\Pi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Gamma = S - A - P = \begin{bmatrix} -0.5 & 0 \\ -0.2 & -0.3 \end{bmatrix},$$

$$G = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

There exists the unique solution to $D(V-AV)=P$

$$V = \begin{bmatrix} 0.625 & 0 \\ -1.875 & 2.5 \end{bmatrix}.$$

The state feedback controller $u(k)=h[x(k)-0.95^k V\omega(k)-(1-0.95^k)\Pi\omega(k)] + (1-0.95^k)(\Gamma\omega(k)+Ga)$. The closed-loop state trackings are plotted in Figure 4.

In each cycle period, two different internal mode principles are applied for a semi-cycle respectively, thus G and V are got and the state-feedback controller $u(k)$ are constructed. State tracking in two cycles are shown in Figure 5, with $x_0=[-0.1 \ -0.01]^T$, $\omega_0=[0 \ 0]^T$, $a=[0.003 \ 0.004]^T$.

References

[1] Francis B A 1976 The linear multivariable regulator problem *IEEE Conf on Decision and Control including the 15th Symposium on Adaptive Processes* NJ 15 pp 873-8
 [2] Francis B A, Wonham W M 1976 The internal model principle of control theory *Automatica* 12(5) 457-65
 [3] Huang J, Chen Z 2002 A general framework for output regulation problem *Proceedings of the American Control Conference. Anchorage* 1 102-9

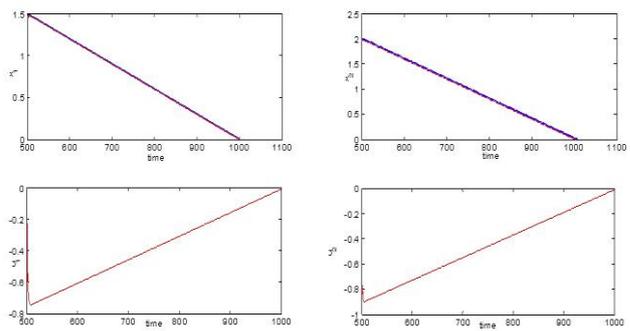


FIGURE 4 closed-loop state tracking in last semi-cycle in Example 2

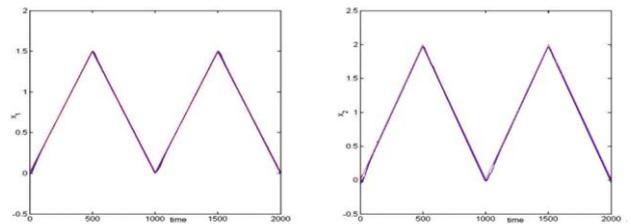


FIGURE 5 State tracking in two cycles in Example 2

6 Conclusions

In this brief, we studied the output regulation problem of saturated linear system under the action of nonlinear exosystem. At the equilibrium point, initial state of the plant and exosystems are restricted in a compact set W_0 . The K -Step asymptotically regulatable region $R_g(K)$ is described by W_0 and K -Step null controllable region $C_K(A,B)$. Segmented control strategies are applied to external disturbances in the case of square signal and triangle signal. The internal principles for each semi-cycle of the exosystem are given. Controller is constructed based on the state feedback laws proposed. Examples has demonstrated the effectiveness of the proposed control methodology.

Acknowledgment

This work is supported by Science and Technology Project of Zhejiang Province (Grant No. 2012C21095), China.

[4] Chung C H, Chen M S, Albert W J 2012 A new adaptive control for periodic tracking/disturbance rejection *Asian Journal of Control* 14(8) 827-34
 [5] Dai ShiLu, Wang Min, Wang Cong, Li Liejun 2014 Learning from adaptive neural network output feedback control of uncertain ocean surface ship dynamics *International Journal of Adaptive Control and Signal Processing* 28(25) 341-65

- [6] Ding Z 2003 Global stabilization and disturbance suppression of a class of nonlinear systems with uncertain internal model *Automatica* **39**(3) 471–79
- [7] Chakraborty A, Arcak M 2009 Time-scale separation redesigns for stabilization and performance recovery of uncertain nonlinear systems *Automatica* **45** 34–44
- [8] Saverio Messineo, Andrea Serrani 2009 Adaptive feedforward disturbance rejection in nonlinear systems *Systems & Control Letters* **58**(8) 576–83
- [9] Qing W J 2009 Disturbance rejection through disturbance observer with adaptive frequency estimation *IEEE Transactions on Magnetics* **45**(6) 2675–8
- [10] Wu H 2013 Adaptive robust stabilization for a class of uncertain nonlinear time-delay dynamical systems *International Journal of Systems Science* **44**(13) 371–83
- [11] Liaw H C, Shirinzadeh B 2011 Robust adaptive constrained motion tracking control of Piezo-actuated flexure based mechanisms for micro/nano manipulation *IEEE Transactions on Industrial Electronics* **58**(4) 1406–15
- [12] Flores J V, Gomes Da Silva J M, Pereira L F A, Sbarbaro D G 2012 Repetitive control design for mimo systems with saturating actuators *IEEE Transactions on Automatic Control* **57**(1) 192–8
- [13] Meda-Campana J A, Gomez-Mancilla J C, Castillo-Toledo B 2012 Exact Output Regulation for Nonlinear Systems Described by Takagi–Sugeno Fuzzy Models *IEEE Transactions on Fuzzy Systems* **20**(2) 235–47
- [14] Serrani A 2005 *The nonlinear output regulation problem local and structurally stable regulation* <http://www.cesos.ntnu.no/activities/workshops/ormls/lecture2.pdf> June 2014
- [15] Isidori A 2005 *Nonlinear Control Systems* Springer Verlag: Berlin pp 67
- [16] Zhao X, Chai L, Xue A 2005 Output Regulation of Linear Systems with Input Constraints *American Control Conference 2005 Portland* **1** 2088–92
- [17] Hu T, Lin Z 2004 Output Regulation of Linear Systems With Bounded *IEEE Transactions on Automatic Control* **49**(11) 1941–53
- [18] Qiu L 2000 Stabilization of Linear systems with input constraints *Proceedings of the 39th IEEE Conference on Decision and Control* Sydney **4** 3272–7

Authors



Wei Huang, born in 1974, Guangyuan, Sichuan, China

Current position, grades: lecturer in school of automation, HDU, China, and a member of the Intelligent robot control institute, where he is engaged in research and development of motion control, robotic applications, and industrial automation.

University studies: B.Sc. degree in electronic engineering from the Hefei University of Technology and Superior Studies of M.Sc. degrees in Computer application from the Hanzhou Dianzi University (HDU), Hangzhou, China.

Approximate trace equivalence of real-time linear algebraic transition systems

Hui Zhang¹, Jinzhao Wu^{2*}, Hongyan Tan³, Hao Yang¹

¹Chengdu Institute of Computer Application, Chinese Academy of Sciences, China

²Guangxi Key Laboratory of Hybrid Computational and IC Design Analysis, Guangxi University for Nationalities, China

³Institute of Acoustics, Chinese Academy of Sciences, China

Received 1 March 2014, www.tsi.lv

Abstract

In allusion to data error and equivalence relation for software program design, the paper proposes approximate trace equivalence of real-time linear algebraic transition systems. Firstly, it leads real-time algebraic program into transition system and establishes real-time linear algebraic transition system. And then, it uses matrix norm and matrix singular value decomposition to analyse approximation of traces. Afterwards, it obtains approximate trace equivalence of real-time linear algebraic transition systems. Finally, the traffic light control vehicle flow system example shows that approximate trace equivalence of real-time algebraic transition systems can optimize real-time linear algebraic programs and reduce states.

Keywords: transition system, approximate, trace equivalence, algebraic program

1 Introduction

Early formal model of software programs are mostly discrete model of concurrent systems, such as various process algebra on the language level, computation tree logic on the logical level, automatic machine and transition system on the structural level. They are made up of abstract actions, discrete states and transition relations between states. The discrete method is unable to exchange the data stream. The data stream exchange is one of the most important functions of software programs. It needs to be created algebraic programs which can describe programs data stream exchange. In 2004, Z. Manna et al. defined Hybrid Systems program model with polynomial equations data stream is expressed as states transition labels and opened up a new way of program design and verification based on polynomial algebra [1].

How to define and determine the behaviour equivalence of programs is one of the most important issues in the field of software system design and verification analysis. The same model structure as an equivalent definition is too strictly. A class of functional behaviour of different structure program model may be exactly the same. Therefore, the program models with the same function behaviour are called equivalent. Functional equivalence makes program model be simplified by removing the duplicate branches. Glabbeek proposed fourteen kinds of linear time-branching time equivalence relations [2]. Trace equivalence is a basic state space equivalence relation and it has been widely used in the

software system program design and verification analysis. But at present there is no research on trace equivalence of real-time linear algebraic transition systems.

In the software program design and verification analysis, experimental data often have errors. We can only take the approximate value and cannot obtain accurate actual value. Within the error range, we can be treated approximate value as actual value. After the value approximation, it makes some same programs and simplifies the structure of real-time linear algebraic transition system.

From what has been discussion above, firstly, it introduces real-time linear algebraic program to transition system and establishes real-time linear algebraic transition system. Secondly, it uses matrix norm and matrix singular value decomposition to judge whether two real-time linear algebraic transition systems are approximate. Thirdly, it simplifies real-time linear algebraic transition system by trace equivalence theory. Finally, the traffic lights control traffic flow system example shows that approximate trace equivalence of real-time linear algebraic transition systems can optimize real-time linear algebraic programs and reduce the states.

2 Real-time linear algebraic transition system

As we know, labelled transition system is a kind of typical model which describes transition relation between states. Many computer system modelling are based on labelled transition system.

* *Corresponding author* e-mail: 415360889@qq.com

Definition 1 (Labelled Transition System [3]) A labelled transition system is a tuple $M = \langle S, L, T, s_0 \rangle$, where

- 1) S is a finite set of states, $s \in S$ is a state.
- 2) L is a finite set of labels, $l \in L$ is a label, a label describes an action.
- 3) $T \subseteq S \times L \times S$ is a set of transitions.
- 4) s_0 is the initial system state.

Definition 2 (Transition of Labelled Transition System) A transition of labelled transition system is a tuple $\langle s_i, l, s_j \rangle$, where s_i represents pre-state of the transition and s_j represents post-state of the transition.

Definition 3 (Trace [3]) In the transition system, a trace is an action sequence $l_1 l_2 \dots l_n$ and satisfies an execution sequence $\eta = s_0 l_1 s_1 l_2 \dots l_n s_n$. $trace(\eta) = l_1 l_2 \dots l_n$.

Definition 4 (Real-Time Linear Algebraic Program) Let \mathbb{R} be the set of real numbers, $x_i (i = 1, \dots, n) \in \mathbb{R}$ and $x'_i (i = 1, \dots, n) \in \mathbb{R}$ be variables, $t \in \mathbb{R}$ is a time variable and $t > 0$. An algebraic program $X' = X + (AX_0 + b)t$ is a real-time linear algebraic program, where $X_0 = (x_{01}, \dots, x_{0n})^T$ is the initial state value of real-time linear algebraic transition system, $X = (x_1, \dots, x_n)^T$ is the pre-state value of real-time linear algebraic program and $X' = (x'_1, \dots, x'_n)^T$ is the post-state value of real-time

linear algebraic program. $A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$ is the $n \times n$ matrix and $A \neq 0$, $b = (b_1, b_2, \dots, b_n)^T$ is the n dimension column vector. The elements of A and b are all real numbers.

In the real-time linear algebraic program, the value of time variable t is a fixed real number. t represents the time from one state transition to another state.

Definition 5 (Real-Time Linear Algebraic Transition System) A real-time linear algebraic transition system is a tuple $M = \langle V, TX, S, P, Q, s_0 \rangle$, where

- 1) $V = \{x_1, \dots, x_n\}$ is a finite set of system variables.
- 2) TX is a finite set of system state values. $X_0 \in TX$ is the initial state value.
- 3) S is a finite set of states, $s \in S$ is a state.
- 4) P is a finite set of real-time linear algebraic program, $p \in P$ is a real-time linear algebraic program.
- 5) $Q \subseteq S \times P \times S$ is a finite set of state transitions, $q \in Q$ is a state transition.
- 6) $s_0 \in S$ is the initial state.

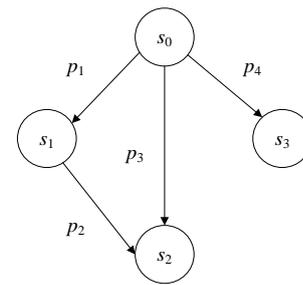


FIGURE 1 An real-time linear algebraic transition system

In the Figure 1, s_0 is the initial state, X_0 is the initial state value. s_1, s_2, s_3 are all system states, X_1, X_2, X_3 respectively represent the state value of s_1, s_2, s_3 . p_1, p_2, p_3, p_4 are real-time linear algebraic programs.

3 Approximation of real-time linear algebraic transition systems

In the software program design, matrix elements and vector elements which are in the real-time linear algebraic programs often have errors. Let $A + \delta A$ be actual matrix, A be approximate matrix, $b + \delta b$ be actual vector, b be approximate vector, $X' = X + [(A + \delta A)X_0 + (b + \delta b)]t$ be actual real-time linear algebraic program, $X' = X + (AX_0 + b)t$ be approximate real-time linear algebraic program. In the actual real-time linear algebraic program and corresponding approximate real-time linear algebraic program, time t is the same. We use matrix norm and matrix singular value decomposition to analyse approximation of real-time linear algebraic transition systems.

Definition 6 (Matrix Singular Value Decomposition) Let $A = U \Sigma V^T$ be the matrix singular value decomposition, the main diagonal elements $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$ of Σ are called singular value of matrix A , the column vectors u_1, u_2, \dots, u_n of matrix U are eigenvectors of matrix AA^T , u_1, u_2, \dots, u_n are left singular vectors of matrix A , v_1, v_2, \dots, v_n are right singular vectors of matrix A .

If $\sigma_{r+1} = \dots = \sigma_n = 0$, then the matrix A singular value decomposition is $A = \sum_{i=1}^r \sigma_i u_i v_i^T$.

Let matrix $A \in \mathbb{R}^{m \times n}$, a norm $\|A\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{\frac{1}{2}}$ of matrix A is called Frobenius norm of matrix A . Matrix norm $\|A\|_F$ and vector norm $\|x\|_2$ are inclusive.

Theorem 1. For any matrix $A \in \mathbb{R}^{m \times n}$, $r = \min\{m, n\}$,

$$\sqrt{\sum_{i=1}^r \sigma_i^2} = \|A\|_F.$$

Let s_0 be the initial state, $p'_1 p'_2 \dots p'_k$ be the trace of actual real-time linear algebraic transition system and it

satisfies a finite execute sequence $\eta=s_0p'_1s_1p'_2 \dots p'_ks_k$, $p_1p_2 \dots p_k$ be the trace of approximate real-time linear algebraic transition system and it satisfies a finite execute sequence $\eta=s_0p_1s_1p_2 \dots p_ks_k$. p'_1, p'_2, \dots, p'_k are actual real-time linear algebraic programs and p_1, p_2, \dots, p_k are approximate real-time linear algebraic programs. The actual value of state s_j is $X'_j = [E + jt(A + \delta A)]X_0 + jt(b + \delta b)$, the approximate value of state s_j is $X_j = (E + jtA)X_0 + jtb$. The actual value X'_j can be written as

$$X'_j = \begin{pmatrix} E + jt(A + \delta A) & jt(b + \delta b) \\ & 1 \end{pmatrix} \begin{pmatrix} X_0 \\ 1 \end{pmatrix}, \quad \text{the}$$

approximate value X_j can be written as

$$X_j = \begin{pmatrix} E + jtA & jtb \\ & 1 \end{pmatrix} \begin{pmatrix} X_0 \\ 1 \end{pmatrix}. \quad \text{Approximate of actual value}$$

X'_j and approximate value X_j is equal to approximate of matrix $\begin{pmatrix} E + jt(A + \delta A) & jt(b + \delta b) \\ & 1 \end{pmatrix}$ and matrix

$$B_j = \begin{pmatrix} E + jtA & jtb \\ & 1 \end{pmatrix}. \quad \text{Let } 1 \leq j \leq k, \quad p'_j \text{ is}$$

$$X' = X + [(A_j + \delta A_j)X_0 + (b_j + \delta b_j)]t_j, \quad p_j \text{ is}$$

$$X' = X + (A_j X_0 + b_j)t_j, \quad \text{the actual value of state } s'_j$$

$$\text{is } X'_j = \left\{ E + \sum_{i=1}^j [(A_i + \delta A_i)t_i] \right\} X_0 + \sum_{i=1}^j [(b_i + \delta b_i)t_i], \quad \text{the}$$

approximate value of state s_j is

$$X_j = \left[E + \sum_{i=1}^j (A_i t_i) \right] X_0 + \sum_{i=1}^j (b_i t_i). \quad \text{The actual value } X'_j$$

can be written as

$$X'_j = \begin{pmatrix} E + \sum_{i=1}^j [(A_i + \delta A_i)t_i] & \sum_{i=1}^j [(b_i + \delta b_i)t_i] \\ & 1 \end{pmatrix} \begin{pmatrix} X_0 \\ 1 \end{pmatrix}, \quad \text{the}$$

approximate value X_j can be written as

$$X_j = \begin{pmatrix} E + \sum_{i=1}^j (A_i t_i) & \sum_{i=1}^j (b_i t_i) \\ & 1 \end{pmatrix} \begin{pmatrix} X_0 \\ 1 \end{pmatrix}. \quad \text{The approximation}$$

of actual value X'_j and approximate value X_j is equal to approximation of matrix

$$\begin{pmatrix} E + \sum_{i=1}^j [(A_i + \delta A_i)t_i] & \sum_{i=1}^j [(b_i + \delta b_i)t_i] \\ & 1 \end{pmatrix} \quad \text{and matrix}$$

$$\begin{pmatrix} E + \sum_{i=1}^j (A_i t_i) & \sum_{i=1}^j (b_i t_i) \\ & 1 \end{pmatrix}.$$

$$B_j + \delta B_j = \begin{pmatrix} E + \sum_{i=1}^j [(A_i + \delta A_i)t_i] & \sum_{i=1}^j [(b_i + \delta b_i)t_i] \\ & 1 \end{pmatrix},$$

$$B_j = \begin{pmatrix} E + \sum_{i=1}^j (A_i t_i) & \sum_{i=1}^j (b_i t_i) \\ & 1 \end{pmatrix}.$$

The singular value decomposition of matrix δB_j is

$$\delta B_j = U_j \Sigma_j V_j^T \quad \text{and} \quad \|\delta B_j\|_F = \sqrt{\sum_{i=1}^{r_j} \sigma_{ji}^2}.$$

$$\|(B_j + \delta B_j) - B_j\|_F = \|\delta B_j\|_F = \sqrt{\sum_{i=1}^{r_j} \sigma_{ji}^2} = W_j. \quad \text{For a given}$$

positive number ε , if it holds $W_j < \varepsilon$, then matrix $B_j + \delta B_j$ and matrix B_j are approximate. If matrix $B_j + \delta B_j$ and matrix B_j are approximate, then actual value X'_j and approximate value X_j are approximate, actual real-time linear algebraic program and approximate real-time linear algebraic program are approximate.

Definition 7 (Approximation of Traces) For a given positive number ε , $\forall j, 1 \leq j \leq k$, if it holds $W_j < \varepsilon$, then actual trace $p'_1 p'_2 \dots p'_k$ and approximate trace $p_1 p_2 \dots p_k$ are approximate. If there exists $W_j \geq \varepsilon$, then actual trace $p'_1 p'_2 \dots p'_k$ and approximate trace $p_1 p_2 \dots p_k$ are not approximate.

Definition 8 (Approximation of Real-Time Algebraic Transition Systems) If all traces of two real-time linear algebraic transition systems are approximate, then these two real-time linear algebraic transition systems are approximate.

Approximation of real-time linear algebraic transition systems can optimize real-time linear algebraic programs and reduce bits of matrix elements and vector elements. It can improve computation speed of real-time linear algebraic transition system. For the actual real-time linear algebraic transition system RS_1 , it gets approximate real-time linear algebraic transition system RS_2 by approximate algorithm.

4 Approximation trace equivalence of real-time linear algebraic transition systems

Definition 9 (Trace Equivalence [2]) If there exists a process q and $p \xrightarrow{\sigma} q$, then $\sigma \in Act$ is a trace of a process p . Let $T(p)$ be a set of traces of process p . If $T(p) = T(q)$, then two processes p and q are trace equivalence, it can be written as $p =_T q$. In trace semantics two processes are identified if they are trace equivalence.

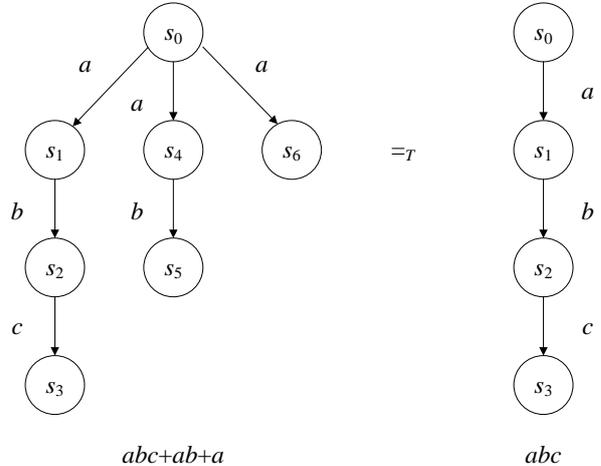


FIGURE 2 A trace equivalence example

In the Figure 2, a, b, c are real-time linear algebraic programs, left real-time linear algebraic transition system has three traces abc, ab, a , right real-time linear algebraic transition system has only one trace abc . The left system and right system have the same function.

We have the approximate real-time linear algebraic transition system RS_2 , through trace equivalence theory we obtain approximate trace equivalence system RS_3 of actual real-time linear algebraic transition system.

Approximate trace equivalence of real-time linear algebraic transition systems has a very important significance. It can optimize real-time linear algebraic programs and reduce states.

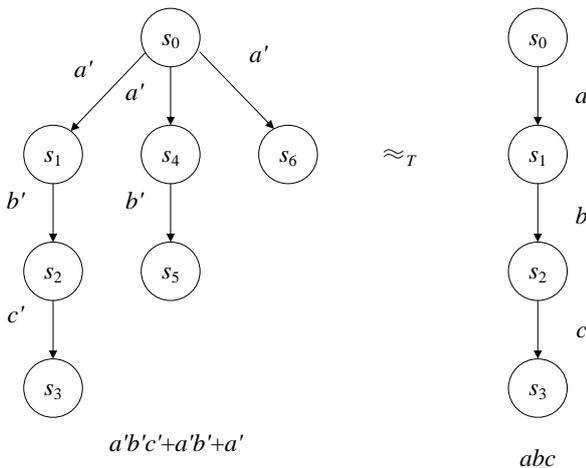


FIGURE 3 An approximate trace equivalence of real-time linear algebraic transition systems example

In the Figure 3, a', b', c' are actual real-time linear algebraic programs, a, b, c are approximate real-time linear algebraic programs. The left system has seven states, the right system has four states.

5 Experiments

Assume that vehicles on a section road are divided into two types of motor vehicle and non-motor vehicle. Exit of road has traffic lights. Traffic light basic transformation sequence is yellow→red→blue→yellow.

When the traffic light is yellow, vehicles which are beyond the stop line can forward pass. When the traffic light is red, vehicles are prohibited passage. When the traffic light is blue, vehicles are allowed passage. Let x_1 be motor vehicle flow, x_2 be non-motor vehicle flow, $X = (x_1, x_2)^T$ be vehicle flow. The initial state value of traffic light control vehicle flow system is $X_0 = (5, 5)^T$. Figure 4 is the actual traffic light control vehicle flow system.

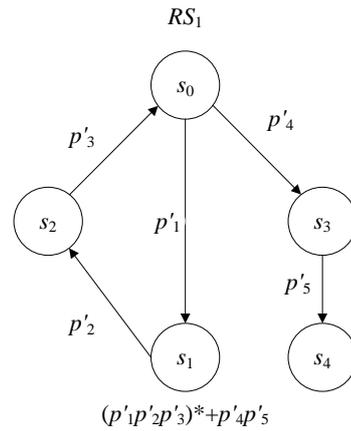


FIGURE 4 Actual traffic light control vehicle flow system

In the Figure 4, actual real-time linear algebraic program p'_1 is $X' = X + \left[\begin{pmatrix} 2.01 & \\ & 2.01 \end{pmatrix} X_0 + \begin{pmatrix} 1.01 \\ 1.01 \end{pmatrix} \right] \times 2$,

p'_2 is $X' = X + \left[\begin{pmatrix} 4.02 & \\ & 4.02 \end{pmatrix} X_0 + \begin{pmatrix} 2.02 \\ 2.02 \end{pmatrix} \right] \times 10$, p'_3 is

$X' = X + \left[\begin{pmatrix} -4.02 & \\ & -4.02 \end{pmatrix} X_0 + \begin{pmatrix} -2.02 \\ -2.02 \end{pmatrix} \right] \times 11$, p'_4 is

$X' = X + \left[\begin{pmatrix} 1.99 & \\ & 1.99 \end{pmatrix} X_0 + \begin{pmatrix} 0.99 \\ 0.99 \end{pmatrix} \right] \times 2$, p'_5 is

$X' = X + \left[\begin{pmatrix} 3.98 & \\ & 3.98 \end{pmatrix} X_0 + \begin{pmatrix} 1.99 \\ 1.99 \end{pmatrix} \right] \times 10$.

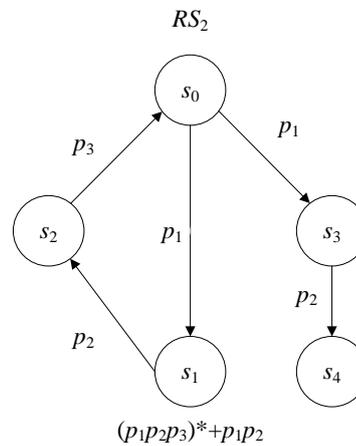


FIGURE 5 Approximate traffic light control vehicle flow system

In the Figure 5, approximate real-time linear algebraic program p_1 is $X' = X + \left[\begin{pmatrix} 2 & \\ & 2 \end{pmatrix} X_0 + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] \times 2$, p_2 is

$X' = X + \left[\begin{pmatrix} 4 & \\ & 4 \end{pmatrix} X_0 + \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right] \times 10$, p_3 is

$X' = X + \left[\begin{pmatrix} -4 & \\ & -4 \end{pmatrix} X_0 + \begin{pmatrix} -2 \\ -2 \end{pmatrix} \right] \times 11$.

For a given positive number $\varepsilon=1$, it uses formula

$$\|(B_j + \delta B_j) - B_j\|_F = \|\delta B_j\|_F = \sqrt{\sum_{i=1}^{r_j} \sigma_{ji}^2} = W_j \quad \text{to obtain}$$

$\|\delta B_j\|_{F \max} = 0.44 < \varepsilon$. Because the corresponding traces of RS_1 and RS_2 are approximate, actual traffic light control vehicle flow system RS_1 and approximate traffic light control vehicle flow system RS_2 are approximate.

Through the trace equivalence theory, we get the approximate trace equivalence system RS_3 .

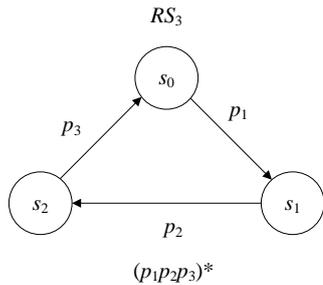


FIGURE 6 Approximate trace equivalence system

Form what has been discussion above, approximate trace equivalence of traffic light control vehicle flow

References

[1] Sankaranarayanan S, Sipma H, Manna Z 2004 Non-Linear Loop Invariant Generation Using Groebner Bases *POPL04* 318-29
 [2] van Glabbeek R J 1990 The Linear Time-Branching Time Spectrum *Inst. für Informatik* 1-86
 [3] Clarke E, Grumberg O, Peled D 2000 *Model Checking* MIT Press
 [4] Chen Y 2000 *Matrix Theory* WestNorth Industry University Press
 [5] Hierons R M 2012 Overcoming controllability problems in distributed testing from an input out transition system *Distributed Computing*, 25(1) 63-81
 [6] Bauer S S, et al 2012 Weighted modal transition systems *Formal Methods in System Design* 1-28
 [7] Shen M 2013 Hinfim filtering of continuous Markov jump linear system with partly known Markov modes and transition probabilities *Journal of the Franklin Institute* 350(10) 3384-99

systems can optimize real-time linear algebraic programs and reduce system states.

6 Conclusion

In this paper, the approximate trace equivalence of real-time linear algebraic transition system is proposed. It can optimize real-time linear algebraic programs and reduce system states. In the future work, we will use approximate trace equivalence of real-time linear algebraic transition systems to study approximate trace equivalence of real-time linear algebraic Hybrid Systems.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No.11371003, the Science and Technology Foundation of Guangxi under Grant No.10169-1, the Natural Science Foundation of Guangxi under Grant No.2011GXNSFA018154 and N0.2012GXNSFGA060003, and the Scientific Research Project No.201012MS274 from Guangxi Education Department.

[8] Fu J, Tanner H, Heinz J, Chandlee J 2014 Adaptive symbolic control for finite-state transition systems with grammatical inference *IEEE Transactions on Automatic Control* 59(2) 505-11
 [9] Bernardo M, Nicola N D, Loreti M 2014 Revisiting trace and testing equivalences for nondeterministic and probabilistic processes *Logical Methods in Computer Science* 10(1)
 [10] Cheval V, Cortier V, Delaune S 2013 Deciding equivalence-based properties using constraint solving *Theoretical Computer Science* 492 1-39
 [11] Bonchi F et al 2012 Final semantics for decorated traces *Electronic Notes in Theoretical Computer Science* 286 73-86
 [12] Weidich M, Mendling J 2012 Perceived consistency between process models *Information Systems* 37(2) 80-98

| Authors | |
|---|--|
|  | <p>Zhang Hui, China</p> <p>Current position, grades: doctoral student University studies: Chengdu Institute of Computer Application, Chinese Academy of Sciences Scientific interest: formal verification and network security</p> |
|  | <p>Wu Jinzhao, China</p> <p>Current position, grades: professor University studies: Institute of System Science, Chinese Academy of Sciences Scientific interest: formal verification and network security</p> |
|  | <p>Tan Hongyan, China</p> <p>Current position, grades: associate professor University studies: Lanzhou University Scientific interest: formal verification and network security</p> |
|  | <p>Yang Hao, China</p> <p>Current position, grades: doctoral student University studies: Chengdu Institute of Computer Application, Chinese Academy of Sciences Scientific interest: formal verification and network security</p> |

A comparative study on artificial neural networks for environmental quality assessment

Yijun Liu^{*}, Sheng He, Yao Wang, Xiumei Wang

Key laboratory of cloud computing & intelligent information processing of Changzhou City, Jiangsu University of Technology, 213001, China

Received 6 March 2014, www.tsi.lv

Abstract

The aim of this study is to use neural network tools as an environmental decision support in assessing environmental quality. A three-layer feedforward neural network using three learning approaches of BP, LM and GA-BP has been applied in non-linear modelling for the problem of environmental quality assessment. The case study shows that the well designed and trained neural networks are effective and form a useful tool for the prediction of environmental quality. Furthermore, the LM network has the fastest convergence speed and the GA-BP network outperforms the other two networks in both predictive and final classification accuracies of environmental quality.

Keywords: neural network model, hybrid ga-bp algorithm, environmental quality assessment

1 Introduction

Environment is the basic premise of human survival and development. Environmental quality is defined as a set of properties and characteristics of the environment, either generalized or local, as they impinge on human beings and other organisms [1]. It is a measure of the environmental condition relative to the requirements of human or other species. Today environmental protection has been a basic national policy in China. It is a difficult task for the government to solve serious environmental problems and make sustainable development strategies. As an essential reference for environmental protection policies and an effective tool for supervision and management, environmental quality assessment, a quantitative description of environmental quality is one of the most challenging areas in environment studies. It is significantly important to make environmental quality assessment more accurate and scientific.

Environment quality assessment can be considered as a pattern recognition task since it deals with a set of input environmental variables which contain information about the environment of a region, mapping to real values indicating environmental quality or a set of discrete and mutually exclusive classes indicating quality degree such as light pollution, heavy pollution, etc. Environmental quality ratings are typically costly to be obtained, since they require investing large amount of time and human resources to perform deep analysis of the environmental status based on various aspects of soil, atmosphere, water, etc. Previous studies focus on mathematical models created by conventional statistical methods [2]. However, in traditional statistical methods researchers are usually required to impose particular structures to

different models, such as the linearity in the multiple regression analysis, and to construct the model by estimating parameters to fit the data or observation [3]. Closely related to social, economic, management and other fields, environmental quality assessment is a complex multiple-objective, multiple-level and multiple-factor engineering problem, and hence does not adhere to common function forms. In recent years, an artificial intelligence technique, namely artificial neural network, abbr. ANN, has attracted many attentions and has been widely used in many real-world applications such as medical diagnosis [4, 5], image classification [6], signal change detection [7], handwriting recognition [8], etc. By using neural networks the structure of the models can be obtained from data automatically. To capture the complexity and the process dynamics of complicated environmental quality system, the neural network method is used to make a comprehensive analysis for environmental quality.

The rest of this paper is organized as follows. In Section 2, the paper defines the problem of environmental quality assessment mathematically and suggests the novel method of neural networks for this problem. In Section 3, three neural network training approaches of backpropagation (abbr. BP), Levenberg-Marquardt (abbr. LM) and backpropagation optimized by the genetic algorithm (abbr. GA-BP) used in this work are introduced. In section 4, a case study is given to show the effectiveness of neural networks in assessing environmental quality. Firstly, the neural network topology is designed and presented according to actual situation. Then the modelling for environmental quality with using the designed network architecture is described. And lastly, the test results of three kinds of neural

^{*} *Corresponding author* e-mail: yijunliu@vip.sina.com

networks are analysed and compared. In section 5, conclusions are drawn and the issue for future works is indicated.

2 Problem definitions

Environmental quality assessment can be defined as a pattern recognition task of regression. Let $P=\{p_1, p_2, \dots, p_t\}$ be a set of monitoring points, $X=\{x_1, x_2, \dots, x_n\}$ be a set of attributes. For attribute value vector $(x_{1i}, x_{2i}, \dots, x_{ni})$ of each monitoring point p_i , there is a corresponding assessment score y_i of p_i , $1 \leq i \leq t$. Assume there is a mapping f as such $y_i=f(x_{1i}, x_{2i}, \dots, x_{ni})$ from attribute values to assessment score. For a monitoring point to be assessed the input of the mapping f is its attribute value vector and the output is the assessment score.

Artificial neural networks can be used as an arbitrary function approximation mechanism that learns from observed data, and hence are applicable to environmental quality assessment. Although neural networks have been criticized for their poor interpretability, they have strong ability of non-linear mapping, high tolerance to noisy data and superior robustness if the learning algorithm and the cost function are appropriately selected for modelling.

The neural network method imitates the way by which the brain processes information [4]. Given an input vector $X=(x_1, x_2, \dots, x_n)$, the network produces an output vector $Y=(y_1, y_2, \dots, y_m)$, where n indicates the number of inputs and m the number of output units. Typically a feedforward neural network is organized into several layers of nodes. The first layer is the input layer and the last layer is the output. In the input layer every node corresponds to an attribute variable and hence the number of nodes equals the number of variables. The input and output layers are usually separated by one or more hidden layers. The nodes in adjacent layers are fully connected. There is a weight associated with each connection. The weight from unit i to unit j is denoted as w_{ij} . For neural network training, learning rules are used to update the weight and to minimize the error function. The learning process repeats until termination condition is met. A degree of nonlinearity is introduced to the model by the activation function to prevent the output from reaching very large values that paralyze neural network models and inhibit training.

A well-trained neural network is capable of exploiting the underlying non-linear relationships that determine the environmental rating of a region. In this paper, we propose a three-layer network based on three learning algorithms of BP, LM and GA-BP for the problem of environmental quality assessment.

3 Neural network approaches

Allowing the modelling of non-linear relationships, neural networks are useful tools for the analysis of large data sets of non-congeneric compounds with unknown or varying modes of action. The standard backpropagation

algorithm is perhaps the most widely used algorithm for supervised training of multi-layered feedforward neural networks. However, the BP algorithm has two significant disadvantages of slow convergence speed and easiness of falling into the local minimum point [9]. Therefore improved training algorithms involving the LM algorithm and the GA-BP algorithm are proposed.

3.1 BACKPROPAGATION ALGORITHM

The BP algorithm is a learning method which minimizes the error of the neural network output compared to the required output. In this algorithm, the performance index $F(w)$ to be minimized is defined as the sum of squared errors between the target outputs and the network's simulated outputs, namely:

$$F(w)=e^T e, \quad (1)$$

where $w = [w_1, w_2, \dots, w_n]$ consists of all weights of the network, e is the error vector comprising the errors for all training samples. The steps involved in training a neural network using the BP algorithm are as follows [8]:

step 1. Initialize each w_i to some small random value.

step 2. Do the step 3 until the termination condition is met.

step 3. For each training sample $\langle (x_1, x_2, \dots, x_n), t \rangle$, where t is the target output, do

1) Input the instance (x_1, x_2, \dots, x_n) to the network and compute the network outputs o_k .

2) For each output unit k , $\delta_k = o_k(1 - o_k)(t_k - o_k)$.

3) For each hidden unit h , $\delta_h = o_h(1 - o_h) \sum_k w_{h,k} \delta_k$

4) For each network weight, do $w_{i,j} = w_{i,j} + \Delta w_{i,j}$, where $\Delta w_{i,j} = \eta \delta_j x_{i,j}$, η is the learning rate defined by users. The algorithm is assumed to have converged when the norm of the gradient is less than some predetermined value, or when the error has been reduced to some error goal.

3.2 LEVENBERG-MARQUARDT ALGORITHM

As a kind of Gradient-based training algorithms, the BP algorithm is not efficient because the gradient vanishes at the solution. The neural networks are allowed to learn more subtle features of a complicated mapping by using Hessian-based algorithms, whose training process converges quickly as the solution is approached due to the fact that the Hessian does not vanish at the solution [10]. To benefit from the advantages of Hessian based training, the Levenberg-Marquardt algorithm is used to train the neural network for environmental quality analysis.

When training with the LM algorithm, Δw , the increment of weights can be obtained as follows:

$$\Delta w = \left[J^T(w)J(w) + \mu I \right]^{-1} J^T(w)e, \quad (2)$$

where $J(w)$ is the Jacobian matrix, $J^T(w)J(w)$ is referred as the Hessian matrix, I is the identity matrix, and μ is the learning rate which is to be updated depending on the

outcome. In particular, μ is multiplied by the decay rate β ($0 < \beta < 1$) whenever $F(w)$ in Equation (1) decreases, whereas μ is divided by β whenever $F(w)$ increases in a new step.

The standard LM training process is illustrated in the following steps.

step 1. Initialize the weights and parameter μ .

step 2. Compute the sum of the squared errors over all inputs $F(w)$.

step 3. Until the termination condition is met, do

1) Solve Equation (2) to obtain Δw , the increment of weights.

2) Use $w + \Delta w$ as the trial w , and judge

if trial $F(w + \Delta w) < F(w)$ in step 2 then $w = w + \Delta w$, $\mu = \mu\beta$

Go back to step 2

else

$$\mu = \mu / \beta$$

Go back to step 3.1.

In the LM algorithm the parameter μ is adjusted automatically at each iteration in order to secure convergence. This algorithm becomes Gauss-Newton method for $\mu = 0$ and becomes steepest decent or the error backpropagation algorithm when μ is very large.

3.3 HYBRID GA-BP ALGORITHM

From mathematical point of view, the BP learning is a non-linear optimization problem and there exists the local minimum point inevitably. The genetic algorithm, abbr. GA, is a highly parallel, stochastic and adaptive optimization technique based on biological genetic evolutionary mechanisms [11]. GA has good global search performance since its search has always been throughout the solution space, while the BP learning algorithm is more effective in local search. Therefore, the hybrid training algorithm combining BP with GA can achieve the goal of network optimization successfully.

GA can be used to optimize the network topology and weights [12]. In this work it is used to optimize the network weights before the BP learning. The optimization problem is described as follows: $\min(e) = f(w_1, w_2, \dots, w_n)$, where w_1, w_2, \dots, w_n are all network weights satisfying $-1 < w_i < 1$, $1 \leq i \leq n$, and n is the number of total weights.

GA simulates the process of natural evolution, performing operations similar to natural selection, crossover and mutation to obtain the final optimization result after repeated iterations. The weight optimization process of the neural network using GA is described in the following steps.

step 1. Start with a randomly generated population comprising N chromosomes, and each chromosome encodes a set of network weights. Set the maximum generation number gen , the crossover probability P_c and the mutation probability P_m .

step 2. If the maximum generation number gen has been reached, go to step 6.

step 3. Calculate the fitness of each individual

chromosome by the fitness formula: $f_i = 1/E_i$, where f_i presents the fitness of chromosome i and E_i presents the MSE, i.e. mean square error of the neural network corresponding to the chromosome i .

step 4. Repeat the following steps until the new population size reaches N .

1) Use the roulette wheel strategy to select a pair of chromosomes from the current population for single-point crossover and mutation with probability of P_c and P_m respectively. For the mutation operator, add a random value between $(-1, 1)$ to each mutation point.

2) Put the new pair of chromosomes in the new population.

step 5. Take the new N chromosomes as the new population. Go to step 2.

step 6. Decode the chromosome with the highest fitness to obtain initial weights of the neural network.

After global optimization with GA, the BP algorithm is used for the local search until the termination criterion is satisfied.

4 A case study

4.1 DESIGN OF NEURAL NETWORK TOPOLOGY

As mentioned in section 2, environmental quality assessment is a pattern recognition problem. The objective of the neural network is to give an accurate comprehensive quality assessment according to environmental input vector. The critical step in building a robust neural network is to create an architecture, which should be as simple as possible and has a fast capacity for learning the data set. The network topology are typically best determined empirically.

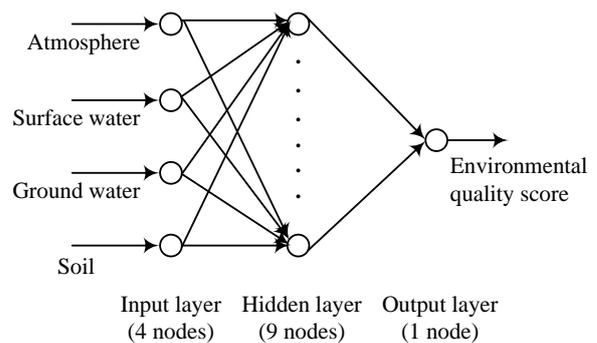


FIGURE 1 Neural network topology for environmental quality assessment

In this study, the selected environmental variables are atmosphere (x_1), surface water (x_2), ground water (x_3) and soil (x_4). Therefore the input layer consists of four nodes representing components of the four-dimensional input vector $X = (x_1, x_2, x_3, x_4)$. The number of nodes in the hidden layer usually is not less than the number of input nodes. Based on Kolmogorov theory, $2N + 1$ hidden nodes should be used for one hidden layer, where N is the number of input nodes. Considering the three-layered topology with one hidden layer, we will have nine hidden

nodes because of four input nodes. The output layer contains a single node giving the score of environmental quality. The architecture of the 4-9-1 feedforward neural network used in this analysis is shown in Figure 1.

4.2 TRAINING OF NEURAL NETWORK

Three learning algorithms of BP, LM and GA-BP described in section 3 together with the neural network architecture presented Figure 1 are used to create, train and test the neural network for environmental quality assessment. The neural networks are implemented using MATLAB 7 (MathWorks, USA) software with its neural network toolbox.

Table 1 tabulates the samples which are from 20 different monitoring points in a certain region of China [13, 14]. Depending on the actual condition of China, environmental quality is divided into four classes of A, B, C and D [13]. An “A” represents high environmental quality without pollution, “B” represents good quality with light pollution, “C” represents ordinary quality with moderate pollution, and “D” signifies a severe pollution. There is a score scope per output pattern, that is, A-[0, 1], B-[1, 2.5], C-[2.5, 5] and D-[5, 10]. To help speed up the

learning process, the input values for each attribute have been scaled so as to always fall within a specified range [0.1, 1] by the following Equation:

$$F_j = \frac{x_j - x_{j\min}}{x_{j\max} - x_{j\min}} \times 0.9 + 0.1, \text{ where } F_j \text{ is the normalization}$$

of the attribute value x_j , $x_{j\min}$ is the minimum value and $x_{j\max}$ is the maximum value of the j^{th} attribute.

In this analysis, the first ten samples in Table 1 have been taken for the training set and the rest for the test set. The transfer function between the input layer and the hidden layer is the tangent Sigmoid function *tansig*, and the one between the hidden layer and the output layer is the linear function *purelin*. The performance curves of the BP, LM and GA-BP networks are shown as Figures 2-5. From Figure 2 we can see that the BP network converges to the preset precision after 15142 epochs, while from Figure 3 the LM network only needs 3 epochs, greatly faster than the BP network. For the GA-BP network we set the initial population size N to 30, the maximum generation number *gen* to 800, the crossing probability P_c to 0.95 and the mutation probability P_m to 0.09. Figure 4 shows the curves of MSE, i.e. Mean Square Error and the fitness in the first process of weight optimization by GA.

TABLE 1 Normalized data set in environmental quality assessment

| No. | Atmosphere | Surface water | Ground water | Soil | Assessment score | Environmental quality rating |
|-----|------------|---------------|--------------|--------|------------------|------------------------------|
| 1 | 0.1010 | 0.1000 | 0.2171 | 0.1130 | 0.9 | A |
| 2 | 0.1000 | 0.1024 | 0.2220 | 0.1130 | 0.9 | A |
| 3 | 0.2173 | 0.2156 | 0.2951 | 0.1000 | 1.8 | B |
| 4 | 0.2006 | 0.1094 | 0.2366 | 0.1097 | 1.6 | B |
| 5 | 0.2168 | 0.1496 | 0.1000 | 0.1173 | 1.7 | B |
| 6 | 0.2284 | 0.2864 | 0.9000 | 0.1336 | 2.2 | B |
| 7 | 0.6198 | 0.1755 | 0.4220 | 0.1227 | 4.4 | C |
| 8 | 0.3554 | 0.2864 | 0.5732 | 0.1195 | 2.9 | C |
| 9 | 0.6082 | 0.1330 | 0.5878 | 0.4735 | 5.2 | D |
| 10 | 0.5329 | 0.5224 | 0.3439 | 0.9000 | 5.9 | D |
| 11 | 0.4322 | 0.2628 | 0.2463 | 0.1217 | 3.3 | C |
| 12 | 0.2067 | 0.1590 | 0.2951 | 0.1152 | 1.7 | B |
| 13 | 0.9000 | 0.5224 | 0.3537 | 0.1314 | 6.6 | D |
| 14 | 0.5880 | 0.3336 | 0.2902 | 0.1758 | 4.5 | C |
| 15 | 0.2775 | 0.1496 | 0.2951 | 0.1173 | 2.1 | B |
| 16 | 0.5243 | 0.1850 | 0.7049 | 0.1693 | 4.1 | C |
| 17 | 0.2896 | 0.1519 | 0.3098 | 0.1227 | 2.2 | B |
| 18 | 0.5637 | 0.9000 | 0.7146 | 0.3349 | 5.4 | D |
| 19 | 0.3073 | 0.1684 | 0.8951 | 0.1520 | 2.7 | C |
| 20 | 0.5030 | 0.7442 | 0.5878 | 0.1671 | 4.5 | C |

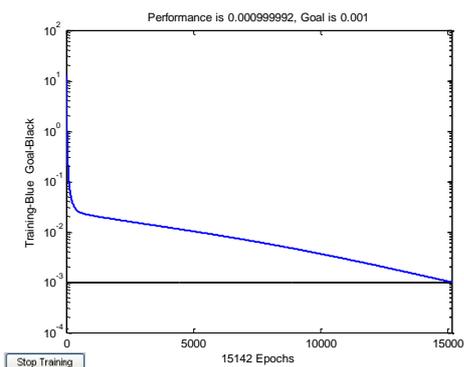


FIGURE 2 Gradient per epoch with the BP algorithm

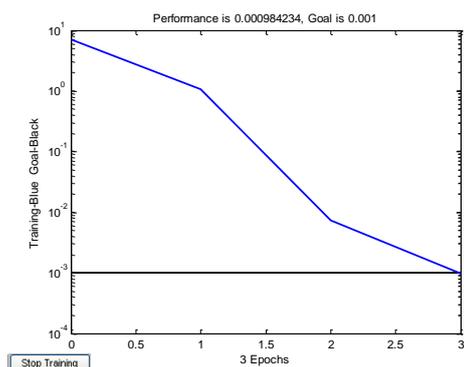


FIGURE 3 Gradient per epoch with the LM algorithm

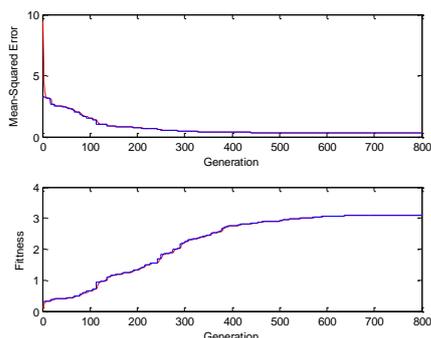


FIGURE 4 MSE and fitness curves of GA optimization

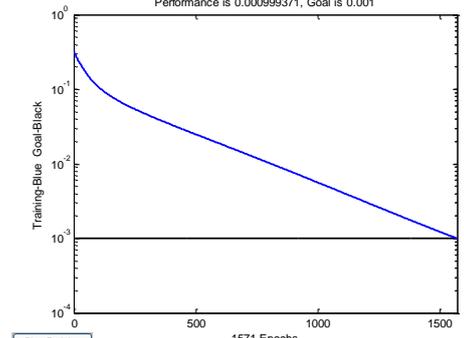


FIGURE 5 Gradient per epoch with the GA-BP algorithm

It can be seen that the fitness of chromosomes tends to stabilize after 500 generations. By this time the error of the GA-BP network reaches to 0.3. In the second process of network training the BP algorithm is used with the error goal of 0.001 and the learning rate of 0.01. The GA-BP network needs 1571 epochs to converge as shown in Figure 5 and it is much faster than the BP network but still slower than the LM network.

4.3 TEST RESULTS AND COMPARISONS

Table 2 tabulates the test results of three neural networks trained by BP, LM and GA-BP algorithms. All test cases are labelled with corresponding quality classes by predictive values. Table 3 tabulates the maximum, minimum and average relative errors and final classification accuracy obtained according to results in Table 2.

Average relative errors of environmental quality prediction of the BP, LM and GA-BP networks are attained 9.05%, 6.13% and 6.12% respectively as tabulated in Table 3. These three neural networks are all effective in assessing environmental quality since their predictive accuracies are all within 10%. The LM and GA-BP networks are both more accurate than the BP network in environmental quality prediction and the GA-BP network has the lowest average relative error.

Comparing quality ratings of three networks in Table 2 with actual quality ratings in Table 1, we can see that for both the LM and GA-BP networks the 18th sample with heavy pollution indicated by “D” is labelled with the wrong quality class of moderate pollution indicated by

“C”, while for the BP network the 14th, 18th and 19th samples are identified wrongly. The classification accuracies of the BP, LM and GA-BP models are 70%, 90% and 90% respectively. Therefore the LM and GA-BP networks are superior in final classification accuracy to the BP network.

Summarily, the LM and GA-BP networks are significantly better than the BP network with faster training speed and better generalization ability. The LM network has the fastest convergence speed, and the GA-BP network has the best performance in assessing environmental quality considering both predictive and final classification accuracies.

5 Conclusions and future works

Environmental quality assessment contributes to decision making in support of sustainable economic and social development, and it has attracted many research interests in the literature. Recent studies show that the neural network method has achieved better performance than the traditional statistical method. Due to the advantages that neural networks can learn a complex nonlinear relationship with limited prior knowledge and perform inferences for an unknown combination of input variables, this paper introduces the popular BP neural network and two other LM and GA-BP networks to environmental quality assessment in attempt to provide a model with good ability of generalization. Experimental results show that the assessment task is successfully accomplished by using neural networks.

TABLE 2 Test results for environmental quality

| No. | BP algorithm | | | LM algorithm | | | GA-BP algorithm | | |
|-----|---------------|----------------|--------|---------------|----------------|--------|-----------------|----------------|--------|
| | Predict score | Relative error | Rating | Predict score | Relative error | Rating | Predict score | Relative error | Rating |
| 11 | 3.7983 | 15.10% | C | 3.0548 | 7.43% | C | 3.2515 | 1.47% | C |
| 12 | 1.6452 | 3.22% | B | 1.7213 | 1.25% | B | 1.7426 | 2.51% | B |
| 13 | 7.0073 | 6.17% | D | 6.0187 | 8.81% | D | 5.5179 | 16.4% | D |
| 14 | 5.1948 | 15.44% | D | 4.3065 | 4.30% | C | 4.2892 | 4.68% | C |
| 15 | 2.0484 | 2.46% | B | 2.1916 | 4.36% | B | 2.2184 | 5.64% | B |
| 16 | 3.5519 | 13.37% | C | 4.4872 | 9.44% | C | 4.0303 | 1.70% | C |
| 17 | 2.1185 | 3.70% | B | 2.2636 | 2.89% | B | 2.3291 | 5.87% | B |
| 18 | 4.9912 | 7.57% | C | 4.9802 | 7.77% | C | 4.7489 | 12.06% | C |
| 19 | 2.3236 | 13.94% | B | 2.9165 | 8.02% | C | 2.7036 | 0.13% | C |
| 20 | 4.9300 | 9.56% | C | 4.8181 | 7.07% | C | 4.0182 | 10.71% | C |

TABLE 3 Result comparisons of three neural networks

| | BP network | LM network | GA-BP network |
|-------------------------|------------|------------|---------------|
| maximum relative error | 15.44% | 9.44% | 16.4% |
| minimum relative error | 2.46% | 1.25% | 0.13% |
| average relative error | 9.05% | 6.13% | 6.12% |
| classification accuracy | 70% | 90% | 90% |
| epochs | 15142 | 3 | 1571 |

However, a weakness of the study is that the created models are relatively hard to explain because neural networks are more concerned about the actual number of variables rather about their nature. Therefore one future direction of the research is to improve interpretability of the neural network models for the problem of environmental quality assessment.

Acknowledgements

This work is supported by the Key Laboratory of Cloud Computing & Intelligent Information Processing of Changzhou City under Grant No. CM20123004 and Applied Basic Research Program of Jiangsu University of Technology under Grant No. KYY10059. Furthermore, we are indebted to the support and encouragements received from the staff and colleagues of School of Computer Engineering.

References

- [1] Environmental Terminology and Discovery Service (ETDS) 2014 *Environmental Quality* http://glossary.eea.europa.eu/EEAGlossary/E/environmental_quality-2014/16 May 2014
- [2] Wang H 2002 Assessment and prediction of overall environmental quality of Zhuzhou city Hunan province China *Journal of Environmental Management* **66**(3) 329-40
- [3] Zan H, Chen H 2004 Credit rating analysis with support vector machines and neural Networks: a market comparative Study *Decision Support Systems* **37**(4) 543-58
- [4] Fidele B, Cheeneebash J 2009 Artificial neural network as a clinical decision-supporting tool to predict cardiovascular disease *Trends in Applied Sciences Research* **4**(1) 36-46
- [5] Sekar B D, Dong M C, Shi J, Hu X Y 2012 *IEEE Sensors Journal* **12**(3) 644-50
- [6] Torres F B J, Cavalcanti G D C, Ren T I 2013 *IEEE Transactions on Cybernetics* **43**(6) 2082-92
- [7] Reznik L, Von Pless G, Al Karim T 2011 *IEEE Sensors Journal* **11**(3) 791-8
- [8] Graves A, Liwicki M, Fernandez S, Bertolami R, Bunke H, Schmidhuber J 2009 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(5) 855-68
- [9] Jin W, Li Z J, Wei L S, Zhen H 2000 The improvements of BP neural network learning algorithm *5th International Conference on Signal Processing Proceedings*, Beijing **3** 1647-9
- [10] Mishra D, Yadav A, Ray S, Kalra P K 2005 Levenberg-Marquardt learning algorithm for integrate-and-fire neuron model *Neural Information Processing* **9**(2) 41-51
- [11] Yuan J, He C, Gao W, Lin J, Yu P 2014 A novel hard decision decoding scheme based on genetic algorithm and neural network *Optik-International Journal for Light and Electron Optics* **125**(14) 3457-61
- [12] Negnevitsky M 2004 Artificial intelligence-A guide to intelligent systems *Addison-Wesley Educational Publishers Inc*
- [13] Yong C, Jian L, Wu C 2009 Environmental quality evaluation based on SVM *Computer Engineering and Applications* **45**(2) 209-11
- [14] Hou Kefu 1992 Environmental system engineering *Beijing Institute of Technology Press (in Chinese)*

| Authors | |
|--|--|
|  | <p>Yijun Liu, born in June, 1978, Changzhou, Jiangsu, China</p> <p>Current position, grades: the lecturer of School of computer engineering, Jiangsu University of Technology, China. University studies: M.Sc. from Nanjing University in China. Scientific interest: machine learning, data mining and intelligent information system. Publications: 10 papers. Experience: teaching experience of 9 years, 4 scientific research projects.</p> |
|  | <p>Sheng He, born in November, 1971, Zongyang, Anhui, China</p> <p>Current position, grades: the associate professor of school of computer engineering, Jiangsu University of Technology, China. University studies: Ph.D. from Jiangnan University in China. Scientific interest: data mining and bioinformatics. Publications: 10 papers. Experience: teaching experience of 15 years, 4 scientific research projects.</p> |
|  | <p>Yao Wang, born in April, 1982, Changzhou, Jiangsu, China</p> <p>Current position, grades: the lecturer of school of computer engineering, Jiangsu University of Technology, China. University studies: M.Sc. from Yangzhou University in China. Scientific interest: computer network and intelligent information system. Publications: 5 papers. Experience: teaching experience of 10 years, 3 scientific research projects.</p> |
|  | <p>Xiumei Wang, born in October, 1971, Xuanhua, Hebei, China</p> <p>Current position, grades: the lecturer of school of computer engineering, Jiangsu University of Technology, China. University studies: B.Sc. in computer science and technology from Langfang Teachers University in China, M.Sc. from Brooklyn College, CUNY in USA. Scientific interest: computability theory, dynamical systems. Publications: 3 papers. Experience: teaching experience of 10 years, 2 scientific research projects.</p> |

Design of Q450 pellet molding machine and force analysis of its molding assembly based on Solidworks

Xiangyue Yuan, Zhongjia Chen*

School Of Technology, Beijing Forestry University, Beijing 100083, China

Received 6 June 2014, www.tsi.lv

Abstract

Energy shortage and environmental pollution is a common serious problem restricting the development of world economy and the society. Biomass energy has become the fourth major energy resource after oil, coal, natural gas energy for the good properties of green, clean, and renewable. So it is important to research biomass energy technology to solve the energy crisis and environmental protection. The technology of biomass densification is a simple solution to make the biomass resource become low cost and high value. In this paper, a new kind of biomass pellet molding mechanism had been deeply studied, and the pellet molding machine, Q450, was designed by the CAD/CAE, Solidworks. The load conditions of three molding assemblies fixed inside the enclosure bodies had also been studied and analysed on the designed machine. Then a method, named FEA (Finite Element Analysis), was conducted to research the mechanical properties of enclosure assembly for the pellet machine in Simulation. Through analysis, the results were obtained that the maximum stress and displacement of enclosure bodies were separately 31.37Mpa and 7.583e⁻²mm, which could provide the reliable strength and stiffness to the enclosure assembly. It convincingly ensured that Q450 pellet molding machine had enough reliability and security.

Keywords: pellet molding mechanism with plunger-roller ring die, Q450 pellet molding machine, enclosure assembly, strength and stiffness with FEA (Finite Element Analysis)

1 Introduction

At present, energy shortage and environmental pollution is a common serious problem restricting the development of world economy and the society. So it is necessary to develop renewable and green energy for the future, which means a low-carbon, sustainable and scientific development [1].

Bio-energy is a kind of energy carried in organic biological resources. The plants absorb solar energy through photosynthesis, convert CO₂ from the atmosphere into fixed organic compounds, and then the solar energy was turned into biomass energy. Compared with fossil fuels, the biomass burning pollutants such as SO₂, NO_x, and dust are much smaller attributed to its good character of low sulfur, nitrogen, high carbon activity, high volatile components and less ash content [2], which would obviously decrease phenomenon of air pollution and acid rain. So the bio-energy is a kind of green, clean and renewable energy.

Bio-energy is widely distributed with a good renewability and richness in natural resources, but the greatest inadequacy lies in the too scattered distribution, the low original density, and low energy density [3], which severely limits the direct utilization of biomass energy. So technologies should be developed to convert biomass into a new form fuel with high unit density and energy density.

Biomass densification is an important kind of technologies to achieve the target of the biomass resource

utilization with low cost, high value and efficiency at present. The process of biomass densification is that compress biomass material physically at heating or room temperature into solid fuels shaped as columnar, cube, or pellet with high unit density after the pretreatment of drying, crushing and etc. [4]. The unit density and combustion value of the solid fuels after processing could respectively reach 0.8~1.4 t/m³ and 16~21 MJ/kg in terms of kinds of biomass materials such as agricultural straws, grasses, woody resources like wood chips and sawdust. Combustion efficiency could be as high as 90%, and the conversion cost is low as well [5].

Currently, there are many kinds of biomass molding machine, but an objective reality that the high power consumption and quick wear of main molding dies is still existed restricting the development of biomass solid fuels. To overcome shortcomings in traditional molding machines, a completely new kind of biomass pellet molding mechanism with plunger-roller ring die was designed after deeply studied the densification of biomass material in this paper, which effectively avoided the extrusion and friction of biomass material out of molding cavities. It would greatly reduce the power consumption of densification (about 37%-40%), and prolong the service life of forming molds.

* *Corresponding author* e-mail: chenzhongjia@bjfu.edu.cn.

2 Biomass pellet molding mechanism with plunger-roller ring die

The biomass pellet molding mechanism with plunger-roller ring die mainly consisted of plunger-roller

component and ring die component, shown in Figure1. Biomass materials were densified into solid pellets by engaging movement of plungers on plunger-roller and forming cavities which were straight through-holes in ring die.

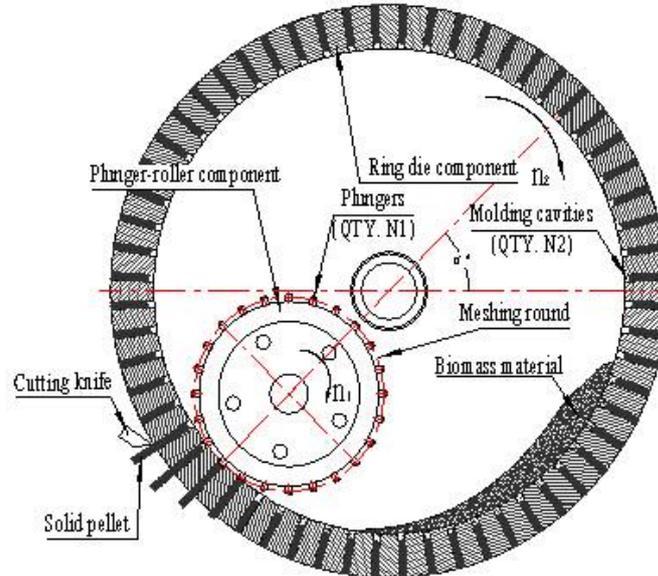


FIGURE 1 The working principle of biomass pellet molding mechanism plunger-roller ring die

Please read through the following sections for more information on preparing your paper. However, if you use the template you do not have to worry about setting margins, page size, and column size etc. as the template already has the correct dimensions.

All of the plungers and forming cavities were uniformly distributed along circumference. Let the quantity of plungers and forming cavities of each row separately be N_1 and N_2 , the revolving speed of plunger-roller and ring die be n_1 and n_2 in turn. Then when the four parameters matched the following Equation (1), the movement between the plungers and forming cavities was similar to the meshing motion of the gear.

$$\frac{N_2}{N_1} = \frac{n_1}{n_2} = \left(\frac{Z_2}{Z_1} : gear \right), \quad (1)$$

where, N_2 and N_1 were the number of plungers on the plunger-roll and forming holes in the ring die respectively. Z_2 and Z_1 meant the number of gear pair, imitated by the molding part of the pellet machine.

The plunger-roller component and ring die component were homodrumy, eccentric mounted in the enclosure bodies, and the angle of their centre line and horizontal direction was α° . When plungers and couple molding cavities rotated to eccentric forming zone, they had the approximately same instantaneous tangential velocities, but different radial velocities due to tangency. So the plungers and molding cavities had a relative velocity at the

stage of engaging, that is, radial extrusion speed for biomass materials. When shattered biomass material filled into cavities due to the action of gravity and centrifugal force, it could be densified through these molding cavities by couple plungers, finally be converted into cylindrical solid pellets, and then be cut off by knife. The process of densification was similar to many pistons extrusion forming molds at the same time, so the designed pellet molding machine in view of meshing motion had a higher productivity and better densification quality than conventional ring die molding mechanism. The biomass material wouldn't gather in the eccentric forming zone because of the angel of α° , which should ensure the meshing plungers only compressing biomass material in forming cavities. There was a gear transmission system between the plunger-roller component and ring die component, guaranteeing their same direction rotation movement with certain speed ratio, shown in the following Figure 2. In order to make sure the better biomass densification of the molding cavities in ring die, the diameter (D) and length (L) of through-hole must maintain a relationship shown in the Equation (2), known from reference [6]:

$$L / D = 5.2 \sim 6.2, \quad (2)$$

where, L and D indicated separately the diameter and length of the through-hole in the ring die.

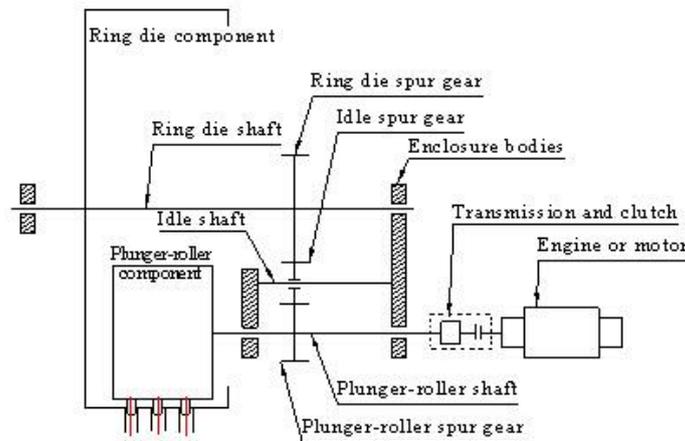
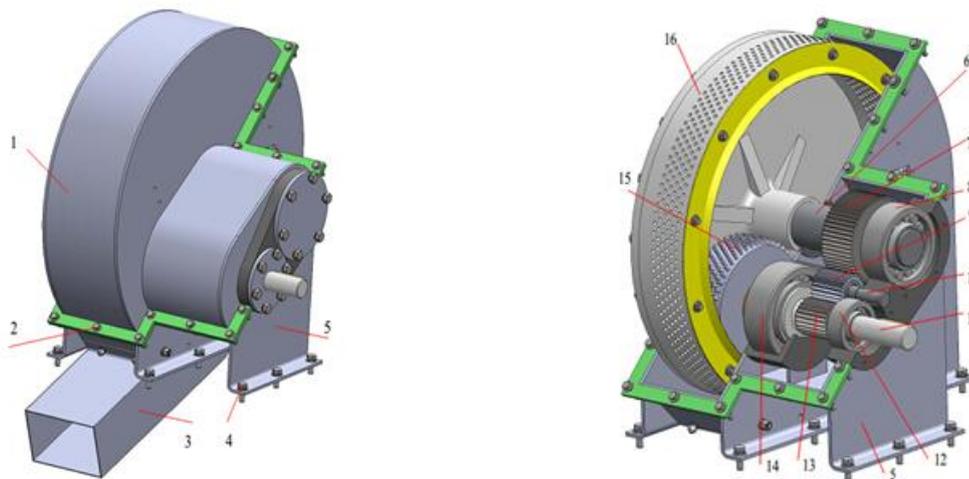


FIGURE 2 The transmission system of biomass pellet molding mechanism with plunger-roller ring die

3 Q450 pellet molding machine

Q450 pellet molding machine was a new type of biomass forming equipment designed with plunger-roller ring die, whose productivity was 450kg/h. It mainly consisted of

three parts, enclosure assembly, outlet hopper, and internal molding assemblies (including plunger-roller assembly, ring die assembly, and idler transmission assembly), seen obviously in Figure 3.



(a) Appearance of the molding machine

(b) The internal structure of the molding machine

FIGURE 3 Q450 pellet molding machine

1. Upper enclosure body, 2. Enclosure body connecting bolts, 3. Outlet hopper, 4. Molding machine fixing bolts, 5. Lower enclosure body, 6. Ring die shaft, 7. Ring die spur gear, 8. Ring die cylindrical roller bearings, 9. Idle spur gear, 10. Idle shaft, 11. Plunger-roller shaft, 12. Right plunger-roller cylindrical roller bearings, 13. Plunger-roller spur gear, 14. Left plunger-roller cylindrical roller bearings, 15. Plunger-roller component, 16. Ring die component

The plunger-roller assembly included plunger-roller component 15, plunger-roller shaft 11, spur gear 13, and two cylindrical roller bearings 12 and 14 (product model was NJ2316E). Plunger-roller component and plunger-roller spur gear were connected to the shaft with flat keys, and fixed in bearing seats of enclosure bodies by bearings. The ring die assembly also included ring die component 16, ring die shaft 6, spur gear 7, two cylindrical roller bearings 8 (product model respectively was NJ2319E and NJ2310E), and its mode of connection and installation was same with plunger-roller assembly. For the idler transmission assembly, idle spur gear 9 was fixed on idle shaft 10 by two bearings (product model was NJ2205E),

and both ends of shaft were installed in bearing seats, which made the idle gear rotate around the shaft freely. The driving force was input to plunger-roller shaft, and then the ring die component was driven to rotate in a same direction with plunger-roller component at certain speed ratio by idle transmission assembly.

4 Design of Enclosure assembly for Q450 pellet molding machine

Figure 4 was the enclosure assembly of Q450 pellet molding machine designed by Solidworks. The meanings of the numbers were shown in the following.

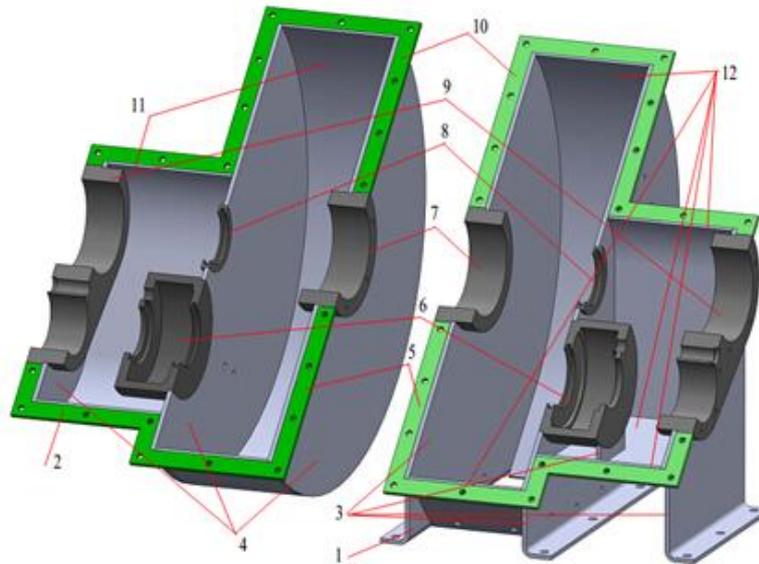


FIGURE 4 Structure of enclosure assembly for Q450 pellet molding machine

1. Lower enclosure body, 2. Upper enclosure body, 3. Vertical plates of lower enclosure body, 4. Vertical plates of upper enclosure body, 5. Connecting flanges (down), 6. Bearing seats of plunger-roller assembly, 7. Bearing seats of ring die assembly, 8. Dustproof felt seats, 9. Integrate bearing seats, 10. Connecting flanges (top), 11. Side plates of upper enclosure body, 12. Side plates of lower enclosure body

From Figure 4, the enclosure assembly was consisting of lower enclosure body 1 and upper enclosure body 2, which were welded together with steel plates. The thickness of all vertical plates and connecting flanges was 8mm, and the one of all side plates was 5mm. To facilitate the installation of internal molding assemblies, 45 split structure was adopted in the whole enclosure assembly. And the material of all steel plates employing in enclosure bodies was Q235-BZ (the mechanical properties were

shown in Table 1, which has a better welding performance. The three pairs of bearing seats 6, 7 and 9 always enduring large alternating loads were made of 45 steel possessing good synthesized mechanical properties, and the dustproof felt seats 8 was made of 20 steel good for welding because of low carbon content. The two enclosure bodies were manual welded with E4303 electrodes, and finally fixed together with M16 bolts.

TABLE 1 The mechanical properties of Q235-BZ [unit: Mpa]

| Tensile strength σ_B | Yield strength σ_S | Bend fatigue strength σ_{-1} | Shear fatigue strength τ_{-1} | permissible bend stress $[\sigma_{-1}]$ |
|-----------------------------|---------------------------|-------------------------------------|------------------------------------|---|
| 400-420 | 225 | 170 | 105 | 40 |

5 Force analysis of molding assemblies

The force distributions of three molding assemblies were shown in Figure 5 and the data calculated listed in Table 2.

TABLE 2 The force values of three molding assemblies

| Plunger-roller assembly | | Ring die assembly | | Idle assembly | |
|-------------------------|------------|-------------------|------------|---------------|----------|
| T_g | -130.242 | T_h | +17.377 | F_{r1} | +87.841 |
| R_a | +11258.133 | F_f | -21191.500 | F_{t1} | -241.341 |
| G_g | -293.050 | G_h | -1519.100 | F_{r2} | -70.273 |
| G_{gz} | -163.635 | G_{hz} | -212.800 | F_{t2} | +193.073 |
| F_{r1} | -87.841 | F_{r2} | +70.273 | G_3 | -20.990 |
| F_{t1} | +241.341 | F_{t2} | +193.073 | G_{dz} | -7.100 |
| G_1 | -12.210 | G_2 | -123.578 | | |

Note: The unit of T_g , T_h , M was [N.m], and $M=-T_g$. The unit of rest items was [N], “+” represented the direction of force was identical to the load distribution coordinate system (d), “-” conversely, and the same to Table 3.

From the force analysis in Table 2, support reactions of three pairs of seats were worked out (data listed in Table 3).

TABLE 3 The reactions of support seats [N]

| | 1 | 2 | 3 | 4 | 5 | 6 |
|-------|-----------|-----------|------------|-----------|---------|---------|
| F_x | -112.282 | -130.060 | -41.809 | -151.264 | +32.742 | +15.527 |
| F_y | +3628.451 | +7629.683 | +15019.920 | +6101.307 | -11.917 | -5.652 |
| F_G | +438.415 | +30.480 | +1610.399 | +245.080 | +18.994 | +9.007 |

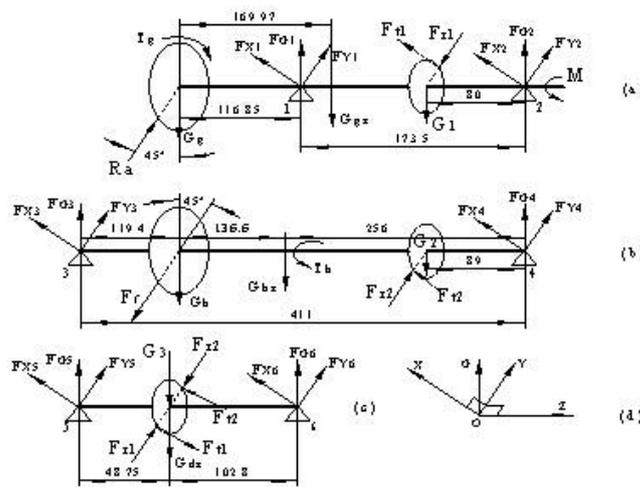


FIGURE 5 Force analysis of three molding assemblies (a)

Force analysis of plunger-roller assembly, (b). Force analysis of ring die assembly, (c). Force analysis of idle transmission assembly, (d). Load distribution coordinate system

Note: To plunger-roller assembly, T_g , M expressed respectively total resistance torque, driving torque. R_a^{total} , G_g was radial force and gravity in turn. G_1 was the gravity of roller gear and F_{r1} , F_{t1} indicated separately the radial force and tangential force of the gear. To ring die assembly, T_h was resistance torque. F_f^{total} , G_h was radial friction force and gravity in turn. G_2 was the gravity of ring die gear and F_{r2} , F_{t2} indicated separately the radial and tangential force of the gear. G_3 was gravity of idle gear and idler bearing, and G_{gz} , G_{hz} , and G_{dz} respectively meant gravities of shaft of the roller, ring die and idler.

6 FEA (Finite Element Analysis) of enclosure assembly

As the important installation carrier of the pellet molding machine, enclosure body must have the properties of sufficient strength and rigidity to ensure the right position of each forming mechanism in the work process and so as to keep normal engage compression of the plunger pair.

Simulation is a integrate FEA module of Solidworks. The designed 3D models for product could be conveniently executed FEA in this CAE module, and a FEA outcome of enclosure assembly for Q450 pellet molding machine was shown in Figure 6. In the static analysis, the bottom faces of vertical plates of lower enclosure body were fixed, and the loads (listed in Table 3) were applied on support seats of enclosure bodies.

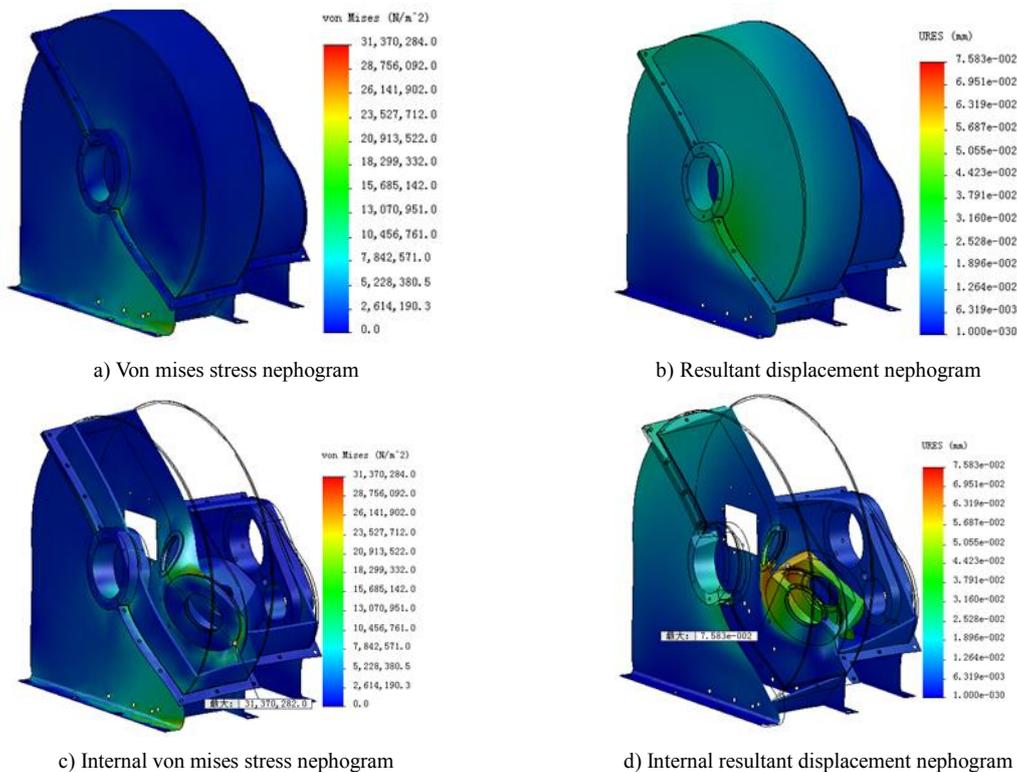


FIGURE 6 FEA result of enclosure assembly

As it is seen from Figure 6c and 6, the maximum stress σ_{\max} and displacement UR_{\max} occurred respectively at the oblique lower and upper position of the left bearing seats of plunger-roller, and the value was 36.44 Mpa and 0.07583mm in turn, less than $[\sigma_{-1}]$ and $[U]$ in the following Equation (3).

$$\begin{aligned} \sigma_{\max} &< [\sigma_{-1}] = 40MPa \\ UR_{\max} &< [U] = 0.15mm \end{aligned} \quad (3)$$

where, $[U]$ was the middle class tolerance value of centre distance (OA=198mm) between molding ring die and plunger-roller.

Known from the Equation (3), strength and stiffness of molding enclosure body respectively matched the demands of working, which ensured the right position when compressing biomass materials by molding mechanism and the accurate engagement between plunger-roller and through-holes on ring die.

7 Conclusion

In the paper, a new kind of biomass pellet molding machine, Q450, was designed to convert biomass

materials into bio-energy efficiently, and the principles of the key three components in the machine were also introduced in detail.

Then force analyses of molding assemblies were conducted, and FEA was processed on the basis of the data from force analyses. According to the FEA result, the maximum stress and displacement were respectively 31.37Mpa and 7.583e-2 mm, which distinctly less than the permissible bend stress (40Mpa) of Q235-BZ and permissible deformation (0.15mm, the Middle Class tolerance of eccentric distance of plunger-roller shaft and ring die shaft which was 198mm). So the whole enclosure assembly had sufficient strength and stiffness to ensure enough reliability and security during working for Q450 pellet molding machine.

Acknowledgments

The financial support of the Fundamental Research Funds for the Central Universities (No. YX2013-26) is gratefully acknowledged.

References

- [1] Tumuluru J S, Wright C T, Hess J R 2011 Biomass densification systems to develop uniform feedstock commodities for bioenergy application *Landtechnik* **54**(1) 28–35
- [2] Dogherty M J 1989 A Review of the mechanical Behaviour of Straw When Compressed to high Densities *Agric Engng Res* **5** (44) 243–85
- [3] Junginger M, Bolkesj T, Bradley D 2008 Developments in international bioenergy trade *Biomass and Bioenergy* **3**(2) 717–29
- [4] Werther J, Saenger M, Hartge E U 2000 Combustion of agricultural residues *Progress in Energy and Combustion Science* **26**(1) 1–27
- [5] Mielenz J R 2001 Combustion of agricultural residues *Current Opinion Inmicrobiology* **1**(4) 324–9
- [6] Song X, Yu G, Jiang C, Du K, Guo X 2011 Parameter Research of Bio-mass Shaping with Open Lineal Mold in Natural Temperature *Heilongjiang Agricultural Sciences* **11** 36–8 (in Chinese)
- [7] Song X 2012 *Parameter Study of Biomass Pellet Making Machine with Rolling Mechanism in Nature Condition* Beijing: Beijing Forestry University (in Chinese)
- [8] Yu G, Song X, Li Z, Zhao X 2011 *Biomass molding machine with plunger-roller* China Patent No. CN102320151A (in Chinese)
- [9] Wang J, Yuan X 2011 A variable diameter conical head of a biomass fuel molding machine at room temperature. *Journal of Beijing Forestry University* **33**(3) 119–21 (in Chinese)
- [10] Yan W 2011 *Research on Biomass Open Compaction at Normal Temperature* Beijing: Beijing Forestry University (in Chinese)

| Authors | |
|---|--|
|  | <p>Xiangyue Yuan, born in September, 1976, Beijing, China</p> <p>Current position, grades: Lecturer at School of Technology, Beijing Forestry University, China. University studies: Doctor's degree from majored in mechanical design and theory in Beijing Forestry University. Scientific interest: development of bio-energy equipment and forestry machinery. Publications: published more than 10 research papers, 5 Chinese patents. Experience: teaching experience of 6 years, 2scientific research projects.</p> |
|  | <p>Zhongjia Chen, born in January, 1981, Beijing, China</p> <p>Current position, grades: Lecturer of School of Technology, Beijing Forestry University, China. University studies: Doctor's degree from majored in mechanical design and theory in Beijing Forestry University. Scientific interest: development of bio-energy equipment. Publications: published more than 8 research papers, 5 Chinese patents. Experience: teaching experience of 4 years, 2 scientific research projects.</p> |

Model driven testing distributed environment monitoring system

Yan Li¹, Zhe Zhang^{2*}, Guihong Jiang¹, Xiaofeng Cui¹

¹*School of Computer, Shandong University of Technology, Zibo 255049, Shandong Province, China*

²*Software College, Nanyang Normal University, Nanyang 473000, Henan Province, China*

Received 6 October 2013, www.tsi.lv

Abstract

Distributed environment monitoring system is more and more widely used, especially the design and verification of embedded system in environmental monitoring is the guarantee of successful use of environmental monitoring. In this paper we demonstrate how test-case prioritization can be performed with the use of model-checkers. For this, different well known prioritization techniques are adapted for model-based use. New property based prioritization techniques are introduced. In addition it is shown that prioritization can be done at test-case generation time, thus removing the need for test-suite post-processing. Several experiments for embedded systems are used to show the validity of these ideas.

Keywords: test case prioritization, software testing, model checking, property testing

1 Introduction

In today's society, environment monitoring system is being paid more attention. Distributed environment monitoring system is more and more widely used [1], especially the design and verification of embedded system in environmental monitoring is the guarantee of successful use of environmental monitoring.

It has been shown [2, 3] that the order in which the test-cases of a test-suite are executed has an influence on the rate at which faults can be detected. In this paper we demonstrate how test-case prioritization can be performed with the use of model-checkers. As common prioritization techniques are based on program source-code, these techniques have to be adapted to the model-based setting. In addition, new property based prioritization techniques made possible by the use of model-checkers are introduced.

Obviously, a model-checker based method to test case prioritization is a useful addition to model-checker based test-case approaches. We therefore show how prioritization can be done at test-case generation time when using model-checkers to create test-cases. That way, no post-processing of the test-suites is necessary while still achieving an improved fault detection ratio of the resulting test-suite. The ideas described in this paper are illustrated using several example applications.

This paper is organized as follows: Section 2 recalls the principles of test-case prioritization and presents different prioritization techniques for model-based use. Then, Section 3 describes how prioritization is performed with the help of a model-checker, while Section 4 shows how prioritization can be done at test-case creation time. Section 5 describes our experiment for several security-

critical embedded systems and presents the results achieved. Finally, Section 6 concludes the paper.

2 Test case prioritization

Test-case prioritization is the task of finding an ordering of the test-cases of a given test-suite such that a given goal is reached faster. The test-case prioritization problem is defined by Rothermel et al. [4] as follows:

Given: T , a test-suite; PT , the set of permutations of T ; f a function from PT to the real numbers.

Problem: Find $T' \in PT$ such that

$$(\forall T'')(T'' \in PT)(T'' \neq T')[f(T') > f(T'')] \quad (1)$$

PT is the set of all possible orderings of T , and f is a function that yields an award value for any given ordering it is applied to. f represents the goal of the prioritization. For example, the goal might be to reach a certain coverage criterion as fast as possible, or to improve the rate at which faults are detected. There are different test-case prioritization techniques that can be used to achieve such goals.

Several different prioritization methods have been discussed in previous works [5, 8]. These methods are generally based on the source code of a program, e.g., the coverage of statements or functions. In contrast, when using a model-checker to determine prioritization we base the techniques on a functional model of the program to test. This section does not provide a complete overview of all available prioritization techniques but selects a representative subset that can be used to illustrate the usefulness of model-checkers in the prioritization process. In addition, the use of a model-checker allows new kinds

* *Corresponding author* e-mail: jdd35@163.com

of prioritization techniques which are introduced in this section.

2.1 TOTAL COVERAGE PRIORITIZATION

There are several code-based prioritization methods that sort test-cases by the number of statements or functions they cover. Model-checker based testing allows the formulation of coverage criteria as properties, as described in the next section. We therefore generalize from different code based methods to a coverage based method which is applicable to any coverage criterion expressible as a set of properties.

For example, the model-based coverage criterion Transition Coverage requires that each transition in an automaton model is executed at least once. Test-case prioritization according to transition coverage sorts test-cases by the number of different transitions executed.

2.2 ADDITIONAL COVERAGE PRIORITIZATION

Total Coverage Prioritization achieves that those test-cases with the biggest coverage are executed first. This does not necessarily guarantee that the coverage criterion is achieved as fast as possible. Additional coverage prioritization first picks the test case with the greatest coverage, and then successively adds those test-cases that cover the most yet uncovered parts.

2.3 TOTAL FEP PRIORITIZATION

This technique orders test cases by the ability to expose faults (fault exposing potential). Mutation analysis [6] is used to determine these values. For a given program a set of mutants is created by the application of a set of mutation operators. Each application of a mutation operator creates a mutant of the source code that differs from the original by a single valid syntactic change. The mutation score represents the ratio of mutants that a test-suite can distinguish from the original program. This mutation score can be calculated for each test-case separately, and then used as an award value for test-case prioritization. Total FEP prioritization uses the mutation score for a total sorting

2.4 ADDITIONAL FEP PRIORITIZATION

Similarly to additional coverage based prioritization test-cases can be sorted by the number of additional, yet undetected mutants. First the test-case with the highest mutation score is chosen, and then successively those test-cases are added that maximize the total number of detected mutants. Traditionally, this FEP based prioritization is computationally more complex than coverage based methods.

2.5 TOTAL PROPERTY PRIORITIZATION

This is a new technique made possible by the use of model-checkers. It is based on the idea of property relevance [7]. A test-case consists of values that are used as input data for the system under test, i.e., they represent the inputs the system receives from its environment. A test-case is said to be relevant to a requirement property if a property violation is possible when the input values are provided to an erroneous implementation. In practice, this can be determined by checking whether there is a mutant that can violate the property. A test-case can of course be relevant to more than one property. Total property prioritization sorts test-cases by the number of properties they are relevant to.

2.6 ADDITIONAL PROPERTY PRIORITIZATION

Similarly to the previous techniques, this method begins with the test-case that is relevant to the most properties and then successively adds test-cases that are relevant to yet uncovered properties.

2.7 HYBRID PROPERTY PRIORITIZATION

If the number of properties is significantly smaller than the number of test-cases, then a property based prioritization can quickly achieve property coverage. In general, the prioritization of the remaining test-cases starts again with the test-case with the highest award value. However, it is also conceivable to combine two different award functions. For example, it can be useful to sort test-cases totally based on the number of relevant properties, and then use a coverage prioritization as a secondary sorting method within test-cases of equal property relevance. We use transition coverage as secondary award value in our experiments.

2.8 RANDOM PRIORITIZATION

Random prioritization is interesting for evaluation of the different techniques. In average, any sorting method should achieve better results than random prioritization in order to be useful. We therefore use random prioritization as a lower bound for our analysis.

2.9 OPTIMAL PRIORITIZATION

The optimal prioritization sorts test-cases such that a given set of faults is detected with the minimum number of test-cases. This technique is not applicable in practice as it requires a-priori knowledge about the faults that are to be exposed. However, in experiments with known mutants it serves as upper bound for improvements that can be achieved with prioritization.

3 Using model checking to determine prioritization

In this section we show how the prioritization methods presented in the previous section can be performed in practice. As mentioned, we use model-checkers for prioritization. In order to do so it is necessary to reformulate test-cases as models, which allows analysis with regard to certain properties. This can be easily done by basing the transition relation of all variables on a special state-counting variable, as suggested by Ammann and Black [1]. As an example, assume a simple test-case

$$t = \{(x = 1, y = 0), (x = 0, y = 1), (x = 1, y = 1)\}.$$

Using the input language of the model-checker NuSMV [5] which we used for our experiments, the test-case can be expressed as:

```

MODULE main
VAR
  x: boolean;
  y: boolean;
  State: 0..2;

ASSIGN
  init(x):=1;
  next(x):= case
    State = 0: 0;
    State = 1: 1;
    1: x;
  esac;

  init(y):=0;
  next(y):= case
    State = 0: 1;
    State = 1: 1;
    1: y;
  esac;

  init(State) := 0;
  next(State) := case
    State<2: State+1;
    1: State;
  esac;

```

3.1 COVERAGE PRIORITIZATION

Model-based coverage criteria can be expressed as trap properties [9, 10]. For each coverable item one such property is formulated, expressing that the item cannot be reached. For example, a trap property might claim that a certain state is never reached or that a certain transition is never taken. Challenging a model-checker with a model and a trap property results in a counter-example, which is a trace illustrating how the item described by the trap property is reached. This principle is used for test-case creation, where it automatically results in test-suites that achieve a given coverage criterion. It is also used to measure the coverage of test-suites. The test-cases are converted to models as described above, and then the model-checker is challenged with the resulting models and the trap properties. For each trap property that results in a counter-example it is known that the test-case covers the according item.

While for overall coverage measurement it is sufficient to check how many trap properties are violated, this can easily be extended such that each test-case is checked against all trap properties. That way the overall coverage of each test-case can be determined. This information can be used in order to sort test-cases according to their coverage, either totally or additionally. The prioritization works as follows:

- 1) Create models from test-cases.
- 2) Create trap properties **TP** from coverage criterion.
- 3) **for** each test-case model **t do**.
- 4) model-check **t** against **TP**.
- 5) each trap resulting in a counter-example is covered.
- 6) **end for**
- 7) sort test-cases by number of covered traps.

3.2 FEP PRIORITIZATION

Fault exposing prioritization is based on mutation analysis. Model and specification mutation was introduced by Ammann and Black [11]. The ability to expose faults can be measured as the mutation score of a test-case.

With model-checkers, this can be done in two ways. One option is to create mutants of a given model, and then symbolically execute the test-cases against these models by combining the mutant model and the test-case model, using the test-case values as input-values for the mutant. A mutant is detected if the model-checker returns a counterexample when queried whether the output values of mutant and test-case are equal along the test-case. Unlike coverage based methods, this requires the model-checker to use the actual model in addition to the test-case model. If the model is complex, then this process is less efficient than the coverage based method.

The alternative is to reflect the transition relation of the model as special properties [12]. Each reflected property refers to one variable. For each possible transition a variable can take, there is one such property. It consists of the transition condition and makes an assertion about the value of the variable in the next state. These reflected properties can then be mutated instead of the original model. When checked against the original model the mutated properties result in efficient test-suites [2]. A mutation score can be efficiently calculated by checking these properties against the test-case models. This prioritization is therefore identical to coverage based prioritization apart from the use of mutated reflected properties instead of trap properties.

3.3 PROPERTY PRIORITIZATION

Property prioritization uses the concept of property relevance. A test-case is relevant to a property if the execution of the test-case can theoretically lead to a violation of the property. As presented in [13], property relevance can be determined with the aid of a model-checker by symbolically executing the test-case against a modified model which is allowed to take one single erroneous transition. The model checker then efficiently determines if a single erroneous transition is sufficient in order to reach a property violating state during the test-case execution. This process has to be repeated for each test-case.

- 1) Create modified model M' from model M .
- 2) Create models from test-cases.
- 3) **for** each test-case model **t do**.

- 4) Combine t and M' such that M' takes input values from t instead of the environment.
- 5) Model-check M' against all requirement properties.
- 6) t is relevant to each property causing a counter-example.
- 7) **end for**
- 8) Sort each test-case by relevance.

While the complexity of this evaluation process can be higher than for coverage or reflection based methods, it is only necessary to challenge the model-checker once with each test-case, so this is still significantly more efficient than the determination of the mutation score using symbolic execution would be. Once the property relevance of each test-case has been determined, this information can be used in order to calculate a total or adding prioritization for the test-cases.

3.4 OPTIMAL PRIORITIZATION

The optimal execution order of a test-suite with regard to a set of mutants is calculated with a greedy algorithm that successively adds the test-case next that detects the most yet undetected mutants.

4 Prioritizing test case at creation

Each test-case is assigned an importance value, initially 1. If a test-case is a prefix of another test-case or equal to it, the importance of this other test-case is increased. If a test-case subsumes other test-cases, then its importance is the sum of the subsumed test-cases plus 1.

- 1) **while** $t = \text{create next test-case}$ **do**
- 2) importance of $t = 1$
- 3) **if** $\exists t' \in T : t = t'$ **then**
- 4) increase importance of t' by 1
- 5) **else if** $\exists t' \in T : t \subset t'$ **then**
- 6) increase importance of t' by 1
- 7) **else if** $\exists t' \in T : t \supset t'$ **then**
- 8) **for all** $\exists t' \in T : t \supset t'$ **do**
- 9) replace t' with t in T
- 10) increase importance of t with importance of t'
- 11) **end for**
- 12) **else**
- 13) inset t in T
- 14) **end if**
- 15) **end while**
- 16) sort test-cases by importance

When creating test-cases automatically it is often the case that redundant test-cases are created. If a new test-case is a prefix of another test-case it is sufficient to retain the subsuming, longer test-case. If a new test-case subsumes other test-cases it is sufficient to retain the new test-case. Redundant test-cases are usually discarded. However, this redundancy information can also be used to prioritize test-cases. If a test-case or part of it is created more than once, this can be seen as an indication that this

test-case is more important than other test-cases. With this information prioritization can be performed without post-processing of the test-suite.

5 Experimental results for three crucial embedded systems

This section presents the results of an empirical evaluation for three security-critical embedded systems aiming to show that model-based test-case prioritization results in a noticeable performance improvement. We also want to analyze the newly defined property coverage techniques in comparison to well-known techniques. Finally we want to determine whether prioritization at test-case generation time results in a measurable improvement.

5.1 APFD

In order to quantify the efficiency gains achieved with a certain test-case prioritization, the metric APFD was introduced by Rothermel et al. [13, 14]. This metric is the weighted average percentage of faults detected over the life of a test suite. The APFD of a test suite T consisting n test cases and m mutants is defined as:

$$APFD = 1 - \frac{TF_1 + TF_2 + TF_3 + \dots + TF_m + \frac{1}{nm}}{nm} . \quad (2)$$

Here, TF_i is the first test-case in ordering T_0 of T which reveals fault i . We use this metric in order to compare the different prioritization techniques.

5.2 EXPERIMENT SETUP

The evaluation is based on a set of three examples. Each example consists of an SMV-model and specification. Different model-checker based methods (various coverage criteria, different mutation operators, property based methods) are used in order to create 23 different test-suites for each model. For each model a set of mutants is created. Unlike for program mutation, a model-checker can efficiently determine whether a model mutant is equivalent to the original or not. The APFD values for each of the test-suites is calculated using the subset of the in-equivalent model mutants that can be detected by the test-suite. Table 1 sums up the results of the test-case generation and the numbers of detected mutants. Only detectable mutants are relevant for the determination of the APFD value, as the test-case execution order has no influence on undetectable mutants.

Car Control (CA) is a simplified model of a car control. The Safety Injection System (SIS) example was introduced in [3] and has since been used frequently for studying automated test-case generation. Environmental Control (CC) is based on [12]. In order to validate the method we also use a set of 25 erroneous mutant implementations for the Cruise Control example applications written by Jeff Offutt.

TABLE 1 Test-suite statistics

| Example | CA | | SIS | | EC | |
|------------|-----|-----|-----|-----|-----|-----|
| | Avg | Max | Avg | Max | Avg | Max |
| Test Cases | 51 | 243 | 22 | 85 | 35 | 246 |
| Mutants | 264 | 311 | 265 | 339 | 535 | 732 |

5.3 RESULTS

Following the tradition of previous papers about test-case prioritization we use box-plots to illustrate the results of the APFD analysis. The box-plots illustrate minimum, maximum, median and standard deviation for each of the used prioritization methods. As can be seen in Figures 1, 3 and 4, there is still a gap between all prioritization techniques and the optimal prioritization.

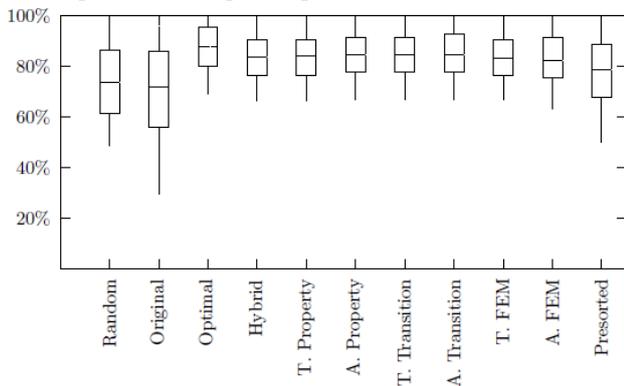


FIGURE 1 APFD of cruise control model

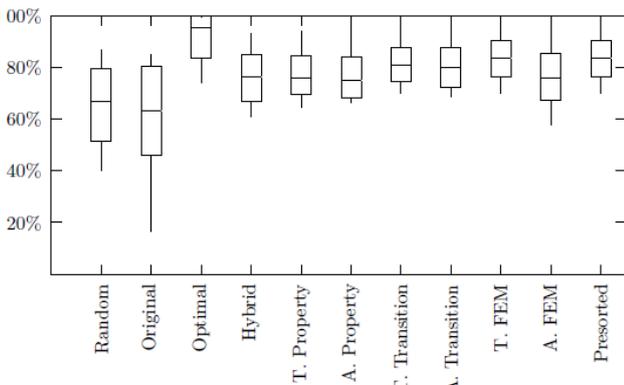


FIGURE 2 APFD of cruise control implementation

However, there is an improvement clearly visible compared to the random sorting of the test-cases and the original sorting, as provided by the test-case generation algorithm. Figure 6 lists the average APFD values for all examples and methods in a concise manner. The improvement is not always as significant as reported in previous works. This is probably because we used test-suites of different sizes, and the improvement is not quite so obvious for large test-suites. In general, a large amount of the mutants is detected with the first couple of test-cases (Figure 5), yet the remaining test-cases and mutants can distort the APFD value, if there are many test cases. Nevertheless, an improvement is visible. Figure 2 illustrates the APFD values for the same test-suites (except the optimal one) as in Figure 1, executed with the 25

erroneous implementations of Cruise Control. The values are comparable and we conclude that model-based prioritization is also valid for real implementations.

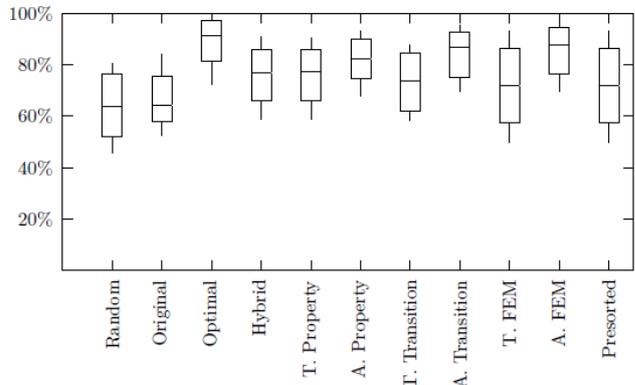


FIGURE 3 APFD of environmental system

The prioritization performed at test-case generation time (labelled presorted in the figures) is clearly better than random ordering, however there is still a gap between presorted test-suites and post-processing prioritization. This gap is also visible in Figure 5. Interestingly, the presorted test-suites performed better than most other prioritization techniques during the evaluation on the cruise control implementations. In general we can conclude that prioritization at test-case generation is definitely useful, especially as it only requires negligible additional computational costs.

There are only minor differences between the various prioritization techniques. In general, those techniques that use adding sorting perform slightly better than those with total sorting. Property prioritization performs good (regard also Figure 5), in fact it sometimes outperforms coverage based prioritization techniques.

However, this case study does not reflect on the quality of the specification. It is conceivable that a specification consisting of more and better properties will result in better property based prioritization.

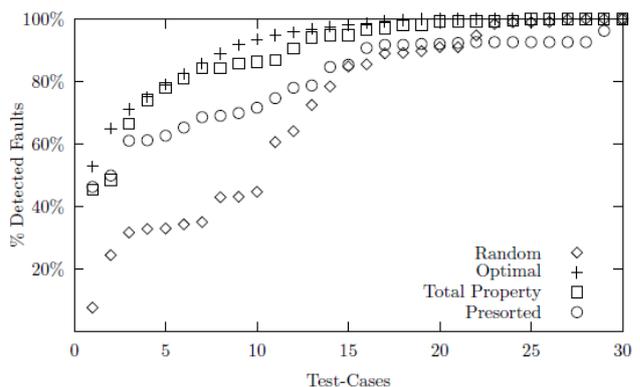


FIGURE 4 Fault detection rates for environmental control example

While model-checkers in general are prone to performance problems this is not a problem for prioritization, as the state space of test-case models is usually significantly smaller than that of related functional models.

6 Conclusion

In this paper we have demonstrated how model-checkers can be used for test-case prioritization for the embedded systems of environmental control.

This makes it possible to efficiently apply prioritization when creating test-cases with model-checkers. We adapted

several well known prioritization methods originally based on source code to models. In addition we introduced new property based prioritization methods. Finally, we showed that test-case prioritization can be performed automatically during test-case generation, without post-processing.

References

- [1] Hu C, Zhu L 2010 The analysis and the evaluation of complicated network software *LNCS* **13**(10) 1-5
- [2] Wang Y, Xia H, Yan R 2008 The analysis of the social network and the study of the application cases of NetDraw *Modern education technology* **18**(4) 85-9
- [3] Pothén A, Simon H, Liou K P 1990 Partitioning sparse matrices with eigenvectors of graphs. *SIAM Matrix Anal. Appl* **11** 430-6
- [4] Girvan M, Newman ME 2001 Community structure in social and biological networks. *Proc. Natl. Acad. Sci* **99**(12) 7821-6
- [5] Newman ME, Girvan M 2004 Finding and evaluating community structure in networks *Phys. Rev. E* **39**(10) 69-84
- [6] Toyoda M, Kitsuregawa M 2003 Extracting evolution of web communities from a series of web archives *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia* 78-87
- [7] Palla G, Derényi I, Vicsek T 2007 The Critical Point of k -groups Percolation in the Erdős–Rényi Graph *Journal of Statistical Physics* **128**(1) 219-27
- [8] Palla G, Barabási A-L, Vicsek T 2007 Community dynamics in social networks *Noise and Stochastics in Complex Systems and Finance* **6601**(3) 273–87
- [9] Xu C, Zhang Y, Dan Y 2011 Ontology based Image Semantics Recognition using Description Logics *IJACT International Journal of Advancements in Computing Technology* **3**(10) 1-8
- [10] Ju C, Wei J 2012 Research on Multi-interest Profile Based on Resource Clustering *JCIT Journal of Convergence Information Technology* **7**(21) 582-90
- [11] Gargantini A, Heitmeyer C 1999 Using Model Checking to Generate Tests From Requirements Specifications *In Software Engineering - ESEC/FSE '99: 7th European Software Engineering Conference, Held Jointly with the 7th ACM SIGSOFT Symposium on the Foundations of Software Engineering* **1687**(1) 146-62
- [12] Kim J-M, Porter A 2009 A history-based test prioritization technique for regression testing in resource constrained environments. *In ICSE '09: Proceedings of the 31th International Conference on Software Engineering* **139**(3) 119-29
- [13] Rothermel G, Untch R H, Chu C, Harrold M J 2009 Test case prioritization: an empirical study *Proceedings of the IEEE International Conference on Software Maintenance* **168**(1) 179-83
- [14] Srikanth H, Williams L 2005 On the economics of requirements-based test case prioritization *Proceedings of the 7th international workshop on Economics-driven software engineering research* **153**(1) 1-3

| Authors | |
|---|---|
|  | <p>Yan Li, born in April, 1977, Zibo County, Shandong Province, China</p> <p>Current position, grades: the lecturer of School of computer, Shandong University of Technology, China. University studies: B.Sc. in application of computer Technology from Northeast Forestry University, M.Sc. from Shandong University in China. Scientific interest: computer modelling, information retrieval. Publications: more than 10 papers. Experience: teaching experience of 14 years, 3 scientific research projects.</p> |
|  | <p>Zhe Zhang, born in January, 1982, Nanyang County, Henan Province, China</p> <p>Current position, grades: the lecturer of School of Software, Nanyang Normal University, China. University studies: M.Sc. in computer applications from Huazhong University of Science & Technology in China. Scientific interest: software engineering, formal modelling. Publications: more than 6 papers. Experience: teaching experience of 10 years, 4 research projects.</p> |
|  | <p>Guihong Jiang, born in November, 1966, Zibo County, Shandong Province, China</p> <p>Current position, grades: the associate professor of School of computer, Shandong University of Technology, China. University studies: B.Sc. in Organic chemical Engineering from Qingdao University of Science & Technology in China. Scientific interest: database design, software engineering. Publications: more than 6 papers, 8 teaching books about database or program design published. Experience: teaching experience of 25 years, 2 scientific research projects.</p> |
|  | <p>Xiaofeng Cui, born in October, 1969, Zibo County, Shandong Province, China</p> <p>Current position, grades: the associate professor of School of computer, Shandong University of Technology, China. University studies: B.Sc. in computer science from China Agricultural University. Scientific interest: data mining, artificial intelligence. Publications: more than 4 papers. Experience: teaching experience of 17 years, 10 education research projects.</p> |

Source enumeration algorithm based on eigenvector: revisit from the perspective of information theory

Wenzhun Huang*, Shanwen Zhang

Department of Engineering Technology, Xijing University, No.1 Xijing Road, Xi'an, 710123, China

Received 11 February 2014, www.tsi.lv

Abstract

In case of low signal to noise ratio (SNR) and small snapshot condition, it is difficult to separate sources and noises, and the performance of classical eigenvector source estimation algorithm drops quickly. To solve the problem, further research is carried out around the characters of eigenvalue and eigenvector, and a novel eigenvalue algorithm is presented based on the theory of source enumeration. In detail, the eigenvectors of sample covariance matrix are employed as the decision factor, which is insensitive to SNR. And an improved Predictive Description Length (PDL) criterion is adopted to enumerate source number. Theoretical analysis and simulation results demonstrate that the proposed algorithm is available and efficient in case of low SNR and small snapshot condition compared with those of Minimum Description Length (MDL) and PDL.

Keywords: source enumeration, eigenvector, signal-to-noise ratio, information theory, predictive description length

1 Introduction

The estimation signal number in noise background is one of the key problems in array signal processing, and widely applied in radar, communication, biomedical, seismic signals and the field of electronic countermeasures and so on. The accuracy number of signals is the prerequisite for many super-resolution array processing algorithm. If the estimated signal number was different with the accurate number of signals, many super-resolutions estimation algorithms would suffer from severe performance degradation. While in the condition of low Signal to Noise Ratio (SNR) and small snapshot, the process of source enumeration becomes much more complex and difficult. Among classic subspace decomposition estimation algorithms for Difference of Arrival (DOA) [1], source number is critical in dividing signal eigenvector from noise eigenvector, which is applied in spanning the independent noise subspace, and signal subspace as well [2]. When estimated source number is distorted or even deviated, the orthogonality between the two subspaces will be damaged [3, 4], which will lead to DOA estimation performance degradation directly. For source number estimation, the algorithm based on information theory criterion is now declared the most common and effective, such as the sequence hypothesis criterion, Akaike information criterion (AIC) [5,6] and minimum description length (MDL) criterion [7]. However, for information theory criterion and algorithms of source number estimation and their improved algorithms, the estimation performance reduction is more serious under the condition of low SNR or missing snapshot. The algorithm based on MDL criterion is used to multiple coherent source number's

estimation and positioning [3], and an improved algorithm applying random geometry-array and covariance matrix noise is proposed [7], but a large number of arrays are required and difficult to realize effective estimation at low SNR. As for the performance and calculation speed, the recursion method of PDL criterion [8-10] provided lately have great improvement than MDL and G&T criterion [11].

All the above algorithms are based on covariance matrix's eigenvalue, and the performance is decided by the eigenvalue capability in dividing signal eigenvalue and noise eigenvalue. However, in case of limited snapshot data, the division process is quite difficult, especially at low SNR. The solution is to change eigenvalue, in which a reconstructed eigenvalue cluster at low SNR is used to extract coherent and non-coherent signal in BEM criterion [7, 8]. The performance of eigenvalue is depressed seriously by various noises, and the eigenvector is also used to replace eigenvalue to estimate source number [12].

In the paper, a source number estimation algorithm is presented based on sampling covariance matrix eigenvector. Eigenvector of sampling covariance matrix is used to construct judgment variable in the algorithm, and the number of source is estimated according to improved predictive description length (PDL) criterion. The paper is organized as follows. The second part gives the preliminary of signal model, and the eigenvector algorithm and source enumeration is provided in third part. The fourth part demonstrates the simulation and the results, and the last gives the conclusion.

* Corresponding author e-mail: wenzhunw@126.com

2 Signal model

Denote N -Array spaced array has M ($M < N$) signals with narrow-band signal and fixed centre frequency from far-field source shooting with angles θ_k ($k = 1, 2, 3, \dots, M$). Denote $X(t)$ the complex envelope receiving vector of antenna arrays [10], and then: $X(t) = [x_1(t), x_2(t), \dots, x_N(t)]^T$ can be expressed as follows at snapshot time t :

$$X(t) = A(\theta)S(t) + N(t) = \sum_{i=1}^M \alpha(\theta_i) \cdot s_i(t) + N(t), \quad (1)$$

$t = 1, 2, 3, \dots$

In the Equation (1), $S(t) = [s_1(t), s_2(t), \dots, s_N(t)]^T$ is the complex envelope vector of M receiving source signals, $N(t) = [n_1(t), n_2(t), \dots, n_N(t)]^T$ complex Gauss noise vector of antenna arrays, and $A(\theta) = [a_1(\theta_1), a_2(\theta_2), \dots, a_N(\theta_M)]$ manifold array matrix with $N \times M$ dimensions. Element $a(\theta_k)$ ($k = 1, 2, \dots, M$) is a steering vector with $M \times 1$ dimensions aiming at wave direction and a_i can be expressed as follows:

$$a_i = [e^{-j\mu_{i1}}, e^{-j\mu_{i2}}, \dots, e^{-j\mu_{iM}}]^T, \quad (2)$$

where $\mu_{mi} = \frac{2\pi}{\lambda} [x_m \cos \theta_i \cos \varphi_i + y_m \sin \theta_i \cos \varphi_i]$,

(x_m, y_m) is the m array signal position, and (θ_i, φ_i) is the parameter of incident angle.

Covariance matrix of original receiving data vector $X(t)$ can be expressed as follow:

$$R = E[X(t)X^H(t)]. \quad (3)$$

In the Equation (3), $E[\]$ denotes the mathematical expectation, and “ H ” denotes Hermitian transformation. And then the covariance matrix [13] can be decomposed as follow:

$$R = U \sum^2 U^H = U_S \sum_N^2 U_S^H + U_N \sum_N^2 U_N^H, \quad (4)$$

where U_S is the matrix consisting of eigenvectors corresponding with top M large eigenvalues, and $\sum^2 = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_N) = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_N^2)$, the diagonal matrix consisting of eigenvalues, U_N the matrix consisting of eigenvectors corresponding with last $N - M$ small eigenvalues. Elements of diagonal matrix $\sum_S^2 = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_M)$ consist of the top M large eigenvalues corresponds with U_S eigenvalues. Elements of diagonal matrices $\sum_S^2 = \text{diag}(\lambda_{M+1}, \lambda_{M+2}, \dots, \lambda_N)$ denote

the last $N - M$ small eigenvalues corresponding to U_N eigenvalues. Eigenvalues meet the condition $\lambda_1 > \lambda_2 > \dots > \lambda_M > \lambda_{M+1} = \lambda_{M+2} = \dots = \lambda_N = \sigma^2$. Source number can be determined by the number of the last $N - M$ eigenvalues.

To determine the number of M eigenvalues, principle component analysis (PCA) and similar theory provide some methods [13]. The popular power method is based on the ratio of eigenvalues, in which the sum of M large eigenvalues is around 85% to sum of all eigenvalues. The ratio can be expressed as follow:

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_M}{\lambda_1 + \lambda_2 + \dots + \lambda_M + \dots + \lambda_N} \approx 85\% \quad (5)$$

and signal-to-noise model provided another method in the case of low noise. In detail, relationship between eigenvalues can be expressed as follow:

$$\lambda_1 > \lambda_2 > \dots > \lambda_M \gg \lambda_{M+1} \approx \lambda_{M+2} \approx \dots \approx \lambda_N \approx \sigma^2 \quad (6)$$

in which the variances of noises are much smaller than those of signals, and around the value of σ^2 .

In practical consideration, for the reason that matrix R is decided by limited snapshot points, and the eigenvalues are different and difficult to be evaluated at low SNR and small snapshot, it is difficult to accurately estimate the source number.

3 Source enumeration via eigenvectors

3.1 TRANSFORM OF EIGENVECTOR

Define signal subspace as the subspace spanned by U_S , and noise subspace spanned by U_N , where the two subspaces are orthogonal. For the subspace spanned by U_S and subspace spanned by A belong to the same signal subspace, and U_S 's column vectors are basic vectors for signal subspace, the steering vector $a(\theta_i)$ ($i = 1, 2, \dots, M$) can be expressed as follow:

$$a(\theta_i) = \sum_{j=1}^M c_{ij} \cdot s_j \quad (7)$$

in which, S_j is the basic vector for signal subspace and column vectors of U_S , c_{ij} ($i, j = 1, 2, \dots, M$) the linear combination index for U_S .

Then it can be obtained that the noise subspace and steering vector subspace spanned from A are mutually orthogonal. That is:

$$n_k^H \alpha(\theta_i) = 0, (i = 1, 2, \dots, N - M). \quad (8)$$

Define vector y_k ($k = 1, 2, \dots, N - M$) the function of column vectors n_k^H and $X(t)$, and then y_k can be expressed as follow:

$$y_k(t) = n_k^H X(t) = n_k^H \cdot \left(\sum_{i=1}^M \alpha(\theta_i) s_i(t) + N(t) \right) = \tag{9}$$

$$n_k^H N(t) = \omega_{Nk}(t).$$

Similarly, eigenvectors of U_s column vectors can be defined as standard orthogonal basis, and a new vector $z_i(t)(i=1,2,\dots,M)$ can be obtained.

$$z_i(t) = S_i^H X(t) = S_i^H \left(\sum_{j=1}^M \alpha(\theta_j) s_j(t) \right) + S_i^H N(t) = \tag{10}$$

$$\sum_{j=1}^M \left[\sum_{\rho=1}^M c_{j\rho} S_i^H s_\rho s_j(t) \right] + S_i^H N(t) =$$

$$\sum_{j=1}^M c_{ji} s_j(t) + \omega_{Ni}(t),$$

where $\omega_{Nk}(t)$ and $\omega_{Ni}(t)$ are independent and identically distributed complex Gauss random vector, with zero mean and σ^2 variance.

From the Equations (9) and (10), it can be concluded that if the steering vector is weight disposed by noise subspace eigenvectors, the output variable will not comprise any signal component. Otherwise, if the steering vector is weighted by signal subspace eigenvectors, the output variable will comprise both signal component and noise component. Therefore, column vector of snapshot time t can be expressed as follow:

$$d(t) = [z_1(t), \dots, z_M(t), y_1(t), \dots, y_{N-M}(t)]^T \tag{11}$$

and observation matrix $D = [d_1, d_2, \dots, d_T]$ is constructed by T snapshots. Each column's power spectrum and peak value P_n of matrix D can be obtained via Fast Fourier Transform (FFT). And then the judgment variable $q_n(n=1,2,\dots,N)$ are defined as follow:

$$q_n = N \cdot \frac{P_n}{\sum_{i=1}^N P_i}. \tag{12}$$

3.2 ALGORITHMS BASED ON INFORMATION THEORY

Information theory criterion is to solve the question of pattern recognition. For given observe signal $X = [X_1, X_2, \dots, X_n]$, the purpose of information theory criterion is to choose the most matching model in the serial of parameter probability model $f(x|\Theta)$ [14,15]. When observed signals are narrow band and satisfy independent and identically distributed (*i.i.d*), and noise is Additive White Gauss Noise (AWGN), the Akaike information criterion (AIC) and minimum description length (MDL) criterion are expressed as follows:

$$MDL(M) = L(N - M) \ln \Lambda(M) + \frac{1}{2} M(2N - M) \ln L, \tag{13}$$

$$AIC(M) = 2L(N - M) \ln \Lambda(M) + 2n(2M - N)$$

where $\Lambda(M)$ can be denoted by the Equation (14).

$$\Lambda(M) = \frac{1}{N - M} \frac{\sum_{i=M+1}^N \lambda_i}{\left(\prod_{i=M+1}^N \lambda_i \right)^{\frac{1}{N-M}}} \tag{14}$$

and when AIC value (or MDL value) is minimized, the corresponding value M is just the number of estimated sources.

3.3 IMPROVED JUDGMENT CRITERION

According to the Equation (2), the covariance matrix of original data $X(t)$ can also be expressed as follow:

$$R = AR_s A^H + \sigma^2 I, \tag{15}$$

where R_s is the covariance matrix, σ^2 noise variance, and matrix A the function of incident azimuth vectors $\theta_k = [\theta_1, \theta_2, \dots, \theta_M]^T$.

Covariance matrix R can also be viewed as the function of parameter set $\Phi = [M, \theta_k, \sigma^2, R_s]$, and the conditional probability density function of receiving data vector is as [5]:

$$f(X | \Phi) = \frac{1}{\pi^n |R|} \exp \left\{ -X^H (R)^{-1} X \right\}, \tag{16}$$

where symbol $|R|$ stands for matrix's determinant. The aim of source number estimation is to estimate source number M and then the incidence angle vector $\theta_k = [\theta_1, \theta_2, \dots, \theta_M]^T$.

Define PDL of L data column vector as follows:

$$P(L) = - \sum_{i=1}^L \log f(X(t) | \Phi_{t-1}), \tag{17}$$

where Φ_{t-1} is Maximum Likelihood (ML) estimator of all receiving data vector before time t . For any given moment, parameter vectors' estimators are based on previous observation, thus it is called prediction of description length estimation.

To estimate source number, the following cost function is designed [5]:

$$PDL(m) = \arg \min_{m \in K} P(L) = \arg \min_{m \in K} \left[- \sum_{i=1}^L \log f(X(t) | \Phi_{t-1}) \right], \tag{18}$$

where $K = \{0, 1, \dots, N-1\}$. And source number m is estimated only when $PDL(m)$ is minimized.

From the classic PDL judgment criterion provided in [8], $P(L)$ can be redefined as follows:

$$P(L) = -\sum_{i=1}^L \log f(x_i(t) | \lambda_i). \tag{19}$$

Among which, λ_i is the eigenvalue of Φ_{t-1} , $x_i(t)$ the i -th receiving data of antenna arrays.

Replace λ_i with judgment variable q_n , and the source number estimation criterion can be expressed as follows:

$$P(L) = -\sum_{i=1}^L \log f(x_i(t) | q_i) \tag{20}$$

Therefore, the process of source number estimation algorithm can be concluded in following steps.

- Step1:** Construct covariance matrix R;
- Step2:** Calculate corresponding eigenvectors by the eigenvalues of R;
- Step3:** Weight $X(t)$ with eigenvectors and obtain observation matrix D;
- Step4:** Calculate each column's power spectrum of D;
- Step5:** Calculate the peak value and q_n in the Equation (12);
- Step6:** Confirm source number according to the Equation (20).

4 Simulation

In the simulations, different noise conditions are considered, such as non-stationary, non-Gaussian or even colored, and the proposed algorithm is compared with those of MDL and PDL. Simulations demonstrate similar results that the proposed algorithm has better performance in the case of low SNR and slow snapshot, and for simplicity only part of simulation process and results are demonstrated here.

4.1 SIMULATION CONDITION

Monte-Carlo numerical simulation method is adopted to estimate source number estimation algorithm's performance. Simulation parameters are set as follows:

- Antenna arrays is an eight uniform array, and array element spacing is 0.5λ .
- At every Monte-Carlo simulation, snapshot number is set as 8192.
- Input signals are three incoherent narrow-band signal (BPSK signal with 32Kbps), and the incidence angles are 45° , 80° and 89° .
- The channel is AWGN channel.

4.2 SIMULATION RESULTS

For given SNR, average value of 200 times Monte-Carlo simulation results is employed to obtain the testing success probability. When $Eb/N0 = 0db$, the distribution result and comparison of judgment variable q_n and eigenvalue in proposed algorithm are shown in Figure 1. From Figure 1, it can be seen that after the arrays output vector weighted, the difference between q is quite obvious. For the value of q array, the top three ones are larger than the last five ones, and the last five ones are almost equal. All the eigenvalue are almost equal expect the first one, which provides guarantee for next source number estimation.

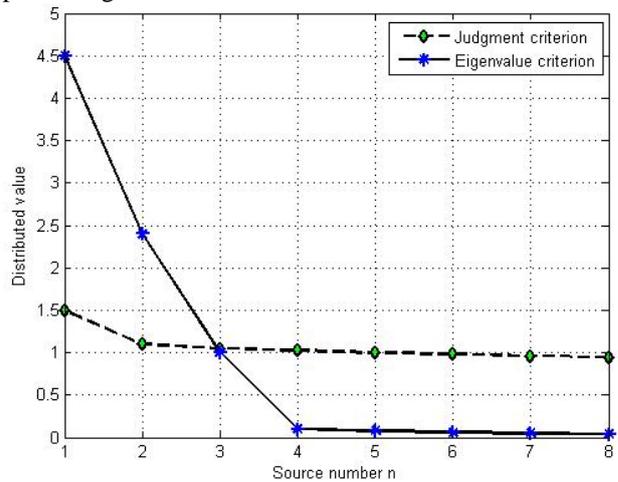


FIGURE 1 Distribution of judgment variance and eigenvalue

Simulation provides the change between provided algorithm's, MDL's and PDL's estimation success probability in Figure 2. There the mean values are used to balance the probability at each $Eb/N0$.

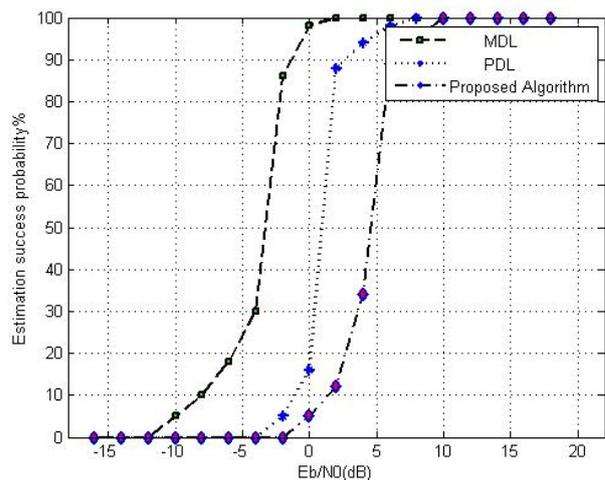


FIGURE 2 Estimation success probability when the value of $Eb/N0$ changes

In detail, Figure 2 shows that when $Eb/N0 \geq 6db$ (signal input SNR is high), the proposed algorithm has similar estimation success probability with that of MDL and PDL. But in the case that the input SNR is low, estimation success probability of proposed algorithm is

higher than those of *MDL* and *PDL*. Under the condition that estimation success probability is 90%, proposed algorithm has almost 4dB improvement than that of *PDL* and 6dB improvement than that of *MDL*.

5 Conclusion

A new source number estimation algorithm is provided to solve the problem in case of low SNR. In the proposed algorithm, eigenvectors of covariance matrix is adopted to replace eigenvalue to obtain a new judgment variable, and then the characteristics of source number are more effectively displayed. Therefore, the improved *PDL* judgment criterion is used to implement the estimation. The simulation results shows that the algorithm provided

has obvious improvement in SNR compared with classic *MDL* and *PDL* at the same estimation success probability.

Acknowledgement

The results discussed in this paper are part of the theses works of UAV data link and satellite communication anti-jamming system. This work is supported by National Natural Science Foundation of China (61174162), and Shaanxi Provincial Science & Technology Department (Program No. 2011K06-36), and Shaanxi Provincial Education Department (Program No. 2013JK1145) and the second batch of College Scientific Research Foundation in 2013 (Program No. XJ130225).

References

- [1] Cozzens J H, Sousa M J 1994 *Signal Processing IEEE Transactions on* **42**(2) 304-17
- [2] Krim H, Cozzens J H 1994 *Signal Processing IEEE Transactions on* **42**(7) 1662-8
- [3] Liavas A P, Regalia P A, Delmas J-P 1999 *Signal Processing IEEE Transactions on* **47**(12) 3336-44
- [4] Green P J, Taylor D P 2002 *Signal Processing IEEE Transactions on* **50**(6) 1307-14
- [5] Valaee S, Kabal P 2004 *Signal Processing IEEE Transactions on* **52**(5) 1171-8
- [6] Huang L, Wu S, Li X 2007 *Signal Processing IEEE Transactions on* **55**(12) 5658-67
- [7] Gu J F, Wei P, Tai H M 2007 *Signal Processing IET* **1**(1) 2-8
- [8] Valaee S, Shabazpanahi S *IEEE Transactions on Communications* **56**(7) 1189-97
- [9] Valaee S, Kabal P 2004 *IEEE Transactions on Signal Processing* **52**(5) 1171-78
- [10] Jiang L, Cai P, Yang J, Wang Y, Xu D 2007 A new source number estimation method based on the beam eigenvalue *Journal of Marine Science and Application* **6**(1) 41-6
- [11] Liu Z-M, Huang Z-T, Zhou Y-Y 2013 *IEEE Transactions on Wireless Communication* **12**(8) 1-12
- [12] Yazdian E, Gazor s, Bastani H 2012 Source enumeration in large arrays using moments of eigenvalues and relatively few samples *IET Signal Processing* **6**(7) 689-96
- [13] Nasser M, Bakhshi H, Sahebdel S, Falahian R, Ahmadi M 2011 PCA Application in Channel Estimation in MIMO-OFDM System *Int'l J of Communications Network and System Sciences* **4**(6) 384-7
- [14] Mei T, Yin F, Wang J 2009 *Audio, Speech, and Language Processing, IEEE Transactions on* **17**(6) 1099-108
- [15] De Luigi C, Moreau E 2013 Optimal combination of fourth-order cumulant based contrasts for blind separation of noncircular signals *Signal processing* **93**(4) 842-55

Authors



Wenzhun Huang, born in February, 1968, Xi'an City, Shanxi Province, P.R. China

Current position, grades: Associate Professor of School of Engineering Technology, Xijing University, China.

University studies: B.S. and M.S. in Electrical Information and Navigation College at Air Force Engineering University in China. Ph.D. at Northwest Polytechnic University in China.

Scientific interest: Information processing, anti-jamming technology, wireless communication.

Publications: more than 81 papers.

Experience: Teaching experience of 17 years, 5 scientific research projects.



Shanwen Zhang, born in August, 1965, Xi'an City, Shanxi Province, P.R. China

Current position, grades: the Associate Professor of Xijing University in China. Teacher in University of China.

University studies: M.Sc. degree at Northwest Polytechnic University in China. Ph.D. degree at Air Force Engineering University in China.

Scientific interest: Wavelet transforms, rough sets, genetic algorithm, pattern recognition and manifold learning.

Publications: more than 50 papers.

Experience: Teaching experience of 25 years, 2 scientific research projects.

A method for improving real-time communication of switched Ethernet

Jun Zhao*, Zhong Ma, Xiangjun Wu

Wuhan Digital Engineering Institute, No.718, Luoyu Road, Wuhan, China, 430074

Received 12 June 2014, www.tsi.lv

Abstract

A method has been proposed for improving real-time communication of Switched Ethernet. Based on virtual link ideas, this method offline plans the whole network traffic under traditional Switched Ethernet hardware conditions, improves network terminal TCP/IP protocol by adding real-time communication interface, traffic shaping and priority queuing etc., and uses IEEE802.1p protocol on the switches of communication path. And it employs the network calculus theory to deduce the equation of calculating maximum end-to-end delay of real-time traffic. Meanwhile, it receives a simulation test of OPNET software. Both theoretical calculation and simulation results show that this method can effectively improve real-time communication of Switched Ethernet.

Keywords: Switched Ethernet, virtual link, real-time, traffic shaping, network calculus

1 Introduction

There has been increasingly extensive application of Switched Ethernet in aerospace, shipping, factory control and many other fields in recent years with its real-time performance as research focus. Although Avionics Full Duplex Switched Ethernet (AFDX) [1] can guarantee real-time communication of key data, it requires a particular network switch and interface card (NIC). In addition, traditional best-effort traffic is not compatible. Although Time-Triggered Ethernet (TTE) [2] can integrate both real-time traffic and best-effort traffic, it also needs a particular switch and demands time synchronization of the whole network. As a master-slave synchronization network, Flexible Time-Triggered Ethernet (FTT-SE) [3] regulates network traffic on the basis of the time division multiple access (TDMA) principle, but it also calls for a separate controller. To improve real-time communication of key information under traditional Switched Ethernet hardware conditions, and make best-effort traffic compatible, this paper proposes a method of improving real-time communication of Switched Ethernet. The first part will elaborate design ideas of this method; the second will employ the network calculus [4-5] theory to deduce the equation of calculating maximum end-to-end delay of real-time traffic; the third part will prove the effectiveness of this method through conducting a simulation test of OPNET software; the final part will draw a conclusion.

2 Method design

Based on virtual link [1] ideas, this method categorizes

whole real-time traffic and best-effort traffic into different virtual links and arranges bandwidth accordingly, then it offline plans the whole network traffic to avoid traffic overload and network congestion. To guarantee effective communication as planned above, it adds real-time communication interface, traffic shaping and priority queuing to the network terminal TCP/IP protocol, and adopts IEEE802.1p standardized as two kinds of network traffic priority assignment on network terminals and switches. Here are some definitions of this method:

Definition 1: suppose that the whole network virtual link set is $VL = \{vl_1, vl_2, \dots, vl_n\}$, the virtual link is $vl_i = (sma_i, dma_i, pr_i)$ in which i is the ordinal of vl_i ; sma_i is the source MAC address of vl_i ; dma_i is the destination MAC address of vl_i ; pr_i is the priority $\in \{p_{rt}, p_{be}\}$ of vl_i , p_{rt} is the real-time traffic priority and p_{be} is the best-effort traffic priority and $p_{rt} > p_{be}$.

Definition 2: suppose the whole network terminal set is $ND = \{nd_1, nd_2, \dots, nd_w\}$, and terminals are $nd_j = (nsma_j, rbma_j, rmma_j, svl_j, rvl_j, LB_j \{LB_j^1, LB_j^2, \dots, LB_j^{jm}\}, SQ_j^{rt}, SQ_j^{be}, RQ_j^{rt}, RQ_j^{be}, (C_j^{send}, C_j^{recv}))$, in which j is the terminal number of nd_j ; $nsma_j$ is the source MAC address of nd_j ; $rbma_j$ is the MAC address to receive broadcast of nd_j ; $rmma_j$ is the MAC address to receive multicast of nd_j ;

* Corresponding author e-mail: 57278201@qq.com

$svl_j = \{vl_i | sma_i = nsma_j\}$ refers to the virtual link set sent by nd_j ; $rvl_j = \{vl_i | dma_i = nsma_j || dma_i = rbma_j || dma_i = rmma_j\}$ refers to the virtual link set received by nd_j ; $LB_j = \{LB_j^k, k \in svl_j\}$ stands for the leaky bucket regulator set of nd_j , which is responsible for shaping the traffic which has been sent to the virtual link. $LB_j^k = (bt_j^k, lbr_j^k, lbb_j^k)$ refers to the leaky bucket regulator of vl_k on nd_j ; bt_j^k is the transmission cycle, lbr_j^k is the arranged bandwidth and lbb_j^k is the maximum packet length; SQ_j^{rt} is the queue of sending real-time traffic of nd_j ; SQ_j^{be} is the queue of sending best-effort traffic of nd_j ; RQ_j^{rt} is the queue of receiving real-time traffic of nd_j ; RQ_j^{be} is the queue of receiving best-effort traffic of nd_j ; C_j^{send} is the bandwidth of sending traffic of nd_j and C_j^{recv} is the bandwidth of receiving traffic of nd_j .

Definition 3: suppose that the virtual link configuration table offline is $VLTable$, then network planners use it to arrange bandwidth for the whole network virtual link to avoid traffic overload and ensure correct and real-time communication.

2.1 IMPROVED MODEL OF NETWORK TERMINAL TRANSMISSION

Compared with the standard TCP/IP protocol, there are two improvements on both the upper layer and the lower layer of the transmission model. First, add real-time traffic transmission interfaces to the application layer to isolate real-time traffic and best-effort traffic, then encapsulate real-time traffic into standard UDP packets for identification and forwarding of the commercial Ethernet switches, finally buffer real-time traffic according to virtual like ideas. Second, add classifiers, leaky bucket regulators [6] and traffic schedulers to the data link layer. To begin with, data stream is first classified and labelled by classifiers, then respective virtual links are calculated and saved into according LB_j^k ; next, traffic shaping is carried out by leaky bucket regulators through the bucket algorithm, and data stream is forwarded to SQ_j^{rt} , SQ_j^{be} ; finally, data stream in SQ_j^{rt} , SQ_j^{be} is scheduled through the strict priority queue scheduling algorithm. The improved model of

network terminal transmission is illustrated in detail in Figure 1.

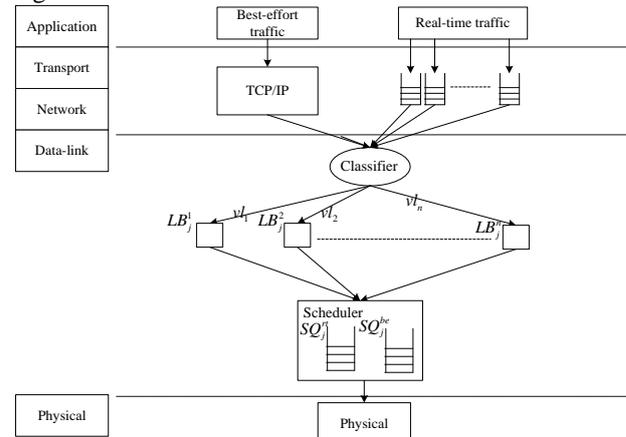


FIGURE 1 The improved model of transmission

Real-time traffic and best-effort traffic are classified according to the Algorithm 1 as follows:

Algorithm1. Packets are classified before transmission

Input: Send the packet $SendPack$, LB_j before classification

Output: LB_j after transmission

Algorithm steps:

- 1) Calculate vl_k it belongs to according to the source MAC address and destination MAC address of $SendPack$, and determine LB_j^k , which must be saved.
- 2) If $SendPack$ is sent by real-time traffic protocol, its $priorityType$ will be labelled as p_{rt} according to IEEE802.1p; otherwise, if it is sent by TCP/IP protocol, its $priorityType$ will be labelled as p_{be} instead.
- 3) Save $SendPack$ into LB_j^k .

After classification, the traffic in LB_j will be shaped by leaky bucket regulators through the bucket algorithm. Then, LB_j^k will arrive at the curve [4]

$\alpha_j^k = lbt_j^k t + lbb_j^k$. To avoid overload of sending traffic, $VLTable$ should meet the requirements of the following formula.

$$\forall nd_j, \sum_{k \in svl_j} lbr_j^k \leq C_j^{send}. \tag{1}$$

To avoid overload of receiving traffic, $VLTable$ should meet the requirements of the following formula.

$$\forall nd_j, \sum_{x \in ND, x \neq j} \sum_{k \in rvl_j} lbr_x^k \leq C_j^{recv}. \tag{2}$$

When virtual link bandwidth is offline arranged, the above two formulas should be abided by. After traffic shaping is realized, data is forwarded into SQ_j^{rt} and SQ_j^{be} . Traffic in SQ_j^{rt} , SQ_j^{be} is scheduled through the strict priority queue scheduling algorithm and is then sent to the physical layer.

2.2 THE IMPROVED MODEL OF NETWORK TERMINAL RECEIVER

TCP/IP protocol, the relative standard of network terminal receiver, consists of improvements on two parts from the upper layer to the lower layer: 1) add dedicated interface to receive real-time traffic at the application layer and isolate the best-effort traffic from real-time traffic with the real-time traffic of each virtual link buffering respectively. 2) add classifier and traffic scheduler at the data link layer. Data traffic firstly passes through the classifier for classification. Then the scheduler uses the strict priority queuing scheduling algorithm for traffic scheduling. See algorithm 2 and algorithm 3 for the specific. The improved model of network terminal receiver is shown in figure 2.

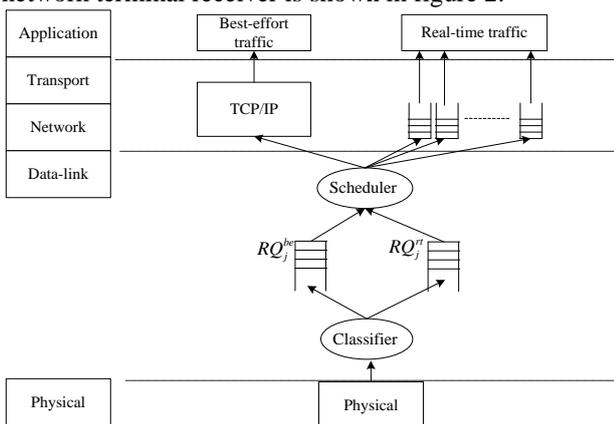


FIGURE 2 The improved model of receiver.

Algorithm 2: Classify and receive messages

Input: Receive the message $RecvPack$.

Output: RQ_j^{rt} , RQ_j^{be} .

Algorithm steps:

1) Analyse the received message $RecvPack$ according to the IEEE802.1p protocol and determine the packet priority $priorityType$.

2) When $priorityType$ is p_{rt} , if RQ_j^{rt} is non-null, store $RecvPack$ at the queue tail of RQ_j^{rt} , Otherwise, abandon $RecvPack$.

3) When $priorityType$ is p_{be} , if RQ_j^{be} is non-null, store $RecvPack$ at the queue tail of RQ_j^{be} , Otherwise, abandon $RecvPack$.

Algorithm 3: The scheduling mainly receives

queue packet.

Input: RQ_j^{rt} , RQ_j^{be}

Output: Scheduling packet

Algorithm steps:

1) If RQ_j^{rt} is non-null, determine vl_k , the message belonging, based on source MAC address and destination MAC address of the RQ_j^{rt} head message and store it in corresponding real-time traffic buffer.

2) If RQ_j^{be} is non-null, store the RQ_j^{be} head message in the TCP/IP protocol buffer.

The switch uses the priority queuing scheduling based on IEEE802.1p.

3 The calculation of end-to-end delay upper bound of real-time traffic

The calculation formula of the maximum end-to-end delay of real-time traffic used in this approach can be deduced according to the network calculus theory. The end-to-end delay of real-time traffic of vl_j is the sum of delays of network transmitting terminal, link transmission, switch and network receiving terminal.

The switch delay $D_{sw}^{rt,j}$ consists of the technical delay $tD_{sw}^{rt,j}$ and the queuing delay $qD_{sw}^{rt,j}$. With respect to specific hardware, $tD_{sw}^{rt,j}$ is a bounded constant and $qD_{sw}^{rt,j}$ relies on the queuing strategy. Suppose that $RTVLE_{sw_i}^x$ is the real-time traffic concentration at the output port x of the switch sw_i and the arrival curve of real-time traffic of vl_j is $\alpha_j^{rt} = r_j^{rt}t + b_j^{rt}$. Then the arrival curve of real-time data streams is

$$\alpha^{rt} = \sum_{j \in RTVLE_{sw_i}^x} \alpha_j^{rt}$$

Suppose l_{max}^{be} is the maximum length of Ethernet data frame of the best-effort traffic and $C_{sw_i}^x$ is the total service bandwidth of the output port x of the switch sw_i . Then the total service curve of the output port x of the switch sw_i is

$$\beta_{sw_i}^x(t) = C_{sw_i}^x [t - 0]^+$$

According to the conclusion 1 of corollary 6.2.1 from literature [7], we can get the service curve of real-time traffic provided by the output port x of the switch sw_i :

$$\beta_{sw_i}^{x,rt} t = \left[\beta_{sw_i}^x t - l_{max}^{be} \right]^+ = C_{sw_i}^x \left[t - \frac{l_{max}^{be}}{C_{sw_i}^x} \right]^+. \quad (3)$$

Then the service rate and service delay of real-time traffic at the output port x of the switch sw_i are as follows:

$$\begin{aligned} R_{sw_i}^{x,rt} &= C_{sw_i}^x, \\ T_{sw_i}^{x,rt} &= \frac{l_{max}^{be}}{R_{sw_i}^{x,rt}}. \end{aligned} \quad (4)$$

Suppose the maximum data backlog of transmission

time is $\sum_{j \in RTVL_{sw_i}^x, j \neq k} b_j^{rt}$ and the maximum frame length

of real-time traffic of vl_j is $l_{max}^{rt,j}$. According to the corollary 6.2.1 from literature [7], then the service rate and service delay of real-time traffic of sw_i at the output port x of the switch sw_i are as follows:

$$\begin{aligned} R_{sw_i}^{x,rt,j} &= R_{sw_i}^{x,rt} - \sum_{k \in RTVL_{sw_i}^x, k \neq j} r_k^{rt}, \\ T_{sw_i}^{x,rt,j} &= T_{sw_i}^{x,rt} + \frac{\left(\sum_{k \in RTVL_{sw_i}^x, k \neq j} b_k^{rt} + l_{max}^{rt,j} \right)}{R_{sw_i}^{x,rt}}. \end{aligned} \quad (5)$$

Suppose SW is the collection of all switches that the virtual link vl_j passes through. Then the service rate and service delay of real-time traffic of vl_j throughout the cascaded network system are as follows:

$$\begin{aligned} R_{sw}^{rt,j} &= \min(R_{sw_i}^{x,rt,j}), \\ T_{sw}^{rt,j} &= \sum_{sw_i \in SW} T_{sw_i}^{x,rt,j} \end{aligned} \quad (6)$$

The switch queuing delay upper bound of real-time traffic of vl_j is:

$$qD_{sw}^{rt,j} = T_{sw}^{rt,j} + \frac{b_j^{rt}}{R_{sw}^{rt,j}} \quad (7)$$

Then the switch delay upper bound $D_{sw}^{rt,j}$ of real-time traffic of vl_j is:

$$D_{sw}^{rt,j} = \sum_{sw_i \in SW} tD_{sw_i}^{rt,j} + qD_{sw}^{rt,j}. \quad (8)$$

Assume that $max_{src} l_{src}^{be}$ and $max_{src} l_{src}^{rt,j}$ represent the maximum frame length of best-effort traffic of the network transmitting terminal src and the maximum frame length of real-time traffic of vl_j respectively. The calculation of network transmitting terminal delay upper bound $D_{src}^{rt,j}$ of real-time traffic of vl_j resembles that of switch delay upper bound. The specific formula is as follows:

$$\begin{aligned} R_{src}^{rt} &= C_{src}^{send}, \\ T_{src}^{rt} &= \frac{max_{src} l_{src}^{be}}{R_{src}^{rt}}, \\ R_{src}^{rt,j} &= R_{src}^{rt} - \sum_{k \in svl_{src}, pr_k = pr_t, k \neq j} r_k^{rt}, \\ T_{src}^{rt,j} &= T_{src}^{rt} + \frac{\left(\sum_{k \in svl_{src}, pr_k = pr_t, k \neq j} b_k^{rt} + max_{src} l_{src}^{rt,j} \right)}{R_{src}^{rt}}, \\ D_{src}^{rt,j} &= T_{src}^{rt,j} + \frac{b_j^{rt}}{R_{src}^{rt,j}} \end{aligned} \quad (9)$$

Suppose that real-time traffic of vl_j passes a total of n_l physical links whose available bandwidth is $\{plb_1, plb_2, \dots, plb_{n_l}\}$ and that the frame length is f_j . Then the link transmission delay $D_{link}^{rt,j}$ of real-time traffic of vl_j is:

$$D_{link}^{rt,j} = \sum_{i=1}^{n_l} \left(\frac{f_j}{plb_i} \right). \quad (10)$$

Assume that the receiving network terminal delay of real-time traffic of vl_j is $D_{des}^{rt,j}$. Then the maximum end-to-end delay $D^{rt,j}$ of real-time traffic of vl_j is:

$$D^{rt,j} = D_{src}^{rt,j} + D_{link}^{rt,j} + D_{sw}^{rt,j} + D_{des}^{rt,j}. \quad (11)$$

Therefore, the maximum end-to-end delay of real-time traffic of vl_j is calculable and the real-time communication under that network bears certainty.

4 The simulation result and analysis

We use the OPNET software to conduct simulation experiments on the method proposed in this paper to verify its instantaneity. The whole simulation network topology is a tree structure. The specific is shown in figure 3. Three switches initiate the IEEE802.1p protocol. The network node includes the client node and the server node which are improved specifically using the method put forward in the first section. The client node which ranges from 1 to 30 is linked with the switch on the second layer while the only one server node is connected with the switch on the first layer. Each client node sends on-demand messages comprised of real-time traffic and best-effort traffic to the server node. Two transmitting modules are established at the client node and two receiving modules established at the server node to distinguish between real-time traffic and best-effort traffic.

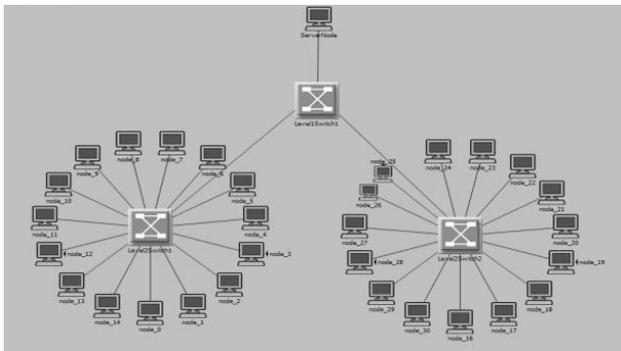


FIGURE 3 Simulation network topology

The flow simulation of client node employs the ON/OFF model. Specific parameters are shown in table 1.

TABLE 1 The parameters of flow simulation

| Parameter | Real-time traffic | Best-effort traffic |
|--------------------------------|----------------------|----------------------|
| Traffic priority | 7 | 0 |
| Start time (s) | constant(0) | constant(0) |
| ON state time (s) | constant(100) | constant(100) |
| OFF state time(s) | constant(0) | constant(0) |
| Packet interval time (s) | exponential(0.00038) | exponential(0.00704) |
| Packet size (bytes) | constant(46) | constant(1500) |
| Leaky bucket time interval (s) | constant(0.00038) | constant(0.00704) |

In experiments, we used the present Ethernet model and this paper's improved model to conduct simulation experiments with the client ranging from 1 to 30. Figure 4 and 5 present respectively the comparison of average end-to-end delays of real-time traffic, and the comparison of maximum end-to-end delays of real-time traffic when the number of clients ranges from 1 to 26. When the

References

[1] ARINC. "Avionics Full Duplex Switched Ethernet (AFDX) Network" ARINC SPECIFICATION 664P7. 2005 June
 [2] SAE. "Time-Triggered Ethernet" AS6802 2011 Nov

number of clients extends from 27 to 30, the difference between the traditional Ethernet model and the improved Ethernet model is larger.

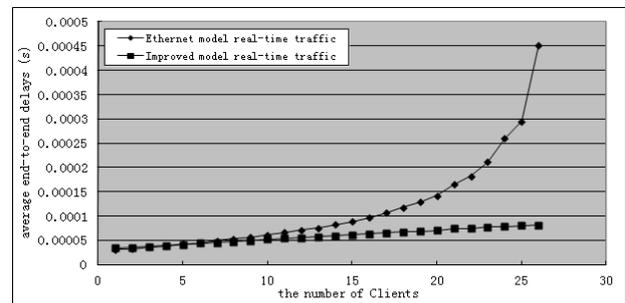


FIGURE 4 Comparison of average end-to-end delays of real-time traffic

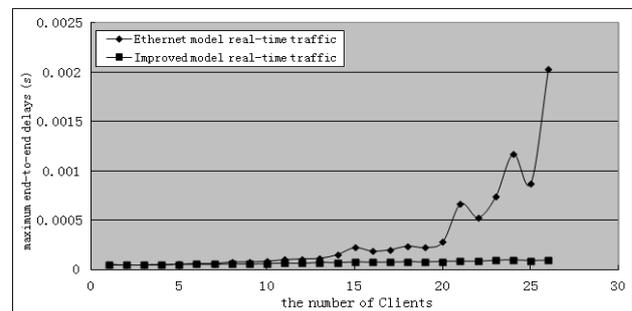


FIGURE 5 Comparison of maximum end-to-end delays of real-time traffic

From figure 4 and 5, we can see that the method proposed in this paper can improve real-time traffic performance over Switched Ethernet and at the same time be compatible with the best-effort traffic.

5 Conclusions

This paper firstly put forward a method for improving real-time communication over Switched Ethernet. Then it employed the network calculus theory to deduce the calculation formula of the maximum end-to-end delay of real-time traffic under this method. Ultimately, it used the OPNET software to conduct simulation experiments. Theoretical calculation and simulation results both show that this method can effectively improve real-time communication over Switched Ethernet. Further research on the traffic scheduling algorithm of network terminal is needed in future.

Acknowledgment

This work was financially supported by Wuhan Digital Engineering Institute, China.

[3] Marau R, Almeida L, Pedreiras P 2006 Enhancing Real-Time Communication over COTS Ethernet switches *IEEE International Workshop on Factory Communication Systems Torino* 295-302
 [4] Cruz R L 1991 *IEEE Transactions on Information Theory* 37(1) 114-31

- [5] Cruz R L 1991 *IEEE Transactions on Information Theory* 37(1) 132-41
- [6] Tranter W, Taylor D, Ziemer R 1993 A generalized processor sharing Approach to flow control in integrated services networks: in the single node case *IEEE Network* 1(3) 344-57
- [7] Boudec J-Y L, Thiran P 2004 *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet* Berlin: Springer Verlag

| Authors | |
|---|--|
|  | <p>Jun Zhao, born in October, 1980, Shulan, Jilin Province, China</p> <p>Current position, grades: He is now a PhD candidate in Wuhan Digital Engineering Institute. University Studies: He received a master in computer architecture from Wuhan Digital Engineering Institute, China. Scientific interest: network architecture theory and cloud computing. Publications: more than 10 articles. Experience: He had been a researcher and an engineer from 2002 to 2014 at Wuhan Digital Engineering Research Institute in China.</p> |
|  | <p>Zhong Ma, born in January, 1962, Wuhan, Hubei Province, China</p> <p>Current position, grades: A senior engineer of Wuhan Digital Engineering Research Institute. University Studies: He received a master in computer architecture from Wuhan Digital Engineering Institute, China. Scientific interest: network architecture theory and computer architecture. Publications: more than 30 articles. Experience: He had been a researcher and an engineer from 1991 to 2014 at Wuhan Digital Engineering Research Institute in China.</p> |
|  | <p>Wu Xiangjun, born in May, 1971, Wuhan, Hubei Province, China</p> <p>Current position, grades: A senior engineer of Wuhan Digital Engineering Research Institute. University Studies: He received a Ph.D. in Information and Communication Engineering from Huazhong University of Science and Technology, China. Scientific interest: network architecture theory, network system design and integrating. Publications: more than 20 articles. Experience: He had been a researcher and an engineer from 2000 to 2014 at Wuhan Digital Engineering Research Institute in China.</p> |

Optimal adaptive wavelet transforms without using extra additional information

Guangchun Gao*, Kai Xiong, Shengying Zhao, Cui Zhang

School of Information Science & Electronic Engineering, Zhejiang University City College, Hangzhou, China, 310015

Received 10 July 2014, www.tsi.lv

Abstract

Wavelet transforms via lifting scheme provides a general and an adaptive flexible tool for the construction of wavelet decompositions and perfect reconstruction filter banks. According to the construction of the lifting wavelet transforms, the optimal filter design method for the adaptive update wavelet transform is proposed by the authors. The optimal update filter coefficients can be acquired based on the Minimum Mean Square Error Criteria (MMSE) in the algorithm. In prediction process, take the case of LeGall 5/3 wavelet, we propose an adaptive version of this scheme that it allows perfect reconstruction without any overhead cost for the smooth signals with the jumps. Compare with other wavelet transform scheme, simulation results show that optimal adaptive wavelet transform proposed by this paper can achieve the detail signals being zero (or almost zero) at big probability and the better linear approximation for the piecewise continuous signal.

Keywords: Lifting scheme, adaptive wavelet transform, optimal update filter, MMSE

1 Introduction

Because of the better temporal and frequency properties, discrete wavelet transforms (DWT) have wide application in signal and image processing [1, 2]. It is well known that wavelet linear approximation (i.e. truncating the high frequencies) can approximate smooth functions very efficiently: it can achieve arbitrary high accuracy by selecting appropriate wavelet basis, it can concentrate the large wavelet coefficients in the low frequencies, and it has a multiresolution framework and associated fast transform algorithms. Standard wavelet linear approximation techniques cannot achieve similar results for functions which are not smooth. The jumps generate large high frequency wavelet coefficients and thus linear approximation cannot get the same high accuracy near the points of discontinuity as in the smooth regions. In fact, the jump points generate oscillations which cannot be removed by mesh refinement.

To overcome these problems within the standard wavelet transform framework, an adaptive ENO-wavelet transform has been presented in [3], which do not generate large high frequency coefficients near the jumps, but the methods use one extra bit for each stencil near the discontinuities to indicate it contains a discontinuity.

Wavelet transforms based on lifting schemes have achieved large recognition in the last years [4]. One of the major reasons for this success is their flexibility: they can be used to construct linear filter banks, but also nonlinear ones [5], e.g. using morphological filter [6]. The lifting framework has lead to designing of adaptive and nonlinear wavelet transforms recently [7-12]. The lifting scheme consists of 3 main steps: Split, Prediction and

Update. In [7, 8], an adaptive prediction step, where the adaptive switching between short and long filters based on the local edges of the input signal has been considered. In this case, the update lifting step, which is fixed, precedes the adaptive prediction step, so that the preserving of the running average of the input signal is not affected by the adaptive prediction. In [9, 10], an adaptive update lifting scheme followed by a fixed prediction has been developed. The main objective of this method is to active adaptive smoothing in the low pass signal. However, the wavelet coefficients, i.e. the high pass subbands are affected by the adaptive update process. In [9, 10], the perfect reconstruction condition for the filter coefficients was presented in the algorithm, but the method for determining the filter coefficient was not proposed. The optimal filter design method for the adaptive update wavelet transform is proposed by the authors. The optimal filter coefficients can be acquired based on the Minimum Mean Square Error Criteria (MMSE) in the algorithm.

In both above cases, either the adaptive update process or the adaptive prediction process has been adopted in the adaptive wavelet transform frameworks based on lifting schemes. Based on the adaptive lifting scheme with perfect reconstruction presented by G. Piella and the author's previous research [9, 13], this paper proposed the improved optimal adaptive wavelet transform (i.e. optimal adaptive update process and adaptive prediction process) without using extra additional information, which can achieve the better linear approximation for the piecewise continuous signals.

* *Corresponding author* e-mail: seesky88@126.com

2 Optimal adaptive wavelet transform

In general, lifting splits a signal into two sub samples, followed by at least two lifting steps, Prediction and Update. A general lifting scheme may comprise any sequence of basic lifting steps being alternatively of prediction and update type. For the adaptive wavelet transform based on lifting scheme, the wavelet transform framework (Fig. 1) by first updating and then predicting has been presented in [7, 8], so the update-then-predict lifting scheme has been adopted in this paper. Considering the previous researches of the adaptive wavelet transforms, the optimal adaptive wavelet transform (i.e. adaptive update and prediction process) based on the lifting schemes is presented.

2.1 ADAPTIVE UPDATE

In [9, 10], an adaptive wavelet transform framework had been provided for building perfect reconstruction filter banks, which did not require any additional bookkeeping to enable inversion. In this approach, a binary map is constructed based on the gradient information and the update operator is selected according to this map. Considering the better features of this approach, the update operator presented by G. Piella has been adopted in the update process of the double adaptive wavelet transform proposed by this paper. The concrete algorithms in [9, 10] has been introduced as follows:

Firstly, define the gradient vector at location n as $(v(n),w(n))=(x(n)-y(n-1),y(n)-x(n))$, In Fig.2, D is the decision set according to the value of x and y. $D(x,y)(n)=d(s(n))$, where $s(n)=|v(n)|+|w(n)|$.

For every possible outcome $d \in D$ of the decision map, we have a different update operator U_d and addition $\oplus d$. Thus, the analysis step of our adaptive update lifting scheme looks as follows:

$$x'(n) = x(n) \oplus_d U_{d_n}(y)(n), \tag{1}$$

where $d_n = D(x, y)(n)$ is the decision at location n.

We denote the subtraction which inverts $\oplus d$ by $\circ, -d$. At synthesis we can invert (1) by

$$x(n) = x'(n) \circ, -d U_{d_n}(y)(n), \tag{2}$$

We assume that the update operator U_d is a 2-tap filter and that $\oplus d$ is the standard addition followed by some scale factor. Now, the analysis step in (1) is of the form

$$x'(n) = \alpha_{d_n} x(n) + \beta_{d_n} y(n-1) + \gamma_{d_n} y(n), \tag{3}$$

And the synthesis step (presumed that d_n is known and $\alpha_d \neq 0$) is given by

$$x(n) = 1/\alpha_{d_n} (x'(n) - \beta_{d_n} y(n-1) - \gamma_{d_n} y(n)), \tag{4}$$

where $\alpha_{d_n}, \beta_{d_n}, \gamma_{d_n}$ are the lifting coefficients of the wavelet transforms. In order to have perfect reconstruction it is necessary that $\alpha_d + \beta_d + \gamma_d$ is constant for all $d \in D$. Based on a simple threshold criterion, to be precise, we assume that $D = \{0, 1\}$ and that the function d in $D(x, y)(n) = d(s(n))$ has the form $d(s) = \begin{cases} 1 & \text{if } s > T \\ 0 & \text{if } s \leq T \end{cases}$, where T is the gradient threshold.

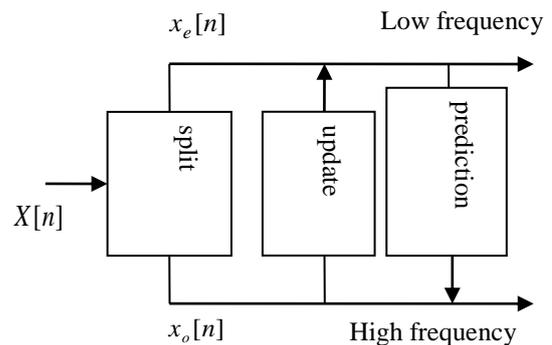


FIGURE 1 Update-then-prediction scheme

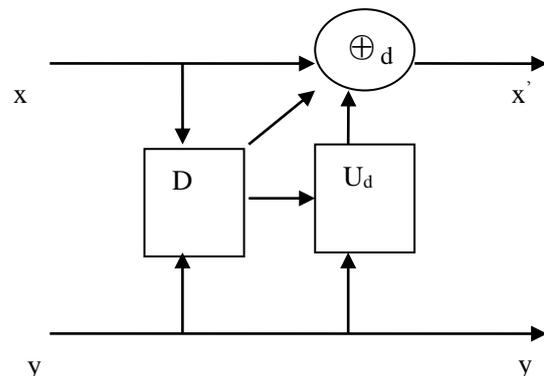


FIGURE 2 Adaptive update scheme

We have $d'(s') = d(s)$ with $T = T'$ if and only if $0 \leq \beta_0, \gamma_0 \leq 1$ and either $\beta_1, \gamma_1 \leq 0$ or $\beta_1, \gamma_1 \geq 1$

If the above conditions hold, then the reconstruction algorithm consists of the following steps:

- Compute $s' = |x' - y| + |z - x'|$,
- Let $d = [s' > T]$ and put $\gamma = \gamma_d$ and $\beta = \beta_d$,
- Compute x from $x = \frac{x' - \beta y - \gamma z}{1 - \beta - \gamma}$,

where (x,y,z) expresses (x(n),y(n-1),y(n)). The previous reconstruction algorithm based on the adaptive update process can be founded in [9, 10].

The perfect reconstruction condition for the filter coefficients is presented in [9, 10], but the method for determining the filter coefficient was not proposed in [9,

10]. For most of the signal, the smooth region is the main part of the signal. So the optimal wavelet transform filter for the smooth region was researched, and the optimal filter coefficients can be acquired based on the Minimum Mean Square Error Criteria (MMSE). The 2-level lifting wavelet transform was illustrated in Fig.3. In this lifting scheme, the mean filter was used in the prediction process. The update wavelet filter coefficients are (a_0, a_1, a_2) , and then updated data are given by

$$X'_{N-1} = a_0 X_{N-2} + a_1 X_{N-1} + a_2 X_N, \tag{5}$$

$$X'_{N+1} = a_0 X_N + a_1 X_{N+1} + a_2 X_{N+2}, \tag{6}$$

Put $X'_N = (X'_{N-1} + X'_{N+1}) / 2$, $e_N = X_N - X'_N$, then

$$X'_N = \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2})}{2}, \tag{7}$$

If μ is the mean value, where $\sum_i a_i = 1$, we can rewrite (7) as follow

$$\begin{aligned} X'_N - \mu &= \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2})}{2} - \mu \\ &= \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2})}{2} - \sum_{i=0}^2 a_i * \mu \\ &= \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2}) - \sum_{i=0}^2 2a_i * \mu}{2} \tag{8} \\ &= \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2}) - 2(a_0 + a_1 + a_2)\mu}{2} \\ &= \frac{a_0(X_N - \mu + X_{N-2} - \mu) + a_1(X_{N-1} - \mu + X_{N+1} - \mu) + a_2(X_N - \mu + X_{N+2} - \mu)}{2} \end{aligned}$$

Put $\tilde{X}'_i = X_i - \mu$, then

$$\tilde{X}'_N = \frac{a_0(\tilde{X}'_N + \tilde{X}'_{N-2}) + a_1(\tilde{X}'_{N-1} + \tilde{X}'_{N+1}) + a_2(\tilde{X}'_N + \tilde{X}'_{N+2})}{2}, \tag{9}$$

Let us simplify the Equation (9) through Omitting the symbol \sim above the X_i . The Equation (9) can be given by

$$X'_N = \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2})}{2}, \tag{10}$$

X_N Mean error (that is the Variance) is $E\{(X_N - X'_N)^2\}$, where $E\{\bullet\}$ is the mathematical expectation. In order to attain the value of the filter coefficients, $E\{(X_N - X'_N)^2\}$ must be the minimum. The optimal wavelet filter coefficients (a_0, a_1, a_2) can be acquired by using the partial differential equations (PDE) for $E\{(X_N - X'_N)^2\}$.

For wavelet filter coefficient a_0 , the partial differential equation is as below.

$$\begin{aligned} \frac{\partial E\{(X_N - X'_N)^2\}}{\partial a_0} &= \frac{\partial}{\partial a_0} E\left\{ \left[X_N - \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2})}{2} \right]^2 \right\} \\ &= -2E\left\{ \left[X_N - \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2})}{2} \right] * \frac{(X_N + X_{N-2})}{2} \right\} \\ &= -2E\left\{ \frac{\frac{X_N * X_N + X_N * X_{N-2} - a_0(X_N * X_N + 2X_N * X_{N-2} + X_{N-2} * X_{N-2})}{2} - \frac{a_1(X_{N-1} * X_N + X_{N-1} * X_{N-2} + X_{N+1} * X_N + X_{N+1} * X_{N+2})}{4} - \frac{a_2(X_N * X_N + X_N * X_{N-2} + X_{N+2} * X_N + X_{N+2} * X_{N+2})}{4}}{2} \right\} = 0 \end{aligned} \tag{11}$$

The covariance for the signal X_i has the form

$$R_{ij} = E\{X_i * X_j\}, \quad i = 1, 2, 3, \dots, N-1.$$

Due to $E(X_N * X_N) = R(0)$, $E(X_N * X_{N-2}) = R(2)$, $E(X_{N-2} * X_{N-2}) = R(0)$, $E(X_{N-1} * X_N) = R(-1)$, $E(X_{N-1} * X_{N-2}) = R(1)$, $E(X_{N+1} * X_N) = R(1)$, $E(X_{N+1} * X_{N-2}) = R(3)$, $E(X_{N+2} * X_N) = R(2)$, $E(X_{N+2} * X_{N-2}) = R(4)$, where $R(-1) \approx R(1)$, $R(-2) \approx R(2)$, the partial differential equation for a_0 is as below.

$$\begin{aligned} [R(0) + R(2)] * a_0 + \left[\frac{R(3)}{2} + \frac{3}{2} R(1) \right] * a_1 \\ + \left[\frac{R(0)}{2} + R(2) + \frac{R(4)}{2} \right] * a_2 = R(0) + R(2) \end{aligned} \tag{12}$$

For wavelet filter coefficient a_1 , the partial differential equation is as below.

$$\begin{aligned} \frac{\partial E\{(X_N - X'_N)^2\}}{\partial a_1} &= \frac{\partial}{\partial a_1} E\left\{ \left[X_N - \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2})}{2} \right]^2 \right\} \\ &= -2E\left\{ \left[X_N - \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2})}{2} \right] * \frac{(X_{N-1} + X_{N+1})}{2} \right\} \\ &= -2E\left\{ \frac{\frac{X_N * X_{N-1} + X_N * X_{N+1} - a_0(X_{N-2} * X_{N-1} + X_{N-2} * X_{N+1} + X_N * X_{N-1} + X_N * X_{N+1})}{2} - \frac{a_1(X_{N-1} * X_{N-1} + X_{N-1} * X_{N+1} + X_{N+1} * X_{N-1} + X_{N+1} * X_{N+1})}{4} - \frac{a_2(X_N * X_{N-1} + X_N * X_{N+1} + X_{N+2} * X_{N-1} + X_{N+2} * X_{N+1})}{4}}{2} \right\} = 0 \end{aligned} \tag{13}$$

Due to $E(X_N * X_{N-1}) = R(1)$, $E(X_N * X_{N+1}) = R(-1)$, $E(X_{N-2} * X_{N-1}) = R(-1)$, $E(X_{N-2} * X_{N+1}) = R(-3)$, $E(X_N * X_{N-1}) = R(1)$, $E(X_N * X_{N+1}) = R(-1)$, $E(X_{N-1} * X_{N-1}) = R(0)$, $E(X_{N-1} * X_{N+1}) = R(-2)$, $E(X_{N+1} * X_{N-1}) = R(2)$, $E(X_{N+2} * X_{N-1}) = R(3)$, $E(X_{N+2} * X_{N+1}) = R(1)$, where $R(-1) \approx R(1)$, $R(-2) \approx R(2)$, $R(-3) \approx R(3)$, the partial differential equation for a_1 is as below.

$$\frac{[3R(1)+R(3)]}{2} * a_0 + [R(0)+R(2)] * a_1 + \frac{[3R(1)+R(3)]}{2} * a_3 = 2R(1). \quad (14)$$

For wavelet filter coefficient a_2 , the partial differential equation is as below.

$$\frac{\partial E\{(X_N - X_{N+2})^2\}}{\partial a_2} = \frac{\partial}{\partial a_2} E\left\{X_N - \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2})}{2}\right\}^2$$

$$= -2E\left\{X_N - \frac{a_0(X_N + X_{N-2}) + a_1(X_{N-1} + X_{N+1}) + a_2(X_N + X_{N+2})}{2}\right\} \cdot \frac{(X_N + X_{N+2})}{2}$$

$$= -2E\left\{\frac{\frac{X_N * X_N}{2} + \frac{X_N * X_{N+2}}{2} - \frac{a_0(X_{N-2} * X_N + X_{N-2} * X_{N+2}) + X_N * X_N + X_N * X_{N+2}}{4}}{4} - \frac{a_1(X_{N-1} * X_N + X_{N-1} * X_{N+2}) + X_{N+1} * X_N + X_{N+1} * X_{N+2}}{4} - \frac{a_2(X_N * X_N + X_N * X_{N+2}) + X_{N+2} * X_N + X_{N+2} * X_{N+2}}{4}\right\} = 0. \quad (15)$$

Due to $E(X_N * X_N) = E(X_{N+2} * X_{N+2}) = R(0)$, $E(X_N * X_{N+2}) = E(X_{N-2} * X_N) = R(-2)$, $E(X_{N+2} * X_N) = R(2)$, $E(X_{N-2} * X_{N+2}) = R(-4)$, $E(X_{N-1} * X_{N+2}) = R(-3)$, $E(X_{N-1} * X_N) = R(-1)$, $E(X_{N+1} * X_N) = R(1)$, $E(X_{N+1} * X_{N+2}) = R(-1)$, $E(X_N * X_{N+2}) = R(-2)$, where $R(-1) \approx R(1)$, $R(-2) \approx R(2)$, $R(-3) \approx R(3)$, $R(-4) \approx R(4)$, the partial differential equation for a_2 is as below.

$$\frac{[2R(2)+R(0)+R(4)]}{2} * a_0 + \frac{[3R(1)+R(3)]}{2} * a_1 + [R(0)+R(2)] * a_2, \quad (16)$$

$$= R(0) + R(2)$$

Equation (10), (12) and (16) are the equation group, In order to acquire the solution of equations, that is the filter coefficient (a_0, a_1, a_2) , the signal covariance R_{ij} must be known. Using the signal sample value, R_{ij} can be obtained, so the optimal filter coefficients can also be acquired.

So far, the design method for the optimal wavelet filter had been carefully illustrated in this section. In the next section, the experiment results will be shown.

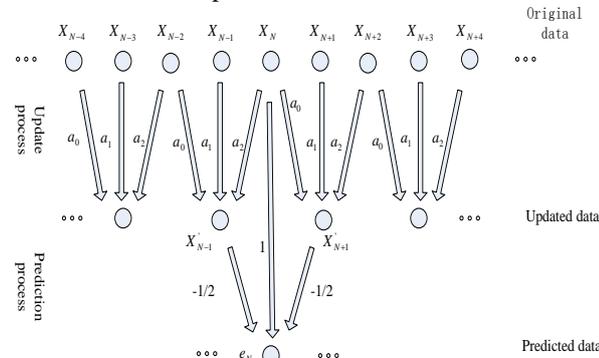


FIGURE 3 2-level lifting wavelet transform

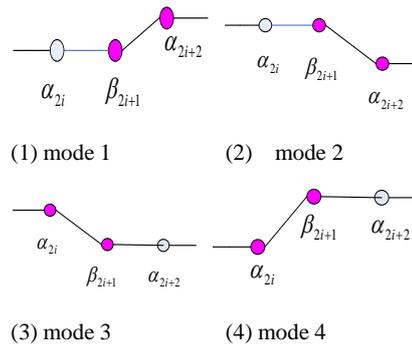


FIGURE 4 Four modes

2.2 ADAPTIVE PREDICTION

Because of the update-then-prediction lifting scheme adopted in this paper, the update process cannot be affected by the prediction process using the adaptive transform scheme. In order to implement adaptive prediction algorithm, there are two crucial points in designing wavelet transform scheme: The first is to detect the jumps in the signal. In order to avoid generating extra edge information, we use the updated data for detecting jumps. This scheme is reversible. When the data has a jump, the position of this jump is preserved after updating. The second is how to use one-sided data near jumps to avoid oscillations. The jumps generate large high frequency wavelet coefficients if we adopt the same wavelet filter as the smooth region. So the jumps can be predicted by the left or the right data of the jumps in this paper. Assuming that β_{2i+1} is the jump (predicted point), α_{2i} is its left side data (updated data) and α_{2i+2} is its right side data (updated data), the three-point relationship can be summed up the four modes in figure 4. The four modes can be introduced carefully as follows:

2.2.1 Mode 1

If $\alpha_{2i+2} > \alpha_{2i}, \beta_{2i+1} < (\alpha_{2i+2} + \alpha_{2i})/2$, the relative of $\alpha_{2i}, \beta_{2i+1}, \alpha_{2i+2}$ can be summed up Mode 1 scheme. For this scheme, β_{2i+1} is the jump of the prediction data, and α_{2i+2} is the jump of the updated data. For the LeGall 5/3, the lifting scheme can be shown in figure 5, so update-prediction process as follow:

$$\text{Update: } \alpha'_{2i} = \alpha_{2i} + (\beta_{2i-1} + \beta_{2i+1})/4$$

$$\text{Prediction: } \beta'_{2i+1} = \beta_{2i+1} - (\alpha'_{2i} + \alpha'_{2i+2})/2$$

If e_{2i} denotes the mean linear error of the updated data the following equations can be acquired.

$$|e_{2i}| = \left| \alpha'_{2i} - (\alpha'_{2i-2} + \alpha'_{2i+2})/2 \right|, \quad (17)$$

$$|e_{2i+2}| = \left| \alpha'_{2i+2} - (\alpha'_{2i} + \alpha'_{2i+4}) / 2 \right|, \tag{18}$$

For the mode 1, α'_{2i} can be updated in the smooth fields, so $\alpha'_{2i} \approx \alpha'_{2i-2}$.for the Equation (17) can be written as follows

$$|e_{2i}| = \left| (\alpha'_{2i} - \alpha'_{2i+2}) / 2 \right|. \tag{19}$$

Through decomposing of the Equation (18), equation (20) can be acquired.

$$|e_{2i+2}| = \left| (\alpha'_{2i+2} - \alpha'_{2i}) / 2 + (\alpha'_{2i+2} - \alpha'_{2i+4}) / 2 \right|. \tag{20}$$

In the update process, α'_{2i+2} can be updated by the value of the β_{2i+1} , the update process as follows:

$$\alpha'_{2i+2} = \alpha_{2i+2} + (\beta_{2i+1} + \beta_{2i+3}) / 4, \tag{21}$$

$$\alpha'_{2i+4} = \alpha_{2i+4} + (\beta_{2i+3} + \beta_{2i+5}) / 4. \tag{22}$$

For the mode 1, α_{2i+2} and α_{2i+4} locate the same smooth fields, that is $\alpha_{2i+2} \approx \alpha_{2i+4} \cdot \beta_{2i+3}$ and β_{2i+5} locate the same smooth fields, that is $\beta_{2i+3} \approx \beta_{2i+5}$. However, β_{2i+1} and β_{2i+3} locate at two sides of the jump, and $\beta_{2i+1} < \beta_{2i+3}$, according to equation (21) and (22), equation (23) can be acquired.

$$\alpha'_{2i+2} < \alpha'_{2i+4}. \tag{23}$$

According to equation (19), (20) and (23), $|e_{2i}| > |e_{2i+2}|$ and $e_{2i} * e_{2i+2} < 0$ can be known, we can get the prediction equations for this scheme $\beta'_{2i+1} = \beta_{2i+1} - \alpha_{2i}$, where β'_{2i+1} is the predicted high frequency coefficient.

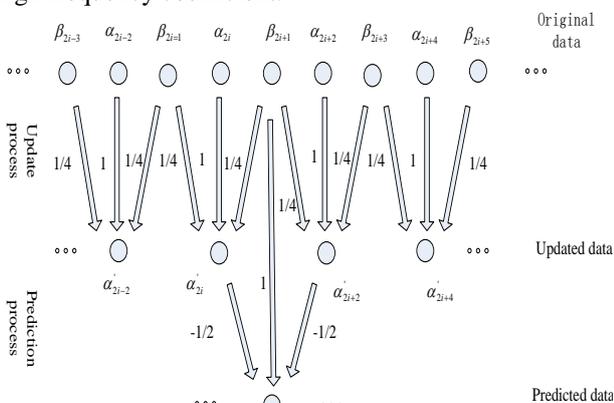


FIGURE 5 2-level lifting wavelet transform

2.2.2 Mode 2

If $\alpha_{2i+2} < \alpha_{2i}, \beta_{2i+1} > (\alpha_{2i+2} + \alpha_{2i}) / 2$, the relative of $\alpha_{2i}, \beta_{2i+1}, \alpha_{2i+2}$ can be summed up Mode 2 scheme. For this scheme, β_{2i+1} is the jump of the prediction data, and α_{2i+2} is the jump of the updated data. In the update process, α_{2i+2} can be affected by the value of β_{2i+1} . When we calculate the linear prediction error of the updated data, the left prediction errors of α_{2i+2} (that is α_{2i+4} point) are usually lager than that of α_{2i} . Based on deduction method as mode 1, we can get the prediction equations for this scheme $\beta'_{2i+1} = \beta_{2i+1} - \alpha_{2i}$, where β'_{2i+1} is the predicted high frequency coefficient.

2.2.3 Mode 3

If $\alpha_{2i+2} < \alpha_{2i}, \beta_{2i+1} < (\alpha_{2i+2} + \alpha_{2i}) / 2$, the relative of $\alpha_{2i}, \beta_{2i+1}, \alpha_{2i+2}$ can be summed up in Mode 3 scheme. For this scheme, β_{2i+1} is the jump of the prediction data, and α_{2i+2} is the jump of the updated data. In the update process, α_{2i} can be affected by the value of β_{2i+1} . When we calculate the linear prediction error of the updated data, the right prediction errors of α_{2i} (that is α_{2i-2} point) are usually lager than that of α_{2i+2} . Based on deduction method as mode 1, we can get the prediction equations for this scheme $\beta'_{2i+1} = \beta_{2i+1} - \alpha_{2i+2}$, where β'_{2i+1} is the predicted high frequency coefficient.

2.2.4 Mode 4

If $\alpha_{2i+2} > \alpha_{2i}, \beta_{2i+1} > (\alpha_{2i+2} + \alpha_{2i}) / 2$, the relative of $\alpha_{2i}, \beta_{2i+1}, \alpha_{2i+2}$ can be summed up in Mode 4 scheme. For this scheme, β_{2i+1} is the jump of the prediction data, and α_{2i+2} is the jump of the updated data. In the update process, α_{2i} can be affected by the value of β_{2i+1} . When we calculate the linear prediction error of the updated data, the right prediction errors of α_{2i} (that is α_{2i-2} point) are usually lager than that of α_{2i+2} . Based on deduction method as mode 1, we can get the prediction equations for this scheme $\beta'_{2i+1} = \beta_{2i+1} - \alpha_{2i+2}$, where β'_{2i+1} is the predicted high frequency coefficient.

According to the previous analysis, the adaptive prediction algorithm consists of the following steps:

For each index i:

- Calculate the linear error e_{2i} sequence of the update data α_{2i} sequence from $e_{2i} = \alpha_{2i} - (\alpha_{2i-2} + \alpha_{2i+2}) / 2$

- For the e_{2i} sequence, the multiplying value of the two adjacent numbers is calculated, that is the value of $e_{2i} \times e_{2i+2}$.

If this value is negative, β_{2i+1} are the jumps, then the next step will be performed.

Else $\beta'_{2i+1} = \beta_{2i+1} - (\alpha_{2i} + \alpha_{2i+2}) / 2$, we get the high frequency coefficient sequence.

- If the α_{2i} is the jump of the updated data, the value of $|e_{2i-2}|$ is larger. Otherwise, the value of $|e_{2i+4}|$ is larger. Comparing with the value between $|e_{2i-2}|$ and $|e_{2i+4}|$, the prediction algorithm using the left side data or the right side data of the jump can be determined.

If $|e_{2i-2}| > |e_{2i+4}|$ then

$$\beta'_{2i+1} = \beta_{2i+1} - \alpha_{2i}.$$

Else If $|e_{2i-2}| > |e_{2i+4}|$ then

$$\beta'_{2i+1} = \beta_{2i+1} - \alpha_{2i+2}$$

Else $\beta'_{2i+1} = \beta_{2i+1} - (\alpha_{2i} + \alpha_{2i+2}) / 2$.

Through the previous discussion, we know that the adaptive prediction algorithm can be reconstructed without using extra additional information. The inverse transform algorithm is the inverse process of the forward wavelet transform.

3 Simulation results

Next, we consider a piecewise smooth function defined by

$$f(x) = \begin{cases} 0 & 0 \leq x < 0.2 \\ -50x - 5 & 0.2 \leq x < 0.4 \\ 10 \sin(4\pi x + 0.8\pi) - 1 & 0.4 \leq x < 1.1 \\ 5e^{2x} - 100 & 1.1 \leq x < 1.6 \\ 0 & 1.6 \leq x \leq 2.0 \end{cases}$$

Fig.6 shows the function $f(x)$. In order to study on the performance of the optimal adaptive wavelet transform, the five different wavelet transform scheme are listed as follows:

Non-adaptive wavelet transform: Update-then-Prediction scheme. The coefficient (1/4, 1/2, 1/4) for the wavelet filter is chosen in the update process, and the filter is the same as that of (5, 3) wavelet in the prediction process.

Adaptive update wavelet transform: Adaptive update and non-adaptive prediction scheme. We adopt the adaptive update scheme proposed in [9, 10]. The filter coefficients are $(a_0 = 1/4, a_1 = 1/2, a_2 = 1/4)$ for smooth region, and the filter coefficients are $(a_0 = 0, a_1 = 1, a_2 = 0)$ for the jumps. The filter is the same as that of (5, 3) wavelet in the prediction process.

Adaptive prediction wavelet transform: update and

adaptive prediction scheme. The filter coefficients are $(a_0 = 1/4, a_1 = 1/2, a_2 = 1/4)$ in the update process. We adopt the adaptive prediction algorithms proposed by this paper.

Adaptive wavelet transform: Adaptive update and adaptive prediction scheme. We adopt the adaptive update scheme proposed in [9, 10]. The filter coefficients are $(a_0 = 1/4, a_1 = 1/2, a_2 = 1/4)$ for smooth region, and the filter coefficients are $(a_0 = 0, a_1 = 1, a_2 = 0)$ for the jumps. The adaptive prediction algorithms proposed by this paper are chosen.

Optimal adaptive wavelet transform: Optimal adaptive update and adaptive prediction scheme. we adopt the optimal adaptive update scheme proposed by this paper. Through calculating the covariance R_{ij} of the tested signal, the filter coefficients $(a_0 = 0.4990, a_1 = 0.3775, a_2 = 0.1241)$ are acquired. Considering the perfect reconstruction condition for the filter in [9, 10], we selected $(a_0 = 5/10, a_1 = 4/10, a_2 = 1/10)$ as the optimal coefficients of wavelet filter for smooth region, and the filter coefficients are the filter coefficients $(a_0 = 0, a_1 = 1, a_2 = 0)$ for the jumps. The adaptive prediction algorithms proposed by this paper are chosen.

Using the non-adaptive wavelet transform scheme, one level wavelet decomposition for the $f(x)$ is shown in Fig.7. The left part corresponds to the low frequency coefficients and the right part the high frequency coefficients. Because we get the similar low frequency coefficients using the wavelet transform scheme listed in this chapter, the transformed high frequency coefficients will be researched. In Fig.8, 9, 10 and 11, we present the adaptive prediction, adaptive update, adaptive wavelet and optimal adaptive wavelet transform 1-level decomposition high frequency coefficients respectively. For the adaptive prediction or adaptive update wavelet transform scheme, we notice that there are some large high frequency coefficients near the discontinuities. On the other hand, no large high frequency coefficients are present in the optimal adaptive wavelet transform. This illustrates that the optimal adaptive wavelet coefficients have better distribution than other wavelet transform listed in this paper, i.e., no large coefficients in the high frequencies and the energy is concentrated in the low frequency end. According to the signal energy calculation

formula: $SignalEnergy = \sum_{i=1}^n X_i * X_i$, where X_i is the

value of the signal sequence, the 1-level decomposition high coefficient energy of the different wavelet transform scheme listed in Table 1. From the table 1, the optimal adaptive wavelet transform has better energy concentration. The performances of this wavelet transform scheme meet the demand of the image compression.

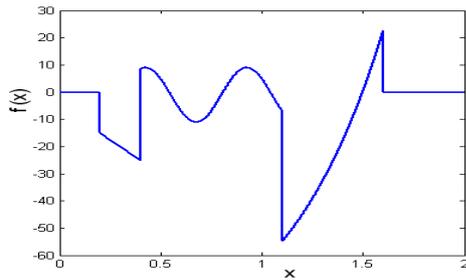


FIGURE 6 A piecewise smooth signal

Using the wavelet transform scheme listed in this chapter, the different wavelet linear approximations are shown in Figure 12, 13, 14, 15 and 16 respectively. From the Figure 12, we notice that the non-adaptive wavelet linear approximation generate the oscillations near the discontinuity (Arrow shows the location in figure). In Figure 13, 14 and 15, it is evident that these schemes can reduce the oscillations near the discontinuity (Arrow shows the location in figures). In Fig. 16, the oscillations near the discontinuity can be basically eliminated. Comparing with the original signal $f(x)$, the linear approximation has the smaller distortion. With studying on the optimal adaptive wavelet transform scheme, the simulation results demonstrate that it can eliminate the oscillations near the discontinuity, and has better linear approximation.

TABLE 1 1-level decomposition high frequency coefficients energy of the different wavelet transform scheme

| Wavelet transform scheme | Energy |
|---------------------------------------|-------------|
| Non-adaptive wavelet transform | 651.053 |
| Adaptive update wavelet transform | 1.0413e+003 |
| Adaptive prediction wavelet transform | 65.0251 |
| Adaptive wavelet transform | 56.3078 |
| Optimal adaptive wavelet transform | 0.8065 |

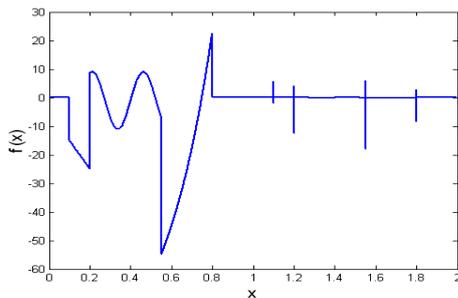


FIGURE 7 Non-adaptive wavelet 1-level decomposition for the piecewise smooth signal

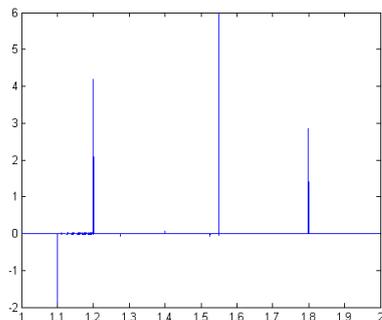


FIGURE 8 1-level decomposition high frequency coefficients for the adaptive prediction wavelet transform

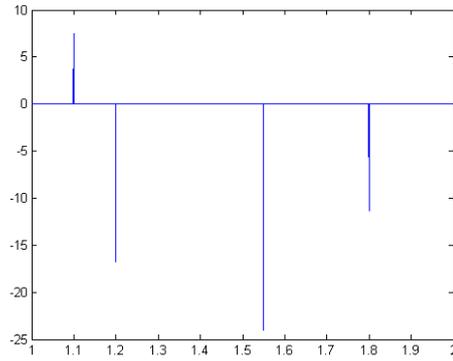


FIGURE 9 1-level decomposition high frequency coefficients for the adaptive update wavelet transform

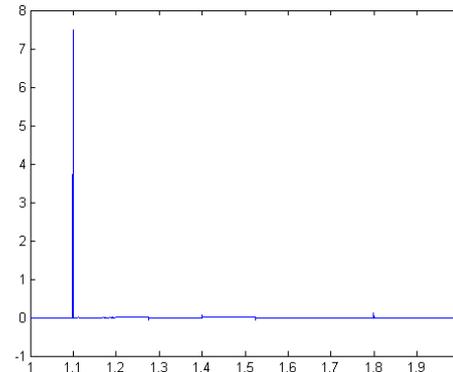


FIGURE 10 1-level decomposition high frequency coefficients for the adaptive wavelet transform

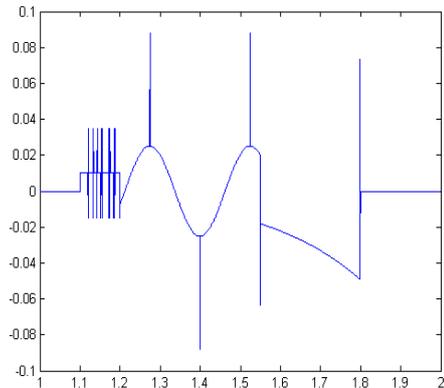


FIGURE 11 1-level decomposition high frequency coefficients for the optimal adaptive wavelet transform

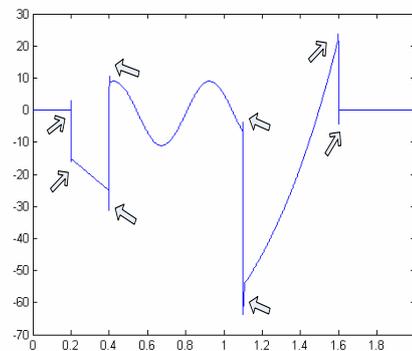


FIGURE 12 1-level linear approximation for the non-adaptive wavelet transform

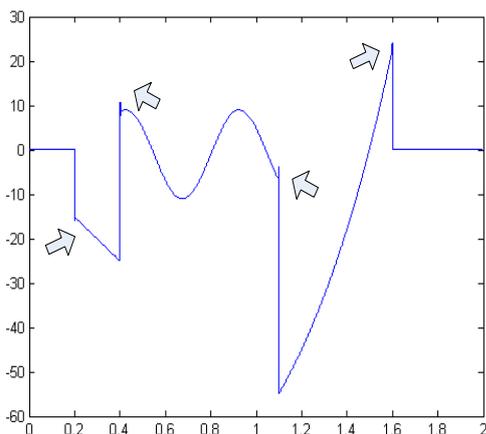


FIGURE 13 1-level linear approximation for the adaptive prediction wavelet Transform

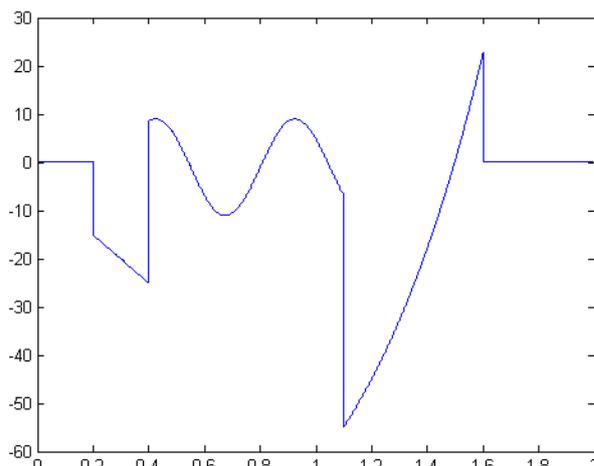


FIGURE 16 1-level linear approximation for the optimal adaptive prediction wavelet transform

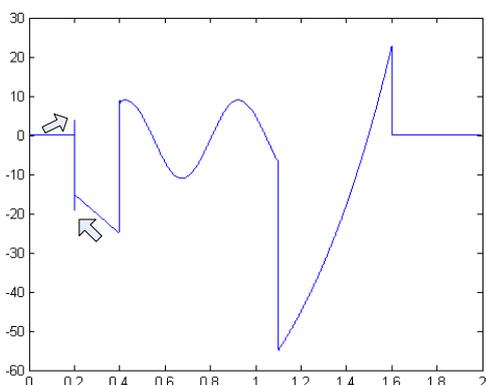


FIGURE 14 1-level linear approximation for the adaptive update wavelet Transform

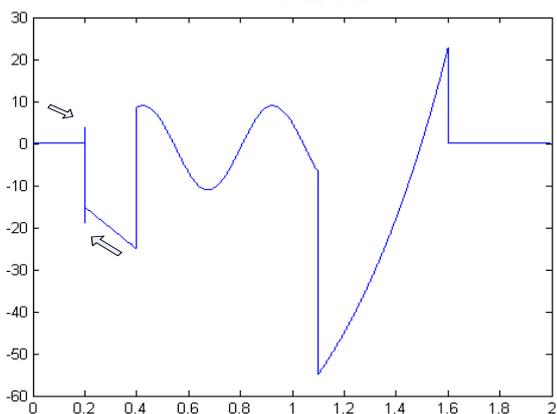


FIGURE 15 1-level linear approximation for the adaptive wavelet Transform

4 Conclusions

Based on the double adaptive wavelet transform proposed in this paper, the simulation results demonstrate that it has better linear approximation of the smooth piecewise function, and can also reduce the high frequency coefficients. At the same time, comparing with the algorithms in [13], the optimal adaptive wavelet transform can be implemented without sending any side information. A wavelet based image compression algorithm usually consists of three steps, namely truncating the real valued wavelet coefficients into a finite set of fixed values so that they can be used in coding process. In this step, the small wavelet coefficients are usually quantized to zero. Therefore, the smaller wavelet coefficients a transform generates the better compression it achieves. The double adaptive wavelet transform presented by the authors meets the demand of image compression. How to use this algorithm together with quantization and coding steps to form complete image compression algorithms will be researched in the future.

Acknowledgments

This work was supported by Zhejiang Provincial Natural Science Foundation of China (No.Y1110632) and (LY12F01017), Supported by the construct program of the key discipline in Hangzhou.

References

- [1] Ramu Satyabama, Annadurai 2013 Image compression using space adaptive lifting scheme *Journal of Computer Science* 7(11) 1704-10
- [2] Kaaniche M, Pesquet-popescu B, Pesquet J-C, Benazza-benyahia A 2012 Adaptive lifting schemes with a global l1 minimization technique for image coding *IEEE International Conference on Image Coding United states*
- [3] Chan T F, Zhou H M 1999 Adaptive ENO-wavelet transforms for discontinuous functions, Technical Report 99-21, Dept. of Math, UCLA
- [4] Sweldens W 1997 The lifting scheme: a construction of second generation wavelets *SIAM J. Math. Anal.* 29(2) 511-46

[5] Zijiang Yang, Ligang Cai, Lixin Gao, Huaqing Wang 2012 Adaptive redundant lifting wavelet transforms based on fitting for fault feature extraction of roller bearing *Sensors* **12** 4381-98

[6] Heijmans H J A M, Goutsias J 2000 Nonlinear multiresolution signal decomposition schemes. Part II: morphological wavelets *IEEE Transaction on Image Processing* **9**(11) 1897-913

[7] Claypoole R, Baraniuk R, Nowak R 1997 Nonlinear wavelet transforms for image coding *In Proceedings of the 31st Asilomar Conference on Signals, Systems, and Computers* **1** pp. 662

[8] Claypoole R, Baraniuk R, Nowak R 1998 Adaptive wavelet transforms via lifting *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 12-15, Seattle, Washington*

[9] Piella G, Heijmans H J A M 2001 *Adaptive lifting schemes with perfect reconstruction* PNA-R0104, CWI, Amsterdam, February 28

[10] Heijmans H J A M, Pesquet-Popescu B, Piella G 2002 *Building nonredundant adaptive wavelets by update lifting* PNA-R0212, CWI, Amsterdam, May 31

[11] Abhayaratne G C K 2003 Spatially adaptive integer lifting with no side information for lossless video coding *Picture Coding Symposium (PCS 03)* pp.495-500

[12] Yu Wu, Guoyin Wang, Neng Nie 2001 Adaptive lifting scheme of wavelet transforms for image compression *Proceeding of SPIE, 4391 on Wavelet Application III*, pp. 154-160

[13] Gao Guang-chun, Yao Qing-dong 2004 New method for calculating lifting coefficients of biorthogonal wavelet *Journal of Zhejiang University: Engineering Science* **38**(12) 1665-8 (in Chinese)

| Authors | |
|---|--|
|  | <p>Guangchun Gao</p> <p>Current position, grades: an associate professor of Zhejiang University City College. University studies: the PhD degrees in Communication and Information System from the Zhejiang University, in 2004. Scientific interests: compressed sensing, image processing and coding, and communication signal processing and system design.</p> |
|  | <p>Kai Xiong</p> <p>Current position, grades: works for the Zhejiang University City College in Hangzhou. University studies: BSc degree from the Zhejiang University in 1999 and MSc and PhD degrees from the Zhejiang University in 2002 and 2005 respectively, all in Optical Engineering Scientific interests: optical imaging, image processing, and optical measurement</p> |
|  | <p>Zhao Shengying</p> <p>Current position, grades: works for the Zhejiang University City College in Hangzhou. University studies: BSc degree in department of Electrical Engineering from Zhejiang University in 1991 and MSc degrees in department of Information and communication engineering from Zhejiang University in 2006. Scientific interests: image processing and compressive sensing</p> |
|  | <p>Zhang Cui</p> <p>Current position, grades: works for the Zhejiang University City College in Hangzhou. University studies: BSc degree in electronics and information engineering from the Northwestern Polytechnical University, Xi'an, China, and the MSc degree in weapon system and application in Engineering from the Northwestern Polytechnical University, China in 2003. Scientific interests: multimedia technology, image processing and Information Processing</p> |

A large-scale MIMO channel information feedback algorithm based on compressed sensing

Jing Jiang*, Shuang Xu, Guangyue Lu, Yongbin Xie

School of Communication and Information Engineering, Xi'an University of Posts & Telecommunications, Xi'an, 710061, China

Received 10 July 2014, www.tsi.lv

Abstract

In order to effectively reduce the feedback overhead of channel state information (CSI), a channel state information feedback algorithm based on compressed sensing was proposed for Large-scale MIMO system. Firstly considering the sparsity of spatial-frequency domain for the large-scale MIMO channel, the channel information was compressed in space domain firstly and in frequency domain subsequently, the receiver acquired the measurement vector based on compressed sensing algorithm; then feedback. CSI observations to the transmitter according to the proposed adaptive feedback protocol, at last the transmitter reconstructed CSI based on the Basis Pursuit (BP) algorithm. It is show in stimulation results that the proposed algorithm can acquire similar BER performance with perfect channel information feedback. The proposed algorithm, which feedbacks the compressed channel information, not only can significantly reduce the feedback overhead, but also ensure that large-scale MIMO performance gain.

Keywords: Large-scale MIMO, Channel State information feedback, Compressed Sensing

1 Introduction

As applications of wireless networks become more and more diverse and users in wireless network increase very rapidly, wireless data grows up dramatically. Enhancing network capacity is still a challenge in the future of wireless communications. The multiple antenna technology has been an important way to improve network capacity. Large-scale MIMO, also known as Massive MIMO, configures a large number (from tens to thousands) of antennas in base station. It can not only greatly improve the system capacity but also reduce transmit power for the large-scale array gain. So Large-scale MIMO has become a hot topic in the research of 5G wireless communication technology [1].

However it is an important issue that how the transmitter acquires the instantaneous and accurate channel state information (CSI). In the time division duplex (TDD) system, the transmitter obtained CSI using the channel reciprocity [2]; in the frequency division duplex (FDD) system, the transmitter obtained CSI by the feedback from the receiver. As the number of antennas increases, the amount of channel state information feedback also grows linearly. Popular CSI feedback schemes, such as vector quantization and codebook-based approaches, may not appropriate for the large-scale MIMO. When the feedback overhead is reduced as to the limited bandwidth and time resource, the accuracy of the feedback will be difficult to ensure [1]. FDD mode is the major mode of current wireless communication systems and will also be one of working modes in the future

wireless communications. Therefore, it is necessary to study CSI feedback mechanisms and algorithm in FDD system of Large-scale MIMO to reduce the feedback overhead significantly and ensure a large-scale antenna array gain.

For the CSI feedback of large-scale MIMO on the FDD mode, the following documents provided a good basis for research. Literature [3] assumed the base station and the user shared a common set of training signals in advance, and then proposed open-loop and closed-loop training frameworks. In the open-loop training, the base station transmitted training signals in a round-robin manner, and the user equipment successively estimated the current channel using long-term channel statistics and the previously received training signals. In the closed-loop training, users only received the training signals with best quality. Numerical results proved that this feedback method could obtain better performance for large-scale MIMO systems, especially when the SNR is low. Literature [4] proposed a non-coherent trellis coded quantization (NTCQ) feedback algorithm which combined channel coding with codebook design. Although the above algorithm can reduce the amount of feedback overhead, their computational complexity or encoding complexity linearly scales up with the number of antennas.

Compressed sensing utilizes the sparsity of signals to reduce the number of sampling and breaks the limit of Nyquist sampling theorem. It can not only reduce the sampling number of signals but also achieve good performances for the immense improvement of the sparse

* *Corresponding author* e-mail: jiangjing18@gmail.com

signal reconstruction [4, 5]. If signals are compressible or sparse in a transform domain, high-dimension signals can be projected onto a low-dimension space through the observation matrix which is uncorrelated with the transformation basis. And then the original signal can be reconstructed from a small number of projections which contains sufficient information of original signals. In this theoretical framework, the sampling rate is not determined by the signal bandwidth, but the structure and contents of information contained in signals. Recently, compressed sensing has been applied to signal processing and communications [7]. For the downlink transmission that services a large number of users, literature [8] proposed a method for channel estimation and user selection which established that full channel state information for each self-selecting user. Full channel state information can be obtained via compressed sensing without increasing the uplink feedback overhead. Literature [9] proposed a compressive sensing feedback method based on the opportunistic feedback protocol. The feedback resources were shared and were opportunistically accessed by high-quality users whose link quality exceeded a certain fixed threshold. Reference [10] proposed channel feedback reduction techniques based on compressive sensing, in which the transmitter can obtain channel information with acceptable accuracy under substantially reduced feedback overhead. At last simulation results showed that CS-based feedback can achieve near optimal rank-1 beamforming performance.

Based on the above-described research, we put forward a compressed sensing feedback algorithm for Massive MIMO system. The channel information is compressed in space domain firstly and in frequency domain subsequently. Compressed channel information is feedback to the receiver according to a new adaptive feedback mechanism. The receiver acquires accurate CSI recovery using Basis Pursuit (BP) reconstruction algorithm. In especial, the adaptive feedback mechanism will modify the compressing rate based on the channel sparsity to improve the feedback efficiency and assure the feedback performance. At last, numerical results proved that the proposed CSI feedback algorithm can not only greatly reduce the feedback overhead, but get similar BER performance to the perfect CSI feedback.

2 Compressed sensing theory

The theory of compressed sensing (CS) mainly includes three steps: get the sparsifying transformation of original signals; acquire the measurement vector of sparsifying signals; and reconstruct original signals.

2.1 SPARSIFYING TRANSFORMATION

Original signal \mathbf{x} of length N can be expressed by the following sparsifying transformation:

$$\mathbf{S} = \Psi \mathbf{x}, \tag{1}$$

where \mathbf{S} is the sparse transformation of \mathbf{x} , Ψ is an $N \times N$ sparsifying-basis. In this case, original signals \mathbf{x} have K non-zero coefficients on this sparsifying-basis and \mathbf{x} are called as K -sparse. Discrete cosine transform (DCT) matrix, discrete fourier transform (DFT) matrix and wavelet transform (DWT) matrix are some typical sparsifying-basis. These transformations are usually orthogonal and (1) can be expressed as following too:

$$\mathbf{x} = \Psi^T \mathbf{S}, \tag{2}$$

where $(\cdot)^T$ represents matrix transpose.

2.2 DESIGN THE MEASUREMENT MATRIX

In the measurement of compressed sensing, it does not directly measure the K -sparse original signals \mathbf{x} itself. Instead, the signal \mathbf{x} is projected with a set of measurement matrix Φ onto CS measurement vector \mathbf{y} :

$$\mathbf{y} = \Phi \mathbf{x}. \tag{3}$$

We substitute (2) into an equation (3), the equation (3) can be rewritten as:

$$\mathbf{y} = \Phi \Psi^T \mathbf{S} = \Theta \mathbf{S}, \tag{4}$$

where \mathbf{y} is an $M \times 1$ CS measurement vector, and M satisfies $M \geq K \log_2(N/K)$. Θ is an $M \times N$ sensing matrix. If original signals are K -sparse and Φ satisfies the Restricted Isometry Property (RIP) [7], K coefficients can be accurately reconstruct from M measurements.

The related literature proved that independent and identically distributed Gaussian random measurement matrix can become universal compressed sensing measurement matrix.

2.3 SIGNAL RECONSTRUCTION

We can obtain the sparse coefficients \mathbf{S} via solving inverse problem of formula (4), then the K -sparse signal \mathbf{x} can be reconstructed from M dimensional measurement vectors \mathbf{y} . This can be formulated as the following l_0 norm (also called 0-norm, that is the number of non-zero elements in the vector) minimization problem:

$$\min \|\mathbf{S}\|_{l_0} \quad s.t. \quad \mathbf{y} = \Phi \Psi^T \mathbf{S}. \tag{5}$$

When the estimation of sparse transformation \mathbf{S}' is

solved by (5), and then original signals \mathbf{x} can be reconstructed as \mathbf{x}' via $\mathbf{x}' = \Psi \mathbf{S}'$. Reconstruction algorithms contain Matching tracking algorithm (MP) Orthogonal Matching pursuit (OMP) algorithm and Linear Programming (LP), basic Pursuit (BP), etc.

3 System Model

In the Large-scale MIMO system, assume that the number of subcarriers is N_c , the number of OFDM symbol is N_t , the number of transmit antennas is $N_t \gg 1$, the number of receive antennas is $N_r > 1$, the received signal of users:

$$\mathbf{g}(t, \omega) = \sqrt{P_d} \mathbf{H}(t, \omega) \mathbf{W}(t, \omega) \mathbf{f}(t, \omega) + \mathbf{n}(t, \omega), \quad (6)$$

where $\mathbf{g}(t, \omega)$ is the received signal of OFDM symbol t and subcarriers ω , the dimension of \mathbf{g} is $N_r \times 1 \times N_c \times N_t$. P_d is the transmission power of the base station, $\mathbf{W}(t, \omega)$ is the $N_t \times N_{sts} \times N_c \times N_t$ downlink precoding matrix., N_{sts} is the number of spatial data streams of transmitted signals. $\mathbf{f}(t, \omega)$ is the $N_{sts} \times 1 \times N_c \times N_t$ transmitted signals. $\mathbf{n}(t, \omega)$ is the additive white Gaussian noise (AWGN) with zero mean and variance σ^2 . \mathbf{H} is $N_r \times N_t \times N_c \times N_t$ dimension channel matrix, and $\mathbf{H}(t, \omega)$ is the $N_r \times N_t$ spatial channel matrix from the transmitter to the receiver of OFDM symbol t and subcarriers ω .

4 Large-scale MIMO channel information feedback based on compressed sensing

We assume that large-scale antenna array of the base station is in a same platform and arranged closely. The correlation of antennas is expected, the channel information can has a sparse representation in both the spatial and frequency domain according to the signal processing theory. Based on this assumption, CS (compressed sensing) techniques can be applied to compress the CSI feedback information. We can carry out the two-dimension compression in space and frequency domain, then feedback the measurement vector to the transmitter, instead of \mathbf{H} .

The detailed process should be explained as follows: Firstly compress channel matrix $\mathbf{H}(t, \omega)$ in the spatial domain and get its sparse transformation; secondly make a secondary compression in frequency domain, then find a suitable measurement matrix to obtain the measured value (this paper employs the random matrix obeying Gaussian distribution); finally reconstruction algorithm

was introduced to reconstruct the original channel matrix from the measured values. So channel matrix $\mathbf{H}(t, \omega)$ should be vectorized into an $N_r N_t \times 1$ vector firstly:

$$\mathbf{h}(t, \omega) = \text{vec}(\mathbf{H}(t, \omega)). \quad (7)$$

After choosing the suitable sparsifying-basis in space domain and compress \mathbf{h} , we can get $\mathbf{S}_1 = \Psi_1 \mathbf{h}(t, \omega)$. The channel matrix of frequency domain is $N_r N_t \times N_c$, while the elements of each column are the elements of \mathbf{S}_1 obtained from spatial domain compression. Then the channel matrix in frequency domain should be vectorized into an $N_r N_t N_c \times 1$ vector, after the second compression, we can obtain sparse coefficients $\mathbf{S}_2 = \Psi_2 \mathbf{h}(t)$.

Then \mathbf{H} is encoded into measurement vector which is used as the content of compressing feedback:

$$\mathbf{y} = \Phi \mathbf{h} = \Phi \Psi_2 \mathbf{S}_2 = \Theta \mathbf{S}_2. \quad (8)$$

Thus, the channel vector $\mathbf{h}(t)$ with dimension $N_r N_t N_c \times 1$ is compressed into $M \times 1$ measurement vector \mathbf{y} through equation (8). M is much smaller than $N_r N_t N_c$ because the channel matrix is sparse in spatial-frequency domain, while CSI recover can be achieved accurately at the transmitter. Feedback load reduced by the compression $\eta = M / N_r N_t N_c$, transmitter and receiver can both get Φ by pre-configured. The transmitter can recover the channel information accurately through reconstruction algorithm after receiving the feedback of \mathbf{y} . The feedback flow chart is depicted in Figure 1.

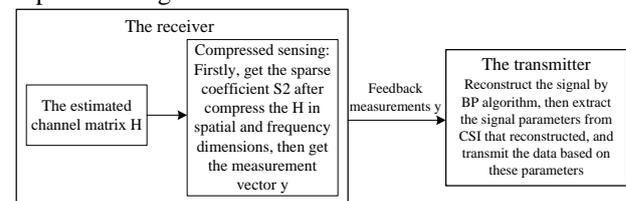


FIGURE1 The flow chart of the proposed CS-based channel feedback method

4.1 SPARSE REPRESENTATION OF THE CHANNEL MATRIX

The choice of sparsifying-basis Ψ plays a key role in sparsifying and reconstructing the information. Generally, an ideal sparsifying-basis would be associated with a more sparse representation or approximated sparse representation for \mathbf{h} . Due to the fact that Wavelet transform (DWT) has a strong ability to remove the

correlation, this paper uses wavelet transform to get the corresponding sparse representation of channel matrix \mathbf{H} in its spatial-frequency domain. Then construct the orthogonal wavelet transform matrix Ψ [11]:

$$\Psi = [P_n] \cdots [P_2][P_1], \tag{9}$$

$$[P_n] = \begin{bmatrix} \mathbf{L}_{(N/2^n) \times (N/2^{n-1})} & \mathbf{0} \\ \mathbf{G}_{(N/2^n) \times (N/2^{n-1})} & \mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{10}$$

where L , G are two matrices which are constructed by low-pass filter l and high-pass filter g . Each row of them is a vector of length $N/2^{n-1}$: $[l(0), l(1), \dots, l(2P-2), l(2P-1), 0, 0, \dots, 0]$ and $[g(0), g(1), \dots, g(2P-2), g(2P-1), 0, 0, \dots, 0]$ that can be obtained by circumference 2 shifts separately. According to the Orthogonality of the filter, it can be proved that $\Psi\Psi^T = \Psi^T\Psi = \mathbf{I}$.

As mentioned above, the sparse representation of \mathbf{H} could be generated after compression in both spatial domain and frequency domain:

$$\mathbf{S}_2 = \Psi_2 \mathbf{h}(t). \tag{11}$$

4.2 RECONSTRUCT THE CHANNEL MATRIX

This paper applies the Basis Pursuit (BP) [12] to reconstruct the channel matrix, which can be formulated as the following l_1 norm minimization problem:

$$\min_s \|\mathbf{S}_2\|_1 \quad s.t. \quad \Theta \mathbf{S}_2 = \mathbf{y}. \tag{12}$$

Solving equation (12) can be equivalent to optimization of the linear programming problem [13], the standard form of which is as follows:

$$\min \mathbf{C}^T \mathbf{S}_2 \quad s.t. \quad \mathbf{A} \mathbf{S}_2 = \mathbf{b} \quad \mathbf{S}_2 \geq 0, \tag{13}$$

where \mathbf{S}_2 is called decision vector that can be used to find the reconstructed channel, \mathbf{C} is called the coefficient vector of the objective function. In this paper, the coefficient vector is specified as a unit vector. In addition, \mathbf{S}_2 and \mathbf{C} are both column vector with dimension N ; \mathbf{b} is the M -dimension column vector called the constant vector of constraint equation; \mathbf{A} is an $M \times N (M \leq N)$ matrix called the coefficient matrix of constraint equations.

The value of vector \mathbf{S}_2 should be decomposed into two parts, positive and negative, in order to solve the above formula. Let \mathbf{u}_2 and \mathbf{v}_2 have the same dimension with \mathbf{S}_2 and $\mathbf{u}_{2,i} = (\mathbf{S}_{2,i})_+, \mathbf{v}_{2,i} = (-\mathbf{S}_{2,i})_+, i = 1, 2, \dots, N$. Then, the vector \mathbf{S}_2 can be rewritten as

$$\mathbf{S}_2 = \mathbf{u}_2 - \mathbf{v}_2, \mathbf{u}_2 \geq 0, \mathbf{v}_2 \geq 0. \tag{14}$$

And let $\mathbf{A} = (\Theta, -\Theta)$, $\mathbf{S}_2 = \begin{pmatrix} \mathbf{u}_2 \\ \mathbf{v}_2 \end{pmatrix}$, $\mathbf{b} = \mathbf{y}$, the value of \mathbf{S}_2 can be solved with the combination of formula (14), then the reconstructed channel matrix $\hat{\mathbf{h}}$ was got by $\hat{\mathbf{h}} = \Psi_2^T \mathbf{S}_2$.

4.3 ADAPTIVE FEEDBACK PROTOCOLS BASED ON SPARSITY OF THE CHANNEL MATRIX

According to the feedback mechanism in large-scale MIMO systems described above, this paper proposed an adaptive feedback protocols based on CS which configures the feedback dynamically according to the sparsity of the channel state matrix to improve system efficiency.

After compressed in both spatial and frequency domain, the channel matrix \mathbf{H} will be transformed into its sparse representation in wavelet domain after the joint compression in space and frequency domain, and wavelet sparsifying-basis satisfies Restricted Isometry Property (RIP), then the K coefficients can be accurately reconstructed from the M measured values ($K < M \ll N$). Since the compression ratio, $\eta = M / N_r N_t N_c$, is proportional to M , K can be used as the threshold of its sparsity (Because the channel matrix is impossible sparsed absolutely, so we need to set smaller sparse coefficients to 0 to get approximated sparse). Only when the channel matrix is $\geq K$ -sparse, the feedback scheme with a larger compression ratio should be employed, if the channel matrix is $< K$ -sparse, a smaller compression ratio should be employed. We adjust the value of M at the receiver, in accordance with the sparsity of instantaneous channel. In this case the feedback can be reduced by using a lower default value of M . In the simplest form, the compression ratio is switched between two possible levels. Note that the extension to adaptation among more than two levels is also possible, where the sparsity of channel range is partitioned into several regions, and each region corresponds to a specific compression ratio.

Consider that M changes between M_1 and

M_2 ($M_1 < M_2$), and Φ is an $M_2 \times N_r N_r N_c$ random measurement matrix stored at both transmitter and receiver. When the lower compression ratio is used ($M = M_1$), only the first M_1 rows of Φ are useful for compression at the receiver and for reconstruction at the transmitter. If $M = M_2$, then the full matrix of Φ is applied for computation. Thus, besides the CS measurements, the feedback should also include an indication of M so that the transmitter is able to determine an appropriate portion of Φ for CSI recovery.

5 Performance analysis and simulation results

In this section, we analysed and evaluated the performance of channel information feedback scheme based on compressed sensing by simulation. Simulation parameters are shown in Table 1. To more clearly evaluate the performance, we compared a fixed compression ratio feedback, an adaptive compression ratio feedback according to the sparsity and a perfect channel information feedback. Simulation results in Figure 2 show the bit error rate (BER) performance of SVD precoding.

TABLE 1 simulation parameters of channel feedback algorithm based on CS

| Simulation parameters | Values |
|------------------------------|---|
| Wavelength | 0.375m |
| Antenna spacing | 0.01m |
| Carrier frequency | 800MHz |
| Subcarrier number | 512 |
| Antenna Configuration | BS:16 antennas UE:16 antennas |
| Channel Model | |
| Simulation Data | 20000bit |
| Modulation and coding scheme | QPSK, 1/2 Convolutional coding |
| Channel estimation | Ideal Channel Estimation |
| Precoding | SVD |
| Channel feedback | Feedback cycle in time domain:5ms;CS-based feedback in Spatial frequency domain, compared with perfect channel information feedback |

As the simulation results shown, the BER performance is better when the compression rate is

References

[1] Rusek F, Persson D, Lau B K, Larsson E G, Marzetta T L, Edfors O 2012 Scaling up MIMO: Opportunities and challenges with very large arrays *IEEE Signal Process. Mag.* **30**(1) 40-60

[2] Jose J, Ashikhmin A, Marzetta T L 2009 Pilot contamination problem in multi-cell TDD systems *IEEE International Symposium on Information Theory (ISIT'09)* July 2009 2184-88

[3] Love D J, Choi J, Bidigare P 2013 Downlink training techniques for FDD massive MIMO systems: open-loop and closed-loop training with memory *IEEE Journal of Selected Topics in Signal Processing (J-STSP)* 2013 7712 1-12

adaptive to the sparsity channel compared with the compression ratio fixed at 40%, and it is very close to the BER of the perfect channel information feedback. The proposed algorithm not only can significantly reduce the feedback overhead, but obtain highly-accurate channel information recovery and ensure large-scale MIMO performance gain. The simulation results are shown in Figure (2):

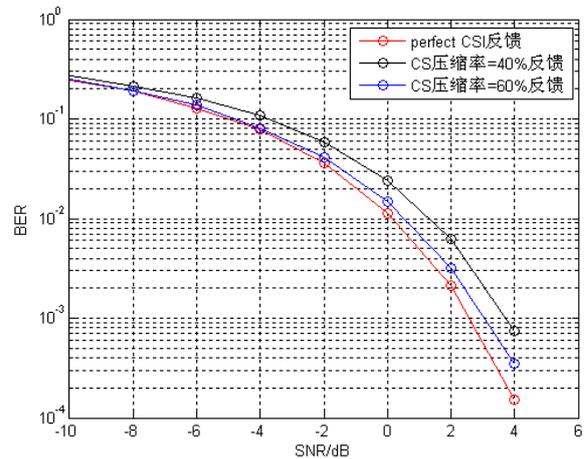


FIGURE 2 Comparison of BER performance

6 Conclusions

To reduce the CSI feedback overhead, we proposed a CS feedback algorithm for Large-scale MIMO systems. CSI Feedback can be compressed by sparsifying projections, feedback with an adaptive compression ratio and reconstructed to highly-accurate channel information. The compression ratio is dynamically adjusted based on the sparsity of instantaneous channel, so the proposed feedback algorithm can acquire a perfect balance between performance and feedback overhead. But in the real implementation process, many practical problems e.g. channel estimation errors, quantization noise, need more investigations in the future.

Acknowledgements

This paper is sponsored by National 863 Project (2014AA01A705), National Natural Science Foundation of China (61102047) and the Science Research Project of Shaanxi Provincial Department of Education (11JK1016).

[4] Choi J, Chance Z, Love D J, Madhow U 2013 Noncoherent trellis coded quantization: apractical limited feedback technique for massive MIMO systems *Communications IEEE Transactions* Dec.2013 **61**(12) 5016-29

[5] Edwards J 2011 Focus on compressive sensing *IEEE signal processing magazine* **28**(2) 11-3

[6] Engelbeg S 2012 Compressive sensing *IEEE instrumentation & measurement magazine* **15**(1) 42-6

[7] Candes E, Romberg J, Tao T 2006 Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information *IEEE Trans. Inform. Theory* **52** 489–509

[8] Davis L M, Hanly S V, Tune P, Bhaskaran S R 2010 Multi-antenna downlink broadcast using compressed-sensed medium access *Proc. IEEE Int. Conf. Commun. (ICC)*, Cape Town, South Africa, May: 1-5

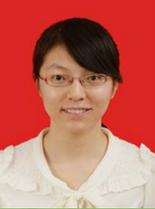
[9] Qaseem S T 2010 Al-Naffouri T Y. Compressive Sensing for Reducing Feedback in MIMO Broadcast Channels. *Communications (ICC), 2010 IEEE International Conference* May 2010 1-5

[10] Kuo P H, Kung H T, Ting P A 2012 Compressive Sensing Based Channel Feedback Protocols for Spatially-Correlated Massive Antenna Arrays *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, April 2012, 492-7

[11] Ge Zhexue, Wei Sha 2007 *Wavelet analysis theory and MATLAB 2007 application* Beijing

[12] Bloomfield P, Steiger W 1983 *Least Absolute Deviations: Theory, Applications, and Algorithms* Birkhauser, Boston

[13] Chen S S, Donoho D L, Saunders M A 2001 Atomic decomposition by basis pursuit *SIAM Review* **43**(1) 129-59

| Authors | |
|---|--|
|  | <p>Jing Jiang, born in October, 1974, Ankang, Shanxi Province</p> <p>Current position, grades: An associate professor of Xi'an University of Posts & Telecommunications</p> <p>University studies: She received M. Sc. from the XiDian University in 2005 and a Ph.D. in Information and Communication Engineering from North Western Polytechnic University, China in 2009.</p> <p>Scientific interest: wireless communication theory and MIMO system design</p> <p>Publications: more than 20 articles</p> <p>Experience: She had been a researcher and a project manager from 2006 to 2012 at ZTE Corporation in China and has been a Member of the IEEE, 3GPP.</p> |
|  | <p>Xu Shuang, born in March, 1989, XinZhou, Shanxi Province</p> <p>Current position, grades: A graduate of Xi'an University of Posts & Telecommunications</p> <p>University studies: She is currently studying as a M.E. in School of Communication and Information Engineering in Xi'an University of Posts & Telecommunications, China</p> <p>Scientific interest: Her research interests include Massive MIMO</p> <p>Publications: two articles</p> <p>Experience: She received the B.E. degree from Tibet University for Nationalities in 2012.</p> |
|  | <p>Lu Guangyue, born in September 1971, NanYang, Henan Province</p> <p>Current position, grades: A professor of Xi'an University of Posts & Telecommunications</p> <p>University studies: He received Ph.D from the key Lab of Radar Signal Processing in XiDian University in 1999, and was a post-doctor from 2004 to 2006 in the University of Uppsala in Sweden.</p> <p>Scientific interest: Radar signal processing and communications signal processing</p> <p>Publications: more than 30 articles</p> <p>Experience: He has worked as a teacher in Xi'an University of Posts and Telecommunications from 2001. Now he is the honorary President of School of Communication and Information Engineering, XUPT.</p> |
|  | <p>Xie Yongbin, born in April 1965, Baotou, Inner Mongolia</p> <p>Current position, grades: A professor of Xi'an University of Posts & Telecommunications</p> <p>University studies: He received M. Sc. and Ph.D from the XI'AN JiaoTong University.</p> <p>Scientific interest: Research and Development of the fifth generation mobile communication technology</p> <p>Publications: more than 10 articles</p> <p>Experience: He had employed Xi'an Datang Group company in 1997, and then served as CEO of Datang Mobile company in 2003. Now he is the honorary President of School of Communication and Information Engineering, XUPT.</p> |

Heterogeneous networks model for lower error using concatenated encoding

Yong Li¹, Jiang Yu^{1*}, Rong Zong¹, Yan Zhang¹, Jihong Shi¹, Jinsong Hu²

¹*School of Information and Engineering, Yunnan University, No.2 of Cuihu north road, Kunming, 650091, China*

²*Communication Branch of Yunnan Power Grid Corporation, Kunming, 650217, China*

Received 1 March 2014, www.tsi.lv

Abstract

Power line communication of the power distribution network requires higher reliability and better standards. In this paper, the ideas of cooperative communication motivate that we have presented a dual heterogeneous networks model with power line communication network and wireless sensor networks. Concatenated channel encoding with cyclic redundancy check code, convolution code, Reed Solomon code and the interleaver are respectively researched, which are used to analyse the performance of dual networks' communication scheme. The simulated results reveal that the dual networks have better communication quality than the one of single network. Compared to dual-network with RS code, the probability of frame error transmission and bit error rate for dual-network with convolution code are lower. On the basis of the characteristics of two different kinds of concatenated codes, we put forward an improved model, which is more close to reality. The results verify the feasibility of the design.

Keywords: power distribution network, power line communication, wireless sensor network, dual heterogeneous networks, concatenated channel encoding

Introduction

In 2009, National Institute of Standards and Technology (NIST) promulgated the smart grid interoperability framework, which proposed the application of wireless communication in the intelligent power grid. In January 2010, Pacific Northwest National Laboratory released a report about wireless communications for the electric power system, which focus on how, where and what type of wireless transmission satisfy requirements of power communication system. A new framework of the next generation power communication transmission network which based on packet transport network (PTN) and optical transport network (OTN) is suggested [1]. Much of studies have been carried on technology of Zigbee, WSN [2]. However, the wireless technologies existed and its standard protocol are difficult to meet the application requirements of network layer, since it has some disadvantages with real time requirement. The performance of bit error rate (BER) in wireless networks could hardly meet to the requirement of 10^{-9} , which is a requirement of industrial control. So the research of reliability and real-time characteristics during the power distribution network communication system is concerned urgently.

Today, electric power networks were built with an integrated and vertical structure. The power energy is mostly generated in centralized power plants and transported over a long distance transmission network to a distribution network before reaching the end users [3].

In this context, the distribution network is regarded as a passive system used to deliver reliable unidirectional power to end users.

With the development of power distribution network, a lot of wired and wireless technologies are relevant to these networks, such as power line communication (PLC) [4-8], wireless sensor networks (WSN) [9], WiMAX, ZigBee and so on, which can potentially be applied to and integrated into power distribution networks. The PLC systems generally operate by transmitting a modulated carrier signal on the wiring system [10]. It benefits from the ubiquity of already existing electrical power delivery networks and promises access to telecommunication service in every corner of a house without requiring installation of new infrastructure. The PLC in Smart Grid has been studied [10]. The PLC itself has low-speed shortcomings for data communication. In order to make the power distribution network communication reliable and robust, we need to improve the transmission media or use certain technologies. Bisceglie et al. argued that [10], as a result of the growing issues related to the quality of power line communication, the system of wide scope voltage checking is more demanding. In sensor networks (WSNs), every node could calculate local, global performances using local information and information exchanged with neighboring nodes.

Multiple Relay Selection Scheme [11] and Reed Solomon (RS) code [12] is concatenated as an inner code to improve the communication performance. Motivated by these ideas, combined with a variety of effective

* *Corresponding author* e-mail: yujiang@ynu.edu.cn

network routing algorithm [13], the communication system designs are analyzed through two representative medium voltage networks (33 and 11 kV) based on Figure 1 modeling approach and assessment framework.

In this paper, a scheme of dual heterogeneous networks of WSN and PLCN based on cyclic redundancy check (CRC) code, Reed-Solomon (RS) code, convolution code and the matrix interleaver respectively is proposed, inspired by the idea of cooperative communication [12]. The dual networks have higher reliability and better reliability than single network, which are suitable for the requirement of power distribution network environment. The operation model of dual networks communication with RS-CRC coding and the matrix interleaver is presented in section 3.1. Section 3.2 describes the simulations and testing results of applying CRC, convolutional code and the matrix interleaver. The performance of error frame transmission with CRC-RS-Inter and CRC-Con-Inter is presented in section 3.3. Finally, a few useful concluding remarks are given in section 4.

2 Communication system model

On the basis of discussion and analysis of the power line communication and wireless sensor networks (WSN), this paper design of dual networks communication module adopting WSN and power line communication network (PLCN). Both networks are introduced with noises of different distribution, so that one of the networks is independent fading network. We consider two binary symmetric channels as WSN and PLC network, the initial seed of which are different. We assume that two networks transmit same information $m(t)$ at the same time. CRC, RS, and interleaving are cascaded, which compose CRC-RS-Inter link. CRC-Inter-Con link is get with CRC, interleaving and convolution cascading. By combining CRC-RS-Inter and CRC-Inter-Con coding mechanisms, we expect that our proposed model can serve power communication with better reliability requirements. In PLC system, the electric facilities are generally stationary installed, they are rarely be moved, so the performance demand of mobility is low. Good mobility of WSN system made up for shortcomings of PLC system very well. Once a sensor node can't detect an event, PLC system can ensure the normal communication. The logic structure of dual-network module is shown in Figure 1.

Random binary numbers $m(t)$ are transmitted in WSN and PLCN network, which are generated by Bernoulli Binary Generator. Firstly, information data are checked with CRC Generator (CRC-32). Then each data frame (or packet) is appended with 32 check bits is encoded into one RS codeword. Cyclic redundancy check (CRC) code is concatenated as an inner code. $RS(n, k)$ codeword used Equation (1) to calculate the length code n , m is m hexadecimal number, $q(q \geq 2)$ is integer number.

$$n = m - 1 = 2^q - 1. \tag{1}$$

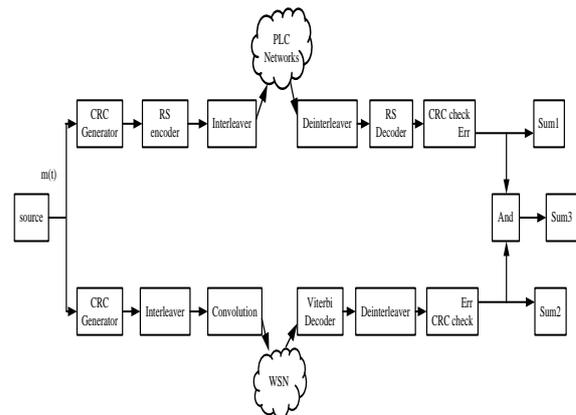


FIGURE 1 Logic structure of dual-network module

RS(31,15) is selected, which has been used in military communication. Since Binary Symmetric Channel (BSC) can reflect the characteristics of two different networks when set up different error probability, so BSC is selected approximately equal to the characteristics of PLCN and WSN. By using the method of corresponding decoding arithmetic, CRC-32 Syndrome Detectors check each error frame and output check result in receiving terminal. If the error of frame happens in data frame, the port of terminal outputs mark information 1. When date frame has no error in course of transmission, it will output mark information 0. Finally, we get transmission results of dual networks by doing logic operation between output values of WSN and PLCN. *Sum1* and *Sum2* terminals applied to calculate the error frames transmitted by WSN and PLCN, respectively. Error frames transmitted in the dual networks' result are got by *Sum3*. The bit error rate (BER) of dual-network is calculated with logic terminal, Error Rate Calculation and Cumulative Sum logic module on the matlab software platform of computer.

3 Communication system design

System study is divided into the following three parts: part 1, both networks use CRC, RS and interleaving code (CRC-RS-Inter). In order to get two single-networks about PLCN and WSN, two BSC are set with different parameters. Part 2, in this section, both networks transmit message with CRC, convolution and interleaving code (CRC-Con-Inter). Part 3, we use CRC-Con-Inter to approximate PLC networks, and CRC-RS-Inter to model WSN. Here we focus on the major parameters of probability of frame error transmission (PET) and bit error rate (BER) for the dual networks and two single-networks in power communication.

PET used Equation (2) to calculate, e is error frame transmission number, t is total frame transmission number. Both e and t are integer number. BER has a similar definition with Equation (2).

$$PET = \frac{e}{t}. \tag{2}$$

3.1 PERFORMANCE OF DUAL NETWORKS WITH RS AND INTERLEAVING

The relationship between probability of frame error transmission and the network error probability in PLC network, wireless sensor networks and dual networks with convolution codes are shown in Figure 2. Here, the frame length is 1468 bits, and 2090 data frames are transmitted continuously. When the network error probability is less than 10⁻², all the frames can be received successfully. As the network error probability increasing, the probability of frame error transmission becomes appear. When the network error probability of WSN and PLCN is 0.03, the error probability of single network system is about 0.3828, 0.3962 respectively, and the dual-network system is 0.1459. When *NEP*>0.03, the curves of error transmission rates are gradually increasing rapidly, but frame error rate of dual networks (i.e. Dual networks-RS in Figure 1) is less than that in single networks (i.e. PLCN-RS, WSN-RS in Figure 1). When *NEP*=0.04, frame error rate of single network is bigger than 0.5, data packets can't be effectively transmitted. When the network error rate increased to 0.06, the frame error rate of single network and dual networks are all almost to 1, at this time, system will not be able to transmit data frames rightly.

The figure below shows the bit error rate under different network error rate condition. The bit error rates of two single-networks are almost consistent (i.e. PLCN-RS, WSN-RS in Figure 2), and the curve of dual networks mechanism (i.e. Dual networks-RS in Figure 2) is lower obviously. For example, when network error probability is 0.01, the bit error rate of PLC network and WSN are 5.53×10⁻⁶, 9.98×10⁻⁶, while dual network is less than 10⁻⁹; when network error probability is 0.05, the bit error rate of PLC network, WSN and dual network are 0.0181, 0.0189 and 3.4288×10⁻⁴. When bit error rate is 0.1, they become 0.0964, 0.0971 and 0.0093 respectively.

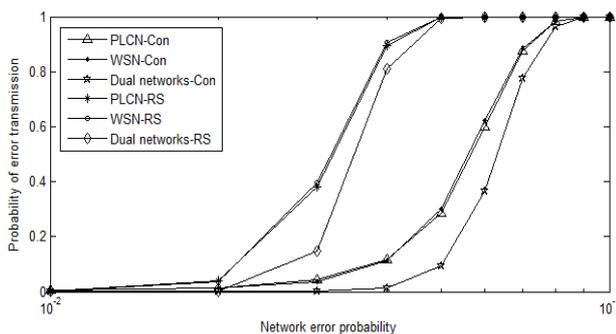


FIGURE 2 The relationship between probability of error frame transmission and the network error probability about WSN, PLCN and dual networks

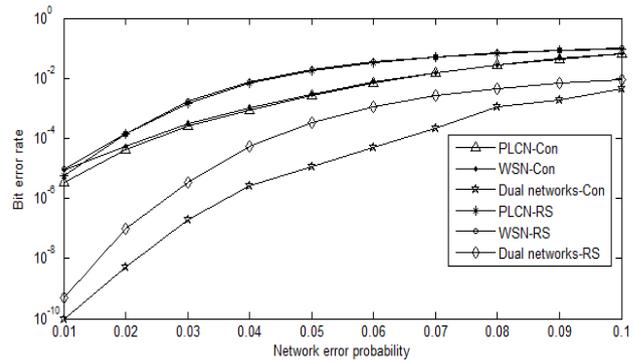


FIGURE 3 The relationship between bit error rate and network error probability about WSN, PLCN and dual networks

Figure 4 shows the relationship curve of probability of error frame transmission and the frame length about WSN, PLCN and dual networks, and the network error probability of PLCN and WSN are set to be 3×10⁻² in the simulation. It can be seen that, with the frame size increasing, the frame error rate has a tendency to increase. What's more, the probability of error packet transmission of dual networks (i.e. Dual networks-RS in Figure 4) is less than the probability of error packet transmission of single WSN (i.e. WSN-RS in Figure 4), PLCN (i.e. PLCN-RS in Figure 4) network. For instance, *PET* for PLCN and WSN scheme are 0.2913 and 0.3288, while *PET* for dual networks system is 0.0928, when the frame length is 1093bits. When frame length is 2893bits, the *PET* for two single network become to be 0.6249, 0.6239, where the reliability of the single network significantly get worse, but The *PET* for dual network is only 0.3893, which still has high reliability. The *PET* of dual network is 0.6086, when frame length increased to 4693. At this packet size, almost no data is transmitted successfully to the destination in single network.

3.2 PERFORMANCE OF DUAL NETWORKS WITH CONVOLUTION CODE AND INTERLEAVING

We can see in Figure 2, probability of frame error transmission increases with network error probability increasing. Compared to single communication links of PLCN and WSN, the dual-network has higher successfully transmission rate. The error probability of single network system is about 0.114 and 0.117, when the *NEP* of WSN and PLCN is 0.04, while the dual-network system is 0.013. When *NEP*>0.04, the communication quality gets bad rapidly. When *NEP*=0.06, the error rate of dual-network is about 0.37, where can communicate with high quality. Under the same condition, dual networks with CRC-RS-Inter coding has no packets receive rightly. The *NEP* of dual networks with CRC-RS-Inter (i.e. Dual networks-RS in Figure 2) is higher than the *NEP* of dual networks with CRC-Con-Inter (i.e. Dual networks-Con in Figure 2), what indicates the CRC-RS-Inter system has better probability of successful transmission.

Figure 3 also describes the PLCN, WSN and dual-network with CRC-Con-Inter coding. Compared with dual CRC-RS-Inter communication (i.e. Dual networks-RS in Figure 3), dual network with convolution code as inner code (i.e. Dual networks-Con in Figure 3) gets smaller bit error rate, which ensure good reliability and robustness. For example, when network error probability is 0.04, the bit error rate of two single networks in CRC-RS-Inter are 8.9×10^{-4} and 9.9×10^{-4} , which in CRC-Con-Inter are 8.9×10^{-4} and 9.9×10^{-4} , and the bit error rate of dual networks in CRC-RS-Inter is 5×10^{-5} , which in CRC-Con-Inter is 2.6×10^{-6} . The bit error rates in CRC-RS-Inter are 0.0681, 0.0688 and 0.0046, and what in CRC-Con-Inter are 0.0269, 0.0277 and 0.0011, when network error probability is set to be 0.08.

The relationship about the probability of frame error transmission and the frame size about WSN, PLCN and dual networks in CRC-Con-Inter are shown in Figure 4. Along with the frame size increasing, the frame error rate has a same tendency of increase. Dual networks (i.e. Dual networks-Con in Figure 4) show better performance than PLCN (i.e. PLCN-Con in Figure 4) and WSN (i.e. WSN-Con in Figure 4) single network. For instance, *PET* for PLCN and WSN scheme are 0.0916 and 0.0969, while *PET* for dual networks system is 0.0096, when the frame length is 1093bits. When frame length is 1693bits, the *PET* for two single network and dual-works become 0.1214, 0.1258 and 0.016. When frame size grows to 4993bits, they become to be 0.2325, 0.2195 and 0.0553. Especially the *PET* of dual networks in CRC-Con-Inter is always smaller than the *PET* for dual networks in CRC-RS-Inter structure under the same frame size. When frame lengths are changed, curve of dual networks-RS change obviously, however, the variety is little.

3.3 PERFORMANCE OF ERROR FRAME TRANSMISSION WITH CRC-CON-INTER AND CRC-RS-INTER

In this section, a dual networks system has been developed for the distribution network. As we known, RS code has a strong ability of correcting burst error, so RS code is commonly used in wireless communication systems. We consider CRC-RS-Inter coding link as WSN, and CRC-Con-Inter link to be PLC network. The simulation analysis is done as following.

The frame size is fitted as 1468 bits. Figure 5 represents the relationship between bit error rate and network error probability about WSN, PLCN and dual networks. The bit error rate of dual networks is less than the bit error rate of single WSN, PLCN network, when the network error probability belongs to [0.01, 0.1], which denotes that dual-network enable high performance. Such as when *NEP*=0.03, bit error rate in WSN is 0.0016, and it in PLCN is 2.2×10^{-5} , but bit error rate for dual networks is only 3.3×10^{-7} . Then we study the mixing dual networks (i.e. Dual networks-Mixing), dual network which both use RS code as inner code (i.e. Dual

networks-RS) and dual network which both consider involution code as inner code (i.e. Dual networks-Con). The curve of dual networks with mixing code method is higher than curve of dual networks-Con, however lower than dual networks-RS. Although the reliability of dual networks with mixing links isn't the best, but this model is closer to reality. The simulation results show that this system has well quality of communication. For example, when *NEP* is 0.01, the bit error rate of dual networks with mixing links is less than 10^{-9} ; when *NEP*=0.05, the bit error rate is about 5.1×10^{-5} . When *NEP*=0.07, the bit error rate become to be 7.5×10^{-4} .

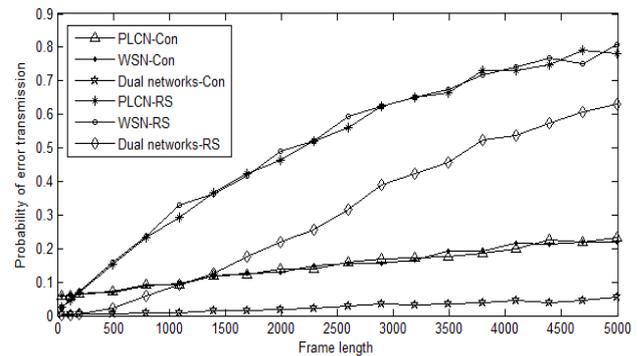


FIGURE 4 The relationship curve of probability of error packet transmission and the frame length about WSN, PLCN network and dual networks

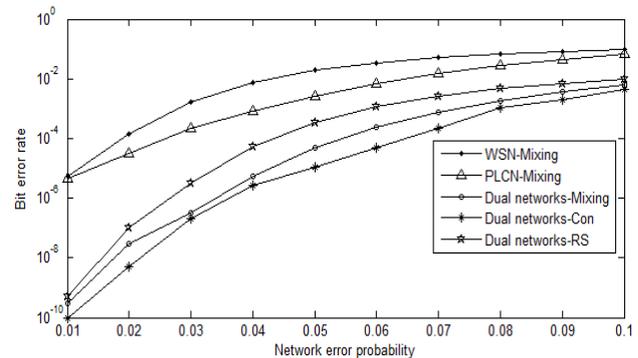


FIGURE 5 The relationship curve between bit error rate and network error probability about WSN, PLCN and dual networks

4 Conclusions

In industrial area, communicate often accompanied by various kinds of interferences. So it usually requires communication system has higher reliability (*BER* less than 10^{-9}). Many communication schemes cannot meet the strict requirements in power distribution network control environment. In this paper, we proposed novel dual heterogeneous networks communication model with WSN and PLCN, which transmit original information at the same time, on which is an improved method based physical layer. The simulation results indicate that the reliability of dual-network is better than that in single-network. Along with network error probability increasing sharply, the probability of frame error transmission of dual-network increases slowly. What's more, the

reliability and robustness of dual-network with the interleaving, CRC code and convolution code concatenated are higher than that in dual-network with CRC code as inner code of RS code. In the end, in order to close to real communication environment, we put forward mixing the interleaving, CRC code and convolution code concatenated respectively to simulate the communication quality of WSN and PLCN. The bit error rate is analysed, and results show that the dual-network obviously improves the service quality of the

Li Yong, Yu Jiang, Zong Rong, Zhang Yan, Shi Jihong, Hu Jinsong
 power distribution network communication. Dual networks with PLCN and WSN is an ideal solution when realized good reliability and seamless coverage.

Acknowledgment

This work is supported by the national natural science foundation (NSF) of China with the grant numbers 61162004.

References

[1] Feng B, Fan Q, Li Y 2012 Research on Framework of the Next Generation Power Communication Transmission Network 2012 International Conference on Computer Science and Electronics Engineering 23-25 March 2012 Hangzhou IEEE 2012 414-7

[2] Li J, Liu S, Wu S 2012 A Design of Remote Computer House Monitoring and Control System Based on ZigBee WSN IJACT: International Journal of Advancements in Computing Technology 4(12) 233-40

[3] Yang Q, Barria J A, Green T C 2011 Industrial Informatics IEEE Transactions on 7(2) 316-27

[4] Oksman V, Zhang J, 2011 Communications Magazine IEEE 49(12) 36-44

[5] Hashmat R, Pagani P, Chonavel T, Zeddami A 2012 Power Delivery IEEE Transactions on 27(4) 2082-9

[6] Schwartz M. 2009 IEEE Communications Magazine 47(1) 14-8

[7] Lai S, Messier G 2012 IEEE Transactions on Communications 60(12) 3865-75

[8] Wang P, Marshall A, Noordn K A, Huo X, Markarian G 2010 Hybrid network combining PLC and IEEE802.16 for hospital environment, Power Line Communications and Its Applications (ISPLC), 2010 IEEE International Symposium March 2010 Rio de Janeiro 267-72

[9] Rout R R, Ghosh S K 2013 IEEE Transactions on Wireless Communications 12(2) 656-67

[10] Gao J, Xiao Y, Liu J, Liang W, Philip Chen C L 2012 A survey of communication/networking in Smart Grids Future Generation Computer Systems 28(2) 391-404

[11] Xu J, Zhang H, Yuan D, 2012 E-Global Optimal Multiple Relay Selection Scheme in Cognitive Relay Networks AISS: Advances in Information Sciences and Service Sciences 4(4) 218-29

[12] Wang S, Reza Soleymani M 2012 Cooperative Communication System with Systematic Raptor Codes CCECE: 25th IEEE Canadian Conference on Electrical and Computer Engineering 2012 Montreal QC 2012 1-6

[13] Saputro N, Akkaya K, Uludag S 2012 A survey of routing protocols for smart grid communications Computer Networks 56(11) 2742-71

| Authors | |
|---|---|
|  | <p>Yong Li, born on November 3, 1984, Henan Province, China</p> <p>Current position, grades: Postgraduate student in the school of Information and Engineering at The Yunnan University of China, doctoral candidate of Northwestern Polytechnical University of China</p> <p>University studies: Information and Communication Engineering in Yunnan University of China</p> <p>Scientific interest: Network communication, network protocol, channel encoding, communication of the power distribution network.</p> <p>Publications: 4 papers.</p> <p>Experience: the National Natural Science Foundation project in 2012.</p> |
|  | <p>Jiang Yu, born on March 14, 1961, Yunnan Province, China</p> <p>Current position, grades: Professor at School of Information Science and Engineering in Yunnan University, China.</p> <p>University studies: University of Electronic Science and Technology of China.</p> <p>Scientific interest: Communication theory, network communication, wireless communication, telecommunication system analysis and design, image processing.</p> <p>Publications: 80 articles, 4 textbooks.</p> <p>Experience: the National Natural Science Foundation project, Yunnan Power Grid Corporation city disaster recovery centre study project, grid communications technology and research methods of emergency Yunnan project.</p> |
|  | <p>Rong Zong, born on December 17, 1962, Yunnan Province, China</p> <p>Current position, grades: Professor at School of Information Science and Engineering in Yunnan University, China.</p> <p>University studies: Yunnan University of China</p> <p>Scientific interest: Network communication and communication system technology, network protocol, channel encoding, communication of the Power Distribution Network</p> <p>Publications: 80 articles, 4 textbooks</p> <p>Experience: the National Natural Science Foundation project, Yunnan Power Grid Corporation city disaster recovery centre study project, grid communications technology and research methods of emergency Yunnan project.</p> |
|  | <p>Yan Zhang, born on December 27, 1991, Hebei Province, China</p> <p>Current position, grades: Undergraduate student in the school of Information and Engineering at The Yunnan University of China.</p> <p>University studies: Communication Engineering in Yunnan University of China</p> <p>Scientific interest: Network communication, network protocol, channel encoding, communication of the power distribution network.</p> <p>Publications: 1 papers</p> <p>Experience: the National Natural Science Foundation project in 2012</p> |

| | |
|---|--|
|  | <p>Jihong Shi, born on December 7, 1963, Yunnan Province, China</p> <p>Current position, grades: Professor at School of Information Science and Engineering in Yunnan University, China. University studies: Huazhong of science and technology of China. Scientific interest: Network communication and communication system technology, including network protocol, channel encoding, communication of the Power Distribution Network. Publications: 40 articles, 6 textbooks. Experience: the National Natural Science Foundation project, Yunnan Power Grid Corporation city disaster recovery centre study project, grid communications technology and research methods of emergency Yunnan project, research with PTN network communication system architecture and technical solutions project.</p> |
|  | <p>Jinsong Hu, born on December 24, 1966, Kunming, China</p> <p>Current position, grades: senior engineer, deputy director of Yunnan Power Grid Corporation dispatch control centre, China. University studies: electronic engineering in Shanghai Jiaotong University. Scientific interest: communication network in power grid, the power of information technology, smart distribution grid communications systems, intelligent control. Publications: 30 papers. Experience: the National Natural Science Foundation project, Yunnan Power Grid Corporation city disaster recovery centre study project, grid communications technology and research methods of emergency Yunnan project, research with PTN network communication system architecture and technical solutions project.</p> |

Improvements and implementation of the permission system based on RBAC model

Zhenrong Deng*, Xingxing Tang, Chuan Zhang, Xi Zhang, Wenming Huang

Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China

Received 14 May 2014, www.tsi.lv

Abstract

Role-based access control as a traditional access control (discretionary access, mandatory access) is a promising place to receive widespread attention. Systematic researches on RBAC models, on one hand, this paper combined with the characteristics of Electronic government affair information management system and added regional filter function to the core RBAC model, besides, the research developed by J2EE framework and this paper presents a high availability and extensibility of RL-RBAC competence management system.

Key words: RBAC, J2EE, permission system

1 Introduction

Along with the rapid development of modern computer and the Internet in recent years, information technology to infiltrate all walks of life, the electronic government affairs information management system obtained high speed development. Although the information management system in e-government brought convenience to people's life and work, the information security problems should be considered. And it not only endanger an individual interests, but also, at the higher level, involves the government and the national security. Access control as an important function of information system security component, its main purpose is to combat the threat of unauthorized operations involving computer or communication system, these threats can be subdivided into the unauthorized use, disclosure, modification, destruction and denial of service, etc. [1,2]. Role-based Access Control (Role -based Access Control, RBAC) model can effectively implement organization security policy, RBAC model greatly alleviate the resource of permissions management problems [3], with the Role of interventional make authority management is more flexible and convenient.

In recent years, people put forward a lot of to improve the model of RBAC model [4-8], but the electronic government affairs information management system is usually diverse from territory. For example, we regard municipal level staff as a user of this system, another county level staff is also a role of this system, they should have the same operation permissions based on the electronic government affairs system module, meanwhile, under the jurisdiction of the municipal agency for is for the whole city area personnel management, and only at the county level staff for his county personnel management, which requires our rights management

system can carry on the limits to the regional data access. However, these models can't solve the problem, this paper focuses on the characteristics of the electronic government affairs information management system, in this paper, on the basis of the classical RBAC model, join the regional limit, put forward a general permission system adapted to the electronic government affairs system model (RL - RBAC).

Current RBAC research mainly focuses on the theoretical research, but lack of the concrete implementation. It restricts the use of research results in engineering practice. In view of the above problems, this paper put forward theoretical model at the same time, the model is given in the current popular open source framework of J2EE implementation method, and implements an easy expansion, easy to use and versatility of adapting to permissions in J2EE system.

2 RBAC model and improved

2.1 THE BASIC MODEL OF RBAC

Role based access control model is put forward by Ferraiolo et al. [9], through continued efforts, RBAC community members in RBAC in February 2004 by the United States standard committee (ANSI, the American National Standards) and IT (INCITS) international Standards committee accepted as ANSI INCITS359-2204 [10]. Basic RBAC model as shown in Figure 1:

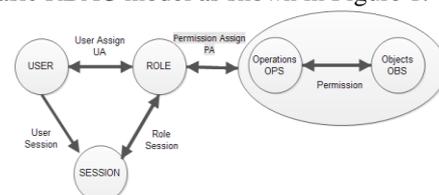


FIGURE 1 Core RBAC figure

* Corresponding author e-mail: zhrdeng@guet.edu.cn

RBAC's basic theory is: gives the user role, and its function will be authorized permissions to roles rather than users. By the user role authorization, a user can be given multiple roles, a role can have multiple permissions, with permissions can be given multiple roles, is also a many-to-many relationship between roles and permissions, access permissions and roles are linked together, roles associated with the user again, achieve the logical separation of user and the access, great convenient for rights management.

Basic RBAC model includes the following parts:

- 1) The basic objects include *USER*, *ROLE*, *OPS*, *OBS*, *SESSION*, (users, roles, action, object, session)
- 2) $UA \subseteq U \times R$, *UA* is the user role assignment, is to be assigned to users to the roles of the many-to-many relationship, a role can be assigned to multiple users, a user can have multiple roles.
- 3) $PA \subseteq PRMS \times R$, shows that a many-to-many relationship between the authority and role, a role can have multiple permissions, a permission can be assigned to multiple roles.
- 4) $USER \rightarrow 2^{SESSION}$, *USER* contacts user session and the session a one-to-many relationship, a user can have multiple sessions, and a session only allow one user participation.
- 5) $SESSION \rightarrow 2^{ROLE}$, shows that the session and the character of a one-to-many relationship, a session can have multiple roles, the same user of a character can only corresponding to a session.
- 6) $2^{OPS \times OBS}$, shows the set of permissions *PRMS*, also can represent the operation and the object of a many-to-many relationship.

2.2 RL-RBAC MODEL

Although RBAC model reduced the workload and the complexity of authorization management, due to the classic RBAC model based on role as the medium of access control, this way in the electronic government affairs information management system has significant limitations, if classic RBAC model is used to implement regional limit is difficult. This paper aimed at the characteristics of the electronic government affairs information management system is put forward based on the regional limit role access control (Region Limited RBAC, RL - RBAC) rights management system, the structure of the model as shown in Figure 2:

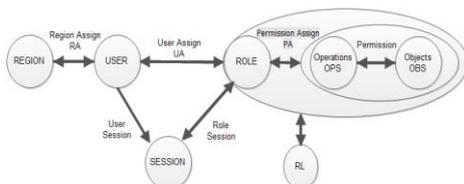


FIGURE 2 RL - RBAC model

Increases the area on the RL - RBAC in the original model defined concept, formal definition is as follows:

REGION on behalf of the REGION, said users and regional many-to-many relationship, a user can be assigned to multiple areas, an area can be assigned to multiple users. After the user login system, according to the users' area, users can only to perform operations on data in the area, a user can belong to multiple regions at the same time, the user is multiple regional data and operating range, RL is the regional limit parameters.

3 System analysis and implementation

3.1 THE IMPLEMENTATION ENVIRONMENT OF RL-RBAC MODEL

During the development process of information management system, J2EE becomes most developers' first choice as its characteristics such as independence, portability, security, and more users. The paper implements a RL - RBAC model permissions system based on the J2EE architecture as the development framework as it has been greatly used. If other framework developers want to apply it in their own system, they just need to modify the permissions system slightly.

3.2 THE DATABASE DESIGN OF RL - RBAC MODEL

The entity relationship model of RL-RBAC is shown in Figure 3:

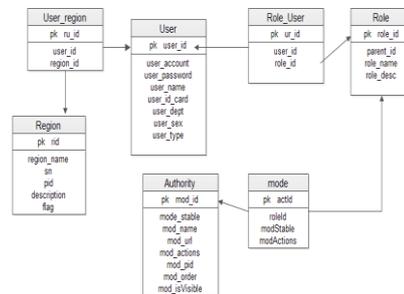


FIGURE 3 The entity relationship model of RL-RBAC

- (1) The user table, mainly to describe the users' information. It is the carrier of access control, corresponding to the users of the model.
- (2) Area table, used to define the limited area, corresponding to the area in the system.
- (3) Area user table, mainly to define the relationship between users and area. It used to limit area for users. The field userid is used to specify the user while regionid used to specify the user area, corresponding to the area assigned in the model.
- (4) Permissions module table, mainly used to define the system function entity and its permissions. It corresponding to the operation and the object of the model as it were set by the system's administrator.
- (5) Role table, it mainly used to define roles, including its name, description and the superior role, corresponding to the model role.
- (6) Role module table, mainly to define the relationship

between roles and system modules, as well as access to the module operation. It used to authorization for roles, corresponding to the permissions assigned in the model.

(7) Role user table, used to define the relationship between roles and users in the system. The relationship is stored in the table, corresponding to the user assignment in the model.

3.3 THE AUTHORIZATION PROCESS

According to the requirements of the electronic government information management system, the authorization process mainly includes three steps as follows.

1) Limit the user area.

Limit user area means distribution area to users: choose user area, then write the incidence relationship between users and area into the area user table.

2) Role's authorization.

Role's authorization is one of the core module of rights management system. Login the system as an administrator, choose the role which need authorize, choose the authorize menu in the system menu tree, granted to the role permissions such as select, delete, modify and so on according to requirement. Then click save menu to write the corresponding information of role and menu into the role menu table. The operation interface are shown in Figure 4, the code is as follows:

```
long roleId=Long.parseLong(req.getParameter("roleId"));
String rmActions = req.getParameter("rmActions");
boolean flag = this.roleActionsService.setting(roleId,
rmActions);
```

3) Enter into the associated modules between roles and users, choose roles assigned to users. This process is used to write the associated relationship between roles and users into the role user table.

3.4 ACCESS FILTER IMPLEMENTATION

After the user login to the system, the system first verify the legitimacy of the user by user name, password, and other ways. For legitimate users, system according to the RL - RBAC model to obtain all of the user's permissions, RL - permissions in RBAC model filter is divided into three parts to processing, to query the data of regional limit, menu display and operating limits.

When user login system, first reads the user role, read all the permissions of the user and writing session, based on the user has permissions, when users use a module during operation, the area of limited information in the form of parameters to the corresponding query statements, realize area is limited. Permissions filtration processes are shown in Figure 5:

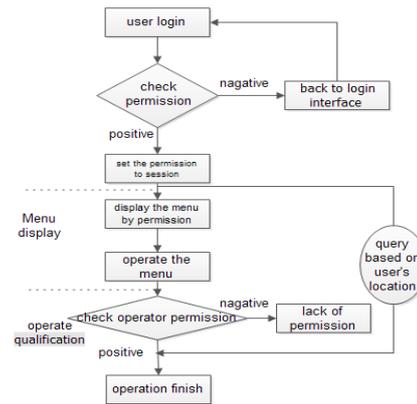


FIGURE 5 permissions filtering process

The following described the concrete implementation of access filter:

3.4.1 User authentication

This function is to validate the user's effective user authentication, is the basis of the authority system, for legitimate users, first by user ID to get all the roles assigned to the user, and then through the character ID table lookup roles module, get all the permissions of the role, the specific code implementation is as follows:

```
List<PermissionRole> rolelist =
this.userLoginService.getRolesByLoginUser(u.getUserId());
/* Get to the user by user ID assigned roles */
List<PermissionRoleActions> perRoleAlist = new
ArrayList();
/* Store user permissions */
For (int i=0; i< rolelist.size;i++)
{
/* get the role */
PermissionRole role = rolelist.get(i);
long roleId = role.getRoleId();
/* get the user's permissions through the role ID */
List<PermissionRoleActions> alist =
this.userLoginService.getActionsByLoginUser(roleId);
/* added to the access list */
perRoleAlist.addall(alist);
}
/* Permissions write to the user session */
request.getSession().setAttribute(Constants.USER_INFO_A
CTION , perRoleAlist);
```

PermissionRoleActions is role of Hibernate persistence of module table, table a PermissionRole is part of the model, rolelist used to store PermissionRole types of objects, are all the logged in user's role, traverse rolelist, role module via the character ID table, insert all permissions perRoleAlist, all permissions through perRoleAlist into the session.

3.4.2 The menu display

Through access control function menu shows, is one of the main work of the authority system, function module in the information system are the form of a tree list, the

main idea of this article is, first of all find out all function modules and deposited in the function module in the system list, and then traverse function module in the list of each object, compared with all of the user's permission to list, if the object in the permissions list, there are marked is proved that the user has permissions to the function module, the function returns the front desk page, according to the specific implementation code is as follows:

```
List<PermissionRoleActions> perRoleAlist =
BaseUtil.getUserAction(req); /* get permissions list from
the session user */
for(Iterator<PermissionModule> it = mList.iterator());
it.hasNext();){
    PermissionModule m = it.next();
    boolean open = false;
    for(PermissionRoleActions ra : perRoleAlist){
        if(m.getModStable().equals(
ra.getModStable())){
            open = true;
            break; }}
        if( !open ){
            it.remove();
        }
    }
return this.buildTreeMenu(mList);
}
```

List mList deposit all the function modules of the system, first of all, get all the permissions through user session, all of the user's permissions are deposited in the permissions list module named perRoleAlist, check the permission of a function module PermissionModule by traverse the perRoleAlist, getModStable () function gets the function module a unique identifier, traverse perRoleAlist is looking for and need to check whether an object function module identifier function module identifier equal to, and it proved that the user has permissions to the function modules.

3.4.3 Operating limit

In this paper, the system is used for the user's basic operation mainly includes input, delete, modify, check, check, import, export, etc., different role requires different permissions, operating limit belongs to the fine-grained access control, the user first needs to be a qualified operation code to display the code, when the user click on the module, through the added to the front desk in advance display code, call the corresponding access control module of backend server, won the users to have all the operation of the module, if the user has the permission operation, will be the user, the operating limit access control on the server side code is given below:

```
/* Returns the menu of the logged in user's authority
information */
public static String
getUserActionByStable(HttpServletRequest req, String
stable){
List<PermissionRoleActions> permissionRoleAlist = null;
```

```
permissionRoleAlist =
(List<PermissionRoleActions>)req.getSession().getAttribut
e(Constants.USER_INFO_ACTION)
for(PermissionRoleActions ra : permissionRoleAlist){
    if(ra.getModStable().equals(stable)){
        return ra.getModActions();
    }
}
return null;
}
```

getUserActionByStable (HttpServletRequest req, String stable) function of the parameter stable represent user access module, current from the session permissionRoleAlist stored all of the user's permissions, by iterating through permissionRoleAlist find stable corresponding module, and then returned to the front desk page processing.

3.4.4 Area limit

Area limited is a very important part of the authority system, the user can operate the data is in the user area within the scope of qualified, first through the user ID assigned to the user of all regions. Data query area will be added to the statement, and have more than one area of the users search multiple regions and sets. The specific code implementation is as follows:

```
List<PermissionRegion> perRegionlist =
this.userLoginService.getRegionByLoginUser(
u.getUserId());
List<Region> regionlist =new ArrayList();
Map<String, Object> searchmap=new HashMap();
for (int i=0; i< perRegionlist.size;i++) {
    Region region = perRegionlist.get(i);
    long regionId= region.getId();
    region = this.userLoginService.getRegionById(regionId);
    regionlist.add (region) ;}
search.put("regionList", regionlist);
IGridModel gmodel =
houseFamilyService.pageQuery(gModel,map);
```

Query users by user ID set into the area perRegionlist, by iterating through perRegionlist query user area set, to which the user belong the regionlist all areas is the user's collection, the last area collection list regionlist deposited in the storage container map query conditions, in the query is executed, the area is limited to join query.

4 Conclusion

Area limit is very important part of the authority system, this paper combined with the characteristics of the electronic government affairs information management system, to improve the classic RBAC model, is proposed based on region (RL – RBAC) limit of RBAC model, classic RBAC model by solving problems on the regional limit, and improve the model at the same time, this article also discusses in detail the implementation steps of the model based on J2EE architecture, the model has good generality and now this model has been applied in civil affairs medical rescue system real-time [11], the urban

planning information system and urban and rural low-income family economic conditions of water use and water verification system, the practice shows that this model can solve the problems of the e-government system of access control well.

References

- [1] Smeureanu I, Diosteanu A 2010 Knowledge Dynamics in Semantic Web Service Composition for Supply Chain Management Applications *Journal of Applied Quantitative Methods* 5 (1) 1-13
- [2] Shrivani D, Suresh P V, Padmaja B R 2010 The Web Services Security Architectures Composition and Contract Design Using RBAC *International Journal on Computer Science and Engineering* 8(2) 2609-15
- [3] Sandhu R S 1996 Role-Based Access Control Models *IEEE Computer* 29(2) 38-47
- [4] Zhang L H, Ahn G-J, Chu B-T 2001 A rule-based framework for role-based delegation *Proceedings of the 6th ACM Symposium on Access Control Models and Technologies* 2001 153-62
- [5] Wainer J, Kumar A 2005 A fine-grained, controllable, user-to-user delegation method in RBAC *Proceedings of the 10th ACM Symposium on Access Control Models and Technologies* 59-66
- [6] Si W, Zeng G, Cheng Q 2006 The fine-grained expansion and application of RBAC model *Computer science* 33(4) 227-80
- [7] Zhai Z D 2006 Quantified-role based controllable delegation model *Journal of Computers* 2006 29(8) 1401-7
- [8] Li H, Guan K 2011 The information terminal kernel model based on RBAC *Computer science* 2011 38(11) 100-3
- [9] Ferraiolo D F, Sandhu R S, Gavrila S 2001 Proposed NIST standard for role-based access control *ACM Transactions on Information and System Security* 4(3) 224-74
- [10] ANSI INCITS 359-2004: American National Standard for Information Technology—Rolebased Access Control 2004
- [11] Deng Z 2010 A real-time medical assistance billing system *International Conference on Intelligent Computing and Integrated System (ICISS 2010)* 757-60 October 2010 Guilin China

Acknowledgment

This work is supported by Guangxi key Laboratory of Trusted Software (No: kx201317), by the Postgraduate's Innovation Project of Guilin University of Electronic Technology under (No: XJY2012017), and also supported by the Research Projects of Education Department of Guangxi Province (No: 2009MS1195).

Authors



Zhenrong Deng, born on July 2, 1977, Guilin, China

Current position: Associate professor of computer science and engineering.
University studies: M.S. degree in computer science and technology in Guangxi University, China, in 2005.
Scientific interest: Grid computing, data mining, trustworthy software.
Publications: 28.



Xingxing Tang, born on February 26, 1988, Beijing, China

Current position: machine learning researcher in qunar.com.
University studies: computer science master in Guiling university of electronic technology.
Scientific interest: machine learning, reinforcement learning and computer vision.
Publications: 2 papers.
Experience: machine learning in qunar.com.



Chuan Zhang, born on July 25, 1988, Guilin, China

Current position: M.S candidate in Guilin University of electronic technology.
University studies: the technology of computer in Guilin University of electronic technology.
Scientific interest: cloud computing, data mining, software engineering.
Publications: 2 papers.
Experience: worked at the largest Chinese search engine company Baidu Inc.



Xi Zhang, born on August 4, 1989, Guilin, China

Current position: M.S candidate in Guilin University of electronic technology.
University studies: the technology of computer in Guilin University of electronic technology.
Scientific interest: recommender systems and data mining.



Wenming Huang, born on July 8, 1963, Suzhou, China

Current position: professor of computer science and engineering, director of department of software engineering.
Scientific interest: grid computing, image processing, software engineering.
Publications: 50.

Anti-spam model based on AIS in cloud computing environments

Jin Yang^{1, 2*}, Lingxi Peng³, Tang Liu⁴

¹*School of Information Science & Technology, Southwest Jiaotong University, Chengdu 610031, China*

²*Department of Computer Science, LeShan Normal University, LeShan 614000, China*

³*Department of Computer and Education software/Guangzhou University, Guangzhou, China*

⁴*College of Fundamental Education/Sichuan Normal University, Chengdu, China*

Received 1 March 2014, www.tsi.lv

Abstract

Cloud computing is becoming a hot research topic. However, there is little attention to cloud computing environment work for anti-spam issues. Spam has become a thorny issue facing with many countries. The overflow of spam not only great wastes the network resources, taking up the user's e-mail resources, reducing the network efficiency, affecting the normal use of the Internet, but also violates the user's individual rights. But the traditional spam solutions for anti-spam are mostly static methods, and the means of adaptive and real time analyses the mail are seldom considered. Inspired by the theory of artificial immune systems (AIS), this paper presents an anti-spam system in cloud computing environment. The concepts and formal definitions of immune cells are given, and the hierarchical and distributed management frameworks of the proposed model are built. The results of evaluation indicate that the proposed model has the features of real-time processing and is more efficient than client-server-based solutions, thus providing a promising solution for anti-spam system for heterogeneous cloud environments.

Keywords: cloud computing, artificial immune systems, anti-spam system

1 Introduction

What is cloud computing? Cloud computing is the delivery of computing as a service rather than a product, whereby shared resources, software, and information are provided to computers and other devices as a metered service over a network (typically the Internet) [1]. It has many advantages such as economy, complex calculations, agility, high scalability, high reliability, and easy maintenance. The concept of cloud computing was born in the 1960s from the ideas of American computer scientist J.C.R. Licklider and John McCarthy stated that computing will become a publicly available service in the future. In 1983, Sun Microsystems brought forward a singular vision that "the network is the computer" [2]. On August 9, 2006, the CEO Google, Eric Schmidt, firstly mentioned the concept of Cloud computing on SES San Jose 2006. On January 30, 2008, Google declared "Cloud Computing Research Plan" in Taiwan and will promote the advanced technology in Taiwan's colleges. February 1, 2008, IBM (NYSE: IBM) announced it will establish the first Cloud Computing Centre for software companies in China. On March 5, 2010, Novell and CSA released a supplier neutral plane, named as Trusted Cloud Initiative. May 22, 2009, China's first Cloud Computing Conference held in Beijing China

World Hotel. January 22, 2010, China cloud computing technology and industry alliance (CCCTIA) announced in Beijing. In the cloud computing trend, including computers, communications, Internet, the entire information technology industry is undergoing a comprehensive updating. Now software industry is facing momentous changing, which the software production organization evolves towards the service-oriented, agile, customized direction and the network terminal equipment begins to show the diversified and personalized features [3, 5].

Email spam, known as unsolicited bulk email, junk mail, or unsolicited commercial email, is the practice of sending unwanted email messages, frequently with commercial content, in large quantities to an indiscriminate set of recipients. Spam is becoming a serious problem since it causes huge losses to the organization, such as wasting the bandwidth, adding the user's time to deal with the insignificance mail, enhancing the mail server processing and causing the mail server to crush [6]. Cloud computing also faces the security issues such as the using of virtualization technology to hide viruses, Trojan horses, especially the spam and other malicious software problems. Anti-spam is the application of data investigation and analysis techniques currently mainly by means of blocking and filtering procedures [7]. However, the current techniques

* *Corresponding author* e-mail: jinyang@163.com

For the convenience using the fields of an antigen x , a subscript operator "." is used to extract a specified field of x , where $x.fieldname$ = the value of filed fieldname x . In the model, all the detectors form a Set Detector called SD .

$$SD = \{ \langle d, age, count \rangle \mid d \in D, age \in N, count \in N \}, \quad (2)$$

where d is the antibody gene that is used to match an antigen, age is the age of detector d , count (*affinity*) is the number of detector matched by antibody d , and N is the set of nature numbers. SD contains two subsets: mature and memory, respectively, the set M and set T . A mature SD is a SD that is tolerant to self but is not activated by antigens. A memory SD evolves from a mature one that matches enough antigens in its lifecycle. Therefore, $SD = M \cup T, M \cap T = \phi$.

$$M = \{ x \mid x \in SD, \forall y \in Self, \langle x.d, y \rangle \notin Match \wedge x.count < \beta \}, \quad (3)$$

$$T = \{ x \mid x \in SD, \forall y \in Self, \langle x.d, y \rangle \notin Match \wedge x.count \geq \beta \}, \quad (4)$$

where $\beta(>0)$ represents the activation threshold. Match is a match relation defined by:

$$Match = \{ \langle x, y \rangle \mid x, y \in D, f_{match}(x, y) = 1 \}. \quad (5)$$

The affinity function $f_{match}(x, y)$ may be any kind of Hamming, Manhattan, Euclidean, and r -continuous matching, etc. In this model, we take r -continuous matching algorithm to compute the affinity of mature Detectors.

2.2 THE DYNAMIC MATURE DETECTOR MODEL

$$M(t) = M(0) = 0, t = 0, \quad (6)$$

$$M(t + \Delta t) = M(t) + M_{new}(\Delta t) + M_{from_other}(\Delta t) - M_{dead}(\Delta t), \quad \text{if } f_{match}(M(t), Ag(t)) \neq 1, \quad (7)$$

$$M_{clone}(t) = \frac{\partial M_{clone}}{\partial x_{clone}} \cdot \frac{\partial M_{active}}{\partial x_{active}} \cdot \Delta(t-1), \quad (8)$$

if $f_{match}(M(t), Ag(t)) = 1$.

$$M.\rho(t + \Delta t) = M.\rho(t) + V_p \cdot \Delta t, \quad (9)$$

$$M.count(t + \Delta t) = M.count(t) + 1,$$

$$M_{new}(\Delta t) = \frac{\partial M_{new}}{\partial x_{new}} \cdot \Delta t = \frac{\partial T_{active}}{\partial x_{active}} \cdot \Delta(t-1), \quad (10)$$

$$M_{dead}(\Delta t) = \frac{\partial M_{death}}{\partial x_{death}} \cdot \Delta t, \quad (11)$$

if $f_{match}(M(t-1), Self(t-1)) = 1$,

$$M_{from_other}(\Delta t) = \sum_{i=1}^k \left(\frac{\partial M_{from_other}^i}{\partial x_{from_other}} \cdot \Delta t \right). \quad (12)$$

Equation (6) depicts the lifecycle of the mature detector, simulating the Agent that the mature detectors evolve into the next generation. All mature detectors have a fixed lifecycle (λ). If a mature detector matches enough antigens ($\geq \beta$) in its lifecycle, it will evolve to a memory detector. However, the detector will be eliminated and replaced by new generated mature detector if they do not match enough antigens in their lifecycle. $M_{new}(t)$ is the generation of new mature SD . $M_{dead}(t)$ is the set of SD that haven't match enough antigens ($\leq \beta$) in lifecycle or classified self antigens as *nonself* at time t . $M_{active}(t)$ is the set of the least recently used mature SD which degrade into memory SD and be given a new age $T > 0$ and count $\beta > 1$. When the same antigens arrive again, they will be detected immediately by the memory SD . In the mature detector lifecycle, the inefficient detectors on classifying antigens are killed through the process of clone selection. Therefore, the method can enhance detection efficiency when the abnormal behaviours intrude the email system again.

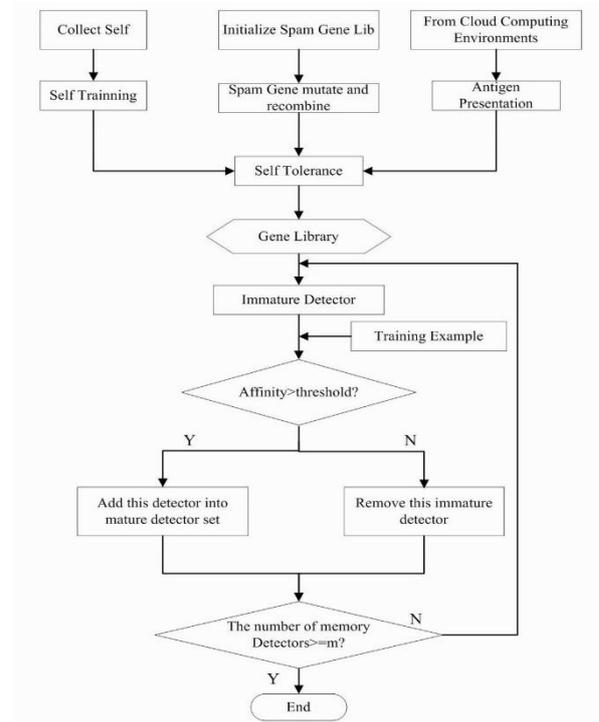


FIGURE 2 The Dynamic Mature Detector Model

As Figure 2 shows, system randomly creates the immature detectors firstly, and then it computes the affinity between the immature detectors and every element of training example. If the affinity of one immature detector is over threshold, it will become a mature detector and will be add into mature detector set.

System repeats this procedure until mature detectors are created.

2.3 THE PROCESS OF EMAIL SURVEILLANCE

Our model uses detector state conversion in the dynamic evolution of mature detector and memory detector, erasing and self matching detector. As the Figure 3 shows, the undetected Emails are compared with memory detectors firstly.

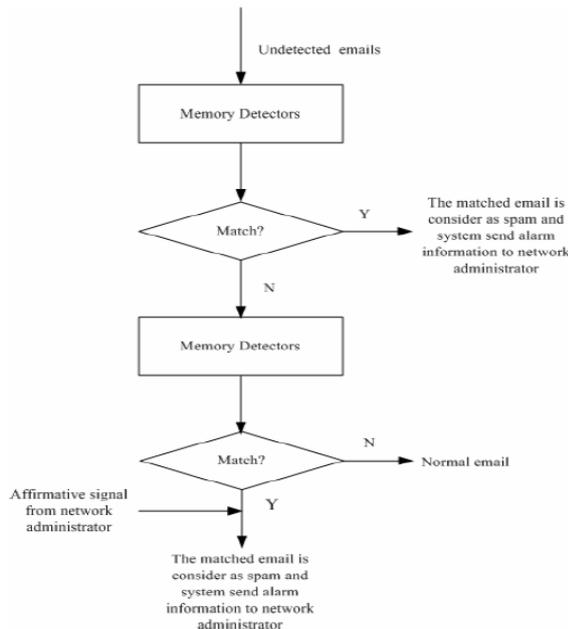


FIGURE 3 The Process of Email Surveillance

If one e-mail match any elements of memory detector set, this Email is classified as spam and send alarming information to user. Then, the remaining Emails which are filtered by memory detectors are compared to mature detectors. Mature detectors must have become stimulated to classify an as junk, and therefore it is assumed the first stimulatory signal has already occurred. Feedback from administrator is then interpreted to provide a co-stimulation signal. If system receives affirmative co-stimulation in fixed period, the matched Email is classified as spam. Or else it is considered as normal Email and delivered to user client in the normal way. During the filtering phase, when a mature detector matches one e-mail, the count field of mature detector will be added. If the value of filed count is over threshold, it will be activated and become a memory detector. Meanwhile, if a memory detector cannot match with any e-mails in fixed period, it will degenerate into a mature detector. When the unsolicited emails and malice intrusions increase, we simulate immune system functions to increase the density of antibody; when they decrease, we simulate immune feedback functions and reduce the density of corresponding antibody, restoring it to normal level.

2.4 THE EVALUATION OF THE EMAIL RISK

Owing to the fact that our model relates to enormous factors for evaluation, on purpose of reasonably and entirely measuring the spam email dangerous status, we classify the involved factors as host dangers, area dangers, cells dangers, and special dangers. Afterwards, we subdivide and arrange all the factors which influence the network dangers, in order to let them locate on different layers, forming a structure model with identify matrix.

1) *Construct Identify Matrix*: First of all, we must construct identify matrix which is result that we compared the relative importance of one group of elements on next layer with some past layer element constraint. That is, it shows the relative importance of any pair of factors. In detail, denote b_{ij} the compared result of the i^{th} factor and j^{th} one, b_{ij} all together form the identify matrix B :

$$B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix},$$

where: $b_{ii} = 1$ if $i = j$ and $b_{ij} = 1/b_{ji}$ if $i \neq j$.

2) *Computing Weights*: Next we obtain the weight of each factor. According to the identify matrix B , we can get the maximum eigenvalue of the matrix λ_{max} . Here, we can get the maximum λ_{max} according with the following condition:

$$\begin{vmatrix} b_{11} - \lambda & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} - \lambda & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} - \lambda \end{vmatrix} = 0.$$

Work out the corresponding eigenvector of maximum eigenvalue of B , $X = (x_1, x_2, \dots, x_n)$, let x_i to be the weight of factor u_i , then we can get unitary weights denote W_i .

$$A = (W_1, W_2, \dots, W_n) = (x_1 / \sum_{i=1}^n x_i, x_2 / \sum_{i=1}^n x_i, \dots, x_n / \sum_{i=1}^n x_i).$$

3) *Test of Consistency*: Because of complexity of evaluation and limit of individual knowledge, the individual identify matrix may not be consistent with the actual one, or the disagreement of any two identify matrixes may result in error of subjective judgment. However, we must test the consistency of the matrix B as follows:

a) Computing consistency value $C \cdot I$:

$$C \cdot I = \frac{\lambda_{\max} - n}{n - 1} \tag{13}$$

b) Computing consistency ratio $C \cdot R$

$$C \cdot R = \frac{C \cdot I}{R \cdot I} \tag{14}$$

where $R \cdot I$ is mean consistency value that can be found in the reference and forms, we often consider that if $C \cdot R$ is smaller than 0.1, the consistency of matrix is acceptable, otherwise we must modify the identify matrix B.

4) *Computing the General Weight Order*: The general weight order means that the weight order comparing the elements in the present layer and the highest layer. We have got each order of element in rule layer to the object layer and the values are W_1, W_2, \dots, W_n , respectively, we also know that order that design layer to the rule layer and the values are $W_1^j, W_2^j, \dots, W_n^j$, then the general order is

$$V = W^j W = \begin{pmatrix} W_1^j & W_1^j & \dots & W_1^j \\ W_2^j & W_2^j & \dots & W_2^j \\ \vdots & \vdots & \vdots & \vdots \\ W_n^j & W_n^j & \dots & W_n^j \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{pmatrix} = \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{pmatrix} \tag{15}$$

2.5 EVALUATING THE DANGER LEVEL

The entire network of spam danger level should fully reflect the value of each of the host facing attacks. Let $n_{ij}(t)$ be the numbers of i^{th} spam detect attacking at time t. Let $\beta_i (0 \leq \beta_i \leq 1)$ be the importance coefficient of i^{th} computer in the network and $\alpha_j (0 \leq \alpha_j \leq 1)$ be the danger coefficient of the j^{th} kind of attack in the network. Therefore, we can get the spam danger $R(t)$ situation and evaluate security at real time.

$$R_j(t) = \frac{2}{1 + e^{-\alpha_j \sum_i \beta_i n_{ij}(t)}} - 1 \tag{16}$$

The conclusion can be shown that the higher value $R(t)$ reaches the more dangerous the network is.

3 Experimental results and analysis

Experiments of simulation were carried out in our Laboratory. The main aim of the experiment was to test the feasibility of the application for anti-spam based on AIS to implement spam detecting. And we developed some series experiments. Here are the coefficients for the model as the Table 1 showing.

TABLE 1 Coefficients for the model

| Parameter | Value |
|---------------------------------|--------|
| r-contiguous bits matching rule | 8 |
| The size of initial self set n | 40 |
| The Initial Scale of Detectors | 100 |
| Match Threshold β | 40~60 |
| Activable Threshold λ | 50~150 |
| Clone Scale | 20 |
| Mutation Scale | 19 |
| Life Cycle of Mature Detectors | 120s |

We prepared the Ling-Spam datasets for analysis and experiments. A mixture of 481 spam messages and 2412 messages sent via the Linguist list, a moderated list about the profession and science of linguistics. Attachments, HTML tags, and duplicate spam messages received on the same day are not included. The whole experiment is divided into two phase: training phase and application phase. The main different between the two phases is that the former does not use filtering module and just generates detectors for system. We partitioned the emails randomly into ten parts and choose one part randomly as a training example, then remaining nine parts are used for test and we can get 9 group recall and precision ratios. The average value of these 9 group values is considered as the model's recall and precision ratio.

Traditional spam filters system and technology almost adopted static measure, however, lack self-adaptation and the ability of parallel distributed processing. In this paper, we have presented a model of spam detection based on the theory of artificial immune system, and we have also illustrated the advantages of this model than traditional models. The concepts and formal definitions of immune cells are given. And we have quantitatively depicted the dynamic evolutions of self, antigens, immune-tolerance, and the immune memory. Additionally, the model utilized a distributed and multi-hierarchy framework to provide an effective solution for the spam. Finally, the experimental results show that the proposed model is a good solution for anti-spam system.

Acknowledgments

This work is supported by the China Postdoctoral Science Foundation (No.2012T50783, No.2011M501419), and Sichuan Provincial Department of Science and Technology Project (No.2014JY0036), and Scientific Research Fund of Sichuan Provincial Education Department (No.13TD0014), and Leshan Normal University of Achievements Transformation Project (No.Z1322), and Science and Technology Key Research Project of Leshan (No.12GZD014).

References

- [1] http://en.wikipedia.org/wiki/Cloud_Computing
- [2] <http://www.mysql.com/news-and-events/sun-to-acquire-mysql.html>
- [3] Garg S K, Versteeg S, Buyya R 2013 A framework for ranking of cloud computing services *Future Generation Computer Systems* **29**(4) 1012-23
- [4] Patel A, Taghavi M, Bakhtiyari K 2013 An intrusion detection and prevention system in cloud computing: A systematic review *Journal of Network and Computer Applications* **36**(1) 25-41
- [5] Luo J-Z, Wu W-J, Yang M 2011 Mobile Internet: Terminal Devices, Networks and Services *Chinese Journal of Computers* **34**(11) 2029-51
- [6] Mezmaz M, Melab N, Kessaci Y 2011 A parallel bi-objective hybrid metaheuristic for energy-aware scheduling for cloud computing systems *Journal of Parallel and Distributed Computing* **71**(11) 1497-508
- [7] Subashini S, Kavitha V 2011 A survey on security issues in service delivery models of cloud computing *Journal of Network and Computer Applications* **34**(1) 1-118
- [8] Kshetri N 2013 Privacy and security issues in cloud computing: The role of institutions and institutional evolution *Telecommunications Policy* **37**(4) 372-86
- [9] Rong C, Nguyen S T, Jaatun M G 2013 *Computers & Electrical Engineering* **39**(1) 47-54
- [10] Villa O, Petrini F 2008 Accelerating real-time string searching with multicore processors *Computer* **41**(4) 42-4
- [11] Burnet F M The 1959 Clonal Selection Theory of Acquired Immunity *Cambridge Cambridge University Press*
- [12] Kepler T B, Perelson A S 1993 Somatic hypermutation in B cells: An optimal control treatment *Journal of Theoretical Biology* 37-64
- [13] Forrest S, Perelson A S, Allen L, Cherukuri R 1994 Self-Nonself Discrimination in a Computer *Proceedings of IEEE Symposium on Re-search in Security and Privacy Oakland*
- [14] Kim J, Bentley P 1999 The Artificial Immune Model for Network Intrusion Detection *the 7th European Congress on Intelligent Techniques and Soft Computing*
- [15] Artin-Herran G, Rubel O 2008 Zaccour G Competing for consumer's attention *Automatica* **44** 361-70 (in Chinese)
- [16] Hanke M 2008 On the effects of stock spam e-mails *Journal of Financial Markets* **11** 57-83
- [17] Li T 2007 An Introduction to Computer Network Security. 1st edition *Publishing House of Electronics Industry Beijing*

| Authors | |
|---|---|
|  | <p>Jin Yang, born on June 9, 1980, Sichuan, China</p> <p>Current position, grades: associate professor at LeShan Normal University, PhD.</p> <p>University studies: Ph.D in computer science at Sichuan University, Sichuan in 2007.</p> <p>Scientific interest: network security, artificial immune, knowledge discovery, expert systems.</p> <p>Publications: 20.</p> |
|  | <p>Lingxi Peng, born on July 23, 1978, Guangzhou, China</p> <p>Current position, grades: professor at department of computer and education software, PhD.</p> <p>University studies: Ph.D in computer science from Sichuan University, Sichuan in 2008.</p> <p>Scientific interest: Network security, artificial immune, knowledge discovery, expert systems.</p> <p>Publications: 20.</p> |
|  | <p>Tang Liu, born on January 20, 1979, Chengdu Sichuan, China</p> <p>Current position, grades: associate professor in Sichuan Normal University, PhD student at the College of Computer Science, Sichuan University, Chengdu, China.</p> <p>University studies: M.S. degree at college of computer science, Sichuan University, China, in 2009.</p> <p>Scientific interest: wireless sensor networks.</p> <p>Publications: 15.</p> |

Research and application on set pair entity similarity model of social network

Yanjun Zhao*, Chunying Zhang

College of Science, Hebei Untied University, Tangshan, 063009, China

Received 1 March 2014, www.tsi.lv

Abstract

In allusion to the certain and uncertain value which exist in the node and relationship attributes of social network, From the attributes and relation angle of the entity to analyse the similarity degree of the affirmative, negative and uncertain between them, then build the set pair entity similarity model based on the set pair analytical method and apply it to the network association detection. First of all, applying the generalized pair close potential and the generalized set loose potential in social network based on set pair analysis method, and see it as the basis of association detection; secondly, giving the set similarity calculation method based on the entity attribute and relation, from the point of view of node attribute and relations attribute to calculate respectively, by setting the weight to consolidated calculate the set pair similarity of the entity; thirdly, utilizing entity set similarity to divide network association into clustering problem, then give the association partitioning algorithm; finally, integrating with the network instance to verify the effectiveness of the new network association partitioning algorithm.

Keywords: set pair, social network, entity similarity, attribute and relation, association partitioning

1 Introduction

The social network is one relation network [1], social network analysis is one analysis method for relation to study [1, 2]. The method has a very wide range application in the social network, owing to the entity is of feature attribute and relation attribute in social network. And the social network analysis aims at studying the relation attribute [3-6], the purpose of network association partitioning algorithm is to excavate the community structure in the network, It is a clustering problem in essence, by defining the similarity degree between the network node, to adopt the clustering algorithm for the node in network to conduct clustering, then to achieve the purpose of dividing network association.

In the similarity constructor method of network node, one constructor method based on the local node information [7], Considering neighbour information of the node, if the two nodes have the same or similar neighbour nodes then regard as them are similar. The Constructor method is based on the network topology information [8] regarding the node in network as the node of receiving and transmitting signals. Guo Jingfeng [9] has proposed one namesake entity partitioning algorithm based on attribute relation figure. By calculating the similarity degree of feature attribute of the entity and the similarity of the link and through the clustering to distinguish the individuals of the same name in the social network, making the accuracy of the analysis of social networks has improved. Zhou Li [10] has proposed one partitioning calculation method of the node similarity

degree in graph. By the node block structure feature in the social network to conduct the similarity degree calculation, and the experiments has proved the effectiveness of the algorithm. Xiang et al has proposed partitioning algorithm based on subgraph similarity degree [11]; this algorithm is superior to the algorithm of optimization module of Clause prominent [12]. Pan Ying has proposed association partitioning algorithm based on the node similarity degree in no weight graph [13]. But Leicht E A et al [14] made amendment for the no weight graph algorithm, and then proposed the structure partitioning algorithm of the weighted network community based on node similarity degree. But these proposed similarity calculation method and entity recognition algorithm can only be confined to one separate issue, the entity in the social network has various attributes, and the uncertain phenomenon exists in these property, How to resolve such problems, according to the authors has propose the entity in social network is constituted of the attribute [15]. Therefore, in this paper proposes the same entity recognition model based on set pair analysis theory, utilizing the attribute of entity set and relation degree of to conduct the similarity calculation. Finally, by the clustering method to set the threshold to identify the entity set, then utilize an example to demonstrate the effectiveness of the algorithm.

2 Set pair social network model

Assuming in the social relation network, any two objects v_k and v_s have the number of attribute is N (that is the

* *Corresponding author* e-mail: teacher_jsj@126.com

number of attribute set of two objects). Assuming the two research object have the number of the same attribute is S , the number of the different attribute is P , the number of the uncertain attribute is F , commanding:

$$\frac{S}{N} = a, \frac{F}{N} = b, \frac{P}{N} = c. \tag{1}$$

In allusion to the complexity of the nodes in social network and the uncertainty in the relation among the node, we give the following definition:

Definition 2.1 [10]: Assuming the domain set of research object $U = \{v_1, v_2, \dots, v_n\}$, thereinto the attribute set of the object v_k and v_s is $A(v_k) = \{x_{k1}, x_{k2}, \dots, x_{km}\}$ and $A(v_s) = \{x_{s1}, x_{s2}, \dots, x_{sm}\}$, so at one moment the contact degree of v_k and v_s is

$$\rho(v_k, v_s)(t) = a_{k,s}(t) + b_{k,s}(t)i + c_{k,s}(t)j \rho(e_k, e_s)(t) = a(t) + b(t)i + c(t)j \tag{2}$$

where $v_k \in U$ and $v_s \in U$, i is the differential label and takes the different in $[-1,1]$ depending on the circumstance. j only plays a marked role and takes the value is -1 .

If the weight of attribute is different, assuming:

$$\omega_k (k = 1, 2, \dots, N, \sum_{k=1}^N \omega_k = 1).$$

Assuming attributes according to the order of S, F, P to align and consecutive number. The contact degree is:

$$\rho(v_k, v_s)(t) = a_{k,s}(t) + b_{k,s}(t)i + c_{k,s}(t)j = \sum_{k=1}^S \omega_k(t) + \sum_{K=S+1}^{S+F} \omega_k(t)i + \sum_{k=S+F+1}^N \omega_k(t)j \tag{3}$$

where $i \in [-1,1]; j = -1; a_{k,s}(t) + b_{k,s}(t) + c_{k,s}(t) = 1$.

Definition 2.2: Assuming the matrix $R = (\rho(v_k, v_s))_{k \times s}$, R is called the set pair contact relation matrix, herein $\rho(v_k, v_s)$ is the element of the set pair contact relation matrix, and behalf of the set pair contact degree between v_k and v_s . $R(t) = (\rho(v_k, v_s)(t))_{n \times n}$ is behalf of the relation degree about every research object in set pair social network, the matrix can be expressed:

$$R(t) = \begin{bmatrix} \rho(v_1, v_1)(t) & \rho(v_1, v_2)(t) & \dots & \rho(v_1, v_n)(t) \\ \rho(v_2, v_1)(t) & \rho(v_2, v_2)(t) & \dots & \rho(v_2, v_n)(t) \\ \dots & \dots & \dots & \dots \\ \rho(v_n, v_1)(t) & \rho(v_n, v_2)(t) & \dots & \rho(v_n, v_n)(t) \end{bmatrix} \tag{4}$$

Do not take the time t into consideration, the R matrix is a static relation matrix, otherwise it is the dynamic relation matrix, with the time goes by, the nodes in social network constantly change, thereby to obtain a relation

matrix is constantly updated. By analyzing the relation matrix at the different moments, then to find the trend of the social network.

Definition 2.3: Assuming the vertex set of social network figure at the moment of t is $V(t) = \{v_1, v_2, \dots, v_n\}$, the attribute of vertex is $VX(t) = \{vx_1, vx_2, \dots, vx_m\}$, the set pair relation matrix at the moment of t is $R(t) = (\rho(v_k(t), v_s(t)))_{n \times n}$, so the dynamic social network analysis model is $GR(t) = (V(t)(VX), R(t))$. Do not take the time t into consideration, the model is the static social network analysis model.

Definition 2.4: The two individuals' relation contact degree in social network is $\rho(v_k, v_s) = a + bi + cj$, the ratio of relative same degree e^a and the relative opposed degree e^c is called the individual generalized set pair potential in the social network, that is:

$$Tread(v_k, v_s)_G = \frac{e^a}{e^c}. \tag{5}$$

Even if $c = 0$, we can still judge the relation trend of the two nodes in network.

Definition 2.5: The two individuals' relation contact degree in social network is $\rho(v_k, v_s) = a + bi + cj$, the ratio of relative same degree e^a and product of the relative opposed degree e^c the relative different degree e^b is called the individual generalized set pair loose potential in social network, that is:

$$Tread(v_k, v_s)_G = \frac{e^a}{e^{c+b}}. \tag{6}$$

The uncertainty term (different degree) is converted to opposition term (opposed degree) to study the trend that the node in social network may withdraw the network.

Definition 2.6: The two individuals' relation contact degree in social network is $\rho(v_k, v_s) = a + bi + cj$, the ratio of relative same degree e^a and product of the relative different degree e^b the relative opposed degree e^c is called the individual generalized set pair tight potential in social network, that is:

$$Tread(v_k, v_s)_G = \frac{e^{a+b}}{e^c} \tag{7}$$

The uncertainties term (different degree) are converted to the same term (same degree) to study the trend that the node and other nodes in social network.

For the generalized set pair loose potential and generalized set pair tight potential to analyse respectively from the angle of node relation may be looser or tighter. They can also be divided again.

The generalized set pair loose potential is the lower limit of the generalized set pair potential, the generalized set pair tight potential is the upper limit of the generalized

set pair potential, within this range, the loose and tight relation between the two nodes is mutual restraint and mutual influence, and mutual conversion under certain conditions.

3 The entity set pair similarity degree calculation method based on attribute relation

The social network is a complex network, the relation of the nodes and among them all have the corresponding attribute feature, this paper will build attribute figure based on the attribute feature, in the meanwhile to describe the relation of entity attributes and among them. To calculate the similarity of the entity in the network, in the meanwhile to concern attribute information of the node and relation information among them. At present, the clustering method of taking the object attribute and relation information among the objects into account and collectively called the clustering based on attribute-relation. Thereinto, the hypertext document clustering, scientist cooperative relation clustering, telecom customer division, and so on. Because of the rich data resource, attribute of the research object and relation information is easier to obtain, and has taken the lead to develop.

An important class algorithm among the clustering method based on attribute-relation is the method based on similarity degree. The representative is the HyPursuit algorithm of Weiss et al [16], the M-S algorithm of Modha et al [17], Wu Lingyu of University of Science and Technology Beijing has proposed the calculation method of the integrated similarity degree among the object and the similarity degree among the class. And design the appropriate strategy to achieve clustering from the bottom up. The algorithm calculated based on the attributes of the object distance and relation distance, assuming these attribute values and relation are certain and completed. In fact, the attribute value or relation is vague, uncertain and incomplete, so only utilize the attribute difference value to calculate the similarity degree among the object is not enough. This section proposes the set pair similarity calculation method of attribute-relation based on set pair analysis theory. In the meanwhile, consider the same, non-same and unsure of the attribute and relation whether is the same three-dimensional similarity, so as to solve the unicity problem in the work [16-18].

3.1 THE ENTITY SET PAIR ATTRIBUTE CONTACT DEGREE BASED ON ATTRIBUTE

In allusion to the complexity of the node in social network, and the uncertainty of node relation, we give the following definition:

Definition 3.1: (entity set pair attribute contact degree) assuming all the entities in the social network as domain set $U = \{v_1, v_2, \dots, v_n\}$, v_k and v_s are entities, they are of

the feature attribute set are $\alpha(v_k) = \{\alpha_{k1}, \alpha_{k2}, \dots, \alpha_{km}\}$ and $\alpha(v_s) = \{a_{s1}, a_{s2}, \dots, a_{sm}\}$, the number of the same attribute value of v_k and v_s is p , the number of the different attribute value is q , the number of the uncertain attribute value is r , and $p+q+r=m$, the similarity contact degree of the entity v_k and v_s in network is:

$$\tau(v_k, v_s) = \frac{p}{m} + \frac{r}{m}i + \frac{q}{m}j = a_{ks} + b_{ks}i + c_{ks}j, \quad (8)$$

$$0 \leq \tau(v_k, v_s) \leq 1.$$

Thereinto $v_k \in U$ and $v_s \in U$, i is the differential label and takes the different in [-1,1] depending on the circumstance. j only plays a marked role and takes the value is -1.

Definition 3.2: (Set pair attribute similarity contact vector) assuming in network, the node v_k has the number of neighbour node is m , the similarity degree of its neighbour node and among them to build the vector $L(v_k) = (\tau(v_k, v_s))_{m \times m}$, $L(v_k)$ is called the set pair attribute similarity contact vector, it is expressed as:

$$R(v_k) = \{\tau(v_k, v_{k1}), \tau(v_k, v_{k2}), \dots, \tau(v_k, v_{km})\}. \quad (9)$$

In allusion to the node in social network, then calculate the set pair attribute contact degree of two entity object of connecting the edge, and signs in network to build the weighted network based set pair attribute contact degree, as showed in Figure 1.

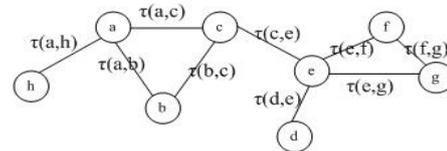


FIGURE 1 The weighted network based set pair attribute contact degree

The set pair attribute similarity contact vector of every node is showed in Table 1.

TABLE 1 The set pair attribute similarity contact vector of Figure 1

| Node | Similar vector |
|----------|---|
| <i>a</i> | { $\tau(a,h)$, $\tau(a,b)$, $\tau(a,c)$ } |
| <i>b</i> | { $\tau(a,b)$, $\tau(a,c)$ } |
| <i>c</i> | { $\tau(a,c)$, $\tau(b,c)$, $\tau(c,e)$ } |
| <i>d</i> | { $\tau(d,e)$ } |
| <i>e</i> | { $\tau(c,e)$, $\tau(d,e)$, $\tau(e,f)$, $\tau(e,g)$ } |
| <i>f</i> | { $\tau(e,f)$, $\tau(f,g)$ } |
| <i>g</i> | { $\tau(f,g)$, $\tau(e,g)$ } |
| <i>h</i> | { $\tau(a,h)$ } |

When to calculate the set pair attribute similarity contact based on the node, we should only consider the node connected with it and ignore the node without association. The vector makes preparation for calculating the comprehensive set pair similarity degree later.

3.2 THE ENTITY SET PAIR ATTRIBUTE SIMILARITY CONTACT DEGREE BASED ON RELATION

Definition 3.3: (the one level neighbour entity) the entity which connected to certain entity v_q in network is called one level neighbour entity of this entity, all the one level neighbour entity of the entity v_q to build the set and denoted as $L(v_q)$.

As showed in Figure 1, the one level neighbour entity of the entity e is c, d, f, g and denoted as $L(e) = \{c, d, f, g\}$.

Definition 3.4: (the common one level neighbour entity) if certain entity v_q and entity v_r, v_s connect directly in network, we call the entity v_q is the common one level neighbour entity for entity v_r, v_s . The common one level neighbour entity of entity v_r, v_s to build the set and denoted as $LQ(v_r, v_s)$. Obviously:

$$LQ(v_r, v_s) = L(v_r) \cap L(v_s). \tag{10}$$

Definition 3.5: (the two level neighbour entity): the one level neighbour entity of the one level neighbour entity of certain entity v_q in network, we call all the two level neighbour entity of the entity v_q to build the set and denoted as $LE(v_q)$.

As shown in Figure 1, the one level neighbour entity of the entity e is a, b and denoted as $L(e) = \{c, d, f, g\}$.

Definition 3.6: (the common two level neighbour entity) if certain entity v_q is the two level neighbour entity of entity v_r and v_s in network, we call entity v_q is the common two level neighbour entity of entity v_r and v_s , the two level neighbour entity of entity v_r and v_s build the set and denoted as $LEQ(v_r, v_s)$. Obviously:

$$LEQ(v_r, v_s) = LE(v_r) \cap LE(v_s). \tag{11}$$

Here, we think that the two neighbour entity may have relation with the entity v_q and may not. So, it is uncertain, yet, the node in addition to one and two neighbour entities have relation with entity is slim chance, so the contact with the other entities is to the contrary.

Definition 3.7: (the entity set pair similarity contact degree based on relation) assuming the number of entity is n in network, the entity v_r and v_s have the common one level entity set is $LQ(v_r, v_s)$ and the common two level entity set is $LEQ(v_r, v_s)$, and denoted as:

$$\mu(v_r, v_s) = \frac{|LQ(v_r, v_s)|}{n} + \frac{|LEQ(v_r, v_s)|}{n} i + \frac{|n - LQ(v_r, v_s) - LEQ(v_r, v_s)|}{n} j. \tag{12}$$

The entity set pair similarity contact degree based on relation, it is abbreviated as:

$$\mu(v_r, v_s) = u_a + u_b i + u_c j. \tag{13}$$

3.3 THE ENTITY SET PAIR ATTRIBUTE SIMILARITY CONTACT DEGREE BASED ON ATTRIBUTE-RELATION

Definition 3.8: (The entity set pair attribute synthesized similarity contact degree based on attribute-relation) assuming v_k, v_s are the two entities based on attribute contact degree in weighted network, so the entity set pair synthesized similarity contact degree based on attribute-relation of v_k, v_s is:

$$Sim(v_k, v_s) = \alpha \times \tau(v_k, v_s) + (1 - \alpha) \times u(v_k, v_s), \tag{14}$$

abbreviated as:

$$Sim(v_k, v_s) = Sim_a + Sim_b i + Sim_c j, \tag{15}$$

where $\tau(v_k, v_s)$ is set pair contact degree for the entity v_k, v_s , $u(v_k, v_s)$ is the entity set pair similarity contact degree for the entity v_k, v_s . α is parameter and $\alpha \in [0, 1]$.

The entity set pair synthesized similarity contact degree based on attribute-relation will make further fusion for attribute and relation of the entity, than the single entity set pair similarity contact degree based on attribute or the entity set pair similarity contact degree based on relation can better reflect the true similarity degree among the entities, and have better entity recognition ability.

3.4 THE SUB GRAPH SET PAIR SYNTHESIZED SIMILARITY CONTACT DEGREE BASED ON ATTRIBUTE-RELATION

Definition 3.9: (The block graph set pair attribute synthesized similarity contact degree based on attribute-relation) assuming $G_p = \{V_1^p, V_2^p, \dots, V_{|G_p|}^p\}$ and $G_q = \{V_1^q, V_2^q, \dots, V_{|G_q|}^q\}$ is the two sub graph based on in weighted network set pair attribute contact degree, so the sub graph set pair synthesized similarity contact degree based on attribute-relation for the sub graph G_p and G_q is:

$$Sim(G_p, G_q) = \frac{\sum_{j=1}^{|G_p|} \sum_{i=1}^{|G_q|} Sim(V_i^p, V_j^q)}{|G_p| \times |G_q|} \quad (16)$$

4 The network associations' detection algorithm

After giving the definition of entity set pair synthesized similarity contact degree based on attribute-relation, the network association partitioning has transformed into a clustering problem.

4.1 THE ALGORITHM IDEA

First, taking each entity as an initial sub graph, iterative selected two sub graph of the highest value of the entity set pair synthesized similarity contact degree based on attribute-relation to merge, until all the entities are divided into one sub graph. The clustering results is final to output one tree, the root node contains all entities, the leaf nodes as a single entity.

When to solve the practical problem can terminate the iterative process, with the fusion of sub graph, the biggest similarity contact value has showed a decreasing trend, So we can set the set pair synthesized similarity degree threshold χ , When the value of Sim_a of set pair synthesized similarity degree current maximum sub graph is less than the threshold value, considering it has finished the clustering of similar entities, the algorithm is over and output sub graph and isolated node (sub graph contains only one entity).

4.2 THE ALGORITHM STEPS

Input:

The social network consists of m entities, the value set of every entity has n attribute feature, the calculation parameter of set pair synthesized similarity degree is α and $\alpha \in [0,1]$, the set pair similarity degree threshold is χ .

Output:

The associations and isolated node.

Steps:

1) Assuming the social network attribute figure of describing the n nodes is $GA = (V(VA, LV), E(EA, LE))$, the vertex attribute threshold as showed in Table 2.

TABLE 2 The vertex threshold table

| | | | | |
|-----------|--------------|--------------|-----|----------------|
| v | va_1 | va_2 | ... | $va_{ v }$ |
| v_1 | Lva_{11} | Lva_{12} | ... | $Lva_{1 v_1 }$ |
| v_2 | Lva_{21} | Lva_{22} | ... | $Lva_{2 v_2 }$ |
| ... | ... | ... | ... | ... |
| $v_{ v }$ | $Lva_{ v 1}$ | $Lva_{ v 2}$ | ... | $Lva_{ v v }$ |

2) From the threshold table to calculate the attribute contact degree $\tau(v_k, v_s)$, ($k, s = 1, 2, 3, \dots, n$) between the two nodes, then to get the set pair attribute contact degree

matrix, The Table 3 shows the set pair attribute contact degree matrix of possessing the 5 nodes.

Table 3 set pair attribute contact degree matrix

| | v_1 | v_2 | v_3 | v_4 | v_5 |
|-------|------------------|------------------|------------------|------------------|------------------|
| v_1 | $\tau(v_1, v_1)$ | $\tau(v_1, v_2)$ | $\tau(v_1, v_3)$ | $\tau(v_1, v_4)$ | $\tau(v_1, v_5)$ |
| v_2 | $\tau(v_2, v_1)$ | $\tau(v_2, v_2)$ | $\tau(v_2, v_3)$ | $\tau(v_2, v_4)$ | $\tau(v_2, v_5)$ |
| v_3 | $\tau(v_3, v_1)$ | $\tau(v_3, v_2)$ | $\tau(v_3, v_3)$ | $\tau(v_3, v_4)$ | $\tau(v_3, v_5)$ |
| v_4 | $\tau(v_4, v_1)$ | $\tau(v_4, v_2)$ | $\tau(v_4, v_3)$ | $\tau(v_4, v_4)$ | $\tau(v_4, v_5)$ |
| v_5 | $\tau(v_5, v_1)$ | $\tau(v_5, v_2)$ | $\tau(v_5, v_3)$ | $\tau(v_5, v_4)$ | $\tau(v_5, v_5)$ |

3) According to the attribute contact degree and the edge adjacency matrix to form weighted network set pair similarity contact vector which based on set pair attribute contact degree.

4) In weighted network, identifying the common one level neighbour entity and the common two level neighbour entity of any two nodes v_k, v_s , calculating the entity set pair similarity contact degree $\mu(v_k, v_s) = u_a + u_b i + u_c j$, where ($k, s = 1, 2, \dots, n$) based on relation.

5) According to the value of α to calculate the entity set pair synthesized similarity contact degree $Sim(v_k, v_s) = Sim_a + Sim_b i + Sim_c j$ of any two nodes v_k, v_s based on attribute-relation.

6) Each node is initialized to a sub graph.

7) Identifying the two looser sub graphs of the current state of the entity set pair synthesized similarity contact degree under the generalized set pair tight potential, calculating the contact degree among the sub graph block.

8) Judging the set pair synthesized similarity contact degree $Sim_a > \chi$ among $Sim(G_p, G_q)$ if it is set up then merge the entity node G_p, G_q and denote

$G_{max} = G_p \cup G_q$, otherwise end the algorithm and output sub graph and isolated node.

9) Applying the Equation (16) to update the set pair synthesized similarity contact degree of G_{max} and other sub graphs. Returning to 8).

By the analysis, the clustering process is actually the process of network association partitioning its time complexity is $O(m)$. Herein, the number of entity node in network is n , the number of iterations is t , the efficiency of the algorithm is higher. In addition, when to calculate the similarity contact degree, at the same time to take into the certain-uncertain of node attribute and the distance of node attribute set pair distance and the object, the result of division of the algorithm has more ideal.

5 The calculation example

Next, we set one simple network figure as example to illustrate the process of partitioning, as shown in Figure 2.

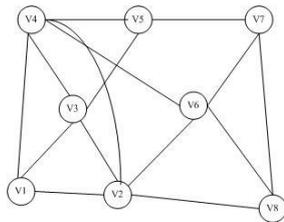


FIGURE 2 One simple network

Assuming the attribute set of eight nodes is shown in Table 4, thereinto, the blank indicates unknown, which is the uncertainty.

TABLE 4 The node attribute set in Figure 3

| Attribute node | A1 | A2 | A3 | A4 | A5 |
|----------------|----|--------|--------|--------|--------|
| V_1 | F | JS | Jsj | Shuxue | |
| V_2 | M | FJS | Jsj | | fayu |
| V_3 | F | | Dianzi | Shuxue | yingyu |
| V_4 | M | JiangS | | Shuxue | Yingyu |
| V_5 | M | Zhuj | Zdh | | fayu |
| V_6 | | JS | Jsj | Shuxue | fayu |
| V_7 | M | Zhuj | | wuli | yingyu |
| V_8 | M | | Zdh | wuli | fayu |

TABLE 5 The set pair similarity vector of the nodes

| node | similarity vector |
|-------|---|
| V_1 | { $1/5+2/5i+2/5j$, $2/5+2/5i+1/5j$, $1/5+2/5i+2/5j$ } |
| V_2 | { $1/5+2/5i+2/5j$, $2/5i+3/5j$, $1/5+2/5i+2/5j$, $2/5+2/5i+1/5j$, $2/5+2/5i+1/5j$ } |
| V_3 | { $2/5+2/5i+1/5j$, $2/5i+3/5j$, $2/5+2/5i+1/5j$, $3/5i+2/5j$ } |
| V_4 | { $1/5+2/5i+2/5j$, $2/5+2/5i+1/5j$, $1/5+2/5i+2/5j$, $1/5+2/5i+2/5j$, $1/5+2/5i+2/5j$ } |
| V_5 | { $3/5i+2/5j$, $1/5+2/5j+2/5j$, $2/5+2/5i+1/5j$ } |
| V_6 | { $2/5+2/5i+1/5j$, $1/5+2/5i+2/5j$, $1/5+2/5i+2/5j$, $1/5+2/5i+3/5j$, $1/5+2/5i+2/5j$ } |
| V_7 | { $2/5+2/5i+1/5j$, $2/5i+3/5j$, $2/5+2/5i+1/5j$ } |
| V_8 | { $2/5+2/5i+1/5j$, $1/5+2/5j+2/5j$, $2/5+2/5i+1/5j$ } |

TABLE 6 The result of the synthesized set pair contact degree of all the node in FIGURE 3

| | V_1 | V_2 | V_3 | V_4 | V_5 | V_6 | V_7 | V_8 |
|-------|-----------|-----------|-------------|-----------|--------------------------|---------------|---------------------|--------------------------|
| V_1 | $1+0i+0j$ | $0.2250+$ | $0.3250+0.$ | $0.2250+$ | $0.1250+0.3250i+0.5500j$ | $0.3625+0.16$ | $0.0000+0.2000i+0.$ | $0.0625+0.2625i+0.6750j$ |
| V_2 | | $1+0i+0j$ | $0.1250+0.$ | $0.2875+$ | $0.3250+0.1000i+0.5750j$ | $0.4250+0.16$ | $0.2250+0.2625i+0.$ | $0.2650+0.3250i+0.4125j$ |
| V_3 | | | $1+0i+0j$ | $0.3875+$ | $0.0625+0.4250i+0.5125j$ | $0.1625+0.20$ | $0.1625+0.2000i+0.$ | $0.2625+0.2000i+0.5375j$ |
| V_4 | | | | $1+0i+0j$ | $0.1625+0.3250i+0.5125j$ | $0.2250+0.20$ | $0.3250+0.1625i+0.$ | $0.2250+0.2000i+0.5750j$ |
| V_5 | | | | | $1+0i+0j$ | $0.2250+0.26$ | $0.2000+0.2625i+0.$ | $0.0625+0.3625i+0.5750j$ |
| V_6 | | | | | | $1+0i+0j$ | $0.0625+0.2625i+0.$ | $0.2875+0.3875i+0.3250j$ |
| V_7 | | | | | | | $1+0i+0j$ | $0.3250+0.3250i+0.3500j$ |
| V_8 | | | | | | | | $1+0i+0j$ |

6) Calculating the generalized set pair tight potential among every node and other nodes, if the generalized set pair tight potential greater than or equal to 1, then divided into a network, calculating the synthesized set pair similarity degree of the network sub graph until all the

- 1) Calculating the attribute set pair contact degree matrix between the two nodes.
- 2) Building the weighted network based on attribute set, as shown in Figure 3:

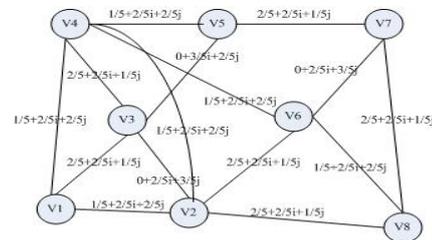


FIGURE 3 the weighted network based on attribute set pair contact degree

- 3) According the weighted network to build the set pair similarity vector, as shown in Table 5.
- 4) According to the weighted network and adjacency matrix to identify the common one level neighbour entity and the common two level neighbour entity, then calculate the set pair contact degree based on relation.
- 5) Setting $\alpha = 0.5$, then calculate the synthesized set pair similarity degree of every node, as shown in Table 6.

nodes are divided into the corresponding network or become isolated nodes.

The result of the final partition in this instance, as shown in Figure 4.

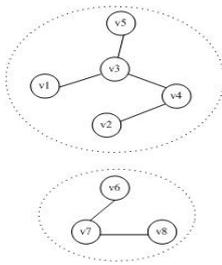


FIGURE 4 the network partition result of Figure 3

6 Conclusions

The better your paper looks, the better the Journal looks. Thanks for your cooperation and contribution. This article aims at the node and its relation in network to conduct analysis based on set pair analysis ideology. Consider the node has attribute and attribute values may exist the uncertainty, and the node relation contains the

References

- [1] Scott J 2000 *Social Network Analysis: a Handbook Sage Publications London*
- [2] Lin J 2009 *The social network analysis, method and its application Beijing Normal University Press Beijing*
- [3] Girvan M, Newman M E J 2002 *Community Structure in Social and Biological Networks PNAS* 12 7821-6
- [4] Budak C, Agrawal D, Abadi A E 2011 *Structural trend analysis for online social networks Proceedings of the VLDB Endowment* 4(10) 101-11
- [5] Deng C, Zheng S, He X, Yan X, Han J 2005 *Mining Hidden Community in Heterogeneous Social Networks LinkKDD'05 Chicago, IL USA*
- [6] Yang D, Xiao Y, Wang W 2010 *Statistical Model-Based Analysis and Predication of Collective Attentions in social networks Journal of Computer Research and Development* 47(Suppl) 378-84
- [7] Liu Z, Li P, Zheng Y, Sun M 2008 *Community detection by affinity propagation, Technical Reports No. 001, 200, Dep of Computer Science and Technology Tsinghua University Beijing China*
- [8] Hu Y, Li M, Zhang P, Fan Y, Di Z 2008 *Community detection by signaling on complex networks Phys Rev E* 78 016115
- [9] Guo J, Hao D 2010 *One Same Name Entity Distinguish Algorithm Based on the Attribute Relationship Diagram ICIC Express Letters (in Chinese)*
- [10] Zhang C, Liang R, Liu L 2011 *The model of set pair social network analysis and its application Journal of Hebei Polytechnic University* 33(3) 99-103
- [11] Xiang B, Chen B H, Zhou T 2009 *Finding community structure based on subgraph similarity Studies in Computational Intelligence* 207(5) 73-81
- [12] Claset A, Newman M E J, Moore C 2004 *Finding community structure in very large networks Physical Review E* 70(6) 066111
- [13] Pan Y, Li D-H, Liu G-J, Liang J-Z 2010 *Detecting community structure in complex networks via node similarity Physica A: statistical Mechanics and its Applications* 389(14) 2849-57
- [14] Leicht E A, Holme P, Newman M E J 2006 *Vertex similarity in networks Physical Review E* 73(2) 026120
- [15] Guo J, Zhang C, Chen X 2011 *Attribute Graph and Its Structure ICIC Express Letters* 5(8A) 2611-6
- [16] Weiss R, Velez B, Sheldon M 1996 *HyPursuit: a hierarchical network search engine that exploits content-link hypertext clustering Proceedings of the 7th ACM conference on Hypertext, ACM Press New York* 180-93
- [17] Modha D, Spangler W 2004 *Clustering hypertext with applications to Web searching USA 2004/0049503AI 3-11*
- [18] Wu L, Gao X, Wu S 2011 *Clustering algorithm based on attribute-relationship integrated similarity Application Research of Computer* 28(1) 44-7

Authors



Yanjun Zhao, born on July 5, 1977, Tangshan, China

Current position, grades: master of engineering, lecturer of Hebei Untied University.

University studies: master's degree at Hebei Technology University.

Scientific interest: data mining, social network, computer network.

Publications: 10 papers.

Experience: Teacher of computer fundamentals for more than ten years in Hebei Untied University.



Zhang Chunying, born on October 3, 1969, Tangshan, China

Current position, grades: Ph.D. candidate, master supervisor, professor of Hebei Untied University.

University studies: Ph.D degree in computer science and its application at Yanshan University in 2012.

Scientific interest: data mining, social network.

Publications: 30 papers.

Comparative study of DXT1 texture encoding techniques

Jizhen Ye¹, Jian Wei², Yan Huang^{1*}, Jingliang Peng¹

¹*School of Computer Science and Technology, Shandong University, Jinan, China*

²*Qualcomm Inc., San Diego, U.S.A.*

Received 1 March 2014, www.tsi.lv

Abstract

In this paper, we make a comprehensive survey of many different methods to implement DXT1 (a widely used lossy texture compression algorithm). Besides that, we propose two new methods that aim for computing speed and image quality, respectively to implement DXT1 texture compression algorithm. For computing speed, we propose a new method called Lsq3d fit which achieves a very fast speed to encode texture images while keeping acceptable image quality. For image quality, we propose a new method called kmeans iteration fit and make a combination of it and the cluster fit from libsquish (an open source lib for DXTC). Kmeans iteration fit performs competitively in the quality of compressed texture images compared with the state-of-the-art DXT1 encoders, and we achieve different levels of quality by controlling the times of iteration. Finally, we test all the methods on Kodak Lossless True Color Image Suite, and CSIQ (Computational Perception and Image Quality Lab) image dataset. Our proposed methods have competitive results of speed and quality in both image datasets. The combination of cluster fit and kmeans iteration fit defeats all other methods in the quality of compressed images.

Keywords: Texture compression, DXTC, DXT1, S3TC, k-means clustering

1 Introduction

Textures play an important role in computer graphics. They are used to increase the realism of the rendered scenes by adding visual details to geometric models. However, textures can not only consume large amounts of system and video memory, but also take up a lot of bandwidth which usually limits the performance of modern rasterizer architectures for computer graphic system [1].

To solve these problems, Knittel et al. [2] and Beers et al. [3] proposed texture compression whose main idea is to conduct lossy compression on texture images, and store the compressed version of the textures. On one hand, as textures have been compressed before they are transferred to video memory by bus, we can save both bandwidth and memory. On the other hand, when accessing the compressed textures during rendering, the compressed textures should be decompressed on-the-fly in real time and support random access. Therefore texture compression is not the same as general image compression. It has its own properties to satisfy peculiar application requirements. We will introduce the main differences between texture compression and image compression in Section 2.1. Most of today's graphics cards allow textures to be stored in a variety of compressed formats that are decompressed in real-time during rasterization [1]. One such format which is supported by most graphics cards is S3TC, also known as DXT compression [4, 5].

The family of DXT compression formats is made up of DXT1, DXT2, DXT3, DXT4 and DXT5. They are different in the way they handle the alpha channel. In this paper, we focus on DXT1 which is the base of other DXT formats. DXT1 [6] is simple, whose basic idea is to divide a texture image into many 4×4 pixel blocks and encode each block independently. Every encoded block is composed of two parts. The first part is used to store two 16-bit RGB565 colours c_0 and c_1 . The second part is used to store 16 2-bit colour indices. The structure of encoded DXT1 block is shown in Figure 1. If the first base colour c_0 as a 16-bit unsigned integer is greater than c_1 , two other colours c_2 and c_3 are calculated as follows: $c_2 = (2c_0 + c_1)/3$ and $c_3 = (c_0 + 2c_1)/3$. Otherwise, $c_2 = (c_0 + c_1)/2$ and c_3 is transparent black. The indices are used to determine the colour value for each pixel. The base colours c_0 and c_1 are the most important to determine the colour quality of each block. How to choose two base colours that can best represent the 4×4 block has been the main focus of DXT1.

In this paper, we make a comprehensive survey of many different encoding techniques conforming to the DXT1 texture compression standard, and test these methods on two widely used image datasets (CSIQ and Kodak). We also propose two new DXT1 texture encoding algorithms that aim for computing speed and image quality, respectively. Experimental results on Kodak image dataset and CSIQ image data set indicate that our methods have outstanding performance in speed or quality.

* *Corresponding author* e-mail: yan.h@sdu.edu.cn

2 Related work

This section introduces the main differences between general image compression and texture compression.

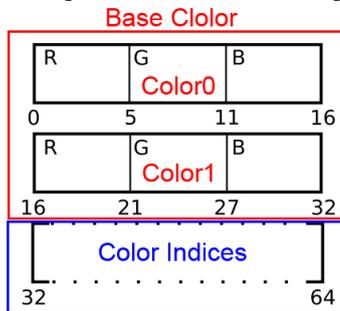


FIGURE 1 The structure of DXT1 encoded block. The red part is two base colours and the blue part is each pixel's colour indices to four colours c_0 , c_1 , c_2 and c_3

2.1 GENERAL IMAGE COMPRESSION VS TEXTURE COMPRESSION

Texture compression is a kind of special compression method which has its own properties to satisfy specific application requirements. Beers et al. [3] list four factors that should be considered when evaluating a texture compression scheme, as described below.

Decoding speed: The accessing of texture data is a critical part in the texturing operations. So it is highly desirable to be able to render directly from the compressed texture data. The decoding algorithm of texture compression should be relatively simple to reduce the cost of hardware and should not impact rendering performance.

Random access: As objects may be oriented and obscured arbitrarily, it is therefore difficult to predict the order that a renderer accesses texels. As such, any texture compression scheme must allow fast random access to compressed texture data.

Encoding speed: In most applications that are related to textures, the majority of textures are compressed well in advance of the rendering, which we often call off-line encoding. It is therefore feasible to employ a scheme where the encoding is considerably slower than the decoding.

Compression rate and visual quality: Because the most important issue in texturing is the quality of the rendered scene rather than the quality of individual textures themselves, some loss of fidelity in the compressed texture is more tolerable than image compression.

From the above descriptions, we know that most general image compression schemes, e.g. JPEG, cannot support direct random access of pixel data within the compressed format because the per-pixel storage rate varies throughout. As such, many texture compression schemes we introduced later, which employ 'fixed rate' encoding.

2.2 HISTORY OF DXTC

In 1979, Delp and Mitchell [7] developed a simple scheme, called block truncation coding (BTC) for image compression. BTC compressed grey scale images in blocks of 4x4 pixels. For each block, two representative 8-bit grey scale values are chosen and each pixel within the block is quantized to either of these two values. This resulted in 2 bits per pixel (2bpp).

Campbell et al. [8] presented colour cell compression (CCC), which is often seen as a simple extension of BTC. CCC stores two colour indices to a palette in a 4x4 block instead of 2 grey scale values in BTC. By using a 256-wide colour palette, the colours can be represented with eight bits each. Thus CCC can encode colour images at 2 bpp. However, the limitation of only two colours in a 4x4 block gives rise to banding artifacts and CCC requires a memory lookup in the palette. Knittel et al. [2] suggested that CCC be implemented in hardware and used in a texturing system.

The S3TC texture compression method by Iourcha et al. [9] is a further adaption of the BTC/CCC method by improving colour data encoding. S3TC (a.k.a DXTC) is the de facto standard in texture compression method. Unlike CCC, it stores two base colours in R5G6B5 format and a 2-bit index for each pixel in a 4x4 block. Each pixel can have four colours to choose, two base colours and two additional colours in-between the base colours.

2.3 EXISTING IMPLEMENTATIONS OF DXT1 ENCODER

From the above descriptions, we know that all DXTC's colours in a block lie on a line in colour space of RGB. To choose two base colours that can best represent all the colours of each block is the main work of DXTC encoder. There are several good DXT compressors available. Such as the ATI Compressorator [10] and the nVidia DXT Library [11], squish library [12], crunch lib [13], LSDxt DXT Compressor [14], Jason Dorie's image library: ImageLib, Mesa S3TC compression library: libtxc_dxtn [16] and so on.

All the above encoders produce different levels of quality to DXT compressed texture images, and some of them are not open source. So it is very meaningful to give a comparative study on these different methods.

It is known that texture compression does not require real-time encoding integral, but in some particular applications real-time compression is also important. A good DXT encoder should provide two different choices for users to choose. One is to compress a texture image very fast, while the quality of the texture can have more error tolerance. The other is to compress a texture with more time and produce high quality of DXT compressed texture images.

3 Encoding approaches

This section includes two parts. Part one introduces some encoding techniques in the open source squish lib. Part two describes two of our novel methods and a combination of cluster fit and kmeans iteration fit which has the best quality of all methods.

3.1 METHODS OF SQUISH LIBRARY

The squish library (abbreviated to libsquish) is an open source DXT compression library that was originally written by Simon Brown et al. Range fit and Cluster fit are based on a concept called principal component [15].

3.1.1 Range fit

For range fit method, it takes the minimum and maximum point along the principal axis as the endpoints directly. Although this method is quite simple and there may be better colour endpoints that are not part of the original point set, it can find the base colour points very fast and keep acceptable quality of compressed texture images. So range fit is a good choice for those applications that require very fast encoding speed, also known as real-time compression applications.

3.1.2 Cluster fit

Range fit takes endpoints from the original colour set, this is not the best choice in most cases. Cluster fit method is under the assumption that: If we assume that the principal axis is very similar to the direction of the line through optimal endpoints, we can also assume that a total ordering of the original colour set in these directions is also very similar. So cluster fit uses the principal axis to define a total ordering on the original colour set. It then tests all possible ways of clustering the original points that preserve this total ordering, and fit endpoints to each generated index set using least squares.

The cluster fit algorithm in squish now forms the core DXT compression algorithm for the NVIDIA Texture Tools.

3.2 OUR NOVEL METHODS AND COMBINATION METHOD

In this section, we will present our novel methods in detail.

3.2.1 Lsq3d fit

It is known that the method of Least Squares is a procedure to determine the best fit line to data.

Line fitting in computational mathematics often fits lines in 2D space with least square method. The fitted equation is $y = ax + b$. However, we want to fit lines in three-dimensional RGB space. We cannot use the linear

least square method directly to determine the linear equation of 3D line. We refer to a math paper: Fitting of the Straight Line Equation in Space [17] to fit space line equation. Its base idea is to convert space line equation to plane line equation.

With the space line, Lsq3d fit takes two endpoints that have the maximum span in the line. Unfortunately, sometimes space line cannot be represented by the above equations. We choose range fit to encode those corresponding blocks. This Lsq3d fit method is proposed for high computing efficiency.

3.2.2 Kmeans iteration fit

Kmeans iteration fit method is based on k-means clustering algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. For each DXT block, we partition 16 colour points into 4 clusters firstly. Then we minimize Equation (1) that calculates the Euclidean distance between the DXT base colour points and cluster centre points.

$$f(x_0, y_0, z_0, x_1, y_1, z_1) = \sum_{i \in \text{clusters}[j]} \sum_{j=0}^3 w_i (\text{Observations}[i] - \text{Centers}[j])^2, \quad (1)$$

where clusters is the result of kmeans clustering, Centers is the centre of each cluster, Observations is the cluster member of each cluster, and w is the weight of each cluster, it is calculated by $\text{clusters}[i].\text{number}/16.0$.

To minimize $f(x_0, y_0, z_0, x_1, y_1, z_1)$, we calculate its partial derivative of each parameter.

Solving this linear equations, we can obtain values of x_0, y_0, z_0, x_1, y_1 and z_1 , which are the coordinates of the DXT base colours.

Kmeans iteration fit can iterate for a user-specified number of times. The $c_0(x_0, y_0, z_0)$ and $c_1(x_1, y_1, z_1)$ are the base values. We feed c_0, c_1, c_2 and c_3 as the unit points to the next round of kmeans clustering. c_2 and c_3 are the interpolated values of c_0 and c_1 .

With the initial points, we can get new cluster results and new values of x_0, y_0, z_0, x_1, y_1 and z_1 . If the distance error between previous endpoints and current endpoints is less than a specified threshold then the algorithm is terminated; otherwise the iteration process continues until the distance error is smaller than the threshold or the number of iteration reaches the maximum iteration number. The definition of the distance error is shown in Equation (2):

$$\text{DisError} = \frac{\sum_{i=0}^3 (\text{curFitpoints}[i] - \text{preFitpoints}[i])^2}{\text{MaxSpan}}, \quad (2)$$

where $curFitpoints[i]$ is the fitted results of this round and $preFitpoints[i]$ is the results of previous round. $MaxSpan$ is the longest distance of any two points in the block.

Figure 2 presents the whole algorithm flow of kmeans iteration fit.

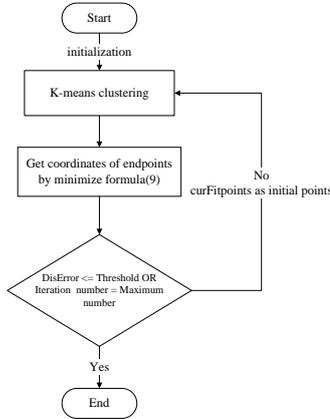


FIGURE 2 The algorithm flow chart of kmeans iteration fit. Initialization specifies some parameters to execute k-means clustering

3.2.3. Combination

We observe many fitting results (as shown in Figure 3) of different encoding methods and find that our method can have a great fitting effectiveness when the colour set in a block has a fine linearity. In order to improve the encoding method, we propose a combination version.

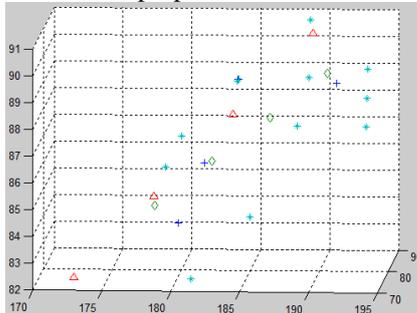


FIGURE 3 The fitting results of a block. Where the cyan stars represent the colour points, red triangles are fitting points of squish's cluster fit, blue add symbols represent cluster centres and the green diamonds are the fitting results of our kmeans iteration fit

For the combination method, we should make a judgment before we encode a block. If the colour points are of high linearity, then we choose kmeans iteration fit to encode. Else the colour points are encoded by cluster fit. We choose two factors to evaluate the linearity of colour points. One is the variance of the distances of cluster centres, and the other is the angle of centre lines. The specific formulas are shown in Equation (3):

$$\begin{aligned}
 Dis_{ij} &= (center[j] - center[i])^2, \\
 Variance &= \frac{(Dis_{ij} - \overline{Dis})^2}{3}, \\
 \theta &= \text{Max}(\angle Dir_{01} Dir_{02}, \angle Dir_{32} Dir_{31}),
 \end{aligned}
 \tag{3}$$

where $i = \{0, 1, 2\}, j = i + 1, Dis_{ij}$ is the distance between $center[i]$ and $center[j]$. Variance represents the variance of Dis_{01}, Dis_{12} and Dis_{23} .

We specify the value of *Variance* and θ by a lot of experiments. The experimental results indicate that this combination method has the best quality of compressed texture images, that is to say this method yields the highest PSNR values and the smallest *RmsError* values.

The flowchart of the combination method are shown in Figure 4.

4 Datasets and quality evaluating metrics

In this section, we introduce two most popular available image datasets that are used to test all DXT1 encoding methods. They are Kodak Image Dataset [18] and CSIQ (Computational Perception and Image Quality Lab) image dataset [19]. The quality evaluation metrics used in our experiment are introduced too.

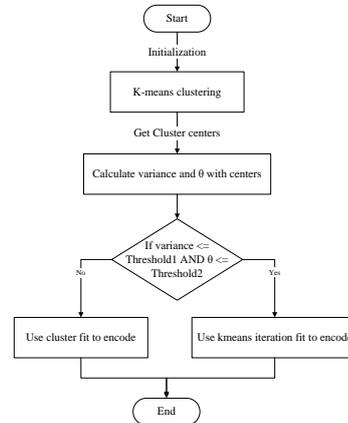


FIGURE 4 The flowchart of combination method. Initialization specifies some parameters to execute k-means clustering

Kodak Image Dataset [18] is released by the Kodak Corporation for unrestricted research usage. There are totally 25 uncompressed PNG true colour images of size 768×512 pixels in it. CSIQ (Computational Perception and Image Quality Lab) image dataset [19] was released by Computational Perception and Image Quality Lab. It consists of 30 original PNG images and six different types of distortions at four to five different levels of distortion. We only use the original images in our experiment. All 30 original images are of size 512×512 pixels and can be divided into five different types and 6 images for each type. They are animals, landscape, people, plants and urban.

In our experiment, we choose *RmsError* which is the square root of *MSE* (Mean Square Error) and *PSNR* (Peak signal-to-noise ratio) as quality evaluating metrics.

PSNR is most commonly used to measure the quality of reconstruction of loss compression codecs (e.g., for image compression). The signal in this case is the original data, and the noise is the error introduced by compression.

MSE is shown in Equation (4):

$$MSE = \frac{1}{3} \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [R_{I(i,j)} - R_{C(i,j)}]^2 + [G_{I(i,j)} - G_{C(i,j)}]^2 + [B_{I(i,j)} - B_{C(i,j)}]^2, \tag{4}$$

where I is the original texture image and C is the compressed texture image. m is the width of the image and n is the height of the image. As we will evaluate the quality of RGB image, we should consider RGB three channels.

PSNR is calculated with Equation (5), where max is 255:

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right). \tag{5}$$

5 Experimental results

In this section, we use Kodak Image Dataset [18] and CSIQ dataset [19] to test all the DXT1 encoders we can find. We present the experimental results with tables and line charts.

Firstly, we will introduce the running environment of our experiment. All the programs run on Microsoft Visual Studio 2008 professional version. The computer used to conduct experiment is DELL OptiPlex 7010, its CPU is Intel® Core™ i5-3470 @3.20GHz, its Operation System is Windows 7 Ultimate Edition and its RAM is 4.00 GB.

TABLE 1 Experimental Results of all the methods on the two image datasets. “Can’t measure accurately” means that we have to make compression operations interactively, so the total time used to compress is not accurate. The red row is Lsq3d fit, which is the fastest method to encode, while the quality is still acceptable and defeat range fit both in time and quality. The blue row is the combination method, which has the best quality of all methods

| Methods | Kodak | CSIQ | Kodak | CSIQ | Kodak | CSIQ |
|--|-----------------|------------------|------------------|----------------------|--------------------------|--------------------------|
| | Mean RmsError | Mean PSNR(db) | Mean PSNR(db) | Running time(second) | Running time(second) | Running time(second) |
| Range fit | 9.615789 | 11.991445 | 33.47758 | 31.713790 | 1.457000 | 1.215000 |
| Lsq3d fit(Range fit for blocks can't express line) | 8.994410 | 11.723098 | 34.057239 | 32.018587 | 0.747000 | 0.650000 |
| Cluster fit + single color fit | 6.982464 | 8.799585 | 36.26763 | 34.383929 | 201.629000 | 167.624000 |
| LSDxt Engine by L. Spiro | 7.621329 | 9.577089 | 35.51408 | 33.662967 | Can't measure accurately | Can't measure accurately |
| AMD: The Compressorator | 7.006528 | 8.860144 | 36.230736 | 34.312864 | Can't measure accurately | Can't measure accurately |
| Crunch lib quality=0 | 19.036631 | 22.901982 | 27.683062 | 26.243883 | Can't measure accurately | Can't measure accurately |
| Crunch lib quality=255 | 6.976277 | 8.866489 | 36.275241 | 34.317618 | Can't measure accurately | Can't measure accurately |
| Kmeans iteration0 fit | 7.673874 | 9.305032 | 35.414205 | 33.866736 | 23.667000 | 19.702000 |
| Kmeans iteration5 fit | 7.352763 | 9.110025 | 35.764885 | 34.049590 | 65.608000 | 54.208000 |
| Combination(cluster fit and kmeans iteration fit) | 6.855624 | 8.619251 | 36.431254 | 34.556220 | 235.573000 | 197.196000 |

Table 1 shows that Lsq3d fit (marked in red) can defeat range fit both in speed and quality. Kmeans iteration fit can have competitive performance compared with other methods, but it cannot be better than cluster fit.

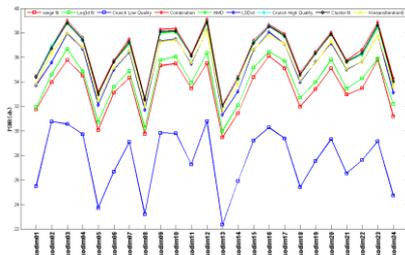


FIGURE 5 The line chart of Kodak dataset. X axis represents image names and Y axis represents the PSNR(db) value of each image

Then, we present parameter settings of our own methods. For kmeans iteration fit, we set the DisError equal to 0.001. To balance the quality and speed, we set iteration times to be five. The result of kmeans iteration fit without iteration is also given to make a comparison. For combination method, the variance threshold is 1.0 and the cosθ threshold is 0.86. We determine these values by a lot of experiments and choose thresholds that have the best effects.

The numerical experimental results can be seen in Table 1. We measure rmsError, PSNR(db) and running time(second) for each image dataset, all these values are average values of corresponding datasets. For each numerical term, the left column is the results of Kodak dataset and the right column is CSIQ’s results. As some methods only offer software or executable files, we cannot measure the running time accurately as we do in methods that have source code and the Compressorator of AMD [10] in quality no matter how many times it iterates. The combination methods (marked in blue) can have the best quality of all methods, but it needs longer time to encode. Improving the speed of this combination method is one of our future works.

In order to give a better visualization of all the methods’ compression performance on every texture image, we present two line charts for Kodak [18] and CSIQ [19] respectively as shown in Figures 5 and 6.

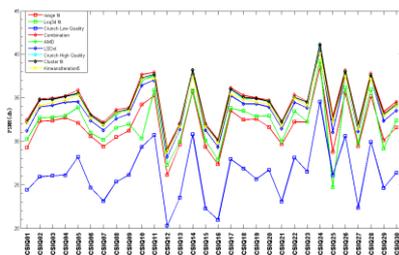


FIGURE 6. The line chart of CSIQ dataset. X axis represents image names and Y axis represents the PSNR(db) value of each image

In these two line charts, we present 9 different methods's PSNR values on each image. The corresponding methods are shown in the legend of the figure. The square symbols represent fast compression methods and the diamond symbols represent the high quality methods. To make it clearer, we use different colours for different methods. The line charts indicate that the combination method has the highest PSNR values for most images.

6 Conclusion and future work

DXTC has been a leading texture compression method for many years and is still widely used today. Correspondingly, many people focus on the implementation of DXT1 encoder. In this paper, we make a comparative study on all the encoders we can find and test them on two image datasets. Besides that, we propose our own method Lsq3d fit and k-means

iteration fit aiming at speed and quality, respectively. The combination of our k-means iteration fit and cluster fit outperforms all the other methods in quality of compressed texture images.

We will extend our work to alpha channel encoding in the future. For textures with alpha channel, DXTC uses the same idea as for colour channels to encode alpha values. The most important thing is also to find the alpha endpoints that can best represent the original 4 x 4 block. New encoding method should be designed specifically for alpha channel in the future.

Acknowledgment

We thank Qualcomm Inc. for funding this project. We also thank Simon Brown who has written the squish lib and shared the source code.

References

- [1] Aila T, Miettinen V, Nordlund P 2003 Delay Streams for Graphics Hardware *ACM Transactions on Graphics*, **22**(3) 792–800
- [2] Knittel G, Schilling A, Kugler A, Strasser W 1996 Hardware for Superior Texture Performance *Computers & Graphics* **20**(4) July 475–81
- [3] Beers A, Agrawala M, Chadda N 1996 Rendering from Compressed Textures *Proceedings of SIGGRAPH* 373–8
- [4] S3 Texture Compression Pat Brown NVIDIA Corporation November 2001 Available Online: http://oss.Sgi.com/projects/oglsample/registry/EXT/texture_compression_s3tc.txt
- [5] Compressed Texture Resources Microsoft Developer Network DirectX SDK, April 2006
- [6] Brown P, Agopian M EXT texture compression dxt1. opengl extension registry. http://opengl.org/registry/specs/EXT/texture_compression_dxt1.txt
- [7] Delp E, Mitchell O 1979 Image Compression using Block Truncation Coding *IEEE Transactions on Communications* **2**(9) 1335–42
- [8] Campbell G, Defanti T A, Frederiksen J, Joyce S A, Leske L A, Lindberg J A, Sandin D J 1986 Two Bit/Pixel Full Color Encoding *Proceedings of SIGGRAPH* **22** 215–23
- [9] Iourcha K, Nayak K, Hong Z 1999 System and Method for Fixed-Rate Block-based Image Compression with Inferred Pixels Values *US Patent* 5,956,431
- [10] ATI Compressorator Library Seth Sowerby Daniel Killebrew ATI Technologies Inc The Compressorator version 1.27.1066
- [11] NVidia DXT Library nVidia nVidia DDS Utilities April 2006 Available Online http://developer.nvidia.com/object/nv_texture_tools.html
- [12] Libsquish, open source DXT compression library writed by Simon Brown 2006 Available Online: <http://code.google.com/p/libsquish/>
- [13] Crunch, advanced DXT texture compression and real-time transcoding library. Available Online: <https://code.google.com/p/crunch/>
- [14] LSDxt DXT compressor by L Spiro October 11 2012. Available Online: <http://lspiroengine.com/?p=516>
- [15] Pearson K, 1901 "On Lines and Planes of Closest Fit to Systems of Points in Space" *Philosophical Magazine* **2**(11) 559–72.
- [16] Mesa S3TC Compression Library Roland Scheidegger libtxc_dxtn version 0.1 May 2006
- [17] Huo X, 2009 Fitting of the Straight Line Equation in Space *Journal Of Huaihua University* **28**(2) Feb 2009
- [18] Kodak Image Dataset released by the Kodak Corporation for unrestricted research usage (Image source: <http://r0k.us/graphics/Kodak>)
- [19] Larson E C, Chandler D M 2010 Most apparent distortion: full-reference image quality assessment and the role of strategy *J Electr Imaging* **19** 001006 1-21 2010

Authors

| | |
|---|--|
|  | <p>Jizhen Ye, born in 1990, Fujian, China.</p> <p>University studies: Master student, School of Computer Science and Technology, Shandong University, Jinan, China</p> |
|  | <p>Jian Wei, born in 1969, Xian, China.</p> <p>Current position: Researcher, Qualcomm multi-media R&D, San Diego, U.S.A.</p> <p>Scientific interest: computer vision, graphic technology.</p> |
|  | <p>Yan Huang, born in 1974, Tongling, Anhui, China.</p> <p>Current position, grades: associate professor in the School of Computer Science and Technology, Shandong University.</p> <p>Scientific interest: Large scale 3D data visualization, intelligent multimedia data analysis.</p> |
|  | <p>Jingliang Peng, born in 1974, Feixian, Shandong, China.</p> <p>Current position, grades: professor in the School of Computer Science and Technology, Shandong University.</p> <p>Scientific interest: digital geometry processing, content-based multi-media data retrieval, image analysis and understanding.</p> |

Real-Time and interactive browsing of massive mesh models

Xian Wu, Yan Huang*

School of Computer Science and Technology, Shandong University, Jinan, China

Received 1 March 2014, www.tsi.lv

Abstract

We present an efficient method for out-of-core construction and real-time interaction of massive mesh models. Our method uses face clustering on an octree grid to simplify and build a Level-of-Detail (LOD) tree for the model. Each octree node leads to a local LOD tree. All the top layers of the local LOD trees are combined together to make the basis of the global LOD tree. At runtime, the LOD tree is traversed top down to choose appropriate local LOD trees given the current viewpoint parameters. The system performance can be dramatically improved by using hierarchical culling techniques such as view-frustum culling and back-face culling. The efficiency and scalability of the approach is demonstrated with extensive experiments of massive models on current personal computer platforms.

Keywords: massive mesh model, out-of-core, level-of-detail, mesh simplification, culling

1 Introduction

3D mesh models are dominant in computer graphics. Applications employing meshes include movies, games, computer aided design, simulation, art and history etc. Today with the fast development of 3D acquisition, modelling and simulation technologies, we have much more complex and accurate mesh models. For instance, mesh models of gigabytes size are not uncommon nowadays.

In the last several decades, the performance of CPU and GPU has improved tremendously. However, the memory bandwidth especially disk bandwidth grows much slower. Therefore the bottleneck lies on the fact that our processor has to wait for the data stored on disk. There is a wide range of simplification methods and multiresolution models have been proposed to solve this problem, but most of them fail to perform either scalable simplification or efficient viewpoint dependent visualization of massive models.

Our contribution of this work is to find a solution for real-time and interactive browsing of massive models on personal computer platforms. In human visual system, the sensitivity to details is inversely proportional to the distance between the view point and the observed point. Thus we can construct a hierarchy of multiple resolution representations of the original model. At run-time, we dynamically and adaptively select the needed level-of-detail (LOD). We build LOD trees through hierarchical face clustering. Our algorithm reveals an out-of-core nature, since we use an octree data structure to partition the model and build the local LOD tree for each octree node. Then we combine all the top layer of the local LOD trees and take it to build the global LOD tree. When in real-time browsing, we mainly interact with the global

LOD tree and use it as an entry point to access the corresponding local LOD trees. Frustum culling and backface culling were used to accelerate the interaction speed. By combining a large set of technologies, our system shows good performance, better visual results, and a highly scalable architecture.

2 Related work

The research on interactive processing of complex models has over 30 years' history [1, 2]. The traditional approaches focus on how to reduce data complexity, manage data organization and utilize the new hardware technology [3]. In recent years, due to the widespread use of massive data sets, no single method could provide satisfying solution. A number of state-of-the-art systems utilizing different sets of technologies have been proposed to tackle this issue.

LOD based mesh visualization. LOD is useful because it is able to adjust the appropriate approximation given some viewing parameters [4]. For example, the Quick-VDR system [5, 6] represents the model as a clustered hierarchy of progressive meshes (CHPM) [7]. It uses the cluster hierarchy for coarse-grained selective refinement and progressive meshes for fine-grained local refinement. The Adaptive TetraPuzzles (ATP) system [8] uses a regular conformal hierarchy of tetrahedra to spatially partition the model. Each tetrahedral cell contains a precomputed simplified version of the original model, which is constructed off-line during a fine-to-coarse parallel out-of-core simplification of the surface contained in diamonds.

Real-time ray tracing. Some other systems diverge from the normal rasterization approach by incorporating a real-time ray tracing algorithm. By using spatial

* *Corresponding author* e-mail: yan.h@sdu.edu.cn

indexing, ray queries can be determined in logarithmic time. The OpenRT real-time ray tracer [9, 10] uses a two-level kd-tree hierarchy as spatial index. It also incorporates a custom memory management subsystem to deal with scenes larger than physical memory.

The volumetric approach. All the above systems assume that the multiresolution models are triangle-based or point-based. But the Far Voxels [11] system adopts a volumetric approach which uses small volume clusters to represent local datasets. By using a coarser granularity in the LOD structure, the cost of data management, traversal and occlusion culling can be reduced dramatically. Tian proposes Adaptive Voxels system [12] based on the Far Voxels system to make use of a novel adaptive sampling method to generate LOD models.

3 Mesh simplification

Mesh simplification is the cornerstone of LOD tree construction. The decimation methods can be classified into two major categories: clustering and incremental decimation. Vertex clustering and face clustering are two major clustering methods. The incremental decimation can have operators such as vertex removal, edge collapse, half-edge collapse etc. We choose face clustering to do mesh simplification for several reasons. First, it is much more efficient to use a clustering algorithm than an incremental decimation one. Second, face clustering provides better visual results than vertex clustering. Third, it is natural to pick face region as the unit not only in LOD tree construction but also in the real-time and interactive browsing of the model.

Suppose the initial mesh model contains N triangles. The overall framework of our algorithm is as follows:

```

Select K representative faces randomly
While(true)
{
  Region Growing
  If (termination criterion)
    Break;
  Update K representative faces
}
Face merging
Edge merging

```

FIGURE 1 Face clustering algorithm

We use a k-means based clustering to do the region growing process. When the K clusters are settled, we merge all the triangles inside one cluster into one super face. We call it super face because it is bounded by the boundary edges and is usually not flat. The effect of face clustering on an example mesh is demonstrated in Figure 2. The original mesh, the clustering result and the face merging result are shown in Figure 2. Looking at Figure 2c carefully, we find that two adjacent super faces normally shares more than one edge. This is a subtle aspect which can affect the overall performance of the whole system. Not significantly in this picture, but imagine if we have a model of millions of triangles and

clusters it into thousands of super faces. Then we'll see large super faces with lots of small edges jagged together.

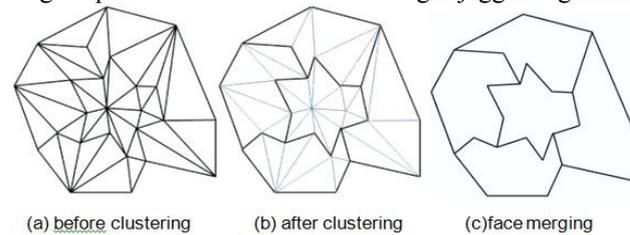


FIGURE 2 Face clustering example

The way to solve this is to do edge merging. We merge those edges that are shared by two adjacent super faces to ensure that only one edge exists between two adjacent upper faces. The process is described as follows: After merging all the triangles, we mark each vertex which is shared by three or more than three super faces as an anchor vertex. Since we need at least three vertices to determine a face region, we require every super face to have at least three anchor vertices. If a super face does not meet this requirement, we just randomly pick a certain number of non-anchor vertices to be anchor vertices. Finally, the boundary edges of super faces are determined by those anchor vertices. Sequentially connecting those anchor vertices will lead to "boundary-straightened" super faces (see Figure 3).

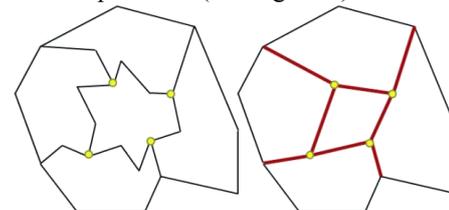


FIGURE 3 Edge merging

4 Multiresolution model

In this paper, we build a Level-of-Detail tree based interactive browsing system of the massive mesh model. We propose a novel approach to LOD tree construction

We use an octree structure to partition the original model to support out-of-core processing. Each local LOD tree corresponds to an octree partition region. We control the octree depth to make the memory usage of each local LOD tree construction under a predefined upper limit. We combine all the top layer of local LOD trees and take it as the bottom layer to construct the global LOD tree. We control the height of all the local LOD trees so that the construction of the global LOD tree can fit in memory. The overall structure of the LOD hierarchy is shown in Figure 4.

By organizing our data in this way, we sort of make a distinction between model's overall look and model's local region display. We can use the global LOD tree to support interactive display of the whole model. When the user is interested in some particular area, the corresponding local LOD trees can be loaded into memory to explicitly show the focused region. The global LOD tree serves as an entry point to find and load the

local LOD trees. Thus the global LOD tree plays a central role in the whole interaction period. During the loading and recycling of the local LOD trees, we can use scheduling policies based on user viewpoint to optimize the overall performance.

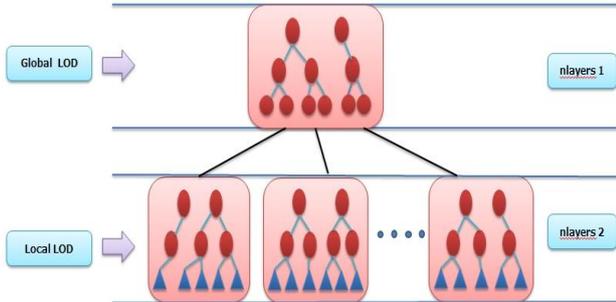


FIGURE 4 The structure of Local and Global LOD trees

5 Construction and serialization of LOD trees

5.1 OCTREE PARTITIONING

We use an octree data structure to partition the original model. We distribute the vertices and triangles of the model according to the following rules:

- 1) To vertex: we distribute it to its corresponding octree cell.
- 2) To triangle: since each triangle has three vertices, there are generally three cases:
 - If three vertices all fall into the same cell, then the corresponding cell is the one contains this triangle.
 - If two of them fall into the same cell, then the corresponding cell is the one contains this triangle.
 - If the three vertices belong to three different cells, we use the first vertex's containing cell to have the triangle.

Figure 5 shows an example for the 2D analogy of octree partition. The red number gives each vertex its appearing order in the triangle, and the blue number inside each triangle represents the number of the cell containing this triangle. During the partitioning process, we also compute each triangle's area, barycenter, normal and store them in disk files for later use.

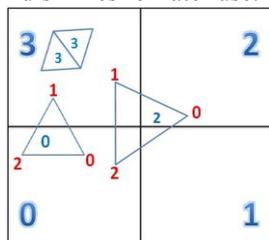


FIGURE 5 Example of 2D analogy for Octree partition

5.2 LOCAL LOD TREE CONSTRUCTION

Once we have done octree partitioning, we construct a local LOD tree for each octree node. Taking all the triangles of one octree node as input, we use K-means based face clustering algorithm to obtain a new simplified representation made up of K super faces to approximate

the original one. Then we take those super faces as input and use face clustering again to get an even more simplified model. By continually doing so, we will get a hierarchy of multiresolution models. Each super face is derived by merging the sub faces it contains. We include this inheritance relationship to build a Level-of-Detail tree structure.

5.3 GLOBAL LOD TREE CONSTRUCTION

For each local LOD tree, we only keep the top layer in memory. The top layer is the simplified representation of the original model in the corresponding spacial region. After constructing local LOD trees for all the octree nodes, we combine all the top layers of those local LOD trees to form a complete simplified representation for the whole model. And we use this layer as the bottom layer to construct the global LOD tree.

There exists some freedom in choosing the octree depth and the height of the local LOD tree. But each octree node must contain at most some maximum number of triangles to make sure that the memory is sufficient in local LOD tree construction. Also, the height of the local LOD tree must be high enough so that when all the local LOD trees' top layers were combined together, there remains sufficient memory to build the global LOD tree.

5.4 SERIALIZATION OF LOD TREES

We need to design a format to represent a LOD tree, such that it can be stored in disk, loaded to memory and interpreted efficiently for real-time viewing. This is achieved through serialization. It is the process of converting an object into a writable format that can be persisted or transported. The complement of serialization is deserialization, which restores an object from a stream. In real-time browsing, we need to load the local LOD trees frequently, so finding an efficient way to do serialization/deserialization is critical.

We use a DFS-based approach for its simplicity to storing the tree structure and it only requires one scan and a small extra stack to do deserialization.

6 View-dependent rendering

We will use the constructed LOD trees to support real-time and interactive browsing of the massive mesh model. The browsing system first loads the global LOD tree. Each level of the global LOD tree represents a certain degree of approximation to the whole model. Given the current viewpoint, we use the model's projected screen space area to choose the suitable level to render. Because the Global LOD tree is resident in memory, we can efficiently doing the traversal and rendering. When the user moves its viewpoint to some particular region of the model, the node of the Global LOD tree cannot provide sufficient accuracy. We have to load the corresponding local LOD tree which represents

the viewer's interested region and select certain level according to the projected screen space area.

Due to the limited size of field-of-view, when we are focused on some particular part of the model, the other parts of the model either are out of the sight or are far enough. In practice, we only need to load several Local LOD trees at the same time. So we can allocate a buffer to contain the currently loaded Local LOD trees. If the viewpoint moves and the buffer is full, we can remove old Local LOD trees and load new Local LOD trees. Because viewers always move their sight in a continuous way, we can optimally remove the Local LOD tree that is furthest to our current viewpoint.

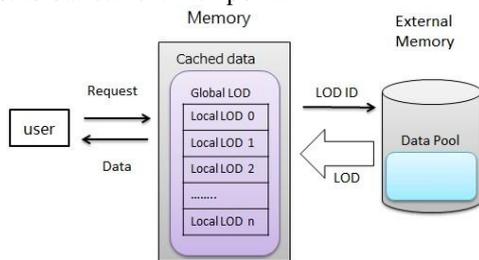


FIGURE 6 Data access framework

Figure 6 shows the whole system's data access framework. The global LOD tree is always in memory and acts as the entry point for accessing the local LOD trees. We have a buffer of size n to contain the currently loaded local LOD trees. A scheduler loads local LOD trees from the disk and removes local LOD trees from the memory if the buffer is full

6.1 VISIBILITY CULLING

Determining the visible parts of the scene is an important graphics problem. It is both inefficient and incorrect to render objects that are unseen. We have to remove those surfaces that are hidden from the viewer. Visibility culling [13] is the process of computing the visible subset of a scene. There are typically three culling techniques: view-frustum culling, back-face culling and occlusion culling. Occlusion culling is mainly used in scenes that contain many models. Because our focus here is on single massive mesh model, we only use view-frustum culling and back-face culling to accelerate our algorithm.

6.2 VIEW-FRUSTUM CULLING

Figure 7 shows a typical camera setup. Models or parts of models outside the frustum cannot be seen by the viewer. Because our LOD tree node represents a super face, we need to test whether this super face is outside the frustum.

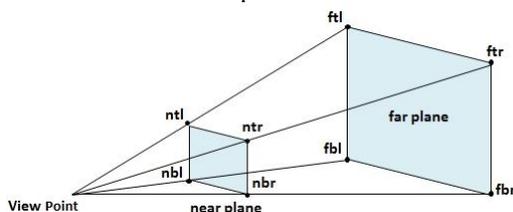


FIGURE 7 View-frustum

In practice, we use the bounding box approach. When we build the LOD tree hierarchically, we also compute the combined bounding box for each node from its children bounding box. This process is quite efficient and visibility test of box against frustum is also easy.

6.3 BACK-FACE CULLING

For a solid opaque object, the back of it is hidden from the viewing ray. Culling primitives that lie on the backs of objects can almost reduce half of the scene geometry to be rendered. Here, we use a clustered backface culling algorithm based on the normal cone [14]. The normal cone is represented by a central cone normal and a cone angle.

Like in Figure 8, we compute the normal cone for each tree node. To every current viewpoint, we also have a viewing normal cone. By comparing these two normal cones we can determine whether the node is back facing or not as show in Figure 9. To find an exact normal cone for a surface patch is a computational geometry problem and is rather slow. Instead, we use a bounding box approach which is fast and approximates well to the exact normal cone. The idea is that a bounding box of all the normal N_i is constructed. The cone normal is defined to be the vector from the origin to the centre of the bounding box. The direction from the origin to the eight corners of the bounding box will have eight angles with the cone normal. We take the largest one as the cone angle.

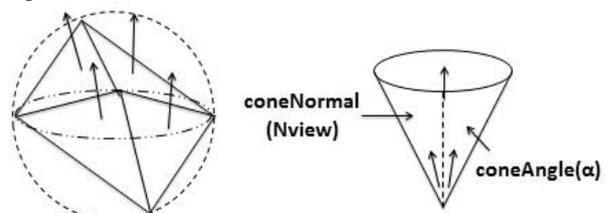


FIGURE 8 The left subfigure shows one node containing 4 triangles, the right subfigure shows the corresponding normal cone

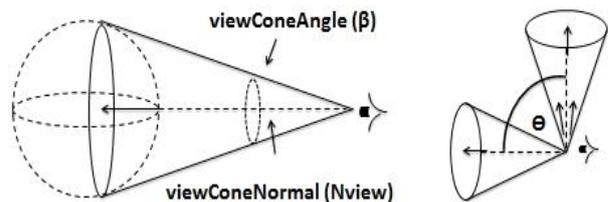


FIGURE 9 The left subfigure shows the viewing normal cone, the right subfigure shows its relation with one node's normal cone

7 Experimental results

7.1 PREPROCESSING

All the experiments were done on a Lenovo PC with 2.83GHz Intel Core 2 Quad CPU Q9500 processors, 4.0 GB of RAM. We have tested a number of massive models. The Figure 10 shows the four models we used for testing our system

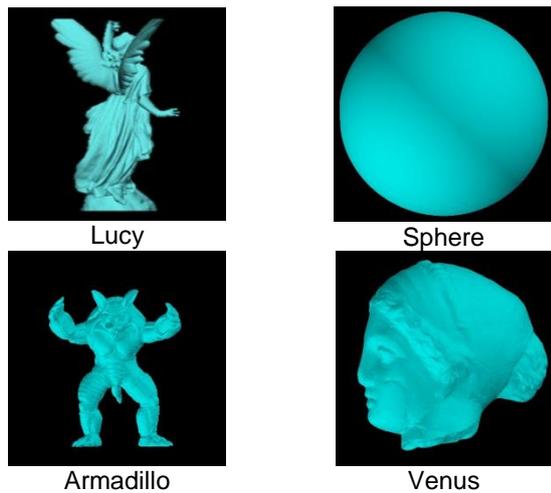


FIGURE 10 Four test models

From Table 1 and Table 2, we can see that our tested models contain tens of millions of and even hundreds of millions of triangles. The model venus has size of over 11GB. The octree partitioning process takes four and a half hours. The construction of LOD takes about six hours. The time used for pre-processing is acceptable since we only have to do it once. Once these LOD trees are built, we simply use these structures in later real-time browsing.

Our algorithm uses face clustering to do mesh simplification. Below, we show some simplified representations of several models in their global LOD tree layers in Figure 11.

From Figure 11, we can see that the simplification algorithm is effective even we simplify the model to only several hundreds of nodes. We compare our algorithm with the Adaptive Voxels system proposed by Tian [12]. Because our algorithm mainly deals with manifold surfaces, we test our system using different models. But we can still compare the two algorithms when the models' sizes were at the same size level.

TABLE 1 Numeric results of pre-processing

| Model | #Vertices | #Triangles | Size (GB) | Maximum Memory (MB) |
|-----------|-------------|-------------|-----------|---------------------|
| Lucy | 14,027,872 | 28,055,742 | 1.05 | 900 |
| Sphere | 31,457,282 | 62,914,560 | 2.55 | 180 |
| Armadillo | 44,280,834 | 88,561,664 | 3.48 | 700 |
| Venus | 135,430,146 | 270,860,288 | 11.49 | 400 |

TABLE 2 Numeric results of pre-processing

| Model | Octree | | LocalLOD | | GlobalLOD | |
|-----------|-----------|-------|----------|-----------|-----------|---------|
| | Time(min) | depth | layers | Time(min) | layers | Time(s) |
| Lucy | 27.82 | 2 | 3 | 43.63 | 4 | 0.35 |
| Sphere | 61.96 | 3 | 3 | 84.93 | 4 | 7.19 |
| Armadillo | 85.59 | 3 | 4 | 121.81 | 4 | 2.16 |
| Venus | 273.24 | 4 | 5 | 370.40 | 4 | 1.01 |

TABLE 3 Numerical comparisons

| Model | Faces (Million) | Pre-processing Time (min) | Size (GB) |
|----------------------------|-----------------|---------------------------|-----------|
| Adaptive Voxels Boeing 777 | 350 | 2,729 | 49.4 |
| Our Method Venus | 270 | 643 | 28.2 |

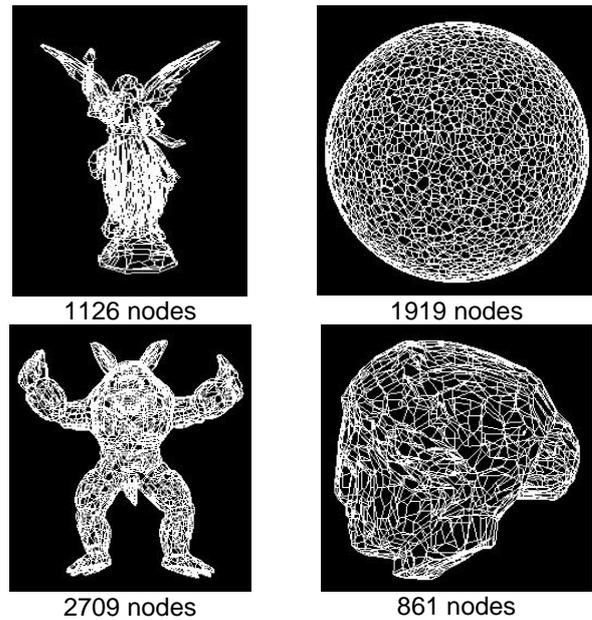


FIGURE 11 Some simplified representations of test models

Table 3 gives the numerical comparisons of our method and the Adaptive Voxels method. Our tested model venus is roughly 77% of the size of the Boeing 777 model. The pre-processing time used in our method is about 25% of the Adaptive Voxels method. The disk space usage is also much smaller in our method. Further, our algorithm is quite scalable and can be easily made parallel. The octree partitioning process and the local LOD tree construction can use parallel computing to largely accelerate the processing. But the Adaptive Voxels method uses BSP tree to construct the scene graph. The structure is intensely correlated, so it is not suitable for parallel computing.

7.2 REAL-TIME AND INTERACTIVE BROWSING

We devised a set of inspection paths to verify the real-time performance of our system. We include the typical tasks such as rotation, translation and scale. We also consider rapid changes from overall scenes to some

specific model regions. The window size is 1024×1024 , the pixel tolerance for each projected node size is 1 pixel. The numeric results are shown in Table 4. The column 4 gives the system setup time. And the last two columns give the average frame per second without and with culling techniques.

From Table 4 we can see that all the average numbers of FPS with culling are above 18. In reality, humans take actions normally three to five times per second. Our system satisfies the real-time interaction rates. The last

two columns of Table 4 shows that, by using hierarchical culling techniques, the system can save almost two fifth of the time. The average FPS of the Adaptive Voxels system is 8. Our system shows a significant improvement in the real-time performance. Another advantage is that our system is particularly efficient at viewing the local regions of the model since we use an optimal scheduling policy to load and recycle the local LOD trees thus minimizing the data access time.

TABLE 4 Numeric results for real-time rendering

| Model | Resolution | Pixel Error | Setup Time | Avg FPS (No Culling) | Avg FPS (With Culling) |
|-----------|------------|-------------|------------|----------------------|------------------------|
| Lucy | 1024×1024 | 1 | 1.57 | 16 | 24 |
| Sphere | 1024×1024 | 1 | 2.57 | 13 | 22 |
| Armadillo | 1024×1024 | 1 | 3.12 | 13 | 21 |
| Venus | 1024×1024 | 1 | 4.86 | 12 | 18 |

8 Conclusions and future work

In this paper, we build an LOD tree based real-time and interactive browsing system for massive mesh models. By using octree partition and face clustering, we construct the local LOD trees and the global LOD tree to provide an efficient out-of-core data access framework. We have tested on a set of massive mesh models and obtained good experimental results.

In the future, improvements could be made to our system. For instance, our current system has not exploited the parallelism inherent in components of the algorithm, such as octree partitioning and local LOD tree

construction. We can also use data compression techniques to further reduce the data to be stored in disk and to be loaded to memory.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant No. 61303083), NSFC Joint Fund with Guangdong under Key Project (Grant No. U1201258) and the Scientific Research Foundation for the Excellent Middle-Aged and Youth Scientists of Shandong Province of China (Grant No. BS2011DX017).

References

- [1] Kasik D J, Manocha D, Slusallek P 2007 *IEEE Computer Graphics and Applications* 27(6) 17-19
- [2] Gobbetti E, Kasik D, Yoon S 2008 Technical strategies for massive model visualization *Proceedings of the 2008 ACM symposium on Solid and physical modeling ACM* 405-15
- [3] Dietrich A, Gobbetti E, Yoon S-E 2007 *IEEE Computer Graphics and Applications* 27(6) 20-34
- [4] Luebke D P 2003 Level of detail for 3D graphics Morgan-Kaufmann
- [5] Yoon S E, Salomon B, Gayle R, Manocha D 2004 *IEEE Computer Society* 67(14) 131-8
- [6] Yoon S E, Salomon B, Gayle R, Manocha D 2005 *IEEE Transactions on Visualization and Computer Graphics* 11(4) 369-82
- [7] Hoppe H. 1996 Progressive meshes *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques ACM* 99-108
- [8] Cignoni P, Ganovelli F, Gobbetti E 2004 *ACM Transactions on Graphics (TOG)* 23(3) 796-803
- [9] Wald I, Purcell T J, Schmittler J, Benthin C, Slusallek P 2003 Realtime Ray Tracing and its use for Interactive Global Illumination *In Eurographics State of the Art Reports*
- [10] Wald I, Dietrich A, Slusallek P 2004 An Interactive Out-of-Core Rendering Framework for Visualizing Massively Complex Models *Rendering Techniques (Proceedings of the Eurographics Symposium on Rendering)* 81-92
- [11] Gobbetti E, Marton F 2005 *ACM Transactions on Graphics (TOG)* 24(3) 878-85
- [12] Fenglin T, Hua W, Dong Z, Bao H. 2010 Adaptive voxels: interactive rendering of massive 3D models *The Visual Computer* 26(6-8) 409-19
- [13] Akenine-Möller T, Haines E, Hoffman N. 2011 Real-time rendering [3rd], *CRC Press* 14 660-70
- [14] Shirmun L A, Abi- Ezzi S S 1993 The cone of normals technique for fast processing of curved patches *Computer Graphics Forum Blackwell Science Ltd* 12(3) 261-72

Authors



Xian Wu, born in 1988, Huainan, Anhui, China

University studies: master student, school of computer science and technology, Shandong University.
Scientific interest: computer graphics.



Huang Yan, born in 1974, Tongling, Anhui, China

Current position, grades: associate professor in the school of computer science and technology, Shandong University.
Scientific interest: mainly in large scale 3D data visualization, intelligent multimedia data analysis.

A kernel induced energy based active contour method for image segmentation

Xiaofeng Li^{1*}, Yanfang Yang¹, Limin Jia²

¹School of Traffic and Transportation, Beijing Jiaotong University, No.3, Shang Yuan Cun, Haidian District, Beijing, China

²State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, No.3, Shang Yuan Cun, Haidian District, Beijing, China

Received 4 March 2014, www.tsi.lv

Abstract

Active contour model is a promising method in image segmentation. However, existing active contour model and its evolution often suffer from slower convergence rates and easily to be trapped in local optima due to the presence of noise. In this paper, a novel curve evolution model based on kernel mapping method is presented. The method first transforms original image data into a kernel-induced space by a kernel function. In the kernel-induced space, the kernel-induced non-Euclidean distance between the observations and the regions parameters is integrated to formulate a new level set based active contour model. The method proposed in this paper leads to a flexible and effective alternative to complex model the image data. In the end of this paper, detailed experiments are given to show the effectiveness of the method in comparison with conventional active contour model methods.

Keywords: Kernel mapping; Active contour; Chan-Vese model; Level-set; Image segment

1 Introduction

Object boundary detection is a key of the fundamental process in image processing that facilitates higher level image analysis, description, recognition and visualization of objects of interest. Realizing the importance of object boundary detection in image processing, a great number of breakthroughs have been made in the past several decades. However, Object boundary detection remains actually problem-centric. Achieving a generic detection method that is universally applicable for broad range of problem domains is difficult [1].

Active contour models have been extensively studied and used in object boundary detection, once it was introduced by Kass et al [2]. However the technique requires the initial contour to lie outside the feature of interest and relies on an inherent contraction force to move the contour towards the feature. Cohen [3, 4] proposed an inflating contour that reduced the sensitivity to initialization. Geodesic active contour model [5]-[9], which is the original snake model in level-set frame, allows changes in topology. Other implementations have also been proposed for capturing more global minimizers by restricting the search space. Dual snakes [10] and dual-band active contour [11] restrict their search spaces exploiting normal lengths on the initial contour. Using of simulated annealing for minimization [12] and dynamic programming [13] has been integrated into the snake model.

All these classical snakes and active contour models rely on image gradient to stop the curve evolutions. Thus the performance of the purely edge-based models is often inadequate. Complex region-based energy functional are researched to likely to yield undesirable local minima when compared to simpler edge-energy functional.

Region-based models have many advantages over edge-based ones. First region-based models utilize image information not only near the evolving contour, but also statistical information inside and outside the contour, which are less sensitive to noise and have better performance for images with weak edges or without edges. Second, they are significantly less sensitive to the location of initial contour then can efficiently detect the exterior and interior boundaries simultaneously. One of the most popular region-based models is the Chan-Vese model [14], inspired by Mumford-Shah functional [15]. This model, as well as most of the region-based energy functional, is computationally onerous.

In order to overcome the limitations, especially the high computational cost, clustering of data is integrated into image segmentation. Vazquez et al [16] showed that image segmentation is spatially constrained clustering of image data. Gibou et al. [17] utilized the simplicity of the k-means algorithm, however, the main drawback is that they are more sensitive to noise.

This paper deals with the above mentioned problems. It presents a novel kernel induced-based active contour, which can handle objects whose boundaries are not necessarily defined by gradient, objects with very smooth

*Corresponding author e-mail: xfengli@bjtu.edu.cn

or even with discontinuous boundaries. Salah et al [18], [19] show that kernel mapping is quite effective in segmentation of various types of images. The kernel function maps implicitly the original image data into a kernel space in a higher dimension and, then, a simpler method in induced space will be possible [20]. The mapping, which transforms the original image data in a higher dimensional space, is implicitly. The transformed data in a higher dimensional space can be expressed via the kernel function because the dot product, the Euclidean norm thereof. The kernel induced method is a tool that has been intensively used in data clustering, but not in active contour methods. Generally, kernel induced methods provide more accurate and robust data clustering, thus, we combine it with active contour methodology, introducing here a model as a kernel induced space-based minimization. The kernel induced method of the energy provides a balanced technique with strong ability to reject “weak” local minima. We use a common kernel function, the radial basis function (RBF), in this paper, and then verify the effectiveness of the method by a quantitative and comparative performance evaluation to Chan-Vese model over a large number of experiments on synthetic images, as well as diverse real image

The remainder of the paper is organized as follows: The description of the model, its kernel induced motivation energy and its properties are presented in Section II. Experimental results are presented in Section III and conclusions are drawn in Section IV.

2 Description of model

2.1 PROPOSED MODEL

The basic idea of the model will be introduced in this section. The image data is generally non-linearly separable. The advantage of kernel mapping is that it maps implicitly the original image data into a kernel space in a higher dimension and, then, a simpler methods in induced space will be possible. In this section, we transform the image data implicitly via a kernel function firstly, and, then, use a kernel-induced non Euclidian distances to form the minimizing functional. To explain the role of the kernel mapping in the segmentation functional proposed in this paper, and describe clearly the ensuing algorithm, we first assume that the original image is formed by two regions of approximately piecewise-constant intensities.

Let $\varphi(\cdot)$ be a non-linear mapping from the observation space \mathcal{I} to a higher (possibly infinite) dimensional feature space \mathcal{F} . Let us defined the evolving the curve C in the image domain Ω . As assumed above, the image I is formed by two regions of approximately piecewise-constant intensities, of distinct values I_1 and I_2 . Assume that the object to be detected is represented by the region with the value I_1 , and its boundary by C_0 . So we have $I \approx I_1$ inside the object

(inside C_0) and $I \approx I_2$ outside the object (outside C_0). Now let us consider the following functional:

$$F_1(C) + F_2(C) = \int_{\text{inside}(C)} \|\varphi(I(x, y)) - \varphi(c_1)\|^2 dx dy + \int_{\text{outside}(C)} \|\varphi(I(x, y)) - \varphi(c_2)\|^2 dx dy \quad (1)$$

where c_1 and c_2 are regions parameters depending on C . The energy terms $F_i(C)$ measure a kernel-induced non Euclidian distances between the observations and the regions parameters $c_i, i = 1, 2$.

Following the Mercer’s theorem conditions [20], the explicit mapping φ has not been known. Instead, the transformed data can be expressed via a continuous, symmetric, positive semi-definite kernel function $K(x, y)$:

$$K(x, y) = \varphi(x)^T \cdot \varphi(y), \forall (x, y) \in \mathcal{I}^2, \quad (2)$$

where “ \cdot ” is the dot product in the feature space \mathcal{F} . Substitution of the kernel functions gives, for $c_i \in \{c_1, c_2\}$:

$$J_k(I(x, y), c_i) = \|\varphi(I(x, y)) - \varphi(c_i)\|^2 = (\varphi(I(x, y)) - \varphi(c_i))^T \cdot (\varphi(I(x, y)) - \varphi(c_i)) \quad (3) = K(I(x, y), I(x, y)) + K(c_i, c_i) - 2K(I(x, y), c_i)$$

where c_1 and c_2 are constants depending on C , expressing the average prototypes of the image regions inside and outside respectively of C . In this simple case, it is obvious that the boundary of the object C_0 , is the minimizer of the fitting term:

$$\inf_C \{F_1(C) + F_2(C)\} \approx 0 \approx F_1(C_0) + F_2(C_0). \quad (4)$$

The five cases are illustrated in Figure 1. If the curve C is outside the object, then $F_1(C) > 0$ and $F_2(C) \approx 0$ or $F_1(C) \approx 0$ and $F_2(C) > 0$ depending on object position (inside or outside the curve). If the curve C is inside the object, then $F_1(C) \approx 0$ and the $F_2(C) > 0$. If C is both inside and outside the object, then $F_1(C) > 0$ and the $F_2(C) > 0$. Finally, the fitting term is minimized when $C = C_0$, i.e., when the curve C is on the boundary of the object. That is, the fitting term is minimized when the curve C is converged to the object boundary C_0 .



FIGURE 1 All possible case in the position of the curve C in relation to the object under consideration

The proposed active contour is based on the minimization of the above fitting term, taking into account the length term of the model C as a regularization term. Therefore, the energy functional $F(C, c_1, c_2)$ is introduced as:

$$F(C, c_1, c_2) = \mu \cdot \text{Length}(C) + \lambda_1 \int_{\text{inside}(C)} \|\varphi(I(x, y)) - \varphi(c_1)\|^2 dx dy + \lambda_2 \int_{\text{outside}(C)} \|\varphi(I(x, y)) - \varphi(c_2)\|^2 dx dy \quad (5)$$

In equation (5), $\mu \geq 0$, $\lambda_1, \lambda_2 > 0$ are fixed parameters. The curve C_0 that minimizes F

$$F(C_0, c_1, c_2) = \inf_C F(C, c_1, c_2) \quad (6)$$

Some common kernel functions are listed in Table I. In this paper, the radial basis function (RBF) kernel, which has been prevalent in pattern data clustering [18, 22, 23] is adopted.

TABLE 1 Examples of prevalent kernel function

| | |
|--------------------------|---|
| RBF Kernel | $K(x, y) = \exp(-\ x - y\ ^2 / \sigma^2)$ |
| Sigmoid Kernel | $K(x, y) = \tanh(c(x \cdot y) + \theta)$ |
| Polynomial Kernel | $K(x, y) = (x \cdot y + c)^d$ |

The radial basis function (RBF) kernel is given as:

$$K(x, y) = \exp(-\|x - y\|^2 / \sigma^2) \quad (7)$$

With an RBF function, the equation (3), $J_k(I, c_i)$ can be simplified to $2(1 - K(I(x, y), c_i))$, $i = 1, 2$. The necessary conditions for a minimum of F with respect to region parameters are:

$$c_i - gR_i(c_i) = 0, i \in \{1, 2\}, \quad (8)$$

where

$$gR_i(c_i) = \frac{\int_{R_i} I(x, y)K(I(x, y), c_i) dx dy}{\int_{R_i} K(I(x, y), c_i) dx dy}, i \in \{1, 2\} \quad (9)$$

2.2 LEVEL-SET FORMULATION OF THE PROPOSED MODEL

In the level set formula, $C \subset \Omega$ is represented by the zero level set of a Lipschitz function $\phi : \Omega \rightarrow \mathbb{R}$, such that

$$\begin{cases} C = \{(x, y) \in \Omega : \phi(x, y) = 0\} \\ \text{inside}(C) = \{(x, y) \in \Omega : \phi(x, y) > 0\} \\ \text{outside}(C) = \{(x, y) \in \Omega : \phi(x, y) < 0\} \end{cases} \quad (10)$$

Using the Heaviside function H_ϵ , and the Dirac measure δ_ϵ as the regularized versions defined, respectively, by

$$\begin{cases} H_\epsilon(z) = \frac{1}{2} \left(1 + \frac{2}{\pi} \arctan\left(\frac{z}{\epsilon}\right) \right), \\ \delta_\epsilon(z) = \frac{1}{\pi} \cdot \frac{\epsilon}{\epsilon^2 + z^2}, z \in \mathbb{R} \end{cases} \quad (11)$$

We express the terms in the energy F in the following way:

$$\begin{aligned} \text{Length}\{\phi = 0\} &= \int_{\Omega} |\nabla H(\phi(x, y))| dx dy \\ &= \int_{\Omega} \delta_\epsilon(\phi(x, y)) |\nabla \phi(x, y)| dx dy \end{aligned} \quad (12)$$

$$\begin{aligned} &\int_{\text{inside}(C)} \|\varphi(I(x, y)) - \varphi(c_1)\|^2 dx dy \\ &= \int_{\Omega} \|\varphi(I(x, y)) - \varphi(c_1)\|^2 H_\epsilon(\phi) dx dy \end{aligned} \quad (13)$$

In equation (13)

$$\begin{aligned} &\int_{\text{outside}(C)} \|\varphi(I(x, y)) - \varphi(c_2)\|^2 dx dy \\ &= \int_{\Omega} \|\varphi(I(x, y)) - \varphi(c_2)\|^2 (1 - H_\epsilon(\phi)) dx dy \end{aligned} \quad (14)$$

Keeping the ϕ fixed and minimizing the energy $F(\phi, c_1, c_2)$ with respect to the constant c_1 and c_2 , it is easy to express these constants function of ϕ by

$$c_1(\phi) = \frac{\int_{\Omega} I(x, y)K(I(x, y), c_1)H(\phi(x, y)) dx dy}{\int_{\Omega} K(I(x, y), c_1)H(\phi(x, y)) dx dy}, \quad (15)$$

$$c_2(\phi) = \frac{\int_{\Omega} I(x, y)K(I(x, y), c_2)(1 - H(\phi(x, y))) dx dy}{\int_{\Omega} K(I(x, y), c_2)(1 - H(\phi(x, y))) dx dy} \quad (16)$$

The region term (13), (14) and the contour smoothness term in (12) are integrated into the energy function given by (5), producing

$$\begin{aligned} F_\epsilon(\phi, c_1, c_2) &= \mu \int_{\Omega} \delta_\epsilon(\phi) |\nabla \phi(x, y)| dx dy \\ &+ \lambda_1 \int_{\Omega} \|\varphi(I(x, y)) - \varphi(c_1)\|^2 H_\epsilon(\phi) dx dy \\ &+ \lambda_2 \int_{\Omega} \|\varphi(I(x, y)) - \varphi(c_2)\|^2 (1 - H_\epsilon(\phi)) dx dy \end{aligned} \quad (17)$$

Keeping c_1 and c_2 fixed and minimizing F_ϵ with respect to ϕ , we deduce the associated Euler-Lagrange equation for ϕ . Parameterizing the descent direction by an artificial time $t \geq 0$, the equation in $\phi(t, x, y)$ ($\phi(0, x, y) = \phi_0(x, y)$) defining the initial contour is

$$\frac{\partial \phi}{\partial t} = [\mu \operatorname{div}(\frac{\nabla \phi}{|\nabla \phi|}) - \lambda_1 J_K(I(x, y), c_1) + \lambda_2 J_K(I(x, y), c_2)] \delta_\epsilon(\phi) \quad (18)$$

3 Experimental results

In this section, we show the performance of the proposed method by presenting numerical results using the kernel-induced model on various synthetic and real images, with different types of contours and shapes. We show the active contour evolving in the original image Ω , and the associated piecewise-constant approximation of Ω (given by constant c_1 and c_2). In our numerical experiments, we generally choose the parameters to be $\lambda_1 = \lambda_2 = 1$. Only the length parameter μ is not same in

our parameters. As shown in [24], μ plays a scaling role in the minimizing functions. μ should be small, if we have to detect all or as many objects as possible and of any size, and be larger when we have to detect large objects.

In the following, we illustrate the flexibility of the proposed method by a representative sample of the tests with various classes of real images and synthetic image with different noise models, and compare the computational time of different models.

Medical image segmentation is challenging and of a rapidly growing interest in recently years. Figure 2 illustrates that the proposed model can detect different objects of different intensities. Figure 2(a) depicts very narrow human vessels with very small contrast within some spots. The curves obtained at convergence are displayed, in Figure 3(f). Segmentation regions, represented by their parameters at convergence, are shown in Figure 2(f) and Figure 3(f). As shown in the two images, the length parameter in Figure 3 is larger than Figure 2, because the object of interest we have to detect, in Figure 3(a), is larger.

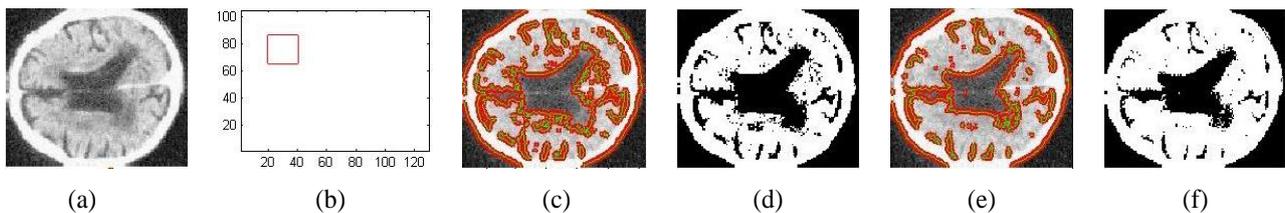


FIGURE 2 (a):Brain images; (b): initialization; (c): final position of the curves with Chan-Vese model; (d): final segmentation with Chan-Vese model; (e): final position of the curves with proposed model; (f): final segmentation with proposed model. Image size: 100×120 , $\mu = 0.1$

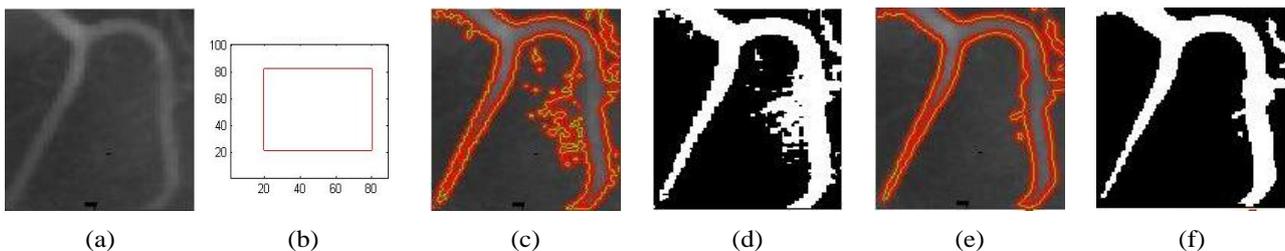


FIGURE 3 (a): Vessel images; (b): initialization; (c): final position of the curves with Chan-Vese model; (d): final segmentation with Chan-Vese model; (e): final position of the curves with proposed model; (f): final segmentation with proposed model. Image size: 100×80 , $\mu = 0.2$

The ability of the proposed model to deal with different noise models allows segmenting regions which require different models. To illustrate this important advantage of the proposed model, we consider a synthetic image with different noise models, as shown in Figure 4 (a), (b). Figure 4 (a) is generated with a salt and pepper noise, the corresponding noise density d is 0.1, the Gaussian noise is added in Figure 4 (b), the corresponding mean u is 0, the variance σ is 0.05. The final segmentation results with Chan-Vese model is in Figure 4 (a1-b2), and final segmentation results with proposed model, in Figure 4 (b1-b2). As shown in

Figures 4 (a1-b2), the Chan-Vese model gives incorrect results as expected. The results demonstrate the ability of our kernel-induced method to insensitive to different noise models.

The computational time for the proposed model and the Chan-Vese model are shown in Table II. The proposed kernel-induced model achieves noticeably lower the Chan-Vese model in all of the medical images. In noisy images, the proposed model costs less time obviously compared to the Chan-Vese model, which demonstrates the ability of proposed model in handling noise.

Figure 5 and Figure 6 demonstrate how the proposed algorithm detects object boundaries on real images, and illustrate the robustness with respect to initial conditions, initial curves were either big rectangle placed arbitrarily about the middle of the image or tiny circles spread all over the image. Segmentation of a natural plane image into two regions is in Figure 5. Original, initial contour and final position of the evolving curve are displayed respectively in Figure 5 (a-c). The final segmentation regions, represented by their parameters at convergence, are illustrated in Figure 5(d). Figure 6 depicts how the proposed method, with the initial contour of tiny circles spread all over the image, detects the boundaries in real image.

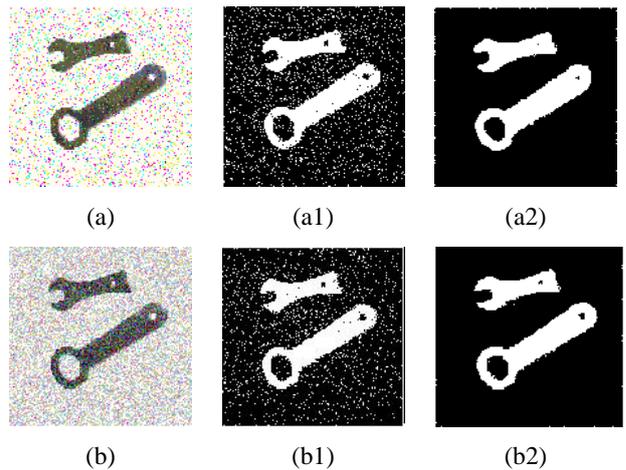


FIGURE 4 Image with different noise models: (a-b) image with salt & pepper , noise density $d = 0.1$ (first row) and Gaussian (second row) noises, mean $\mu = 0$ and variance $\sigma = 0.05$; (a1-b1) segmentation results with Chan-Vese model; (a2-b2) segmentation results with proposed model. Image size: 200×200 , $\mu = 0.12$

TABLE 2 Comparison of computation time (Seconds)

| Model | Without noise | | With noise | |
|-----------------|---------------|--------|---------------------|----------------|
| | Brain | Vessel | Salt & Pepper noise | Gaussian noise |
| Proposed model | 15.43 | 10.98 | 90.45 | 61.73 |
| Chan-Vese model | 19.97 | 22.77 | 271.2 | 522.40 |

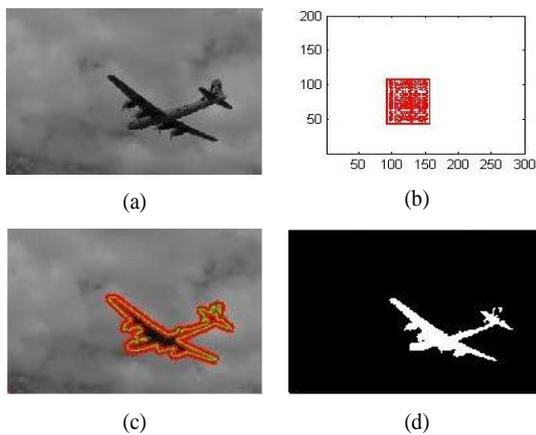


FIGURE 5 (a) Real plane image; (b) initialization; (c) final position of the curve; (d) final segmentation. Image size: 200×300 , $\mu = 0.1$

4 Conclusion

In this paper, a novel fast and robust model for active contours to detect objects in an image was introduced. The model can detect objects whose boundaries are not necessarily defined by gradient, due to the fact that it is based on an energy minimization algorithm, and not on an edge-function as the most classical active contour models. This energy is based on kernel mapping, which can be seen as a particular case of a minimal partition problem, and is used as the model motivation power

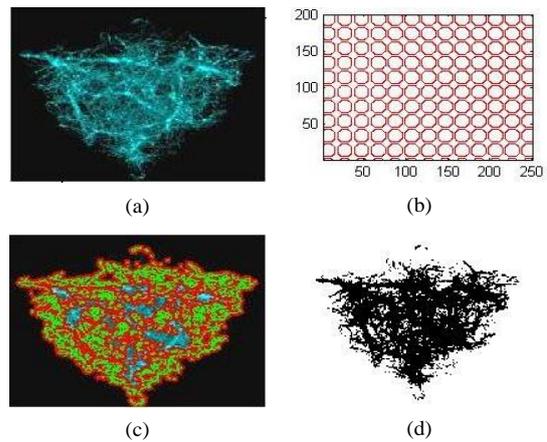


FIGURE 6 (a) Real image; (b) initialization; (c) final position of the curve; (d) final segmentation. Image size: 200×250 , $\mu = 0.02$

evolving the active contour until to catch the desired object boundary. We presented several experiments on synthetic with different noise model and real data which showed the effectiveness and the flexibility of the method.

Acknowledgments

This paper is supported by the National High-tech Research and Development Program of China (2011AA110503) and the Fundamental Research Funds for the Central Universities (T11JB00530).

References

- [1] Eason G, Noble B, Sneddon I N 1995 *Phil. Trans. Roy. Soc.* **A247** 529-51
- [2] Kass M, Witkin A, Terzopoulos D 1988 *Int. J. Comput. Vis.* **1**(4) 321-31
- [3] Cohen L D, Cohen I 1993 *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(11) 1131-47
- [4] Cohen L D 1991 *CVGIP: Image Understanding* **53**(2) 211-8
- [5] Caselles V, Catta F, Coll T, Dibos F 1993 *Numer. Math.* **66**(1) 1-31
- [6] Malladi R, Sethian J A, Vernuri B C 1995 *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(1) 158-75
- [7] Osher S, Sethian J A 1988 *J. Comput. Phys.* **79** 12-49
- [8] Caselles V, Kimmel R, Sapiro G 1997 *Int. J. Comput. Vis.* **22**(1) 61-79
- [9] Yezzi A, Kichenassamy S, Kumar A, Olver P, Tannenbaum A 1997 *IEEE Trans. Med. Imag.* **16**(2) 199-209
- [10] Gunn S, Nixon M 1994 *Proc. Brit. Machine Vision Conf.* 305-14
- [11] Dawood M, Jiang X, Schafers K 2004 *Lecture Notes Comput. Sci.* **3212** 544-51
- [12] Sorvik G 1994 *IEEE Trans. Pattern Anal. Mach. Intell.* **16**(10) 976-86
- [13] Geiger D, Gupta A, Costa L A, Vlontzos J 1995 *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(3) 294-302
- [14] Chan T, Vese L 2001 *IEEE Trans Image Process* **10**(2) 266-77
- [15] Mumford D, Shah J 1989 *Comm. Pure Appl. Math* **42** 577-685
- [16] Vazquez C, Mitiche A, Ayed I B 2004 *IEEE Int. Conf. Image Processing* 3467-4370
- [17] Gibou F, Fedkiw R 2005 *Proc. 4th Annu. Hawaii Int. Conf. Statistics and Mathematics* 281-291
- [18] Salah M B Mitiche A, Ayed I B 2010 *IEEE Trans. Image Process* **19**(1) 220-32
- [19] Salah M B, Mitiche A, Ayed I B 2009 *Image Process. Int. Conf.* 2997-3000
- [20] Cover T M 1965 *Electron. Comput* **EC-14** 326-34
- [21] Muller K R, Mika S, Ratsch G, Tsuda K, Scholkopf B 2001 *IEEE Trans. on Neur. Net.* **12**(2) 181-202
- [22] Dhillon I S, Guan Y, Kulis B 2007 *IEEE Trans. Pattern Anal Mach. Intell.* **29**(11) 1944-57
- [23] Wu K L, Yang M S 2002 *Pattern Recognit* **35** 2267-78
- [24] Krinidis S, Chatzis V 2009 *IEEE Trans. Image. Process* **18**(12) 2747-55

Authors



Xiaofeng Li, born on February 26, 1977, Suixian Henan, China

Current position, grades: Doctor of Control Science and Engineering, lecturer and master supervisor in Beijing Jiaotong University

University studies: Control Science and Engineering in Harbin Institute of Technology (HIT)

Scientific interest: Image processing and analysis, Computer vision and Complex electromechanical system fault diagnosis

Publications: 3 Patents, 10 Papers

Experience: He received Ph.D degrees on Control Science and Engineering from Harbin Institute of Technology (HIT) of China, 2008. Now he is a Lecture and master supervisor at Beijing Jiaotong University, Beijing, China. He mainly focuses on digital image processing and analysis, computer vision and complex electromechanical system fault diagnosis.



Yanfang Yang, born on August 9, 1985, Liuzhou Guangxi, China

Current position, grades: PhD Candidate

University studies: Safety Science and Engineering in Beijing Jiaotong University

Scientific interest: Image processing and analysis

Publications: 6 Papers

Experience: She is a safety science and engineering major PhD Candidate. She mainly focuses on digital image processing and analysis.



Limin Jia, born on January 18, 1963, Altai Xinjiang, China

Current position, grades: Doctor of Traffic Information Engineering & Control, professor and doctoral supervisor in Beijing Jiaotong University

University studies: Traffic Information Engineering & Control in Graduate School of China Academy of Railway Sciences

Scientific interest: Railway traffic Control and safety, intelligence transport system

Publications: 10 Patents, 100 Papers

Experience: He received Ph.D degrees on Traffic Information Engineering & Control from China Academy of Railway Sciences, 1991. Now he is a professor and doctoral supervisor in Beijing Jiaotong University. He mainly focuses on Railway traffic Control and safety, intelligence transport system

Image fusion based on MPCNN and DWT in PCB failure detection

Yiming Yuan*, Ming Jiang, Wengen Gao

College of Electrical Engineering, Anhui Polytechnic University, Anhui, China

Received 6 June 2014, www.tsi.lv

Abstract

The traditional contact-type printed circuit board (PCB) test methods have been unable to meet the needs of the fault detection and maintenance of a variety of increasingly complex electronic equipment. The visible and infrared respectively reflects the background information and the radiation information of PCB, so we can fuse the visible image and infrared image of the board together, and use the new fusion image to locate and identify the abnormal high temperature components or areas of the circuit board. A novel fusion algorithm of multi-sensor image is proposed based on Discrete Wavelet transform (DWT) and pulse coupled neural networks (PCNN) in this paper. Firstly, the IR and visible images are decomposed by DWT, then a fusion rule in the DWT is given based on the PCNN. This algorithm uses the local entropy of wavelet coefficient in each frequency domain as the linking strength, then its value can be chosen adaptively. After processing PCNN with the adaptive linking strength, new fire mapping images are obtained. According to the fire mapping images, the firing time gradient maps are calculated and the fusion coefficients are decided by the compare-selection operator with firing time gradient maps. Finally, the fusion images are reconstructed by wavelet inverse transform. The proposed algorithm of image fusion using modified pulse coupled neural networks (MPCNN) and DWT results in better quality of fused image with Entropy, Average grads, Cross-Entropy as compared to conventional image fusion Algorithms.

Keywords: PCNN image-fusion, DWT, PCB, failure detection

1 Introduction

Image fusion is an active research field as an aspect of data and information fusion, which is widely applied in remote sensing, computer vision, medical image processing .etc. It combines sensory data from multiple sensors to provide more reliable and accurate information [1] Image fusion is introduced into PCB failure detection, more abundant and comprehensive complementary information could be obtained through infrared and visible image, which could improve the accuracy and validity of PCB failure detection. Image fusion is the process of combining information from two or more images of a scene into a single composite image, which is more informative and suitable for human visual perception or computer processing [2, 3]. Most image fusion algorithms are based on multi-resolution analysis including Wavelet Transform, PCNN, etc. Conventional PCNN image fusion algorithms have been successful used in image fusion and could retain more details. However, in these algorithms, the value of single pixel is used to motivate on neuron. In fact, humans are sensitive to edges, directional features, etc. So, a pure use of single pixels is not enough. For the second question, the linking strength of PCNN neurons based on experiment or experience is great and has caused great inconvenience to the application of image fusion. Due to joint information representation at the spatial spectral domain, wavelet transform becomes the most popular multi-scale decomposition domain method in image fusion. This paper

proposed a new method for image fusion based on DWT and PCNN. Experimental results demonstrate that the proposed algorithm outperforms typical DWT-based, and conventional PCNN-based in terms of objective criteria [4, 5].

2 Conventional PCNN model

PCNN is a novel biologically neural network, which was developed by Eckhorn based on the experimental observations of synchronous pulse bursts in cat and monkey visual cortex. The basic model of PCNN neuron is shown in Figure 1, which comprises three parts: receptive field, modulation field and pulse generator. The neuron receives input signals from other neurons and external sources from two channels viz. F channel and L channel in the receptive field. F channel is the feeding input F_{ij} , which receives the input from external source and output of other neurons. L channel is the linking input L_{ij} which receives the input from other neurons output. The feeding field is modulated by linking field to calculate the internal activity U_{ij} . θ is the dynamic threshold. Matrixes M/V_F , W/V_L and V_θ are the linking weight/magnify coefficient of the feeding back field, the linking weight/magnify coefficient of the linking field, the threshold magnify coefficient, respectively; α_F and α_θ are the decayed constants associated with F , L , u ; b , n and Y_{ij} are the linking strength, the iteration number and the

* Corresponding author e-mail: yymblesky@163.com

pulse output, respectively. The discrete mathematical equations of each neural can be described as follows:

$$F_{ij}[n] = \exp(-\alpha_F) F_{ij}[n-1] + V_F \sum m_{ijkl} Y_{kl}[n-1] + I_{ij}, \quad (1)$$

$$L_{ij}[n] = \exp(-\alpha_L) L_{ij}[n-1] + V_L \sum W_{ijkl} Y_{kl}[n-1], \quad (2)$$

$$U_{ij}[n] = F_{ij}[n](1 + \beta L_{ij}[n]), \quad (3)$$

$$\theta_{ij}[n] = \exp(-\alpha_\theta) \theta_{ij}[n-1] + V_\theta Y_{ij}[n-1], \quad (4)$$

$$Y_{ij} = \begin{cases} 1, & U_{ij}[n] \geq \theta_{ij}[n] \\ 0, & U_{ij}[n] < \theta_{ij}[n] \end{cases}, \quad (5)$$

where (1), (2) and (3) are the mathematical models of the receptive feeding input, receptive linking input and modulating coupler, respectively; (4) and (5) are the step function of the pulse generator and the expression of variable threshold function respectively [6, 7]

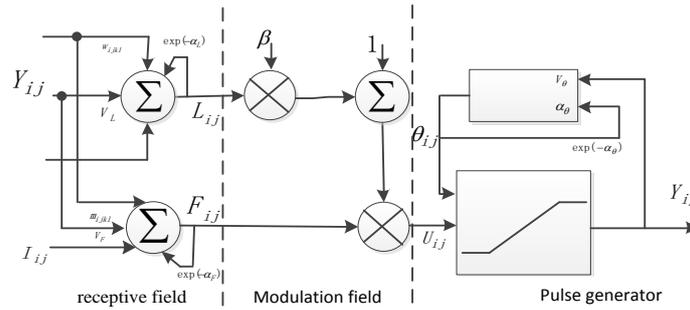


FIGURE 1 The Basic Model of PCNN Neuron

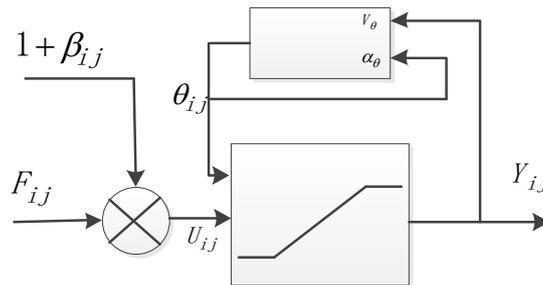


FIGURE 2 The MPCNN Model

3 Modified PCNN

The conventional PCNN model has the limitation of slow processing because of large number of iterations and computational complexity, which makes it unsuitable for image processing applications where large amount of data are to be handled. In practical application, to meet the demand of the situation should be simplified, easy for hardware implementation in saving cost and overhead at the same time, therefore according to different purposes, many scholars put forward different degrees of the simplified PCNN model, this paper uses the modified model according to the actual needs of image fusion [8-11]. The MPCNN model is shown in Figure 2. The expressions of MPCNN are listed as follows:

$$F_{ij}[n] = I_{ij}, \quad (6)$$

$$L_{ij}[n] = \sum_{kl} w_{ijkl} Y_{kl}[n-1], \quad (7)$$

$$U_{ij}[n] = F_{ij}[n](1 + \beta L_{ij}[n]), \quad (8)$$

$$U_{ij}[n] = F_{ij}[n](1 + \beta L_{ij}[n]), \quad (9)$$

$$Y_{ij} = \begin{cases} 1, & U_{ij}[n] \geq \theta_{ij}[n] \\ 0, & U_{ij}[n] < \theta_{ij}[n] \end{cases}. \quad (10)$$

The number of parameters to be determined will decrease if the PCNN model is simplified, but for a special image the key parameters are still needed to be selected by experiments or empirically.

4 Image fusion based on DWT and MPCNN

Image fusion based on Discrete Wavelet transform (DWT) is to decompose the original images into a series of frequency channels and combine the different features and details at multiple decomposition levels and in multi-frequency bands, which is suitable for multi-scale properties of the human vision system. In this paper, DWT and MPCNN are combined effectively to display their own advantages. Suppose that there are two original images with the same size denoted I and A, which are both accurately registered and F is the final fusion image. The new fusion scheme is shown in Figure 3.

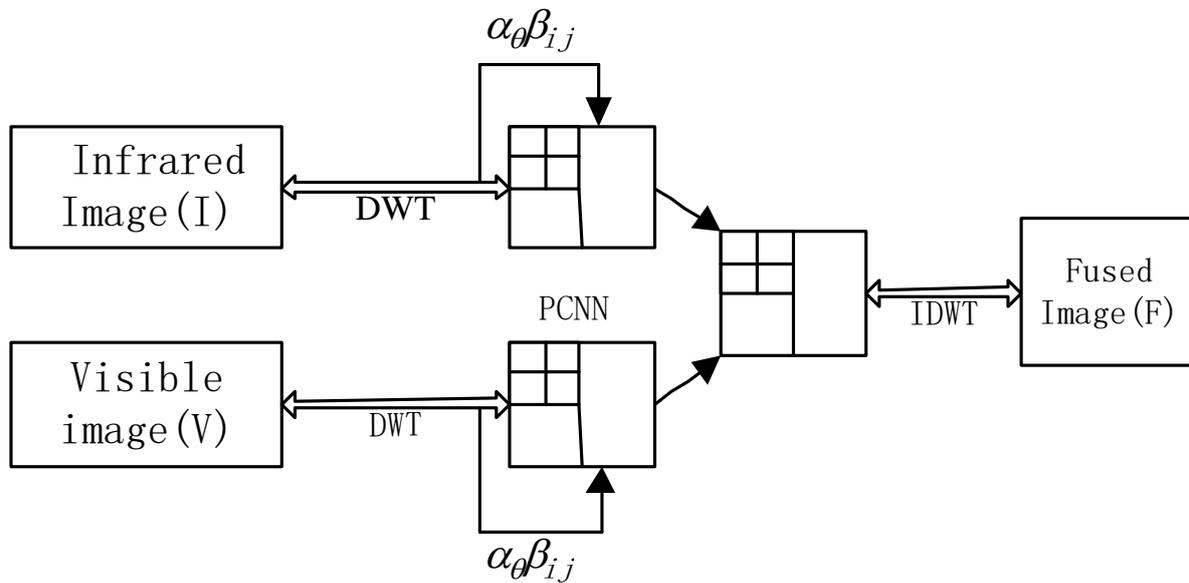


FIGURE 3 The new fusion scheme

The concrete steps of the new fusion algorithm can be described below [8-10].

Step 1: Get Wavelet Pyramid by DWT to infrared and visible images which are matched. Let $C(i,j)$ denote the wavelet coefficients of wavelet domain (i,j) . Let β_{ij} denote the local entropy of window. The clearer the pixel is, the larger the linking strength β is, and accordingly, the greater the linking extent of the corresponding neuron is. The parameter β_{ij} can be expressed with the following formula:

$$\beta_{i,j} = -\sum_{i=1}^M \sum_{j=1}^N P_{i,j} \log P_{i,j}, \tag{11}$$

$$P_{i,j} = \frac{c(i,j)}{\sum_{i=1}^M \sum_{j=1}^N C_{i,j} \log C_{i,j}}, \tag{12}$$

where β_{ij} is the linking strength of the neuron ij

Step 2: Let β_{ij} denote the linking strength of the PCNN, the wavelet coefficients are mapped to the corresponding gray range; the output of threshold function decay with time to the minimum gray when all the pixels in the image are the ignition.

Step 3: Setting the initial values of the MPCNN's parameters;

Step 4: For each iteration the following steps were done to MPCNN. Let V_{min} and V_{max} denote the minimum and V_{min} fire threshold, let t denote iteration time. The parameter α_θ can be expressed with the following formula

$$\alpha_\theta = -\frac{\ln\left(\frac{V_{min}}{V_{max}}\right)}{t}. \tag{13}$$

The input neurons of MPCNN were computed according to Equations (6), (7), (8), (9), (10) calculate the linking strength β_{ij} according to (11) and (12).

Step 5: Reconstruct the original image by using an inverse DWT, thus obtaining the fused image F [8-18].

5 Experimental results and conclusion

The performance evaluation criteria of image fusion are still a hot topic in the research of image fusion. Besides visual observation, objective performance evaluation criteria are used in this paper, such as Entropy, Average grads, Cross-Entropy etc.

In this section, the example, which is conducted by MATLAB 2012a on a PC with Intel P8700, is given to prove the validity of the proposed fusion technique. The related source images are PCB infrared and visible images whose size is 512×512 . To illustrate the proposed fusion method, several experimental results are presented in this section. Parameters of PCNN is set as $t = 20$ PCNN

$$\text{iteration time: } w = \begin{bmatrix} 0.707 & 1 & 0.707 \\ 1 & 0 & 1 \\ 0.707 & 1 & 0.707 \end{bmatrix} \text{ linking synapse.}$$

To evaluate the performance of the proposed fusion algorithm, it is compared with DWT-based fusion algorithm, and PCNN-based. The fusion results are shown in Figure 4 and the objective performance evaluation criteria are shown in Table 1.

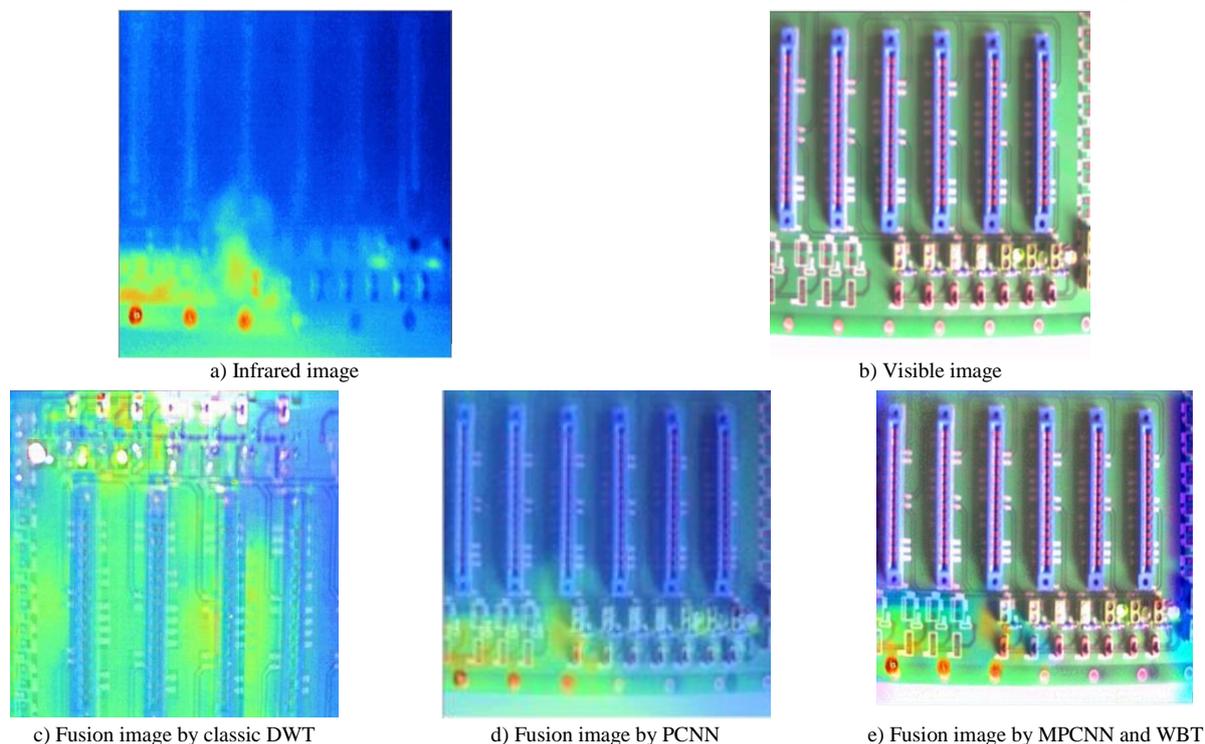


FIGURE 4 Fusion results using different algorithms

TABLE 1 Evaluation of statistical parameters

| Method | Entropy | Average grads | Space frequencies | Standard deviation | Cross-Entropy |
|-------------|---------|---------------|-------------------|--------------------|---------------|
| classic WBT | 5.8199 | 3.9318 | 13.7498 | 66.5730 | 0.4709 |
| PCNN | 7.4947 | 5.2122 | 12.1433 | 62.4376 | 0.5022 |
| MPCNN&WBT | 7.7698 | 8.2097 | 17.0698 | 63.6202 | 0.7029 |

PCNN is a mammal visual cortex-inspired neuron networks and has been widely employed in image processing. The combination of DWT and MPCNN can make full use of the multi-resolution characteristics of DWT and the global couple and pulse synchronization characteristics of PCNN. The experimental results in Figure 4 and Table 1 show that the new method presented in this paper can improve the fusion effect. the entropy of the new algorithm are larger indicating the fused image contains more information, and the average gradient of the new algorithm is larger indicating that the fused image is more clear and contains more details and texture features.

Fused image of the proposed algorithm extracts almost all clear parts in source images and get good-focus on whole fused image than the fused image of wavelet-based algorithm and PCNN-based algorithm.

Acknowledgement

This paper is supported by National Natural Science Foundation of China under grant no.61271377 and Anhui provincial scientific research projects in Universities under grant no KJ2013B021

References

- [1] Zhong Z, Blum R S 1999 *Proceedings of the IEEE* **87**(8) 1315-26
- [2] Johnson J L, Padgett M L 1999 *IEEE Transaction on Neural Networks* **10**(3) 480-98
- [3] Keenan E, Wright R G, Zgol M, Mulligan R, Tagliavia V 2004 *IEEE A&E Systems Magazine* **19**(6) 9-15
- [4] Hak J, Yong L, Kim S, Lee D 2010 Robust CCD and IR image registration using gradient-based statistical information *IEEE Signal Processing Letters* **16**(4) 347-51
- [5] Malyutenko V K 2003 High resolution infrared "vision" of dynamic electron processes in semiconductor devices *Review of Scientific Instruments* **74**(1) 655
- [6] Jung Min-Suk, Lee Shin-Bok, Lee Ho-Young, Ryu Chang-Sup, Ko Young-Gwan, Park Hyung-Wook, Joo Young-Chang 2014 Improvement of electrochemical migration resistance by Cu/Sn intermetallic compound barrier on Cu in Printed Circuit Board *Device and Materials Reliability* **14**(1) 382-89
- [7] El-taweel G S, Helm A K 2013 Image fusion scheme based on modified dual pulse coupled neural network *Image Processing, IET* **7**(5) 407-14
- [8] Kong W, Lei Y, Lu X N 2011 Image fusion technique based on non-subsampled contourlet transform and adaptive unit-fast-linking pulse-coupled neural network *IET Image Process* **5**(2) 113-21
- [9] Agrawal D, Singhai J 2010 Image fusion technique based on non-subsampled contourlet transform and adaptive unit-fast-linking pulse-coupled neural network *IET Image Process* **4**(6) 443-51
- [10] Gonzalez-Audicana M, Saleta J L, Catalan R G, Garcia R 2004 Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition *IEEE Geoscience and Remote Sensing* **42**(6) 1291-99
- [11] Qu X, Yan J, Xie G, Zhu Z 2007 A novel image fusion algorithm based on bandelet transform *Chinese Optic Letters* **5**(10) 569-72

[12] Amolins K, Dare Z 2007 Wavelet-based image fusion techniques an introduction, review and comparison *ISPRS Photogramm Remote Sensing* 62 249-63

[13] Ellmauthaler A, Pagliari C L, 2013 Multiscale Image Fusion Using the Undecimated Wavelet Transform With Spectral Factorization and Nonorthogonal Filter Banks *Image Process* 22(3) 1005-17

[14] Wang Z J, Ziou D, Armenakis C, Li D, Li Q Q 2005 A comparative analysis of image fusion methods *Geoscience and Remote Sensing* 43(6) 1391-402

[15] Lewis J J, Nikoloc S G, Bull D R 2007 Pixel and region based image fusion with complex wavelets, *Inform fusion* 8(2) 119-30

[16] Ma Y D, Zhang, H J 2008 A new image denoising algorithm combined PCNN with gray-scale morphology *Beijing Univ. Posts Telecomm* 31(2) 108-12

[17] Wu Z G, Wang Y J, Li G J 2010 Application of adaptive PCNN based on wavelet transform to image fusion *Optics and Precision Engineering* 18(3) 708-15

[18] Zhao Z B, Guang Z J, Gao Q, Wang K Q 2013 Infrared and visible images fusion of Electricity Transmission Equipment using CT-domain Hidden Markov Tree Model *Gaodiyuan Jishu/High Voltage Engineering* 39(11) 2642-9

| Authors | |
|--|--|
|  | <p>Yuan Yiming, born in December, 1982, Anhui Province, China</p> <p>Current position, grades: the lecturer of School of Anhui Polytechnic University, China. University studies: B.S. in electrical information engineering from HeFei college, M.S. from Anhui Polytechnic University in China. Scientific interest: signal analysis and processing, detection technology and automatic equipment. Experience: teaching experience of 8 years, 2 scientific research projects.</p> |
|  | <p>Jiang Ming, born in March, 1965, Anhui Province, China</p> <p>Current position, grades: the professor of School of Anhui Polytechnic University, China. University studies: B.S. in automation from Mechanical & Electrical College, M.S. Shanghai University in China. Scientific interest: information fusion, detection technology and automatic equipment. Experience: teaching experience of 28 years, 10 scientific research projects.</p> |
|  | <p>Gao Wengen, born in May, 1973, Anhui Province, China</p> <p>Current position, grades: the lecturer of School of Anhui Polytechnic University, China. University studies: B.S. in electrical engineering and automation from Mechanical & Electrical College, M.S. from Anhui Polytechnic University in China. Scientific interest: information fusion, detection technology and automatic equipment Experience: teaching experience of 8 years, 5 scientific research projects.</p> |

Research on virus transmission of online social network

Min Yang*, Yaoliang Song, Qianmu Li

School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China

Received 3 April 2014, www.tsi.lv

Abstract

Online social networks (OSN) are up-and-coming complex network systems. Experiments indicate that it is difficult for simple complex network theory to describe virus transmission behaviour. Based on comprehensive research into current virus transmission, this paper combines user behaviour with social engineering theory and builds a model of virus transmission on OSN. Key factors affecting virus transmission on OSN are then analysed. Lastly, in light of public opinion transmission theory, this paper refers to social reinforcement factors concepts to describe computer virus transmission on OSN and analyses transmission disciplines in regular and random networks.

Keywords: online social network, virus transmission, epidemic spreading

1 Introduction

The classical virus transmission model SI/SIR/SIS is frequently used to simulate the spreading process of viruses, which is through nodes. On social networks, the relationship between nodes depends on the on-line status. Besides, we are more likely to receive information from friends. So, whether from the macro- or microscopic angle, virus transmissions on OSN differ greatly from those on off-line.

Using research into on-line social networks, Centola investigated the transmission process of public opinion on various network structures [1]. The research argues that social reinforcement can be a great impact factor in terms of transmission of behaviours on networks. If an individual is subject to certain behaviours or opinions from multiple friends then, from a macro point of view, this behaviour or opinion will tend to propagate faster on networks. Experiments show that a single signal cannot influence the decision of a node. Only when the node receives more signals can it receive information or produce behaviours. Information and behaviours can propagate faster on a regular network with a high clustering coefficient than on a random network, because people can receive signals more easily from other nodes on a regular network with a high clustering coefficient. This article describes the research of Centola, considers the differences between virus transmission and information behaviour transmission, and builds the social network virus transmission SEIR model combining with communication opinion [2].

2 Model description

2.1 MECHANISM OF VIRUS SPREADING IN THE MODEL

According to the model described in this article, the virus transmission process is as follows: Firstly, some virus infected nodes exist on social networks. They deliver the signal with a virus to all nodes on the friends list. However, only some of the friend nodes of this node will be infected by the virus. Then the infected nodes will spread the virus signal to all their friend nodes on their friends list. Thus these signals can be links to the virus or the Auto-run file of the virus in real life.

The nodes in social networks can be described by these four states:

- 1) S status. S status indicates the node has not received any signals and it can be infected [3].
- 2) I status. I status indicates the node has been infected by a virus and it will spread the virus signal to infect other nodes.
- 3) R status. R status indicates this node recovers from I status. It develops immunity, and it may receive the virus signal but not be infected.
- 4) E status. E status indicates the node has received the virus, but it is not infected and it will not spread the virus. E status can describe the issues that users receive malice information, but not believe and be infected. During the spread process, the node which receives more signals is more likely to be infected [4]. Through this process, the virus achieves spread on the social networks. The model is shown in Figure 1.

Firstly, randomly select a node as "seed". The seed node must be at I status, other nodes at S status. The seed node deliver virus signals to other nodes. Then it will recover to R status and will never infect or spread the virus further.

* *Corresponding author* e-mail: ym8670435@126.com

At time t , if a node at status S or status E receives the virus signal, then the probability for it to transform to status I is λ_m , m is the frequency the node receiving the signals of the virus, m and λ_m have a positive correlation. If the node does not receive any signal, the status will not change, irrespective of how many times the node received the signals before.

At time t , if the status of the node is I , during the next Δt , the node will deliver the virus to all its neighbours, and itself will transform to status R .

The changes of status of all the nodes are synchronous in the model, i.e., at time t all nodes assess their status at the next time simultaneously and make the changes during $t + \Delta t$.

When there is no node to change status, transmission of the virus stops.

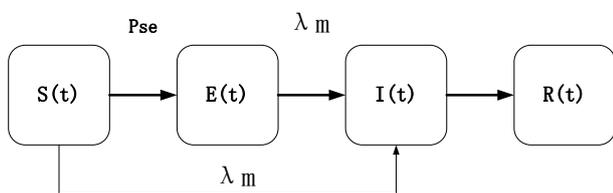


FIGURE 1 Mechanism of virus transmission in the model

The differences between the model provided by this article and other existing models are that we use the link between the nodes at most one time. For social networks, if one node delivers the information with the virus to its neighbour nodes many times, it will increase their vigilance. So, as social networks, few users will deliver the same information to their neighbours many times [5].

2.2 THE MATHEMATICAL MODEL OF THE VIRUS SPREADING PROCESS

When node j is at status S , it receives the virus signals the first time. We assume the probability of infection is λ_1 , λ_1 is the initial spread rate. When node j is at status S , the rate of infection when it receives the virus signals the second time is λ_2 , simply, when node j is at status S , the m -th time it receives the signals, the infected rate is λ_m , Here is the expression:

$$\begin{aligned} \lambda_1 &= \lambda_1, \\ \lambda_2 &= \lambda_1 + \alpha(1 - \lambda_1), \\ \lambda_3 &= \lambda_2 + \alpha(1 - \lambda_2), \\ &\dots \\ \lambda_m &= \lambda_{m-1} + \alpha(1 - \lambda_{m-1}). \end{aligned} \tag{1}$$

In the expression, $\alpha \in [0, 1]$ it means social reinforcement factors, the bigger α , the higher and the rate at which the other node can be infected. From the Equation (1), we know the infection rate for a node after receiving m -th times is $\alpha(1 - \lambda_{m-1})$ higher than if it receives $(m-1)$ -th

times. $\alpha(1 - \lambda_{m-1})$ can be regarded as the rate when the node isn't infected at the $(m-1)$ -th virus signal $1 - \lambda_{m-1}$ transforming to I status influenced by social reinforcement factors α . There are two special situations in the Equation (1), when $\alpha = 1$, social reinforcement factors have a great influence, and when $m > 2$, the rate of virus infection approaches 1. That means the node will be infected as soon as it receive the virus signals two times. When $\alpha = 0$, it means social reinforcement factors have no influence on the model. That is, the model provided by this article degenerates to the normal SIR model [6].

Simplifying and deforming the Equation (1) we can get the Equation (2):

$$\lambda_m = \begin{cases} \lambda_1 & \alpha = 1, m = 1 \\ 1 & \alpha = 1, m \geq 2 \\ 1 - (1 - \lambda_1)(1 - \alpha)^{m-1} & 0 \leq \alpha < 1, m \geq 1 \end{cases} \tag{2}$$

In the Equation (2), when α and λ_1 are fixed, λ_m will increase monotonically by m . That is, a node in social networks is more likely to be infected if it receives virus signals more times. As α increases, so does λ_m . In Figure 2, we describe the relationship among, m and α when $\lambda_1 = 0.2$.

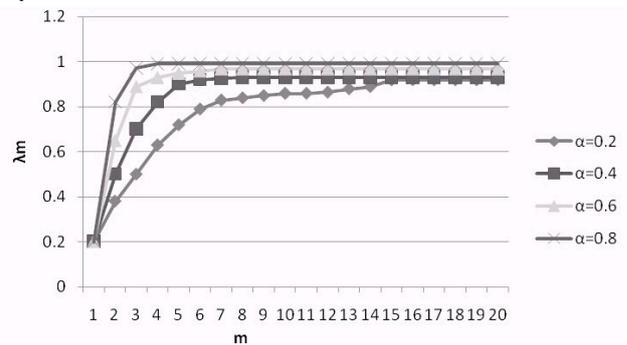


FIGURE 2 the relationship between λ_m , m and α

During the spreading process, a node turns from status S to status E , meaning the node received the virus signals but is not infected and will not spread the virus. The rate P_{SE} is decided by each neighbouring node.

$$P_{SE} = 1 - \prod_{c \in N_j} \prod_{n=1}^{m-1} (1 - \lambda_n) \tag{3}$$

Node i is the neighbour of node j . $\prod_{c \in N_j} \prod_{n=1}^{m-1} (1 - \lambda_n)$ means at time $m-1$, the neighbour of node j still has not attained the rate of status S . which are the neighbours of node j , the j node must receive the virus signals next time.

Combining Figure 1 with the description of the mechanism of the model, we can get the node status changing the sum formula:

$$S(t + \Delta t) = S(t)(1 - \lambda_m), \tag{4}$$

$$E(t + \Delta t) = S(t) \left(1 - \prod_{c \in N_j} \prod_{n=1}^{m-1} (1 - \lambda_n) \right) + E(t)(1 - \lambda_m), \tag{5}$$

$$I(t + \Delta t) = S(t)\lambda_m + E(t)\lambda_m, \tag{6}$$

$$R(t + \Delta t) = I(t). \tag{7}$$

3 Simulation experiment and analysis

3.1 EXPERIMENTAL NETWORK SELECTION AND CONSTRUCTION

The regular network that simulation experiments mainly select is the Moore network, with the network boundary conditions periodic. The network structure is shown in Figure 3:

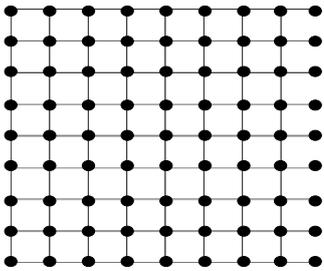


FIGURE 3 Nodal neighbour structure in the Moore network

Referring to experimentation in which Centola researched the dissemination of public behaviour in the network, in the simulation experiments in this paper, we select the network structure in which the length and width ratio is two. It is a uniform random network using the Maslov-Sneppen small world model [1]. The process of setting it up is as follows:

- 1) At time t , we randomly select a pair of edges A-B, C-D, then we reconnect this pair of edges to A-D, B-C. In the process of reconnection, we do not allow self-connection and reconnection.
- 2) In order to establish a completely random network structure, the reconnection process must be repeated many times. In the experiment in this article, the number of repeating the above process of selection is pN_E . In pN_E , p can be used to describe the randomness of constructed random networks, N_E is the number of connections in random networks.
- 3) Strictly speaking, only when $p \rightarrow \infty$, it is a random network really but as the literature shows simulations of topological information and true random network are very

approximate when $p > 1$, so in this paper $p = 10$.

The main tool for virus transmission simulation experiments on a regular Moore network is cellular automata, so we set up a cellular automata model $T(C, Q, V, f)$: Cellular space is selected as a quadrangle cellular space; Cellular discrete state is combined as $Q = \{S, E, I, R\}$; Cellular neighbour set's selection is as shown in Figure 1, the cellular boundary's selection cycle type; Cellular transformation rules as shown in Equation (7).

3.2 DESIGN OF SIMULATION EXPERIMENT AND ANALYSIS OF THE RESULTS

3.2.1 Experiment about the impact on the experimental results and analysis of different network topologies

Let $\rho_t = \frac{I_t}{N_t}$, ρ_t represent the proportion of nodes whose state is I to the total nodes in the network space at time t , ρ_{sreg} and ρ_{srand} respectively represent the proportions of nodes whose state is I to the total nodes in the regular networks and random networks in the final stable state. In order to compare the spread of the virus in the two different network topological structures, we define $\delta_p = \rho_{sreg} - \rho_{srand}$. If $\delta_p > 0$, the virus spread in the dissemination of a wider range in a regular network than in a random network; if $\delta_p < 0$, the spread of the virus is faster in the random network.

Figure 4 shows the relation diagram of α , λ_1 and ρ_t in the regular networks and random networks. In Figure 4, the longitudinal coordinates are α the abscissa is λ_1 , different colors represent different ρ_∞ .

In Figure 5, when the $\lambda_1 = 0.1, 0.2, \dots, 1$, we take different social enhancement factor α 's values of ρ_∞ ; it will help us more clearly observe the change trend of ρ_∞ and provide a supporting role to explain Figure 4.

In order to more intuitively represent the spread of the virus in two different kind of networks, we use along with the changes of α and λ_1 , and we use Matlab visual effects to draw Figure 6. In Figure 6 the longitudinal coordinates are α , the abscissa is λ_1 , different colours represent different δ_p .

As shown as Figure 4, when α is larger but λ_1 is smaller, the range of the spread of the virus is bigger in a regular network than in a random network. However, when α is smaller and λ_1 is larger (about 0.3), the range of the spread of the virus is larger in a random network than in a regular network.

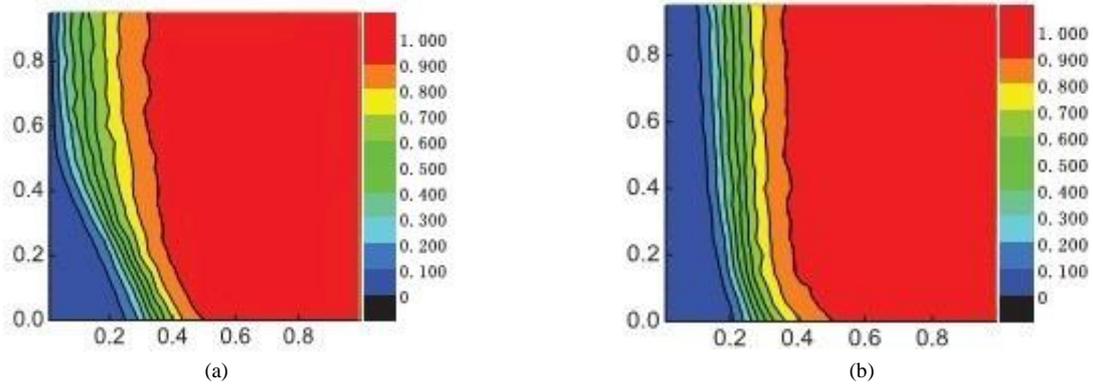


FIGURE 4 ρ_∞ in the regular networks and random networks

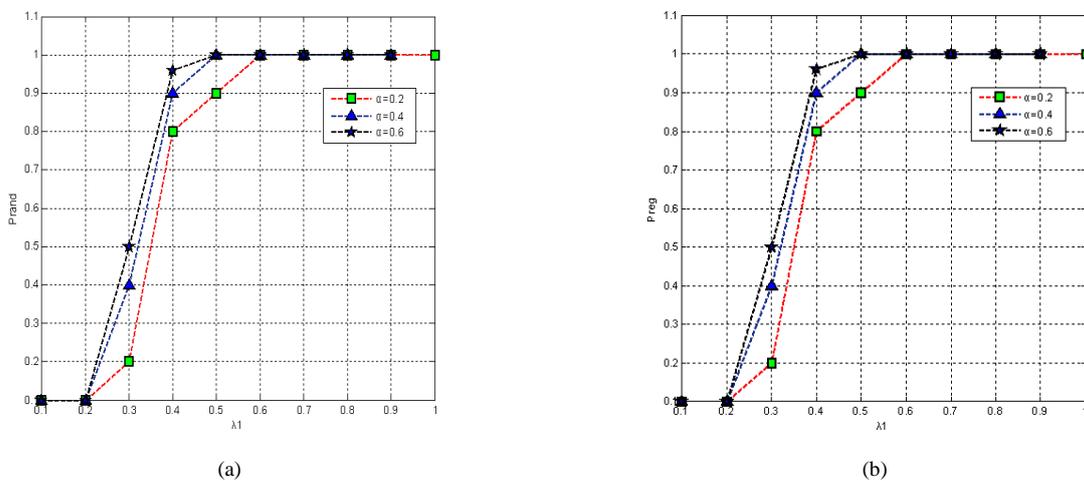


FIGURE 5 ρ_∞ in the regular networks and random networks when λ_1 and α are given

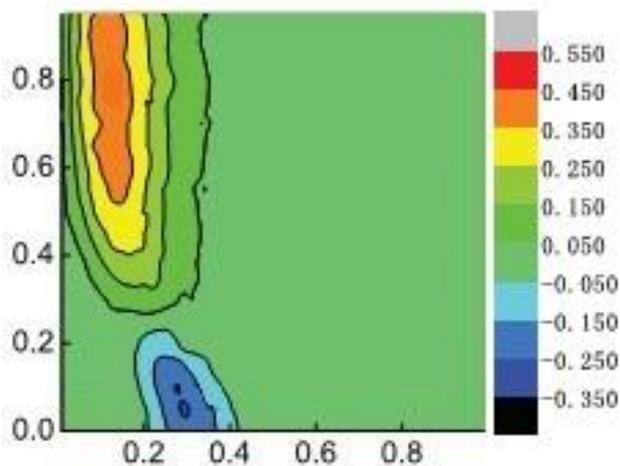


FIGURE 6 δ_p in two different kinds of networks

Figure 6 shows the distribution of δ_p in regular networks and random networks in the space (λ_1, α) . In Figure 6, we can clearly see that there are two isolated “island” shapes in this figure: one is that when λ_1 is small and α is big, the scope of the spread of the virus is much

greater in a regular network, which is consistent with Centola’s experimental results. The other is when λ_1 is relatively larger and α relatively smaller, the scope of the spread of the virus is much greater in random networks than in regular networks. This is fully consistent with the conclusion of previous researchers: viruses spread faster in random networks than in regular networks. δ_p has a gap except in two ranges, the other times the value of δ_p is zero.

Analysis of two areas is worthwhile, the first region is located in the top right corner of Figure 6 areas: When λ_1 is very large, regardless of whether social reinforcement factor α is big or small, the virus will spread almost throughout all the network range, the expression of experiment shown that there have not any basic difference of virus between regular networks and random networks; the second region is located in the lower left corner of Figure 5 areas: when λ_1 is very small and social enhancement factor α is very small, the virus in the two kinds of networks are not spread very well apart, in Figure 4 we can see that in the steady state I node occupies a small proportion of all nodes in the network in 0.1.

3.2.2 Experiment about evolution of I state nodes in the model

In order to illustrate the effectiveness of the models, in Figure 7 we give the change curve of the infected nodes' amount with the curve of time in regular networks and random networks, and in Figure 8 we present the experimental results of Centola. In Figure 8, the black solid circular and hollow triangle respectively represent the number of individuals receiving public opinion behaviour dissemination in regular networks and random networks.

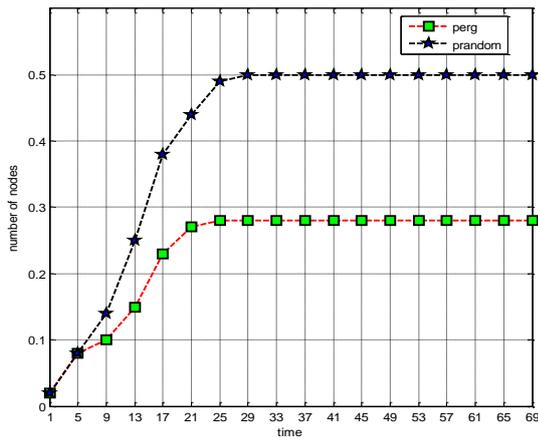


FIGURE 7 The change curve of the infected nodes' amount with the curve of time

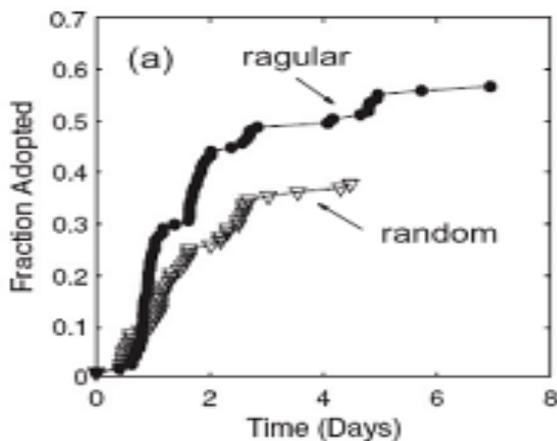


FIGURE 8 The experimental results of Centola

The parameters in Figure 7 is that the value of λ_1 is 0.18 and the value of α . By comparing Figures 7 with 8, we can find that from the preliminary transmission rate and the final steady-state communication range, regular networks spread faster and wider than random networks. In order to describe each time t infected nodes' density, we use $\rho_t - \rho_{t-1}$ to describe the increased amount of the I node's density in regular networks and random networks in each time. Thus we can draw Figure 9:

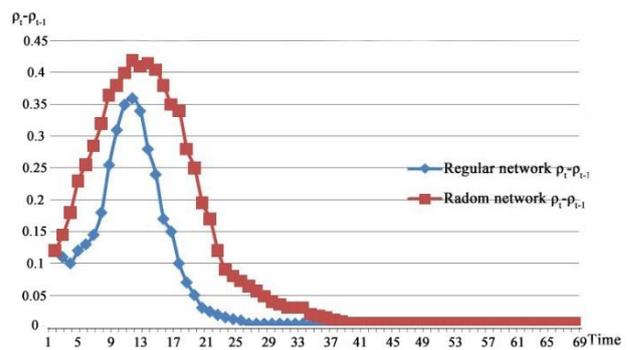


FIGURE 9 The change of $\rho_t - \rho_{t-1}$ in regular network and random networks

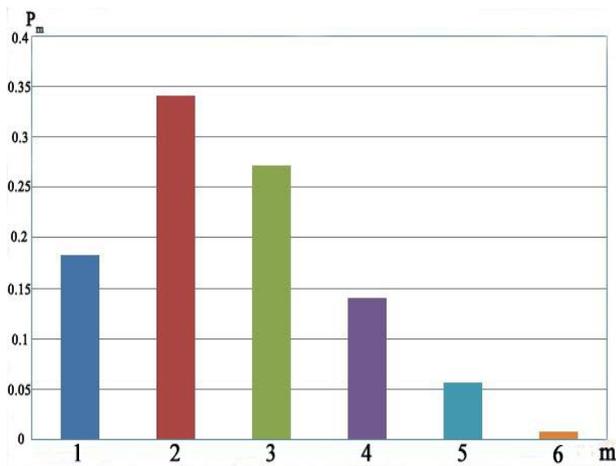
In Figure 9 we can see that when virus transmission process begins, the rate of spread of the virus in the two networks grows very fast. After the two curves are almost simultaneously at a peak of their own, this corresponds to virus outbreak events in the real world. After reaching the peak value, the two curves begin to decline rapidly with the passage of time. In the propagating process of the virus, $\rho_t - \rho_{t-1}$ is always bigger in regular networks than in random networks, which means that the transmission rate of the virus is bigger in regular networks than in random networks. The value of $\rho_t - \rho_{t-1}$ is zero, meaning that with the termination of the propagating process, the range of transmission of the virus reaches the maximum value.

3.2.3 Experiment about the critical value of the social enhancement factor when the virus spreads in the social network

In the propagating process of the virus, α plays an important role in the transmission process of the virus in the social network. In Experiment 2, when α is equal to 0.52 and λ_1 is equal to 0.14, finally there is a δ_ρ difference of about 0.35. Therefore, an issue emerges: what is the social enhancement factor's influence in the spread of the virus, namely, how many does a node need to receive a signal to be infected?

Parameters of Experiment 2 were selected as follows: λ_1 is 0.18, α is 0.4. The model is simulated by cellular automata, degree of nodes is four. The cellular space is 1000, the experiment is carried out 10 times, the statistic selection for the network reaches the steady state. Each cell has a count value to calculate the number of signals of the virus which are to be received to change into the R state, according to the rules of virus transmission, count subtract one to calculate the number of signals of the virus which are to be received to change into the I state [7].

We define P_m as the probability of being infected by the virus after the user receives the virus's signal m times. The P_m statistics are as follows:

FIGURE 10 P_m statistics

As shown in Figure 10, only about 17% of people after they receive the virus's signal only once will accept and become infected with the virus. When the node receives two signals of the virus, more than 30% of people will choose to believe the information and will become infected with the virus. Although when m is equal to 3 or 4, the degree of users' adoption of virus information is very high, the experiment proved that the second virus signal is the most important. This conclusion is verified by the experiment of Centola [9].

4 Verify and description of the validity of the model

The first experiment forms the basis for all other tests. In Experiment 1, we tested the relationship between λ_1 and ρ_t in the case of different topologies and different values of α . In Figures 3 and 4, what are shown are the values of ρ_{reg} and ρ_{rand} in the hexagonal network topology. There are two examples to prove the validity of the model presented in this paper: One is under the condition of small λ_1 and large α the scope of transmission of the virus is large in the regular network, which is consistent with the conclusions of Centola experiments; another is under the condition of relatively large λ_1 and smaller α , the scope of transmission of the virus in the random network is wider than the one in the regular network. This is in line with the research conclusions of previous scholars: A virus travels faster in a random network than in a regular network [10].

In Experiment 2, we compared the experimental results with Centola. In addition to good description of the spread of the virus on social networks, this model can also describe well the Centola experiment when $\lambda_1 = 0.18$, $\alpha = 0.4t$, which confirms the validity of this model from the side.

In Experiment 3, we obtained statistical information on

the social reinforcement factor parameter that makes users infected with the virus. By averaging the testing values of repeated measurements, we can infer that the second time users receive the virus information is crucial for whether or not the node is infected with the virus, as the same point is also verified in the Centola experiment.

5. Conclusion

This paper combines innovatively with social public opinion communication and establishes a SEIR virus spread on a social network model. Using Matlab and quadrilateral cellular space and periodic boundary conditions of cellular automata as a tool, the model was put forward by the simulation experiment. In the experiments we first established regular and uniform random networks, then studied the effects of network topology, the social reinforcement factor parameters α , and the initial virus infection rate λ_1 on the process of the spread of the virus.

Studies have shown that in the process of the spread of the virus in a social network, social reinforcement factor α and the initial transmission rate λ_1 played a very important role. The main conclusions and results of the model are as follows:

- 1) Even when the initial transmission rate λ_1 is very small, the virus can still be spread on the regular network if social reinforcement factor α relatively large, but on the same condition, the virus cannot be spread extensively on uniform random networks. That is to say, the virus travels faster and wider in the random network than in the regular network. This conclusion supports the results of the Centola experiment in certain cases. When social reinforcement factor α is 0.4 and the initial transmission rate λ_1 is 0.18, the model proposed in this paper can be used to simulate the Centola experimental network.
- 2) When the initial transmission rate λ_1 is bigger, social reinforcement factor α is smaller, the spread of the virus travels faster and wider in the uniform random network than in the regular network. This conclusion is consistent with the traditional conclusion of virus spread on the network.
- 3) When the initial transmission rate λ_1 is very large, all nodes in the network have a high probability of infection no matter what social reinforcement factor α is. As a result, the virus spreads quickly to the whole social network, and the spread of the virus has nothing to do with the network structure.
- 4) When social reinforcement factor $\alpha = 0$ the proposed model can be degraded as the standard SIR model for the spread of the virus in complex networks.

References

- [1] Zou C, Gong W, Towsley D 2007 *IEEE Transactions on Dependable and Secure Computing* 4(2) 105-118
- [2] Anderson R., May R M 2002 *Infectious Diseases in Humans Oxford University Press: Oxford* 56-77
- [3] Yang M, Li Q, Song Y 2014 A SEIR Model Epidemic of Virus on the Online Social Network *Journal of Digital Information Management* 12(2) 102-7
- [4] Jin C, Liu J 2009 Network virus transmission model based on effects of removing time and user vigilance *International Journal of Network Security* 9(2) 156-63
- [5] Li Q, Zhang H 2012 Information Security Risk Assessment Technology of Cyberspace: A Review *Information An International Interdisciplinary Journal* 15(11) 4677-84
- [6] Li Q, Li J 2007 Rough Outlier Detection Based Security Risk Analysis Methodology *China Communications* 5(7) 14-21
- [7] Maslov S, Sneppen K 2002 Specificity and Stability in Topology of Protein Networks *Science* 296 910-3
- [8] Chen Z, Ji C 2005 *IEEE Transactions on Neural Networks* 16(5) 1291-303
- [9] Balthrop J, Forrest S, Newman M E J 2004 Computer Science: Technological Networks and the Spread of Computer Viruses *Science* 304(5670) 527-9

| Authors | |
|--|---|
|  | <p>Min Yang, born in April, 1978, Nanjing, Jiangsu Province, P. R. China</p> <p>University studies: Ph.D at Nanjing University of Science and Technology. Scientific interest: Computer and electronic engineering. Experience: Teaching experience of 13 years.</p> |
|  | <p>Yongliang Song, born in June, 1960, Nanjing, Jiangsu Province, P. R. China</p> <p>University studies: Ph.D at Nanjing University of Science and Technology. Scientific interest: Communication and information systems. Experience: Teaching experience of 30 years, 8 scientific research projects.</p> |
|  | <p>Qianmu Li, born in January, 1979, Nanjing, Jiangsu Province, P. R. China</p> <p>University studies: Ph.D at Nanjing University of Science and Technology Scientific interest: Computer applications Experience: Teaching experience of 10 years, 10 scientific research projects</p> |

SVM classification of hyperspectral images based on wavelet kernel non-negative matrix factorization

Lin Bai*, Meng Hui

School of Electronics and Control Engineering, Chang'An University, 710064, Xi'An, P.R.China

Received 1 March 2014, www.tsi.lv

Abstract

This paper presents a new kernel framework for hyperspectral images classification. In this paper, a new feature extraction algorithm based on wavelet kernel non-negative matrix factorization (WKNMF) for hyperspectral remote sensing images is proposed. By using the feature of multi-resolution analysis, the new method can improve the nonlinear mapping capability of kernel non-negative matrix factorization. The new classification method of hyperspectral image data combined with the novel kernel non-negative matrix factorization and support vector machine (SVM). The simulations results show that, the method of WKNMF reflect the nonlinear characteristics of the hyperspectral image. Experimental results on Airborne Visible Infrared Imaging Spectrometer 220 bands data in Indian pine test site and HYDICE 210 bands hyperspectral imaging in Washington DC Mall are both show that the proposed method achieved more strong analysis capability than comparative algorithms. Compared with the PCA, non-negative matrix factorization and kernel PCA method, classification accuracy of WKNMF with SVM can be improved over 5%-10%.

Keywords: hyperspectral, non-negative matrix factorization, classification, support vector machine, kernel method

1 Introduction

It is well know that each material has its own specific electromagnetic radiation spectrum characteristic. Using hyperspectral sensors, it is possible to recognize materials and their physical states by measuring the spectrum of the electromagnetic energy they reflect or emit. The spectral data which consist of hundreds of bands are usually acquired by a remote platform, such as a satellite or an aircraft, and all bands are available at increasing spatial and spectral resolutions. After 20 years of development, hyperspectral technology has not only been widely used in military, but also has been successfully applied in ocean remote sensing, vegetation surveys, geological mapping, environmental monitoring and other civilian areas [1, 2].

Due to the state of art of sensor technology developed recently, an increasing number of spectral bands have become available. Huge volumes of remote sensing images are continuously being acquired and archived. This tremendous amount of high spectral resolution imagery has dramatically increased the information source and increased the volume of imagery stored. For example, hyperspectral imagery captured by Airborne Visible Infrared Imaging Spectrometer (AVIRIS, operated by NASA) includes 224 bands, which contains up to 140Mbytes [2, 3].

However, the excessive hyperspectral data increase the difficulty of image processing and analysis. Such as supervised classification of hyperspectral images is a very challenging task due to the generally unfavourable ratio

between the large number of spectral bands and the limited number of training samples available a priori, which results in the 'Hughes phenomenon'. Without the supports of new scientific concepts and novel technological methods, the existing large volumes of data prohibit any systematic exploitation. This has led to great demands to develop new concepts and methods to deal with large data sets [2-4].

Hyperspectral image classification has been a very active area of research in recent years [5]. Given a set of observations, the goal of classification is to assign a unique label to each pixel vector so that it is well-defined by a given class.

There are several important challenges when performing hyperspectral image classification. Supervised classification faces challenges related with the unbalance between high dimensionality and limited availability of training samples, or the presence of mixed pixels in the data. Another relevant challenge is the need to integrate the spatial and spectral information to take advantage of the complementarities that both sources of information can provide [5].

Over the last years, many feature extraction techniques have been integrated in processing chains intended for reduce the dimensionality of the data, thus mitigating the Hughes phenomenon. These methods can be unsupervised or supervised. Classic unsupervised techniques include principal component analysis (PCA), or independent component analysis (ICA). Supervised approaches comprise discriminant analysis for feature extraction (DAFE), decision boundary feature extraction

* *Corresponding author* e-mail: Bai1981@sina.com.cn

(DBFE), and non-parametric weighted feature extraction (NWFE), among many others [4-7].

Recently, it was shown by Lee and Seung that positivity or non-negativity of a linear expansion is a very powerful constraint that also seems to yield sparse representations [8, 9]. Their technique, called non-negative matrix factorization (NMF), was shown to be a useful technique in approximating high dimensional data where the data are comprised of nonnegative components. However, NMF and many of its variants are essentially linear, and thus can't disclose nonlinear structures hidden in the hyperspectral data. Besides, they can only deal with data with attribute values, while in many applications we do not know the detailed attribute values and only relationships are available. The NMF cannot be directly applied to such relation data. Furthermore, one requirement of NMF is that the values of data should be non-negative, while in many real world problems the non-negative constraints cannot be satisfied.

Since the mid-1990s, nuclear method has been successfully applied in the future, there are many scholars have proposed Nonlinear feature extraction method based on kernel method [10-13].

In this paper, a novel study is proposed for the feature extraction of high volumes of remote sensing images by using wavelet kernel non-negative matrix factorization (WKNMF). We propose the WKNMF, which can overcome the above limitations of NMF. Classification experiments on AVIRIS and HYDICE data sets by combination of feature extraction method and support the vector machine (SVM). The proposed method is applied to experiment data sets, compared with the other algorithms the classification accuracy can be increased over 5%-10%. The outline of this paper is as follows. Section 2 presents the proposed feature extraction based on WKNMF. Experimental results are reported in section 3. Finally, conclusions are given in section 4.

2 Methodology

2.1 NON-NEGATIVE MATRIX FACTORIZATION

NMF imposes the non-negativity constraints in learning the basis images. Both the values of the basis images and the coefficients for reconstruction are all non-negative. The additive property ensures that the components are combined to form a whole in the non-negative way, which has been shown to be the part based representation of the original data. However, the additive parts learned by NMF are not necessarily localized [8, 9].

Given the non-negative $n \times m$ matrix V and the constant r , the non-negative matrix factorization algorithm finds a non-negative $n \times r$ matrix W and another non-negative $r \times m$ matrix H such that they minimize the following optimality problem: $\min f(W, H)$.

$$\text{Subject to } W \geq 0, H \geq 0, \tag{1}$$

This can be interpreted as follows: each column of matrix W contains a basis vector while each column of H contains the weights needed to approximate the corresponding column in V using the basis from W . So the product WH can be regarded as a compressed form of the data in V . The rank r is usually chosen $r \ll \min(n, m)$. $f(W, H)$ is a loss function. In this paper, we set loss function as follow:

$$f(W, H) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m (V_{ij} - (WH)_{ij})^2. \tag{2}$$

Solving the multiplicative iteration rule function as follows:

$$H_{bj} \leftarrow \frac{(W^T V)_{bj}}{(W^T W H)_{bj}}, W_{ib} \leftarrow W_{ib} \frac{(V H^T)_{ib}}{(W H H^T)_{ib}}. \tag{3}$$

The convergence of the process is ensured. The initialization is performed using positive random initial conditions for matrices W and H .

2.2 KERNEL NON-NEGATIVE MATRIX FACTORIZATION

Given m objects $\Theta_1, \Theta_2, \Theta_3, \dots, \Theta_m$, with attribute values represented as an n by m matrix $\Omega = [\omega_1, \omega_2, \dots, \omega_m]$, each column of which represent one of the m objects. Define the nonlinear map from original input space Ω to a higher or infinite dimensional feature space Φ as follows

$$\phi: x \in \Omega \rightarrow \phi(x) \in \Phi. \tag{4}$$

From the m objects, denote

$$\phi(\Omega) = [\phi(\omega_1), \phi(\omega_2), \dots, \phi(\omega_m)]. \tag{5}$$

Similar as NMF, KNMF finds two non-negative matrix factors W_ϕ and H such that

$$\phi(\Omega) = W_\phi H. \tag{6}$$

W_ϕ is the bases in feature space Φ and H is its combining coefficients, each column of which denotes now the dimension-reduced representation for the corresponding object. It is worth noting that since $\phi(\Phi)$ is unknown. It is impractical to directly factorize $\phi(\Omega)$. From Equation (6), we obtain

$$(\phi(\Omega))^T \phi(\Omega) = (\phi(\Omega))^T W_\phi H. \tag{7}$$

A kernel is a function in the input space and at the same time the inner product in the feature space through the kernel-induced nonlinear mapping. More specifically, a kernel is defined as

$$k(x, y) = \langle \phi(x), \phi(y) \rangle = (\phi(x))^T \phi(y). \tag{8}$$

From Equation (8), the left side of Equation (7) can be rewritten as:

$$\begin{aligned} (\phi(\Omega))^T \phi(\Omega) &= \{(\phi(\omega_i))^T \phi(\omega_j)\}_{i,j=1}^m \\ &= \{k(\omega_i, \omega_j)\}_{i,j=1}^m = K, \end{aligned} \tag{9}$$

Denote

$$Y = (\phi(\Omega))^T W_\phi. \tag{10}$$

From Equation (9) and (10), Equation (7) can be changed as:

$$K = YH. \tag{11}$$

Comparing Equation (11) with Equation (6), it can be found that the combining coefficient H is the same. Since W_ϕ is a learned base of $\phi(\Omega)$, similarly we call Y in Equation (11) as the bases of the kernel matrix K . Equation (11) provides a practical way for obtaining the dimension-reduced representation H by performing NMF on kernels.

For a new data point, the dimension-reduced representation is computed as follows

$$\begin{aligned} H_{new} &= (W_\phi)^+ \phi(\omega_{new}) \\ &= (W_\phi)^+ ((\phi(\Omega))^T)^+ (\phi(\Omega))^T \phi(\omega_{new}). \\ &= Y^+ K_{new} \end{aligned} \tag{12}$$

Here A^+ donates the generalized (Moore-Penrose) inverse of matrix A , and $K_{new} = (\phi(\Omega))^T \phi(\omega_{new})$ is the kernel matrix between the m training instance and the new instance. Equation (11) and (12) construct the key components of KNMF when used for classification, it is easy to see that, the computing of KNMF need not to know the attribute values of objects, and only the kernel matrix K and K_{new} are required.

Obviously, KNMF is more general than NMF because the former can deal with not only attribute value data but also relational data. Another advantage of KNMF is that it is applicable to data with negative values since the kernel matrix in KNMF is always non-negative for some specific kernels.

2.3 WAVELET KERNEL NON-NEGATIVE MATRIX FACTORIZATION

The purpose of building kernel function is project hyperspectral observed data from low dimensional space to another high dimensional space. This WKNMF method uses the kernel function into the non-negative matrix factorization and improved it by replaced the traditional kernel function with the wavelet kernel function. By the feature of multi-resolution analysis, the nonlinear mapping capability of kernel non-negative matrix factorization method can be improved.

Assuming $h(x)$ is a wavelet function, parameter α represent stretch and β represent pan. If there $x, x' \in R^N$, then we get dot product form of wavelet kernel function:

$$K(x, x') = \prod_{i=1}^N h\left(\frac{x_i - \beta_i}{\alpha}\right) h\left(\frac{x'_i - \beta'_i}{\alpha}\right). \tag{13}$$

Meet the reasonable expression product approved under the condition of translation invariance, the Equation (13) can be rewritten as:

$$K(x, x') = \prod_{i=1}^N h\left(\frac{x_i - x'_i}{\alpha}\right). \tag{14}$$

In this paper Morlet wavelet function was selected as generating function, according to the theory of translation invariance wavelet function, kernel function constructed as:

$$h(x) = \cos(1.75x) e^{(-x^2/2)}. \tag{15}$$

From Equation (13), (14) and (15) a wavelet kernel function meets the requirements of Mercer kernel function build as:

$$K(x, x') = \prod_{i=1}^N \left(\cos\left(1.75 \frac{(x_i - x'_i)}{\alpha}\right) e^{-\frac{\|x_i - x'_i\|^2}{2\alpha^2}} \right). \tag{16}$$

Use Equation (16) in kernel non-negative matrix factorization, we can get Wavelet kernel non-negative matrix factorization.

2.4 SUPPORT VECTOR MACHINE CLASSIER INTRODUCTION

In machine learning, SVM are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two

possible classes forms the output, making it a non-probabilistic binary linear classifier.

Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. The basic mathematical formula of SVM is:

$$\min_{w,b} \Phi(w,b) = \frac{1}{2} \|\omega\|^2$$

$$s.t. y_i(w \cdot x_i - b) \geq 1 (i = 1, \dots, n) . \tag{17}$$

For more information about SVM see reference [14, 15].

3 Experimental results

3.1 EXPERIMENTAL ON AVIRIS DATA SET

The experiments were carried out on hyperspectral images produced by the AVIRIS. In order to simplify the logistics of marking this example analysis available to others, only a small portion of data set was chosen for this experiment. It contains 145 lines by 145 pixels (21025 pixels) and 190 spectral bands selected from a June 1992 AVIRIS data set of a mixed agriculture/forestry landscape in the Indian Pine Test Site in Northwestern Indiana.

For verification the feature extraction algorithm effect to hyperspectral data classification application, SVM classifier used in this paper. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into

that same space and predicted to belong to a category based on which side of the gap they fall on.

We select corn-min, corn-notill, soybean-min, soybean-notill and woods from AVIRIS images for classification experiment. Each object classes include 1434, 834, 968, 2468 and 1294 sample point respectively. The 3-bands (20, 80, 140 band) false colour synthesis image used in experiment and the ground truth are shown in Figure 1.

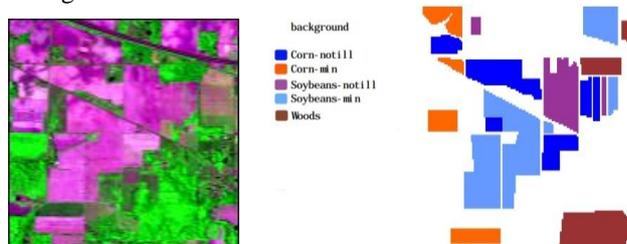


FIGURE 1 False colour images and ground truth of AVIRIS

Experiments using PCA, NMF, polynomial kernel KPCA (Poly-KPCK) comparison with WKNMF respectively, which Poly-KPCK coefficient kernel function is 5. To verify the classification capabilities of different feature extraction algorithm, we use Euclidean distance as the sum of the difference between the experimental data points in each band to take images of the same type of experimental data.

Take the Euclidean distance difference of surface features points between the different categories as a distance between the classes. The ratio of distance between the classes and distance within classes' values can reflect the degree to distinguish between different data. The experimental result was shown in Figure 2. From the experimental results, we can see WKNMF can get lowest ratio value than other algorithms. The result proves the new feature extraction method in this paper can effectively improve the discrimination between hyperspectral images category.

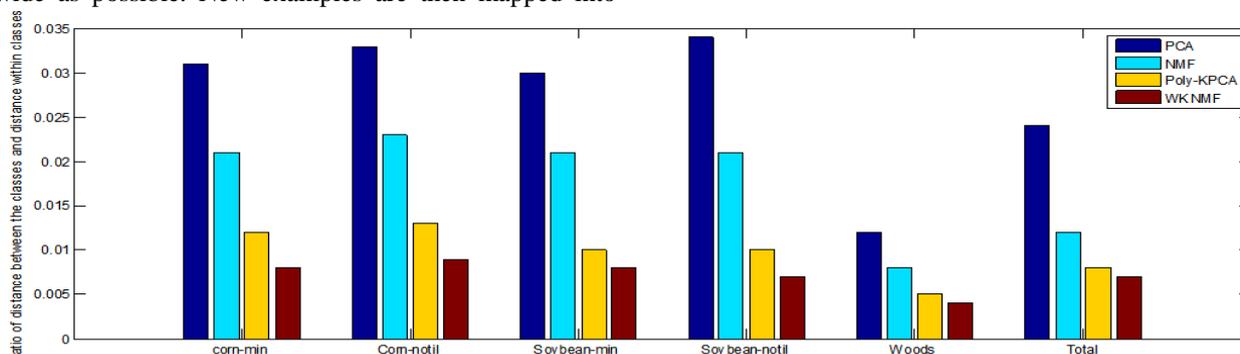


FIGURE 2 Ratio of distance between the classes and distance within classes

In order to verify the classification performance of feature extraction algorithm, experiments using the SVM method as a classifier, respectively PCA, NMF, polynomial KPCA (Poly-KPCA coefficient kernel function is 5) as feature extraction was compared with

WKNMF. We use the overall accuracy (OA), as the evaluation index in experiment results.

Experiment randomly select 10% samples as training data on original hyperspectral data and the remaining 90% of sample as test data. The classification experiment

was repeated 10 times, taking the statistical average for final results.

Experiment with feature extraction algorithm, feature dimensions taken before 15 feature components as input, the energy of the total energy accounted for more than

96%. The classification result was shown as Table 1. An impact of feature dimensionality to the SVM classifier for hyperspectral remote sensing images was shown as Figure 3.

TABLE 1 Classification results use 10% training sample data

| Methods | corn-min | corn-notil | soybean-min | soybean-notil | woods | Total (OA) | Kappa |
|-----------|----------|------------|-------------|---------------|-------|------------|-------|
| PCA | 85.22 | 46.21 | 50.14 | 96.05 | 98.81 | 78.96 | 0.763 |
| NMF | 88.69 | 69.05 | 67.32 | 96.17 | 99.92 | 85.03 | 0.837 |
| Poly-KPCA | 88.13 | 73.54 | 65.89 | 97.11 | 99.92 | 86.81 | 0.849 |
| WKNMF | 92.93 | 82.12 | 76.88 | 96.61 | 99.81 | 91.91 | 0.893 |

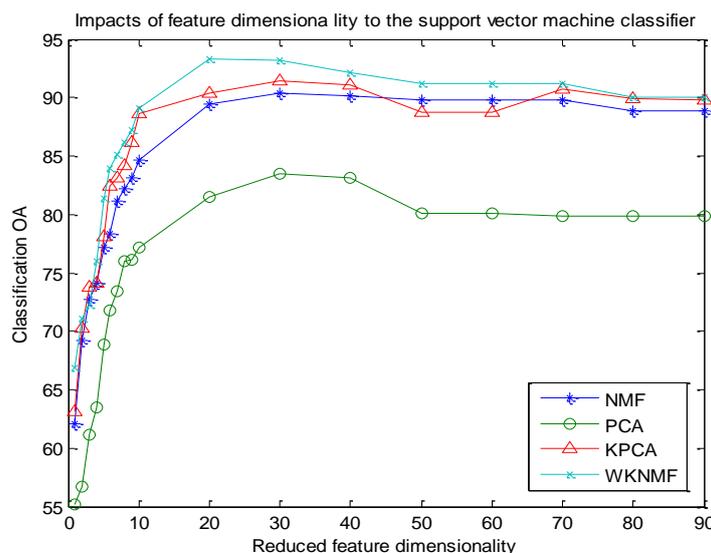


FIGURE 3 Classification OAs with respect to reduced dimensionality in AVIRIS

The overall classification accuracy of test samples show that the use of a few training samples, WKNMF method can achieve higher classification accuracy. The classification accuracy enhance effect is very obvious, especially in corn-notil and soybean-min. Compared with the other algorithms WKNMF can improve the overall classification accuracy over 10%. We can see from table 1, variety of feature extraction algorithms not work very well in corn-notil and soybean-min, because of their spectral are similar and easily misclassification. Even under such adverse circumstances that the proposed method can still get higher classification accuracy than others.

3.2 EXPERIMENTAL ON HYDICE DATA SET

The Figure 4 shows a simulated colour IR view of an airborne hyperspectral data flightline over the Washington DC Mall provided with the permission of Spectral Information Technology Application Centre of Virginia who was responsible for its collection. The sensor system used in this case measured pixel response in 210 bands in the 0.4 to 2.4 μm region of the visible and infrared spectrum. Bands in the 0.9 and 1.4 μm region where the atmosphere is opaque have been omitted from the data set, leaving 191 bands. The data set contains 1208 scan lines with 307 pixels in each scan line. It totals

approximately 150 Megabytes. The image at left was made using bands 60, 27, and 17 for the red, green, and blue colours respectively. The HYDICE data set include Roofs, Street, Path (gravelled paths down the mall centre), Grass, Trees, Water, and Shadow.



FIGURE 4 False colour images of HYDICE

Experimental test data and training data are selected as shown in Table 2.

TABLE 2 Experimental data

| HYDICE data set(Washington DC Mall) | | | |
|-------------------------------------|------------|---------|------|
| classification | | samples | |
| Class No. | Class name | Train | Test |
| 1 | Roofs | 400 | 3434 |
| 2 | Street | 168 | 248 |
| 3 | Path | 36 | 139 |
| 4 | Grass | 814 | 1114 |
| 5 | Trees | 80 | 325 |
| 6 | Water | 224 | 1000 |
| 7 | Shadow | 11 | 86 |

In order to verify the classification performance of feature extraction algorithm, experiments using the SVM method as a classifier, respectively PCA, NMF, polynomial KPCA (Poly-KPCA coefficient kernel function is 5) as feature extraction was compared with WKNMF. We use the overall accuracy (OA), as the evaluation index in experiment results. The classification experiment was repeated 10 times, taking the statistical average for final results.

Experiment with feature extraction algorithm, feature dimensions taken before 20 feature components as input, the energy of the total energy accounted for more than 97%. The classification result was shown as Table 3. An impact of feature dimensionality to the SVM classifier for hyperspectral remote sensing images was shown as Figure 5.

TABLE 3 Classification results on HYDICE data set

| Class No. | Class name | Classification Algorithms | | | | |
|------------------|------------|---------------------------|---------|---------|----------|-----------|
| | | SVM | PCA+SVM | MNF+SVM | KPCA+SVM | WKNMF+SVM |
| 1 | Roofs | 62.1% | 64.8% | 66.4% | 70.7% | 78.4% |
| 2 | Street | 98% | 100% | 94.8% | 98.4% | 98.6% |
| 3 | Path | 100% | 100% | 100% | 100% | 100% |
| 4 | Grass | 97.2% | 98.1% | 97.7% | 100% | 99.8% |
| 5 | Trees | 98.8% | 98.8% | 98.8% | 95.4% | 97.8% |
| 6 | Water | 99.9% | 99.9% | 99.9% | 99.8% | 99.8% |
| 7 | Shadow | 82.6% | 79.1% | 84.9% | 89.5% | 89.8% |
| overall accuracy | | 78.6% | 80.7% | 81% | 84.1% | 89.5% |
| Kappa | | 0.717 | 0.744 | 0.745 | 0.787 | 0.853 |

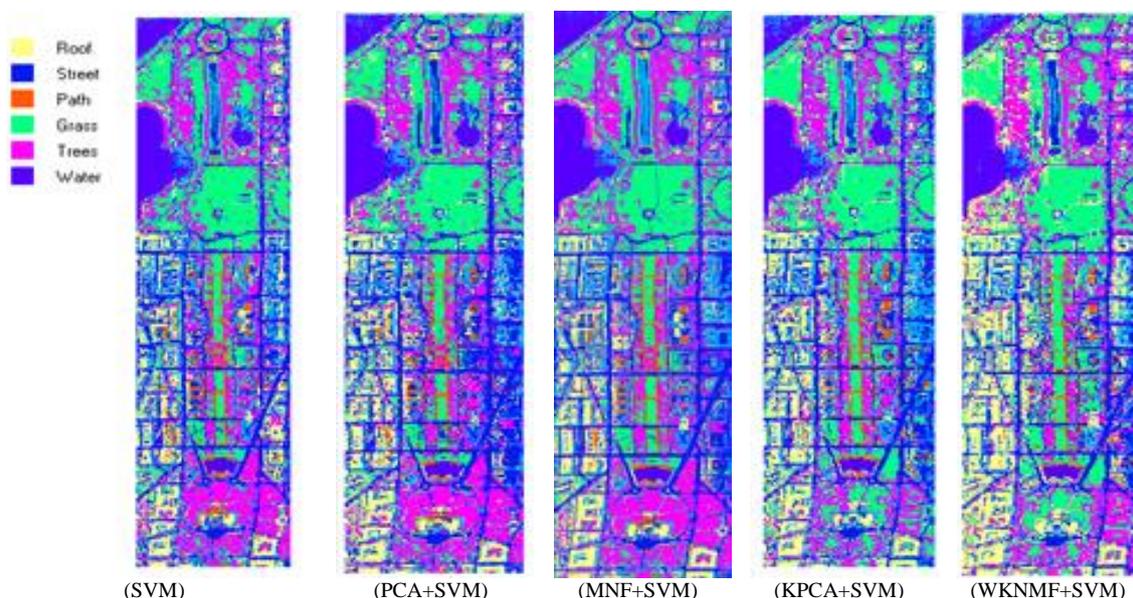


FIGURE 5 Classification images of different algorithms on HYDICE data set

The overall classification accuracy of test samples show that the use of a few training samples, WKNMF method can achieve higher classification accuracy on HYDICE data set. Compared with the other algorithms WKNMF can improve the overall classification accuracy over 5%-10%.

4 Conclusions

In this paper, we propose a feature extraction of hyperspectral images by using WKNMF. The idea of using WKNMF techniques to find a set of basic functions to represent image data where the basic functions enable the identification and classification of intrinsic "parts" that make up the object being imaged by multiple observations. Experimental results on AVIRIS 220 bands

data set in the Indian pine test site and HYDICE data sets in Washington DC Mall are both show that the proposed method achieved more strong analysis capability than comparative algorithms. Compared with the PCA, MNF and Poly-KPCA method, classification accuracy can be increased over 5%-10%. The WKNMF balance algorithm efficiency and performance very well.

Acknowledgments

The project was supported by national natural science foundation (No.41101357) and the special fund for basic scientific research of central colleges, Chang'an University, number CHD2011JC170 and CHD2011TD018.

References

- [1] Muñoz-Marí J, Tuia D, Camps-Valls G 2012 Semisupervised classification of remote sensing images with active queries *IEEE Trans. Geosci. Remote Sensing* **50**(10) 3751–63
- [2] Bioucas-Dias J M, Plaza A, Scheunders C-V G, Nasrabadi P, Chanussot N M 2013 Hyperspectral Remote Sensing Data Analysis and Future Challenges *Geoscience and Remote Sensing Magazine IEEE* **1**(2) 6-36
- [3] Tuia D, Volpi M, Copa L, Kanevski M, Muñoz-Marí J 2011 A survey of active learning algorithms for supervised remote sensing image classification *IEEE J. Select. Topics Signal Processing* **5**(3) 606–17
- [4] Dopido, I.Villa, A., Plaza, A., Gamba, P. 2012 A Quantitative and Comparative Assessment of Unmixing-Based Feature Extraction Techniques for Hyperspectral Image Classification *IEEE Journal of selected topics in applied earth observations and remote sensing* **5**(2), pp 421-435
- [5] José M. Bioucas-Dias, Antonio Plaza, G. Camps-Valls, Paul Scheunders, Nasser Nasrabadi, and Jocelyn Chanussot 2013 Hyperspectral Remote Sensing Data Analysis and Future Challenges *IEEE Geoscience and Remote Sensing Magazine* **1**(2) 6-36
- [6] Liu F, Gong J Y 2009 A classification method for high spatial resolution remotely sensed image based on multi-feature *Geography and Geo-Information Science* **25**(3) 19-41
- [7] Su H, Yang H, Du Q, Sheng Y H 2011 Semisupervised band clustering for dimensionality reduction of hyperspectral imagery *IEEE Geoscience and Remote Sensing Letters* **8**(6) 1135-9
- [8] Sen Jia, Yuntao Qian 2009 Constrained Nonnegative Matrix Factorization for Hyperspectral Unmixing *IEEE Transactions on Geoscience and Remote Sensing* **47**(1) 161-73
- [9] Xuesong Liu, Wei Xia, Bin Wang, Liming Zhang 2011 An approach based on constrained nonnegative matrix factorization to unmix hyperspectral data *IEEE Transactions on Geoscience and Remote Sensing* **49**(2) 757-72
- [10] Song Xiangfa, Jiao Licheng 2012 Classification of hyperspectral remote sensing image based on sparse representation and spectral information *Journal of Electronics & Information Technology* **34**(2) 268-73
- [11] Fauvel M, Chanussot J, Benediktsson J A 2012 A Spatial spectral Kernel based Approach for the Classification of Remote sensing Images *Pattern Recognition* **45**(1) 381-92
- [12] Chen Yi, Nasser M, Trac D T 2013 Hyperspectral Image Classification via Kernel Sparse Representation *IEEE Transactions on Geoscience and Remote Sensing* **51**(1) 217-31
- [13] Wang J, Jabbar M A 2012 Multiple Kernel Learning for adaptive graph regularized nonnegative matrix factorization *national association of science and technology for development* **41**(3) 115-22
- [14] Yushi Chen, Xing Zhao, Zhouhan Lin 2014 Optimizing Subspace SVM Ensemble for Hyperspectral Imagery Classification *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **7**(4) 1295-305
- [15] Bor-Chen Kuo, Hsin-Hua Ho, Cheng-Hsuan Li, Chih-Cheng, Hung A 2014 Kernel-Based Feature Selection Method for SVM With RBF Kernel for Hyperspectral Image Classification *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **7**(1) 317-26

Authors



Lin Bai, born on August 16, 1981, Shannxi, PR.China

Current position, grades: Lecturer

University studies: 1999/09-2003/06 Northwestern University gets BS. degree. Major: Electronic Information Science and Technology. 2006/09-2006/03 Xidian University gets MS. degree. Major: Circuits and Systems.

2006/03-2011/07 Northwestern Polytechnical University get Ph.D. degree. Major: Signal and information processing.

Scientific interest: Signal Processing, Image Processing, Remote sensing information processing.

Publications: The author has published over 10 papers including 6 in EI index and 1 in SCI index. **Experience:** 2011/07- Chang'an University, school of Electronic and Control Engineering.



Meng Hui, born on October 19, 1981, Shannxi, PR. China

Current position: Lecturer

University studies: 2000/09-2004/07 Xi'an JiaoTong University gets BS. degree. Major: Mechanical Engineering.

2004/09-2007/01 Xi'an JiaoTong University gets MS. degree. Major: Mechanical Engineering.

2006/03-2011/07 Xi'an JiaoTong University get Ph.D. degree. Major: Electrical Engineering.

Scientific interest: Nonlinear system, Chaos, Electrical Engineering.

Publications: The authors has published 14 papers including 6 papers in SCI index. **Experience:** 2012/04 - Chang'an University, school of Electronic and Control Engineering.

The optimal promised quality defect model for service guarantees

Wenlong Wang^{1, 2*}, Xinmei Liu^{1, 2}, Xiaojie Zhang^{1, 2}

¹School of Management, Xi'an Jiaotong University, 28 Xiannin Road, Xi'an Shaanxi, China

²The key lab of the Ministry of Education for Process Control & Efficiency Engineering, 28 Xiannin Road, Xian Shaanxi, China

Received 6 October 2013, www.tsi.lv

Abstract

Service quality guarantee is an important tool for firms to boost demands, put up prices, and enhance profits. However, when promised quality defect is too high or low, the impact on the organization and the customer is usually negative. Therefore, determining the level of promised quality defect is of critical strategic and tactical importance in businesses. Yet, systematic quantitative methods aren't found to help managers determine promised quality defect. We propose a simple but powerful model in finding the optimal promised service quality defect. The model makes trade-offs between benefits and costs of service defect guarantees. Firstly, the decision of promised quality defect is analysed when service price is exogenous. We secondly investigated when service price is endogenous, how can a service provider make decisions on service price and promised quality defect simultaneously to maximize its profit. Thirdly, comprehensive analysis of how service providers promise the optimal quality defect from two aspects of demand and supply is given. Numerical analysis is conducted to illustrate the interactive effect of endogenous service price and affected service supply. In the end, we conclude the paper and suggest areas for future research. With only definitional changes, the model can be applied to other guarantee contexts.

Keywords: quality guarantees; promised quality defect; service providers; affected service supply

1 Introduction

Service quality has received considerable attention in the rapidly developing service economy. Primarily due to the intangibility in the process of service production and consumption, firms have tended to adopt certain kind of service quality guarantee policy as the differentiation strategy to attract customers. A service quality guarantee policy assures the customer that during the transaction, if the actual service quality defect exceeds the promised level, the firm will have to bear the cost of service failure, such as compensation to customers, service recovery, and loss of goodwill.

Several studies indicate that service guarantees are a signal of quality and that customers follow this signal to judge product quality [1-3]. Similarly, many researchers argue that service guarantees decrease the perceived risk of customers [4]. As a result, the demand for the service will be increased. A great number of companies, especially those that are service-oriented, adopt this strategy to provide service in accordance with the service quality guarantees that they make in advance. The service quality guarantee policy has efficiently promoted service demand, sales and reputation, and helped those firms win customers. However, when the promised quality defect exceeded, the service provider incurs a substantial cost. Thus, how to make quality guarantee policies to balance the potential profits and costs is an issue for companies.

According to ref [5], a typical service guarantee policy includes two elements: a meaningful promise of a certain service quality defect and a compensation or pay-out offer. The extant literatures [6-8] primarily focus on the compensation for quality defect, providing little insight on the promised quality defect. For example, comparative static analysis was employed to derive optimal decisions of service price and compensation cost for quality defect [6], leaving promised quality defect being ignored, while defect commitment for service quality is the foundation for compensation. Only when clear defect commitment is determined, can service providers compensate to customers for excessive quality defects. Thus, the promised quality defect has to be taken into account when providers make compensation in their guarantee policy.

In addition, the effect of quality guarantees on service demand is thoroughly analysed in extant literature, yet, without considering that on service supply [6, 9, 10]. Specifically, when the promised quality defect is quite low, providers will have to bear more risk on quality defect. In order to mitigate such risks, service output will be reduced inevitably, making it difficult to meet customers' service demand. Conversely, in high promised quality defect condition, less risk will lead providers to supply more service to customers, exceeding actual market demand. Therefore, when making service quality guarantee policies, service providers have to consider the

* Corresponding author e-mail: wwl.1986@stu.xjtu.edu.cn

impact of quality defect commitment on their supply capacities.

In a word, promised quality defect, as a part of the quality guarantee policy, has not been systematically studied in prior research. Thus, this paper tends to analyse the optimal decision of promised quality defect. In the next section, we present a review of the existing literature on service quality guarantees. Then, we explain the conceptual model of the problem, identifying and rationalizing service demand and cost of quality defect. In section 4 a model of service quality guarantee is developed and analysed. In this section the decision of promised quality defect is analysed when service price is exogenous firstly. This section secondly investigated when service price is endogenous, how can a service provider make decisions on service price and promised quality defect simultaneously to maximize its profit, namely the joint optimal decision on service price and promised quality defect. Thirdly, in a competition-intense market where the service price can be taken as fixed, comprehensive analysis of how service providers promise the optimal quality defect from two aspects of demand and supply is given. In other words, when making the optimal decision on promised quality defect, the service provider will in advance take into account the impact of promised quality defect on its service supply, which will influence its profit by its relative deficiency or excess to the demand. In section 5 numerical analysis is conducted to illustrate the interactive effect of endogenous service price and affected service supply. Finally, Section 7 concludes the paper and suggests areas for future research.

2 Literature review

Service quality guarantee is an extension of product warranties, and it primarily can reduce the risk perceived by customers [11]. "Service failure" will occur when the service fail to meet the promise provided in a service guarantee program and then service remedy for the customer is needed according to the guarantee [4, 12]. Service quality guarantee issues have been continuously been studied in recent years. The adopted setting in these studies is a single enterprise that provides a service quality guarantee to customers, with various methods, including theoretical model analysis [13, 14], experimental design [15, 16], and industry investigation [17, 18].

For instance, ref [15] employed a before-after experimental design with a role-playing approach to investigate the impact of a service guarantee on an outstanding versus a good service provider in the hotel industry. This research indicates that an explicit service guarantee does not negatively affect the outstanding service provider, and the impact on the good service provider is more significant than that on the outstanding service provider. With a conceptual model, ref [19] empirically examined the effects of service guarantees. They found that service reliability is customers' primary

interest and coming to the second is the interest in compensation for service failures. Their findings provide support to the idea that including service process evidence can lead to significantly increased customers' willingness to purchase from the service provider. Furthermore, when service process evidence is listed with detail in the service quality guarantee, the compensation is more persuasive. Ref [13] developed a framework of service guarantee strength, in which they posited that high service guarantee effort can improve service quality, customer satisfaction, and customer loyalty. Ref [6] generalized existing blanket delivery-time guarantee models by drawing on concepts from other field. They relaxed simplifying assumptions to provide a comprehensive and practical model, and found that pricing policies are less critical than previously thought when the payment made for late delivery is included as part of the delivery-time guarantee policy. Ref [14] proposed a resource allocation and pricing mechanism for a service system that is subject to a class-dependent quality of service (QoS) guarantee. They suggested that the pricing scheme with QoS guarantee depends on the scheduling policy implemented and is characteristically different from that without the QoS guarantee. Ref [18] also empirically found that the type of service guarantee can significantly influence customers' perceived quality and perceived risk. Ref [10] studied the quality decisions of the functional logistics service provider (FLSP) and the logistics service integrator (LSI) with a service quality defect guarantee promised by the FLSP. The optimal quality defect guarantee of the FLSP and the optimal quality supervision effort of the LSI are presented fewer than three typical game modes: Nash game, Stackelberg game, and centralized decision. Ref [7] proposed a quantitative model, the Economic Pay-out Model for Service Guarantees (EPMSG), for determining the optimal pay-out level for the service industry. Based on ref [7], ref [8] took a service guarantee level into consideration to obtain the optimal pay-out. They considered a generic model to provide insights into the dynamic interaction between the service guarantee and optimal pay-out levels.

It can be seen that previous research mainly focused on the positive effect of a service quality guarantee policy, such as customers' satisfaction and loyalty. Promised quality defect has occasionally been mentioned to some extent in modelling research. For example, ref [10] did not systematically analyse promised quality defect (e.g., how to make decisions on quality defect when service price is endogenous and when the decisions can affect service supply) although they noticed the issue of promised quality defect in quality decisions.

To date, there are few researches focusing on optimizing the promised quality defect. The most relevant researches are ref [6] and ref [8]. However, ref [6] discussed the optimal quality guarantee policy from the demand perspective without considering the impact of promised quality defect on service supply capacity.

Although ref [8] included quality commitment in optimal compensation for the quality defect, they viewed the actual service quality as exogenous, ignoring the nature of randomness in service production and delivery. Thus, this paper aims to contribute in two ways. Firstly, the actual service quality defect is taken as a random variable in order to better describe the real business practice, and then we try to solve how to make joint optimal decision of service price and promised quality defect when service price is endogenous. Secondly, the impact of promised quality defect on supply is included in the model. In other words, the optimal promised quality defect is decided combining the influence of demand and supply.

3 The conceptual model

Consider the situation where a service provider that has adopted a quality guarantee policy supplies service to the market. The service provider guarantees a quality defect level q . On one hand, when the actual quality defect X is greater than q , the provider will have to bear the defect cost, including the compensation for the customer and the reputation loss resulting from higher quality defect than its promised level. On the other hand, promised quality defect will also enhance service demand by attracting more customers through less risk perceived by customers. In brief, how to balance the revenue increased from boomed demand and cost from taking the risk of quality defect when designing a service quality guarantee policy is the primary concern for the service provider.

3.1 SERVICE DEMAND

Ref [20] used the exponential function to depict the relationship between service demand and quality guarantees. Ref [10] also adopted the exponential function when studying the quality guarantee policy in supply chain. However, their service demand function only includes promised quality defect while ignored the effect of service price, which is one of the main factors when customers purchase services, on service demand. Therefore, besides promised quality defect, this paper also considers the impact of price on service demand. The function of service demand expresses as $D(p, q) = \eta e^{-\varepsilon p - wq}$, where η signifies the total service demand, p is the service price, and ε and w are elasticity of service price and promised quality defect respectively.

3.2 COST OF QUALITY DEFECT

The production and delivery of service is randomly influenced by some factors such as adverse weather and mistakes of front-line service staff. Consequently, service received by customers is virtually unreliable. The quality

defect cost will be incurred if the actual service quality defect X exceeds the promised quality defect q .

Explicit and implicit costs are supposed to be both included in the cost of quality defect. The explicit cost mainly refers to the payment to customers when the actual quality defect exceeds q . Of course, if there is a specific payment that is being made, these are appropriate measures. In addition to those payments, unrecorded or "hidden" quality costs such as customer dissatisfaction and loss due to bad reputation should also be included as part of the defect cost; these types of costs are generally not part of current accounting systems [21, 22] and must be incorporated separately. Also, a firm may make a quality promise without a guaranteed monetary payment; still, customer dissatisfaction is an indirect cost if the promise is not met. However, since the implicit quality defect cost is difficult to measure and obtain, we merely focus on the explicit one similar to the studies of ref [6].

The cost of exceeding guaranteed quality defect has previously been modelled in two ways: (1) using a fixed payment to the customer regardless of how severe the quality defect is, and (2) using a payment that is a function of the degree of quality defect, which is the difference between the actual quality defect and promised quality defect. In general, the latter way, namely variable defect cost, can not only help service providers to improve their service, but also better remedy the reputation loss by compensating customers who suffered from service quality defect. Given this, the quality defect cost in this paper is in accordance with the variable quality defect cost. The product quality defect literature has embraced the well-known quadratic loss function as an appropriate measure of the second type of quality defect cost. Ref [23] analysed the rationality of the adoption of quadratic function when the actual quality defect exceeds the promised quality defect. Thus, based on the work of ref [23], the function of quality defect cost in this paper is $C(q) = c \int_q^{\infty} (X - q)^2 f(X) dX$, where $(X - q)$ signifies the degree of quality defect, $f(X)$ is its probability density function, and c is the unit cost for quality defect.

Without loss of generality, it is assumed that there is no fixed production cost and the variable cost per unit is v ($v < p$). The profit function of a service provider is as following.

$$\begin{aligned} \Pi(p, q) &= D(p, q) [p - v - C(q)] \\ &= \eta e^{-\varepsilon p - wq} \left[p - v - c \int_q^{\infty} (X - q)^2 f(X) dX \right] \end{aligned} \quad (1)$$

4 The analytic model

4.1 WITH EXDOGENENOUS SERVICE PRICE

We begin by analysing the promised quality defect q of a service provider in this section, where q_b^* signifies the

optimal promised quality defect, and subscript b is the benchmark. Suppose that service price p is exogenous. The service provider tries to maximize its profit by promising the level of quality defect q .

Consider $\Pi_b(q)$ signifies the profit function of the service provider. As p is constant, the optimal promised quality defect q_b^* can be derived from the first order condition (FOC) of $\Pi_b(q)$. Then, the concavity of $\Pi_b(q)$ can be obtained from the second order condition (SOC) of $\Pi_b(q)$. The function expression of $f(X)$ is needed for FOC and SOC. Drawn on the work of ref [10], assume that X is a random variable with the exponential distribution, and the mean is $\frac{1}{\lambda}$.

Proposition 1 There exists one and only one optimal q_b^* that maximizes the service provider's profit function $\Pi_b(q)$. Specifically, $\Pi_b(q)$ increases concavely in $(0, q_b^*)$, decreases concavely in $[q_b^*, q_b^*]$, and decreases convexly in (q_b^*, ∞) , where $q_b^* = -\frac{1}{\lambda} \ln \frac{(p-v)w\lambda^2}{2c(w+\lambda)}$,

$$q_b^* = -\frac{1}{\lambda} \ln \frac{(p-v)w\lambda^2}{2c(w+\lambda)^2}.$$

Proof in Appendix.

From Proposition 1, it can be seen that when service price p is exogenous, there is an optimal promised quality defect for the service provider. To be specific, $\Pi_b(q)$ is concave-convex in q , with the inflection point of q_b^* , and the unique optimal value is q_b^* . That is to say, when faced with fixed service price p , the provider can maximize its profit by promising quality defect q_b^* .

There are two effects of the promised quality defect on the profit of the service provider. The first is called commitment-demand effect, which is the negative effect of the promised quality defect on the provider's profit via service demand. The second one is commitment-marginal-profit effect, referring to the positive effect of promised quality defect on provider's profit through the marginal profit of the service. In $(0, q_b^*)$, the commitment-demand effect is weaker than the commitment-marginal-profit effect, leading to a continually increased service profit. However, the commitment-demand effect becomes stronger than the commitment-marginal-profit effect in $[q_b^*, +\infty)$. Thus, the service profit in $[q_b^*, +\infty)$ is decreasing. At the critical point q_b^* where the two kinds of effects reach a balance, the service provider can maximize its profit.

4.2 WITH ENDOGENOUS SERVICE PRICE

Service price is assumed to be endogenous in this section. Joint optimal decisions on service price and promised quality defect need to be made. In other words, optimal service price p_p^* and optimal promised quality defect q_p^* to maximize the provider's profit are derived simultaneously, where the subscript p signifies endogenous service price. Since the provider's profit function $\Pi_p(p, q)$ is not jointly concave in p and q , it is impossible to find the optimal joint decision by negative definite Hessian Matrix. But the two-stage optimization method can solve the joint decision problem of service price and promised quality defect [24, 25]. In the first stage, promised quality defect is assumed to be constant, and then the optimal service price function of promised quality defect $p^*(q)$ can be obtained through the FOC of $\Pi_p(p, q)$ to service price. In stage 2, substituting $p^*(q)$ into the original profit function $\Pi_p(p, q)$, we can derive a new profit function $\Pi_p(p^*(q), q)$. If the new profit function $\Pi_p(p^*(q), q)$ has a maximum in q , then the original profit function $\Pi_p(p, q)$ can also be maximized in p and q , which is the optimal joint decision on service price p and promised quality defect q .

4.2.1 The optimal service price function of promised quality defect

In this section, the optimal service price p^* is obtained when promised quality defect q is given. It means that when certain quality guarantee policy is adopted, quality defect can be considered as exogenous. Thus, the problem for the service provider is how to pricing the service to maximize its profit.

As promised quality defect is exogenous, the profit function of the service provider is $\Pi_p(p)$. The optimal service price can be obtained through the FOC of $\Pi_p(p)$. Then the SOC will show concavity of the profit function $\Pi_p(p)$.

Lemma 1 Given the promised quality defect q , the unique optimal service price is p^* , which can maximize the provider's profit function $\Pi_p(p)$. Specifically, the provider's profit function $\Pi_p(p)$ increases in $(0, p^*)$, decreases convexly in $[p^*, p^*]$, and decreases concavely in (p^*, ∞) , where $p^* = \frac{1}{\varepsilon} + v + c \int_q^\infty (X - q)^2 f(X) dX$, $p' = \frac{2}{\varepsilon} + v + c \int_q^\infty (X - q)^2 f(X) dX$.

Proof in Appendix.

Lemma 1 shows that when the promised quality defect is given, there exists an optimal pricing policy. To be specific, the service provider's profit function $\Pi_p(p)$ is convex-concave in the service price p with the inflection point of p' , and the unique maximum value is p^* . That is to say, when faced with fixed promised quality defect, the service provider can maximize its profit by pricing at p^* .

As for the service price, there are also two kinds of effects on the service profit. One can be called price-demand effect, which is the negative effect of the service price on provider's profit via service demand. The other is price-marginal-profit effect, referring to the positive effect of the service price on the provider's profit through the marginal profit of the service. In $(0, p^*)$, the price-demand effect is weaker than the price-marginal-profit effect, leading to a continually increased service profit. However, in $[p^*, +\infty)$, the price-demand effect becomes stronger than the price-marginal-profit effect. Thus, the service profit in $[p^*, +\infty)$ is decreasing. At the critical point p^* where the two effects reach a balance, the service provider can maximize its profit.

4.2.2 The optimal quality defect for profit function

Substituting $p^* = \frac{1}{\varepsilon} + v + c \int_q^\infty (X - q)^2 f(X) dX$ into the service provider's original profit function $\Pi_p(p, q)$, the new profit function is expressed as $\Pi_p(p^*(q), q)$. The monotonicity and concavity of the profit function $\Pi_p(p^*(q), q)$ in promised quality defect q_p can be obtained from the FOC and SOC. Consequently, the optimal promised quality defect q_p^* can be derived.

Lemma 2 when $p^* = \frac{1}{\varepsilon} + v + c \int_q^\infty (X - q)^2 f(X) dX$, there exists one and only one optimal promise quality defect q_p^* that can maximize the provider's profit function $\Pi_p(p^*(q), q)$. When $\frac{1}{\lambda} \geq \frac{w}{2c\varepsilon}$, $q_p^* = \frac{1}{\lambda} \ln \frac{2c\varepsilon}{w\lambda}$ and when $\frac{1}{\lambda} < \frac{w}{2c\varepsilon}$, $q_p^* = 0$.

Proof in Appendix.

Although the provider's profit function $\Pi_p(p^*(q), q)$ is not concave in promised quality defect, there always exists one and only one maximum value q_p^* by analysing the monotonicity of the profit function. Since $\frac{1}{\lambda}$ is the expected value of actual service quality, the provider can

make optimal promised quality defect according to its actual service quality. Specifically, increased demand brought by decreased quality defect, will have a positive effect on the provider's profit at an increasing rate; however, increased cost from less quality defect will have a negative effect on the provider's profit at an increasing rate. The net effect determines how the provider will act. When $\frac{1}{\lambda} < \frac{w}{2c\varepsilon}$, the negative effect of cost on profit is far weaker than the positive effect of demand on profit due to the provider's high qualified service. Thus, if the actual quality defect is at a low level, the policy of promised quality defect is beneficial for the provider's profit. That is to say, promising zero quality defect is the optimal choice for the provider in this condition. However, when $\frac{1}{\lambda} \geq \frac{w}{2c\varepsilon}$, the negative effect of cost on profit is much stronger than the positive effect of demand on profit due to the provider's poor service quality. Hence, if the actual quality defect increased to a high level, the effect of promised quality defect on the provider's profit is changing from positive to negative. It means that promising appropriate quality defect is the optimal choice for the provider in this condition.

Based on the two-stage optimization method, Proposition 2 is obtained combining Lemma 1 and Lemma 2.

Proposition 2 There exists one and only one joint optimal p and q that can maximize the provider's profit function $\Pi_p(p, q)$, where the optimal service price is $p^* = \frac{1}{\varepsilon} + v + \frac{w}{\lambda\varepsilon}$ and the optimal promised quality defect is $q_p^* = -\frac{1}{\lambda} \ln \frac{w\lambda}{2c\varepsilon}$.

Proposition 2 demonstrates that although the provider's profit function $\Pi_p(p, q)$ is not jointly concave in p and q , the two-stage optimization method can help to solve the joint decision problem of service price p , and promised quality defect q . Proposition 2 also exhibits that quality guarantee is a two-dimensional strategy. Service price and promised quality defect both need to be taken into consideration when the provider adopts quality guarantee as a differentiation strategy to compete in the market. Neither the service price nor the promised quality defect alone can maximize the provider's profit. Making quality guarantee maybe incur increased cost for quality defect to some extent, but also can boost demand from market due to promised quality defect. Thus, the provider can realize its maximized profit through joint optimal decision of service price and promised quality defect.

To make it more visualized, numerical analysis by MAPLE 17 software to verify the validity of proposition 2 is shown in Figure 1. Assume that $\eta = 20000$, $\varepsilon = 0.8$, $w = 0.2$, $c = 0.5$, $v = 1$, $\lambda = 1$.

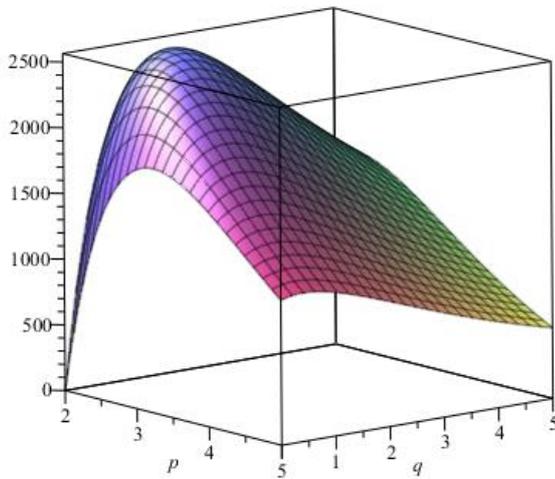


FIGURE 1 Profit function of the service provider on the service price p and promised quality defect q

From Figure 1, it can be seen that the profit function $\Pi_p(p, q)$ is jointly quasi-concave on the service price p and promised quality defect q . There is a unique $p^* = 2.5$ and unique $q^* = 1.386$ that can maximize the provider's profit, which is $\Pi_p(p^*, q^*) = 2564$.

4.3 WITH AFFECTED SERVICE SUPPLY

Promised quality defect can affect the service supply for the following two reasons, occupancy of resources for service production and the avoidance of quality risks. Firstly, when the promised quality defect is at a low level, part of resources will be used to improve and maintain high service quality, such as training programs for front-line employees and procedure improvement; otherwise, these resources should have been used to expand the production scale. Secondly, the lower the promised quality defect is, the more the risk is inherent. From the view of intrinsic preference to mitigate risks, the provider will supply less service to the market. Thus, the supply capacity of the provider is more limited when the service is of low promised quality defect than when that of high promised quality defect. The supply S increases monotonically with q , $\frac{\partial S}{\partial q} > 0$, at an increasing rate,

$\frac{\partial^2 S}{\partial q^2} > 0$ based on theory of increasing marginal cost. For the purpose of simplicity and consistency (with the early mentioned form of service demand function), the supply function is expressed as $S(q) = \xi e^{\mu q}$, where ξ is the total service supply of the provider, $e^{\mu q}$ is the proportion of service volume to the provider's total supply, and μ is

the elasticity of the provider supply to the quality defect guarantee. On this occasion, the supply does not necessarily have to satisfy the demand from the market. The effective supply is influenced by the provider's promised quality defect. On one hand, although the service with low promised defect quality is very much in demand, the actual supply is quite small due to the high requirement of the service for the provider. It means that the effective supply is the actual supply regardless of the demand. On the other hand, however, the supply of high promised defect quality will be great while the demand is rather small. That is to say, the supply that exceeds the demand is meaningless for customers. In this case, the effective supply is the demand. Thus, the effective supply can be expressed as $\min(D(q), S(q))$. Drawn on the model in Section 3, the provider's profit function can be expressed as $\Pi_s(q) = \min(D(q), S(q))[p - v - C(q)]$, where the subscript S represents the effective service supply.

The demarcation point $q^\# = \frac{\ln(\frac{\eta}{\xi}) - \varepsilon p}{\mu + w}$ can be derived from $D(q) = S(q)$. When the promised quality defect is less than the demarcation point, $q \leq q^\#$, service supply is less than its demand. The new profit function of the service provider now is $\Pi_s(q) = S(q)[p - v - C(q)]$, called the supply profit function. Otherwise, the new profit function is $\Pi_s(q) = D(q)[p - v - C(q)]$, called the demand profit function, when $q > q^\#$.

Thus, the optimal promised quality defect will be identified by whether the supply is greater than the demand. Local optimum is firstly derived in order to obtain the global optimal in the final step.

4.3.1 Supply is less than demand

The derivation analysis of the provider's profit function is used in this section to figure out the local optimal decision in the condition of $q \leq q^\#$.

Lemma 3 When the supply is less than the demand, $q \leq q^\#$, there is an optimum in $[0, q^\#]$, which is $q_s^* = q^\#$, that can maximized the provider's profit function $\Pi_s(q)$.

Specifically, If the actual quality defect is low ($\frac{1}{\lambda} \leq \frac{1}{\mu}$), the provider's profit function monotonically increases with promised quality defect. If the actual quality defect is high ($\frac{1}{\lambda} > \frac{1}{\mu}$), the provider's profit first decreases and then increases with the increase of promised quality defect.

Lemma 3 shows the condition where the service supply is less than the demand caused by the low level

promised quality defect made by the service provider. If the actual quality defect is low ($\frac{1}{\lambda} \leq \frac{1}{\mu}$), the provider's profit function monotonically increases with promised quality defect. In this case, the marginal profit $p - v - C(q)$ is positive because the quality defect cost of the provider $C(q)$ is quite small when the actual quality defect is at a low level. Therefore, the service provider will continually enlarge its supply because the marginal profit increases with the promised quality defect. If the actual quality defect is high ($\frac{1}{\lambda} \leq \frac{1}{\mu}$), the provider's profit first decreases and then increases with the increase of promised quality defect. The reason is that When the promised quality defect is rather low ($q < q_s^*$, and $q_s^* = \frac{1}{\lambda} \ln \frac{2c(\mu - \lambda)}{\mu \lambda^2 (p - v)}$), the quality defect cost for the provider $C(q)$ is rather great, or even greater than the marginal revenue $p - v$, resulting in a negative marginal profit. With the increase of the promised quality defect, the supply $S(q)$ increases while the marginal profit decreases. That is to say, the more the service provider supplies to the market, the more it will lose. When the promised quality defect is rather high ($q \geq q^c$), the quality defect cost of the provider is far less than the marginal revenue, resulting in a positive marginal profit. With the increase of the promised quality defect, the supply and the marginal revenue both increase. Thus, the provider's profit increases with promised quality defect.

4.3.2 Supply is more than demand

When the supply is greater than the demand, the provider's profit function is $\Pi_s(q) = D(q)[p - v - C(q)]$. From Proposition 1, it can be seen that the provider maximizes its profit when $q_b^* = -\frac{1}{\lambda} \ln \left(\frac{pw\lambda^2}{2c(w + \lambda)} \right)$. As $q > q^\#$, the relationship of magnitude between q_b^* and $q^\#$ will impact the optimal decision of the promised quality defect q . From Proposition 1, Lemma 4 is obtained.

Lemma 4 When the supply is greater than the demand, $q > q^\#$, consider two cases: if $q_b^* > q^\#$, the optimal promised quality defect is obtained when $q_s^* = q_b^*$; otherwise, the optimal solution is $q_s^* = q^\#$.

The global optimal decision can be obtained by synthetically analysing the local optimal in the two above mentioned parts.

Proposition 3 When the supply is affected by the promised quality defect, consider two cases: if $q_b^* > q^\#$,

the optimal promised quality defect is obtained when $q_s^* = q_b^*$; otherwise, the optimal solution is $q_s^* = q^\#$.

Proposition 1 in section 4.1 indicates that if, regardless of the supply, the quality defect can only affect the demand, there is an unique optimal service quality defect. However, Proposition 3 also demonstrates that taking the supply and the demand simultaneously into account, there are two cases for the provider's optimal decision on service quality defect. Specifically, if $q_b^* > q^\#$, the provider's profit function, $\Pi_s(q)$, increases convexly in $[0, q^\#]$, increases concavely in $(q^\#, q_s^*]$ and decreases concavely in $(q_s^*, +\infty)$. Thus, the provider can maximize its profit when $q = q_s^*$. Intuitively, when $q_b^* > q^\#$, the intersection of the supply profit function curve and the demand profit function curve is to the left of the maximum of the original demand profit function, which is not included in the impacted area of the provider's profit function from the supply (the actual supply profit function). In addition, in most area affected by the supply, the provider's profit is less than that when there is no affect from the supply. Thus, the optimal of the original demand profit function is the same as that with simultaneous influence from the supply and the demand (see figure 2). If $q_b^* \leq q^\#$, the provider's profit function, $\Pi_s(q)$, increases convexly in $[0, q^\#]$, and decreases concavely in $(q^\#, +\infty)$. Thus, the provider can maximize its profit when $q = q^\#$. Intuitively, when $q_b^* \leq q^\#$, the intersection of the supply profit function curve and the demand profit function curve is to the right of the maximum of the original demand profit function, which is included in the impacted area of the provider's profit function from the supply), leading to the difference between the optimal of the original demand profit function and that with simultaneous influence from the supply and the demand. Besides, the local optimal in the area affected by supply becomes the global optimal in this condition (see Figure 3).

To make it more visualized, numerical analysis by MAPLE 17 software to verify the validity of proposition 3 is shown in Figure 2 and Figure 3. Assume that $\eta = 3000000$, $\varepsilon = 0.8$, $w = 0.6$, $c = 2.5$, $v = 1$, $\lambda = 0.5$, $\mu = 1.2$, $\beta = 0.5$, $p = 4$, $\xi = 30$ or $\xi = 10$.

Figure 2 and Figure 3 provide support for Proposition 3. Figure 2 shows that if $q_b^* > q^\#$, the local optimal solution when the supply is more than the demand is better than that when the supply is less than the demand. Thus, the service provider can maximize its profit at $q_s^* = q_b^*$. The unique optimal promised quality defect is $q_s^* = 5.007$ and the maximized profit is $\Pi_s^*(q_b^*) = 8182.357$. Otherwise, as indicated in Figure 3,

if $q_b^* \leq q^\#$, the local optimal solution when the supply is less than the demand is better than that when the supply is more than the demand. Thus, the service provider can maximize its profit at $q_s^* = q^\#$. The unique optimal promised quality defect is $q_p^* = 5.223$ and the maximized profit is $\Pi_s^*(q_b^*) = 8070.553$.

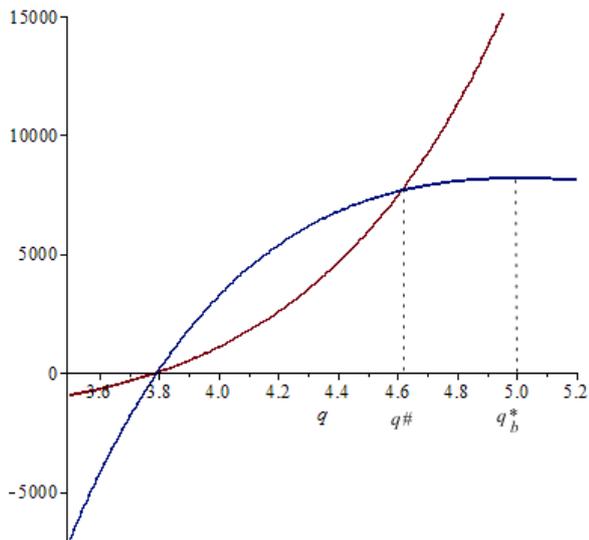


FIGURE 2 The provider's profit function of promised quality defect when $q_b^* > q^\#$

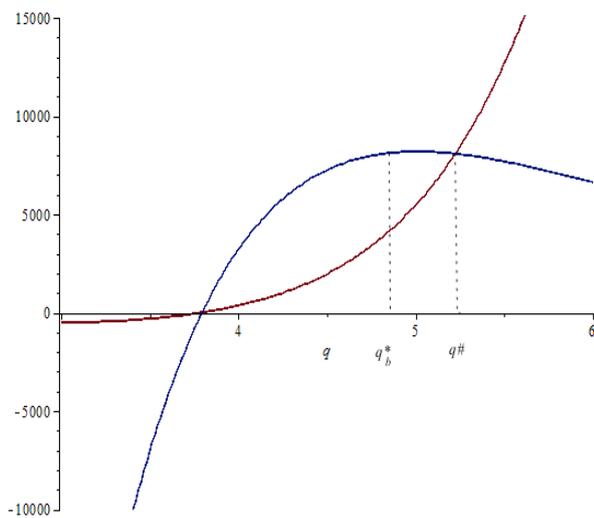


FIGURE 3 The provider's profit function of promised quality defect when $q_b^* \leq q^\#$

5 Numerical analysis

Section 4 provided the quantitative analysis of the optimal promised quality defect when service price is endogenous and when service supply is affected by the quality guarantee policy, respectively. At the same time, the case with the interaction of endogenous service price and affected service supply is also taken into consideration. However, the closed form solution cannot be mathematically derived due to the complexity of the

model. Thus, the numerical analysis is used to obtain the joint optimal decision of service price and promised quality defect when the supply is affected by the service guarantee policy.

In the similar vein with section 4.3, the supply function of the service provider is $S(p, q) = \xi e^{\mu q + hp}$, where h represents the sensitivity of supply to service price. Then, the effective supply is $\min(D(p, q), S(p, q))$. The profit function of the provider can be acquired based on the model in Section 3 as $\Pi_s(p, q) = \min(D(p, q), S(p, q)) [p - v - C(q)]$. Reasonable assignment is chosen ($\eta = 30000$, $\varepsilon = 0.8$, $w = 0.2$, $c = 0.5$, $v = 1$, $\lambda = 1$, $\xi = 5$) for the numerical analysis in order to intuitively get the joint optimal decision of service price and promised quality defect. As the sensitivity of the supply function significantly influences the joint optimal decision, the low and the high sensitivity of the supply function are considered separately.

5.1 LOW SENSITIVITY OF THE SUPPLY FUNCTION

When the sensitivity of the supply function to the service price and promised quality defect is low ($\mu = 0.6$, $h = 0.8$), the profit function $\Pi_s(p, q)$ on p and q can be drawn as following.

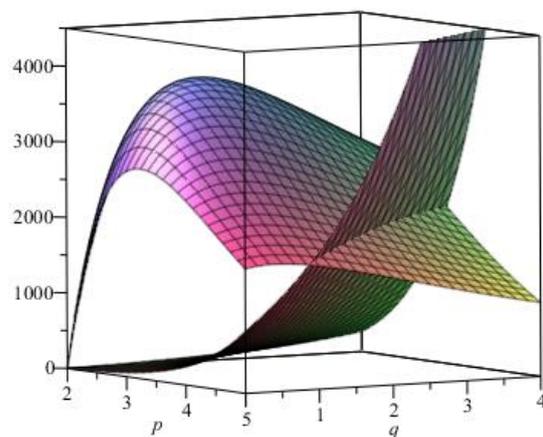


FIGURE 4 Overall profit function with low sensitivity

Figure 4 shows that on account of the affected supply, the joint optimal decision of service price and promised quality defect is codetermined by the supply profit surface and the demand profit surface. In the case where the intersection curve of the two surfaces lies in the outside of the optimal point of the demand profit surface, the optimal point is on the intersection curve. Further analysis of the intersection curve indicates that the FOC solution of profit function on p and q at the intersection

curve gives the joint optimal decision of service price and promised quality defect, $p^* = 3.4276$ and $q^* = 4.0127$. Thus, the maximized profit of the service provider is $\Pi_S(p^*, q^*) = 2076.2$.

As a result of the low sensitivity of the supply function to service price and promised quality defect, the supply rises slowly with the increase of the service price and promised quality defect, leading to an unsatisfied demand, which means that the optimum has not reached the extreme point of the demand profit surface. Only when the service demand is met, meaning on the intersection curve, the profit of the provider can be maximized.

5.2 HIGH SENSITIVITY OF THE SUPPLY FUNCTION

When the sensitivity of the supply function to service price and promised quality defect is high ($\mu = 2.5$, $h = 2.5$), the profit function $\Pi_S(p, q)$ on p and q is shown as following.

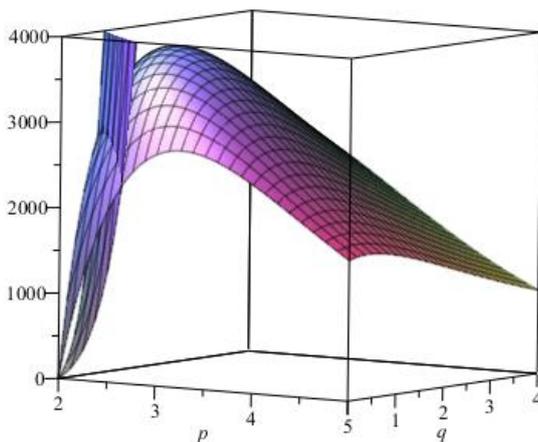


FIGURE 5 Overall profit function with high sensitivity

Figure 5 demonstrates that in comparison with the condition where there is no affected supply, the profit function of the provider has changed in this condition, yet owing to the much less vulnerability to the affected supply, the optimal point is on the demand profit surface. Although the overall profit surface changed, the intersection curve of the demand profit surface and supply profit surface is located at the inside of the optimal point of the demand profit surface. Thus, the changed part of the supply surface is still below the extreme point of the unchanged part of the demand surface, meaning that the optimum on the demand profit surface is the point that maximizes the overall profit function of the provider in this condition. The joint optimal decision is $p^* = 2.5$ and $q^* = 1.3863$. Thus, the

maximized profit of the service provider is $\Pi(p^*, q^*) = 3825.2237$.

Due to the high sensitivity of the supply function to the service price and promised quality defect, the supply raises rather quickly with the increase of the service price and promised quality defect, leading to an effective satisfied demand. Therefore, the optimum is the extreme point of the demand profit surface.

Combing the analysis in 5.1 and 5.2, it is known that when the effect of service price and promised quality defect on the supply function is quite low, the joint optimal decision of the provider is on and can be derived from the intersection curve of the supply profit surface and the demand profit surface; however, when the effect of service price and promised quality defect on the supply function is quite high, the joint optimal decision of the provider depends on the extreme point of the demand profit surface. In other words, the joint optimal decision of the provider can be obtained from the FOC of the demand profit function on the service price and promised quality defect.

6 Conclusions

Service quality guarantee is an important tool for firms to boost demands, put up prices, and enhance profits. This paper presents a simple but powerful model in finding the optimal promised service quality defect. The model makes trade-offs between benefits and costs of service defect guarantee. With only definitional changes, the model can be applied to other guarantee contexts in which the demand and supply are influenced by service guarantees and actual service defect variable follows the exponential distribution.

We adopt an analytical approach for optimal quality defect promise of a firm making quality guarantees on their service. The proposed model generalizes existing service quality guarantee models in two primary aspects:

(1) The existing literature concerning service quality guarantees mainly focuses on service price and payment made for defect. The proposed model also includes the promised quality defect as a main decision variable and incorporates it as a part of the demand function. Further, the joint decision of price and promised quality defect is discussed when the service price is endogenous.

(2) Previous studies on service quality guarantees neglect the impact of promised quality defect on service supply capacity, which is included in the proposed model to further analyse its influence on provider's profit. Consequently, the optimal decision of promised quality defect is derived in this condition. This finding will better guide service providers to make the quality guarantee policy.

In addition, numerical analysis provides an intuitive joint optimal decision on service price and promised quality defect when the supply is vulnerable to the latter. When the sensitivity of supply function is high, the affected supply can hardly change the overall profit

surface of the service provider. In this case, the joint optimal decision is the extreme point of the demand profit surface, which is consistent with joint optimal decision in 4.2 where the affected supply has no effect. However, when the sensitivity of supply function is low, the overall profit function is, to a great extent, impacted by the affected supply. Then the joint optimal decision is on the intersection curve of the demand profit surface and the supply profit surface, which is inferior to the condition in 4.2 where the affected supply has no effect.

An interesting direction of future research would be how a service provider can make promised quality defect to differentiate itself from its rivals in a competitive market. Moreover, what is the effect of industrial characteristics on service quality guarantee in some special service industry, either quite new or with greater risks (e.g. finance and information security) is merely studied. Last but not the least, from the perspective of supply chain, it is also intriguing that how the promised quality defect of a service provider can affect decisions of upstream and downstream members, the profit of the whole service supply chain, and further the coordination mechanism of the service supply chain based on quality guarantee.

Appendix

Proof of proposition 1

Because X is a random variable with the exponential distribution and the mean is $1/\lambda$, the service provider's profit function is

$$\begin{aligned} \Pi(q) &= \eta e^{-\varepsilon p - wq} \left[p - v - c \int_q^\infty (X - q)^2 f(X) dX \right] \\ &= \eta e^{-\varepsilon p - wq} \left[p - v - \frac{2ce^{-\lambda q}}{\lambda^2} \right]. \end{aligned} \tag{2}$$

Then, according to the first-order condition of the service provider's profit function, we can obtain that

$$q_b^* = -\frac{1}{\lambda} \ln \frac{(p-v)w\lambda^2}{2c(w+\lambda)}. \tag{3}$$

Then, when $q \leq q_b^*$, $\frac{d\Pi(q)}{dq} \geq 0$; and when $q > q_b^*$,

$\frac{d\Pi(q)}{dq} < 0$. So there exists a unique optimal promised quality defect q_b^* , which can maximize the profit of the service provider.

According to the second-order condition of the service provider's profit function, we can obtain that

$$q_b' = -\frac{1}{\lambda} \ln \frac{(p-v)w^2\lambda^2}{2c(w+\lambda)^2}. \tag{4}$$

Then, we have that

$$\begin{cases} \frac{d^2\Pi(q)}{dq^2} \leq 0, & \text{if } q \leq q_b' \\ \frac{d^2\Pi(q)}{dq^2} > 0, & \text{if } q > q_b' \end{cases}. \tag{5}$$

Since $\frac{w}{w+\lambda} < 1$, $\ln \frac{w}{w+\lambda} < 0$, we can obtain that

$$q_b' - q_b^* = -\frac{1}{\lambda} \ln \frac{w}{w+\lambda} > 0. \tag{6}$$

$\Pi_b(q)$ increases concavely in $(0, q_b^*)$, decreases concavely in $[q_b^*, q_b']$, and decreases convexly in (q_b', ∞) .

Proof of lemma 1

Since the service provider's profit function is

$$\begin{aligned} \Pi_p(p) &= D(p)[p - v - C(q)] \\ &= \eta e^{-\varepsilon p - wq} \left[p - v - c \int_q^\infty (X - q)^2 f(X) dX \right]. \end{aligned} \tag{7}$$

Then, according to the first-order condition of the service provider's profit function, we can obtain that

$$p^* = \frac{1}{\varepsilon} + v + c \int_q^\infty (X - q)^2 f(X) dX. \tag{8}$$

$$\text{Since } \begin{cases} \frac{d\Pi_p(p)}{dp} \geq 0, & \text{if } p \leq p^* \\ \frac{d\Pi_p(p)}{dp} < 0, & \text{if } p > p^* \end{cases}, \tag{9}$$

there exists a unique optimal service price p^* that can maximize the profit of the service provider.

$$\frac{d^2\Pi_p(p)}{dp^2} = \left[-2 + \varepsilon \left(p - v - c \int_q^\infty (X - q)^2 f(X) dX \right) \right] \eta e^{-\varepsilon p - wq}. \tag{10}$$

According to the second-order condition of $\Pi_p(p)$, we can obtain that

$$p' = \frac{2}{\varepsilon} + v + c \int_q^\infty (X - q)^2 f(X) dX. \tag{11}$$

Besides, $p' - p^* = \frac{1}{\varepsilon} > 0$.

So, the provider's profit function $\Pi_p(p)$ increases convexly in $(0, p^*)$, decreases convexly in $[p^*, p']$, and decreases concavely in (p', ∞) .

Proof of lemma 2

Because X is a random variable with the exponential distribution and the mean is $1/\lambda$, the optimal service price is $p^* = \frac{1}{\varepsilon} + v + \frac{2ce^{-\lambda q}}{\lambda^2}$.

Then, we can obtain that

$$\Pi(p^*(q), q) = \frac{\eta}{\varepsilon} e^{-1-\varepsilon v - \frac{2c\varepsilon e^{-\lambda q}}{\lambda^2} - wq} \quad (12)$$

The first derivative of service provider's profit function $\Pi(p^*(q), q)$ is

$$\frac{d\Pi(p^*(q), q)}{dq} = -\frac{\eta}{\varepsilon} e^{-1-\varepsilon v} \left(w - \frac{2c\varepsilon}{\lambda} e^{-\lambda q} \right) e^{-\frac{2c\varepsilon e^{-\lambda q}}{\lambda^2} - wq} \quad (13)$$

According to the first-order condition of $\Pi(p^*(q), q)$, we can obtain that $q^* = -\frac{1}{\lambda} \ln \frac{w\lambda}{2c\varepsilon}$.

$$\begin{cases} \frac{d\Pi(p^*(q), q)}{dq} \geq 0, & \text{if } q \leq q^* \\ \frac{d\Pi(p^*(q), q)}{dq} < 0, & \text{if } q > q^* \end{cases} \quad (14)$$

So service provider's profit function ($\Pi(p^*(q), q)$) is increasing-to-decreasing with q .

Since $q \geq 0$, when $\frac{1}{\lambda} \geq \frac{w}{2c\varepsilon}$, $q^* = \frac{1}{\lambda} \ln \frac{2c\varepsilon}{w\lambda}$; when $\frac{1}{\lambda} < \frac{w}{2c\varepsilon}$, $q^* = 0$.

Proof of lemma 3

Since supply is less than demand, the service provider's profit function is

$$\Pi_s(q) = S(q)[p-v-C(q)] = \xi e^{\mu q} \left(p-v - \frac{2ce^{-\lambda q}}{\lambda^2} \right) \quad (15)$$

Then, the first derivative of $\Pi_s(q)$ is

$$\frac{d\Pi_s(q)}{dq} = \frac{\xi e^{\mu q}}{\lambda^2} [2ce^{-\lambda q}(\lambda - \mu) + \mu\lambda^2(p-v)] \quad (16)$$

(1) Since $p-v > 0$, $2ce^{-\lambda q}(\lambda - \mu) + \mu\lambda^2(p-v) > 0$.

When $\lambda - \mu \geq 0$. Then, $\frac{d\Pi_s(q)}{dq} > 0$, which means that the service provider's profit function maximizes at $q_s^* = q^\#$.

(2) When $\lambda - \mu < 0$, according to the first order condition of $\Pi_s(q)$, we can obtain that

$$q_s^* = \frac{1}{\lambda} \ln \frac{2c(\mu - \lambda)}{\mu\lambda^2(p-v)} \quad (17)$$

When $q < q_s^*$, owing to $2ce^{-\lambda q}(\lambda - \mu) + \mu\lambda^2(p-v) < 0$, $\frac{d\Pi_s(q)}{dq} < 0$. When $q \geq q_s^*$, owing to $2ce^{-\lambda q}(\lambda - \mu) + \mu\lambda^2(p-v) \geq 0$, $\frac{d\Pi_s(q)}{dq} \geq 0$. So the service provider's profit function is decreasing in $[0, q_s^*)$, and increasing in $[q_s^*, q^\#]$. Since

$$\begin{cases} q^\# > 0 \\ \Pi_s(q=0) = \xi \left(p-v - \frac{2c}{\lambda^2} \right) \\ \Pi_s(q=q^\#) = \xi e^{\mu q^\#} \left(p-v - \frac{2ce^{-\lambda q^\#}}{\lambda^2} \right) \end{cases} \quad (18)$$

and $\Pi_s(q=0) < \Pi_s(q=q^\#)$, the service provider's profit function maximizes at $q_s^* = q^\#$.

Combining (1) and (2), it can be inferred that when the supply is less than the demand, $q \leq q^\#$, there is an optimum in $[0, q^\#]$, which is $q_s^* = q^\#$, that can maximize the provider's profit function $\Pi_s(q)$. Specifically, If the actual quality defect is low ($\frac{1}{\lambda} \leq \frac{1}{\mu}$), the provider's profit function monotonically increases with promised quality defect. If the actual quality defect is high ($\frac{1}{\lambda} > \frac{1}{\mu}$), the provider's profit first decreases and then increases with the increase of promised quality defect.

References

- [1] Jens H, Dwayne D G 2009 Twenty years of service guarantee research a synthesis *Journal of Service Research* **11**(4) 322-43
- [2] Steven A C, Daniel C Q 2008 Exotic reservations-Low-price guarantees *International Journal of Hospitality Management* **27**(2) 162-69
- [3] Jehn Y W, Sheng H T, Chih H W 2009 *The Service Industries Journal* **29**(9) 1261-72
- [4] Sara B L, Per S 2003 *International Journal of Service Industry Management* **14**(1) 36-58
- [5] Rajiv K 2001 The effects of service guarantees on external and internal markets *Academy of Marketing Science Review* **5**(10) 1-19
- [6] Timothy L U 2009 *European Journal of Operational Research* **196**(3) 959-67
- [7] Tim B, David A C 2005 *Decision Sciences* **36**(2) 197-220
- [8] Wen-Chyuan Chiang, Gangshu Cai, Xiaojing Xu, Xiangfeng Chen 2013 Service guarantee and optimal payout models *International Journal of Production Economics* **141**(2) 519-28 (In chinese)
- [9] Kut C S, Jing S S 1998 *European Journal of operational research* **111**(1) 28-49
- [10] Liu W H, Xie D 2013 *International Journal of Production Research* **51**(5) 1618-34
- [11] Gordon H M, Terrence L, Peter V 1998 Designing the service guarantee: unconditional or specific *Journal of Services Marketing* **12**(4) 278-93
- [12] Christopher W H, Leonard A S, Dan M 1991 Guarantees come to professional service firms *Sloan Management Review* **33**(3) 19-20
- [13] Julie M H, Arthur V H 2006 *Journal of Operations Management* **24**(6) 753-64
- [14] Vernon N H, Susan H X, Boris J 2009 *Manufacturing & Service Operations Management* **11**(3) 375-96
- [15] Jochen W, Doreen K, Khai S L 2000 *Journal of Services Marketing* **14**(6) 502-12
- [16] Julie M H, Arthur V H 2001 *Production and Operations Management* **10**(4) 405-23
- [17] Hemant K B, Daewon S 2008 *European Journal of Operational Research* **191**(3) 1189-204
- [18] Cedric H W, Hsiao C L, Kuang P H, Yi H H 2012 *International Journal of Hospitality Management* **31**(3) 757-63
- [19] Howard M, Dan S, Walfried M L 2001 *Journal of Services Marketing* **15**(2) 147-59
- [20] Arthur V H, Julie M H, Eitan N 2000 *Journal of Service Research* **2**(3) 254-64
- [21] Michael W K, Woody M L 1994 Estimating hidden quality costs with quality loss functions *Accounting Horizons* **8**(1) 8-18
- [22] Andrea S, Vince T 2006 *International Journal of Quality & Reliability Management* **23**(6) 647-69
- [23] Li M C 2003 Quality loss functions for the measurement of service quality *The International Journal of Advanced Manufacturing Technology* **21**(1) 29-37 (In chinese)
- [24] Chang H L, Byong D R, Cheng T C 2013 *European Journal of Operational Research* **228**(3) 582-91
- [25] Stephen P B, Lieven V 2004 *Convex optimization Cambridge university press*

Author



Wang Wenlong, born in September, 1986, Baoji County, Shaanxi Province, P.R. China

Current position, grades: the Doctoral Student of School of Management, Xi'an Jiaotong University, China.

University studies: received his B.Sc. in Electric Business from Northwest A&F University in China. He received his M.Sc. from Xi'an Jiaotong University in China.

Scientific interest: His research interest fields include service operation, service innovation.

Experience: He participated in many research projects about Service Management from 2010 to 2013, mainly responsible for modelling work.



Liu Xinmei, born in July, 1962, Langfang County, Hebei Province, P.R. China

Current position, grades: the Professor of School of Management, Xi'an Jiaotong University, China.

University studies: received her B.Sc. in Metallograph from Xi'an Jiaotong University in China. She received her M.Sc. and PHD. from Xi'an Jiaotong University in China.

Scientific interest: Her research interest fields include service management and innovation.

Publications: more than 50 papers published in various journals.

Experience: She has teaching experience of over 30 years, has completed several scientific research projects.



Zhang Xiaojie, born in November, 1989, Weifang County, Shandong Province, P.R. China

Current position, grades: the Doctoral Student of School of Management, Xi'an Jiaotong University, China.

University studies: received her B.Sc. in Management from Lanzhou University in China. She received her M.Sc. from Xi'an Jiaotong University in China.

Scientific interest: Her research interest fields include service innovation and creativity.

Experience: She participated in many research projects about service innovation and creativity.

Research on supply chain competition advantage under repeated games

Yu Yue^{1*}, Hu Yong-shi², Xu Ming-xing^{1, 3}

¹*School of Economics & Management, Fuzhou University, Fuzhou City, Fujian Province, China, 350108*

²*Department of Traffic and Transportation, Fujian University of Technology, Fuzhou City, Fujian Province, China, 350108*

³*Concord University College of Fujian Normal University, Fuzhou City, Fujian Province, China, 350117*

Received 8 January 2014, www.tsi.lv

Abstract

To reveal whether the order of supply chains' competition exerts an effect on their profits and whether the repeated game interferes with this effect, the paper builds a Stackelberg game model constructed by two supply chains with each containing a supplier and a retailer based on the previous studies. Through comparing respective profits of the leading and following supply chain represented by 'Copycat', this paper concludes that the following supply chain is more likely to gain more profits than the leading one in this case, and this advantage is determined by the order of decision-making itself. Under repeated games, the possibility of the following supply chain to be more profitable and the approaches to make decisions will be related to the substitutable coefficient.

Keywords: leading-following supply chain, Stackelberg game, repeated games, later-mover advantage

1 Introduction

Along with the development of economy in China, the competitions between the enterprises have become the rivalries between the supply chains. In the real market, some leading companies in the industry often dominate in the decision-making and set up the standards for the smaller companies to follow. In other words, the competition between supply chains is actually the competition between leading and following supply chains. Tengxun, which built up by imitating ICQ, provides a good example. A game called 'vegetable stealing' launched in Kaixin network was once a sweeping trend in the virtual community, followed by QQ farm which has similar features. It turned out that Tengxun achieved an outright win in this battle with Kaixin network, ending up taking over it reversely. Another prototypical example is that the copycatting mobile phones, which participate in the competition with low pricing strategy, have strongly impacted the industry and hit the brand products when they entered the market. However, the copycatting netbooks are not in a close game, the low shipments and profit margin, the long cycle length as well as the high risk are the contributing factors to the bankruptcy of new entrants who have been forced to retreat from the market. Even within the copycatting industry, the results of market selection and elimination can be widely divergent. For instance, Tianyu, once an unknown manufacturer, which started out doing copycatting products, used to be one of the most promising emerging domestic mobile phone brands

due to its rapid growth of market share. Its sales volume once outnumbered those of Samsung, Sony Ericsson, Motorola and LG. However, some brands such as TCL, Konka and Bird might enjoy the popularity for a time but they were doomed to fail and eventually disappeared from the scene. Therefore, a clear understanding of the main competition mechanism and the profit analysis of the leading and following supply chain is of significant importance to companies that participate in this fierce competition, especially for those being dominated.

The research on the competition mechanism between the supply chains can be divided into two categories including single game and repeated game. Abundant theoretical results regarding research on the single game between leading and following supply chains have been obtained. According to Yang Daojian and Qi Ershi [1], through the analysis of the effect of the product substitution degree has on the overall profit of the supply chain in the competitive environment, the extent to which this effect influences the information sharing can be assessed. Based on Zhang Jiangnan and Yuan Zuofang [2], by comparing the performance of the Bertrand game conducted among different supply chains under the market demand of power function structure, a conclusion that traditional 'the first-mover advantage' could not hold can be drawn. Although this literature provides a good insight on analysing competition and performance between supply chains, the authors failed to consider the fact that not all companies are doing the game at the same time because leading companies are more likely to make decisions in advance, leaving other smaller companies no

* *Corresponding author* e-mail: yu_fish86@126.com

choices but to follow. As such, the paper still needs to be improved. A research conducted by Li Boxun [3] accentuates how to choose between centralized and decentralized strategy under the leading-following Stackelberg game and identifies the relationship between strategy selection and product substitution degree. This paper analyses the existing competition between the leading and following supply chains in the marketplace, however, it concerns more about the impact of business decisions made by each member in a supply chain has on the performance than unveiling the profit differences that result from the leading-following supply chain competition mechanism itself. Moreover, the authors also failed to discuss the consequences in the case where the market demand is uncertain.

Researches on repeated game between the leading and following supply chain are still relatively unnoticed, while more attention are paid to the repeated game within one single supply chain or between duopoly supply chains. A study conducted by Qi Guiqing [4] illustrates the competitive and cooperative relationship between the supply chains in a cluster network under the repeated game. Besides, through the application of the dynamic repeated game theory, this paper carries a prudent study on the members of one single supply chain among the various supply chains in a cluster network; it also demonstrates the competing and cooperating status of node firms in the parallel supply chains with the same value chain and illustrates the conditions of the cooperative equilibrium. According to Wang Ruoying and Chen Hongmin [5], the dynamic contact equation has been confirmed under the repeated game in the duopoly market. Moreover, a conclusion that the outcomes of the finitely repeated game will deviate from the Nash equilibrium to cooperation is verified. Through analysing repeated game towards synergistic competing and collaborating relationship between the node firms in a supply chain under the conditions of asymmetric information, Yan Guangquan [6] obtains a payoff matrix in relation to the likelihood of cooperation among the node firms and proposes a hypothesis regarding the settings of the ruthless strategy and incentive mechanism under the repeated game.

Based on the previous studies, this paper seeks to build a Stackelberg game model constructed by two supply chains with each containing a supplier and a retailer. By comparing revenues generated from the existing leading-following supply chain in a common market, the paper aims to demonstrate whether the competition mechanism has an effect on the supply chain revenues and to discuss whether it will continue its influence under the repeated game in order to provide the following companies with future references.

2 Model assumptions

This paper analyses the leading-following supply chain competition in the real market. To further discuss the

main impact of the supply chain competitive mechanism has on the profits and to eliminate the interference of corporate decision-making in the supply chain such as information sharing, the study is under the condition of complete information where the game between leading-following supply chains takes place; this means that retailers would always faithfully report the market demand to the suppliers. Ideally, each supply chain includes a supplier S and a retailer R. As such, this article assumes that supply chain *i* (SC_{*i*}) is the leading supply chain that dominates the market to make the first moves, while supply chain *j* (SC_{*j*}) is the following supply chain that observes and follows the trend accordingly. In other words, both of them are participating in the Stackelberg game. Hypothetically, suppliers on these two supply chains offer products to the retailers with price *w_i*, *w_j* respectively, and accordingly their retailers provide consumers with the retailing price *p_i*, *p_j*, while the market demands present *q_i*, *q_j* correspondingly. The structures of these two supply chains are shown below.

$$\begin{matrix} S_i \xrightarrow{w_i} R_i \xrightarrow{p_i} C \\ S_j \xrightarrow{w_j} R_j \xrightarrow{p_j} C \end{matrix} \quad (1)$$

Based on the principles of economics [7], this paper assumes that the basic needs of the market are corresponding to a linear uncertain demand,

$$q_i = \alpha_{Di} - p_i + \gamma * p_j, \quad q_j = \alpha_{Dj} - p_j + \gamma * p_i, \quad (2)$$

where γ represents the alternative coefficient of the two price supply chains while α_D acts as the initial needs of the market. As such, the dynamic contact equation is as follows.

$$\begin{aligned} \alpha_{Dj}^1 &= \alpha_{Dj}^1 = \alpha_D \\ \alpha_{Dj}^{t+1} &= \alpha_{Dj}^t - \frac{p_j^t - p_i^t}{2}, \quad t = 1, 2, \dots, n. \\ \alpha_{Di}^{t+1} &= \alpha_{Di}^t - \frac{p_i^t - p_j^t}{2}, \quad t = 1, 2, \dots, n \end{aligned} \quad (3)$$

Recording Π as the revenues that correspond to each part, as such the retailer's earnings are as follows:

$$\Pi_{Ri} = (p_i - w_i) * q_i, \quad \Pi_{Rj} = (p_j - w_j) * q_j. \quad (4)$$

Suppliers' earnings are:

$$\Pi_{Si} = w_i * q_i, \quad \Pi_{Sj} = w_j * q_j. \quad (5)$$

Supply chain overall revenues are:

$$\Pi_{SCi} = \Pi_{Si} + \Pi_{Ri}, \quad \Pi_{SCj} = \Pi_{Sj} + \Pi_{Rj}. \quad (6)$$

3 Model solutions of repeated games

After observing the pricing and sales of leading supply chain *i*, supply chain *j* will make the decisions correspondingly; later in the supply chain *i* / *j*, given the wholesale price of supplier *i* / *j*, the retailer *i* / *j* then will make the decisions accordingly.

So reverse induction is used to calculate the maximum benefits of the retailer *j*, that is $\Pi_{Rj} = 0$, it can be formulated as follow:

$$p'_j = \frac{\alpha'_{Dj} + \gamma * p'_i + w'_j}{2} \tag{7}$$

To maximize the benefits of supplier *j*, that is when $\Pi_{Sj} = 0$, therefore

$$w'_j = \frac{\alpha'_{Dj} + \gamma * p'_i}{2} \tag{8}$$

Substituting equation (7) into equation (8) it can be obtained that,

$$p'_j = \frac{3}{4}(\alpha'_{Dj} + \gamma * p'_i), \quad q'_j = \frac{1}{4}(\alpha'_{Dj} + \gamma * p'_i) \tag{9}$$

For retailer *i*, given that:

$$q'_i = \alpha'_{Di} - p'_i + \gamma * p'_j \tag{10}$$

In order to maximize the benefits of retailer *i*, that is when $\Pi_{Ri} = 0$, it can be obtained that

$$p'_i = \frac{\alpha'_{Di} + \frac{3}{4}\gamma\alpha'_{Dj} + (1 - \frac{3}{4}\gamma^2)w'_i}{2(1 - \frac{3}{4}\gamma^2)} \tag{11}$$

Based on the results summarized above in reverse order, we forward projected and substituted equation 3-5 into supplier *i*'s revenue equation expression. According to the principle of maximizing the interests of supplier *i*, that is when $\Pi_{Si} = 0$, it can be obtained that

$$w'_i = \frac{\alpha'_{Di} + \frac{3}{4}\gamma\alpha'_{Dj}}{2(1 - \frac{3}{4}\gamma^2)} \tag{12}$$

Substituting equation (12) into equation (11), it can be obtained that

$$p'_i = \frac{3}{4} \frac{(\alpha'_{Di} + \frac{3}{4}\gamma\alpha'_{Dj})}{(1 - \frac{3}{4}\gamma^2)}, \quad q'_i = \frac{1}{4}(\alpha'_{Di} + \frac{3}{4}\gamma\alpha'_{Dj}) \tag{13}$$

Therefore, the revenues of each part of the supply chain *i* are:

$$\begin{aligned} \Pi_{Si} &= \frac{(\alpha'_{Di} + \frac{3}{4}\gamma\alpha'_{Dj})^2}{8(1 - \frac{3}{4}\gamma^2)}, \\ \Pi_{Ri} &= \frac{(\alpha'_{Di} + \frac{3}{4}\gamma\alpha'_{Dj})^2}{16(1 - \frac{3}{4}\gamma^2)}, \\ \Pi_{S_{Ci}} &= \frac{3(\alpha'_{Di} + \frac{3}{4}\gamma\alpha'_{Dj})^2}{16(1 - \frac{3}{4}\gamma^2)}. \end{aligned} \tag{14}$$

Substituting the result of equation (13) into equation (9), it can be obtained that

$$\begin{aligned} p'_j &= \frac{3}{4} \frac{(1 - \frac{3}{16}\gamma^2)\alpha'_{Dj} + \frac{3}{4}\gamma\alpha'_{Di}}{(1 - \frac{3}{4}\gamma^2)}, \\ q'_j &= \frac{1}{4} \frac{(1 - \frac{3}{16}\gamma^2)\alpha'_{Dj} + \frac{3}{4}\gamma\alpha'_{Di}}{(1 - \frac{3}{4}\gamma^2)}, \\ w'_j &= \frac{1}{2} \frac{(1 - \frac{3}{16}\gamma^2)\alpha'_{Dj} + \frac{3}{4}\gamma\alpha'_{Di}}{(1 - \frac{3}{4}\gamma^2)}. \end{aligned} \tag{15}$$

Therefore, the revenues of each part of the supply chain *j* are:

$$\begin{aligned} \Pi_{Sj} &= \frac{1}{8} \left[\frac{(1 - \frac{3}{16}\gamma^2)\alpha'_{Dj} + \frac{3}{4}\gamma\alpha'_{Di}}{(1 - \frac{3}{4}\gamma^2)} \right]^2, \\ \Pi_{Rj} &= \frac{1}{16} \left[\frac{(1 - \frac{3}{16}\gamma^2)\alpha'_{Dj} + \frac{3}{4}\gamma\alpha'_{Di}}{(1 - \frac{3}{4}\gamma^2)} \right]^2, \\ \Pi_{S_{Cj}} &= \frac{3}{16} \left[\frac{(1 - \frac{3}{16}\gamma^2)\alpha'_{Dj} + \frac{3}{4}\gamma\alpha'_{Di}}{(1 - \frac{3}{4}\gamma^2)} \right]^2. \end{aligned} \tag{16}$$

According to equation (3), it can be concluded that

$$\alpha_{Dj}^{t+1} = \frac{(\frac{5}{8} + \frac{9}{32}\gamma - \frac{87}{128}\gamma^2)\alpha'_{Dj} + (\frac{3}{8} - \frac{9}{32}\gamma)\alpha'_{Di}}{1 - \frac{3}{4}\gamma^2},$$

$$\alpha_{Di}^{t+1} = \frac{(\frac{3}{8} - \frac{9}{32}\gamma - \frac{9}{128}\gamma^2)\alpha'_{Di} + (\frac{5}{8} + \frac{9}{32}\gamma - \frac{3}{4}\gamma^2)\alpha'_{Dj}}{1 - \frac{3}{4}\gamma^2}. \quad (17)$$

4 Analyses of product substitutable factor and initial market demand

By comparing the respective profits of the leading and following supply chain, this article targets the revenue difference as the objective function to achieve the equation shown below.

$$F(t, \gamma) = \Pi_{SCj} - \Pi_{SCi} = p_j^t q_j^t - p_i^t q_i^t. \quad (18)$$

4.1 INFLUENCE OF PRODUCT SUBSTITUTABLE FACTOR

Substituting the objective function and dynamic contact equation into MATLAB to iterate repeatedly for 50 times, the outcome can be seen in the following diagrams. To calculate conveniently, we assign value 1 to the initial demand α_D . Besides, for the purpose of distinguish one from another readily, the blue lines in the figures are used to represent the game proceed from the 7th to the 50th.

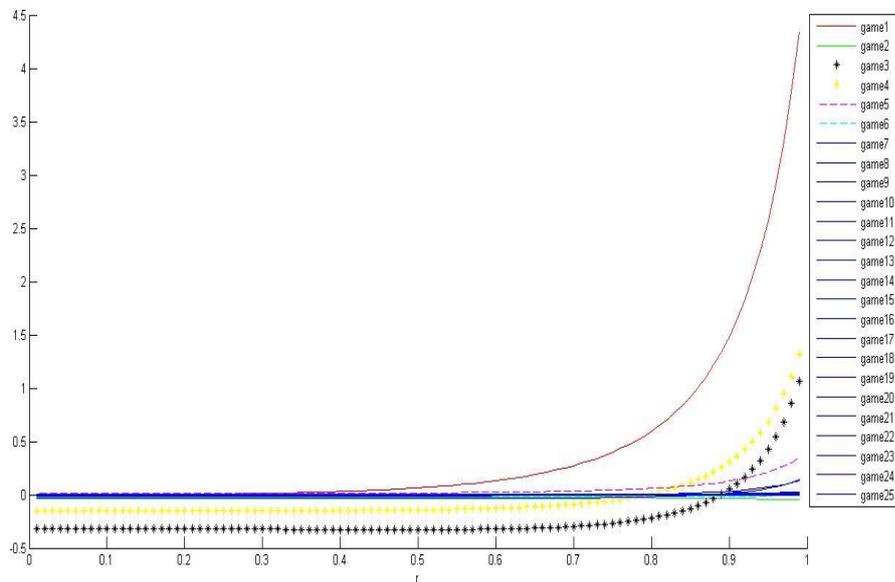


FIGURE 1 Games repeated 25 times while initial demand is 1

As it can be seen clearly from the chart, when the product substitution coefficient $\gamma \in (0, 0.5)$, which means the products of two supply chains do not possess high similarity, it is difficult to achieve a higher profit or late-mover advantage via competitions. This result is consistent with our cognition that rivalry is not likely to occur between supply chains that yield different types of products.

When the product substitution coefficient $\gamma \in (0.5, 0.9)$, that is, the products generated from two supply chains have relatively high similarity, the following supply chain is able to take advantage of the late moves in an extremely limited number of the game, and moreover this advantage will be offset or even outstripped by the revenue effects resulted from the price differences. Consequently, the profits of the first-mover and the late-mover supply chain will reach an equilibrium state.

When the product substitution coefficient $\gamma \in (0.9, 1)$, which means there are striking similarities in the products of two supply chains (products can basically be considered to be identical), it is feasible for late-mover supply chain to observe the market responses aroused by the first movers in order to reduce the market uncertainty. This action enables the followers to attain more profits until a state of equilibrium is reached ultimately.

4.2 INFLUENCE OF INITIAL MARKET DEMAND

Considering the alterations of initial market demand, we assume the original demand α_D to be 1, 10 and 100, respectively. Substituting these values into the objective function to calculate over 25 iterations, the profits difference between leading and following supply chains can be acquired as exhibited in the following figures.

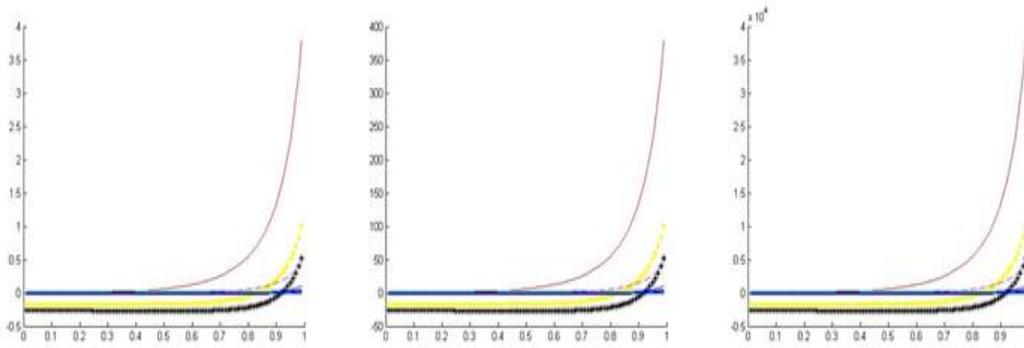


FIGURE 2 Games repeated 25 times while initial demand is 1, 10 and 100

The figure 2 reveals that the function curves of revenue differences between leading and following supply chains are not affected by the variations on the initial market demand, which means the changes are without impact on the late-mover advantage. However, the relationship between the values of profit differences and the alterations of initial market demand demonstrates the quadratic complexity. On the other hand, due to the

deficiency of product awareness for the first entry into the market, the original demand of those followers is normally lower than that of the dominant ones. Given that the level of initial demand has no impact on the function curve of late-mover advantage, we presume the value of original demand of the leading and following supply chain to be 10 and 0-10, accordingly. The function curves are illustrated by the diagrams.

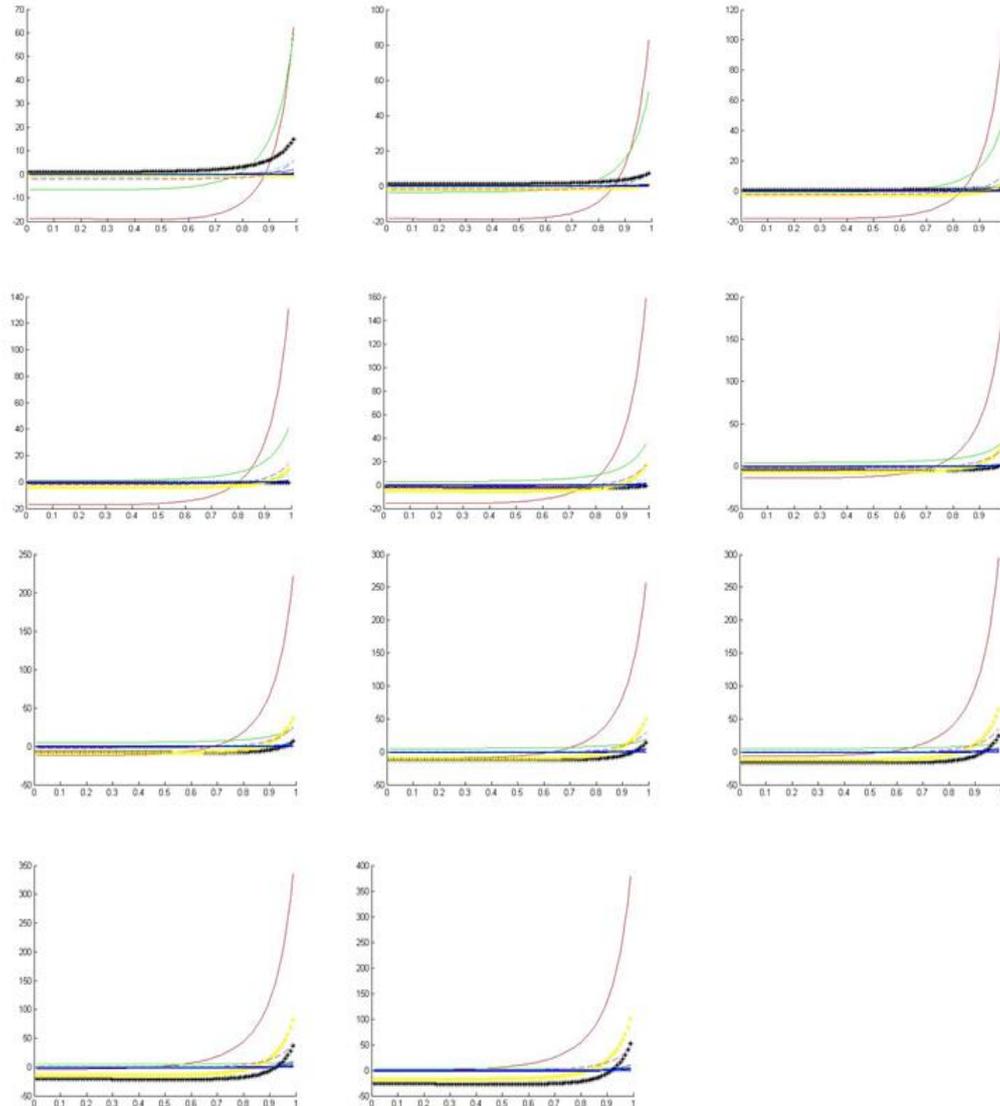


FIGURE 3 Games repeated 25 times while the initial demands of leading and following supply chain are 10 and 0-10 respectively

It can be obviously observed from the Figure 3 that the lower recognition of the following supply chains is while entering into the market, that is, the weaker the initial demand is, the less likely for them to attain the late-mover advantage on previous occasions of the game. However, as for long-term repeated game, the profits of leading and following supply chains will ultimately revert to the equilibrium state, verifying parts of conclusion in terms of product substitution coefficient as mentioned above.

4.3 ANALYSIS OF PRACTICAL PROBLEMS

As we go back to the copycatting issues proposed in the foreword, we can find out the causes behind the failure of copycatting netbooks to secure a competitive foothold in the marketplace and to duplicate the success of the copycatting mobile phones. The primary reason is that the attributes of copycatting netbooks with higher technical contents and greater differentiation are alien from those of the copycatting mobile phones, leading to the difficulty in imitating. Therefore, when companies are in a certain industry with high-tech or wide product differentiation, more efforts should be directed towards the technological innovation and R&D on products to distinguish themselves in the established industries through technological breakthrough rather than price competition.

Generally speaking, the profits of leading and supply chains will ultimately achieve a balanced state in the case of repeated game, namely, the observation of information will counteract the late-mover advantage of the followers attained from surveying the market response and restricting the market ambiguity under the repeated game. Therefore, in order to secure the long-term success, both supply chains and enterprises are supposed to strengthen the awareness of innovation even in some industries with high product similarities such as mobile phone, battery and commodity, and additionally to further reinforce the late-mover advantage derived from the low price strategy in the preliminary stage of competition through imitative innovation.

5 Conclusions

By comparing the revenues of leading and following supply chains under the condition of repeated game, we can reach the following conclusions:

Generally, the later-mover advantage of the following supply chain has been seen in the leading-following supply chain competition. This is mainly because the following supply chain is able to observe the outcomes of the decisions and actions made by the leading supply chain, thereby reducing the market uncertainty to gain more revenues.

In the case of repeated game, whether the following supply chain is able to increase benefit and what takes to make proper decisions largely relate to the substitution coefficient γ . When $\gamma \in (0, 0.5)$, the difficulty in

augmenting revenues or acquiring the late-mover advantage via competition enables each supply chain to concentrate on their own business to avoid unnecessary rivalry. When $\gamma \in (0.5, 0.9)$, the following supply chain can take advantage of the late moves, but in the long term companies ought to ensure the enhancement of the service and technology. When $\gamma \in (0.9, 1)$, through observing the market response of the leading supply chain, those followers are capable of decreasing the market uncertainty to achieve more profits. As such, costs on the marketing to compete with rivals should be increased. However, the respective revenues generated from the leading and following supply chains will reach an equilibrium state eventually, which means the late-mover advantage of the following supply chain obtained from observing the market reaction and lessening the market uncertainty will be offset by the information observation under the repeated game.

The change of initial market demand are without impact on the late-mover advantage and the relationship between the values of profit differences and the alterations of initial market demand demonstrates the quadratic complexity. The lower recognition of the following supply chains is while entering into the market, that is, the weaker the initial demand is, the less likely for them to attain the late-mover advantage on previous occasions of the game. And as for long-term repeated game, the profits of leading and following supply chains will ultimately revert to the equilibrium state.

The above conclusions provide a theoretical verification for some economic theories such as freebie and sunk costs and also offer inspiration for specific companies to make decisions. In the low product-differentiated industries, those followers who face the dilemma in entering or surviving in the business dominated by a certain number of leading companies with technological advantages can manage to decrease the market uncertainty through observing the consequences of the first moves taken by leading companies, and hence take advantages of appropriate later moves in order to develop the capabilities to catch up. Therefore, technological innovation and high quality of the products are the secure guarantees to the development and success of the copycatting industry in the long run.

Thru the comparison of revenues generated from the leading-following supply chains in the repeated game, a profound finding that the following supply chain can take advantage of the late moves, with profits and decision-making varying in relation to the product substitution coefficient, can be acquired. This can provide references for the further research regarding the competition between supply chains and for the followers that urge to gain footholds in the target market. However, there are still a number of incomplete information and multi-chains games in the real world that need to be taken into consideration in the future studies. Additionally, for further research it should also be interesting to investigate

how to adjust reasonably the contracts offered by suppliers, how to boost the competitive advantages of a

supply chain and how to coordinate effectively within a supply chain.

References

- [1] Yang Daojian, Qi Ershi 2009 Product differentiation and information sharing under supply chain competition *Systems engineering and electronics* **31**(9) 2141-5 (*In chinese*)
- [2] Zhang Jianghan, Yuan Zuofang 2010 Analysis of behaviour performance and competition mechanism between two supply chains *System Engineering* **28**(8) 81-4 (*In chinese*)
- [3] Li Boxun, Zhou Yongwu, Wang Shengdong 2012 The longitudinal structure decision model between supply chains under Stackelberg game *The management of scientific research* **33**(12) 50-7 (*In chinese*)
- [4] Qi Guiqing, Yang Xihuai, Li Sen 2006 Coopetition analysis of clustered network supply chain by repetitive game theory *Journal of North-eastern University* **27**(2) 233-6
- [5] Wang Ruoying, Chen Hongmin 1998 Countermeasures of duopoly in repeated game: an experimental study *Quantitative and technical economic* **15**(11) 23-8
- [6] Yan Guangquan, Wu Qinglie, He Yong 2008 Analysis of repeated game of supply chain coordination under asymmetric information *Industrial economy* **27**(2) 33-7 (*In chinese*)
- [7] Choi S C 1991 Price Competition in a Channel Structure with a Common Retailer *Marketing Science* **10**(4) 271-97 (*In chinese*)
- [8] Holt C A 1996 *Journal of Economic Perspectives* **10**(1) 193-203
- [9] Pasternack B A 1985 *Marketing Science* **4**(2) 166-76
- [10] Bernstein F, Federgruen A 2005 *Management Science* **51**(1) 18-29
- [11] Zhou Y W 2007 *European Journal of Operational Research* **181**(2) 686-703

| Author | |
|---|---|
|  | <p>Yu Yue, born in August, 1986, Fuzhou City, Fujian Province, P.R. China</p> <p>Current position, grades: Ph. D. student of School of Economics & Management, Fuzhou University, Fuzhou, China.</p> <p>University studies: received his Bachelor of Science from Nanjing University of Science and Technology in China. He received his Master of management from Fuzhou University in China.</p> <p>Scientific interest: His research interest fields include logistics, supply chain management</p> <p>Experience: He has completed four scientific research projects.</p> |
|  | <p>Hu Yong-shi, born in April, 1982, Fuzhou City, Fujian Province, P.R. China</p> <p>Current position, grades: Teacher of Fujian University of Technology, Fuzhou, China.</p> <p>University studies: received his Bachelor of Management from Fuzhou University in China. He received his Master of management from Fuzhou University in China. He received his Doctor of management from Fuzhou University in China.</p> <p>Scientific interest: His research interest fields include logistics, supply chain management</p> <p>Experience: He has completed twenty scientific research projects.</p> |
|  | <p>Xu Ming-xing, born in February, 1982, Fuzhou City, Fujian Province, P.R. China</p> <p>Current position, grades: Ph. D. student of School of Economics & Management, Fuzhou University, Fuzhou, China; teacher of Fujian Normal University, Fuzhou, China.</p> <p>University studies: received his Bachelor of Engineering from Fujian Normal University in China. He received his Master of Management from Fuzhou University in China.</p> <p>Scientific interest: logistics and supply chain management, E-economics</p> <p>Experience: He has completed a dozen scientific research projects and many academic papers.</p> |

The development and evolution of bridge in Chongqing China

Yan Li^{1*}, Dong Wei², Yangyang Chen¹

¹Faculty The College of Civil Engineering, Chongqing University, 400030, Chongqing, China

²The College of Architectural Engineering Vocational, Chongqing, 400038, Chongqing, China

Received 28 June 2014, www.tsi.lv

Abstract

In this paper, the development and evolution of bridge in Chongqing and around the world are summarized particularly. Besides, the categories of bridges, the development of bridge design theories as well as the breakthroughs of bridge construction with the passage of time are also introduced systemically. With the introduction of the historical stone-arch bridges, the recent reinforced concrete slab bridges, modern pre-stressed concrete bridge, various arch bridges, suspension bridges and cable-stayed bridges, the prosperity and progress made by human beings in the process of transformation of nature are gradually revealed in this paper, the development and evolution of bridge is also revealed with the introduction of new techniques. The construction of bridge promotes the economic development and strengthens the connection of different areas, brings a booming market. The role of mechanics in bridge design is analysed and the development of Chongqing bridges can also be experienced in this paper. Bridge is not only a construction but also the creator of the soul of a city, showing the fighting spirit and braveness of a generation.

Keywords: Chongqing, Bridges, Development, Evolution model

1 Introduction

Bridge is a construction built to overcome natural or artificial obstacles. It usually comprises of five big parts and five small ones. The big parts refer to superstructure and substructure of bridge span that bears the load of cars and other vehicles. They assure the safety of bridge structure including (1) bridge span structure (also called bridge opening structure or superstructure), (2) support system, (3) bridge pier, bridge abutment, (4) cushion cap, (5) pile foundation [1]. The five small parts refer to those directly related to the service function of bridge, which used to be called bridge floor structures including (1) bridge floor pavement, (2) waterproof and water drainage structure, (3) bridge railing, (4) expansion joint, (5) lighting [2]. Bridge is also part of a road. From technologic perspective, the bridge development can be divided into ancient bridge, early modern time bridge and modern bridge [3].

Before 17th century, the ancient bridges are usually made of wood and stone, and accordingly, there were wooden bridge and stone bridge [4-6]. Bridge construction in the early modern times speeds up the rising and developing of bridge theories. In 1857, based on the precious researches on arch theories, statics and mechanics of materials, St. Venant put up a more comprehensive theory about girder theory and torsion theory. Meanwhile, theories about continuous girder and cantilever girder were also brought up. The development of these theories promotes the development of truss, continuous girder and cantilever girder. Modern bridges can be divided into pre-stressed rebar concrete bridge, the

rebar concrete bridge and steel bridge.

The construction of bridge promotes the economical development and strengthens the connection of different areas, brings a booming market. The role of mechanics in bridge design is analysed and the development of Chongqing bridges can also be experienced in this paper. Bridge is not only a construction but also the creator of the soul of a city, showing the fighting spirit and braveness of a generation. Hence, it is very necessary and significant to investigate the development and evolution of Bridge.

Chongqing, which is honoured as a city of water and mountains, has to largely depend on bridges to connect rivers and hills. Because of various kinds of mountains and rivers, every connection must rely on innovation of bridge construction techniques. Chongqing bridges have outnumbered other cities and its construction numbers and difficulties are not a frequent sight around the world. So, study the development process of bridges at Chongqing has its scientific and social value. Based on the above reasons, this paper mainly generalizes the development of bridges in China and around the world.

2 Evolution of Chongqing bridges from a historical view

Because of the special terrain, bridges at Chongqing city play an important role in connecting mountains with rivers and thus improve the traffic. At 2005, Chongqing is also identified as the only "City of Bridges" by Mao Yishen Bridges Committee. Bridges at Chongqing have reached the number of 4900. Among all bridges, there are

* Corresponding author e-mail: 379791595@qq.com

stone arch bridges with long history, magnificent steel and suspension bridges which across Yangtse River. Chongqing has a long history of bridge construction and the wisdom of its people is also well reflected in nature transformation and transport improvement. Because of too many valleys, rivers and rich sources of stones with high quality, many stone arch bridges were built in history and some of them are classic ones. Shiji Bridge, which is also called Guangji Bridge, located in the west part of Rongchang County and cross Laixi River. Shiji Bridge was built in AD 1050 when Wen Boyan, a chancellor in Song dynasty, named it Siji Bridge, which is the earliest bridge in record at Chongqing. At the end of Qing dynasty, for the consequence of Taiping Rebellion, salt in Sichuan has to be transported to Hubei Province, in which Siji Bridge provided a good guarantee. After that, it was honoured as the safeguard of east Sichuan Province by Majesty Cixi and earned its reputation.



FIGURE 1 Shiji Bridge

Dragon and phoenix Kezhai Bridge, which was built in Yuan dynasty, lies to 25 km from west part of Xiushan County and cross Pingjiang River. The bridge is oriented north and south and composed of six stone and wood arches. It is 58.2 meters long and 8.95 meters high. The bridge heads are doorways made up of blue bricks and gable walls and there are seven and ten steps respectively at each side. Kezhai village lies in west and cottages in east of bridge. It is the connecting road for villages Qingxi, Longfeng and Tangao and also serves as a place for divine activity for local Tujia nationality.



FIGURE 2 Kezhai Bridge

Except for the ancient stone arch bridge, there are still distinctive rope bridges at Ningchang, Wuxi County among all old bridges of Chongqing. People in Wuxi County are proud of Ningchang because the most glorious moment in the history of Wuxi County was created here. Back then, the people who used to live in deep mountains enjoyed great prosperity and richness for salt manufacturing there. With a history of over 4000 years, salt manufacturing at Ningchang is honoured as the place where the manual workshop in the world started. Thanks to the prosperity at ancient time, the rope bridge at Ningchang came into being.



FIGURE 3 Rope Bridge at Ningchang, Wuxi country

The construction of bridges at Chongqing mainly depends on the needs of city road construction. Less than 100 bridges with short spans are widely spread and most of them are arch ones. The most significant bridge must be Hualong Bridge built in 1930 since it is the first road stone arch bridge in Chongqing.



FIGURE 4 Hualong Bridge

Besides, there is Number One Bridge, which is the earliest and greatest one in scale. It was a five-span 16m rebar concrete girder bridge that was built in 1940s.

After the foundation of PRC, people in Chongqing made the most of their talents and abilities to construct many cross-river bridges with large span and high technology, seven of which are most representative ones. Baishatuo Bridge, which was built in 1959, is the first cross-river bridge in Chongqing. Niujiatuo Jialing River Bridge, which was built in 1966, is the only steel truss girder bridge in southeast China.

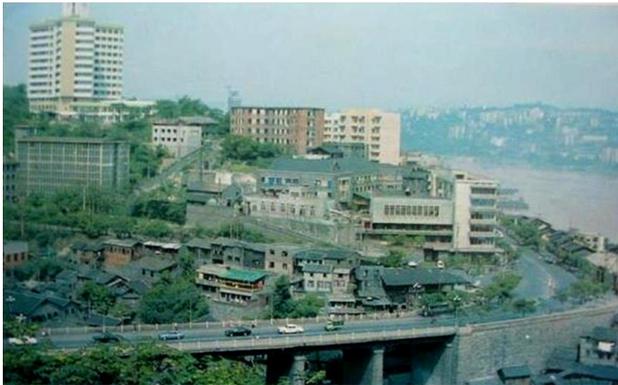


FIGURE 5 No.1 Bridge

Chaoyang Bridge at Beipei district, which was built in 1969, is the only double chain suspension bridge in China; Tangxihe cable-stayed bridge, which was built in 1975 at Yun'an, a small town in Yunyang County, is the first cable-stayed bridge in China. It marks the beginning of cable-stayed bridge construction in China. Shibampo Yangtze River Bridge, which was the last work of Mao Yishen, was a prestressed concrete T-shaped rigid frame bridge built in 1980. Shimen Bridge, which was built in 1988, was the single tower and single plane cable-stayed bridge with the greatest span in China. Chaotianmen Bridge, which was built in 2009, was honoured as the "number one" steel trussed arch bridge in the world.



FIGURE 8 Chaoyang Birge at Beipei



FIGURE 9 Shibampo Bridge

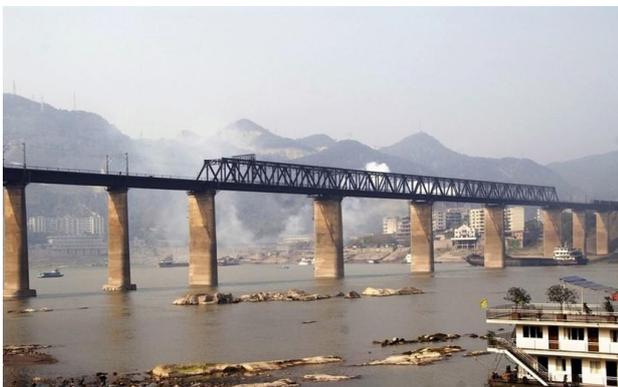


FIGURE 6 Baishatuo Bridge



FIGURE 10 Tangxihe Cable-stayed Bridge at Yun'an, Yunyang County



FIGURE 7 Jialing River Bridge at Njiaotuo



FIGURE 11 Shimen Bridge

3 The development and evolution of bridges at Chongqing from economic and technological view

Chongqing, which is honoured as a city of water and mountains, has to largely depend on bridges to connect rivers and hills. Because of various kinds of mountains and rivers, every connection must rely on innovation of bridge construction techniques. Before modern times, most bridges at Chongqing were arch ones because of the underdeveloped economy and construction techniques. With the foundation of PRC and development of economy and technology, bridges with different structures and materials have been gradually built over rivers and valleys.

3.1 ARCH BRIDGE

Arch bridge depends on the arch to be the main load bearing structure. The earliest arch bridge is a stone one transmitting the load of bridge and carrying capacity put on bridge horizontally to piers at the both end of bridge with small units like trapezoid stone. When the small units push and press, they increase intensity of bridge in the meantime. Many ancient bridges are arch bridges not only because it arch formed by arc is steadier than straight girder, but also because arch bridge looks beautiful. The most famous arch bridge is Zhaozhou Bridge. After withstanding thousands of years of wind, rain and earthquake, it is still playing an important role today.

Wushan Yangtze River Bridge, which is honoured as the first bridge at east Chongqing, is an arch bridge with half-through concrete-filled steel tube. The chord is straight steel pipes with major diameter. Elevation web member is composed of vertical and horizontal way of arrangement to improve the angle between web member and chord member. Wanxiang county Yangtze River Bridge, which is the largest concrete arch bridge in the world at present, is 856.12 meters long, 24 meters wide and 147 meters high. The main span is 420 meters. The main arch ring is of composite structure with high strength concrete and steel-tube concrete stiff framework. Chaotianmen Bridge, which is 552 meters long, is an arch bridge with the biggest steel truss. It proposes a composite of rigid and flexible structure. In construction, cantilever construction is achieved through counter weight, cable-stayed fasten and hang system. The techniques of unstressed fold of arches in the mid-span are also adopted.

The construction of Wushan Yangtze River Bridge, Wanxiang County Yangtze River Bridge and Chaotianmen Bridge fully indicates that Chongqing has reached a world-class in building concrete arch bridge, steel-tube concrete bridge and steel arch bridge and that the construction of arch bridge in Chongqing has stepped into the international level.

3.2 SUSPENSION BRIDGE

Suspension Bridge is also called hanging bridge. The towers at the both ends of bridge bears the strength of suspension bridge and the suspension cable between two towers hold bridge floor. Suspension bridge is the main form of long-span bridges. Except for Sutong Bridge and The stonecutters bridge in Hong Kong, other bridges with above 1000m span are all suspension bridges. China has the earliest recorded history of suspension bridge. Until today, suspension bridge is still influencing development of hanging bridges in the world.

Zhongxian County Yangtze River Bridge locates in Xiadukou, Zhongzhou Town, Zhongxian County, Chongqing. It is a bridge with the main span of 560 m and the length of 1200 m. Opened to traffic in 2001. The bridge is built into spatial steel tube truss and stiffening girder structure and is characterized by higher stiffness, stiffness of torsion, wind stability and low investment. Besides, the roof falling structure of tunnel anchor and stone anchor is also domestically unprecedented. The stress distribution of surrounding rocks is improved and security increased.

Egongyan Bridge, which located in the downtown of Chongqing city serves as the main channel connecting Chengdu-Chongqing expressway and Chongqing-Guizhou expressway. Built in 2000, the bridge with the 600m main span is 1420 meters long and 35.5 meters wide. It is a two-way bridge with six lanes and the place for light rail is also pre-reserved. Since it is the only continuous stiffening steel box girder suspension bridge.

The Second Wanzhou Yangtze River Bridge, which located in Juyutuo, Wanzhou District, Chongqing city, is 1153.86 meters long and 20.5 meters wide with the main span of 580m. The main cable of suspension bridge is pre-ordered galvanized parallel wire with high stiffness. Stiffening beam is of spatial truss structure of steel tube. The bridge is also characterized by largest tunnel anchorage structure adopting anchorage pre-stressing anchor system whose way of force transmitting is largely improved compared with the previous steel anchor system and therefore, the intensity of high-strength material can be fully exerted.

3.3 CONTINUOUS RIGID FRAME BRIDGE

Continuous rigid frame bridge is a consolidated continuous bridge of pier and girder. It falls into multi-span rigid frame bridge with continuous girder and multi-span continuous rigid frame bridge. Both uses pre-stressed concrete structure. More than two main piers are consolidated by pier and girder. It has the advantage of T-shaped rigid frame bridge.

Continuous Rigid Bridge is one of the main types of bridges with 100m-300m span, it is characterized by simple structure, convenient construction and low initial investment. In Chongqing, continuous rigid bridges have outnumbered other places, so its design, construction and

research rank high in China. Among, the most representative one is double line bridge of Shibampo Changjiang River Bridge designed by Lin Tongyan International Engineering Consulting Co, Ltd. Steel box girder section is designed in the 330m main span of this bridge, which is a great innovation. Considering that too much Dead load stress of large-span continuous rigid frame with pre-stressed concrete is bad for span capability, the composite steel-pre-stressed concrete system is adopted, which helps to reduce the dead load and improve the span. The composite steel-pre-stressed concrete system also weakens the negative influence of concrete shrinkage on bridge structure.

3.4 CABLE-STAYED BRIDGE

Cable-stayed Bridge uses one or more main towers and wire rope to support the bridge floor. Among a large number of various cable-stayed bridge, there are some representative ones such as Fuling Yangtze River Bridge, Dafosi Yangtse River Bridge, Lijiatuo Yangtse River Bridge, Masangxi Yangtse River Bridge, Fengjie Yangtze River Bridge, Jiangjin Guanyinyan Yangtze River Bridge, Zhongxian county Yangtze River Bridge over Shizhong expressway, etc. Fuling Yangtze River Bridge adopts a double longitudinal beam rib-plank structure with pre-stressed concrete. The beam is 2.3 meters high and the plate is 25 cm thick. It is the thinnest main beam of same bridge types in record. Lijiatuo Yangtse River Bridge firstly brought up and successfully adopted flat double main beam structure of pre-stressed concrete at home. The cable-stayed zone of upper pylon is pre-stressed anchor system. The front-point proposes cantilever cradle quick construction techniques. Jiangjin Guanyinyan Yangtze River Bridge, which is a cable-stayed bridge with concrete composite girder, is also the biggest composite girder bridge. Another difficulty lies in the construction of bridge foundation of main pier in deep-water. Fast-flowing water makes the construction even more difficult because work has to be done during the dry season. The design and construction of these cable-stayed bridges broke so many domestic records and also filled many technical blanks. Some of indexes and performance rank first in Asia and even in the world, which indicates that the construction of cable-stayed bridges at Chongqing has reached a world-class level.

4 The development and evolution of bridges at Chongqing from cultural and artistic view

It is the mountains and rivers that decide the soul of Chongqing because bridge is built over the shock of people's soul. There is deep culture connotation behind the name of "bridge city".

Chongqing, a city which carries the 3000 years civilization. When it gets dark, the lights on the bridge create such a beautiful arch that it feels like a fairyland to

linger on. People also composed a poem to praise Shibampo Yangtse River Bridge.

Bridge witnesses history because every old bridge brings us into a time tunnel to feel Chongqing civilization. Every bridge is endowed with different spirit and souls for its exquisitely carved stone and delicately designed bridgehead buildings. Place yourself in "city of bridge" to appreciate different views of ancient and modern bridges, see ancient culture, feel vicissitudes of the years, look at the views of times, sing the praise of prosperity of society, pass through ancient and modern times to continue the glory of Chongqing.

5 Application of mechanics in bridge engineering

Before the middle 19th century, the bridge construction uses support to assemble steel girder and to pour concrete girder, making the girder be free from stress in the whole construction process. The cantilever construction of bridge has moved from steel bridge to pre-stressed concrete bridge, which provides a strong construction technique support for the development of pre-stressed concrete cantilever Girder Bridge, continuous girder bridge, continuous rigid frame bridge, arch bridge, cable-stayed bridge and other long-span bridges and starts a new era of development of long-span bridges [12-15].

Construction is growing out of nothing, from natural to artificial, from low level to advanced level. Mechanics development also serves as the foundation to construction. The stability is one of the most important factors of all constructions. Although some building like the Leaning Tower of Pisa and hanging garden which look unstable, but they are firm from mechanics perspective. Now, people care about three basic factors when design bridges, namely, practicability, economization and beauty. Especially, because of complicated construction forms and great population pressure, the rationality of structure must be taken into consideration. While the construction serves its practicability, it also must serve economical applicability.

The mechanic theory in arch bridge is to transfer all or most of bending moment stress caused by load to compressive stress through horizontal thrust. Simply speaking, the longitudinal force is transferred into lateral force through internal force of bridge to work on basis of two sides of bridge, so the bridge must be constructed in areas with solid foundation. The bending moment of strut beam is the largest, followed by cantilever beam, statically determinate multi-span beam, three-hinged rigid frame, composite structure, truss and three-hinged arch with reasonable arch axis have zero bending moment) [16-18].

For example, when the superstructures (segmental beams, arch rib) need to cross deep water, valleys and navigable rivers or when bridge must be built during flood season, bridges are usually constructed without support hoisting segment by segment with cable rope. When hoist place and assemble the prefabricated parts,

the stress state of these parts is different from the working condition of finished bridge. The location and number of suspension point of parts should be assured and calculated before construction. Segment of beam and arch rib usually uses two-point suspension point. When segment or curvature of part is large, it is better to use four-point suspension point. The best location of suspension point should be based on the stability and reasonable load carrying capability when hoisting these parts although the influence of other secondary factors should be taken into consideration. The selection of the most suitable suspension point in hoisting of cable rope should be based on the maximum tensile stress as control objectives to identify the four-suspension point of section. As for indeterminate problem of variable cross-section part of two-point suspension and four-point suspension, optimization and iteration methods and summing trial methods [19] can be used to calculate the change rule of suspension point and practical results.



FIGURE 12 Hoisting of Rigid Frame Arch Bridge

If section parts are similar, then uses straight girder to calculate. When the upper and lower allocation of the section of part is equal and when two-point suspension is used, the stress feature of part is like that of double overhanging and simply supported beam. Suspension points should be symmetrically arranged. The controlling object is the absolute value of maximum negative bending moment M_1 equals the absolute value of maximum positive bending moment M_2 at mid-span, which leads to $x=0.207L$ (L is the length of arch rib part). Taking the factors such as the eccentricity tensile pressure produced by cable, numbers of reinforcement of upper and lower reason, the connector of the tip and the location of gravity of bending arch rib in to consideration, the actual location of suspension point is located $0.22L$ from the top of arch rib. (L is the length of arch rib part).

When the bridge span is long and shipping ton of hoisting equipment permits, the rib can be prefabricated according to the longer segment (especially the arch part) to avoid air operation. When the curvature of arch is large, four point hoisting and deflecting pulley is also applied. Stress bearing of four-point hoisting of parts belongs to statically indeterminate problem [21-23] and it can be analysed with statically indeterminate girder after using symmetrical semi structure. Choose the basic system as showed in figure 13, the unnecessary unknown force X_1 is the positive bending moment M_1 in the

middle of component. The force method equation is $\delta_{11}X_1 + \Delta_{1P} = 0$. Suppose distance between outside suspension point and rod end is X , for the convenience of construction, usually the distance between external suspension point and internal suspension point is $0.2L$ (when it is less than $0.27L$, it bending moment of span is small). Using graphic multiplication method to calculate key coefficient δ_{11} and free term Δ_{1P} , it can be known that:

$$M_1 = -\frac{\Delta_{1P}}{\delta_{11}} = \left(\frac{43}{6000} - \frac{13x}{200} + \frac{x^2}{5} - \frac{x^3}{6} \right) / \left(\frac{11}{30} - x \right) \quad (1)$$

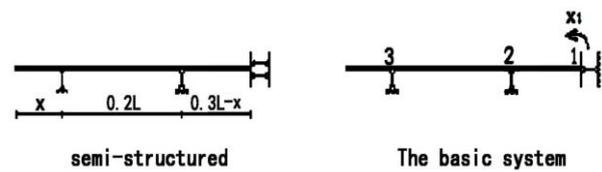


FIGURE 13 Simplification of basic system

Taking

$$|M_1| = |M_2| = |M_3|, \quad (2)$$

as controlling objective, statically indeterminate can remain unknown. Though

$$(0.3L - x)^2 / 4 = x^2 / 2, \quad (3)$$

we can directly get $x = 0.124L$.

To adapt to the change of internal force, the solid web section of vault of continuous girder, overhanging girder and trussed arch and rigid frame arch usually uses vertical variable cross-section. The rules of variation of sections are usually straight line or second degree parabola and two-point and four-point suspension are usually used. The identification of suspension point of variable cross-section, even it is two-point suspension, is a statically indeterminate Problem. Taking two-point suspension of variable cross-section part of parabola as an example (graph 14), thickness $B=1$ and unit weight $\gamma=1$ to calculate.

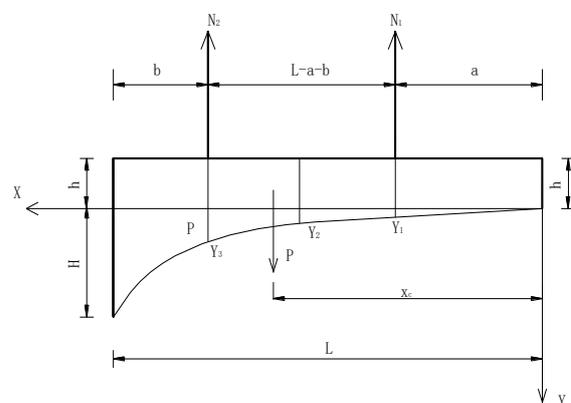


FIGURE 14 Two-point suspension of variable cross-section part

Here, the optimal controlling object is to keep the upper and lower reason of largest tensile stress of section of suspension point and section in the span as small and equal as possible, which is especially important for lightly reinforced concrete part.

Location of the centre of gravity:

$$X_c = \frac{3(2h+H)L}{4(3h+H)} \tag{4}$$

Controlling bending moment of section:

$$M_1 = \frac{a^2}{12}(6h+y_1), \quad M_3 = \frac{hb^2}{2} + \frac{HL}{12}(4b-L) + \frac{y^2}{12}(L-b)^2, \tag{5}$$

$$M_2 = -\frac{M_1+M_3}{2} + \frac{L^2}{96}(12h+y_1+10y_2+y_3). \tag{6}$$

In which:

$$y_1 = \frac{Ha^2}{L^2}, \quad y_3 = \frac{H(L-b)^2}{L^2}, \quad y_2 = \frac{H(L+a-b)^2}{4L^2}, \quad L_1 = L-a-b. \tag{7}$$

As for lightly reinforced or plain concrete, modulus of bending resistance section is

$$W_i = \frac{(h+y_i)^2}{6}, \quad (i=1, 2, 3). \tag{8}$$

Optimal object:

$$\sigma_1 = \sigma_2 = \sigma_3 \tag{9}$$

that is to say

$$\frac{|M_1|}{W_1} = \frac{|M_2|}{W_2} = \frac{|M_3|}{W_3}. \tag{10}$$

Here, the largest bending moment should be replaced with section bending moment to avoid the valuation of bending moment one by one. Error should be less than 2%. For the stability of calculation and easy controlling of preciseness of calculation, after input of constant L, H and h , the non-dimensional method is used to process. $H=0$ is uniform section. Try to calculate a and b value on computer with reasonable step size. Firstly, using starting value of a to calculate σ_1 within $0 \sim 0.5L$ or even smaller given area and using 0.618 method to iterate one by one to shorten interval. $\sigma_2 = \sigma_3$ is the objective every time you take a value an method of advance and retreat is used to get the b value, moreover, $|\sigma_1 - \sigma_3| < \varepsilon$ (ε is given preciseness) is used to control iteration. After getting required preciseness of two suspension points with possibly less times of iteration, the vertical component of cable force of two cables can be determined with the location of centre of gravity. As for variable cross-section part, the cable force of two suspension cables cannot be equal.

The reasonable section stress (some sections of suspension point and weak and sensitive sections of part) should be the controlling objective to determine the best suspension point of four-point suspension of variable section parts. Solid web section of vault usually uses four-point suspension. Using summation method to trial gets the result that satisfies the preciseness of construction [20].

Suspension bridge has long span, but it is vulnerable to wind and can even be damaged by strong wind. Wind-resistance is the key problem in the construction of this flexible bridge. Stability can be improved by the careful design in fluid mechanics. From 1818 to the end of 19th century, vibration caused by wind at least destroyed 11 suspension bridges.

After the Second World War, people started to study the destruction of Tacoma Bridge by wind. Some Aeronautical engineers believed that the vibration of Tacoma Bridge is like vibration of wing and they reproduced this wind-induced torsion divergence vibration through wind tunnel testing of bridge models. In the meantime, hydromechanics expert represented by Von Karman believes that different from stream-line wing, blunt-pointed H section of the main girder of Tacoma Bridge has obvious vortex loss, so it should be explained with vortex shedding vibration. In 1950s, the two ideas are controversial until 1963 when American professor R. Scanlan put up separated flow self-excitation flutter theory of blunt body section, the vibration mechanism causing wind destruction of Tacoma Bridge was successfully explained, which also laid the theoretical foundation for bridge flutter. Canadian professor Davenport started a method of analysing bridge buffeting with random vibration theory. This theory became more workable after being revised by Scanlan in 1977. It is safe to say that Scanlan and Davenport laid a good foundation for wind vibration theory of bridge.

Because of the participation of pier, the working condition between continuous rigid bridge and continuous girder bridge is different. As for continuous rigid bridge, positive bending moment in mid-span area caused by live load is smaller than continuous girder bridge with same span. When pier reaches a certain height, the internal force of superstructure of two bridges is nearly the same. By comparing the span between continuous rigid with three spans and continuous girder superstructure with three spans, it is found that the spans of dead and live load at the root of girder are basically the same. When the pier is 40m high, the difference between the dead and live load of mid-span of girder is less than 10%. The span of dead and live load at the root of pier of continuous rigid bridge decreases with the increase of the height of pier, but when the pier is above 40m, the rate of decrease becomes low. The axial tension of dead and live load of inner part of continuous rigid girder decreases with the increased height of pier, but when the pier is above 30 m, the rate of decrease becomes low.

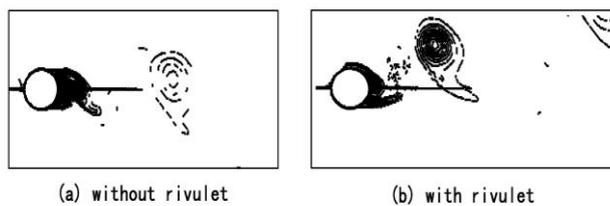


FIGURE 15 Contour line of streaming vorticity

Except for the whole vibration of bridge structure, the cable of cable-stayed bridge also vibrates with the strong wind and heavy rain, which means that the rope of cable-stayed bridge vibrates heavily with the coming of rainstorm and strong wind. This interesting phenomenon was firstly found in 1990s. It can be explained that the rain flows along cable to make a raised waterway called rivulet on the cable, destroying the rounded profile of section of cable. With rainstorm and strong wind, when the raised shape of rivulet suits its position on the cable, a self-induced mechanism is formed, producing self-induced vibration. Professor Ren Wenmin from department of engineering mechanics of Qinghua University used the method of calculating aerodynamic to guide his postgraduates to calculate two-dimension flow field around the section of cable. It is found that the rising of rivulet changes aerodynamic performance of cable section, which on the one hand, leads to the sharp decrease of frequency of alternating falls off in wake zone without rivulet, on the other hand, leads to the trace of fall off of vortex drifts horizontally. The former causes change of resonance wind speed and the later leads the rising of negative damping, increasing amplitude continuously. The following pictures are contour map of streaming vorticity calculated by Qinghua University. 15a is contour map of streaming vorticity without rivulet. 15b is the contour map of streaming vorticity with rivulet.

References

- [1] Chen Shen, Chen Huanjing 1998 Mechanical Principle in the Design of Pile Frame-Bent Abutment *Fuzhou University* 19(2) 81-4 (In chinese)
- [2] Chen Shen, Chen Peijian 1999 Design scheme optimization and modification of Mountainous Highway Bridge Structure *Engineering Mechanics* 25(8) 174-9 (In chinese)
- [3] Lin Zhangchuan 1993 A Discussion on the Applicability and Structure Type of concrete suspension bridge *Chinese Journal of Highway* 93(4) 39-44 (In chinese)
- [4] Huang Wenji, Qi Guangrong, Liang Tianxi 1994 Fujian Province Suspension Bridge Engineering *China Civil Engineering Society of Bridge and Structural Engineering 11th Annual Meeting Proceedings 12, Shantou* (In chinese)
- [5] Ding Han Shan, Huang Wen Ji 1989 The design of Yangtougou Grade Ring Bridge by Bi-Spline Subdomain Method *Chinese Journal of Highway* 89(4) 60-8 (In chinese)
- [6] Qin Rong 1985 Spline Function and Its Network Progress on Structural Mechanics Nanning: *Guangxi People's Publishing House* (In chinese)
- [7] Fujian 2007 Communications Planning and Design Institute *Research Report on Application of spatial pre-stressed concrete technology* in October, 2007 (In chinese)
- [8] Chen Shen, Liu Liesheng, Cao Ying Xuan. 2001 Cast-in-place box beam formwork system of high pier continuous rigid frame *Engineering Mechanics Supp.* (In chinese)
- [9] Institute of Structural Engineering 2009 Fuzhou University *The Qingzhou viaduct support test report* 2001 in October 2009. (In chinese)
- [10] Fan Lichu 1988 Pre-stressed concrete continuous beam bridge *Beijing: People's Communications Press* (In chinese)
- [11] Guo Jinqiong 1991 *Design theory of box girder* People's Communications Press: Beijing (In chinese)
- [12] Luo Ying 1959 *Chinese stone Bridges* People's Communications Press: Beijing (In chinese)
- [13] Qian Lingxi Bearing Capacity Analysis of Zhaozhou Bridge *Civil Engineering Journal* 1987(4) 39-48 (In chinese)
- [14] Zheng Zhenfei, Peng Dawen 1985 Non-Linear Analysis of Indeterminate Reinforced Concrete Arches *Journal of Fuzhou University* 5(2) 116-27 (In chinese)
- [15] Chen Shen, Liu Liesheng, Cao Ying Xuan 2003 Study on prestressed composite truss-type corbelling system for cast-in-place box girders *Civil Engineering Journal* 28(15) 185-91 (In chinese)
- [16] Liu Zhi. Study 2001 Vehicle vibration problems of rigid arch bridge and Study on the static and dynamic characteristics *Journal of Fuzhou University* 14(2) 187-92 (In chinese)

The two pictures show the situation when a vibration period gets started. The raised part above the cable of rounded section is rivulet.

In 21st century, China will build long span bridges, which calls for us to do deeper and further basic and applied researches on identification of aerodynamic parameter of long span bridge, nonlinear wind vibration theory, wind tunnel simulated experiment and data analysis.

6 Conclusions

The development of bridges is summarized from two aspects in this chapter with emphasis on world bridges and Chongqing bridges. Firstly, by a brief introduction and summary of bridge types, theories of bridge design and evolution of bridge construction techniques and the development and evolution of bridges around the world are introduced. Furthermore, the types of bridges, development of theories of bridge design and the breakthrough of bridge construction techniques with time move on are all systemically clarified. Secondly, the process of Chongqing bridge development and evolution is emphasized. On the one hand, when fighting with nature, man remoulded nature and constructed bridge for the convenience of their trips. The progress the man made in the process of remoulding nature is revealed. On the other hand, from the perspective of economy and techniques, this paper clarified the positive role that economical and technical development plays in promoting bridges construction. Finally, the function of mechanics in bridge design is analysed and the development of bridge is experienced from the aspect of culture and art and aesthetics and spirit it brings to people.

[17]Chen Shen, Li Zhengliang 1999 Studies of Capacity Formula for Digging and Filling Pile *Journal of Fuzhou University* 14(3) 62-7 (In chinese)

[18]Lin Zhiliang 1997 Design of main part of Wulongjiang Bridge Highway 5(6) 6-12 (In chinese)

[19]Wang Quanfeng 1995 Basic structural optimization method and its application in high-rise buildings *Journal of Xiamen University* 4(4) 47-52 (In chinese)

[20]Chen Shen 1999 The problem of the optimum lifting point of the suspension cable *Journal of Zhejiang University* 1(45) 112-6 (In chinese)

[21]Lei Junqing 1998 cantilever bridge construction and design *Journal of Chongqing Jianzhu University* 65(11) 35-46 (In chinese)

[22]Xiang Zhongfu 2001 Bridge construction control technique *Journal of Peking University* 24(8) 17-25 (In chinese)

[23]Xu Yongming, Li Jian 1995 Construction control of Long Span Prestressed Concrete Bridges *Journal of Tongji University* 7(21) 24-8 (In chinese)

[24]Xu Junlan 2000 Construction control of long span bridges *Journal of Chongqing University* 17(8) 247-52 (In chinese)

| Authors | |
|--|--|
|  | <p>Yan Li, born in April, 1979, ChongQing City, P.R. China</p> <p>Current position, grades: The College of Civil Engineering, Chongqing University, University studies: received her B.Sc. from Anhui Normal University in China. She received her M.Sc. from Southwestern University in China. Scientific interest: Her research interest fields include Aesthetics. Experience: She has teaching experience of 10 years, has completed one scientific research projects.</p> |
|  | <p>Dong Wei, born in March, 1962, ChongQing City, P.R. China</p> <p>Current position, grades: the senior engineer of The College of Architectural Engineering Vocational, Chongqing University studies: received his B.Sc. from Southwestern University in China. Scientific interest: His research interest fields include Landscape plants, landscape design, garden construction Publications: more than 20 papers published in various journals. Experience: He has teaching experience of 20 years, has completed ten scientific research projects.</p> |
|  | <p>Yangyang Chen, born in March, 1988, Jing County, HeBeing Province, P.R. China</p> <p>Current position, grades: the student of School of Civil Engineering, Shijiazhuang Tiedao University, China. Scientific interest: His research interest fields include Healthy Monitoring</p> |

A comparative study on efficiency of two different circulation modes of agricultural products based on DEA model: wholesale market and logistics distribution centre

Yaoting Chen¹, Xiaowei Lin^{2*}

¹*School of history and social development of Minnan Normal University, No.36, Xian Qian Road Zhangzhou in Fujian province, China*

²*School of Management of Minnan Normal University No.36, Xian Qian Road Zhangzhou in Fujian province, China*

Received 6 June 2014, www.tsi.lv

Abstract

The purpose of this study is to find out the relatively efficient circulation mode of agricultural products through a comparative analysis on the operating efficiency of two different circulation modes of agricultural products: wholesale market and logistics distribution centre. Based on the input and output data collected from the survey of the main representatives of enterprises in the two modes in Zhangzhou, Fujian, including: fixed assets, number of employees, main business cost, main business net profit, gross margin, the paper uses Data Envelopment Analysis to conduct the analysis. The results show that the third party logistics mode based on logistics distribution centre is relatively more efficient, comparing with the traditional wholesale market mode. Therefore, in order to reduce circulation cost of agricultural products, and to promote the development of agricultural industry, it is necessary to make policies to encourage the development of the third party logistics mode based on logistics distribution centre.

Keywords: data envelopment analysis model, wholesale market mode, logistics distribution centre mode, operation efficiency

1 Introduction

Bain [1] analysed industry performance by using the S-C-P mode in the industrial organization, which is used to evaluate circulation efficiency by agricultural sector. Among them, S (structure) reflects the organizational characteristics of market; C (conduct) is the marketing strategies adopted by enterprises; P (performance) is the result of the overall operation of industry. Stern [7] pointed out that the circulation channel performance included effect, fairness and efficiency, in which the effect reflected accessibility and incentives, while efficiency reflected productivity and profitability. Chen [2] proposed the efficiency of food marketing system could be studied through the analysis of system and organization structure. Subsequently, Enke [3] proposed the spatial price equilibrium theory, based on which Samuelson [6], Takayama [8] put forward the commodity's price is decided by the transfer cost in a competitive market. Clark [4] considered that market efficiency could be studied from two aspects, namely the individual enterprises and the public. From the perspective of the enterprises, market efficiency were evaluated mainly through profit or cost; while from the perspective of the public, market efficiency were evaluated through circulation service level and circulation cost.

At present, both at home and abroad, many scholars have conducted fruitful researches on efficiency of circulation channels of agricultural products [2, 9-11].

Taking chickpea as an example, [8] used empirical analysis approach to analyse the circulation efficiency of different distribution channels, and the results of the study indicated that: farmers were facing a lot of problems in the circulation, and the circulation cost was the main part of the price of agricultural products. Based on the field survey of 112 retailers randomly selected from 6 farmers markets in 6 districts of Nanjing City, Zhou and Lu [11] chose some indicators such as the circulation level and cost, loss rate, producers share ratio and so on to analyse the circulation efficiency of fresh vegetables in different circulation channels in Nanjing City. Wang [9] combined with the results of the survey on large-scale circulation enterprises of fresh agricultural products in Wuxi and conducted the efficiency evaluation model of the supply chain management and the traditional circulation mode of fresh agricultural products, through the construction of input-output, efficiency evaluation system of circulation mode of fresh agricultural products and comprehensive preference cone model. Yang and Xiao [10] investigated and analysed the subject of circulation in each circulation link of the three kinds of circulation channels of grape circulation in Jinzhou city. Through the in-depth interview on the subject of circulation, she analysed the circulation cost and efficiency in different circulation channels. Chen, Cai and Dai [2] conducted a detailed analysis on the interior and exterior transaction cost of the three major circulation modes: farmer-spontaneous circulation mode, cooperatives-oriented circulation mode and contract-

* *Corresponding author* e-mail: 170000369@qq.com

oriented vertical circulation mode, and then found out the factors influencing the transaction efficiency. Finally, it put forward some countermeasures to enhance the transaction efficiency of each circulation mode.

Owing to the difficulty to obtain the detail research data, the study on the comparative efficiency evaluation of wholesale market or the third party logistics of agricultural products mainly stays in the qualitative research, namely, the research on quantitative analysis is relatively less. Based on the previous studies, the paper uses Data Envelopment Analysis (DEA) approach to analyse the efficiency level of wholesale markets mode and the third party logistics mode of agricultural products, hoping the result can provide the reference basis for the management decision-making.

2 The basic situation and sample selection

Zhangzhou City, located in the south-east of Fujian Province, is the largest plain in Fujian - a total area of 12,600 square kilometres with nearly 4,810,000 residents. As the "Flower and Fruit City", Zhangzhou City has a superior geographical position. It is the largest agricultural export base in Fujian Province and the key national export-oriented agriculture demonstration area. Its total agricultural output value in 2012 reached 55,885,000,000 Yuan. Its per capita of agricultural products such as fruit, aquatic products, vegetables, edible mushroom (fresh) and the flower ranks the forefront of districts, especially the fruit production in 2011 reached 2,726,600 tons. Its output and export volume exceeded more than half of the similar products of the province, while the banana is the greatest in production in Zhangzhou. The fruit of Zhangzhou export a lot every year. Therefore, the scale of the wholesale, transportation, warehousing, distribution of agricultural products promote the rapid development of agricultural products logistics in Zhangzhou City. The "Twelfth Five Year Plan" modern agricultural development planning of Zhangzhou City shows that there are 25 wholesale markets of all types of agricultural products in Zhangzhou City, covering the main agricultural products of the region. However, there exist some problems in the traditional wholesale mode in the actual operation, such as: the irrational allocation of resources, not timely and comprehensive information feedback, and the difficulty in meeting the needs of the development of modern logistics of agricultural products. Meanwhile, the third party logistics transformed from large wholesalers are developing rapidly. The logistics distribution centre of agricultural products sets acquisition, storage, distribution processing and distribution as one, and has gradually become an important bridge for agricultural super docking of supermarket. Many scholars believed that the development of the logistics distribution centre mode in China could adapt to the agricultural super docking, could improve circulation efficiency and reduce the circulation cost. However, many of these studies were qualitative descriptions, lack of empirical analysis.

Therefore, this paper chooses 10 representatives: 6 wholesale markets and 4 third party logistics companies of agricultural products in Zhangzhou City, using Data Envelopment Analysis to evaluate the operating efficiency of the traditional wholesale market mode and the third party logistics. The 10 economies are: ZQN wholesale market, ZPB wholesale market, MLN wholesale market, ZLS wholesale market, ZSS wholesale market, ZCT wholesale market, ZYS company, ZTL company, ZJX company and ZWG company.

3 Main circuit structure

3.1 RESEARCH APPROACH

Data Envelopment Analysis can be used for relative efficiency evaluation of complex system with multiple input and output. There are two basic models: CCR and BCC. In the paper, the CCR model is used, a multi-input and multi-output efficiency evaluation model under the assumption of a certain scale of reward. The basic ideas of the model are as follows: the model assumes there are n objects to be evaluated, called the decision making units (DMU), and there are m input and s output in each decision making unit. X_{ij} indicates the i input of the j decision-making unit, and Y_{rj} indicates the r output of the j decision-making unit, so the input and output vector of DMU $_j$ respectively can be expressed as:

$$X_j = (x_{1j}, x_{2j}, \dots, x_{mj})^T, j = 1, 2, \dots, n, \quad (1)$$

$$Y_j = (y_{1j}, y_{2j}, \dots, y_{sj})^T, j = 1, 2, \dots, n, \quad (2)$$

where v_i stands for input weight in i , and u_r stands for output weight in r . Therefore, the weight vector of input and output can be expressed as the following:

$$v = (v_1, v_2, \dots, v_m)^T, \quad (3)$$

$$u = (u_1, u_2, \dots, u_s)^T. \quad (4)$$

So the efficiency evaluation index for DMU $_i$ is:

$$h_j = \frac{u^T y_j}{v^T x_j}, j = 1, 2, \dots, n. \quad (5)$$

Among them, h_j is variable. Adding u and v , it makes $h_j \leq 1$. The larger h_{j0} is, the higher the efficiency it shows. In other words, less input can get more output. Therefore, the relative efficiency of DMU_{j0} can be made out through changes in u and v , until the maximum value of h_{j0} can be got. CCR model can be used to evaluate efficiency of each DMU. Weight coefficient is variable; the efficiency index of DMU_{j0} is the target; and all efficiency index of DMU h_j

$\leq 1, j = 1, 2, \dots, n$ are the constraints; therefore, CCR model can be constructed as the following:

$$C^2R = \begin{cases} \max \frac{u^T y_i}{v^T x_j} = V_{C^2R}^1 \\ s.t. \frac{u^T y_i}{v^T x_j} \leq 1, j = 1, \dots, n \\ u \geq 0, v \geq 0 \end{cases} \quad (6)$$

This is a fractional programming model. By using Charnes-Cooper transformation, it will be transformed into a linear programming model: if

$$t = \frac{1}{v^T x_0}, \omega = tv, \mu = tu,$$

then

$$P_{CCR} = \begin{cases} \max \mu^T y_0 = V_{C^2R}^1 \\ s.t. \omega^T x_j - \mu^T y_j \geq 0, j = 1, \dots, n \\ \omega^T x_0 = 1 \end{cases} \quad (7)$$

According to the dual nature, the dual model of the programming model is obtained. Because it is difficult to directly judge DEA efficiency, the concept of non-Archimedean infinitesimal is usually introduced, which makes up the CCR duality programming model with non-Archimedean infinitesimal:

$$D_{CCR} = \begin{cases} \min [\theta - \varepsilon (e^{-T} S^- + e^{T^+} S^+)] \\ s.t. \sum_{j=1}^n X_j \lambda_j + S^- = \theta X_0 \\ \sum_{j=1}^n Y_j \lambda_j - S^+ = Y_0 \\ \lambda_j \geq 0, j = 1, 2, \dots, n \\ S^+ \geq 0, S^- \geq 0 \end{cases} \quad (8)$$

Among them, ε is the amount of non-Archimedean infinitesimal, while $e^{-T} = (1, 1, \dots, 1)$, $e^{T^+} = (1, 1, \dots, 1)$, S^- and S^+ are the slack variables.

In the case of the CCR, DEA's economic meaning is: it is the technology efficiency, but also the scale efficiency. The former indicates that the output has reached the maximum, comparing with the input; and the latter means that the input is in the state of constant returns to scale. θ ($0 \leq \theta \leq 1$) the relative efficiency of DMU_j , is the reflection of the state of resource allocation efficiency in j decision-making unit. The bigger θ is, the higher relative resource allocation efficiency is, which means that the allocation of resources is more reasonable, and vice versa. Specifically, supposing $\lambda^*, S^{+*}, S^{-*}, \theta^*$ as the optimum solution of the model, there are:

1. when $\theta^* = 1, S^{+*} = 0, S^{-*} = 0$, the decision-making unit is DEA efficient, which shows that the output has reached the optimal state under the present input level.
2. when $\theta^* = 1, S^{+*} \neq 0, S^{-*} \neq 0$, the decision-making unit is weak DEA efficient, which shows that it is necessary to adjust the input or output.
3. when $\theta^* < 1$, the decision-making unit is non DEA efficient, which shows that there must be a redundant input or output deficiency [5].

3.2 SETTING OF EVALUATION INDEX AND DATA SOURCES

Based on qualitative and quantitative principles, and systematic and comprehensive principle, this paper selects the total value of fixed assets, number of employees and the main business cost as the input index, main business net profit and gross margin of the main business as the output index. The specific meanings of each index are as follows:

1. The total value of fixed assets: it reflects the amount of capital investment. For the circulation subject of agricultural products, it includes transportation, warehouse, related equipment etc.
2. The number of employees: it reflects the labour input. The agricultural products circulation process requires a lot of manpower, therefore the number of employees in the enterprise is considered as an important input variable.
3. The main business cost: it reflects the direct input, the cost of resource value to get the business income.
4. The main business net profit: it reflects the output efficiency under a fixed input. The higher the main business profit is, the more efficient business shows.
5. The gross margin of the main business: this paper selects the economy's gross margin of the main business to reflect the proportion of the management expenses. Combined with the research results of these 10 economies, the data are sorted out as shown in Table 1.

Note: the data are from the corporate annual reports and the research results of project group of the 10 economies in 2009.

TABLE 1 Input and Output Indicators of 10 Economies in Zhangzhou

| Decision making unit | Fixed assets (millions of Yuan) | Number of employees | Main business cost (millions of Yuan) | Main business net profit (millions of Yuan) | Gross margin of the main business (%) |
|----------------------|---------------------------------|---------------------|---------------------------------------|---|---------------------------------------|
| 1 | 2200 | 265 | 1200 | 600 | 15.65 |
| 2 | 2400 | 184 | 1380 | 600 | 19.01 |
| 3 | 2450 | 140 | 735 | 560 | 20.55 |
| 4 | 5850 | 280 | 1100 | 1300 | 25.29 |
| 5 | 960 | 162 | 600 | 600 | 27 |
| 6 | 3200 | 120 | 520 | 600 | 30.26 |
| 7 | 800 | 80 | 520 | 689 | 31.31 |
| 8 | 668 | 80 | 420 | 534 | 32 |
| 9 | 1890 | 157 | 881 | 675 | 21.03 |
| 10 | 596 | 114 | 340 | 350 | 24.52 |

The numbers represent respectively the related decision-making units:

1. ZPB wholesale market;
2. MLN wholesale markets;
3. ZLS wholesale markets;
4. ZJX Company;
5. ZCT wholesale market;
6. ZQN wholesale market;
7. ZWG Company;

8. ZYS Company;
9. ZSS wholesale market;
10. ZTL Company.

4 Empirical results

Take the decision-making unit 1 for example, then the following equation can be obtained:

$$D_{CCR} = \begin{cases} \min \left[\theta - 10^{-10} (S_1^- + S_2^- + S_3^- + S_1^+ + S_2^+) \right] \\ 2200\lambda_1 + 2400\lambda_2 + 2450\lambda_3 + 5850\lambda_4 + 960\lambda_5 + 3200\lambda_6 + 800\lambda_7 + 668\lambda_8 + 1890\lambda_9 + 596\lambda_{10} + S_1^- - 2200\theta = 0 \\ 265\lambda_1 + 184\lambda_2 + 140\lambda_3 + 280\lambda_4 + 162\lambda_5 + 120\lambda_6 + 80\lambda_7 + 80\lambda_8 + 157\lambda_9 + 114\lambda_{10} + S_2^- - 265\theta = 0 \\ 1200\lambda_1 + 1380\lambda_2 + 735\lambda_3 + 1100\lambda_4 + 600\lambda_5 + 520\lambda_6 + 520\lambda_7 + 420\lambda_8 + 881\lambda_9 + 340\lambda_{10} + S_3^- - 1200\theta = 0 \\ 600\lambda_1 + 600\lambda_2 + 560\lambda_3 + 1300\lambda_4 + 600\lambda_5 + 600\lambda_6 + 689\lambda_7 + 534\lambda_8 + 675\lambda_9 + 350\lambda_{10} - S_1^+ = 600 \\ 15.65\lambda_1 + 19.01\lambda_2 + 20.55\lambda_3 + 25.29\lambda_4 + 27.00\lambda_5 + 30.26\lambda_6 + 31.31\lambda_7 + 32.00\lambda_8 + 21.03\lambda_9 + 24.52\lambda_{10} - S_2^+ = 15.65 \\ \lambda_j \geq 0, j = 1, 2, \dots, 10 \\ S_1^+, S_2^+, S_1^-, S_2^-, S_3^- \geq 0 \end{cases} \quad (9)$$

The results of other decision making units can be obtained in the same way. By using Matlab software, the calculated results are shown in Table 2. From the results, it is found that θ of ZTL Company, ZJX Company, ZWG Company and ZYS Company equal to 1, and $S^+=0, S^-=0$, which constitute the DEA efficient frontier of the whole evaluation model. That is to say, from the point of view of the above three inputs and two outputs, the operation conditions of the 4 economies are DEA efficient; the resources are fully utilized, and the output is high; therefore, this is a kind of ideal condition. Similarly, according to Table 2, the evaluation results of the 6 DEA economies can be obtained: θ of ZQN wholesale market, ZPB wholesale market, MLN wholesale market, ZLS wholesale market, ZSS wholesale market, ZCT wholesale market are respectively: 0.853, 0.804, 0.763, 0.816, 0.732, 0.928, compared with the other 4 economies, the operation

conditions of the 6 economies are relatively inefficient.

In addition, according to the slack variables in Table 2, it comes out that it is necessary for the economies to continue to improve in each input and output indicators. Specifically, there exist problems in ZLS wholesale market such as employee redundancy and too high main business cost, but low main business net profit; there exist problems in MLN wholesale market and ZQN wholesale market such as employee redundancy, low main business net profit and low gross margin of the main business. While the input and output of the other 4 economies are more proportionate.

5 Analysis and Discussion

From the above evaluation results, it is not difficult to find out that, among the 10 economies, only the operation

efficiency of ZJX Company, ZWG Company, ZYS Company and ZTL Company is efficient, while the operation efficiency of the other 6 economies is relatively low. However, how much is the distance away from DEA efficiency? How to improve it? Therefore, the paper conducts the projection analysis. Taking the project analysis of the input cost as an example, the specific

approach is: taking 4 efficient decision making units as a reference, the DEA evaluation values of the other 6 decision making units are set to be 1, and suppose that the value of the fixed assets, number of employees, the main business net profit and the gross margin of the main business of the 6 units were unchanged.

TABLE 2 The results of DEA Efficiency Evaluation of 10 Decision Making Units

| Decision making unit | θ | Input slack variables | | | Output slack variables | | Relative efficiency |
|----------------------|----------|-----------------------|-------|-------|------------------------|--------|---------------------|
| | | S1 | S2 | S3 | S*1 | S*2 | |
| 1 | 0.386 | 0.124 | 0.022 | 0.000 | 0.000 | 16.056 | inefficient |
| 2 | 0.435 | 0.287 | 0.000 | 0.113 | 38.213 | 12.526 | inefficient |
| 3 | 0.594 | 0.766 | 0.003 | 0.000 | 0.000 | 11.334 | inefficient |
| 4 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | efficient |
| 5 | 0.771 | 0.016 | 0.045 | 0.000 | 0.000 | 4.706 | inefficient |
| 6 | 0.890 | 2.122 | 0.027 | 0.000 | 0.000 | 1.446 | inefficient |
| 7 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | efficient |
| 8 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | efficient |
| 9 | 0.580 | 0.308 | 0.011 | 0.000 | 0.000 | 10.342 | inefficient |
| 10 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | efficient |

Project the input cost of the main business to the efficient frontier, and it is obvious to notice the gap between the actual input cost of the decision making units

and DEA efficient frontier. The results of the projection analysis are shown in Table 3.

TABLE 3 Projection Analysis Results of the Main Business of 10 Decision Making Units

| Decision making unit | Fixed assets (millions of Yuan) | Number of employees | Main business cost / (millions of Yuan) | Main business net profit / (millions of Yuan) | Gross margin of the main business (%) |
|----------------------|---------------------------------|---------------------|---|---|---------------------------------------|
| 1 | 2200 | 265 | 463.9 | 600 | 15.65 |
| 2 | 2400 | 184 | 600.2 | 600 | 19.01 |
| 3 | 2450 | 140 | 436.6 | 560 | 20.55 |
| 4 | 5850 | 280 | 1100 | 1300 | 25.29 |
| 5 | 960 | 162 | 462.6 | 600 | 27 |
| 6 | 3200 | 120 | 462.8 | 600 | 30.26 |
| 7 | 800 | 80 | 520 | 689 | 31.31 |
| 8 | 668 | 80 | 420 | 534 | 32 |
| 9 | 1890 | 157 | 510.9 | 675 | 21.03 |
| 10 | 596 | 114 | 340 | 350 | 24.52 |

Figure 1 is the comparison of ideal input cost and actual input cost of 10 decision making units. It is obviously that there exists a gap between the actual main business input cost and the ideal one of each decision making unit. Similarly, projection analysis can be done to other input or output indicators, in order to study the gap between the actual value of each index and the ideal one, which also can provide decision-making reference for business subject to improve the efficiency.

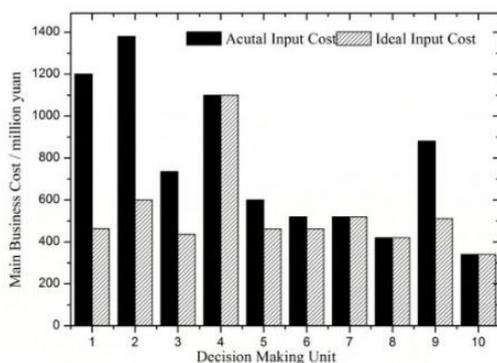


FIGURE 1 Comparison of Ideal Input Cost and Actual Input Cost of 10 Decision Making Units

The results of the efficiency evaluation of DEA show that the scale and technology of ZJX Company, ZWG Company, ZYS Company and ZTL Company are relatively efficient. In contrast, there exist some problems in the operation efficiency of the other 6 economies such as input redundancy or output shortage in different degree. These 6 economies are the traditional logistics economies of agricultural products -- wholesale markets. Through the projection analysis, it can be found that there are some problems in the 6 traditional economies such as employee redundancy, high operating cost, and low margin. Through comparative analysis, it can be found that the traditional mode of wholesale markets is of smaller scale and more dispersed, while the third party logistics is relatively efficient. Therefore, it is suggested that Zhangzhou City should devote much more effort to developing the third party logistics of agricultural products, so as to improve the overall operation efficiency of the logistics of agricultural products.

6 Conclusion

Based on DEA, the paper evaluated the circulation efficiency of the traditional wholesale market mode and the third party logistics. The circulation input and output data were collected from the 10 representative wholesale markets and the third party logistics companies of agricultural products in Zhangzhou City in 2009. The results showed that the third party logistics mode based on logistics distribution centre was more efficient comparing

with the traditional wholesale market mode, while the problems that restricted the circulation efficiency of traditional wholesale market mode mainly existed in the employee redundancy, high operating cost and low gross margin.

7 Acknowledgement

This work was supported by Fujian Science & Technology Project (No: 2014R0080)

Reference

- [1] Bain J S, 1959 Industrial Organization *New York John Wiley & Sons* 40
- [2] Chen Y, Cai X, Dai J 2013 An Analysis on Transaction Efficiency of Different Circulation Modes of Agricultural Products *Agricultural Economy* **11**(1) 120-22 (in Chinese)
- [3] Enke S 1951 Equilibrium among Spatially Separated Markets: Solution by Electric Analogue *Econometrica* **19**(1) 40-7
- [4] Clark F E 1978 Principles of Marketing *New York Ayer Co Pub*
- [5] Qu H, Gao W 2009 On the Application of Data Envelopment Analysis in the Evaluation of Input and Output Efficiency of Graduate Education *Journal of Beijing Institute of Technology (Social Sciences Edition)* **53**(6) 26-30 (in Chinese)
- [6] Samuelson P A 1952 Spatial Price Equilibrium and Linear Programming *American Economic Review* **24**(42) 283-303.
- [7] Stern L W, El-Ansary A L 1982 Marketing Channels Englewood Cliffs NJ Prentice Hall 56-60 (in Chinese)
- [8] Takayama T, Judge G G 1964 Equilibrium among Spatially Separated Markets: a Reformulation *Econometrica* **32**(10) 510-24
- [9] Wang B, 2008 Research on Efficiency Evaluation of Fresh Agricultural Products Circulation Modes Based on the Comprehensive Cone of DEA-Preference Model *Journal of Anhui Agricultural Science* **36**(12) 5176-81 (in Chinese)
- [10] Yang Y, Xiao Q 2011 A Comparative Study on the Circulation Cost and Efficiency of Agricultural Products under Different Circulation Channels-based on the Case of Grape Circulation of Jinzhou, *Agricultural Economy Issue* **36**(2) 79-88 (in Chinese)
- [11] Zhou Y, Lu L 2008 Study on Supply Chain Efficiency of Fresh Vegetable - a Case Study of Nanjing *Jiangsu Agricultural Sciences* **36** (1), 69-72 (in Chinese)

| Authors | |
|---|--|
|  | <p>Chen Yaoting, born in September, 1979, Xiangcheng District, Fujian Province, China</p> <p>Current position, grades: Associate Professor of School of history and social development Minnan Normal University, China University studies: Ph.D in management at Fujian Agricultural and forestry University in China. Master Degree in management at Tianjin University in China. Bachelor Degree in management from Fujian Normal University in China. Scientific interest: Circulation of agricultural products Publications: 13 papers published in various journals Experience: Teaching experience of 9 years, has completed 10 scientific research projects</p> |
|  | <p>Xiaowei Lin, born in November, 1974, Xiangcheng District, Zhangzhou City, Fujian Province, China</p> <p>Current position, grades: Associate Professor of School of Management, Minnan Normal University, China University studies: Ph.D. in Management at Jiangxi University of Finance and Economics in China. B.Sc. at Fujian Normal University in China Scientific interest: LSCM, Cloud Service and Cloud computing Publications: More than 20 papers published in various journals Experience: Teaching experience of 20 years, has completed 5 scientific research projects</p> |

A study on mechanism of environmental protection industry innovation under open innovation - the intermediary effect based on the enterprise network dynamic capability

Qing-huang Huang*, Ming Gao

College of Economics and Management, Fuzhou University, Fuzhou City, Fujian Province, China, 350108

Received 8 July 2014, www.tsi.lv

Abstract

In the dynamically changing external environment, it is the core issue of enterprise innovation strategy that how enterprises maintain a sustained level of innovation by creating their own capabilities. In addition, the open innovation proposed by Chesbrough provides a new way of thought for innovation management. This essay constructs conceptual models of several sets of variables relationships between environmental protection industry innovation performance and external innovation resources, which is based on 85 environmental protection enterprises as the questionnaire objects and a path analysis of the model is conducted. The results show that the cooperation with horizontal and vertical enterprises can significantly affect innovation performance only by virtue of the intermediary effect of the enterprise network dynamic capability, and government-industry-academy-research cooperation can directly improve innovation performance. Mechanistic study not only reveals that the joint action by external innovation resources and network dynamic capabilities can influence the innovation and motivation of environmental protection enterprises, but also reflects that a major source of environmental protection innovation is the internal resources. This provides theoretical guidance for enterprises to effectively implement the open innovation strategy in the innovation practice.

Keywords: open innovation, environmental protection industry, innovation performance, network dynamic capability

1 Introduction

In the knowledge innovation era, technical innovation is the key to enterprises' sustainable development. However, the market requirements and high uncertainty of the environment gradually make the innovation a complexity activity, and this urges the transformation from the traditional closed innovation model to a new open innovation model. Under the current situation of the global innovation, technical innovation of environmental protection (hereinafter referred to as EP) enterprises is facing opportunities and challenges. With the gradually increasing national focus on environmental pollution, ecological destruction and comprehensive utilization of resources, EP enterprises also introduce more new technologies, new products and new concept to satisfy the requirements of the country and the society for environmental protection, which greatly promotes the technical innovation level of EP enterprises. However, there are still many problems in such enterprises' technical innovation. Most EP enterprises are not capable to take the initiative for high-level innovation. Compared with other enterprises, internal research and development of such environmental protection enterprises are of more uncertainty mainly in: uncertainty of the achievement of the R & D goals, uncertainty of the business application of final products, skills and technologies, uncertainty of the project profitability. For the purpose of solving these

problems, this essay adopts the Chesbrough open innovation theory based on strategic alliance theory [1], that is, to obtain innovation resources outside the enterprises to make up for the inadequacy of innovation resources and capability, to decrease the uncertainty in R & D process and of the results, so as to enhance the enterprises' innovation level.

Open innovation comprehensively utilizes the marketed channels inside and outside enterprises to serve the innovation activities and coordinates the internal and external resources to create innovation ideas, so as to realize the innovation activities in the shortest time and at the least cost. Therefore, open innovation is a dynamic process of synergy, interaction and integration of various elements. Open innovation model means that the enterprises will carry out little independent innovation, but cooperate with horizontal enterprises and vertical enterprises and base on the government-industry-academy-research cooperation to obtain new products, skills, technologies and ideas. Meanwhile, acquisition and integration of external innovation elements require the enterprises to be equipped with some Internet dynamic capability. And the Internet dynamic capability requires a flexibly organization and cooperation by the enterprises. With the open innovation as a new model of enterprises innovation activities, the enterprises now can A more effectively integrate innovation resources among the social resources, without any needs to enter into formal.

* *Corresponding author* e-mail: 15980210627@163.com

With the rapid development of national economy and increasing EP investment in our country, as well as policies and measures formulated by the country to encourage development of EP industry, the environmental protection industry experiences a rapid development. However, compared EP enterprises with other countries such as America, Japan and Finland, the EP enterprises in China are characterized by small size, insufficient R & D funds, low level of opening, weak capability of resource integration and lack of technical personnel, which are far from requirements of EP enterprises as the technology-intensive and capital-intensive enterprises and constrains the innovation level of EP industry. Research on innovation of EP industry in existing literature mainly focuses on policy and market. For example, Luo Jianhua proposes to promote innovation of EP industry from the aspect of policy [2]; Fu Tao indicates the swift of environmental technology innovation application from engineering, capitalization and marketization to industrialization [3]; Jiang Hongqiang and Zhang Jing proposes to promote the industrialization of EP high and new technology through the EP technology system combining industry, academy and research [4]; Dong Ying indicates that the government shall deeply understand regularity of R&D activities in EP industry and the confronting problems, establish virtuous investment mechanism for the industry innovation and gradually improve and gradually improve the Intermediate Organization of technology [5]. These researches are carried out rarely from the aspect of enterprises' micro-mechanism. EP industry has transited from the "government-led and regulations driven phase" and "mandatory institutional change and trying to utilize market mechanism phase" to "inducing institutional change and deepening development phase", therefore we should pay more attention on the independent innovation of enterprises in local markets. However, this is contradictory with the difficulty of EP enterprises to carry out independent innovation on their own, and Chesbrough open innovation model provides an idea for development of EP industry. This essay analyses the action mechanism of innovation performance of EP industry and provides theoretical guidance for EP enterprises to utilize internal and external innovation resources in accordance with their own network properties with external innovation resources as input variable, innovation performance as output variable, process from input variable to output variable as a black-box and based on the enterprise network dynamic capability as the intermediary.

2 Research model and hypothesis for open innovation mechanism

2.1 THEORETICAL MODEL

Through combing relevant domestic and foreign literature and documents and inspecting the EP enterprises' development practices, it is found that the innovation capability is the motive force for sustainable development of EP industry. With the increasing emphasis on environment protection by the country and the society, requirements for EP products and technology put increasing pressure on innovation. Insufficiency of fund and uncertainty of market profitability make the independent innovation increasingly difficult. Therefore it is necessary for EP enterprises to implement open innovation strategy to make up for the inadequate innovation capability. Open innovation emphasizes the utilization of external market-oriented channels and coordinates external innovation resources to achieve spill over of technology and knowledge through cooperation with horizontal enterprises and vertical enterprises and government-industry-academy-research cooperation, and absorb, allocate and integrate externally acquired resources with the aid of enterprise network dynamic capability, so as to accelerate the innovation of EP enterprises.

From the aspect of enterprise resource theory, discuss the role of enterprise network dynamic capability as an intermediary of external innovation resources and innovation performance. Thus, a resource meaning is attached to external innovation resources and enterprise network dynamic capability and innovation. External innovation resources are the main source of innovation elements of open innovation strategy and the efficiency of the absorption, allocation and integration of external resources by the enterprise relies on network dynamic capability. Network dynamic capability restricts the scope of external resources to be acquired by enterprises and intensity of inter-organizational exchange, which will necessarily affect the efficiency of innovation accumulation through external integration approach and then further affect the innovation performance [6]. However, for open innovation, during the process of acquiring external resources, the synergy with internal innovation resources should be emphasized. Lack of acquisition of external innovation resources makes closed innovation, while lack of cultivation of internal innovation resources makes it difficult for enterprises to own technologies and products of independent intellectual property rights. Therefore, during the research on the effect of innovation performance in EP industry under the open innovation, the innovation performance serves as the explained variable, cooperation with horizontal enterprises and vertical enterprises, government-industry-academy-research cooperation and internal resources as the explanatory variable, and enterprise network dynamic capability as the intermediary variable. Construct the

conceptual model under this essay in accordance with above analysis (see, Figure 1).

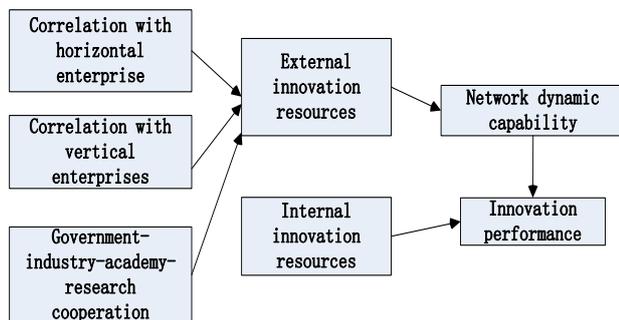


FIGURE 1 Conceptual model of open innovation mechanism and action path

2.2 HYPOTHESIS

2.2.1 Leading logic and hypothesis of external innovation resources and innovation performance

Enterprises' innovation not only needs to utilize internal resources, but also coordinate external resources. Enterprises searches for, identifies and absorbs innovation resources from outside and establishes the innovation relationship network with horizontal and vertical enterprises and government-industry-academy-research cooperation through the construction of external innovation resources searching mechanism, so as to acquire new resources, technologies, skills and concepts necessary for enterprises and ultimately promotes the improvement of innovation performance. Therefore, external innovation resources are the main channel for enterprises to acquire new technologies, markets and knowledge and the original point of innovation sources [7]. As to the dimension of external innovation, scholars divide, based on the prospective of innovation resources, the external innovation elements into horizontal cooperative enterprises, vertical cooperative enterprises, special technical organizations and other organizations which include government and venture capital enterprises [8]. EP industry is of the characteristics of special industry, and is the kind of industry form integrating policy-guiding, capital-intensive and technology-intensive. Thus government policies and measurements, venture capital enterprises and technical agencies are critical to EP enterprises. This essay divides the external innovation resources into horizontal enterprise cooperation, vertical enterprise cooperation and government-industry-academy-research cooperation in accordance with characteristics of external resources and the position in the industrial structure and based on the specialty of EP industry.

Many scholars analyse the impact of open innovation performance from the perspective of complementarities of specific cooperative object resources. Horizontally, Ritala uses game theory model to demonstrate that higher performance can be achieved by cooperation with

competitive enterprises than simply by competition, and in high-tech industries, cooperation with competitors will create breaking through innovation and incremental innovation [9]. Hagedoorn believes cooperation with complementary enterprises will realize the sharing of complementary innovation resources and innovation risks and cost, so as to shorten innovation cycle and improve innovation performance [10]. Vertically, Clark points out that enterprises can improve innovation performance by integrating technologies and standard resources of upstream providers and market resources of downstream customers in the vertical industrial chain [11]. For government-industry-academy-research cooperation, Laursen & Salter point out that intermediary organizations are capable to effectively solve the problems of diseconomy and asymmetry of information in and out of markets, and reduce the information search cost and transaction expense in market, policy, technology, funds and others, so as to improve enterprises' motivation of innovation [12]. Based on above analysis, it is hypothesized as follows:

H1a: positive correlation between innovation performance and cooperation with horizontal enterprises;

H1b: positive correlation between innovation performance and cooperation with vertical enterprises;

H1c: positive correlation between innovation performance and government-industry-academy-research cooperation.

2.2.2 Leading logic and hypothesis of external innovation resources and enterprise network dynamic capability

Selection process of external innovation resources by enterprises is not static, but varying with the change of factors such as development phase, innovation awareness and innovation capability. Enterprise network dynamic capability requires that the network, resources and capabilities of enterprises shall be able to cope with the change of external innovation environment, so as to acquire favourable external resources. Tether believes enterprises are comparatively in the external innovation environment and innovation process persists throughout the innovation system consisting of several enterprises [13], which requires enterprises be equipped with the capability to manage and coordinate innovation resources outside enterprises, that is, enterprise network dynamic capability [14]. Therefore, network dynamic capability, in essence, is closely connected with dynamics of external innovation environment. This is further confirmed by Gulati who points out that constant changing of external innovation environment objectively enhances the cooperation of technologies, knowledge and ideas between enterprises and external organizations such as providers, users and competitors. Certain network dynamic capability is required for enterprises to transform the technical spillover generated by such cooperation into actual innovation performance [15]. It is hypothesized as follows based on above analysis:

H2a: positive correlation between enterprise network dynamic capability and cooperation with horizontal enterprises;

H2b: positive correlation between enterprise network dynamic capability and cooperation with vertical enterprises;

H2c: positive correlation between enterprise network dynamic capability and government-industry-academy-research cooperation.

2.2.3 Leading logic and hypothesis of enterprise network dynamic capability and impact of innovation performance

In the fiercely competitive environment, enterprises need introduce external innovation resources constantly to make up the shortage of internal resources to improve innovation capability. And the efficiency to acquire innovation resources from external environment depends on the network dynamic capability of enterprises. The network dynamic capability is utilized to acquire, integrate and reallocate internal and external resources and cope with rapidly changing environment. Therefore, competitive advantages of enterprises are derived from network dynamic capability. Only by rapidly acquiring external innovation resources based on development level, innovation level and market environment, can enterprises effectively develop new products to satisfy various market requirements [16]. It is hypothesized as follows based on above analysis:

H3a: positive correlation between enterprises innovation performance and network dynamic capability.

Combining the above 3 groups of correlation analysis of external innovation resources and innovation performance, external innovation resources and network dynamic capability, and network dynamic capability and innovation performance, it is hypothesized as follows:

H3b: Network dynamic capability plays an intermediary role in the process of external innovation resources' positive impact on innovation performance.

3 Research design

3.1 MEASUREMENT OF VARIABLES

In innovation performance correlation model of EP industry under open innovation, measurement of variables of cooperation with horizontal enterprises and vertical enterprises, government-industry-academy-research cooperation, network dynamic capability and innovation performance. As these variables cannot be objectively and quantifiably measured, the 5-Liked Scale is adopted for subjective scoring in this research. Score of 1-5 is used to refer to the level of compliance of items from complete non-compliance to complete compliance. Based on reference, expertise and on-site investigation and research, this research adopts several items of variables for measurement.

In this research, measurement of cooperation with horizontal enterprises in external innovation resources is mainly according to Dorsey's definition of horizontal enterprises as enterprises and organizations in the same link of market system [17] as well as Chen Yufen and Chen Jing's division of horizontal enterprises [8]. It is adjusted based on broad meaning of EP industry. Consequently, 2 items of cooperation with competitive enterprises and cooperation with complementary enterprises with total 6 sub-items are determined. For the measurement of cooperation with vertical enterprises, 2 items of providers and customers with total 6 sub-items are determined based on Harabi's verification by virtue of statistic data that 84% of emerging enterprises carry out cooperation and R&D with customers or providers in vertical industrial chain direction selection [18], Liu Wei's analysis of vertical dimension in external innovation resources [19] and empirical analysis. For measurement of government-industry-academy-research cooperation, 3 items of academy, government and intermediary organization with 9 sub-items are determined based on Kong Xianghao's "Four Wheel Driven" structure model of synergy innovation for government-industry-academy-research cooperation [20], Chen Hongxi's viewpoint that academy, industry and university relationships are the three helices of innovation network [21] with communication as its core during the analysis of government-industry-academy-research cooperation model and the development phase of EP industry.

Measurement of enterprise network dynamic capability refers to the division of enterprise network capability dimensions by Damanpour F & Gopalakrishnan [22], and is adjusted based on research by Xin Qing and Yang Huixing [23]. Acquisition of external innovation resources is derived from identification of new market requirements and cognition of new technical capability; screen and assess items based on the enterprises' own resources stock and innovation level, with the standards of assessment including that there are market requirements for innovation results, enterprises are capable to carry out innovation activities and the economic feasibility of the innovation process; after the screening, transform and integrate the external innovation resources inside the enterprises to obtain network dynamic capability and achieve the constant improvement of organizational innovation capability. Based on above analysis, this essay constructs the theoretical measurement scale of enterprise network dynamic capability from three dimensions of resources searching identification, screening and assessment and transformation with 9 sub-items.

As to the measurement of innovation performance, Lichtenthaler's two items of utilizing the chance to enter new market and improving enterprises' technical level are adopted to measure the open innovation performance [24]. Cai Ning and Yan Chun points out that innovation performance is a multi-dimension structure [25], however, assessment of innovation performance in current researches emphasizes on the financial perspective

excessively. Lichtenthaler advises that enterprises should take the factor of strategic motive which cannot be neglected into consideration besides material rewards during the implementation of open innovation strategy [26]. Based on above analysis, this essay identifies 6 sub-items to construct theoretical measurement scale of innovation performance, including success rate of new projects, quantity of new products, quantity of new patents, innovation culture, leading posture of innovation management capability and skills.

To verify the effectiveness of the scale, this research carries out a pre-research before the formal questionnaire. T test is performed for key variables (average) of pre-research and formal questionnaires, and there is no significant difference. Meanwhile, carry out correlation analysis of various indicators of pre-research and formal questionnaires, and the results show significant correlation, that is, the answers by objects of research are effective.

3.2 SAMPLING

This research is targeted at EP enterprises and distributes and collects questionnaires mainly through the following 3 channels. The first channel is to utilize social relationships to distribute questionnaires to 33 EP enterprises in 3 provinces of Zhejiang, Jiangsu and Guangdong and distribute 66 questionnaires to the targeted enterprises in the way of E-mail, 59 of which are collected, including 54 effective questionnaires. The second channel is to select 45 EP enterprises and related enterprises from the member list of China Environmental Protection Association during 2012 to 2013. In this way, totally 90 questionnaires are distributed, 43 of which are collected, including 36 effective questionnaires. The third channel is to visit 7 EP enterprises listed in the Demonstration Technical Category for National Advanced Pollution Control (in 2012), deeply interview the technical professionals and managers, and deliver 14 questionnaires on site, all of which are collected. Through above three channels, there are 170 questionnaires are distributed in this research, 116 of which are collected, achieving an overall collection rate of 68.2%. Taking out 12 ineffective questionnaires, there are 104 effective questionnaires suitable for subsequent research, making up 89.6% of the total questionnaires.

3.3 VERIFICATION OF RELIABILITY AND VALIDITY

Firstly, carry out descriptive statistic analysis of variables, including maximum and minimum value, average value and standard deviation (Table 1). For external innovation resources, the maximum average of government-industry-academy-research cooperation is 3.846, which is confirmed in research. EP industry is led by policy and the industrial development is closely related to the

environmental protection of the government. As a capital-intensive industry, most enterprises in EP industry are of small size and shortage of funds, so they usually choose to cooperate with high schools and academies on innovation of products and technologies. Average of cooperation with vertical enterprises is also high ($M=3.821$), which is compliant with development tendency of EP industry, that is, beginning to pay attention to vertical industry integration strategy unifying EP technology, plan project, R&D, construction and operation, strengthening communication and cooperation among various links to effectively extend the industrial chain and obtain competitive advantages of value chain. The minimum average of cooperation with horizontal enterprises is lowest ($M=3.519$), that means, EP enterprises are usually reluctant to cooperate with competitive or complementary enterprises, so as to avoid the disclosure of core technology capability and consequently affect the competitive advantages. Average of enterprise network dynamic capability ($M=3.789$) is related to the low utilization of external innovation resources. Most enterprises have not realized the impact of open innovation on enterprises' innovation capability, and the core of enterprises is to introduce technologies, thus leading to high R&D expenses and failure to give consideration to external innovation resources. Average of innovation performance is 4.029, which means, the innovation performance is at a comparatively ideal position and the EP industry has developed from government-leading phase to combination of government and market phase. Polluters' requirements and demands for EP products are increasing which strengthens EP enterprises' motivation for investment and development.

This research studies the internal consistency reliability of the six sub-scales of cooperation with horizontal enterprises, cooperation with vertical enterprises, government-industry-academy-research cooperation, internal resources, network dynamic capability and innovation performance based on Cronbach's α standards. The results of reliability analysis are as follows: Corrected Item-Total Correlation (CITC) value falls within the scope of 0.412 ~ 0.848 and all CITC values are more than 0.35 as required; measurement variable consistency index (Cronbach's α) falls within the scope of 0.649 ~ 0.848, basically compliant with the requirements of Cronbach's α to be more than 0.7 [27]. Therefore, under open innovation, internal consistency is high among external innovation resources, network dynamic capability and innovation performance and the scale design complies with the requirements. The results generated in the way of principal components and factors analysis show: the factor load capacity of variables are all more than 0.5 and average KMO is more than 0.7, which show the scales are of comparatively high validity.

TABLE 1 Descriptive statistics and validity analysis of research variables

| Variables | Innovation Performance | Cooperation with Horizontal Enterprises | Cooperation with Vertical Enterprises | Government-industry-academy-research Cooperation | Network Dynamic Capability | Internal Resources |
|-----------------------|------------------------|---|---------------------------------------|--|----------------------------|--------------------|
| Quantity of Sub-items | 6 | 6 | 9 | 10 | 6 | 4 |
| Minimum Value | 3.000 | 2.330 | 3.000 | 2.000 | 2.670 | 3.000 |
| Maximum Value | 5.000 | 4.330 | 5.000 | 5.000 | 4.670 | 5.000 |
| Average | 4.077 | 3.519 | 3.821 | 3.846 | 3.789 | 4.029 |
| Standard Deviation | 0.383 | 0.440 | 0.548 | 0.531 | 0.415 | 0.453 |
| CITC | 0.724 | 0.412 | 0.500 | 0.591 | 0.848 | 0.584 |
| Cronbach alpha | 0.788 | 0.694 | 0.712 | 0.791 | 0.848 | 0.784 |
| Factor Load Capacity | 0.524 | 0.654 | 0.773 | 0.852 | 0.751 | 0.839 |
| Effective N | 104 | 104 | 104 | 104 | 104 | 104 |

Note: * refers to be significant at the level of 0.05 (both sides), ** refers to be significant at the level of 0.01 (both sides)

3.4 HYPOTHESIS TESTING

This essay adopts path analysis to test the conceptual model as shown in Figure 2. The variables in the model can be divided into three categories: endogenous variables, exogenous variables and unmeasured variables. Cooperation with horizontal enterprises (X_1), cooperation with vertical enterprises (X_2), government-industry-academy-research cooperation (X_3) and internal resources (X_5) are endogenous variables, as they will not be affected by other variables in the model. Network dynamic capability (X_4) and innovation performance (X_6) are exogenous variables as they will be affected by endogenous variables. At last, R_i represents the variables not appearing in the model. Path coefficient P_{ij} is used to show the relation among variables in the model, which is equivalent to the standard regression coefficient β .

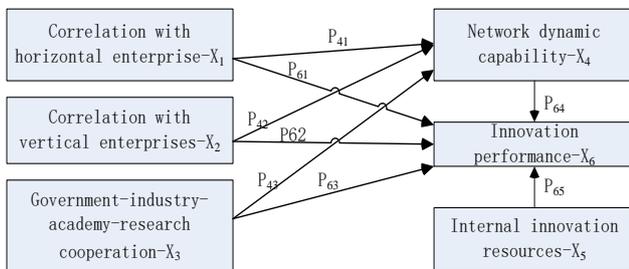


FIGURE 2 Path model

In the section of Hypothesis, we have clarified the impact of cooperation with horizontal enterprises, cooperation with vertical enterprises, government-industry-academy-research cooperation and network

dynamic capability on innovation performance. The path coefficient P_{ij} represents the regression coefficient of No. j variable to No. i variable. To get the path coefficient P_{ij} , this essay constructs the following 5 equations in combination of SPSS 19.0 to calculate the adjacent correlation coefficient between paths of external resources acquisition and innovation performance action path and standard regression coefficient of other variables.

$$X_2 = P_{21}X_1 + P_{2u}R_1, \tag{1}$$

$$X_3 = P_{31}X_1 + P_{3v}R_2, \tag{2}$$

$$X_4 = P_{41}X_1 + P_{42}X_2 + P_{43}X_3 + P_{4w}R_3, \tag{3}$$

$$X_6 = P_{61}X_1 + P_{62}X_2 + P_{63}X_3 + P_{6x}R_4, \tag{4}$$

$$X_6 = P_{65}X_5 + P_{6y}R_5. \tag{5}$$

In the above equations, X_1 represents the cooperation with horizontal enterprises; X_2 represents the cooperation with vertical enterprises, X_3 represents government-industry-academy-research cooperation; X_4 represents enterprise network dynamic capability; X_5 represents internal resources; X_6 represents innovation performance; P_{ij} represents standard regression coefficient; R_i represents standard residual. The regression results are shown as follows:

TABLE 2 Regression results

| Variables | Path Coefficient | Coefficient Value | t Value | Significance (p value) |
|---|------------------|-------------------|---------|------------------------|
| Equation 1: $X_2 = P_{21}X_1 + P_{2v}R_1$ | | | | |
| X ₁ Cooperation with Horizontal Enterprises | P ₂₁ | 0.180 | 1.542 | 0.126 |
| R-squared =0.023, Adjusted R-squared=0.013, p=0.126 | | | | |
| Equation 2: $X_3 = P_{31}X_1 + P_{2v}R_2$ | | | | |
| X ₁ Cooperation with Horizontal Enterprises | P ₃₁ | 0.027 | 0.266 | 0.791 |
| X ₂ Cooperation with Vertical Enterprises | P ₃₂ | 0.234 | 2.740 | 0.007 |
| R-squared =0.073, Adjusted R-squared=0.055, p=0.022 | | | | |
| Equation 3: $X_4 = P_{41}X_1 + P_{42}X_2 + P_{43}X_3 + P_{4v}R_3$ | | | | |
| X ₁ Cooperation with Horizontal Enterprises | P ₄₁ | 0.174 | 2.513 | 0.014 |
| X ₂ Cooperation with Vertical Enterprises | P ₄₂ | 0.267 | 4.723 | 0.000 |
| X ₃ Government-industry-academy-research Cooperation | P ₄₃ | 0.364 | 6.475 | 0.000 |
| R-squared=0.534, Adjusted R-squared=0.520, p=0.000 | | | | |
| Equation 4: $X_6 = P_{61}X_1 + P_{22}X_2 + P_{63}X_3 + P_{6v}R_4$ | | | | |
| X ₁ Cooperation with Horizontal Enterprises | P ₆₁ | 0.078 | 1.183 | 0.240 |
| X ₂ Cooperation with Vertical Enterprises | P ₆₂ | 0.086 | 1.388 | 0.168 |
| X ₃ Government-industry-academy-research Cooperation | P ₆₃ | 0.167 | 2.154 | 0.034 |
| X ₄ Network Dynamic Capability | P ₆₄ | 0.259 | 3.127 | 0.000 |
| R-squared=0.528, Adjusted R-squared=0.505, p=0.000 | | | | |
| Equation 5: $X_6 = P_{65}X_5 + P_{6v}R_5$ | | | | |
| X ₅ Internal Resources | P ₆₅ | 0.507 | 7.907 | 0.000 |
| R-squared=0.480, Adjusted R-squared=0.465, p=0.000 | | | | |

From above table, there exists correlation between adjacent sequence in conceptual model, and such correlation may be the indirect factor to affect the interaction of variables in next link. Therefore, direct, indirect, implied, and unanalysed effect relationship and correction shall be comprehensively taken into consideration. The indirect impact coefficient in the model means the effect of the variable on other variables to

impact innovation performance; implied correlation coefficient should refers to the direct effect of external innovation resources on innovation performance by the aid of network dynamic capability; unanalysed impact coefficient refers to the correlation between external innovation resources which act on network dynamic capability.

TABLE 3 Decomposition of relationship between variables

| Variable Combination | Correlation Coefficient | Decomposition of Relationship | | | Total impact Coefficient |
|--------------------------------|-------------------------|-------------------------------|--|--|--------------------------|
| | | Direct Impact Coefficient | Indirect Impact Coefficient | Implied Correlation Coefficient | |
| X ₁ →X ₂ | R ₁₂ | P ₂₁ =0.180 | | | 0.180 |
| X ₁ →X ₃ | R ₁₃ | P ₃₁ =0.027 | | | 0.027 |
| X ₂ →X ₃ | R ₂₃ | P ₃₂ =0.234 | | | 0.234 |
| X ₁ →X ₄ | R ₁₄ | P ₄₁ =0.164 | | | 0.222 |
| X ₂ →X ₄ | R ₂₄ | P ₄₂ =0.267 | | $P_{42}R_{12} + P_{43}R_{13} = 0.058$ | 0.381 |
| X ₃ →X ₄ | R ₃₄ | P ₄₃ =0.364 | | $P_{41}R_{12} + P_{43}R_{23} = 0.114$ | 0.431 |
| X ₁ →X ₆ | R ₁₆ | P ₆₁ =0.078 | $P_{62}R_{12} + P_{63}R_{13} + P_{64}R_{14} = 0.076$ | $P_{61}R_{13} + P_{62}R_{23} + P_{64}R_{34} = 0.134$ | 0.154 |
| X ₂ →X ₆ | R ₂₆ | P ₆₂ =0.086 | | | 0.202 |
| X ₃ →X ₆ | R ₃₆ | P ₆₃ =0.167 | | | 0.301 |
| X ₄ →X ₆ | R ₄₆ | P ₆₄ =0.259 | | $P_{61}R_{14} + P_{62}R_{24} + P_{63}R_{34} = 0.122$ | 0.381 |

As to the verification of relationship between 3 dimensions of external innovation resources and innovation performance, based on the path analysis in Table 3, it is found that cooperation with horizontal and vertical enterprises have significant positive effect on innovation performance ($R_{16}=0.154, R_{26}=0.202$). However, the observed value consists of two parts, among which direct relationship with innovation performance is not significant ($P_{61}=0.078, P_{62}=0.086$). Therefore, it is assumed that 1a and 1b are false, that is, the cooperation with horizontal and vertical enterprises cannot directly improve the EP enterprises innovation performance.

Government-industry-academy-research cooperation is significantly correlated with innovation performance ($R_{26}=0.301, P<0.1$) and also significantly correlated in direct relationship. Therefore, it is assumed that 1c passes the validation and government-industry-academy-research cooperation is beneficial to the improvement of innovation performance.

As to the verification of relationship between 3 dimensions of external innovation resources and enterprise network dynamic capability, based on the path analysis in Table 3, it is found that the direct correlation coefficients of cooperation with horizontal and vertical enterprises and

government-industry-academy-research cooperation and network dynamic capability are significant ($P_{41}=0.164$, $P_{42}=0.267$, $P_{43}=0.364$). Therefore, it is assumed that 2a, 2b and 2c are verified, that is, horizontal and vertical enterprises and government-industry-academy-research cooperation have direct positive effect on enterprise network dynamic capability.

As to the verification of relationship between 3 dimensions of external innovation resources and enterprise network dynamic capability, based on the path analysis in Table 3, it is found that enterprise network dynamic capability has significant effect on improvement of innovation performance ($P_{64}=0.259$, $P<0.01$). Therefore, it is assumed that 3a is verified.

Verification of enterprise network dynamic capability's intermediary effect in the relationship between external innovation resources and innovation performance: in the above relationship disposition of variables, it is found that indirect impact of cooperation with horizontal enterprises on innovation performance is achieved through the intermediary of network dynamic capability ($P_{64}R_{14}=0.056$). Similarly, indirect impact factors of network dynamic capability on cooperation with vertical enterprises and government-industry-academy-research cooperation are respectively 0.063 and 0.112. External innovation resources have significant positive effect on innovation performance by virtue of enterprise network dynamic capability. Therefore, it can be illustrated that integration levels of external innovation resources have indirect positive impact on innovation performance by improving the enterprise network dynamic capability. Therefore, H_{3b} is verified. Enterprise network dynamic capability plays an intermediary role for external innovation resources to improve innovation performance.

4 Research results and discussion

4.1 RESULT DISCUSSION FOR EXTERNAL INNOVATION RESOURCES AND ACTION MECHANISM OF INNOVATION PERFORMANCE OF ENTERPRISES

4.1.1 Dimension of cooperation with horizontal enterprises

Correlation coefficient between cooperation with horizontal enterprises and innovation performance is 0.154, in which the direct correlation coefficient is only 0.078. The significance probability is $0.240 > 0.1$ and the path relationship is not obvious with path coefficient not significant even under the significance level of 0.1, which shows that there is no direct correlation between cooperation with horizontal enterprises and innovation performance. To test the action mechanism of external innovation resources to innovation performance, in line with research results in the documents at home and abroad, the intermediary variable of network dynamic capability is established in this essay. By the action of network dynamic

capability, cooperation with horizontal enterprises obtains a significant correlation ($R_{16}=0.154$, $P<0.01$) with innovation performance, which shows cooperation with horizontal enterprises is beneficial for the improvement of innovation performance by the action of intermediary network dynamic capability. This conclusion is confirmed by many scholars. Bayona believes cooperation between horizontal enterprises can provide technical combination advantages to achieve coordination effect and R&D scale effect, so as to ultimately achieve the technology breakthrough [28]. Although the cooperation with horizontal enterprises promotes, to some degree, the innovation performance, such effect of promotion is the weakest compared with the effect of cooperation with vertical enterprises and government-industry-academy-research cooperation. This is confirmed by the empirical research. Under the co-effect of policy guidance and market coordination, EP industry has positive externality and polluters do not take the initiative to bear negative externality costs, consequently resulting in big potential demands and small real demands in domestic market. For capital-intensive and technology-intensive EP industry, because of big investment and long payback period, EP industry, in most cases, is not willing to make capital and personnel investment to develop new technologies and new products, but imitates and replicates the existing technologies and products of relevant enterprises, which makes decreasing communication between peers to avoid the disclosure of technologies, information and products.

4.1.2 Dimension of cooperation with vertical enterprises

Correlation coefficient between cooperation with vertical enterprises and innovation performance is 0.202, in which the direct correlation coefficient is only 0.086. The significance probability is $0.168 > 0.1$ and the path relationship is not obvious with path coefficient not significant even under the significance level of 0.1, which shows that there is no direct correlation between cooperation with vertical enterprises and innovation performance. However, by the action of network dynamic capability, cooperation with vertical enterprises can achieve a significant correlation ($R_{26}=0.202$, $P<0.05$) with innovation performance, which shows that under the open innovation strategy, enterprises will transform the knowledge and technical spillover into innovation elements by improving their own network dynamic capability, so as to maximize the acquisition of external resources and improve the innovation performance. This is confirmed by Deng Yingxiang and Zhu Guilong. They propose that cooperation with enterprises in vertical industrial chain can be beneficiary for the improvement of innovation performance, and network dynamic capability mostly directly acts on searching for, absorbing and integrating new knowledge in vertical industrial chain to be transformed into the innovation elements [29]. Based on the analyses in Table 3, the impact of cooperation with vertical enterprises on network dynamic capability and

innovation performance falls between the horizontal dimension and government-industry-academy-research dimension. It is found in field research that cooperation of EP enterprises and vertical enterprises is beneficiary for acquisition of market information, standards and technical resources, so as to achieve the innovation of incremental products and technologies.

4.1.3 Dimension of government-industry-academy-research

Correlation coefficient between government-industry-academy-research cooperation and innovation performance is 0.301, in which the direct correlation coefficient is only 0.259. The significance probability is $0.000 < 0.05$ and the path relationship is not obvious with path coefficient not significant even under the significance level of 0.05, which shows that there is direct positive correlation between government-industry-academy-research cooperation and innovation performance. This primarily show that government-industry-academy-research cooperation acts directly on innovation, without any need of intermediary of network dynamic capability. This is compliant with the current development status of EP industry, that is, most EP enterprises don't have their own R&D departments due to small size or sufficient network dynamic capability to acquire external innovation resources, therefore in most cases, they will choose to cooperate with academies, designing institutions and public service platforms of government for technical innovation. Empirical results also show that, government-industry-academy-research plays a comparatively important role in innovation performance, mostly because industrialization of technical and scientific R&D and results is of high market risk, while most EP enterprises' low technical R&D makes low conversion ratio of industrialization of technical and scientific R&D results. Therefore, in most cases, EP enterprises will tend to choose government-industry-academy-research cooperation for innovation.

4.2 ANALYSIS OF INTERNAL RESOURCES' EFFECT ON INNOVATION PERFORMANCE

Internal innovation resources manly include R&D investment. A large quantity of empirical researches and the results prove that internal innovation resources have significant positive impact on innovation performance. It is found in this essay that, by the way of path analysis, acquisition of external innovation resources have significant positive correlation with innovation performance with a path coefficient of 0.381, while path coefficient between internal innovation resources and innovation performance is 0.507, which is bigger than that of external innovation resources and innovation performance. This is incompliant with Chesbrough's opinion that the external innovation resources share the same importance with internal resources. This shows that,

EP enterprises still implement the open innovation strategy at a low level and make insufficient use of external innovation resources. Veugelers & Cassiman's research shows that, technology-intensive and capital-intensive enterprises are more likely tend to acquire external resources. Open innovation does not mean to give up internal R&D, but to effectively utilize and integrate internal resources of enterprises [30]. Therefore, enterprises not only need the external resources related to basic science, but also depend on their own R&D activities. It is also found in mechanism analysis that internal resources have significant correlation with network dynamic capability, which shows the structure of internal resources of EP enterprises affects the capability to acquire external resources and the improvement of network dynamic capability which can helps more effectively acquire the resources necessary for enterprises and enrich the enterprises' internal resources.

5 Political meaning

This essay constructs the conceptual models between cooperation with horizontal enterprises, cooperation with vertical enterprises, government-industry-academy-research cooperation and internal resources with innovation performance based on the deficiencies in current researches and in line with implementation of open innovation strategy by EP enterprises, and carry out empirical research on internal effect mechanism between variables by questionnaires to 85 EP enterprises. The results reveal that cooperation with horizontal and vertical enterprises require the EP enterprises to be equipped with certain network dynamic capability to improve innovation performance, while government-industry-academy-research cooperation can directly promote the improvement of enterprises' innovation capability. At the meantime, empirical researches also demonstrate that main factor for EP enterprises innovation is internal resources and external innovation resources act only as a supplementary to internal resources. Combing the above analysis results, we can summarize the idea of sustainable development for EP innovation as:

- 1) During the development process, EP enterprises shall lay emphasis on the construction of multi-layer network structure and can acquire resources from external innovation environment in accordance with different development phases and external environment and in line with their own requirements and capability, reduce the redundancy of network structure, and improve efficiency to absorb external innovation resources.
- 2) Constantly perfect innovation service system of EP enterprises by establishing new organizations or adjusting the service scope of existed service organizations. The system shall serve not only for the acquisition of external resources, but also for the R&D of internal technologies.
- 3) EP enterprises establish more extensive social relationship with horizontal competitors or partners, enterprises in vertical industrial chain, agencies,

academies, high schools, governments and financial institutions by establishing information platform.

4) The government can enhance the relations among EP enterprises and extend social network of informal

communication by establishing agencies, such as environmental protection associations.

References

- [1] Chesbrough H 2004 Managing open innovation *Research Technology Management* **42**(1) 23-26
- [2] Luo J, Ma H, Zhang J 2010 The Development and Innovation of China's Environmental Industry *Bulletin of Chinese Academy of Sciences* **5**(2) 146-52 (in Chinese)
- [3] Fu T 2013 Technical innovation leads the development of environmental *Protection industry* **12**(21) 17-24
- [4] Jiang G, Zhang J 2012 Environmental technology innovation and development of environmental *Protection industry* **15**(15) 31-4
- [5] Dong Y 2007 Characteristics of Innovation System of Environmental Industry and Its Countermeasures *Ecological Economics* **16**(9) 134-7
- [6] Xing X, Tong Y 2007 Study on relationship between enterprise network capability and technical capability under innovation perspective *Science and Management of Science and Technology* **28**(12) 182-6
- [7] Chi R, Tang L 2008 The linking features between enterprise external innovative network and innovation sources *Science & Technology Progress and Policy* **25**(11) 38-40
- [8] Chen Y, Chen J 2009 A study on the mechanism of open innovation promoting innovative performance *Scientific research management* **30**(4) 1-9
- [9] Ritala P, Hurmelinna-Laukkanen P What's in it for me? Creating and appropriating value in innovation—related cooperation *Technovation* **29**(12) 819-28
- [10] Hagedoorn, J 1990 Organizational modes of inter—firm cooperation and technology transfer *Technovation* **10**(1) 17-30
- [11] Clark K B 1989 Project scope and project performance, the effect of parts strategy and supplier involvement on product development *Management Science* **35**(10) 1247-63
- [12] Laursen K, Salter A 2006 Open for innovation: The role of openness in explaining innovation performance among UK manufacturing firms *Strategic Management Journal* **12** (2) 16-25
- [13] Tether B 2002 Who co-operates for innovation and why an empirical analysis *Research Policy* **31**(6) 947-67
- [14] Tidd J, Bessant J, Pavitt K 2008 Managing innovation: integrating technological, market and organizational change *Singhua University Press*
- [15] Gulati R 1999 Network location and learning: the influence of network resources and firm capabilities on alliance formation *Strategic Management Journal* **20**(5) 397-420
- [16] Miao G, Chen W, Tang C 2014 A study on relationship between external innovation search, knowledge integration and innovation performance *Science & Tech Progress and Policy* **31**(1) 130-4
- [17] Dorsey S G 2006 Measuring the impact of integration and diversification on firm value in the food industry Ph.D dissertation *Kansas state university*
- [18] Harabi N 1997 Channels of R&D spillovers: An empirical investigation of Swiss firms *Technovation* **17**(11) 627-37
- [19] Liu W, Zhang Z, Zhang W 2009 A study on common R&D investment mechanism in vertical cooperation *Journal of Industrial Engineering and Engineering Management* **23**(1) 19-22
- [20] Kong X, Xu Z, Suzhou 2012 A study on "Four Wheel Driven" structure and mechanism of synergy innovation for government-industry-academy-research cooperation *Science & Technology Progress and Policy* **29**(22) 15-8
- [21] Chen H 2009 A study on government-industry-academy-research cooperation model and mechanism based on 3 helices theory *Science & Technology Progress and Policy* **26**(24) 6-8
- [22] Damanpour F, Gopalakrishnan S 2001 The dynamics of the adoption of product and process innovation in organization *Journal of Management Studies* **38**(1) 45-65
- [23] Xin Q 2012 How does knowledge network impact enterprise innovation: empirical research from dynamic capability perspective *Research and Development Management* **24**(6) 12-21
- [24] Lichtenthaler U 2007 Developing reputation to overcome the imperfections in the markets for knowledge *Research Policy* **36**(1) 37-55
- [25] Cai N, Yan C 2013 Measurement of open innovation performance: Theoretical model and empirical testing *Studies in Science of Science* **31**(3) 469-77
- [26] Lichtenthaler U 2008 Integrated roadmaps for open innovation *Research Technology Management* **51**(3) 45-9
- [27] Ma Q 2008 Research method in management science *Higher Education Press*
- [28] Bayona C, Garcia-Marco T, Huerta E 2001 Firms' motivation for cooperative R&D: an empirical analysis of spanish firms *Research Policy* **30**(8) 1289-1307
- [29] Deng Y, Zhu G 2009 A study of intermediary effect of absorption capability in innovation process—experience evidence from Pearl River Delta enterprises *Science of Science and Management of S.&T* **30**(10) 85-9
- [30] Veugelers R, Cassiman B 2006 Make and buy in innovation strategies: Evidence from Belgian manufacturing firms *Research Policy* **28**(1) 63-80

| Authors | |
|---|---|
|  | <p>Huang Qing-huang, born in July, 1987, Fuzhou City, Fujian Province, China</p> <p>Current position, grades: Doctor student of College of economics and management, Fuzhou University, China.</p> <p>University studies: Bachelor of Management from Fujian Agriculture and Forestry University, Master of management from Fuzhou University in China.</p> <p>Scientific interest: industrial economy, resource and environmental management.</p> <p>Experience: 6 scientific research projects.</p> |
|  | <p>Gao Ming, born in May, 1965, Fuzhou City, Fujian Province, China</p> <p>Current position, grades: Professor of College of economics and management, Fuzhou University, China.</p> <p>University studies: Master of management from Northeast Agricultural University in China, Doctor of management from Renmin University of China.</p> <p>Scientific interest: environment and resource management, regional development, industrial economy.</p> <p>Experience: 18 scientific research projects.</p> |

An analysis on the growth and effect factors of TFP under the energy and environment regulation: data from China

Jiansheng Zhang*

The School of Economics and Management, Chongqing Three Gorges University, Wanzhou, Chongqing, China, 404120

Received 6 May 2014, www.tsi.lv

Abstract

The paper analyses the growth and effect factors of TFP (Total factor productivity) under the energy and environment regulation with the data of China from 2002 to 2012. The results show that: in the past 10 years, without considering the energy and environmental regulation, the average annual growth rate of TFP is 3.2%, but it is 2.7% when considering them. The technological progress is the major contributor to TFP under the energy and environment regulation. From the comparison of various provinces, the growth difference of TFP was great. The TFP value in eastern coastal region is higher than that in the central and western regions. From the time trend, the average growth rate of TFP is in the lower. After the financial crisis of 2008, the TFP starts to decline and the average annual growth rate is -0.3%. The three variables of the FDI, environmental regulation intensity and industrial structure have a negative impact on TFP growth, but the two variables of R&D investment and energy consumption structure have a positive impact on it.

Keywords: environmental pollution, energy regulation, TFP, effect factors

1 Introduction

Since more than 30 years' reform and opening up policy, the annual growth rate of economic reaches 10% in China. But by the rapid economic growth, there are a number of problems, such as low resource utilization efficiency, environmental degradation and loss of environmental health which make the sustainable development face severe challenges. The energy consumption had increased by more than four times in 2000-2008 than that in the 1990s. At the same time, the polluted environment and changing climate by a lot of energy consumption which also bring a huge ecological and environment pressure to the social development. In 2010, the world's environmental performance index (EPI) ranking, China had the score of 49 which was 121st in 163 countries and regions, the international community gives a growing awareness of environmental problems in China [1].

According to the theory of modern economic growth, economic growth comes from two aspects: one is the inputs such as capital and labour, the second is from the improvement of TFP (Total factor productivity). Lack of per capita resources, environment pressure, and the growth of output relies on the input which is not sustainable. Therefore, the future economic growth must rely on the improvement of TFP in China. But the traditional measure of TFP does not consider environmental factors, also does not take the energy factors into account. With the resource and environmental problems in the process of economic development, more and more scholars think the resources and environment are not only the endogenous variables, but also the rigid constraints. Therefore, assessing

economic performance by the TFP does not only consider the traditional factors of capital and labour, but also consider the resources and environmental factors which have a huge impact on economic growth.

Although Zhang [2], and other scholars measure the TFP in each province of China under energy and environment regulation, but the study does not consider the energy regulation, and it does not analyse the influence factors of TFP. On this basis, this paper researches the TFP in China on energy and environmental constraints, and analyses its reasons. Meanwhile, this study plays an important role for the Chinese government to guide transformation of economic structure and adjust the green GDP accounting target. This paper gives an empirical analysis by the data of China.

2 Literature review

At present, there are a large number of literatures on the research of total factor productivity. These literatures can be divided into two categories: the first kind of literatures does not consider environment pollution and energy input when measuring productivity. Most of these studies analyse the TFP by the Solow residual method, Malmquist index method and the stochastic frontier production function method. The capital and labour are the inputs, and the GDP is the output. The research results show that the TFP has been increasing in China, and it has more and more influence on the economy [3-6]. The second kind of literatures which put the environmental factor into the TFP framework. The related results show that the different provinces of China are as the research object, and it

* *Corresponding author* e-mail: asheng0124@126.com

analyses the TFP growth under the constraints of SO₂ and CO₂ emissions. The results show that when considering environmental factors, the growth rate of TFP was only 1/3 of the conventional measurement value [7]. It estimates the TFP in manufacturing industry by the Directional Distance Function and Malmquist-Luenberger productivity indicator Function, and the results shows that the TFP presents a growth trend when considering the environmental factors. The factors of capital deepening, industry scale, spending of R&D and environmental pollution have different degree of influence on the TFP in light industry and heavy industry [8]. It analyses the agricultural TFP when considering the pollution in agricultural as a "bad" output, the results show that the agricultural TFP growth obviously under the environment constraint in China, and the growth is mainly driven by the agricultural technology progress. The agricultural TFP in each region appears different degree of deterioration. From the point of regional differences, the TFP presents the decreasing in east, west and the centre under environment regulation [9]. It estimates the TFP in energy-intensive industry by the directional distance function and non-parametric DEA method, and the results shows that the growth of TFP is mainly driven by technological progress. The status quo of China shows that the TFP in current energy intensive industry has greater room for improvement. The growth of TFP in each province presents different degree of convergence. Market-oriented reform, FDI inflows and the decline in energy intensity are all conducive to the growth of TFP [10].

3 Research method

In order to put the environmental factors into the productivity analysis framework, it need to construct a production possibility set which includes good and bad output, namely the environmental technology. Suppose each region using *N* kind of input, $X = (X_1, \dots, X_N) \in R_+^N$, and produce *M* kind of good output, $Y = (Y_1, \dots, Y_M) \in R_+^M$. At the same time, it also produces *I* kind of bad output, $b = (b_1, \dots, b_I) \in R_+^I$, so the production possibility set of environmental technology is:

$$p(x) = \left[(y, b) : (y, b) \in p(x), x \in R_+^N \right]. \tag{1}$$

Due to the purpose of this study is that keeping the bad output to decrease and the good output to grow. Therefore, the bad output in technology has weak disposability. By the directional distance function [11], the equation is expressed as:

$$\overline{D}_0(x, y, b; g) = \sup \left\{ \beta : (y, b) + \beta g \in p(x) \right\}, \tag{2}$$

where $g = (y, -b)$ is the Direction Vector and β is the directional distance function. It measures the increasing value of good output while maintaining the bad output reduces under the condition of certain inputs. The

directional distance function can be represented by the following mathematical equation:

$$\begin{aligned} \overline{D}_0^t(x_k^t, y_k^t, b_k^t; y_k^t, -b_k^t) &= \max \beta, \\ s.t. \sum_{k=1}^K \lambda_k^t y_{km}^t &\geq (1 + \beta) y_{km}^t, m = 1, \dots, M, \\ \sum_{k=1}^K \lambda_k^t u_{ki}^t &= (1 - \beta) u_{ki}^t, i = 1, \dots, I, \\ \sum_{k=1}^K \lambda_k^t x_{kn}^t &\leq x_{kn}^t, n = 1, \dots, N, \\ \lambda_k^t &\geq 0, k = 1, \dots, K. \end{aligned} \tag{3}$$

To solve the evaluated problem of TFP when considering the case of bad output, Chung et al. [12] put forward the Malmquist-Luenberger productivity indicator by the environmental DEA technology and direction distance function [12]. The ML productivity indicator from *t* to *t* + 1 period is as following:

$$ML_t^{t+1} = \left[\frac{1 + \overline{D}_0^t(x^t, y^t, b^t; g^t)}{1 + \overline{D}_0^t(x^{t+1}, y^{t+1}, b^{t+1}; g^{t+1})} \times \frac{1 + \overline{D}_0^{t+1}(x^t, y^t, b^t; g^t)}{1 + \overline{D}_0^{t+1}(x^{t+1}, y^{t+1}, b^{t+1}; g^{t+1})} \right]^{1/2}. \tag{4}$$

The ML productivity indicator can be decomposed into the technical changing efficiency (MLEC) and technological progress index (MLTC):

$$ML_t^{t+1} = MLEC_t^{t+1} \times MLTC_t^{t+1}, \tag{5}$$

$$MLEC_t^{t+1} = \frac{1 + \overline{D}_0^t(x^t, y^t, b^t; g^t)}{1 + \overline{D}_0^{t+1}(x^{t+1}, y^{t+1}, b^{t+1}; g^{t+1})}, \tag{6}$$

$$MLTC_t^{t+1} = \sqrt{\frac{1 + \overline{D}_0^{t+1}(x^t, y^t, b^t; g^t)}{1 + \overline{D}_0^t(x^t, y^t, b^t; g^t)} \times \frac{1 + \overline{D}_0^{t+1}(x^{t+1}, y^{t+1}, b^{t+1}; g^{t+1})}{1 + \overline{D}_0^t(x^{t+1}, y^{t+1}, b^{t+1}; g^{t+1})}}. \tag{7}$$

The ML indicator measures the change of productivity from *t* to *t* + 1. If *ML*=1, it shows the productivity declines. If *ML*>1, it shows the productivity improve. If *ML*<1, it shows the productivity remains the same. If *MLEC*>1, it shows the decision-making unit to be near the production frontier and the efficiency improves. If *MLEC*<1, it shows the efficiency declines. If *MLEC*=1, it shows the efficiency remains the same. If *MLTC*>1, it shows the technological progress of decision-making unit. If *MLTC*<1, it shows the technological retrogression of decision-making unit. If *MLTC*=1, it shows the technical level is constant.

4 Empirical analysis

The input indicators include the capital deposit, number of employed persons and total energy consumption. It calculates the capital stock by perpetual inventory method, and the formula is $K_{it} = K_{it-1}(1-\lambda) + I_{it} / P_{it}$. In the formula, λ is the depreciation rate, the value is 10%, I is the newly increased fixed assets, P is the price indices for investment in fixed assets in each province, and this index reduced for the year of 2000.

The paper references the method of Dai Yongan (2010) [13], the initial capital stock is equal to the total investment in fixed assets divided by 10% in 2002. The labours are the number of staff for each region. The energy consumption refers to the various energy consumption in various regions, including oil, coal, natural gas, electricity, etc. For the unified unit, this paper converts the consumption of various energy into 10000 tons of standard coal (SCE).

The output indicators include good output and bad output. The good output refers to the GDP, and the GDP data converts into the constant in 2000. The bad output is represented by industrial s SO_2 in each region.

The paper bases on the data of 30 provinces in China from the year 2002 to 2012. Because of the lacking of data in Tibet, it doesn't analyse.

4.1 THE TFP UNDER THE ENERGY AND ENVIRONMENT REGULATION

The paper calculates the TFP under energy and environment regulation by the output data of 30 provinces in China. Meanwhile, the TFP is divided into the MLTC and MLEC. The data in Table 1 is the geometric mean for 2002 to 2012 which reflects the regional differences and average growth. The Figure 1 is the geometric average value in each region, which reflects the changing trend of ML, MLTC and MLEC over time.

From Table 1 and Figure 1 is is seen:

1) Overall, there is a rapid growth of TFP under energy and environment regulation in China. In the past 10 years, the average growth has reached 2.7%, among them, the average annual growth of technological progress is 4.2%, while the technical efficiency has reduced by 1.5% per year on average. The technological progress is the main contributor to TFP.

2) The growth of TFP has great difference among regions in China. The fastest growth is Liaoning, with an average annual growth rate of 15.1%, and 12.4% above the average. In addition, the growing bigger include the eastern provinces of Jiangsu, Guangdong, Beijing, Shandong, Shanghai, Tianjin, etc. The low growth rate are the central and western provinces of Xinjiang, Henan, Ningxia, overall, Heilongjiang, etc. The average annual growth rate is less than 1%. It can be seen that the TFP value in eastern region is far higher than that in the centre and west from 2002 to 2012.

3) The TFP continuously grows in most provinces, but the TFP appears backwards a few provinces. The TFP have

reduced include the midwest provinces of Shanxi (-4.6%), Guangxi (-3.7%), Chongqing (-2.0%), Qinghai (-0.9%), Hunan (-0.4%), Neimenggu (-0.2%).

4) From the time trend, growth rate of TFP gradually reduces in China. The average annual growth rate in 2003 was 8.4%, 6.6% in 2004, 4.1% in 2007, and it appeared a significant reduction in 2008, only 0.9%. It fell by 3.2% comparing with 2007. The growth rate in 2009 was -0.1%, while it increased slightly by 0.1% in 2010. It was a negative growth in 2011 and 2012, the growth rate of 1.5% and 0.7%. This paper argues that it is mainly due to the financial crisis in 2008, which made the exports and economic downturn in China.

TABLE 1 The TFP and its decomposition under energy and environment regulation (2002-2012)

| Region | ML | MLTC | MLEC |
|--------------|-------|-------|-------|
| Beijing | 1.083 | 1.083 | 1 |
| Tianjin | 1.064 | 1.078 | 0.987 |
| Hebei | 1.024 | 1.063 | 0.963 |
| Shanxi | 0.954 | 1.016 | 0.94 |
| Neimenggu | 0.998 | 1.027 | 0.972 |
| Liaoning | 1.151 | 1.155 | 0.997 |
| Jilin | 1.024 | 1.042 | 0.983 |
| Heilongjiang | 1.008 | 1.019 | 0.99 |
| Shanghai | 1.066 | 1.066 | 1 |
| Jiangsu | 1.101 | 1.101 | 1 |
| Zhejiang | 1.044 | 1.04 | 1.004 |
| Anhui | 1.045 | 1.058 | 0.988 |
| Fujian | 1.013 | 1.013 | 1 |
| Jiangxi | 1.005 | 0.976 | 1.03 |
| Shandong | 1.072 | 1.072 | 1 |
| Henan | 1.003 | 1.046 | 0.958 |
| Hubei | 1.023 | 1.032 | 0.99 |
| Hunan | 0.996 | 1.021 | 0.975 |
| Guangdong | 1.091 | 1.091 | 1 |
| Guangxi | 0.963 | 1.005 | 0.958 |
| Hainan | 1.014 | 1.03 | 0.985 |
| Chongqing | 0.98 | 1.034 | 0.948 |
| Sichuan | 1.025 | 1.046 | 0.98 |
| Guizhou | 1.022 | 1.021 | 1.002 |
| Yunnan | 1.019 | 1.048 | 0.972 |
| Shanxian | 1.029 | 1.034 | 0.995 |
| Gansu | 1.012 | 1.015 | 0.998 |
| Qinghai | 0.991 | 1.019 | 0.972 |
| Ningxia | 1.004 | 1.007 | 0.997 |
| Xinjiang | 1.002 | 1.021 | 0.981 |
| Average | 1.027 | 1.042 | 0.985 |

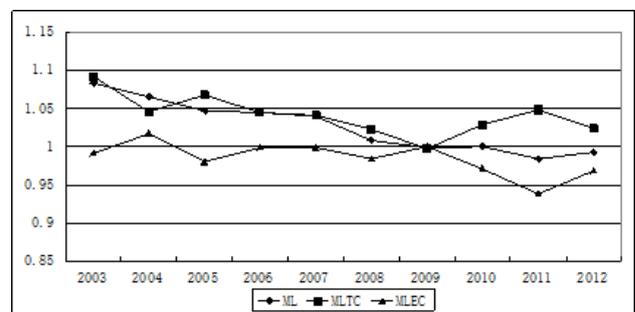


FIGURE 1 The changing trend of ML, MLTC and MLEC over time

4.2 THE COMPARISON OF TFP IN TWO CIRCUMSTANCES

If it does not consider the energy constraints and bad output, there may be a large error for the calculated results of TFP. Therefore, this paper gives a comparison of the results in two circumstances. Figure 2 shows the comparison of productivity index, the technical progress and technical efficiency in both cases for 30 provinces:

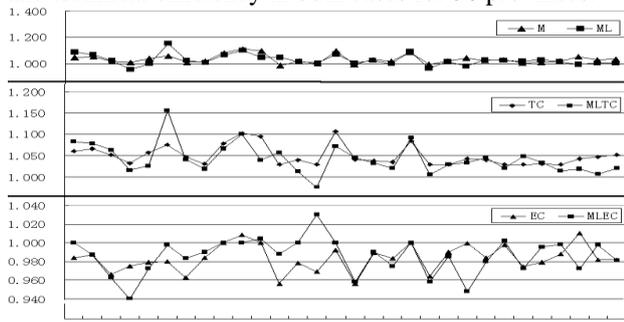


FIGURE 2 The comparison of TFP in two circumstances

As can be seen from the Figure 2, when considering the energy constraints, and bad output, the annual growth rate of TFP declines in most provinces. The growth rate of TFP falls from 3.2% to 2.7%, which shows that the economic growth in most regions of China still basing on the guidance of GDP growth. It consumes a lot of energy and emissions of pollutants. But there are also some provinces such as Beijing, Liaoning, Anhui and other regions, after considering energy constraints and bad output, the growth rate of TFP has been increasing. These provinces appear the *potter win-win* situation. When considering energy and environmental factors, the average annual growth rate of technical progress is from 5.0% to 4.2%, reducing by 8% a year. Among them, the technological progress index falls in 19 provinces which show the measure result of technological progress is overvalued. The average annual growth rate of technical efficiency in each region decreases from -1.7% to -1.5% which shows the energy and environment have less effect on the technical efficiency. Only a few provinces of Shanxi, Anhui, overall and Chongqing appear a larger fluctuation.

5 The analysis of influence factors

What factors influence on TFP? According to the research achievements of other scholars, this paper gives an empirical analysis selecting the following indicators for influencing the growth of TFP:

- 1) The foreign direct investment. In developing countries, the inflow of FDI can bring the advanced production and management technology which will promote the growth of efficiency.
- 2) The environmental regulation intensity. After implementation of environmental regulation, the enterprise needs to increase the investment related to environmental protection which will increase the cost of enterprise. But if the enterprise does the innovation

activities of technology, and use the new technology and new equipment which will reduce pollution. The intensity index of environmental regulation is expressed by the industrial SO₂ emissions compliance rate.

3) The R&D investment. The investment funds of science and technology can promote the rapid economic growth in a country. The R&D index can be represented by the proportion of expenditure for science and technology in government expenditure.

4) The energy consumption structure. The bad production by different energy is also different. The energy consumption structure can be represented by the proportion of coal consumption in total consumption of energy.

5) The industrial structure. There is a great difference between input and output in different industries which will affect green TFP in a region. The industrial structure index is represented by the ratio of added value in the second industry to GDP.

The model of influencing on TFP factors is as follows:

$$ML_{it} = \alpha + \beta_1 fdi_{it} + \beta_2 eri_{it} + \beta_3 rd_{it} + \beta_4 ecs_{it} + \beta_5 str_{it} + \epsilon_{it}, \quad (8)$$

where ML is the green TFP. FDI is the foreign direct investment. The *eri* is the environmental regulation intensity. The *rd* is the R&D investment. The *ecs* is the energy consumption structure. The *str* is the industrial structure.

Due to unable to get data of some year, the paper gives an analysis by the data from 2005 to 2010. The regression analysis results are shown in Table 2:

TABLE 2 The regression analysis results for panel data

| Variable | coefficient | t-statistic |
|-------------------------|-------------|----------------|
| fdi | -0.002 | -0.03 |
| Eri | -0.146 | -5.75 |
| rd | 0.011 | 1.52 |
| ecs | 0.137 | 4.41 |
| str | -0.051 | -0.54 |
| adr ² =0.612 | F=9.303 | D.W. stat=1.44 |

1) The negative influence of FDI on the green TFP. The conclusion is the same as Li ling's [14]. Meanwhile, the conclusion verifies the "pollution haven hypothesis" which means the degree of environmental regulation of developed countries is higher than that in the developing countries. Therefore, a large amount of FDI flows to the developing countries. The FDI can promote the economic growth in developing countries, at the same time, it also brings a lot of pollution.

2) The negative influence of environmental regulation intensity on the TFP and it is through the test of significance. The conclusion shows that the implement of environmental protection measures in Chinese governments which increases the cost of enterprise and hinders the growth of green TFP. The conclusion also shows that the environment and economy without achieving common development in China, and the "potter win-win" situation is only in a few provinces.

3) The positive effect of the R&D investment on TFP, and it is through the test of significance at the 15% level. The R&D can improve the level of technology, improve the energy efficiency in regional economic growth, and reduce the pollution emissions.

4) The positive effect of the energy consumption structure on TFP. It does not agree with the expectations, and the possible reasons lie in the choice of inappropriate metrics.

5) The positive effect of the industrial structure on TFP, but it is not through the test of significance. The result shows that the high energy consumption and high pollution in industrial development is bad for TFP growth. The higher proportion of GDP, the slower of the green TFP.

6 Conclusion

1) If it doesn't consider the environmental regulation, the traditional method can lead to a great deviation for the TFP measurement. So the paper puts the environmental factor into the TFP analysis framework, and analyses the growth of TFP and influence factors with the data of 30 provinces in China. In the past 10 years, the average growth has reached 2.7%, and the average annual growth of technological progress is 4.2%, while the technical efficiency has reduced by 1.5% per year on average. The technological progress is the main contributor to TFP.

2) The growth of TFP has a great difference among regions in China. The TFP value in east is much higher than that in the midwest. The TFP in most provinces has been increasing, but the 6 western provinces of Shanxi, Guangxi, Chongqing, Qinghai, Hunan, and Neimenggu are backwards.

3) The growth rate TFP is in the lower in China. After the 2008 financial crisis, The TFP starts to decline, and the average annual growth rate is -0.3%.

4) When considering the energy regulation and bad output, the annual growth rate of TFP declines in most provinces. The growth rate of TFP falls from 3.2% to 2.7%, which

shows that it invests a large amount of energy and emissions of pollutants for the economic growth in most provinces of China.

5) The three variables of FDI, environmental regulation intensity and industrial structure have a negative effect on green TFP, but the two variables of R&D investment and energy consumption structure have a positive impact.

At present, it is in a stage of rapid development in China. But for a long time, the disadvantages of economic growth path for "high investment, high pollution, high output" is more and more obvious. The living environment for Chinese residents deteriorates. According to the research conclusion, the following suggestions are put forward: First, in order to promote the growth of green TFP, the Chinese government must change the old development model. At the same time of relying on technological progress, it should strengthen the application of existing technology and improve the technical efficiency. Second, the government makes policy for transferring and diffusing the advanced environmental technology between different provinces which can effectively promote the environmental protection technology level in west. Third, it will optimize the industrial structure, promote the development of the third industry and reduce the proportion of secondary industry. Fourth, the local government should attach great importance to the serious pollution of FDI. Fifth, it will continue to implement the environmental regulation measures, play the advantages of market competition, integrate of the resources and factors of polluting industries, shut down the enterprises of backward technology, high energy consumption and high pollution and encourage the development of large enterprises of high level technology, less pollution and good benefit.

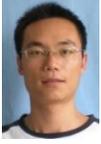
Acknowledgements

This research was supported by the national social science fund projects (12xgl019) of China.

References

- [1] Qu X, Xi Y 2012 Total factor productivity in China under the dual regulation of the resources and the environment *Journal of business economics* **24**(5) 89-97
- [2] Zhang J, Kopytov C, Qifang M 2014 Study on the Change and Regional Differences About Total Factor Productivity Considering the Environmental Pollution in China *Nature Environment and Pollution Technology* **13**(2) 327-32
- [3] Li G, Zhou C, Jiang J 2010 The Estimation of Regional TFP and Its Role in China's Regional Disparities *The Journal of Quantitative & Technical Economics* (5) 49-61
- [4] Chen H, Li G, Chen H 2010 Calculated and compared about total factor productivity of three industrial in China *Research on Financial and Economic Issues* **2** 28-31
- [5] Tu Z 2007 Total factor productivity and source of regional economic growth *Nankai economic studies* **4** 14-36
- [6] Zhong H, Xu P 2011 The determinants of TFP Growth in China: A Comprehensive Analysis Based on BACE Approach *Journal of Beijing Institute of Technology (Social Sciences Edition)* **13**(6) 1-8
- [7] Wu J, Da F, Zhang J 2010 Environmental Regulation and Regional TFP Growth of China *Statistical Research* **27**(1) 83-9
- [8] Yuan T, Shi Q, Liu Y 2012 Total factor productivity and its determinants under environmental constraint in China's Manufacturing industry *Wuhan University of Technology (Social Science Edition)* **25**(6) 860-7
- [9] Han H, Zhao L 2013 Growth and convergence of agricultural total factor productivity in China under environmental regulations *China Population, resources, and environment* **23**(3) 70-6
- [10] Shen K, Gong J 2011 Environmental pollution, technical progress, and productivity growth of energy-intensive industries in China *China industrial economics* **12** 25-34
- [11] Fare R, Grosskopf S, Pasurka C A Jr 2007 Environmental Production Functions and Environmental Directional Distance Functions *Energy* **32**(7) 1055-66
- [12] Chung Y H, Fare R, Grosskopf S 1997 Productivity and Undesirable Outputs: A Directional Distance Function Approach *Journal of Environmental Management* **51**(3) 229-40
- [13] Dai Y 2010 The Efficiency and Determinants of China's Urbanization *The Journal of Quantitative & Technical Economics* **12** 103-18
- [14] Li L, Tao F 2011 An analysis on the Green TFP and reasons of Pollution intensive industries *Economist* **12** 32-9

Authors



Jiansheng Zhang, born in January, 1981, Chongqing, China

Current position, grades: the Associate Professor of Economics and Management, Chongqing Three Gorges University, China.

University studies: D.Sc. at Xinan Jiaotong University in China.

Scientific interest: Operations research, pollution of the environment and economic development.

Publications: More than 20 papers published in various journals.

Experience: Teaching experience of 8 years, 2 research projects.

Analysis of the public satisfaction index of public cultural services based on the Grey Correlation AHP method

Liping Fu¹, Juan Li^{1, 2*}

¹Public Resource Management Research Center, College of Management and Economics Tianjin University, Tianjin, China

²Hebei United University, 46 Xinhua Road, Tangshan, Hebei, China

Received 19 February 2014, www.tsi.lv

Abstract

The public is the service object of the public cultural services while the public satisfaction index is the main indicator in the judgment of the public cultural service effect. The Grey Correlation Method is applied to selecting the main factors which influence the public satisfaction index of public cultural services, and the number of public library, the public cultural activities of organizations, and the number of staff in the public cultural service institutions is the most three important factors. After that, the paper builds public satisfaction model based on grey correlation AHP, applies the method to evaluating the current public satisfaction of public cultural services in China, and proposes the specific measures to improve and promote public cultural services in China on the basis of the evaluation result.

Keywords: public cultural services, public satisfaction index, Grey Correlation AHP method

1 Introduction

The system of public cultural services is an important part of the governmental public services as well as a significant way to realize citizens' cultural rights. Public satisfaction is an exclusive criterion for judging whether the public cultural services are effective or not. Only the satisfaction level of public cultural services obtained from the objective measurement and analysis can inspect whether the public services provided are effective or not, and only the corresponding suggestions based on public needs can improve public cultural services, so as to serve the public better.

The public satisfaction evaluation originated from the enterprise's customer satisfaction index. In 1989, Sweden took the lead in establishing the evaluation model of Swedish Customer Satisfaction Barometer (SCSB) [1], followed by the European countries, USA and other developed countries, which made various improvements and innovations on the basis of the original model and in combination with the actual conditions and applied it to the government, bank, hospital and other fields [2-6]. In 2004, You Jianxin and other public administration scholars formally introduced the concept of Public Satisfaction into the field of the Chinese Public Administration [7] and they believed that the government performance evaluation, based on public satisfaction, is an inevitable requirement for the construction of a modern and efficient government [8], which required the government to attach great importance to the public service satisfaction and adopt the down-top assessment and measurement method [9]. The assessment of public satisfaction of the Chinese

government services were of great importance to guide the construction of standardized performance standards for the public services of the Chinese government and promoted the reform of the governmental administrative system and the government construction [10]. The research on public satisfaction was mainly focused on the concrete application of the research methods [11-14], the construction of public satisfaction evaluation system taking the city or community as the research object, and the main factors of influencing the results through the research methods and proposing the policies and suggestions based on the main factors [15-18]. In recent years, more and more researches have been conducted on rural public service satisfaction. Li Yanling made an analysis on the satisfaction with the rural public goods supply and agricultural information and their influence in Hunan province by means of questionnaire survey [19-20]. Yang Weijing constructed the evaluation index system of the rural public goods supply and established a corresponding fuzzy synthetic evaluation model to evaluate the satisfaction level and put forward the corresponding countermeasures and suggestions [21].

The public satisfaction index model based on Grey Correlation Analytic Hierarchy Process was established in this paper. Based on the research on the public cultural services and the use of the improved grey correlation method, the quantitative analysis of the public cultural services was carried out to achieve the public's satisfaction degree for the current China's public cultural services, so as to evaluate the current China's public cultural services and pinpoint the problems existing in the development of China's public cultural services in the days to come.

* Corresponding author e-mail: lijuanzw@126.com

2 Selection of the main factors by the Grey Correlation Method

Covering a wide range of contents, the public cultural services are involved in multiple factors that influence the satisfaction degree of the public cultural services. The

factors include the public cultural venues and facilities, the public cultural service contents, the public cultural service quality, other primary influence factors and the corresponding secondary influence factors. The factors divided into primary evaluation index and secondary evaluation index are as shown in Figure 1:

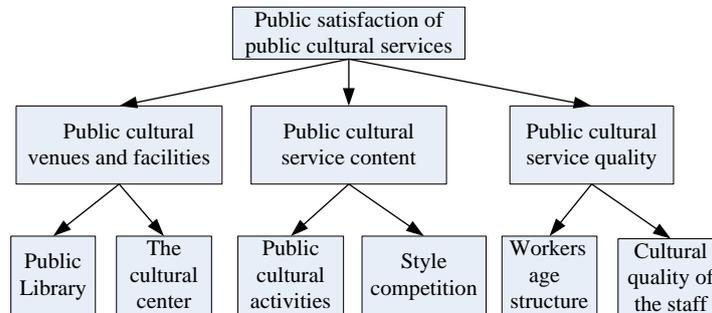


FIGURE 1 Public cultural services evaluation index

The secondary indexes, in close relations to primary indexes above, were selected, according to the multiple factors above, by means of the Grey Relation Method and further improved model. Thus the public satisfaction index based on the Grey Correlation AHP Method can be analysed.

In the public satisfaction index system of public cultural services, there are many factors that influence different primary evaluation indexes. The grey correlation degree of above factors will be analysed respectively to determine an index which has the largest correlation with the system.

2.1 THE FACTORS OF PUBLIC CULTURAL VENUES AND FACILITIES

The public cultural venues and facilities are the factors of direct influence and the basis of evaluating public cultural services. The venue construction quality is relevant to the numbers of visitors and subsequently influences the public satisfaction index. The number of library, cultural centre and museum constructed and the total number of visitors' circulation in China from 2008 to 2012 are as shown in Table 1.

TABLE 1 Public cultural venues and facilities from 2008 to 2012

| Year | Total number of circulation(10,000) | Library | Cultural centre(station) | Museum |
|------|-------------------------------------|---------|--------------------------|--------|
| 2008 | 31295.7 | 2820 | 41156 | 1893 |
| 2009 | 31468.7 | 2850 | 41959 | 2252 |
| 2010 | 32167.5 | 2884 | 43382 | 2435 |
| 2011 | 32823.3 | 2952 | 43675 | 2650 |
| 2012 | 38151.0 | 3076 | 43876 | 3069 |

*Source: Statistical Yearbook 2013

1) The characteristic behaviour sequence of influence factor is as follow:

$$x_i' = (x_i'(1), x_i'(2))^T, i = 1, 2, 3, \tag{1}$$

in which the behaviour sequences of related factors are as follows:

$$x_1' = (2820, 2850, 2884, 2952, 3076),$$

$$x_2' = (41156, 41959, 43382, 43675, 43876),$$

$$x_3' = (1893, 2252, 2435, 2650, 3069),$$

$$x_i' = \begin{pmatrix} 2820 & 41156 & 1893 \\ 2850 & 41959 & 2252 \\ 2884 & 43382 & 2435 \\ 2952 & 43675 & 2650 \\ 3076 & 43876 & 3069 \end{pmatrix}.$$

2) Determination of reference sequence – take the sequence of the total number of circulation x_0' as the reference sequence:

$$x_0' = (31295.7, 31468.7, 31267.5, 32823.3, 38151).$$

3) Data processing by initialization method – the behaviour sequence of related factors is processed by Equation (2) as follow:

$$x_i(k) = \frac{x_i'(k)}{x_i'(1)} \tag{2}$$

and the calculation result is as follows:

$$x_1(k) = \frac{x_1'(k)}{x_1'(1)} = \frac{(2820, 2884, 2850, 2952, 3076)}{2820} = (1, 1.02, 1.01, 1.05, 1.09)$$

$$x_2(k) = \frac{x_2'(k)}{x_2'(1)} = \frac{(41156, 41959, 43382, 42675, 43876)}{41156} = \min_{1 \leq i \leq 3} \min_{1 \leq k \leq 3} |x_0' - x_i(k)|, \max_{1 \leq i \leq 3} \max_{1 \leq k \leq 3} |x_0' - x_i(k)|, \quad (3)$$

(1,1.02,1.05,1.04,1.07)

$$x_3(k) = \frac{x_3'(k)}{x_3'(1)} = \frac{(1893, 2252, 2435, 2650, 3069)}{1893} = \min_{1 \leq i \leq 3} \min_{1 \leq k \leq 3} |x_0' - x_i(k)| = 21118.11,$$

(1,1.19,1.29,1.40,1.62)

4) Calculate.

$$\zeta_i(k) = \frac{\min_{1 \leq i \leq n} \min_{1 \leq k \leq m} |x_0'(k) - x_i(k)| + \rho \times \max_{1 \leq i \leq n} \max_{1 \leq k \leq m} |x_0'(k) - x_i(k)|}{|x_0'(k) - x_i(k)| + \rho \times \max_{1 \leq i \leq n} \max_{1 \leq k \leq m} |x_0'(k) - x_i(k)|}, \quad (4)$$

where ρ is resolution ratio, $\rho \in (0,1)$ and $\rho = 0.5$. The higher value is ρ , the closer relation is.

Calculate Equation (4) with different values of $|x_0'(k) - x_i(k)|$ and then get:

$$\zeta_1 = (0.851, 0.849, 0.841, 0.833, 0.774)$$

$$\zeta_2 = (0.850, 0.837, 0.840, 0.829, 0.767).$$

$$\zeta_3 = (0.847, 0.835, 0.841, 0.827, 0.756)$$

6) Calculation of correlation degree – take above calculation result into the following formula of correlation degree calculation:

$$r_i = \frac{1}{m} \sum_{k=1}^m \zeta_i(k), \quad (5)$$

Then get $r_1 = 0.8296, r_2 = 0.8246, r_3 = 0.8212$, the details are as shown in Table 2.

TABLE 2 Value of the grey correlation degree

| | Library | Cultural centre (station) | Museum |
|--------------------|---------|---------------------------|--------|
| Correlation degree | 0.8296 | 0.8246 | 0.8212 |

The correlation degree sheet above shows that the correlation degree value of library is the largest, and that of cultural centre (station) and museum are the second and the third. The difference among them is slight, which means that they are all closely related to the construction of venues and facilities for public cultural services; however, the library has the relatively closest relation. Therefore, the factor of the largest correlation degree value-the library, is selected as the factor of model analysis, for the improvement of the following model.

2.2 THE FACTORS OF PUBLIC CULTURAL SERVICE CONTENTS

The public cultural service contents involves in rich and colourful activities of many forms, including artistic performance, public cultural activities, skill training,

$$\max_{1 \leq i \leq 3} \max_{1 \leq k \leq 3} |x_0' - x_i(k)| = 74410.2.$$

5) Calculate the correlation coefficient.

The calculation formula of correlation coefficient is shown as follows:

popular science propaganda, entertainment, and so on. The frequency of citizen participation in the public cultural activities is the indirect influencing factor of the citizen's satisfaction degree for the public cultural services. Select two representative factors: public cultural activities and artistic performance, as the main analysis factors, the relevant data is shown in Table 3.

TABLE 3 Forms of public cultural services

| Year | Person-time of participation | Public cultural activities (times) | Artistic performance hall |
|------|------------------------------|------------------------------------|---------------------------|
| 2008 | 307529000 | 41814 | 1944 |
| 2009 | 308746000 | 41828 | 2137 |
| 2010 | 308769000 | 42749 | 2112 |
| 2011 | 318745000 | 42958 | 1956 |
| 2012 | 319580000 | 43876 | 2364 |

Source: China Statistical Yearbook 2013

The correlation degree values of public cultural activities and artistic performance are calculated by the Grey Correlation Method following the same procedures in 2.1.1, calculation results are as shown in Table 4 below:

TABLE 4 Correlation degree value

| | Public cultural activities (times) | Artistic performance |
|--------------------|------------------------------------|----------------------|
| Correlation degree | 0.8275 | 0.8223 |
| Correlation degree | 0.8275 | 0.8223 |

According to the analysis of the data above, the public cultural activities of the organization have the largest correlation degree value as 0.8275 and the correlation degree value of artistic performance is 0.8223, so the difference is also slight. However, for the purpose of further establishment of the following models, the index with a large correlation degree, that is, the public cultural activities of the organization, is selected as the main factor for future model analysis.

2.3 THE FACTORS OF PUBLIC CULTURAL SERVICE QUALITY

As a public cultural service spreader, the public cultural servicer should have good image, professional qualification and higher cultural level to spread the

positive information to the public, and to enhance the possibility of selecting the service of the public. The servicer has a direct contact to the public, whose working attitude, working enthusiasm; age structures and so on will become the key factors of influencing the public satisfaction index gradually and affect the served people to a certain extent.

According to the document literature and previous

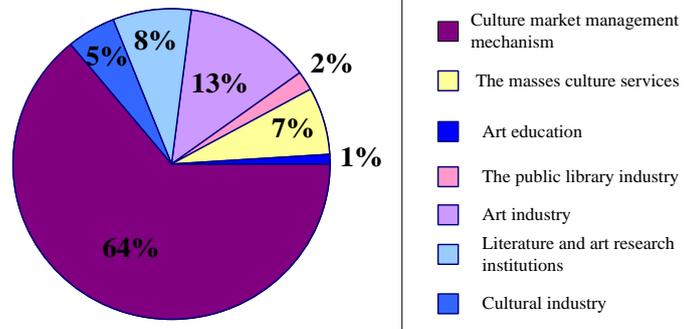


FIGURE 2 Public Cultural Service Personnel

Figure 2 shows that in the Chinese public culture market, the number of people in marketing organization is the largest, accounting for 64% of the total one, thus the number of people in other public cultural service institutions, such as public library, art education institution, social media institution, etc is relatively small, which restricts the development of public cultural services and influences the public satisfaction index to a certain extent. Therefore, the number of people in public cultural servicers was selected as another main factor for the analysis of subsequent model establishment.

3 Public satisfaction index model based on the Grey Correlation AHP Method

On the basis of the grey correlation degree model of the public satisfaction indexes above, the main factors of influencing the public satisfaction index are the number of public library on the number of pavilion, public cultural activities of the organization, and the number of public cultural services. Moreover, the model is improved and the public satisfaction index model based on the grey correlation analytic hierarchy process is established on the footing of the analysis above.

3.1 INITIAL MODEL ESTABLISHMENT

Target layer: public satisfaction index (PSI):

Criterion layer: scheme influence factor C_1 is the public library on the number of pavilion; C_2 is the public cultural activities of the organization; and C_3 is the number of public cultural services.

Scheme layer: A_1 is great satisfaction; A_2 is ordinary and A_3 is not very satisfied.

TABLE 5 Implication of 1~9 ratio scales

research results, the public's requirements for public services, demands for participating in public cultural activities, and needs for promoting spiritual life are becoming higher and higher. However, China is in the shortage of public cultural servicers. According to China's Statistical Yearbook 2013, the number of staff in the Chinese public cultural service institutions is shown as follows:

The layers of relevant influence factors may be separated from top to bottom and the upper layer is influenced by the lower layer, but the factors in various layers are relatively independent. The hierarchical structure is shown as follows:

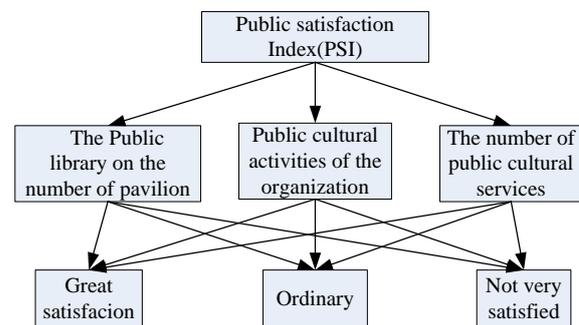


FIGURE 3 Hierarchical structural model

3.2 FACTOR ANALYSIS

An analysis by the Grey Correlation Method shows that the public library on the number of pavilion, the public cultural activities of the organization and the number of public cultural services are the principal indexes which influence the public cultural services, so they are selected as the scheme influence factors for further judgment of the public satisfaction index of the public cultural services.

3.3 CONSTRUCTION OF COMPARATIVE MATRIX

Take the pair-wise comparison of factors respectively and express the importance degree of each factor in layers corresponding to the factors in upper layer by matrix. The 1~9 ratio scales proposed by the operational research experts are quoted in this paper.

| Scale a_{ij} | Definition |
|------------------------------------|--|
| 1 | Factor i is equally important to factor j. |
| 3 | Factor i is slightly important than factor j. |
| 5 | Factor i is more important than factor j. |
| 7 | Factor i is quite important than factor j. |
| 9 | Factor i is absolutely important than factor j. |
| 2,4,6,8 | The scale values in the intermediate state between the judgments above |
| Reciprocal of various values above | If factor i is compared with factor j, the judgment value: $a_{ji}=1/a_{ij}, a_{ij}=1$ |

According to the scale table above, set the judgment matrix as A:

$$A = \begin{pmatrix} 1 & 1 & 5 \\ 1 & 1 & 3 \\ \frac{1}{5} & \frac{1}{3} & 1 \end{pmatrix},$$

where A is a positive reciprocal matrix obviously. The judgment matrix of scheme layers corresponding to the different criterion layers is constructed as follows:

TABLE 6 Judgment matrix of criterion layer of C_1

| C_1 | A_1 | A_2 | A_3 |
|-------|-------|-------|-------|
| A_1 | 1 | 3 | 5 |
| A_2 | 1/3 | 1 | 4 |
| A_3 | 1/5 | 1/4 | 1 |

TABLE 9 Random consistency index

| n | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|------|---|---|------|------|------|------|------|------|------|------|------|
| RI | 0 | 0 | 0.58 | 0.90 | 1.12 | 1.24 | 1.32 | 1.41 | 1.45 | 1.49 | 1.51 |

Consistency ratio: if $CR = \frac{CI}{RI} < 0.1$, the pair-wise comparison matrix constructed passes the consistency check.

2) Weight calculation:

At first, $A = \begin{pmatrix} 1 & 1 & 5 \\ 1 & 1 & 3 \\ \frac{1}{5} & \frac{1}{3} & 1 \end{pmatrix}$ shall be processed as follows:

$$\begin{aligned} &\xrightarrow{\text{Column vector normalization}} \begin{pmatrix} 0.699 & 0.670 & 0.845 \\ 0.699 & 0.670 & 0.507 \\ 0.140 & 0.228 & 0.169 \end{pmatrix} \\ &\xrightarrow{\text{According to the row sum}} \begin{pmatrix} 2.214 \\ 1.876 \\ 0.537 \end{pmatrix} \xrightarrow{\text{Normalized}} \begin{pmatrix} 0.738 \\ 0.625 \\ 0.179 \end{pmatrix} = W^0 \end{aligned}$$

TABLE 7 Judgment matrix of criterion layer of C_2

| C_2 | A_1 | A_2 | A_3 |
|-------|-------|-------|-------|
| A_1 | 1 | 3 | 5 |
| A_2 | 1/3 | 1 | 4 |
| A_3 | 1/5 | 1/4 | 1 |

TABLE 8 Judgment matrix of criterion layer of C_3

| C_3 | A_1 | A_2 | A_3 |
|-------|-------|-------|-------|
| A_1 | 1 | 2 | 3 |
| A_2 | 1/2 | 1 | 3 |
| A_3 | 1/3 | 1/3 | 1 |

3.4 CALCULATION OF THE RELATIVE WEIGHT OF COMPARED FACTOR TO THE CRITERION

1) Consistency check, consistency index:

$$CI = \frac{\lambda_{\max} - n}{n - 1} \tag{6}$$

Random consistency index: generate many matrixes at random, add the consistency index of each matrix and then get the following average RI in Table 9.

$$\text{Then } A \times W^0 = \begin{pmatrix} 2.258 \\ 1.900 \\ 0.533 \end{pmatrix}$$

So $\lambda_{\max}^0 = 3.02610$.

In the same way, the maximum eigenvalue and eigenvector corresponding to the judgment matrix of criterion layer are shown as follows:

$$\lambda_{\max}^{(1)} = 2.874, \omega_1^1 = \begin{pmatrix} 0.883 \\ 0.413 \\ 0.140 \end{pmatrix},$$

$$\lambda_{\max}^{(2)} = 2.874, \omega_2^1 = \begin{pmatrix} 0.883 \\ 0.413 \\ 0.140 \end{pmatrix},$$

$$\lambda_{\max}^{(3)} = 2.865, \omega_3^1 = \begin{pmatrix} 0.875 \\ 0.406 \\ 0.141 \end{pmatrix}.$$

According to the calculation, the maximum eigenvalue of pair-wise comparison matrix $\lambda_{\max} = 3.026$, $RI = 0.58$.

According to consistency index $CI = \frac{\lambda_{\max} - n}{n - 1}$, take the result calculation into Equation (7) to get $CI = \frac{3.026 - 3}{3 - 1} = 0.013$.

Consistency ratio $CR = \frac{CI}{RI} = \frac{0.013}{0.58} = 0.022 < 0.1$, so the

pair-wise comparison matrix A constructed passes the consistency check. In the same way, the judgment matrix of criterion layer passes the consistency check too.

3) Combined weight vector calculation:

$W^1 = (\omega_1, \omega_2, \omega_3)$ and:

$$W = W^1 \times W^0, \quad (7)$$

$$W = \begin{pmatrix} 0.507 \\ 0.273 \\ 0.220 \end{pmatrix}.$$

3.5 RESULT ANALYSIS

According to the calculation results of the combined weight above, for the evaluation of public cultural services, the "great satisfaction" is 50.7%, the "ordinary satisfaction" is 27.3%, and the "not very satisfied" is 22.0%. It can be concluded that most of the public are satisfied with the current Chinese public cultural services and only a small part of them are not satisfied, so the government should still keep a high enthusiasm for work, strengthen the construction of public cultural facilities, make efforts to solve the problems reflected by the public in the process of public cultural services, try to improve public cultural services and better serve the public for the purpose of further promoting the public satisfaction level.

References

- [1] Fornell C 1992 A national customer satisfaction barometer: the Swedish experience *Journal of Marketing* 56(1) 6-21
- [2] Fornell C, Larcker D F 1981 Evaluating structural equation models with unobservable variables and measurement error *Journal of Marketing Research* 18(1) 39-50
- [3] Oliver R L 1980 Dissatisfaction and complaining behavior *Journal of Consumer Satisfaction* 15(2) 1-6
- [4] Liu H Y, Li J, Ge Y X 2006 Design of customer satisfaction measurement index system of EMS service *Journal of China Universities of Posts and Telecommunications* 13(1) 109-13
- [5] Hsu S H 2008 Developing an index for online customer satisfaction: adaptation of american customer satisfaction index *Journal of Expert Systems with Applications* 34(3) 3033-42
- [6] Johnson M D, Gustafsson A, Andreassen T W, Lervik L, Cha J 2001 The evolution and future of national customer satisfaction index models *Journal of Economic Psychology* 22(1) 217-45
- [7] Jianxin Y, Luning S, Miao Y 2004 The public satisfaction philosophy and the evaluation of public satisfaction *Shanghai Management Science* 5(2) 59-61 (in Chinese)
- [8] Li Z, 2006 Government performance evaluation based on public satisfaction *Academic Forum* 29(6) 48-51 (in Chinese)
- [9] Wu L, Xue Y 2006 Measuring users, satisfaction for governmental public services *Journal of Northeastern University (Social Science)* 8(2) 129-132 (in Chinese)
- [10] Mingke S, Guizhong L 2006 Model and method of measurement public satisfaction about government service *Hunan Social Sciences* 19(6) 36-40 (in Chinese)
- [11] Yangqing O, Feng L, Kaifeng C 2006 Application of fuzzy analysis hierarchy process to the customer's satisfaction *Applied Science and Technology* 33(5) 40-2 (in Chinese)
- [12] Rong M, Yu Z, Zhiyong D 2005 Application of satisfaction degree in teaching quality evaluation *Journal of Hebei Institute of Architectural Science and Technology (Social Science Edition)* 22(1) 114-6 (in Chinese)
- [13] Wen T, Aizu Ch 2005 Questionnaire test of customer satisfaction evaluation *Application of Statistics and Management* 24(1) 55-61 (in Chinese)

4 Conclusions

The public is the service object of the public cultural service and the public satisfaction index is the main indicator in the judgment of the effect of public cultural services. In this paper, the Grey Correlation Degree Method was applied to selecting the main factors which influence the public satisfaction index of public cultural services. The public satisfaction model of the Grey Correlation AHP Method was established to evaluate the public satisfaction of the current public cultural services in China and the specific countermeasures are also put forward.

1) Three main factors from many evaluation indexes of the public cultural services involved are selected in this paper and the Grey Correlation Method is applied to analysing the correlation degree among different primary indexes and secondary indexes, concluding that the numbers of public library, the public cultural activities of organizations, and the number of staff in public cultural service institutions are the most three important factors.

2) The public satisfaction index model based on the Grey Correlation AHP Method can be improved and reconstructed to evaluate the current public cultural services in China, drawing a conclusion that most of the public are satisfied with the current public cultural services and only a small part of them are not satisfied with it.

3) Reliable and convincing suggestions based on the quantitative evaluation results are raised, recommending that the government should strengthen the support for public cultural services, improve the management mechanism; increase the number of services and promote the construction of public cultural venues and facilities, such as public library, cultural centre and museum and so forth; and actively organize various public cultural activities to enhance the public satisfaction degree to the maximum extent.

- [14] Daqing K, Xumei Zh 2003 Evaluation architecture and method of customer satisfaction for product *Computer Integrated Manufacturing Systems* 19(5) 407-11 (in Chinese)
- [15] Liangyu L, Ping D, Weiwu Y 2011 Research of public cultural service system based on the public satisfaction analysis *China Economist* 26(6) 7-9 (in Chinese)
- [16] Lingyun Zh, Ping D, Weiwu Y 2011 The empirical research of public cultural facilities: satisfaction - case study of shanghai *China Economist* 26(7) 9-11 (in Chinese)
- [17] Quanhua Z, Yan L 2012 The Research of Evaluation Index System on Public Cultural Service Object *Social Science Review* 27(9) 244-7 (in Chinese)
- [18] Ying Ch 2013 Survey on satisfaction degree of demand for the public cultural services in new guangming road of shenzhen city *Library Work and Study* 35(4) 2-6 (in Chinese)
- [19] Yanling L, Fusheng Z 2008 The analysis of farmers's satisfaction index to rural public goods supply and it's influencing factors *The Journal of Quantitative & Technical Economics* 25(8) 3-16 (in Chinese)
- [20] Yanling L, Miao Zh 2013 The satisfaction research of farmers' needs in rural information public services *Chinese Public Administration* 28(10) 119-23 (in Chinese)
- [21] Weijing Y 2010 The evaluation research of rural public goods supply satisfaction *Modern Science and Technology in Rural Areas* 39(4) 54-5 (in Chinese)

| Authors | |
|---|--|
|  | <p>Fu Liping, born in February, 1963, Tianjin City, China</p> <p>Current position, grades: Professor of College of Management and Economics, Tianjin University, China. University studies: B.Sc. in political economics at Beijing Normal University in China. M.Sc. and Ph.D. in world economy from Nankai University in China. Scientific interest: public management, technological innovation. Publications: More than 50 papers published in various journals. Experience: Teaching experience of 29 years, more than 10 scientific research projects.</p> |
|  | <p>Li Juan, born in March, 1979, Tangshan City, Hebei Province, China</p> <p>Current position, grades: Associate professor of College of Yisheng, Hebei United University, China. University studies: B.Sc. in industrial analysis at Jilin Institute of Chemical Technology in China. M.Sc. from Yanshan University in China. Scientific interest: public management, technological innovation. Publications: More than 20 papers published in various journals. Experience: Teaching experience of 8 years, 3 scientific research projects.</p> |

Asymmetric effects of exchange rate pass-through: an empirical analysis among China, the United States and Japan

Yezheng Liu, Jun Liu*

School of Management, Hefei University of Technology, No.193, Tun Xi Road Hefei, Anhui, China

Received 29 January 2014, www.tsi.lv

Abstract

From the perspective of exchange rate direction fluctuations, this paper comparatively studied the asymmetric effect of movements in the nominal exchange rate on consumer prices among China, the United States, and Japan. To this end, the paper used the error correction model (ECM) to conduct an empirical analysis from the first season of 1994 to the last season of 2010 period. The results showed that: (1) the pass-through of exchange rate movements to consumer prices was incomplete; (2) exchange rates fluctuated in different directions, meaning that when the exchange rate appreciated and depreciated, the pass-through of exchange rate movements to consumer prices was asymmetric. However, the direction varied among the three countries. The influence of depreciation on consumer prices was higher in both China and the United States, while Japan was the opposite; (3) exchange rate pass-through was different in the three countries. The level of exchange rate pass-through in China was higher than the other two countries; (4) when short-term fluctuations deviated from long-term equilibrium, the adjustment was higher in the United States, followed by Japan, and China was relatively lower. These results had important implications for current monetary policies and practices.

Keywords: exchange rate, consumer prices, exchange rate pass-through, asymmetry

1 Introduction

In 2005, China reformed the exchange rate regime by moving into a managed floating exchange rate system based on market supply and demand with reference to a basket of currencies. Exchange rate movements increased, and the impact on the national economy increased. Exchange rate changes may cause fluctuations in domestic price levels, i.e. the exchange rate pass-through. Exchange rate pass-through is a hot issue in the field of international economics researches. It usually assumes that the relationship between the price level and the exchange rate is symmetric. This means that appreciation and depreciation will be transmitted to the final price with the same magnitude. However, in real situations the asymmetry is widespread within the prices of final goods, and is often manifested in the fact that prices rise easier than they fall. Such phenomenon is contrary to traditional assumptions, meaning the assumption of symmetric pass-through does not match with reality. Meanwhile, omitting the asymmetric effect will lead monetary policy effects to serious deviations. Therefore, the symmetric pass-through assumption has been relaxed in some new empirical researches [1].

Current studies have mostly focused on the asymmetric pass-through in developed countries; few studies have been done on developing countries. Moreover, the literature generally focuses on import prices, and the analysis of consumer prices is quite sparse. However, compared with other prices, there are large differences in

the pass-through mechanism associated with consumer prices. Furthermore, due to different backgrounds, there may be large differences in asymmetry across different countries. Most importantly, as the largest developing country in the world, and due to the fact that it is in a context of transition, it is very important for China to analyse asymmetric exchange rate pass-through to consumer prices. Based on the above considerations, this paper selected consumer prices as the research object. It comparatively analysed the asymmetric effects of exchange rate pass-through on consumer prices by looking at China, the United States and Japan. On one hand, we can measure changes in the domestic price level caused by exchange rate fluctuations. On the other hand, CPI is a measure of inflation. Therefore, to analyse the impact of exchange rate changes on CPI, we can uncover the impact of exchange rate changes on domestic inflation. This allows us to provide policy guidance for inflation forecasts as well as policy recommendations. A comparative analysis of these representative countries can contribute to a better understanding of the status, and also provide some references for current monetary policy practice.

This paper selected the three above countries for the following considerations: first, these three countries have an important influence around the world, and they have important positions in the global economy. Second, the level of economic development differs among these countries. The United States and Japan belong to developed countries, and China is a developing country. Third, these countries are at different stages of economic

* *Corresponding author* e-mail: liujun83@163.com

development. The United States has a mature economy with a higher level of economic development. Its development is more robust. China has an emerging and developing economy that features rapid economic development. Meanwhile, Japan's economy has been experiencing a prolonged slump; the pace of economic development here is slow. Fourth, there is a great deal of differences in the exchange rate systems of these countries. The United States and Japan implement the free floating exchange rate system, while China is implementing a managed floating exchange rate system. Fifth, China has close economic ties with both the United States and Japan; they are China's important trading partners. Sixth, the availability of data is quite good for all of these countries.

Early studies of asymmetric pass-through adopted differential equations for analysis [2]. Although differential conversion can overcome the non-stationary problems associated with time series, the transformation will cause a loss of long-term information on the relationship between economic variables. It can also lead to a serial correlation of the regression model error, resulting in a failure of the regression analysis. There are also studies which use a distributed lag model [1]. However, because the model uses time-series data, it is possible for residuals to create autocorrelation problems within this regression method. Based on the above considerations, we used the error correction model (ECM). This method can better characterize both the long-term and short-term dynamic changes in variables. It maintains the long-term dynamic information of the relationship between variables. It also ensures the effectiveness of the regression analysis. In specific applications, this paper will introduce dummy variables into the model for analysis. Furthermore, it introduces a monetary policy variable into the model, and selects a foreign price index constructed to measure the cost of foreign exporters. By taking into account the lagged impacts of the exchange rate and monetary policy changes on the price index, this model takes a lag respectively.

In conclusion, this paper selected three representative countries: China, USA and Japan. It empirically studied the asymmetric effect of nominal exchange rate changes on consumer prices. The results showed that the pass-through of exchange rate changes to consumer prices was incomplete. For different directions of fluctuations the asymmetry of pass-through existed, but there were differences in the direction. In the United States and China, the impacts of exchange rate depreciation on consumer prices were higher; Japan was exactly the opposite. When the short-term fluctuations deviated from the long-term equilibrium, the adjustment was higher in the United States, followed by Japan, while China was relatively lower.

The remaining parts of the article are organized as follows: Section 2 reviews and summarizes the existing literature; Section 3 presents a theoretical model that has been constructed for studying the asymmetric effect of exchange rate changes on consumer prices; Section 4

presents the empirical results, comparatively studying asymmetric effects of exchange rate changes on consumer prices in China, the U.S. and Japan; Section 5 provides the main conclusions.

2 Literature review

2.1 THEORETICAL STUDIES

According to the causes of asymmetric pass-through in the direction of exchange rate movements, there are currently four accepted theoretical explanations.

Market share: When the goal of manufacturers is to build up market share, an appreciation in the importing country's currency will cause manufacturers to reduce import prices in order to increase their market share while maintaining mark-up. However, when there is depreciation, exporters will offset the potential increase in prices to maintain their market share by lowering mark-up. Therefore, the pass-through effect is higher in the case of appreciation than for that of depreciation [3].

Production switching: In this model, when making use of inputs, foreign manufacturers will switch between imported and domestically produced inputs depending on prices. When the importing country's currency appreciates, foreign manufacturers will only use domestically produced inputs. At that time, the level of pass-through depends on the elasticity of the mark-up. Foreign manufacturers will use inputs from the depreciating country during instances of devaluation, and then no pass-through takes place [4].

The two above theories both believe that the pass-through effect is higher for appreciation than depreciation. The following two explanations provide exactly the opposite interpretation. Here, the extent of exchange rate pass-through is higher during the depreciation.

Quantity constraints: This theory is also known as capacity constraints. Here, an appreciation of the importing country's currency will cause foreign manufacturers to reduce import prices. However, quantity rigidities will limit the expansion of its sales through low prices. Therefore, foreign manufacturers will increase mark-up to keep import prices unchanged in the currency of the importing country so as to guarantee sales and raise profit margins. In the case of depreciation, foreign firms are likely to increase the price of products in the importing country in order to reduce losses caused by devaluation. Therefore, the pass-through effect is higher for depreciation than appreciation [3].

Market structure: This theory holds that different levels of the pass-through effect during appreciation and depreciation are caused by monopoly. When the home currency appreciates (if foreign firms have monopoly power within the domestic market), they are likely to keep commodity prices unchanged in the local currency (a corresponding increase in the price level measured in the currency of foreign firms). This can increase the profits of foreign firms (measured in the currency of foreign firms), then, it shows a lower exchange rate pass-through effect or

no exchange rate pass-through effect at all. Conversely, if the home currency depreciates, foreign firms will raise the price of goods in order to maintain their profits (measured in foreign currency). This shows higher exchange rate pass-through effect [1].

2.2 EMPIRICAL STUDIES

Early asymmetry studies generally focused on the direction of exchange rate fluctuations to conduct tests. Studies usually used two methods. One analysed whether the pass-through effect differed during periods of appreciation and depreciation. For example, [5] studied whether there were differences in the level of pass-through to U.S. import prices in the periods of exchange rate depreciation (1977-1980) and appreciation (1981-1985). The other introduced dummy variables to identify exchange rate appreciation and depreciation. These studies were generally performed from the two levels of the industry and aggregated price.

Industry level studies have discovered that, in the United States, the level of pass-through is higher during periods of depreciation [6, 7]. Other studies found that the extent and direction of asymmetric pass-through varied between the industries and countries. Even within the same industry, there were also differences in the direction of asymmetric pass-through across countries [2]. By contrast, only a few studies have investigated the issue at the aggregated price. Studies found asymmetry in pass-through, the extent was higher for appreciation than depreciation [5, 8]. Instead, [4] found the extent of pass-through was higher for depreciation than appreciation in seven Asian countries.

At the same time, the literature involving consumer prices is relatively sparse. Studies found the existence of asymmetric responses. Here, the effect of exchange rate depreciation on consumer prices was higher than for appreciation [1].

3 Theoretical model

3.1 MODEL SPECIFICATION

Making use of the profit-maximization behaviour of an exporting firm, we analysed the relationship between the price level for exporting firms and exchange rate fluctuations.

$$\pi = P(Q) \cdot Q - C(Q) \tag{1}$$

Given profit-maximization, we know

$$MR = MC,$$

$$MR = \partial P(Q) \cdot Q / \partial Q = P + Q \cdot dP / dQ = P + P \cdot (Q / P) \cdot dP / dQ$$

That is $MR = P + P \cdot (1 / E_d) = MC,$

where E_d denotes the price elasticity of demand. We can get

$$P = MC / [1 + 1 / E_d], \tag{2}$$

$$\text{Set } \mu = 1 / [1 + 1 / E_d],$$

where π denotes profits, expressed in the exporting country's currency. P is the price of goods in the foreign currency. e is the exchange rate, adopting the direct quotation method, measured in units of home currency per unit of the exporting country's currency. $C(\cdot)$ is the cost function in the exporting country's currency, Q is the quantity demanded for goods. P^d denotes the price of goods in domestic currency. From Equation (2) we can get:

$$P^d = eC_q \mu, \tag{3}$$

where C_q is the marginal cost and μ is the mark-up, which depends on the price elasticity of demand for goods. Therefore, the price of goods in domestic currency depends on exchange rate, marginal cost, and mark-up. The marginal cost depends on local input cost. The mark-up depends on the demand conditions in the importing country.

Taking Equation (3) in logarithm, the price equation can be expressed as follows:

$$P_t^d = \alpha_0 + \alpha_1 e_t + \alpha_2 P_t^* + \alpha_3 Y_t + \varepsilon_t, \tag{4}$$

where P^* denotes the exporting firm's marginal cost and Y denotes demand conditions in the importing country. From the reviewed literature on pass-through, it can be seen that Equation (4) is generally used for model specification. Equation (4) is generalized as follows:

$$P_t = \alpha + \delta X_t + \gamma E_t + \psi Z_t + \varepsilon_t, \tag{5}$$

where P_t represents the price index, X_t means the control variables abroad (usually using the cost or price to be measured according to the type of research). E_t refers to the exchange rate, Z_t refers to domestic control variables.

With regard to P_t , we selected consumer prices as a research object. Given that consumer prices are different from other prices, and the pass-through mechanism varies. We adopted the consumer price index (CPI) as a proxy variable for consumer prices. For a control variable X_t , by referring to [9], we utilized foreign price index as a proxy variable to measure the cost for foreign exporters. $WPI_t = (CPI_t \cdot Neer_t / Reer_t)$, $Neer$ and $Reer$ denote nominal effective exchange rate and real effective exchange rate. For control variable Z_t , we selected GDP and money supply; the former reflects domestic demand conditions, while the latter reflects the monetary policy

factor. By referring to [10] and introducing monetary policy, we established the following model:

$$LCPI_t = \beta_0 + \beta_1 \cdot LNER_t + \beta_2 \cdot LM2_t + \beta_3 \cdot LGDP_t + \beta_4 \cdot LWPI_t + \varepsilon_t, \quad (6)$$

$(t=1,2,3\dots n)$

where *LCPI*, *LNER*, *LM2*, *LGDP*, and *LWPI* respectively refer to the consumer price index, RMB nominal effective exchange rate, money supply, gross domestic product, and foreign price index. All of the variables are expressed in logarithms.

LCPI, *LNER*, *LM2*, *LGDP*, and *LWPI* and certain linear combinations of the variables are stationary. We can make use of ε_t generated from Equation (6) to construct an error correction model (ECM). The error correction model settings are as follows:

$$\Delta LCPI_t = \sum_{i=0}^p \alpha_{1i} \cdot \Delta LNER_{t-i} + \sum_{i=0}^q \alpha_{2i} \cdot \Delta LM2_{t-i} + \alpha_3 \cdot \Delta LGDP_t + \alpha_4 \cdot \Delta LWPI_t + \alpha_5 \cdot AR(1) + \alpha_6 \cdot \varepsilon_{t-1} + \mu_t, \quad (7)$$

where *p*, *q* represents the *LNER*, *LM2* lags respectively. Here, the lags are taken because of the lagging impact that the exchange rate and monetary policy changes have on the price index. This often manifests itself over several periods.

In order to analyse asymmetry in exchange rate pass-through to consumer prices, we introduced a dummy variable. This was done in order to investigate whether the effect of exchange rate pass-through varied in the case of appreciation and depreciation.

Assuming that,

$$D_1 = \begin{cases} 0, & \text{if effective exchange rate appreciates} \\ 1, & \text{if effective exchange rate depreciates} \end{cases}$$

we constructed the model as follows:

$$\Delta LCPI_t = \alpha_0 + \alpha_1 \cdot \Delta LNER_t + \gamma_1 \cdot D_1 \cdot \Delta LNER_t + \sum_{i=1}^p \alpha_{1i} \cdot \Delta LNER_{t-i} + \sum_{i=0}^q \alpha_{2i} \cdot \Delta LM2_{t-i} + \alpha_3 \cdot \Delta LGDP_t + \alpha_4 \cdot \Delta LWPI_t + \alpha_5 \cdot AR(1) + \alpha_6 \cdot \varepsilon_{t-1} + \mu_t, \quad (8)$$

We presented following hypotheses:

$H_0: \gamma_1 = 0$, meaning there are no statistical differences, and exchange rate pass-through is symmetric.

$H_1: \gamma_1 \neq 0$, meaning there are statistically significant differences, and exchange rate pass-through is asymmetric.

The above hypotheses will be tested in the empirical analysis. If the coefficients of the dummy variable are significant statistically then the null hypothesis will be rejected.

3.2 SOURCES OF DATA AND DESCRIPTION

3.2.1 Sources of data and processing

With the use of Eviews 6.0, we selected quarterly data from the first season of 1994 to the last season of 2010 to conduct our empirical analysis. We selected the consumer price index as a proxy variable for consumer prices, and selected the nominal effective exchange rate as a proxy variable for the exchange rate, taking the indirect quotation, an increase of the index means exchange rate appreciation. Given the different base periods for the original data, we converted various indexes to the first season of 1994 as the base period. In order to eliminate the impact of seasonal factors, the variables were seasonally adjusted. China's consumer price index, money supply and GDP were derived from the China Economic Information Network statistics database. The effective exchange rate was obtained from the IMF International Financial Statistics (IFS). The U.S. and Japanese data was mainly collected from the IFS.

3.2.2 Data description

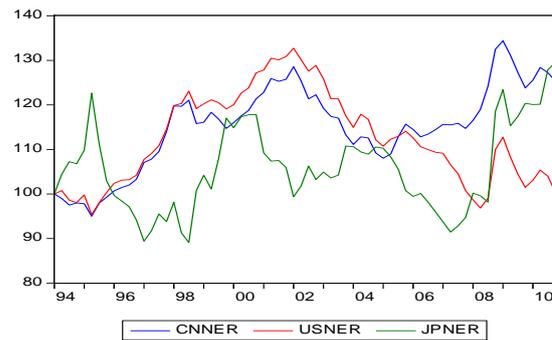


FIGURE 1 Exchange rate chart showing China, the U.S. and Japan

Figure 1 shows that, after 1995, the RMB nominal effective exchange rate experienced a waved rise, reaching its peak in 2002, and then fell to its lowest level in 2005. After the exchange rate system reform, the RMB exchange rate experienced a wave of rapid rise, reaching its maximum in 2008, and after the financial crisis it declined. From the figure, we can see that the magnitude of exchange rate appreciation was higher than depreciation. In the United States, the exchange rate also experienced a waved rise from 1995 to 2002, and then it declined until 2009. After that, it rebounded after the financial crisis and later fell again. On the whole, the sizes of appreciation and depreciation were considerable. The Yen exchange rate appreciation and depreciation were staggered. Exchange rate depreciation lasted a long time. The magnitudes of appreciation and depreciation were considerable. After both the 1998 Southeast Asian economic crisis and the 2008 financial crisis, the yen exchange rate showed a wave of rapid rise. It reached its maximum in 2010.

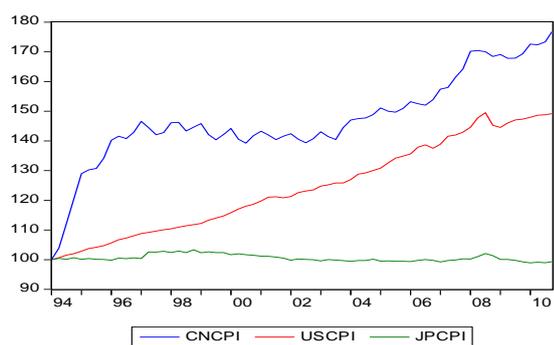


FIGURE 2 CPI chart showing China, the U.S. and Japan

As can be seen from Figure 2, prior to 1996, China experienced a wave of rapid rise in CPI and then it tended to be stable. After that, it entered an upward trend, and the exchange rate declined during the financial crisis. Overall, there has been an upward trend in U.S. CPI and the exchange rate declined briefly during the financial crisis. Japan's CPI showed low level fluctuations generally. The trend was relatively stable. During Southeast Asian

economic crisis and the financial crisis, CPI fluctuated greatly.

4 Empirical results

4.1 UNIT ROOT TEST

In order to produce more robust results when studying the relationship between variables, it is necessary for the *LCPI*, *LNER*, *LM2*, *LGDP*, and *LWPI* sequences to conduct a unit root (ADF) test. According to Information Criterion (AIC) and Schwarz Criterion (SC) we chose lags according to the principle of minimum. The null hypothesis is that a unit root is present. Table 1 shows the unit root test results.

The unit root test shows that at the 5% significance level, the *LCPI*, *LNER*, *LM2*, *LGDP* and *LWPI* original series are non-stationary series. However, after the first difference they all become stationary series. Therefore, we can consider that the *LCPI*, *LNER*, *LM2*, *LGDP* and *LWPI* are integrated with order 1 processes.

TABLE 1 ADF test results

| Variable | China | | The U.S. | | Japan | |
|---------------|----------------|----------------|----------------|----------------|----------------|----------------|
| | ADF statistics | Result | ADF statistics | Result | ADF statistics | Result |
| LCPI | 0.624 | Non-stationary | -0.894 | Non-stationary | -2.214 | Non-stationary |
| LNER | -2.017 | Non-stationary | 0.381 | Non-stationary | -2.001 | Non-stationary |
| LM2 | 0.788 | Non-stationary | 0.231 | Non-stationary | -1.147 | Non-stationary |
| LGDP | 1.871 | Non-stationary | -2.424 | Non-stationary | -1.101 | Non-stationary |
| LWPI | -0.547 | Non-stationary | -1.931 | Non-stationary | -2.129 | Non-stationary |
| Δ LCPI | -4.541 | Stationary | -8.697 | Stationary | -3.455 | Stationary |
| Δ LNER | -5.527 | Stationary | -6.017 | Stationary | -3.764 | Stationary |
| Δ LM2 | -6.68 | Stationary | -8.908 | Stationary | -3.028 | Stationary |
| Δ LGDP | -3.865 | Stationary | -6.264 | Stationary | -6.955 | Stationary |
| Δ LWPI | -5.567 | Stationary | -3.209 | Stationary | -9.583 | Stationary |

4.2 CO-INTEGRATION ANALYSIS

Since the premise of the error correction model is the existence of a co-integration relationship between variables, we needed to conduct a co-integration test. This is because the variables are first difference stationary

series. Then, we applied the EG two-step method to do a co-integration test for the variables to determine the long-term stable relationship among them. First, we performed the OLS regression model. Then we performed an ADF test for the residual sequence. Table 2 shows the results.

TABLE 2 Co-integration equation results

| Variable | China | | The U.S. | | Japan | |
|-------------------------|-------------|--------------|-------------|--------------|-------------|--------------|
| | Coefficient | T-statistics | Coefficient | T-statistics | Coefficient | T-statistics |
| LNER | -0.249** | -2.458 | -0.158*** | -17.67 | -0.049*** | -4.92 |
| LM2 | 0.215* | 1.907 | 0.125*** | 4.98 | -0.155*** | -4.39 |
| LGDP | -0.549*** | -3.483 | 0.118*** | 2.809 | -0.249*** | -8.04 |
| LWPI | 1.952*** | 5.304 | 0.52*** | 11.98 | 0.225*** | 6.78 |
| C | -0.29 | -0.191 | 0.297 | 0.957 | 10.01*** | 20.39 |
| R ² | 0.88 | | 0.997 | | 0.72 | |
| Residual ADF statistics | -2.68*** | | -3.1** | | -3.05** | |

NOTES: ***, **, * indicate the statistical significance at the 1%, 5% and 10% levels, respectively.

The residuals of the regression equation above are stationary. Therefore, the model specification is reasonable. It indicates the presence of a long-term co-

integration relationship between *LCPI*, *LNER*, *LM2*, *LGDP* and *LWPI*. The results also indicate that the long-term exchange rate pass-through to consumer prices is not

complete. The pass-through effects of exchange rate changes on consumer prices are: -0.249, -0.158, -0.049 in China, the U.S. and Japan respectively. Meanwhile, the pass-through effect of the RMB exchange rate on consumer prices is -0.249. This means that the RMB nominal effective exchange rate appreciates by 1% and consumer prices fall 0.249 percentage points. In this regard, our argument is that this is because the consumer goods contain a lot of non-tradables on their own. The delivery of non-tradable goods to consumers require many distribution chains. This further increases the ingredients of non-tradables in the consumer goods. However, the

impact of exchange rate changes on non-tradables is smaller. This results in the incomplete exchange rate pass-through to consumer prices.

4.3 ERROR CORRECTION MODEL ESTIMATE

From the direction of the exchange rate changes in Model (8), during the appreciation and depreciation we tested whether the degree of exchange rate pass-through to consumer prices varied. Namely, we tested for the existence of asymmetric pass-through. Table 3 gives the results.

TABLE 3 Error correction model estimate results

| Variable | China | | The U.S. | | Japan | |
|-------------------|-----------------------|----------------------|---------------------|-----------------------|---------------------|------------------------|
| | Appreciation | Depreciation | Appreciation | Depreciation | Appreciation | Depreciation |
| $\Delta LNER$ | -0.106*** (-2.997) | -0.168** (-2.644) | -0.021 (-1) | -0.051** (-2.417) | -0.04* (-1.904) | -0.02** (-2.235) |
| $\Delta LNER(-1)$ | 0.012 (0.36) | 0.051 (0.693) | 0.01 (0.487) | -0.02 (-0.782) | -0.031 (-1.321) | 0.035*** (3.558) |
| $\Delta LM2$ | -0.019 (-0.491) | -0.176** (-2.747) | -0.081* (-1.804) | -0.174*** (-2.939) | -0.063 (-0.464) | -0.071 (-1.412) |
| $\Delta LM2(-1)$ | -0.027 (-0.732) | 0.105* (2.086) | -0.004 (-0.095) | 0.019 (0.369) | -0.068 (-0.596) | -0.057 (-1.157) |
| $\Delta LGDP$ | -0.025 (0.881) | -0.021 (-0.694) | -0.09 (-1.547) | -0.327*** (-3.602) | -0.139* (-1.977) | -0.361*** (-10.101) |
| $\Delta LWPI$ | 0.867*** (10.451) | 0.728*** (3.764) | 0.426*** (3.262) | 0.254 (1.538) | 0.439*** (3.537) | 0.208*** (4.276) |
| $\Delta ECM(-1)$ | -0.013 (0.202) | -0.211* (-1.865) | -0.174 (-1.112) | -0.291* (-1.731) | -0.092 (-0.706) | 1.005*** (-6.574) |
| AR(1) | 0.362** (2.456) | 0.728** (2.505) | 0.663** (2.651) | 0.578** (2.613) | 0.505* (1.676) | 0.733*** (7.513) |
| AR(2) | 0.318** (2.608) | 0.711** (2.335) | | | | |
| R ² | 0.92 | 0.97 | 0.75 | 0.92 | 0.67 | 0.81 |
| D.W. | 1.93 | 2.23 | 2.01 | 2.11 | 1.52 | 1.77 |

NOTES: ***, **, * indicate the statistical significance at the 1%, 5% and 10% levels, respectively. The values in the bracket are t-statistics

The results show that when exchange rate depreciates, the pass-through elasticity of the RMB exchange rate to consumer prices is -0.168, meanwhile, the pass-through rate is -0.106 in the case of appreciation. For the United States, when depreciating, the pass-through rate is -0.051, the value is -0.021 in the case of appreciation. For Japan, the pass-through rate is -0.02 for depreciation, and conversely, -0.04 for appreciation. It is clear that there is a negative relationship between prices and exchange rate changes. In other words, exchange rate depreciation (appreciation) will drive prices to increase (decrease). This result is consistent with traditional economic theory. Secondly, responses from prices according to exchange rate fluctuations are statistically significant. However, there are differences in the level of pass-through. For appreciation and depreciation, the degree of exchange rate pass-through to consumer prices varied. Namely $\gamma_1 \neq 0$, meaning the pass-through of exchange rate changes to consumer prices is asymmetric. Therefore, the hypothesis H_1 holds.

Thus, there are asymmetric effects of exchange rate changes on consumer prices. This means that the effect of exchange rate pass-through on consumer prices is influenced by the direction of exchange rate fluctuations. For different directions of fluctuations, the exchange rate pass-through is asymmetric, but the direction varies for both. When the yen appreciates, the pass-through effect on consumer prices is higher. For these results, we argue that it may be related to the long-term downturn in the economic environment. Since the 1990s, the Japanese real estate bubble has burst and the economy declined. This decline has lasted a long time and Japan has still not emerged from a state of economic downturn. In a recession, the price level generally shows a downward trend. Even though exchange rate depreciates, firms do not adjust the price in proportion to the rise in costs. Here, the smaller the change in price, the lower the exchange rate pass-through effect is. On the contrary, appreciation will depress price levels. Furthermore, the recession will accelerate the declines in the price level. The larger the change in prices, the higher the exchange rate pass-through

effect is. By contrast, in China and the United States, when the currency depreciates, the pass-through effect on consumer prices is larger. In this regard, we argue that this phenomenon should be consistent with market structure theory. This is because monopoly power exists in the market. This results in differences in the degree of exchange rate pass-through for appreciation and depreciation. Specifically, and in the short-term, when exchange rate depreciates there is upward pressure on prices. Here, production costs increase, firms have incentive to adjust prices upward, prices are vulnerable to adjustment and the pass-through effect on consumer prices is greater. Meanwhile, for appreciation there is downward pressure on prices, but because of incomplete competition, firms have some monopoly power in prices. Here, price adjustments are smaller or they remain unchanged, and the effect of exchange rate pass-through on consumer prices is smaller.

In addition, the effect of exchange rate pass-through on consumer prices varies across countries. China's exchange rate pass-through stays at the higher level, followed by the United States, while Japan is the lowest. Among them, when the RMB exchange rate depreciates, the pass-through rate is -0.168. Meanwhile, in Japan, this value is only -0.02. Why is China's pass-through rate higher than two other countries? Regarding this, our argument is that this may be related to the composition of imports. Within China's imports, primary products are relatively high. Here, the proportion increased from 12.4% in 1985 to 25.4% in 2007. Recently, according to the needs of economic development, China imported a large proportion of resources products. The proportion of mineral fuels, lubricants and related raw material imports accounted for primary products increased from 6.5% in 2002 to 11% in 2007. Here, the proportion of non-food raw materials increased from 7.7 % to 12.4%. The prices of primary products are more sensitive to exchange rate changes. This results in a higher level of exchange rate pass-through in China than in the U.S. and Japan.

Finally, from the coefficients of the error correction term, it can be seen that the coefficient is higher in America, followed by Japan, and China is relatively low. This means that short-term fluctuations deviate from the long-term equilibrium in China. Here, the magnitude of adjustment is weak. On the contrary, in the United States the adjustment from non-equilibrium to equilibrium is relatively high. These results may be related to the fact that the U.S. has a relatively mature market environment. On the one hand, the U.S. implements a freely floating exchange rate system. When shocks burst out in the market it causes a deviation from the long-run equilibrium level. Here, the floating exchange rate system can adjust the exchange rate to absorb these shocks, tending to the equilibrium level. On the other hand, there are relatively mature participants and a perfect organizational system in the U.S. market. Here, the responses to shocks are high, and through their reactions, this tends to make the deviations tend towards equilibrium. Therefore, relative to

the U.S., it is relatively backward for China in the participants or organization system involved in, it is natural that the magnitude and speed of adjustment is lower than that of the United States.

5 Conclusions

Exchange rate pass-through is a hot issue in the field of international economics researches. From the perspective of the direction of exchange rate fluctuations, we comparatively studied the asymmetric effect of nominal exchange rate changes on consumer prices in China, the United States and Japan. We applied the error correction model (ECM) to perform an empirical analysis on quarterly data from the first season of 1994 to the last season of 2010. The conclusions are summarized as follows:

First, exchange rate pass-through to consumer prices was incomplete. Regardless of whether we looked at long-term or short-term, exchange rate changes had an effect on consumer prices. However, the degree was incomplete. For different directions of exchange rate changes, exchange rate pass-through to consumer prices was asymmetric. However, the direction varied in all three countries. In the United States and China, the effect on consumer prices was higher during the depreciation, but in Japan exactly the opposite was true. For the former, this paper has argued that the results may be related to imperfect competition within the market. Meanwhile, the latter may be related to a long-term downturn in the economic environment. In order to further reveal an intrinsic relationship, micro-economic corporate pricing behaviour should be the key point of focus. Based on this assumption, industry-level data would provide more valuable information.

Second, exchange rate pass-through varied across countries. The level of China's exchange rate pass-through was higher, followed by the United States, and Japan was the lowest. Among the first two countries, when the RMB exchange rate depreciated, the pass-through rate was -0.168, whereas this value was only -0.02 for Japan. We argued that this may be related to the composition of China's imports. When short-term fluctuations deviated from the long-term equilibrium, the magnitude of adjustment was higher in the United States, followed by Japan, and China was relatively low. This may be related to the U.S.'s relatively mature market environment. This indicates that China needs to further improve its own market environment, and improve policy efficiency. On one hand, China should improve the market organization system. It should also strive to improve the transparency and efficiency of policy formulation, and improve its policy transmission mechanism. Furthermore, China should improve its policy transmission efficiency and strengthen legal constructions. It should also regulate market operations more effectively. On the other hand, China needs to foster and enhance the ability of market participants. It should also regulate the behaviour of

various parties, and improve the maturity of market participants.

Finally, the asymmetric effect of exchange rate pass-through will have a significant impact on monetary policy, driving policymakers into dilemmas as they try to pursue price stability and export competitiveness. Under equivalent conditions, currency depreciation assists in reducing export prices. It also strengthens export competitiveness. However, depreciation has inflationary effects on domestic price levels. The results also imply that when making monetary policy rules in the context of incomplete exchange rate pass-through, the direction of exchange rate variations also needs to be considered.

References

- [1] Delatte A-L, López-Villavicencio A 2012 Asymmetric exchange rate pass-through: Evidence from major countries *Journal of Macroeconomics* **34**(3) 833-44
- [2] Gil-Pareja S 2000 Exchange rates and European countries' export prices: An empirical test for asymmetries in pricing to market behaviour *Weltwirtschaftliches Archiv* **136**(1) 1-23
- [3] Knetter M M 1994 Is export price adjustment asymmetric? Evaluating the market share and marketing bottlenecks hypotheses *Journal of International Money and Finance* **13**(1) 55-70
- [4] Webber A G 2000 Newton's gravity law and import prices in the Asia Pacific *Japan and the World Economy* **12**(1) 71-87
- [5] Mann C L 1986 Prices, profit margins, and exchange rates *Fed Reserve Bull* **72**(6) 366-79
- [6] Goldberg P K 1995 Product Differentiation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry *Econometrica* **63**(4) 891-951
- [7] Olivei G P 2002 Exchange rates and the prices of manufacturing products imported into the United States *New England Economic Review* **1** 3-18
- [8] El Bejaoui H J 2013 Asymmetric effects of exchange rate variations: An empirical analysis for four advanced countries *International Economics* **135-136** 29-46
- [9] Campa J M, Goldberg L S 2005 Exchange rate pass-through into import prices *Review of Economics and Statistics* **87**(4) 679-90
- [10] Parsley D C, Popper H 1998 Exchange rates, domestic prices, and central bank actions: recent US experience *Southern Economic Journal* **64**(4) 957-72

When the exchange rate fluctuates in different directions, the pass-through effect varies, and the impact on domestic prices is asymmetric. Therefore, the policy departments should consider these differences when setting monetary policy rules.

Acknowledgements

We would like to thank Chunhua Sun, Xiumin Chu and Chuanqi She, as well as the participants in the seminar during the weekend at Hefei University of Technology, China. We are also grateful to them for their useful and constructive comments and suggestions.

| Authors | |
|---|--|
|  | <p>Yezheng Liu, born in September, 1965, Hefei County, Anhui Province, China</p> <p>Current position, grades: the professor and doctoral supervisor of School of Management, Hefei University of Technology, China. University studies: M.Sc. in applied physics from Hefei University of Technology in China, Ph.D. in management science and engineering from Hefei University of Technology in China. Scientific interest: business intelligence, data mining. Publications: more than 41 papers. Experience: teaching experience of 28 years, 30 scientific research projects.</p> |
|  | <p>Jun Liu, born in January, 1983, Hefei County, Anhui Province, China</p> <p>Current position, grades: a Ph.D. student of School of management, Hefei University of Technology, China. University studies: B.E. in economics from Anhui University in China, M.E. in finance from Anhui University in China. Scientific interest: econometrics. Publications: more than 3 papers. Experience: teaching experience of 1 year, 1 scientific research project.</p> |

Deformation forecasting with a novel high precision grey forecasting model based on genetic algorithm

Ning Gao*, Cai-Yun Gao

School of Geomatics and City Spatial Information, Henan University of Urban Construction, Pingdingshan City, Henan Province, China, 467036

Received 1 June 2014, www.tsi.lv

Abstract

The precision of prediction of grey forecasting model depends on the conformation of background value and the selection of the initial condition. Existent literatures optimized grey forecasting model just from one side, respectively. Therefore, a novel model named BIGGM (1,1) is proposed in this paper by integrated optimizing background value and initial condition. In addition, genetic algorithm has also been integrated into the new model to solve the optimal parameter estimation problem. An illustrative example of deformation of Lianzi cliff dangerous rock along the Yangtze River in china is adopted for demonstration. Results show that the BIGGM (1,1) model can increase the prediction accuracy, and it is suitable for use in modelling and forecasting of deformation.

Keywords: GM (1,1) model, background value, initial condition, genetic algorithm, integrated optimization, deformation forecasting

1 Introduction

Deformation monitoring (also referred to as Deformation survey) is the systematic surveying and tracking of the distortion in the shape or dimensions of an object (such as bridges, dams, landslides, high rise buildings, etc.) as a result of stresses induced by applied loads or factors such as changes of ground water level, tectonic phenomena, tidal phenomena, etc. Deformation monitoring is a major component of logging measured values that may be used to for further computation, deformation analysis, predictive maintenance and alarming. Deformation forecasts play an important role in protecting the safety of people's lives and property. Such forecasts results of deformation data is a main basis for decision-making, its quality can directly influence the effect of the whole monitoring work. Therefore, increasing the accuracy of forecasts of deformation is an important issue [1].

Numerous researches have studied the deformation forecasting for a quite long time, they developed various models, including the linear regression model, time series model, artificial neural network (ANNS), etc. Since these methods are easy to use and accurate, they have been used in a wide range of applications. However, these methods have limitations. For instance, the linear regression model assumes that variables are independent and that samples have normal distribution. This model, therefore, requires a larger number of samples. The time series model requires stable trends and patterns of historical data. ANNS demands a great amount of training data and relatively long training period for robust generalization, and the hidden layers in ANNS are difficult to explain [2-5].

Meantime, owing to the complexity of deformation system, only part of the system structure can be fully

realized. Therefore, based on the character of deformation system and the advantage of the grey model, this study applies the grey forecasting model to forecast deformation [1]. A deformation monitoring point in Lianzi cliff dangerous rock along the Yangtze River of Hubei province in china is chosen as the research object to explore the applicability of various grey forecasting methods.

The grey system theory was first proposed by Professor Deng [6]. The theory is mainly deals with systems that are characterized by poor information or for which information is lacking. Generally, the grey model is written as GM (m, n), where m is the order and n is the number of variables of the modelling. The first-order one-variable grey model GM (1,1) is most widely employed in various fields and achieved satisfactory results [6, 7].

In order to make the grey forecasting model more precise, a large number of researchers concentrate upon improvements of GM (1,1) model mainly in four aspects [8-13]:

- 1) improvement of data modelling approach to reduce the variability of data to effectively improve prediction accuracy;
- 2) application of optimal approach to estimate development coefficient and grey input to improve prediction equation;
- 3) optimization the initial condition of GM (1,1) model;
- 4) combination of the GM (1,1) model and the advantages of other forecasting models to increase the accuracy of forecasts.

These types of improvements for GM (1, 1) model can increase prediction precision in some practical applications. However, there still exists some space to improve it. A novel high precision model is proposed to improve its precision by the following three aspects.

* *Corresponding author* e-mail: gaoninghaoyun@163.com

Firstly, the error term resulted from the calculation of background value in the conventional grey forecasting model is discussed, and then an integration term is used to substitute the original calculation of background value to eliminate the error term. Second, based on latest information priority principle, we adopt the *n*-th tem of $\underline{x}(1)$ as the initial condition of the grey differential model to increase prediction precision. Third, genetic algorithm (GA) used for search development coefficient and grey input.

The remaining paper is organized as follows. A brief introduction to the traditional GM (1,1) model is presented and discussed. Based on the concept in Section 2, a novel grey forecasting model named BIGGM (1,1) is proposed in Section 3. Section 4 will illustrate the application of BIGGM (1,1) by a numerical example. Finally, Section 5 concludes this paper.

2 Introduction to the traditional GM (1, 1) model

Grey system theory was first proposed by professor Deng in 1982. From then on the grey model (GM) has been concentrated on by a large number of scholars and adopted in various fields. GM mainly focuses on such systems as partial information known and partial information unknown. GM (1, 1) model is constructed as follows [6]:

Step 1: Establish the one-time AGO series.

Supposing surveying a certain observation point at *n* different times, thus forms a time observation series $x^{(0)} = \{x^{(0)}(i), i = 1, 2, 3 \dots n\} (n > 4)$. $x^{(0)}$ is a non-negative sequence of raw data, where $x^{(0)}(i)$ is the datum at *i*-th time and *n* is the total number of modelling data. $x^{(1)}(k)$ is the generated datum of $x^{(0)}(i)$ and can be obtained by:

$$x^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i), (k = 1, 2, \dots, n) . \tag{1}$$

Then, it alters the raw data into data with, which an exponential curve can be produced and it's effective for decreasing noise influence.

Step 2: Define the grey parameters.

$x^{(1)}(k)$ can be modelled by a first order differential equation given as:

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = u , \tag{2}$$

where, parameter *a* is called the developing coefficient and *u* is called the grey input.

Step 3: Estimate the values of *a* and *u*.

In fact, parameters *a* and *u* cannot be calculated directly from Equation (2). For practice, the algorithm of GM (1,1) is used first to obtain their approximate values through its difference equation:

$$x^{(0)}(k) + az^{(1)}(k) = u , \tag{3}$$

where $z^{(1)}(k)$ is the background value of the *k*-th datum and defined as:

$$z^{(1)}(k) = \frac{1}{2}(x^{(1)}(k) + x^{(1)}(k-1)) . \tag{4}$$

Then, values of grey parameters *a* and *u* can be estimated by using least square (LS) method as:

$$(a, u)^T = (A^T A)^{-1} A^T B , \tag{5}$$

where $A = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(n) & 1 \end{bmatrix}$, $B = (x_{(2)}^{(0)}, x_{(3)}^{(0)}, \dots, x_{(n)}^{(0)})^T$.

Step 4: Define the predicting model.

Substituting *a* and *u* in Equation (3) with Equation (5), the approximate equation becomes the following:

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{u}{a} \right) e^{-ak} + \frac{u}{a} , \tag{6}$$

where $\hat{x}^{(1)}(k+1)$ is the forecasted value at the time (*k*+1). After the completion of an inverse accumulate generating operation on Equation (6) the predicted value of the primitive data of $\hat{x}^{(0)}(k+1)$ can be obtained as:

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) . \tag{7}$$

3 The proposed GM (1, 1) model

3.1 OPTIMIZATION OF THE BACKGROUND VALUE

According to the description of the traditional GM (1,1) model in Section 2 we can know that the prediction accuracy depends on the development coefficient *a* and grey input *u*, and the solution *a* and *u* depends on the structure of background value $z^{(1)}(k)$. Thus, background value $z^{(1)}(k)$ has direct influences on the precision of GM(1,1), as shown in Equation (4) it is a smoothing formula, Equation (4) is chosen to describe the background value for a very short period of time Δt , however, being a short period of time, Δt is only a relative conception. In this period of time, for a deformable body, it may include mutations or the deformation does not occur at an average rate, under this situation, GM (1,1) model often performs very poor and makes delay errors, so the error term resulted from the original calculation of background value needs to be eliminated [8-10].

Take the integral of both sides of the grey differential Equation (2) in the interval [*k*-1, *k*], we can get:

$$\int_{k-1}^k \frac{dx^{(1)}(t)}{dt} dt + a \int_{k-1}^k x^{(1)}(t) dt = u . \tag{8}$$

By comparing Equation (3) with Equation (8), it can be found that usage of $\int_{k-1}^k x^{(1)}(t)dt$ as the background value is more adaptive than Equation (4). The error term will exist when the value of Equation (4) is not equal to the $\int_{k-1}^k x^{(1)}(t)dt$. So how to obtain the accurate value of $\int_{k-1}^k x^{(1)}(t)dt$ is the key factor to enhance the performance of GM(1,1).

In order to eliminate the error term resulted from the original calculation of background value, then the $\int_{k-1}^k x^{(1)}(t)dt$ is directly taken as the background value:

$$z^{(1)}(k) = \int_{k-1}^k x^{(1)}(t)dt \tag{9}$$

Owing to Equation (2) gives the result of the exponential function, so in this paper, set:

$$x^{(1)}(t) = Ae^{Bt} + C, \tag{10}$$

where A, B and C are constants. It is substituted in Equation (9), and use $k, k-1, k-2$ to substitute t in Equation (10), we can get:

$$x^{(1)}(k) - x^{(1)}(k-1) = A \cdot e^{Bk} - A \cdot e^{B(k-1)}, \tag{11}$$

$$\begin{aligned} x^{(1)}(k-1) - x^{(1)}(k-2) \\ = A \cdot e^{B(k-1)} - A \cdot e^{B(k-2)}. \end{aligned} \tag{12}$$

Calculations:

$$\begin{aligned} \frac{x^{(1)}(k) - x^{(1)}(k-1)}{x^{(1)}(k-1) - x^{(1)}(k-2)} &= \frac{Ae^{B(k-1)}(e^B - 1)}{Ae^{B(k-2)}(e^B - 1)}, \\ &= e^B = \frac{x^{(0)}(k)}{x^{(0)}(k-1)} \end{aligned}$$

$$B = \ln x^{(0)}(k) - \ln x^{(0)}(k-1).$$

Substitution of B into $x^{(1)}(k), x^{(1)}(k-1)$, we can obtain A and C :

$$A = \frac{x^{(0)}(k)}{\left(\frac{x^{(0)}(k)}{x^{(0)}(k-1)}\right)^{k-1} \cdot \left(\frac{x^{(0)}(k)}{x^{(0)}(k-1)} - 1\right)},$$

$$C = x^{(1)}(k) - \frac{x^{(0)}(k)^2}{x^{(0)}(k) - x^{(0)}(k-1)}.$$

Thus, we can obtain the optimized background value as:

$$\begin{aligned} z^{(1)}(k) &= \frac{x^{(0)}(k)}{\ln x^{(0)}(k) - \ln x^{(0)}(k-1)} + \\ &x^{(1)}(k) - \frac{x^{(0)}(k)^2}{x^{(0)}(k) - x^{(0)}(k-1)}. \end{aligned} \tag{13}$$

3.2 OPTIMIZATION OF THE INITIAL CONDITION

In the real world, the developments of a deformation system are always instable. So, under such situation, the author suggests paying more attention on the latest datum rather than on other previous data because the latest datum can offer more useful information about development tendency of deformation.

However, in the above-mentioned algorithm of GM(1,1), the initial condition is set as $x^{(1)}(1)$. In fact, from the description of the original GM(1,1) model we can find that the general solution of the Equation (2) can be expressed as following [11-13]:

$$x^{(1)}(t) = ce^{-at} + \frac{u}{a}, \tag{14}$$

where c is a constant.

For the general solution of we can let $t = 1$ and $t = n$, respectively, then we can get the following equations:

$$x^{(1)}(1) = ce^{-a} + \frac{u}{a}, \tag{15}$$

$$x^{(1)}(n) = ce^{-an} + \frac{u}{a}. \tag{16}$$

Then we can obtain:

$$c = \left(x^{(1)}(1) - \frac{u}{a}\right)e^a \text{ and } c = \left(x^{(1)}(n) - \frac{u}{a}\right)e^{an}.$$

When $t = 1$, which is the originally initial condition proposed by professor Deng. This way will pay more importance on the farthest datum of modelling data, so the use of the latest datum may not be enough. Therefore, GM(1,1) model cannot have enough ability to catch the instable variance. Hence, this paper, suggests setting the initial condition as $x^{(1)}(n)$ to fully catch the latest tendency. Thus, the specified solution would be:

$$\hat{x}^{(1)}(k+1) = \left(x^{(1)}(n) - \frac{u}{a}\right)e^{-a(k-n+1)} + \frac{u}{a}. \tag{17}$$

3.3 GA BASED OPTIMIZED GM (1, 1) MODEL

From above discussion, in this paper we take advantage of the ability of the genetic algorithm (GA) to solve optimization problems. This article builds a GA based GM (1,1) model, termed as BIGGM (1,1) model, by combining GA and optimization integrated background value with initial condition to estimate the parameters a and u , then, enhance prediction ability.

Genetic algorithm is a global random search technique invented by Holland in 1975, which is on the basis of Darwin's theory of nature evolutions and the theory of genetic mutations. It is especially useful for complex optimization problems where the number of parameters is large and the analytical solutions are difficult to obtain. Therefore, GA provides a common framework that is suitable for solving the optimization problems in complex systems, and it does not depend on the problem fields in itself [14-18].

To make the GM (1,1) more adaptive and precise, we use the GA to find the optimal parameters a and u of the BIGGM (1,1) model. The steps of BIGGM (1,1) can be shown as follows.

Steps 1: Same as the traditional GM (1,1) model.

Steps 2: Calculate background values through Equation (13) in Section 3.1.

Steps 3: Search the optimal parameters a and u of BIGGM (1,1) by using GA. The procedure is as follows.

1) Define the fitness function. In GA, each chromosome is decoded as network parameter. Input the training samples and calculate the fitness of each individual. From the previous studies, the MAPE (mean absolute percentage error, MAPE) is used as the fitness function.

2) Roulette selection. The roulette selection method is applied for the selection operation in BIGGM (1, 1). Selection is a process which selects the individuals with higher fitness to the next population.

3) Adaptive crossover. Crossover is carried out by recombining genetic material in two father individuals to produce two child chromosomes that share the characteristics of their parents. Crossover operation can produce fresh individuals, so some new points in the searching space can be checked. The frequency is determined by the probability in crossover operation, the higher is the frequency, and the faster will be the convergence speed. But a high frequency will cause genetic algorithm degenerate to be a stochastic search, and probably result in premature convergence that a local optimum solution is reached. The common ranges of crossover rate P_c are 0.01-0.1. In this paper set crossover rate $P_c = 0.5$.

4) Mutation. Mutation is implemented by the conversion of a child individual with a minor probability, the probability of conversion is inversely proportional to the number of variable, and has nothing to do with the size of population. The mutation operator provides a way of recovery of the loss of genetic diversity. The common

ranges of mutation rate P_m are 0.01-0.1. In this paper set mutation rate $P_m = 0.02$.

5) Terminal condition. The terminal condition of the optimal procedure is checked at the end of each generation and the process is terminated when the condition is satisfied. Two termination criteria can be chosen: maximum generation criterion and setting boundary criterion. We select the first way as termination criteria, namely, set the maximum generation $T = 1000$.

By using the GA, the optimal parameters a and u are obtained based on the criterion of minimum MAPE. The development coefficient and grey input becomes \tilde{a} and \tilde{u} .

Steps 4: Setting the initial condition of the grey differential model as $x^{(1)}(n)$.

Steps 5: The forecasting equation of the BIGGM (1,1) model is stated as:

$$\tilde{x}^{(1)}(k+1) = (x^{(1)}(n) - \frac{\tilde{u}}{\tilde{a}})e^{-a(k-n+1)} + \frac{\tilde{u}}{\tilde{a}}. \quad (18)$$

Steps 6: Taking the IAGO on series $\tilde{x}^{(1)}(k+1)$ we get:

$$\tilde{x}^{(0)}(k) = \tilde{x}^{(1)}(k+1) - \tilde{x}^{(1)}(k), k = 1, 2, \dots, n. \quad (19)$$

4 Illustrative examples

In this section, in order to demonstrate the effectiveness of the proposed model, we use the deformation of Lianzi cliff dangerous rock along the Yangtze River of Hubei province in china as an illustrating example. Lianzi cliff is one of the most dangerous rocks along Yangtze River, situated in Zigui county of Hubei province in china, and geodetic surveying method used as the most effective ways to monitoring its deformation. In this study, we use the historical deformation of Lianzi cliff dangerous rock from 1978 to 1993 as our research data. There are 10 observations, where 1978 -1987 is used for model fitting and 1988-1993 are reserved for testing. For the purpose of comparison, four forecasting models are used as follows:

- 1) GM (1,1) forecasting model, termed as GM (1,1),
- 2) Background value improved GM (1,1) model, termed as BGM (1,1),
- 3) Initial condition improved GM (1,1) model, termed as IGM (1,1),
- 4) BIGGM (1,1) model.

In order to examine the precision of grey model, error analysis is necessary to understand the difference between modelled value and actual value. Generally, two criteria are used to validate the GM (1,1) model. The first one is the absolute mean error criterion:

$$AME = \frac{1}{n} \sum_{k=1}^n |x^{(0)}(k) - \hat{x}^{(0)}(k)| \quad (20)$$

and the second one is the mean absolute percentage error criterion:

$$MAPE = \frac{1}{n} \sum_{k=1}^n \frac{|x^{(0)}(k) - \hat{x}^{(0)}(k)|}{x^{(0)}(k)} \quad (21)$$

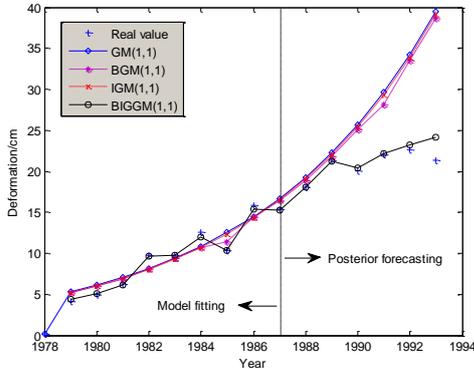


FIGURE 1 Real values and modelling values together with forecasts of deformation by four models

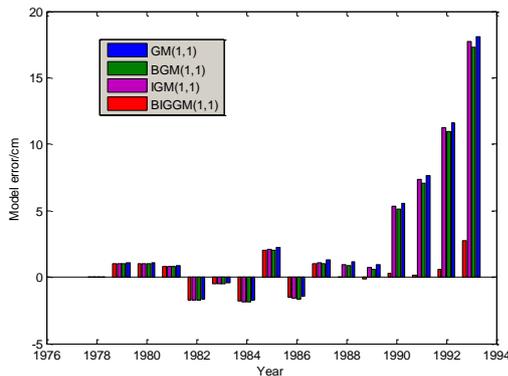


FIGURE 2 Model error distribution from 1978 to 1993

The predicted results obtained by the four models are shown in Figure 1. The model error distribution is also shown in Figure 2. For the purpose of comparison, Table 1 demonstrates that the performance of four models.

References

[1] Zhang Z L 2007 Deformation monitoring analysis and prediction for engineering construction *Surveying and mapping Press: Beijing* 55-69
 [2] Gao N, Cui X M, Gao C Y 2011 Novel high-precision grey forecasting model and its application of deformation of tunnel surrounding rock *Applied Mechanics and Materials* **90-93** 2869-74
 [3] Gao N, Cui X M, Gao C Y 2012 An effective hybrid approach for processing deformation monitoring data *Advanced Materials Research* **446-449** 3247-51
 [4] Gao C Y, Pan C J, Gao N 2012 An effective combined approach based on GM and AR for deformation analysis of high-rise buildings *Advanced Materials Research* **368-373** 2123-7
 [5] Gao N, Cui X M 2012 Grey forecasting model refining in deformation prediction based on semi-parametric regression *Applied Mechanics and Materials* **204-208** 2731-5
 [6] Deng J L 1982 Control problems of grey systems *Systems Control and Letters* **1(5)** 288-94

According to the results shown above, our proposed BIGGM (1,1) model seems to obtain the lowest forecasting errors among these models. So we can say that the BIGGM (1,1) model is an appropriate model for deformation forecasting.

TABLE 1 Performance evaluations of four models

| | GM (1,1) | BGM (1,1) | IGM (1,1) | BIGGM (1,1) |
|------------------|----------|-----------|-----------|-------------|
| AME (1978-1987) | 1.313 | 1.129 | 1.306 | 0.217 |
| MAPE (1978-1987) | 15.081 | 13.068 | 14.811 | 2.237 |
| AME (1988-1993) | 7.504 | 6.741 | 7.213 | 0.613 |
| MAPE (1988-1993) | 34.872 | 31.285 | 33.494 | 2.851 |

5 Conclusions

From the results in this study, the following three conclusions can be drawn,

- 1) The first main contribution of the paper is to modify the calculation algorithm of background value of GM (1,1). A new integration equation is used to substitute the original calculation of background value to eliminate the error term, such that the precision of prediction can be apparently increased.
- 2) The second main contribution is based on latest information priority principle, we adopt the *n*-th term of *x*(1) as the initial condition of the grey differential model to increase prediction precision.
- 3) The third main contribution is proposed a new model named BIGGM (1,1). The BIGGM (1,1) model has the advantage of grey forecasting model and genetic algorithm; it performs well concerning tendency data. When there are large variations in the data set that will decrease forecasting accuracy. However, no matter what kind of data, the forecasting performance of BIGGM (1,1) model is better than the original GM (1,1) in model forecasting. Therefore, it is a suitable method for forecasting problems with small data sets.

[7] Deng J L 1993 Grey Forecasting and Grey Decision-Making *Huazhong University of Science and Technology Press: Wuhan*
 [8] Tan G J 2000 The structure method and application of background value in grey system GM(1,1) Model (I) *Journal of systems engineering & practice* **20(4)** 98-103
 [9] Lin Y H, Lee P C, Chang T P 2009 Adaptive and high-precision grey forecasting model *Expert Systems with Applications* **36(6)** 9658-62
 [10] Deng Y H 2011 Study improved background value functions of non-equidistance GM (1,1) model in surveying and mapping engineering *Applied Mechanics and Materials* **90-93** 2887-90
 [11] Dang Y G, Liu S F, Chen K 2004 The GM models that *x*(*n*) be taken as initial value *Kybernetes* **33(2)** 247-54
 [12] Wang Y H, Dang Y G, Li Y Q 2010 An approach to increase prediction precision of GM (1,1) model based on optimization of the initial condition *Expert Systems with Applications* **37(8)** 5640-4
 [13] Yao T X, Liu S F, Xie N M 2009 On the properties of small sample of GM (1,1) model *Applied Mathematic Modelling* **33(4)** 1894-903

- [14]Holland J H 1975 Adaptation in natural and artificial system
University of Michigan Press: Ann Arbor 25-33
- [15]Ou S L 2012 Forecasting agricultural output with an improved grey forecasting model based on the genetic algorithm *Computers Electronics in Agriculture* 85(6) 33-9
- [16]Hsu L C 2010 A genetic algorithm based nonlinear grey Bernoulli model for output forecasting in integrated circuit industry *Expert Systems with Applications* 37(6) 4318-23
- [17]Lee Y-S, Tong L-I 2011 Forecasting energy consumption using a grey model improved by incorporating genetic programming *Energy Conversion Management* 52(1) 147-52
- [18]Hsu L C 2011 Using improved grey forecasting models to forecast the output of opto-electronics industry *Expert Systems with Applications* 38(11) 13879-85

Authors



Ning Gao, born in February, 1982, Wangdu County, Hebei Province, China

Current position, grades: lecturer at School of Geomatics and City Spatial Information, Henan University of Urban Construction, China.
University studies: Master of Engineering in Geodesy and Surveying Engineering at East China Institute Of Technology in China, Doctor of Engineering at China University of Mining and Technology (Beijing) in China.
Scientific interest: Formation monitoring and Surveying and mapping data processing.
Publications: more than 20 papers published in various journals.
Experience: Teaching experience of 8 years, scientific research projects.



Cai-Yun Gao, born in October, 1980, Jinzhou County, Hebei Province, China

Current position, grades: lecturer at School of Geomatics and City Spatial Information, Henan University of Urban Construction, China.
University studies: Bachelor of Engineering in Surveying and mapping engineering from Institutes Of Technology Of Hebei in China, Master of Engineering in Geodesy and Surveying Engineering at East China Institute Of Technology in China.
Scientific interest: Intelligence algorithm in the application of Geodesy and Surveying Engineering and disaster prediction.
Publications: more than 15 papers published in various journals.
Experience: Teaching experience of 8 years, 6 scientific research projects.

Effect judgment and effectiveness estimation of anti-dumping duty – an example of the case of canned mushroom

Hua Zhang, Shunchang Liu*

College of Economics and Management, China Jiliang University, Hangzhou, China

Received 6 June 2014, www.tsi.lv

Abstract

The paper aimed to provide a method for accurate estimation of the Anti-Dumping (AD) tax. Having used the case of canned mushrooms, exported from Indonesia to the United States as an example, the paper presented methods for effective judgment and accurate estimation of the AD. By having done so, it provided AD policy makers with a scientific and fair AD tax which, at the same time, would incorporate legislation that will prevent abuse of AD taxation system. The paper further analysed the impact of AD tax on the trend of export; the spread of related indicators between a taxable situation and a non-taxable one through the Chow test and other methods. Final results provided AD users with logical basis and method support.

Keywords: SOFC, discrete sliding mode, control, DC/AC, converter

1 Introduction

Anti-Dumping (AD) is the main reason that leads to trade conflict and even trade war in the international business environment. From 1995 when the World Trade Organization (WTO) was established to the end of 2012, 4230 AD complaints were filed in the world. AD measures including duties were taken against an astonishing 2719 of these cases. The AD measures vary from case to case, some of which are original, triggered, for relief, protective, or even retaliatory. But no matter what the purpose is, each country believes they have just been caused for its punitive actions while the defending countries believe they did nothing wrong. Thus, AD measures have become very controversial or even the most controversial trade measures. However, it is still a licensed trade remedy means in terms of WTO law and rule. In our opinion, there are several reasons that AD measures have been criticized. First, anti-dumping is defined despite its obscured evidences. Second, one country took AD measures into their own hands while another used it as revenge. Lastly, one country took AD measures so aggressively the targeted country could not reasonably accept the conditions. The first and second reasons we just mentioned exemplify how these countries violate WTO regulations. Therefore, they shouldn't fall under WTO jurisdiction. This paper focuses on the analysis of AD caused by the third reason. Even though this type of AD is legitimate in reality, these actions are ineffective because they crossover certain trade relief limits. So, evaluating AD duty's effect based on confirming AD duty-caused export variation of Countries accused of dumping will provide theoretical guidance for reasonable estimating of the appropriate rate.

2 Research literature review

As one of the important trade measures in the latter half of the 20th century, AD has become a hot topic discussed in both academia and industry from domestic and overseas. Among numerous of studies related to AD, we can sum them up as follows:

Firstly, the study of the AD trade restrictions effect. The AD trade restriction effect is the direct effect of AD measures on the exports of the countries which are accused of dumping. The research has shown the decline of export volume in the accused country. The direct and indirect evidences can be found in the research of Vandebussche and Zanardi [1] who used the gravity model to analyse the gross trade volume data between 1980 and 2000 and Bao's [2] research on the impact of China's AD policy on GDP.

Secondly, the study of AD trade diversion and trade deflection effects. Although there has been an increasing number of research on this, the conclusions are not all the same. For example, Durling and Prusa [3] paper on hot rolled steel anti-dumping case (1996-2001) and Feng Zongxian & Xiang Hongjin (2010) 's research on European and American AD measures have shown opposite conclusion.

Thirdly, the study of AD relief effect. Similar to the above mentioned studies, such studies also have shown quite different conclusions. For example, the study of Bao [5] in the 1997 - 2004 Chinese anti-dumping case analysis and Corinne M. Krupp and Susan Skeath (2002) used seven TSUSA (United States Tariff) and eight SIC (Standard Industrial Classification) tariff empirical data have shown different conclusions. Despite of the different study objectives, we cannot arbitrarily consider the two conclusions are totally opposite to each other. The reason

* *Corresponding author* e-mail: 522535422@qq.com

is that single variable like market share or import volume cannot fully measure the relief effect.

Lastly, the study of the impact of AD research methods. Eaton and Grosman [7] first used the metrology method to analyse the effect on negative impact on certain industries. Since then, scholars such as Prusa [8], Bao [4], Xiang, Ke and Feng [9] have done such researches from different perspectives with different techniques to analyse the impact of AD on trade, which have provided evidences and references for other researchers.

Obviously, the existing studies of the AD trade effect have already provided good references. However, the study can be furthered because of the imperfect setting of AD duty. A statistic method should be applied in this case. Therefore, this paper utilizes Chow test in the AD effect study. Besides, VAR model is applied to obtain the time features from different AD duties to get the general idea of how to set the AD duty.

3 Research ideas and steps

The effect of AD duty is the key to judge if AD duty rulings trade remedy function works and if it functions properly, the former is a matter of judgment, the latter is measurement of the problem. Therefore, this paper will follow the logic of "effect judgment – efforts to measure".

3.1 CHOW TEST ON THE EFFECT OF AD DUTY

Although the analysis on simple data changing across time can provide evidence that if a country's export-related industries are affected by the import country's anti-dumping duty, it still does not seem sufficient and convincing. The reason is that even if there is a decline in export volume or product price in the accused country, there is no way to prove such decrease is directly related to AD. In other words, if the decline happened before being imposed AD duty, the AD duty just maintained the downward trend, and thus we cannot prove the AD duty has significant impact on export at time series level.

Under such circumstances, unless we can see a spike in time series trend, we cannot make the argument that AD duty is the cutoff point on the change of export trend for accused Dumping country. Therefore, Chow test should be applied in this case. Chow Test is used to determine the variation of independent variables when given a series of time period. It breaks the whole time period into two parts, before and after, and F-test is used to test both parts to determine if the cutoff point is significant on the change of trend.

For the anti-dumping duty, if we take the point of time when AD duty is imposed as cut-off point, some parameters such as import volume associated with the products involved in the import party will be used. Chow test is used to analyse such time series data and we observe the changes to determine its statistical significance on AD duty trade effects.

Firstly, we choose parameters such as price, export volume etc. of the country whose products are imposed AD duty;

Secondly, we apply logarithmic transformation on such data and use OLS regression. Then, we conduct heteroscedasticity and correlation test on the regression equation.

Lastly, in the adjusted regression model, we apply Chow test on multiple time points, and observe the P value and F value at 1% significance level. If the P value is less than 0.01, then it is considered trend mutation effect occurs, meaning the AD duty has changed the original development trend of the product concerned.

3.2 THE VAR MODEL IN AD DUTY EFFECT MEASUREMENT

Chow test is used to measure if AD duty will change the structure and trend of original model, and VAR model is applied to further analyse the impact of AD duty. VAR model uses lag value transformation to change the univariate autoregressive model to multivariate time series self-regression model. With Chow test results, the VAR model can be used to predict the normal trend of export volume and price for a country if they were not being imposed AD duty. Then we compare the predicted value to actual value to measure the effectiveness of AD duty. Below are the test steps in details:

Firstly, take export price and export volume as endogenous variables and constant term as exogenous variable, we use AIC and SC criteria to determine the lag periods P;

Secondly, build autoregression model with lag periods p, and obtain the estimation results;

Lastly, we take Chow trend mutation point of time as cut off point, then use VAR model to predict the original trend with transformed price and export volume. With the comparison of predicted value and actual value, we can obtain the basic idea of the effectiveness of AD duty.

4 The determination of AD duty effect and Chow Test

4.1 DATA SELECTION AND PREPARATION

We take the 1998 AD cases (average tax rate of 14.1%) US against Indonesia canned mushrooms (HS Code: 2003100053) as an example to investigate the mutation effect on the product involved and estimate the impact of Indonesia canned mushroom export decline caused by AD duty. We apply Chow test on total 36 periods of time series data including export volume and export price of the canned mushrooms before/after AD duty (See Tables 1 and 2). The data source is from the U.S. Department of Commerce Bureau of Foreign Trade Statistics Database, and the data analysis tool used here is Eviews7.0.

The original data is listed in Tables 1 and 2.

TABLE 1 Mushroom export quantity from Indonesia to the U.S. (Tons)

| | | | | | | | | | | | |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| t ₁ | t ₂ | t ₃ | t ₄ | t ₅ | t ₆ | t ₇ | t ₈ | t ₉ | t ₁₀ | t ₁₁ | t ₁₂ |
| 64.8 | 49.1 | 106.2 | 167.5 | 225.6 | 605.9 | 424.8 | 541.9 | 418.9 | 49.2 | 253.7 | 240.0 |
| t ₁₃ | t ₁₄ | t ₁₅ | t ₁₆ | t ₁₇ | t ₁₈ | t ₁₉ | t ₂₀ | t ₂₁ | t ₂₂ | t ₂₃ | t ₂₄ |
| 215.1 | 116.7 | 195.8 | 179.5 | 225.7 | 44.6 | 79.9 | 16.9 | 11.7 | 36.2 | 25.6 | 65.7 |
| t ₂₅ | t ₂₆ | t ₂₇ | t ₂₈ | t ₂₉ | t ₃₀ | t ₃₁ | t ₃₂ | t ₃₃ | t ₃₄ | t ₃₅ | t ₃₆ |
| 56.2 | 94.5 | 76.2 | 179.6 | 154.9 | 129.8 | 166.4 | 65.8 | 155.7 | 146.9 | 135.8 | 77.2 |

TABLE 2 Mushroom export price from Indonesia to the U.S. (Thousands of USD)

| | | | | | | | | | | | |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| t ₁ | t ₂ | t ₃ | t ₄ | t ₅ | t ₆ | t ₇ | t ₈ | t ₉ | t ₁₀ | t ₁₁ | t ₁₂ |
| 1417.4 | 1887.7 | 1810.5 | 1744.7 | 1,956.8 | 1,956.3 | 2,084.3 | 1,970.6 | 1,858.9 | 2,359.1 | 1,406.9 | 1,547.7 |
| t ₁₃ | t ₁₄ | t ₁₅ | t ₁₆ | t ₁₇ | t ₁₈ | t ₁₉ | t ₂₀ | t ₂₁ | t ₂₂ | t ₂₃ | t ₂₄ |
| 1,422.4 | 1,651.5 | 1,254.0 | 1,267.9 | 1,204.6 | 1,359.9 | 1,277.0 | 1,280.2 | 1,210.4 | 1,593.5 | 1,643.6 | 1,542.6 |
| t ₂₅ | t ₂₆ | t ₂₇ | t ₂₈ | t ₂₉ | t ₃₀ | t ₃₁ | t ₃₂ | t ₃₃ | t ₃₄ | t ₃₅ | t ₃₆ |
| 1,304.9 | 1,607.5 | 1,476.3 | 1,538.6 | 1,897.8 | 1,741.6 | 2,134.4 | 2,063.5 | 1,954.9 | 1,997.4 | 1,772.3 | 1,836.5 |

Now we build a model in which the dependent variable is export volume Y , and independent variable is export price X , the regression model is $y = \alpha + \beta x_i + \varepsilon_i$, after Least-squares regression based on the log transformation, the parameter estimation results is below:

$$\ln \hat{y}_i = -7.0530 + 1.5909 \ln \hat{x}_i \quad (1)$$

(-1.1915) (1.9902)

$$\hat{R}^2 = 0.1043 \quad DW = 1.0755 .$$

From the Equation (1) we can find that \hat{R}^2 is small, and DW is around 1, this indicates that the fitting effect is

not so good meaning there are some possible problems of heteroscedasticity and autocorrelation. Now we turn to deal with heteroscedasticity and autocorrelation.

4.1.1 Heteroscedasticity test

The validity of parameter estimation subject to the heteroscedasticity.

Here we use method of Breusch-Pagan-Godfrey through Eviews 7.0 to test heteroscedasticity. The result is listed in Table 3. The results in Table 3 show that we cannot reject the original hypothesis, namely, there is no heteroscedasticity.

TABLE 3 Results from Breusch-Pagan-Godfrey test

| | | | |
|---------------------|----------|---------------------|--------|
| F-statistic | 0.716611 | Prob. F(1,34) | 0.4032 |
| Obs*R-squared | 0.743102 | Prob. Chi-Square(1) | 0.3887 |
| Scaled explained SS | 0.415589 | Prob. Chi-Square(1) | 0.5191 |

4.1.2 Autocorrelation test and correction

Here we adopt Correlogram-Q-statistics in Eviews 7.0 to test autocorrelation. The result is shown Figure 1.

It is not difficult to find that only the partial correlation coefficient's histogram of the first period of the model is beyond the dashed part, showing the first-order auto-correlation phenomenon.

So, introduce $AR(1)$ in the model, the corresponding parameter estimation results is below:

$$\ln \hat{y}_i = 15.4230 - 1.4362 \ln \hat{x}_i + 0.7412AR(1) \quad (2)$$

(2.4150) (-1.6760) (6.2823)

$$\hat{R}^2 = 0.1043 \quad F = 12.5577 \quad DW = 1.0755 .$$

Though the absolute number of the impacts of export price and initial export quantity on export is not the same, its basic trend is analogous to the basic results by Zhang Hua (2010) through building the following model with panel datas.

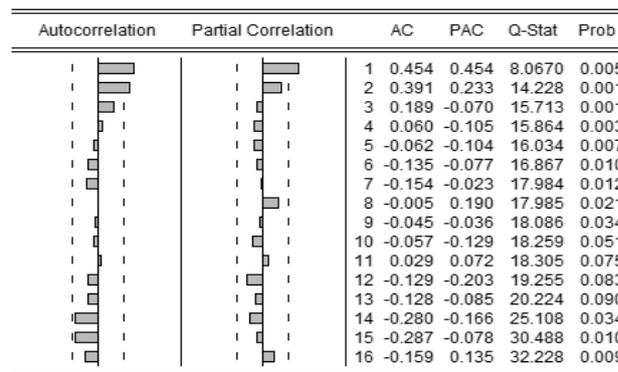


FIGURE 1 Autocorrelation and partial correlation

$$y_{it} = \alpha_{0i} + \alpha_{1i}t_{it} + \alpha_{2i}x_{it} + \alpha_{3i}y_{i(t-1)} + \varepsilon_{it} . \quad (3)$$

Its eventual regression results is as following:

$$\hat{y} = 1.050372 - 0.453271\hat{t} - 0.412267\hat{x} + 0.722975\hat{y}_{t-1} . \quad (4)$$

From the adjusted OLS model above we can see that F value is significant. DW is around 2, which eliminates the problem of the first order autocorrelation. One thing to notice is $\hat{R}^2 = 0.4393$, which is not very significant. The Reasonable explanation for this is that due to the short-term volatility of the export volume and export price for the products involved, the goodness of fit is not very satisfactory in the entire 36 periods of time, which reflects

the fact that the occurrence of unexpected events potentially break the sequence of the original law. Therefore, to adjust the OLS model in a correct way lays the foundation for Chow test.

4.2 CHOW TEST

Based on the adjusted OLS model, we choose 10-27 period of time to conduct Chow test, the result is below:

TABLE 4 Chow mutation point test

| Time point | F test value | P test value | Time point | F test value | P test value |
|------------|--------------|--------------|------------|--------------|--------------|
| 10 | 0.7704 | 0.5199 | 19 | 3.7679 | 0.0213 |
| 11 | 2.3940 | 0.0887 | 20 | 4.5222 | 0.0102 |
| 12 | 3.1167 | 0.0413 | 21 | 3.0433 | 0.0446 |
| 13 | 3.2556 | 0.0358 | 22 | 3.7308 | 0.0221 |
| 14 | 3.0108 | 0.0461 | 23 | 1.0204 | 0.3980 |
| 15 | 2.8977 | 0.0519 | 24 | 1.7576 | 0.1773 |
| 16 | 3.4633 | 0.0289 | 25 | 1.5592 | 0.2205 |
| 17 | 3.7328 | 0.0220 | 26 | 1.3930 | 0.2648 |
| 18 | 6.0587 | 0.0025 | 27 | 1.0425 | 0.3885 |

We can see from Table 4 that the p value is less than 0.01 at the 1% level of significance at period 18. Therefore we reject the null hypothesis which is interpreted as no structure change in the regression equation, and period 18 is the time when US imposed AD duty against Indonesia canned mushrooms. Therefore, we believe the AD duty against Indonesia canned mushroom directly lead to a change in the export price and volume. In addition, after taking the differences between two of the regression sequence data before and after the cut-off point, we discovered that slopes of the regression equations changed greatly. All the evidences have confirmed that the AD duty against Indonesia canned mushroom has caused a significant impact. Therefore, we can draw the conclusion that the AD duty against Indonesia canned mushroom has statistical significance on the trade between US and Indonesia.

5 The measure of anti-dumping duties' effectiveness based on VAR model

The traditional econometric approach is based on economic theories to describe the relationships among variables. However, economic theories are usually not sufficient to provide rigorous descriptions for these dynamic relationships. Moreover, endogenous variables can appear in either the left or right side of the equation, making estimation and inference become more complex. To solve these problems, a non-structural method is applied to building the model describing relationships among variables. Therefore, the vector auto-regression model, represented by the unstructured equation, is selected in this thesis to analyse the problem.

Chow mutation-point test is to test whether the impact from external events, e.g. anti-dumping duties, would change the structure and trend of the original model. So based on Chow test, vector auto-regression (VAR) model can be used to measure AD's effect.

5.1 THE DETERMINATION OF LAG PERIODS

In an ideal model, the AIC and SC are supposed to be very small. So by comparing the values of AIC and SC in models adopting different hysteresis, the ideal lag period, which makes the model's both AIC and SC values are relatively small, can be selected to adjust the explanatory variable. In this case, the AIC and SC values of each lag period are listed in Table 5.

TABLE 5 AIC and SC values of each lag period

| | lag period one | lag period two |
|-----|----------------|----------------|
| AIC | 26.0049 | 26.1048 |
| SC | 26.6770 | 26.5583 |

Obviously, the minimum AIC value (26.0049) can be obtained when adopting lag period one, while the SC value also reaches the minimum (26.6770) simultaneously. So the ideal lag period is one.

5.2 PARAMETER ESTIMATION VIA VAR MODEL

To establish VAR model, this thesis sets export volume as the explanatory variable Y while export price as the independent variable X , regarding these two as endogenous variables while the acquiesce constant term C in Eviews 7.0 as the exogenous variable. When adopting lag-period-one endogenous variable, VAR model built has the following results:

$$Y = 0.5500Y(-1) + 0.1322X(-1) - 145.8603 \quad (5)$$

(4.1248) (2.1499) (-1.4700)

$$\hat{R}^2 = 0.4529 \quad F = 15.0708,$$

$$X = 0.6162Y(-1) + 0.5225X(-1) + 700.4231 \quad (6)$$

(2.0791) (3.8459) (3.1760)

$$\hat{R}^2 = 0.421173 \quad F = 13.36972.$$

Identifying and testing the model is regarded as an important segment to determine whether they meet previous hypothesis and satisfy economic significance. Testing the above VAR model via Granger causality test is to examine whether one certain lagged variable can be introduced to the equation. If the introduction of one lagged variable had influence on the other variable, then these two variables have Granger causality. Table 6 below shows the test results:

TABLE 6 Results from Granger causality tes

| | Excluded | Chi-sq | df | Prob. |
|-----------------------|----------|----------|----|--------|
| Dependent variable: Y | X | 4.622031 | 1 | 0.0316 |
| | All | 4.622031 | 1 | 0.0316 |
| Dependent variable: X | Y | 4.322450 | 1 | 0.0376 |
| | All | 4.322450 | 1 | 0.0376 |

Table 6 presents the test statistics and probability of each endogenous variable with respect to the other endogenous variable via Granger causality. When first viewing variable Y, from the test results we know that the Granger causality (Y to X) is significant (significance level is 0.05). Similarly, the Granger causality of X to Y is also significant.

5.3 THE PREDICTION OF THE ORIGINAL TREND AFTER CHOW MUTATION POINT VIA VAR MODEL

When introducing lag-period-one export volume and price of Indonesian canned mushrooms to the VAR model, regarding Chow trend-mutation point as time-demarcation point, and applying the VAR model to predicting the original Indonesian canned mushroom export volume and price trend (under the condition of no anti-dumping duties) after the mutation point (Period 18-36), we can obtain the trend curves shown in Figures 2 and 3 respectively.

In Figures 2 and 3, after the trend-mutation point 18, the blue line represents the actual export volume and export price trend respectively when anti-dumping duties were imposed, while the red line represents the estimated export volume and export price trend with no imposition of anti-dumping duties.

Thus the distance between these two kinds of lines either in terms of export volume or in terms of export price can indicate the extent of AD's trade-restriction effect. Further speaking, the relative position, shapes and interrelationships of these two kinds of lines after the trend-mutation point 18 represent features of AD duties' impacts from the perspective of the time, structure and degree.

Specifically in Figure 2, after Point 18, the estimated original trend of Indonesian canned mushrooms export rises slowly. At Point 18 when the U.S. started to impose anti-dumping duties on Indonesian canned mushrooms, the distance between the red and blue lines increase in the following 4 time units, indicating the AD's short-term impact on export volume was significant in this time period, which can be attributed to the price pressure

brought by anti-dumping duties on canned mushrooms export business in Indonesia.

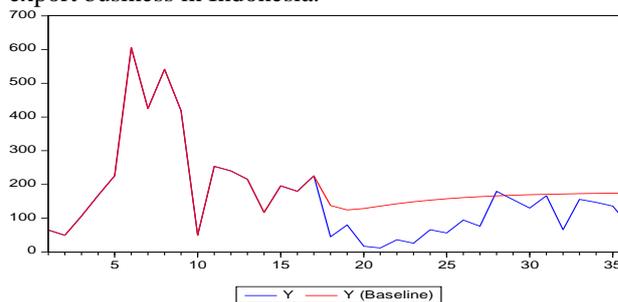


FIGURE 2 Actual export volume and the estimated original export volume (Tonsn)

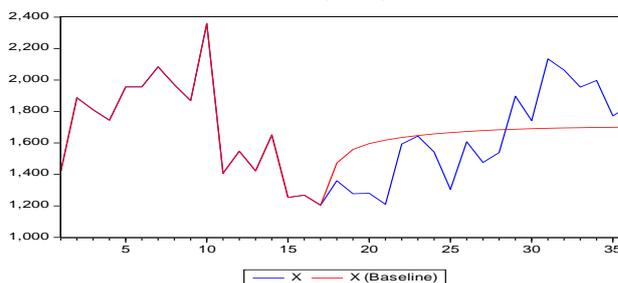


FIGURE 3 Actual export price and the estimated original export price (Thousands of USD)

As time goes on, even though the distance is gradually reduced, the blue line is always under the red line, indicating the existence of the AD's long-term effect on Indonesian canned mushrooms export volume to the U.S. While in Figure 3, Period 28 can be regarded as the cutoff time point--during the period 18-28 the spread between red and blue lines is positive but afterwards it turns negative, indicating the actual price fluctuations caused by imposing anti-dumping duties were lagging. It is noteworthy that before the AD's imposition (Period 18), both the export volume and price fell sharply the significant decline of export price did not bring the growth of export volume and the export volume and price in the period 8-18 showed a positive correlation, indicating the AD's imposition had trade-investigation effect, which started to appear 10 periods before the anti-dumping policy implementation and lasted for this period. Thus the AD's impact from the import country on the product export in this case was not only originated from anti-dumping duties, but also derived from the anti-dumping measures in broad terms, e.g. the previous case investigation. So more comprehensive and accurate assessments of AD's effects from various aspects are urgent, and further sustained attention and in-depth research from both the academic sector and real industries are highly needed.

6 Conclusions and implications

Chow mutation-point test can examine whether structural changes will appear at a certain point. Such principle and function can be applied to judge a certain economic policy's function. In this paper, we use Chow mutation-point test to analyse the data of US-Indonesia canned mushroom case. We find that the original export trend was

interrupted by the external event at the mutation point, namely, imposing anti-dumping duties. Further analysis show that the negative short-term impact the accused companies suffered was particularly significant within the 4 time periods since imposing anti-dumping duties, then the negative impact gradually weakened but never disappeared, the actual export level was still below the normal development level.

Trade remedy is approved by WTO as an option for maintaining trade justice. How to play its positive role precisely largely depends on whether the implementation of specific measures is able to accurately reach the critical point of two opposite trade effects that is to maximize the trade-remedy effect but not to go beyond it. This is an ideal solution in theory but quite difficult to achieve in practice especially for those quite controversial trade-remedy measures such as the anti-dumping and countervailing ones for various aspects should be fully considered, from whether the measures can play the right role, to the eventual effects on trade justice and order. The idea of Chow point-mutation test and its extensions can be used to not only determine the trade effects, but also measure the

significance of such effects based on the structural variation from the test results. Furthermore, in this way you can also pre-design trade-remedy effectiveness. Therefore, based on specific circumstances, it can be widely applied to pre-design and effect judgment of trade and even other economic and social policies, thus minimizing the negative influence of such policies.

Acknowledgment

This paper is both a phase result of NNSFC (National Natural Science Foundation of China) Project in 2013 (Project No.71373249), a staged achievements of ZPNSF (Zhejiang Provincial Natural Science Foundation) Project in 2012 (Project No. LY12G03034) and a phase achievements of a youth fund project of MOE (Ministry of Education in China) Humanities and Social Sciences in 2010 (Project No.10YJC790375). It is also sponsored by Zhejiang Industrial Development Policy Key Research Centre of Philosophy and Social Science of Zhejiang Province and Zhejiang Provincial Key Research Base of Management Science and Engineering.

References

- [1] Vandenbussche H, Zanardi M 2010 The chilling trade effects of antidumping proliferation *European Economic Review* **54** (6) 760-77
- [2] Bao X 2004 Analysis on effects of China's AD measures *Economic Review* (1) 9-11 (in Chinese)
- [3] Durling J P, Prusa T J 2006 The trade effects associated with an antidumping epidemic: The hot-rolled steel market, 1996-2001 *European Journal of Political Economy (special issue on antidumping)* **22**(3) 675-95
- [4] Xiang H, Lai M 2012 Industry remedy and welfare effects of China's anti-dumping and countervailing measures: A research based COMPAS model *Industrial Economics Research* **2** 1-8 (in Chinese)
- [5] Bao X 2007 Trade-remedy effects of Antidumping measures: empirical analysis based China data *Economic Research Journal* **2** 71-84 (in Chinese)
- [6] Prusa T J, Skeath S 2002 The economic and strategic motives for antidumping filings *Review of World Economics* **138**(3) 389-413
- [7] Eaton J, Grossman M 1986 Optional trade and industry policy under oligopoly *Quarterly Journal of Economics* **101**(2) 383-406
- [8] Prusa T J 1997 *The Effects of U.S. Trade Protection and Promotion Policies* University of Chicago Press: Chicago
- [9] Xiang H, Ke K, Feng Z 2009 Industry injury test in antidumping: theoretical and empirical research based on COMPAS model *China Industrial Economics* (1) 49-52 (in Chinese)

| Authors | |
|---|---|
|  | <p>Hua Zhang, born in December, 1973, Xi'an City, Shaanxi Province, China</p> <p>Current position, grades: Doctor, Associate Professor at College of Economics and Management, China Jiliang University, China. University studies: B.Sc. in Economics from Northwest A & F University of Shaanxi in China. M.Sc. from Northwest A & F University in China. Scientific interest: international trade, international finance. Publications: more than 30 papers published in various journals. Experience: teaching experience of 12 years, 6 scientific research projects</p> |
|  | <p>Shunchang Liu, born in December, 1989, Shiyan City, Hubei Province, China</p> <p>Current position, grades: Postgraduate of College of Economics and Management, China Jiliang University, China. University studies: B.Sc. in Nanjing University of Finance and Economics from Jiangsu in China. Scientific interest: international trade, financial engineering. Publications: 2 papers. Experience: 3 scientific research projects.</p> |

Identification of key subsystems for urban rail vehicles based on fuzzy comprehensive evaluation

Zongyi Xing^{1*}, Lingli Mao², Limin Jia³, Yong Qin³

¹*School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China*

²*School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*

³*State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China*

Received 6 May 2014, www.tsi.lv

Abstract

Identification of key subsystems for urban rail vehicles is important for the selection of maintenance strategy. The fuzzy comprehensive evaluation technique is applied to determine the key subsystems of urban rail vehicles. Firstly, the vehicle is divided into nine subsystems according to the module partition method. Then, the degrees of occurrence, severity, detection and maintenance cost are chosen as the evaluation factors that are quantified based on fuzzy theory and collected historical data. Finally, the calculation model of critical degree is established based on the fuzzy comprehensive evaluation method. The proposed approaches are applied to Guangzhou Metro Corporation, and five key subsystems are selected. The experiment results, which are consistent with those of most knowledgeable engineers and experts, indicate the validity of the proposed method.

Keywords: key subsystem, urban rail vehicle, fuzzy comprehensive evaluation

1 Introduction

Urban rail vehicles served as efficient tools for transporting large volumes of passengers often operate in a closed environment, which may cause great inconvenience to passengers and lead to personal injury or property loss when a traffic accident occurs. With the increase in rail lines and vehicle complexity, vehicle maintenance is more challenging and complicated, so it is essential to determine a valid and effective method to identify the key subsystems of urban rail, which ensures the reliability and operational safety of urban rail vehicles under existing resources.

The most commonly-used methods to identify key subsystems or components are as follows: importance degree evaluation, risk priority number assessment, and etc. Birnbaum [1] proposed an importance degree-based method to identify the most important components by quantifying their contribution to the performance of the whole system. The risk priority number method [2] classifies and scores the severity, occurrence, and detection of failure modes according to on-site experience, and then the key subsystems are identified by those three factors. Zhike [3] proposed that the particular set of kinetic parameter values of the model closely approximates the corresponding biological system, and globally identified the key components and steps in signal transduction networks at a systems level by applying multi-parametric sensitivity analysis. Pan [4] used the fault numbers as evaluation criteria to determine the important subsystems of urban rail vehicles. In the paper, fuzzy comprehensive

evaluation method was put forward to analyse the key subsystems of urban rail vehicles.

In next section, we summarize the evaluation factors and main identification methods of the key subsystems. Section 3 discusses a case study, in which the door subsystem was chosen as an illustration example to calculate the key degree, and then key subsystems of urban rail vehicles are constructed. Finally, conclusions are offered in Section 4.

2 Identification of key subsystems based on fuzzy comprehensive evaluation

There are two fundamental principles of fuzzy comprehensive evaluation [5] as follows: fuzzy linear transformation and maximum membership degree, which quantify the related various factors of the evaluated object and judge the allocation weights of the factors according to their impact on the object to make a reasonable comprehensive evaluation. With the introduction of fuzzy math into evaluation process, the complex uncertain problem can be solved better and evaluation results are more objective and accurate.

2.1 EVALUATION FACTORS OF IDENTIFICATION OF CRITICAL SYSTEM

The occurrence, severity, detection, and maintenance cost of failure modes are combined to confirm a key subsystem based on deterministic calculation and uncertainty evaluation using fuzzy sets.

* *Corresponding author* e-mail: xingzongyi@163.com

The failure occurrence degree λ_i of the subsystem S_i is represented as

$$\lambda_i = n_i / N, \tag{1}$$

where n_i is the fault number of subsystem S_i , and N is the fault number of the whole urban rail vehicle.

The j -th failure mode probability of the subsystem S_i is written as

$$\alpha_{ij} = n_j / n_i, \tag{2}$$

where n_j is the appearance number of the failure mode M_j and n_i is the fault number of subsystem S_i .

The failure occurrence degree λ_i and failure mode probability α_{ij} can be quantified based on the collected historical data.

The severity, detection and maintenance costs are evaluated comprehensively based on fuzzy theory [6]. The severity s_{ij} is determined according to the failure mode's influence on the normal operation of the vehicle and its damage to the vehicle's functions. The detection d_{ij} indicates the degree that the failure mode can be detected in advance. The maintenance costs c_{ij} contain the human and material costs. From relevant literatures [7,8] and on-site practical experience, the evaluation criteria and membership functions about severity, detection, and maintenance costs are detailed in Tables 1-3.

TABLE 1 Rating scales of severity

| Severity s_{ij} | Membership function |
|--|---------------------|
| Inevitably having a huge impact on normal operation of vehicles, resulting in casualties (A) | [0.8 0.9 0.95 1.0] |
| Probably having a large impact on normal operation of vehicles, causing a great damage to the functional realization of the vehicles (B) | [0.6 0.75 0.8 0.85] |
| Probably having an impact on the normal operation of vehicles, causing damage to the functional realization of the vehicles (C) | [0.35 0.4 0.5 0.65] |
| Having no impact on the normal operation of vehicles, causing damage to the functional realization of the vehicles (D) | [0.2 0.25 0.3 0.4] |
| Having little/ no impact on the normal operation of vehicles and functional realization of the vehicles (E) | [0 0.05 0.15 0.25] |

TABLE 2 Rating scales of detection

| Detection d_{ij} | Membership function |
|--------------------------|---------------------|
| Hardly detectable (A) | [0.8 0.9 0.95 1.0] |
| Hard to detect (B) | [0.6 0.75 0.8 0.85] |
| Possible to detect (C) | [0.35 0.4 0.5 0.65] |
| Easy to detect (D) | [0.2 0.25 0.3 0.4] |
| Inevitable to detect (E) | [0 0.05 0.15 0.25] |

TABLE 3 Rating scales of cost

| Cost c_{ij} | Membership function |
|---------------|---------------------|
| Very high (A) | [0.8 0.9 0.95 1.0] |
| High (B) | [0.6 0.75 0.8 0.85] |
| Moderate (C) | [0.35 0.4 0.5 0.65] |
| Low (D) | [0.2 0.25 0.3 0.4] |
| Very low (E) | [0 0.05 0.15 0.25] |

2.2 CONSTRUCTION OF DECISION MATRIX

After confirming the scale of the failure modes according to expert experience, the fuzzy comprehensive evaluation decision matrix can be established [9]. Supposing the system S_i has m kinds of failure modes, the initial decision matrix F_i can be constructed as follows:

$$F_i = \begin{bmatrix} s_1 & d_1 & c_1 \\ s_2 & d_2 & c_2 \\ \vdots & \vdots & \vdots \\ s_m & d_m & c_m \end{bmatrix}, \tag{3}$$

where s_j , d_j , and c_j are trapezoidal fuzzy numbers of rating scales for severity, detection, and maintenance costs of the failure mode M_j respectively.

The weight coefficient W is determined in consideration of research needs and on-site experience, and the weighted decision matrix V_i is rewritten using fuzzy composite operator:

$$V_i = W \circ F_i = [v_1 \quad v_2 \quad \dots \quad v_m]^T, \tag{4}$$

where the v_j is a trapezoidal fuzzy number.

2.3 CALCULATION OF COMPREHENSIVE QUANTITATIVE VALUES

The decision matrix in the form of fuzzy numbers should be transformed to crisp values using defuzzification methods. The center of gravity method [10] is one of the most commonly-used defuzzification methods and can solve ambiguity of the weighted fuzzy evaluation value. Given a fuzzy number $v_j=(a_j,b_j,c_j,d_j)$, the defuzzification value can be defined as:

$$Z_{ij} = v_j = \begin{cases} a_j, & a_j = b_j = c_j = d_j \\ \frac{d_j^2 + c_j^2 - b_j^2 - a_j^2 + d_j \cdot c_j - b_j \cdot a_j}{3(d_j + c_j - b_j - a_j)}, & \text{else} \end{cases}, \tag{5}$$

where Z_{ij} is the comprehensive quantitative value for j -th failure mode M_j of i -th system S_i considering the severity, detection, and maintenance costs.

2.4 PARETO DIAGRAM

A Pareto diagram [11] is an intuitive chart for analysis and selection of main factors for complex systems. All candidate factors are arranged from left to right as the horizontal axis, and the percentage or cumulative percentage of each factor is taken as the vertical axis value. According to Pareto's law, the factors whose vertical axis values accounts for 80 percent are identified as the main factors.

To measure the weight of different subsystems, the key degree of a subsystem is denoted as

$$K_i = \lambda_i \cdot \sum_{j=1}^m \alpha_{ij} Z_{ij}, \tag{6}$$

where K_i is the importance measure of the i -th subsystem. The bigger K_i is, the more important i -th subsystem is.

2.5 MAIN STEPS OF THE PROPOSED METHOD

The proposed approach used to identify key subsystems of urban rail vehicles can be summarized as follows:

- 1) Calculate the failure occurrence degree λ_i of i -th subsystems.
- 2) Calculate the failure mode probability α_{ij} of j -th failure mode.
- 3) Determine the scale of severity, detection, and maintenance costs.
- 4) Establish the decision matrix F_i .
- 5) Defuzzify the weighted decision matrix to get the comprehensive quantitative value.
- 6) Calculate the key degree K_i of the i -th subsystem.
- 7) Normalize the critical degree K_i and identify the key subsystems of urban rail vehicle.

3 Experiments and results

3.1 SUBSYSTEM OF URBAN RAIL VEHICLES

Given that there is no uniform standard for the partitioning of urban rail vehicles, metro corporations often classify different subsystems of rail vehicles. In our experiment, urban rail vehicles were divided into nine subsystems according to such items as function, principle, and behavioral and structural characteristics [12], including door S_1 , air brake S_2 , auxiliary system S_3 , body S_4 , running gear S_5 , passenger information S_6 , air-conditioner S_7 , traction/electronic brake S_8 , train control and diagnosis S_9 .

3.2 KEY SUBSYSTEM IDENTIFICATION OF URBAN RAIL VEHICLES

The failure modes were counted, and the failure occurrence degrees of each subsystem were calculated based on three-year historical failure data collected from Guangzhou Metro Corporation. The obtained failure occurrence degree λ_i of subsystem S_i is listed in Table 4.

Due to space limitations and the large number of subsystems failure modes, the door subsystem S_1 was selected as an example and was analyzed in detail. Table 5 shows the statistics of failure modes and its probability of the door subsystem.

TABLE 4 Failure occurrence degree of subsystems

| | | | |
|-------------|-------|-------|-------|
| S_i | S_1 | S_2 | S_3 |
| λ_i | 0.129 | 0.076 | 0.136 |
| S_i | S_4 | S_5 | S_6 |
| λ_i | 0.117 | 0.079 | 0.111 |
| S_i | S_7 | S_8 | S_9 |
| λ_i | 0.156 | 0.167 | 0.029 |

TABLE 5 Failure mode probability of door subsystem

| Failure Mode M_j | α_{ij} |
|--|---------------|
| Door switch indicator does not light M_1 | 0.079 |
| Red dot display M_2 | 0.118 |
| Yellow dot display M_3 | 0.02 |
| Wear/deformation/loss of parts M_4 | 0.047 |
| Display system error M_5 | 0.17 |
| Abnormal noise M_6 | 0.02 |
| Cannot be closed properly M_7 | 0.012 |
| Cannot be opened properly M_8 | 0.098 |
| Malfunction of cab door M_9 | 0.266 |
| Activation of obstacle detection M_{10} | 0.17 |

Knowledgeable engineers and experts of the metro corporation were invited to evaluate scales of severity, detection, and maintenance costs for each failure mode. The final results are shown in Table 6. The decision matrix F_1 was established according to the trapezoidal fuzzy numbers defined in Section 2.1:

$$F_1 = \begin{bmatrix} (0.20 & 0.25 & 0.30 & 0.40) & (0.00 & 0.05 & 0.15 & 0.25) & (0.20 & 0.25 & 0.30 & 0.40) \\ (0.35 & 0.40 & 0.50 & 0.65) & (0.00 & 0.05 & 0.15 & 0.25) & (0.35 & 0.40 & 0.50 & 0.65) \\ (0.20 & 0.25 & 0.30 & 0.40) & (0.00 & 0.05 & 0.15 & 0.25) & (0.20 & 0.25 & 0.30 & 0.40) \\ (0.00 & 0.05 & 0.15 & 0.25) & (0.35 & 0.40 & 0.50 & 0.65) & (0.20 & 0.25 & 0.30 & 0.40) \\ (0.35 & 0.40 & 0.50 & 0.65) & (0.20 & 0.25 & 0.30 & 0.40) & (0.35 & 0.40 & 0.50 & 0.65) \\ (0.00 & 0.05 & 0.15 & 0.25) & (0.35 & 0.40 & 0.50 & 0.65) & (0.20 & 0.25 & 0.30 & 0.40) \\ (0.35 & 0.40 & 0.50 & 0.65) & (0.20 & 0.25 & 0.30 & 0.40) & (0.35 & 0.40 & 0.50 & 0.65) \\ (0.35 & 0.40 & 0.50 & 0.65) & (0.20 & 0.25 & 0.30 & 0.40) & (0.35 & 0.40 & 0.50 & 0.65) \\ (0.35 & 0.40 & 0.50 & 0.65) & (0.20 & 0.25 & 0.30 & 0.40) & (0.20 & 0.25 & 0.30 & 0.40) \\ (0.35 & 0.40 & 0.50 & 0.65) & (0.20 & 0.25 & 0.30 & 0.40) & (0.35 & 0.40 & 0.50 & 0.65) \end{bmatrix}$$

TABLE 6 Scales of door subsystem failure modes

| M_j | s_{ij} | d_{ij} | c_{ij} |
|----------|----------|----------|----------|
| M_1 | D | E | D |
| M_2 | C | E | C |
| M_3 | D | E | D |
| M_4 | E | C | D |
| M_5 | C | D | C |
| M_6 | E | C | D |
| M_7 | C | D | C |
| M_8 | C | D | C |
| M_9 | C | D | D |
| M_{10} | C | D | C |

The weight coefficient of severity, detection, and maintenance costs was chosen as [0.5 0.3 0.2], which means that the weight of severity is 0.5, the weight of detection is 0.3, and the weight of maintenance costs is 0.2. The quantitative value Z_1 and the weighted decision matrix V_1 were calculated as follows:

$$Z_1 = [0.2372 \quad 0.3790 \quad 0.2372 \quad 0.2728 \quad 0.4317 \quad 0.2728 \quad 0.4317 \quad 0.4317 \quad 0.3846 \quad 0.4317]^T,$$

$$V_1 = W \cdot F_1 = \begin{bmatrix} (0.140 & 0.190 & 0.255 & 0.355) \\ (0.255 & 0.315 & 0.405 & 0.530) \\ (0.140 & 0.19 & 0.255 & 0.355) \\ (0.160 & 0.225 & 0.300 & 0.400) \\ (0.315 & 0.375 & 0.450 & 0.575) \\ (0.160 & 0.225 & 0.300 & 0.400) \\ (0.315 & 0.375 & 0.450 & 0.575) \\ (0.315 & 0.375 & 0.450 & 0.575) \\ (0.275 & 0.325 & 0.400 & 0.525) \\ (0.315 & 0.375 & 0.450 & 0.575) \end{bmatrix}$$

The key degree of the door subsystem K_1 was computed from:

$$K_1 = \lambda_1 \cdot \sum_{j=1}^n \alpha_{1j} Z_{1j} = 0.0494.$$

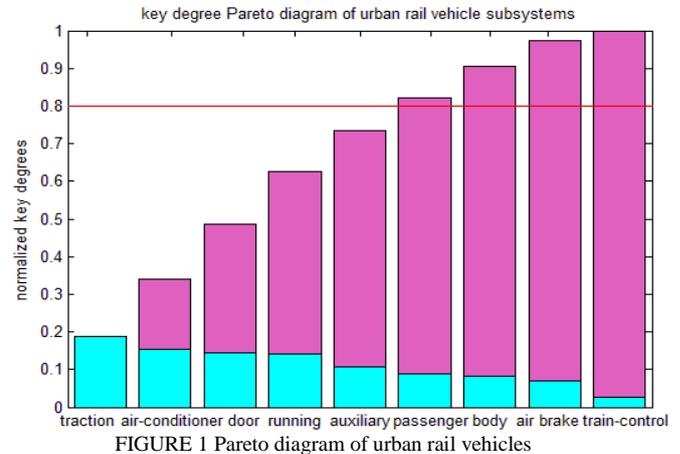
The above process was repeated until all the key degrees of other subsystems were calculated. Table 7 shows the obtained results.

TABLE 7 Key degree of rail vehicle subsystems

| | | | |
|-------|--------|--------|--------|
| S_i | S_1 | S_2 | S_3 |
| K_i | 0.0494 | 0.0241 | 0.0371 |
| S_i | S_4 | S_5 | S_6 |
| K_i | 0.0278 | 0.0480 | 0.0301 |
| S_i | S_7 | S_8 | S_9 |
| K_i | 0.0521 | 0.0647 | 0.0087 |

The key degrees of the subsystems were normalized and sorted by size, and then the Pareto diagram was drawn in Figure 1, which indicates that when the number of

accumulated normalization key degrees reached 0.8, the key subsystems were identified as the first five subsystems, including traction/electric brake, air conditioner, door, running gear and auxiliary system. The obtained results are consistent with most knowledgeable engineers and experts.



4 Conclusions

The fuzzy comprehensive evaluation method was applied in this paper to identify the key subsystems of urban rail vehicles. Occurrence, severity, detection, and maintenance costs were selected as evaluation factors, and these qualitative and quantitative indicators were combined to determine the key subsystems more comprehensively and reasonably. The proposed approaches were applied to Guangzhou Metro Corporation, and the five subsystems whose normalization accumulated key degrees reached 0.8 were identified: traction/electric brake subsystem, air conditioning subsystem, door subsystem, running gear subsystem, and auxiliary subsystem.

References

- [1] Xing L 2008 *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 38(1) 105-15
- [2] Li Y, Yu Y 2012 Quality risk analysis of submarine pipeline in construction period base on FEMA and fuzzy theory *China Safety Science Journal* 22(1) 112-7 (in Chinese)
- [3] Zi Z, Cho K-H, Sung M-H, Xia X, Zheng J, Sun Z 2005 In silico identification of the key components and steps in IFN- γ induced JAK-STAT signaling pathway *FEBS Letters* 579(5) 1101-8
- [4] LiSha P, Ling G, LingLi M 2012 Study on reliability of key systems of urban rail vehicles *China Railways* 51(7) 80-3 (in Chinese)
- [5] Xian S 2010 A new fuzzy comprehensive evaluation model based on the support vector machine *Fuzzy Information and Engineering* 2(1) 75-86
- [6] Chang K-H, Chang Y-C, Lai P-T 2011 Applying the concept of exponential approach to enhance the assessment capability of FMEA *Journal of Intelligent manufacturing* 22(2) 113-29
- [7] Awasthi A, Chauhan S S, Omrani H, Panahi A A 2011 Hybrid approach based on SERVQUAL and fuzzy TOPSIS for evaluating transportation service quality *Computers & Industrial Engineering* 61(3) 637-46
- [8] Ouédraogo A, Groso A, Meyer T 2011 Risk analysis in research environment – Part I: Modeling Lab Criticality Index using Improved Risk Priority Number *Safety Science* 49(6) 778-84
- [9] Liang Z, Yang K, Sun Y, Zhang H, Zhang Z 2006 Decision support for choice optimal power generation projects: Fuzzy comprehensive evaluation model based on the electricity market *Energy Policy* 34(17) 3359-64
- [10] Hao G, Yang Y, Wang Y, Cao Z, Guo H 2012 Research on the spares provisioning of the breech system based on fuzzy multiple attribute decision making *Value Engineering* 31(29) 21-3 (in Chinese)
- [11] Hill L G, Maucione K, Hood B K 2007 A Focused Approach to Assessing Program Fidelity *Prevention Science* 8(1) 25-34
- [12] Li G, W J, Zhang M, Gong, J, Chang P 2009 Approach to product modular design based on FPBS *Journal of National University of Defense Technology* 31(5) 75-80 (in Chinese)

| Authors | |
|---|---|
|  | <p>Zongyi Xing, born in February, 1974, Linyi City, Shandong Province, P. R. China</p> <p>Current position, grades: Associate Professor at School of Automation, Nanjing University of Science and Technology, China. University studies: Doctor degree at China Academy of Railway Science and Technology in China. Scientific interest: Fuzzy modeling, industry process control and rail traffic safety. Publications: more than 30 papers published in various journals. Experience: teaching experience of 10 years, 5 scientific research projects.</p> |
|  | <p>Lingli Mao, born in April, 1990, Nantong City, Jiangsu Province, P. R. China</p> <p>Current position, grades: the Graduate student of School of Mechanical Engineering, Nanjing University of Science and Technology, China. University studies: Bachelor degree from Nanjing University of Science and Technology in China. Scientific interest: Reliability, mechatronics and rail traffic safety. Publications: more than 4 papers published in various journals.</p> |
|  | <p>Limin Jia, born in January, 1963, Altay City, Xinjiang Province, P. R. China</p> <p>Current position, grades: Professor at School of Transport of Beijing Jiao tong University, China. University studies: Doctor degree at Academy of Railway Science and Technology in China. Scientific interest: Intelligent transportation system, rail traffic control and safety. Publications: more than 150 papers published in various journals. Experience: teaching experience of 20 years, 20 scientific research projects.</p> |
|  | <p>Yong Qin, born in December, 1971, Xuzhou City, Jiangsu Province, P. R. China</p> <p>Current position, grades: Professor at School of Transport of Beijing Jiao tong University, China. University studies: Doctor degree from Academy of Railway Science and Technology in China. Scientific interest: Intelligent transportation system, rail traffic control and safety. Publications: more than 50 papers published in various journals. Experience: teaching experience of 12 years, 6 scientific research projects.</p> |

Reputation risk contagion and control of rural banks in China based on epidemic model

Wu Yu*

School of Economics Henan University of Science and Technology, Luoyang City, Henan Province, China, 471000

Received 6 June 2014, www.tsi.lv

Abstract

Rural bank reputation risk is the negative evaluation formed in stakeholders' minds as a result of events which pose both internal and external risks. Regardless of whether or not these risk events have actually occurred, any resulting negative evaluations tend to propagate and accumulate in both the public's mind and within the main financial system. The growing negative opinion can create a herd effect, ultimately creating a reputation crisis. This paper attempts to research the contagion mechanism of rural bank reputation risk based on epidemic model, then explores a simulation study under different situations. The results show that the key to prevent or regulate reputation risk contagion is to reduce the unit available contact rate and the re-entry ratio, as well as the lurker infected rate. Finally, this paper puts forward management and control strategies from the perspective of the entire process. These strategies specifically focus on constructing an early warning mechanism, a dissolving mechanism and a long-term mechanism.

Keywords: reputation risk, reputation risk contagion, reputation risk control, epidemic model, rural bank

1 Introduction

The establishment of rural bank is one product of increment reform of rural finance in China, with a total of over 1000 within less than seven years. Nowadays, rural banks spread to all 31 provinces and more than half of 1880 counties, which have become the new force to support agriculture and SME. However, as the function of a financial enterprise is mainly to manage currency and credit services, rural banks also have the innate characteristics of financial fragility. This is especially true in the county areas, where the banks must cope with the inherent problems of small scale operations, manpower shortages and a lack of capital strength, these deficiencies make it more difficult to cope with liquidity risks, operational risks and credit risks. Once any of these negative events occurs, a reputation crisis is likely to surface as well, even infecting other financial institutions and posing a risk to the entire banking system. Therefore, rural bank reputation risk can be regarded as being the result of other types of risks concentrated to a certain degree, which creates dynamic changes. Another factor is whether or not the events which triggered the reputation risk -whether accurate or not- are controlled in a timely fashion.

Domestic and foreign scholars tend to research commercial bank reputation problems mainly from the perspective of reputation connotation. They emphasize that the root of a reputation crisis is information asymmetry, and that the basis of the crisis is the lessening of other parties' trust. These scholars reveal the mechanism of reputation as one of incentive and

punishment by building a reputation model. Bushman and Wittenberg investigated the importance of the role of reputation for borrowers and banks respectively. They found that borrowers with a more reputable standing led the loan arrangers to assume superior performance even before the loan was given, compared to the assumptions made with regard to borrowers with a lesser reputation. Banks with better reputations were associated with greater profitability and superior credit quality in the three years subsequent to the loans' initiation [1]. Reputation risk was not considered to be one of the eight risks associated with commercial banks by the Basel Committee prior to 1997. After that date, reputation risk is defined as a concept whereby, due to the deviations in the understanding of customers, partners, stock holders, bond holders and other relevant personnel, the sustainable management and business development capacity of a financial institution is affected by the subsequent negative impact. CBRC published its "commercial bank reputation risk management guidelines" in 2009, which clearly pointed out that, not only due to the internal factors but also the external factors affecting a bank, reputation risk reflected the perceptions of the stakeholders in the market. Such definitions are both discussed from the cause and effect, which can be showed in Figure 1.

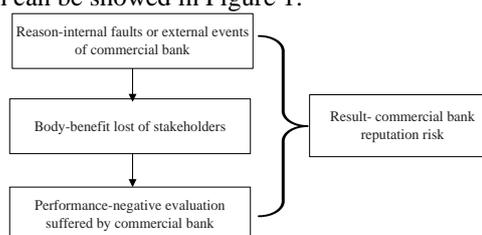


FIGURE 1 Reputation risk formation of commercial bank

* Corresponding author e-mail: meiyoyu1987@sina.com

The identification of reputation risk is closely linked to attempts to manage such risks. Reputation surveys suggest that a bank’s reputation is constructed around five elements: financial performance, quality of management, social and environmental responsibility, employee quality and the treatment of employees, and the quality of the bank’s goods and services [2]. Therefore, from a managerial disciplinary point of view, reputation risk management is a comprehensive approach which involves the anticipation of all potential variables, including taking different stake holders and circumstances into consideration [3, 4]. Ultimately, good management requires foreseeing crises, considering the associated opportunities and threats, and putting control plans into practice. Such an approach is both necessary and valuable to a bank’s improving the early warning mechanism and defusing mechanism of reputation risk [5, 6].

The studies mentioned above focus on theoretical analysis, and put forward the corresponding prevention measures through an analysis of the reputation risk causes for commercial banks, rather than dissecting and separating the infectious process and operation mechanism of reputation risk from the dynamic perspective. For rural banks, just as with community banks, whose ability to cope with reputation risk is much less than that of other commercial banks, the key to ensuring steady management is discovering how to master the reputation risk changes effectively and to take timely measures to limit that risk.

2 Reputation risk characteristics of rural bank

Rural bank reputation risk not only has the same characteristics as that of commercial banks, but it also possesses some unique features.

2.1 COMPLEXITY

Many factors induce the reputation risk of a rural bank, the risk may be caused by internal or external factors that cause concern among the public, or the risk can even be caused by the interaction of multiple factors. The effects of these factors may change in terms of their form over time. The main risk factors can be summarized as follows.

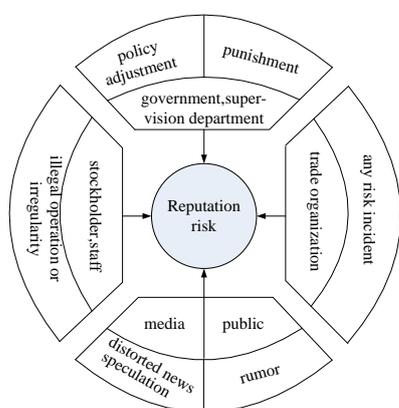


FIGURE 2 Main factors of reputation risk

2.2 INVISIBILITY

Compared with operational risk and credit risk, reputation risk is invisible and therefore extremely difficult to measure. Though some reputation risk evaluation models have been constructed, like the one presented by Harris-Fombrun [7], most banks have not accepted these quantitative analysis methods, and particularly not rural banks with poor technology resources.

2.3 TRANSMISSIBILITY

Reputation risk has strong negative externality, which can be spread via several formats, including television, newspaper, the internet and through the general population. For a rural bank located in a county where the informational development level is relatively low, the latter is the most common way for reputation risk to be disseminated. Apart from the fact that rural residents are gregarious and quick to follow suit, risk incidents are very likely to be distorted or exaggerated in the process of word-of mouth communication, and this can also exacerbate public panic. On March 24 of this year, Sheyang Rural Commercial Bank in Jiangsu suffered a run on the bank by one thousand depositors, merely because of a rumor that the bank was about to fail.

2.4 SUDDENNESS

The suddenness of reputation risk can be reflected in two aspects. On the one hand, there is a process of moving from quantitative changes to qualitative changes in the occurrence of reputation risk without prior warning, which cannot be detected in advance by bank managers. These changes can lead to a sudden outbreak when risk accumulates to a certain level or tipping point. On the other hand, the service objective of rural bank is to give priority to farmers and individual businesses who have a lesser ability to identify information.

3 Reputation risk contagion model of rural bank

The essence of any infectious disease is the transmission of germs from the carrier of pathogens to other individuals through contact. This method of infection is similar to that of rural bank reputation risk contagion. That is, the risk recipients transfer their negative evaluation of a rural bank to other stakeholders via various communication channels. Therefore, rural bank reputation risk contagion is the process of spreading and diffusing a negative evaluation of the bank, which in turn leads to the loss of both capital and customers, and which can even precipitate a rural financial system crisis.

Studies of mathematical models of the spread of infectious diseases began in the twentieth century, and began to flourish in the middle of that period. According to different infectious disease transmission mechanisms, a large number of mathematical models have been produced

in a bid to reveal the causes and systems of disease epidemics to predict development trends, and then to seek the optimal strategies for the prevention and control of epidemics [8]. In general, individuals are divided into several categories in such models, whose basic conditions include susceptible, infected and recovered. Based on the ideas of the infectious disease spread model, this paper establishes parallel dynamic models of rural bank reputation risk contagion.

3.1 SIR MODEL

In this model, the group conditions are divided into three categories, and the classifications can be shown as below.

TABLE 1 Classification of group conditions

| Category | Connotation |
|-----------|--|
| Spreader | Stands for infector, who has a perceived risk and spreads the negative evaluation to others |
| Credulous | Stands for susceptible, who is unable to identify the negative evaluation he receives but who continues to spread it to others |
| Rational | Stands for a recovered ,who can distinguish the authenticity of a negative evaluation |

Assumption one: The number of individuals in the model remains unchanged and is given as N .

Assumption two: in the beginning, the number of communicators holding a negative evaluation for a rural bank is i_0 ; the number of susceptible people who have not been informed of the negative evaluation is s_0 and $i(t)$, $s(t)$, $r(t)$ stands for the proportion of spreader, credulous and rational individuals with time changes respectively.

Assumption three: Each spreader contacts with some credulous individuals in a unit of time, where λ is defined as the unit available contact rate and stays fixed, due to the assumption that each individual can be uniform mixing and contact completely.

Assumption four: A spreader can be transformed into a rational, but the rational also has the chance of being infected again, so using μ to stand for the unit effective removal rate.

Assumption five: The number of the actual dependence of each spreader in unit time can be defined as σ , where $\sigma = \lambda/\mu$.

On the basis of the above assumptions, the model can be built as follows:

$$\begin{cases} di/dt = \lambda si - \mu i \\ ds/dt = -\lambda si \\ dr/dt = \mu i \\ s(t) + i(t) + r(t) = 1 \end{cases}, \tag{1}$$

where $i(0)=i_0$, $s(0)=s_0$. Since the number of rational individuals in the beginning is very small, $i_0+s_0 \approx 1$. In view of the difficulty of solving this problem, we can discuss their properties through the use of phase trajectory. In the

phase plane $s-i$, its domain can be given as: $D = \{(s,i) | s \geq 0, i \geq 0, s + i \leq 1\}$, and eliminating dt can get:

$$\begin{cases} di/ds = 1/\sigma s - 1 \\ i|_{s=s_0} = i_0 \end{cases}, \tag{2}$$

where

$$i(s) = (s_0 + i_0) - s + \frac{1}{\sigma} \ln \frac{s}{s_0}. \tag{3}$$

Define s_∞ , i_∞ , r_∞ as the limit value of s , i , r when $t \rightarrow \infty$, so from $ds/dt \leq 0$ and $s(t) \geq 0$, s_∞ exists, from $dr/dt \geq 0$ and $r(t) \leq 1$, r_∞ exists, also i_∞ exists. Then through simulation and MATLAB software programming, we can draw Figure 3 and Figure 4.

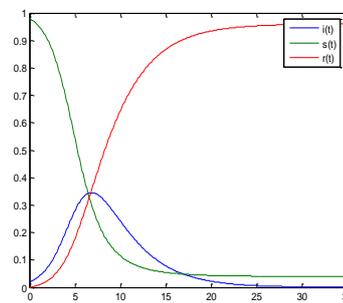


FIGURE 3 $i(t)-s(t)-r(t)$

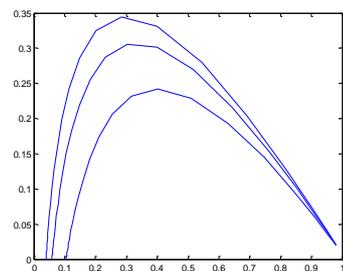


FIGURE 4 $i-s$

The parameters in Figure 3 are set as $\lambda=1$, $\mu=0.3$, $s_0=0.98$, $i_0=0.02$; the parameters in Figure 4 are set as $\lambda=0.9$, $\mu=0.3$, and $\lambda=1$, $\mu=0.4$, in addition, with s_0 , i_0 unchanged.

Remark one: no matter how much the value of the initial condition changes, the negative evaluation spreader will eventually disappear, due to the assumption of the constant number and no re-entry.

Remark two: Though the results of this model are contradictory to reality, the significance lies in discussing when the spreader number reaches maximization and the seriousness of the spread. When $s_0 > 1/\sigma$, $i(t)$ is initially on the increase and maximizes at $s(t)=1/\sigma$, then gradually decreases and eventually reaches 0; when $s_0 \leq 1/\sigma$, $i(t)$ and $s(t)$ both have monotone decreases, where s_0 is fixed and close to 1.

Remark three: The key to the control of reputation risk contagion caused by the spread of negative evaluations is to increase the value of $1/\sigma$, which can be called the "threshold". If meeting the condition that $s_0 \leq 1/\sigma$, the

negative evaluation will not be extended. Also, from $\sigma = \lambda/\mu$, it is reflected in the sense that we should try to reduce the unit available contact rate of the spreader and improve the unit effective removal rate, in order to effectively prevent the reputation risk contagion.

3.2 SIRS MODEL

In practice, individuals who are removed from the combination could enter it again and hence have a chance of being re-infected. Therefore, against the limitation of the SIR model, we can enhance it, that is to add assumption six based on the assumptions above.

Assumption five: The re-entry ratio of individuals removed from the combination in unit time is set by δ .

The model can be built as follows:

$$\begin{cases} di/dt = \lambda si - \mu i \\ ds/dt = -\lambda si + \delta r \\ dr/dt = \mu i - \delta r \\ i(t) + s(t) + r(t) = 1 \end{cases} \quad (4)$$

Since it is difficult to solving, we can use parameter analysis, the values are shown in Table 2.

TABLE 2 Simulated values of different parameters

| Parameter | λ | μ | δ | $i(0)$ | $s(0)$ |
|----------------|-----------|-------|----------|--------|--------|
| Simulation one | 1 | 0.3 | 0.2 | 0.02 | 0.98 |
| Simulation two | 1 | 0.3 | 0.3 | 0.02 | 0.98 |

Through MATLAB software programming, we can draw Figure 5 and Figure 6.

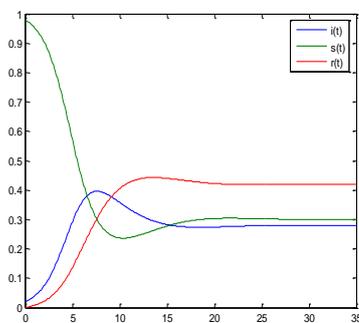


FIGURE 5 $i(t)-s(t)-r(t)$ simulation one

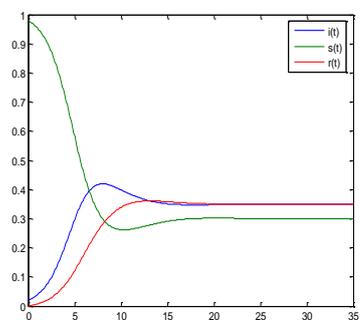


FIGURE 6 $i(t)-s(t)-r(t)$ simulation two

Remark four: It is reflected that the number of spreaders first increases, then decreases, and finally

stabilizes, both from Figure 5 and Figure 6. Along with the rise of the re-entry ratio, the number of spreaders also becomes larger in succession, not to mention the maximum value. However, the re-entry ratio does not have a significant effect on the occurrence time of the infectious peak. When reaching a steady state, the final number of credulous individuals would be approximately the same, accounting for 30 percent no matter what the re-entry ratio is. Whereas the number of spreader individuals is proportional to the re-entry ratio, which is inversely proportional to the number of rational individuals. From these assumptions, we can see that the re-entry individuals are the rational persons transformed from the spreaders, who may themselves increase the contagion effect.

3.3 SIRS MODEL WITH QUANTITY CHANGE

The number of rural bank stakeholders can change at any time, so adding assumption seven is based on the assumptions above.

Assumption seven: The increased rate of the number of individuals in the combination every unit time is set by α .

The model can be built as follows:

$$\begin{cases} di/dt = \lambda si - \mu i \\ ds/dt = \alpha - \lambda si + \delta r \\ dr/dt = \mu i - \delta r \end{cases} \quad (5)$$

Using parameter analysis, the values are shown in Table 3.

TABLE 3 Simulated values of different parameters

| Parameter | λ | μ | δ | $i(0)$ | $s(0)$ | α |
|------------------|-----------|-------|----------|--------|--------|----------|
| Simulation three | 1 | 0.3 | 0.2 | 0.02 | 0.98 | 0.1 |
| Simulation four | 1 | 0.3 | 0.2 | 0.02 | 0.98 | 0.2 |
| Simulation five | 1 | 0.3 | 0 | 0.02 | 0.98 | 0.2 |

Through MATLAB software programming, we can draw Figure 7 and Figure 8.

Remark five: It is shown that the number of spreader and rational persons will increase proportionately with that of individuals in the combination raises in Figure 7, but the number of credulous individuals basically remains unchanged after the eighth period. It is also found that we can effectively control the increase of spreader numbers by reducing the re-entry number under the condition that the increase rate of the number of individuals in the combination stays the same when comparing Figure 7 with Figure 8.

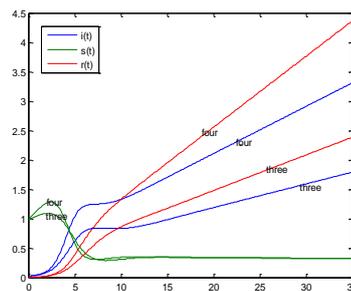


FIGURE 7 $i(t)-s(t)-r(t)$ simulation three and four

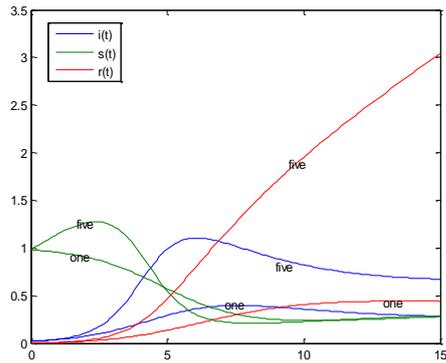


FIGURE 8 $i(t)$ - $s(t)$ - $r(t)$ simulation one and five

$$\begin{cases} di/dt = \epsilon e - \mu i \\ ds/dt = \alpha - \lambda s e + \delta r \\ dr/dt = \mu i - \delta r + \nu e \\ de/dt = \lambda s e - \epsilon e - \nu e \end{cases}$$

Using parameter analysis, the values are shown in Table 4.

TABLE 4 Simulated values of different parameters

| Parameter | λ | μ | δ | $i(0)$ | $s(0)$ | α | ν | ϵ | $e(0)$ |
|------------------|-----------|-------|----------|--------|--------|----------|-------|------------|--------|
| Simulation six | 1 | 0.3 | 0.2 | 0 | 0.98 | 0.1 | 0.1 | 0.6 | 0.02 |
| Simulation seven | 1 | 0.3 | 0.2 | 0 | 0.98 | 0.1 | 0.2 | 0.6 | 0.02 |

3.4 SEIRS MODEL

In general, there is always an incubation period before the credulous becomes the spreader after receiving negative evaluation, and the credulous individual will not infect others during that incubation period. If controlled in a timely manner, the credulous individual could be turned into a rational individual, therefore, adding assumption eight and nine based on the assumptions above.

Assumption eight: The credulous individual during an incubation period can be called a lurker; $e(t)$ stands for the proportion of lurkers with time changes.

Assumption nine: The lurker removal rate is defined as ν , and the lurker infected rate is defined as ϵ .

The model can be built as follows:

Through MATLAB software programming, we can draw Figure 9.

Remark six: Depending on the actual situation, the credulous individual should initially be a lurker, so it can be found after correcting the parameter values that the changes of different category numbers slows down. If adding the lurker removal rate under the condition that the other parameters remain unchanged, the number of spreaders would reduce in all periods, as well as the peak value, and even the peak period would be delayed. On the whole, the process of rural bank reputation risk contagion can be divided into about six stages, namely 1) incubation period, 2) start period, 3) outbreak period, 4) recession period, 5) fluctuation period and 6) stability period, and this conforms precisely to reality.

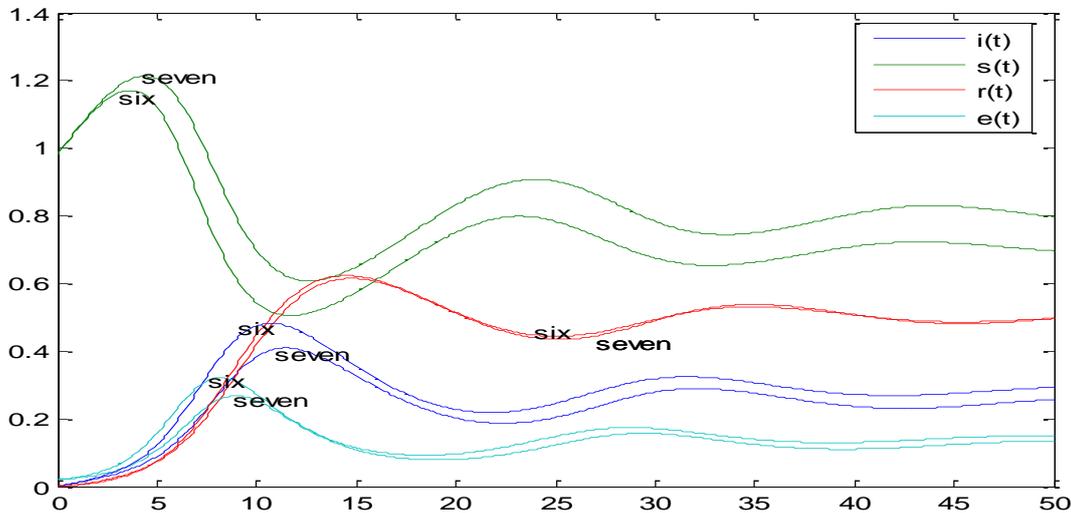


FIGURE 9 $i(t)$ - $s(t)$ - $r(t)$ simulation six and seven

4 Reputation risk control strategy of rural bank

As a new subject in the rural financial market, rural banks have a greater chance of suffering reputation risk, since given the limits of low recognition and credibility. Moreover, rural banks always tend to focus on business development while ignoring the importance and necessity

of reputation risk management. Therefore, it is essential for these banks to establish a sound set of reputation risk management systems and procedures. According to the study above, we can propose the corresponding countermeasures from three aspects, which are showed in Figure 10

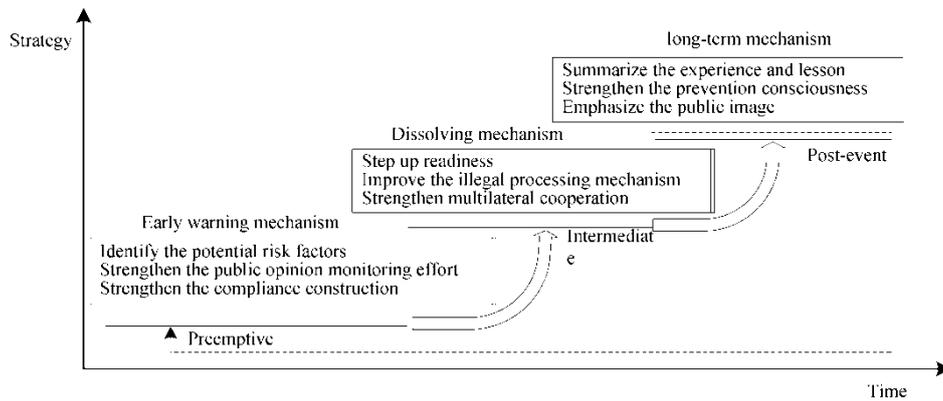


FIGURE 10 Countermeasures of rural bank reputation risk

4.1 PREEMPTIVE CONTROL - EARLY WARNING MECHANISM

Identify the potential risk factors. Rural banks should make regular investigations into every business unit or sector where reputation risks may be found. Based on the combined methods of quantitative analysis and qualitative diagnosis, the bank can identify the potentially dominant risk factors, as well as the recessive ones, and issue a warning signal, which would enhance the bank’s ability to react rapidly and make appropriate decisions.

Strengthen the public opinion monitoring effort. Rural banks should be responsible for the daily dynamic and real-time monitoring of public opinion. They should also strive to excavate and process the spreader individuals while they are still in the incubation period, in order to reduce the contact opportunities and decrease the proportion of spreaders. Such actions could significantly lessen the scale of risk infection and delay the peak time, or even act to curb the reputation risk contagion [9]

Strengthen the compliance construction. The internal reasons which cause rural bank to receive negative evaluation may include illegal operations, poor attitudes with regard to customer service, or withholding information. Therefore, the banks should increase the supervision and detection of abnormal or bad behaviours by their employees.

4.2 INTERMEDIATE CONTROL - DISSOLVING MECHANISM

Step up readiness. When negative evaluations spread, rural banks should promptly and persistently release timely, accurate and scientific information to the general public, to eliminate the negative impact of the poor evaluation as far as possible.

Improve the processing mechanisms for dealing with illegal activity. For those who act maliciously or spread inaccurate information, rural banks should quickly take legal action. Those who break the law must be prosecuted [10].

Strengthen multilateral cooperation. When negative events occur, rural banks should change their approach from one of escape to one of actively responding,

regardless of whether the information is false, partially accurate or completely true. The bank must clarify the true facts via the authorities or the media. Bank executives, village cadres, regulators, the government, TV, newspapers and so on should be used to take practical actions to answer questions and divert discontent. In short, the bank can control the further spread of poor evaluations through good public relations practices.

4.3 POST-EVENT CONTROL - LONG-TERM MECHANISM

Summarize the experience and the lessons to be learned. Rural banks should create a summary and report when reputation risk happens, in order to accumulate work experience which can be used in the future to prevent and defuse the similar crises, as well as to improve their ability to comprehensively dispose of a reputation crisis.

Strengthen the prevention consciousness. The staffs of rural banks are typically new graduates who have limited capacity and experience. Therefore, rural banks should organize for all their staff to participate in professional training or lectures, in order to actively cultivate their awareness of reputation risk prevention and reputation management skills, as well as to sublimate reputation risk management to the business philosophy of the banks’ culture. Moreover, the staff should strengthen their service consciousness and disseminate their financial knowledge to the local populace, as this could enhance their own legal awareness and information recognition abilities.

Emphasize the public image. Firstly, rural banks should try to improve customer satisfaction through the creation of obligation informs, to address solutions proactively and to protect the legitimate rights and interests of customers. Secondly, the banks should increase external publicity, using the power of public media to improve their social reputation. Thirdly, the bank should exchange and share their reputation risk management experiences with their peers. This could result in the mutual supervision of risks and enhance the anti-risk abilities of the entire industry. Ultimately, building a good reputation for all rural banks would improve customer brand loyalty and help reduce losses to the lowest possible level.

5 Conclusions

Rural bank reputation risk can be defined as negative evaluations by stakeholders that are not only due to internal factors but also to external ones. Since the essence of rural bank reputation risk contagion is similar to that of infectious diseases, this paper uses epidemic models to deduct contagion mechanisms under different situations, in order to excavate the key indicators to prevent or regulate reputation risk contagion, then puts forward an

early warning mechanism, dissolving mechanism and long-term mechanism to realize full process control, which could act as an important and significant reference.

Acknowledgment

The research was supported by the National Natural Science Foundation Project "Research on the Conduction Mechanism and Control of Operational Risk in Commercial Banks" (grant number: 70771085)

References

- [1] Bushman R. M, Wittenberg M R 2012 The role of bank reputation in 'certifying' future performance implications of borrowers' accounting numbers *Journal of Accounting Research* **50**(4) 883-930
- [2] Bebbington J, Larrinaga C, Moneva J M 2008 Corporate social reporting and reputation risk management *Accounting, Auditing & Accountability Journal* **21**(3) 337-61
- [3] Okur M E, Arslan M L 2014 Reputation management in global financial institutions *Managerial Issues in Finance and Banking* 173-84
- [4] Scandizzo S 2011 A framework for the analysis of reputational risk *Journal of Operational Risk* **6**(3) 217-35
- [5] Lu M, Pan X 2010 Construction of reputation risk-resolving mechanisms of commercial banks *Journal of Shijiazhuang University of Economics* **33**(3) 7-11 (In Chinese)
- [6] Liu M 2013 Research on reputation risk management of commercial banks in post-crisis era *Accounting and Finance* **28**(1) 57-60
- [7] Fombrun C J, Gardberg N A, Sever J M 2000 The reputation quotient: a multi-stakeholder measure of corporate reputation *Journal of Brand Management* **7**(4) 241-55
- [8] Hethcote H W 2000 The mathematics of infectious diseases *SIMA Review* **42**(4) 599-653
- [9] Lan Y 2012 Research on network rumor diffusion model and countermeasures in the emergency *Information Science* **30**(9) 1334-8
- [10] Qiao H, Li Y 2013 Research on the conduction and effect of reputation risk of financial enterprises *On Economic Problems* **36**(9) 87-92

Author



Yu Wu, born in March, 1987, Luoyang County, Henan Province, China

Current position, grades: the lecturer of school of economics, Henan University of Science and Technology, China.

University studies: doctor's Degree in Management from Wuhan University of Technology in China.

Scientific interest: financial engineering, rural finance.

Publications: more than 7 papers.

Experience: teaching experience of 4 years in spare time, 1 scientific research project.

Research on the influential factor of consumer model based on online opinion leader

Fei Meng¹, Jianliang Wei^{2*}

¹Department of Public Foundation Zhejiang Police College, Hangzhou, Zhejiang, China, 310053

²School of computer and information engineering / contemporary business and trade research center Zhejiang Gongshang University, Hangzhou, Zhejiang, China, 310018

Received 26 March 2014, www.tsi.lv

Abstract

With the development of internet and e-commerce, online opinion leader becomes an important information resource which influences the purchase decision and behaviour model of online consumer, although the influential mechanism is still uncertain. In order to obtain a more precise user model, Grounded Theory is adopted in this paper and an interview table is designed according features of online opinion leader. Then, more than 20 online consumers concerning on opinion leader frequently in internet communities such as Taojianghu and Douban are selected for interview. After open coding, axial coding and selective coding on the interview materials, several findings are obtained: professional knowledge and interactive features of opinion leader influenced the purchase intension of consumer; characters such as visual cues and timeliness of recommended information from opinion leader have impact on consumer intension; consumer perceived value of product recommended by opinion leader influenced their purchase behaviour; and trust is the principle reason for consumer' acceptance on product information recommended by opinion leader.

Keywords: online opinion leader, grounded theory, consumer behaviour model

1 Introduction

The influence of word-of-mouth (WOM) on consumer behaviour has long been concerned. There are many researchers examine the effect of WOM in different fields, and they find that the impact brought by WOM in brands selection was far greater than newspaper ads, marketing staff, radio ads. With the popularity of online consumption, effect of e-WOM imposes on online consumer purchase decision have become a topic of great concern, especially the emergence of online opinion leaders, whose influence is more powerful and far-reaching.

Now the questions are: What kind of online opinion leader has more powerful impacts? What kind of recommendation information has more influential? To the first question, important characteristics between online opinion leader and ordinary consumer should be identified, and instinct characteristics of online opinion leader need to be described. To the second question, distinct features of the persuasive recommendation information should be recognized, as well as the dimensions to describe them. By now, these two questions are not solved yet in academic. This paper is dedicate to analyse the mainly characteristics of online opinion leader and features of the information their recommended, and Grounded Theory is adopted in this exploratory analysis.

2 Related research

2.1 INFLUENTIAL FACTORS OF CONSUMER BEHAVIOR MODEL

Studies on the effect of WOM origins from the communication persuasion theory made by Hovland (1953), which summarize factors affecting the results of communication into three categories: source, message and receiver. They change the persuasion result by affecting various factors around the disseminators, information content, communication channels and recipients [1]. Basis on this, factors which affecting the results of communication are divided into three kinds, namely, disseminators, the information itself, and recipients. According the Communication Persuasion Theory, consumers can be influenced in following aspects by WOM: First, the features of disseminator, especially the professional level and opinion leaders; second, the homophily of disseminators and recipient; Third, the characteristics of recipient, including the professional level and personal preferences [2].

2.2 INFLUENTIAL MECHANIZE OF CONSUMER BEHAVIOR MODEL

WOM plays an important role in the process of consumer's attitude form and behaviour model, and this is one of ideas that has been widely accepted in consumption behaviour

* *Corresponding author* e-mail: jianliang@zjgsu.edu.cn

[3]. Ennew, Banerjee & Li (2000) have indicated that WOM had a significant impact on the choice of products and services [4], since WOM is one of the most influential market information sources to consumers [5]. Many traditional WOM researchers, such as Engel (1969), confirmed a remarkable power that WOM had on consumer decision-making [6], including consumer purchase behaviour, brand selection and businesses choose [7].

With the development of Internet activities, researches on online WOM spread and consumer behaviour has become a hot issue. Smith (2002) studied the effect mechanism of information recommended by ordinary consumers to the decision of other consumer decisions [8], in which variable trust is adopted, as the mediating variable between consumer decision and three influential factors, including individual differences of recipient, characteristics of disseminators recommendation and purchasing target. Then, Cheung et al. (2008) further examined the homogeneous variables on this basis, and suggested that consumers were more inclined to accept the comments made by similar consumers and editors, the homogeneity between disseminators and recipients was positively correlated with WOM influence [9].

2.3 WOM AND ONLINE PURCHASE

Park, Lee & Han (2007) suggest that online information search behaviours of consumers before buying are always positively related to purchase decisions [10]. Based on this, Shaver (2007) found that online information searching behaviour is increasing while conducting a study monitor the difference of U.S. consumer attitude from 2000 to 2005, and found the purpose of people search for online product information is to improve their decision-making capacity [11]. Compared with consumer searching, the impact of online recommendations is more significant on consumers' final purchase decisions. Some researchers found that online book reviews was very important on book sales based on the data of Amazon.com and BN.com [12]. Compared with ordinary products, decision-making on experience goods can be influenced much easier by online reviews, as well as new coming products. Meanwhile, positive WOM can strengthen the tie between consumers trust and their desire to purchase online, and enhance the reliability of consumer perception as well [13].

2.4 OPINION LEADERS AND ONLINE PURCHASING BEHAVIOUR

Chevalier & Mayzlin's research indicated that online comments made by participants with high reputation and exposure in platform such as Amazon had a deep influence on product sales [12]. Lim and Chung (2014) pointed out that the influence of opinion leaders was positively related to their familiarity of product. When opinion leaders have more related knowledge and more familiar with the

product, indicating a higher expertise degree, they will more likely become the consulting person for consumer who is seeking for product recommendation information [14]. Meanwhile, influence of the recommended information transmitted by opinion leaders was closely related with transfer distance, i.e., the number of users in information transmission [15].

In summary, although many researchers have paid great attention on the relationship between WOM and consumer purchasing behaviour, but rarely literatures is from the perspective of opinion leaders, let alone a systematic study on the influential mechanism is lacked. Therefore, it is necessary to resort to a qualitative research to summarize a theoretical context close to reality of online opinion leader and its influential factors.

3. Research methods

3.1 METHOD SELECTION

3.1.1 Segregation index

With the development and application of computer and information technology, traditional methods rely on experience are not important as usual, researchers begin to use computer and network method to work out problems which traditional methods are hard to resolve [16, 17]. Refer to our topic, social network is becoming more and more popular to analyse the relationship between opinion leader and consumer behaviour, and method such as segregation index is suitable for finding the influential of opinion leader's attribute on its follower, i.e. the number and structure of follower's community. Its basic steps are as follows:

$$Seg = \frac{E(X) - X}{E(X)}, \quad (1)$$

in which $E(X)$ present the expected number of cross-group ties, X presents the number of observed cross-group ties. Mixing matrix need created first, and next is two indicator matrixes which can be used to generate the Mixing matrix (M).

$$M = \begin{matrix} I' & A & I \\ (k \times k) & (k \times n) & (n \times n) \end{matrix} \quad (n \times k) \quad (2)$$

According to different principles, there are two ways to get segregation index.

(1) Freeman Segregation Index.

Here $A=+1$ and $B=-1$ represent the nodes types, p is the mixing matrix where entry p_{ab} with $a, b \in \{A, B\}$ counts the number of links connecting a a-type node with a b-type node. Freeman Segregation Index can be expressed as:

$$FSI = \frac{[E(p_{AB}) + E(p_{BA})] - [p_{AB} + p_{BA}]}{[E(p_{AB}) + E(p_{BA})]} \quad (3)$$

(2) Spectral Segregation Index.

For each type $X \in \{-1, +1\}$, define the Spectral Segregation Index with $SSI(x)$ as the largest eigenvalue of A_x . Since that

$$d_{\min}(x) \leq \bar{d}(x) \leq SSI(x) \leq d_{\max}(x), \quad (4)$$

in which, $d_{\min}(x)$ refers the minimum degree of the nodes associated with agents of type x in the sub-graph W composed of nodes hosting an agent of type x , $\bar{d}(x)$ means the average degree and $d_{\max}(x)$ means the maximum degree. Then, we have Spectral Segregation Index as follow:

$$SSI(x) = \frac{SSI(x) - d_{\min}(x)}{d_{\max}(x) - d_{\min}(x)}. \quad (5)$$

3.1.2 Grounded analysis

In this study, Grounded Theory (GT) is adopted for selecting opinion leader and ordinary consumer to conduct in-depth interviews, analysing and mining the interview data, and core category reflecting the opinion leaders' impact on consumers in the online environment is eventually extracted. Considering the relationship in this scope, a suitable theoretical model is built. In other words, using grounded analysis method must have a field to be studied first. Then, through a systematic collection and procession of data, concepts and theories sprout out from the field are extracted, thereby a new theory can be explored and obtained. In this process data collection and procession is very important. Walker and Myrick pointed out that the Grounded Theory must be based on realistic data, and theoretical framework can obtained only through in-depth analysis, thus grounded analysis has a strong dependence on the data [18]. Simultaneously, the target of grounded analysis is to confirm a statement of a certain phenomenon which required to be studied, which is very suitable for the situation of online opinion leader and its impact on consumer purchase intention, thus we choose Grounded Theory as our research method.

3.2 GROUNDED ANALYSIS STEPS

The procession of Grounded Theory can be classified into five stages and nine steps as following: research and design, collection of information, data compilation, data analysis and documentary comparison [19].

1) Stage I: Research and design

Step 1: Literature discussion. The literature discussion is a pre-understanding in order to improve the theoretical sensation. In the process of conducting literature discussion, questions and preliminary ideas of the research should be clearly defined, and unrelated variables should be excluded in order to improve the external validity of the study.

Step 2: Case selection. The purpose of Grounded Theory is forming a theory which can be explained through data analysis of purposive sampling.

2) Phase II: Information collection

Step 3: Design a stringent data collection methodology to build a database of case studies. Variable data collection methods can be applied to gather information, such as participant observation, interview and text analysis.

Step 4: Enter research place. Researchers can use flexible data collection methods when collecting data in the practical place, focusing on the immediate theme and particularity cases.

3) Phrase III: Data compilation

Step 5: Data compilation. Data compilation includes related notes, records and filled data analysing, in which events can be sorted by time.

4) Phrase IV: Data analysis

Step 6: First case analyse. Data analysis includes three sub-steps, namely, open coding, axial coding and selective coding.

a) Open coding: open coding includes data decomposition, inspection, comparison, conceptualizing and subcategorizing. Firstly, the collected data is conceptualized, then is the sub-scope and nominate process. The process of sub-scope means combining set of concepts into a scope which include many sub-scopes.

b) Axial coding: the main purpose of axial coding is to establish links between the scope and sub-scopes, find out the causal relationship between scopes, and establish a theoretical framework. The axial coding comes after open coding, which analyse conditions and context of the phenomenon, the strategy and result in action (or interactive).

c) Selective coding: The purpose of selective coding is to identify the core themes. At first, core category and link the core category with other scope systematically are selected. Then we supplement and improve the unrefined scope after obtaining the relationship between the scopes. Core category is the centre, which combines all other areas.

Step 7: Theoretical sampling. The purpose of sampling is to form a new theory, which gathering commons in different samples to form certain theory. After open coding, axial coding and selective coding analysis, the first case could be transmitted into a preliminary theory. Then, to analyse the second case to obtain another theory and verify whether the second theory is coincide with the first case. If not, comparing and amending the two theories to form a more complete theoretical model. After that, return to the second step, and purposive sampling numbers of typical cases to analysis, until the new theory becomes saturation.

Step 8: Theoretical saturation. The initial theory obtained must be tested and fitted with new samples. If the theory is consistent in different situation, we call it is theoretical saturation.

5) Phrase V: Documentary comparison

Step 9: Comparing the new theory and the existing documents. If they conflict, theory revision should be implemented. If not, such comparison can help find universalization deductive recognition which is helpful to improve the external validity.

It can be perceived that the analysis process of grounded method is critical to refine conclusions. As Walker and Myrick said, grounded analysis is a circle process of data collection and analysis. Based on the theory in pre-stage form, using the cases and other theories to validate the new theory, we can revise and improve the theory constantly, and ultimately form a theory which can reflect the object phenomenon [18].

4. Grounded research design and data collection

The purpose of grounded analysis is to demonstrate the typical object of aimed phenomenon. According to the theme of this paper, the investigators need to satisfy conditions as follows: be familiar with Internet, visit various social networking sites and e-commerce sites frequently, be interested in opinion leaders and have times of their information gathering history.

Investigators are interviewed by face-to-face, and we found that opinion leaders with high degree of attention are mainly from online community and blog forum. According to the frequency of use and the tendency of respondents, four typical platforms are ultimately selected, including Taobao, Taojianghu, Douban and Tianya. Moreover, two fashion websites: Onlylady and YOKA are selected as well, as a data implement. Numbers of followers and relevant information recommended by the opinion leaders in the above sites are also chosen, in order to enrich the collection of information.

Twenty five people are finally interviewed, including college students, enterprises and workers, freelancers and network sellers, etc. The respondents are quite young, mainly at the age of 18 to 40 years old. Ten people are less than 22 years old, twelve people are 23 to 30 years old, and three people are above 30 years old. To ensure consistency of data collection, and facilitate the compilation and comparison process, all objects are interviewed by the authors personally. The time for each interview maintains from 20 to 40 minutes. The main content of the interview is shown in Table 1.

TABLE 1 Interview table based on Grounded Theory

| | Main content |
|----------------------------|--|
| Part 1 Basic information | ①Your gender, age, level of education, profession? |
| Part 2 Attention behaviour | ①How long have you been searching for product information by Internet? |
| | ②How long have you noticed the online opinion leaders? |
| | ③Remember the latest impressive product you have bought since of opinion leaders, which places did you get the recommendation? |
| | ④What advantages of the product have? |
| | ⑤Would you usually communicate with online opinion leaders? Or concern hers/his interaction with others? |
| Part 3 The main interview | ①How much does opinion leaders' recommendation influence your purchase decisions? These effects mainly present in which aspects? |
| | ②What factors influence your evaluation to the recommendation information? |
| | ③Do you have a lot interaction with opinion leader? Do you think they have professional knowledge? |
| | ④Do you think there is some similarity between you and the opinion leader? Whether this would affect your acceptance to hers/his advice? |

5. Data analysis and factors affecting model construction

On the basis of the interviews, open coding, axial coding and selective coding are conducted gradually. The whole interviewed case is named "A", and the English letter "A" is adopted in order to facilitate subsequent analysis.

Open coding is the foundation of all the coding in Grounded Theory. After original data gathering through in-depth interviews and observations, comparative analysis is used to develop concept. We finally extracted 40 concepts by analysing and comparing the recorded data on "opinion leaders' impact on consumers purchase intention", which concluded based on the existing literatures and interviews' terminology.

As Larossa (2005) described, the concept here could be a word, a phrase or a short sentence, but no matter what

form of expression, it must be able to accurately reflect the nature of the interview connotation [20]. According to this principle, the concepts are classified, and 12 scopes (A1-A12) are ultimately obtained as following: professional knowledge of opinion leader, product involvement of opinion leaders, interactivity of the opinion leaders, popularity of the opinion leaders, visual clues of recommended information, timeliness of recommended information, consistency of recommended information, homogeneity of consumers, functional value, emotional value, trust and purchase intension. Details are shown in Table 2.

Based on this, Table 3 illustrates the process of open coding, the whole process involves a lot of information and contents, but partial process of the coding is presented since of space limitation.

TABLE 2 Interview table based on Grounded Theory

| No. | Categories | Concepts |
|-----------------|------------------------------|---|
| A ₁ | professional knowledge of OL | product knowledge, expert, authority |
| A ₂ | product involvement of OL | consumer experience, product preferences |
| A ₃ | interactivity of OL | interaction, tie strength, familiarity |
| A ₄ | popularity of OL | social status, public image, celebrity, forum leader |
| A ₅ | visual clues of RI | present form, visually appealing, impact |
| A ₆ | timeliness of RI | information update, timeliness, fashion |
| A ₇ | consistency of RI | recommend consistency, similar product, similar viewpoint |
| A ₈ | consumer homogeneity | personality, hobbies, interests and values |
| A ₉ | functional value | product quality, performance, reliability, durability |
| A ₁₀ | emotional value | pleasure, excitement, happiness, sense of identity |
| A ₁₁ | trust | capacity, goodwill, honesty |
| A ₁₂ | purchase intension | product certainty, product interest, product desire |

Note: OL refers to opinion leader, RI refers to recommended information

5.2 AXIAL CODING AND PARADIGM MODEL

In this chapter, linkages between the scopes are analysed step-forward, especially the link of the scope and their corresponding sub-scopes. Then, a further information re-integration to develop the key scope is executed. There are six parts in axial coding: causality conditions, phenomenon, context, intermediary conditions, action/interaction strategies, and results. In this phrase, a paradigm model is needed to further excavate the meaning of scope, which is an important analysis tool in axial coding. After sorted and analysed the six parts in online

opinion leader case, the main scope paradigm model of this study is finally obtained, as shown in Figure 1.

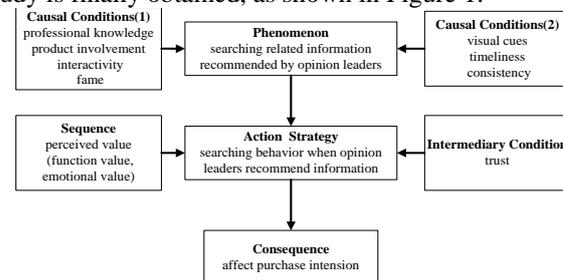


FIGURE 1 Main scope paradigm model

TABLE 3 An example of open coding on opinion leaders' influence

| Data records | Open coding | | | |
|--|------------------------------------|---|---|--|
| | conceptualization | categorization | category property | property dimension |
| This OL know many brands, familiar with product feature and function(a ₁) | product knowledge a ₁ | 1. a ₁ , a ₂ , a ₃ conceptualization as: professional knowledge (A ₁) | 1. nature of professional knowledge: comprehension, type, status | comprehension: strong/weak type: many/few status: high/low |
| This OL is very professional, and a well-known expert in this area (a ₂) | expert a ₂ | | | |
| | | | | |
| This OL bought too many similar products, very experienced(a ₄) | purchase experience a ₄ | 2. a ₄ , a ₅ conceptualization as: product involvement of OL (A ₂) | 2.nature of product involvement: familiarity, likeness, input | familiarity: high/low likeness: strong/weak input: much/little |
| She have a preference for such products, particularly to one of them (a ₅) | product preference a ₅ | | | |
| | | | | |
| This recommendation is very impressive, with video and detailed usage (a ₁₄) | forms a ₁₄ | 5. a ₁₃ , a ₁₄ , a ₁₅ conceptualization as: visual cues of RI (A ₅) | 5. nature of visual cues: forms, clearness, content arrangement | forms: many/few clearness: high/low content arrangement: good/bad |
| Pretty product pictures, clearly show product details, and very attractive(a ₁₅) | visual appeal a ₁₅ | | | |
| | | | | |
| She recommended some new products, which I do not familiar with(a ₁₇) | timeliness a ₁₇ | 6. a ₁₆ , a ₁₇ , a ₁₈ conceptualization as: RI timeliness (A ₆) | 6. nature of timeliness: update frequency, products follow up | update frequency: high/low products follow up: fast /slow |
| This net advisor is very fashion, his recommendation reflect the latest vogue (a ₁₈) | latest vogue a ₁₈ | | | |
| | | | | |

| | | | | |
|---|----------------------------------|---|---|---|
| This OL's recommendation always have a good quality (a ₂₇) | product quality a ₂₇ | 9. a ₂₇ , a ₂₈ , a ₂₉ , a ₃₀ conceptualization as: perceived function value of consumer (A ₉) | 9. nature of function value: quality, durability | quality: good/bad durability: good/bad |
| Since he have used a lot products, his recommendation is always easy and suit to use (a ₂₈) | performance a ₂₈ | | | |
| | | | | |
| The recommendation of experts is more reliable, and helpful to understand the advantages and disadvantages (a ₃₅) | ability a ₃₅ | 11. a ₃₅ , a ₃₆ , a ₃₇ conceptualization as: trust (A ₁₁) | 11. nature of trust: ability, standard, honest | ability: high/low standard: high/low honest: reliable/unreliable |
| | | | | |
| The product in F' blog is nice, as well as its packaging (a ₃₈) | product affirm a ₃₈ | 12. a ₃₈ , a ₃₉ , a ₄₀ conceptualization as: purchase intention (A ₁₂) | 12. nature of purchase intention: willingness, desire | willingness: strong /weak desire: strong/weak |
| I have never try this, it seems an interesting recommendation, I would try it next time (a ₃₉) | product interest a ₃₉ | | | |
| Seeing lots of professional are using this, I am ready to buy one (a ₄₀) | product desire a ₄₀ | | | |
| | (40 concepts total) | (12 scopes total) | | |

5.3 SELECTIVE CODING AND MODEL VARIANTS

The purpose of selective coding is to develop the core category. Depending on the interactive comparison of the original data, concepts and categories, especially relationships among scopes, Core category is refined which can best reflects the essence of the phenomenon, as

well as the baseline around the core scope. The analysis in this process is quite abstract, and four core scopes are extracted, including professional knowledge, product involvement, interactivity and visual cues. Based on this, variable definitions and dimensions between traditional e-commerce environment and online social business environment are compared, as shown in Table 4.

TABLE 4 Variant definition and dimension contraction in different environment

| | | Traditional environment | Socialized environment |
|---------------------|------------|--|--|
| Expertise | definition | professional knowledge of recommender | comprehensive and systematic product knowledge of recommender |
| | dimension | professional knowledge, authority, expert | professional knowledge, experiences, product knowledge |
| Product Involvement | definition | product preference, time spending | product preferences and involvement |
| | dimension | product preferences, energy cost and time concern concerning | product preferences, energy cost, online information concerning and transmission |
| Interactivity | definition | social tie strength between individuals | tie strength between individual and website, individual and opinion leader |
| | dimension | importance, contact frequency, relationship type | online linkage, interactive information, website relationship |
| Visual Cues | definition | picture display | various types of information presented by recommender |
| | dimension | clear, well organized | various forms, detail picture, clear video |

On this basis, the mainline of this study can be stated as follows: due to the shortage of product-related information, consumers tend to take the recommendation from experienced opinion leaders on related products. Opinion leaders' effect on consumer purchase intention is subjected to their characteristics such as professional knowledge and product involvement, as well as features of information recommended, like interactivity and visual cues.

6 Discussions and results

Currently, e-commerce is becoming more and more popular and wide spreading, and role of online opinion

leaders on consumer purchase and behaviour model is increasing significant, especially in the field of online purchase decision-making. Through the interviews of online community users, and qualitative analysis based on Grounded Theory, several results which are significant for online user model are found:

1) Characteristics of opinion leader.

Compared to traditional opinion leaders, the online opinion leaders have many new features, their product knowledge is richer and social network is wider. Just as the words of two grounded interviewers: "Opinion leaders" comments on product contains much of professional knowledge, letting me aware systematic product knowledge and insight understanding about the

product” and “When you are hesitate in selection, it is better to ask others or directly to the opinion leaders for advice”, suggesting that the professional knowledge and interactivity of opinion leaders plays an important role on consumer behaviour, as well as in user model construction.

2) Features of recommendation information.

Compared with general public reputation, recommendation of opinion leaders is distinct in richer visual cues and stronger timeliness. Just as the viewpoints of two grounded interviewers: “Product recommendation has a detailed description and a very clear picture” and “When I see a new product of this brand, the immediate comments from opinion leaders help me know better about this product”, suggesting that features such as visual clues and timeliness of recommendation influence consumer behaviour, and the corresponding model.

3) Consumer perceived value.

Consumer perceived value is an important psychological factor which influences consumer purchase intension and model construction. When the product recommendation makes a high perceived value, consumers tend to generate an interest for the product. Just as the looking of two grounded interviewers: “The product he recommended in his blog seems interesting and quite simple to use, and the raw material is also quite reliable, it should be in good quality” and “I really like the package she bought, and now a lot of people want to buy one, I would be fashionable if I have one”, indicating that consumers perceived value of the products recommended

by the opinion leaders also have an impact on consumer behaviour.

4) Intermediary role of trust to behaviour model

Opinion leaders’ impact on consumer purchase intention need an intermediary mechanism, depending on Grounded Theory, we summarized that opinion leaders often bring a sense of trust to consumers, which would further affect their purchase intension. The interviews show that the most important reason of the interviewee’s acceptance of recommendation is their trust on opinion leaders. The professional knowledge, product involvement, interactivity and fame of opinion leaders, the visual cues, consistency and timeliness of the recommended information, and the functional value and emotional value consumer perceived, may affect the generation of trust.

Acknowledgement

The work was supported by the supported by the National Social Science Foundation of China (No.12CTQ028), Humanities and Social Science Foundation of Chinese Ministry of Education(No. 13YJC870019), Zhejiang Nonprofit Technology Research Projects (No. 2013C33060), Special Foundation on Innovation Team Building and Personnel Training of Zhejiang (No.2013F20008), Contemporary Business and Trade Research Center of Zhejiang Gongshang University which is the Key Research Institute of Social Sciences and Humanities Ministry of Education.

References

- [1] Fang D, Fang C-L, Tsai B-K, Lan L-C, Hsu W-S 2012 Relationships among trust in messages, risk perception, and risk reduction preferences based upon avian influenza in Taiwan *International Journal of Environmental Research and Public Health* **9**(8) 2742-57
- [2] Chu S-C, Kim Y 2011 Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites *International Journal of Advertising* **30**(1) 47-75
- [3] Sandes F S, Urdan A T 2013 Electronic Word-of-Mouth impacts on consumer behavior: exploratory and experimental studies *Journal of International Consumer Marketing* **25**(3) 181-97
- [4] Ennew C T, Banerjee A K, Li D 2000 Managing word of mouth communication: empirical evidence from India *International Journal of Bank Marketing* **18**(2) 75-83
- [5] Khalid S, Ahmed M A, Ahmad Zr 2013 Word-of-Mouth communications: a powerful contributor to consumers decision-making in healthcare market *International Journal of Business and Management Invention* **2**(5) 55-64
- [6] Engel J F, Blackwell R D, Kegerreis R J 1969 How information is used to adopt an innovation *Journal of Advertising Research* **9**(4) 3-8
- [7] Richins M L 1983 Negative word-of-mouth by dissatisfied consumers: a pilot study *Journal of Marketing* **47**(1) 68-78
- [8] Goldsmith R E 2002 Explaining and predicting consumer intention to purchase over the internet: an exploratory study *Journal of Marketing Theory and Practice* **10**(2) 22-8
- [9] Cheung C M K, Lee M K O, Rabjohn N 2008 The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities *Internet Research* **18**(3) 229-47
- [10] Park D-H, Lee J, Han I 2007 The effect of on-line consumer reviews on consumer purchasing intention: the moderating role of involvement *International Journal of Electronic Commerce* **11**(4) 125-48
- [11] Shaver D 2007 Impact of the Internet on consumer information search behavior in the United States *Journal of Media Business Studies* **4** (2) 27-39
- [12] Chevalier J A, Mayzlin D 2006 The effect of word of mouth on sales: online book reviews *Journal of Marketing Research* **43**(3) 345-54
- [13] Cheunga C M K, Thadanib D R 2012 The impact of electronic word-of-mouth communication: a literature analysis and integrative model *Decision Support Systems* **54**(1) 461-70
- [14] Lim B C, Chung C M Y 2014 Word-of-mouth: the use of source expertise in the evaluation of familiar and unfamiliar brands *Asia Pacific Journal of Marketing and Logistics* **26**(1) 39-53
- [15] Kimura M, Saito K, Nakano R, Motoda H 2010 Extracting influential nodes on a social network for information diffusion *Data Mining and Knowledge Discovery* **20**(1) 70-97
- [16] Wu H, Ying S, Jia X, Zhang L, Chen X 2013 Robustness analysis of social network based on a dynamic model *Journal of Digital Information Management* **11**(6) 453-61
- [17] Chen B, Huang X 2013 Environment compatibility of services interaction and its reachable analysis *Computer Modelling and New Technologies* **17**(4) 162-7
- [18] Walker D, Myrick F 2006 Grounded Theory: an exploration of process and procedure *Qualitative Health Research* **16**(4) 547-59
- [19] Wolfswinkel J F, Furtmueller E, Wilderom C P M 2013 Using grounded theory as a method for rigorously reviewing literature *European Journal of Information Systems* **22**(1) 45-55
- [20] Larossa R 2005 Grounded Theory methods and qualitative family research *Journal of Marriage and Family* **67**(4, 837-57

| Authors | |
|---|--|
|  | <p>Fei Meng, born in January, 1980, Yining, Xinjiang Province, China</p> <p>Current position, grades: the assistant professor of Zhejiang Police College, China. University studies: Ph. D degree in Information Science from Nanjing University in China. Scientific interest: e-commerce and information service. Publications: more than 20 papers. Experience: teaching experience of 3 years, 2 scientific research projects.</p> |
|  | <p>Jianliang Wei, born in May, 1980, Hangzhou, Zhejiang Province, China</p> <p>Current position, grades: the associate professor of Zhejiang Gongshang University, China. University studies: Ph. D degree in Information Management Engineering from Nanjing University in China. Scientific interest: e-commerce and information service. Publications: more than 30 papers. Experience: teaching experience of 5 years, 5 scientific research projects.</p> |

Service and revenue sharing strategies in a dual-channel supply chain with fairness concerns

Ming-xing Xu^{1, 2*}, Yue Yu¹, Yong-shi Hu³

¹*School of Economics & Management Fuzhou University, Fuzhou City, Fujian Province, China*

²*Concord University College of Fujian Normal University, Fuzhou City, Fujian Province, China*

³*Department of Traffic and Transportation, Fujian University of Technology, Fuzhou City, Fujian Province, China*

Received 26 May 2014, www.tsi.lv

Abstract

This paper incorporates the concept of fairness in a dual-channel supply chain to examine the effect of fairness concerns on the supply chain partners' service and revenue-sharing strategies in three different scenarios: only the retailer is concerned about fairness, only the manufacturer is concerned about fairness, and both parties are concerned about fairness. Though applying the equilibrium analysis, the results show that (1) Fairness concerns strongly influence the manufacturer's and the retailer's decision-making and utility. (2) The revenue sharing ratio increases with the strengthening of channel members' fairness concerns. (3) If only the retailer is concerned about fairness, the retailer's service is unaffected by his fairness concerns. (4) There exists a Pareto improvement for channel members' utility when the manufacturer without fairness concern becomes fair-minded.

Keywords: fairness concerns, dual-channel supply chain, service level, revenue sharing

1 Introduction

With rapid development and wide use of the Internet and related information technology, more and more consumers are accustomed to shopping online, many well-known manufacturers in a variety of industries such as Apple, HP, Estee Lauder, Nike, Lenovo, etc., have already redesigned their sales channel structures by engaging in direct online sale in order to meet different customer requirements that cannot be met by the bricks-and-mortar retail channel. Manufactures that adopt the online direct channel can remove the intermediary, increase the potential market demand and improve the efficiency of supply chain. It also can make high profits by directly controlling the distribution and price [1]. Meanwhile, consumers prefer hybrid channel, they can choose freely in the online direct channel and retail channel according to their preference [2]. However, a side effect of this trend is that the retailers, manufacturer's traditional retailer partners may feel disenfranchised and thus tend to resist the direct channel initiative because they perceive that the direct channel is bound to cannibalize their market shares [3, 4].

To mitigate this "channel conflict", some manufacturers use consistent pricing scheme (e.g., ZARA, Apple, etc.) by selling the products in both channels at the same price [5], and the traditional retailers continuously improve their retail service to survive, thus the service level in dual-channel is higher than it in single channel [6, 7]. Retail services have significant effects on customers' channel choice, demand and loyalty [4, 8, 9], it also

strongly influences the manufacturer and the retailer's pricing strategies and profit [10-12]. Therefore, retail services play a strategic role in a dual-channel supply chain, which is the research subjects of this paper. Both the manufacturer and the retailer can be always benefited no matter who provides service in the Stackelberg game [13]. Although manufacturers may provide consumers with services such as product information and consultations, graphic pictures and sound, traditional retailers play a key role in providing diverse transactional and post-sales services such as personal inspections, additional expertise, advice and technical support, as well as the easy and prompt replacement or refund for defective parts [14]. Therefore, in most cases, manufacturer free-rides retailer's sales effort, the free riding effect reduces retailer's desired effort level, and thus undermines the manufacturer's profit and the overall supply chain performance [15]. To motivate the retailer to improve service level and thus enhance the performance of the supply chain, a supply chain coordination mechanism with service cost sharing between the manufacturer and the retailer should be established [16].

Most of the previous studies assume that channel members are rational-economic men, who always try to maximize their own profits. However, abundant evidence show that decision makers not only care about their own profits, but also the profit difference between the two sides, meaning they concern about fairness. "There is a significant incidence of cases in which firms, like individuals, are motivated by concerns of fairness" in business relationships including channel relationships

* *Corresponding author* e-mail: chn_ming@qq.com

[17]. Each member's own judgment of fairness may play a role on the contract's channel performance [18]. Researches in economics and marketing have demonstrated that fairness is of significant importance in developing and maintaining channel relationships [19-22]. Many empirical or experimental studies illustrate that channel member would sacrifice their own margins for the benefit of their counterpart because of fairness concerns [23, 24]. The channel members should attach importance to fairness concerns, which may bring benefits to them [25].

Therefore, it is critical to study retailers' service strategy when channel members have fairness concerns. This paper considers a manufacturer that sells a product through his direct online channel and a traditional retail store during a sales season. In order to avoid channel conflict and improve service efficiency, the manufacturer outsources his direct channel's service to retailer and gives the retailer a proper percentage of the revenue generated by the dual channel. This paper analyses the retail service and revenue sharing decisions in the following four cases: (1) neither channel member has fairness concerns; (2) only the retailer has fairness concerns; (3) only the manufacturer has fairness concerns; (4) both channel members have fairness concerns.

The rest of the paper is organized as follows. Section 2 describe the notation and formulates the decision models for the manufacturer and the retailer. Section 3 considers the basic situation that neither channel members have fairness concerns. Section 4 discusses the case when channel members have fairness concerns. Section 5 reports the results of numerical experiments carried out to investigate the impacts of fairness concerns on the manufacturer and retailer's decisions and revenue (or utility). Section 6 gives the concluding remarks and directions for future research.

2 The model

A manufacturer who produces a single product at a unit cost c and distributes it through his wholly owned direct online channel and independent retailer channel at a price p in a perfectly competitive market, p is determined by the market. The manufacturer outsources the direct channel's services to the retailer so that the retailer provides services for the dual channel. The level of the services is denoted by s , determining by the retailer, which includes immediate customer support, presale advice, in-store advertising and promotions, technical and shopping assistance, post-sale service, channel assembly services, etc. The manufacturer and the retailer share the total revenue of the dual-channel supply chain, and the manufacturer prorates the total revenue of the dual-channel.

Let D_r and D_d denote the market demand from the retail channel and the direct channel respectively, they mainly depend on the level of service devoted by the retailer. We adopt a linear demand function which has been widely used

in Yue and Liu [26], Huang and Swaminathan [27], Dan et al. [12], it was described as follow:

$$D_d = \theta a - p + \lambda_1 s, \quad (1)$$

$$D_r = (1-\theta)a - p + \lambda_2 s. \quad (2)$$

The parameter a represents the base level of demand rate, supposing $a - 2p > 0$. θ ($0 < \theta < 1$) represents the degree of customer loyalty to the direct channel. The parameter λ_1 and λ_2 ($\lambda_i > 0; i=1,2$) are the coefficient of service elasticity of D_d and D_r .

With the above notation, the total revenue of the supply chain is:

$$\pi_T = (p-c)(D_d + D_r), \quad (3)$$

the retailer's profit is:

$$\pi_r = k + \varphi\pi_T - c(v), \quad (4)$$

and the manufacturer's profit is:

$$\pi_m = (1-\varphi)\pi_T - k, \quad (5)$$

where k represents the case when the manufacturer pays the retailer a fixed fee to maintain channel relationship which is irrelevant to the level of service, φ ($0 < \varphi < 1$) acts as the proportion that the retailer gets of the revenue which is decided by the manufacturer, reflecting the significance of the retailer's service. It will rise with the increasing importance of the retailer's service and decline when service makes less impact on demand. $c(s)$ represents the cost of retail service, which has the properties of $dc(s)/ds > 0$ and $d^2c(s)/ds^2 > 0$, and its most common form is: $c(s) = \eta s^2 / 2$.

The paper considers a decentralized dual-channel supply chain under the Stackelberg game; both the manufacturer and the retailer make their own decisions to maximize the utility, where the manufacturer takes the leader role and sets the revenue sharing ratio; in response to the manufacturer's decision, the retailer selects service level, which affects the market demand.

3 Model without fairness concerns

In the traditional decentralized setting, all parties maximize their own monetary payoffs without considering any fairness issues. Under the Manufacturer Stackelberg, the manufacturer takes the retailer's reaction function into consideration for her revenue sharing decisions. Using backward induction, the retailer's best response for any given revenue sharing ratio $\varphi \in (0,1)$, the retailer chooses $s > 0$ to maximize his profits, and his decision problem can be described as below:

$$\max_s \pi_r = k + \varphi(p-c)[a - 2p + (\lambda_1 + \lambda_2)s] - \frac{\eta s^2}{2}. \quad (6)$$

Taking the first-order partial derivatives of π_r with respect to s , and letting the derivatives be zero, that is:

$$\frac{\partial \pi_r}{\partial s} = -\eta s + \varphi(\lambda_1 + \lambda_2)(p - c) = 0.$$

Taking the second-order partial derivatives of π_r with respect to s , we have $\frac{\partial^2 \pi_r}{\partial s^2} = -\eta$.

Since the second order condition for π_r with respect to s is negative, the retailer's best response to the φ is:

$$s = \frac{\varphi(\lambda_1 + \lambda_2)(p - c)}{\eta}. \tag{7}$$

This function implies that the retailer's service depend on the given proportion of revenue that dictated by the manufacturer.

The manufacturer anticipates the retailer's best response and incorporates it into his optimization problem, which is given by $\pi_m^* = \max \pi_m(\varphi, s(\varphi))$, which can be written:

$$\max_{\varphi} \pi_m = (1 - \varphi)(p - c)[a - 2p + (\lambda_1 + \lambda_2)s] - k. \tag{8}$$

Substituting Equation (7) into Equation (8), then taking the first-order and second-order partial derivatives of π_m with respect to φ :

$$\frac{\partial \pi_m}{\partial \varphi} = -\frac{(\lambda_1 + \lambda_2)^2(p - c)^2 \varphi + (\lambda_1 + \lambda_2)^2(p - c)^2 - \eta(a - 2p)(p - c)}{\eta},$$

$$\frac{\partial^2 \pi_m}{\partial \varphi^2} = -\frac{(\lambda_1 + \lambda_2)^2(p - c)^2}{\eta}.$$

since $\frac{\partial^2 \pi_m}{\partial \varphi^2} < 0$, it implies that there exists a unique optimal revenue sharing ratio φ^* :

$$\varphi^* = \frac{1}{2} - \frac{\eta(a - 2p)}{2(\lambda_1 + \lambda_2)^2(p - c)}. \tag{9}$$

Substituting Equation (9) into Equation (7), the retailer's best service strategy is given by:

$$s^* = \frac{(\lambda_1 + \lambda_2)(p - c)}{2\eta} - \frac{a - 2p}{2(\lambda_1 + \lambda_2)} \tag{10}$$

Corollary1. The proportion of revenue sharing φ and the level of service s were determined by the significance of service in the dual channel, and they will rise with the increasing importance of the retailer's service and decline with the diminishing importance of the retail service in dual-channel.

Proof:

$$\frac{\partial \varphi}{\partial (\lambda_1 + \lambda_2)} = \frac{\eta(a - 2p)}{(\lambda_1 + \lambda_2)^3(p - c)} > 0,$$

$$\frac{\partial s}{\partial (\lambda_1 + \lambda_2)} = \frac{(p - c)}{2\eta} + \frac{a - 2p}{2(\lambda_1 + \lambda_2)^2} > 0.$$

4 Model with fairness concerns

4.1 ONLY THE RETAILER IS CONCERNED ABOUT FAIRNESS

This section considers the case that the retailer has fairness concerns, but the manufacturer does not. The manufacturer maximizes his own profit whereas the retailer maximizes his utility depending on the profits of both members. Cui et al. [22] capture fairness in the members' objectives through the following utility function:

$$U_i = \pi_i + f_i, \quad i \in \{m, r\}, \tag{11}$$

where, U_i stands for the channel member's utility, while π_i represents the monetary profit and f_i denotes channel member's disutility due to the unfairness or inequity. This means that channel member's utility consist of the monetary payoff and the utility of fairness. For the sake of simplification, the paper adopts a modified model [28], which was firstly constructed by Charness and Rabin [29]. Which is:

$$U_r^r = \pi_r - \alpha(\pi_m - \pi_r). \tag{12}$$

In the text, the subscript "r", "m" means the parameters correspond to the retailer and the manufacturer, while the superscript "r", "m", "b" means the parameters are corresponding only when the retailer has fairness concern or the manufacturer has fairness concern, or both the retailer and the manufacturer have fairness concerns. α represents the retailer's sensitivity on fairness, α measures the retailer's utility(or disutility) of earning more (less) than the manufacturer, that means if the retailer's monetary payoff is higher (or lower) than the manufacturer's, an advantageous (or disadvantageous) equality occurs, which will result in a utility (or disutility) for the retailer in the amount of α per-unit difference in the two payoffs. When $\alpha = 0$, it means the retailer is fairness neutral, he does not concern fairness, and he only cares about his monetary payoff. When $\alpha = 1$, the retailer concerns fairness extremely, such that the retailer would give up some monetary payoff to move in the direction of more equitable outcomes. Therefore, when the retailer is concerned about fairness, his optimization problem is given below:

$$\begin{aligned} \max U_r^r &= \max_s \pi_r - \alpha(\pi_m - \pi_r) = \\ \max_s (1 + 2\alpha)k + [(1 + 2\alpha)\varphi - \alpha] \cdot \\ [a - 2p + (\lambda_1 + \lambda_2)s](p - c) - \frac{(1 + \alpha)\eta}{2} s^2. \end{aligned} \tag{13}$$

The retailer's best service strategy s^{r*} satisfies:

$$\begin{aligned} \frac{\partial U_r^r}{\partial s^r} \Big|_{s^r=s^{r*}} = \\ [(1 + 2\alpha)\varphi - \alpha](\lambda_1 + \lambda_2)(p - c) - (1 + \alpha)\eta s^{r*} = 0 \end{aligned} \tag{14}$$

and $\frac{\partial^2 U_r^r}{\partial s^{r2}} = -(1 + \alpha)\eta < 0$.

Thus, the retailer's best response to the given φ is:

$$s^r = \frac{[(1 + 2\alpha)\varphi - \alpha](\lambda_1 + \lambda_2)(p - c)}{(1 + \alpha)\eta}. \tag{15}$$

Substituting Equation (15) into Equation (8), solving the optimal decision for the manufacturer, the manufacturer's optimal decisions as follows:

$$\varphi^{*} = \frac{1 + 3\alpha}{2(1 + 2\alpha)} - \frac{\eta(1 + \alpha)(a - 2p)}{2(1 + 2\alpha)(\lambda_1 + \lambda_2)^2(p - c)}. \tag{16}$$

Substituting Equation (16) into Equation (15), the retailer's best service strategy is given below:

$$s^{r*} = \frac{(\lambda_1 + \lambda_2)(p - c)}{2\eta} - \frac{a - 2p}{2(\lambda_1 + \lambda_2)}. \tag{17}$$

Corollary 2. When only the retailer is concerned about fairness, $\varphi^{*} > \varphi^*$, $s^{r*} = s^*$; φ^{*} rise with the increasing of α .

Proof:

$$\varphi^{*} - \varphi^* > 0, \frac{\partial \varphi^{*}}{\partial \alpha} = \frac{1}{2(1 + 2\alpha)^2} + \frac{\eta(a - 2p)}{2(\lambda_1 + \lambda_2)^2(p - c)} > 0.$$

Corollary 2 illustrates that if only the retailer cares about fairness, the manufacturer has to consider the retailer's fairness feelings, he will distribute more revenue to the retailer to satisfy the retailer's fairness concerns. If the retailer has stronger fairness concerns, he will obtain a higher proportion of revenue sharing from the manufacturer. However, since the manufacturer has no fairness concerns, the retailer's best service level remains unchanged and stays the same level as the level of neither channel members have fairness concerns.

Corollary 3. When only the retailer has fairness concern, the manufacturer's profits maybe shrinkage because the market demand which is influenced by the retailer's services has not increased, leading to the total supply chain revenue remain unchanged. With the retailer's revenue increased, the manufacturer's revenue will reduce and the retailer's utility will increase as well.

Since the expressions of π_m^* , π_m^{r*} , π_r^* and U_r^{r*} are complicated, in section 5, the paper will present numerical examples to compare π_m^* with π_m^{r*} and compare π_r^* with U_r^{r*} to prove the corollary 3.

4.2 ONLY THE MANUFACTURER IS CONCERNED ABOUT FAIRNESS

When only the manufacturer cares about fairness, which is similar to section 4.1, the manufacturer's utility function is:

$$\begin{aligned} U_m^m &= \pi_m - \beta(\pi_r - \pi_m) = [1 + \beta - (1 + 2\beta)\varphi] \\ [a - 2p + (\lambda_1 + \lambda_2)s](p - c) - (1 + 2\beta)k + \frac{\beta\eta}{2} s^2, \end{aligned} \tag{18}$$

β represents the manufacturer's sensitivity on fairness, when $\alpha = 0$, it means the manufacturer has no fairness concerns; when $\alpha = 1$, he concerns fairness extremely, such that the manufacturer would give up some monetary payoff to get a fair shake.

Because the retailer is indifferent to the fairness, his utility function is just the same as the profits function, that is:

$$\begin{aligned} U_r &= \pi_r = \pi^r - c(s) = k + \varphi(\pi_r + \pi_r) - \frac{\eta s^2}{2} \\ &= k + \varphi(p - c)[a - 2p + (\lambda_1 + \lambda_2)s] - \frac{\eta s^2}{2}. \end{aligned} \tag{19}$$

With similarity to the solution process of section 4.1, the retailer's response function is

$$s^m = \frac{\varphi(\lambda_1 + \lambda_2)(p - c)}{\eta} \tag{20}$$

Substituting Equation (20) into Equation (18), solving the optimal decision for the manufacturer:

$$\varphi^{m*} = \frac{(1 + \beta)(\lambda_1 + \lambda_2)^2(p - c) - \eta(1 + 2\beta)(a - 2p)}{(\lambda_1 + \lambda_2)^2(p - c)[2(1 + 2\beta) - \beta(p - c)]}. \tag{21}$$

Substituting Equation (21) into Equation (20), the retailer's best service strategy is:

$$s^{m*} = \frac{(1 + \beta)(\lambda_1 + \lambda_2)^2(p - c) - \eta(1 + 2\beta)(a - 2p)}{\eta(\lambda_1 + \lambda_2)[2(1 + 2\beta) - \beta(p - c)]}. \tag{22}$$

Corollary 4. When only the manufacturer has fairness concern, (a) if $p > p^{m_2}$, then $\varphi^{m*} > \varphi^*$, $s^{m*} > s^*$, $\frac{\partial \varphi^{m*}}{\partial \beta} > 0$, $\frac{\partial s^{m*}}{\partial \beta} > 0$; (b) if $p^{m_1} < p < p^{m_2}$, then $\varphi^{m*} > \varphi^*$, $s^{m*} > s^*$, $\frac{\partial \varphi^{m*}}{\partial \beta} < 0$, $\frac{\partial s^{m*}}{\partial \beta} < 0$; (c) if $p < p^{m_1}$,

then $\varphi^{m^*} < \varphi^*$, $s^{m^*} < s^*$, $\frac{\partial \varphi^{m^*}}{\partial \beta} < 0$, $\frac{\partial s^{m^*}}{\partial \beta} < 0$, where

$$p_{1}^m = \frac{(c+2)(\lambda_1 + \lambda_2)^2 + a\eta}{(\lambda_1 + \lambda_2)^2 + 2\eta},$$

$$p_{2}^m = \frac{[c+2(1+2\beta)](\lambda_1 + \lambda_2)^2 + a\eta}{(\lambda_1 + \lambda_2)^2 + 2\eta}.$$

Proof:

$$\varphi^{m^*} - \varphi^* = \frac{\beta(\lambda_1 + \lambda_2)^2(p-c-2) - \beta\eta(a-2p)}{2(\lambda_1 + \lambda_2)^2[2(1+2\beta) - \beta(p-c)]},$$

$$s^{m^*} - s^* = \frac{\beta(\lambda_1 + \lambda_2)^2(p-c-2) - \beta\eta(a-2p)}{2\eta(\lambda_1 + \lambda_2)[2(1+2\beta) - \beta(p-c)]},$$

when

$$\frac{\beta(\lambda_1 + \lambda_2)^2(p-c-2) - \beta\eta(a-2p)}{2(1+2\beta) - \beta(p-c)} > 0, \varphi^{m^*} - \varphi^* > 0,$$

$$s^{m^*} - s^* > 0, \text{ then } p > \frac{(c+2)(\lambda_1 + \lambda_2)^2 + a\eta}{(\lambda_1 + \lambda_2)^2 + 2\eta} = p_{1}^m.$$

$$\frac{\partial \varphi^{m^*}}{\partial \beta} = \frac{-2(1+2\beta)(\lambda_1 + \lambda_2)^2 + (\lambda_1 + \lambda_2)^2(p-c) - \eta(a-2p)}{(\lambda_1 + \lambda_2)^2[2(1+2\beta) - \beta(p-c)]^2},$$

$$\frac{\partial s^{m^*}}{\partial \beta} = \frac{-2(1+2\beta)(\lambda_1 + \lambda_2)^2 + (\lambda_1 + \lambda_2)^2(p-c) - \eta(a-2p)}{\eta(\lambda_1 + \lambda_2)[2(1+2\beta) - \beta(p-c)]^2},$$

$$\varphi^{b^*} = \frac{[\alpha(1+\alpha)(1+2\beta) + (1+\alpha)(1+2\alpha)(1+\beta) - \alpha\beta(1+2\alpha)(p-c)](\lambda_1 + \lambda_2)^2(p-c) - \eta(1+\alpha)^2(1+2\beta)(a-2p)}{(1+2\alpha)(\lambda_1 + \lambda_2)^2(p-c)[2(1+\alpha)(1+2\beta) - (1+2\alpha)\beta(p-c)]}, \quad (25)$$

$$s^{b^*} = \frac{(1+\alpha+\beta)(\lambda_1 + \lambda_2)^2(p-c) - \eta(1+\alpha)(1+2\beta)(a-2p)}{\eta(\lambda_1 + \lambda_2)[2(1+\alpha)(1+2\beta) - (1+2\alpha)\beta(p-c)]}. \quad (26)$$

Corollary 5. When both channel members care about fairness, the impact of φ on s is greater than it when both channel members have no fairness concerns.

Proof:

$$\frac{\partial s}{\partial \varphi} = \frac{(\lambda_1 + \lambda_2)(p-c)}{(1+\alpha)\eta},$$

$$\frac{\partial s^r}{\partial \varphi} = \frac{(1+2\alpha)(\lambda_1 + \lambda_2)(p-c)}{(1+\alpha)\eta} > \frac{\partial s}{\partial \varphi},$$

$$\frac{\partial s^m}{\partial \varphi} = \frac{(\lambda_1 + \lambda_2)(p-c)}{\eta} > \frac{\partial s}{\partial \varphi},$$

$$\frac{\partial s^{b^*}}{\partial \varphi} = \frac{(1+2\alpha)(\lambda_1 + \lambda_2)(p-c)}{(1+\alpha)\eta} > \frac{\partial s}{\partial \varphi}.$$

In this section, due to the fact that the expressions of φ^{b^*} and s^{b^*} are explicated, it is not easy to observe their variation which affected by α and β . It is also difficult

when

$$-2(1+2\beta)(\lambda_1 + \lambda_2)^2 + (\lambda_1 + \lambda_2)^2(p-c) - \eta(a-2p) > 0,$$

then

$$\frac{\partial \varphi^{m^*}}{\partial \beta} > 0, \frac{\partial s^{m^*}}{\partial \beta} > 0,$$

note that

$$p > \frac{[c+2(1+2\beta)](\lambda_1 + \lambda_2)^2 + a\eta}{(\lambda_1 + \lambda_2)^2 + 2\eta} = p_{2}^m.$$

4.3 BOTH THE MANUFACTURER AND THE RETAILER ARE CONCERNED ABOUT FAIRNESS

When both channel members have fairness concerns, they no longer strive to maximize only their monetary payoff. Instead, their objective is to maximize their utility, which define as

$$\max U_r^b = \max_s \pi_r - \alpha(\pi_m - \pi_r) \quad (23)$$

$$\max U_m^b = \max_\varphi \pi_m - \beta(\pi_r - \pi_m) \quad (24)$$

As the manufacturer is the Stackelberg leader, which is similar to the solution procedure of section 4.1, the optimal strategies for both channel members are given by

to compare φ^{b^*} (or s^{b^*}) with φ^* (or s^*), the paper will analyse by numerical example.

5 Numerical analysis

According to the above theoretical analysis, the result shows that the fairness concern has an important impact on the optimal decisions and utilities of the channel members as well as the profits of the dual-channel supply chain. In this section, by conducting several numerical examples, some related issues will be illustrated.

Under the constraints of $0 < \varphi < 1$ and $s > 0$, assuming $a=100$, $p=5$, $c=1$, $\lambda_1=2$, $\lambda_2=3$, $\eta=0.5$, $k=20$, $\alpha, \beta \in [0,1]$.

When neither the manufacturer nor the retailer cares about fairness, the retail service level $s^* = 11$, the revenue sharing ratio $\varphi^* = 0.275$, the channel members' profits are $\pi_m^* = 400.5$ and $\pi_r^* = 149.25$, the total profits

$\pi_T^* = 549.75$. In this case, the dual-channel supply chain's total utility is equal to its total profits, where $U_T^* = \pi_T^* = 549.75$.

5.1 ONLY THE RETAILER IS CONCERNED ABOUT FAIRNESS

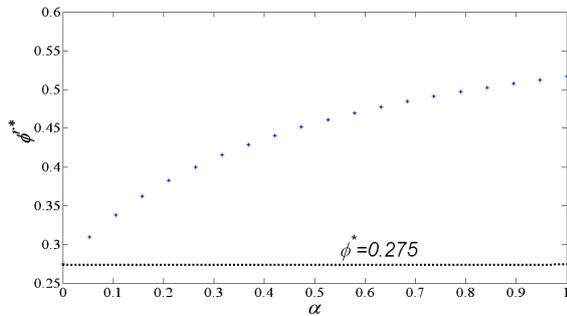


FIGURE 1 The impact of α on ϕ

Figure 1 shows the impact of the retail service on the manufacturer's revenue sharing decision, $\phi^* \in [0.275, 0.517]$, it is higher than the case where both channel members are completely rational economic man, this is because the retailer cares about fairness and the manufacturer needs to consider the retailer's preference by giving more revenue to the retailer. So ϕ^* augments with the increase of α , but the growth rate is decline, which reflects that the manufacturer is sensitive to the retailer's fairness concerns at first, however, as the retailer's fairness concern enhances, the manufacturer is gradually accustomed to the retailer's concerns as such the ϕ will grow slowly. As the retailer gets more revenue from the manufacturer, the retail service remains unchanged and stays the same as the situation that neither them has fairness concerns. As a result, the market demand affected by the retailer's services has not increased, which leads to the supply chain's total revenue keeps unchanged. Then, with the growing of ϕ^* , the manufacturer's revenue is decreasing as the retailer's revenue is increasing.

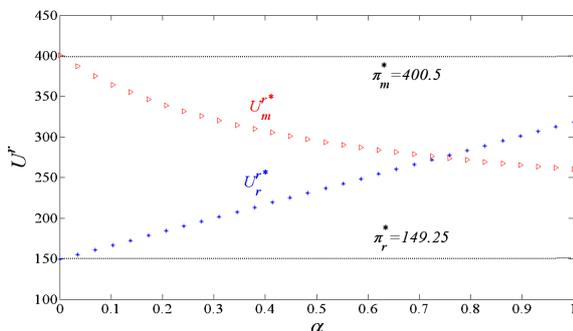


FIGURE 2 The impact of α on the channel members' utilities

As shown in Figure 2, the manufacturer's utility $U_m^* \in [400.5, 260.33]$ is equal to his profit, reducing with the α increasing; the retailer's utility $U_r^* \in [149.25, 318.5]$ exceeds his profit, increasing with

the α increasing. The total utility of the dual-channel supply chain $U_T^* \in [549.75, 578.83]$ exceeds the total revenue of the supply chain, where $U_T^* > \pi_T^*$. This scenario is a zero-sum game, and it maybe prejudicial to the stability of the supply chain.

5.2 ONLY THE MANUFACTURER IS CONCERNED ABOUT FAIRNESS

Only when the manufacturer has fairness concern, the revenue sharing ratio (ϕ^m) is rising with the enhancement of manufacturer's fairness concerns (β), so the retailer will get more revenue from the manufacturer, as shown in Figure 3, where the revenue sharing ratio $\phi^{m*} \in [0.275, 0.325]$.

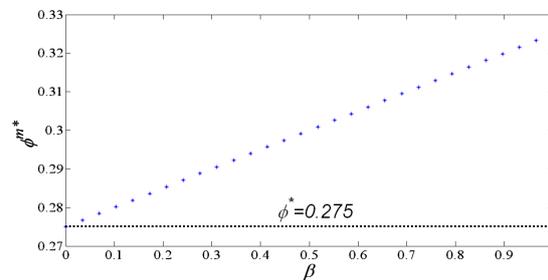


FIGURE 3 The impact of β on ϕ

Comparing to the case where only the retailer is concerned about fairness, the manufacturer's fairness concern has less impact on the revenue sharing ratio than the retailer's fairness concern, so $\phi^* < \phi^{m*} < \phi^{**}$. To satisfy the manufacturer's fairness concern and increase the market demand, with the increase of β , the retailer needs to improve his retail service (s^m), as shown in Figure 4, the retailer's service $s^{m*} \in [11, 13]$, $s^{m*} > s^{r*} = s^*$.

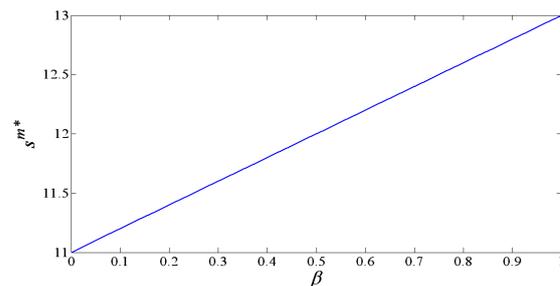


FIGURE 4 The impact of β on the retailer's service

Retail service will promote sales, bringing about an increase in total revenue of the dual-channel supply chain. From Fig.5, with the increase of β , both the manufacturer's and the retailer's utility increase, the retailer's utility ($U_r^{m*} \in [149.25, 179.25]$) is equal to his profit, and the manufacturer's utility $U_m^{m*} \in [400.5, 617.75]$ is greater than his profit, so the supply chain's total utility $U_T^m \in [549.75, 797]$ exceeds its

total profits, where $U_T^{m*} > \pi_T^{m*}$, and it exceeds U_T^* and U_T^{r*} .

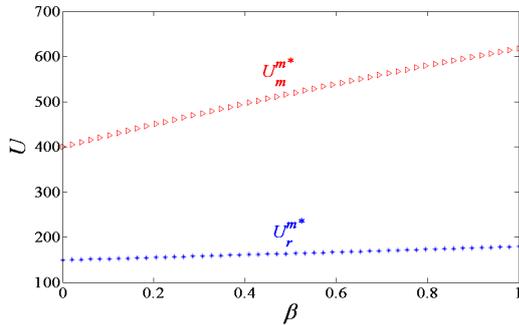


FIGURE 5 The impact of β on the channel members' utilities

It reflects that the manufacturer has fairness concern is beneficial to both sides, which leads to a “win-win” situation. Besides, it also means that it is a Pareto improvement when the manufacturer converts from complete rationality to fairness concern.

5.3 BOTH THE MANUFACTURER AND THE RETAILER ARE CONCERNED ABOUT FAIRNESS

Since the manufacturer and the retailer are fair-minded, the revenue sharing ratio $\varphi^{b*} \in [0.275, 1)$, a higher proportion than φ^* , φ^{r*} and φ^{m*} . As shown in Figure 6, the revenue sharing ratio φ^{b*} rises with α and β , when α and β below a certain threshold, that is $\alpha, \beta < 0.6$. α has a strong effect on φ^{b*} than β , and φ^{b*} increases slightly with the rise of β . When both channel members have strong sense of fairness, where $\alpha, \beta > 0.6$, φ^{b*} increases sharply with the rise of α or β .

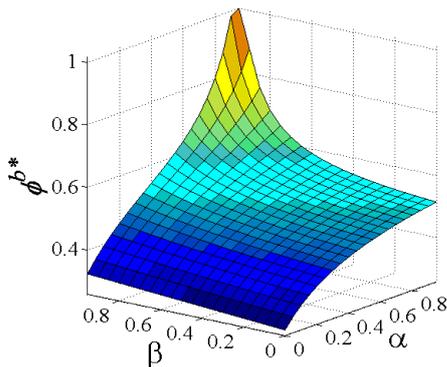


FIGURE 6 The impact of α and β on φ

This demonstrates that when both channel members are concerned about fairness, the manufacturer will take full account of the retailer's fairness concern; the retailer's revenue relies heavily on his fairness concern, when the retailer is obviously fair-minded, he will get much revenue from the manufacturer, and otherwise he will get less when he cares little about fairness.

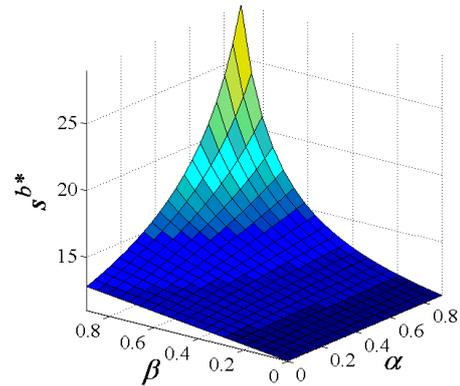


FIGURE 7 The impact of α and β on the retailer's service

Just as the variation of the revenue sharing ratio, the retailer's service is affected by both channel members' fairness concerns; it will increase with α and β , as shown in Figure 7, $s^{b*} \in [11, 40)$. When $\beta < 0.6$, the retailer lacks the inclination to improve service, and the retail service is not nearly affected by his fairness concern, so the service level remains largely unchanged when $\alpha \in [0, 1]$. If $\alpha < 0.6$, even though the manufacturer is very concerned about fairness, the retail service increases slowly with the rise of β . When $\alpha, \beta > 0.6$, the retail service is increasing sharply with the rise of α or β , and it is obviously higher than any of the above case.

As shown in Figures 8 and 9, the manufacturer's utility and the retailer's utility are affected by their fairness concerns. The manufacturer's utility (U_m^{b*}) decreases with the rise of α and β , $U_m^{b*} \in [-768.84, 605.96]$, and it mainly affected by β .

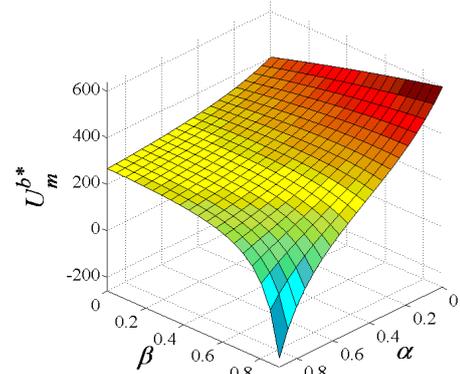


FIGURE 8 The impact of α and β on the manufacturer's utility

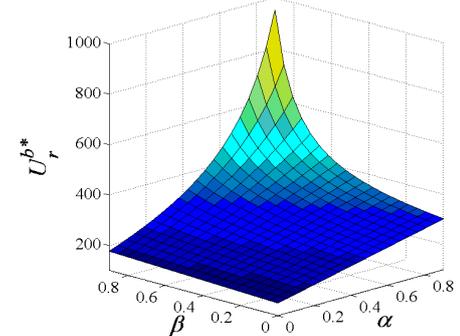


FIGURE 9 The impact of α and β on the retailer's utility

When β is low, $U_m^{b^*}$ decreases slowly with the rise of α ; when β is high, $U_m^{b^*}$ decreases sharply with the rise of α . $U_m^{b^*}$ increases with the rise of β only when α is quite low. Contrary to $U_m^{b^*}$, $U_r^{b^*}$ increases with the rise of α and β , when both channel members are strongly concerned about fairness, $U_r^{b^*}$ will increase quickly with the rise of α or β . This implies that channel members' strong fairness concerns are beneficial to the retailer but disadvantageous to the manufacturer. Most firms in reality care about the fair outcomes in business relations, and the impact of fairness concerns can be expected to be most significant for members' decision making and utility, so this scenario can well reflect the real-world conditions. It enlightens us that the retailer should lower his fairness concern and the manufacturer should enhance his fairness concern, which will be good for both channel members and the stability of the dual-channel supply chain.

6 Conclusions

This paper takes an initial step to incorporate fairness concerns of channel members into the study of revenue sharing and cooperative service in a dual-channel supply chain, and examine the effect of fairness on the supply chain partners' strategies and utility, as well as compare the situations when neither channel members have fairness concern. The study finds that only when the retailer is concerned about fairness, he will get more revenue from the manufacturer and remain his retail service the same level when both sides are complete rationality, resulting in the increase of retailer's utility and the decrease of manufacturer's utility. Unlike the retailer's fairness concern, only when the manufacturer is fair-minded, the retailer will improve service level, which leads to the increase in the total revenue of the supply chain, bringing

about a good result for both channel members. When both the manufacturer and the retailer are strongly concerned about fairness, the retailer is likely to provide a high level of service and maximizes the total revenue of the supply chain, while the manufacturer divides a great part of the total revenue to the retailer and he gets little. In this case, the retailer has a great utility but the manufacturer has a negative one, as such the retailer is the great beneficiaries.

The channel members' fairness concerns may heighten the revenue sharing ratio, the service level and the equilibrium utility of the manufacturer as well as the whole channel. Interestingly, there exists a Pareto improvement of both the utilities of the manufacturer and the retailer when a manufacturer without fairness concern becomes fair-minded. These results suggest that both the manufacturer and the retailer should attach importance to fairness concerns, and the best arrangement is that the manufacturer care strongly about fairness and the retailer cares little about it, which is beneficial to both channel members.

Although the analysis has derived some useful insights, it is worth mentioning that this research can be extended in several directions. First, the paper assume the manufacturer to be the Stackelberg leader in this paper, but there are practical examples of large retailers (e.g., Wal-Mart, Amazon) as channel leaders. Thus it is an interesting direction that the retailer acts as the Stackelberg leader of the channel. Second, the paper does not examine how incomplete information may affect channel interactions in the presence of fairness concerns. For instance, a manufacturer may not know a retailer's service cost to estimate whether it has attained its equitable payoffs or not. Other extensions are worthy of studying including investigating channel members' decision making under fairness concerns when the market demand is uncertain or incorporating consumer fairness concern to explore their implications for channel coordination.

References

- [1] Louis S, Adel E A Anne C 1996 *Marketing channels (5th Ed)* New Jersey Prentice Hall
- [2] Takahashi K, Aoi T, Hirotsu D, Morikawa K 2011. Inventory control in a two-echelon dual-channel supply chain with setup of production and delivery *International Journal of Production Economics* **133**(1) 403-15
- [3] Chiang W K, Chhajed D, Hess J D 2003 Direct marketing, indirect Profits: A strategic analysis of dual-channel supply chain design *Management Science* **49**(1) 1-20
- [4] Yan R L, Pei Z 2009 Retail services and firm profit in a dual-channel market *Journal of Retailing and Consumer Services* **16**(4) 306-14
- [5] Liu B, Zhang R, Xiao M D 2010 Joint decision on production and pricing for online dual-channel supply chain system *Applied Mathematical Modelling* **34**(12) 4208-18
- [6] Littler D, Melanthiou, D 2006 Consumer perceptions of risk and uncertainty and the implications for behaviour towards innovative retail services: the case of internet banking *Journal of Retailing Consumer Service* **13**(6) 431-43
- [7] Kaya M 2006 Essays in supply chain contracting: Dual channel management with service competition and quality risk in outsourcing California Stanford University
- [8] Devaraj S, Fan M, Kohli R 2002 Antecedents of B2C channel satisfaction and preference Validating e-commerce metrics *Information Systems Research* **13**(3) 316-33
- [9] Rohm A J, Swaminathan V 2004 A typology of online shoppers based on shopping motivations *Journal of Business Research* **57**(7) 748-57
- [10] Dumrongsiri A, Fan A, Jain A, Moizadeh K 2008 A supply chain model with direct and retail channels *European Journal of Operational Research* **187**(3) 691-718
- [11] Chen Y G, Liu N 2010 Dual-channel supply chain competition strategy with service differentiation *Computer Integrated Manufacturing System* **16**(11) 2484-9
- [12] Dan B, Xu G Y, Liu C 2012 Pricing policies in a dual-channel supply chain with retail services *International Journal of Production Economics* **139**(1) 312-20
- [13] Xu M H, Yu G, Zhang H Q 2006 Game analysis in a supply chain with service provision *Journal of Management Science In China* **9**(2) 18-27
- [14] Chun S H, Rhee B D, Park S Y, Kim J C 2011 Emerging dual channel system and manufacturer's direct retail channel strategy *International Review of Economics and Finance* **20**(4) 812-25

- [15] Xing D H, Liu T M 2012 Sales effort free riding and coordination with price match and channel rebate *European Journal of Operational Research* **219**(2) 264-71
- [16] Luo M L, Li G, Sun LY 2011 Competition in a dual-channel supply chain with service spill-over effect *Journal of System & Management* **20**(6) 648-57
- [17] Kahneman D, Knetsch J L, Thaler R H 1986 Fairness and the assumptions of economics *Journal of Business* **59**(4) 285-300
- [18] Caliskan-Demirag O, Chen Y H, Li J B 2010 Channel coordination under fairness concerns and nonlinear demand *European Journal of Operational Research* **207**(3) 1321-6
- [19] Frazier G L 1993 Interorganizational exchange behavior in marketing channels: A broadened perspective *Journal of Marketing* **47**(4) 68-78
- [20] Anderson E, Weitz B 1992 The use of pledges to build and sustain commitment in distribution channels *Journal of Marketing Research* **29**(1) 18-34
- [21] Geyskens I, Steenkamp, J-B E M, Kumar N 1998 Generalizations about trust in marketing channel relationships using meta-analysis *International Journal of Research in Marketing* **15**(3) 223-48
- [22] Cui T H, Raju JS, Zhang ZJ 2007 Fairness and channel coordination. *Management Science* **53** (8) 1303-14
- [23] Kumar N 1996 The power of trust in manufacturer-retailer relationships *Harvard Business Review* **74** (6) 92-106
- [24] Loch C H, Wu Y, 2008 Social preferences and supply chain performance: an experimental study *Management Science* **54** (11) 6-38
- [25] Yang J, Xie J X, Deng X X, Xiong H C 2013 Cooperative advertising in a distribution channel with fairness concerns *European Journal of Operational Research* **227**(2) 401-7
- [26] Yue X, Liu J 2006 Demand forecast sharing in a dual-channel supply chain *European Journal of Operational Research* **174**(1) 646-67
- [27] Huang W, Swaminathan J M 2009 Introduction of a second channel: implications for pricing and profits *European Journal of Operational Research* **194** (2) 258-79
- [28] Du S F, Du C, Liang L, Liu T Z 2010 Supply chain coordination considering fairness concerns *Journal of Management Science in China* **13**(11) 41-8
- [29] Charness G, Rabin M 2002 Understanding social preferences with simple tests *The Quarterly Journal of Economics* **117**(3) 817-69

| Authors | |
|---|---|
|  | <p>Ming-xing Xu, born in February, 1982, Fuzhou City, Fujian Province, P. R. China</p> <p>Current position, grades: Ph. D. student at School of Economics & Management, Fuzhou University, Fuzhou, China. Teacher at Fujian Normal University, Fuzhou, China.</p> <p>University studies: Bachelor of Engineering at Fujian Normal University in China, Master of Management from Fuzhou University in China.</p> <p>Scientific interest: logistics and supply chain management, E-economics.</p> <p>Experience: 12 scientific research projects.</p> |
|  | <p>Yue Yu, born in August, 1986, Fuzhou City, Fujian Province, P. R. China</p> <p>Current position, grades: Ph. D. student at School of Economics & Management, Fuzhou University, Fuzhou, China.</p> <p>University studies: Bachelor of Science at Nanjing University of Science and Technology in China. Master of Management at Fuzhou University in China.</p> <p>Scientific interest: logistics, supply chain management.</p> <p>Experience: 4 scientific research projects.</p> |
|  | <p>Yong-shi Hu, born in April, 1982, Fuzhou City, Fujian Province, P. R. China</p> <p>Current position, grades: Teacher of Fujian University of Technology, Fuzhou, China.</p> <p>University studies: Bachelor of Management at Fuzhou University in China. Master of Management at Fuzhou University in China. Doctor of Management at Fuzhou University in China.</p> <p>Scientific interest: logistics, supply chain management.</p> <p>Experience: 12 scientific research projects.</p> |

Time-varying decision-making for hazardous chemical transportation in a complex transportation network

Yibo Du^{1, 2*}, Jin Zhang^{1, 2}

¹*School of Transportation and Logistics, Southwest Jiao Tong University, Chengdu, Sichuan, China, 610031*

²*National United Engineering Laboratory of Integrated and Intelligent Transportation, Chengdu, Sichuan, China, 610031*

Received 26 April 2014, www.tsi.lv

Abstract

The transit and storage of hazardous chemicals are harmful. A distributed decision model for hazardous chemicals is developed in this study, with the time window established, to improve the efficiency of transportation and storage. The route, mode, time, and volume of each demand can be determined by this model. The model minimizes the total transportation risk and cost. The model is divided into two parts, and the corresponding ant colony algorithm is designed and achieved. The feasibility and efficiency of the model are illustrated through a numerical example with eight transfer nodes, six origin–destination (OD) demands, and multiple transportation mode alternatives. The developed model provides an effective approach for hazardous chemical substance transportation.

Keywords: hazardous chemicals, transportation decision, nonlinear mixed integer programming model, complex transportation network, ant colony algorithm

1 Introduction

A hazardous chemical substance, regardless of its physical or biochemical feature, contains materials that are harmful to humans. Such substance poses high danger during transport. Industries that use hazardous chemical substances have elicited much attention in recent years because of the high risk, diversity, and high perniciousness of these substances. Route selection is generally considered to achieve efficiency and low cost during the transport of these substances. Meanwhile, most scholars consider the transport process a low-risk event.

Many studies on the transport of hazardous chemical substances have been conducted in recent years. Current studies generally adopt double-route planning to reduce the risk and cost of the carriage [3–7]. Ren Chang-xing selected a rational transit line by adjusting the weight of the side boundary [8]. Zhang Jin proposed a nonlinear network transport model based on integrative storage [9]. Ma Chang-xi built a road transportation route multi-objective decision model for hazardous chemical substances after considering the transportation risk, service time, and population [10, 11]. We Hang constructed a route preference model based on a time-varying network condition [12]. Zou Zong-feng established a transportation route for a hazardous chemical substance based on the condition of the mixing time window. Fank considered shipment distance, transit time, human risk, accident probability, and accident consequence in the route selection index [13]. Huang employed a geographical information system (GIS) and a genetic algorithm (GA) to evaluate path risk [14]. Liang

Qi-chao established an index that includes carriage risk, cost, and time and considered carriage cost the general objective [15]. Wu Feng considered accident risk, disaster, and remedy as key factors that influence security evaluations [16].

The results of most previous studies on the transportation problem can be mainly regarded from two aspects: risk and cost. However, these studies were merely under the condition of a sole mode of transportation or an established environmental implication. Only a few studies have considered transportation decision making under the conditions of multiple transportation modes.

In this study, a dynamic hazardous chemical transportation decision-making model was established in consideration of the complexity and time-varying feature of transport networks. This model optimizes control methods at low transportation risk and costs. The model not only considers the changed population conditions but also confirms the transportation route, shipping type, transit time, and cost.

2 Description of a hazardous chemical substance

The hazardous chemical substance transport network is the basis of the transport decision. In the transport network $G = (V, E, \Omega, \Gamma)$, V is the peak gather, E is the directed arc gather, and Γ is the arc cost gather. The urban and connected nodes of the transport network represent the network peak and directed arc, respectively. The hazardous chemical substance transport network is shown in Figure 1.

* *Corresponding author* e-mail: 6855051@qq.com

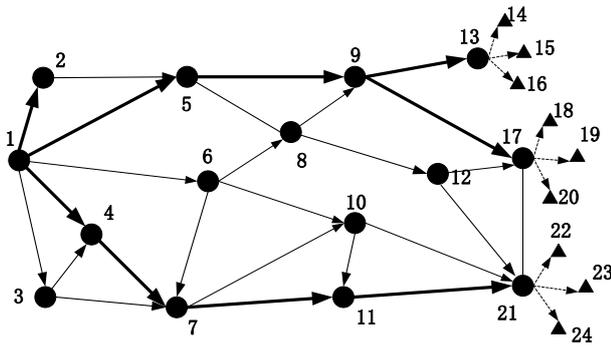


FIGURE 1 Modern logistics distribution network

Vertex set V includes the origin–destination, transfer, and goal nodes. The vertex of the total network can be divided into two stages. The first stage $V_1 = \{1, 2, \dots, n\}$ represents the origin–destination and transfer nodes in the transport network. The second stage $V_2 = \{n+1, n+2, \dots, n+m\}$ represents the goal node.

The directed arc includes gather E_1 and gather E_2 , where gather E_1 includes the various modes of transportation in the two nodes.

The arc parameters include the transportation capability (Ω) and generalized transport costs (Γ). Generalized transport costs mainly include the transportation risk (Γ_1), transportation cost (Γ_2), and transit time (Γ_3), where the transportation cost is a constant value. However, the value of transportation risk may differ in different periods. The initial networks are changed when multiple transportation modes exist between the two nodes in a transport network. If the two nodes have two transportation modes, the line should link the two nodes that correspond to different transportation modes. If a transit shipment task is involved in the transport network, the node must be splitted. The converted transport network is shown in Figure 2.

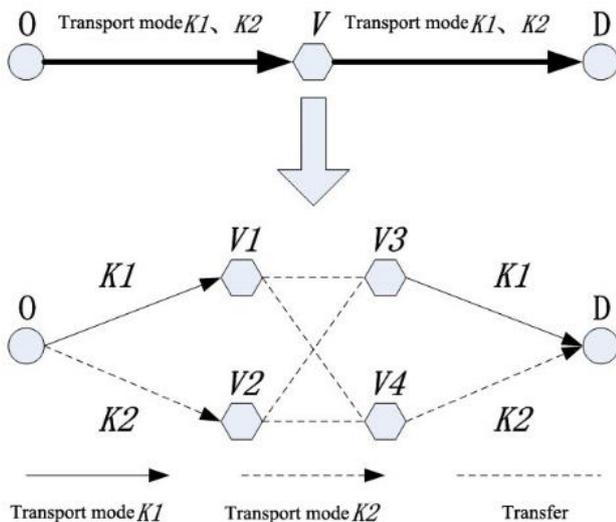


FIGURE 2 Transformation of a transport network

3 Model building

3.1 MODEL ASSUMPTION

Diverse flow directions exist in the hazardous chemical substance network; multiple transportation modes also exist among different ODs. We propose several assumptions to simplify the model:

- 1) We assume that the system considers only road traffic capacity during the hazardous chemical substance transportation process.
- 2) We assume that the transported hazardous chemical substances are of the same type; hazardous substances with the same flow direction cannot be segmented in the transportation process.
- 3) The random factor (e.g., operate miss, climate, natural environment) is disregarded.

3.2 PARAMETER SPECIFICATION

F is defined as the gather of the transport demand (OD), f is the variable of OD, and $f \in F$. $o(f)$ and $d(f)$ are the origin–destination and end point, respectively. Q_f is the freight volume for the f flow direction, and $f = n+1, \dots, n+m$. P^f is the gather of the selectable transportation route. p_l^f is the l selectable transportation route in P^f , and $p_l^f \in P^f$. $\psi_{ij}^k = 1$ or ε is a feasible coefficient. K is the gather of the different transport modes for the hazardous chemical substance. T is the gather of the time frame of the hazardous chemical substance. $R_{ij}^{k\delta}$ is the social risk in section (i, j) with k transport mode. c_{ij}^k is the transportation cost in section (i, j) with k transport mode. t_{ij}^k is the transit time in section (i, j) with k transport mode. c_i^k is the switching cost after altering the transport mode at node i . t_i^k is the time consumed after altering the transport mode at node i . θ_{ij}^k is the maximum carrying capacity of section (i, j) with k transport mode. $[T_1, T_2]$ is the time window restraint for delivering the hazardous chemical substance, T_1 is the earliest delivery time, and T_2 is the latest delivery time. δ , γ and η are dimension conversion coefficients. $x_{ij}^{t,k}$ is a variable (ranging from 0 to 1) that indicates whether the hazardous chemical substance passed section (i, j) with k transport mode during period t . x_i^k is a variable (ranging from 0 to 1) that indicates whether the hazardous chemical substance altered the transport mode at node i . s_o^f is the departure time in the f flow direction.

3.3 MODEL BUILDING

3.3.1 Objective function

The transportation route, mode, and time are generally considered in the decision making process for hazardous chemical substance transport. The optimization objectives are (1) minimize the social risk and (2) minimize the transportation expenses, which include the reloading fee.

The social risks involved in the transport of a hazardous chemical substance can be calculated as:

$$Z_1 = \sum_{f \in F} \sum_{k \in K} \sum_{(i,j) \in P_f^k} \sum_{t \in T} Q_f x_{ij}^{tk} R_{ij}^{tk} / \psi_{ij}^k . \tag{1}$$

The expense generated from hazardous chemical substance transport can be calculated as

$$Z_2 = \sum_{f \in F} \sum_{k \in K} \sum_{(i,j) \in P_f^k} Q_f x_{ij}^{tk} c_{ij}^k + \sum_{f \in F} \sum_{k \in K} \sum_{j \in o(f)} Q_f x_i^{tk} c_i^k . \tag{2}$$

3.3.2 Constraint condition

The constraint condition mainly includes the node flow, traffic capacity, and delivery time.

$$\sum_{(i,j) \in P_f^k} x_{ij}^{fk} - \sum_{(i,j) \in P_f^k} x_{ji}^{fk} = \begin{cases} 1, i \in o(f) \\ -1, i \in d(f), f \in F, P_f^f \in P^f, \\ 0, else \end{cases} \tag{3}$$

$$\sum_{k \in K} x_{ij}^{fk} \leq 1, f \in F, \tag{4}$$

$$Q_{ij} \leq \min_{(i,j) \in P_f^k} (\theta_{ij}), \tag{5}$$

$$T_1 \leq s_o^f + \sum_{k \in K} \sum_{(i,j) \in P_f^k} x_{ij}^{fk} t_{ij}^{Tk} + \sum_{i \in o(f)} \sum_{k_1 \in K} \sum_{k_2 \in K} x_{ij}^{fk} t_{k_1 k_2}^i \leq T_2, \tag{6}$$

$$x_{ij}^{fk} \in \{0,1\}. \tag{7}$$

The constraint condition in Equation (3) is utilized to guarantee directivity and flow balance during transportation. The constraint condition in Equation (4) represents only one mode or path of transportation. The constraint condition in Equation (5) indicates that traffic cannot exceed the capacity of the road segment. The constraint condition in Equation (7) shows that the value range of the decision variable is from 0 to 1. The multi-objective optimization model, M1, can be briefly expressed as:

$$M1 = \min\{Z_1, Z_2\}. \tag{8}$$

Subject to Equations (1) – (7).

4 Ant colony algorithm design

In computer science and operations research, the ant colony optimization algorithm is a probabilistic

technology for solving computational problems that can be reduced to finding good paths through graphs.

This algorithm is a member of the ant colony algorithms family in swarm intelligence methods and constitutes some meta-heuristic optimizations. The algorithm searches for an optimal path in a graph based on the behaviour of ants seeking a path between their colony and a food source. The original idea has since been diversified to solve a wider class of numerical problems. As a result, several problems that draw on the various aspects of the behaviour of ants have emerged.

4.1 HEURISTIC INFORMATION

Multiple selections for transport can be considered in deciding the pathway for a hazardous chemical substance. Under this circumstance, the original transport network should be converted through Figure 2. The decision making for hazardous chemical substance transport not only affects the population and transportation cost but also involves the mode of transport. Therefore, the heuristic information can be described as follows:

$$\eta_{ij} = \frac{1}{[R_{ij}^{tk} \oplus (c_{ij}^k + c_i^k)]}. \tag{9}$$

4.2 STATE TRANSITION

The trail level represents a posteriori indication of the desirability of a particular move. Trails are usually updated when all ants have completed their solutions. The level of trails that correspond to moves that are part of “good” or “bad” solutions is increased or decreased, respectively.

Generally, the k^{th} ant moves from state i to state j with probability:

$$p_{ij}^k = \begin{cases} \frac{[\tau_{ij}]^\alpha \times [\eta_{ij}]^\beta}{\sum_{j \in N^k(i)} [\tau_{ij}]^\alpha \times [\eta_{ij}]^\beta}, & \text{if } j \in N^k(i) \\ 0, & \text{other} \end{cases}, \tag{10}$$

where $N^k(i) = U / Tabu^k$ is the selectable gather for ant k , τ_{ij} is the amount of pheromones deposited for transition from state i to state j , $\alpha (\alpha > 0)$, $\beta (\beta > 0)$, $0 \leq \alpha$ is a parameter to control the influence of τ_{ij} , η_{ij} is the desirability of state transition ij , and $\beta \geq 1$ is a parameter to control the influence of η_{ij} . τ_{ij} and η_{ij} represent the attractiveness and trail level for the other possible state transitions.

4.3 PHEROMONE UPDATE

When all the ants have completed a solution, the trails are updated by:

$$\tau_{ij} \leftarrow \rho \cdot \tau_{ij} + \sum \Delta \tau_{ij}^k, \Delta \tau_{ij}^k = Q / f^k \quad (11)$$

where τ_{ij} is the amount of pheromones deposited for state transition ij , ρ is the pheromone evaporation coefficient, and $\Delta \tau_{ij}^k$ is the amount of pheromones deposited by the k^{th} ant.

4.4 PHEROMONE MAINTENANCE

To avoid arithmetic stagnation, the pheromone concentration section was set to $[\tau_{min}, \tau_{max}]$. We ordered τ_{max} when the concentration exceeded τ_{max} ; otherwise, we ordered τ_{min} when the concentration exceeded τ_{min} .

The implementation of the dynamic route optimization selection process is summarized below:

- 1) The values of each parameter (α, β, ρ, Q) are set. The number of ants is m , the maximum iteration number is N_{max} , the present iteration number is $n \leftarrow 1$, pheromone $\tau_{ij} \leftarrow 1$, and the departure of the hazardous chemical substance is t .
- 2) m ants are placed in the origin–destination of each direction. The ants select the next node according to Equation (8) and repeat the transition rule until the capacity restraint is met.
- 3) The pheromone is updated globally.
- 4) The departure time of the hazardous chemical substance is updated as $t \leftarrow t + \Delta t$. Return to step 2 until the departure time quantum has traversed. The iterations are updated as $n \leftarrow n + 1$.
- 5) End the process when $n = N_{max}$ and the optimal solution has produced an output. Otherwise, return to step 2.

5 Applications

The hazardous chemical substance transport network is shown in Figure 3.

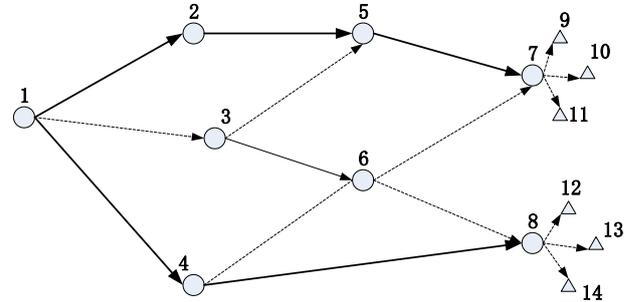


FIGURE 3 Hazardous chemical substance transport network

$V_1 = 1$ represents the output node of the hazardous chemical substance, and $V_1 = \{2, 3, 4, 5, 6, 7, 8\}$ represents the transfer node in the transport network of the hazardous chemical substance. $V_2 = \{9, 10, 11, 12, 13, 14\}$ represents the goal node of the hazardous chemical substance. We divided one day into T sections $T = [3, 6, \dots, 21]$. Two transportation modes exist in the hazardous chemical substance transport network. k_1 represents the railway, and k_2 represents the road. The thick line in the figure represents the transportation mode between the railway and the road. The filament and imaginary line represent the existing road transportation only.

The restraint of the time window is $[6, 18]$, and the quantity demanded is $R = [23, 16, 11, 25, 13, 32]$. The line parameters are described in Table 1.

Under the condition of different transportation periods on the railway and road, the populations referred to in each section are described in Table 2.

The calculated transportation route, mode, and volume in each flow direction are shown in Table 3.

TABLE 1 Reference value in each road segment

| Type | | Directed edge | | | | | | | | | | |
|----------------|---------|---------------|---------------|-------|---------------|---------------|---------------|---------------|-------|---------------|---------------|---------------|
| | | (1,2) | (1,3) | (1,4) | (2,5) | (3,5) | (3,6) | (4,6) | (4,8) | (5,7) | (6,7) | (6,8) |
| Feasible Index | Railway | 1 | \mathcal{E} | 1 | 1 | \mathcal{E} | \mathcal{E} | \mathcal{E} | 1 | 1 | \mathcal{E} | \mathcal{E} |
| | Road | 1 | 1 | 1 | \mathcal{E} | 1 | 1 | 1 | 1 | \mathcal{E} | \mathcal{E} | 1 |
| Ability | Railway | 30 | — | 45 | 35 | — | — | — | 35 | 20 | — | — |
| | Road | 20 | 15 | 25 | 25 | 20 | 20 | 25 | 15 | 20 | 25 | 25 |
| Cost | Railway | 10 | — | 9 | 12 | — | — | — | 9 | 10 | — | — |
| | Road | 12 | 11 | 8 | 14 | 10 | 10 | 10 | 11 | 12 | 11 | 9 |
| Time | Railway | 1.7 | — | 1.6 | 2.0 | — | — | — | 1.8 | 2.1 | — | — |
| | Road | 1.8 | 1.6 | 1.7 | 1.8 | 2.1 | 2.0 | 2.1 | 2.1 | 2.0 | 1.9 | 1.8 |

TABLE 2 Population referred to in each section (railway/road)

| Directed Edge | Time interval (h) | | | | | | | |
|---------------|-------------------|--------|--------|---------|---------|---------|---------|---------|
| | [0,3) | [4,6) | [7,9) | [10,12) | [13,15) | [16,18) | [19,21) | [22,24) |
| (1,2) | 40/60 | 40/60 | 60/60 | 80/60 | 50/60 | 60/60 | 80/60 | 40/60 |
| (1,3) | 60/— | 60/— | 140/— | 160/— | 140/— | 150/— | 170/— | 90/— |
| (1,4) | 50/70 | 50/70 | 80/70 | 100/70 | 70/70 | 80/70 | 100/70 | 50/70 |
| (2,5) | 20/40 | 20/40 | 60/40 | 80/40 | 50/40 | 60/40 | 80/40 | 20/40 |
| (3,5) | 30/— | 30/— | 50/— | 70/— | 50/— | 60/— | 70/— | 40/— |
| (3,6) | 60/— | 60/— | 80/— | 100/— | 80/— | 90/— | 110/— | 70/— |
| (4,6) | 50/— | 60/— | 50/— | 70/— | 60/— | 60/— | 80/— | 50/— |
| (4,8) | 100/80 | 100/80 | 120/80 | 150/80 | 130/80 | 130/80 | 150/80 | 100/80 |
| (5,7) | 70/90 | 70/90 | 90/90 | 120/90 | 100/90 | 100/90 | 120/90 | 80/90 |
| (6,7) | 85/— | 80/— | 90/— | 120/— | 100/— | 110/— | 125/— | 90/— |
| (6,8) | 90/— | 90/— | 120/— | 140/— | 120/— | 120/— | 150/— | 90/— |

TABLE 3 Path choice in each flow direction

| Flow direction | Pathway and shipping type | Departure time | Population | Cost |
|----------------|--|----------------|------------|------|
| 1→9 | 1→ (Railway) →2→ (Railway) →5→ (Road) →7 | [8,9) | 210 | 782 |
| 1→10 | 1→ (Road) →2→ (Railway) →5→ (Railway) →7 | [5,6) | 170 | 544 |
| 1→11 | 1→ (Road) →3→ (Road) →5→ (Road) →7 | [4,5) | 160 | 363 |
| 1→12 | 1→ (Road) →4→ (Road) →6→ (Road) →8 | [2,3) | 190 | 675 |
| 1→13 | 1→ (Road) →3→ (Road) →6→ (Road) →8 | [5,6) | 210 | 390 |
| 1→14 | 1→ (Railway) →4→ (Railway) →8 | [7,8) | 150 | 576 |

Table 3 shows the following: the transportation route of the hazardous chemical substance is 1→2→5→7→9, the departure time is [8,9), transportation route 1→2→5 is for railway transportation, and transportation route 5→7 is for road transportation. The transportation route of the (1→11) direction is 1→3→5→7→11, and the departure time is [7, 8). The transport mode is road transportation. The transportation route of the (1→13) direction is 1→3→6→8→13; the departure time is [7, 8). The transportation route of the (1→14) direction is 1→4→8→14; the departure time is [8, 9), and the transport mode is road transportation. The overall populations influenced by the hazardous chemical

substance are 109 million tons, and the cost is 3300 million dollars.

6 Conclusions

A transportation decision-making optimization model was established in this study. The complexity of the transport network and the transportation route, mode, and time were defined in the model. The decision-making optimization model was created based on the ant colony algorithm. The results indicate that the model is feasible and provides an effective approach for hazardous chemical substance transportation.

References

[1] Hu Y Current Situation 2013 Causes and Strategies Analysis of China's dangerous goods logistics development-Reference from Management Experience of Dangerous Goods Transportation in Developed Countries *Practice in Foreign Economic Relations and Trade* 05(4) 90-2 (in Chinese)

[2] Zor Z, Zhang B 2012 Route Optimization of Hazardous Chemicals Transportation with Mixed Time Windows *China Safety Science Journal* 22(4) 134-9

[3] Current J R 1988 The minimum-covering shortest-path problem *Decision Science* 19(7) 490-503

[4] Abkowitz M, Cheng P 1988 Developing a risk-cost frame work for routing truck movement of hazardous materials *Accident Annual Prevent* 20(15) 39-51

[5] Erkut E, Gzara F 2008 Solving the hazmat transport network design problem *Computers & Operations Research* 35(7) 2234-47

[6] Vedat V, Bahar Y K 2008 A Path-Based Approach for Hazmat Transport Network Design *Management Science* 1(54) 29-40

[7] Kai Y X, Wang H Y 2009 Research on optimization of transportation mode and route for hazardous materials transportation network *Journal of Safety Science and Technology* 5(1) 37-41

[8] Ren C X, Wu Z Z 2006 On route - choice analysis of hazardous materials transportation *Journal of Safety and Environment* 6(2) 84-8

[9] Zhang J, Ma X, Du W 2004 Nonlinear Programming Model and Algorithm of Two-Layer Distribution Network Based on Integrated Transit-Inventory Concept *Journal of Southwest Jiao Tong University* 39(3) 301-5

[10] Ma C X, Guang X P 2009 Highway Transportation Route Decision-Making of Hazardous Material in Developed Transportation Network. *Journal of Transportation Systems Engineering and Information Technology* 9(4) 134-9

[11] Ma C, Wen J, Li C 2009 Highway transportation route multiple attribute decision-making of hazardous material under the certain environment *Journal of Lanzhou Jiaotong University* 28(3) 115-8

[12] Wei H, Li J, Pu Y 2006 Route Planning for Hazardous Materials Transportation in Time-varying Network *Engineering-Theory & Practice* 17(10) 107-12

[13] Frank W C, Thill J C, Battar 2000 Spatial decision support system for hazardous material truck routing *Transportation Research Part C: Emerging Technologies* 8(1-6) 337-59

[14] Huang B, Cheu R L, Liew Y S 2004 GIS and genetic algorithms for HAZMAT route planning with security considerations. *International Journal of Geographical Information Science* 18(8) 769-87

[15] Liang Q 2010 Research on the liquid hazardous chemicals route selection by road transportation. *Beijing Jiaotong University* 7(7)

[16] Wu F, Wang X 2011 A safety evaluation model for dangerous goods transportation based on fuzzy Petri nets and its application *China Safety Science Journal* 21(1) 93-8

| Authors | |
|---|---|
|  | <p>Yibo Du, born on February 3, 1984, Sichuan, China</p> <p>Current position, grades: Ph.D Candidate at Southwest Jiao Tong University, China. University studies: Master degree at School of Public Administration, Southwest Jiao Tong University, China. Scientific interest: Logistics System Optimization and Informatization. Publications: more than 5 papers.</p> |
|  | <p>Jin Zhang, born on June 23, 1963, Sichuan, China</p> <p>Current position, grades: Professor at the School of Transportation and Logistics, Southwest Jiao Tong University, China. University studies: Ph.D degree of Traffic and Transportation Engineering from Southwest Jiao Tong University, China. Scientific interest: Logistics System Planning. Publications: about 60. Experience: Working experience of 28 years, completed 30 scientific research projects.</p> |

Research and implementation on integration information platform in China tobacco industry enterprise

Hailong Lu, Yong Cen*

China Tobacco Zhejiang Industrial Co., Ltd. Hangzhou, PR. China

Received 1 March 2014, www.tsi.lv

Abstract

For better information technology combined with enterprise management mechanism to solve the information technology into the enterprise production and operation of each link, the integration of information platform for building modern tobacco industry enterprise is proposed. First of all, this study combines the principal business process for building market-driven enterprise, and proposes based on self-assembled dynamic fifth-order business model for enterprise business operation model. According to the actual situation of enterprise information system construction, the integration information platform application architecture and integration architecture designs is developed. This information platform uses SAP XI as the ESB transforms the formats of all data coming from source systems to realize the seamless integration among different systems. With this integration information platform construction it can better support enterprise development strategies, optimizing resource allocation, improve business and management efficiency, and promote scientific enterprise sustainable development.

Keywords: tobacco industry enterprise, business process model, integration information platform

1 Introduction

As the China tobacco industry continues to promote market-oriented reforms, re-quires the tobacco industry enterprises to improve the ability of adapting to the market, to advance the tobacco industry continues to upgrade and promote the tobacco industry sustained, stable and coordinated and healthy development. At the present, how to promote market-oriented business transformation, improve the integration of management and control capability, and improve operational efficiency, follow-up industry development strategic planning and strategy, to cope with more open and more intense market competition, become the important challenges facing tobacco industrial enterprise.

At present the domestic tobacco industry enterprises to innovative development model, combining information technology and management and system mechanism innovation, to deepen the depth of industrialization and information technology integration, Focus on information technology transformation and using information technology into research and development, manufacture, marketing, procurement and other aspects. Promoting industry transformation and upgrading of industry, and build the modern tobacco industry enterprise.

The final aim is to adhere to the market demand as the leading factor, market orientation reform practice and management model innovation, to construct integrated information platform meet the research and development, manufacture, marketing, procurement and other business, which to support enterprise development strategies,

optimizing resource allocation, improve business and management efficiency, and promote scientific enterprise sustainable development.

2 The Information platform current situation in tobacco industry enterprise

Nowadays, the research and application of integration information platform meet the enterprise business model has been the focus of manufacturing industry informatization in home and abroad enterprises. Since the main features that reflect the supply chain-oriented operation mode, management processes, business practices and refining, support changing business models and business models continuous improvement, etc.

Domestic tobacco industry enterprise information platform research direction and trend mainly reflected in the following aspects:

- 1) Management ideas and technology to further blend in epitaxial aspects: On the basis of enterprise resource planning management thinking, continue to absorb the latest advanced management ideas or patterns, such as agile manufacturing, lean production, concurrent engineering, total quality management and agile virtual enterprise organization and management model, based on e-commerce enterprise collaborative management model, and cross-enterprise collaborative project management model, combines the management ideas and business processing model.
- 2) Integration information platform capabilities continue to expand in connotation aspects: The increasingly

*Corresponding author e-mail: ceny@zjtobacco.com

fierce market competition requires enterprises to dynamically adjust timely and transformation. There are two main trends of the development, on the one hand, the introduction of Dynamic Enterprise Modelling (DEM), which means to introduction business process continuous improvement as the goal of dynamic information system in support of Internet/Intranet technical environment. On the other hand, the Intelligent Resource Planning (IRP) breaking the previous all those "transaction processing-oriented" management model, help managers followed even ahead of market changes quickly make the right decisions, changed the original plan, and fastest implementation of these changes to solve previously unsolvable "collaborative manufacturing" and "resource constraints" and other issues.

- 3) Integration information platform for the transition to the intelligent information processing: Integration information platform for business intelligence features include intelligent filtering and processing capabilities, program optimization, intelligent data analysis, etc. Integrated data warehousing, data mining and online analytical processing (OLAP), business intelligence, decision will support to strengthen enterprise knowledge management capabilities. Constitute a set of integrated query, reporting as the intelligent decision-making information systems to help business managers to find a best solution for macro decision-making and business strategies, and improving the resilience of the market and on-site management capabilities.
- 4) Integration information platform combined with Internet business model, mobile terminals and other cutting-edge technology: Internet-based integration information platform will incorporate the latest Internet technologies and business management

concepts in the browser/server structure, make the enterprise marketing management, logistics network across the enterprise management, product lifecycle management, and decision support system extended to computers, terminals, mobile phones, PAD and other new types of end products. To realization of the front office, business intelligence, electronic commerce, office automation system and the integration of supply chain management applications.

3 Tobacco industry enterprise business analysis

3.1 KEY BUSINESS PROCESSES

Enterprise constructed market driven management model based on market-oriented, means from market to market in the operation of the process. Its contents are demand forecasting as the leader, production planning, raw materials preparation plan, quality inspection plan, marketing plan coordination linkage, sensitive reflect changes dynamically adjust promptly to ensure the most timely reflect market demand and market supply to the maximum extent satisfied.

Mainly is implemented based on the market, manufacturing, procurement, technology research and development as four centres, key business processes can be summarized as "Five Rolling", respectively are Rolling Forecasts, Rolling Production, Rolling Supply, Rolling Maintenance and Rolling Service. According to market-driven enterprise management mode, business status, departments, functions characteristics, as well as information on the role of operations management, aiming at process performance, combing involves the whole business process of the operation of the enterprise management system to form a new process (see Figure 1).

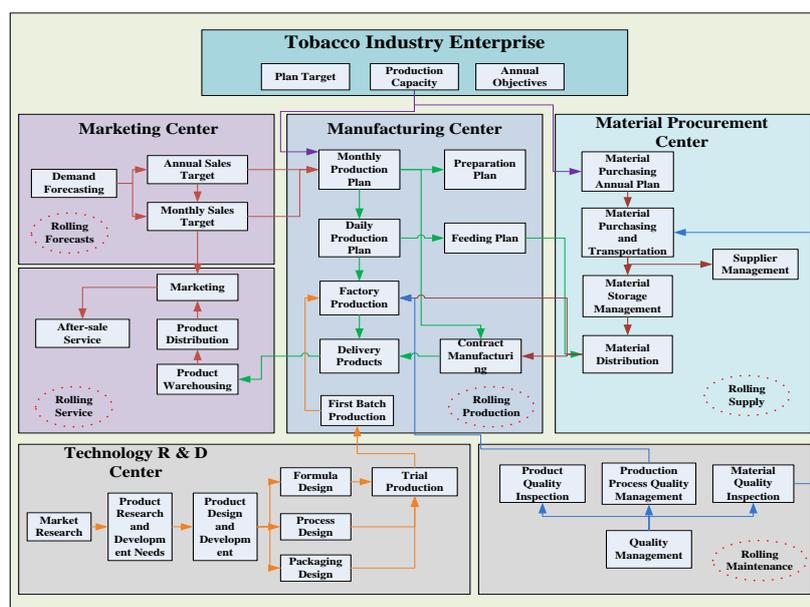


FIGURE 1 Key business processes in enterprise

According to business processes, induction, to extract five vertical and five horizontal ten main line (see Figure the enterprise's core business and management support of 2).

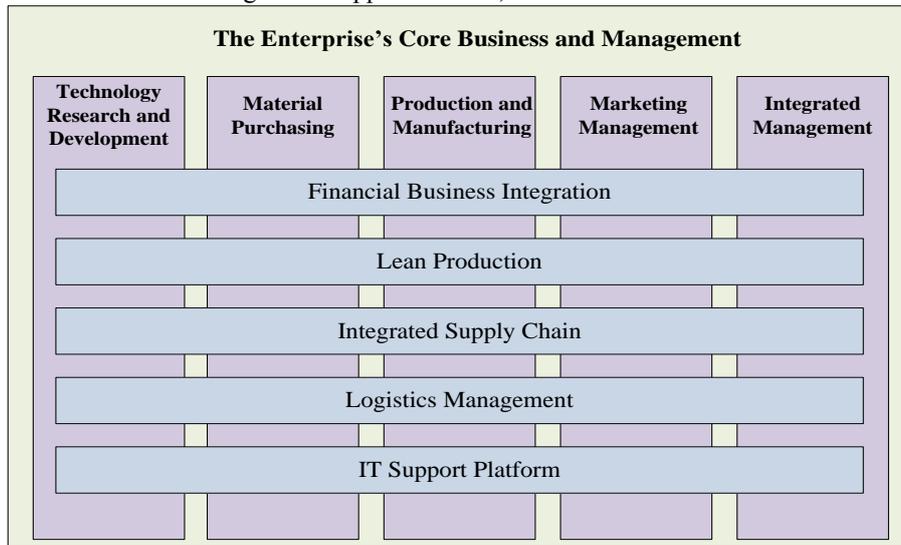


FIGURE 2 The enterprise's core business and management support

3.2 TOBACCO INDUSTRY MARKET-DRIVEN ENTERPRISE INTEGRATION MODE OF OPERATION

Enterprises based on market-driven business model and management features, build a unified organizational structure and improve adapt to the market, in response to the market's ability to supply chain operations. By building the Self-Assembled Dynamic fifth-order business model, inherent characteristics of the tobacco industry's supply chain operation mode, better achieved in the monopoly planning system, flexible organization, dynamically adaptive market demand, and achieved the balance between the company's production and supply coordination.

With the continuous development of information technology and the improvement of supply chain theory, the game between enterprises became the game between supply chain and the supply chain enterprises. How to create an efficient, integrated supply chain system become the key to the success of enterprise.

Enterprises to achieve planning, production, supply, service, maintenance, and other key aspects of the integration of the supply chain work together, across the various business links and nodes in the network and resources the effective aggregation, it can adapt to the changing requirements of supply chain operation. On this basis, we built the five order type and the dynamic network model in conformity with the characteristic. The model is divided into three layers (see Figure 3).

The first layer is a dynamic five-order management system. Respectively, from project management, production management, supply management, maintenance management, service management, and other aspects of the dynamic five-stage system for the

overall program, dynamic coordination, operation and optimization.

The second layer is the business network. The business network automatically according to certain rules and gather resources within the network nodes, while providing resource sharing across organizational capabilities and the ability of transparent access to resources. It will contribute to the overall operation of enterprise management solutions, such as supply network planning, coordination problems and some practical applications. Model every aspect of a business involved both as a self-governing region, extraterritorial autonomy exists between the ministries of collaborative relationships between its internal collaborative relationships also exist.

The third layer is the introduction of the supply chain lifecycle management. From the core of supply chain strategy perspective, until the dynamic five-order system of the whole process of construction, operation and continuous improvement, it will reduce duplication and redundancy; standardize work processes, efficient use of resources for the node enterprises to reduce operating costs, reducing operational risk.

The three layer architecture of Self-Assembled Dynamic fifth-order model to better overall management of the company's business characteristics and requirements of a broad and concise, inherent characteristics of the tobacco industry to build the supply chain operations and management and control model. To design and implement integrated information platform to provide a clear demand-oriented and construction requirements. Simultaneously, promote innovation and technological innovation business management joint development and promote the formation of a unique competitive advantage of enterprises.

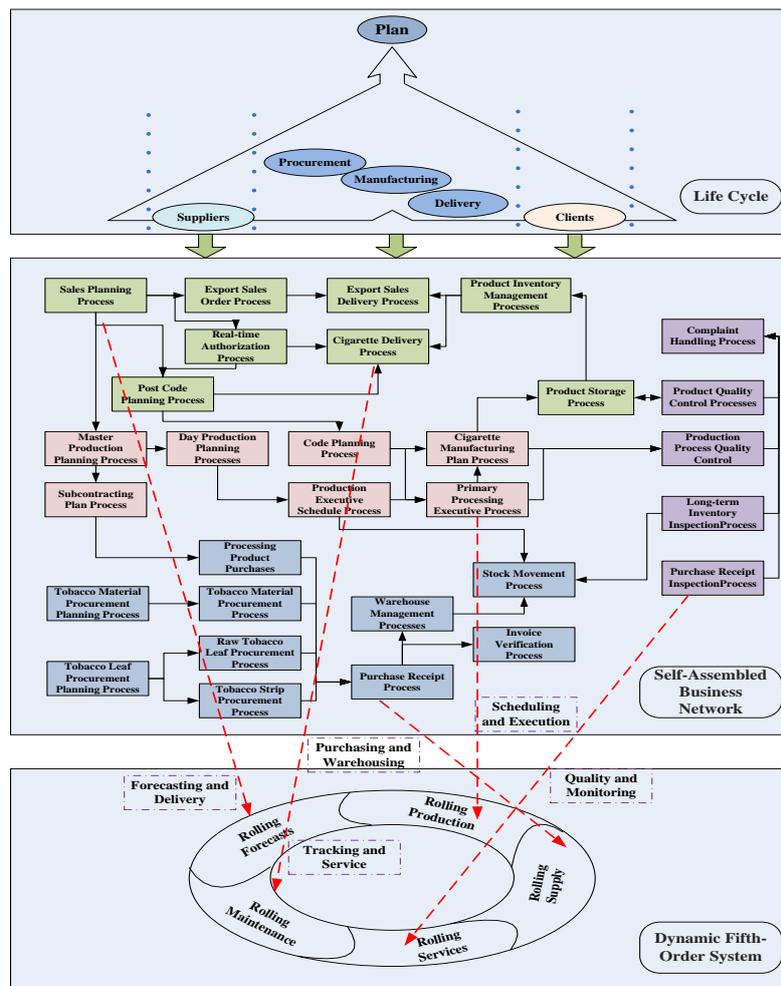


FIGURE 3 The self-assembled dynamic fifth-order business model

3.3 SUPPLY CHAIN MODEL BASED ON PUSH-PULL TYPE BUSINESS PROCESS

The meaning of the business process management is the business process can be automatic control and processing, make the enterprise to realize the automatic operation and management. It makes the enterprise can according to the actual situation of change and the demand of the market, to adjust the business process, improve the flexibility of enterprise business process. Business process management (BPM) provides a broad range of facilities to enact and manage operational business processes. Increasingly, more and more organizations use BPM techniques and tools to promote business effectiveness and efficiency [1, 2]. BPM Not only covers the traditional "Work flow" of the processes to pass, process monitoring the scope of, usually in order to Internet way to achieve information transmission, data synchronization, the business monitoring, and enterprise business processes' continued to upgrade optimization of, breaking the traditional "Work Flow" technology the bottleneck.

Business process models are abstract representations of business process, also is the abstract representation of the business process. Push-pull business process models from the establishment of the purposes is to achieve

business process automation, the model includes not only the required number of discrete activities and their interconnected relationships, also define many other information, such as organizations, resources, data, roles, and relationships of these elements describe rules. Description of the activities and routing is the main content of the business process model, the process can be broken down to atomic activity, eventually routing nodes. Model is convenient, comprehensive and intuitive description of the process, help process optimization analysis, depends on the node type and semantics, semantic richness node will directly affect the ability to express the model.

According to the actual circumstances of the business process, proposed push-pull type business process model based on Petri nets. Petri nets have turned into one of the most widely used formalism for workflows model. Their expressive power and readability, especially in their high-level version, has proven sufficient to represent most of control-flow patterns of workflows [3]. The Petri net (PN) is a graphical and mathematical model tool that has such characteristics as concurrent, asynchronous, distributed, parallel, nondeterministic, and stochastic. It can be used to model and analyse various systems [4]. Petri nets have formal semantics definitions; graphical representations of

intuitive, graph theory and mathematical rigor phase support the theoretical advantages. Characteristic is that it focuses on the system changes, including changes of conditions, the result of the change and the inner link between changes. By analysing the business model and the operational characteristics of the supply chain, to form a single point model, multi-start model for business process model.

After business process model was constructed by using Petri nets, we using 6-R model (Role, Relationship, Regions, Resources, Risks, Reconfiguration) to describe the supply chain nodes and build supply chain model. According to the evolution of manufacturing systems models, different types of networked manufacturing system present the typical pattern for the corresponding.

Network's important characteristic is its distribution characteristics and relationships between resources or ability. Follow this evolution path; the 6-R model defines the basic elements of a network, which provides a basic structure for the network design. In this model, Role, Relation-ship and Region mainly for the static structure, and describe the network's main static and operational characteristics. Resource, Risk and Reconfiguration mainly for dynamic structure, describe the performance of the network and the adaptive mechanism. Meanwhile, in each R, also includes many substructure properties.

The 6-R model as the Figure 4 shown, this model has a number of dimensions greater than three, better openness and interoperability, and also, its number of views can be determined according to their needs.

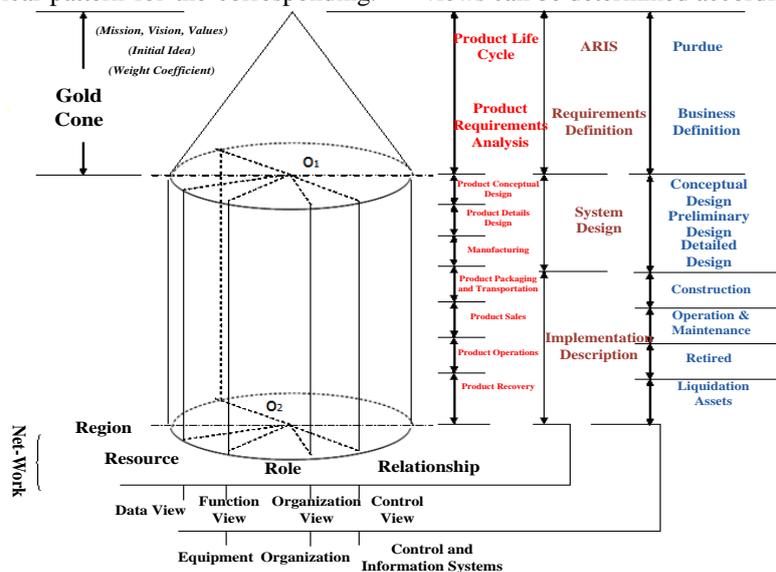


FIGURE 4 The 6-R supply chain model

The 6-R model provide new utility for enterprise supply chain node classification, description, publication, discovery and reuse, help enterprise from a new perspective of whole life cycle to re-examine its own position in the supply chain, optimize from each aspect, so as to speed up promote its sustainable development ability. It also can help enterprises optimize the integration of massive resources and other advantages of personalized, and make the enterprise in the unpredictable environment make agile response to rapidly changing markets and opportunities, the advantages of lower production costs, etc.

4 Integration information platform construction

4.1 APPLICATION ARCHITECTURE DESIGN

According to the National Bureau of standards, enterprise standards for requirements, to establish standard system of enterprise information resources, ensure application system integration architecture of advanced, stable, flexible, and guarantee data sharing timely, efficient and accurate.

Through the understanding of business operations and management requirements, based on enterprise business management model, analysis of the work of information technology. Under the condition of the core information system basically, the requirements for information platform focus on several aspects:

- 1) Information coverage to be more complete and information construction need to enterprise exterior supply chain.
- 2) Information management content to be more precise, in the supply chain must manage the smallest packaging unit.
- 3) To dig deeper into the data mining to improve data analysis capabilities, enhance the role of decision support information.
- 4) Further integration of various information systems; improve information management capacity and efficiency.

Enterprise put forward the information construction of three-tier architecture, respectively are control layer, manufacturing execution layer and enterprise management layer (see Figure 5).

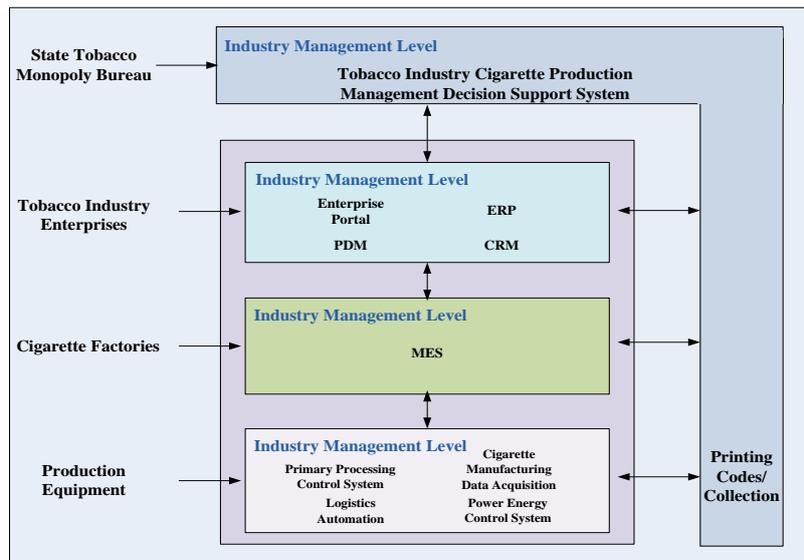


FIGURE 5 The three-tier architecture of the information construction

Enterprise Management Layer: Mainly includes that core systems construction in company, such as enterprise portal, ERP, PDM [5], CRM, and quality management system [6], etc. These application systems fully integrated enterprise logistics, information flow and capital flow, support the production and operation of enterprises.

Manufacturing Execution Layer: The core applications are manufacturing execution system, panoramic display technology embodied in flexible multi-specification, refinement, dynamic characteristics, comprehensive support production, quality, equipment,

three major lines of business; achieve intelligent scheduling, digital equipment operation and maintenance, quality control, manufacturing process traceability, real-time performance analysis and integrated plant management.

Control Layer: Mainly includes primary processing control system, cigarette manufacturing department data acquisition, logistics automation and power energy control system. Implements the data collection automation, digital equipment operations, make data acquisition activity to realize scientific and rationalization and carry out the benefit maximization.

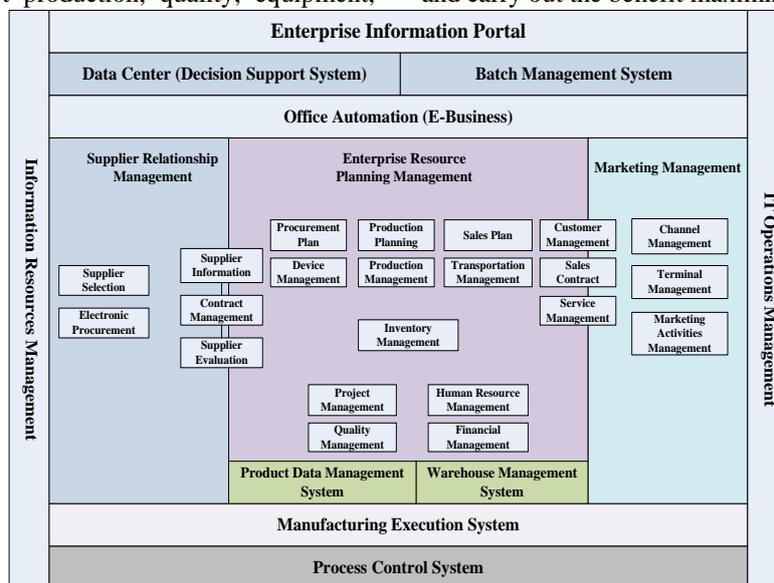


FIGURE 6 The application architecture design of the integration information platform

4.2 INTEGRATION ARCHITECTURE DESIGN

According to the actual system application and development trend of the future, design the overall top-level of information platform in enterprise. It must be meet the application system standardization, integration,

visualization, real-time principles. At last, technical requirements mainly reflected in the following aspects: Introduction of data service bus and establishment of service-oriented enterprise technical architecture; Support heterogeneous systems, unified data storage management

and security management; On the basis of the data center, and gradually establish decision auxiliary analysis system.

The integration information platform was constructed by using service oriented architecture (SOA) and SAP XI technology. SAP XI as the ESB transforms the formats of all data coming from source systems to realize the seamless integration among different systems. The service bus of SAP XI links different systems in enterprises and factories and realizes the seamless integration among multiple business systems of different levels. The application of information integration platform ensures the flexible and controllable information communication among application systems in enterprise at architectural level and solves the problems of information island, data quality and flow integration [7].

Enterprise adhere to the business requirements, and process and data integration planning, adoption of SOA architecture design, using the SAP XI (Exchange Infrastructure) as the enterprise service bus technology to form an integrated whole as shown Integration Architecture. The technical architecture mainly reflected in the following aspects, namely, with its special loose coupling, coarse granularity, reusability and interoperability, SOA has become the effective method of information resources integration management, and is the important

development direction of realizing the fusion of tobacco industry information and IT industry. The essence of SOA is to integrate reusable services with distinct boundary and self-contained functions by the centralized management platform. The platform can link different services through the interfaces and the contracts defined among application services [8].

By using the SAP XI as the Enterprise Service Bus (ESB), and through by using the SOAP protocol and star-shape structure can realize the integration of various heterogeneous system. In the technical implementation, established star-shape connection with SAP XI as a central Hub, systems simply can make a connection with the SAP XI. In order to effectively avoid the sharp rise of complexity caused by the one-to-one connection between the middleware, system and system interconnect system integration.

Means for using the service transmission heterogeneous systems, rather than directly access each database, increase efficiency while also reducing the technical risk.

Using data bus technology has the advantage of architecture expansionary, the re-usability of data and ease of maintenance.

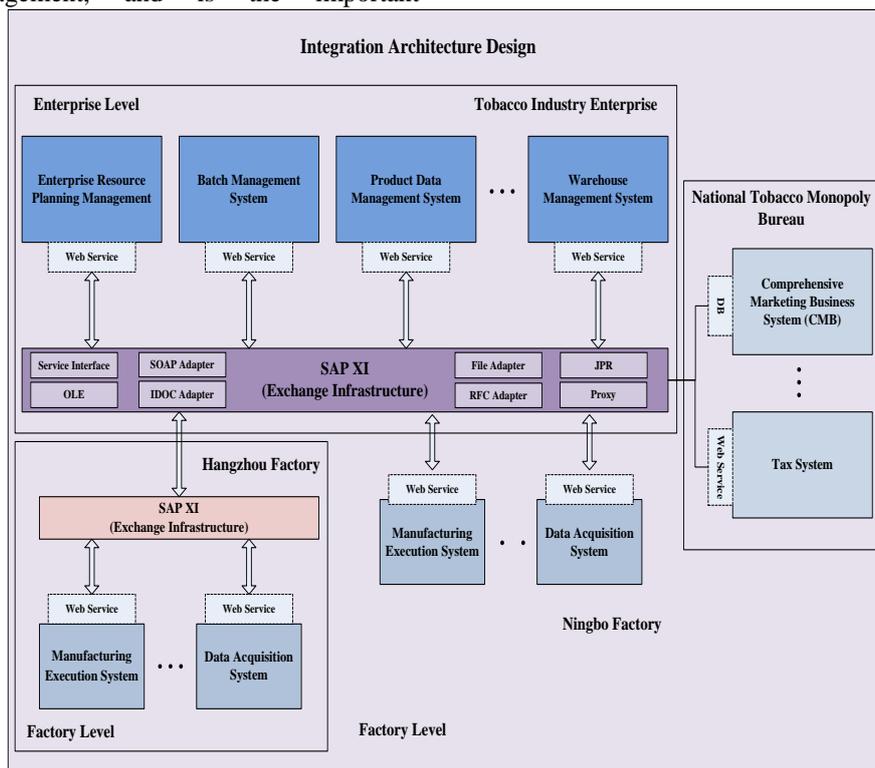


FIGURE 7 The integration architecture design of the integration information platform

5 Conclusions

In this paper, an integrated information platform for business and process engineering framework for synthesis and design in tobacco industry enterprise has been presented. The end goal of the integration information platform is to develop the information

technology into the design, manufacture, marketing, procurement and so on each link, promote transformation and upgrading of enterprises, and build the modern tobacco industry enterprise. This paper illustrates the key business processes and builds the enterprise business model based on the self-assembled dynamic fifth-order business model. On the basis of the self-assembled

dynamic fifth-order business model, design and implement information integration platform, will promote the business management innovation and technology

innovation with integration and development together, and improve enterprise's comprehensive competitiveness.

References

- [1] Huang Z X, van der Aalst W M P, Lu X D, Duan H L 2010 *Expert Systems with Applications* 37(12) 7533-41
- [2] Huang Z X, Lu X D, Duan H L 2012 *Expert Systems with Applications*, 39(7) 6458-68
- [3] Vidal J C, Lama M, Bugarin A 2012 *Expert Systems with Applications* 39(17) 12799-813
- [4] Shen V R L, Yang C Y, Wang Y Y, Lin Y H 2012 *Expert Systems with Applications* 39(17) 12935-46
- [5] Cen Y, Wang H L 2013 Research and Implementation Cigarette Product Design Management System in Chinese Tobacco Industry Enterprise. *2013 International Conference on Computer Science, Electronic Technology and Intelligent System (CSETIS) March 22-23 2013 Hangzhou, China* 396-400
- [6] Cen Y, Wang H L, Zhang Z H, Wang H W 2013 Design and Implementation of Quality Management Information System in Chinese Tobacco Industry Enterprise *2013 International Conference on Information, Business and Education Technology (ICIBIT) March 14-15 2013 Beijing, China*, 279-83
- [7] Wang H L, Cen Y 2013 Implementation of Information Integration Platform in Chinese Tobacco Industry Enterprise Based on SOA *The 2nd International Conference On Systems Engineering and Modelling (ICSEM) April 19-20 2013 Beijing China* 0291-95
- [8] Duan Y E 2012 Research about Based-SOA Agriculture Management Information System *2012 International Conference on Information and Automation (ICIA) June 6-8 2012 Shenyang China* 78-82

Authors



Hailong Lu, born on July 28, 1976, Hangzhou, China

Current position, grades: Chief of System Operation in China Tobacco Zhejiang Industrial Co., Ltd.
University studies: Management Information System in Kunming University of Science and Technology
Scientific interest: Informatization Planning, Information System Construction, Project Management



Yong Cen, born on July 15, 1981, Hangzhou, China

Current position, grades: Staff of System Operation in China Tobacco Zhejiang Industrial Co., Ltd.
University studies: Pattern Recognition and Intelligent Systems in Xiamen University
Scientific interest: Data Mining, Information System Construction, Project Management
Publications: He has published four papers

A fuzzy clustering approach of the customers' demands, which influences the e-banking service quality

Xilong Liu, Yizeng Chen*

School of Management, Shanghai University, Shanghai, 200444, China

Received 1 March 2014, www.tsi.lv

Abstract

The interest rate liberalization have a huge influence for commercial banking management in China, the net interest margin (NIM) is more and more low, but the competition is becoming increasingly fierce, so it is a very necessary and urgent work to strength the management of banks by the key financial innovation. Some researches showed that the e-banking service quality plays an important role during competition among banks as well as the core competence of banks' sustainable development. In order to improve the service quality, the first task is to really master and understand the customer demands, which influence the e-banking service quality. The paper proposed a fuzzy clustering method for customer demands and empirical analysis, and the results showed that all the customer demands can be classified into two clusters according to the maximum value of F-statistics, one of which indicated that the new trend of customer demands in e-commerce environment and have great influence on the decisions of users to use the service of e-banking.

Keywords: e-banking, service quality, customers' demands, fuzzy clustering

1 Introduction

Some researches have shown that the Interest Expense is the most important operating expense for banks (about 85%) in china, and the narrowing of spreads creates a huge challenge for commercial bank management, the interest rate liberalization have also impacted on the profitability of commercial banking [1], so commercial banking have to innovate the payment business and strength the management of banks in order to seek for new sources of profit. In the People's Bank of China (PBC) work conference 2013, which have proposed to promote the reform of interest rate liberalization steadily, to encourage and guide the payment business innovation, and further enhance the level of financial services and management.

With the rapid development of Network Economy and the optimization of electronic payment environment, Electronic banking, i.e., e-banking, as the key channel of business innovation not only can reduce the operation cost, but also can increase the intermediary business income and promote the business innovation, it is more convenient for service the customer, and so on. All variableistics have increasingly become the focus of banks service innovation. According to the "2012 China e-banking research report" of CFCA, which indicated that the e-banking services of china have sustained growth in the past three years, 68% of users use e-banking services instead of the more than half of the counter business, and the replacement rate of some internet-banks are more than 85%. How to improve the e-banking service quality and strengthen the management of banks is an urgent

subject in the background of interest rate liberalization in China.

2 Literature review

Many researches have proved that the service quality determines the Customer Satisfaction [2], and the Customer Satisfaction determines Customer Loyalty [3], which directly or indirectly brings in the profit increase. For example, if the customer loyal rises by 5%, the company's net profit will increase by 25%~95% [4]. As mentioned above, the logic can be concluded as: Service Quality → Customer Satisfaction → Customer Loyalty → Profit [5]. Therefore, when evaluating and improving the banking service quality, it is the primary and essential factor to recognize the customer demands, which has been studied by many domestic and foreign professors.

Davis has given a detailed measure, one of the key decisive factors is the Technology Acceptance Model (TAM) [6], and the model is widely synthesized to a set of methods to test users' acceptance psychology among several fields. Karjaluoto et al. have made an investigation on the attitude of users to use e-banking. They held the point of view that users' proficiency to technique is good to use e-banking. Providing customers with fast, convenient and trustworthy service influenced the e-banking development, which is the primary consideration factor for customers to be pushed to accept the e-banking services [7]. Minocha et al. believed that banks should cultivate a positive customer experience for e-Banks and the real atmosphere to attract more users. Lacking trust for the e-business channels is the

* *Corresponding author* e-mail: mfcyz@shu.edu.cn

major cause that customer lead to the restriction of extended application of e-bank [8]. Mills et al. had done a research from 92 potential users of internet banking. The results showed that security problem was one of main obstacles that cause e-banking services were refused by users in developing countries [9]. Those security measures that cause inconvenience to the user of the e-banking may result in safety weakness, and the reason is that user lacked of experience. Lee M C. explores and integrates the various advantages of online banking to form a positive factor named perceived benefit with the technology acceptance model (TAM) and theory of planned behavior (TPB) model to propose a theoretical model to explain customers' intention to use online banking. The results indicated that the intention to use online banking is adversely affected mainly by the security/privacy risk, as well as financial risk and is positively affected mainly by perceived benefit, attitude and perceived usefulness [10]. About the factors of perceptive service quality, both Jabnoun and Al-Tamimi thought that the service quality played an important role in protecting market share, and customers believed that staff's personal professionalism was the biggest factor of service quality [11]. Sadiq Sohail et al. emphasized the importance of customers' perception, and they pointed that the reasonable navigation property, search tools and higher interactivity of e-banking websites have positive influence on perceptive degree of e-banking website-friendly [12]. Li and Worthington also emphasized that the transaction cost, the connection speed and their confidence of e-banking trades have influenced on whether they will use e-banking or not [13]. Chu, Po-Young et al. proposed a research model to examine the relationships between service quality, customer satisfaction, customer trust, and loyalty on Taiwanese e-banking, It was found that e-banks must focus on service quality to increase customer satisfaction and trust to obtain customer loyalty [14]. By studying domestic and foreign research conclusions, Zhang Wei et al. have put forward 10 pieces of assuming dimensions, including appearance, reliability, responsibility, guarantee, empathy, sensibility, technicality, information function, price property and defect degree. They have investigated 2409 users of e-banking. They adopted confirmatory factors to analyse and got 26 questions to support these 10 pieces of assuming dimensions. After testing the relationship between 10 pieces of assuming dimensions and total service quality, they have proved that these 10 pieces of assuming dimensions have positive correlation with service quality and those were factors that influenced service quality [15]. According to the above information, we find that there are many papers of studying e-banking service quality, but it still need to improve on scientific recognition and classification of customers' demands. This paper tries to distinguish customers' demands by fuzzy cluster method, and confirm crucial customers' demands for the e-banking QFD [16], as a system based on customer demands and changing the demands into technical requirements at

various stages. Quality Function Deployment (QFD) is now playing a more and more important role in the service design.

3 The fuzzy clustering algorithm

In the fields of science-technology and economy management, which are classified by certain criteria (for example similarity degree or affinity-disaffinity relationship). Because of the boundary of science technology and economy management are ambiguous, it's practical to adopt fuzzy-cluster. It is one of basic methods for making the pattern classification and system modelling [17], it is very simple, direct and easy for the basic thought. It is classified by similarities degrees of each modeling features, that is to say, the similarities will be classifies as a cluster, and the different another [18]. Because fuzzy-cluster can obtain an uncertainty degree of samples that belonging to different classes and present intermediacy of a sample, namely, create an uncertainty description of sample to classification. In this way, it can better present the real world cases so that it will become the main stream of cluster study [19].

The steps of cluster analysis method are described below.

3.1 UNIVERSE VARIABLEIZATION

Assume that the universe U were the clustered case, $U = \{x_1, x_2, \dots, x_n\}$, where x_i is the i th case, can be indicated as:

$$x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}, (i = 1, 2, \dots, n),$$

where m is the m^{th} variable of each case, so the original data matrix can be obtained as follows:

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix},$$

where x_{im} is the m^{th} variable of n^{th} case.

3.2 DATA NORMALIZATION [20]

In order to make the different data were compared, it is need to data normalization, the data were compressed to interval [0,1], these methods includes moving-standard deviation normalized, moving-range normalized, logarithmic normalized. In this paper, we use the methods of translation standard difference alternate and translation limit difference alternate.

3.2.1 Moving-standard deviation normalized

$$x'_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}, \quad (i=1,2,\dots,n; k=1,2,\dots,m), \quad (1)$$

where $\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}, s_k = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}$.

x'_{ik} can be obtained from the Equation (1), the mean of each case is 0, the standard deviation is 1, and which eliminated the influence of dimension, but the x'_{ik} is not necessarily in the interval [0,1].

3.2.2 Moving-range normalized

$$x''_{ik} = \frac{x'_{ik} - \min_{1 \leq i \leq m} \{x'_{ik}\}}{\max_{1 \leq i \leq m} \{x'_{ik}\} - \min_{1 \leq i \leq m} \{x'_{ik}\}}, \quad (k=1,2,\dots,n) \quad (2)$$

x''_{ik} can be obtained from the Equation (2), it is clearly that $0 \leq x''_{ik} \leq 1$, the mean of each case is 0, the standard deviation is 1, which have really eliminated the influence of the dimension.

3.3 CALCULATE THE FUZZY SIMILARITY MATRIX

Assume that the universe $U = \{x_1, x_2, \dots, x_n\}$, where $x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$. According to the traditional clustering algorithm to determine the similarity coefficient, the fuzzy similarity matrix can be built, r_{ij} is the similarity degree between x_i and x_j , $r_{ij} = R(x_i, x_j)$. There are several methods to calculate that including the similarity coefficient, distance method and other methods of the traditional clustering algorithm. In this paper, we mainly use the absolute value of the arithmetic subtraction as follows:

$$r_{ij} = 1 - cd(x_i, x_j), \quad (3)$$

where, c is the chosen properly parameter, $0 \leq r_{ij} \leq 1$, $d(x_i, x_j)$ is the distance between x_i and x_j :

$$d(x_i, x_j) = \sum_{k=1}^m |x_{ik} - x_{jk}|. \quad (4)$$

3.4 CALCULATE THE FUZZY EQUIVALENCE MATRIX

According to the fuzzy similarity matrix R are given to, the fuzzy equivalence matrix R^* which can be obtained by transformation use the transitive closure method. The

transitive closure of R can be calculate by square arithmetic as follows:

$$t(R) = R^*. \quad (5)$$

The $t(R)$ is the fuzzy equivalence matrix, which meet the reflexivity, symmetry and transitivity.

3.5 CLUSTERING

For $t(R)$, which cutting matrix can be calculated by different threshold λ , Through analysis the similarity degree of cutting matrix elements, the factors of each customer demand will be carried out clustering.

Assume that the $\tilde{R} = (r_{ij})_{m \times n}$ is fuzzy matrix, for any $\lambda \in [0,1]$, the $\tilde{R}_\lambda = (r_{ij}^{(\lambda)})_{m \times n}$ indicates the λ - cutting matrix of $\tilde{R} = (r_{ij})_{m \times n}$, where,

$$r_{ij}^{(\lambda)} = \begin{cases} 1, & r_{ij} \geq \lambda \\ 0, & r_{ij} < \lambda \end{cases}$$

3.6. DETERMINE THE OPTIMAL THRESHOLD λ

During fuzzy clustering analysis, we can get different classification with different threshold, $\lambda \in [0,1]$, When the threshold is decreased gradually from big to small, classification is changed from thin to thick, so it forms a dynamic clustering analysis shown in Figure 1. Through the cluster analysis chart, according to the actual problem, select the best threshold, it can get the sort of case classification and realize the identifying and analysis for the customer demand of e-banking service quality. In this paper which use F-Statistic to determine the threshold λ .

Assume that the universe $U = \{x_1, x_2, \dots, x_n\}$ is the case set, each case have m variables, $x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$, ($i=1,2,\dots,n$), so the original data matrix can be expressed as follows:

TABLE 1 The original data matrix

| | V_1 | V_2 | ... | V_k | ... | V_m |
|-----------|-------------|-------------|-----|-------------|-----|-------------|
| C_1 | x_{11} | x_{12} | ... | x_{1k} | ... | x_{1m} |
| C_2 | x_{21} | x_{22} | ... | x_{2k} | ... | x_{2m} |
| ... | ... | ... | ... | ... | ... | ... |
| C_i | x_{i1} | x_{i2} | ... | x_{ik} | ... | x_{im} |
| ... | ... | ... | ... | ... | ... | ... |
| C_n | x_{n1} | x_{n2} | ... | x_{nk} | ... | x_{nm} |
| \bar{x} | \bar{x}_1 | \bar{x}_2 | ... | \bar{x}_k | ... | \bar{x}_m |

* C_n indicates the n^{th} case, V_m indicates the m^{th} variable.

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}, \quad (k=1,2,\dots,m), \quad (6)$$

where, \bar{x} was called the centre vector of case.

Assume that r is the number of clusters, which corresponds to λ , the n_j is the j^{th} number of case which can be indicated as $x_1^{(j)}, x_2^{(j)}, \dots, x_{n_j}^{(j)}$, the j^{th} clustering center can be indicated as $\bar{x}^{(j)} = (\bar{x}_1^{(j)}, \bar{x}_2^{(j)}, \dots, \bar{x}_m^{(j)})$, where $\bar{x}_k^{(j)}$ is the mean of the k^{th} variable,

$$\bar{x}_k^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ik}^{(j)}, (k=1,2,\dots,m). \tag{7}$$

F-Statistic can be calculated as follows,

$$F = \frac{\sum_{j=1}^r n_j \frac{\|\bar{x}^{(j)} - \bar{x}\|^2}{r-1}}{\sum_{j=1}^r \sum_{i=1}^{n_j} \frac{\|x_i^{(j)} - \bar{x}^{(j)}\|^2}{n-r}}, \tag{8}$$

$$\|\bar{x}^{(j)} - \bar{x}\| = \sqrt{\sum_{k=1}^m (\bar{x}_k^{(j)} - \bar{x}_k)^2}. \tag{9}$$

It is the distance between $\bar{x}^{(j)}$ and \bar{x} , $\|x_i^{(j)} - \bar{x}^{(j)}\|$ is the distance between the i^{th} case and the centre in j^{th} cluster. F-Statistic is a F-distribution which complies with $r-1, n-r$ degrees of freedom. Its molecular indicated that the distance between the different classes, its denominator indicated that the distance between the different cases. That is to say, the value of F-Statistic is bigger, the diversity between the different classes is bigger, and the classes are better.

4 Empirical research

4.1 QUESTIONNAIRE-INVESTIGATION

In order to analyse scientifically the customer demands which influence the e-banking service quality, we investigated 10 commercial banking which include 7000 customers. The investigation cases include students, the workers, senior director, individual operators, retired workers. We use the method of questionnaire-investigation to assess the attention degree of relevant indicators. The investigation questionnaire is mainly focus on 11 dimensional to set which including the appearance x_1 , the information function x_2 , response x_3 , guarantee x_4 , price x_5 , sensitivity x_6 , technical x_7 , convenience x_8 , empathy x_9 , defect degree x_{10} , safety x_{11} , the questionnaire use the answer of yes or no as the choice of customers whether focus on some question, which is convenient for the customers to choose.

The questionnaires were sent by email. Those persons who take part in the questionnaires were provided and contacted by Shanghai Quality Association, because they are familiar with the situation which involved in the questionnaire. A total of 7000 questionnaires were sent to

the users of 10 banks (B_1, B_2, \dots, B_{10}). A total of 5000 questionnaires were responded, in which, there were a total of 4698 valid questionnaires, the response rate was 71.43%, and the effective rate was 67.1%. The relevant problems and measures as shown in Table 2.

TABLE 2 The questionnaire of customer demands which influence the e-banking service quality

| | B_1 | B_2 | B_3 | B_4 | B_5 | B_6 | B_7 | B_8 | B_9 | B_{10} |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|
| x_1 | 274 | 250 | 191 | 245 | 290 | 465 | 257 | 452 | 157 | 323 |
| x_2 | 322 | 285 | 431 | 230 | 309 | 156 | 325 | 363 | 269 | 404 |
| x_3 | 156 | 346 | 287 | 240 | 499 | 221 | 429 | 354 | 407 | 232 |
| x_4 | 409 | 340 | 559 | 277 | 384 | 174 | 397 | 448 | 304 | 516 |
| x_5 | 287 | 305 | 474 | 262 | 325 | 159 | 356 | 379 | 278 | 437 |
| x_6 | 252 | 448 | 496 | 304 | 404 | 197 | 375 | 414 | 404 | 514 |
| x_7 | 304 | 278 | 214 | 266 | 345 | 495 | 294 | 475 | 194 | 351 |
| x_8 | 295 | 443 | 212 | 307 | 259 | 312 | 405 | 220 | 171 | 335 |
| x_9 | 166 | 393 | 311 | 276 | 594 | 231 | 393 | 351 | 421 | 242 |
| x_{10} | 233 | 297 | 401 | 317 | 280 | 268 | 189 | 306 | 332 | 272 |
| x_{11} | 309 | 459 | 221 | 341 | 281 | 339 | 410 | 241 | 174 | 360 |

4.2 DATA NORMALIZATION

From the Equation (1), the Table 1 was transformed using the method of moving-standard deviation normalized, we can obtained the Table 3 as follows:

TABLE 3 Translation standard difference alternate

| | B_1 | B_2 | B_3 | B_4 | B_5 | B_6 | B_7 | B_8 | B_9 | B_{10} |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|
| x_1 | 0.01 | -1.33 | -1.17 | -0.96 | -0.68 | 1.62 | -1.22 | 1.06 | -1.26 | -0.4 |
| x_2 | 0.68 | -0.86 | 0.65 | -1.39 | -0.5 | -1 | -0.31 | -0.01 | -0.14 | 0.43 |
| x_3 | -1.64 | -0.05 | -0.44 | -1.1 | 1.33 | -0.45 | 1.08 | -0.12 | 1.24 | -1.33 |
| x_4 | 1.9 | -0.13 | 1.62 | -0.05 | 0.22 | -0.85 | 0.65 | 1.01 | 0.21 | 1.57 |
| x_5 | 0.19 | -0.59 | 0.98 | -0.48 | -0.34 | -0.98 | 0.1 | 0.18 | -0.05 | 0.76 |
| x_6 | -0.3 | 1.31 | 1.14 | 0.72 | 0.41 | -0.66 | 0.36 | 0.6 | 1.21 | 1.55 |
| x_7 | 0.43 | -0.95 | -0.99 | -0.36 | -0.15 | 1.87 | -0.72 | 1.33 | -0.89 | -0.12 |
| x_8 | 0.3 | 1.25 | -1.01 | 0.81 | -0.98 | 0.32 | 0.76 | -1.73 | -1.12 | -0.28 |
| x_9 | -1.5 | 0.58 | -0.26 | -0.08 | 2.24 | -0.37 | 0.6 | -0.16 | 1.38 | -1.23 |
| x_{10} | -0.57 | -0.7 | 0.42 | 1.1 | -0.78 | -0.05 | -2.13 | -0.7 | 0.49 | -0.92 |
| x_{11} | 0.5 | 1.46 | -0.94 | 1.78 | -0.77 | 0.55 | 0.83 | -1.48 | -1.09 | -0.02 |

From the Equation (2), the Table 2 was transformed using the method of moving-range normalized, we can obtained the Table 4 as follows:

TABLE 4 Translation limit difference alternate

| | B_1 | B_2 | B_3 | B_4 | B_5 | B_6 | B_7 | B_8 | B_9 | B_{10} |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|
| x_1 | 0.47 | 0.00 | 0.00 | 0.14 | 0.09 | 0.91 | 0.28 | 0.91 | 0.00 | 0.32 |
| x_2 | 0.66 | 0.17 | 0.65 | 0.00 | 0.15 | 0.00 | 0.57 | 0.56 | 0.42 | 0.61 |
| x_3 | 0.00 | 0.46 | 0.26 | 0.09 | 0.72 | 0.19 | 1.00 | 0.53 | 0.95 | 0.00 |
| x_4 | 1.00 | 0.43 | 1.00 | 0.42 | 0.37 | 0.05 | 0.87 | 0.89 | 0.56 | 1.00 |
| x_5 | 0.52 | 0.26 | 0.77 | 0.29 | 0.20 | 0.01 | 0.70 | 0.62 | 0.46 | 0.72 |
| x_6 | 0.38 | 0.95 | 0.83 | 0.67 | 0.43 | 0.12 | 0.78 | 0.76 | 0.94 | 0.99 |
| x_7 | 0.59 | 0.13 | 0.06 | 0.32 | 0.26 | 1.00 | 0.44 | 1.00 | 0.14 | 0.42 |
| x_8 | 0.55 | 0.92 | 0.06 | 0.69 | 0.00 | 0.46 | 0.90 | 0.00 | 0.05 | 0.36 |
| x_9 | 0.04 | 0.68 | 0.33 | 0.41 | 1.00 | 0.22 | 0.85 | 0.51 | 1.00 | 0.04 |
| x_{10} | 0.30 | 0.22 | 0.57 | 0.78 | 0.06 | 0.33 | 0.00 | 0.34 | 0.66 | 0.14 |
| x_{11} | 0.60 | 1.00 | 0.08 | 1.00 | 0.07 | 0.54 | 0.92 | 0.08 | 0.06 | 0.45 |

4.3 BUILT THE FUZZY SIMILARITY MATRIX

From the Equation (3), the values in Table 3 were calculated using Matlab software, the fuzzy similar matrix can be obtained as follows: $R(r_{\alpha\beta})_{11 \times 11}$.

quality, which included the appearance, the information function, response, guarantee, price, sensitivity, technical, reliability, empathy, defect degree, safety. Not only classifies the attention degree of relevant indicators which have similarly features into one class, but also distinguishes the optimal classification with F-Statistic. The results of classification fairly showed that the emphases of attention degree are empathy and safety, and the customer demands have an important role to influence the e-banking service quality. From the literature review, we have known that the improved service quality can indirectly increase the profit of commercial banking, so it should be an important input factor for QFD (Quality Function Deployment), and the e-banking should pay attention to its unique advantages and get the trust of customers.

Fuzzy clustering method is a developing analysis approach, and it has a close relationship with computer

software. It is practical in technique to use fuzzy mathematical principles in classification and choosing of the customer demands which influence the e-banking service quality. But there will constantly appear some new and uncertain factors which influences e-banking service quality with the development of science and economy, which will influence the classification results. Therefore, how to make further choice and reduce deviation is valuable for us to study.

Acknowledgments

This paper is supported by National Nature Science Foundation of Chinese (project no. NSFC 71272177 /G020902) and Innovation Program of Shanghai Municipal Education Commission (project no.12ZS101).

References

- [1] Lina L 2013 The traditional modes of banks development are not to continue *China Finance* **04** 68-70
- [2] Claes F, Johnson M D, Anderson E W, Cha J, Bryant B E 1996 The American customer satisfaction index: nature, purpose and findings *The Journal of Marketing* 7-18
- [3] Chunxiao W, Xiaoyun H, Biyan W 2003 An Empirical Study of the Relationship between Customer Satisfaction and Loyalty *Nankai Business Review* **6**(4) 70-4
- [4] Reichheld F F 1996 The Loyalty Effect *Harvard Business School Press*
- [5] Liu X, Chen Y 2013 A FAHP-FUZZY Approach of Evaluating Banking Service Quality *International Journal of Business and Management* **8**(14) 158-67
- [6] Davis F D 1986 A technology acceptance model for empirically testing new end-user information systems: Theory and results *Massachusetts Institute of Technology*
- [7] Karjaluoto H, Mattila M, Pentto T 2002 Electronic banking in Finland: consumer beliefs and reactions to a new delivery channel *Journal of Financial Services Marketing* **6**(4) 346-61
- [8] Minocha S, Millard N, Dawson L H 2003 Integrating customer relationship management strategies in (B2C) e-commerce environments *Proceedings of the INTERACT 2003 Conference* 2003 335-42
- [9] Mills A, Tennant V, Mansingh G, Rao-Graham L 2013 Internet Banking: Enablers and Inhibitors for Developing Economies – A Study of Potential Users in Jamaica *Proceedings of the 19th Americas Conference on Information Systems(AMCIS 2013)* Chicago Illinois August 1-9
- [10] Lee M C 2009 Factors influencing the adoption of internet banking: An integration of TAM and TPB with perceived risk and perceived benefit *Electronic Commerce Research and Applications* 2009 **8**(3) 130-41
- [11] Jabnoun N, Al-Tamimi H A H 2003 Measuring perceived service quality at UAE commercial banks *International Journal of Quality & Reliability Management* **20**(4) 458-72
- [12] Sadiq Sohail M, Shanmugham B 2003 E-banking and customer preferences in Malaysia: an empirical investigation *Information Sciences* **150**(3) 207-17
- [13] Li S C Y, Worthington A 2004 The relationship between the adoption of Internet banking and electronic connectivity: an international comparison Queensland University of Technology School of Economics and Finance
- [14] Chu P Y, Lee G Y, Chao Y 2012 Service quality, customer satisfaction, customer trust, and loyalty in an e-banking context *Social Behavior and Personality* **40**(8) 1271-83
- [15] Wei Z, Jinchen Z, Hao Z 2007 Key Dimensions of E-service Quality to Internet Banking *Modern Finance and Economics* **10** 13-19
- [16] Gharakhani D, Eslami J 2012 Determining customer needs priorities for improving service quality using QFD *Management* **1**(6) 21-8
- [17] Bezdek J C 1981 Pattern recognition with fuzzy objective function algorithms *Kluwer Academic Publishers*
- [18] Grabisch M, Nguyen H T, Walker E A 1995 Pattern Recognition and Computer Vision *Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference* Springer Netherlands 261-92
- [19] Wuqi L 2004 Research on Fuzzy Cluster analysis and application in data mining *Journal of Anqing Teachers College Natural Science Edition* **10**(2) 65-7
- [20] Jinglong W 2008 Multi-variate statistical analysis Beijing Science Press

Authors



Xilong Liu, born on August 11, 1977, Qingdao, China

Current position, grades: Ph.D. student of School of Management, Shanghai University.
University studies: School of Economics and Management, University of Science and Technology Beijing (USTB).
Scientific interest: Service quality of e-banking.
Publications: 3 papers.



Yizeng Chen, born on September 2, 1971, Mudanjiang, China

Current position, grades: professor of school of management, Shanghai University.
University studies: mechanical and electronic engineering in Northeastern University.
Scientific interest: The fuzzy quality function deployment (QFD) modelling, optimization research.
Publications: More than 40 papers.

An autonomous decision making algorithm applied for the evaluation of power quality

Yun-jun Yu*, Sui Peng, Yun-tao Xue, Chao Tong, Zi-heng Xu

College of Information Engineering, Nanchang University 330031, China

Received 1 May 2014, www.tsi.lv

Abstract

An autonomous decision making algorithm applied for the evaluation of the field power quality is proposed. This algorithm can reflect to the characteristics of evaluation objects, develop evaluation objects initiatives, weakens the influence of the subjective weight of index on evaluation results and implements the comparison of different power qualities of the assessed in the area. The paper introduces the implementation steps of autonomous decision making algorithm, analyses the competition scope of the power quality of the assessed with this algorithm. The competition model is established, which output the comprehensive evaluation results of the assessed. The simulation demonstrates the effectiveness and practicability of this method.

Keywords: power quality, autonomous decision, algorithm, evaluation

1 Introduction

With the increasing concern about power quality, many power quality assessment methods have been put forward. Actually, we often need to assess the power quality in a certain region, so as to compare the result with those of other parts in that region. At present, many assessment methods about power quality have been proposed by science researchers, including Entropy Weight Analysis [1], Neural Network Analysis [2], Extentics Cloud Theory [3], Fuzzy Mathematics [4, 5], Analytic Hierarchy Process [6, 7]. Such methods are easy to use and simple in principle [8]. But for the determination of subjective weights of the evaluation indexes, some controversies are still needed to be resolved:

- 1) When assessing power quality, the characteristics of the objects have been omitted [9];
- 2) Subjective preferences influence the weights determination of the assessment criteria [10].

In allusion to the problems mentioned above, an autonomous decision making algorithm applied for the evaluation of the field power quality is proposed in the paper. This algorithm can reflect the characteristics of evaluation objects, develop evaluation objects initiatives, weakens the influence of the subjective weight of index on evaluation results, which maintains algorithm implementation steps, Evaluation model is established. Simulation results illustrate the efficiency of this autonomous decision making algorithm.

2 Algorithm

2.1 DESIGN STEPS OF THE ALGORITHM

- 1) Decide the technical and non-technical indicators of the electric energy first.
- 2) Convert the historical energy data into a matrix with certain rules.
- 3) Through the judgment of property weight value, get the non-authoritarian conditions to constrain weight.
- 4) Establish the optimized rules of competition range to reflect the autonomous decision-making of evaluators.
- 5) Construct models in accordance with the above rules, and work out the competitive range of the evaluate.
- 6) Finally, obtain the evaluated conclusions which can reflect the independently decision-making of evaluate.

2.2 ESTABLISH THE RULES OF THE POWER QUALITY MATRIX

- 1) The indexes of power quality are considered in the paper is Frequency offset, Voltage deviation, Short-term flicker, Unbalanced three-phase, Harmonic distortion rate, Power supply reliability, Long flicker [11].
- 2) Using the method of extremum of historical electricity data regularization processing rules matrix A Maximization of index formula:

$$x_{ij_M} = \frac{M_j - x_{ij}}{M_j - m_j}, \quad (1)$$

Minimization of the index formula:

$$x_{ij_m} = \frac{x_{ij} - m_j}{M_j - m_j}, \quad (2)$$

* *Corresponding author* e-mail: yuyunjun@ncu.edu.cn

x_{ij_M} - maximization of index; x_{ij_m} - minimization of the index; x_{ij} - the assessed sample; M_j - maximum value in sample; m_j - the minimum value in sample.

2.3 COMPETITION MODEL

2.3.1 Condition of weights nondictatorship

Condition of weights nondictatorship refers to the circumstance when any indicator does not play a leading role in terms of all the relatively insignificant indicators within domain [12, 13]. M represents the total number of samples, and $x \in M, w_j$ denotes the weight of x_j . When the condition that $w_1 \geq w_2 \geq w_3 \geq \dots \geq w_m$ is satisfied, the satisfiable Equation (3) is named condition of weak weights nondictatorship [14].

$$w_j \in [0.5^{m-1}, 0.5], \tag{3}$$

and the satisfiable Equation (4) is named condition of strong weights nondictatorship:

$$\begin{cases} w_1 + w_2 + w_3 + \dots + w_m = 1 \\ 0.5^{m-1} \leq w_m \leq 0.5 \\ w_1 \leq w_2 + w_3 + \dots + w_m \\ w_2 \leq w_3 + \dots + w_m \\ \dots \\ \dots \\ w_{m-1} \leq w_m \end{cases} \tag{4}$$

When weights are assigned to the maximum indicator value after being sorted by indicator value as far as possible under constraint condition, evaluation value would reach the maximum; when weights are assigned the other way around, evaluation value would reach the minimum. Based on such principles, the following conclusions are drawn: under condition of weak weights nondictatorship, comprehensive evaluation value y_i achieves its maximum and minimum value, when the optimal descending weight vector and the worst descending weight vector are w_{i_u} and w_{i_f} respectively, and it can be seen in Equations (5), (6) under condition of strong weights nondictatorship, the comprehensive evaluation value y_i achieves its maximum and minimum value when the optimal descending weight vector and the worst descending weight vector are w_{i_u} and w_{i_f} respectively, and it can be seen in Equations (7), (8).

$$w_{i_u} = (0.5, 0.5 - (m-2)0.5^{m-1}, 0.5^{m-1}, \dots, 0.5^{m-1}), \tag{5}$$

$$w_{i_f} = (0.5^{m-1}, 0.5^{m-1}, \dots, 0.5 - (m-2)0.5^{m-1}, 0.5), \tag{6}$$

$$w_{u_u} = (0.5^1, 0.5^2, 0.5^3, \dots, 0.5^{m-2}, 0.5(1 - \sum_{j=1}^{m-2} 0.5^j)), \tag{7}$$

$$w_{u_f} = (0.5(1 - \sum_{j=1}^{m-2} 0.5^j), 0.5^{m-2}, \dots, 0.5^3, 0.5^2, 0.5^1). \tag{8}$$

The research question in this paper using two rules mentioned above to get the comprehensive evaluation value range Y under certain conditions.

2.3.2 Optimality principle of competition range

Competition range refers to the set of objects evaluated and those who may compete with the evaluated. In the above, the object evaluated can be represented by $r(r \in N)$, where N represents the total number of the objects evaluated. When the following two conditions are met, the objects will lie in the competition range.

- 1) $x_{ik} \geq x_{jk} (i, j \in N, i \neq j)$ is false. As $k \in M$ for at least one k there is strict inequality established.
- 2) $Y_i \cap Y_j \neq \emptyset$, where Y_i refers to the value range of power quality evaluation of r_i , and Y_j refers to that of r_j .

From the optimality principle of competition range, it can be found that all the objects evaluated want to enhance their competitiveness and weaken that of their competitors. When competition range is C :

$$C = \{r_1^i, r_2^i, r_3^i, \dots, r_m^i\}, N_i = (1, 2, 3, \dots, n_i),$$

$$i \in N, x_{ij}^i \in x_j(r_i^i), j \in M, l \in N_i$$

the expected weight value w_i is the solution of the following multi-objective linear programming.

$$\begin{aligned} \max & \left[\gamma_1 \sum_{j=1}^m x_{ij} w_j^i - \gamma_2 \sum_{j=1}^m u_l^i \sum_{j=1}^n x_j^i w_j^i \right], \\ & \sum_{j=1}^m w_j^i = 1, w_j^i \geq 0, w_j \in \alpha, j \in M, l \in N_i, \end{aligned} \tag{9}$$

α represents the constraint set of weight vector w_i meeting with weights non-dictatorship condition; $\gamma_1 + \gamma_2 = 1$, among them γ_1 refers to the weight coefficient of enhancing their own competitiveness, γ_2 that of weakening the competitiveness of their competitor u_l^i is the competitiveness-focused coefficient of r_i to competitiveness evaluation objects r_l^i in competition range C ; $\sum_{l=1}^{n_i} u_l^i = 1$ stands for the competitiveness-focused coefficient vector of $u_i = (u_1^i, u_2^i, u_3^i, \dots, u_{n_i}^i)$.

Therefore, the value range of r_i is $Y_i = [y_i^L, y_i^U]$ and that of r_i^i is $Y_i^i = [y_{ii}^L, y_{ii}^U]$.

If $Y_i \cap Y_i^i \neq \emptyset$ and C_i is the competitive interval, $C_i = Y_i \cap Y_i^i$.

If d_{ii} is the competitive intensity of r_i and r_i^i ,

$$d_{ii} = \frac{e(Y_i \cap Y_i^i)}{e(Y_i \cup Y_i^i)};$$

competitiveness-focused-coefficient $u_i^i = dil / \sum_{l=1}^n d_l^i$; and

stands for the calculation function of interval width.

In the constraint set of weight vector w_i meeting with weights non-dictatorship condition, the above method can be adopted to solve the linear programming problem and the weight vector in the weights non-dictatorship condition. In the same way, the weight vectors in the competitive view of other objects can be determined and

finally the optimal weight vector matrix W can be combined.

2.4 ESTABLISH COMPREHENSIVE EVALUATION MATRIX

Comprehensive evaluation matrix B :

$$B = AW^T. \tag{10}$$

Based on comprehensive evaluation matrix B , comprehensive evaluation conclusion m is proposed:

$$m = [m_1, m_2, m_3, \dots, m_m].$$

3 Case study

This paper conducts simulation based on the monitoring data, which is provided in reference [15], of eight quality indexes of power in the 12 months of 2009 from six 220kV transformer substations of the Power Supply Bureau of some region. The regularization matrix obtained according to Equations (1) and (2) is as shown in Table 1.

TABLE 1 The regularization matrix of power quality indices at each spot

| Monitoring point | Frequency offset | Voltage deviation | Short-term flicker | Unbalanced three-phase | Harmonic distortion rate | Power supply reliability | Long flicker |
|------------------|------------------|-------------------|--------------------|------------------------|--------------------------|--------------------------|--------------|
| 1 | 0 | 0.53 | 0.92 | 0.49 | 0.00 | 0.50 | 0.90 |
| 2 | 1 | 0.28 | 1.00 | 0.28 | 0.58 | 0.00 | 1.00 |
| 3 | 0 | 1.00 | 0.47 | 0.70 | 1.00 | 1.00 | 0.44 |
| 4 | 1 | 0.00 | 0.75 | 0.00 | 0.00 | 0.00 | 0.56 |
| 5 | 1 | 0.00 | 0.75 | 0.84 | 0.84 | 0.49 | 0.69 |
| 6 | 0 | 0.10 | 0.58 | 1.00 | 1.00 | 0.50 | 0.50 |

According to the competition model of autonomous decision-making, regularized matrix is given. Under the condition of non-dictatorship weight, the optimized competition model is built. Specific steps are as follows.

- 1) Selecting the non-dictatorship weight as the rule for autonomous decision-making algorithm.
- 2) According to the Equation (7), optimal descending weight vector w_{u-u} is:

$$w_{u-u} = (0.5, 0.25, 0.125, 0.0625, 0.03125, 0.015625)$$

- 3) Comprehensive evaluation value range of the assessed $r_1, r_2, r_3, r_4, r_5, r_6, \dots$:

$$Y_1 = [0.13, 0.83],$$

$$Y_2 = [0.10, 0.92],$$

$$Y_3 = [0.22, 0.94],$$

$$Y_4 = [0.04, 0.79],$$

$$Y_5 = [0.29, 0.88].$$

- 4) Competition intensity matrix D :

$$D = \begin{bmatrix} 0.75 & 0.71 & 0.84 & 0.93 & 0.83 & 0 & 0.77 & 0.59 \\ 0.84 & 0.71 & 0 & 0.93 & 0 & 0.84 & 0.72 & 0.67 \\ 0 & 0.81 & 0.84 & 0.80 & 0.77 & 0.75 & 0.57 & 0.76 \\ 0.62 & 0.58 & 0 & 0.77 & 0 & 0.83 & 0 & 0.48 \\ 0.81 & 0 & 0.71 & 0.77 & 0.58 & 0.71 & 0.52 & 0.83 \\ 0.81 & 0.77 & 0.83 & 0 & 0.77 & 0.93 & 0.92 & 0.64 \end{bmatrix}$$

- 5) The competitiveness-focused coefficient vector U

$$U = \begin{bmatrix} 0.14 & 0.13 & 0.16 & 0.17 & 0.15 & 0 & 0.14 & 0.11 \\ 0.18 & 0.15 & 0 & 0.18 & 0 & 0.18 & 0.16 & 0.15 \\ 0 & 0.16 & 0.16 & 0.16 & 0.12 & 0.14 & 0.11 & 0.15 \\ 0.19 & 0.18 & 0 & 0 & 0 & 0.25 & 0 & 0.16 \\ 0.16 & 0 & 0.14 & 0.14 & 0.12 & 0.16 & 0.11 & 0.16 \\ 0.14 & 0.13 & 0.15 & 0.15 & 0.13 & 0 & 0.16 & 0.11 \end{bmatrix}$$

6) Assuming that target coefficient $\gamma_1 = \gamma_2 = 0.5$, the optimal weight vector matrix W is:

$$W = \begin{bmatrix} 0.008 & 0.063 & 0.250 & 0.015 & 0.008 & 0.031 & 0.500 & 0.125 \\ 0.500 & 0.031 & 0.125 & 0.016 & 0.063 & 0.008 & 0.250 & 0.008 \\ 0.008 & 0.500 & 0.008 & 0.063 & 0.125 & 0.250 & 0.016 & 0.031 \\ 0.500 & 0.008 & 0.250 & 0.016 & 0.063 & 0.008 & 0.125 & 0.031 \\ 0.500 & 0.008 & 0.250 & 0.063 & 0.016 & 0.008 & 0.125 & 0.031 \\ 0.008 & 0.008 & 0.031 & 0.500 & 0.063 & 0.063 & 0.016 & 0.250 \end{bmatrix}$$

7) $B=AW^T$, comprehensive evaluation matrix B is

$$B = \begin{bmatrix} 0.458 & 0.404 & 0.376 & 0.488 & 0.381 & 0.825 & 0.289 & 0.557 \\ 0.259 & 0.904 & 0.925 & 0.271 & 0.918 & 0.784 & 0.367 & 0.179 \\ 0.940 & 0.257 & 0.284 & 0.650 & 0.271 & 0.491 & 0.772 & 0.852 \\ 0.080 & 0.765 & 0.761 & 0.095 & 0.785 & 0.479 & 0.245 & 0.071 \\ 0.304 & 0.862 & 0.830 & 0.740 & 0.853 & 0.660 & 0.656 & 0.426 \\ 0.374 & 0.309 & 0.278 & 0.827 & 0.301 & 0.542 & 0.745 & 0.430 \end{bmatrix}$$

8) Comprehensive evaluation conclusion m is:

$$m = [0.324, 0.393, 0.192, 0.282, 0.444, 0.229]$$

9) Comprehensive evaluation conclusion is shown in Table 2:

TABLE 2 The result of evaluation of the substations

| Monitoring point | r ₁ value ranges | r ₂ value ranges | r ₃ value ranges | r ₄ value ranges | r ₅ value ranges | r ₆ value ranges | comprehensive ordering |
|------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|------------------------|
| 1 | 1 | 5 | 5 | 4 | 2 | 4 | 3 |
| 2 | 5 | 1 | 4 | 5 | 4 | 2 | 4 |
| 3 | 3 | 6 | 1 | 3 | 6 | 5 | 5 |
| 4 | 6 | 3 | 6 | 1 | 3 | 6 | 6 |
| 5 | 4 | 2 | 2 | 6 | 1 | 3 | 2 |
| 6 | 2 | 4 | 3 | 2 | 5 | 1 | 1 |

From Table 2, it can be seen that

- 1) the entire accessed can give full play to his superiority within his own competitive scope, getting the highest rank among comparing matters and attaining the goal of weakening the competitors;
- 2) the Comprehensive Assessment Table is relatively fair, which is able to stand for the features of every competitor.

4 Conclusions

An autonomous decision making algorithm applied for the evaluation of the field power quality is proposed. Using of autonomous decision-making algorithm, the

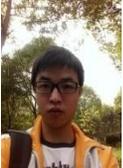
competitive interval and view of power quality the assessed are analysed and an evaluation model is established; finally rational sorting results of the assessed are obtained. Results of the simulation show that the autonomous decision making algorithm can incarnate the autonomous action of the assessed and both fairness and effectiveness of evaluation results can be ensured.

Acknowledgement

This work is supported by International Science and Technology Cooperation Project (2014DFG72240), and Jiangxi Science Support Project (2013BBE50102)

References

- [1] Zhengyuan J, Liang Z 2010 Comprehensive evaluation of power quality based on the model of entropy weight and unascertained measure *Power System Protection and Control* 38(15) 33-7 (in Chinese)
- [2] El-Zonkoly A M 2005 Power system model validation for power quality assessment applications using genetic algorithm *Expert Systems with Applications* 29(4) 941-4
- [3] Li R, Su H 2012 A synthetic power quality evaluation model based on extension cloud theory *Automation of Electric Power Systems* 36(1) 66-70
- [4] Farghal S A, Kandil M S, Elmitwally A 2002 *IEEE Proceedings of Generation: Transmission and Distribution* 149(1) 44-4
- [5] Huihui T, Jianchun P 2003 Research on Synthetic and Quantificated Appraisal Index of Power Quality Based on Fuzzy Theory *Power System Technology* 27(12) 85-8
- [6] Chatzimouratidis A I, Pilavachi P A 2008 Multicriteria evaluation of power plants impact on the living standard using the analytic hierarchy process *Energy Policy* 36(3) 1074-89
- [7] Tian W, Bai J, Meisun H 2013 Application of the analytic hierarchy process to a sustainability assessment of coastal beach exploitation *Journal of Environment Management* 115(30) 251-8
- [8] Lei C 2005 Discussion about the Methods of Evaluating Power Quality *Electrotechnical Journal* 24(1) 58-62 (in Chinese)
- [9] Yangsen Q, Yijian L, Yang L 2012 An Autonomous Decision-Making Based Method for Comprehensive Evaluation of Power Quality and Its Application *Power System Technology* 36(2) 70-4 (in Chinese)
- [10] Lei C, Yonghai X 2005 Discussion about the methods of evaluation power quality *Electric Application* 24(1) 58-65
- [11] Qingquan J, Jiahua S, Hua L 2000 Quality of Electricity Commodity and Its Fuzzy Evaluation *Power System Technology* 24(6) 46-9 (in Chinese)
- [12] Ortega M J, Hernandez J C, Garcia O G 2013 Measurement and assessment of power quality characteristics for photovoltaic systems: Harmonics, flicker, unbalance and slow voltage variations *Electric Power System Research* 96(2) 3-35 (in Chinese)
- [13] Pingtao Y, Yajun G 2007 Multi-attribute decision-making method based on competitive view optimization under condition of weights nondictatorship *Control and Decision* 22(11) 1259-68 (in Chinese)
- [14] Yi Z, Ning C, Longyun H 2010 Evaluation of bulk cargo transportation organization competitiveness of the Yangtze river based on self-determination decision-making *Journal of Wuhan University of Technology Transportation Science & Engineering* 34(6) 1185-9 (in Chinese)
- [15] Jinjing Y 2012 Research of power quality dynamic comprehensive evaluation method on regional power grid *Guangzhou, China: South China University of Technology (in Chinese)*

| Authors | |
|--|---|
|  | <p>Yunjun Yu, born on June 28, 1978, Jiangxi, China</p> <p>Current position: doctor, lecture at Nanchang University Nanchang, Jiangxi, China. University studies: automation in Nanchang University. Scientific interest: photovoltaic inverter, fault diagnosis, micro-grid. Publications: 30 papers. Experience: Ph.D degrees from Chinese Academic of Science, 2013.</p> |
|  | <p>Sui Peng, born on November 15, 1992, Jiangxi, China</p> <p>Current position: junior student in NanChang university, Jiangxi, China. University studies: NanChang university, Jiangxi, China. Scientific interest: power system and its automation.</p> |
|  | <p>Yuntao Xue, born on December 18, 1991, Henan, China</p> <p>Current position: junior student in NanChang university, Jiangxi, China. University studies: NanChang university, Jiangxi, China. Scientific interest: power system and its automation.</p> |
|  | <p>Tong Chao, born on July 14, 1992, Jiangxi, China</p> <p>Current position: junior student in NanChang university, Jiangxi, China. University studies: NanChang university, Jiangxi, China. Scientific interest: power system and its automation.</p> |
|  | <p>Xu Ziheng, born on September 20, 1992, NanChang, China</p> <p>Current position: junior student in NanChang university, Jiangxi, China. University studies: computer science and technology in School of Electronics and Information Engineering. Scientific interest: optimization and application of the algorithm.</p> |

A simulation model on the formation of knowledge-based collaborative networks

Shanshan Shang^{1*}, Jianxin You²

¹*College of International Business, Shanghai International Studies University, No. 550, Dalian Road, Shanghai, China*

²*School of Economics and Management, Tongji University, No. 1239, Siping Road, Shanghai, China; School of Management, Shanghai University, No. 99, Shangda Road, Shanghai, China*

Received 1 March 2014, www.tsi.lv

Abstract

Collaborative network has been a hot topic in the related research field. This paper proposes a simulation model on the formation of knowledge-based collaborative networks mainly based on the Set theory. The paper proposes that formation process as follows: (1) find the key skills and the core members; (2) classify the organizations; (3) establish the relationship between organizations in different classifications.

Keywords: Knowledge-Based, Collaborative Networks, Set theory

1 Introduction

Collaboration between companies in collaborative networks has been widely accepted as an effective approach to cope with the challenges [1]. In order to be competitive, companies need to decrease their product's time-to-market, share information, and shift from standardization to a customization approach [2]. Rapid changes in technology often force such firms to depend on external technological knowledge and skills in addition to internal technological resources. Many firms today are relying more extensively on external linkages to acquire new technological knowledge using strategies such as technology licensing and collaborative agreements. Inter-firm collaboration is an important vehicle for the creation of technological competencies [3]. Collaborative Networks have emerged as a new and prominent paradigm to improve organizations competitiveness in a sustainable way in the increasing globalised and dynamic businesses [4]. Therefore, research on collaborative models such as collaborative networks has attracted more and more attention from experts and scholars. And also the concept of collaborative networks has risen as an organizational alternative in order to fast react to market changes and turbulences associated to the globalised economy [5].

Researchers focus on the topic of collaborative networks mainly from the perspective of motives for the collaboration, evaluation on the impact of different types of collaborative networks on product innovation performance, and value systems in collaboration networks [6-8]. However, the diffusion of knowledge and its effect on innovation is of major importance to ensure productivity growth, thus, this paper mainly talks about

the formation of the collaboration networks from the perspective of knowledge, for network structure impacts the function of the community, improving or impeding the flow of information and ideas, opinion formation, and the spread of effective technologies.

2 Collaborative networks

A Collaborative network is that business entities work collaboratively to support the different processes and activities [9]. A Collaborative network are the entities which are geographically distributed or heterogeneous with respect to their operating environment, culture, social capital and goals collaborate to achieve common goals, supported by Information and Communications Technologies [10]. The collaborative network consists of heterogeneous and autonomous partners and this business model permits the rapid integration of competencies to establish an experience-centric network. Within the collaborative network each member has its own core values and the success of the collaboration network is the appropriate alignment of these values amongst the partners [11].

The purpose of building a collaboration network is to benefit from the inter-organization links that connects people and knowledge from diverse fields [12]. It is obvious that networks hold many different characteristic, which make different forms of networks suited for very different purposes and functions [13]. There is no universal network-model that fits all collaborative purpose and suitable to all situation. However, the core factors that affect the design of the collaborative project and the way it is carried out are the size of the collaborative network measured by number of active

* *Corresponding author* e-mail: sss336699@hotmail.com

participants and the proximity of partners in relation to geographical and disciplinary scope [14]. The large scale and very diverse networks are especially well suited for projects with the aim of searching for new knowledge, exploring new collaborative opportunities, or creating associations [15]. However, it also needs cross-unit coordination activities to keep the network parts together, which requires strong management, and clear structures of the network. Therefore, Large scale network have the advantages of easier knowledge search for the pool of knowledge to search from is more diverse and easier exploration activities, with the disadvantages of easier for partners to violate an obligation to provide resources, management challenges, hard to get rid of non-performers. To the contrast, small scale networks have the advantages of easier to build trust, easier knowledge transfer and easier exploitation activities, with the disadvantages of redundant partner knowledge and difficult to ensure a diverse pool of knowledge [16].

Knowledge Networks is defined as [17] “A Knowledge Network signifies a number of people and resources, and the relationships between them that are able to capture, transfer and create knowledge for the purpose of creating value. An Integrated Knowledge Network spans all domains communities, and trust relationships with the goal of fostering sustainable innovation that will continue to promote the competitiveness of its users.” Each member in the network will have impacts on the success of innovation projects by knowledge sharing and collaboration [18].

3 The simulation model of knowledge-based network formation

The formation of a knowledge-based collaborative network requires collaborative network members have access to both internal and external knowledge resources. So the structures of collaborative networks differ markedly according to the characteristics [9].

3.1 PROBLEM DESCRIPTION

Suppose individual or organization possess some kind of skills, but in order to complete the task or accomplish the goal, they skills they have is not enough, they need to collaborate together to form a network, so it is important to define the formulation of the network. Therefore, the problem can be described as: the input is the skills possessed by different individuals or organizations, the output is the network, and the important point is how to choose and organize these organizations to formulate an effective network.

$X = \{x_1, x_2, \dots, x_N\}$ denotes the set of the individuals or the organizations, and $i=\{1, 2, 3, \dots, N\}$ means the i^{th} individual or organization. $S=\{s_1, s_2, \dots, s_M\}$ denotes the finite set of all skills. An individual i 's skill set is the subset of those skills she possesses, $S_i \subseteq S$.

Each individual or organization is endowed with a copy of a problem requiring a subset of the skills.

3.2 SKILL REDUCTION AND CORE MEMBER

The organizations and the skills they possessed can be seen as a whole knowledge system, the organizations are the objects in the system, and the skills that are needed to complete the task are the attributes.

Definition 1: (U, A, F, V) is a knowledge system, U means the object

$U = X = \{x_1, x_2; \dots, x_N\}$ denotes the set of the individuals or the organization;

A is the attribute set, $A=\{a_1, a_2, a_3, \dots, a_m\} \subseteq S$ means that in order to complete the task, the skills that are needed. F is the information function set;

$F=\{f_{a1}, f_{a2}, f_{a3}, \dots, f_{am}\}$, for each $a_i \in A$, f_{a_i} is a mapping function from U to V_{a_i} , that is, $f_{a_i}: U \rightarrow V_{a_i}$

V_{a_i} is the range of attribute a_i ($1 \leq i \leq m$), $V=\{(v_{a1}, v_{a2}, v_{a3}, \dots, v_{am}) \mid v_{a_i} \in V_{a_i}, 1 \leq i \leq m\}$, here we define that $v_{a_i}=\{0, 1\}$.

Therefore, actually a knowledge system is a data table, in which columns are labelled by attributes while rows are labelled by objects. For example, suppose a task need a_1, a_2, a_3, a_4, a_5 five kind of skills, and $x_1 \sim x_6$ organizations are going to collaborate so that to complete the task, if x_i owns the skill a_j , then the value in the convergence of the x_i row and the a_j column will be 1, otherwise 0, as shown in table 1.

TABLE 1 A knowledge system

| U | a ₁ | a ₂ | a ₃ | a ₄ | a ₅ |
|----------------|----------------|----------------|----------------|----------------|----------------|
| x ₁ | 1 | 0 | 0 | 1 | 0 |
| x ₂ | 0 | 1 | 0 | 0 | 1 |
| x ₃ | 1 | 0 | 1 | 0 | 1 |
| x ₄ | 1 | 0 | 0 | 1 | 0 |
| x ₅ | 0 | 1 | 1 | 1 | 0 |
| x ₆ | 0 | 1 | 0 | 0 | 1 |

Base on the knowledge system, in order to get the key skills, algorithm as follows is taken use of.

The reduction algorithm:

Step 1: calculate the matrix $M_{n \times n}$

$$M_{n \times n} = (c_{ij})_{n \times n} = \{\alpha \mid (\alpha \in A) \wedge (f_{\alpha}(x_i) \neq f_{\alpha}(x_j))\}, \forall i, j = 1, 2, 3, \dots, n$$

n is the total number of objects in U , that is, $n = |U|$;

Step 2: for all $c_{ij} \neq \Phi$, get the disjunctive normal form

$$L_{\wedge(\vee)} = \bigwedge_{\forall c_{ij} = \alpha(x_i, x_j) \neq \Phi \in M_{n \times n}} \alpha(x_i, x_j)$$

Step 3: convert the disjunctive normal form to conjunctive normal form $L_{\vee(\wedge)} = \bigvee_{L_k \neq \Phi} L_k$

Step 4: get the RED(C) = $\{L_k \mid \forall L_k \in L_{\vee(\wedge)}\}$

From the algorithm above, we can get the key skills in the knowledge system, and the organizations who own the key skills will be the core numbers in the network.

3.3 NODE CLASSIFICATION

Each organization can be seen as a node in the network. The paper classified the organizations according to their attributes which mean the skills.

Definition 2: Assume R is a equivalence relation on the non-empty finite set U, for $\forall x \in U, [x]_R = \{y \mid yRx\}$, $[x]_R$ is a classification of U according to relation R.

Definition 3: $|S|$ means the number of elements in the set S.

Definition 4: Suppose S is the object set including n subsets which are represent by $C_1, C_2, C_3, \dots, C_n$, then the entropy of S is

$$entropy(S) = -\sum_{i=1}^n p_i \log_2 p_i \tag{1}$$

p_i means the probability of C_i , that is $p_i = \frac{|C_i|}{|S|}$

Definition 5: Suppose S is partitioned into m subsets by attribute A, then

$$entropy(S, A) = \sum_{i=1}^m \frac{|S_i|}{|S|} entropy(S_i) \tag{2}$$

$$entropy(S_i) = -\sum_{j=1}^n \frac{|S_i \cap C_j|}{|S_i|} \log_2 \frac{|S_i \cap C_j|}{|S_i|} \tag{3}$$

$$gain(S, A) = entropy(S) - entropy(S, A) \tag{4}$$

S_i is the i^{th} partition subset of set S.

Therefore, the larger $gain(S, A)$ is, the more important attribute A is.

Then the paper classified the organizations by the following steps:

Step 1: For all the $a_i \in A$, each a_i is seen as a set A, and calculate $gain(U, A)$ according to definition 5 so that to get the relative importance of the skill, and arrange A in descending order according to the relative importance, that is, after rearrangement, for $A = \{a_1, a_2, a_3, \dots, a_m\}$, $gain(U, a_1) \geq gain(U, a_2) \geq gain(U, a_3) \dots \geq gain(U, a_m)$;

Step 2: According to the skill reduction algorithm introduced in 3.2, get the key skills set $K = \{k_1, k_2, \dots, k_s\}$, $K \subseteq A$;

Step 3: Get the partition of U according to attribute K, $PU = \{SPU_1, SPU_2, SPU_3, \dots, SPU_h\}$, SPU_i is the i^{th} partition subset of set U;

Step 4: Calculate $SKN(SPU_i)$, which is the number of skills owned by the members in $SPU_1 \sim SPU_h$, arrange $SPU_1 \sim SPU_h$ in descending order according to

$SKN(SPU_i)$, that is, if the members in PU_i own skill more than PU_j , then $i < j$. For example, according to table 1, if key skill set $K = \{a_1, a_4\}$, then $SPU_1 = \{x_1, x_4\}$ because x_1 and x_4 owns both a_1 and a_4 skill, $SKN(SPU_1) = 2, |K| = 2$;

Step 5: $A' = A - K$, and $A' = \{A_1', A_2', \dots, A_f'\}$, A' is also in descending order according to the relative importance of skills. Let $A_0' = K$, then $|A_i'| = |A_{i-1}'| + k$, k is a constant set by people, and $\sum_{i=1}^f |A_i'| = |A - K|$.

Step 6: Get each partition of U according to A_1', A_2', \dots, A_f' , that is according to A_1' get a partition PU_1' of U, according to A_2' get a partition PU_2' of U, ..., total get f partition.

3.4 NODE DISTANCE

The paper adopts Euclidian distance to calculate the node distance. For the m dimensional space, the Euclidian distance is

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \tag{5}$$

For the organizations as the objects, each object can be seen as a vector, and the attributes can be seen as its dimensions. So each organization is a m dimensional vector, the i^{th} organization in the form of vector is $x_i = [a_{1i}, a_{2i}, a_{3i}, \dots, a_{mi}]$, and the distance between x_i and x_j is

$$d(x_i, x_j) = \sqrt{(a_{1i} - a_{1j})^2 + (a_{2i} - a_{2j})^2 + \dots + (a_{mi} - a_{mj})^2} \tag{6}$$

Obviously, from the Euclidian distance, we can see that the more different skill owned by the two organizations, the larger Euclidian distance is.

3.5 EDGE GENERATION

The relationship between the organizations can be represented by edges between nodes. So how to establish the relationship between the organizations so that to form the network is a quite important issue.

The core members in the network who own the key skills should first establish some relationship with other members who own the skills the core member don't have. Actually, the core members may in the same classification, so the relation establishment is between the members in different classification, as shown in figure 1.

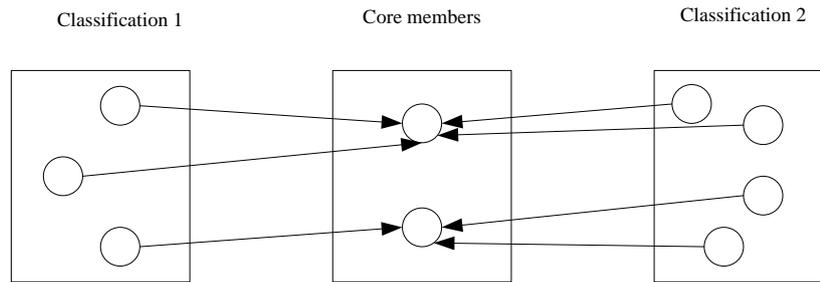


FIGURE 1 Relationship establishment

The relationship establishment between organizations, in other words, the edge generation between nodes, follows the steps below:

Step 1: Get all the node classification $PU, PU_1', PU_2', PU_3', \dots$, according to the algorithm introduced in 3.3.

$$PU = \{SPU_1, SPU_2, SPU_3, \dots, SPU_h\},$$

$$PU_1' = \{SPU_{11}', SPU_{12}', SPU_{13}', \dots\},$$

$$PU_2' = \{SPU_{21}', SPU_{22}', SPU_{23}', \dots\},$$

.....

Step 2: Set $PUM = PU, PUM = \{PUM_1, PUM_2, PUM_3, \dots\}$

If the classifications in PUM cannot cover all the skills, for $j=1, 2, 3, \dots$, once at a time, $PUM = \cup (PUM_i \cap SPU_j')$.

NOTES:

1. the skill of each classification is the least skills owned by the members in the classification, for example, classification 1 have x_1, x_2 two members, x_1 owns a_1, a_2, a_3 , x_2 owns a_1, a_2 , then the skills owned by this classification are a_1 and a_2 ;

2. "once at a time" means that if $PUM = PU$ cannot cover all the skills, $PUM = \cup (PUM_i \cap SPU_1')$, and if PUM still can't cover all the skills, $PUM = \cup (PUM_i \cap SPU_2')$, just like this, until classifications in PUM can cover all the skills.

Step 3: Get the final PUM set, arrange the subsets in PUM in descending order according to the relative importance, $PUM = \{PUM_1, PUM_2, PUM_3, \dots\}$, that skills owned by PUM_1 is more important than skills owned by PUM_2 , and each subset PUM_i in PUM is actually a classification,

Step 4: Calculate the each node distance between every two classifications.

Step 5: For each node x_i , find the node x_j that is most far away from it, establish the relationship from x_i in PUM_i to x_j in PUM_j if $i < j$, or establish the relationship from x_j in PUM_j to x_i in PUM_i if $i > j$.

3.6 THE SIMULATION PROCESS OF NETWORK FORMULATION

So the process of the network formation is shown in figure 2.

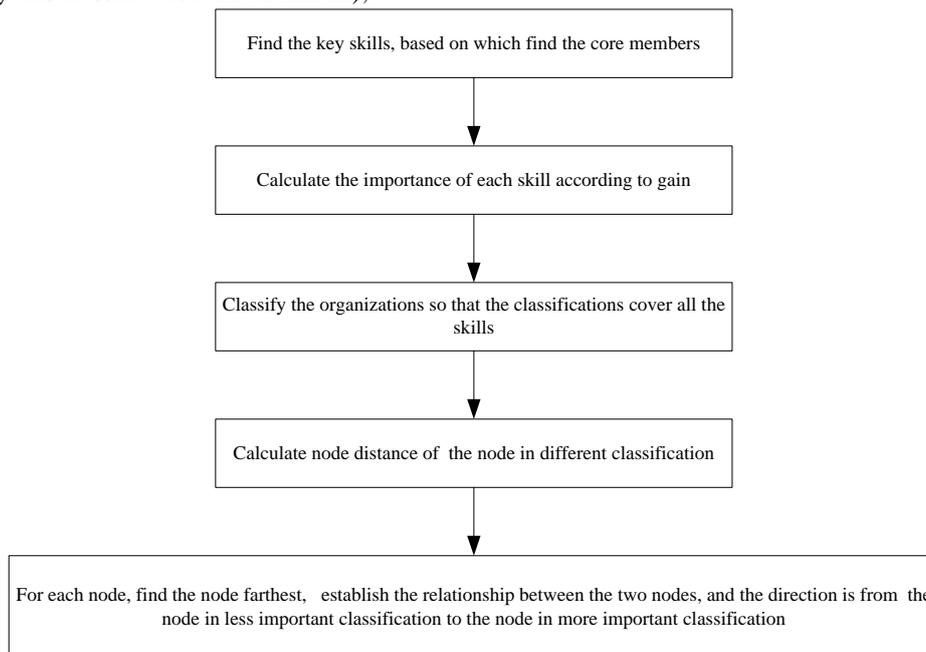


FIGURE 2 Network formation process

4 Conclusions

Recently, the research topics on collaborative networks are mainly on the motives for the collaboration, evaluation on the impact of different types of collaborative networks on product innovation performance, and value systems in collaboration networks. However, there are few researches on the formation of the network. Therefore, this paper mainly

talks about the network formation and proposes the simulation model on the formation process based on the set theory.

Acknowledgements

This study is supported by the foundation of youth scholar, Shanghai international studies university, No.KX181118 and No.2013114038.

References

- [1] Heiko Thimm, Karsten Boye Rasmussen 2011 Adaptable Information Provisioning in Collaborative Networks: An Object Modeling Framework and System Approach *International Journal of Distributed Systems and Technologies* 2(4) 44-56
- [2] Roberto da Piedade Francisco, Américo Azevedo, António Almeida 2012 Alignment prediction in collaborative networks *Journal of Manufacturing Technology Management* 23(8) 1038-56
- [3] Kuen-Hung Tsai 2009 Collaborative networks and product innovation performance: Toward a contingency perspective *Research Policy* 38 765-78
- [4] Romero D, Rabelo R J, Molina A 2013 Collaborative Networks as Modern Industrial Organisations: Real Case Studies *International Journal of Computer Integrated Manufacturing* 26 1-2
- [5] Leandro Loss, Servane Crave 2011 Agile Business Models: an approach to support collaborative networks *Production Planning & Control* 22(5) 571-80
- [6] Camarinha-Matos L M, Macedo P 2010 A conceptual model of value systems in collaborative networks *Journal of Intelligent Manufacture* 21 287-99
- [7] Fornasiero R, Zangiacomì A 2013 A structured approach for customised production in SME collaborative networks *International Journal of Production Research* 51(7) 2110-22
- [8] Campos P, Brazdil P, Mota I 2013 Comparing Strategies of Collaborative Networks for R&D: An Agent-Based Study *Computer Economy* 42 1-22
- [9] Lyons A C, Everington L, Hernandez J, Li D, Michaelides R, Um J 2013 The application of a knowledge-based reference framework to support the provision of requisite variety and customisation across collaborative networks *International Journal of Production Research* 51(7) 2019-33
- [10] Macke J, Vargas Vallejos R, Kadgia Faccin, Genari D 2013 Social capital in collaborative networks competitiveness: the case of the Brazilian Wine Industry Cluster *International Journal of Computer Integrated Manufacturing* 26(1) 117-24
- [11] Jardim-Goncalves R, Agostinho C, Sarraipa J, Grilo A, Pedro Mendona J 2013 Reference framework for enhanced interoperable collaborative networks in industrial organizations *International Journal of Computer Integrated Manufacturing* 26(1) 166-82
- [12] Rui Pinto Ferreira, Jorge Neves Silva, Faimara do Rocio Straus 2011 Performance Management in Collaborative Networks: a Methodological Proposal *Journal of Universal Computer Science* 17(10) 1412-29
- [13] Romero D, Galeano N, Molina A, 2009 Mechanisms for assessing and enhancing organisations' readiness for collaboration in collaborative networks *International Journal of Production Research* 47(17) 4691-710
- [14] Choudhary A K, Harding J, Camarinha-Matos L M, Koh S C L, Tiwari M K 2013 Knowledge management and supporting tools for collaborative networks *International Journal of Production Research* 51(7) 1953-7
- [15] Rosas J, Macedo P, Camarinha-Matos L M 2011 Extended competencies model for collaborative networks *Production Planning & Control: The Management of Operations* 22(5) 501-17
- [16] Abreu A, Macedo P, Camarinha-Matos L M 2009 Elements of a methodology to assess the alignment of core-values in collaborative networks *International Journal of Production Research* 47(17) 4907-34
- [17] Manz'uch Z 2011 Collaborative networks of memory institutions in digitisation initiatives *The Electronic Library* 29(3) 320-43
- [18] Ovidiu Noran 2013 Collaborative networks in the tertiary education industry sector: a case study *International Journal of Computer Integrated Manufacturing* 26(1) 29-40

Authors



Shang Shanshan, born in November, 1983, HeNan

Current position, grades: lecturer

University studies: Shanghai International Studies University

Scientific interest: Knowledge management, E-government

Publications: The method of selecting critical successful factors to knowledge management and its automation; The research on the key influential factors of knowledge management and their relationship

Experience: a lecturer of information system and information management at school of business and management, Shanghai International Studies University, China. She received her Ph.D from the Tongji university.



You Jianxin, born in April, 1961, JiangSu

Current position, grades: professor

University studies: Tongji University, Shanghai university

Scientific interest: operation management, innovation management

Publications: Research on science and technology project performance management based upon technological innovation; Review on the research on intellectual property strategy evaluation for high-tech corporations; Reciprocity-based knowledge conversion mechanism of knowledge-based enterprises

Experience: a professor of business management and doctoral supervisor at school of economics and management, Shanghai Tongji University, and at school of management, Shanghai University, China. He received her Ph.D from the Tongji university.

Analysis on road traffic accidents spatial distribution based on the multi-fractal theory

Rende Yu^{1*}, Juan Shen²

¹School of Transportation and Vehicle Engineering, Shandong University of Technology, Zhangzhou Road 12, Zibo, Shandong, China

²Linyi Science and technology cooperation and Application Research Institute, Jinqushang Road 46, Linyi, Shandong, China

Received 1 March 2014, www.tsi.lv

Abstract

After analysing the characteristics on spatial distribution of road traffic accidents in some areas in China, this paper took road traffic accidents of some provinces/cities in China as an example and thought those provinces/cities as cells. Then the fractal spectrum of road traffic accidents spatial distribution in two-dimensional space and $\ln \varepsilon - \ln \chi_q(\varepsilon)$ curve was obtained by MATLAB programming based on multi-fractal theory. Because of the preferable linear relation $\ln \chi_q(\varepsilon)$ between and $\ln \varepsilon$, the conclusion of which road traffic accidents spatial distribution satisfies power-law form and accord with multi-fractal distribution was obtained. By calculating the relation of related parameters, this paper analysed the characteristics of road traffic accidents spatial distribution further.

Key words: Road traffic accidents, spatial distribution, multi-fractal spectrum, deaths, injuries

1 Introduction

The fractal theory that combine traditional determinate theory with Stochastic theory is one of subjects established and developed in the late 19th century. It prompts people to know the complex natural phenomena more profoundly and has become one of important branches on scientific research about nonlinear (CHENG Q et al., 1995; Harte D., 2001; SUN Hongjun et al., 2005).

The fractal theory applies to establish model on natural phenomena and to forecast, but it is difficult to use fractal geometry in real world. Because genuine fractal is not nonexistent in nature, accurate fractal in mathematics is changed as approximative fractal in nature for applying (CHENG Q., 1999; EVERTSZ C J G et al., 1992). For example, in physical phenomenon, Brownian Movement is persuasive on fractal model (EVERTSZ C J G et al., 1992). So far, the fractal theory has been applied to many fields, including chemistry, life science, astronomy, geography, geology, economics and biology etc. other than physics (Kantelhardt J W et al., 2006; Godano C et al., 1997; MOLCHAN G et al., 2007; Kantelhardt J W et al., 2002). Especially, the fractal science has significant help in many researches of the practical problem, such as earthquake analysis, analysis of the characteristics of rivers and lakes and stock trend analysis etc (LIU Yan et al., 2007; FU Qiang et al., 2010; NIU Zhi-guang et al., 2010; TANG Lizhong et al., 2010; LI Min et al., 2011).

The fractal theory is applied not only in natural science, but also in social science. It have been proved by practical studies that the stock market and changes in profits are related to the fractal, so economists researched the complex economic phenomenon by the fractal theory (YU Jian-ling et al., 2006; YANG Ni et al., 2008). Studies have shown that the road net in traffic system has fractal features also.

In recent years, many researchers applied the fractal theory to study traffic system in China. LIU-Miaolong in Tongji University taken the urban traffic network in Shanghai as a case, the fractal dimension of the traffic network in Shanghai and its districts have been measured and calculated so as to research on the spatial changes of the fractal features (LIU-Miaolong et al., 2003). XU Zhi-hai in Information Engineering University studied several fractal dimensions of characterizing the structure of spatial networks base on the factual geometry theory, and stressed on both the definition and the geographical meaning of similitude dimension, and illustrated its application on the research of traffic networks distribution character with an example. According to the result analysis, some conclusions are given (XU Zhi-hai et al., 2006). JIANG Kejin Southwest in Jiaotong University proposed the consistent evolution between urban morphology dimension and road network dimension and put forward the coordinate between gene conception and model for evaluating the matching relationship between them, finally taken Chengdu city as an example to analyse it's the evolution process of urban morphology and road network (JIANG Kejin et al.,

* Corresponding author e-mail: yuwenrende@tom.com

2008). ZHANG Xiao-hong, YU Ren-de in Shandong University of Technology studied the multi-fractal spectrum of traffic accident time series in Shandong province from 2008 to 2009 by the partition function and obtained the monthly multi-fractal spectrum of time series based on the multi-fractal theory (ZHANG Xiao-hong, YU Ren-de et al., 2013). CHEN Peng in Southeast University demonstrated the feasibility to analyse traffic accident using fractal theory by power spectrum analysis, and extended outwards the fractal property of inner interval by self similarity and scale invariance of fractal, and on this basis proposed a fractal extrapolation algorithm to realize the prediction of traffic accidents (CHEN Peng et al., 2008). CHEN Zhiyu in Shanghai Maritime University used the fractal interpolation to handle discrete data aggregation of temporal series and evaluate the IFS iteration function system and its attractors, and predicted the quantity trend of marine traffic accidents in the coming years (CHEN Zhiyu et al., 2009).

By studying and analysing, road traffic accident time series has multi-fractal distribution characteristics. However, the law of road traffic accident reflects not only on time series distribution but also on spatial distribution. This paper applies the fractal theory to study the characteristics on spatial distribution of road traffic accidents.

2 Multi-fractal spectrum on two-dimensional space

2.1 DEFINITION OF BINARY FRACTAL SPECTRUM

X is assumed to represent multi-measure fractal limit set, then definition of binary fractal spectrum is given by

$$\begin{cases} d_s^{(u,v)}(X) = h_s^{(u,v)}(X) \times d(\mu), \\ d_R^{(u,v)}(X) = h_R^{(u,v)}(X) \times d(\mu), \end{cases} \quad (1)$$

where, $d(\mu)$ as the dimension of radical measure μ , often as n .

Especially, if μ, m is same measure on R^n , and $\mu = q, v = 1$, binary fractal spectrum is converted into one dimensional fractal spectrum as formula (2). It is Renyi dimension, information dimension as $q = 1$, capacity dimension as $q = 0$.

$$\begin{cases} h_{S/R}^{(q,1)}(X) = h_{S/R}^{(q)}(X), \\ d_{S/R}^{(q,1)}(X) = d_{S/R}^{(q)}(X). \end{cases} \quad (2)$$

Conventional fractal dimension include box counting dimension, Renyi dimension, information dimension, capacity dimension.

2.2 THE CALCULATION OF MULTI-FRACTAL SPECTRUM ON BINARY DIMENSION

$(x_i, y_i, z_i), i = 1, 2, \dots, n_i$ is denoted as distribution of a set of points on binary dimension, (x_i, y_i) as plane coordinates of points, z_i as measuring data, n_i as number of points.

Rectangular area including all points $S = \{(x, y) | \min(x_i) \leq x \leq \max(x_i), \min(y_i) \leq y \leq \max(y_i)\}$ is divided into a lot of small squares in size ε . Small squares are called unit. Sum of measuring data on i^{th} unit is denoted as $\mu_i(\varepsilon)$. Formula (3) is defined.

$$\chi_q(\varepsilon) = \sum_i \mu_i^q(\varepsilon). \quad (3)$$

If $\mu_i(\varepsilon)$ meets multi-fractal condition, and relationship between formula (3) and size ε . Then formula (4) can be obtained for any q meeting $-\infty \leq q \leq \infty$.

$$\chi_q(\varepsilon) = k \times \varepsilon^{\tau(q)}, \quad (4)$$

where $\tau(q)$ as function of q , and q as integer only.

Formula (4) is converted into formula (5).

$$\ln \chi_q(\varepsilon) = \ln k + \tau(q) \ln \varepsilon. \quad (5)$$

The linear relation with slope $\tau(q)$ between $\ln \varepsilon$ and $\ln \chi$ is indicated on two logarithm planes. In calculation, part of q is obtained to carry linear fitting by least square method. Then $\tau(q)$ is obtained. Finally multi-fractal spectrum $f(\alpha)$ is obtained by $\tau(q)$.

3 Analysis on road traffic accident spatial distribution based on the multi-fractal theory

As a whole, road traffic accident declines slowly by statistic data in China. For traditional analysis on road traffic accident, object of study is time series. As a result, the time distribution can be obtained.

3.1 SITUATION ANALYSIS ON ROAD TRAFFIC ACCIDENTS SPATIAL DISTRIBUTION

This paper selects road traffic accidents statistic data in some provinces/cities of China from 2009 to 2011 for analysing. Road traffic accidents, deaths, injuries spatial distribution are showed as Figure 1 – Figure 3 in some provinces/cities of China from 2009 to 2011.

The following conclusions are obtained according Figure1-Figure3.

- (1) In China, the curve shape of road traffic accidents spatial distribution is almost similar every year.
- (2) The difference is clear on road traffic accidents, deaths, injuries in every provinces/cities.
- (3) It is subject to the level of economic development and location (such as eastern coast or western interior).

For example, road traffic accidents, deaths, injuries in less developed district and western interior are less than in developed district and eastern coast.

Thus, road traffic accidents spatial distribution takes on the inhomogeneity of the characteristics in China. So, it is necessary to study further for grasping the characteristics of road traffic accidents spatial distribution.

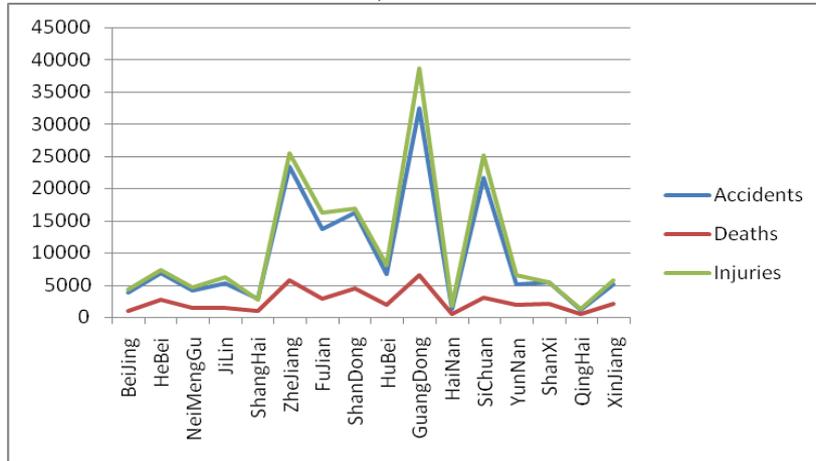


FIGURE 1 Road traffic accidents statistic data in some provinces/cities of China in 2009

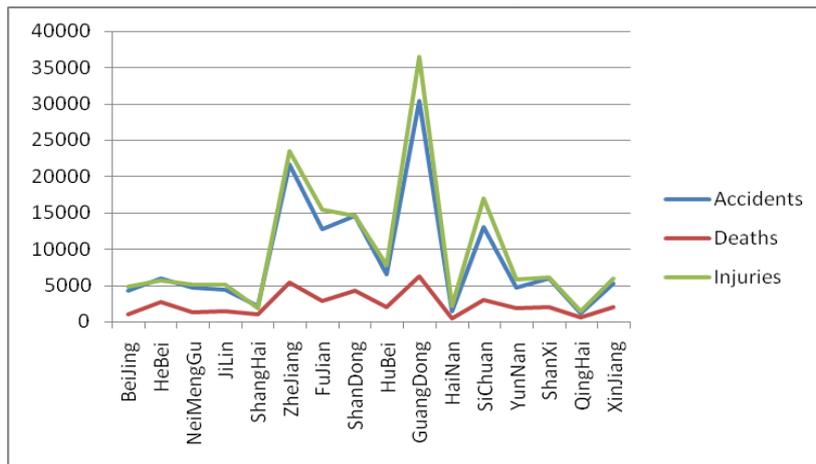


FIGURE 2 Road traffic accidents statistic data in some provinces/cities of China in 2010

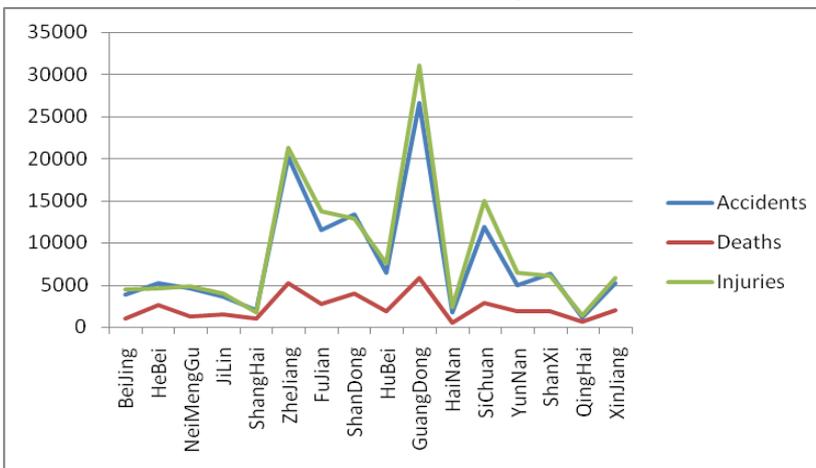


FIGURE 3 Road traffic accidents statistic data in some provinces/cities of China in 2011

3.2 ANALYSIS ON ROAD TRAFFIC ACCIDENTS SPATIAL DISTRIBUTION BASED ON MULTI-FRACTAL SPECTRUM

1) Choosing and analysing data

Road traffic accidents spatial distribution is almost the same in some provinces/cities of China from 2009 to 2011, so we choose road traffic accidents in some

provinces/cities of China in 2011 and look provinces/cities are as cells for studying.

2) Multi-fractal spectrum on road traffic accidents spatial distribution

Road traffic accidents in some provinces/cities of China in 2011 is analysed as Table 1

Program is written as Figure 4, then $\lg \varepsilon - \lg \chi$ curve is obtained as Figure 5.

TABLE 1 Road traffic accidents in some provinces/cities of China in 2011

| provinces/cities | Road traffic accidents | provinces/cities | Road traffic accidents | provinces/cities | Road traffic accidents |
|------------------|------------------------|------------------|------------------------|------------------|------------------------|
| Beijing | 3934 | Anhui | 14005 | Sichuan | 11860 |
| Tianjin | 2600 | Fujian | 11517 | Guizhou | 1566 |
| Hebei | 5197 | Jiangxi | 3354 | Yunnan | 5022 |
| Shanxi | 6239 | Shangdong | 13375 | Xizang | 943 |
| Neimenggu | 4591 | Henan | 6877 | Shanxi | 6362 |
| Liaoning | 6446 | Hubei | 6490 | Gansu | 3027 |
| Jilin | 3639 | Hunan | 8118 | Qinghai | 1163 |
| Heilongjiang | 3435 | Guangdong | 26586 | Ningxia | 1829 |
| Shanghai | 2085 | Guangxi | 4290 | Xinjiang | 5182 |
| Jiangsu | 13436 | Hainan | 1737 | | |
| Zhejiang | 20178 | Chongqing | 5729 | | |

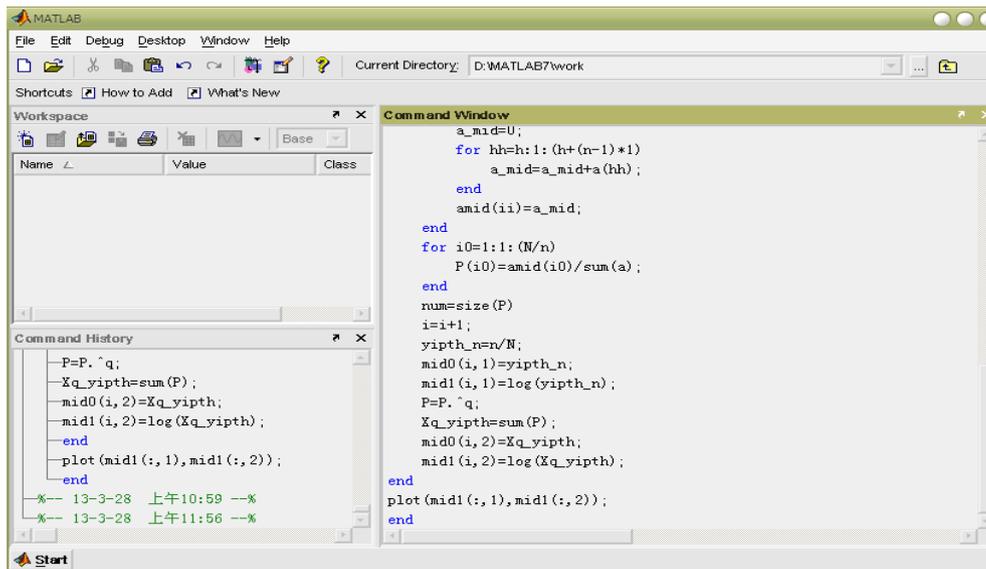


FIGURE 4 Matlab program analysis

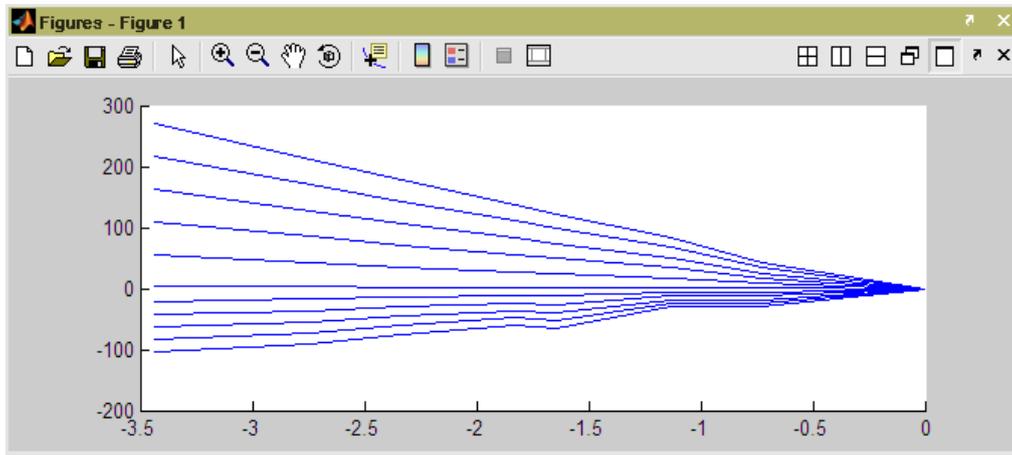


FIGURE 5 $\ln \varepsilon - \ln \chi_q(\varepsilon)$ curve

It is known that there is better linear relation between $\ln \chi_q(\varepsilon)$ and $\ln \varepsilon$ by $\ln \varepsilon - \ln \chi_q(\varepsilon)$ curve in Figure 5.

4 Conclusion

4.1 RELATIVE PARAMETER RELATION

1) $q - \tau(q)$ relation

Road traffic accidents spatial distribution has characteristic of multi-fractal distribution, so first $\tau(q)$, slope of $\ln \varepsilon - \ln \chi_q(\varepsilon)$ curve, must be obtained, then $q - \tau(q)$ relation is confirmed as Figure6.

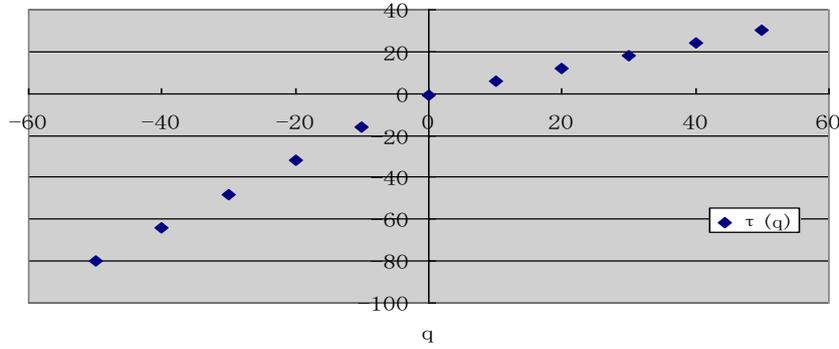


FIGURE 6 $q - \tau(q)$ relation

2) $q - \alpha(q)$ relation

$\alpha(q)$ is obtained by the differential of the $\tau(q)$, and then $q - \alpha(q)$ relation is obtained as Figure7.

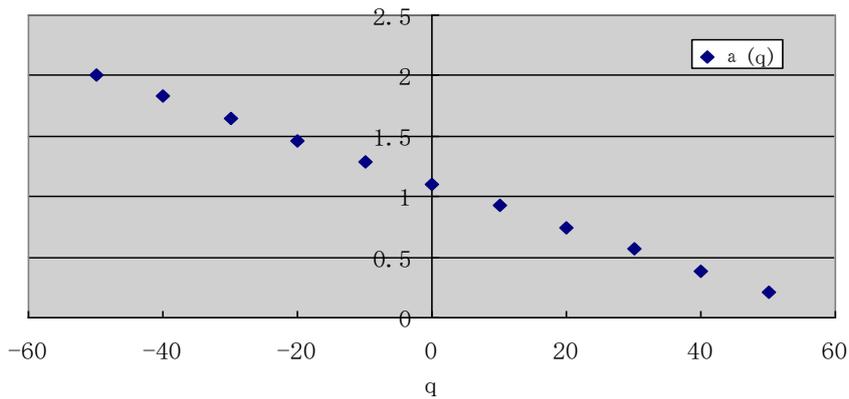


FIGURE 7 $q - \alpha(q)$ relation

3) $f(\alpha) - \alpha(q)$ relation

$f(\alpha)$ is obtained by $\alpha(q)$, and then is obtained $f(\alpha) - \alpha(q)$ relation is confirmed as Figure 8.

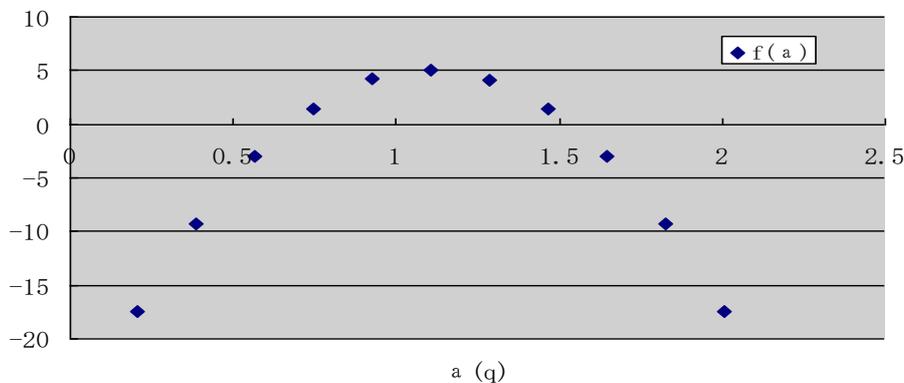


FIGURE 8 $f(\alpha) - \alpha(q)$ relation

4.2 CONCLUSION

By analysing Figure 6 – Figure 8, conclusion follows: $\tau(q)$ is an increasing function of q , $\alpha(q)$ as derivative of $\tau(q)$ decreases as q increases, $f(\alpha)$ is convex function of $\alpha(q)$.

It is known $f(\alpha)-\alpha(q)$ curve has broad and relaxed characteristic by Figure 8. The size of $\Delta\alpha = \alpha_{\max} - \alpha_{\min}$ indicates the size of strangeness on spatial distribution. Because of $\Delta\alpha = \alpha_{\max} - \alpha_{\min} = 2.007 - 0.207 = 1.8$, the strangeness of road traffic accidents spatial distribution is greater in example in this paper.

In fact, economy development level and geographical location are different, there is obvious difference on road traffic accidents among different provinces/cities in China. In general, road traffic accidents in undeveloped and

inland provinces/cities are relatively less than in developed and coastal provinces/cities. So studying result based on multi-fractal theory accords with fact in China. It follows that road traffic accidents special distribution has characteristic of continuous and multi-fractal distribution.

Acknowledgments

We owe the greatest thanks to the Shandong Traffic Management Bureau for providing related road traffic accidents data.

References

- [1] Harte D 2001 *Multi-fractals: Theory and Application* New York: Chapman & Hall/ CRC Press
- [2] Sun Hongjun, Zhao Lihong 2005 Creation and Application of the Fractal Theory *Journal of Liaoning Institute of Technology* **25**(2) 113-7
- [3] Cheng Q, Agterberg F P 1995 Multi-fractal modeling and spatial point processes *Math Geol* **27** 831-45
- [4] Cheng Q 1999 Multi-fractality and spatial statistics *Computer & Geosciences* **25**(9) 949-62
- [5] Evertsz C J G, Mandelbrot B 1992 *Multi-fractal measures (Appendix B)* In: PEITGEN H-O, JURGENS H, SAUPE D. Chaos and fractals New York: Springer Verlag 922-53
- [6] Kantelhardt J W, Koscielny - Bude Eva, Braun P, et al 2006 Long-term Persistence and Multi-fractality of Precipitation and River Runoff *Journal of Geophysical Research* **322**(1-4) 120- 37
- [7] Godano C, Alonzo M L, Vilaro G 1997 Multi-fractal Approach to Time Clustering of Earthquakes, Application to Mt. Vesuvio Seismicity *PAGEOPH* **149** 375 - 90
- [8] Molchan G, Kronrod T 2007 Seismic interevent time: a spatial scaling and multi-fractality *Pure and Applied Geophysics* **164**(1) 75-96
- [9] Kantelhardt J W, Zschiegner S A, Koscielny-Bunde Eva, et al. 2002 Multi-fractal Dredged Fluctuation Analysis of Non-stationary Time Series *Physical A* **316**(1-4) 87 - 114
- [10] Liu Yan, Mu De-jun, Zhang Jia-zhong 2009 Traffic Prediction for LAN Based on Multifractal Spectrums *Journal of System Simulation* **21**(12) 3743-7
- [11] Fu Qiang, Zhou Yu-huan, Wang Liang 2010 Multi-fractal spectrum for signal time series of turbulence and worm DNA *Journal of PLA University of Science and Technology(Natural Science Edition)* **11**(5) 572 - 7
- [12] Niu Zhi-guang, Lu Ren-qiang, He Xiao-yun 2010 Study on spatial characteristic of coastal pollutant by multi-fractal theory *Marine environmental science* **29**(1) 99 - 103
- [13] Tang Lizhong, Xia K W, Li Xibing 2010 Seismic Multi-Fractal Characteristics in Mines and Seismicity Prediction *Chinese Journal of Rock Mechanics and Engineering* **29**(9) 1818 - 24
- [14] Li Min, Li Yi 2011 Local fractal and multifractal characteristics of soil number-based particle size distributions *Journal of Northwest A&F University(Nat. Sci. Ed.)* **39**(11) 216 – 22
- [15] Yu Jian-ling, Zang Bao-jiang, Shang Peng-jian 2006 Multifractal Analysis of Stock Market Time Series *Journal of Beijing Jiaotong University* **30**(6) 69-72
- [16] Yang Ni, Xie Chi, Sun Bo, Zhang Yuanyuan, Ding Hui 2008 Empirical Research on the Multifractal Characteristics of Exchange Rate Time Series *Journal of Hunan University (Natural Sciences)* **35**(8) 89 - 92
- [17] Liu Miaolong, Huang Peibei 2003 Application of Fractal Theory to Research on the Tempo-spatial Changes of Urban Traffic Network-Taking Shanghai as a Case *Geomatics and Information Science of Wuhan University* **28**(6) 749-53
- [18] Xu Zhi-hai, Zhang Zhao-yun 2006 Application of fractal theory in the study of characterizing the structure of traffic networks *Engineering of surveying and mapping* **151**(1) 27-30
- [19] Jiang Kejin, Zhang Dianye 2008 Research on the Integrative Evolution of Urban Morphology & Road Network Based on the Fractal Theory **21**(3) 388-91
- [20] Chen Peng, Li Xu-hong, Sun Hua-can 2008 Analysis of Traffic Accident Based on Fractal Theory *Journal of Highway and Transportation Research and Development* **25**(3) 130-3
- [21] Zhang Xiao-hong, Yu Ren-de, Zhang Qiang 2013 Traffic accidents analysis based on multi-fractal spectrum *Journal of Shandong University of Technology (Natural Science Edition)* **27**(2) 62-4
- [22] Chen Zhiyu, Hu Shenping, Hao Yanbin 2009 Prediction of marine traffic accidents based on fractal theory *Journal of Shanghai Maritime University* **30**(3) 18-21

| Authors | |
|---|--|
|  | <p>Yu Rende, born on May 5, 1961, Yantai, Shangdong, China</p> <p>Current position, grades: associate professor, master University studies: road traffic safety Scientific interest: transportation planning, road traffic safety, traffic simulation, intelligent transportation Publications: >50 papers Experience: Member of Shandong road traffic safety association .Works in Shandong University of technology. is engaged in education and scientific research in road transportation. main research domain: traffic planning, road traffic safety, traffic simulation, intelligent transportation. 50 papers is published in journal of domestic and international almost. Great contribution in transportation planning and road traffic safety and traffic impact analysis</p> |
|  | <p>Shen Juan, born on January 10, 1984, Jining, Shangdong, China</p> <p>Current position, grades: archives members, Bachelor University studies: Chinese Linguistic Literature Scientific interest: Road traffic safety, City traffic management Publications: >18 papers Experience: Shen Juan is now working at Linyi Institute of Sci-Tech Cooperation and Application. She has been engaged in agricultural machinery application engineering, archive management, road traffic accidents statistical analysis, and urban transportation management. She has published 18 journal articles both in China and abroad.</p> |

Research on the principal-agent problems in China's low-carbon ecological urban construction

Shuiping Zhang*

School of Economy and Management, Anhui University of Science and Technology, Anhui Province, China

Received 1 July 2014, www.tsi.lv

Abstract

In the game of the interest bodies of low-carbon ecological urban construction, the central government, as a principal, will lose some interests in some ways because of information disadvantages, whereas the local governments, as agents, will make use of their information advantages to make profitable action choices for more interests. As a result, moral risks will appear for the latter. This paper attempts to construct a mathematical model of the game theory for the principal-agent problems in the low-carbon ecological urban construction and analyses the choice actions involved. The conclusion is drawn that for the optimal balance of the game to be realized between the central and local governments, a relevant system must be established. This system is expected to change the information asymmetry by increasing the central government's ability to acquire information while stimulating or restraining the local governments' choice actions so that the external pressure on the local governments will be turned into their internal actions in a low-carbon ecological urban construction.

Keywords: low-carbon ecological cities and towns, agency by agreement, information asymmetry, system

1 Introduction

Urbanization refers to the shifts and aggregations of the population and production factors from rural areas to old or new cities and towns, including the increase of a city's or town's population and number and the modernization of the urban and rural economy and society. Since the reform and opening toward the outside, Chinese urbanization construction has achieved a remarkable development. The urbanization rate increased to 53.73% in 2013 from 12.5% in the early 1980s, which is an increase of 41% (2013). However, it is evident that with the increased rate of urbanization, the huge increase in resource consumption and the rapid expansion of the city scale have caused significant impacts on the original functions and structures of cities and towns. For example, some external problems, such as resource shortages and environmental pollution, are increasingly serious. Chinese urbanization has faced great pressure from population, resource and environment aspects, so China must choose the development path of low-carbon ecological urbanization.

The theory of ecological environment management is mainly based on the European practice, and the early theory includes three points as follows: effective economic development, social justice development and environmentally friendly development (Fan, 2011). It is a double-win model of the economy and environment by nature, holding that economic increases and environmental protection should be coordinated. There is a view that the realization of ecological management

relies largely on the innovation of science and technology, not changes to the basic social system, holding that social structural change is made to promote environmentally friendly production and consumption (Christoff, P., 1996). This view holds that ecological management refers to the adjustment of a capitalist political and economic structure to promote environmentally benign development and that only when the capitalist internal unreasonable structure is fully adjusted under the current basic political and economic system will the environmentally benign development occur (White, D. F., 2002). The adoption of cleaner production technologies and preventive environmental protection measures are cases of this view. Such views belong to the preventive strategic theory. Another view proposes setting up theoretical and practical models of ecological management from the perspective of environmental protection and industrial transformation (Gerald, B. et al., 2001). It belongs to the selective reform theory.

At the beginning of the 21st century, the concept of "low-carbon city" has appeared in foreign countries. In 2004, Japanese governments and scholars began studying the model of a low-carbon city and its development paths. For example, Japanese experts held that the system design and construction of a low-carbon city should be combined with the status quo of the local system, economy, environment, history and value (Aoki, 2009). The UK is the forerunner in the design and practice of the low-carbon city. English scholars think the design of the low-carbon city should be part of overall planning, taking

* *Corresponding author* e-mail: zhsp310@163.com

seven aspects into consideration, namely the centre of cities and towns, the centre of the city edge, the inner city area, the industrial zone, the suburbs, the stretch area, and the rural area (Thuli, N. M. et al., 2010). Glaeser and Kahn (2001) noted in their study of carbon emission, city scale and land development that the city scale is directly proportional to carbon emissions, i.e., with the expansion of the city scale, the per capita carbon emissions of the newly added population are higher than the average emissions of the original population. Jenny Crawford and Will French (2008) studied the relation between English space planning and the low-carbon goal, thinking that the understanding and adaptability of new technologies in English planning are crucial to making low-carbon space come true and pointing out the best effective development of a low-carbon city is to adopt flexible measures according to specific conditions. In addition, some works provide specific measures to solve the environmental pollution of cities and towns. For instance, based on recycling agriculture, Diana, M. L., et. al (2006) proposed "sustainable agriculture" to solve the environmental pollution of American small towns in the country; other experts, such as Elizabeth Economy (2006), combine economic methods with material balance theory to propose their environmental management measures for cities and towns.

Currently, based on a large number of Chinese domestic research literature about the ecological environment of urbanization, scholars have performed studies on environmental pollution management theories of urbanization, but most of them are limited to the simple application analysis of economic theories. To sum up, their studies focus mainly on the following two aspects:

The first is analytical investigation based on macro-economics. Hu A. G., et. al., (2012) held that the low-carbonization city is an important aspect in the process of Chinese economic transformation from high-carbon to low-carbon, including low-carbon energy, increasing gas popularity rate, increasing city greening rate, increasing waste processing rate, and so on. Qiu Baoxing (2012) suggested rethinking the city construction ideas and development models to explore city development paths suitable for China's actual conditions and ecological civilization construction. Li Kexin (2009) held that the theoretical basis of low-carbon city construction is "environmental harmony theory", i.e., to build liveable cities with sustainable development. Hong Dayong (2012) noted that the existing binary structure of the social system has much to do with the fact that medium and small towns self-pollute, and some pollution is uncontrollable for a long time.

The second is the analytical investigation based on micro-economics. Such research works mainly cover the external analysis, the social economic factors and the economic loss evaluation of the small town environmental pollution problems. For example, Meng Xuejing and Shang Jie (2012) proposed with their

economic analysis that the environmental pollution of rural urbanization is caused mainly by three factors: market failure, governmental failure and the simultaneous failure of the market and government. Jang, H. L., et. Al. (2010) argued that there are prominent institutional obstacles and specific implementation problems in the prevention and control of the environmental pollution of Chinese cities and towns, which is mainly manifested in the lack of an effective management system of environmental pollution and an incentive mechanism of environmental and economic policy, in the serious interest conflicts between interest bodies involved in the prevention and control of small towns' pollution, in the ignorance of residents' dominant role in protecting the environment and preventing pollution, and so on. Cai Yuqiu and Yu Xiaochen (2011) proposed a management model of city/town ecological environment, i.e., to strengthen residents' awareness of environmental protection, to establish a sound early warning and monitoring network project of the ecological environment, to integrate small villages and establish the management model with the top-down vertical leadership. Li Yinming and Song Jianxin (2011) studied the key operating factors, management levels and typical models of the environmental self-governance system of small towns and put forward a frame system that can be used for reference by the environmental self-governance system of Chinese small towns.

The above reviews of domestic and foreign research have revealed that domestic and foreign scholars have presented many opinions on the ecological environmental management of cities and towns from different perspectives and have made some achievements. However, generally speaking, those studies seldom systematically analyse the ecological environmental management on the background of urbanization, and they rarely mention any game behaviours between the central government and local governments in the construction of low-carbon cities/towns. Therefore, this paper attempts to construct a mathematical model of the game theory for the principal-agent problems between the central and local governments in low-carbon ecological urban construction and analyses their choice actions and corresponding results with the aim of establishing an institutional and policy system to solve practical problems more systematically.

2 The analysis of the interest demands of local governments in the ecological management game during the new-type urbanization

An urbanization construction and an urban ecological environment are a unity of opposites, so their mutual relations should be treated carefully. The former needs a great number of material resources, which resorts to a good ecological environment as a supply security. However, urbanization construction inevitably destroys the ecological environment, which will do harm to its

own development basis; therefore, ecological environmental management in urbanization is necessary (Yang, 2007). The participants of urbanization construction are made up of diverse interest bodies, so the ecological environmental management in the process of urbanization is a cooperative management of all interest bodies. The ecological environmental problems in the process of urbanization are the external cost of the urbanization construction, which is caused by the interest demand motivations of different interest bodies. The utilitarianism of interest bodies' demands will distort their decision-making behaviours and will eventually cause increasingly serious ecological environmental problems for cities.

According to the principal-agent theory, the central government is the principal and the local governments are the agents of the central government, as well as of their local people. The local governments must ensure their social public interests so that they can be accepted by the principals—the central government and the local people. However, the “economic men’s” feature of politicians has caused local governments to be selfish, as well as to be of social public interest. Their pursuits of political capital cause the local governments to have too many selfish behaviours, so the rapid growth of GDP and financial revenue become the basic goals of their policies and behaviours, whereas the ecological environment is out of their consideration or only counts very little, especially in some remote and backward areas that face the huge pressure of economic development and financial gaps. Local governments often neglect or even permit the ecological pollution of some factories.

In the 1990s, when China began the reform of decentralization and social economic transformation, policy instability and the imperfection of the market economy system existed in China. So the local governments as agents had information advantages over the central government as a principal. They were motivated enough to make use of the long information chain to conceal information from the central government for their private interests. The lower the level of the local governments are, the more private interests they demand, so they have more serious ecological problems in the process of urbanization.

3 Research methods

In an agency by agreement, the principal will ask the agent to act for his own interests, but because of information asymmetry, the former cannot comprehensively master the latter’s choice behaviours. The principal can only know partial information about the agent’s actions, so the problem the principal needs to solve is how to encourage the agents to choose actions for the principal’s interests with rewards and punishments according to the partial information. Here are two formulations for the principal to restrain the agents with a dominant motivation system:

3.1 STATE-SPACE FORMULATION

Suppose A is the combination of all of the agent’s choice behaviours; α represents the one-dimensional variable of the agents’ effort level; θ is an exogenous random variable beyond the formulation (natural state). θ ’s value scope is Θ ; $G(\theta)$ and $g(\theta)$ are the distribution function and density function, respectively ($\theta \in \Theta$). When an agent chooses an action α , the principal will get his own income result $\pi(\alpha, \theta)$, and at the same time can know the actions’ result of the agents $x(\alpha, \theta)$. Assume that π is α ’s increasing concave function (under a given θ , agents’ gains are proportional to the degree of its own efforts, but their efforts’ marginal productivity decreases), and π is also an increasing function of θ . Now the principal needs to design a contract $S(X)$ and rewards or punishes the agents according to the results of their actions $x(\alpha, \theta)$.

Then, the principal and agents’ expected utility function can be expressed as: $v(\pi - s(x))$, $u(s(\pi)) - c(a)$, and $c(a)$ is the agent’s cost function.

From the previous statements, we get: $v' > 0, v'' \leq 0; u' > 0, u'' \leq 0; c' > 0, c'' > 0$. When both the principal and the agent are risk-neutral and their effort’s marginal productivity decreases, $v' > 0$ indicates that the principal expects the agent to make greater efforts, and $c' > 0$ indicates that the agent does not want to give more to the principal. In this case, these two parties are in an interest conflict with each other. Therefore, the principal needs to design a reasonable contract to stimulate the agent to give more. On the other hand, the expected utility function of the agent (P) can be expressed as:

$\int v(\pi(a, \theta) - s(x(a, \theta)))g(\theta)d\theta$. To maximize the principal’s utility function, the agent needs to meet two constraints. One is the individual rationality constraint (IR), which means that the agent will get more utility if he accepts the principal’s contract than if he refuses. If \bar{u} indicates the total utility the agent gets when he refuses the principal’s contract and acts according to his own willingness, then IR can be expressed as:

$\int u(s(x(a, \theta)))g(\theta) - c(a) \geq \bar{u}$. The second constraint is the incentive compatibility constraint (IC), which means that under the natural state θ , the principal cannot know the agent’s behaviours a when the agent is likely to maximize his own interests by choosing the low effort behaviour a^L , though the principal expects the agent to choose the high effort behaviour a^H . The condition of solving their conflicts is that the agent can get more utility if he chooses the high effort behaviour a^H than if he chooses the low effort behaviour a^L . In this case, IC can be expressed as:

$$\int u(s(x(a^H, \theta)))g(\theta)d\theta - c(a^H) \geq \int u(s(x(a^L, \theta)))g(\theta)d\theta - c(a^L), \forall a^L \in \Lambda$$

To sum up, if the expected utility function of the principal (P) wants to get the maximum value, the two constraints (IR and IC) must be satisfied, namely:

$$\begin{cases} \max_{s(x)} \int v(\pi(a, \theta) - s(x(a, \theta)))g(\theta)d\theta \\ \int u(s(x(a, \theta)))g(\theta)d\theta - c(a) \geq \bar{u} \\ \int u(s(x(a^H, \theta)))g(\theta)d\theta - c(a^H) \geq \int u(s(x(a^L, \theta)))g(\theta)d\theta - c(a^L), \forall a^L \in \wedge \end{cases}$$

3.2 PARAMETERIZED DISTRIBUTION FORMULATION

The parameterized distribution formulation is used to transform the distribution function under the natural state of the state-space formulation into the distribution function with x and π as results. For every a , there is a distribution function of x and π . From the original distribution function $G(\theta)$, we can derive a new distribution function $F(x, \pi, a)$ and a new density function $f(x, \pi, a)$. In the state-space formulation, the utility function gets the expected value from the θ of the natural state, whereas in the parameterized distribution formulation, the utility function gets the expected value from the variable x . In the latter formulation, the expected utility function (P) of the principal gets its maximum on the conditions as follows:

$$\begin{cases} \max_{s(x)} \int v(\pi - s(x))f(x, \pi, a)dx \\ \int u(s(x))f(s, \pi, a)dx - c(a) \geq \bar{u} \\ \int u(s(x))f(s, \pi, a^H)dx - c(a^H) \geq \int u(s(x))f(s, \pi, a^L)dx - c(a^L), \forall a^L \in \wedge \end{cases}$$

4 The principal-agent models of the central and local governments in the construction of low-carbon ecological cities

4.1 THE PRINCIPAL-AGENT MODEL

According to the previous analysis, the central government and local governments have formed the principal-agent relationship in the process of urbanization construction.

The variable $a(a > 0)$ is a one-dimensional effort variable, representing the effort level of the local governments in conducting the contract of the central government. $A(A > 0)$ is the subjective action ability level of the local governments. $B(B > 0)$ is a constant, standing for the scale of the local government resources. t is a time variable, and I is an input variable. $\eta(\eta > 0)$ is the adjustment coefficient. θ is the uncertain factor beyond the formulation, the random variable of the normal distribution with the mean value 0 and variance σ^2 . The utility linear function of local governments can be expressed as: $\pi = \eta t I(A + B)a + \theta$, so the utility expected value is: $E\pi = \eta t I(A + B)a$.

Suppose the central government is risk-neutral and the local governments are risk-aversers. When the local governments choose to conduct the contract of the central

government, their contract's equation is $s(\pi) = \alpha + \beta\pi$ (α is the fixed gain of the local governments in urbanization construction and it is not related with the output π ; β is the contract's incentive strength, namely the proportional coefficient between the effort level of the local governments and the output π ; $\beta = 0$ means that the local governments choose the maximum cost). Therefore, if the equation of a given commission contract is $s(\pi) = \alpha + \beta\pi$, then the expected gain of the central government can be expressed as: $E v(\pi - s(\pi)) = (1 - \beta)\eta t I(A + B)a - \alpha$.

On the other hand, suppose the equation of the effort cost for the local governments to perform the central government's contract is $c(a) = ba^2 / 2$ ($b > 0$; b is the cost coefficient). If the cost the local governments pay with the same efforts is directly proportional to b , then the actual gain of the local governments can be expressed as: $w = s(\pi) - c(a) = \alpha + \beta(tI(A + B)a + \theta) - ba^2 / 2$.

The certain equivalent gain is: $W = Ew - \rho\beta^2\sigma^2 / 2 = \alpha + \beta\eta t I(A + B)a - ba^2 / 2 - \rho\beta^2\sigma^2 / 2$.

In this equation, Ew is the expected income of the local governments, $\rho\beta^2\sigma^2 / 2$ is the cost risk of the local governments, ρ is the degree of the local governments' risk-aversion ($\rho > 0$ means risk-aversion and $\rho = 0$ means risk-neutral). From the above derivation, it can be obtained that the central government's goal is to maximize its expected gain $E v(\pi - s(\pi))$, and the local governments' goal is to maximize the certain equivalent gain W . Suppose the retained gain of the local governments is a constant \bar{w} . If the certain equivalent gain the local governments obtain after performing the central government's contract is less than their retain gain ($W < \bar{w}$), the moral risk will appear. In conclusion, the individual rationality constraint (IR) of the local governments can be expressed as: $W = \alpha + \beta\eta t I(A + B)a - ba^2 / 2 - \rho\beta^2\sigma^2 / 2 \geq \bar{w}$.

The incentive compatibility constraint (IC) of the local governments can be expressed as: $\alpha + \beta\eta t I(A + B)a^H - b(a^H)^2 / 2 - \rho\beta^2\sigma^2 / 2 \geq \alpha + \beta\eta t I(A + B)a^L - b(a^L)^2 / 2 - \rho\beta^2\sigma^2 / 2, \forall a^L \in \wedge$.

4.2 INFORMATION ASYMMETRY

On the condition of information asymmetry, the principal (the central government) can observe the effort level of the agent (the local government) in performing the contract. In such a case, the local governments cannot wantonly choose effort levels to maximize their own interests, and the incentive compatibility constraint (IC) becomes invalid. Therefore, as long as the individual rationality constraint (IR) is met to perform the central

government's contract $s(\pi)$, the local government can choose any effort level a . However, in urbanization construction, the realization of the principal's (the central government's) interests depends on the agent's (the local government's) performance, and the principal does not usually participate in the administration decisions of urbanization. In addition, China is in the reform of decentralization and social economic transformation, and some things like policy instability and the imperfection of the market economy system exist in China and cause significant information asymmetry. For those reasons, the central government cannot observe the local governments' efforts level a . In this condition, the incentive compatibility constraint (IC) is valid, and the central government aims to realize the contract $s(\pi)$ by solving the following optimization problems through choosing (α, β) :

$$\begin{cases} \max_{\alpha, \beta, a} Ev(\pi - s(\pi)) = \max_{\alpha, \beta, a} [(1 - \beta)\eta t I(A + B)a - \alpha] \\ \alpha + \beta \eta t I(A + B)a - ba^2 - \rho \beta^2 \sigma^2 / 2 \geq \bar{w} \\ \alpha + \beta \eta t I(A + B)a^H - b(a^H)^2 / 2 - \rho \beta^2 \sigma^2 / 2 \geq \\ \alpha + \beta \eta t I(A + B)a^L - b(a^L)^2 / 2 - \rho \beta^2 \sigma^2 / 2, \forall a^L \in \Lambda \end{cases}$$

Through the above formulae, the IR and IC can be expressed as follows (IR): $\alpha + \beta \eta t I(A + B)a - ba^2 / 2 - \rho \beta^2 \sigma^2 / 2 = \bar{w}$, (IC): $a = \beta \eta t I(A + B) / b$.

Now, put IR and IC into the objective function of the central government's expected income and get:

$$\max_{\alpha, \beta, a} Ev(\pi - s(\pi)) = \max_{\alpha, \beta, a} \left[\frac{\beta \eta^2 t^2 I^2 (A + B)^2 / b - \beta \eta^2 t^2 I^2 (A + B)^2 / 2b - \rho \beta^2 \sigma^2 / 2 - \bar{w}}{\beta \eta^2 t^2 I^2 (A + B)^2 / 2b - \rho \beta^2 \sigma^2 / 2 - \bar{w}} \right]$$

Take the derivative of β and let the first derivative be zero, then you can get the first order conditions:

$$\frac{\partial Ev}{\partial \beta} = \eta^2 t^2 I^2 (A + B)^2 / b - \eta^2 t^2 I^2 (A + B)^2 \beta / b - \rho \beta \sigma^2 = 0$$

The solution is:

$$\begin{cases} \beta' = \frac{1}{1 + \rho b \sigma^2 / \eta^2 t^2 I^2 (A + B)^2} \\ a = \beta' \eta t I(A + B) / b \\ \alpha' = \bar{w} + ba^2 / 2 + \rho \beta'^2 \sigma^2 / 2 - \beta' \eta t I(A + B)a \end{cases} \quad (1)$$

Then, the function of the central government's development contract is: $s(\pi) = \alpha' + \beta' \pi$ and the marginal cost and the marginal expected utility of the local government efforts, respectively, are:

$$c'(a) = (ba^2 / 2)' = ba = \beta' \eta t I(A + B).$$

And $[E\pi(a)]' = \eta t I(A + B)$. The local government always tries to avoid risks, so $\rho > 0, 0 < \beta' < 1$. Thus, the marginal cost of the local government's effort is less than its marginal expected utility, and its highest effort level a^H will not be reached, namely: $a = \beta' \eta t I(A + B) / b < a^H$. In such a case, the expected effectiveness of the central government, the actual effectiveness of the local government and the agency cost of the central government are, respectively, as follows:

$$Ev = \beta' \eta^2 t^2 I^2 (A + B)^2 / 2b - \bar{w} = \frac{\eta^2 t^2 I^2 (A + B)^2}{2b(1 + \rho b \sigma^2 / \beta'^2 t^2 I^2 (A + B)^2)} - \bar{w} < Ev^* \quad (2)$$

$$w = \bar{w} + \frac{\beta'^2 \rho \sigma^2}{2(1 + \rho b \sigma^2 / \eta^2 t^2 I^2 (A + B)^2)} > w^* \quad (3)$$

$$AC = Ev^* - Ev = \frac{\rho \sigma^2}{2(1 + \rho b \sigma^2 / \eta^2 t^2 I^2 (A + B)^2)} \quad (4)$$

4.3 THE INFORMATION ASYMMETRY AFTER ADDING VARIABLES

It follows that the central government's expected effectiveness will not reach its objective effectiveness, whereas the local government's actual effectiveness will be larger than its expected effectiveness. From this analysis, it can be concluded that because of information asymmetry, the central government as the principal (with the information disadvantages) will suffer a certain loss, whereas the local government as the agent (with the information advantages) will get more actual effectiveness by choosing low-effort-degree actions. Therefore, the question for the central government is how to design an incentive system to increase the local government's effort degree and to avoid the moral risks for the purpose of avoiding the central government's agency risks and increasing its expected effectiveness. Next, we analyse the previous principal-agent model. In urban construction, the local government will get the fixed gains α as well as share the surplus gains $\beta \pi$. Here, the gain π not only depends on the its effort level but is also influenced by the uncertain factor θ . The previous effectiveness function of the local government shows that the gain π is simultaneously influenced by the local government's capacity level A, hardware strength B, effort level a and the uncertain factor θ . Thus, π is not a sufficient statistic but a noisy signal about coefficient A, B and a . Therefore, the agency contract about $s(\pi)$ makes the local government face

extra risks, which disagrees with the optimal risk allocation theory. Therefore, such a contract is a suboptimal solution. Now we assume that the central government can observe a new variable k (k has nothing to do with the effort level of the local government but is associated with the uncertain factor B), and the mean value of k is zero with a normally distributed variance. The new contract can be expressed as: $s(\pi, k) = \alpha + \beta(\pi + \phi k)$ (here ϕ means the coefficient between the local government's gain and the variable k .)

The expected gain of the central government is $Ev = (1 - \beta)\eta I(A + B)a - \alpha$, and the fixed equivalent gain of the local government is: $W = \alpha + \beta\eta I(A + B)a - ba^2 / 2 - \rho\beta^2(\sigma^2 + \varphi^2\sigma_k^2 + 2\varphi\text{cov}(\pi, k)) / 2$.

On the same condition of information asymmetry, the central government's goal is to choose α and β to solve the following optimization equations:

$$\begin{cases} \max_{\beta, \varphi} Ev = \max_{\beta, \varphi} [(1 - \beta)\eta I(A + B)a - \alpha \\ \alpha + \beta\eta I(A + B)a - ba^2 / 2 - \rho\beta^2(\sigma^2 + \varphi^2\sigma_k^2 + 2\varphi\text{cov}(\pi, k)) / 2 \geq \bar{w} \\ \alpha + \beta\eta I(A + B)a^H - b(a^H)^2 / 2 - \rho\beta^2(\sigma^2 + \varphi^2\sigma_k^2 + 2\varphi\text{cov}(\pi, k)) / 2 \geq \\ \alpha + \beta\eta I(A + B)a^L - b(a^L)^2 / 2 - \rho\beta^2(\sigma^2 + \varphi^2\sigma_k^2 + 2\varphi\text{cov}(\pi, k)) / 2, \forall a^L \in \wedge \end{cases}$$

So now IR and IC are as follows:

(IR):

$$\alpha + \beta\eta I(A + B)a - ba^2 / 2 - \rho\beta^2(\sigma^2 + \varphi^2\sigma_k^2 + 2\varphi\text{cov}(\pi, k)) = \bar{w}$$

(IC): $a = \beta\eta I(A + B) / b$

Put them into the objective function, and then we get the equation as follows:

$$\max_{\beta, \varphi} Ev = \max_{\beta, \varphi} [\beta\eta^2 t^2 I^2(A + B)^2 / b - \beta^2 \eta^2 t^2 I^2(A + B)^2 / 2b - \rho\beta^2(\sigma^2 + \varphi^2\sigma_k^2 + 2\varphi\text{cov}(\pi, k)) / 2 - \bar{w}]$$

Take the first derivative of β and φ from the central government's objective function, let it be zero, and we get the two optimized first-order conditions:

$$\begin{cases} \frac{\partial(Ev)}{\partial\beta} = \eta^2 t^2 I^2(A + B)^2 / b - \beta^2 \eta^2 t^2 I^2(A + B)^2 / b \\ -\rho\beta(\sigma^2 + \varphi^2\sigma_k^2 + 2\varphi\text{cov}(\pi, k)) = 0 \\ \frac{\partial(Ev)}{\partial\varphi} = \varphi\sigma_k^2 + \text{cov}(\pi, k) = 0 \end{cases}$$

The solution is:

$$\begin{cases} \varphi = -\text{cov}(\pi, k) / \sigma_k^2 \\ \beta = \frac{1}{1 + \rho b(\sigma^2 - \text{cov}^2(\pi, k) / \sigma_k^2) / \eta^2 t^2 I^2(A + B)^2} \end{cases}$$

So $\sigma^2 \sigma_k^2 \geq \text{cov}^2(\pi, k)$ and $\sigma^2 \geq \text{cov}^2(\pi, k) / \sigma_k^2$.

Combine the above two equations and we get:

$$\beta = \frac{1}{1 + \rho b(\sigma^2 - \text{cov}^2(\pi, k) / \sigma_k^2) / \eta^2 t^2 I^2(A + B)^2} \geq \frac{1}{1 + \rho b\sigma^2 / \eta^2 t^2 I^2(A + B)^2} = \beta'$$

$$a = \beta\eta I(A + B) / b$$

$$= \frac{\eta I(A + B)}{b(1 + \rho b(\sigma^2 - \text{cov}^2(\pi, k) / \sigma_k^2) / \eta^2 t^2 I^2(A + B)^2)}$$

From the comparison between these two equations and equation (1), it can be concluded that by putting a noticeable variable k into the agency contract of the central government, and with the function of the incentive system, the local government will enhance its effort level of participating in urban construction, and the corresponding surplus portion it shares will increase. Now the expected gain of the central government, the actual gain of the local government, and the agency cost of the central government are, respectively, as follows:

$$Ev = \frac{\eta^2 t^2 I^2(A + B)^2}{2b(1 + \rho b(\sigma^2 - \text{cov}^2(\pi, k) / \sigma_k^2) / \eta^2 t^2 I^2(A + B)^2)} - \bar{w}$$

$$w = \bar{w} + \frac{\rho\sigma^2}{2(1 + \rho b(\sigma^2 - \text{cov}^2(\pi, k) / \sigma_k^2) / \eta^2 t^2 I^2(A + B)^2)}$$

$$AC = \frac{\rho(\sigma^2 - \text{cov}^2(\pi, k) / \sigma_k^2)}{2(1 + \rho b(\sigma^2 - \text{cov}^2(\pi, k) / \sigma_k^2) / \eta^2 t^2 I^2(A + B)^2)}$$

Compare the previous equations (2), (3) and (4) with the above equations and we can see that both the central government's expected gain and the local government's actual gain are much greater than when they rely on only the one variable π and that the total agency cost of the central government also decreases significantly.

5 Conclusion and policy suggestions

The above analysis of the principal-agent model shows that in the process of low-carbon ecological urban construction, the central government has the disadvantage of information asymmetry. For any new noticeable variable k (only if observing k does not cause any cost or if the cost of observing k is less than the reduced agency cost it brings about will k be meaningful), as long as k contains more information about θ or a than the given variable π contains, and k is put into the incentive contract as a new term, the central government's agency cost will be greatly reduced and the central government's expected gain will be increased. For the central government, both the rationality of awarding or punishing the local government and the executive cost of supervision and inspection are mainly determined by the information content the central government acquires. Sufficient information not only ensures that the central government makes reasonable regulations but also prevents the local governments from the "moral risks" in low-carbon urban construction, which is helpful for promoting the development of the urban recycling economy. Therefore, to realize the optimal balance in the game of low-carbon ecological urban construction between the central government and the local government, we must design a system that can change the information asymmetry status to strengthen the central government's ability to acquire information and to stimulate or restrict

the local government's action choice. It can turn the external pressure of low-carbon ecological urban construction into the local government's own internal actions. The policy suggestions are as follows:

(1) Strengthen the awe and reliability of the agency policy and perfect the central government's monitoring mechanism. Now that the individual in the game is rational, whether the agent would provide the true information is decided by whether it can get more gains than if it hides the true information. Therefore, if the principal wants to achieve maximum utility, it pays the agent in such a way that the latter knows that providing the true information is its optimal choice. With a lack of a direct or indirect information channel, the central government can get a low-information-cost contract by changing the game rules, i.e., seeking an optimal institutional arrangement.

(2) Establish the "reputation" of a city's ecological environment and the mechanism of publicly releasing the environmental information. In some sense, for the final establishment of an urban environmental management system, only the tangible system of restricting the local government is not enough, and it is necessary to establish an intangible system - "reputation". By strengthening the propaganda of the urban ecological environmental "reputation", the local government is forced to disclose the environmental information and to regard the construction of its own social image as one of the construction goals, with a huge pressure from public opinion, which will unify the economic and social benefits.

(3) Establish an effective mechanism of information exchange and promotion. This plays an extremely important role in reducing transaction costs and in realizing the optimal game balance. Firstly, better information openness of the local governmental affairs means asking the relevant departments to timely provide environmental information, such as information about the companies that produce waste, information about the waste's features and potential value, information about the circulation companies and their market access, information about the latest recycling technologies, and information about the policy support. This will promote the spread of environmental information and entirely change the current information asymmetry status in the process of developing a recycling economy. Secondly, gradually perfect the operation mechanism of marketizing the environmental information, greatly promote the rapid

development of the medium organizations of environmental protection information, and establish a highly efficient exchange platform of recycling economy information to make up for the deficiency of the local governmental information openness.

(4) Foster the public subject consciousness of recycling economy and perfect the information feedback mechanism. The complexity of the resource and environment problems determines the necessity of the wide participation in the development of the recycling economy, and the public participation scale and quality is determined by the public ideas about ecological protection. First, the central government should create an environmental information publishing network and a green service centre. They can be used to strengthen the propaganda and education of green knowledge and the effective communication with the public. Finally, make full use of news media and non-governmental organizations. The news media widely represent the public opinions and have a non-obligatory supervision function, so their reports of hot issues and market quotations can speed the development of information openness to a certain degree. Non-governmental organizations can also timely reflect the public demands. The environmental protection they launch helps to promote green consumption and to increase the level of market openness and norms. Therefore, encourage their relevant propagandas and fully exert their function of conveying information. They can introduce the community and market incentive mechanism into the control of environmental products and set up a communication bridge between social members.

Acknowledgments

Sincerest thanks are extended to the following funding programs for their support: (1) The National Fund Program of Social Science: Research on the "Ecological Co-management" Mechanism between the Interest Bodies in the Process of the New-style Urbanization of Central Plain Economic Region (13CY036); (2) The National Key Fund Program of Social Science: Research on the New-style People-oriented Urbanization Path in China (13&ZD025); and (3) The Philosophy and Social Science Program of Anhui Province: The Empirical Research on the Growth Theory of Green TFP (total factor productivity) and the Economic Growth in Northern Region of Anhui Province (AHSK11-12D105).

References

- [1] China Statistical Yearbook 2013 Beijing: China Statistics Press
- [2] Fan J P 2011 Review of low-carbon city *China population, resources and environment* 21(3) 478-81
- [3] Christoff P 1996 Ecological modernisation, ecological modernities *Environmental Politics* 5(3) 476-500
- [4] White D F 2002 A Green Industrial Revolution? Sustainable Technological Innovation in a Global Age *Environmental Politics* 11(2) 1-26
- [5] Gerald B, Andrew F, Frances H, Richard J 2001 Ecological Modernization as a Basis for Environmental Policy: Current Environmental Discourse and Policy and the Implications on Environmental Supply Chain Management *Innovation: The European Journal of Social Science Research* 14(1) 55-72
- [6] Kuroki K, Usui H, Onari S, Arita R, Aoki H 2009 Pnictogen height as a possible switch between high- T_c nodeless and low- T_c nodal pairings in the iron-based superconductors *Physical Review B* 79(22) 34-47

- [7] Thuli N M, Coleen H V 2010 Challenges to achieving a successful transition to a low carbon economy in South Africa: examples from poor urban communities *Mitigation and Adaptation Strategies for Global Change* 15(3) 205-22
- [8] Glaeser E L, Kahn M E 2001 Decentralized Employment and the Transformation of the American City *NBER Working Paper No. 8117* 2001(4) 346-65
- [9] Crawford J, French W 2008 A low-carbon future: Spatial planning role in enhancing technological innovation in the built environment *Energy policy* 36(12) 4575-9
- [10] Diana M L, Silvina V 2006 Neoliberalism and the Environment in Latin America *Annual Review of Environment and Resources* 31 327-63

Authors



Zhang Shuiping, born on May 24, 1974, Anhui Province, China

Current position, grades: Master's Degree

Scientific interest: Environment Economy

Experience: Zhang Shuiping, male, Born in Anhui Province, May 24, 1974, associate professor, head of economics department, Graduate teacher research supervisor.

A multi objective optimization algorithm for recommender system based on PSO

Zhao-Xing Li^{1, 2*}, Li-le He¹

¹*School of Mechanical and Electrical Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China*

²*School of Information Engineering, Yulin University, Yulin 719000, China*

Received 3 March 2014, www.tsi.lv

Abstract

In order to follow the development of Internet information service and improve the accuracy of recommender systems and recommendation algorithm. An optimal selection approach of multi-objective and particle swarm optimization (MOP-PSO) was put forward based on PSO algorithm. Furthermore, through two sets are combined and repeated dynamic adjustments, to achieve a better balance in algorithm efficiency and accuracy. Proposed a weighted cosine similarity method to calculate the user similarity, and then optimizing the weight by the PSO algorithm. Simulation results show that the algorithm has a better effective and can effectively improve the scoring accuracy, effectively improve the quality of the recommendation system.

Keywords: particle swarm optimization, recommended system, multi objective optimization

1 Introduction

Along with the degree rise of Internet service socialization, as well as the popularity of the mobile Internet, the one-way service mode of traditional Internet is affected by new Internet service mode, Users are no longer satisfied with traditional passive network service information received in the role, but also hope that the manufacture and dissemination of information in the network activities, such network applications are driving the demand for services in the form of bi-many Internet applications transition mode, thus make the network information services into web 2.0 application stage [1]. In particular, such as blog and micro blogging and other social networking applications in the form of attracting a large number of individual users and business users and even has become an important way to national government agencies as well as politicians and the public to communicate and exchange.

In order to retrieve information on the Internet resources as soon as possible, search engines technology has become the preferred solution for users [2]. The major search engines are based on information retrieval technology, user manual input to keyword based on information and search [3]. In addition, the traditional information service model is a passive mode of service; information service providers can only passively wait for the user's service request and to provide users with personalized information free.

In response to these problems, and limitations of information overload, academia and industry search engine and information service mode passive proposed recommendation system solution [4]. Recommended

system can be based on user preferences historical interest, the initiative to provide users with information resources in line with their needs and interests. Recommended system has become a hot research direction of data mining, machine learning and human interface areas. However, with the further development of Internet applications, the traditional recommendation system and its algorithm is difficult to adapt to the user scale, the concept of the rapid growth of the number of projects recommended data and user history score, score data sparsity and user interest drift problems caused Recommend decreased quality, user satisfaction, reduced, or even the loss of a large number of users, which have seriously hampered the further promotion and application of the recommendation system. Therefore, the current recommendation algorithm to solve the problems and improve the recommendation accuracy of the recommendation system theory and practice has very important significance [5]. In this background, this paper will be recommended for the target accuracy of the recommendation system, drift issues and concepts for the sparsity recommendation system to carry out research work, and propose an improved recommendation algorithm has some innovative. In order to improve the accuracy of prediction, we propose a new similarity calculation method-weighted cosine similarity, and through PSO optimization calculation it weights. Experiments show that this method can effectively improve the accuracy of predictions. Figure 1 shows the general model of recommendation systems; recommender systems play an important role in modern daily life.

* *Corresponding author:* e-mail: 39812767@qq.com

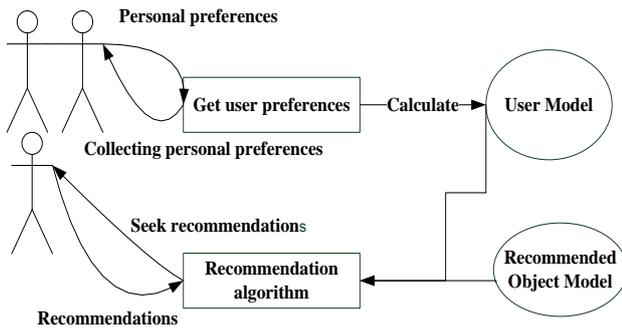


FIGURE 1 Recommended system model

2 Recommendation algorithm

2.1 PARTICLE SWARM OPTIMIZATION ALGORITHM

Particle swarm optimization algorithm (PSO) is proposed by Kenney and Eberhart in 1995 population parallel search algorithm based on global optimization [6, 7], through cooperation and competition between groups in the community to achieve optimal particle. Mathematical description of PSO: a population size is n, the i particles in m dimensional search space representation of $X_i = (X_{i1}, X_{i2}, \dots, X_{ij}, \dots, X_{im})$, flight speed is $V_i = (V_{i1}, V_{i2}, \dots, V_{ij}, \dots, V_{im})$, the optimal position of individual so far to search is $P_i = (p_{i1}, p_{i2}, \dots, p_{ij}, \dots, p_{im})$. The particle swarm optimal position is $P_{gbest} = (p_{gbest1}, p_{gbest2}, \dots, p_{gbestm})$. It can update the particle velocity and position according to the formula (1) and (2):

$$v_{ij}(t+1) = \omega \cdot v_{ij}(t) + c_1 \cdot r_1 \cdot (p_{ij} - x_{ij}(t)) + c_2 \cdot r_2 \cdot (p_{gbestj} - x_{ij}(t)), \tag{1}$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1). \tag{2}$$

where t represents the t iteration, $j=1,2,\dots,n$; $j=1,2,\dots,m$; $c_1, c_2 > 0$ are respectively the individual learning factor and social learning factors; t is the current number of iterations, r_1 and r_2 are uniformly distributed random numbers in the range of [0,1]. ω is the inertia weight coefficient, used to control the effect of history on current speed. In order to balance the global and local search ability, make the ω along with the increase in the number of iterations decreases linearly, can significantly improve the performance of the PSO algorithm, it is given

$$\omega = \omega_{\min} + (\omega_{\max} - \omega_{\min}) \times \frac{(\omega_{\max} - \omega_{\min})}{iter_{\max}}, \tag{3}$$

wherein ω_{\min} , ω_{\max} respectively the maximum and minimum weighting factor, iter is the current iteration number, $iter_{\max}$ is the total number of iterations.

In the formula (3), ω_{\max} is the initial inertia weight; ω_{\min} is the last inertia weight; t_{\max} is the maximum number of iterations. Flight speed is $v_i \in [-V_{\max}, V_{\max}]$, the constraint conditions to prevent particle speed missed optimal solutions, through the improvement of the algorithm further improves the global searching ability of particle swarm.

2.2 MULTI-OBJECTIVE OPTIMIZATION

Objective optimization problem was originated in the design of many complex systems, modelling, planning issues. Since the 1960s, multi-objective optimization problem attracted the attention of a growing number of researchers from different backgrounds [8]. Especially in recent years, multi-objective evolutionary algorithm is to obtain the optimization of the more widely used and studied, resulting in a series of novel algorithms and get a good application. Multi-objective optimization proposition is generally no unique global optimal solution, so this is actually a multi-objective optimization proposition is often how the process of seeking Pareto set. Traditional multi-objective algorithm is often converted into a single objective proposition after the use of sophisticated single-objective optimization algorithm, the drawback is that the optimal solution can only be determined once a solution. And now the multi-objective evolutionary strategy tends to be more parallel computing can be solved once a sufficient number of solutions distributed on the Pareto front provides decision-makers to the next decision. Which PSO as a novel evolutionary computing strategy has been more and more widely used in multi-objective optimization problem. Multi-objective optimization is described as follows:

Definition 1:

$$\min f(x) = [f_1(x), f_2(x), \dots, f_n(x)]$$

$$st. g_i(x) \leq 0, i = 1, 2, \dots, m$$

$$h_j(x) = 0 \quad j = 1, 2, \dots, n,$$

where $x = [x_1, x_2, \dots, x_D] \in R^D$,

$f_i(x) (i = 1, 2, \dots, n)$ is the objective function;

$g_i(x) (i = 1, 2, \dots, m)$ is the inequality constraint;

$h_j(x) (j = 1, 2, \dots, p)$ is the equality constraints.

Definition 2: global optimum

Given a multi-objective optimization of the overall proposition $f(x), f(\bar{x})$ is called optimal if and only if $\forall x \in \Omega, have f(\bar{x}) \leq f(x)$.

3 Multi objective optimization recommendation algorithm based on PSO

3.1 MULTI OBJECTIVE PARTICLE SWARM OPTIMIZATION ALGORITHM

In order to improve the quality of recommendation system, the paper-based multi-objective optimization PSO algorithm recommended by the relevant principles, we propose a means of optimizing the recommendation system. Entire score prediction algorithm roughly as follows:

(1) Initial population, the initial population size is denoted by n . Based on the idea of fitness domination, it is divided into two sub-groups, and the initial population size is denoted by n . A subset of non-dominated is called (P_Set), referred to dominate another subset is (NP_Set), wherein (P_Set) and (NP_Set) are referred to as a base, respectively n_1, n_2 , and satisfy $n_1 + n_2 = n$ ($1 \leq n_1, n_2 < n$).

Apparently satisfied: $\forall x_i \in NP_Set$, at least $\exists x_j \in P_Set$, then x_i control x_j , where $1 \leq j \leq n_1, 1 \leq i \leq n_2$.

Nearest neighbour choice is personalized recommendation system is a very important step, multi-target system depends largely on the quality of the recommendation algorithm to find the neighbour effect. During each iteration, only the particles NP_Set and position update rate, and updated based on the particles NP_Set disposable Thought fitness particle P_Set comparison, if $\forall x_i \in NP_Set, \exists x_j \in P_Set$, made x_j control x_i . Then the switch x_i, x_j position, that is $x_j \in NP_Set, x_i \in P_Set$.

Then build the project configuration file, build the project configuration file contains basic user information, project information, rating, and other features of values based on the characteristics of the project, the project configuration file is a comprehensive reflection characteristics of the project.

(2) Nearest neighbour extract candidate set. Using the maximum intersection method using matrix operations to extract maximum current projects of common user items rated as the nearest neighbour candidate sets.

(3) The optimal project similarity calculation. The idea of this paper is that the weight of each set of eigenvalues first project configuration file attributes heavy initial project, and then build the weighted cosine similarity function between the current project and the project's neighbours, and through PSO optimization algorithm in the training set of weights value optimization, prediction error when the score reaches the most hours of the nearest weight value, calculate the current project and the project's neighbours according to a recent weights recently similarity solution.

(4) Current Ratings forecasts. According to the characteristics of the optimization project between the neighbours being the best score and after the two projects

worth similarity to recent experiments focused solutions to existing projects predicted score. In the prediction process, first select the current project does not score and score on a neighbour project users, according to a recent candidate set of previously established neighbours to the right of the current project is the recent similarity profiles and project configuration files between neighbours weight score to predict the current project.

(5) Make recommendations. In the user - after item rating matrix to fill, select a user forecast projects a higher score to make recommendations for the current user.

Realization of the program algorithm is as follows:

Step 1: Initial population, x_1, x_2, \dots, x_n , ($n = Pop_Max$), where set the dimension is D .

Step 2: Initialization fitness populations, for $i = 1$ to n .

Calculate $f(x_i) = [f_1(x_i), f_2(x_i), \dots, f_n(x_i)]$.

Based on the concept of fitness dominance, the initial population is divided into two subgroups, referred to as non-dominated subset of a subset of P_Set and NP_Set domination, P_Set subset of cardinality Pareto denoted n_1 , N_Pareto subset NP_Set of cardinality denoted n_2 .

Step 3: Velocity and position of each particle in the conduct of particles NP_Set update, location update: $V_p = \omega \times V_{id} + c_1 \times r_1 \times (P_{id} - x_{id}) + c_2 \times r_2 \times (P_d - x_{id})$.

Speed update: $x_{id} = x_{id} + V_{id}$, where P_{gd} is the entire population from a randomly selected subset of P_Set .

Step 4: Dynamic exchange strategy: NP_Set subset of each particle and each particle P_Set comparing each subset, and the subset of NP_Set . Vocabulary particles is $x_1, \dots, x_i, \dots, x_{n_2}$, P_Set is a subset of particles is $x_1, \dots, x_j, \dots, x_{n_1}$.

for ($k = 1$ to n_2)

for ($t = 1$ to n_1)

if ($f(x_i) \leq f(x_j)$)

Swap x_i and x_j , and update their number and location in the collection.

End if

End for

Comparison of x_i and x_j after all, there is j made the $f(x_i) \leq f(x_j)$ established, and then clearly, x_i may be a non-dominated solutions. Therefore, the x_i also joined P_Set subset. It is to Update P_Set and NP_Set subset, if P_Set has k elements duplicate, delete and re-initialized k particles in NP_Set ; if there are duplicate k elements in NP_Set , for a similar operation.

Update n_1, n_2 .

If $n_1 \neq n$ or the maximum number of iterations is not reached, then jump to Step 3.

3.2 WEIGHT OPTIMIZATION MULTI-OBJECTIVE PARTICLE SWARM ALGORITHM

In the experiment, it is assumed that each item characteristic is stable weight value, and therefore optimization of the process, an optimal solution this project there is only the weight value, i.e., there is only a single fitness function. This adaptation function, which is defined as the average of the prediction error rates between the two items. Thus, the goal of the optimization is to calculate each current project W . Adapted as a function of the expression (4) below, where n is the number of users of the T and j exists between the joint score. T representative of the current project, j represents a neighbourhood of T , av_T on behalf of T being the mean score, $similar(T, j)$ is the similarity of T and j , $VT(i, j)$ represents for i users on the project of the j score, av_i on behalf of all users of the mean score of i ,

$$fitness = \frac{1}{n} \left\{ (av_T + u \sum_{i=1}^m similar(T, j) \times (VT(j, i) - av_i)) - VT(T, i) \right\}, \tag{4}$$

where $similar(T, j) = \sum_{i=1}^n \frac{TW \cdot jW}{|TW| \times |jW|}$, (5)

where T represents the current project configuration file; j on behalf of the project selection process to select the configuration file out of the neighbourhood project configuration file, and $T \neq j$; W is a $K \times K$ diagonal matrix, the diagonal matrix for each value represents the value of the heavy weight of each feature, n represents the total number between two projects have a common score users.

PSO optimization procedure is divided into three steps, first, initialization of the particle velocity and position, using an experimental paper initialize random process; Secondly, the rules according to the dynamic movement of particles, establishing a new position of particles and particle velocity iteration; then establish each step local optimum and the global optimum particle, by the position of each particle, and the fitness function to get the value of the particle positions, and decide whether to update the local optimum and the global optimum.

3.3 USER RATING PREDICTION AND RECOMMENDATION

After establishing the nearest neighbour recent and current projects right profile heavy, you can start score prediction.

(1) Score prediction. Score prediction formula as formula (6) are shown, Wherein $C_VT(T, i)$ is a

prediction of the current score of the item, av_T Q is the average of all the scores of the project T , $VT(T, i)$ Q is the user i to the project T scores of the configuration file, j is the user i to the project j scores of the configuration file, $similar(T, j)$ is the training set optimized project configuration files T and j recently cosine similarity value, $VT(i, j)$ is user i to the project of the j score, av_i is the average user i all items on the score, n is the number of neighbour selection, k as the standard parameters.

$$C_VT(T, i) = av_i + u \sum_{j=1}^n similar(T, j) \times (VT(j, i) - av_i) \tag{6}$$

(2) Be recommended. Recommended system recommended by different algorithms based on differ mainly in two ways: First, when extracting the most similar to the current nearest neighbour user interest, users will be interested in the project's neighbours to recommend to the current user; its Second, the presence of score recommendation system, the nearest neighbour first extracted, and then the prediction of the current project, or user rating score according to the neighbour, then the current user of the prediction score ranking, for extracting several high recommendation rating.

4 Experiment and Analysis

In this paper, experiments using Jingdong online shopping evaluation data set (<http://www.jd.com/>). The dataset contains 123,883 users and 2190 project (this paper is to mobile phones) and those users detailed information on the project 200,000 votes.

During the experiment, in order to verify the accuracy of the predicted score, the whole data set were randomly divided into A, B two parts, crossover turns twice. In each experiment, the selected item as a two-part test ratings training set for training right weighted cosine similarity weight; remainder Grading set of experiments to verify the validity as to optimize the results.

Specifically, we use a randomly selected set of methods currently selected item from the data sequence A, B, as the training set, B, A set of experiments was performed twice crossover, divided into two groups in each experiment as follows:

(1) Randomly extracted 20 project as the active item, using the maximum number of items extracted intersection nearest neighbour candidate set of n ($n = 5, 15, 35$).

(2) The number of items randomly extracted 20 items as the active item, randomly extracted nearest neighbour candidate set of n ($n = 5, 15, 35$). When the experiment 1 and experiment 2, the nearest neighbour is different extraction candidate sets to verify the maximum extraction of the intersection nearest neighbour candidate

set, the accuracy of the prediction score increased. By experiments 1 and 2, a different number of neighbours recently extracted to determine the size of the nearest neighbour extraction.

In the training set, the same set of parameters for the PSO, as follows:

(1) The number of particles: $N = 50$; learning factor: $c_1 = 1$, $c_2 = 1$; inertia weight: $W = 0.6$; maximum number of iterations: $M = 2000$.

Figure 2 represents in A, B is the training set, B, A is the cross-experiment experimental set twice; In the two experiments, each randomly selected 30 current phones, and compare the intersection of law in accordance with the maximum of n bits select the nearest neighbours and randomly selected candidate n-bit nearest neighbour candidate scores received predictable results. In Figure 2, the horizontal axis represents the size of the nearest neighbour candidate set from 4-35, and the vertical axis represents the 30 current project a combined average score prediction error value MAE. Experimental results show that the maximum intersection nearest neighbour method is selected when the prediction accuracy of the obtained score is higher than the overall results of randomly selected, and the number of candidates is 21, 22, respectively, the prediction accuracy results stabilized.

Comprehensive experimental results show that the intersection of law and the maximum weighted cosine similarity score improved the accuracy of prediction, which can improve the quality of the system recommended by the recommendation.

5 Conclusions

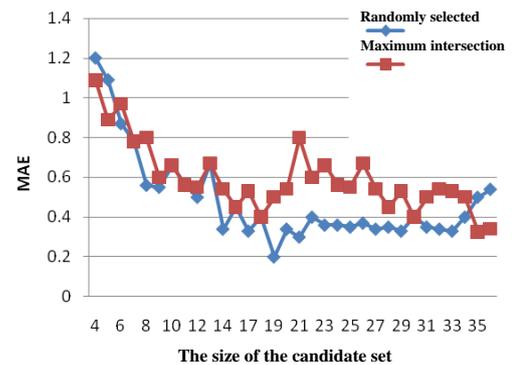
This paper starts from the status quo of Internet information service, introduces the application background of the recommender system, then the multi-objective particle swarm optimization algorithm of recommendation system is improved.

This paper makes the algorithm improvement goal focused mainly on the recommendation algorithm prediction accuracy, although the improvement and perfection of the recommendation algorithm increases the calculation task recommendation system in a certain extent, but from the computational complexity point of view, these improvements in computational performance

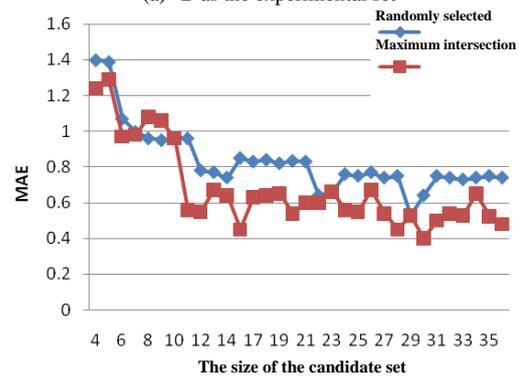
References

- [1] Adomavicius G, Tuzhilin A 2005 *IEEE Transactions on Knowledge and Data Engineering* 17(6) 734 - 49
- [2] Kurkovsky S, Harihar K 2006 *Personal and Ubiquitous Computing* 10(4) 227-40
- [3] Cherkassky V, Yunqian Ma 2004 *Neural Networks* (S0893-6080) 17(1) 113-26
- [4] Zheng Y, Burke R 2012 Mobasher B. Optimal feature selection for context-aware recommendation using differential relaxation *ACM RecSys* 12
- [5] Üstün B, Melssen W J 2005 *Analytical Chimica Acta* (S0003-2670) 544(1-2) 292-305
- [6] Yaoyu Feng, Zhimin He 2003 *Enzyme and Microbial* 32(2) 282-9
- [7] Zhang Yong-Heng, Zhang Feng 2013 *Sensors and Transducers* 151(4) 95-100
- [8] Cao X, Chen S Y 2009 *Computer Standards and Interfaces* 31(3) 579-85

of the recommendation system is very limited, remained in the recommended time the algorithm complexity level. In order to validate the proposed algorithm in this paper carry high recommendation accuracy, by comparing the prediction accuracy of each algorithm and the traditional recommendation algorithm, experimental results show that the algorithm proposed in this paper are different degree improves the prediction accuracy of the recommended, show that these algorithms are effective measures for improvement.



(a) B as the experimental set



(b) A as the experimental set

FIGURE 2 The maximum intersection method and random extraction experiments compare the results

Acknowledgments

This work is partially supported by the Research and Cooperation Project for Department of Yulin city of (#Gy13-15). Thanks for the help.

| Authors | |
|---|---|
|  | <p>Li Zhao-Xing, born on March 9, 1982, in Shannxi Yulin</p> <p>Current position, grades: lecturer in Yulin University. University studies: MS degree in Computer science from Xidian University in 2010. Scientific interest: Cloud integrated manufacturing technology, big data, Evolutionary Algorithm</p> |
|  | <p>He Li-Le, born on March 18, 1966, in Shannxi Xi'an</p> <p>Current position, grades: professor in Xi'an University of Architecture and Technology University University studies: MS degree in Computer science from Xi'an University of Technology University in 2006. Scientific interest: Image Understanding and Change Detection, Evolutionary Computation, Fuzzy Systems and Neural Networks</p> |

The effectiveness of using methods two-stage for cross-domain sentiment classification

Hoan Manh Dau^{*}, Ning Xu

School of Computer Science, Wuhan University of Technology, China

Received 14 May 2014, www.tsi.lv

Abstract

Traditional sentiment classification approaches perform well in sentiment classification but traditional sentiment classification approaches does not perform well with learning across different domains. Therefore, it is necessary to build a system which integrates the sentiment orientations of the documents for every domain. However, this needs much labelled data involving and much human labour as well as time consuming. Thus, the best solution is using labelled data in one existed in source domain for sentiment classification in target domain. In this paper, a two-stage approach for cross-domain sentiment classification is presented. The First Stage is building a bridge between the source domain and the target. The Second Stage is following the structure. The study shows that the mining of intrinsic structure of the target domain brings a considerable effectiveness during the process of sentiment transfer. This is a typical mining approach comparing to previous approaches basing on information from the source domain to address the task of sentiment transfer, which does not depend on intrinsic structure of the target domain. Experimental results on sentiment classification with a two-stage approach indicate that the effectiveness outperforms other traditional methods.

Keywords: cross-domain, sentiment classification, sentiment transfer, opinion mining

1 Introduction

Sentiment classification is attracting more and more people's attention because of its great benefits to social and human life. Automatic sentiment classification aims to predict automatically sentiment polarity (e.g., positive or negative) of users who publish sentiment data. Sentiment classification is the area of sentiment analysis can help human analyse, synthesize, organize, summarize and forecast for determining the sentiment orientation of subjective text. This is an important sub-task of sentiment analysis. It plays an important role in numerous applications like opinion mining, market analysis and opinion summarization. Today, when internet services bloom as mushrooms with many social networking sites, handsets can connect to the network and many people create sentiment data to share on the Web. Users express and share their opinions about many topics on Websites and blogs. Researching of sentiment classification has contributed to text classification research and therefore it has an important significance for those who want to forecast information from text document data. Researchers have pointed out that sentiment classification has been applied effectively, such as [1-4].

In most cases, supervised learning methods for sentiment classification have been studied popularly and applied rather successfully. Researches have showed that standard machine learning techniques definitively outperform human-produced baselines. However, a disadvantage of sentiment is expressed differently in different domains, and it is the requirement for labelled in

domain data for training. Supervised learning methods for sentiment classification require two conditions to ensure the accuracy in classification. The first condition is that training data is sufficient and labelled well; and the second is that training data and test data should have the same distribution. However, in reality these two conditions cannot be met. The main reason is that labelling data involves much human labour and it is time-consuming; apart from that the labelled and unlabelled data are often from different domains, and often have different distributions. Therefore, the problem is how to use labelled sentiment data in source domain for sentiment classification in target domain. This is the main task of cross-domain sentiment classification. When performing cross-domain sentiment classification (or sentiment transfer), many researchers gave some techniques to improve them in order to make the process of sentiment transfer more effective such as [5, 6]. However, the two problems of sentiment transfer are that the distribution of the target domain is not same with that of the source domain and the intrinsic structure of the target domain is static. To solve these two problems, a bridge needs building to share information between source domain and target domain and the intrinsic structure needs used to carry out for target domain. Selecting technique to build a bridge between the source and the target domain will impact on the effectiveness of the process of sentiment transfer. Transfer learning aims to use data from other domains to help current learning task. Transfer learning plays important role in research field in machine learning. There were some typical

^{*} *Corresponding author* e-mail: daumanhhoan@yahoo.com

researches such as: introducing a statistical formulation to domain adaptation in terms of a simple mixture model [7] introducing a two-stage approach to domain adaptation for statistical classifiers [8] proposing a bridged algorithm, which takes the mixture distribution of the training and test data as a bridge to better transfer from the training data to the test data [9] presenting an adapting naive Bayes to domain adaptation for sentiment analysis [10]. However, some researchers based only on the labelled documents to improve the performance of sentiment transfer [11, 12]. Most of these researchers used information from the source domain to address the task of sentiment transfer but ignored the intrinsic structure of the target domain.

In this paper, technique for transfer learning in the context of sentiment classification is presented. The effectiveness of applying the SentiRank algorithm in the process of sentiment transfer and mining intrinsic structure of the target domain which brings feasible effectiveness are shown. The testing results are presented and showed that the effectiveness of this approach when making sentiment transfer is considerable.

2 The proposed approach

2.1 PROBLEM DEFINITION

There are two document sets in this paper: D^U denotes the test data, and D^L denotes the training data. Assign every document a sentiment score ("1" denotes positive, and "-1" denotes negative) to represent its degree of sentiment orientation and call it sentiment score. S^U denotes the sentiment score set of D^U , and S^L denotes the sentiment score set of D^L . It is assumed that the training dataset D^L is from the related but different domain with the test dataset D^U . The aim is to maximize the accuracy of assigning a label in D^U utilizing the training data D^L in another domain.

2.2 OVERVIEW

A two-stage approach for sentiment transfer is given.

The implementations of this approach are shown in [6]. The process consists of two stages (two-stage):

- the first stage: building a bridge;
- the second stage: following the structure.

In the first stage, there are 3 steps:

- the first step is to use SentiRank algorithm to get the sentiment scores of the target domain documents.
- the second step is to get Initial Sentiment Scores of the Target Domain Data.
- the third step is to choose Seed Set of confidently labelled documents as high-quality.

In the second stage, there are two steps:

- the first step is to apply a manifold-ranking algorithm to follow the structure of the target domain.
- the second step is to use the manifold-ranking scores to label the target-domain data.

2.3 THE FIRST STAGE: BUILDING A BRIDGE

In this stage, firstly the SentiRank algorithm is used to build a bridge between the source domain and the target domain. In order to get the labels of the target domain documents, the information of the source domain is used. The SentiRank method [13] is an algorithm for sentiment transfer and it is used to get the sentiment orientations of the target-domain documents using the similarity between the documents from both the source domain and the target domain. The implementation of the algorithm is that if one document has a strong relationship with positive documents or negative documents, it can probably be positive or negative. The implementation of SentiRank is described as follow: A weighted graph is built from the data, and a sentiment score is assigned for every labelled and unlabelled document to denote its extent to "negative" or "positive", then the score is iteratively calculated making use of the accurate labels of source domain data as well as the "pseudo" labels of target domain data via the weighted graph. The final score for sentiment classification is achieved when the algorithm is converged, so the target domain data can be labelled based on these scores. The SentiRank process is described in details in [6].

In this algorithm, α and β show the relative importance of source domain and target domain to the final sentiment scores, and $\alpha + \beta = 1$. Algorithm achieves the convergence when the changing between the sentiment scores computed at two successive iterations for all documents in the target domain falls below a given threshold. Secondly, in order to find high quality documents from the target domain, sentiment score needs creating and using to denote the "negative" or "positive" correlation of documents. Next, the target domain documents is sorted in descending order according to their sentiment scores. So the more forward the document is sorted, the more likely it is positive; the more backward the document is sorted, the more likely it is negative. Then, the first K documents and last K documents as the high quality documents are chosen. Thirdly, Seed Set of confidently labelled documents as high quality is chosen. This algorithm proves that results produce high quality seeds. Sorting the target domain documents according to their opinion extent is effective and the proof is shown in Table 1.

TABLE 1 Seed accuracies on six tasks [6]

| Do main | K | | | | | | |
|---------|--------|--------|--------|--------|--------|--------|--------|
| | 50 | 90 | 130 | 170 | 210 | 250 | 290 |
| B→H | 0.9500 | 0.9222 | 0.9230 | 0.9294 | 0.9333 | 0.9340 | 0.9240 |
| B→N | 0.8200 | 0.8778 | 0.8923 | 0.8912 | 0.8905 | 0.8820 | 0.8860 |
| H→B | 0.8000 | 0.8055 | 0.8115 | 0.8117 | 0.8024 | 0.7540 | 0.7431 |
| H→N | 0.9300 | 0.9277 | 0.9230 | 0.9235 | 0.9214 | 0.9100 | 0.9086 |
| N→B | 0.7400 | 0.7500 | 0.7461 | 0.7264 | 0.7142 | 0.7120 | 0.6810 |
| N→H | 0.9167 | 0.9111 | 0.9000 | 0.8976 | 0.8990 | 0.8980 | 0.8972 |

TABLE 2 Accuracy comparison of different methods [6]

| Domain | Proto | TSVM | SentiRank | EM based on Proto | Manifold based on Proto | Main Approach |
|---------|-------|-------|-----------|-------------------|-------------------------|---------------|
| B→H | 0.735 | 0.749 | 0.772 | 0.765 | 0.761 | 0.790 |
| B→N | 0.651 | 0.769 | 0.714 | 0.667 | 0.745 | 0.776 |
| H→B | 0.645 | 0.614 | 0.671 | 0.723 | 0.677 | 0.683 |
| H→N | 0.729 | 0.726 | 0.749 | 0.657 | 0.784 | 0.784 |
| N→B | 0.612 | 0.622 | 0.638 | 0.763 | 0.665 | 0.650 |
| N→H | 0.724 | 0.772 | 0.764 | 0.765 | 0.779 | 0.791 |
| Average | 0.683 | 0.709 | 0.718 | 0.723 | 0.735 | 0.746 |

Table 1 shows that the effectiveness of the algorithm carried out from domains $B \rightarrow H$, $H \rightarrow N$ and $N \rightarrow H$ has the accuracy of above 89%, and the effectiveness from domains $B \rightarrow N$ and $H \rightarrow B$ has the accuracy of above 75%. This high accuracy demonstrates that the effectiveness of algorithm is enough to choose high-quality seeds. In the case of transfer from domain $N \rightarrow B$, the accuracy is not particularly good. The main reason is due to the too big difference between the two domains notebook (N) and book reviews (B). However, this shortcoming can be overcome and can improve the performance of sentiment transfer exploiting these seeds.

2.4 THE SECOND STAGE: FOLLOWING THE STRUCTURE

In this stage, although the algorithm can build a bridge between the source domain and the target domain, the distribution of the target domain is not used but the intrinsic structure of the target domain is used for sentiment transfer. It starts with a small amount of high quality seed set, this is the number of seeds representing for intrinsic structure of the target domain. Manifold-ranking method is used to make better use of the seeds, and it can improve the performance of sentiment transfer.

The manifold-ranking method [14] is a universal ranking algorithm and it is initially used to rank data points along their underlying manifold structure. The prior assumption of manifold-ranking is nearby points which are likely to have the same ranking scores and points on the same structure are likely to have the same ranking scores. The implementation of the algorithm is as follow: a weighted network is formed on the data, and a positive rank score is assigned to each known relevant point and zero to the remaining points which are to be ranked. All points then spread their ranking score to their nearby neighbours via the weighted network. The spread process is repeated until a global stable state is achieved, and all points obtain their final ranking scores [6].

With a high quality seed set, first, the weighted network whose points denote documents in D^U is built.

And then integration the sentiment scores of the seeds into the manifold-ranking process is carried out. Then the sentiment manifold-ranking process is implemented. Finally, label the documents in target domain according to their ranking score vector. Each document is labelled with positive or negative labels.

3 Experiments

3.1. BASELINE SYSTEMS

In this part, testing results of chosen method is shown and compared to the results of other baseline methods.

Table 2 shows that accuracy comparison of different methods [6]:

- Method Proto: the results from column 2 show that the accuracy ranges from 61.25% to 73.5%. It is result of method which applies a traditional supervised classifier, prototype classifier for the sentiment transfer [15]. This technique only uses source domain documents as training data.

- Method Transductive Support Vector Machine (TSMV): the results from column 3 show that the accuracy ranges from 61.42% to 77.17%, which is better than that of method Proto. This method applies transductive SVM for the sentiment transfer [16]. This is a widely used method for improving the classification accuracy. This method uses both source domain data and target domain data.

- Method SentiRank: column 4 shows the results that the accuracy ranges from 63.7% to 77.2%, which is much better than method Proto and TSVM. The implementation of this method is to run SentiRank algorithm at places initializing the sentiment scores by prototype classifier.

- Method Expectation Maximization (EM) based on Proto: the column 5 shows that results of the method of EM algorithm [17] based on prototype classifier is similar to the above apart from changing the training classifier from SentiRank to prototype classifier, and its results are accuracy ranges from 65.7% to 76.5%, better than the first three baselines.

- Method Manifold based on Proto [14]: the column 6 shows that the accuracy ranges from 66.5% to 78.4%, which is better than all other baselines. This method begins by training a prototype classifier on the training data, then by use the similarity scores between the documents and the positive central vector and the similarity scores between the documents and the negative central vector to separately initial the ranking score vectors of the test data. Finally, it is carried out to choose KM documents that are most likely to be positive and KM documents that are most likely to be negative as seeds for manifold-ranking.

3.2 THE MAIN APPROACH

The proposed approach is compared with 5 baseline methods. The column 7 in Table 2 shows the mentioned approach [6]. The approach recommended in this paper is better performed than all the method baselines. Table 2 shows that greatest increase of accuracy is achieved by about 12.7%, when implementing $H \rightarrow N$ compared to method EM based on Proto. The second greatest increases of accuracy is achieved by about 12.5%, when performing $B \rightarrow N$ and the third greatest increases of accuracy is achieved by about 6.7%, when implementing $N \rightarrow H$ compared to method Proto respectively. The greatest average increase of accuracy is achieved by about 6.3% compared to method Proto. The experiment results show that this method can dramatically improve the accuracy when transferred to a new domain. The results in Table 2 also show that the average accuracies of method SentiRank and TSVM are higher than method Proto. The problem is that method SentiRank and TSVM use information of both source domain and target domain while method Proto not. This proves that using the

information of two domains is better than using the information of only one domain for improving the accuracy of sentiment transfer. In addition, it is clear that the average accuracies of three last methods are higher than that of the three first methods. Three last methods use two-stage approaches, while three first methods do not, which proves that two-stage transfer approach is more effective for sentiment transfer. The above results indicate that the approach which is recommended in this paper has feasible effectiveness.

4 Conclusions

In this paper, the effectiveness of implementing two-stage approach for sentiment transfer is presented. The effectiveness of this approach is proved by comparing its testing results to other basic approaches' results. In order to carry out this approach, a bridge between the source domain and the target domain is built and then the intrinsic structure of the target domain to improve the performance of sentiment transfer is used. The typical characteristic of this approach is using the "pseudo" labels technique to create sentiment scores of the target-domain documents by applying the SentiRank algorithm, then using sentiment scores to identify the best domains with labelled documents as high-quality seeds, in the meanwhile using manifold-ranking algorithm for ranking score for every unlabelled document, finally implementing label the target-domain data based on these scores. Testing results on data prove that this approach improves the accuracy, and can be employed as a high-performance sentiment transfer system. Exploiting good points and advantages and extending this approach for other text classification tasks are potential for further research.

References

- [1] Pang B, Lee L, Vaithyanathan S 2002 Thumbs up? sentiment classification using machine learning techniques *Proceedings of EMNLP* 79–86
- [2] Pang B, Lee L 2004 A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts *Proceedings of ACL* 217–78
- [3] Pang B, Lee L 2008 Opinion mining and sentiment analysis *Foundations and Trends in Information Retrieval* 2(1-2) 1–135
- [4] Hu M, Liu B 2004 Mining and summarizing customer reviews *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining* Seattle WA USA ACM 168–77
- [5] Liu K, Zhao J 2009 Cross-Domain Sentiment Classification Using a Two-Stage Method *Proceedings of the 18th ACM conference on information and knowledge management* 1717–20
- [6] Qiong Wu, Songbo Tan 2011 A two-stage framework for cross-domain sentiment classification *Proceedings Expert Systems with Applications* 38(11) 14269–75
- [7] Daumé III H, Marcu D 2006 Domain adaptation for statistical classifiers *Journal of Artificial Intelligence Research* 26 101–26
- [8] Jiang J, Zhai C 2007 A Two-Stage Approach to domain adaptation for statistical classifiers *Proceedings of the 16th ACM conference on Conference on information and knowledge management* Pages 401–10
- [9] Xing D, Dai W, Xue G, Yu Y 2007 Bridged refinement for transfer learning *Proceedings of the 11th European Conference on Practice of Knowledge Discovery in Databases (PKDD)* Springer 324–35
- [10] Tan S, Cheng X, Wang Y, Xu H 2009 Adapting naive Bayes to domain adaptation for sentiment analysis *Proceedings of the 31st European Conference on IR Research (ECIR)* Toulouse France April 2009 337–49
- [11] Tan S, Wang Y, Wu G, Cheng X 2008 Using unlabelled data to handle domain-transfer problem of semantic detection *Proceedings of the 2008 ACM symposium on Applied computing* 896–903
- [12] Dasgupta S, Ng, V 2009 Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification *ACL-IJCNLP 2009 Proceedings of the Main Conference* 701–9
- [13] Wu Q, Tan S, Cheng X 2009 Graph ranking for sentiment transfer *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* 317–320
- [14] Zhou D, Weston J, Gretton A, Bousquet O, Schölkopf B 2003 *Ranking on data manifolds* Advances in neural information processing systems (NIPS) 16 169–76
- [15] Han E, Karypis G 2000 Centroid-based document classification: Analysis & experimental results *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery* 424–31

[16]Joachims T 1999 Transductive inference for text classification using support vector machines *Proceedings of the Sixteenth International Conference on Machine Learning* 200–9

[17]Dempster A P, Laird N M, Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm *Journal of the Royal Statistical Society* 39(B) 1–38

| Authors | |
|---|---|
|  | <p>Hoan Manh Dau, born in June, 1976, Vietnam</p> <p>Current position, grades: Ph.D student at computer field at School of Computer Science from 2011, Wuhan University of Technology, China. University studies: M.A degree at Informatics at Hue University of Sciences in Vietnam in 2004. Experience: lecturer of informatics for 12 years at University.</p> |
|  | <p>Ning Xu, China</p> <p>Current position, grades: Professor at the Computer Science Department of Wuhan University of Technology, senior member of China Computer Federation and the Chinese Institute of Electronics. University studies: Ph. D. degree in electronic science and technology at the University of Electronic Science and Technology of China in 2003. Scientific interest: Computer-aided design of VLSI circuits and systems, computer architectures, data mining, highly combinatorial optimization algorithms. Publications: Over 50 research papers. Experience: 5 research projects.</p> |

Multi-objective hub location problem in hub-and-spoke network

Ying Lu^{*}, Junping Xie

School of Automotive and Traffic Engineering, Jiangsu University, Xuefu Str. 301, 212013 Zhenjiang, China

Received 1 May 2014, www.tsi.lv

Abstract

Through observations from the construction of Chinese national emergency material reserve system, we introduce the multi-objective hub location problem. We provide a mathematical model for finding the optimal hub locations to minimize the total transportation cost and maximize the coverage of the hubs simultaneously in the whole network. Then, a procedure for solving this model is proposed. By using a numerical example, we discuss the efficiency of the tabu-search-based algorithm compared to the complete enumeration method and the impact of cost discount factor on the performance of hub-and-spoke network. The results show that the heuristic algorithm based on tabu search may be better than the complete enumeration research method for big size multi-objective hub location problem and as the cost discount factor is increased, the cost savings in the hub-and-spoke network compared to the direct connect network would decrease while the covering rate remains the same unless the cost discount factor is close to 1. Finally, we set future research directions on the multi-objective hub location problem.

Keywords: hub location problem, multi-objective programming, hub-and-spoke network, tabu search

1 Introduction

Our research is motivated by the practices of the construction of Chinese national emergency material reserve system. In order to achieve quick response to the urgent need of emergency materials in affected areas right after disasters, China began to build a national emergency material reserve system in 1998. Until now, there have been 18 central-level warehouses in the whole country, and each province has established a provincial-level warehouse, as well as 92 percent of cities and 60 percent of towns own town-level warehouses. In the next ten years, China plans to build more emergency material warehouses with different levels to form a perfect-served emergency material reserve system.

In China, once a natural or man-made disaster occurs, the local town-level and provincial-level warehouses would transport the emergency materials to the affected areas as quickly as possible. If the amount of the materials required exceeds the available stock in the local town-level and provincial-level warehouses, the nearest central-level warehouses would be involved in supplying the materials to the local lower-level warehouses. The other central-level warehouses would gather the materials, receive the donations and transfer them to the disaster-affected area if needed. To a certain extent, Chinese emergency material reserve system works like a hub-and-spoke network since a certain portion of materials flowing among provincial-level warehouses and town-level warehouses is transferred via the central-level warehouses which can be regarded as hubs.

The one basic difficulty for associating the hub-and-spoke network with emergency logistics is that in the

emergency logistics system the central-level warehouses need to cover as many emergency material flows as possible if the disaster occurs and, at the same time, the transportation cost should be acceptable when considering that the flows via central-level warehouses tend to generate many detours. As a result, the construction of emergency logistics system has multiple objectives, including the total transportation cost and the coverage of the warehouses. Consequently, the location of central-level warehouses in the emergency material reserve system is a *multi-objective hub location problem* (MOHLP) in the hub-and-spoke network.

Hub-and-spoke network is widely used in many transportation networks where hubs usually act as sorting, transshipment, and consolidation terminals. Instead of sending flows directly between all origin-destination pairs of nodes, hub facilities consolidate flows in order to take advantage of the economies of scale. Hub location problem (HLP) is an important issue arising in the design of hub-and-spoke network [1]. Much research has focused on presenting discrete hub median and related models to better capture behaviour observed in practice. However, to our knowledge there is very limited research that studies multi-objective multiple allocation hub location problem. Farahani et al. [2] reviewed all variants of HLP and discussed the mathematical models, solution methods, main specifications, and applications of HLPs, which including multi-objective HLP. Wang et al. [3] developed a fuzzy bi-objective programming model for emergency logistics systems by considering fuzzy demand of relief materials, timeliness and limited resources. The goal of their model is to minimize the total cost and the relief time of system. However, they focused

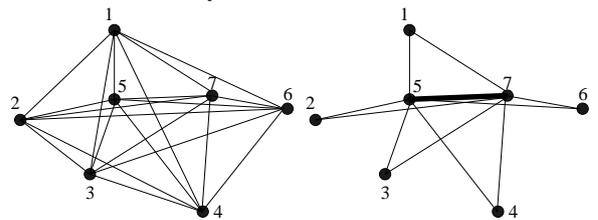
^{*} *Corresponding author* e-mail: luying@mail.ujs.edu.cn

on location-routing problem (LRP) instead of HLP. Mirzaei and Bashiri [4] used multiple objective approach for hub location to minimize total cost of the networks and minimize maximum travel time between nodes, whereas they only presented a brief comparison between their formulation and the other existing models in terms of the number of variables and constraints of the models. Tavakkoli-Moghaddam et al. [5] presented a multi-objective mathematical model for a capacitated single-allocation hub location problem. The multiple objectives are to minimize the total cost of the networks and minimize the maximum travel time between nodes. They solved the model by a multi-objective imperialist competitive algorithm (MOICA). However, the algorithm they proposed need more time to solve the models due to introducing many comparison metrics. Barzinpour et al. [6] developed a mathematical programming formulation for the bi-objective non-strict single allocation hub location problem. The objectives include minimizing the total system-wide cost and minimizing the maximum commodity routing distance between pairs of nodes in the network. They proposed a heuristic algorithm based on tabu search approach to solve the model. Geramianfar et al. [7] considered a multi-objective hub covering location problem under congestion. The first objective is to minimize total transportation cost and the second one is to minimize total waiting time for all hubs. They used simulated annealing (SA) to solve the model and compared the performance of the model against two other alternative methods, that is, particle sward optimization and NSGA-II. Tajbakhsh et al. [8] proposed a hub location model with both qualitative and quantitative objectives. To achieve better solutions, infeasible regions were also taken into account and a graded penalty term was used to penalize infeasible solutions. However, they did not present any numerical examples to prove the effectiveness of the algorithm they suggested. Costa et al. [9] proposed a multi-objective HLP in which the first objective is to minimize the total transportation cost, and the second one is to minimize the maximum service time of the hub nodes, whereas they did not provide any heuristic algorithm.

The purpose of this study is to develop a mathematical mode in the context of MOHLP in hub-and-spoke network and also present an effective algorithm to work out the best hub locations. The rest of this paper is organized as follows. In Section 2 we build a mathematical model for finding the optimal hub locations to minimize the total transportation time and to maximize the coverage of the hub(s). Section 3 proposes a heuristic algorithm to find the optimal solutions of the model. The results of a numerical example are presented in Section 4. Section 5 discusses the impact of the cost discount factor on the performance of the hub-and-spoke network and provides some managerial insights. Finally, Section 6 presents summary comments and discusses promising areas for future research.

2 Model formulation

We consider a hub-and-spoke network consisting of several nodes. Traditionally, each origin-destination pair of nodes can be connected directly, which is so-called direct connect network (see Figure 1a). However, in the context of a hub-and-spoke network, each origin-destination flow must be routed via the hubs (see Figure 1b). The main problem in this network is to decide on the location of the hubs and the allocation of the non-hub nodes to these hubs. Our objectives are to minimize the total transportation cost and maximize the coverage of the hubs simultaneously.



(a) Direct connect network (b) Hub-and-spoke network

FIGURE 1 Direct connect network and hub-and-spoke network

2.1 NOTATIONS

Consider a complete graph $G(V, A)$ with node set $V = 1, 2, \dots, N$ where nodes correspond to origins and destinations as well as potential hub locations. The notations we used are as follows.

Index and parameters:

A : set of all arcs;

N : the number of nodes;

p : the number of hubs;

i, j : index for nodes ($i, j = 1, \dots, N$);

k, m : index for potential hub locations;

c_{ij} : standard cost per unit from origin i to j ;

h_{ij} : the amount of flows between nodes i and j ;

J : set of origin-destination pairs of nodes,

$$J = (i, j) | h_{ij} > 0, i, j \in V ;$$

C_{ij}^{km} : the transportation cost of a unit of flow from origin i to destination j via hubs k and m on path $i - k - m - j$. Note that C_{ij}^{km} is composed of three parts: a cost of communication from the source node to its respective hub, a cost of communication to a sink node from its respective hub, and the cost of communication between the two hubs. Consequently, we have

$$C_{ij}^{km} = c_{ik} + \alpha c_{km} + c_{nj}, \tag{1}$$

where α is the cost discount factor for the inter-hub transportation due to heavy traffic [10].

V_{ij}^{km} : the binary variable. V_{ij}^{km} is equal to 1 if hubs k and m cover origin-destination pair (i, j) , and zero otherwise. Let β_{ij} be the maximum cost for origin-destination pair (i, j) . The interpretation of coverage is that origin-destination pair (i, j) would be covered by hubs k and m if the transportation cost from i to j via k and m does not exceed a specified value, that is,

$$V_{ij}^{km} = \begin{cases} 1, & \text{if } C_{ij}^{km} \leq \beta_{ij} \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

Decision variables:

X_{ij}^{km} : the binary variable. X_{ij}^{km} is equal to 1 if shipments from origin i to destination j are assigned to hubs k and m on path $i-k-m-j$, and zero otherwise.

Y_k : the binary variable. Y_k is equal to 1 if node k is chosen to be a hub, and zero otherwise.

2.2 ASSUMPTIONS

The assumptions of our model are as follows:

(1) The transportation between any two nodes in the network is performed only via hubs.

(2) The network has enough arcs with sufficient capacity to enable all the flows generated at the origin nodes to reach all the destination nodes regardless of the amount of flows received by the hubs, that is, we do not take capacity constraints into account.

(3) There are two objectives under consideration in this study. The first one is to minimize the total cost in the hub-and-spoke network, and the second one is to maximize the coverage of the hubs. These two objectives are of roughly comparable importance. As a result, the MOHLP is a non-preemptive objective programming.

2.3 MATHEMATICAL MODEL

In our model, the decision involves choosing X_{ij}^{km} , Y_k to minimize the total transportation cost and maximize the coverage of the hubs. The mathematical formulation for this MOHLP is as follows:

$$\min z_1 = \sum_i \sum_j \sum_k \sum_m h_{ij} X_{ij}^{km} C_{ij}^{km}, \quad (3)$$

$$\max z_2 = \sum_i \sum_j \sum_k \sum_m h_{ij} X_{ij}^{km} V_{ij}^{km} \quad (4)$$

subject to

$$\sum_k \sum_m X_{ij}^{km} = 1, \quad \forall (i, j) \in J, \quad (5)$$

$$X_{ij}^{km} \leq Y_k, \quad \forall i, j, k, m, \quad (6)$$

$$X_{ij}^{km} \leq Y_m, \quad \forall i, j, k, m, \quad (7)$$

$$\sum_k Y_k = p, \quad (8)$$

$$X_{ij}^{km} \in 0,1, \quad \forall i, j, k, m, \quad (9)$$

$$Y_k \in 0,1, \quad \forall k. \quad (10)$$

In the above formulation, z_1 is the total transportation cost in the network and z_2 is the coverage of the hubs. Constraint (5) ensures that each origin-destination flow is sent via some hub pair (possibly a single hub as in X_{ij}^{kk}). Constraints (6) and (7) ensure that nonhub nodes can only be allocated to the hubs which work as transfer terminals. Constraint (8) requires exactly p hubs are selected. Constraints (9) and (10) define the decision variables to be binary.

3 Solution procedure for the model

3.1 SOLUTION PROCEDURE FOR SHLPOMIC

Before giving the algorithms for solving MOHLP, we define a single hub location problem of minimizing cost (SHLPOMIC) as follows:

$$\min z_1 = \sum_i \sum_j \sum_k \sum_m h_{ij} X_{ij}^{km} C_{ij}^{km} \quad \text{subject to constraints (5)-(10).}$$

SHLPOMIC is NP-hard. A heuristic algorithm based on tabu search is widely used to solve this kind of problem [11, 12]. The elements of tabu search include initial solution, neighbourhood structure, tabu lists, and so on. Remember that V represents the set of all nodes. Let T be the set of hubs, $V-T$ be the set of nonhub nodes, $N(T)$ be the set of neighbouring solutions for T . For simplicity, we adopt a 'single location exchange' rule to generate a neighbourhood solution, denoted by T^i , for a given T , that is, replacing exactly one hub in T with one nonhub node in $V-T$. During the process of 'single location exchange', the node leaving T is denoted by V^i and the node leaving $V-T$ and entering T is denoted by W^i . Under this rule, there are $p(n-p)$ possible neighbouring solutions for a given T , i.e., $N(T) = T^i, i=1,2,\dots,p(n-p)$. The node leaving T and generating a new T^i at the current iteration is recorded in the tabu list, named as tabu status. In order to forbid the reversal of this replacement unless the move leads to a solution better than the best found so far (this is the so-called aspiration criterion), the tabu status at the

current iteration cannot be selected to enter T again in a number (tabu list size) of future iterations. For instance, setting $tabu_tag(i) = t$ means that node i acts as a tabu status and cannot be an element of T in the next t iterations. The value of the objective function is expressed in terms of just the current T as

$$Z_1(T) = \sum_i \sum_j h_{ij} \min_{k,m \in T} C_{ij}^{km} \quad (11)$$

In order to improve the efficiency of tabu search, we would obtain the initial solution where the search starts by employing a local search method. The solution procedure can be summarized as follows.

Step 1. Choose arbitrarily p nodes as an initial solution T . Designate this initial solution as the optimal solution, that is, set $T^0 = T$.

Step 2. Generate $N(T)$ for current T and calculate $Z_1(T^i)$ via (11) for all $T^i \in N(T)$. Find the smallest one and let T^* denote corresponding T^i .

Step 3. If $Z_1(T^*) < Z_1(T^0)$, set $T^0 = T^*$, $T = T^*$, and then go to Step 2. Otherwise, go to Step 4 with the initial solution T .

Step 4. Initialize the tabu lists and the number of iterations, that is, set a value for the maximal number of iterations denoted by max_itm and set $t = 0$, $tabu_tag(i) = 0$ for all $i \in V$. Update the current optimal solution T^0 and set $T^0 = T$.

Step 5. Generate $N(T)$ for current T . Calculate $Z_1(T^i)$ via (11) for all $T^i \in N(T)$.

Step 6. Choose index l such that $Z_1(T^l) = \min_{i=1, \dots, p(n-p)} Z_1(T^i)$. If $tabu_tag(W^l) = 0$ or $Z_1(T^l) < Z_1(T^0)$ (the aspiration criterion), then set $T = T^l$, record V^l in the tabu list, i.e., set $tabu_tag(V^l)$ equal to a uniform random number over the interval $[\sqrt{n}, 2\sqrt{n}]$, and go to Step 7. Otherwise, delete T^l from $N(T)$ (that is, set $N(T) = N(T) - T^l$), and return to Step 6.

Step 7. Set $t = t + 1$. If $Z_1(T) < Z_1(T^0)$, then update the current optimal solution (that is, set $T^0 = T$).

Step 8. If $t < max_itm$, then update the tabu list, i.e., set $tabu_tag(i) = tabu_tag(i) - 1$ for all i such that $tabu_tag(i) > 0$, and return to Step 5. Otherwise, the best solution of SHLPOMIC is T^0 .

3.2 SOLUTION PROCEDURE FOR SHLPOMAC

Similarly, we define a single hub location problem of maximizing covering (SHLPOMAC) as follows:

$$\max z_2 = \sum_i \sum_j \sum_k \sum_m h_{ij} X_{ij}^{km} V_{ij}^{km} \quad \text{subject to constraints} \quad (5)-(10).$$

The value of the objective function is expressed in terms of just the current T as

$$Z_2(T) = \sum_i \sum_j h_{ij} \min \left(1, \sum_{k \in T} \sum_{m \in T} V_{ij}^{km} \right). \quad (12)$$

The solution procedure can be summarized as follows.

Step 1. Set the initial solution $T = \varphi$, and also set $k = 0$.

Step 2. Choose a node l^* from $V - T$ by which the amount of the new origin-destination flows covered is maximal, i.e., $V_{l^*} = \max_{l \in N-T} [Z_2(T+l) - Z_2(T)]$. Set $T = T + l$ and $k = k + 1$.

Step 3. If $k < p$, then go to Step 2. Otherwise, go to Step 4 with the initial solution T .

Step 4. Initialize the tabu lists and the number of iterations, that is, set a value for the maximal number of iterations denoted by max_itm and set $t = 0$, $tabu_tag(i) = 0$ for all $i \in V$. Update the current optimal solution T^0 and set $T^0 = T$.

Step 5. Generate $N(T)$ for current T . Calculate $Z_2(T^i)$ via (12) for all $T^i \in N(T)$.

Step 6. Choose index l such that $Z_2(T^l) = \max_{i=1, \dots, p(n-p)} Z_2(T^i)$. If $tabu_tag(W^l) = 0$ or $Z_2(T^l) > Z_2(T^0)$, then set $T = T^l$, record V^l in the tabu list, i.e., set $tabu_tag(V^l)$ equal to a uniform random number over the interval $[\sqrt{n}/2, \sqrt{n}]$, and go to Step 7. Otherwise, delete T^l from $N(T)$ (that is, set $N(T) = N(T) - T^l$), and return to Step 6.

Step 7. Set $t = t + 1$. If $Z_2(T) > Z_2(T^0)$, then update the current optimal solution (that is, set $T^0 = T$).

Step 8. If $t < max_itm$, then update the tabu list, that is, set $tabu_tag(i) = tabu_tag(i) - 1$ for all i such that $tabu_tag(i) > 0$, and return to Step 5. Otherwise, the best solution of SHLPOMAC is T^0 .

3.3 SOLUTION PROCEDURE FOR MOHLP

MOHLP will be harder to solve to optimality with the addition of multiple objectives. Thus, there will be a need to develop efficient heuristic algorithms for it.

One of the most common approaches to multi-objective optimization is the goal programming method [13]. Let $X = X_{ij}^{km}, Y_k$ be the vector of decision variables and F represent the feasible set of decision

vectors for which all the constraints are satisfied (that is, $X \in F$). Also let $z_l(X)$ denote the l^{th} objective function, X_l^* represent the optimum of the l^{th} single-objective function subject to the constraints in the multi-objective problem, z_l^* be the corresponding objective function values, and b_l be the target value for the l^{th} objective function $z_l(X)$. In the absence of any other information, we can set $b_l = z_l^*$. Then, according to the theory of goal programming method, we should minimize the total deviation from the goals $\sum_l |d_l|$, where d_l is the deviation from the goal b_l for the l^{th} objective. To model the absolute values, d_l is split into positive and negative parts such that $d_l = d_l^+ - d_l^-$, with $d_l^+ \geq 0$, $d_l^- \geq 0$, $d_l^+ d_l^- \geq 0$. We have $|d_l| = d_l^+ + d_l^-$. d_l^+ and d_l^- represent underachievement and overachievement, respectively, where achievement implies that a goal has been reached. In the case of our model, we have the following parameters: $l \in 1, 2$, F is determined by constraints (5) ~ (10), $z_1(X)$ and $z_2(X)$ are given by (3) and (4) (note that $X = (X_{ij}^{km}, Y_k)$), z_1^* and z_2^* are the optimal objective function values for SHLPOMIC and SHLPOMAC, respectively. Consequently, the optimization problem, named as Problem (MOP), is formulated as follows:

Problem (MOP):

$$\min z' = \sum_{l=1}^2 P_l \lambda_l d_l^+ + P_l \lambda_l d_l^- \tag{13}$$

subject to

$$z_l(X) - d_l^+ + d_l^- = z_l^*, l=1,2, \tag{14}$$

$$X \in F, \tag{15}$$

$$d_l^+, d_l^- \geq 0, \tag{16}$$

where P_l is the weighting coefficients and λ_l is the normalization constant for the l^{th} objective. According to the assumption that all the objectives have the same priority level, we set the weighting coefficients $P_l = 1$ for $l = 1, 2$.

A major difficulty of adopting the goal programming method lies in the incommensurability, which occurs when deviational variables measured in different units are summed up directly. This simple summation will cause an unintentional bias towards the objectives with a larger magnitude, which may lead to erroneous or misleading results. To overcome the incommensurability, we suggest using the percentage normalization method where the

normalization constant is hundred divided by the target value: $\lambda_l = 100/b_l$ for $l = 1, 2$ [14]. This ensures that all deviations are measured on a percentage scale.

However, (14) is a nonlinear equality constraint, which makes it not easy to find the optimal solution with larger problems. Here, we propose applying the tabu search method again to solve (MOP) approximately. The solution procedure can be summarized as follows:

Step 1. Solve SHLPOMIC and SHLPOMAC by means of the algorithms mentioned earlier in this section and denote the best achieved objective function values by z_1^* and z_2^* , respectively. Set $\lambda_l = 100/z_l^*$ for $l = 1, 2$.

Step 2. Choose arbitrarily p nodes as an initial solution T . Designate this initial solution as the optimal solution, that is, set $T^0 = T$.

Step 3. Calculate $Z_l(T^0)$ via (11) and (12) for each $l = 1, 2$, respectively. Substitute $Z_l(T^0)$ for $z_l(X)$ in (MOP). Note that once $Z_l(T^0)$ is given, (MOP) becomes a linear programming problem. Solve (MOP) and denote the optimal objective function value by $z'(T_0)$.

Step 4. Generate $N(T)$ for current T . For all $T^i \in N(T)$, Calculate $Z_1(T^i)$ and $Z_2(T^i)$ via (11) and (12), Substitute $Z_1(T^i)$ and $Z_2(T^i)$ for $z_1(X)$ and $z_2(X)$ in (MOP). Solve (MOP) to obtain the objective function value $z'(T^i)$. Find the smallest one and let T^* denote corresponding T^i .

Step 5. If $z'(T^*) < z'(T^0)$, set $T^0 = T^*$, $T = T^*$, and then go to Step 4. Otherwise, go to Step 6 with the initial solution T .

Step 6. Initialize the tabu lists and the number of iterations, that is, set a value for the maximal number of iterations denoted by \max_itm and set $t = 0$, $\text{tabu_tag}(i) = 0$ for all $i \in V$. Update the current optimal solution T^0 and set $T^0 = T$.

Step 7. Generate $N(T)$ for current T . For all $T^i \in N(T)$, Calculate $Z_1(T^i)$ and $Z_2(T^i)$ via (11) and (12), Substitute $Z_1(T^i)$ and $Z_2(T^i)$ for $z_1(X)$ and $z_2(X)$ in (MOP). Solve (MOP) to obtain the objective function value $z'(T^i)$.

Step 8. Choose index l such that $z'(T^l) = \min_{i=1, \dots, p(n-p)} z'(T^i)$. If $\text{tabu_tag}(W^l) = 0$ or $z'(T^l) < z'(T^0)$ (the aspiration criterion), then set $T = T^l$, record V^l in the tabu list, i.e., set $\text{tabu_tag}(V^l)$ equal to a uniform random number over the interval $[\sqrt{n}, 2\sqrt{n}]$, and go to Step 9. Otherwise, delete T^l from $N(T)$ (that is, set $N(T) = N(T) - T^l$), and return to Step 8.

Step 9. Set $t = t + 1$. If $z'(T) < z'(T^0)$, then update the current optimal solution (that is, set $T^0 = T$).

Step 10. If $t < \max_itm$, then update the tabu list, that is, set $tabu_tag(i) = tabu_tag(i) - 1$ for all i such that $tabu_tag(i) > 0$, and go to Step 7. Otherwise, the best solution of the MOHLP is T^0 .

4 Numerical examples

In this section we adopt the AP data set that is used in various hub location studies and is available from OR Library (<http://mscmga.ms.ic.ac.uk/info.html>). AP data set consists of 200 nodes. Without loss of generality, we only consider the first 7 nodes in the data set corresponding to 7 cities (that is, $n = 7$). All the 7 cities are the candidates of hub locations. Set the transportation cost per kilometre per unit of flow to be 3. Consequently, c_{ij} is taken to be equal to 3 times the distance between i and j in AP set, as shown in Table 1. The flows in and out of these cities are shown in Table 2. We set the flows within the same node to be zero, i.e., $h_{ii} = 0$, to avoid the unreasonable detour between nodes and hubs. We also set $\max_itm = 100$, $\alpha = 0.4$, $p = 2$, and $\beta_{ij} = 1.2c_{ij}$.

TABLE 1 Standard cost per unit between pairs of cities c_{ij} in thousand dollars

| Origin i | Destination j | | | | | | |
|------------|-----------------|-------|------|-------|-------|------|-------|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 2 | 0 | 5.21 | 0.43 | 5.35 | 10.03 | 1.73 | 5.62 |
| 3 | 5.21 | 0 | 5.19 | 0.87 | 15.16 | 5.35 | 1.74 |
| 4 | 0.43 | 5.19 | 0 | 5.26 | 9.98 | 1.30 | 5.47 |
| 5 | 5.35 | 0.87 | 5.26 | 0 | 15.13 | 5.21 | 0.87 |
| 6 | 10.03 | 15.16 | 9.98 | 15.13 | 0 | 9.95 | 15.16 |
| 7 | 1.73 | 5.35 | 1.30 | 5.21 | 9.95 | 0 | 5.21 |
| 8 | 5.62 | 1.74 | 5.47 | 0.87 | 15.16 | 5.21 | 0 |

TABLE 2 The amount of flows between pairs of cities h_{ij}

| Origin i | Destination j | | | | | | |
|------------|-----------------|------|------|------|------|------|------|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 2 | 0 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| 3 | 0.01 | 0 | 0.20 | 0.01 | 0.05 | 0.26 | 0.16 |
| 4 | 0.01 | 0.09 | 0 | 0.01 | 0.05 | 0.26 | 0.16 |
| 5 | 0.01 | 0.01 | 0.01 | 0 | 0.01 | 0.02 | 0.01 |
| 6 | 0.01 | 0.03 | 0.07 | 0.01 | 0 | 0.10 | 0.06 |
| 7 | 0.01 | 0.14 | 0.31 | 0.01 | 0.07 | 0 | 0.25 |
| 8 | 0.01 | 0.05 | 0.12 | 0.01 | 0.03 | 0.15 | 0 |

We first program the algorithm described in Section 3 with MATLAB 7.1 for SHLPOMIC and SHLPOMAC. Under SHLPOMIC, the nodes 5 and 7 are designated as hubs and the corresponding optimal total transportation cost in this hub-and-spoke network z_1^* is 12.185. Under SHLPOMAC, nodes 4 and 5 are designated as hubs and the amount of flows that are covered by the hubs is 2.8535.

After obtaining the z_1^* and z_2^* , we then program the algorithm for MOHLP described in Section 3 with MATLAB 7.1. The results show that nodes 5 and 7 are designated as hubs. The corresponding objective function value z' is 0.701. In this situation, the total transportation cost z_1 is 12.185 and the coverage of hubs z_2 is 2.8335. The route that starting from origin i to destination j via two hubs are given below in Table 3. Here the symbol | indicates the pair of hubs. For instance, according to Table 3, the flow from origin 3 to destination 6 is sent on path 3-5-7-6.

TABLE 3 The routes for origin-destination flows

| Origin i | Destination j | | | | | | |
|------------|-----------------|-----|-----|-----|-----|-----|-----|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 2 | - | 7 5 | 7 7 | 7 5 | 7 7 | 7 7 | 7 5 |
| 3 | 5 7 | - | 5 7 | 5 5 | 5 7 | 5 7 | 5 5 |
| 4 | 7 7 | 7 5 | - | 7 5 | 7 7 | 7 7 | 7 5 |
| 5 | 5 7 | 5 5 | 5 7 | - | 5 7 | 5 7 | 5 5 |
| 6 | 7 7 | 7 5 | 7 7 | 7 5 | - | 7 7 | 7 5 |
| 7 | 7 7 | 7 5 | 7 7 | 7 5 | 7 7 | - | 7 5 |
| 8 | 5 7 | 5 5 | 5 7 | 5 5 | 5 7 | 5 7 | - |

By multiplying each of the individual terms in Table 1 by the corresponding term in Table 2, and summing up these individual products, we also calculate the total transportation cost in the situation where the flows of materials are sent directly through the arcs linking origin-destination nodes, denoted by z^d , as follows.

$$z^d = \sum_i \sum_j c_{ij} h_{ij}, \tag{17}$$

let $\Delta z' = z^d - z_1$ be the cost savings in the hub-and-spoke network compared to direct connect network. We have $z^d = 15.07$ and $\Delta z' = 2.885$. This result shows that compared to direct connect network, the hub-and-spoke network can save 2.885 thousand dollars.

Meanwhile, we also calculate the covering rate, denoted by β , as follows

$$\beta = z_2 / \sum_i \sum_j h_{ij}. \tag{18}$$

The resulting value of β is 99%, which demonstrates that the majority of flows is covered by the hubs.

In order to observe the performance of the heuristic algorithm on a different data set, we generate a set of test instances with different parameters. Let t_{tabu} and t_{enum} denote the respective CPU time requirements of the heuristic algorithm presented in Section 3 and the complete enumeration method. Table 4 shows the CPU times in different methods.

TABLE 4 Performance of the heuristic algorithm on the network

| p | Hubs | z' | $\Delta z'$ | β | t_{tabu} | t_{enum} |
|-----|-----------|-------|-------------|---------|------------|------------|
| 2 | 5,7 | 0.701 | 2.885 | 99.3% | 20.014 | 7.83 |
| 3 | 6,5,7 | 0.701 | 5.885 | 99.2% | 21.221 | 9.865 |
| 4 | 6,7,4,5 | 0 | 7.412 | 1 | 34.563 | 10.074 |
| 5 | 7,3,4,8,6 | 0 | 8.901 | 1 | 24.233 | 11.072 |

From Table 4, we can see that, generally, the CPU time requirement of the heuristic algorithm for the MOHLP increases with the increase of the candidate number of hub locations. However, this outcome seriously depends on the selected initial solution.

It can also be seen that the solutions provided by the algorithm we proposed and the complete enumeration research method are the same as well as the complete enumeration research method seems more effective than the algorithm presented in Section 3 when the number of the nodes is small. However, due to the fact that the number of hub arc combinations increases faster than linearly, we recommend using the heuristic algorithm based on tabu search instead of the complete enumeration research method for big problems.

5 The impact of cost discount factor on the performance of hub-and-spoke network

Intuitively, the cost discount factor may have the effect on the performance of hub-and-spoke network. Keeping the other parameters presented in Section 4 unchanged, for different values of α , we calculate the corresponding $\Delta z'$, β , z_1^* , z_1 , z_2^* , z_2 . Figure 2 and Figure 3 display $\Delta z'$ and β with different values of α , respectively. Figure 4 shows the corresponding z_1^* and z_1 . Figure 5 shows the corresponding z_2^* and z_2 .

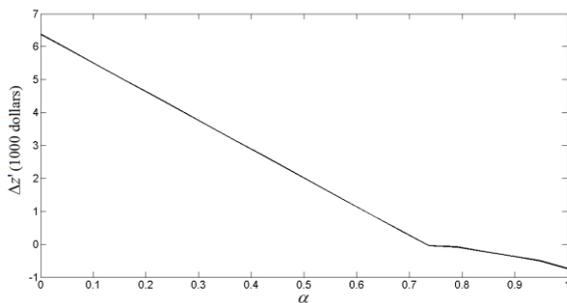


FIGURE 2 The relation between the cost savings in the hub-and-spoke network and the cost discount factor

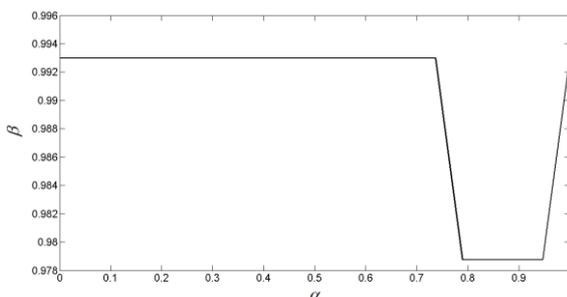


FIGURE 3 The relation between the covering rate and the cost discount factor

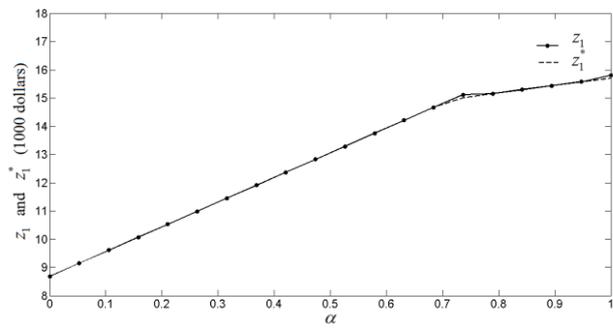


FIGURE 4 The total transportation costs in MOHLP and SHLPOMIC with changes in the transportation cost discount factor

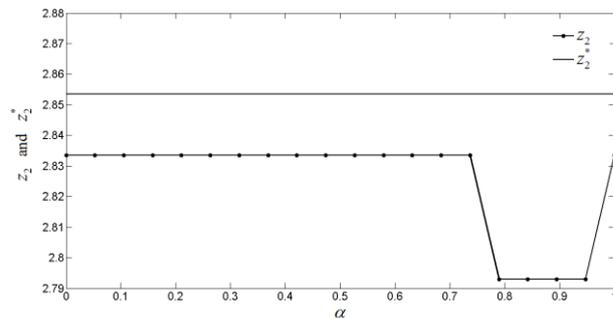


FIGURE 5 The total transportation costs in MOHLP and SHLPOMIC with changes in the transportation cost discount factor

From Figures 2 and 3, it can be seen that, as the cost discount factor is increased, the cost savings in the hub-and-spoke network compared to direct connect network would decrease, which means that the cost discount factor has negative effects on the performance of hub-and-spoke network. Meanwhile, the cost discount factor seems likely to have no significant effect on the covering rate because the latter depends much more upon the maximum cost for origin-destination pair (that is β_{ij}) than upon the cost discount factor. Note that when the cost discount factor is close to 1, the covering rate would sharply decrease to a certain value and then quickly increase to the value which is the same as that with a small α . The reason is that in a situation where α is large, there is no improvement in the total transportation cost in the hub-and-spoke network compared to the direct connect network so that few origin-destination flows would be sent via hubs.

From Figures 4 and 5, we can see that in the context of our example, generally, the goal of minimizing the total transportation cost can be achieved in MOHLP, whereas there is a difference between the coverage of the hubs achieved in MOHLP and that in SHLPOMAC. Note that when the cost discount lies in the interval between 0.7 and 0.8, the total transportation cost in MOHLP would be larger than that in SHLPOMIC. This result indicates that when there are multiple objectives in the hub-and-spoke network, one objective may be achieved better than the others. If we want to obtain the cost objective, the cost discount factor should be taken on a small value near zero.

6 Conclusion

In this paper we introduce a multi-objective hub location problem in the hub-and-spoke network. We provide a mathematical model for minimizing the total transportation cost and maximizing the coverage of the hubs. We then propose a heuristic algorithm based on tabu search for finding the optimal hub locations. Additionally, we compare our algorithm with the complete enumeration method and investigate the impact of cost discount factor on the performance of hub-and-spoke network. The results of a numerical example show that our heuristic algorithm based on tabu search may be better than the complete enumeration research method for big size MOHLP, and as the cost discount factor is increased, the cost savings in the hub-and-spoke network compared to direct connect network would decrease while the covering rate remains the same unless the cost discount factor is close to 1. Another interesting finding is that when there are multiple objective in the hub-and-spoke network, one objective may be achieved better than the others.

References

- [1] Campbell J F 1994 Integer programming formulations of discrete hub location problems *European Journal of Operational Research* **72**(2) 387–405
- [2] Farahani R Z, Hekmatfar M, Arabani A B, Nikbakhsh E 2013 Hub location problems: A review of models, classification, solution techniques, and applications *Computers & Industrial Engineering* **64**(4) 1096–109
- [3] Wang S, Ma Z, Zhuang B 2014 Fuzzy location-routing problem for emergency logistics systems *Computer Modelling & New Technologies* **18**(2) 265-73
- [4] Mirzaei M, Bashiri M 2010 Multiple objective multiple allocation hub location problem *Proc. Int. Conf. Comput. Ind. Eng.: Soft Comput. Tech. Adv. Manuf. Serv. Syst., CIE40* (Awaji) IEEE Computer Society: Piscataway, NJ, United States pp 1-4
- [5] Tavakkoli-Moghaddam R, Gholipour-Kanani Y, Shahramifar M 2013 A multi-objective imperialist competitive algorithm for a capacitated single-allocation hub location problem *International Journal of Engineering* **26**(6) 605-20
- [6] Barzinpour F, Ghaffari-Nasa N, Saboury A 2011 Bi-objective non-strict single allocation hub location problem: mathematical programming model and a solution heuristic *Proc. Int. Conf. Comput. Ind. Eng., CIE41 (Los Angeles)* pp 86-91
- [7] Geramianfar R, Pakzad M, Golhashem H, Moghaddam R T 2013 A multi-objective hub covering location problem under congestion using simulated annealing algorithm *Uncertain Supply Chain Management* **1**(3) 153-64
- [8] Tajbakhsh A, Haleh H, Razmi J 2013 A multi-objective model to single-allocation ordered hub location problems by genetic algorithm *International Journal of Academic Research in Business and Social Sciences* **3**(2) 374-9
- [9] Costa M G, Captivo M E, Climaco J 2008 Capacitated single allocation hub location problem – A bi-criteria approach *Computers and Operations Research* **35**(11) 3671-95
- [10] Limbourg S, Jourquin B 2009 Optimal rail-road container terminal locations on the European network *Transportation Research Part E* **45**(4) 551-63
- [11] Skorin-Kapov D, Skorin-Kapov J 1994 On tabu search for the location of interacting hub facilities *European Journal of Operational Research* **73**(3) 502–9
- [12] Saboury A, Ghaffari-Nasab N, Barzinpour F, Jabalameli M S 2013 Applying two efficient hybrid heuristics for hub location problem with fully interconnected backbone and access networks *Computers & Operations Research* **40**(10) 2493–507
- [13] Marler R T, Arora J S 2004 Survey of multi-objective optimization methods for engineering *Structural and Multidisciplinary Optimization* **26**(6) 369-95
- [14] Zhao J 2012 Multi-objective location-routing problem in emergency logistics as time varies *Journal of Highway and Transportation Research and Development* **29**(4) 137-42 (in Chinese)

Acknowledgments

This study is supported by National Natural Science Foundation of China (51208232), Science Foundation of Jiangsu University (09JDG078).

Authors



Ying Lu, born in January, 1981, Zhenjiang city, Jiangsu province, P. R. China

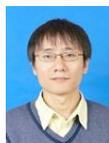
Current position, grades: The lecturer in School of Automotive and Traffic Engineering, Jiangsu University, doctoral degree.

University studies: Ph.D. degree was earned in major of management science and engineering, Sun Yat-sen university in 2009.

Scientific interest: production planning, supply chain management.

Publications: He has published 1 book and more than 10 papers in various journals. Four of papers were included in the international Engineering Index (EI).

Experience: He received his Bachelor and Master degrees in Automotive Engineering from Jiangsu University in 2002 and in 2005, respectively. In 2009 he received Ph.D. degree in Management Science and Engineering from Sun Yat-sen University. He serves as a lecturer in School of Automotive and Traffic Engineering, Jiangsu University.



Junping Xie, born in November, 1980, Zhenjiang city, Jiangsu province, P. R. China

Current position, grades: The lecturer in School of Automotive and Traffic Engineering, Jiangsu University, doctoral degree.

University studies: Ph.D. degree was earned in major of Communication and Transportation Programming, Southeast University in 2010.

Scientific interest: transportation system design, logistics management, operations research

Publications: He has published 1 book and more than 8 papers in various journals. Two of papers were included in the international Engineering Index (EI).

Experience: He received his Bachelor degree in Civil Engineering from Southeast University in 2003. In 2010, he received Ph.D. degree in Communication and Transportation Programming from Southeast University. He serves as a lecturer in School of Automotive and Traffic Engineering, Jiangsu University.

Analysis of necessary investments in the production and warranty service of innovative products considering the necessity of their backup

Ilana Ter-Saakova^{1*}, Nataly Podolyakina²

¹*Baltic International Academy, Lomonosova 1, Riga, LV-1019, Latvia*

²*Transport and Telecommunication Institute, Riga, Latvia, Lomonosova 1, Riga, LV-1019, Latvia*

Received 26 April 2017, www.tsi.lv

Abstract

Redundancy is one of the commonly used methods to improve the reliability of industrial products and is used in various designs. Another way to increase the reliability is to use more reliable components during the production. This work provides a feasibility study of the redundancy during the manufacturing as well as a comparative analysis of the conditions under which one or another methods are chosen to improve the reliability of a product as a function of its value.

Key words: costs structure, warranty service, reliability level, probability of no-failure operation of products

1 Introduction

Nowadays there is tough competition in the market, competition for customers, for cost-cutting of the production process and issues of reliability of the made products are as relevant as ever. It is no secret that there is an enormous amount of products, the reliability of which does not influence the person's safety and their premature failure only harms the image of its producer. But there are some types of products that must not be unreliable. The issues of reliability require special attention in the production of innovative component parts, assemblies and mechanisms of vehicles. Their sudden failure can result in catastrophic consequences and not only economic losses or loss of the business reputation of the producer but also the unmitigatable death of people. At the same time, the increase in the reliability of products entails a cost escalation for the producer. Thus, it is necessary to solve the contradiction between a desire to reduce the price of the production process of a product and maintain a certain level of reliability.

2 Essence and key reliability indices

If we speak about the fact that the made products must offer a certain level of reliability, it is necessary to define the essence of the concept of reliability and state its characteristic indices.

Reliability is a system or component property which performs the set functions, providing fail-free operation, durability and serviceability [1, 3, 4].

Depending on the conditions of the current task, one and the same item can be named a system or component. Under the term system we understand an aggregate of the jointly operating components (spare parts, associated parts, devices), performance of the set functions.

In order to evaluate the reliability properties (fail-free operation, serviceability, storageability, durability), it is necessary to introduce quantitative reliability indices.

Quantitative reliability indices of nonrepairable items: 1). The probability of no-failure operation $P(t)$. Under the probability of no-failure operation of an item we understand that within the limits of the set operating time the item will not fail. Mathematically this index can be determined as the probability that time T of no-failure operation is a random variable and will exceed the set t [2]:

$$P(t) = \frac{N_0 - n(t)}{N_0}, \quad (1)$$

where N_0 – total number of products;

$n(t)$ – number of failed products till the beginning of the interval of time under investigation.

2). Failure probability $Q(t)$. Here failure probability means the probability that failure of an item will happen during a period of time not exceeding the set value t [2]:

$$Q(t) = 1 - P(t). \quad (2)$$

3). Rate of failures $\lambda(t)$ – rate of failure of a nonrepairable product in a unit of time after the current moment upon the condition that the failure could not happen till that moment [2]: The probability of no-failure operation will be determined in the following way:

* *Corresponding author* e-mail: ilana@alida.lv

$$P(t) = e^{-\lambda t} \tag{3}$$

4). Mean time to failure T_0 – mathematical expectation of the running time of a product till the first failure [2]:

$$T_0 = M[T] = \frac{\sum_{i=0}^N \tau_i}{N_0} \tag{4}$$

On the whole, the task for the calculation of reliability: determination of fail-free operation indices of a system, consisting of non-repairable items, according to data on component reliability and interactions between them.

In turn, the purpose of the calculation of reliability is a choice of one or another structural solution, as well as the determination of possibility and, mainly, the economic efficiency of backing up.

3 Costs structure for production and post-sale maintenance of innovative products

When evaluating necessary investment amounts, first of all, it is necessary to determine the cost value of the enterprise when planning to implement the investment project. If such an investment project represents the industrial production of an innovative product, we can state that the total costs of the enterprise in manufacturing the innovative product on the one hand, include the costs of its development, production, sale and post-sale maintenance, i.e. maintenance of use, but on the other hand, these costs depend on the degrees of reliability of the innovative product characterised by the probability of no-failure operation $P(t)$.

Formally combined costs of the whole above-stated process can be presented as follows:

$$C_{\Sigma} = C_{prod} + C_{use} = f[P(t)] \tag{5}$$

where C_{prod} – costs of development and production of the innovative product; C_{use} – costs of maintenance of use of the innovative product.

Obviously, the higher the probability of no-failure operation of the product, the higher its price. It is possible to accept the following expression as a model of this dependence:

$$C_{prod} = a \cdot P^{\alpha} \tag{6}$$

where a and α are corresponding factors, determining C_{prod} , $P = P(tr)$ – probability of no-failure operation during the warranty period tr .

The model of dependence for operating costs can be presented as follows:

$$C_{use} = b \cdot (1/P)^{\beta} \tag{7}$$

Quite often the made products can be used in devices requiring enhanced reliability due to a direct influence on

the safety of vital activity security or high importance of performed tasks. As an example it is possible to mention sea, air, rail and road transport, where switching of the systems responsible for vital activity security to backup ones should be carried out with a probability close to 1.

Therefore backing up is presently one of the most in-use methods for the increase in no-failure operation of innovative products, especially for nonrepairable (impossible to repair) devices.

But in this case of backing up a problem regarding the contradictions between mass-dimensional and cost limitations often appears, as well as a question about the economic efficiency of such a decision, especially with the creation of an innovative product. The practical possibility of backing up at the level of components, associated parts and devices on the whole meets the challenge of the increase in fail-free operation of a device thanks to backing up of the weak component of the basic kit.

It is generally known, that the failure of a product happens due to the breakdown of one or several components of this product. Thus, as experience shows, in the vast majority of cases other components work smoothly for quite a long period. In this context, duplication of a product on the whole means, that for the sake of one or several failed components we include one more of the same product with the same component with a high probability of failure. Therefore, the larger the product is, the less confidence we have in the justification of the backup. The price is too high to pay for not-knowing which component of the device exactly will fail during its use.

Therefore, first of all, it is necessary to discuss the economic model of the cost of duplication of products as the simplest and most widespread type of backing up.

4 Evaluation of the economic efficiency of the increase in reliability of an innovative product by means of backing up

The working efficiency of systems without backing up requires working efficiency of all of the components of the system. In complex technical devices, without backing up, it is never possible to reach a high level of reliability even in the case of using components with high fail-free operation indices. System with backing up is a system with a redundancy of components, i.e. with backup parts, which are redundant in relation to the minimum necessary (main) construction and executing the same functions as the main components. In systems with backing up, the operation of the system is guaranteed while there are enough backup components that can start their work when the original components fail.

It is necessary to state that failure of the main component or component duplicating the main one, does not mean failure of the duplicated device in general. If a backing up device is in standby mode, it is easy to show

that the total probability of no-failure operation of the whole duplicated device of $P_g(t)$ will be as follows:

$$P_g(t) = 1 - [1 - P(t)]^2 = P(t)[2 - P(t)], \tag{8}$$

where $P(t)$ – probability of no-failure operation of the main or backing up device.

Duplication of a product prolongs its fail-free operation, i.e. $P_g(t) > P(t)$. Thus, it is obvious, that the cost of the duplicated device exceeds the cost of the main or backing up device by at least two times and, naturally, it is necessary to expect that the total expenses will be, at least, two times higher compared to the expenses for the creation of non-backed up devices. However, this seemingly obvious conclusion requires deeper analysis and, as it will be presented, it is not always justified.

Let us explain the relevance of raising the issue of the economic efficiency of the backing up of devices using an example of the simplest case of backing up duplication.

For example, the probability of no-failure operation of the developed innovative device considering the fail-free operation of the existing components base, is evaluated at the level of $P = P(t_g) = 0.9$ for a guaranteed period of time t_g , but the required reliability value of this device for the same period of time is $P = P_{required}(t_g) = 0.99$.

There are two obvious solutions to this issue.

The first way represents an increase in the fail-free operation value of the device due to the increase in reliability of its components without backing up of the components or the device in general, i.e. thanks to using a more expensive and more reliable component base, more attentive selection of components, more careful input and output control, etc.

The second way is duplication of the device, which also represents a solution to the set problem, because the probability of fail-free operation $P_g = P_g(t_g)$ of the duplicated device according to expression (2) and the probability of no-failure operation $P(t_g) = 0.9$ set above, will be equal to the required value.

$$P_g = 1 - [1 - P]^2 = 1 - 0.1^2 = 0.99$$

If here, the number of required devices with the probability of no-failure operation $P_g(t_g) = 0.99$, is equal to N_g , the general production of the duplicated devices will be:

$$N_{\Sigma g} = N_g + (1 - P_g(t_g))N_g \tag{9}$$

Because the number of duplicated devices $(1 - P_g(t_g))N_g$ with the probability of no-failure operation $P_g = P_g(t_g)$ during their use can fail.

Here $(1 - P_g)$ – is the probability of failure of a product when probability of no-failure operation is $P_g = P_g(t_g)$ during the guarantee period of its use $t = t_g$.

One duplicated product represents two identical nonredundant devices, therefore the cost of every

duplicated device makes $2C_1$, where C_1 – the cost of one nonredundant device.

Accordingly, taking into account formula (8) the total production costs $N_{\Sigma g}$ of the duplicated products make:

$$C_{\Sigma g} = [N_g + (1 - P_g(t_g))N_g]2C_1 = 2N_g C_1(2 - P_g) \tag{10}$$

If our aim is to increase fail-free operation of a product without its duplication, then total expenses C_{Σ} will make:

$$C_{\Sigma} = N_1 C_1 [1 + K(a/k)](2 - P_1) \tag{11}$$

where $K(a/k)$ – factors, characterising the cost of the increase in fail-free operation from the value of probability of no-failure operation P to the value P_1 .

Increase in the probability of fail-free operation is equivalent to the decrease in fault intensity:

$$P_1 = e^{-\lambda t} \cdot tr$$

According to Equation (7) the value of the production costs of the product in this case will make:

$$C_{prod} - aP_1^a = aP^{a/k}$$

Using Equations (11) and (12), let us determine the $K(a/k)$ value, in the case of which the number of goods $N_1=N_g$ and the total costs for the increase in reliability without duplication are equal to the total costs when duplication is used and here the probability of fail-free operation of a device with duplication P_g and a device of enhanced reliability without duplication P_1 are equal. Solving these equations together we will get:

$$K(a/k) = \ln P_1 / \ln P \tag{12}$$

It coincides with the found dependence $K = \ln P_1 / \ln P$. If the actual probability of fail-free operation of the nonredundant device we used above as an example $P=0.9$, and the required $P_1=P_g=0.99$, than, by inserting these values in (11), we get:

$$K(a/k) = \ln 0.99 / \ln 0.9 \approx 0.1$$

The obtained result means that in the case of AN increase of fail-free operation of a product without backing up, evaluated by means of the value $K(a/k)=0.1$, the total costs of production and use during the guaranteed period of time t_g of the duplicated products or nonredundant products with the increase of their fail-free operation without duplications are identical, i.e. $C_{\Sigma}=C_{\Sigma g}$. At $K(a/k)>0.1$ duplication of a device requires less expenses than the corresponding increase in reliability of a device without duplication in other conditions, described above, i.e. $C_{\Sigma g} < C_{\Sigma}$. In case of other values of reliability of a product the P and P_1 used in calculation of increase of fail-free operation of a product, and determined by means of (11), will have other values.

For example, at a primary level of reliability (probability of fail-free operation $P=0.7$) and required

level of probability of fail-free operation $P_1=0.9$, the factor determining the increase in cost of a device of enhanced reliability will be:

$$K(a/k) = \ln 0.99 / \ln 0.7 \approx 0.3$$

It means that if the cost of the increase in fail-free operation of a product reached by the enterprise-manufacturer $K(a/k) > 0.3$, then manufacturer's total expenses on production and replacement of the faulty devices for new ones in the case of duplication of devices will be less than the total expenses in the case of an increase in fail-free operation of products without their backing up. Indeed, let us suppose that the factor determining the reached cost of increase in fail-free operation of products makes $0.4 > 0.3$ for the manufacturer, and initial terms remain the same: $P=0.7$, $P_1=0.9$.

As follows from (9), the relative total costs of the enterprise in the case of duplication of the product are equal to:

$$C_{\Sigma g} / C = 2(2 - P_g) = 2(2 - 0.91) = 2.18$$

It is easy to show that if primary reliability is equal to 0.7, P_g will be

$$P_g = 1 - [1 - P]^2 = 1 - 0.3^2 = 0.91$$

In the case of an increase in fail-free operation of the product without duplication, from expression (5) we get:

$$C_{\Sigma g} / C = (1 + 0.4(\ln 0.7 / \ln 0.9))(2 - 0.91) = 2.589$$

As a result

$$\Delta C_{\Sigma} = C_{\Sigma} - C_{\Sigma g} = 2.589 - 0.409C.$$

Taking into account that the total cost of the initially manufactured products $C = N_1 C_1$, where N_1 – number of manufactured products, and C_1 – cost of one product.

Thus, the economic efficiency of the duplication of a product depends on the actual safety level of the product, the required reliability level P_1 , as well as the cost of the increase in fail-free operation of the product, evaluated by means of a/k value in case backing up is not made.

Thus, the higher the cost of the increase in fail-free operation of the product without its duplication, the higher the probability that duplication of the product will require lower costs regarding the production and warranty service from the consumer compared to the increase in fail-free operation of the product without duplication.

Duplication of products at low values of cost for an increase in fail-free operation is economically inefficient, except for the cases when duplication is obligatory for an enterprise-manufacturer according to the specifications of the produced device.

The results stated above on the comparative analysis of the cost of production and warranty service of the backed up and non-redundant devices of innovative products, allow justification of the necessary amount of

money for the financing of an innovative project by an investor.

The approach applied to the analysis of these requirements allows certain methods for a decrease in the expended monetary means regarding the duplication of products to be offered, if the strategy of duplication is accepted. If a non-redundant device offers the probability of fail-free operation P , and the required probability is equal to P_1 , then in a number of cases it is possible to do as follows.

If products of two quality classes are manufactured, one of them with a probability of fail-free operation P , and the other one with a probability $P_2 < P$. If $P_2 < P$, the cost of the second-class quality product will be lower according to the presented in equations. In the case of duplication of these two devices, the resulting probability of fail-free operation will be:

$$P_g = 1 - (1 - P)(1 - P_2). \quad (14)$$

From the presented dependence it is evident, that if $P_2 < P$, and the production costs of such a product are lower, then later during decision-making it is necessary to discuss requirements applied to level of reliability.

If applying the requirement $P_g = P_1$ of the required probability of fail-free operation, using (9) we get:

$$P_g = (P_1 - P) / (1 - P). \quad (15)$$

For example we will perform the calculation of the required reliability of the cheaper device P_2 , if the reliability of the first-class quality product $P = 0.95$, and the required probability of fail-free operation of the duplicated product $P_1 = 0.99$. Using the given numbers in (8) we get $P_2 = 0.8$.

Obviously the product with lower reliability will have a lower price and will allow requirements for investments to be reduced without a decrease in the requirements for the reliability of the produced innovative product.

The only inconvenience in this case is that two assembly lines may be required for the first- and second-class quality products, but due to the obvious economic efficiency of it, such an approach has a right to exist.

Modern technological processes allow this problem to be solved. It is far from certain that for the production of products of a different quality class two different production lines will be required. Such a requirement is only possible in the case if they are absolutely technologically different, or if the identical number of such products must be manufactured at the same time. If there are no such requirements, the organisation of production can be performed consistently using one and the same production line.

5 Conclusions

Thus, it is determined that the higher the cost for an increase in fail-free operation of the product without its duplication, the higher the probability that duplication of

the product will require lower total costs regarding the production and warranty service of the products compared to the increase in fail-free operation of the product without duplication. The higher the cost for an increase in fail-free operation of the product without its duplication, the higher the probability that the total costs

consumer will be lower than the costs for duplication of the products.

Economic efficiency of duplication of a product at low values of cost for an increase in fail-free operation is only possible in the case of a comparatively high level of fail-free operation of the nonredundant product.

References

- [1] Levin V I 1985 The Logical Theory of Reliability of Complex Systems Moscow: Energoatomizdat 1985 *(in Russian)*
- [2] Ostreykovsky V A 2008 Reliability Theory Man for High Ed Inst Moscow Vyssh shc *(in Russian)*
- [3] Nechiporenko V I 1977 Systems structural analysis (efficiency and reliability) Moscow Sov Radio *(in Russian)*
- [4] Ryabinin A, Cherkosov G N 1981 Logic and probabilistic investigation methods of systems reliability. – Moscow Radio and svjaz *(in Russian)*

Authors

| | |
|--|---|
|  | <p>Ilana Ter-Saakova, born on April 20, 1974, Riga, Latvia</p> <p>Current position, grades: Ph.D. student Baltic International Academy. University studies: MBA Information System Management Institute, Riga, Latvia, 2011. Scientific interest: economy, reliability of system.</p> |
|  | <p>Nataly Podolyakina, born on March 3, 1967, Russia</p> <p>Current position, grades: Pro-Rector for Economy Transport and Telecommunication Institute, Riga, Latvia University studies: Ph.D. Moscow Civil Aviation Technological State University. Scientific interest: Economy Publications: More than 17 papers published in various journals. Experience: Teaching experience of 12 years.</p> |

The continuous-time optimal portfolio using a multivariate normal inverse Gaussian model

Xing Yu*, Guohua Chen

Department of Mathematics & Applied Mathematics Hunan university of humanities, science and technology, Loudi, 417000, P.R. China

Received 1 March 2014, www.tsi.lv

Abstract

This paper develops the continuous-time portfolio model using a multivariate normal inverse Gaussian model. Though the weighted average of lognormal variables is no longer lognormal, it can be approximated by other distributions, such as a multivariate normal inverse Gaussian model. Our method belongs to the analytic approximation class. By comparing to Monte Carlo experiments, it illustrates the computational efficiency and accuracy of our approach.

Keywords: Continuous-time portfolio, Normal inverse Gaussian, Approximation, Monte Carlo, Optimization

1 Introduction

The corns torn of portfolio selection problem is stem from Markowitz (1952) [1] on mean-variance model for single period portfolio selection problem. After Markowitz's pioneer work, numerous scholars extended the single period case to multi-period ones, and continuous-time framework. Bielecki et al (2005) [2] considered bankruptcy prohibition in continuous time with martingale approach. Czichowsky and Schweizer (2011) [3] proposed cone- constrained continuous-time mean-variance portfolio problem with price processes being semi-martingales. Some literates aim to the market condition under continuous-time environment. Li et al (2002) [4] supposed the price processes of assets are continuous Ito process, and derived the optimal portfolio for the continuous-time mean-variance model with no shorting using duality method. Fu (2010) [5] derived explicit closed form solutions for the dynamic mean-variance portfolio selection problem with borrowing constraint, the method used is the HJB equation of stochastic programming. Cui (2014) [6] considered the mean-variance formulation in multi-period portfolio selection under no shorting constraint.

To the best our knowledge, few of all the existing research focus on the price process of weighted sum of assets, in fact, the continuous-time portfolio payoff depends on the value of a portfolio of assets. The challenge in describing the portfolio stem from the fact that there is no explicit closed form for the weighted sum of correlated assets. There are two categories approximation techniques to solve this problem, numerical methods and approximations. Although numerical methods such as Monte Carlo simulation is a very flexible method, it is very time-consuming. Jarrow and Rudd (1982) [7] is the first to introduce Edeworth

expansion. Turnbull and wakeman (1991) [8] used an Edeworth series expansion to approximation the density function of the weighted sum. Mileusky (1998) [9] adopted the reciprocal Gamma distribution for alternative. Because a normal inverse Gaussian process incorporates an idiosyncratic drift, characteristics volatility, correlated Brownian motion, and a common inverse Gaussian time change, the multivariate normal inverse Gaussian model should provide more realistic diffusion of assets. This paper adopts a multi-normal inverse Gaussian (MNIG) process approximation to the weighted sum of correlated assets.

The plan for the paper is as follows. It is details the MNIG process in section 2. In section 3, we propose the approximation method. Section 4, introduces the optimal portfolio selection model and give the compared results of our method to Monte Carlo experiment in a numerical example. Section 5 contains our conclusion.

2 A multi-normal inverse Gaussian distribution

In the following, we will introduce the notions of ING and MING process according to Wu (2009) [10]. Suppose that G follows an inverse Gaussian distribution with parameter $a, b > 0$, whose density function is as follows

$$f(x) = \sqrt{\frac{1}{2\pi x^3}} \exp\left(-\frac{b}{2a^2 x} (x-a)^2\right), x \geq 0.$$

And its characteristic function is given by

$$\varphi(u) = \exp\left(\frac{b}{a} \left(1 - \sqrt{1 - \frac{2a^2 u i}{b}}\right)\right).$$

* *Corresponding author* e-mail: hnyuxing@163.com

A direct calculation yields the mean and variant are a and $\frac{a^3}{b}$, respectively. Due to the stochastic process is dependent on time t . It needs to introduce G_t whose parameters are at and bt . So $E G_t = at$, $Var G_t = \frac{a^3 t^2}{b}$. For simply, let $a=1, b=\frac{1}{\gamma}$.

We define a normal inverse Gaussian process $X^j_t = \theta_j G_t + W^j_{G_t}$, where W_{G_t} is Brownian motion [11], θ determines the tendency of the sample paths. The mean and variant of X^j_t are θ_j and $\sigma_j^2 + \gamma \theta_j^2$, σ_j is the volatility rate, and the characteristic function is $exp \gamma^{-1} [1 - \sqrt{1 + u^2 \sigma_j^2 \gamma - 2u \theta_j \gamma i}]$.

3 Model formulations

It is considered a portfolio consisting of n assets with price $S_t^i = S_0^i exp [r - q_i] t + X_t^j + d_j t$, where $d_j = exp \gamma^{-1} \sqrt{1 + u^2 \sigma_j^2 \gamma - 2 \theta_j \gamma i} - 1$.

The log-value of portfolio is $s_t = ln S_t = \sum_{i=1}^n \omega_i ln S_t^i$. The characteristic function of s_T is as $\varphi_{s_T}(u) = exp [iu aT - qT + s_0 T + k^{-1} T \sqrt{1 + u^2 \sigma^2 k - 2i \theta u k}]$, $d = \sum_{i=1}^n \omega_i d_i, \theta = \sum_{i=1}^n \omega_i \theta_i,$ $q = \sum_{i=1}^n \omega_i q_i, S_0 = \sum_{i=1}^n \omega_i S_0^i, k = \gamma T^{-1}$, $\sigma = \sqrt{\sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j \rho_{ij} \sigma_i \sigma_j},$ $w = exp [k^{-1} \sqrt{1 + \sigma^2 k - 2 \theta k} - 1]$. $a = r + d + w$

The detailed derivation refer to Wu. The relationship of the characteristic function and moment [12] is as follow proposition.

Proposition. Suppose that there exist n order moments of variable X , then the characteristic function of X also exist, and its n order derivative exist,

$$E X^k = \frac{\phi_X^k(0)}{i^k}.$$

In fact, $\phi_X^k(t) = \int_{-\infty}^{+\infty} i^k exp [itx] f(x) dx = i^k E X^k exp [itX]$.

Let $t = 0, E X^k = \frac{\phi_X^k(0)}{i^k}.$

So $E X = \frac{\phi_X'(0)}{i}, E X^2 = -\frac{\phi_X''(0)}{i^2}.$

From the characteristic function of s_T , we get $E s_T = aT - qT + T\theta + S_0,$ $E s_T^2 = aT - qT + \theta T + S_0^2 + T\sigma^2.$

4 Numerical example

In this section, we firstly describe the optimal portfolio model, then give an numerical example to illustrate the accuracy and computational efficiency of our method.

4.1 THE OPTIMAL PORTFOLIO MODEL

We consider the relatively simple continuous-time mean-variance portfolio selection model refers to the problem of finding the optimal admission strategy to minimize the variance while attaining a given level of the expectation

$$\begin{cases} \min_{\omega} Var s_T = E s_T - u^2 \\ s.t E s_T = u \end{cases}, \tag{1}$$

where u is a given constant, representing the expected level, which the investor requires to achieve, $E s_T, Var s_T = E s_T^2 - E s_T^2$ is from section 3.

According to the optimal portfolio model (1), the problem can be deal with by the Lagrange method. Introducing the Lagrange multiplier λ leads to the following problem

$$\min E X T - u^2 + 2\lambda [E X T - u]. \tag{2}$$

Let $\pi T = g(x, T), \lambda$ be the optimal solution of the Lagrangian problem (2) and $G(x_0, \lambda)$ be the optimal value. According to the Lagrange duality theory, if λ^* satisfies $\max_{\lambda} G(x_0, \lambda)$, then $\pi^* T = g(x, T), \lambda^*$ is the optimal shares of (1) and $G(x_0, \lambda^*)$ is its optimal value.

Problem (2) is equivalent to $\min E x T + a^2$, where $a = \lambda - u.$

4.2 ILLUSTRATIVE EXAMPLE

The goal of our numerical experiments is to test t the computational efficiency and accuracy of our approach. We therefore set up a simulation study and compare the results to the calculated results using Monte Carlo simulations.

First, we define a two-asset portfolio with a one year, a constant continuously compounded risk-free rate of 0.1. Each asset is given its own dynamic parameters: the drift parameters θ_1 and θ_2 are 0.1 and 0.2, and the volatility parameters σ_1 and σ_2 are 0.2 and 0.3. We also consider two different levels of correlation and weights between the underlying assets: the correlation ρ is set to either 0.5 or 0, and the weights are set to either (0.7, 0.3) or (0.3, 0.7) under two different economic states corresponding to $\gamma = 0.1$ and 0.2. To solve problem (1), the Monte Carlo simulation results of portfolio weights is $\omega_1 = 0.2813, \omega_2 = 0.7187$, and the results from the approximation method proposed in our study is $\omega_1 = 0.2752, \omega_2 = 0.7248$.

From the compared results, we find that the portfolio shares from our model and Monte Carlo simulation is nearly the same.

5 Conclusion

The difficult to solve the continuous-time portfolio selection is a closed-form formula of the weights of assets is not available. The main contribution of this

paper is to develop a closed-form analytic expression for the portfolio. Our approach is based on a multivariate normal inverse Gaussian (mNIG) model, which is a more appropriate representation of asset dynamics than the geometric Brownian motion (GBM) model. Because an NIG model has economic meaning: θ and σ represent the drift and volatility of the individual assets respectively, while γ represents the effect of an economic state shared by all assets. Numerical example results for two-asset shows the accuracy and computational efficiency compared to Monte Carlo simulation results. It provides a new way to deal with the continuous-time portfolio selection.

Acknowledgment

This research is supported by the key project of Hunan province department of education (12A077).

Reference

- [1] Markowitz H M 1952 Portfolio selection *Journal of Finance* **7** 77-91
- [2] Bileleck T R, Jin H Q 2005 Continuous-time mean-variance portfolio selection with bankruptcy prohibition *Mathematic finance* **15** 213-44
- [3] Czichowsky C, Schweizer M 2011 *Cone-constrained continuous-time Markowitz problems* NCCR FINRISK working paper **683** ETH Zurich
- [4] Li X, Zhou X Y, Lim A E B 2002 Dynamic mean-variance portfolio selection with no-shorting constraints *SIAM Journal of control optimization* **40** 1540-55
- [5] Fu C, Lari-Lavassani A, Li X 2010 Dynamic mean variance portfolio selection with borrowing constraint *European journal of operational research* **200** 312-9
- [6] Xiangyu Cui, Gao J J, Li D 2014 Optimal no-shorting constraint *European journal optimal multi-period mean-variance policy of operational research* **234** 459-68
- [7] Jarrow R, Rudd A 1982 Approximation option valuation for arbitrary stochastic process *Journal of financial economic* **10** 347-69
- [8] Turnbull S, Wakeman L 1991 A quick algorithm for pricing European average options *Journal of financial and quantitative analysis* **26** 377-89
- [9] Milevsky M A, Posner S E 1998 A closed-form approximation for valuing basket options *Journal of derivatives* **5**(4) 54-61
- [10] Wu Y C, Liao S L, Shyu S D 2009 Closed-form valuations of basket options using a multivariate normal inverse Gaussian model *Mathematics and Economics* **44** 95-102
- [11] Carr P, Geman H, Madan D B, Yor M 2003 Stochastic volatility for levy processes *Mathematical finance* **13** 345-82
- [12] Morris H, DeGroot, Schervish M J 2005 *Probability and Statistics* Higher Education Press

Authors



Xing Yu, born on February 15, 1981, in Xianning City of Hubei province

Current position, grades: Hunan university of humanities, Science and technology (China); Lector

Professional interests: Applied mathematics; Finance model

Research interests: the optimal portfolio model; mathematical model



Guohua Chen, born on July 24, 1969, in Xinhua City of Hunan province

Current position, grade: Hunan university of humanities, Science and technology (China); doctor

Professional interests: Applied mathematics; Finance model

Scientific interests: the optimal portfolio model; mathematical model

Computer information technology and agricultural logistics management system

Peng Ma*

Zhongzhou University, 450044, Henan, China

Received 15.08 2014, www.tsi.lv

Abstract

At present, there are kinds of problems on circulation pattern of agricultural products' supply chain, leading to high cost of agricultural logistics system and unreasonable planning. In order to solve the problem of agricultural product circulation pattern, we need to put forward a circulation pattern of agricultural products' supply chain, which takes the agricultural product logistics as a core enterprise. This thesis introduces the idea of combination between computer information technology and logistics management system. It also analyses how to better complete the modules of logistics management system and key points of them, based on the technology of computer, automation, bar code, etc., which focus on analysing the module of farm, customer relations and decision.

Keywords: supply chain; logistics management; circulation pattern; information technology; module

1 Introduction

IT application in agricultural logistics management is a new content of agricultural modernization and inexorable trend of world agricultural development. With a background of rapid expansion of economic globalization, agricultural products market at home and abroad needs to be integrated, which demands us to improve the level of agricultural informatization construction. From the view of agricultural producer, operator, or government sector, there is vitally important significance on agricultural logistics management information system construction. It is the need of optimizing allocation of resources of agricultural producer and expanding the market of agricultural products, the need of reducing the agricultural risk (including the market risk and natural risk), the need of increasing the agricultural products international competitiveness (including the price competitiveness, quality competitiveness and brand credibility competitiveness), the need of government support and agriculture protection [1].

In recent years, with the rapid expansion of computer networks and communication technology, agricultural products logistics management system evolved from an easy mode into automated management whose main feature automatic logistics equipment, such as automatic guided vehicle automatic storage and extraction system, sky-rav-rail automated vehicles, stockers, etc., as well as the appearing of logistics computer management and control system. They change people's life-style, work-style and thinking method with leap-type to bring immeasurable social profit and economical profit to the whole society.

2 Summary on computer information technology and logistics management

Analysing the present stage of circulation pattern of agricultural products' supply chain, we discovered that there are many defects [2]. In order to solve the disadvantages and deficiency, we put forward a circulation pattern that takes the agricultural product logistics as a core enterprise, focusing on the profit of agricultural products factory and customer. Take the measure of combination with computer technology to improve the supply chain pattern, which makes the logistics and information stream circulate more smoothly.

The chief application of computer information technology in logistics information technology includes device collecting and transporting, storage equipment automation, as well as information collecting, transmitting, and processing based on bar codes technique, EDI technique and network technique. With the social development and the approaching of e-commerce, computer technology makes the logistics informationized, automated, networked, intelligentized and flexible [3]. Multifunctional modern logistics highly depends on some demand, which contains mass of data and information collecting, analysing, processing and immediate updating. That's why logistics informatization is the inevitably demand of society imformatization. The application of these modern techniques and equipment highly improves the efficiency of logistics activities and enlarges the field of logistics activities, forming into supply chain.

*Corresponding author e-mail: mpzdx@163.com

3 Logistics management system

There is an idea, which refers to system integration and the total cost controlling in modern logistics. Economic activity includes supply, production, market, transportation, inventory and other relevant information flow. Modern logistics see all these as a dynamic system master, concerning the operation efficiency and costs in this system [4]. The whole system specifically includes these modules: purchasing management, farm, customer relation and decision, inventory management and supply chain management system [5].

3.1 PURCHASING MANAGEMENT

With the system providing the correct and immediate purchasing information, we help business manager make a scientific acquisition strategy, provide them purchasing management in a good time, defined amount and price and know the performance of suppliers in time. System provides the function from purchasing requisition to cargo checkups and acceptance, data quality monitoring, etc., realizing the general management of purchasing business.

3.2 FARM, CUSTOMER RELATION AND DECISION

The agricultural buying and selling system does not run smoothly means it does not obey the marketing rule to make the production-supply-marketing into a dragon (a mode of agricultural products supply chain). Therefore, the agriculture cannot be developed rapidly, usually leading to a situation of 'hard to buy or selling'. The market of agricultural products becomes the main part of the whole supply chain. For this reason, we came up with the bi-le module programming module of agricultural logistics system, to optimize the supply chain.

3.2.1 The basic idea of Bi-level programming module of supply chain

From the supply chain integration's perspective, it uses bi-level programming module [6] to describe the optimization problem of bi-level distribution network, which takes a full consideration on the both profit of agricultural products factory and customers and design a heuristic solution algorithm to solve as well.

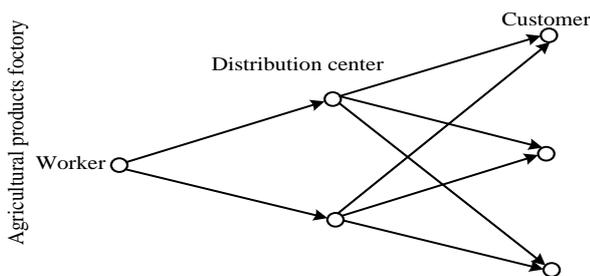


FIGURE 1 Agricultural products distribution framework

We could regard the optimization problem on distribution network as a mentor (henchman) problem. Therein to, the decision-making section of manufacturing concern is the mentor, and customer's choice behaviour to distribution centre or his need on allocation of distribution centres is the henchman. Decision-making section could change the position and distribution costs of a distribution centre through policy and management, affecting the customer's choice to distribution centre, but not controlling. Customers could make a comparison between available distribution centres and make the choice based on their need and behavioural habits. This kind of relationship can be described by bi-level programming, whose basic idea are two submodels of following mathematic model bi-level programming model:

$$\begin{aligned} \min_x & F(x, y) \\ \text{s.t.} & G(x, y) \leq 0 \end{aligned}$$

Among them: $y = y(x)$ can be got from following project (L_0):

$$\begin{aligned} \min_y & f(x, y) \\ \text{s.t.} & g(x, y) \leq 0 \end{aligned}$$

Bi-level programming model is consist of two submodel of (U_0) and (L_0). (U_0) is called superstratum project. (L_0) is called substratum project. F is the objective function decided by superstratum. x is the decision variable of superstratum. G is constraint of variable. f is the objective function decided by substratum project. y is the decision variable of substratum. g is constraint of variable y . The superstratum decision maker affects the substratum decision maker by setting x , so it limits the feasible constraint set of substratum decision maker. On the contrary, the activity of substratum decision maker will affect superstratum decision by y . Therefore, the variable y is the function of x , which means $y = y(x)$. It is usually called reaction function. The optimization problem of supply chain's distribution network relates to 2 decision makers (the agricultural products factory and customers) of obviously different objective function, so it is proper to describe this relation by bi-level programming module [7].

3.2.2 The optimization of agricultural products supply chain's distribution Bi-level module

Considering the superstratum decision maker is the owner of agricultural products factory (the peasant households), the problem they are concerning about involves 2 aspects, costs and profits. It means the costs can be less, while the profit can be more. Taking this as an objective, we could set a project of superstratum decision maker. Due to the distribution centre built by investment of peasant households, the costs they are concerning about are not only production costs, but also costs of delivery centre. The substratum decision maker

is the customers. Their objective to distribute the quantity demanded among the distribution centres to chasing the least total costs. The module can be set by (U):

$$\min F = \sum_{i=1}^m \sum_{j=1}^n C_{ij}x_{ij} + \sum_{j=1}^n f_j u_j + \sum_j \left(t_j \sum_i x_{ij} \right) - \theta \sum_i \sum_j x_{ij}, \quad (1)$$

$$\text{s.t.} \sum_{j=1}^n u_j \geq 1, \quad (2)$$

$$u_j \in \{0,1\}. \quad (3)$$

The first item in this objective function is the generalized unit cost of the *i*-th customer served by the distribution centre of *j* place, which mainly means the transportation expanses. It can be estimated in practical application with the increase of quantity demanded. x_{ij} is the quantity demanded of *i* customer satisfied by distribution centre of *j* place. θ is a harmonic coefficient. As it mentioned before, the peasant households ask for the best profit while taking the least costs. In many works, the optimization objective is the least costs but ignoring the factor of profit. Taking these factors into consideration, the author takes a way to make the module better. Due to the unfinished distribution centre, how can we judge the profit? There are many unpredictable factors in the market which we don't take into account. Under this circumstance, if the total demanded customers are large, it reflects that the market share of this distribution is large. The market share is an important indicator to measure the profit. Therefore, we use this indicator to measure the profit expected. The first 3 items of this objective function is the cost objective. In order to balance the disunity between costs and profit indicator, we bring in harmonic coefficient θ . f_j makes the fixed investment on building the distribution centre at *j* place. t_j is the transportation costs from agricultural products factory to the *j* distribution centre. X is the constant. The transportation costs could be got according to the data of local transportation and carriage former years. Formula (2) ensures to set at least one distribution centre. 0 and 1 in formula (3) is variable. You should set it to 1 when building distribution centre at *j* place, otherwise you should choose 0.

$$(L) \min T = \sum_i \sum_j \int_0^{x_{ij}} D^{-1}(w) dw, \quad (4)$$

$$\text{s.t.} \sum_{j=1}^n x_{ij} = w_i, \quad i = 1, \dots, m; \quad (5)$$

$$D^{-1}(x_{ij}) \geq 0 \quad i = 1, \dots, m; \quad j = 1, \dots, n; \quad (6)$$

$$\sum_{i=1}^m x_{ij} \leq s_j \quad j = 1, \dots, n; \quad (7)$$

$$x_{ij} \leq s_j u_j \quad i = 1, \dots, m; \quad j = 1, \dots, n; \quad (8)$$

$$x_{ij} \geq 0 \quad i = 1, \dots, m; \quad j = 1, \dots, n. \quad (9)$$

$D_{ij}^{-1}(\cdot)$ in the formula is the inverse function of demand function, which means cost function. The expression of quantity demanded is $x_{ij} = D_{ij}^{-1}(f_{ij})$. In this formula, x_{ij} is the quantity demanded of *i* customer satisfied by distribution centre at *j* place. f_{ij} is the generalized cost function. It is the demand of *i* customer that the least cost should be provided by the contribution centre at *j* place, which usually use power function or log function to express. w_i is the quantity demanded of *i* customer spot. s_j is the best supply ability of the distribution centre at *j* place.

The substratum programming is the customer's distributed quantity demanded of each distribution centres to make least to the total costs. Formula (5) ensures customer's quantity demanded could be satisfied by distribution centres. Formula (6) is nonnegativity restrictions of cost function. Formula (7) ensures the quantity demanded will not be beyond the best supply ability. Formula (8) ensures the quantity demanded could be distributed only at built distribution centre. The quantity demanded will be 0 if the distribution centre is not built. Formula (9) is nonnegativity restrictions of variable. Using the centre quantity u given by substratum programming could figure out that hessian matrix of objective function is positive definite. Therefore, there is the only solution to module (L).

By analysing the formula (8), we could see if $u_j=0$, $x_{ij}=0$ is correct. If $u_j=1$, $x_{ij} \leq s_j$ of formula (7) is obviously correct. It is to say, if u is fixed, formula is useless which could be omitted.

3.2.3 The solution method for the model

Through analysis, the superstratum of this module is a normal module of location problem, and the substratum is a module of nonlinear programming. It is hard to get the solution. Therefore, as the bi-level programming module mentioned in this thesis, the scholars like Sun Huijun came up the heuristic algorithm: Take the Formula (8) as a heuristic method to get the solution. In the meanwhile, we could make following assumption: The production-supply-marketing situation of agricultural products factory in a given period conforms to the module condition; Presumed inventory and extra lost is little; optimize one kind agricultural products logistics system at one time.

From the module (L), we could see formula (8) expresses the relation between the distributed quantity demanded of customers and address selection planning of each distribution centres. But from the formulas mentioned, formula (8) could be omitted. However, in order to get the reaction function, we should keep it but not in the module, to simplify it into the following form:

$$x_{ij} = S_j u_j - y_{ij}, \quad i = 1, \dots, m; j = 1, \dots, n. \quad (10)$$

In this formula, y_{ij} is a slack variable. When $u_j=0$, we could get x_{ij} and y_{ij} directly. When $u_j=1$, we could use available way to solve the substratum module (L) to get the quantity demanded at each distribution centres x_{ij}^* when in a balanced stage. Then use formula (10) to get y_{ij}^* , and the relations of all reaction function could be expressed:

Bring that relation into superstratum objective function and solve by an available way, like branch and bound method. For the solution solved from superstratum problem, we could solve the substratum problem and get the distribution of customers' demanded quantity. And repeat that, we could get another new address selection planning. Finally, it is possible to get a best solution though bi-level programming. The solution algorithm is in fact a heuristic algorithm based on formula (11). Here are the steps:

Step 1: Set a initial solution u_j^0 , make the iteration $k=0$.

Step 2: For the given u_j^k , get the substratum problem solution x_{ij}^k .

Step 3: Base on formula (10), calculate y_{ij}^k , and bring $x_{ij}^k = S_j u_j^k - y_{ij}^k$ into superstratum objective function. Get the solution of superstratum problem and get a new u_j^{k+1} .

Step 4: Stop if $|F^{k+1} - F^k| \leq \varepsilon$; otherwise, make $k=k+1$ and back to step 2: ε is iteration accuracy.

This method is heuristic algorithm which is hard to prove the styplicity, but the calculation improves the method is convergent. Those modules are basic agricultural products logistics bi-level programming module, which is appropriate for the agricultural products that won't go bad in a short time, like wheat, corns, beans, peanuts and so on. As for the vegetables and fruit which go bad easily, it is only to add the timing constraint into the original module. This constraint is substratum constraint, which is not appropriate to add into available module. We could make the judgment before calculating the substratum programming in following way.

In the following formula, $e_{ij} \leq E_{ij}$ is the delivery time which is asked not to surpass the given time by I

customer at j distribution centre. e_{ij} means the delivery time. E_{ij} is the given time.

For $i=1$ to m

For $j=1$ to n

If $e_{ij} \leq E_{ij}$ then

$x_{ij} > 0$

Else

$x_{ij} = 0$.

3.3 INVENTORY MANAGEMENT

Modern storage, instead of static storage, is a process of logistics which operate through the application of buffer stand, accumulation area, and some related operations. Storage is a process that materials have not only a short stay. Using this system, managers can obtain real-time dynamic material inventory information. In order to better control inventory, improve efficiency and timeliness of delivery and improve customer service levels, we could finally reduce inventory costs, production costs, feedback logistics information timely, and accelerate capital operation through the intelligent analysis [8].

3.4 SUPPLY CHAIN MANAGEMENT SYSTEM

- 1) Take advanced data encryption technique to ensure the security when transferring the data.
- 2) On the basis of satisfying the market needs, the excellent architectural design has good scalability to satisfy the enterprise on business development in the future.
- 3) Support a variety of business models and integrated supply chain; support the ports of original background of enterprise ERP or MIS system interface; support the integration of CRM, OA, e-commerce, quality management system and so on, forming a powerful enterprise information centre.
- 4) Gathering the total quality management theory, the thought of ISO quality management system and so on, we will promote the development of the enterprise quality management and provide powerful tools for continuous improvement in quality. Quality management system regulates the quality management process to effectively monitor quality, improving product percent of pass, and eventually reducing the production cost, improving the economical benefits of enterprises [9].
- 5) B/S structure reduces the workload of maintenance, and supports telecommuting. This system totally adopts B/S structure. Customers completely don't need to install the software or configurate. Only through the IE can we realize all functions [10].

4 Conclusions

This paper shows the bi-level optimization module for agricultural logistics system. Take the whole agricultural supply chain as the research object on the purpose of obtaining the perfect effect of the whole supply chain. It involves the purchase of agricultural products supply chain, distribution, final agricultural products for customers, inventory management, supply chain management system and other aspects. Emphatically analysing the distribution model, it sets the bi-level

optimization module, which takes farmers as superstratum, customers as substratum. It plays a reference role on optimization decision problems of agricultural logistics system. The next research direction is to further explore the optimization method of supply chain, and get real data to argue.

Acknowledgment

Science and technology planning project of transportation office in Hunan (2012P63).

References

- [1] Qingzhen Z, Wenbin W, Peixin Z 2005 The ANP Method to Evaluate The Agricultural Logistics Performance *Scientific and comprehensive research on agricultural system* **21**(2) 237-240
- [2] Lambert D M, Cooper M C 2000 Issues in supply chain management *Industrial Marketing Management* **7**(29) 65-83
- [3] Deqiang Z 2013 The Optimization on Supply Chain And Logistics System of Agricultural products *Mechanical Manufacture and Automation in Inner Mongolia University of Science and Technology (in Chinese)*
- [4] Thomas E 2010 Principle of Service Design *Beijing, Beijing Electronic Industry Press*
- [5] Shaojun W 2013 The Enterprise Logistics Management under Supply Chain System *Management World* **49**6 150
- [6] Zhigang Z, Xinyi G 2006 The Method on Getting Solution of Supply Chain Distribution Module **28**(3) 299-302
- [7] James E S, Detlof V W 2004 Decision Analysis in Management *Science Management science* **50**(5) 561-74
- [8] Li Z 2013 The Influence of Computer Network techniques on Logistics informatization *Science and Technology Guide* **14** 50
- [9] Thomas E 2010 Service-Oriented Architecture *Beijing, Beijing Electronic Industry Press*
- [10] Zhenfei M 2011 The Design And Implementation on Comprehensive Logistics Supply Chain Management Platform *The Software engineering of Electronic Science and Technology University (in Chinese)*

Author



Peng Ma, Henan Province of China

Current position, grades: lecturer

University studies: PHD degree was earned in major of management science and engineering, Henan Agricultural University in 2003

Scientific interest: energy management

Research on speed regulation system for matrix converter fed induction motor

Junmei Zhao, Zhijie Zhang*, Yifeng Ren

School of Computer and Control Engineering, North University of China, 3Xue Yuan Road, Taiyuan, China

Received 1 June 2014, www.tsi.lv

Abstract

This study presents the application of a Matrix Converter (MC) and an active disturbance rejection controller (ADRC) to Direct Torque Control (DTC) system based on an induction motor. Matrix Converter (MC) is applied to Direct Torque Control (DTC) system based on an induction motor in order to reduce power grid harmonic pollution which is caused by AC-DC-AC converter in conventional DTC system. Then a PID controller and an ADR controller are both designed to regulate the speed of the system. Design procedures for ADRC are given in detail. Finally, corresponding results are compared. The simulation results show that the novel DTC system has combined the advantages of both MC and DTC--stable running, strong anti-jamming, good dynamic and static performance.

Keywords: DTC, MC, ADR controller, space vector, PI controller

1 Introduction

Direct Torque Control (DTC) has obtained widespread concern of scholars and has developed rapidly because of the simple control method and the fast system response. Generally, DTC method is used to AC-DC-AC converter, in which DC link not only increases system burden and reduces power factor, but also brings power grid harmonic pollution problem. For these problems, many improved methods have been proposed. An alternative method to reduce torque ripples based on space vector modulation (SVM) technique was proposed [1]. An adaptive DTC control for induction motor drive with a fixed switching frequency and a low torque ripple was reported [2]. A fuzzy logic controller was used to select voltage vector in a conventional DTC system [3]. A fuzzy adaptive controller was used to reduce torque ripple [4]. The principle of variable structure was used in DTC system for IPM Synchronous Motor [6]. A novel control method for DTC based on SVPWM was applied by using all voltage vectors of inverter to give a constant torque switching frequency and reduce torque ripple [7]. A high-performance direct torque control of an induction motor was proposed [8].

The use of matrix converter (MC) to DTC system for induction motor was introduced [5]. It has obtained rapid development; some improved methods were put forward. The induction motor's performance was improved because of these advanced methods, but speed regulator used PI controller in these systems. If the environment changes seriously when the system is activated, PI controller will not be able to adjust the system in real time, which leads

to poor system performance. In this paper, ADRC is used to take place of PI controller, in order to obtain better system performance.

In this paper, we will use MC instead of AC-DC-AC converter. It is superior to conventional inverter because it does not have bulky dc-link capacitors and offers bidirectional power flow capacitors, sinusoidal input or output current, and an adjustable input power factor. Furthermore, because of high integration, MC topology is recommended for extreme temperatures and critical volume or weight applications [9]. The modulation for MC includes four methods: double space-vector pulse width modulation, switching function modulation, double line voltage modulation and output circuit hysteretic current. In this paper, double space-vector pulse width modulation is presented, because it has the capability to achieve full control of both output voltage vector and input current vector [10].

In this paper, the combination of DTC and MC has important theoretical significance and provides some engineering reference for frequency converter. Furthermore, Active Disturbance Rejection Control (ADRC) was used to the system, which was invented by Professor Jingqing Han who serves in Chinese Academy of Science. It is a new control method which doesn't depend on the system precision model. It can estimate and compensate the influences of all internal and external disturbances in real time when the system is activated. Combining with the special nonlinear feedback structure, it can realize good control quality, such as small overshoot, fast response and strong robustness.

*Corresponding author e-mail: zzzhaojunmei@163.com

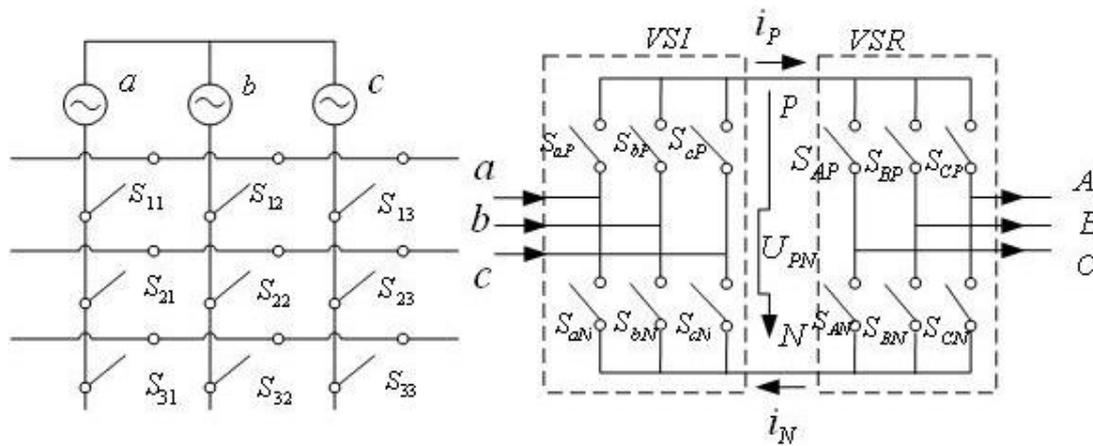


FIGURE 1 Matrix converter topology structure and equivalent AC-DC-AC structure

2 Modulation strategy of matrix converter

2.1 PRINCIPLE

Three phase MC topology is shown in Figure 1. Input phase voltage of MC is V_a, V_b, V_c respectively, output line voltage is V_{AB}, V_{BC}, V_{CA} respectively, there is:

$$\begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} = V_{im} \begin{bmatrix} \cos(\omega_i t) \\ \cos(\omega_i t - 2\pi/3) \\ \cos(\omega_i t - 4\pi/3) \end{bmatrix}, \tag{1}$$

$$\begin{bmatrix} V_{AB} \\ V_{BC} \\ V_{CA} \end{bmatrix} = \sqrt{3}V_{om} \begin{bmatrix} \cos(\omega_o t - \phi_o + \pi/6) \\ \cos(\omega_o t - \phi_o + \pi/6 - 2\pi/3) \\ \cos(\omega_o t - \phi_o + \pi/6 - 4\pi/3) \end{bmatrix}. \tag{2}$$

There is a transfer matrix T of duty ratio, which makes:

$$\begin{bmatrix} V_{AB} \\ V_{BC} \\ V_{CA} \end{bmatrix} = T \begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix}. \tag{3}$$

The matrix T can be written as follow:

$$T = m \begin{bmatrix} \cos(\omega_o t - \phi_o + \pi/6) \\ \cos(\omega_o t - \phi_o + \pi/6 - 2\pi/3) \\ \cos(\omega_o t - \phi_o + \pi/6 - 4\pi/3) \end{bmatrix} \begin{bmatrix} \cos(\omega_i t) \\ \cos(\omega_i t - 2\pi/3) \\ \cos(\omega_i t - 4\pi/3) \end{bmatrix}^T, \tag{4}$$

where m is modulation factor, $0 \leq m \leq 1$, ω_i is input power frequency, ω_o is output power frequency, ϕ_i is arbitrary power factor angle, and there is

$$U_{om} = \frac{\sqrt{3}}{2} U_{im} m \cos \phi_i. \tag{5}$$

So the MC input power factor can be any value, it can be positive, negative and even one. In actual operation, duty ratio transfer matrix T can be calculated according to modulation strategy in real time.

2.2 MODULATION STRATEGY OF MATRIX CONVERTER

Space vector modulation of matrix converter was put forward by L. Hubera and D. Borojevic in 1989. It is a kind of indirect modulation algorithm. The space vector modulation technology is being used to pre-virtual rectifier and rear virtual inverter. The sinusoidal wave with adjustable angular displacement can be obtained on the input side of MC, and fundamental voltage with adjustable amplitude, phase and frequency can be obtained on the output side of MC. Finally overall control is realized through combining the two parts, which is known as dual space vector method.

For inverter part, output voltage vector distribution and output voltage reference vector synthesis are shown in Figure 2. Output voltage vector is synthesized by two non-zero voltage vectors and one zero voltage vector, its expression is $V_{oL} = \frac{T_\alpha}{T_s} V_\alpha + \frac{T_\beta}{T_s} V_\beta + \frac{T_0}{T_s} V_0$, so duty ratio of effective vector V_α, V_β and zero vector V_0 can be calculated according to sine theorem and space vector modulation principle as follows [11]:

$$\begin{cases} d_\alpha = \frac{T_\alpha}{T_s} = m_u \sin(60^\circ - \theta_j) \\ d_\beta = \frac{T_\beta}{T_s} = m_u \sin \theta_j \\ d_0 = 1 - d_\alpha - d_\beta \end{cases}. \tag{6}$$

and $\theta_j = \theta_0 + k\omega_0 T_s$, $m_u = U_{im} / U_{dc}$, m_u is voltage modulation coefficient, T_s is switch cycle.

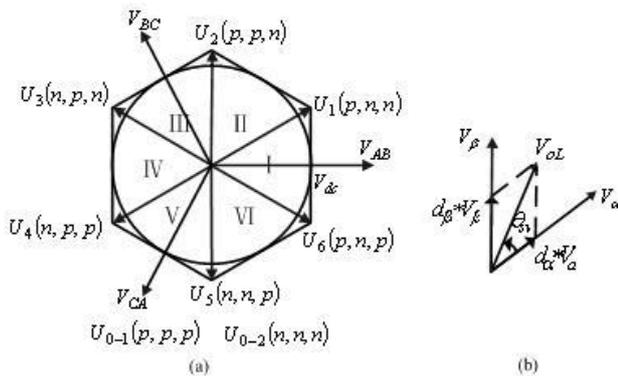


FIGURE2 Output voltage vector distribution and the reference vector synthesis

For rectifier part, its space vector modulation is similar to inverter part, input current vector is synthesized by two non-zero current vectors and one zero current vector, its expression is $I_i = \frac{T_\mu}{T_s} I_\mu + \frac{T_\nu}{T_s} I_\nu + \frac{T_0}{T_s} I_0$, so duty ratio of effective vector I_μ , I_ν and zero vector I_0 can be calculated according to sine theorem and space vector modulation principle as follows:

$$\begin{cases} d_\mu = \frac{T_\mu}{T_s} = m_c \sin(60^\circ - \theta_k) \\ d_\nu = \frac{T_\nu}{T_s} = m_c \sin \theta_k \\ d_0 = 1 - d_\mu - d_\nu \end{cases} \quad (7)$$

and $\theta_k = \theta_0 + k\omega_i T_s$, $m_c = I_{im} / I_{dc}$, m_c is current modulation coefficient, T_s is switch cycle.

AC-AC converter control law is gotten by synthesizing input current vector and output voltage vector. Four basic voltage and current vectors V_α , V_β , I_μ , I_ν , zero voltage vector and their action time are shown in Equations (8)-(12) then sent to the gate terminal of MC in the form of pulse signal, and modulate required amplitude and frequency.

In T_α time, duty ratio of switch state of two adjacent output voltage vectors are:

$$d_{\alpha\mu} = d_\mu d_\alpha = m \sin(60^\circ - \theta_k) \sin(60^\circ - \theta_j), \quad (8)$$

$$d_{\beta\mu} = d_\mu d_\beta = m \sin(60^\circ - \theta_k) \sin \theta_j. \quad (9)$$

In T_β time, duty ratios of switch state of two adjacent output voltage vectors are:

$$d_{\alpha\nu} = d_\nu d_\alpha = m \sin \theta_k \sin(60^\circ - \theta_j), \quad (10)$$

$$d_{\beta\nu} = d_\nu d_\beta = m \sin \theta_k \sin \theta_j. \quad (11)$$

The duty ratio of zero switch vectors is

$$d_0 = 1 - d_{\alpha\mu} - d_{\beta\mu} - d_{\alpha\nu} - d_{\beta\nu}. \quad (12)$$

The final switch state is obtained through synthesizing Equations (4)-(7) and input modulating switch of input current vectors.

3 The fusion of matrix converter and direct torque control structure

In order to improve performance of speed response and ensure higher input power factor, space vector modulation strategy and DTC constitute a new control strategy. The control principle diagram is shown in Figure 3 [12].

In Figure 3, switch time calculating unit determines working time of switch combination of output voltage vector of MC. PWM unit gives signal to control switch state. Switch converter unit provides safe conversion. Direct torque control link is composed of torque controller, flux controller, torque and flux observers, which control and observe torque and flux of induction motor. PI regulator realizes speed response of induction motor. But it cannot adjust quickly when loads and parameters of induction motor change. So in this paper, ADRC is used instead of PI regulator.

ADRC equations of induction motor are described as follows:

Differentiation-tracker

$$\begin{aligned} \varepsilon_0 &= z_{11} - v \\ \dot{z}_{11} &= -r f_{al}(\varepsilon_0, \alpha_0, \delta_0) \end{aligned} \quad (13)$$

Extended state observer

$$\begin{aligned} \varepsilon &= z_{21} - y \\ \dot{z}_{21} &= z_{22} - \beta_{01} f_{al}(\varepsilon, \alpha, \delta) + bu(t) \\ \dot{z}_{22} &= -\beta_{02} f_{al}(\varepsilon, \alpha, \delta) \end{aligned} \quad (14)$$

Nonlinear state error feedback

$$\begin{aligned} \varepsilon_1 &= z_{11} - z_{21} \\ u_0 &= \beta_1 f_{al}(\varepsilon_1, \alpha_1, \delta_1) \end{aligned} \quad (15)$$

$$u(t) = u_0(t) - \frac{z_{22}}{b}$$

The expression of optimal control function f_{al} is

$$f_{al}(\varepsilon, \alpha, \delta) = \begin{cases} |\varepsilon|^\alpha \operatorname{sgn}(\varepsilon) & |\varepsilon| > \delta \\ \frac{\varepsilon}{\delta^{1-\alpha}} & |\varepsilon| \leq \delta \end{cases}, \quad (16)$$

where v is a given signal for ADRC; z_{11} is tracking signal to v ; r is tracking speed factor; y is system output; z_{21} is tracking signal to y ; z_{22} is tracking signal to disturbance signal $\omega(t)$; ε is error signal; α is nonlinear factor; δ is ESO filter factor; β_{01} , β_{02} are correction gain for output error; β is gain error.

The structure of ADRC controller is shown in Figure 4:

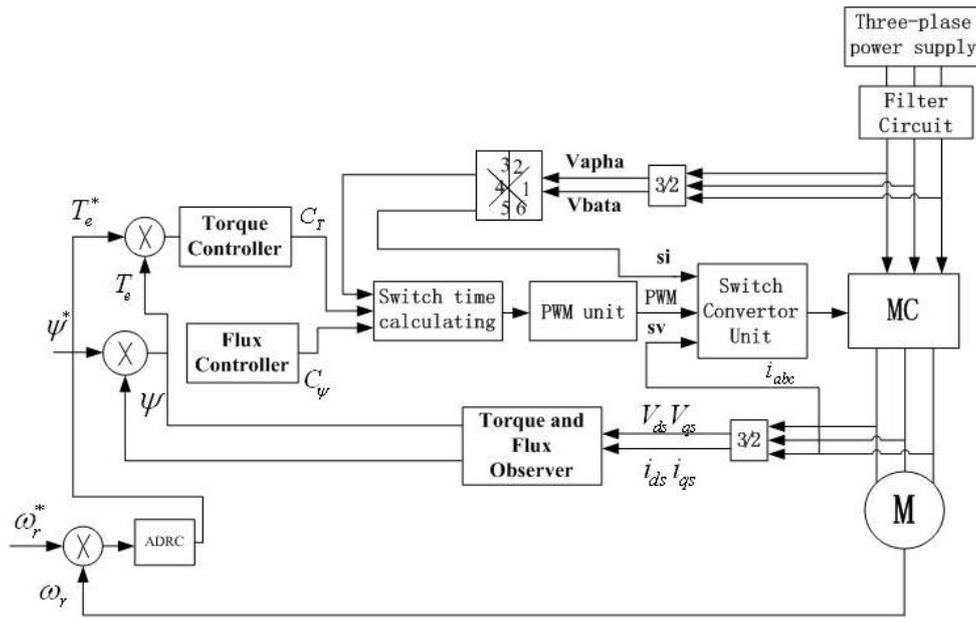


FIGURE 3 Direct torque control system structure based on matrix

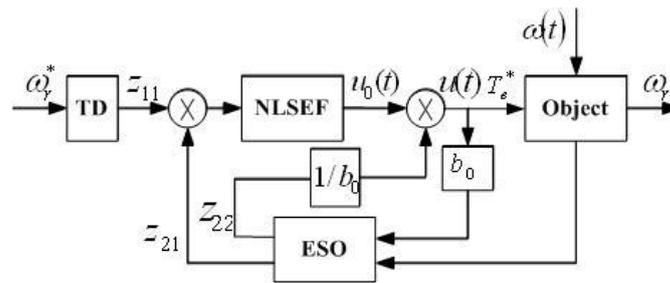


FIGURE 4 Structure of ADRC controller

In Figure 4, ω_r^* and T_e^* are given speed and given torque; ω_r is system speed feedback signal; z_{11} is tracking signal to ω_r^* ; z_{21} is tracking signal to ω_r ; TD arranges transition process for ω_r^* , obtaining smoothing input signal, causing system quick response without overshoot. ESO link estimates all state variables in real time, and observes internal and external disturbances and uncertain model accurately, feedback linearization of dynamic system can be come true to compensate uncertainty of controlled object in the feedback so that we can achieve the goal of refactoring object; NLSEF can realize integrated disturbance compensation and nonlinear control of “small error and great gain” in order to improve system steady-state accuracy.

4 The simulation experiments and analysis

According to the previous analysis, direct torque system model for induction motor can be built based on PI control and ADRC control respectively.

Motor parameters are:

$R_s=0.435\Omega$, $R_r=0.816\Omega$, $L_s=0.02H$, $L_r=0.02H$, $L_m=0.69H$, motor pole $p=2$, load torque $T_g=25N\cdot m$, flux reference $\Psi=0.56Wb$.

ADRC parameters are: $b=1385$, $\beta_{01}=\beta_{02}=4300$, $\beta_1=3$, $\alpha=0.8$, $\delta\delta=0.05$, $\alpha_1=0.8$, $\delta_1=0.04$; PI control parameters are: $K_p=3$, $K_I=0.45$.

Figure 5 is ADRC system steady-state waveform when speed and torque are given; Figure 6 is PI control system and ADRC control system dynamic waveform when load mutates; Figure 7 is PI control and ADRC control steady-state waveform under the disturbance.

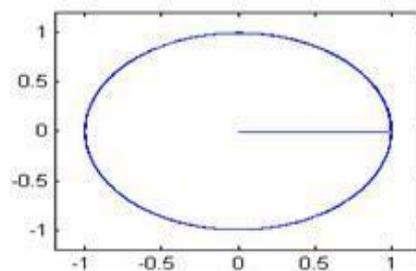


FIGURE 5a Stator flux linkage

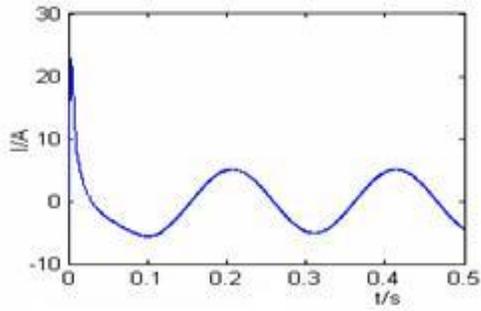


FIGURE 5b Stator current

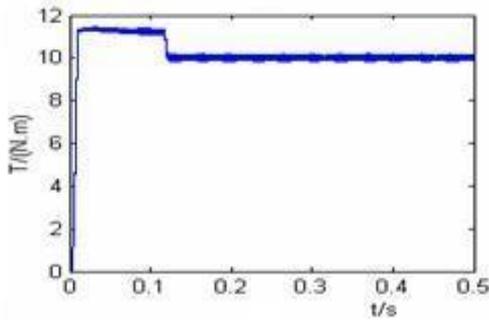


FIGURE 5c Electromagnetic torque

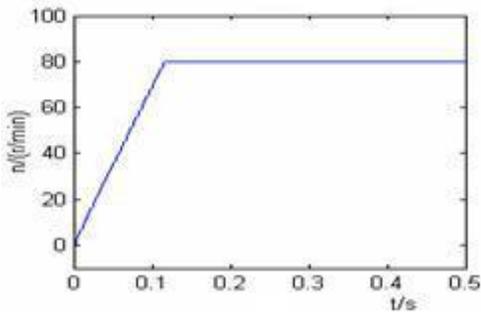


FIGURE 5d Speed

*Figure 5 shows that low performance changes obviously, torque ripple is well improved

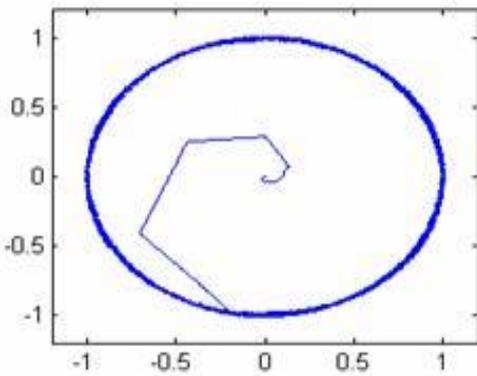


FIGURE 6a Stator flux based on PI controller

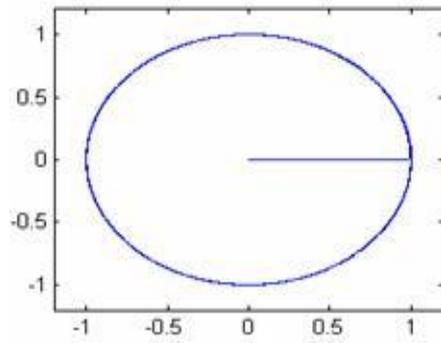


FIGURE 6b Stator flux based on ADRC controller

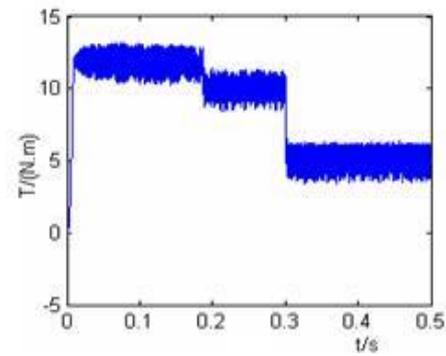


FIGURE 6c Electromagnetic torque based on PI controller

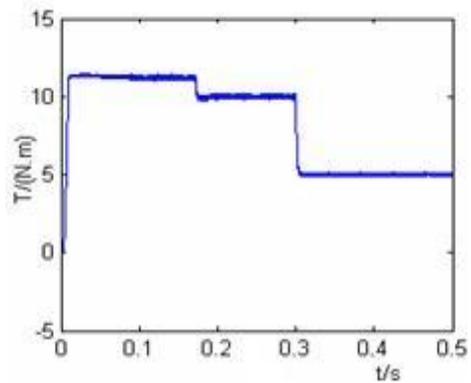


FIGURE 6d Electromagnetic torque based on ADRC controller

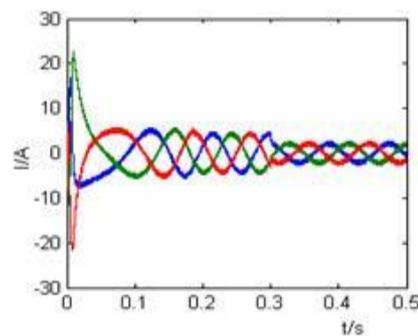


FIGURE 6e Three phase stator currents based on PI controller

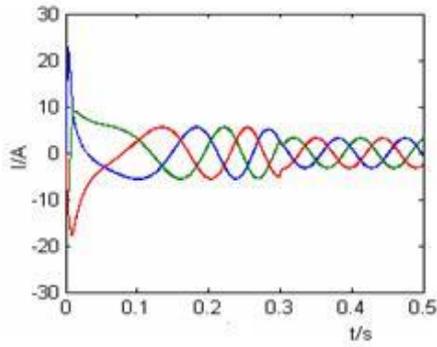


FIGURE 6f Three phase stator currents based on ADRC controller

Figure 6 shows that stator flux is more close to circular, stator flux and stator current disturbance is small, torque ripple is effectively suppressed when the load torque mutates in direct torque control system based on ADRC.

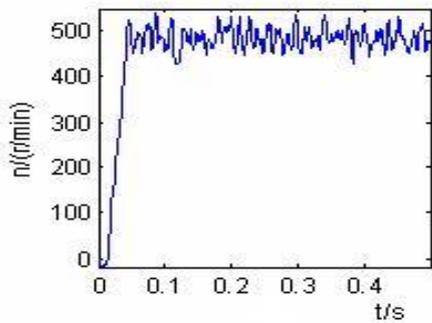


FIGURE 7a Speed based on PI controller in interference environment

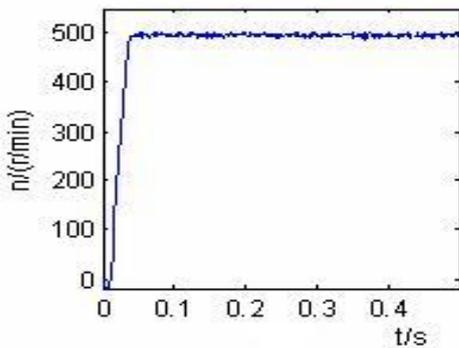


FIGURE 7b Speed based on ADRC controller in interference environment

5 Conclusion

This paper has adopted ADRC algorithm for induction motor based on MC. It has shown that ADRC control strategy is independent of system model and external disturbance. The performance of ADRC controller and

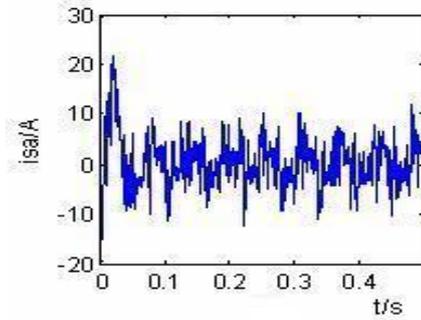


FIGURE 7c Stator current based on PI controller in interference environment

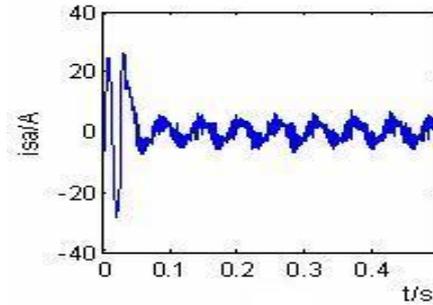


FIGURE 7d Stator current based on ADRC controller in interference environment

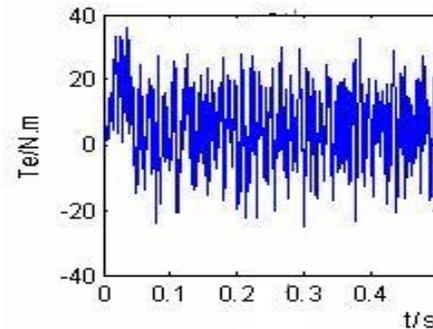


FIGURE 7e Electromagnetic torque based on PI controller in interference environment

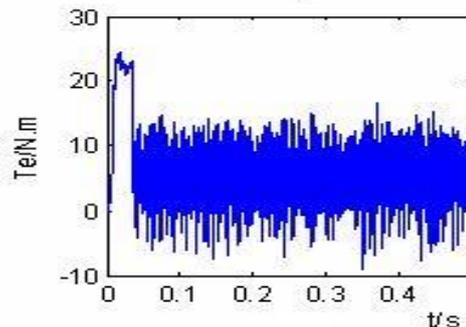


FIGURE 7f Electromagnetic torque based on PI controller in interference environment

conventional PI controller are compared under the same situation. Simulation results show that the ADRC controller has good dynamic and static characteristics, while the traditional PI controller is hard to guarantee high precision and high disturbance rejection ability with the existence of disturbance.

References

- [1] Lai Y S, Chen J H 2001 *IEEE Transactions on Energy Conversion* **16**(3) 220-7
- [2] Lascu C, Boldea I, Blaabjerg F 2000 *IEEE Transactions on Industry Applications* **36**(1) 122-30
- [3] Mir S A, Elbulukand M E, Zinger S 1994 *IEEE Transactions on Industry Application* **30**(3) 729-35
- [4] Romeral L, Arias A, Aldabas E, Marcel J 2003 *IEEE Transactions on Industry Electronics* **50**(3) 487-92
- [5] Lee H H, Nguyen H M, Chun T W 2008 *Journal of power electronics* **8**(2) 74-80
- [6] Zhuang X, Faz Rahman M 2007 *IEEE Transactions on Power Electronics* **22**(6) 2487-96
- [7] Profumo F, Pastorelli M, Tolbert L M 1992 *IEEE Transactions on Industry Application* **28**(5) 1045-53
- [8] Takahashi.I, Ohmori.Y 1989 *IEEE Transactions on Industry Applications* **25**(2) 257-64
- [9] Huber L, Borojevic D 1995 *IEEE Transactions on Industry Applications* **31**(6) 1234-46
- [10] Casadei D, Serra G, Tani A 1998 *IEEE Transactions on Industrial Electronics* **45**(3) 401-11
- [11] Casadei D, Giovanni S 2002 *IEEE Trans on Industry Electronics* **49**(2) 370-81
- [12] Casadei D, Serra G, Tani A 2001 **48**(6) 1057-64

| Authors | |
|--|---|
|  | <p>Junmei Zhao, born in June, 1979, Taiyuan City, Shanxi Province, China</p> <p>Current position, grades: the instructor of School of Computer and Control Engineering, North University of China, China. University studies: B.Sc. in Electrical engineering and automation from North University of China, M.Sc. from North University of China in China. Scientific interest: nonlinear control theory, complex nonlinear control system. Publications: more than 15 papers. Experience: teaching experience of 8 years, 3 scientific research projects.</p> |
|  | <p>Zhijie Zhang, born in March, 1965, Taiyuan City, Shanxi Province, China</p> <p>Current position, grades: the professor of School of Instrument and Electronics, North University of China, China. University studies: B.Sc. in Automation Instrument from Tianjin University, M.Sc. from North University of China in China. Scientific interest: the modern test theory and technology, dynamic testing and intelligent instruments. Publications: more than 30 papers. Experience: teaching experience of 28 years, 20 scientific research projects.</p> |
|  | <p>Yifeng Ren, born in August, 1968, Taiyuan City, Shanxi Province, China</p> <p>Current position, grades: the professor of School of Computer and Control Engineering, North University of China, China. University studies: B.Sc. in physics from Nankai University, M.Sc. from North University of China in China. Scientific interest: complex nonlinear control system, computer control Publications: more than 30 papers Experience: has teaching experience of 25 years, 15 scientific research projects.</p> |

Research on the anisotropy of the coal rock under different bedding direction

Xiaohui Liu^{1, 2*}, Feng Dai¹, Jianfeng Liu¹

¹College of Water Resources and Hydropower, Sichuan University, Chengdu, Sichuan, China, 610065

²College of Energy and Environment, Xihua University, Chengdu, Sichuan, China, 610039

Received 6 June 2014, www.tsi.lv

Abstract

The Fu Rong mining area was selected to analyse the micro view characteristics of the coal, the ultrasonic acoustic characteristics and the uniaxial compression feature from different directions (the parallel direction and the vertical direction). The results show that: (1) Coal rock has large discreteness with strong anisotropic properties. (2) The impulse wave velocities have obvious anisotropic characteristics. The parallel and vertical wave velocities are different. The parallel bedding velocity is greater than the vertical of coal rock no matter the longitudinal wave or transverse wave of coal rock. (3) The uniaxial compressive strength of parallel bedding coal rock is less than the vertical bedding of coal rock. The uniaxial compressive strength is normally distributed vertical wave velocity, obeying exponential functions or power functions. (4) Failure pattern of the coal rock in the parallel bedding direction is splitting, while the vertical bedding direction of coal rock is shearing. The uniaxial compression strength and deformation parameters in two directions are obviously different. In other words, the anisotropic is apparent.

Keywords: coal rock, bedding, anisotropy, ultrasonic velocity, uniaxial compression test

1 Introduction

Coal rock is a kind of micro heterogeneity, which has original injury. The different direction of bedding within the coal rock will heterogeneity and these properties will have a certain effect on the mechanical properties. Sui Wang-hua found that there is an obvious anisotropic with mechanical properties of the rock through variance analysis for the engineering geological property index of the rock [1]. Wang Yun found that there is a linear correlation between the velocity and density by ultrasonic measuring six coal samples collected from different regions [2]. Zhao Qun found that the coal has an obvious anisotropy when testing the ultrasonic speed and decay using the pulsed transmission technology [3]. Wu Ji-wen tested the velocity and the tensile in the coal using wave velocity determination of tensile strength of coal seam and found there is anisotropy within the coal rock. So the ultrasonic acoustic can be selected for measuring the anisotropic characteristics of the coal rock. The coal rock ultrasonic wave velocity anisotropy was used to reflect the anisotropic feature of the coal structure and for further exploration of coal structure [4]. Yin Guang-zhi made a dynamic CT test on the micro damage evolution law in the process of coal rock failure and drew a conclusion that the coal rock has a constitutive relationship [5]. Liu Bao-xian calculated a coal rock damage evolution curve equation after uniaxial compression of coal rock's acoustic emission

characteristics analysis has conducted [6]. Lai Xing-ping analysed the relationship with acoustic emission energy on the stage of coal fracture [7]. So far, there are many researches about the energy and strength change of uniaxial compression destruction of the coal. However, the mechanical properties in different direction of coal are still less considered. Therefore, it is very important to make a further research in different directions of the coal rock. The two different directions (Parallel and perpendicular) were selected to make the ultrasonic acoustic test, the uniaxial compressive strength test and the deformation analysis. These researches will clarify the mechanical properties under different directions of the coal rock and will provide a theoretical basis and the parameters for coal mine disaster prevention and safe and efficient exploitation.

2 Microscopic parameter of coal

The anthracite from Furong Baijiao coal mine was selected as the experimental sample in this research; three coal samples were randomly selected for analysis. The X-ray diffraction, electron microscope scanning and the X-ray fluorescence test were selected to determine the microscopic parameter of coal. The results show that Furong Baijiao coal mineral contents are quartz (11.74%), Kaolinite (5.58%), Calcite (5.04%) and other amorphous mineral (77.64%).

* Corresponding author e-mail: lxh_1001@tom.com

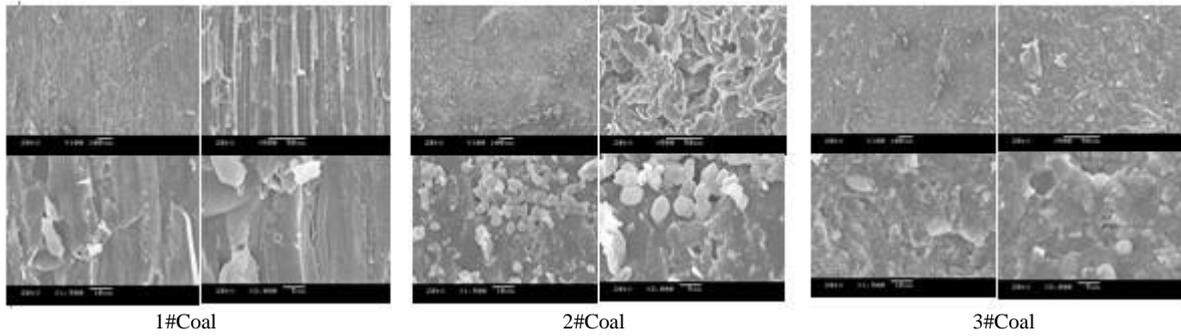


FIGURE 1 Coal mineral scanning electron microscope

The coal rocks were magnified with 100, 500, 1500 and 3000 times by using scanning electron microscopy (shown in Figure 1). We find the coal rocks are porous materials and the pores are rich which mainly include the original hole, the exogenous hole, the metamorphic hole and the mineral hole [8]. There are plurality groups of micro cracks in the coal rock which are almost developed in parallel and perpendicular to the surface [9].

3 Ultrasonic testing

3.1 THE BASIC PHYSICAL PROPERTIES OF COAL

According to "Method for determination the physical and mechanical properties of coal rock (GB/T 23561.7-2009), the Fu Rong mining area which was made into a cylindrical according to vertical and parallel bedding

direction, the diameter is 50mm and the height is 100 mm. The non-parallelism of the two end surface is less than 0.05 mm and perpendicular to the sample axis. The maximum deviation should be less than 0.25°, the samples are numbered according to the different directions after the samples are processed (Figure 2).

The basic coal parameters are measured (Table 1), the data show that the density of rock is different ($\rho_{max}=1541.94 \text{ kg/m}^3$). We find that the value (Ratio between the coal rock specimen density and the largest coal rock specimen density) is between 0.93 and 1; we conclude this distribution feature may be with respect to the origin, component and the internal micro cracks of coal. This conclusion may help us to definite the discrete initial grouping on the basis of different density.

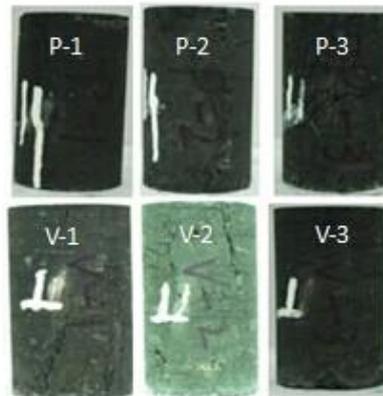


FIGURE 2 Pictures of coal rock

TABLE 1 The basic parameters of coal rock

| Coal No. | Diameter/mm | Height/mm | Volume/ cm^3 | Weight/g | Density kg/m^3 | ρ/ρ_{max} |
|------------|-------------|-----------|-----------------------|----------|-------------------------|-------------------|
| P-1 | 50.33 | 81.86 | 162.78 | 233.50 | 1434.47 | 0.93 |
| P-2 | 50.07 | 100.18 | 197.15 | 304.00 | <u>1541.94</u> | 1.00 |
| P-3 | 50.43 | 90.19 | 180.06 | 258.54 | 1435.89 | 0.93 |
| Mean value | 50.28 | 90.74 | 180.00 | 265.35 | 1470.77 | 0.95 |
| V-1 | 50.06 | 101.93 | 200.52 | 293.11 | 1461.76 | 0.95 |
| V-2 | 49.86 | 100.32 | 195.78 | 291.46 | 1488.73 | 0.97 |
| V-3 | 49.72 | 100.09 | 194.23 | 293.77 | <u>1512.46</u> | 0.98 |
| Mean value | 49.88 | 100.78 | 196.84 | 292.78 | 1487.65 | 0.96 |

3.2 DYNAMIC MECHANICAL ULTRASONIC TESTING MODEL

The coal physical mechanics index is essential basis for the mining and coal sample design. The elastic parameter method and the acoustic method are selected to test the anisotropy of mechanical properties of the rock [10]. Considering the acoustic velocity can reflect the strength and structure of the coal [11], the coal acoustic characteristics was used to reflect the feature of the coal and to seek what caused the external differences [12]. The ultrasonic testing system consists of three system (The fluorescence oscilloscope (TDS3014), the ultrasonic pulse receiver (5077PR) and the data acquisition instrument (3499B). The centre frequency of emission and receiving transducer is 50 kHz. The system can complete testing and

data recording combined with the MTS rock mechanics test system. Equation of dynamic mechanical ultrasonic parameters can be expressed as follows:

$$E_d = v_{mp}^2 \rho \frac{(1 + \mu_d)(1 - 2\mu_d)}{1 - \mu_d} = 2v_{ms}^2 \rho(1 + \mu_d)$$

$$\mu_d = \frac{v_{mp}^2 - 2v_{ms}^2}{2(v_{mp}^2 - v_{ms}^2)}$$

$$G_d = \frac{E_d}{2(1 + \mu_d)} v_{ms}^2 \rho$$

The calculation results are shown in Table 2.

TABLE 2 Physical and mechanical parameters of coal rock

| Coal sample No. | Density kg/m^3 | P-wave velocity m/s | Shear velocity m/s | Wave velocity ratio | Poisson's ratio ν | Dynamic elastic modulus E/MPa | Shear modulus G/MPa | Bulk modulus K/MPa |
|-----------------|------------------|-----------------------|----------------------|---------------------|-----------------------|---------------------------------|-----------------------|----------------------|
| P-1 | 1434.471 | 899.450 | 458.210 | 1.963 | 0.325 | 797.973 | 590.881 | 758.933 |
| P-2 | 1541.941 | 958.720 | 527.570 | 1.817 | 0.283 | 1101.101 | 767.672 | 845.041 |
| P-3 | 1435.892 | <u>1274.100</u> | <u>844.930</u> | 1.508 | 0.107 | 2270.571 | 1272.022 | 964.137 |
| Mean value | 1470.768 | 1044.090 | 610.237 | 1.763 | 0.238 | 1389.882 | 876.858 | 856.037 |
| V-1 | 1461.764 | 817.280 | 485.120 | 1.685 | 0.228 | 844.894 | 547.208 | 517.695 |
| V-2 | 1488.734 | 804.370 | 393.320 | 2.045 | 0.343 | 618.554 | 470.657 | 656.150 |
| V-3 | 1512.463 | <u>864.860</u> | <u>503.280</u> | 1.718 | 0.244 | 953.129 | 630.369 | 620.506 |
| Mean value | 1487.654 | 828.837 | 460.573 | 1.816 | 0.272 | 805.526 | 549.411 | 598.117 |
| Total | 1478.005 | 951.839 | 546.095 | 1.786 | 0.253 | 1139.443 | 736.524 | 745.500 |

3.3 EXPERIMENTAL RESULTS AND ANALYSIS

Figure 3 is the longitudinal wave velocity distribution diagram. The value V_p/V_{Pmax} is the ratio between the coal wave velocity and the sample coal maximum P-wave velocity. From the Figure 3a we can see that the coal wave P-wav velocity has two obvious segmented regions according to the different layering surface. The wave velocity with parallel layer coal distributes in the range from 890m/s to 1300m/s and the wave velocity from the parallel direction is greater than the wave velocity from the perpendicular direction. From Figure 3b, we can find that

there is a certain distribution relationship between the density and the wave velocity of the coal. We find that the wave velocity from the longitudinal wave has a discrete phenomenon and the coal density from the longitudinal appears a less discrete phenomenon. The coal density from parallel layer has a large dispersion and the wave velocity from the longitudinal direction also has a large difference at the similar density. We can draw a conclusion that this phenomenon may due to the presence and the hole exists in the coal. The results also show that the large differences with the wave velocity from the longitudinal direction may attribute to the different directivity.

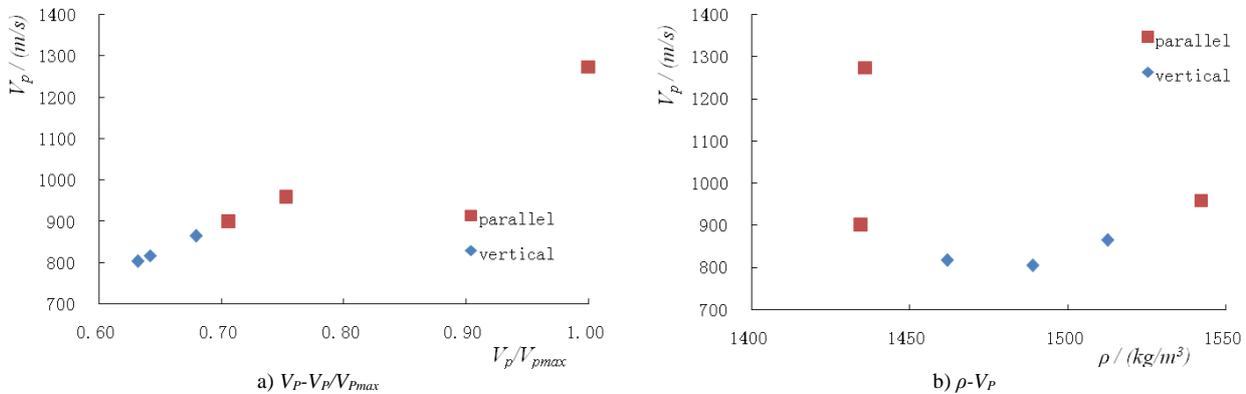


FIGURE 3 The distribution diagram of longitudinal wave velocity for coal rock

Figure 4 is the distribution relationship of transverse wave velocity. The value $V_s/V_{s,max}$ is the ratio between the transverse wave velocity and the sample coal maximum transverse wave velocity. From Figure 4a we find that the shear wave has a larger discreteness when it crosses the bedding surface.

The transverse wave velocity of the parallel bedding coal distributes from 450m/s to 850m/s, the transverse wave velocity of the vertical bedding coal distributes from 390 to 510m/s. From the above we can see that the

transverse wave velocity of the coal from the vertical direction remains larger than the transverse wave velocity from the parallel direction. The Figure 4b is the distribution relationship between the density and the transverse wave velocity of the coal and we find that the Figure 3b has a similarly changing regulation with Figure 3a and Figure 3b. The results above sufficiently prove that there is a certain relationship between the ultrasonic wave velocity, density and the micro crack distribution of the coal rock.

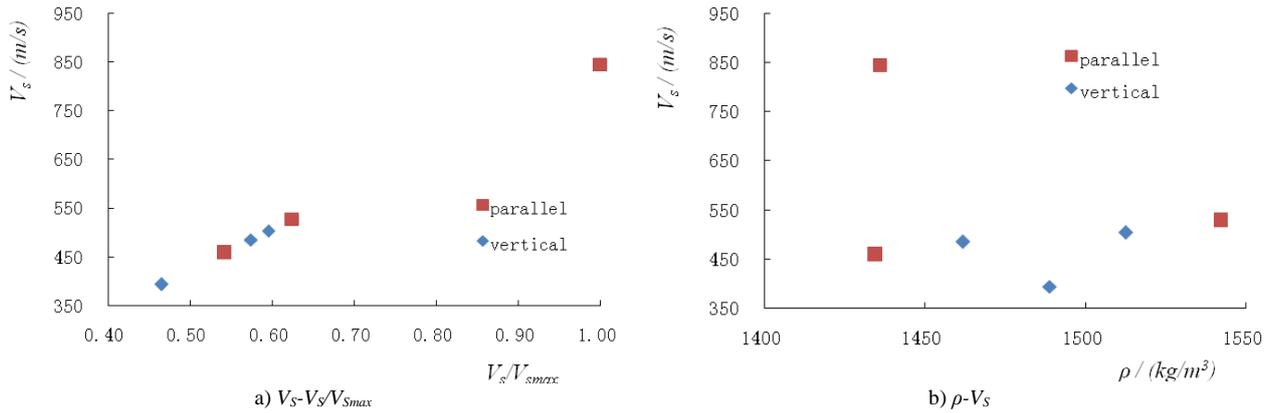


FIGURE 4 The distribution diagram of shear wave velocity for coal rock

4 The uniaxial compression experiment of coal rock

4.1 EXPERIMENTAL METHOD

The rock mechanics test system (MTS815 Flex Text GT) is selected in this research which is from rock mechanics laboratory, Si Chuan University. During the experiment, when the data reaches 10kN/min, the system has loaded and turns to circumferential control with 0.08mm/min when the data reaches 7kN.

The samples are divided into two groups, namely the parallel directions and the vertical directions. There are three samples in each group and each sample records the destroyed morphology themselves. The aspect ratio (L/D) of the coal has a great influence on the experimental results. Equation (2) is revised due to the coal P-1 and P-3 is nonstandard sample:

$$\sigma_c = \frac{\sigma_c'}{0.788 + 0.22 \frac{D}{L}}, \tag{2}$$

where σ_c is actual rock uniaxial compressive strength, σ_c' is measured rock uniaxial compressive strength.

4.2 RESULTS

Coal rock under uniaxial compression results and the statistical parameters are shown in Table 3.

The Uniaxial compression strength test and the deformation of coal are shown in the Figure 5 and Figure 6 respectively. The axial strain of coal, the volumetric strain, the transverse strain and the failure modes graph are shown in Figures 5 and 6 (a-d) respectively.

TABLE 3 Test results of coal samples under uniaxial compression

| Coal No. | Peak load /kN | Compressive strength /MPa | Modulus of elasticity /MPa | Deformation modulus /MPa | Poisson's ratio (50%) | Poisson's ratio (100%) |
|------------|---------------|---------------------------|----------------------------|--------------------------|-----------------------|------------------------|
| P-1 | 11.95 | 6.571 | 1914 | 1011.940 | 0.04 | 0.54 |
| P-2 | 17.68 | 8.986 | 1201 | 1006.311 | 0.05 | 0.82 |
| P-3 | 23.12 | 12.596 | 1710 | 975.574 | 0.05 | 0.37 |
| Mean value | 17.58 | 9.384 | 1608.33 | 997.94 | 0.05 | 0.58 |
| V-1 | 28.76 | 14.618 | 1761 | 1062.915 | 0.14 | 0.39 |
| V-2 | 18.70 | 9.582 | 1413 | 1012.894 | 0.13 | 0.33 |
| V-3 | 30.32 | 15.624 | 1921 | 1349.610 | 0.09 | 0.17 |
| Mean value | 25.93 | 13.275 | 1698.33 | 1141.81 | 0.12 | 0.30 |

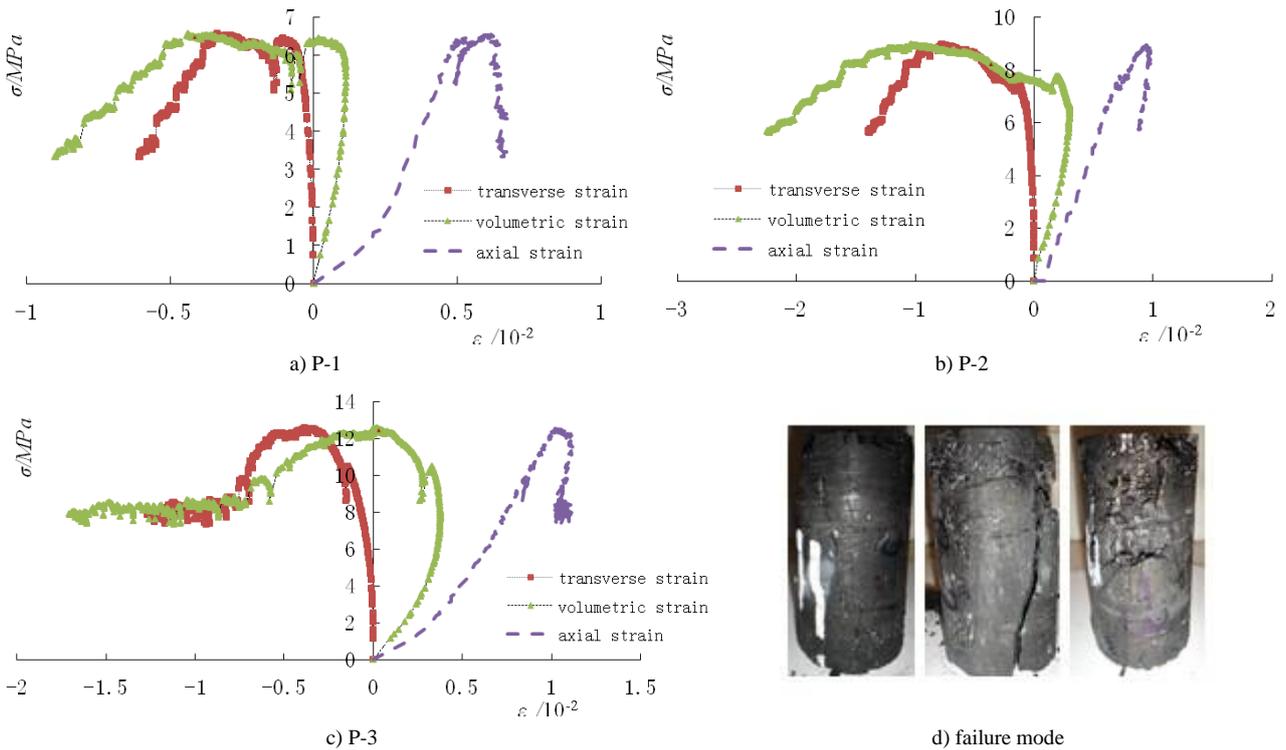


FIGURE 5 Test results of coal rock with parallel bedding under uniaxial compression

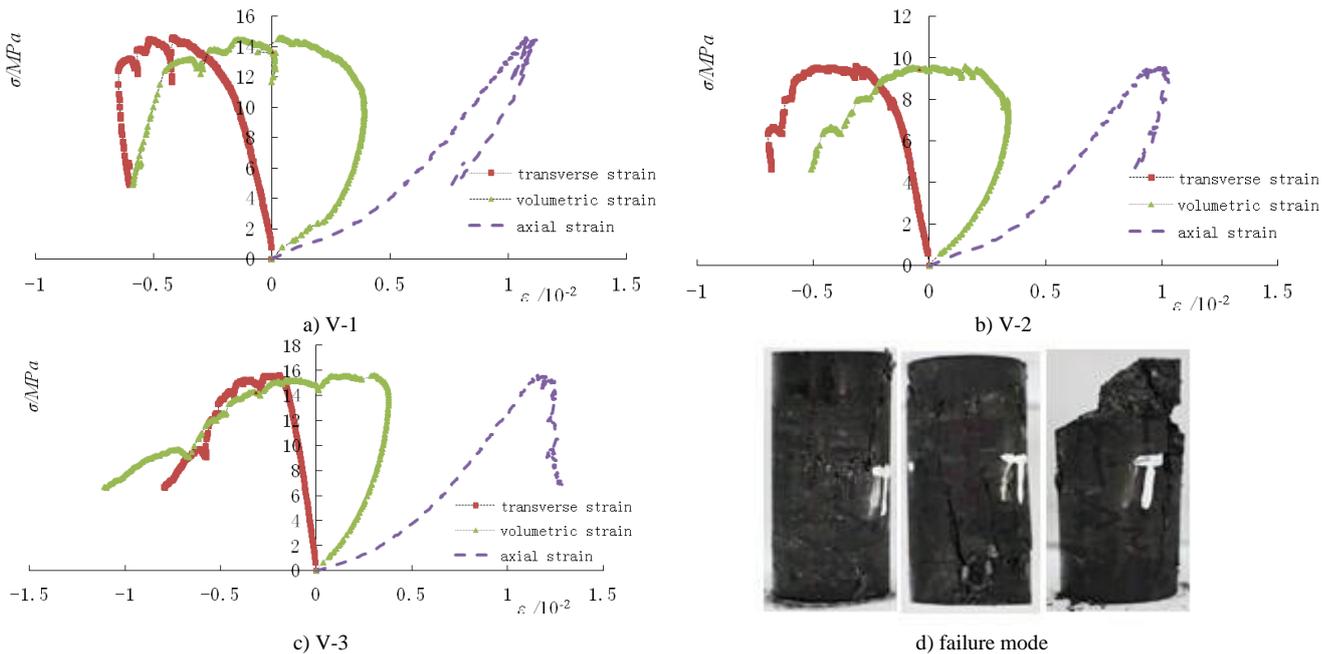


FIGURE 6 Test results of coal rock with vertical bedding under uniaxial compression

4.3 EXPERIMENTAL RESULTS AND ANALYSIS

The corresponding situations between the acoustics and the uniaxial compression properties are analysed from two different directions of coal (the vertical and parallel bedding direction).

1) The uniaxial compressive strength of the coal from parallel bedding is 6.5~12.6MPa and the mean value is

9.38MPa. The uniaxial compressive strength of the coal from vertical direction is 9.5~15.7MPa and the average value is 13.27MPa. We find that the compressive strength from the parallel bedding is less than that from the direction of vertical bedding.

2) From Figure 5d and Figure 6d we can see that the uniaxial compression failure mode from the parallel

direction is splitting and forms a sheet damage phenomenon.

3) Figure 7 represents the uniaxial compressive axial stress and the strain curve at the different directions. We can find that the compression test of the coal has experienced four stages: compacting, elastic, yield and destruction [13]. There is a great difference between stress and strain curves of coal rock at different directions:

The pre-peak deformation with parallel bedding coal has a great difference; the P-1 presents the yield deformation in the earlier time. However, the yielding time of the other two samples presents a time lag; the stress drop phenomenon will appear after reaching the peak value and have a different extent. The pre-peak deformation with vertical bedding coal is similar and presents stable mechanical properties.

4) We make a comparison between Table 3 and Figure 7 and get some conclusions: The modulus of elasticity in

the parallel bedding coal is 1200 ~ 1914MPa and the average value is 1608.33MPa; The deformation modulus is 975 ~ 1012MPa and the average value is 997.94MPa. The Poisson's ratio is 0.17 ~ 0.39 and the mean value is 0.30. The research above demonstrates that the in-depth analysis can be made for anisotropic properties of coal by using modulus of elasticity, deformation modulus and Poisson's ratio.

5) The correlation analysis is made for both compressional wave velocity and uniaxial compressive strength of coal [14]. The Figure 8 is the relationship curve between the ultrasonic P-wave velocity and uniaxial compressive strength within the coal rock and we find that the uniaxial compressive strength σ_c will enlarge along with the increasing p-wave velocity V_p [15].

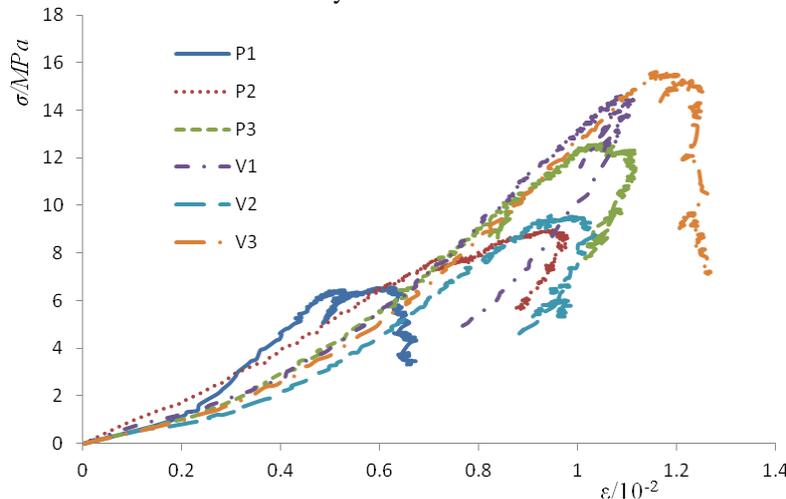


FIGURE 7 Axial stress-strain curves of coal rock in different directions to the bedding plane under uniaxial compression

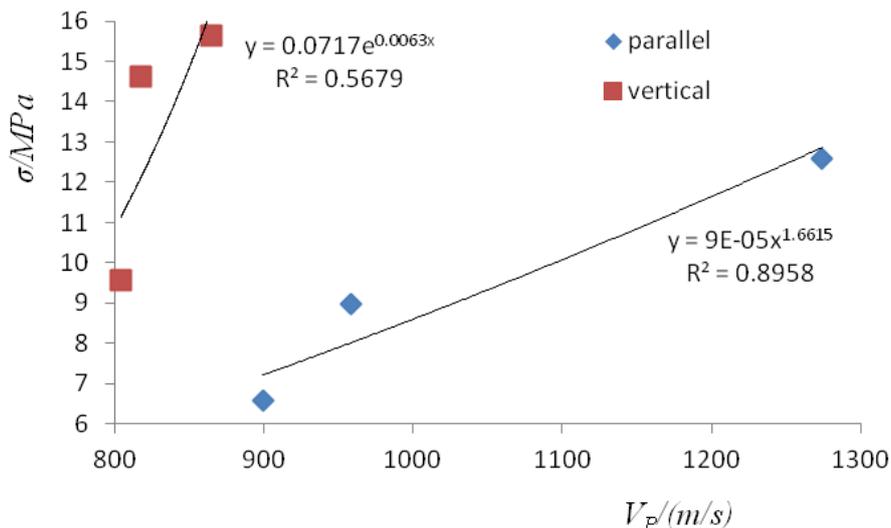


FIGURE 8 The relationship curve between ultrasonic velocity and compression strength

5 Conclusions

- 1) The coal rock is a porous materials, dispersion phenomenon in the coal may be caused by the micro cracks and the laminar distribution.
- 2) The longitudinal and transverse wave velocity are different, the transverse wave velocity from the parallel bedding direction is greater than that from the vertical direction.
- 3) The uniaxial compressive strength from the parallel bedding direction is less than that from the vertical direction and the compressive strength of the coal has a

normal distribution with the compressional wave velocity. 4) Failure pattern of the coal rock in the parallel bedding direction is splitting, while the vertical bedding direction of coal rock is shearing. There are obvious differences with the uniaxial compression and the deformation parameters.

Acknowledgements

The study was supported by the national key basic research development plan (937) funded project (2011CB201201, 2010CB226802, 2010CB732005)

References

- [1] Sui W-H, Zhou S-W, Jiang Z-Q 1998 The variability of indices of engineering geologic properties in coal-measures rocks *Coal Geology & Exploration* **26**(5) 45-48 (in Chinese)
- [2] Wang Y, Xu X-K, Zhang Y-G 2012 Characteristics of P-wave and S-wave velocities and their relationships with density of six metamorphic kinds of coals *Chinese Journal of Geophysics* **55**(11) 3754-61 (in Chinese)
- [3] Zhao Q, Hao S-L 2006 Anisotropy test instance of ultrasonic velocity and attenuation of coal sample *Progress in Geophysics* **21**(2) 531-4 (in Chinese)
- [4] Wu J-W, Jiang Z-Q, Fan C 2005 Study on tensile strength of coal seam by wave velocity *Chinese Journal of Geotechnical Engineering* **27**(9) 999-1003 (in Chinese)
- [5] Yin G-Z, Dai G-F, Pi W-L 2003 CT Real-time analysis of damage evolution of coal under uniaxial compression *Journal of Chongqing University (Natural Science Edition)* **26**(6) 96-100 (in Chinese)
- [6] Liu B-X, Huang J-L, Wang Z-Y 2009 Study on Damage Evolution and Acoustic Emission Character of Coal-rock Under Uniaxial Compression *Chinese Journal of Rock Mechanics and Engineering* **28**(z1) 3234-8 (in Chinese)
- [7] Lai X-P, Wu X-M, Gao X-C 2008 Characteristics analysis of coal-rock damage based on MTS-AE uniaxial compression *Journal of Xi'an University of Science and Technology* **28**(2) 375-8 (in Chinese)
- [8] Zhang H 2001 Genetical type of proes in coal reservoir and its research significance *Journal of China Coal Society* **26**(1) 40-4 (in Chinese)
- [9] Yang Y-J 2006 Basic Experimental Study on Characteristics of Strength, Deformation and Micro Seismic Under Compression of Coal *Qingdao, Shandong: Shandong University of Science and Technology* (in Chinese)
- [10] Yan R-G 1982 Measurement of the anisotropic mechanical properties and stress in mine rocks *Journal of China Coal Society* **1**(3) 30-45 (in Chinese)
- [11] Yan Li-Hong 2006 Relationship study between characteristics and strength of coal and rock wave velocity in Yangzhuang Mine *Coal Science and Technology* **34**(6) 57-60 (in Chinese)
- [12] Cheng L, Wang Y, Zhang Y-G 2013 The present situation and prospect of the acoustic properties research in coal *Progress in Geophysics* **28**(1) 452-61 (in Chinese)
- [13] Liu K-D, Liu Q-S, Zhu Y-A 2013 Experimental study of coal considering directivity effect of bedding plane under Brazilian splitting and uniaxial compression *Chinese Journal of Rock Mechanics and Engineering* **32**(2) 308-16 (in Chinese)
- [14] Zhou X-D 1999 Study on velocity of elastic wave in rock& its application *Guizhou Water Power* **13**(2) 10-3 (in Chinese)
- [15] Zhao M-J, Wu D-L 2000 The ultrasonic identification of rock mass classification and rock mass strength prediction *Chinese Journal of Rock Mechanics and Engineering* **19**(1) 89-92 (in Chinese)

| Authors | |
|---|---|
|  | <p>Liu Xiaohui, born on October 1, 1977, Chengdu, China</p> <p>Current position, grades: on-the-job Doctor; the lecturer of Xihua University, Sichuan, Chengdu. University studies: Master degree in Geotechnical Engineering from Sichuan University. Scientific interest: geotechnical engineering, water resources and hydropower engineering. Publications: more than 10 papers. Experience: 8 years teaching experience.</p> |
|  | <p>Dai Feng, born on June 1, 1978, Anhui, China</p> <p>Current position, grades: Sichuan University, Sichuan, China University studies: Dr. Degree in Department of Civil Engineering at Toronto University at 2010. Scientific interest: rock dynamic mechanics. Publications: more than 40 papers.</p> |
|  | <p>Liu Jianfeng, born on August 27, 1979, Henan, China</p> <p>Current position, grades: Sichuan University, Sichuan, China. University studies: Dr. Degree in Geotechnical Engineering. Scientific interest: geotechnical engineering, rock mechanics and engineering. Publications: about 60 papers.</p> |

Research on the laser transmission simulation based on random phase screen in atmospheric turbulent channel

Yanli Feng^{1, 2*}, Dashe Li^{2, 3}, Shue Liu⁴

¹Department of Computer Foundation Studies, Shandong Institute of Business & Technology, China

²Key Laboratory of Intelligent Information Processing in Universities of Shandong (Shandong Institute of Business and Technology), China

³School of Computer Science and Technology, Shandong Institute of Business & Technology, China

⁴School of Computer Science and Technology, Binzhou Medical University, China

Received 30 December 2013, www.tsi.lv

Abstract

On the basis of collimated Gaussian beams, the paper focused on the modelling and simulating of the transmission of laser beams using two-dimension random phase screens in the atmospheric turbulence channel. Firstly, with the analysis of the transmission model of Gaussian beams through the phase screens, the simulation theory of random phase screens and the depth range model of the phase screens were proposed. Then, in accordance with Kolmogorov atmospheric turbulence theories, a two-dimension random phase screen was built using Fourier transform. Numerical simulation experiments were conducted with low frequency compensation to simulate the propagation of Gaussian collimated beam in Kolmogorov turbulence. Finally, the two-dimension random phase screen was testified by the phase structure function. The results showed that the approach of simulating the random phase screen using Fourier transform was appropriate after compensating the low frequency.

Keywords: random phase screen, atmospheric turbulence, Gaussian beam, Fourier transform, Kolmogorov

1 Introduction

Atmospheric turbulence is one of the important factors affecting the beam propagation. Numerical simulation method for beam propagation is an effective way to study the atmospheric turbulence besides experimental and theoretical research. Several numerical simulation methods have been proposed to generate the random phase screen for numerically simulating the atmospheric turbulence [1, 2]. Numerical simulating methods can be basically divided into two categories. The first one was proposed by Mc Glamery, which was indirect simulation of the frequency field using Fourier transform. The other was direct simulation of the spatial domain, which can represent the phase front using an orthogonal complete set of Zernike polynomial [3]. Moreover, Yan put forward a random numerical simulation method of the atmospheric turbulence based on fractals, Wang et al, proposed a simulating model on laser transmission in the atmosphere through any thick random phase screen, and Andrews et al, studied the statistical characteristics of the transmission in thin random phase screen [4].

In order to study effect of atmospheric turbulence on the propagation properties of the laser beam, in the paper, the random phase screen established by the Fourier transform is simulated in compliance with Kolmogorov atmospheric turbulence theories. A new method for laser beam propagation research in atmospheric turbulence is

put forward to overcome the limitations of experimental and theoretical approaches.

2 Simulating theories on random phase screen

As to the collimated Gaussian beams propagation through the atmospheric turbulence, let ω_0 be the beam-waist radius at the input end. And denote beam-waist radius as ω after the beams transmit a distance of Lkm . In this process, if the changes caused by the fluctuation of atmospheric refractivity is sufficiently small, the continuous atmospheric turbulence can be divided into a series of phase screens (sampling grid) with Δ_z per thickness. The collimated Gaussian beams located in the front surface of Z_i screen will be transmitted to the back surface of the screen through the atmospheric turbulence with Δ_z thickness. Then the phase modulation caused by the phase screen in the atmospheric turbulence forms the ultimate optical field distribution E_i . After Field E_i passes through the same atmospheric turbulence and is modulated, it arrives at the back surface of Z_{i+1} . There are three steps to generate the two-dimension phase screens using Fourier transform. First, a matrix with random numerals obeying the Gaussian distribution is generated. Second, the air power spectral function adhering to the Kolmogorov turbulence distribution filters the matrix generated in the first step. Finally, the new filtered

* Corresponding author e-mail: fengyanli@sdibt.edu.cn

complex Gaussian matrix is computed with the inverse Fourier transform to obtain the random phase. In this process, the phase distortion occurring on each phase screen is accumulated on Gaussian optical field E , and the optical fielding after passing through i phase screens is expressed as follows [5, 6]:

$$E_{i+1} = F^{-1}\{F[E_i \times \exp(i\phi)] \times \exp(-is_i)\} \tag{1}$$

In order to ensure that the phase changes caused by each phase screen is sufficiently small and meanwhile the propagation distance L in the atmospheric turbulence can be substituted by calculus of Δz , Thus, the thickness of the two-dimension random phase screen should be infinitely thin so that the generated optical waves will only affect the phase of the Gaussian optical waves while with no obvious influence on the amplitude. Therefore, the following condition must be satisfied.

$$\Delta z \ll \lambda / \sigma_n, \tag{2}$$

where, λ is the wavelength of Gauss beam and σ_n^2 is the average variance of the refraction rates fluctuation [5].

The adjacent phase screens should be mutually independent and meanwhile spatially connected. And the front and back phase screen should have some extent of correlation. Therefore, the depth of the phase screen Δz should exceed the outer scale of the turbulence. That is:

$$\Delta z > L_0, \tag{3}$$

where L_0 is the outer scale of the turbulence [5].

In Fourier transform and inverse Fourier transform, the thickness Δz of a random phase screen replaced the thickness calculus of the whole screen. However, the prerequisite to do so is that the refraction in the phase screen is evenly distributed and the transmission of optical lights follows the principles of geometric optics. Hence, the scale of Fresnel should be smaller than that inside the turbulence. That is:

$$\Delta z < l_0^2 / \lambda, \tag{4}$$

where l_0 is the inner scale of the turbulence.

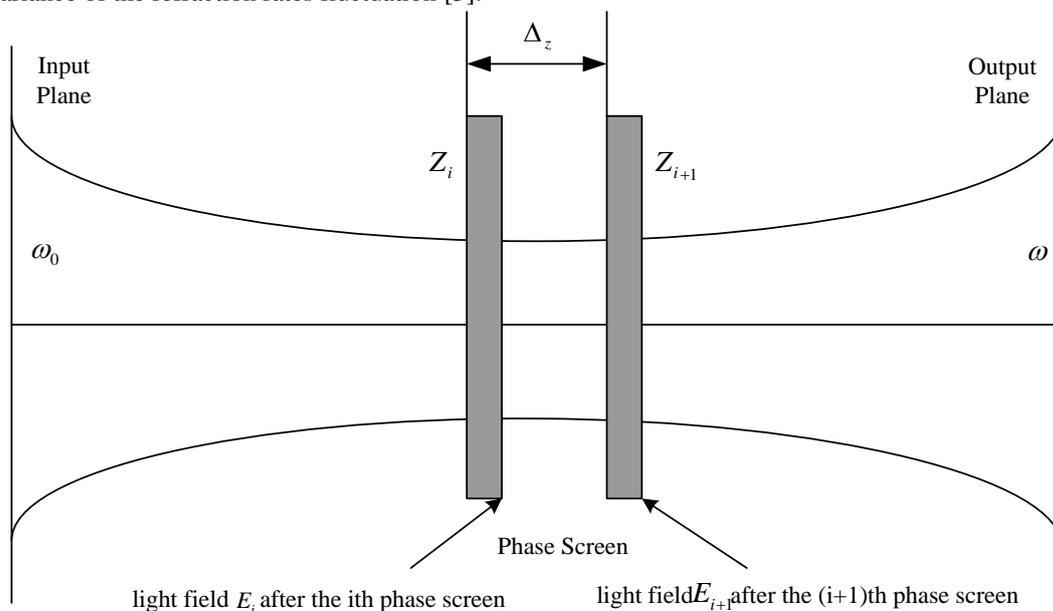


FIGURE 1 Model of the transmission of Gaussian beams through the phase screen

3 The construction of the random phase screens

Along the propagation direction of the Gaussian beams, the atmospheric turbulence $(0, L)$ is uniformly divided into N phase screens. Thus, the thickness of each phase screen is the same as $\Delta z = L / N$. First, using the Fourier transform and the Kolmogorov turbulence distribution theories, each phase screen grid with no spatial correlation is calculated to obtain the complex Gaussian random matrix. Then the modified Von Karman model is used for filtering. Moreover, the random phase is obtained based on the spatially correlated phase screens after inverse Fourier transformation. Finally, the same method is used to generate the next new phase screen with the same spatial

distribution. The newly generated random numbers of the screens are not completely new and thus the new screens have a spatial correlation with both the front and the back screens [7]. The above process can be defined as [8, 9]:

$$X(m\Delta_x, n\Delta_y) = \sum_{m'=-N_x/2}^{N_x/2-1} \sum_{n'=-N_y/2}^{N_y/2-1} [a(m, n) + jb(m, n)] \times \exp[2\pi j(\frac{m'm}{N_x} + \frac{n'n}{N_y})] \tag{5}$$

where N_x and N_y represent the dimensions in the direction of x, y in the matrix, Δ_{xk} and Δ_y are the intervals of the sampling grid in the direction of x, y . $a(m, n)$ and $b(m, n)$ are mutually independent Hermitian

Gaussian random numbers with the mean zero. The variance is [10]:

$$\langle a^2(m,n) \rangle = \langle b^2(m,n) \rangle = \frac{\sqrt{0.00058} r_0^{-5/6}}{\sqrt{G_x G_y}}, \quad (6)$$

$$\times \Delta k_x \Delta k_y F(m\Delta k_x, n\Delta k_y, z) \Phi_n(\Delta k_x, \Delta k_y)$$

where $F(m\Delta k_x, n\Delta k_y, z)$ is the filtering function and can be rewritten as the following equation:

$$F(m\Delta k_x, n\Delta k_y, z) = [(2\pi)^3 / \lambda^2 \times \Delta z \times 0.33 C_n^2 (k^2 + k_0^2)^{-11/6} \exp(-k^2 / k_m^2)]^{1/2}, \quad (7)$$

where $k = 2\pi[1 / (m\Delta k_x)^2 + 1 / (n\Delta k_y)^2]^{-1/2}$, and $\langle \cdot \rangle$ refers to the overall average. $F(\cdot)$ is the spatial filtering function of the phase screen, and it is also the function of the propagation distance z . Δk_x and Δk_y are the grid intervals on the phase screens. G_x and G_y represent the size of the phase screen, r_0 is the atmospheric coherence length and Δz is the depth of the turbulence layer. $\Phi_n(\cdot)$ is the function calculating the refraction power density. Here the power spectral density function derived from the Kolmogorov model is adopted.

If the modified Von Karman power spectral density is substituted for power spectral density function $\Phi_n(\cdot)$. And the plane wave can be given by the following Equation [11]:

$$\Phi_n(\cdot) = 0.49 r_0^{-5/3} \frac{\exp(-k^2 / k_m^2)}{(k^2 + k_0^2)^2}, \quad (8)$$

where $k = 2\pi f$, $k_0 = 2\pi f_0$ and $k_m = 2\pi f_m$.

As seen from above equations, it is easy to use inverse Kolmogorov to construct the two-dimension random phase screens, but the lack of samples on the spatially low frequency part leads to the absence of power spectrum at the low frequency components in this phase screen, thus resulting in a relatively low accuracy of the generated phase screen. Hence low frequency compensation is necessary for improving the accuracy. With the help of Lane's ideas, interpolation merge is conducted based on the re-sampling of the Fourier low frequency subharmonics so as to make low frequency compensation to the subharmonics in this screen. The equation can be rewritten as follows:

$$\Phi_{SH}(m\Delta x, n\Delta y) = \sum_{m'=-1}^1 \sum_{n'=-1}^1 \sum_{p=1}^{N_p} [a(m,n) + jb(m,n)] \times \exp[2\pi j \times 3^{-p} (\frac{m'm}{N_x} + \frac{n'n}{N_y})], \quad (9)$$

where p refers to the subharmonic series.

The inverse transform method is adopted to simulate the Kolmogorov spectrum phase screens under the conditions that the wave length is $1.06 \times 10^{-5}m$, the size of

the phase screen is $4.8m \times 4.8m$, $W_0 = 0.8 \times 10^{-6}m$, the propagation distance is $L_0 = 10km$, the sampling points is $N_x = N_y = 1024$, and the interval between each phase screen is $\Delta z = 500m$.

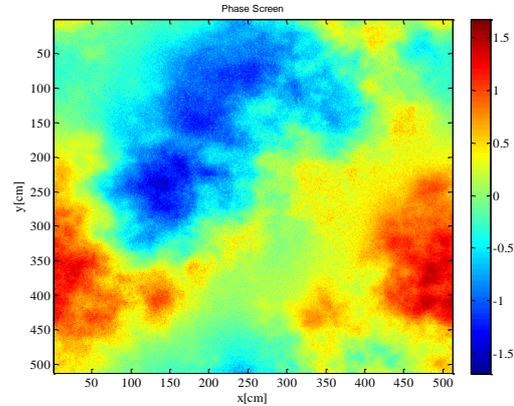


FIGURE 2a Two-dimension figure of the random phase screen after the harmonics are added

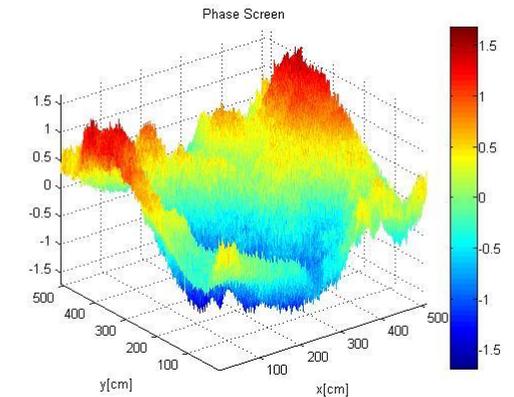


FIGURE 2b Three-dimension figure of the random phase screen after the harmonics are added

From Figure 2a and 2b the low frequency part of phase screen is more apparent after the overlay of sub-harmonics, which shows the overlay of harmonics can effectively compensate the lack of low frequency caused by the Fourier transform.

4 Propagation step by step

As illustrated in Figure 1, in the beam propagation process, the collimated Gaussian beams are quite similar to the Gaussian beams at the transmitting terminal (the input plane) with the same beam-waist radius ω_0 . After the beams transmit for a distance of L in the atmospheric turbulence, they remain similar to Gaussian beams at the receiving terminal with the beam-waist radius ω .

According to the Gauss equation, the optical field distribution in the input plane is [12]:

$$U_G(x, y, z) = \frac{A}{q(z)} \exp ik \left(\frac{x^2 + y^2}{2q(z)} \right). \quad (10)$$

We can obtain the following equation from Collins integral equation:

$$\frac{1}{q(z)} = \frac{1}{R(z)} - \frac{i\lambda M^2}{\pi W^2(z)}, \tag{11}$$

were:

$$W^2(z) = W_0^2(z) \left[1 + \left(\frac{\lambda z}{\pi W_0^2} \right)^2 \right], R(z) = z \left[1 + \left(\frac{\lambda z}{\pi W_0^2} \right)^2 \right].$$

are the isophase surface curvature radius and the beam radius of the Gaussian beam respectively. M^2 is defined as the beam quality factor and for the fundamental-mode Gaussian beam, its value is 1. W_0 is the beam waist radius, λ is the wavelength of the laser being transmitted and z is the position where the laser is located on its transmitting route.

From the equation of $R(z)$, we can find $R(z) \rightarrow \infty$. Substituting $R(z) \rightarrow \infty$ into Equation (11), the result is:

$$q(z) = \frac{i\pi W_0^2}{\lambda M^2}. \tag{12}$$

If $q(z)$ is substituted into Equation (10), the optical field distribution is as follows:

$$U_G(x, y, z) = -\frac{iA\lambda M^2}{\pi W_0^2} \exp\left(M^2 \frac{x^2 + y^2}{W_0^2} \right). \tag{13}$$

Diffraction occurs after the beams are collimated, and the diffracted beam continue to transmit for a distance of Δz , in the end we can represent the optical field distribution as follows based on Fresnel diffraction integral:

$$U_G(x_2, y_2, z_2) = -\frac{\exp(ikz_2)}{i\lambda\Delta z} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_G(x_1, y_1, z_1) \times \exp\left[\frac{i\pi M^2}{\lambda\Delta z} \frac{(x_2 - x_1)^2 + (y_2 - y_1)^2}{W_0^2} \right] dx_1 dx_2. \tag{14}$$

TABLE 1 The values of the simulation parameters and their physical significance

| Parameter | Value | Physical significance |
|------------|-------------------------------------|---------------------------------------|
| W_0 | 0.05m | Beam waist |
| λ | 1.06 μ m | The wavelength of the reflected laser |
| Δz | 500m | Interval between the phase screens |
| N_x, N_y | 1024 | Sampling points |
| L_0 | 20m | Outer turbulence scale |
| l_0 | 5m | Inner turbulence scale |
| C_n^2 | 10 ⁻¹⁷ m ^{-2/3} | Atmosphere structure parameters |
| M^2 | 1 | Gaussian beam quality factors |
| G_x, G_y | 4.8m | Size of the phase screen |
| r_0 | 0.810mm | Atmospheric coherence length |
| D_1 | 0.1m | Transmitter aperture diameter |
| D_2 | 0.2m | Receiver calibre diameter |

5 Simulating

When the quasi-Gaussian beam propagates some distance, beam expander will occur. For expanded beam, the emission aperture and receiving aperture will change, its new diffraction diameter and receiving aperture is defined as [13]:

$$D_1' = D_1 + c \frac{\lambda\Delta z}{r_{0,rev}}, \tag{15}$$

$$D_2' = D_2 + c \frac{\lambda\Delta z}{r_{0,rev}}, \tag{16}$$

where $r_{0,rev}$ is atmospheric coherence diameter for back-propagation and c is an adjustment factor of turbulence sensitive.

In order to simulate the beam propagation process with more accuracy, sampling points of the transmit and receive aperture plane aperture plane, sampling interval, maximum allowable interval between planes and minimum number of transmission steps should be selected [14]. The phase difference between two adjacent points on the phase screen should be smaller than π according to the Nyquist law. In other words, the grid sampling intervals Δk_x and Δk_y on the phase screen satisfy the conditions [7]:

$$|\Psi(k_x + \Delta k_x, k_y, z) - \Psi(k_x, k_y, z)| < \pi, \tag{17}$$

$$|\Psi(k_x, k_y + \Delta k_y, z) - \Psi(k_x, k_y, z)| < \pi. \tag{18}$$

The parameters of the system are selected as follows for the purpose of simulation:

After the Kolmogorov atmospheric turbulence is introduced into the optical propagation route, the light intensity and phase distribution in receiving aperture are shown in Figures 3a and 3b.

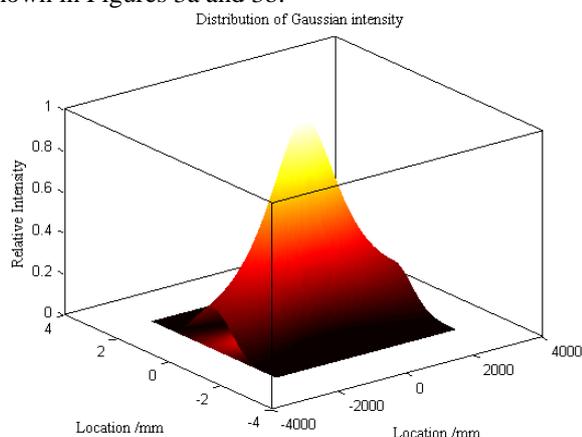


FIGURE 3a Three-dimension figure of Optical field distribution of collimated Gaussian beams in the receiving plane

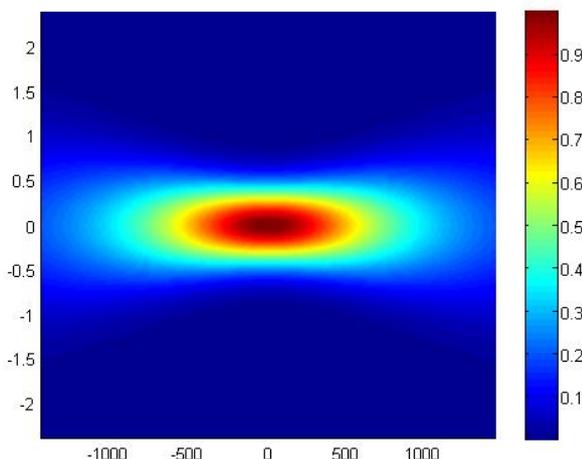


FIGURE 3b Two-dimension figure of Optical field distribution of collimated Gaussian beams in the receiving plane

6 Testifying the results after simulating the turbulent atmosphere

The statistical characteristics of the atmospheric turbulence phase can be depicted by the phase structure function, thus the structure function can be used as a benchmark to testify if the simulated phase screen is correct. Thus, Fried offered the definition equation of the structure function corresponding to Kolmogorov spectrum [15]:

$$D(r) = 6.88(r / r_0)^{5/3} \tag{19}$$

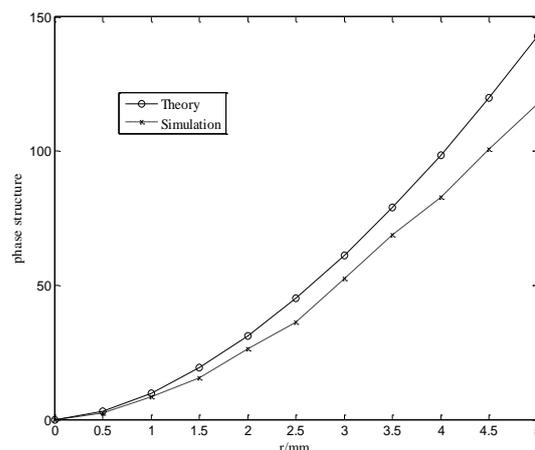


FIGURE 4 Comparison between values of the phase structure function

In the experiment, $r_0=0.810\text{mm}$ and Y-axis are the values of the phase structure function. It can be found from Figure 4 that due to the sampling frequency of the Fourier transform, part of the low frequency is lost in the phase screen generated. Thus the structure function of the phase screen obviously lacks low frequency part compared to the theoretical situation, while the performance is the same in the high frequency parts.

7 Conclusions

Based on Kolmogorov turbulence theory and power spectral inversion method, the model and simulation method on laser beams propagation through two-dimension random phase screens in the atmospheric turbulence channel were proposed and testified by phase structure function. In the method, the random phase screen was established by the Fourier transform, and step-by-step transmitting approach was used to simulate the propagation of collimated Gaussian beams in Kolmogorov turbulence. According to the divergence between phase structure function and theoretical results, the accuracy of simulated phase screen was analysed. Simulation results showed the proposed method in the paper was appropriate after compensating the low frequency and can be used to calculate the light propagation, which will be more practical meaningful to evaluate and test the phase screen.

Acknowledgments

This research was supported in part by National Natural Science Foundation (No.61070175), Shandong Province Natural Science Foundation (ZR2013FL017, ZR2013FL018), Shandong Province University Science and Technology (J12LJ03) of China, project development plan of science and technology of Yantai (2013ZH347, 2013ZH091). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

- [1] Tomlinson A, Hemenger R P, Garriott R 1993 Method for estimating the spherical aberration of the human crystalline lens in vivo *Ophthalm. Vis. Sci* **34**(6) 621-29 (in Chinese).
- [2] Wang R, Wang T 2013 Simulation Model of Laser Atmospheric Transmission Characteristics Using Arbitrary Thickness Random Phase Screen *Chinese Journal of Lasers* **10**(8) 1-6 (in Chinese).
- [3] Duan J, Wang X, Jing W 2010 The Atmosphere Turbulence Simulation Based on Zernike Polynomial *Journal of Changchun University of Science and Technology* **33**(3) 63-6 (in Chinese).
- [4] Andrews R C, Phillips R L 2005 Laser Beam propagation through Random Media (2nd Edition) *Bellingham: SPIE PRESS*, Bellingham, WA chapter 6
- [5] Li H., Yan C., Hua Z 2009 Numerical Simulation of the Atmospheric Turbulence Phase Screen based on Fractal. *Journal of Atmospheric and Environmental Optics* **4**(3) 171-7 (in Chinese)
- [6] Wang L, Li Q, Wei H 2007 Numerical simulation and validation of phase screen distorted by atmospheric turbulence *Opto-Electronic Engineering* **34**(3) 1-5
- [7] Ma D, Wei J, Zhuang Z 2004 Performance Evaluation and Channel Modeling of Multiple-Beam Propagation for Atmospheric Laser Communication *Acta Optica Sinica* **24** (8) 1020-5 (in Chinese)
- [8] Lane R G, Glindemann A, Dainty C 1992 Simulaiton of a Kolmogorov phase screen *Waves in Random Media* **2**(3) 209-24
- [9] Zhang L, Wu Z, Gao S, Cui M 2013 Arrival-time model with femtosecond precision for simulating laser pulses propagating through atmospheric turbulence *Optik* **124**(14) 1758-62
- [10] Gan X J, Guo J, Fu X Y 2006 The simulation turbulence method of laser propagation in the inner field *Physics* **48**(1) 907-10
- [11] Bloom S, Korevaar E, Schuster J, Willebrand H 2003 Understanding the performance of free-space optics *Journal of Optical Networking* **2**(6) 178-200
- [12] Garcia-Zambrana A 2007 *IEEE Communication Letters* **11**(5) 390-2
- [13] Chatzidiamantis N D, Uysal M, Tsiftsis T A, Karagiannidis G K 2010 *IEEE Journal of Lightwave Technology* **28**(7) 1064-70
- [14] Voelz D G, Roggemann M C 2009 Digital simulation of scalar optical diffraction: revisiting chirp function sampling criteria and consequences *Applied Optics* **48**(32) 6132-42
- [15] El Hage S G, Berny F 1973 Contribution of the crystalline lens to the spherical aberration of the eye *Journal of Optical Society of America* **63**(3) 205-11

| Authors | |
|---|---|
|  | <p>Yanli Feng, born in August, 1963, Yantai County, Shandong Province, P. R. China</p> <p>Current position, grades: Professor at the Department of Computer Foundation Studies, Shandong Institute of Business & Technology, China. University studies: B.S. in computer application at Shandong University in China. M.S. at Dalian University of Technology in China. Scientific interest: Wireless communication, computer networks. Publications: more than 30 papers. Experience: teaching experience of 31 years, 12 scientific research projects.</p> |
|  | <p>Dashe Li, born in February, 1978, Yantai County, Shandong Province, P. R. China</p> <p>Current position, grades: Associate Professor at the Department of School of Computer Science and Technology, Shandong Institute of Business & Technology, China. University studies: B.S. in wireless communication at Qufu Normal University. M.S. at China University of Mining & Technology (Beijing) in China. Scientific interest: Wireless communication, computer networks. Publications: more than 20 papers. Experience: teaching experience of 31 years, 8 scientific research projects.</p> |
|  | <p>Shue Liu, born in November, 1977, Yantai County, Shandong Province, P. R. China</p> <p>Current position, grades: Lecturer at School of Computer Science and Technology, Binzhou Medical University, China. University studies: B.S. in wireless communication at Qufu Normal University in China. M.Sc. at Yantai University in China. Scientific interest: Wireless communication, computer networks. Publications: more than 10 papers. Experience: teaching experience of 12 years, 3 scientific research projects.</p> |

The design of a dynamic slope compensation circuit for boost DC-DC converter

Zhao Han, Yuan Rao, Wentao Chen, Junkai Huang*

School of Information Science and Technology, Jinan University, Guangzhou, Guangdong Province, 510632, China

Received 6 April 2014, www.tsi.lv

Abstract

This paper proposes a structure of peak current mode Boost DC-DC converter with slope compensation circuit, and designs a dynamic slope compensation circuit applied to this converter. With the utilization of the voltage controlled resistance characteristics of MOS transistor and the introduction of a clamp circuit consist of cascade current mirror, a dynamic slope compensation circuit is realized. The circuit is simulated on Cadence Spectre using SMIC 0.18 μ m CMOS technology. Results show that it can provide proper slope compensation following the variation of input and output. The load capacity of DC-DC converter reaches 550mA and the transient response lows to 10 μ s. By eliminating the problem of instability caused by the peak current mode switching power supply of double loop control, the design improves the stability of switching power supply.

Keywords: boost DC-DC, slope compensation, current mirror, voltage controlled resistor

1 Introduction

With the rapid development of electronic devices, the requirements of the power supply with higher storage capacity and stability are growing rapidly. Though battery technologies have been developing fast, they still cannot meet these requirements. So more and more attentions have been draw to DC-DC switching power supply [1-3], which has high conversion efficiency and stability. However, in the peak current mode, the deviation between the peak inductive current and the average output current lowers the precision. Especially, there are several drawbacks such as sub-harmonic oscillation, open-loop instability and worse loop response when the duty cycle is over 50%. Therefore, the slope compensation circuit [4, 5] used to improve system performance has practical value. The fixed slope compensation and the piecewise linear slope compensation have been proposed in [6] and [7]. Nevertheless, these two slope compensation methods cannot adjust the slope compensation current dynamically with the changes in the duty cycle, and may cause excessive compensation or under compensation, which may lead to lower the transient response and load capability of the switching power supply.

Based on DC-DC converter, a low power, MOS tube consisted dynamic slope compensation circuit is proposed, which can automatically adjust the slope compensation current following the variation of input and output voltage, and the simulation of it is also conducted.

2 Peak Current Mode boost DC-DC converter

Based on the technical requirements of the Boost DC-DC converter [4], the proposed structure of peak current mode Boost DC-DC converter with slope compensation circuit is shown in Figure 1. Synchronous rectification with dead time is used in this structure to prevent the main power tubes and freewheeling tubes from conducting simultaneously. Moreover, the structure is controlled by two feedback loops. One is voltage feedback loop composed by the error amplifier which receives the output voltage sampling signal, and the other is current feedback loop composed by the peak current comparator which receives current sampling signal and slope compensation signal.

In Figure 1, error amplifier amplify the difference between the output voltage sampling signal V_{FB} and the reference voltage V_{REF} , and input the resulting voltage signal V_E to the peak current comparator. The inductive current signal V_S extracted by current sampling circuit and the slope compensation signal generated by slope compensation circuit are superimposed, and the result is inputted to the peak current comparator. The output control signals adjust conduction time of the power MOSFET to realize stability. So, the performance of slope compensation circuit plays a decisive role in realizing stability of the whole system. It can reduce or even eliminate the open-loop instability and other negative phenomena caused by inductive current disturbance.

* *Corresponding author* e-mail: hjkeed@163.com

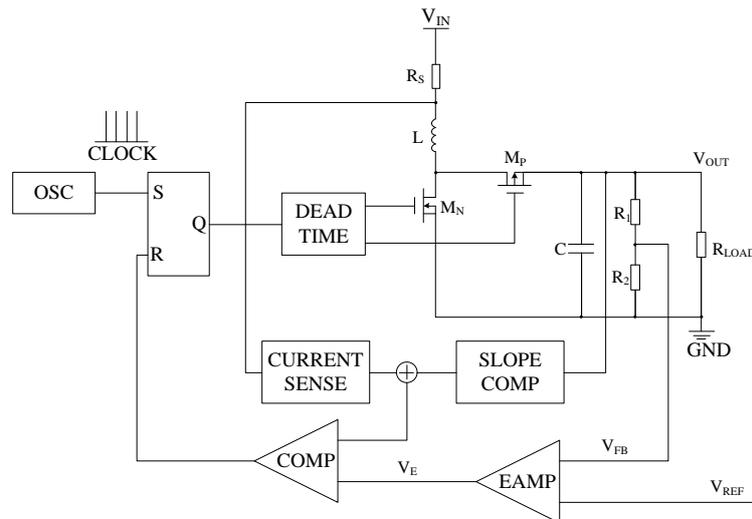


FIGURE 1 Structure of peak current mode Boost DC-DC converter

3 Dynamic Slope Compensation Method and Design

3.1 THE BASIC PRINCIPLE OF SLOPE COMPENSATION

Figure 2 shows the open-loop instabilities phenomena [2, 5] in the Boost DC-DC converter when duty cycle D is over 50%. Where V_E is the error amplifying signal of the voltage feedback loop, m_1 and m_2 represent the ascending slope and descending slope respectively, and ΔI_0 is the initial disturbance current.

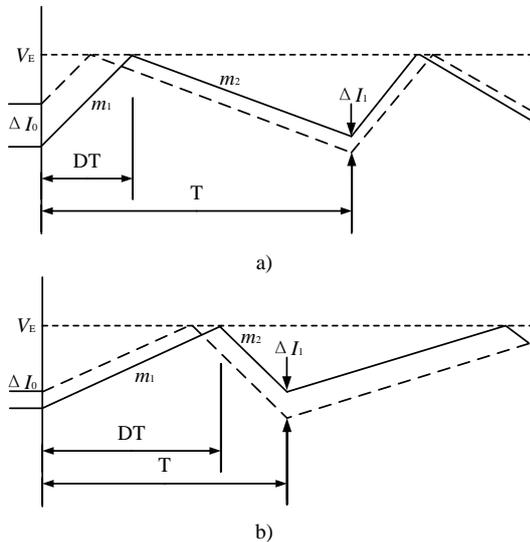


FIGURE 2 The instabilities of Boost DC-DC converter without slope compensation: a) $D < 50\%$, b) $D > 50\%$

In Figure 2, the solid lines stand for the inductive current waveform when the circuit is stable, while the dashed lines stand for the same waveform when the circuit is disturbed. As we can see, when $D < 50\%$, the ratio $m_2/m_1 < 1$, the disturbance current is lowering, the system is stable. While $D > 50\%$, the ratio $m_2/m_1 > 1$, the disturbance current is rising, the open-loop is unstable.

Figure 3 presents the introduction of slope compensation when $D > 50\%$. To eliminate the open-loop

instability, the slope m of the introduced signal should satisfy the equation below:

$$\left| \frac{m_2 - m_1}{m_1 + m} \right| < 1 \tag{1}$$

In the Continue Conduction Mode (CCM) mode, the relationship between the slope m of the slope compensation and the duty cycle D can be expressed as [7]:

$$m > (1 - \frac{1}{2D})m_2 \tag{2}$$

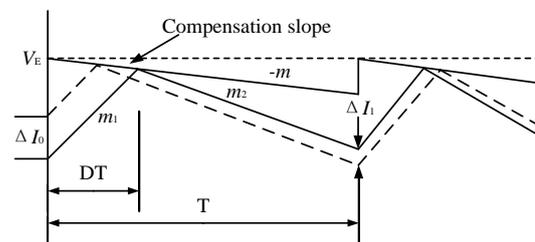


FIGURE 3 In the situation of $D > 50\%$, the stability of system is realized when slope compensation is introduced

This is the basic principle of slope compensation. And two conventional ways are the fixed slope compensation and the piecewise linear slope compensation. However, both of them may bring excessive compensation (or under compensation), resulting in lower transient response and load capability of switching power supply.

3.2 THE BAISC METHOD OF DYNAMIC SLOPE COMPENSATION

According to the drawbacks of fixed slope compensation and piecewise linear slope compensation, this paper presents a method of dynamic slope compensation. Specifically, for the Boost DC-DC converter shown in Figure 1, the rise and fall slope of inductive current can be expressed as [8]:

$$m_1 = \frac{V_{IN}}{L} R_{DSON}, \tag{3}$$

$$m_2 = \frac{V_{OUT} - V_{IN}}{L} R_{DSON}, \tag{4}$$

where L is the inductive value and R_{DSON} is the resistance of switch.

Using Equations (3) and (4), Equation (1) can be reversed as:

$$m > \frac{0.5V_{OUT} - V_{IN}}{L} R_{DSON}. \tag{5}$$

The system can remain stable as long as the input voltage V_{IN} , the output voltage V_{OUT} and the slope m of the slope compensation signal satisfy Equation (5). Therefore, we can achieve dynamic slope compensation by adjusting compensation slope following the variation of input and output voltage, so that excessive compensation can be eliminated and the load capacity of the system can be enhanced.

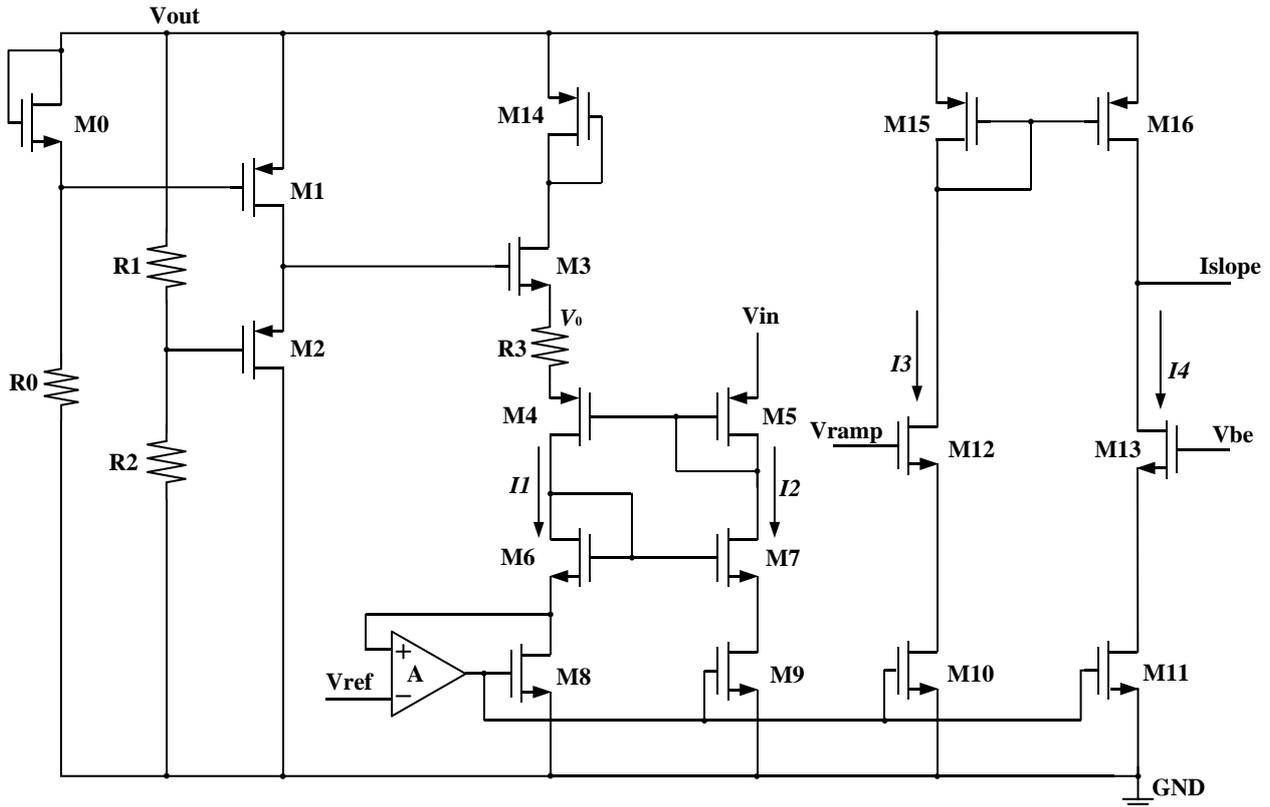


FIGURE 4 Dynamic slope compensation circuit

3.3 DYNAMIC SLOPE COMPENSATION CIRCUIT DESIGN

Figure 4 is the dynamic slope compensation circuit designed in this paper, where M0, M1, M2, M3, R0, R1, R2, R3 constitute a voltage divider network producing a regulated output voltage, and M4, M5, M6, M7 constitute cascode current mirror producing current I_1 following the variation of input and output voltage. M8, M9, M10, M11 work in the linear region and can be equivalent to resistances whose values are controlled by the overdrive voltage. Operational amplifiers A working as a voltage follower adjust the equivalent resistance of M8. The slope compensation current I_{SLOPE} is outputted after the slope conversion of sawtooth signal V_{RAMP} . M14, M15, M16 constitute current mirror.

Specifically, M0 and R0, M1 and M2, M3 and R3 form a source follower, R1 and R2 form a voltage divider

network. By adjusting the W/L of M0, M1, M2, M3 and other related parameters to get $|V_{GS2}|=V_{GS3}$, we obtain:

$$V_0 = \frac{R_2}{R_1 + R_2} V_{OUT}. \tag{6}$$

To get the current following the variation of input and output voltage, we introduce cascode current mirror as clamp circuit and let the W/L of M4, M5 and M6, M7 equal respectively, which make $I_1=I_2$ and source voltage of M4 equal to V_{IN} , therefore:

$$I_1 = \frac{(V_0 - V_{IN})}{R_0}. \tag{7}$$

With Equation (6), Equation (7) can be rewritten as:

$$I_1 = \frac{((R_2 / (R_1 + R_2)) V_{OUT} - V_{IN})}{R_0}. \tag{8}$$

The inverse of the operational amplifier A is provided by a 0.2V reference voltage, which makes the drain voltage of M8 equal to 0.2V, so that it can work in linear region. Then, the resistance of it can be expressed as:

$$R_{M8} = \frac{0.2}{I_1} = \frac{0.2R_0}{(R_2/(R_1 + R_2))V_{OUT} - V_{IN}} \quad (9)$$

The on-resistances of MOS tubes working in linear region can be expressed as [9]:

$$R_{DS} = \frac{1}{\mu C_{OX}(V_{gs} - V_{th})(W/L)} \quad (10)$$

When the gate voltage and source voltage of M8, M9, M10, M11 equal respectively, they have the same V_{gs} . As long as the W/L of these four tubes are equal, they can work in linear region and have the same on-resistance, which can be expressed as:

$$R_{M8} = R_{M9} = R_{M10} = R_{M11} = \frac{0.2R_0}{(R_2/(R_1 + R_2))V_{OUT} - V_{IN}} \quad (11)$$

In Figure 4, M10, M11, M12, M13 constitute slope conversion circuit with a source negative feedback, if $R_{M10} \gg \frac{1}{g_{m12}}$, drain current of M12 can be thought as a linear function of its own input gate voltage. We obtain:

$$I_3 \approx \frac{V_{RAMP} - V_{th}}{R_{M10}} \quad (12)$$

Similarly, if $R_{M11} \gg \frac{1}{g_{m13}}$, we obtain:

$$I_4 \approx \frac{V_{BE} - V_{th}}{R_{M11}} \quad (13)$$

Make the W/L of M15, M16 equal, as $R_{M10} = R_{M11}$, according to the current mirror relationship, we obtain:

$$I_{SLOPE} = I_3 - I_4 = \frac{V_{RAMP} - V_{BE}}{R_{M10}} \quad (14)$$

$$= \frac{(V_{RAMP} - V_{BE})((R_2/(R_1 + R_2))V_{OUT} - V_{IN})}{0.2R_0}$$

We can get the slope of compensation signal by (14):

$$m_{SLOPE} = \frac{(R_2/(R_1 + R_2))V_{OUT} - V_{IN}}{0.2R_0} m_{RAMP} \quad (15)$$

As a result, we can choose resistances of R0, R1, R2 according to the slope m_{RAMP} of sawtooth signal so that the following formula can be satisfied:

$$\frac{1}{0.2R_0} m_{RAMP} > \frac{1}{L} R_{DSON} \quad (16)$$

Thereby, we can adjust the slope of slope compensation circuit following the variation of input and output voltage to eliminate excessive compensation phenomenon and achieve the stability of the current loop.

4 Simulation results and analysis

Based on peak current mode Boost DC-DC converter shown in the Figure 1, the dynamic slope compensation circuit is simulated on Cadence Spectre using SMIC 0.18μm CMOS technology. Figure 5a) is the changes of compensation slope when the input voltage is 1.8 V and the range of output voltage is 4 V to 7 V. Figure 5b) shows changes of the compensation slope when the output voltage is 4 V and the range of input voltage is 2.3 V to 3.3 V. As we can see, the compensation slope has been increased with the duty cycle changing from 55% to 75%, and the compensation slope decreases with the increasing of input voltage.

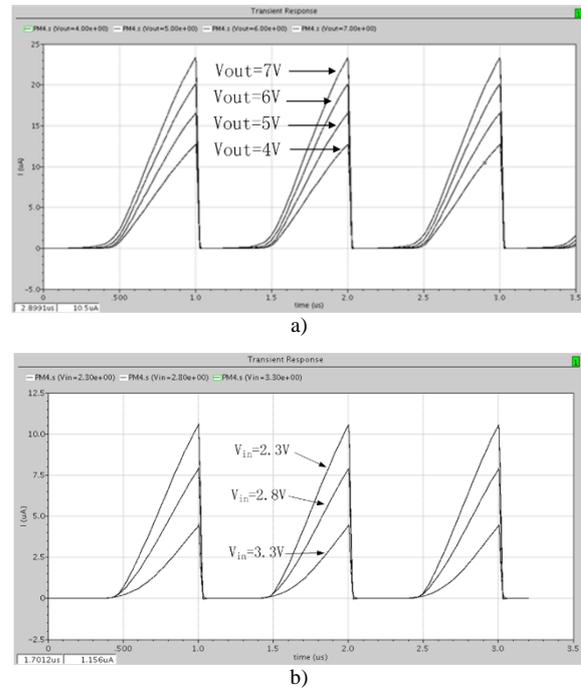


FIGURE 5 Slope compensation current with the changes in V_{IN} and V_{OUT} : a) $V_{IN}=1.8V$, b) $V_{OUT}=4V$

The PWM waveforms shown in Figure 6 are simulated in peak current mode Boost DC-DC converter with the dynamic slope compensation circuit working in steady state. Where, the simulation parameters are $V_{IN}=1.8V$, $V_{OUT}=4V$, $L=2.2\mu H$, $C=10\mu F$, $f=1MHz$. As we can see, under the condition of duty cycle $D=55\%$, the inductive current is constant at 550mA and sub-harmonic oscillation phenomenon did not occur. Therefore, the compensation circuit designed in this paper works properly and the loop is stable.

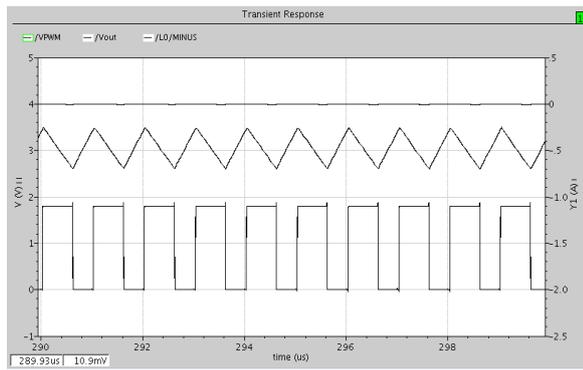


FIGURE 6 Simulation waveform under PWM

Figure 7 shows the transient response simulation waveform of DC-DC converter when rise and fall time are $1\mu s$ and amplitude of square load current is $200mA$ in $V_{IN}=1.8V$, $V_{OUT}=4V$. The response time of the output voltage is less than $10\mu s$, voltage overshoot is only $80 mV$, and there is no ringing. The system works stable.

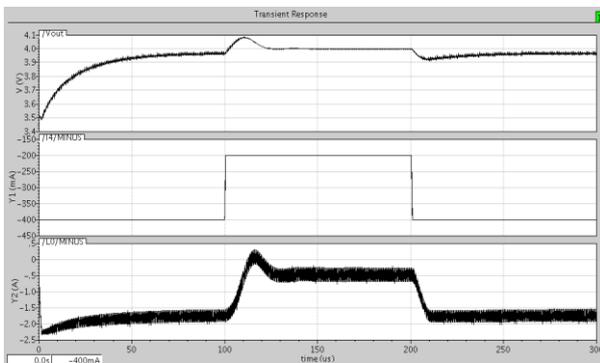
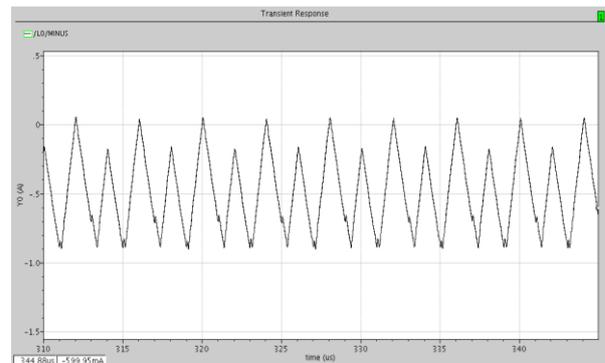


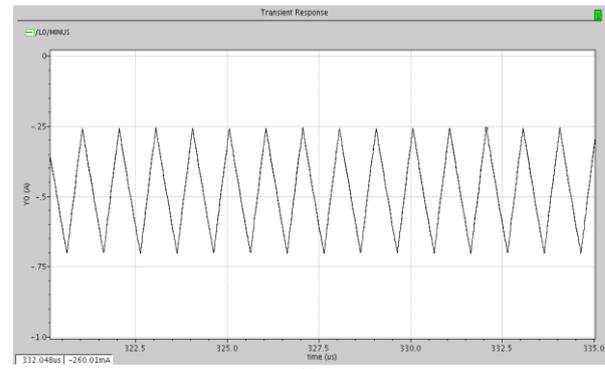
FIGURE 7 Load transient response waveform

Finally, in the case of load current is constant at $550mA$, we simulate the whole DC-DC converter with the slope compensation circuit and without the slope compensation circuit respectively to further verify the stability of the

system when the duty cycle is over 50%. The simulation result is shown in Figure 8. The oscillation of inductive current is serious and the inductive current is instable when slope compensation circuit is not introduced. On the contrary, when the slope compensation circuit is introduced, the inductive current is constant at $550mA$.



a)



b)

FIGURE 8 Inductor current simulation waveform: a) Without slope compensation, b) With slope compensation

The comparison of the main technical parameters of the circuit designed in this paper and other typical slope compensation circuits is shown in Table 1.

TABLE 1 Comparison between circuit in this paper and other typical slope compensation circuits

| Parameter | This paper | Reference [9] | Reference [10] |
|-------------------------------------|------------|---------------|----------------|
| technology (μm) | 0.18 | 0.8 | 0.25 |
| input voltage (V) | 1.8-4 | 2.5-5 | 3.3-12 |
| output voltage (V) | 4 | 6 | 3.3 |
| load current (mA) | 550 | 500 | 800 |
| voltage overshoot (mV) | 80 | 100 | 150 |
| transient response time (μs) | 10 | 30 | 70 |

5 Conclusions

This paper proposed a new method of dynamic slope compensation and designed a dynamic slope compensation circuit with simple structure based on the traditional slope compensation. Simulation results show that this circuit can provide proper slope compensation signal for peak current Boost DC-DC converter, as well as faster transient response, slighter voltage overshoot and higher load

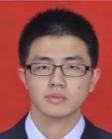
current. More importantly, the results show that the circuit can eliminate excessive compensation and open-loop instability, and realize stable operation of the system.

Acknowledgments

This work is supported by Promotion of Development in Science and Technology Services Project in Guangdong Province, China (No.2011B040300035).

References

- [1] Qun Z, Lee F C 2003 *IEEE Transactions on Power Electronics* **18**(1) 65-73
- [2] Lee C F, Mok P K T 2004 *IEEE Journal of solid-state circuits* **39**(1) 3-14
- [3] Liaw C M, Chiang S J, Lai C Y, Pan K H, Leu G C 1994 *IEEE Transactions on Industrial Electronics* **41**(2) 231-40
- [4] Hu S, Zou X, Zhang J Kong L 2007 A new dynamic slope compensation circuit applied to boost DC-DC converter *Computer and Digital Engineering* **35**(10) 159-62 (in Chinese)
- [5] Chen F, Lai X, Li Y 2008 Design and implementation of an adaptive slope compensation circuit *Journal of Semiconductors* **29**(3) 593-7 (in Chinese)
- [6] Canesin C A, Barbi I 1996 *IEEE Applied Power Electronics Conference and Exposition San Jose* **2** 807-13
- [7] Bryant B, Kazimierczuk M K 2005 *IEEE Transactions on Circuits and Systems* **52**(11) 2404-12
- [8] Lu J, Wu X 2007 *IEEE Electron Devices and Solid-State Conference* 929-32
- [9] Wang L, Wang S, Lai X, Tian J 2007 Design of a piecewise linear slope compensation circuit for peak current mode boost DC-DC converter *Research & Progress of SSE* **27**(2) 269-74 (in Chinese)
- [10] Guoding Dai, Yang Xu, Weimin Li and Bo Hu. 2012 The design and realization of internal compensation circuit for current mode PWM step-down DC-DC converters *Electron Devices* **33**(1) 53-7 (in Chinese)

| Authors | |
|---|--|
|  | <p>Zhao Han, born in February, 1990, Jincheng, Shanxi, P.R. China</p> <p>Current position, grades: Master student at the School of Information Science and Technology, Jinan University, China. University studies: B.Sc. at College of Information and Business at North University of China in China (2008-2012). Scientific interest: power management units and analogue integrated circuits designs. Experience: studying in the field of analogue integrated circuit design for 2 years, 1 scientific research project.</p> |
|  | <p>Yuan Rao, born in May, 1989, Wuzhou, Guangxi, P.R. China</p> <p>Current position, grades: Master student at the School of Information Science and Technology, Jinan University, China. University studies: B.Sc. at the International School at Beijing University of Posts and Telecommunications in China (2007-2011). Scientific interest: analogue integrated circuits designs, system on chip designs, power management units. Publications: 3 papers. Experience: studying in the field of analogue integrated circuit design for 3 years, 2 scientific research projects.</p> |
|  | <p>Wentao Chen, born in February, 1990, Longyan, Fujian, P.R. China</p> <p>Current position, grades: Master student at the School of Information Science and Technology, Jinan University, China. University studies: B.Sc. at the school of microelectronics at Xidian University in China (2008-2012). Scientific interest: analogue integrated circuits designs and system on chip designs. Publications: 1 paper submitted. Experience: studying in the field of analogue integrated circuit design for 2 years.</p> |
|  | <p>Junkai Huang, born in October, 1963, Shantou, Guangdong, P.R. China</p> <p>Current position, grades: Ph.D. degree, Professor at the School of Information Science and Technology, Jinan University, China. Director of the Department of Electronic Engineering and the Vice President of the School of Information Science and Technology in Jinan University. University studies: B.Sc. and M.Sc. at the School of Information Science and Technology from Jinan University in China. Ph.D. at South China University of Technology in China. Scientific interest: Thin-film material and devices, application specific integrated circuit chips. Publications: more than 50 papers. Experience: teaching experience of 30 years.</p> |

The time-frequency analysis of the train Axle box acceleration signals using empirical mode decomposition

Xingjie Chen, Xiaodong Chai*, Xining Cao

College of Urban Railway Transportation, Shanghai University of Engineering Science, Shanghai, China, 201620

Received 6 May 2014, www.tsi.lv

Abstract

Rail defects usually result in lots of problems such as affecting the comfort of passengers, increasing the wheel-rail forces, exacerbating the train axle boxes vibration and track wear, even threatening the safe operation of trains. In this paper, the characteristic frequency distribution of the changing axle box acceleration caused by defects is analysed by empirical mode decomposition and Hilbert-Huang Transform is used to analyse the time-frequency changes of axle box acceleration. As a result, rail defects can be effectively positioned and the short wave irregularities within a certain degree can be detected. The research provides timely protection for the maintenance of the track.

Keywords: track detection, empirical mode decomposition, time-frequency analysis, Axle box acceleration

1 Introduction

Condition monitoring of railway tracks, vehicles are essential in ensuring the safety of railways [1]. Early track defects maintenance can not only prevent further deterioration of the state of the track, but also to save the track testing and maintenance costs. Traditional railway track detection methods, special track inspection cars, light rail detection cars do not adapt to the frequent detection of rapid measurement of high-density urban railway lines.

Track detection methods based on the online running vehicles are becoming a key research direction. Li [2] research on early detection of track defects based on the method of axle box vibration acceleration mutations, using validated finite element model to simulate the axle box acceleration changes caused by the track defects to analysis of the type and the location of track defects for early maintenance. Italy M. Boccione [3] et al studied vehicular measurement device to collect vehicle axle box acceleration, and study the wear condition of the track by the method of the mean square value of the collected data analysis. The graduate school of Nihon University Mori et al research [4] collected noise signal of the train compartment caused by rail defects by installing the microphone on the train when it running, and combined simulation to detect track status online.

The aforementioned researches on the rail defect detection and data processing have lay a solid foundation for our research. In this paper, Hilbert-Huang transform is used to analyse and process the axle box acceleration signal. And the characteristic frequency distribution of the changing axle box acceleration caused by track defects is analysed to get track irregularity frequency (wavelength)-amplitude time frequency distribution. The rest of this

paper is organized as follows. Firstly, the empirical mode decomposition method is discussed in section 2. Then, Axle box acceleration signals are collected in section 3. Based on this, the time-frequency analysis of axle box acceleration is researched in section 4. Finally, conclusions are discussed in section 5.

2 The Empirical Mode Decomposition method

As the acceleration collected on the railway vehicle axle box are non-stationary and nonlinear signals, And Hilbert-Huang Transform [5] is the most effective method to process the signals. This analysis method was a new signal analysis theory, it first proposed by NASA NE Huang et al in 1998. The main innovations include the proposed of the concept of Intrinsic Mode Function (IMF) and the introduction of Empirical Mode Decomposition (EMD). EMD decompose a signal based on the data itself characteristic time scales, it is a kind of self-adaptive and efficient data processing methods, and have very obvious advantage in nonlinear non-stationary signals processing.

EMD decompose the signal into some intrinsic mode functions (IMFs) with different scale features, and the IMFs satisfy the following two conditions:

- 1) In the whole data set, the number of extreme points and the number of zero-crossings must be either equal or differ at most by one;
- 2) At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

To extract the IMF from a given data set, the sifting process is implemented as follows. First, identify all the local extreme, and then connect all of the local maxima by a cubic spline line as the upper envelope. Then, repeat the

* *Corresponding author* e-mail: cxdyj@163.com

procedure for the local minima to produce the lower envelope. The upper and lower envelopes should cover all the data between them. Their mean is designated as $m_1(t)$, and the difference between the data and $m_1(t)$ is $h_1(t)$, i.e.:

$$h_1(t) = x(t) - m_1(t). \tag{1}$$

Ideally, $h_1(t)$ should be an IMF, for the construction of $h_1(t)$ described above should have forced the result to satisfy all the definitions of an IMF by construction. To check if $h_1(t)$ is an IMF, we demand the following conditions:

(i) $h_1(t)$ should be free of riding waves i.e. the first component should not display under-shots or over-shots riding on the data and producing local extremes without zero crossing.

(ii) To display symmetry of the upper and lower envelopes with respect to zero.

(iii) Obviously the number of zero crossing and extremes should be the same in both functions.

The sifting process has to be repeated as many times as it is required to reduce the extracted signal to an IMF. In the subsequent sifting process steps, $h_1(t)$ is treated as the data; then:

$$h_{11}(t) = h_1(t) - m_{11}(t), \tag{2}$$

where $m_{11}(t)$ is the mean of the upper and lower envelopes of $h_1(t)$.

This process can be repeated up to k times; $h_k(t)$ is then given by:

$$h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t). \tag{3}$$

After each processing step, checking must be done on whether the number of zero crossings equals the number of extreme points.

The resulting time series is the first IMF, and then it is designated as $c_1(t) = h_{1k}(t)$. The first IMF component from the data contains the highest oscillation frequencies found in the original data $x(t)$.

This first IMF is subtracted from the original data, and this difference, is called a residue $r_1(t)$ by:

$$r_1(t) = x(t) - c_1(t). \tag{4}$$

The residue $r_1(t)$ is taken as if it was the original data and we apply to it again the sifting process. The process of finding more intrinsic modes $c_i(t)$ continues until the last mode is found. The final residue will be a constant or a monotonic function; in this last case it will be the general trend of the data.

$$x(t) = \sum_{j=1}^n c_j(t) + r_n(t). \tag{5}$$

Thus, one achieves a decomposition of the data into n -IMFs, plus a residue, $r_n(t)$, which can be either the mean trend or a constant.

Then using Hilbert transform to every IMF component $c_i(t)$, we get:

$$H[c_i(t)] = \frac{1}{\pi} P \int_{-\infty}^{+\infty} \frac{c_i(\tau)}{t - \tau} d\tau, \tag{6}$$

where P indicates the Cauchy Principle Value integral. And the analytic signal can be constructed by $c(t)$ and $H[c(t)]$,

$$z_i(t) = c_i(t) + jH[c_i(t)] = a_i(t)e^{j\theta_i(t)}, \tag{7}$$

where $a(t) = (c^2(t) + H[c(t)]^2)^{\frac{1}{2}}$ is instantaneous

amplitude, $\theta(t) = \arctan\left(\frac{H[c(t)]}{c(t)}\right)$ is instantaneous

phase, and the instantaneous frequency can be achieved by

$f(t) = \frac{1}{2\pi} \left[\frac{d\theta(t)}{dt} \right]$, therefore, the original signal $x(t)$ can

be presented as follows:

$$x(t) = \text{Re} \sum_{i=1}^n a_i(t) \exp \left[j \int f_i(t) dt \right]. \tag{8}$$

Equation (7) enables us to represent the amplitude and the instantaneous frequency, in a three-dimensional plot, in which the amplitude is the height in the time-frequency plane. This time-frequency distribution is designated as the Hilbert-Huang spectrum $H(\omega, t)$:

$$H(\omega, t) = \text{Re} \sum_{i=1}^n a_i(t) \exp \left[j \int \omega_i(t) dt \right]. \tag{9}$$

The Hilbert spectrum offers a measure of amplitude contribution from each frequency and time.

3 Axle box acceleration collection

Rail corrugation and sleeper spacing irregularity would cause great changes in the wheel-rail forces when vehicle running, wheel vibration will become more apparent especially the vehicle in the case of a high speed. Axle box and the wheel are rigid connection, axle box acceleration not only able to characterize the state of wheel vibration, but also to reflect the excitation formed by short track irregularity. Acceleration axle box contains larger bandwidth, with a variety of information of track defects. By measuring the acceleration of the axle box and its amplitude frequency variation analysis, we can make a fairly credible judgment for the track status.

Figure 1 is the schematic diagram of the axle box acceleration collection, accelerometer collect the changing axle box acceleration, the collected acceleration and the

train GPS data were transmitted to the signal collection devices via the data line simultaneously. Axle box acceleration, GPS data and other information of the train collected by ATS receiver connect with an external computer via data fusion system. Computer process the data to analysis track status. Figure 2 is the sensor layout of the test vehicle.

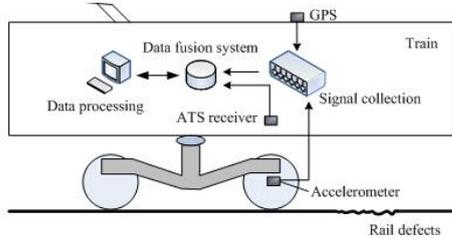


FIGURE 1 The schematic diagram of the Axle box acceleration collection



FIGURE 2 Sensor layout of the test vehicle

4 Axle box acceleration time-frequency analysis

Track irregularity is the main source of excitation of vehicle vibration, different spatial frequency components of the track will effects vehicle running status differently [6]. Shortwave irregularity is one of the main reasons to generate vehicle noise and cause changes in wheel-rail interaction force. Tiny rail surface defects generate huge wheel-rail interaction forces when train at high speed, and it cause axle box vibrate severely. Meanwhile, axle box acceleration can reflect the status of the track. In this paper, EMD and Hilbert transform combined with time-frequency analysis were used to process axle box acceleration.

B(Zhangjiang High Technology station)



A(longyang road station) Run line: A-B(2.9km)

FIGURE 3 The run line of test vehicle

The red line in Figure 3 is the run line of test vehicle, the train run between the longyang road station (point A) and Zhangjiang High Technology station (point B), in shanghai metro line 2. Vertical acceleration of the axle box collected in 92.5m long track, on a section of track between A and B. The speed of the test vehicle was 33.3Km / h, 270 sampling points per meter, the sampling frequency is 2500Hz. Figure 4 is the vertical acceleration of the axle box processed by a low-pass filter, figure 5 is its corresponding spectrogram.

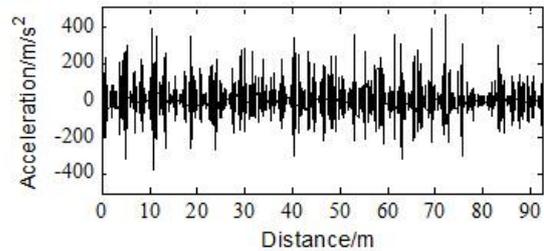


FIGURE 4 The vertical acceleration of the axle box

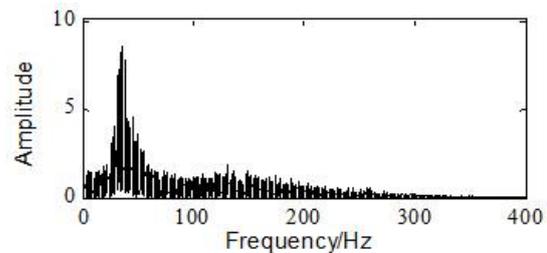


FIGURE 5 Spectrogram of acceleration

Shortwave irregularity wavelength typically distribute in the range of 0.05m-1m, therefore, it will cause vibration frequency distribution in the 9.25-185Hz on the vehicle axle box when the train speed is 33.3Km/h. The frequency distribution of axle box vertical acceleration can be a good reaction in vibration between wheel and rail. Using EMD to decompose the measured vertical acceleration of the axle box, figure 6 is the decomposition results, the acceleration was adaptively decomposed into 14 intrinsic mode function (IMF) based on the EMD algorithm, the last of the IMFs is trend signal. Figure 7 is the corresponding spectrum of IMFs, it is obviously that the frequency of IMFs is descend from high to low. In entire frequency band, the frequency amplitude between 0 and 200Hz are larger and it gradually weakening when frequency over 200Hz. This is because the vertical acceleration of axle box was processed by a low-pass filter before it decomposed by EMD.

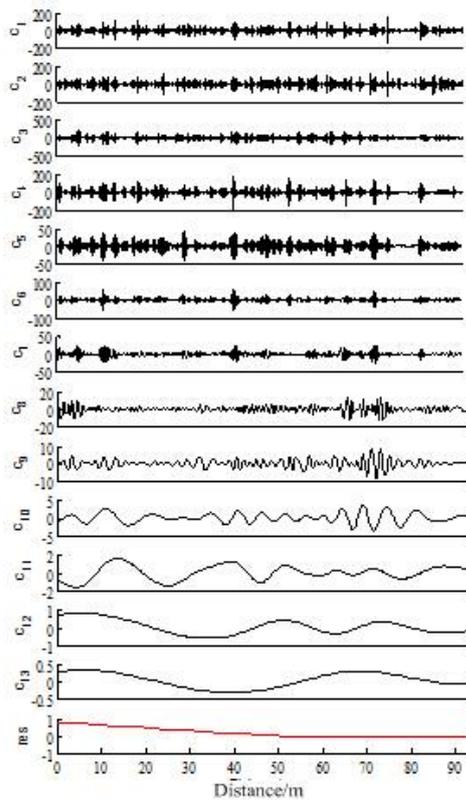


FIGURE 6 The decomposition results

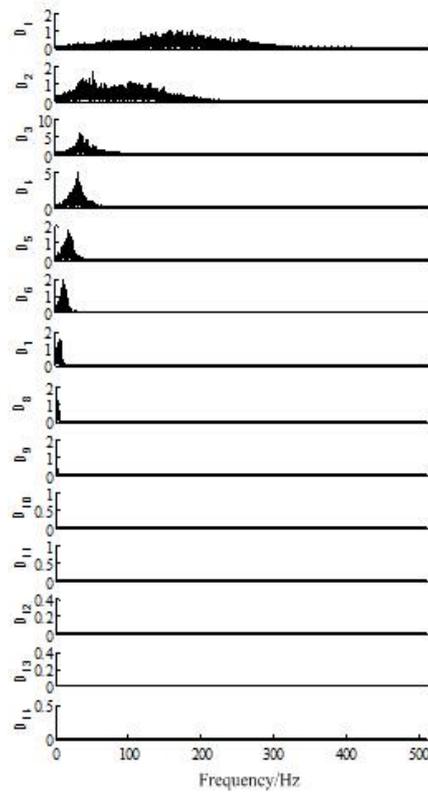


FIGURE 7 The corresponding spectrum of IMFs

The collected axle box acceleration includes not only part of the axle box vibrations caused by shortwave irregularity, but also includes other vibration from mechanical parts itself and surrounding environment. Due to the different vibration intensity, the energy of vibration on the performance of axle box is different, in which the vibration caused by the shortwave irregularity is far greater than any other case vibration. Because the acceleration have non-stationary characteristics and the shortwave irregularity randomness, it may appear that in acceleration sometime the signal intensity and the SNR are small, and another time the signal intensity and the SNR are high. Hilbert-Huang Transform can select the higher signal intensity to process in time series and improve the quality of data processing. It also can characterize random signals time-frequency distribution characteristics at the same time.

Figure 8 is the three-dimensional HHT time-frequency spectrum of axle box acceleration, it is obtained by calculating IMFs shown in Figure 6 by the method of Hilbert transform. Points in the figure indicates the energy, the brighter the colour, which means the higher the energy, and vice versa, the lower the energy. The horizontal axis represents the track mileage, and the vertical axis represents the frequency distribution of acceleration. As shown in figure 8, the frequency of the acceleration mainly within 200Hz, the frequency spectrum with good aggregation, and the instantaneous energy spectrum intuitively and unevenly distributed in several track

mileage (Figure 8, points 1-4). We can well identify energy changes with time and frequency.

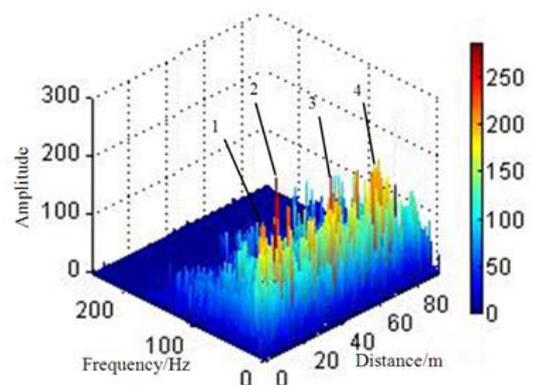


FIGURE 8 Three-dimensional time-frequency spectrum

Observing acceleration spectrum from the three-dimensional in Figure 8, in the frequency range of 10Hz-100Hz, the time-frequency energy spectrum is relatively concentrated obviously, especially in track mileage of 5m, 12m, 41m, 68m (1-4 point) the energy peaks are larger, the corresponding frequency and shortwave wavelength of track are mainly concentrated in the range of 11-27Hz and 0.343-0.841m. Wheel-rail interaction force will cause axle box greater vibration when vehicle running in these track shortwave irregularities, and it will be clearly reflected on energy peaks of axle box acceleration. By comparing the actual situation in the field section of track testing, we found the track shortwave irregularities position and the

wavelength range distribution are better match the analysis in this paper. The type of hundreds of millimetres track shortwave irregularities we detected, including sleeper spacing irregularities and rail corrugation, these would be the focus of the track maintenance.

5 Conclusion

This paper pre-processed the collected axle box acceleration, and decomposed the vertical acceleration of axle box using the EMD method to analyse the characteristics frequency distribution of track shortwave irregularity. Hilbert transform was then used to calculate the IMFs to get the three-dimensional HHT time-frequency spectrum of axle box acceleration. By analysing the changes of the frequency of axle box acceleration and

its energy distribution variation combined with the trains operating parameters, we can position the distribution of the track shortwave irregularity. It is beneficial for the track maintenance, and ensuring the safety of urban rail transit operation.

Acknowledgements

This work was supported by the Scientific Research Innovation Project of Shanghai Education Commission (12YZ149, 12ZZ184), Shanghai Municipal Natural Science Foundation (12ZR1412300), Key Technology R&D Project of Shanghai Committee of Science and Technology (13510501300) and Construction Project for Transportation Engineering(13SC002)

References

- [1] Bruni S, Goodall R M, Mei T X 2007 Control and monitoring for railway vehicle dynamics. *Vehicles System Dynamics* **45**(7-8) 765-71
- [2] Molodova M, Li Z, Dollevoet R 2011 Axle box acceleration: Measurement and simulation for detection of short track defects *Wear* **271**(1-2) 349-56
- [3] Bocciolone M, Caprioli A, Cigada A 2007 A measurement system for quick rail inspection and effective track maintenance strategy *Mechanical Systems and Signal Processing* **21**(23) 1242-54
- [4] Mori H, Tsunashima H, Kojima T 2010 Condition Monitoring of Railway Track Using In-service Vehicle *Journal of Mechanical Systems for Transportation and Logistics* **3**(1) 154-65
- [5] Huang N E, Shen Z, Long S R 1998 The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis *Proceedings of the Royal Society of London Series A Mathematical Physical and Engineering sciences* **454**(1971) 903-995
- [6] Chen Xianmai, Wang Lan, Tao Xiaxin. 2008 Study on the Judgment Method for Track Regularity of the Main Railway Lines in China. *China Railway Sciences*, **29**(4) 21-2

| Authors | |
|---|---|
|  | <p>Xingjie Chen, born in December, 1975, Qidong County, Jiangsu Province, P.R. China</p> <p>Current position, grades: Lecturer at the School of city railway transportation, Shanghai University of Engineering Science, China. University studies: B.Sc. in mechanical engineering at Nanjing University of Science and Technology, China. M.Sc. at Nanjing University of Science and Technology, China. Scientific interest: automatic control, image processing. Publications: more than 6 papers. Experience: Teaching experience of 7 years, 3 scientific research projects.</p> |
|  | <p>Xiaodong Chai, born in January, 1962, Shanghai, P.R. China</p> <p>Current position, grades: Professor at the School of Urban Railway Transportation, Shanghai University of Engineering Science, China. University studies: B.Sc., M.Sc. and Ph.D. in Electrical Engineering from Anhui University in China. Scientific interest: signal processing, hologram image processing. Publications: more than 30 papers. Experience: teaching experience of 25 years, 10 scientific research projects.</p> |
|  | <p>Xining Cao, born in November, 1988, Shanghai, P.R. China</p> <p>Current position, grades: Lecturer at the School of Urban Railway Transportation, Shanghai University of Engineering Science, China. University studies: B.Sc. in Automobile Service Engineering at Yancheng Institute of Technology in China. Scientific interest: track status detection. Publications: 1 paper in Instrument technique and sensors. Experience: teaching experience of 2 years, 1 scientific research projects.</p> |

Research into voltage sag online detection technology based on wavelet tree

Xianfeng Zheng^{1, 2*}, Zheng Fan²

¹Faculty of Electrical Engineering Henan Mechanical and Electrical Engineering College, No.699 Ping Yuan Road, Xinxiang, Henan, China

²Faculty of Automatic Control Engineering Henan Mechanical and Electrical Engineering College, No. 699 Ping Yuan Road, Xinxiang, Henan, China

Received 6 February 2014, www.tsi.lv

Abstract

A general process model is established using the real-time requirements of data stream processing, and the data is constantly processed with a sliding window. This paper selects the recursion-based complex wavelet as the detecting algorithm for voltage sag, and tries to detect when the voltage sag occurs and ends with amplitude and phase information contained in the wavelet analysis results. Meanwhile, this paper seeks to improve the precision of detection by looking for optimal wavelet scales with information entropy. The shifted wavelet tree-based data flow anomaly detection algorithm and data update method of shifted wavelet tree have been improved to make rapid detection possible. Finally, this paper reports the experimental simulation, which proved the instantaneity and accuracy of this method.

Keywords: Data Stream, Voltage Sag, Recursive Complex Wavelet Transform, Shifted Wavelet Tree

1 Introduction

Power quality is a common concern in certain parts of the world. Voltage sag - a major quality problems affecting the normal and safe operation of electrical equipment - is a fast short-term decline of voltage effective value caused by a system short circuit fault, overload or a large motor starting [1-2].

According to one survey on power supply: with the exception of outages, voltage sag is the greatest power quality issue, accounting for more than 80% of the problems. Voltage sag causes widespread equipment failure, causing economic losses, such as damage to or malfunctioning of electronic, computer and control equipment and other sensitive devices.

In response to market demand, accurate real-time detection of magnitude, duration and occurrence of voltage sag is required. At present, the most commonly used method is, according to the definition of continuous periodic signal effective value, to obtain sag amplitude by calculating voltage effective values, among which the voltage effective value can be obtained with digital RMS operation in one cycle of time domain [3-4]. As the

voltage sag problem becomes increasingly evident, an effective method is needed for a real-time detection. Accordingly, real-time detection of occurrence time/end time, duration, frequency of voltage sag is a primary subject of the anomalous voltage detection research.

2 Data stream processing model

A windowed process with a sliding window was developed in light of the very obvious temporal characteristics of voltage sag information detection, the special technical requirements of one access, limited storage, continuous processing, and rapid response during the real-time processing of the data stream formed by detection-related data, and the higher demands for instantaneity [5]. The deadline time of data processing is defined to ensure real-time constraints on data stream processing. The created process system model is shown in Figure 1.

Figure 1 shows that the system core is the data stream processing engine. Different processing algorithms should be adopted according to function. In this system, data flow anomaly detection is the main system function.

* Corresponding author e-mail: hnxzxf@126.com

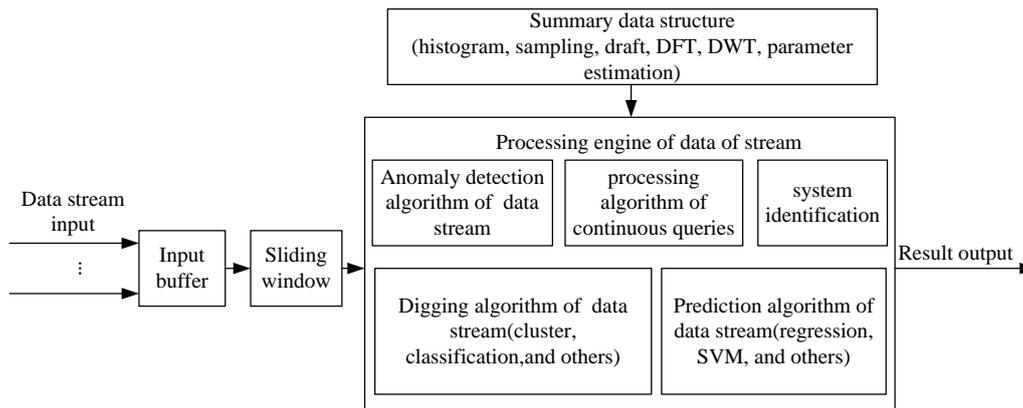


FIGURE 1 Data stream processing model

3 Detection of data flow anomaly

The essence of anomaly detection is to find significantly different data from a large number of data, and the definitions of anomalous data vary in different fields of application. The detection method we used for data stream anomaly with mobile wavelet tree data structure uses the SWT structure to do the anomaly detection, and is capable of detecting anomalies of different length.

In the anomaly detection process, firstly, the data stream is $x_i (i=1... n)$, the aggregation function H (including data maximum, minimum, average, sum, standard deviation, and etc.) can be obtained by calculation, and according to different wavelet resolution, several sliding windows w_j and corresponding threshold $H (w_j)$ are set, and then compared with the threshold value, we screen out the anomalous data so that the aggregation function is greater than the corresponding threshold. The algorithm, based on the anomalous duration, detects corresponding wavelet levels. Once the value of a window in this level is found to exceed the predefined threshold, then the corresponding lower window search is continued, until the anomalous position is found.

In the above process, the data in different levels should use a different sliding window length and different wavelet scale factor, so that the time series is decomposed into a multi-level wavelet tree structure, in which the original data sequence consisting of zeroth layer wavelet decomposition tree (Level0), wherein the average and difference of adjacent data constitutes the first layer (Level1) wavelet coefficients. And then the process repeats itself, the wavelet coefficients of the next layer will be obtained through successive recursion, until only one average value and difference are left in the top layer. Therefore, the wavelet coefficients in wavelet tree contain the average value and difference of each layer, and the original signal can be reconstructed from these coefficients without distortion [6].

For a specific data aggregation, each definite anomalous length can be detected at a corresponding level. If it is over the threshold, it can be assumed that each data aggregation contained in this window exceeds the threshold, and concluded that the data includes an

anomaly. The anomaly detection algorithm filters windows that are found with no anomaly, and does not detect it. Instead, this detection algorithm only detects a few windows aggregations that are over the threshold, which thus greatly reduces the detection range and improves detection efficiency, thereby making the detection omission rate relatively low. And by changing the length of elastic windows, we can carry out anomaly detection of an arbitrary length of data flow.

The summary data of the above aggregation function is obtained through a wavelet tree and a real-time complex wavelet algorithm; an improved recursive wavelet is used here to process voltage sag information [7].

First, the selected base wavelet is equation (1)

$$\varphi(t) = \left(-\frac{\sigma^3 t^3}{3} - \frac{\sigma^4 t^4}{6} - \frac{\sigma^5 t^5}{15} \right) e^{(\sigma + j\omega_0)t} u(-t). \tag{1}$$

Here, $\sigma = 2\pi/\sqrt{3}$; $\omega_0 = 2\pi$; at this point, $\varphi(0) = 0$, which ensures that the selected base wavelet satisfies the admissibility condition. The base wavelet can be represented in the frequency domain

$$\psi(\omega) = \left[\frac{6\sigma^5 - 2\sigma^3(\omega - \omega_0)^2}{\sigma + i(\omega - \omega_0)} \right]^*. \tag{2}$$

Discretizing the above expressions, we can get equation (3)

$$\varphi(fnT) = \left(-\frac{\sigma^3 (fnT)^3}{3} - \frac{\sigma^4 (fnT)^4}{6} - \frac{\sigma^5 (fnT)^5}{15} \right) e^{(\sigma - \omega_0)fnT} u(-fnT). \tag{3}$$

Among them, $f=1/a$, a is the wavelet scale factor, T is sampling period. After transformation and sorting out via Z , the following equation (4) can be obtained

$$\Psi(Z) = \frac{\delta_1 Z^{-1} + \delta_2 Z^{-2} + \delta_3 Z^{-3} + \delta_4 Z^{-4} + \delta_5 Z^{-5}}{1 + \lambda_1 Z^{-1} + \lambda_2 Z^{-2} + \lambda_3 Z^{-3} + \lambda_4 Z^{-4} + \lambda_5 Z^{-5} + \lambda_6 Z^{-6}}. \tag{4}$$

Among them, $\delta_n = [a_n(\sigma fT)^3 + b_n(\sigma fT)^4 + c_n(\sigma fT)^5]A^n$, a_n, b_n, c_n is the coefficient; after calculation, the following can be obtained $a_1=1/3; a_2=2/3; a_3=-2; a_4=2/3; a_5=1/3; b_1=-1/6; b_2=-5/3; b_3=0; b_4=5/3; b_5=1/6; c_1=1/15; c_2=26/15; c_3=22/5; c_4=26/15; c_5=1/15$. $A = e^{-fT(\sigma-j\omega_0)}$, $\lambda_1=-6A; \lambda_2=15A^2; \lambda_3=-20A^3; \lambda_4=15A^4; \lambda_5=-6A^5; \lambda_6=A^6$.

And available expressions

$$W_{s,\phi} = \sqrt{fT}[\delta_1s((k-1)T, f) + \delta_2s((k-2)T, f) + \delta_3s((k-3)T, f) + \delta_4s((k-4)T, f) + \delta_5s((k-5)T, f)] - \lambda_1W_{s,\phi}((k-1)T, f) - \lambda_2W_{s,\phi}((k-2)T, f) - \lambda_3W_{s,\phi}((k-3)T, f) - \lambda_4W_{s,\phi}((k-4)T, f) - \lambda_5W_{s,\phi}((k-5)T, f) - \lambda_6W_{s,\phi}((k-6)T, f) \quad (5)$$

According to different application requirements, the signal, by decomposing wavelet packet subspaces, can be decomposed into a high frequency part and a low frequency part as the summary data signal with the above wavelet decomposition.

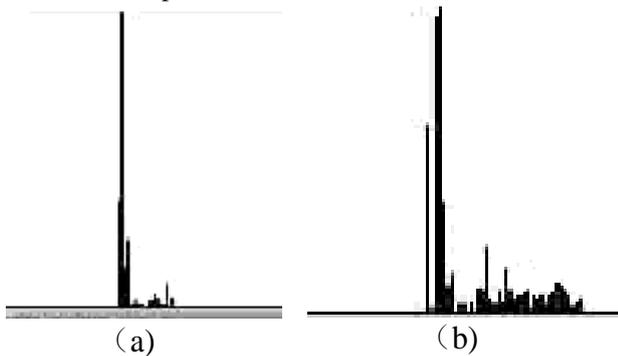


FIGURE 2 The wavelet analysis results of the end point of voltage sag (a) $f=3000$ (b) $f=5000$

4 Establishment and search of wavelet error tree

In certain data stream applications (e.g., correlation analysis) only a low frequency coefficient is needed for synoptic data of the data stream and analysis. In other applications (such as power quality disturbance identification), detailed coefficients are required for summary purposes [8]. The algorithm must be able to handle the real-time, continuous, unlimited data stream; accordingly, the wavelet tree is constructed as shown in Figure 3.

Based on the wavelet tree shown above, the source data can be reconstructed, and the range, threshold value and other parameters will also be rebuilt; thereby aggregation query will be achieved.

In the actual detection process, because the anomaly detection algorithm of the data stream based on the shifted wavelet tree needs to simultaneously detect the entire length of the sliding windows, which takes much time and space, causing deterioration of system

performance; therefore this search method need to be optimized [9]. The binary search method adopted here can greatly reduce system overhead.

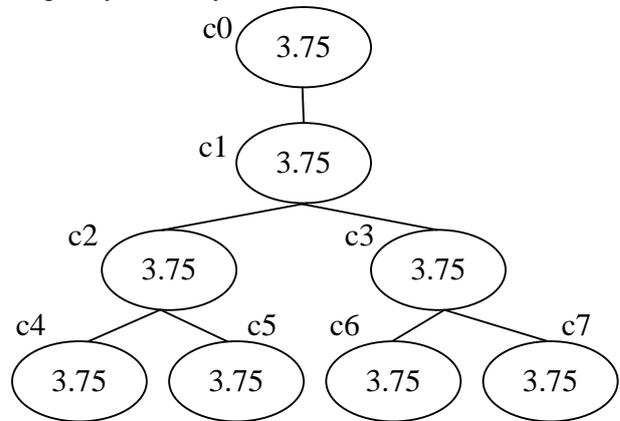


FIGURE 3 Wavelet Tree Structure Figure

First, the monotonic search space will be established, and all the sliding windows, which need to be detected are sorted according to length. Based on this, a binary search is carried out to find the largest anomaly window length.

According to the methods above, and considering the fact that the given data may include anomaly data, whose length is W_i (where in, $i=1,2,\dots, m$, m is various window length of abnormal data to be detected), the monotonic search space will be constructed first, that is to sort space (W_1, \dots, W_m) according to detection window size. And then, the binary search is carried out on the basis of this space sequence. First, the detection begins from the intermediate position window of entire search space. If the anomaly is found, then the latter half part will be searched. If no anomaly is found, then the former will be searched. Then the process is repeated ceaselessly, until the anomaly window is found. And then, the anomaly starting position will be located in the window.

5 Data pre-processing

In order to eliminate the jolts data and improve detection accuracy, the original data stream should be properly pre-processed first [10].

The aim of data pre-processing is to eliminate the jolts data and improve detection accuracy. The jolts data are data in the data stream that reach certain intensity, which, however, is not enough to become anomaly data. The jolts data of different window length has a varied effect on detection results. Some is not considered anomaly data in a small window. When an anomaly detected in a larger window, these cumulative small changes are likely to be mistaken as an anomaly.

In order to improve the accuracy of anomaly detection, the ratio threshold anomaly detection method is adopted here to remove the interference of jolts data. The ratio threshold, using the recent two equal length window aggregation ratio in the data stream, determines whether the anomaly has occurred. It is defined as follows

$$(x_{w+1} + \dots + x_{2w}) > \beta(x_1 + \dots + x_w) \quad \beta > 1, \quad (6)$$

$$(x_{w+1} + \dots + x_{2w}) < \beta(x_1 + \dots + x_w) \quad 0 < \beta < 1. \quad (7)$$

In the above formula, β , corresponding to the upper limit threshold value and lower limit threshold value, is determined by the ratio of the two same windows before the current time point.

In practical application, if no anomaly occurs in the data stream, it is necessary to consider the possibility that the size of the time window might lead to a false anomaly and try to prevent it from happening. Since the large window anomaly caused by jolts data does not appear in a small window, therefore, the window whose length is 1 should be detected first with the above ratio threshold method. If no anomaly occurs in this window, then it will not be considered as an anomaly in the large window, which reduces the probability of misinformation.

After processing the jolts data, no anomaly caused by jolts data appears in the large window, so the maximum length of the anomaly sliding window is the actual length of the detected anomaly window.

6 Update of data of shifted wavelet tree

With the continuous updating of the data stream in the sliding window, calculation of the DWT coefficient of the data stream in the sliding window is done in two ways: one is direct convolution calculation of all the data in the sliding window, which is called direct update; the other is incremental update, that is, incremental calculation on the entire sliding DWT coefficient with the DWT coefficients before window sliding and the newly arrived data. The incremental update algorithm can avoid the disadvantages of the traditional algorithm; the wavelet summary data structure in the sliding window need not be rebuilt when new data comes into the sliding window, thus improving time cost.

The direct update algorithm can ensure accurate completion of each node update when new data arrive. But taking into account the window sliding, and as long as the update data length is less than the sliding window size, data redundancy exists before and after sliding, and the direct update of reconstruction of the wavelet summary reduces processing efficiency.

When receiving new data in the sliding window, it is only necessary to carry out wavelet decomposition on new arrival data with the incremental update method, and the other part can be obtained by shifting the wavelet decomposition results before window sliding. This avoids time overhead of re-computing. Compared with the direct update algorithm, the short calculation time of the SWAT algorithm is particularly prominent in the analysis application of a large data stream. One limitation is when only half of the new data of the sliding window arrives; the summary has to be completely updated again. And

when the amount of new arrival data is less than half of the sliding window, the summary data structure cannot accurately represent the window data, thus precision cannot be guaranteed in some applications. Another limitation is that because an update operation is carried out during one unit of time, the large data stream and frequent update operation need a certain time overhead in practical application, which affects process efficiency.

Considering the limitation of the above two kinds of incremental updating algorithm, an improved algorithm is proposed in this paper, whose main idea is as follows: the sliding window is divided into several basic windows (sub windows), wavelet approximation trees are constructed respectively in each sub window and transition wavelet approximation trees (similar to the intermediate nodes) are built between the adjacent two sub windows. When new data arrives, it is not only updated in each tree node, but also updated between trees. Therefore, the complete update of the entire wavelet summary only needs the data of half the amount of the basic window length.

After analysis, the update process of the improved algorithm can be divided into two cases.

(1) Update of internal nodes in the first tree. It can be further divided into two types: node update in the same layer. After one update cycle, some node value is shifted to the left node, and the right node value is shifted to this node; to achieve the node update between different layers, as for the lower node, the value of the upper node located in the left and right sides of this node can be calculated, the result is stored in this node.

(2) Update between different trees. Like the node update inside the tree, the entire tree achieves dynamic incremental update through shifting.

The specific approach is that we add more redundant data windows to the wavelet tree, and store one summary data set for 2^i number data. $w-2^i+1$ summary data is saved in each layer, and w is the length of the selected sliding window. According to the real time incremental updating algorithm, incremental update involves carrying out a small amount of data summary update in each wavelet level when one data set arrives. It is in the i layer of the wavelet tree where one summary data set for 2^i number data is saved. Whenever new data arrives, the summary corresponding to the wavelet level of the oldest data as the starting point is expired and, at the same time, the upper data summary of the new data as the end point is generated. So when one data set arrives, a small amount of the summary corresponding to wavelet level is updated, which reduces the time complexity of algorithm detection. The real time continuous updating algorithm can detect the anomaly more quickly and accurately, meeting the requirement of real-time detection of data stream and improving the efficiency of anomaly detection.

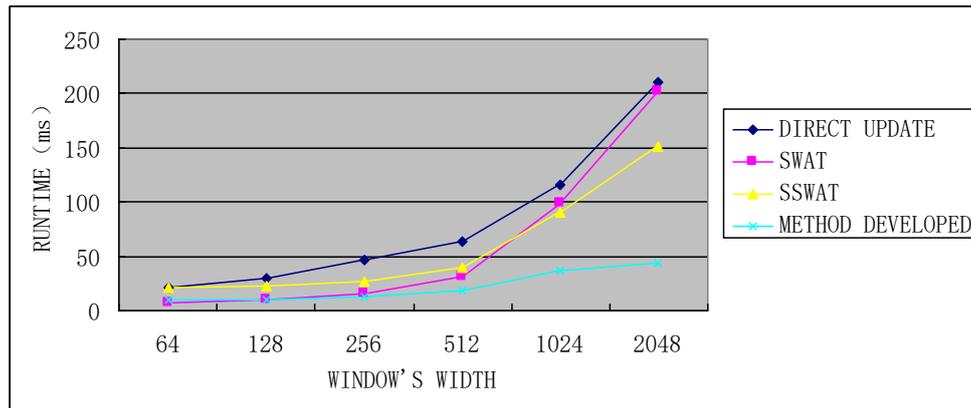


FIGURE 4 Running Time Ratio of updating algorithm for All Kinds of Data

The full update can be completed when two new data sets (half-length of base window) arrive in the improved algorithm. Compared with the foregoing two traditional algorithms, the necessary data length of completing one full update is shorter, and suitable for higher requirements on the accuracy and speed of data stream processing system.

The specific data is shown in Figure 4. It can be seen that the required time of the SWAT algorithm to complete one update is much shorter than a direct update, but the time of completing one update is equal to half the sliding window size, which is not suitable for the case of: (i) an update interval of less than one-half of the sliding window size, or (ii) an accuracy requirement.

When the sliding window is small, the time it takes for SWAT to be completely updated should be less than SSWAT. But when the sliding window length is greater than 1024, the time required for SSWAT is less than for SWAT, and real-time data streams are relatively large. Therefore, SSWAT is more suitable for high-speed data stream processing.

Compared to SSWAT, the improved algorithm only needs one half of the base window to complete the full update, which is suitable for the application of a small time interval of output result, with better accuracy than SSWAT.

Given that the implementation of wavelet decomposition is extraction from convolution of the input sequence and wavelet decomposition filter, so when the length of wavelet decomposition filter coefficients is too long and the input sequence is limited, it will cause

boundary distortion, and the above discussed wavelet decomposition incremental update algorithm will also sometimes cause boundary distortion. The solution is to adopt extension, such as zero extension, periodic extension, symmetric extension and so on. In this case, the wavelet decomposition that results after window shifting cannot be simply obtained by the above method of shifting. As to the edge part, we will obtain it by direct convolution calculation, while the middle part can be obtained by the wavelet decomposition results before window shifting.

7 Example analysis

The anomaly detection algorithm of the data stream with improved shifting wavelet tree carries out detection on simulation signal which involves voltage sag, and compares it to the common error tree algorithm to show the relative effectiveness of this method.

First, the typical power quality disturbance signal model is established, and the signal is generated in Matlab, which includes four common kinds of transient signal in an electric power system (voltage swell, voltage sag, and voltage interruption). In order to simulate an actual situation, the parameters of voltage sag, voltage swell and interruption are allowed to vary randomly within the permitted range (the parameters that characterize the disturbance signal fluctuate randomly in a certain range), and random white noise of 10-30dB noise ratio is added to them [11-12].

TABLE 1 Disturbance signal model

| | | |
|-----------------------------|---|--|
| Voltage swell | $0.1 \leq \alpha \leq 0.8, T \leq t_1 - t_2 \leq 9T$ | $v(t) = A(1 + \alpha(u(t_2) - u(t_1)))\sin(\omega t)$ |
| Voltage sag | $0.1 \leq \alpha \leq 0.8, T \leq t_1 - t_2 \leq 9T$ | $v(t) = A(1 - \alpha(u(t_2) - u(t_1)))\sin(\omega t)$ |
| Voltage interruption | $0.9 \leq \alpha \leq 1, T \leq t_1 - t_2 \leq 9T$ | $v(t) = A(1 - \alpha(u(t_2) - u(t_1)))\sin(\omega t)$ |
| Harmonic | $0.05 \leq \alpha_3(\alpha_5, \alpha_7) \leq 0.15, \sum \alpha_i^2 = 1$ | $v(t) = A(a_1 \sin(\omega t) + a_3 \sin(3\omega t) + a_5 \sin(5\omega t) + a_7 \sin(7\omega t))$ |

In the analysis process, the input signal analysis time length adopts 10 sine wave cycles (0.2S), 6.4kHz sampling rate, 50Hz voltage frequency. Given the noise condition, the accuracy of detection result is shown in Figure 5.

From the results of the table above, the accuracy of the shifting wavelet tree algorithm is much better than that of the error tree algorithm. Lower accuracy of detection results when the detection is affected by noise.

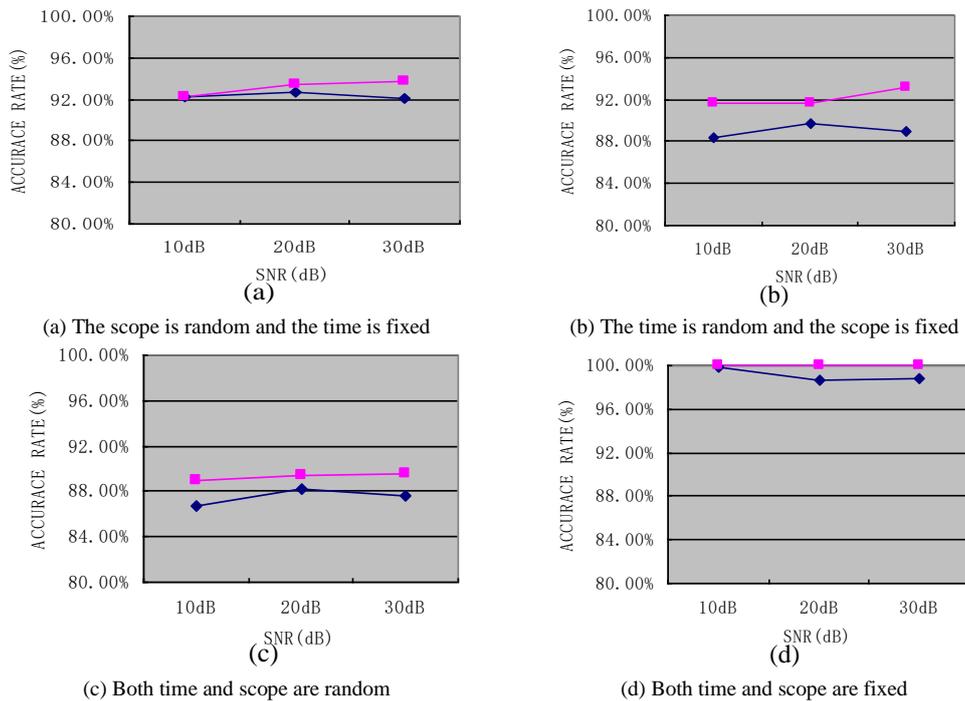


FIGURE 5 Data table of detection result under different disturbances

As for the required time of detection, under the same condition, one time of detection needs 0.8ms with the algorithm of the shifting wavelet tree, while the error tree algorithm needs about 1.2ms. The shifting wavelet tree algorithm is obviously better than the error tree algorithm.

8 Conclusion

By using an improved detecting method of data stream based on the shifted wavelet tree data structure, we constructed a wavelet tree data model and reduced the time complexity with binary search error of the wavelet tree. We designed a voltage sag detecting algorithm based on the adapted recursive wavelet, which is capable of

rapid and precise detection of the start-stop of voltage sag. Because it is simple and effective - not at the cost of precision - this method guarantees instantaneity of the system.

In addition, with this real-time incremental updating algorithm to detect wavelet tree data, we can meet the requirement of anomaly detection, and improve detecting efficiency and instantaneity, only through updating a small amount of data summary of the relevant wavelet levels. Analysis of the actual example proved that this algorithm, with advantages like low time consuming, high real-time and so on, provides accurate results for the range, duration and frequency of voltage sag, offering a new method for detecting voltage sag.

References

- [1] Babu S, Widom J 2001 Continuous queries over data streams *ACM SIGMOD Record* **30**(3) 109-20
- [2] Chakrabarti K, Garofalakis M, Rastogi R, et al 2001 Approximate query processing using wavelets *VLDB Journal* **10**(2-3) 199-223
- [3] Gerbec D, Gasperic S, Smon I, Gubina F 2005 *IEEE transactions on power systems* **20**(2) 548-55
- [4] Guha S, Meyerson A, Mishra N, Motwani R, O'Callaghan L 2003 *IEEE Transactions on Knowledge and Data Engineering* **15**(3) 515-28
- [5] Barbara D 2003 Requirements for clustering data streams *In Proceedings of ACM SIGKDD Explorations Newsletter* **3**(2) 23-27
- [6] Abdel-Galil T K, Kamel M, Youssef A M, et al 2004 Power quality disturbance classification using the inductive inference approach *IEEE Transactions on Power Delivery* **19**(4) 1812-7 (In chinese)
- [7] Fan Zheng, Tian Xiaowu, et al 2008 Design of virtual instrument for voltage sag detection based on recursive complex wavelet transform *Electric power automation equipment* **28**(6) 100-2 (In chinese)
- [8] Fan Zheng, Tian Xiaowu, et al 2008 Design of virtual instrument for voltage sag detection based on recursive complex wavelet transform *Electric power automation equipment* **28**(6) 100-2 (In chinese)
- [9] Figueiredo V, Rodrigues F, Vale Z, Gouveia J B 2005 *IEEE Transactions on Power Systems* **20**(2) 596-602
- [10] Wang Yongli, Zhou Jinghua, Xu Hongbing, Dong Yisheng, Liu Xuejun 2007 Adaptive prediction of time series data stream *Journal of automation* **2**(33) 197- 201 (In chinese)
- [11] Monedero I, Leon C, Roperio J, Garcia A, Elena J M, Montano J C 2007 *IEEE Transactions on Power Delivery* **22**(3) 1288-96
- [12] Wu Shanshan, Gu Yu, Lv Yanfei, Yu Ge 2007 Sliding window processing strategy of sensitive deadline in data stream *Computer science* **34**(7) 99-102 (In chinese)
- [13] Liu Lianguang 2008 Achieving accurate measurement of voltage sag with wavelet transform and RMS algorithm *Automation of electric power systems* **27**(11) 30-3 (In chinese)

| Author | |
|---|--|
|  | <p>Zheng Xianfeng, born in March, 1972, Xinxiang County, Henan Province, P.R. China</p> <p>Current position, grades: the Associate Professor of Henan Mechanical and Electrical Engineering College, China. University studies: received his B.Sc. in Industrial Electrical Automation from Shanxi Institute of Mining Technology in China. He received his M.Sc. from Xi'an Jiaotong University in China. Scientific interest: His research interest fields include measurement and control technology and electrical insulation testing. Publications: more than 31 papers published in various journals. Experience: He has teaching experience of 18 years, has completed ten scientific research projects.</p> |
|  | <p>Fan Zheng, born in June, 1973, Xinxiang County, Henan Province, P.R. China</p> <p>Current position, grades: the Associate Professor of Henan Mechanical and Electrical Engineering College, China. University studies: received his B.Sc. in Electrical Engineering and Automation from Yan Shan University in China. He received his M.Sc. from Xi'an Jiaotong University in China. Scientific interest: His research interest fields include computer measurement and control technology, electrical testing and analysis of power quality. Publications: more than 28 papers published in various journals. Experience: He has teaching experience of 19 years, has completed five scientific research projects.</p> |

Modelling and simulation of marine rudder system in a unified M&S platform

Chang Chen^{*}, Guojin Chen, Shaohui Su, Haiqiang Liu

School of Mechanical Engineering, Hangzhou Dianzi University, 310018, Hangzhou, China

Received 18 June 2014, www.tsi.lv

Abstract

For modelling and simulating of marine rudder system, there are lots of along with their model libraries, such as AMEsim could be used. But the models in these tools lack of flexibility and are not open to the end-user. And these tools could not model the whole marine rudder system consisted of mechanical, hydraulic and control sub-system in a unified form. In order to solve those problems, a flexible and extensible marine rudder system library was constructed, based on the Modelica, by the object-oriented strategy. It supports the reuse of knowledge on different granularities: physical phenomenon, component model and system model. A conventional model of marine rudder system was built and calculated using the library, and the results shows that the object-oriented modelling strategy is effective; the framework of the library is reasonable.

Keywords: Marine Rudder System, Modelica, M&S Unified Platform, Object-oriented modelling strategy

1 Introduction

Marine rudder system, which is the actuator of operation control system, is important part of ship. Its basic task is accurately turning the rudder according to the given rudder angle. Electronic hydraulic position valve system colligating the advantages of electric and hydraulic is the typical marine rudder system. It has advantages of high control accuracy, fast response speed, flexible signal processing, etc. Up to present, lots of modelling platform could be used to design marine rudder system and analyse its performance in single domain. Performance analysis of steering control system could be realized with Matlab/Simulink. The AMEsim could be used to design and analyse the hydraulic sub-system of marine rudder system [1]. However, the whole marine rudder system is difficult to model and simulate because of the shortcoming as follows.

(1) The model is commonly described as a black box which users just know how to use but don't know the details of. So it is very difficult to modify an existing model and introduce a new one in these simulators. In another word, the models lack of flexibility and are not open to the end-user.

(2) And the worse is the models developed in one tool cannot be easily used in another one because of the different model description, translation, data organization etc. among these simulators.

(3) The model is usually expressed in an explicit state-space form and iterative process of model solution has to be given out if the explicit expression could not be found. Consequently, the topology of the system gets lost

and any future extension and reuse of the mode is also tedious and error-prone.

Fortunately, among the last dozen years' research in modelling and simulation, there are two concepts that closely related to these problems: object oriented modelling language and non-causal modelling. Modelica [2-4], developed in an international effort, is such a kind of physical system modelling language. It supports encapsulation, composition and inheritance facilitating model development and update. Elementary models of physical elements are defined in the declarative expression by their constitutive physical principles, and their interface with the outer world is described by physical connectors without any implied causality, rather than by writing assignments relating inputs to outputs. This makes the description of physical systems much more flexible and natural than it is possible with causal or block-oriented modelling languages, or by directly writing simulation code using procedural languages such as C or FORTRAN. Complex models can then be built by connecting elementary models through their ports. The ports are non-causal, so any connection which is physically meaningful is allowed without restrictions. The Modelica is applied to the modelling and simulation of vehicle [5-6], craft [7], machine [8], fluid system [9-10], control system [11] and so on.

The Modelica language includes graphical annotations, which allow to use graphical user interfaces (such as the one provided by the tool MWorks [12]) to select components from a library, drag them into a diagram, connect them, and set their parameters, thus making the process of model development highly intuitive for end users. In order to minimize the

^{*} *Corresponding author* e-mail: chenchang@hdu.edu.cn

development time, Modelica allows the definition and reuse of intermediate elements, common to more models of the same component. The key feature to extensive model reusability and flexibility is given by the extensive exploitation of two Modelica constructs: extend and redeclare, and their associated keywords: partial and replaceable. The partial/extends binomial permits the extension of partial (i.e., partial-complete) models into complete, fully detailed models.

In this article, a Marine Rudder System library was built based on the novel multidisciplinary physical modelling language Modelica and the Modelica-supported tool MWorks, aiming at providing a framework of Marine Rudder System model library with the object oriented technology and equation-based model description.

2 Structuring idea of library

2.1 OBJECT-ORIENTED MODELLING STRATEGY

Traditionally, a physical system model development has followed the top-down design approach, which applies the method of functional decomposition to establish the model structure. This method has been successfully applied in many physical engineering domains, but it fails to reflect the real world. As a result many attempts have been made to tackle this problem by applying object-oriented technology.

So the library of Marine Rudder System is designed as an object-oriented work, which is a set of classes that embodies an abstract design for solution to a family of related problems. The set of classes define “partial-complete” application that captures the common characteristic of object structures and functionality. Specific functionality in new applications is realized by inheriting from, or composing with, framework components. In this paper, new object-oriented modelling language Modelica and a Modelica-supported tool MWorks are used to develop the model library which can be later used to assemble system-level Marine Rudder System model.

During the simulation, the system’s mathematical model is mapped to collections of interacting objects, rather than decomposed into segments of different functions that can implement certain algorithms. Each object mimics the behavioural and structural characteristic of a physical or conceptual entity models. And it represents an instance of a software class, while the classes are united into a hierarchy via inheritance relationships.

According to the modelling idea mentioned above, the Marine Rudder System is decomposed into the units according to the physical reality and further the process units are decomposed into the basic physical phenomenon, such as conservation of mass or of energy, as the left branch shows in Fig. 1. Afterwards, models of basic physical phenomenon, unit and system are built

level by level using language Modelica and tool MWorks, as the right branch shows in Fig.1. As a result, the different levels’ models: from basic physical phenomenon models, over physical unit modes, up to system models could be reused flexibly in the process of constructing a new model. So the professional engineering knowledge can be solidified, propagated and reused in varied granularity.

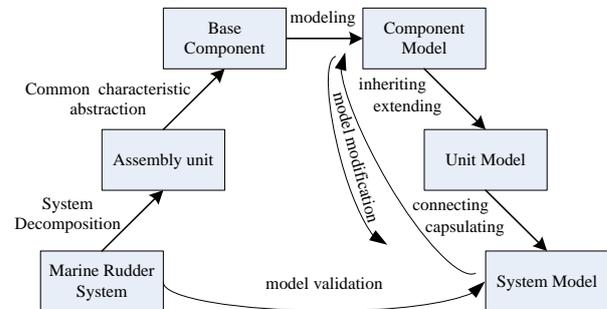


FIGURE 1 System decomposing and modelling with Modelica and Mworks

2.2 MODEL LIBRARY FOR MARINE RUDDER SYSTEM

Following the level progressive modelling strategy mentioned above and considering the maximum reuse of codes in the library, the Marine Rudder System could be divided into several sub-systems. Each sub-system consists of various components that were created through inheriting and expanding the basic physical phenomenon models. Fig.2 presents the architecture of the library. The library consisted of fourteen parts.

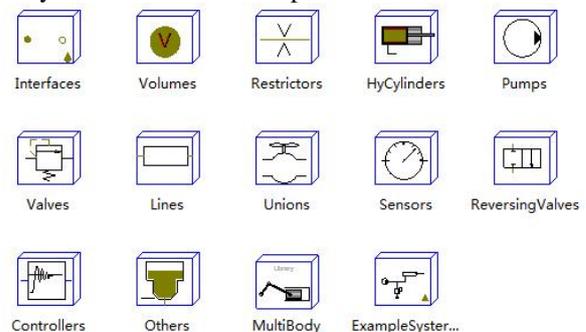


FIGURE 2 Structure of model library for marine rudder system

Interfaces: the connectors among the components are defined here.

Volumes: package of tank and boundary in hydraulic system.

Restrictors: package of restrictors.

HyCylinders: package of hydraulic cylinders.

Pumps: package of pumps, such as centrifugal pump, constant pressure pump and so on.

Valves: package of valves, like check valve, dropping valve, etc.

Lines: the pipes with lumped parameter and distributed parameter method.

Unions: connection like bend, buffer valve.

Sensors: sensors like pressure sensor, flow sensor.

ReversingValves: multi-channel directional control valve consist of models in package Valves.

Controllers: controller model like PID, PI.

Others: filter.

MultiBody: models of mechanical part, nsuch as rudder blade.

ExampleSystems: demos of marine rudder system.

3 Common Models

3.1 INTERFACE

A special-purpose class connector as an interface defines the variables of the model shared with other models, without implied causality, rather than by writing assignments relating inputs to outputs [13]. In this way the connections can be, besides inheritance concepts, thought of as one of the key features of object oriented modelling, enabling effective model reuse. There are two types of built-in variables: potential variable and flow variable with no prefix, the prefixes flow respectively.

The potential and flow variables follow the Generalized Kirchhoff's law. For examples, the connector Port for hydraulic oil flow in Marine Rudder System has two variables as follows: pressure p, volume flow rate q.

```
connector Port
SI.AbsolutePressure p;
flow SI.VolumeFlowRate q; // "positive is flowing from outside
into the component "
end Port;
```

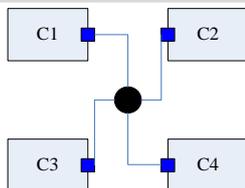


FIGURE 3 Connecting of four components

Figure 3 shows the connection of four components that has the connector Port. The connection point is treated as an infinitely small control volume. All the pressures in each connector must be equalized while the sum of all the molar flows should be zeros:

$$C_1 \cdot p = C_2 \cdot p = C_3 \cdot p = C_4 \cdot p, \tag{1}$$

$$C_1 \cdot q + C_2 \cdot q + C_3 \cdot q + C_4 \cdot q = 0. \tag{2}$$

During the model translation, Eq. (1) and Eq. (2) originating from the connector definitions, are automatically generated and added to the other equations of the model.

3.2 PARTIAL MODEL

Partial model is a kind of semi-complete component which abstracts the common character shared by a group

of models that have some common properties and behaviours or the same structure. It is a basic section of the component-model sub-library. For example, the pipe, restrictor and pump that could be considered as models include two Ports (one for inlet and another for outlet) as presented in section 3.1. They have the common properties and behaviours as follow:

$$q_{in} + q_{out} = 0, \tag{3}$$

where, q_{in} means the mole flow rate of inlet; q_{out} the mole flow rate of outlet.

The pressure drop equation

$$p_{in} - p_{out} = \Delta p, \tag{4}$$

where, p_{in} means pressure of inlet; p_{out} pressure of outlet; Δp the pressure loss between two ports.

So a partial model *OneInOneOut* was built to describe the properties and physical behaviours shared by these three models and other similar models as follow:

```
partial model OneInOneOut
Port In"inlet";
Port out"outlet";
SI.Pressure ploss" pressure loss between inlet and outlet";
equation
In.q+Out.q =0; //mass balance
In.p-Out.p =ploss; // pressure drop equation
end OneInOneOut;
```

At the beginning of the codes mentioned above, the prefix 'partial' means model *OneInOneOut* is semi-complete. When a higher level model like pipe is wanted to be introduced, *OneInOneOut* will be inherited and expanded. So the embedded engineering knowledge could be reused on physical behavioural level but not only on model level. In the "equation" region, the behaviours described in Eqs. (3-4) are coded in non-causal expression. When the model is needed to be simulated, MWorks will give out the calculation sequence automatically.

```
model pipe
extends OneInOneOut; // extend from model OneInOneOut
parameter Real m,n;
equation
ploss = m*q^n;
end pipe;
```

3.3 COMPONENT: PIPE AS AN EXAMPLE

In this section, a pipe model will be taken as an example to illustrate how to use a partial model to build a complete component model. Besides the information included in *OneInOneOut*, an equation describes relation between the pressure loss and volume should be add.

$$\Delta p = m \cdot q^n, \tag{5}$$

where, m and n are parameters obtained from empirical data. The follows codes show how to build a complete pipe model through inheritance and expansion.

3.4 ASSEMBLY UNIT

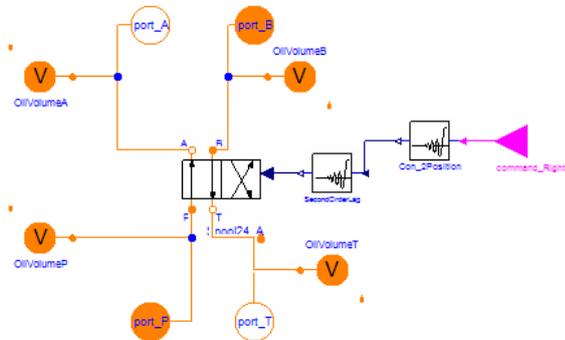
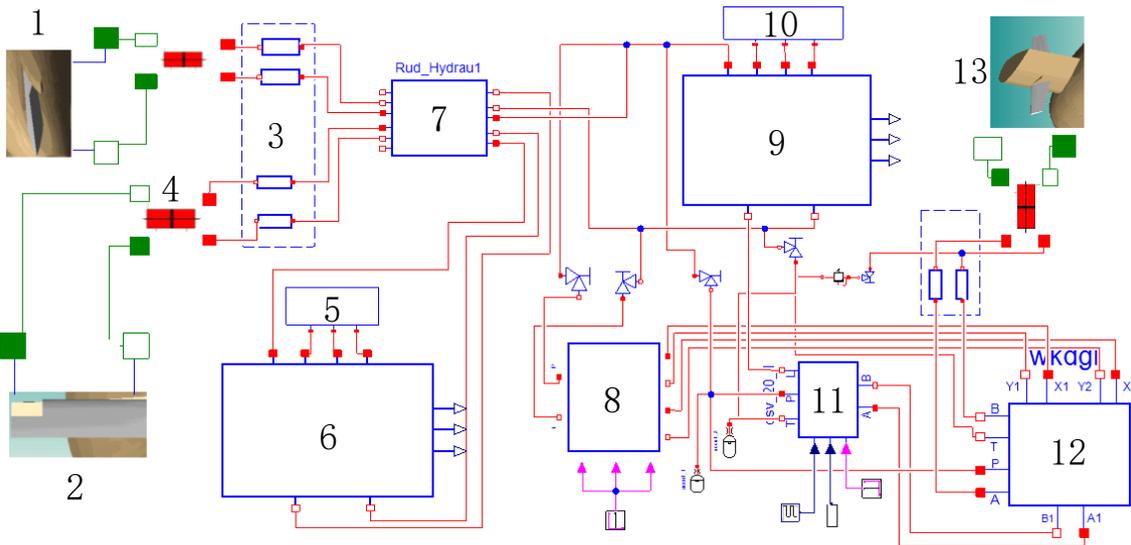


FIGURE 4 the diagram of two-position-four-channel reversing valve Assembly unit means the assembly unit made up of component model mentioned in section 3.2. According to the practical topological relation, an assembly unit model is constructed by connecting several related component

model. As shown in Figure 4, a two-position-four-channel electro-hydraulic reversing valve contains a valve body, a controller, one signal connector, four f ports, four control volumes and the connections between these components.

4 AN ILLUSTRATIVE EXAMPLE

In this section, a marine rudder system model of a typical submarine is constructed using the library graphically. As shown in Figure 5, the system model contains mechanical sub-system and electro hydraulic servo sub-system. The mechanical system is consisted of rudder, elevator and fairwater plane and the hydraulic system is consisted of pump station, energy accumulator, silencer, air system, steering gear hydraulic system, emergency manual valve group, electro-hydraulic servo valve and so on. Respectively, Figure 6 and Figure 7 show the pressure response of drainage port of DSV-20 and pressure on outlet port of pump station when the control input of DSV-20 is a pulse signal.



1-rudder, 2-elevator, 3- silencer, 4-actuator, 5-air system, 6-pump station, 7-steering gear hydraulic system, 8-emergency manual valve group, 9-pump station of ship, 10-air system, 11-electro-hydraulic servo valve (DSV-20), 12-insulated hydraulic valve group, 13- fairwater plane
FIGURE 5 Marine rudder system model of a typical submarine

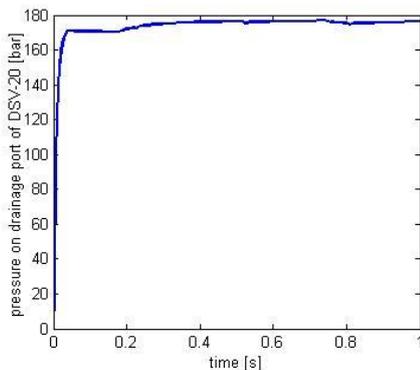


FIGURE 6 Pressure on drainage port of DSV-20

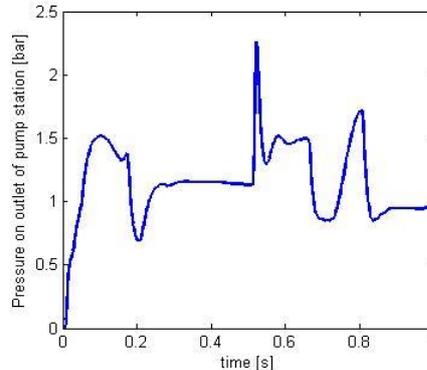


FIGURE 7 Pressure on outlet port of pump station

5 Conclusions

The whole marine rudder system is difficult to model and simulate in a unified platform because of the shortcoming as follows.

Model is described as a black box which is not open to the end-user.

Model in different simulators could not be shared with each other.

Only supporting reuse of knowledge on unit model level.

In this paper, we tried to utilize the multidisciplinary physical modelling language Modelica to build an object-oriented modular marine rudder system library in the Modelica-supported tool MWorks. The aim is to explore the modelling methods and implement of marine rudder

system model based on the object-oriented technology and to provide a marine rudder system library. The library mainly includes restrictor models, hydraulic cylinder models, pump models, pipe models, valve models and son on. A model of marine rudder system of submarine shows that the object-oriented modelling strategy is effective; the framework of the library is reasonable.

Acknowledgements

The paper was supported by Science and Technology Plan Projects of Zhejiang Province, China (2010R50003), Zhejiang province education department scientific research projects (Y201327060) and the National Natural Science Foundation of China (51305113).

References

- [1] Xiao D Z 2007 *Ship Science and Technology* 29(S1) 142-5 (In Chinese)
- [2] Elmqvist H, Mattsson S E 1997 *ESS'97 European Simulation Symposium* Passau, Germany 19-22
- [3] Tiller M 2001 *Introduction to Physical Modeling with Modelica* Kluwer Academic Publishers: Boston
- [4] Mattsson S E, Elmqvist H, Otter M 1998 *Control. Eng. Pract* 6(4) 501-10
- [5] Andreas D, Johannes G 2011 *Proceedings of the 8th International Modelica Conference* Linköping University Electronic Press 13-7
- [6] Andreasson J 2011 *Proceedings of the 8th International Modelica Conference* Linköping University Electronic Press: Sweden 414-20
- [7] Looye G 2008 *Proceedings of the 6th International Modelica Conference* Press: The Modelica Association 193-202
- [8] Dressler I, Schiffer J, Robertsson A 2009 *Proceedings of the 7th International Modelica Conference* Linköping University Electronic Press: Sweden 261-9
- [9] Viel A 2011 *Proceedings of the 8th International Modelica Conference* Linköping University Electronic Press: Sweden 256-65
- [10] Casella F, Sielemann M, Savoldelli L 2011 *Proceedings of the 8th International Modelica Conference* Linköping University Electronic Press: Sweden 86-96
- [11] Baur M, Otter M 2009 *Proceedings of the 7th International Modelica Conference* Linköping University Electronic Press: Sweden 593-602
- [12] Zhou F L, Chen L P, Wu Y Z, et al. 2006 *Proceedings of the 5th International Modelica Conference* Press: The Modelica Association 725-31
- [13] Peter F 2004 *Principles of Object-Oriented Modeling and Simulation with Modelica2.1* Wiley IEEE Press: New York

| Authors | |
|---|---|
|  | <p>Chang Chen, born on December 10, 1983, Taizhou City, Zhejiang Province, China</p> <p>Current position, grades: Ph.D., lecturer of Department of Mechanical Engineering in Hangzhou Dianzi University University studies: Huazhong University of Science and Technology Scientific interest: Multi-domain unified modelling theory and technology Experience: Huazhong University of Science and Technology, 2008/9-2013/1, Mechanical Design and Theory, PhD, research subjects engaged in Multi-domain unified modelling theory and technology, well known about modelling methodology and simulation strategy</p> |
|  | <p>Guojin Chen, born on May 2, 1961, Ningbo City, Zhejiang Province, China</p> <p>Current position, grades: Ph.D., Professor of Department of Mechanical Engineering in Hangzhou Dianzi University University studies: XiDian University Scientific interest: Mechatronics theory and technology, Control theory and technology Experience: XiDian University, 2004/9-2007/6, Mechanical Manufacturing and Automation PhD, research subjects engaged in the auto-focusing technology of digital image</p> |
|  | <p>ShaoHui Su, born on September 7, 1978, Ruyang City, Henan Province, China</p> <p>Current position, grades: Ph.D., Associate Professor of Mechanical Engineering in Hangzhou Dianzi University University studies: Zhejiang University Scientific interest: Product Digitalize Design, Manufacturing Information Engineering Experience: Zhejiang University, 2002/9-2007/12, Mechanical Manufacturing and Automation PhD, research subjects engaged in theory and method of Product Data Management well known about PLM methodology and research integrated techniques of CAX/PDM, and focused on build the integrated product data model</p> |
|  | <p>Haiqiang Liu, born on April 19, 1980, JiangXi Province China</p> <p>Current position, grades: Ph.D., lecturer of Department of Ocean Engineering in Hangzhou Dianzi University University studies: Zhejiang University Scientific interest: Intelligent design and digital product design Experience: Zhejiang University 2005/9-2010/9 Mechanical Manufacturing and Automation PhD research subjects engaged in theory and method of Product Data Management well known about PLM methodology and research integrated techniques of CAX/PDM, and focused on build the integrated product data model</p> |

Study on quantitative diagnosis method of valve clearance based on cylinder head vibration signal of diesel engine

Zhangming Peng^{1, 2, 3}, Guojin Chen^{1, 2}, Shaohui Su^{1, 2*}

¹Hangzhou Dianzi University, Hangzhou, China

²Zhejiang Provincial Key Laboratory of Ship and Port Machinery Equipment Technology, Hangzhou, China

³Yangfan Group CO., LTD, Yangfan Ship Design & Research Institute, Zhoushan, China,

Received 1 March 2014, www.tsi.lv

Abstract

The vibration signal of cylinder head contains abundant performance information of diesel engine, and it is inseparable from injection advance angle and valve timing in time domain, so it is easy to separate the response signal of each exciting force from vibration signal. In this paper, the vibration signals of the exhaust valve closing were cut out by extract time interval sampling, and the energy information of feature frequency range was extracted by HHT transform, the corresponding relationship between valve clearance and energy information was established after normalization, so that it is realized to quantitatively diagnose the valve clearance.

Keywords: cylinder head vibration, diesel engine valve, extract time interval sampling, quantitative diagnosis

1 Introduction

There has to be vibration when diesel engine works, the vibration signal contains abundant performance information of the diesel engine. There are many vibration sources, and the moving parts are many and various shapes, thus the contribution of excitation forces to the body vibrations are inconsistent, so the diesel engine vibration is very complex with nonlinear and nonstationary characters [1-2]. The motion of the diesel engine is complex, and is a combination of rotary motion and reciprocating motion. The vibrations include combustion vibration, the opening and closing collisions of intake and exhaust valves, and the collisions produced by all mechanical moving parts, etc., so the online fault diagnosis on the diesel engine with vibration signal is very difficult.

There are many diagnosis methods on diesel engine valve with the vibration signal [3-4], and those focus only on qualitative research on the state of the diesel engine valve, the engineering applications of valve clearance online quantitative monitoring present great difficulties, so this paper analyses the vibration mechanism of the engine valve. According to intercepting the vibration signal of the exhaust valve closing on time domain, extracting the feature energy information of the vibration signal with HHT transform, and normalization process, the correspondence between the valve clearance and the vibration information is researched, that lays the technical foundation for the online quantitative diagnosis on diesel engine valve clearance.

2 Valve vibration mechanism

The valve moves on plunger role, it can be simplified as single degree of freedom model [5-6]. Figure 1 shows a single degree of freedom dynamic model of the valve.

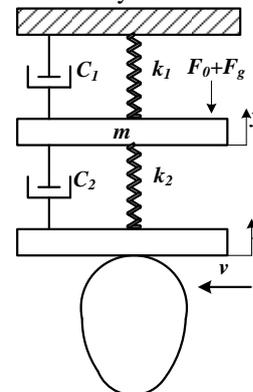


FIGURE 1 Valve train dynamic model of single degree of freedom

The differential equation of the valve motion can be described as

$$\frac{d^2 y}{d\alpha^2} + \frac{c_1 + c_2}{m\omega} \frac{dy}{d\alpha} + \frac{k_1 + k_2}{m\alpha^2} y = \frac{c_2}{m\omega} \frac{dx}{d\alpha} + \frac{k_2}{m\omega^2} x - \frac{1}{m\omega^2} (F_0 + F_g) \quad (1)$$

where m is the system quality equivalent, k_1 the valve spring stiffness, k_2 the system equivalent stiffness, c_1 the system external damping, c_2 the damping, F_0 the valve

* Corresponding author e-mail: 123151021@qq.com

spring preload, F_g the force of gas on valve, α the cam angle, ω the cam angular velocity, y the valve lift, x the equivalent cam lift.

$$x(\alpha) = \zeta h(\alpha) - e, \tag{2}$$

where ζ is the rocker ratio, e the valve clearance, h the plunger lift and the function of cam angle.

The equation (1) meets the initial condition

$$\begin{cases} y|_{\alpha=\alpha_0} = x(\alpha_0) \\ \frac{dy}{d\alpha}|_{\alpha=\alpha_0} = \frac{dx}{d\alpha}|_{\alpha=\alpha_0} \end{cases}, \tag{3}$$

α_0 is the valve opening angle, at this point the valve in upward force and the downward force exactly is in balance, α_0 can be seen as the following equation on α :

$$c_2\omega \frac{dx(\alpha)}{d\alpha} + k_2x(\alpha) = F_0 + F_g(\alpha), \tag{4}$$

The force balance of the valve in the open moment meets equation (4), the first item on the left is smaller damping force, if it is ignored, then:

$$e = \zeta h(\alpha_0) - \frac{F_0 + F_g}{k_2}, \tag{5}$$

According to the above formulas, when the valve clearance increases, the valve seating angle advances, seating speed and acceleration increase, these will inevitably lead seating impact energy to increase, it mean that there is a certain relationship between the valve clearance and vibration energy, therefore it is possible to quantitatively detect valve clearance with the valve seating energy.

3 HHT transform principle

The quantitative diagnosis on diesel engine valve clearance needs to extract the energy information on the characteristic frequency range, it can be realized by HHT transform.

On 1998, Norden E.Huang put forward a new nonlinear and non-stationary signal processing method, Hilbert - Huang transform (HHT). With empirical mode decomposition (referred to as EMD) method, the signal is decomposed into several intrinsic mode functions (referred to as IMF), and then for the IMFs, Hilbert transform is used to obtain the instantaneous frequency

and amplitude, thus the time-frequency distribution of signal can be thus fully expressed.

For the real signal $x(t)$, According to HHT on each c_i , ignoring the residual r_n , $x(t)$ can be described as

$$\begin{aligned} x(t) &= \text{Re} \sum_{i=1}^n a_i(t) e^{j\Phi_i(t)} \\ &= \text{Re} \sum_{i=1}^n a_i(t) e^{j \int \omega_i(t) dt} \end{aligned}, \tag{6}$$

$a_i(t)$ is the amplitude function, $\Phi(t)$ is the phase function, which are the analytic signal built by Hilbert transform on each intrinsic mode function.

$$H(\omega, t) = \text{Re} \sum_{i=1}^n a_i(t) e^{j \int \omega_i(t) dt}. \tag{7}$$

The Hilbert marginal spectrum can be defined as

$$h(\omega) = \int_0^T H(\omega, t) dt. \tag{8}$$

T is the total length of signal.

4 Fault test of diesel engine valve

4.1 THE ARRANGEMENT OF MEASURING POINT

Experimental system is shown in Figure 2. Low frequency interference signal has larger amplitude in diesel engine vibration signal. In order to reduce the interference of other factors, improve the signal-to-noise ratio of the vibration signal, and reflect the true state of the diesel engine, the acceleration sensor is installed over the cylinder head to measure the vertical vibration signal.

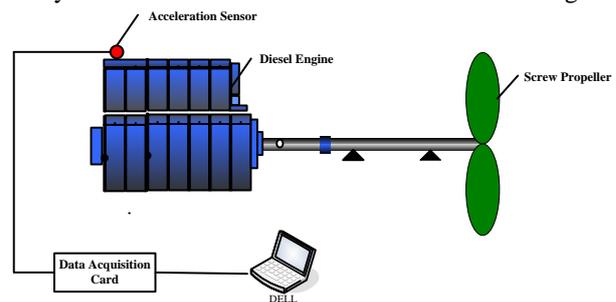


FIGURE 2 Sketch map of test system

4.2 EXTRACTION AND NONMALIZATION OF VIBRATION SIGNAL

The monitoring method on diesel engine state with the vibration signal needs to separate the vibration signals, there are differences among responses caused by the excitation sources of diesel engine in time phase, and it is easy to separate them on time domain. According to the working characteristics of diesel engine, in a work cycle,

vibration signal contains valve characteristics in certain time period, so in the certain time period, vibration signal is intercepted to analyse.

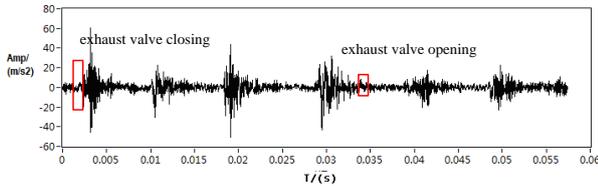


FIGURE 3 Cylinder head vibration of the 6# cylinder

Test was achieved on a certain diesel engine, and the vibration signal of 6# cylinder was extracted to analyse. The vibration signal of valve seating was used to diagnose the valve clearance. According to the valve timing of the diesel engine, the exhaust valve closed from 26° CA to 31° CA, as shown in Figure 3, the vibration signal of valve seating was intercepted with the extract time interval sampling, it contained the information of valve seating, and was decomposed by EMD, see Figure 4 a), and its Hilbert marginal spectrum as shown in Figure 4 b).

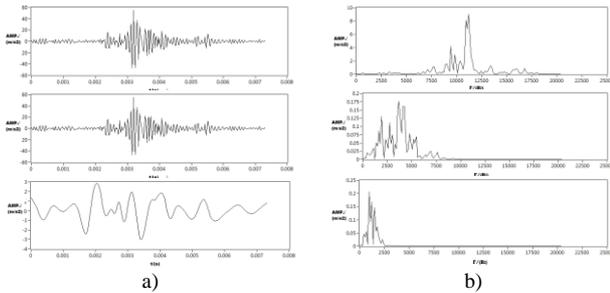


FIGURE 4 IMFs and Spectrums of intercepted signals

According to the vibration characteristics of the diesel engine valve, the frequencies of valve seating signal mainly were above the 6 KHZ. The frequencies of IMF0 were in this range by HHT, It meant the energy of valve seating signal was concentrated in the decomposition signal IMF0. As shown in Figure 4b, the frequencies of IMF0 were mainly 8000 HZ~14000 HZ, it indicated that the energy of valve seating was mainly concentrated in 8000 HZ~14000 HZ, This frequency band could be taken as characteristic frequencies at valve shutdown. When the speed of diesel engine was 1250 r/min, and the exhaust valve clearances were 0.3 mm, 0.4 mm, 0.7 mm and 1.0 mm, normalization was achieved with formula (9) and (10).

$$E = \left(\sum_{i=1}^n |E_i|^2 \right)^{1/2} \quad (9)$$

The following equation gave an value for the normalized value η :

$$\eta = \frac{E_i}{E} \times 100\% \quad (10)$$

In order to reduce the workload of test and calculation, the normalized values of 0.5 mm, 0.6 mm, 0.8 mm and 0.9 mm were achieved with interpolation, and the normalized curve of valve clearance was established, as shown in figure 5.

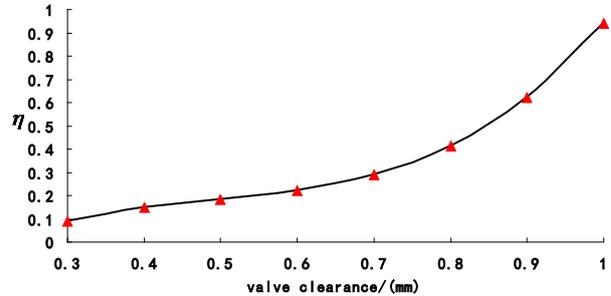


FIGURE 5 Normalization curve

4.3 EXPERIMENTAL VERIFICATION

In order to validate the method, at the speed of 1250r/min, setting valve clearances to 0.5 mm, 0.6 mm, 0.8 mm and 0.9 mm, test was completed. Table 1 shows the normalized data. Figure 6 shows the comparison between the test normalized curve and standard normalized curve, the error is less than 5%, it basically meets the requirement of engineering application.

TABLE 1 The comparison data of test

| exhaust valve clearance (mm) | Standard normalized value (a) | Test normalized values (b) | Error c (%) |
|------------------------------|-------------------------------|----------------------------|-------------|
| 0.3 | 0.088 | 0.092 | 4.55 |
| 0.4 | 0.147 | 0.154 | 4.76 |
| 0.7 | 0.287 | 0.277 | 3.48 |
| 1.0 | 0.942 | 0.903 | 4.14 |

Where: $c = \frac{|a-b|}{a} \times 100\%$.

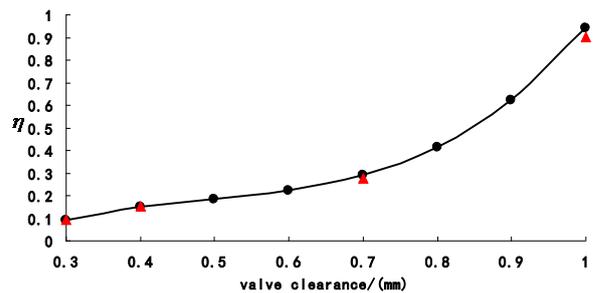


FIGURE 6 The comparison chart of test

5 Conclusions

Vibration signal of valve seating was cut out with extract time interval sampling, According to extracting the cylinder cover vibration information with HHT, and establishing the normalized standard curve, it can be achieved to quantitatively monitor valve clearance of diesel engine. The error is less than 5% that basically meets the requirement of engineering application.

References

- [1] Cai Yong 2007 *The Research on Monitoring and Diagnosing for Stirling Engine* Wuhan University of Technology: Wuhan (In chinese)
- [2] Zhao Juan 2006 *Study on Fault Diagnosis of Fuel System of Diesel Engine Based on Fractal Theory* Central South University: Changsha (In chinese)
- [3] Wang Chunlin, Mei Weijiang, Bian Jinying 2010 *Journal of Shihezi University: Natural Science* 28(1) 113-7 (In chinese)
- [4] Wang Chun-tao, Lu Jin-ming 2010 *Journal of Vibration* 30(4) 465-8
- [5] Li Shu-man 2006 *The New Diesel Engine Design, Fault Diagnosis and the Domestic and Foreign Standards Manual* Knowledge Press: Beijing (In chinese)
- [6] Xiao Jian-kun, Yu Fei 2004 *Journal of East China Shipbuilding Institute: Natural Science Edition* 18(2) 81-5 (In chinese)

| Authors | |
|--|---|
|  | <p>Peng Zhangming, born on September 17, 1977, Hubei Province, China</p> <p>Current position, grades: Lecturer of Hangzhou Dianzi University, doctor's degree University studies: Wuhan university of Technology Scientific interest: Monitoring and control of diesel engine Publications: 2 Experience: Wuhan university of Technology, 2007/3-2010/12, marine engineering PhD, research subjects engaged in diesel engine</p> |
|  | <p>Guojin Chen, born on May 2, 1961, Ningbo City, Zhejiang Province, China</p> <p>Current position, grades: Ph.D., Professor of Department of Mechanical Engineering in Hangzhou Dianzi University University studies: XiDian University Scientific interest: Mechatronics theory and technology, Control theory and technology Experience: XiDian University, 2004/9-2007/6, Mechanical Manufacturing and Automation PhD, research subjects engaged in the auto-focusing technology of digital image</p> |
|  | <p>Shaohui Su, born on September 7, 1978, Ruyang City, Henan Province, China</p> <p>Current position, grades: Associate professor University studies: Zhejiang University Scientific interest: Product data management, innovation design Publications: 10 Experience: Zhejiang University, 2002/9-2007/12, Mechanical Manufacturing and Automation PhD, research subjects engaged in theory and method of Product Data Management well known about PLM methodology and research integrated techniques of CAX/PDM, and focused on build the integrated product data model</p> |

A water quality changing prediction model for agricultural water-saving irrigation based on PSO-LSSVR

Jian-Gang Dong*, Feng Zhang, Yong-Heng Zhang

School of Information Engineering, Yulin University, 719000, Yulin, China

Received 3 March 2014, www.tsi.lv

Abstract

In order to improve the prediction of early warning and agriculture information processing level of water quality for agricultural water-saving irrigation, using mathematics and information theory model to predict and estimate the possibility of future changes in water quality based on getting the quality data by using sensor device. The basic process, model for water quality prediction of agricultural water-saving irrigation, forecasting and early warning method of establishing process is designed. Finally, was using the PSO-LSSVR forecasting method to predict the water quality in the agricultural water-saving irrigation of water quality changes prediction. Simulation results show that the parameters of LSSVR were optimized by PSO algorithm, and overcome the cross validation to determine the influence of subjective factors of LSSVR parameters, has better prediction accuracy and generalization ability, its precision can satisfy the need for intensive irrigation production management.

Keywords: prediction model, PSO-LSSVR, water quality, water-saving irrigation, ZigBee

1 Introduction

It is one of the many field's applications of agricultural information processing method for agricultural forecast, It is using mathematics and information theory model to predict and estimate the possibility of future changes in water quality based on getting the quality data by using sensor device. This paper is discussed the agricultural water-saving irrigation forecast method, basic principles and basic steps, forecasting and early warning method is used, and was using the PSO-LSSVR algorithm prediction water quality of agricultural water-saving irrigation [1]. There were accurate prediction of the water quality parameters of agricultural water-saving irrigation according to the information and online monitoring data acquired, to grasp the variation of water quality in time and space and the development trend of the estimates and projections [2]. In advance of the irrigation water quality changes make proper identification, for agricultural water-saving irrigation management and water management departments in the water, making water quality planning adjustment measures and prevention of sudden deterioration of water quality events and provide scientific basis for decision making [3].

Prediction is refers to the various information and data on the history of investigation and statistics as the basis, from the point of view of things presented phenomena, using scientific methods and means, the possibility of future development of things projections and estimates, the development and changes of things in the future to make scientific analysis; thus the past and present to speculate about the future, by known to

extrapolate, a science which reveals the trend of future development and law of objective facts or things. At present, there is extensive application in financial industry, commercial, meteorology, along with the development and popularization of agricultural technology, obtain more agricultural data, it's worthy to study agriculture forecast will become the future applications in the field [4].

It is in soil, environmental, meteorological data, growth, crop or animal conditions for agricultural production, fertilizer and pesticide, feed, aerial or satellite images and other practical agricultural data for agriculture forecast, based on economic theory, by means of mathematical model, projections and estimates the possibility of future development of the study object [5]. It is an important basis for scientific decision of precision fertilization, irrigation, sowing, weeding, pest control and other farming and agricultural production plan, supervise the implementation of the situation, but also improve the effective means of agricultural management. Therefore, agriculture prediction is one of the important technical means to support all aspects of agricultural production, sales activities based on the future agriculture Internet.

Agriculture forecast accuracy directly affects the quality of decision-making errors, and processing scheme thus, how to predict the various agricultural decision making process, and how to predict with high accuracy is an important problem in modern precision agriculture research. It's varied of prediction methods, all the effects of each prediction method cannot be completely contained to predict the target factor, in view of the complexity prediction target in agriculture and diversity,

* *Corresponding author* e-mail: donguncn@126.com

selection of prediction methods, it is the key what to establish to what mathematical model is more suitable for modern precision agriculture for prediction.

2 Water quality prediction method

In the agricultural water hydrology monitoring, the monitoring area wide, there are multiple items and unattended or duty personnel shortage, especially in the unexpected bad climate and environment is very difficult to ensure that the data transmitted to the monitoring centre in time, and may cause significant losses. For monitoring and early warning related hydrological data using wireless sensor networks, network mode is simple, time-saving, and real-time. Low cost, low power consumption, a sensor node dormancy and wake-up mechanism, can guarantee that the network can work stably for a long time in the field environment. Low cost, low power consumption, a sensor node dormancy and wake-up mechanism, can guarantee that the network can work stably for a long time in the field environment. The network comprises a water level gauge, gauge, anemometer and the gate level gauge according to the terminal node, can be installed in the rivers, reservoirs or farmland designated place of actual needs, work in the field of unattended [6]. ZigBee node through GPRS/CDMA wireless network or ADSL data collected will be sent to the hydrologic and water resources monitoring and management centre, has high reliability and expandability, combined with the superiority of the GPRS technology, is to achieve an ideal solution of Wireless Hydrological monitoring network.

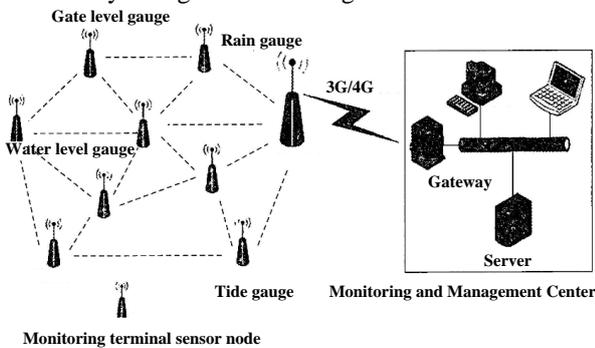


FIGURE 1 Agricultural water hydrology monitoring based on WSN

2.1 PARTICLE SWARM OPTIMIZATION ALGORITHM

Particle swarm optimization algorithm (PSO) is proposed by Kenney and Eberhart in 1995 population parallel search algorithm based on global optimization [7], through cooperation and competition between groups in the community to achieve optimal particle. Mathematical description of PSO: a population size is n , the i particles in m dimensional search space representation of $X_i = (X_{i1}, X_{i2}, \dots, X_{ij}, \dots, X_{im})$, flight speed is $V_i = (V_{i1}, V_{i2}, \dots, V_{ij}, \dots, V_{im})$, the optimal position of

individual so far to search is $P_i = (P_{i1}, P_{i2}, \dots, P_{ij}, \dots, P_{im})$. The particle swarm optimal position is $P_{gbest} = (P_{gbest1}, P_{gbest2}, \dots, P_{gbestm})$. It can update the particle velocity and position according to the formula (1) and (2):

$$v_{ij}^{t+1} = \omega \cdot v_{ij}^t + c_1 \cdot r_1 \cdot (p_{ij} - x_{ij}^t) + c_2 \cdot r_2 \cdot (p_{gbestj} - x_{ij}^t), \tag{1}$$

$$x_{ij}^{t+1} = x_{ij}^t + v_{ij}^{t+1}, \tag{2}$$

where $j=1,2,\dots,n$; $j=1,2,\dots,m$; $c_1, c_2 > 0$ are respectively the individual learning factor and social learning factors; t is the current number of iterations, r_1 and r_2 are uniformly distributed random numbers in the range of $[0,1]$. ω is the inertia weight coefficient, used to control the effect of history on current speed. In order to balance the global and local search ability, make the ω along with the increase in the number of iterations decreases linearly, can significantly improve the performance of the PSO algorithm, it is given by

$$\omega = \omega_{max} - t \times \frac{\omega_{max} - \omega_{min}}{t_{max}}. \tag{3}$$

In the formula (3), ω_{max} is the initial inertia weight; ω_{min} is the last inertia weight; t_{max} is the maximum number of iterations. Flight speed is $v_i \in [-V_{max}, V_{max}]$, the constraint conditions to prevent particle speed missed optimal solutions, through the improvement of the algorithm further improves the global searching ability of particle swarm.

2.2 LEAST SQUARES SUPPORT VECTOR REGRESSION ALGORITHM

Suykens proposed a least squares support vector regression (LSSVR) in 1999, is used to solve the problem of function estimation [8]. LSSVR is replaced by equality constraints and inequality constraints, the function of the error square and loss experience loss as training set, the traditional support vector machine in the solution of two quadratic programming problem is transformed into solving linear equation group, effectively improves the calculation speed and convergence precision, has better generalization performance. The mathematical model of the least squares support vector regression is given by:

$$\min J(\omega, \xi) = \frac{1}{2} \omega^T \omega + \frac{C}{2} \sum_{i=1}^l \xi_i^T \xi_i, \tag{4}$$

$$s.t \ y_i = \omega^T \varphi(x_i) + b + \xi_i. \tag{5}$$

In the formula, $X_i \in R^l$ and $Y_i \in R^l$ are the input and output vector system, $\xi_i \in R$ is the empirical error, b is offset, $C \in R^+$ is the regularization parameter, $\varphi(\cdot)$ is a nonlinear mapping of the input space to the feature space. To solve the constrained optimization problems, Lagrange polynomial function of the dual problem is given by

$$L(\omega, b, \xi, a) = J(\omega, \xi) - \sum_{i=1}^l a_i (\omega^T \varphi(x_i) + b + \xi_i - y_i) \quad (6)$$

In the formula (6), $a = [a_1, a_2, \dots, a_l]^T$ is the Lagrange multiplier. According to the Karush-Kuhn-Tucher (KKT) conditions, respectively, for ω, b, ξ_i, a_j partial derivative, and let it equal to 0, linear system can be obtained as follows:

$$\begin{cases} \frac{\partial L}{\partial \omega} = 0 \rightarrow \omega = \sum_{i=1}^l a_i \varphi(x_i) \\ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^l a_i = 0 \\ \frac{\partial L}{\partial \xi_i} = 0 \rightarrow a_i = C \xi_i \\ \frac{\partial L}{\partial a_i} = 0 \rightarrow \omega^T \varphi(x_i) + b + \xi_i - y_i = 0 \end{cases} \quad (7)$$

In the formula (7), to eliminate of ω, ξ_i , we can get the following linear equations

$$\begin{bmatrix} 0 & I^T \\ I & \Omega + C^{-1}E \end{bmatrix} * \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (8)$$

In the formula, $I = [1, 1, \dots, 1]^T$, E is a unit matrix of dimension $l \times l$; $y = [y_1, y_2, \dots, y_l]^T$, $\Omega_{ij} = \phi(x_i, x_j)$, $\phi(x_i, x_j) = K(x_i, x_j)$ as the kernel function satisfying the Mercer conditions, the least square for the support vector regression model is given by

$$y = f(x, a) = \sum_{i=1}^l a_i k(x, x_i) + b \quad (9)$$

3 Create the prediction model

3.1 THE LSSVR PARAMETER OPTIMIZATION BASED ON PSO

The study found the penalty factor C and kernel function parameter determines the performance of the LSSVR regression model. In order to improve the prediction

performance of C and σ , it is the key of getting the best parameter combination. On the combination of parameter optimization of LSSVR model, there is no effective method, often through cross validation or gradient descent method, time-consuming and anthropogenic influence. Therefore, using the PSO algorithm for the parameters of LSSVR for automatic optimization, not only overcome the randomness of artificial selection, but also through particle fitness function settings, realize the automatic selection of objective parameters. Mean square error can directly reflect the performance of the LSSVR model (MSE) as the inverse function Fitness PSO algorithm's fitness (ε), its expression is shown as follows:

$$Fitness(C, \sigma^2, \varepsilon) = \frac{1}{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}} \quad (10)$$

In the formula (10), y_i, \hat{y}_i were real and predicted values. Improved PSO LSSVR parameter optimization algorithm based on the procedure described as follows:

1) Particle swarm (C, σ) initialization. Set the number of particles size n , the maximum number of t_{max} iterations, range, the inertia weight ω particle velocity v limits, learning factor c_1, c_2 and other parameters, and a set of randomly generated initial particle velocity and position.

2) To train the LSSVR with the training set, by formula (10) are calculated for each particle's fitness value $Fitness(C, \sigma)$, then according to the particle's fitness value of individual extremum p_i and global extreme update p_{gbesti} .

3) By the formula (1), the formula (2) and formula (3), the speed and position of each particle is updated.

4) Checking algorithm termination conditions, such as the number of iterations is equal to T or optimal solution will not change, then the algorithm ends, output the optimal combination of parameters. Otherwise, proceed to step 3) optimization.

3.2 PREDICTION MODEL OF WATER QUALITY BASED ON PSO-LSSVR

The basic idea of the model for the nonlinear robust multi parameter water quality prediction system based on PSO-LSSVR is to make full use of PSO - LSSVR has the ability to establish the complex nonlinear relationship between the change of water quality and the reduced set of agricultural water-saving irrigation waters influence factors, mining internal variations of water quality, so as to realize the change of the historical factors make accurate prediction of future water quality change.

The model not only fully consider various factors with different times give different weights, optimized the parameters of LSSVR using improved PSO, effectively

improve the accuracy of the prediction model and the generalization ability. The water quality prediction model building steps is shown in figure 2.

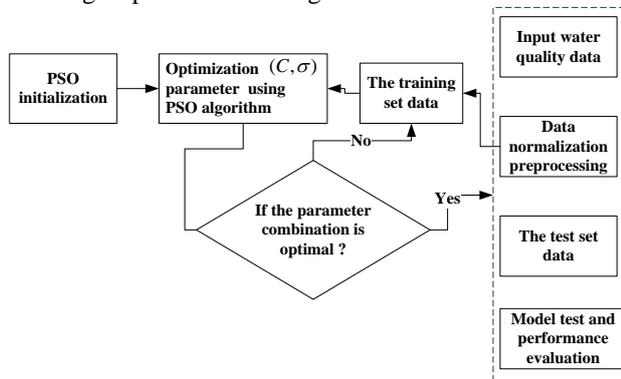


FIGURE 2 PSO-LSSVR water quality prediction model create step

4 The implementation of the model prediction

4.1 INTENSIVE WATER-SAVING IRRIGATION ECOLOGICAL ENVIRONMENT DATA SOURCE

In order to effectively forecast model for agricultural water saving irrigation dissolved oxygen inspection is presented in this paper, taking Yulin University network application demonstration base of agricultural water-saving irrigation reservoir ecological environment data as sample data, each sample including dissolved oxygen, temperature, pH, and rainfall index.

The sampling period is from 2014 May 10 to June 10th, sampled once every 30 minutes, a total of 561 samples, 500 samples collected before 7 days of training sets, the remaining 61 samples as a test set, taking the pond dissolved oxygen concentration on the next moment of quantitative prediction. The ecological environment monitoring part of the original data is shown in Table 1.

TABLE 1 Water saving irrigation ecological environment of original data

| Time | Water temperature / °C | pH | Dissolved oxygen /(mg/L) | Rainfall / mm |
|-------|------------------------|------|--------------------------|---------------|
| 06:00 | 25.132 | 7.08 | 6.6842 | 729 |
| 07:30 | 25.543 | 6.89 | 7.1231 | 456 |
| 08:10 | 26.113 | 7.23 | 6.9087 | 1295 |
| 09:20 | 26.453 | 7.12 | 7.9764 | 4324 |
| 12:00 | 30.211 | 6.98 | 7.9753 | 433 |
| 14:30 | 31.653 | 6.09 | 7.0984 | 543 |
| 17:30 | 32.756 | 6.34 | 6.0644 | 356 |
| 20:20 | 31.675 | 6.98 | 5.9745 | 453 |
| 22:10 | 30.087 | 6.91 | 4.5357 | 345 |
| 23:30 | 30.011 | 6.92 | 3.7543 | 344 |

4.2 DATA PRE-PROCESSING

Intensive water-saving irrigation dissolved oxygen is influenced by many factors, is a sequence of data changes with time, with different dimension.

If the direct use of the original data for combinatorial optimization for training the LSSVR parameters based on

PSO algorithm, not only affects the learning speed, but also seriously restricts the accuracy and robustness of the predictive model, it is necessary to establish prediction model by type (10) of the original data pre-treatment, to reduce the data the dimension of different influence on the prediction model, so we can write

$$\bar{x}_m^n = \frac{x_m^n - \min(x_m^n |_{k=l})}{\max(x_m^n |_{m=1}) - \min(x_m^n |_{m=1})}, d = 1, 2, \dots, m. \quad (11)$$

In the formula (11), l is the total sample, n is the dimension of the sample vector, \bar{x}_m^n and x_m^n respectively the original data of water saving irrigation of the ecological environment and the normalized data.

4.3 ANALYSIS OF ALGORITHM AND PERFORMANCE

Algorithm using Matlab11 language programming, PSO algorithm is initialized: population size is $n = 50$, $c_1 = c_2 = 1.23$, the maximum number of iterations of $t_{max} = 98$, the scope of the ω is set to $[0.9, 0.35]$, particle velocity v limits $[0.3, 8]$. The dissolved oxygen, temperature, pH values as input of half an hour before, prediction of dissolved oxygen of the next moment values as output, the mean square error of prediction values of dissolved oxygen and the next time the actual value as the particle fitness function. According to the optimization algorithm is used to train the PSO-LSSVR LSSVR parameters based on PSO, the training times of $t = 80$, the best combination of parameters for LSSVR: $\sigma = 0.0075$, $C = 78.36$. The combined prediction of dissolved oxygen concentration parameter generation PSO-LSSVR models, as is shown in figure 3.

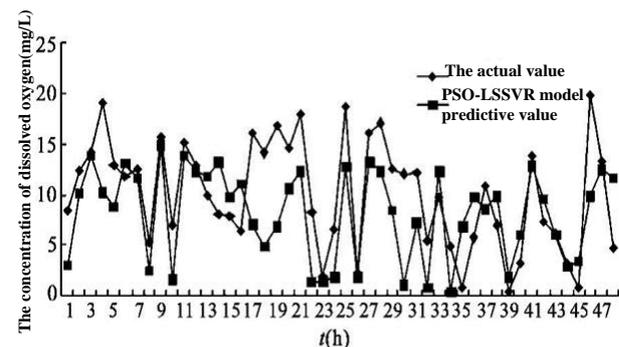


FIGURE 3 The comparison of prediction value and real value of dissolved oxygen based on PSO-LSSVR

It can be seen from Figure 3, the predicted curve are basically consistent with the measured curve, and with the combination of parameters of PSO algorithm for LSSVR optimization, overcomes the cross validation to determine the influence of subjective factors combination parameters of the LSSVR, has better prediction accuracy and generalization ability, the prediction accuracy can

meet the need for intensive production of water-saving irrigation management.

5 Conclusions

This paper discusses the basic method for prediction of agricultural irrigation water quality, detailed analysis of the basic principles of prediction model of agricultural irrigation water quality prediction model, the selection principle, basic steps and methods of forecasting. The construction of the research method and the experimental method of model is predictive the PSO-LSSVR intensive agricultural water-saving irrigation based on water quality. The construction of intensive agricultural water-saving irrigation water quality prediction model of based

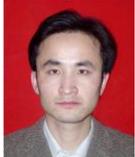
on PSO-LSSVR is not necessarily the best forecasting model, the need for further research of intelligent computer technology and the improvement of the algorithm of support vector machine in the future research, in order to improve the robustness of water quality prediction.

Acknowledgments

This work is partially supported by the Funding Project for Department of Education of Shaanxi Province in 2013 #2013JK1151, Research and Cooperation Project for Department of Yulin city of (#Gy13-16, #Sf13-08, #Sf13-23). Thanks for the help.

References

- [1] Avci E 2009 Selecting of the optimal feature subset and kernel parameters in digital modulation classification by using hybrid genetic algorithm support vector machines: HGASVM *Expert Systems with Applications* **36**(2) 1391-402
- [2] Babaoglu I, Findik O, Ulker E 2010 A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine *Expert Systems with Applications* **37**(4) 3177-83
- [3] Cherkassky V, Ma Y 2004 Practical selection of SVM parameters and noise estimation for SVM regression *Neural Networks (S0893-6080)* **17**(1) 113-26
- [4] Hsieh T J, Hsiao H F, Yeh W C 2012 Mining financial distress trend data using penalty guided support vector machines based on hybrid of particle swarm optimization and artificial bee colony algorithm *Neurocomputing* **82** 196-206
- [5] Üstün B, Melssen W J 2005 Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization *Analytical Chimica Acta (S0003-2670)* **544**(1-2) 292-305
- [6] Yaoyu Feng, Zhimin He 2003 Optimization of agitation, aeration, and temperature conditions for maximum β -mannanase production *Enzyme and Microbial* **32**(2) 282-9
- [7] Zhang Yong-Heng, Zhang Feng 2013 A New Time Synchronization Algorithm for Wireless Sensor Networks Based on Internet of Things *Sensors and Transducers* **151**(4) 95-100
- [8] Cao X, Chen, Sun Y 2009 An Interface Designed for Networked Monitoring and Control in Wireless Sensor Networks *Computer Standards and Interfaces* **31**(3) 579-85

| Authors | |
|---|---|
|  | <p>Dong Jian-Gang, born on November 28, 1974, in Shannxi Yulin</p> <p>Current position, grades: associate professor in Yulin University. University studies: MS degree in Computer science from Xidian University in 2009. Scientific interest: Data mining technology, the Internet of Things applications.</p> |
|  | <p>Zhang Feng, born on June 26, 1980, in Shannxi Yulin</p> <p>Current position, grades: associate professor in Yulin University. University studies: MS degree in Computer science from Xidian University in 2009. Scientific interest: Cloud integrated manufacturing technology, the modelling of complex systems, the Internet of Things applications.</p> |
|  | <p>Zhang Yong-Heng, born on October 25, 1968, in Shannxi Yulin</p> <p>Current position, grades: associate professor in Yulin University. University studies: MS degree in Computer science from Xidian University in 2010. Scientific interest: Data mining technology, Mass data processing technology.</p> |

Intelligent data-collaboration mechanism under the distributed application environment

Jian-Long Ding^{1*}, Weifang Chen¹, Ji Gao²

¹Zhejiang Shuren University, College of Information Science, Hangzhou, China

²Zhejiang University, The Artificial Intelligence Research Institute, Hangzhou, China

Received 1 March 2014, www.tsi.lv

Abstract

Because of the complexity, the dynamic and uncertainty of the distributed applications environment, the data-collaboration crisis caused by isolated information island is serious day by day. Through the establishment of Data Cooperation-based Virtual Organization (DCVO), is conducive to meet the realistic demand of the on-demand dynamic data collaboration, which led to the distributed application to carry out the intelligent data collaboration in effective control, as well as realize the intelligent data retrieval across application domain. Through research of the Distributed Application System-based Data Cooperation Architecture (DASDA), to straighten out the related technology and method of distributed collaborative, from the semantic specification (including the application of domain ontology, relational databases and ontology mapping mechanism, cooperative data transmission standard), rational of data-collaboration (policy representation and configuration), collaborative service personalization, data structure and model of collaborative content level and so on, to provides an important reference to solve the eliminate problem such as semantic fuzzy, dynamic expansion, uncontrollable, cooperative security, recall and precision of conflict which caused in the process of the data-collaboration.

Keywords: ontology, virtual organization, data collaboration, policy configuration

1 Introduction

With the development of information construction, the information system for kinds of application field become perfect gradually, except the application system under distributed environment witch face to information island and information gap problem. Service collaboration and data exchange requirements between application systems be increasingly urgent [1]. Because of the complexity, dynamic and uncertainty of the distributed environment, the distributed software system for specific application domains need to configure data-collaboration ability, in order to ensure that the system can be reliable, stable, accurate, and to provide cross-system data collaboration service for end users [2]. But because of the complexity of the distributed system and its lack of coordination ability, make the data-collaboration research difficult without the theoretical model support, and can't meet data cooperative crisis caused by growing information islands. In recent years, as the rising and developing of the cloud computing, data research and service architecture (SOA), promoting changing of distribution system development method, from the static sealing process of high cost, low efficiency, depends on the specific hardware environment to on-demand combined virtual business service process of dynamic, fast, low cost [3-6]. These changes promote the development of Data Cooperation-based Virtual Organization (DCVO) [7-10].

On this basis, to carry out the study of the Distributed Application System-based Data Cooperation Architecture (DASDA) research [11, 12], in order to eliminate the data-collaboration dilemma, which caused by the closure between business systems. Thus, to make sure the original distributed application have strong adaptability to face the all kinds of data-collaboration problem caused by non-deterministic data centre distributed operating.

2 Data collaborative DCAgent model

In order to improve the intelligent characteristics of each application point, One method is providing intelligent Agent called Data-Cooperation Agent (DCAgent) corresponding for each application point, which can improve the coordination ability. As the main body of DCVO, DCAgent's structure can be divided into two levels: application domain development layer and Agent engine) running layer as shown in Figure 1.

Development layer can provide the humanized man-machine interface to assist the developers to build the Agent model and application domain ontology, relational data and domain ontology mapping configuration file and external call component as collaborative service components which can provide basic services as personalized data synchronization, the domain ontology ConceptInstance dynamic construction etc. It can also support application developers to design all kinds of

* Corresponding author's e-mail: zjuding@163.com

DCAgents for different purposes, to meet the needs of organizations and Individual user. Running Layer provides Agent engine, which achieve the following functions.

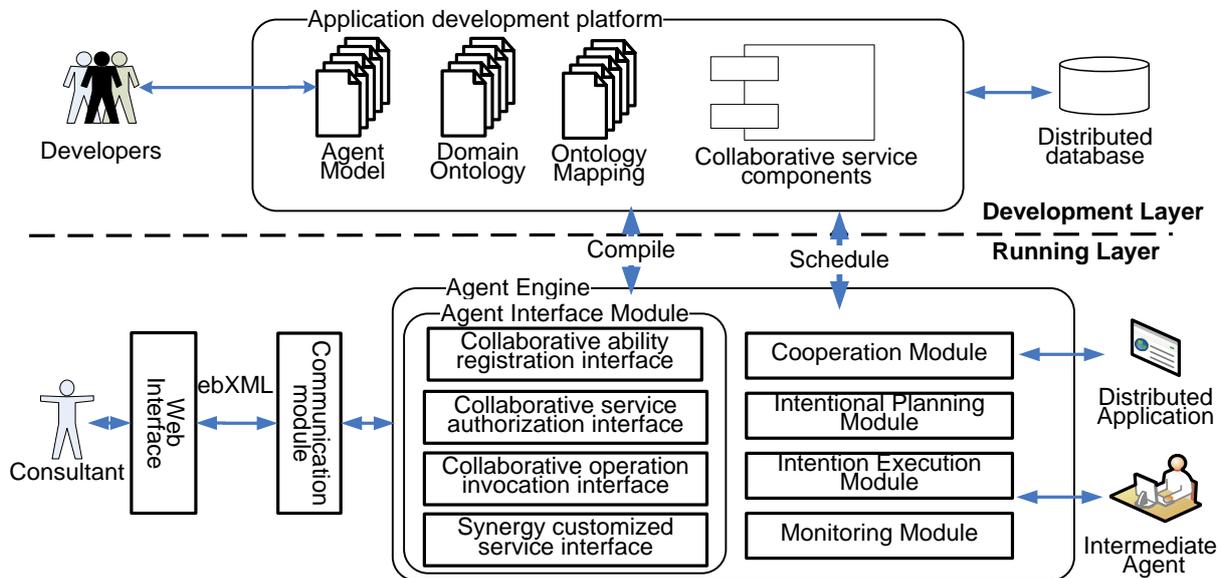


FIGURE 1 DCAgent Model diagram

2.1 PROVIDES THE INTERFACE OF AGENT SERVICE

This service is to support the interaction between Agent and human, mainly includes four aspects: collaborative ability registration interface, collaborative service authorization interface, cooperative operation interface, Synergy customized service interface. The four interface support human to manage and control Agent effectively.

Collaborative ability registration interface, this interface response to submit the concept set of collaborative field data to the intermediary DCAgent based on the domain ontology, and set the scope of cooperative operation authorization which can opening to the other DCAgents.

Collaborative service authorization interface mainly provides such functions as user management, role management, collaborative policy, configuration, etc. So as to effectively ensure the safety and reliability of the DCAgent-service calling.

Cooperative operation interface is mainly responsible for the user to invoke the DCAgent service execution process, including DCAgent service release, and authority audit, call, execution, tracking process at the levels of DCAgent, service, and operation. It can make the Agent convenient, reliable, and controllable.

Synergy customized service interface mainly provides the policy custom interface for personalized data-query, to realize the semantic level queries across distributed application by policy customization, which is used for the constraint and customization to query data.

2.2 PROVIDES THE UNITED SCHEDULING MECHANISM

The United Scheduling Mechanism consists of Cooperation Module, Intentional Planning Module, Intention Execution Module and Monitoring Module. According to the DCAgent model effectively develop and coordinate social activities, including the intermediary service request. To Establish and optimizes the cooperation relationship through rational negotiation, as well as properly handling exceptions occurred while processing of cooperation, so that, their behaviour comply with collaborative authorization policy constraints, and regulate it's the collaborative by customized control policy [13].

3 DCVO architecture under the distribute application environment

Distributed application system includes database and file system, application, in order to make the distribution of the application nodes with intelligent collaboration, need to develop and deploy DCAgent for each application node. The DCAgent will improve the data-coordination ability of application point. At the same time, in order to improve the data-coordination's performance, stability and failure recovery ability, also need to deploy high performance database (Crash level) for each application node [14]. An intermediary DCAgent should also be Set up to realize the data collaboration, data-collaboration rights, which include synchronization-role auditing, synchronization-authorization validating, synchronization-data cache, synchronous data assembly and forwarding, and synchronization commands issue and

execution. Thus the distribution of application nodes will Dynamic combine into a virtual organization called DCVO according to the specific needs of collaboration-data [15].

4 Data collaboration execution process under the distributed environment

Data collaborative virtual organization DCVO is not a static organization, but a dynamic forming organization according to data collaborative demand. Therefore the distributed application data collaborative process started by application point which data changing occurred. The triggering process is divided into two levels, the database

level and process level. For the existing island type of distributed application system for data collaboration capability upgrading situation, due to the complexity of existing application system, can consider to trigger collaborative process by adding triggers in the database level. For the new distributed application system development situation, can consider to trigger collaborative process at the process level through the assembly calling way [16]. The initiator DCAgent must submit data-collaboration Role authorization application to the intermediary DCAgent in order to become the main organizer and then launched a series of coordinated operation. The specific process is shown in Figure 2.

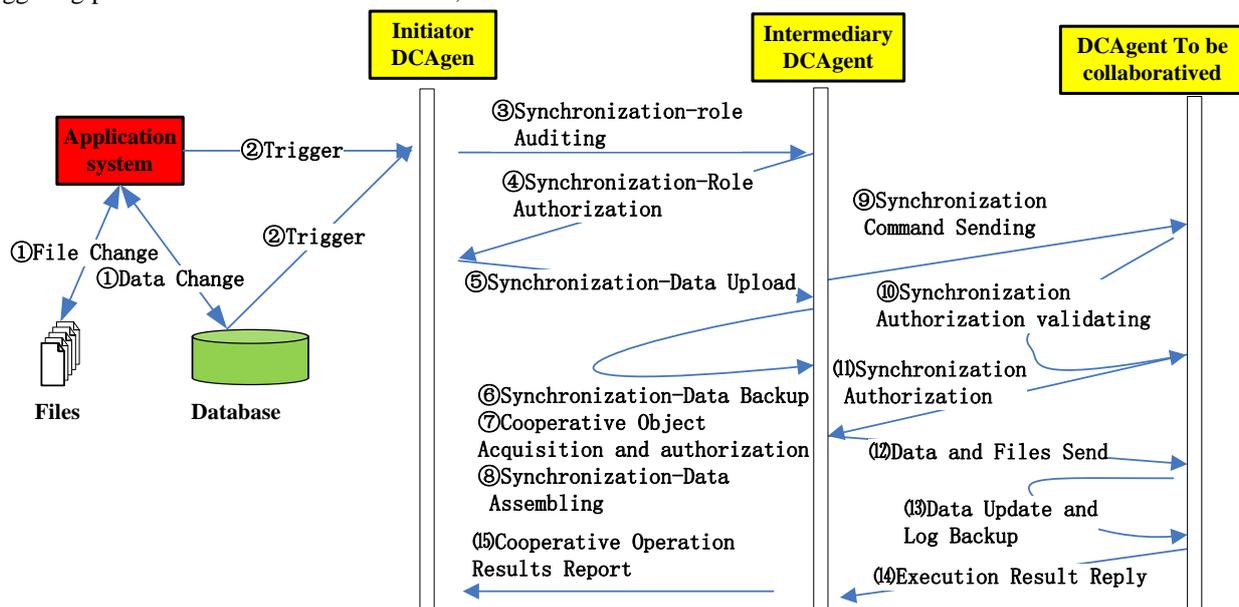


FIGURE 2 Data collaboration process description

5 Collaborative Dasda system based on distributed application data

The data collaboration under distributed application environment is a complex problem, the study of the DASDA method system will help straighten out the related technology and methods of distributed collaborative.

DASDA is defined as the following 5 elements: **DASDA = (CS, IS, RP, CI, DP).**

CS—Clear Semantics. By the research on sharing information modelling and ontology-based representation mechanism, and application domain relation database and domain ontology mapping mechanism, and data transmission format specification, etc., the DCVO will have clear semantics when data storage and exchanging.

IS—Intermediary Services. Through the establishment of DCVO oriented collaborative system of intermediary services, so that DCVO can accurate and convenient to obtain the relevant data changes to the application system at anytime, anywhere. And provides collaborative authorization registration, role application,

and semantic information retrieval service cross distributed systems.

RP—Rational Process. To establish data-coordination specification set, So that, DCVO can Carry out authorization policy description and allocation, cooperative agreement, rollback model, data retrieval policy generation operation, rational and flexible propulsion data cooperative process, to achieve the Intelligent Collaborative higher.

CI—Cooperation Individuation. Book personalized service for each DCVO through data coordination authorization policy customization. Provide personalized data query service cross distributed application point by service-customization interface, man-machine interface generation component and Data query policy.

DP—Development Plat for DCVO. Through the application of domain ontology editor, local cooperative service component development guide, configuration files auxiliary generator, VO model assisted editing platform to provide convenient, transparent development assistance service for DCVO to the user.

5.1 CLEAR SEMANTICS CS

As the intelligent agent and executive body of application domain, DCAgent need experts to define the Agent service and the implementation process by constructing a DCAgent model, but the actual operation is realized by the cooperative service components which can be scheduled by DCAgent [17]. In order to realize the intelligent data - collaboration and semantic level data-retrieval, works such as completing semantic description transformation and building corresponding mapping mechanism for application domain by ontology technology must be finished. All those works is depend on domain ontology and ontology mapping file. To make the modeling information semantic content clear, CS can describe as following 3 elements.

CS = (DCL, OML, Mapping):

Mapping: DCL→Realtion DataBase;

the "→" is a the symbol in Z language, indicating the injective function.

DCL—As knowledge representation language based-on ontology, which using in the Agent.

OML—A markup language based on Ontology, also be a communication language between Agent.

Mapping—An one-to-one mapping mechanism between DCL language and relation database.

5.1.1 Domain ontology description specification

Because of the variety of distributed application nodes on the data describing structure, property, storage way, and etc. An unified concept description standard for application domain must be established by ontology description language. As a domain concept language, DCL can be used to represent the domain ontology by the form of concept-relationship-attribute and constraint rules [17]. Its BNF form descript as follows:

```

Ontology <Ontology-Name> [<Version-declaration>]
 [<Ontology-Citation>][<Synonymous-Concepts>]
 [<Synonymous-Properties>] [<Property-Definitions>]
 [<Concept-Definitions>] [<Type-Definitions>]
   <Concept-Definitions> := {Concept <Concept-Name>
 [Super: {<Super Class Name>}+ ]
   {<Slot-Name>: {<Aspect-Name> <Aspect-
Content>,*};}*
   [Constraint: <Condition-Expression>]*
   <Aspect-Name> := val /type /mode /number /derive
 /restriction /unit /inverse /superSlot

```

Ontology is consist of the set of concepts, the super slot is used to establish the relationships between concepts when slot value is a single super class name. Slot is used to define the attributes or the relation and function parameters of object. Each slot include Type and Mode side, which respectively indicating the type of concept instance slot values (types can be another concept) and the provide way. Constraints express then relations between different slot.

5.1.2 Collaboration-data Transmission of language OML

Each DCAgent for application node need for information and data transfer during collaborative processing, which can realized by the high performance distributed middleware platform called ICE (Internet Communications Engine). Ontology Based Markup Language (OML) is designed for constrained ebXML, contain the DCL representation. Collaborative command and collaborative data is the main content of the communication, the bottom element is composed of concept instances, and therefore we need to use the ebXML to extend ICE to Expand ICE called E-ICE. Original message will by packaged into **ebXMLMessage** by **E-ICE**. It is defined as follows:

Concept ebXMLMessage

MessageType: type ebXMLMessageType mode necessary;

PartyID: type string;

Service: type string;

Action: type string;

ConversationId: type string;

CooperationRole: type CooperationRoleType mode necessary;

*MessagePayload: type*string;*

End ebXMLMessage

The message will be serialized and then be send through ICE. and then be deserialized after be received by the DCAgent terminal.

5.1.3 Mapping mechanism between database and domain ontology

In order to shield the structure and scope difference between data-synchronization nodes. the data in relational database must be processed and packed into concept instance mode before synchronization. Data-ConceptMapping configuration file will help DCAgent to identify the transformation relationship automatically [18].

Mapping: DCL→Realtion DataBase.

*<Mapping> ::= {< ConceptMap >}**

<ConceptMap> ::= ({Table/view}+, <Concept>) /<

SlotMaps >

*<SlotMaps > ::= {<SlotMap>}**

<SlotMap > ::= (<Field>, <Slot>) /<weight>

Mapping defines the relationship between the relational database and the application domain concepts. One **<ConceptMap>** node represents one mapping relationship, **<Table>** indicates data table be mapped **<Concept>** indicates the corresponding concept if the mapping object relates to multiple data tables, needs to establish a view first, then establishes the mapping relationship with the corresponding Concept. One **<SlotMap>** node indicates relationship between the data fields and slot, the weight property indicates the matching weight.

5.2 RATIONAL COLLABORATIVE PROCESS RP

There is a big difference in the operate scope and mode during the data-collaborative process, the operation authority and data range of the Data-synchronization must be declared for each application node. We can solve the above problems through policies, and using PolicyAssignment to specify the scope of authorization. So that the DCAgent's operation can be executed under the premise and controllable situation.

5.2.1 Policy Representation and Assignment Language PRAL

PRAL (Policy Representation and Assignment Language) is designed to be easy for users to understand and use, has broad application demand coverage of common language, to support the policy of declarative representation and configuration description. PRAL provides a structured and object oriented representation, is used to define the configure of policy. The first-order logic based on concept instance is used to represent the policy definition and configuration details. The PRAL entity class (including the characteristics and relationship) and complexity (with multiple characteristics) are defined as "concept".

5.2.2 Policy definition

Representation structure of the Policy is defined as a policy embedded in PRAL.

```
Type PolicyType: base_type string, restriction enumeration
(Select, Insert, Update, Delete, RollBack, Bake, Restore);
Concept Policy
```

```
Name: type string mode necessary;
PolicyType: type PolicyType mode necessary;
Processing: type ProcessingType;
Target: type condition;
Update: type date mode necessary;
End Policy
```

PolicyType can be divided into 7 categories: query, insert, modify, delete, rollback, backup, recovery. Target refers to the target range using conditional expression. Corresponding to data synchronization service that DCAgent provide.

5.2.3 Policy Assignment

Policy definition and configuration separately, facilitates the reuse policy, policy allocation and revocation of convenience. The policy configuration with PRAL embedded Policy Assignment concept definition independent representation of Policy definition and Policy assignment is help to promote the performance of policy on reuse and convenient The policy Assignment defined by with embedded **PolicyAssignment** concept in PRAL.

```
Type ModalityType: base_type string, restriction
enumeration (+, -);
```

```
Concept PolicyAssignment
```

```
Policy: type string mode necessary;
```

```
Modality: type ModalityType val "+";
```

```
Subject :type *RoleAssignment mode necessary;
```

```
Delegator: type string;
```

```
TemporalConstraint: type TemporalLogic;
```

```
End PolicyAssignment
```

5.3 DEVELOPMENT PLAT FOR DCVO

5.3.1 Research on the DCVO development platform

At present, the unit to which I have developed the application domain and modelling platform (ARMF), DCAgent development platform (RADF). In order to make the develop and deploy of DCVO. Simple, we also need to design following tool, the way of the research on mapping file (for relational database and Domain ontology) generation and verification tool, PRAL language develop tool.

5.3.2 Research on Data-synchronization component

The development of data system components mainly involves two key technologies: high performance distributed middleware ICE and high performance of real-time database Berkeley DB. Use high performance distributed middleware platform ICE (Internet Communications Engine) to realize files and database synchronization [18]. Use Berkeley DB to realize the data cache in all the data synchronization node [19].

6 Conclusion

With the continuous deepening of information technology, the way of system development from Independent development into system integration so as to provide experience to solve problems occurs during data-synchronization such as semantic fuzziness, uncontrollable, security, synchronous efficiency. This study is mainly face to the common problem of distributed application system, to enhance the utility of the mechanism, need to further study on the specific application environment, especially how to improve the automation ability configuration after a the new synchronization node is increased.

Acknowledgements

Foundation items: 1. The Education Department of Zhejiang Province, general scientific research project (Y201018715). 2 The National Natural Science Fund Project (61375071). 3 The priority theme emphases project of Zhejiang province, China (Grant 2010C11045).

References

- [1] Li X 2003 The role of digital resources integration based on the information island *Library Forum* 6(23) 121-2
- [2] Beners-Lee T, Hall W, Hendler J, Shadbolt N, Weitzner D J 2006 Creating a science of the web *Computer Science* 313(5788) 769-71
- [3] Chen Q 2009 Cloud computing and its key techniques *Journal of Computer Application* 9(26) 2562-7
- [4] Papazoglou M P, Traverso P, Dustdar S, Leymann F 2007 *IEEE Computer* 40(11) 38-45
- [5] Mike P, Papazoglou M P, van den Heuvel W J 2007 Service oriented architectures: approaches, technologies and research issues *The VLDB Journal* 16 389-415
- [6] Demirkan H, Kauffman R J, Vayghan J A, Fill H-G, Karagiannis D, Maglio P P 2008 Service-oriented technology and management: Perspectives on research and practice for the coming decade *Electronic Commerce Research and Applications* 7 356-76
- [7] Wei I, Blake M B 2010 Service-Oriented Computing and Cloud Computing Challenges and Opportunities *IEEE Internet Computing* 14(6) 72-5
- [8] Lv J, Ma X 2006 Research and development of Internetware, Science in China (Series E) 36(10) 1037-80
- [9] Jones J, Parnin C, Sinharoy A, Rugaber S, Goel A K 2009 Teleological Software Adaptation *The 3rd IEEE International Conference on Self-Adaptive and Self-Organizing Systems* 198-205
- [10] Paulo L, Paul V, Emmanuel A, 2009 Self-Adaptation for Robustness and Cooperation in Holonic Multi-Agent Systems *In Hameurlain Transaction on Large-Scale Data- & Knowl-Cent Sys I LNCS 5740* 267-88
- [11] Ghosh D, Sharman R, Rao H R, Upadhyaya S 2007 Self-healing systems-survey and synthesis *Decision Support Systems* 42 2164-85
- [12] Gao J, Yuan C-X, Wang J 2005 SASA5: A Method System for Supporting Agent Social Activities, Chinese Journal of computers 5(28) 838-48 (in Chinese)
- [13] Hu J, Gao J, Zhou B, Liao B-S, Chen J-J 2004 Ontology based agent services compatible matchmaking mechanism *Proceedings of the 3rd International Conference on Machine Learning and Cybernetics* Shanghai 111-6
- [14] Dou D, Qin H, LePendu P 2010 Ontograte: Towards automatic integration for relational databases and the Semantic Web through an ontology-based framework *International Journal of Semantic Computing* 4(1) 123-51
- [15] Weyns D, Georgeff M 2010 *IEEE Software* 27(1) 86-91
- [16] Aly W H F, Lutfiyya H 2007 Dynamic adaptation of policies in data center management *Proceedings of the 8th IEEE International Workshop on Policies for Distributed Systems and Networks* 2007 266-72
- [17] Calvanese D, De Giacomo G, Lembo D, Lenzerini M, Rosati R, Ruzzi M 2009 Using OWL in data integration in R. De Virgilio, F. Giunchiglia, L. Tanca ed. *Semantic Web Information Management - a Model Based Perspective* Chapter 17 Springer 397-424
- [18] Xu S, Johnson A 2009 Comparing the Performance of EPICS Channel Access with a New Implementation Based on ICE (the Internet Communications Engine) *the 16th IEEE - NPSS Real Time Conference (RT 2009)* 2009 113-6
- [19] Langley N 2008 Embedded database Berkeley DB offers speed and management gains *Computer Weekly* 2008 34

Authors



Jianlong Ding, born on September 15, 1980, Hangzhou, China

Current position, grades: lecturer of computer science and application in Zhejiang Shuren University.
University studies: Master's degree in the computer science and application from Zhejiang University.
Scientific interest: intelligent software, distributed computing.
Publications: 6 patents, 7 papers.



Weifang Chen, born on June 22, 1966, Ningbo, Zhejiang, China

Current position, grades: lecturer of computer science and application in Zhejiang Shuren University.
University studies: electronic university of science and technology of Hangzhou.
Scientific interest: artificial intelligence, network computing.
Publications: 1 patents, 4 papers.
Experience: teaching and research work, in recent years.



Ji Gao, born in August, 1946, Shangdong, China

Current position, grades: professor at the department of computer science and engineering college of computer science and technology of Zhejiang University, doctoral tutor.
Scientific interest: network computing and pervasive computing, intelligent software, agent technology, software engineering, middleware technology.
Publications: 2pPatents, 131 papers.
Experience: dean at the school of information science and technology in Zhejiang Shuren university in 2000 -2008. Institute Director of Internet Computing and Intelligent Systems Research Institute.

Lifetime forecasting for hemispherical resonator gyroscope with wavelet analysis-based GM(1,1)

Yunxia Zhang¹, Chenglong Dai^{2*}, Jifeng Cui²

¹De Network and Information Center, Southeast University, Nanjing 210018, China

²College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Received 15 June 2014, www.tsi.lv

Abstract

Because of high cost and small batch of spacecraft like flywheel and gyroscope, how to estimate their reliability and lifetime becomes a tough task. A method to predict the lifetime of hemispherical resonator gyroscope (HRG) is put forward in this paper. This method utilizes grey correlation and mean absolute percent error (MAPE) to estimate the reliability of predictive data sequence. For reducing noise, Daubechies wavelet is used to decompose and reconstruct the test data in the paper as well. After pre-processing, predictive data sequences are gained by using GM(1,1) prediction model and then according to grey correlation and MAPE of each predictive data sequence, the threshold value meets conditions can be gained. Finally, the lifetime of HRG is predicted with using the threshold value. In this paper, the method is applied to the data of one type of HRG provided by a research institute in China and the result shows the gyroscope can normally run 4780 days at least, namely about 13.10 years.

Keywords: hemispherical resonator gyroscope (HRG), lifetime prediction; wavelet analysis, GM(1,1), MAPE, grey correlation

1 Introduction

With ever-developing space technologies of China, spacecraft need higher reliability, longer lifetime and higher efficiency now, say, satellites can work normally in orbit for 3 years or more, even 5-10 years. High cost and small number of flywheel and gyroscope, the indispensable parts in attitude control and attitude measuring unit of spacecraft makes it hard to estimate their lifetime which is a tough problem to solve. For gyroscope, it needs not only to detect the subtle angular displacement change and display rational response signal, but also to guarantee high reliability especially in unstaffed modern spacecraft, like satellites. A related survey [1] shows: 60% failure distribution of inertial system is from electronic circuit and 40% from inertial platform in which 60% is from gyroscope. Meanwhile, for domestic technological level, the failure rate of inertial platform and electronic circuit is half to half. Overall, researching reliability and lifetime prediction of gyroscope is important to inertial system. But worse still is that reports on the HRGs are not as many as other types of gyros, like dynamically tuned gyroscope (DTG) [2], MEMS gyro [3], fibre optic gyro [4]. For hemispherical resonator gyros, reference [5] analysed performances of HRGs with drift data and [6] researched the influences of temperature complement on navigation accuracy of hemispherical resonator gyros. In these studies, they didn't study the lifetime of the HRG, even no prediction methods mentioned in them. Therefore, we bring in grey system to analyse lifetime of the HRG.

The grey system theory is fairly appropriate for prediction. The accumulated generating operation [7] is the most important characteristic for the grey system theory and its purpose is to reduce the randomness of data. The main feature of grey theory is its capability of using as few as four data items to forecast the future data [7]. The grey prediction has been widely used in engineering sciences, social sciences [8, 9], power consumption [10, 11], as well as other fields [12, 13]. Moreover, various wavelets are also widely used to in signal processing and time series prediction, for example, reference [14] uses Haar filters to decompose time series data and predict, and reference [15] also adapts wavelet decomposition to time series prediction. Meanwhile, Daubechies wavelet is used to process biomedical signal in [16].

2 Lifetime prediction method with GM(1,1)

2.1 GREY PREDICTION MODEL

When modelling with grey system, pre-processing test data sequence is necessary and this process has two directions [12, 17]:

- 1) Provides intermediate information for modelling, i.e., finds out the law of raw data sequence.
- 2) Reduces the randomness of raw data sequence, that is, guarantees the correctness of the law.

Data producing mainly includes two operations: Accumulated Generating Operation (AGO) and Inverse Accumulated Generating Operation (IAGO).

* Corresponding author's e-mail: drizzlynight@163.com

2.1.1 Accumulated generating operation

The AGO is a process that every data amasses its previous one and then gain a monotonically increasing data sequence.

Let $X^0 = [x^0(1), x^0(2), \dots, x^0(n-1), x^0(n)]$ be the raw data sequence, where $x^0(i)$ is drift data at time i .

Do 1-AGO for X^0 , then a new data sequence X^1 is generated, $X^1 = [x^1(1), x^1(2), \dots, x^1(n-1), x^1(n)]$ and its derivation function is:

$$x^1(k) = \sum_{i=1}^k x^0(i), \quad k = 1, 2, \dots, n. \tag{1}$$

If do m-AGO for X^0 , its derivation function is

$$x^m(k) = \sum_{i=1}^k x^{m-1}(i), \quad k = 1, 2, \dots, n. \tag{2}$$

Generally speaking, for nonnegative data sequence, more time it amasses, more remarkable the randomness reduces. Namely, when data amasses many times, the data sequence would become nonrandom. Also, 1-AGO for test data sequence is normally enough in grey model.

2.1.2 Inverse accumulated generating operation

IAGO is the inverse operation of AGO, that is, IAGO is a process that every collected data subtracts its previous one and then gain a new data sequence.

According to Equation (1), it's easy to regain X^0 from X^1 using 1-IAGO, and its process is

$$x^0(k) = x^1(k) - x^1(k-1), \quad 2 \leq k \leq n, \tag{3}$$

where $x^0(1) = x^1(1)$.

Accordingly, do m-IAGO:

$$x^{m-1}(k) = x^m(k) - x^m(k-1), \quad 2 \leq k \leq n, \tag{4}$$

where $x^{m-1}(1) = x^m(1)$.

2.2 GM(1,1) MODEL

1) Set sequence $X^0 = [x^0(1), x^0(2), \dots, x^0(n-1), x^0(n)]$ denote the drift of the gyro, where $x^0(i)$ is the output at time i .

2) When X^0 is subjected to 1-AGO, then the following monotonically increasing sequence X^1 is obtained:

$$X^1 = [x^1(1), x^1(2), \dots, x^1(n-1), x^1(n)]. \tag{5}$$

3) The whitening equation of X^1 is therefore, as follows:

$$\frac{dx'(k)}{dk} + ax'(k) = u, \tag{6}$$

in above, $[a \ u]^T$ is the parameters matrix that can be got as step 4) shows.

4) Parameters \hat{a} can be obtained by using least square method:

$$\hat{a} = \begin{bmatrix} a \\ u \end{bmatrix} = (B^T B)^{-1} B^T y_N, \tag{7}$$

where

$$B = \begin{bmatrix} -\frac{1}{2}(x'(1)+x'(2)) & 1 \\ -\frac{1}{2}(x'(2)+x'(3)) & 1 \\ \dots & 1 \\ -\frac{1}{2}(x'(n-1)+x'(n)) & 1 \end{bmatrix} \quad y_N = \begin{bmatrix} x^0(2) \\ x^0(3) \\ \dots \\ x^0(n) \end{bmatrix}.$$

5) According to Equation (6), the solution of $\hat{x}^1(k)$ at time k is:

$$\hat{x}^1(k) = (x(1) - \frac{u}{a})e^{-ak} + \frac{u}{a}. \tag{1}$$

To obtain the predicted value of the primitive data, \hat{X}^0 , the IAGO is used for \hat{X}^1 , and then

$$\hat{X}^0 = [\hat{x}^0(1), \hat{x}^0(2), \dots, \hat{x}^0(n)]. \tag{2}$$

where, $\hat{x}^0(k+1) = \hat{x}^1(k+1) - \hat{x}^1(k)$, $k = 1, 2, \dots, n-1$, and $\hat{x}^0(1) = \hat{x}^1(1)$.

2.3 GREY CORRELATION

Grey correlation analysis can evaluate the compact degrees of two data sequences by using similarity degrees of their curves. The more similar the two curves they are, the larger correlation they have. And grey correlation is adopted to forecast short-term power in [18]. Grey correlation $\gamma(X_0, X_i)$ or γ_{0i} can be obtained as follows:

1) First of all, calculating range profile of each sequence:

$$X'_i = X_i / x_i(1) = (x'_i(1), x'_i(2), \dots, x'_i(n)), \tag{10}$$

$$i = 0, 1, 2, \dots, m.$$

2) Then calculating difference sequence:

$$\Delta_i(k) = |x'_0(k) - x'_i(k)|, \tag{11}$$

$$\Delta_i = (\Delta_i(1), \Delta_i(2), \dots, \Delta_i(n)), \quad i = 0, 1, 2, \dots, m.$$

3) Next step: Calculating maximum difference and minimum difference:

$$M = \max_i \max_k \Delta_i(k), \quad m = \min_i \min_k \Delta_i(k). \quad (12)$$

4) And calculating correlation coefficient $\gamma(x_0(k), x_i(k))$ or $\gamma_{0i}(k)$ at point k :

$$\gamma_{0i}(k) = \frac{m + \xi M}{\Delta_i(k) + \xi M}, \quad \xi \in (0,1), \quad k = 1, 2, \dots, n, \quad (13)$$

$i = 1, 2, \dots, m.$

5) At last, calculating grey correlation:

$$\gamma_{0i} = \frac{1}{n} \sum_{k=1}^n \gamma_{0i}(k). \quad (14)$$

2.4 MEAN ABSOLUTE PERCENT ERROR AND PREDICTION ACCURACY

To guarantee the reliability of prediction data, mean absolute percent error (MAPE) is used to estimate the reliability and it can be gained:

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%. \quad (15)$$

where y is primitive data, \hat{y} is predictive data, and N is length of \hat{y} .

Therefore, the prediction accuracy can be obtained:

$$P = 100\% - MAPE. \quad (16)$$

3 Data pre-processing with wavelet analysis

3.1 DATA

The test data of the HRG used in the paper is provided by China Electronics Technology Group Corporation 26th Research Institute. Besides, the duration of experiments is from June 26, 2009 to February 8, 2012, 956 days. And there is 1590 data points in total and all of them are positive with which GM(1,1) can be used to forecast.

3.2 DATA PREPROCESSING

To reduce randomness and noise of test data, Daubechies wavelet is adopted to pre-process test data. Because of low sampling frequency, test data is decomposed by 'db5' wavelet function in 3-scale, 6-scale, 9-scale respectively and then reconstructed low-frequency part respectively. The filtering results are shown in Figure 1.

Figure 1 shows Daubechies wavelet filters some high frequency away and expresses law of test data well. Result of 3-scale filtering is over fitting as well as that of 9-scale is under fitting, and 6-scale is properly fitting, therefore, the reconstruction data of 6-scale is adopted to predict lifetime of this HRG.

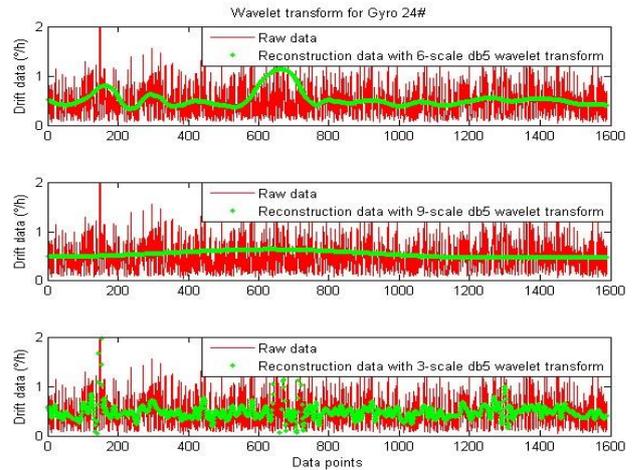


FIGURE 1 Filtering with Daubechies wavelet analysis

4 Predictive results and analysis

Predict 6 groups of data sequences using original data and pre-processed data with Daubechies wavelet respectively, and two residual sums of squares (RSS) are worked out. The results show that the residual sum of squares with Daubechies wavelet is 22.61 and MAPE is 11.3% as while as that of original data is 3421.33 and MAPE is 39.5% (Shown in Table 1).

TABLE 1 RSS and MAPE

| | RSS | MAPE |
|------------------------------------|---------|-------|
| Prediction with original data | 3421.33 | 39.5% |
| Prediction with pre-processed data | 22.61 | 11.3% |

According to Table1, it's known that the prediction using Daubechies transform is much better than that without pre-processing, so in the paper, predictive data sequences got from pre-processed data are used in lifetime prediction of HRG, not using those of without pre-processing.

To predict lifetime of the gyro, grey correlation and mean absolute percent error are used to estimate whether the gyro is invalid or not. It is obvious the further prediction is away from samples, the lower the grey correlation is. Namely, grey correlation goes down while prediction data is generated. Therefore, when grey correlation suddenly raises, so the value could be the threshold which means the gyro is invalid after the point. Meanwhile, if the average accuracy at that value is not less than 60%, so the value is the threshold. Otherwise, let the former grey correlation compare with 60%, if it is also less than 60%, do this process until one grey correlation's average accuracy is equal to or more than 60%. Finally, the value meets the two conditions is the threshold.

4.1 GREY CORRELATIONS OF 24#HRG

6 groups of predictive data are obtained by using grey prediction model and each grey correlation is worked out as well (Shown in Table 2).

TABLE 2 Grey correlations of 6 groups of prediction data sequence

| Prediction Data | 1st time | 2nd time | 3rd time | 4th time | 5th time | 6th time |
|------------------|----------|----------|----------|----------|----------|----------|
| Grey Correlation | 0.814 | 0.766 | 0.741 | 0.723 | 0.749 | 0.768 |

4.2 AVERAGE ACCURACY

According to Equations (15) and (16), calculate MAPEs of 6 groups of predictive data sequences, and then work out average accuracy of every group, as Table 3 shows below.

TABLE 3 Average Accuracy of 6 groups of prediction data sequence

| Prediction Data | 1st time | 2nd time | 3rd time | 4th time | 5th time | 6th time |
|----------------------|----------|----------|----------|----------|----------|----------|
| Average Accuracy (%) | 88.3 | 80.0 | 70.7 | 60.3 | 48.4 | 35.8 |

Moreover, grey correlations and average accuracies of 6 groups of prediction data sequence also shows in Figure 2.

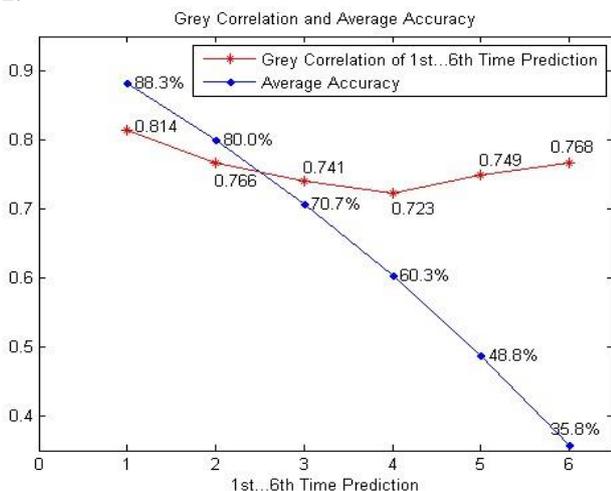


FIGURE 2 Grey Correlations and Average Accuracy of 6 Groups of Prediction Data Sequence

Reference

- [1] Feng P D, Zhang B J, Fu T 1996 Application of Reliability Engineering Technology to Aerial Inertial Navigation Systems *Journal of Chinese Inertial Technology* 4 56-65
- [2] Cain J, Heppler G, McPhee J, Staley D 2006 Stability Analysis of a Dynamically Tuned Gyroscope *Journal of Guidance, Control, and Dynamics* 29(4) 965-9
- [3] Kirkko-Jaakkola M, Collin J, Takala J 2012 Bias Prediction for MEMS Gyroscopes *IEEE Sensors Journal* 12(6) 2157-63
- [4] Lloyd SW, Fan S, Digonnet M J F 2013 Experimental observation of low noise and low drift in a laser-driven fibre optic gyroscope *Journal of Lightwave Technology* 31(13) 2079-85
- [5] Peng H, Fang Z, Lin K, Zhou Q, Jiang C Q 2008 Error Analysis of

Figure 2 shows the average accuracy of 4th group is 60.3%. Both of 5th and 6th are less than 60%. Meanwhile, the grey correlation of 4th group predictive data sequence is under the threshold (0.723). Therefore, we predict that the gyro can normally work as 4 times long as it already ran for now. Namely, the gyro can at least run 3824 days ($956 \cdot 4 = 3824$). And then add test period 956 days to predictive time, finally we calculate that the gyro can work at least 4780 days, that is, 13.10 years.

5 Conclusions

This paper uses Daubechies wavelet to decompose test data sequence and reconstruct it for reducing noise in it. And then GM(1,1) is applied to predict several sequences. Moreover, the predictive result of this method is much better than that of directly using original data. Meanwhile, grey correlation and MAPE are adopted to estimate prediction data and then find out the threshold value which helps determine the lifetime of the hemispherical resonator gyroscope. Applying the test data of one type of gyro experimented by one research institute to this prediction process, this type of gyro can normally work at least 4780 days, namely 13.10 years. According to the 10 global longest spacecraft: Voyager 2 (1977.8-), Voyager 1 (1977.9-), GOES 3 (1978.6-), ATS-3 (1967.11-2001), Mirasat F2 (1976.6-2008.10), Landsat 5 (1984.3-2012.12), TDRS-1 (1983.4-2009), GOES 7 (1987.2-2012.4), TDRS-3 (1988.9-), and GOES 2 (1977.6-2001), all of them can work more than 24 years, it means some gyroscopes, as the inertial unit in satellite can also run more than 24 years, therefore, in the paper, our predictive result is receivable and our method is reliable as well.

Acknowledgements

The authors wish to acknowledge the assistance of China Electronics Technology Group Corporation 26th Research Institute for providing the test data of HRG.

- [6] Li B, Wu Y, Wang C 2010 The Identification and Compensation for Temperature Model for Hemispherical Resonator Gyro Signal *Proceedings of the 3rd International Symposium on Systems and Control in Aerospace and Astronautics (ISSCAA 2010)* Harbin China 398-401
- [7] Liu S, Lin Y 2010 Grey systems: Theory and Practical Applications *London: Springer-Verlag London Ltd*
- [8] Yin M S, Tang H W 2013 On the fit and forecasting performance of grey prediction models for China's labour formation *Mathematical Hemispherical Resonator Gyro Drift Data Proceedings of the 2nd International Symposium on Systems and Control in Aerospace and Astronautics (ISSCAA 2008)* Shenzhen China Dec 1-4

- and *Computer Modelling* 57(3-4) 357-65
- [9] Chang T S, Ku C Y, Fu H P 2013 Grey theory analysis of online population and online game industry revenue in Taiwan *Technological Forecasting and Social Change* 80(1) 175-85
- [10] Li G D, Masuda S, Nagai M 2013 The prediction model for electrical power system using an improved hybrid optimization model, *International Journal of Electrical Power & Energy Systems* 44(1) 981-7
- [11] Wang J. Z, Ma, X L, Wu J, Dong Y 2012 Optimization models based on GM(1,1) and seasonal fluctuation for electricity demand forecasting *International Journal of Electrical Power & Energy Systems* 43(1) 109-17
- [12] Tien T 2009 A new grey prediction model FGM(1,1) *Mathematical and Computer Modeling* 49(7-8) 1416-26
- [13] Kayacan E, Ulutas B, Kaynak O 2010 Grey system theory-based models in time series prediction *Expert Systems with Applications* 37(2) 1784-9
- [14] Soltani S 2002 On the use of the wavelet decomposition for time series prediction *Neurocomputing* 48(1-4) 267-77
- [15] Mabrouk A B, Abdallah N B, Dhifaoui Z 2008 Wavelet decomposition and autoregressive model for time series prediction *Applied Mathematics and Computation* 199(1) 334-40
- [16] Rafiee J, Rafiee M A, Prause N, Schoen M P 2011 Wavelet basis functions in biomedical signal processing *Expert Systems with Applications*, 38(5) 6190-201
- [17] Kayacan E, Ulutas B, Kaynak O 2010 Grey system theory-based models in time series prediction *Expert Systems with Applications* 37(2) 1784-9
- [18] Jin M, Zhou X, Zhang Z, Tentzeris M M 2012 Short-term power load forecasting using grey correlation contest modeling *Expert Systems with Applications* 39(1) 773-9

| Authors | |
|---|--|
|  | <p>Yunxia Zhang, born on November 26, 1974, Nanjing, China</p> <p>Current position, grades: Network and Information Centre, Southeast University.</p> <p>University studies: M.S. degree on Computer Science and Technology from Nanjing Normal University, China, 1989.</p> <p>Scientific interest: data mining, data analysis.</p> <p>Publications: 6 papers.</p> |
|  | <p>Chenglong Dai, born on August 4, 1989, Chengdu, China</p> <p>Current position, grades: Computer Science and Technology in Nanjing University of Aeronautics and Astronautics.</p> <p>University studies: M.S. degree on Computer Science and Technology from Nanjing University of Aeronautics and Astronautics, China, 2011.</p> <p>Scientific interest: data mining, data analysis.</p> <p>Publications: 4 papers.</p> |
|  | <p>Jifeng Cui, born on March 16, 1988, Tsingtao, China</p> <p>Current position, grades: Computer Science and Technology in Nanjing University of Aeronautics and Astronautics.</p> <p>University studies: M.S. degree on Computer Science and Technology from Nanjing University of Aeronautics and Astronautics, China, 2011.</p> <p>Scientific interest: data mining, data analysis.</p> <p>Publications: 3 papers.</p> |

Automatic license plate detection based on colour gradient map

Xiaodong Huang*

Capital Normal University, Beijing 100048, China

Received 1 March 2014, www.tsi.lv

Abstract

License plate detection plays a key role in traffic surveillance, speeding vehicles ticketing and vehicle detecting, and so on. However, most of the previous approaches to detect license plate experience difficulties in handling license plate with the uneven illuminations changes, complex background or tilted alignments. In this paper, we propose a method of license plate detection. License plate regions contain plate characters, frames and screws. First we propose to build the Colour Gradient Map (CGM) based on the colour gradient method. Then we perform the Niblack's method on the Colour Gradient Map (CGM) to retrieve the candidate license plate regions. Finally, we use the template matching to remove most of background noises. Experimental results show that this approach is robust and can be effectively applied to license plate detection.

Keywords: license plate detection, colour gradient, template matching

1 Introduction

With the rapid growth of city traffic, there is an urgent demand for intelligent transportation systems. The automatic license plate detection normally can be applied in various applications of intelligent transportation systems, such as traffic surveillance, speeding vehicles ticketing, vehicle detecting and stolen vehicle verification, and so on. As a result, automatic license plate detection is vital importance for intelligent transportation systems.

Although some papers (e.g. [1-12]) proposed some methods to detect the license plate, they have difficulties in detecting license plate in the situation, such as the uneven illuminations changes, complex background or tilted license plate. License plate regions contain plate characters, frames and screws. However, due to various cameras observation angles, the frames and screws will connect the plate characters with other regions, which is difficult to accurately detect the license plate. Therefore, we propose to build the Colour Gradient Map (CGM) (to be described in Section 3) based on the colour gradient method [13]. Then we perform the Niblack's method on the Colour Gradient Map (CGM) to retrieve the license plate regions.

The rest of this paper is organized as follows. Section 2 reviews the related work. Colour Gradient Map produced by our proposed method is described in Section 3. License plate detection is described in Section 4. Experimental results are presented and discussed in Section 5. Finally, in Section 6, we draw conclusion.

2 Related work

Current approaches on the license plate detection can be classified into three classes: Morphology-based methods, local features-based method, and Learning-Based methods.

The first class uses morphology-based methods [1-3] to detect license plate. Hsieh et al. [1] proposed a morphology-based method for detecting license plates. First, they used a morphology-based method to extract contrast features to search the desired license plates. Then, they applied a recovery algorithm for reconstructing a license plate if the plate is fragmented into several parts. Finally, they performed the license plate verification.

The second class uses the local features-based methods [4-7] to detect license plate. Zhou et al. [4] proposed a license plates detection method by principal visual word (PVW). They automatically discover the PVW characterized with geometric context. Given a new image, the license plates are extracted by matching local features with PVW. Due to the relatively expensive time cost in feature extraction, Zhou's approach is suitable for applications without strong requirement of real-time efficiency. Chen et al. [5] proposed a license plates detector based on a modified convolutional neural network (CNN) verifier. In the proposed verifier, a single feature map and a fully connected MLP were trained by examples to classify the possible candidates. They applied the Pyramid-based localization techniques to fuse the candidates and to identify the regions of license plates. Then, geometrical rules filtered out false alarms in license plate detection. Clemens Arth et al. [6] proposed a full-featured license plate detection system. They detect the license plate using the detector based on the

* *Corresponding author's* e-mail: dawn_hxd@yeah.net

AdaBoost approach. Detected license plates are segmented into individual characters by using a region-based approach.

The third class uses the learning-based methods [8-11] to detect license plate. Zhang et al. [9] proposes a license plate detection algorithm using both global statistical features and local Haar-like features. Classifiers using global statistical features are constructed firstly through simple learning procedures. Then the AdaBoost learning algorithm is used to build up the other classifiers based on selected local Haar-like features. Combining the classifiers using the global features and the local features, they obtain a cascade classifier. They construct the cascade classifier for license plate detection using both global and local features.

Different from the above three kinds of detecting license plate methods, some proposed other approaches recently. Lin et al. [12] proposed a license plate detection algorithm based on image saliency. The proposed algorithm consists of two parts. The first part segments out the characters on a license plate using an intensity saliency map with a high recall rate. The second part applies a sliding window on these characters to compute some saliency-related features to detect license plates.

3 Colour gradient map

License plate regions contain not only plate characters but also various adornments such as frames, screws. However, due to various cameras observation angles, the frames and screws will connect the plate characters with other regions, which is difficult to accurately detect the license plate. Therefore, we propose to build the Colour Gradient Map (CGM) based on the colour gradient method [13].

We use the colour gradient method to process the image. We use f to represent a colour image, R, G, B are the three colour bands of colour space RGB, respectively.

$$f(x, y) = \begin{bmatrix} R(x, y) \\ G(x, y) \\ B(x, y) \end{bmatrix}. \tag{1}$$

Then we define g_{xx} , g_{yy} , g_{xy} as follows:

$$g_{xx} = \left(\frac{\partial R}{\partial x}\right)^2 + \left(\frac{\partial G}{\partial x}\right)^2 + \left(\frac{\partial B}{\partial x}\right)^2, \tag{2}$$



a)

$$g_{yy} = \left(\frac{\partial R}{\partial y}\right)^2 + \left(\frac{\partial G}{\partial y}\right)^2 + \left(\frac{\partial B}{\partial y}\right)^2, \tag{3}$$

$$g_{xy} = \frac{\partial R}{\partial x} \frac{\partial R}{\partial y} + \frac{\partial G}{\partial x} \frac{\partial G}{\partial y} + \frac{\partial B}{\partial x} \frac{\partial B}{\partial y}. \tag{4}$$

The gradient orientation in coordinate (x,y) is $\theta(x, y)$; the gradient magnitude in coordinate (x,y) is $F_\theta(x, y)$, they can be calculated by [13]:

$$\theta(x, y) = \frac{1}{2} \tan^{-1} \left(\frac{2g_{xy}}{g_{xx} - g_{yy}} \right), \tag{5}$$

$$F_\theta(x, y) = \sqrt{\frac{1}{2} [(g_{xx} + g_{yy}) + (g_{xx} - g_{yy}) \cos 2\theta + 2g_{xy} \sin 2\theta]}. \tag{6}$$

We convolves the $f(x,y)$ with the averaging filters s via Equation (7) to get the mean colour image $fa(x,y)$.

$$s = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \tag{7}$$

$$fa(x, y) = \begin{bmatrix} Ra(x, y) \\ Ga(x, y) \\ Ba(x, y) \end{bmatrix}, \tag{8}$$

$$cgm(x, y) = \begin{bmatrix} \|Ra(x, y) - F_\theta(x, y)\| \\ \|Ga(x, y) - F_\theta(x, y)\| \\ \|Ba(x, y) - F_\theta(x, y)\| \end{bmatrix}. \tag{9}$$

Because the colour gradient magnitude can represent the colour differences remarkably, we use the mean colour image subtract the gradient magnitude $F_\theta(x, y)$. As a result, we can get the Colour Gradient Map (CGM) via the Equation (9). The CGM can keep the license plate character regions completely and remove the colour difference, which can make the character edge details clearly. As a result, on the CGM the edges of character do not connect with the frames or screws. The Figure 1b is the CGM, compared with the original image Figure 1a, we can find that the license plate character has whole contour and do not connect with the screw or frames in the CGM.



b)

FIGURE 1 a) Original image, b) Colour Gradient Map on original image

4 License plate detection

License Plate detection is difficult due to uneven illuminations changes, complex background or tilted license plate. Niblack’s method [14] presents a low-complexity method for automatically detecting text of any sizes, fonts, and alignments from images. However, Niblack’s method relies on the local mean and standard deviation, which is sensitive to local abnormal intensity change. Because Colour Gradient Map (CGM) has remarkably made the character edge details clearly, it is suitable for the Niblack’s method. Therefore, we perform the Niblack’s method not on the original image but on the Colour Gradient Map (CGM). After performing the Niblack’s method, we use the connected component analysis to remove the background noises.

4.1 NIBLACK’S METHOD

Niblack method can segment image into three different layers WonB, BonW and EonB. WonB refers to the White foreground on Black background. BonW refers to the Black foreground on the White background. The EonB refers to Edge on the Black background.

$$WonB(x, y) = \begin{cases} 1 & f(x, y) > T_+ \\ 0 & otherwise \end{cases}, \tag{10}$$

$$BonW(x, y) = \begin{cases} 1 & f(x, y) < T_- \\ 0 & otherwise \end{cases}, \tag{11}$$

$$EonB(x, y) = \begin{cases} 1 & T_- < f(x, y) < T_+ \\ 0 & otherwise \end{cases}, \tag{12}$$

$$T_+(x, y) = \mu(x, y) + a\sigma(x, y), \tag{13}$$

$$T_-(x, y) = \mu(x, y) - a\sigma(x, y), \tag{14}$$

where Niblack threshold, T_+ and T_- , are calculated based on μ and σ , which are the mean and standard deviation in a neighbourhood window ($h \times w$), and a is the constant which can be got by experiments. Figure 2 shows the WonB, BonW, EonB which is segmented by Niblack’s method. Because the license plate license plates must be very salient to human visual observation, the license plate will always keep high contrast on the background. Therefore, the license plate will always identified by the WonB.

4.2 CONNECTED COMPONENT ANALYSIS

We perform the connected component analysis (CCA) on the WonB which is got by the Niblack’s method. We use the following simple rules to perform the CCA on WonB.

Rule 1: We assume that the license plate will not occupy the whole image or occupy only small regions. As a result, we will remove some too small regions or too big regions.

Rule 2: Normally, the license plate will be surrounded by the frames, which will produce some backgrounds interference. So we will scan the WonB in horizontal line, and remove the lines which width is bigger than the one twentieth of the image width.

Rule 3: The license plate aligns in horizontal way, and generally the license plate contains at least five to seven characters. Therefore, we will remove the candidate regions when its width is bigger than the one tenth of the image width.

After the CCA on the WonB, we can remove some background interference, which is shown in Figure 2d.

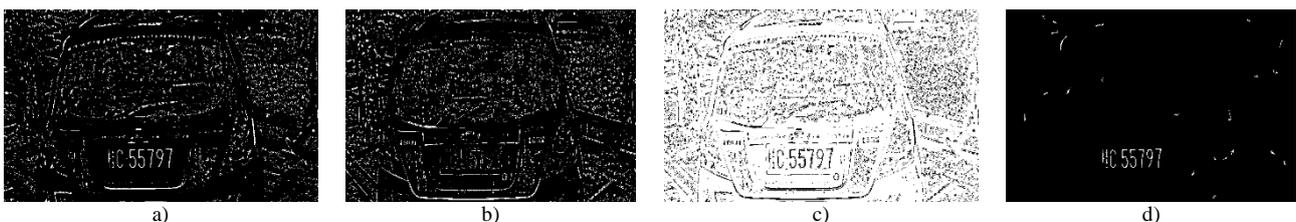


FIGURE 2 a), b), c) Niblack’s method segmentation results, d) CCA on the WonB

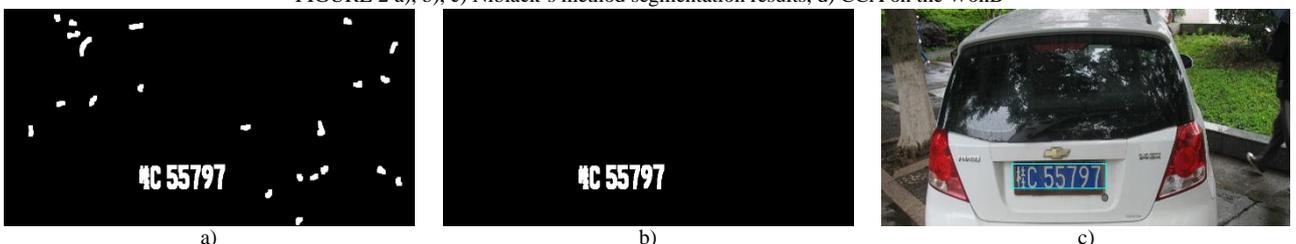


FIGURE 3 a) dilation morphological operation, b) template matching filtered results, c) final license plate detection results

4.3 CANDIDATE REGIONS FILTERED BASED ON TEMPLATE MATCHING

We use the matching of correlation [15] to realize the template matching. According to the matching of

correlation [15], the spatial correlation can be obtained as the inverse Fourier transform of the product of the transform of one function times the conjugate of the transform of the other, which is shown in Equation (15).

$$f(x, y) \circ w(x, y) \Leftrightarrow F(u, v)H^*(u, v), \tag{15}$$

where “o” indicate the correlation and “*” indicate the complex conjugate.

Gonzalez at el. [15] proposed that given an image $f(x,y)$, the correlation is to find all places in the image that

match a given template $w(x,y)$. The best match of $w(x,y)$ in $f(x,y)$ is the location of the maximum value. As a result, we can get the matching of correlation in the frequency domain.



FIGURE 4 License Plate Detection Results



FIGURE 5 Examples with Missed and False Detections

TABLE 1 Performance Comparison for License Plate Detection

| | Total License Plate | Total Missed Textboxes | Total False Alarm | Detection Rate | False Alarm Rate | Detection Speed (Second/Per Image) |
|----------------|---------------------|------------------------|-------------------|----------------|------------------|------------------------------------|
| Zhou's Method | 300 | 29 | 11 | 90.3% | 3.7% | 0.35 |
| Hsieh's Method | 300 | 36 | 9 | 88.0% | 3.0% | 0.43 |
| Our Method | 300 | 23 | 7 | 92.3% | 2.3% | 0.29 |

We find that license plate character is composed of Arabic number character and alphabetic character. As a result, we build a template set which include the 0-9 Arabic number character and A-Z alphabetic character. For some Chinese license plate, when we implement the template matching between the Chinese characters with the template set, the best matching value of Chinese character is far bigger than that of background noises. As a result, we can remove most of background noises based on the template matching. Before the template matching for the candidate regions, we perform the dilation morphological operation, which make the template

matching performance more accurately. Figure 3 shows the whole process. In Figure 3a, we implement the dilation morphological operation on the CCA results. Figure 3b is the filtered results got by the template matching. Compared Figure 3b with Figure 3a, we can find that the background interference can be removed by the template matching.

5 Experiments and discussion

We have collected 300 license plate images. Our test data is composed of the Chinese license plate. Figure 4 shows

the experimental results of license plate detection. In Figures 4(a-c), the license plate has different alignment. The detected results demonstrate that our approach is robust to detect such kind of license plate. The license plate which have light illumination changes Figures 4(d-f) are correctly detected. Figures 4 g and 4h demonstrate that our approach is robust to detect license plate with low resolution. Figure 4i shows that our method can detect the license plate when the license plate was blurring. These results also confirm that the proposed license plate detection algorithms are capable of handling the light uneven illuminations changes, complex background or tilted license plate.

Figure 5 shows some examples of misses or false detections. In Figure 5a the license plate character is too big to detect by our method. In Figure 5b, due to license plate regions is too dark, although our method can detect the license plate correctly, we also detect some other background regions. In Figure 5c, the license plate image is too blurred and the license plate character is too small, so our method cannot detect the license plate correctly. In Figure 5d, those false detected regions have similar texture as the text, meanwhile the license plate has serious stains which result the false detection. As a result, we can conclude that our method will cause some misses in license plate which have too big or small character, low contrast character and over-blurring character, and cause some false alarms which the license plate have serious stain.

References

- [1] Hsieh J W, Yu S-H, Chen Y-S 2002 Morphology-based License Plate Detection from Complex Scenes *International Conference on Pattern Recognition* 3 176-9
- [2] Qiu Y, Sun M, Zhou W 2009 License plate extraction based on vertical edge detection and mathematical morphology *International Conference on Computational Intelligence and Software Engineering* 1-5
- [3] Yichuan L, Chunhong Z 2011 Vehicle license plate location based on mathematical morphology and variance projection *International Conference on Image Analysis and Signal Processing (IASP)* 360-3
- [4] Zhou W, Li H, Lu Y, Tian Q 2012 *IEEE Transaction on IP* 21(9) 4269-79
- [5] Chen Y-N, Han C-C, Wang C-T, Jeng B-S, Fan K-C 2006 The Application of a Convolution Neural Network on Face and License Plate Detection *International Conference on Pattern Recognition* 3 552-5
- [6] Arth C, Limberger F, Bischof H 2007 Real-Time License Plate Recognition on an Embedded DSP-Platform *IEEE Conference on Computer Vision and Pattern Recognition* 1-8
- [7] Wang W, Jiang Q, Zhou X, Wan W 2011 Car license plate detection based on MSER *International Conference on Consumer Electronics, Communications and Networks* 3973-6
- [8] Lee Y, Song T, Ku B, Jeon S, Han D K, Ko H 2010 License plate detection using local structure patterns *International Conference on Advanced Video and Signal Based Surveillance* 574-9
- [9] Zhang H, Jia W, He X, Wu Q 2006 Learning-Based License Plate Detection Using Global and Local Features *International Conference on Pattern Recognition* 2 1102-5
- [10] Lim H W, Tay Y H 2010 Detection of license plate characters in natural scene with MSER and SIFT unigram classifier *IEEE Conference on Sustainable Utilization and Development in Engineering and Technology* 95-8
- [11] Lim H W, Tay Y H 2009 Two-stage license plate detection using gentle Adaboost and SIFT-SVM *First Asian Conference on Intelligent Information and Database Systems* 109-14
- [12] Lin K-H, Tang H, Huang T S 2010 Robust license plate detection using image saliency *IEEE International Conference on Image Processing* 3945-8
- [13] Zenzo S D 1986 A Note on the Gradient of a Multi-image *Computer Vision, Graphics and Image Processing* 33(1) 116-125
- [14] Winger L L, Robinson J A, Jernigan M E 2000 Low-complexity character extraction in low-contrast scene images *International Journal of Pattern Recognition and Artificial Intelligence* 14(2) March 113-35
- [15] Gonzalez R C, Woods R E 2008 Digital Image Processing *Prentice Hall* 490-2

6 Conclusions

A novel approach to detect license plate on the basis of the colour gradient map is proposed in the paper. Our experimental results and the comparisons with other methods show that our method is robust to detect license plate with the light uneven illumination changes, complex background or tilted alignments. The known limitation of license plate detection is that license plates with severe illumination changes cannot be detected. These issues will be addressed in our future research.

Authors



Xiaodong Huang, born on December May, 1974, Beijing, China

Current position, grades: doctor of computer science, lecturer in Capital Normal University.

University studies: M.S. in computer science (2003-2006) at Beijing University of Posts and Telecommunications, Ph.D. degree in Computer Science (2007-2010).

Scientific interest: pattern recognition and computer vision, video text detection, localization and extraction scene text detection and people tracking.

Publications: 2 patents, 10 papers

Prediction model of recast layer thickness in die-sinking EDM process on Ti-6Al-4V machining through response surface methodology coupled with least squares support vector machine

Jun Li*, Xiaoyu Liu, Shiping Zhao

School of Manufacturing science and Engineering, Sichuan University, Chengdu 610065, P. R. China

Received 10 July 2014, www.tsi.lv

Abstract

Ti-6Al-4V is widely applied in frontier for its excellent properties such as a high strength-weight ratio, great heat stability and exceptional corrosion resistance. Electrical discharge machining (EDM) is suitable for machining titanium alloys, because it is the technical that removal materials by discharge energy and non-contact in processing progress. The recast layer is formed by the solidification of molten metal on the machined surface during the EDM process. In the present investigation, a hybrid approach using Least squares support vector machines (LS-SVM) and response surface methodology (RSM) for predication the recast layer thickness is proposed. Experimental plan is performed by response surface method with 20 experimental runs. The different machining parameters of pulse current, pulse on-time, and pulse off-time are selected as input factors. The white layer thickness (WLT) is response variable. The LSSVM method is applied to construct the predication model based on the orthogonal experiment swatches. The randomly 15 experimental runs were utilized to train the LS-SVM model to predict the WLT. Finally, support vector machine is used to compare with the proposed method. The proposed model can be good performance in prediction of white layer thickness of the complex EDM process.

Keywords: Electrical discharge machining, Least squares support vector machines, Response surface methodology, Recast layer

1 Introduction

Electrical discharge machining (EDM) is directly to use the electrical energy and heat energy to fabricate the workpiece. In the machining process, the material is wiped out of the workpiece just by a succession of electrical discharges occurring between the workpiece and the electrode which is not contacted with each other and produce local and instantaneous high temperatures [1]. EDM is used widely in machining special structure and complex shape parts in aerospace and nuclear sector by reason of it can machine every conductive material effectively and economically with no obvious mechanical cutting force, which has no limit on the hardness, brittleness tenacity and melting point of the workpiece material. Its typical applications include the processing of cooling holes on turbine blades and fuel nozzles [2, 3]. Because the material is removed by melting and vaporization, the resolidified/recast layer is inevitable to produce on the top surface of the workpiece by subsequently resolidifies and cools at a high rate. When the recast layer is observed by scanning electron microscope, the layer is white and can be called the white layer. It contains numerous pock marks, globules, cracks and microcracks and will influence the fatigue life of parts. Various researchers have made a great deal work to

optimize and reveal the relationship between the input parameters and output parameters like metal removal rate (MRR), tool wear rate (TWR), and surface finish. However, the efforts are less concentrated towards the white layer thickness and tool wear ratio. According to the research of Ti-6Al-4V alloy machining by EDM about recast layer/white layer is less. Because of the Ti-6Al-4V alloy properties such as high strength-to-weight ratio, high temperature stability and good corrosion resistance are classified as difficult-to-cut materials [4]. However, the Ti-6Al-4V alloy is commonly used in the important industries such as aerospace; the recast layer/white layer machined by EDM will have a great effect on the finished workpiece.

Some investigations have been conducted on MRR, EW and WLT in the EDM/micro-EDM process. H. Ramasawmy [5] made an attempt to investigate the relationship between the EDM process factors (current and pulse on time) and the thickness of the white layer. It correlates the thickness of the white layer with 3D surface roughness parameters and reveals a better correlation between the average thick ness of the white layer and the spatial parameters. Ahmet [6] carried out the experiments to machine the Ti-6Al-4V with different electrode materials (graphite, electrolytic copper and aluminium) using the process parameters (pulse current and pulse

* *Corresponding author* e-mail: lijunj08@163.com

duration). It was noted that the value of material removal rate, surface roughness, electrode wear and average white layer thickness increase accompanying with the increasing current density and pulse duration. Among the different electrode materials, the graphite electrode is best choice on material removal rate, electrode wear and surface crack density although the poorer surface finish. Unfortunately, this experiment just reveals the relationship between the average white layer thickness and the process parameters with no mathematic model. Ulas caydas and ahmet hascalik [7] made an attempt to model electrode wear and recast layer thickness through response surface methodology (RSM) in a die-sinking EDM process. Analysis of variance (ANOVA) was applied to study and pointed out the pulse current was the most important factor related to the EW and WLT, but the pulse off time is not important factor. B. Jabbaripour [8] changed the main machining parameters just as pulse current, pulse on time and open circuit voltage during EDM tests. Analysis of variance (ANOVA) was done and revealed the current and voltage have significant effect on the MRR, the current. In the same way the pulse on time, voltage and current have significant effect on the tool wear ratio. It was reported that the recast layer thickness has great relationship with the pulse energy based on pulse on time and pulse current variations. Zhang [9] performed support vector machine (SVM)/genetic algorithm (GA) to settle of the optimal micro-EDM processing (discharge pulse, pulse on time, pulse off time, capacitance, electrode rotating speed, and servo reference speed) to minimum processing time and electrode wear. It was reported that a new multi-objective optimization GA based on the idea of non-dominated sorting had a great performance on the micro-EDM processing. Tzeng and Chen [10] applied response surface methodology and genetic algorithm approach to model and optimize EDM process parameters for SKD61. Meanwhile the result had been compared with the SVM and BPNN/GA. Somashekhar [11] established the parameter optimization model to analyse the material removal of Micro-EDM by making use of the artificial neural network (ANN). The genetic algorithms (GAs) have been applied to optimize the best process parameters.

In this study, 20 experiments were carried out which was based on the design of response surface methodology, the least squares support vector machines and response surface methodology were proposed and applied to model and optimize EDM processing parameters for Ti-6Al-4V. Simultaneously, a mathematical predictive model based on the statistical learning theory was selected to predict the white layer thickness (WLT). The WLT was observed through scanning electron microscope. Finally, the comparison of the different approaches of LS-SVM/RSM and SVM were also conducted.

2 Description of the experimentation

2.1 MATERIAL

Ti-6Al-4V is a widely material applied to the aerospace, automotive and biomedical for its excellent properties in mechanical and thermal. The composition of the Ti-6Al-4V is 89.464wt%Ti, 6.08wt%Al, 4.02wt%V, 0.22wt%Fe, 0.18wt%O, 0.02wt%C, 0.01wt%N, 0.053wt%H. The hardness of Ti-6Al-4V is 600, the yield strength of it is 745MPa and elastic modulus is 113GPa.

2.2 EXPERIMENTAL INSTALLATIONS

The series of experiment were performed on a die-sinking EDM machine of type MITSUBISHI ELECTRIC-EX22 shown in Figure 1 and the model was FP60E of 8.7 KVA machine unit input. The electrode was made of a pure cylindrical copper (99.9% Cu) rod 500 μm in diameter and 15 mm in height, which machined was shown in Fig.2, and Commercial grade EDM oil (specific gravity = 0.763, freezing point = 94° C) was used as a dielectric fluid.

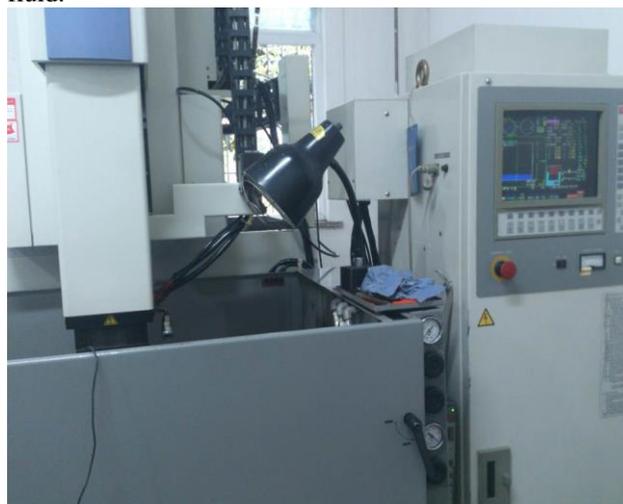


FIGURE 1 Experimental set-up used for experimentation

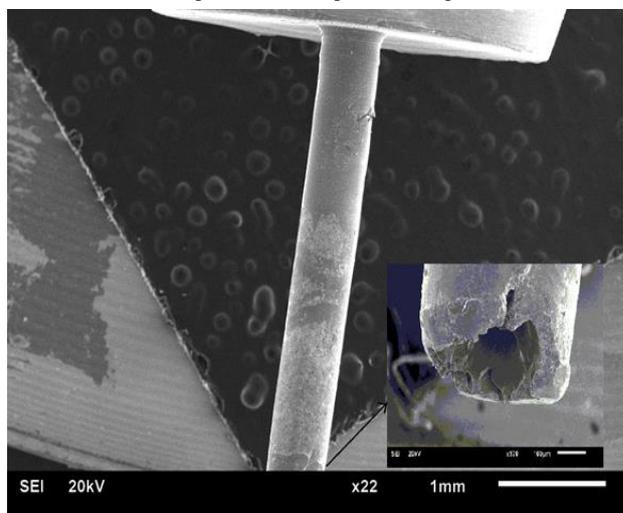


FIGURE 2 Copper electrode

2.3 EXPERIMENTAL PARAMETERS AND DESIGN

The white layer thickness produced by the EDM is mainly affected by the process parameters like discharge current, pulse on time and pulse off time. The proper selection of the process parameters which include the lower discharge current and longed pulse on time can cause a less recast layer thickness. In this study, experiments were planned using face-centred design with three variables that is based on a response surface methodology. The design of machining parameters and their levels for the CCD used was shown in Table 1. The pulse on time (t_i), pulse off time (t_s) and discharge current (I_p) were selected as the input parameters for the EDM process.

TABLE 1 Design scheme of machining parameters and their levels

| Parameters | Unit | Symbol | Levels | -1 | 0 | 1 |
|-----------------------------|---------|--------|--------|----|-----|---|
| Pulse on time(t_i) | μs | X_1 | 32 | 64 | 96 | |
| Pulse off time(t_s) | μs | X_2 | 64 | 96 | 128 | |
| Discharge current (I_p) | A | X_3 | 3 | 6 | 9 | |

2.4 RESPONSE VARIABLES EVALUATION

The white layer thickness after the EDM operation was observed using scanning electron microscope (JSM-7500F) with high magnification. The average thickness is measured by the image process software (Image-Pro version 6.0). The average white layer thickness is calculated at $34.20 \mu m$ in Fig.3.

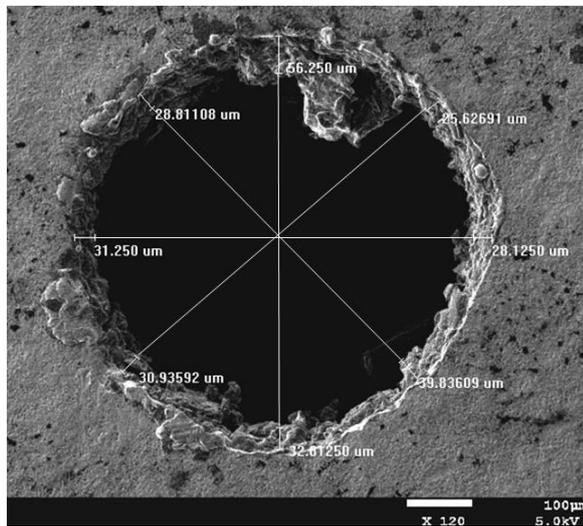


FIGURE 3 Micro-holes corresponding to $I = 9A$, $t_s = 128\mu s$ and $t_i = 96\mu s$

3 Analysis method

3.1 EXPERIMENTAL DESIGN WITH RSM

The response surface method is by constructing a clear form of implicit polynomials to approximate expression

function, which use a limited test by regression analysis to fit the analytical expression to replace the real response surface. Response surface method is an interaction of mathematical and statistical techniques for modelling and analysis of machining parameters in the EDM process which contains the discharge current, pulse on time and pulse off time in order to obtain the relationship to the WLT. In this study, the central composite design (CCD) is used to finish the experimental design.

In general, the response of the system and design factors (x_1, x_2, \dots, x_n) can be represented as following:

$$y = f(x_1, x_2, \dots, x_n) + \varepsilon, \tag{1}$$

where y is the response of the system, f is the response function (or response surface), x_1, x_2, \dots, x_n are the independent input variables and ε is the fitting error. In present, most of all use the quadratic model to demonstrate the second-order effect of each variable and the two-way to find the interaction between combinations of these design factors. The quadratic model of y can be written as follows:

$$y = \alpha_0 + \sum_{i=1}^n \alpha_i x_i + \sum_{i=1}^n \alpha_{ii} x_i^2 + \sum_{i < j} \alpha_{ij} x_i x_j + \varepsilon, \tag{2}$$

where α_0 is constant, α_i, α_{ii} and α_{ij} are the coefficients of linear, quadratic and cross product terms, respectively. For three variables ($n=3$), the experimental runs number is 20, which consists 2^3 factor points, 6 axial points and six centre points. Table 2 shows that the central composite design composes three input variables; X_1 (Pulse on time), X_2 (Pulse off time) and X_3 (Discharge current).

TABLE 2 Design layout and experiment results

| Run | Coded factors | | | Actual factors | | | |
|-----|---------------|-------|-------|----------------|-------|-------|----------------|
| | X_1 | X_2 | X_3 | t_i | t_s | I_p | WLT_{actual} |
| 1 | 0 | 0 | 0 | 64 | 96 | 6 | 14.64 |
| 2 | -1 | 1 | 1 | 32 | 128 | 9 | 28.67 |
| 3 | 1 | -1 | 1 | 96 | 64 | 9 | 31.24 |
| 4 | 1 | 0 | 0 | 96 | 96 | 6 | 16.24 |
| 5 | 0 | 0 | -1 | 64 | 96 | 3 | 13.04 |
| 6 | 0 | -1 | 0 | 64 | 64 | 6 | 14.33 |
| 7 | 1 | -1 | -1 | 96 | 64 | 3 | 13.22 |
| 8 | 1 | 1 | -1 | 96 | 128 | 3 | 15.93 |
| 9 | 0 | 0 | 0 | 64 | 96 | 6 | 14.64 |
| 10 | 0 | 0 | 0 | 64 | 96 | 6 | 14.64 |
| 11 | 0 | 0 | 1 | 64 | 96 | 9 | 29.90 |
| 12 | 0 | 1 | 0 | 64 | 128 | 6 | 20.01 |
| 13 | 0 | 0 | 0 | 64 | 96 | 6 | 14.64 |
| 14 | 0 | 0 | 0 | 64 | 96 | 6 | 14.64 |
| 15 | -1 | -1 | -1 | 32 | 64 | 3 | 11.16 |
| 16 | 0 | 0 | 0 | 64 | 96 | 6 | 14.64 |
| 17 | -1 | 1 | -1 | 32 | 128 | 3 | 13.78 |
| 18 | 1 | 1 | 1 | 96 | 128 | 9 | 34.20 |
| 19 | -1 | -1 | 1 | 32 | 64 | 9 | 28.40 |
| 20 | -1 | 0 | 0 | 32 | 96 | 6 | 13.96 |

3.2 DATA ANALYSIS USING RSM

The design expert software (version 7.0.0) is used to design and analysis the process parameters of the response equation, and subsequent analysis of variance (ANOVA) was assessed. In this study, the analysis of variance (ANOVA) is utilized to summary the above tests performed and analyse the results of the experimental runs. As per this technique, the response variable WLT was evaluated by the F-test of ANOVA shown in Table 3, respectively. The model should be considered to be

significant when the p-values were less than 0.05 and 0.001 when using 5% and 1% significance levels. In the Table 3, the p-values are less than 0.05 which indicate that the model for WLT is significant. In the same way, the effect of the discharge current, pulse on time and pulse off time were significant which can be seen in the Table 3. It can be seen that the effects of X_1, X_2, X_3, X_2^2 and X_3^2 were statistically significant. In the model WLT, the discharge current (X_1) played the important role in the machining process.

TABLE 3 Analysis of variance for WLT (RLT)

| Source | Sum of squares | Degrees of freedom | Mean square | f-value | Prob.>F | |
|-------------------------------|----------------|--------------------|-------------|---------|---------|-------------|
| Model | 1006.58 | 9 | 111.84 | 129.10 | <0.0001 | significant |
| X ₁ | 22.08 | 1 | 22.08 | 25.49 | 0.0005 | |
| X ₂ | 20.28 | 1 | 20.28 | 23.41 | 0.0007 | |
| X ₃ | 727.27 | 1 | 727.27 | 839.50 | <0.0001 | |
| X ₁ X ₂ | 0.97 | 1 | 0.97 | 1.12 | 0.3158 | |
| X ₁ X ₃ | 2.16 | 1 | 2.16 | 2.50 | 0.1451 | |
| X ₂ X ₃ | 0.55 | 1 | 0.55 | 0.64 | 0.4436 | |
| X ₁ ² | 0.46 | 1 | 0.46 | 0.53 | 0.4847 | |
| X ₂ ² | 7.60 | 1 | 7.60 | 8.78 | 0.0142 | |
| X ₃ ² | 97.77 | 1 | 97.77 | 112.86 | <0.0001 | |
| Residual | 8.66 | 10 | 0.87 | | | |
| Lack of fit | 8.66 | 5 | 1.73 | | | |
| Pure Error | 0.000 | 5 | 0.000 | | | |
| Correlation total | 1015.25 | 19 | | | | |
| R ² =0.9915 | | | | | | |

3.3 LEAST SQUARES SUPPORT VECTOR MACHINES (LS-SVM)

As a new learning machine, Support Vector Machine based on statistical learning theory proposed by Vapnik [12] is known as an excellent tool for the classifying regression problems of good generalization. In the following, the learning theory has been developed by many researchers and it has various types and Least square support vector machine (LS-SVM) is widely used in the pattern recognition and nonlinear regression. In this paper, we briefly introduce the principle of LS-SVM:

For a given training set of S data points $\{x_k, y_k\}_{k=1}^l$, $x_k \in R^m$, $y_k \in R$, x_k is the input data and y_k is the output data. A nonlinear function φ is utilized to map the input data x to high dimensional feature space G and linear approximation in this space. By statistical theory, this function can be written as follow:

$$f(x) = w^T \varphi(x) + b \tag{3}$$

In Equation (3) where w is weight vector and b is deviation value.

LS-SVM utilizes the quadratic loss function to transform the inequality constraints to equality constraints. Then the following optimization problem is formulated:

$$\min J(w, e) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_k^n e_i^2, \tag{4}$$

$$\text{s.t. } y_i = w^T \varphi(x_i) + b + e_i \quad i = 1, 2, \dots, n, \tag{5}$$

where the regularization factor is γ and e_i is the difference between the desired and the actual output.

In order to solve this constrained optimization, a Lagrangian is constructed:

$$L(w, b, e, a) = J(w, e) - \sum_{i=1}^p (w^T \varphi(x_i) + b + e_i - y_i), \tag{6}$$

where α_i is Lagrangian multiplier.

According to the optimization theory, the conditions are given by:

$$\begin{cases} \frac{\partial L}{\partial w} = 0, \rightarrow w = \sum_{i=1}^p \alpha_i \varphi(x_i) \\ \frac{\partial L}{\partial b} = 0, \rightarrow \sum_{i=1}^p \alpha_i = 0 \\ \frac{\partial L}{\partial e_i} = 0, \rightarrow \alpha_i = \gamma e_i \\ \frac{\partial L}{\partial \alpha_i} = 0 \rightarrow w^T \varphi(x_i) + b + e_i - y_i = 0 \end{cases}, \tag{7}$$

From Equation (7), the following linear equations can be obtained after elimination of the variables w and e :

$$\begin{bmatrix} 0 & \vec{1}^T \\ \vec{1} & \Omega + \gamma^{-1}I \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{8}$$

where $y = [y_1, \dots, y_n]^T$, $\vec{1} = [1, \dots, 1]^T$, $\alpha = [\alpha_1, \dots, \alpha_p]^T$, $\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$, $i, j = 1, 2, \dots, p$.

The resulting LS-SVM model can be evaluated as follows:

$$f(x) = \sum_{i=1}^p \alpha_i K(x_i, x_j) + b, \tag{9}$$

where b and α_i are the solutions to Equation (8) and $K(x_i, x_j)$ is the kernel function which meet Mercer condition. In LS-SVM, the kernel function is different and has many choices. In this work, the radial basis function (RBF) is selected as the kernel function with its stronger learning ability.

RBF kernel functions as follows:

$$K_{RBF}(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right), \tag{10}$$

where σ^2 is the kernel function parameter.

3.3.1 Data pre-processing

Data pre-processing is the method of transferring the original sequence to a comparable sequence which method is applied to cancel the difference of the orders of magnitude. After data pre-processing, the data is adjusted to between a range of 0 and 1. In general, the data normalizations include two kinds. In this study, the origin data is normalized as in the Equation (11).

The larger-the-better

$$x_k = \frac{x_k - x_{\min}}{x_{\max} - x_{\min}}, \tag{11}$$

The mean variance method

$$x_k = \frac{x_k - x_{\text{mean}}}{x_{\text{var}}}. \tag{12}$$

3.3.2 The LSSVM regression model establishment and test

In order to establish the LS-SVM regression model, the Matlab software (version R2012b) is used to estimate the LS-SVM model. Hence, the mathematical modelling of EDM process is needed. Many empirical, statistical and regression techniques have been used in literature [13-16]. The data pre-processing and normalization are very important for training and testing. The 15 groups are randomly selected as the training data and used to establish the training model and the remaining 5 groups

are adopted as the testing samples. When the regression model is applied to model the EDM machining process, the kernel function should be selected firstly which can be suitable for the nonlinear and complex EDM process. In this paper, the Gaussian function kernel is chosen as the kernel function which has better performance comparing with the other linear kernel. The Gaussian function is expressed in Eq.8 which has less hypermeter that influences the complexity than the polynomial kernel. The predict data after the LS-SVM model is reversed to origin data. In order to make the LS-SVM has great performance; the best set of hyperparameters such as γ and σ^2 are found by using the grind search and leave-one-out cross-validation method [17, 18] where γ is the regularization parameter and σ^2 is the kernel function parameter. The mean squared error is adopted as the model error by using leave-one-out method. The evaluation function correlation coefficient R^2 is utilized to discuss the predict accuracy of the LS-SVM model.

$$MSE = \frac{\sum_{i=1}^N e_i^2}{N}, \tag{13}$$

$$R^2 = 1 - \sum_{i=1}^N \left(\frac{Y - X}{X}\right)^2, \tag{14}$$

where the N is the experimental number, Y is the testing data and X is the predict result of the model.

After repeated tests, the regularization parameter γ and the kernel function parameter σ^2 are chosen as 43201.2 and 120.764 for the LS-SVM model by using the proposed parameter chosen method. Then the LS-SVM model is trained by using training data and the remaining 5 groups is selected as the testing data. The experimental white layer thickness is compared with the LS-SVM model predict result in Figure 4. That $MSE = 4.55653$ and $R^2 = 0.997029$. The comparison results between the original and predict values are shown in Table 4. From the Table 4, the trained LS-SVM model has small output errors and can correctly reflect the causality between inputs and outputs. Comparing with the RSM method, the R^2 can reflect the LS-SVM method having great performance.

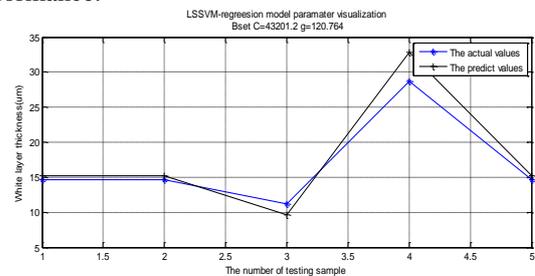


FIGURE 4 The predicted white layer thickness versus the actual white layer thickness

TABLE 4 LSSVM model test result using the remaining five groups of parameter combinations

| Initial no | t_i | t_s | I_p | WLT_{actual} | $WLT_{predicted}$ | Residual |
|------------|-------|-------|-------|----------------|-------------------|----------|
| 1 | 64 | 96 | 6 | 14.64 | 15.23 | -0.59 |
| 9 | 64 | 96 | 6 | 14.64 | 15.23 | -0.59 |
| 15 | 32 | 64 | 3 | 11.16 | 9.65 | 1.51 |
| 17 | 32 | 128 | 9 | 28.67 | 32.84 | -4.17 |
| 10 | 64 | 96 | 6 | 14.64 | 15.23 | -0.59 |

4 Result and discussion

4.1 OBSERVATION OF THE MACHINED MICRO-HOLE

Figure 5 and 6 show the micrographs in the 120 magnification that can be observed the micro-hole finish of Ti-6Al-4V after EDM processing. Figure 5 displays the micro-hole finish under the EDM with a discharge current of 9A, a pulse on time of 32 μs and a pulse off time of 128 μs . Similarly, Fig. 6 displays the micro-hole finish under the EDM with a discharge current of 9A, a pulse on time of 64 μs and a pulse off time of 96 μs .

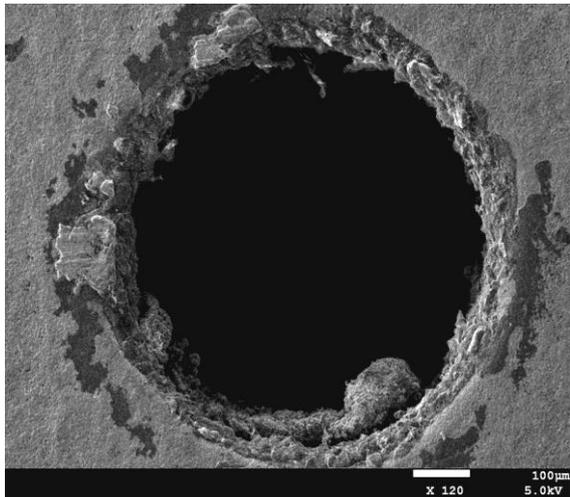


FIGURE 5 Micro-holes corresponding to $I = 9A$, $t_s = 128\mu s$ and $t_i = 32\mu s$

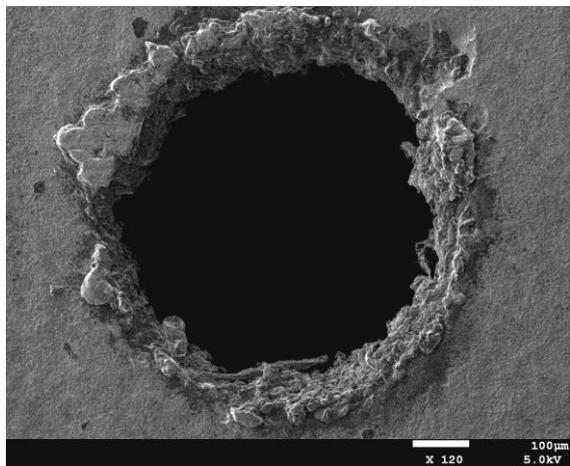


FIGURE 6 Micro-holes corresponding to $I = 9A$, $t_s = 96\mu s$ and $t_i = 64\mu s$

4.2 COMPARING WITH SINGLE SUPPORT VECTOR MACHINE

According to the data in the Table 2, the SVM is applied to model the nonlinear EDM processing and the Gaussian function kernel is selected as kernel function. K-fold Cross Validation (K-CV) is employed to select the best set of hyperparameters such as C and g . In the K-CV method, the training set has been divided to groups. Each subset is selected as the validation set and the remaining set as the test set. Basic pairs of (C and g) are tried and the best coefficient and are chosen as 2.8284 and 0.0625 after repeated tests. The selection result of the SVR parameter (3D view) is shown in Figure 7. For example, the growing sequences of adjustment parameters as:

$$C = e^{-4}, e^{-2}, \dots, e^4 \quad g = e^{-4}, e^{-2}, \dots, e^4, \tag{15}$$

The MSE is selected as the evaluation index as in the Equation (13) and the evaluation function correlation coefficient R^2 as in the Equation (14).

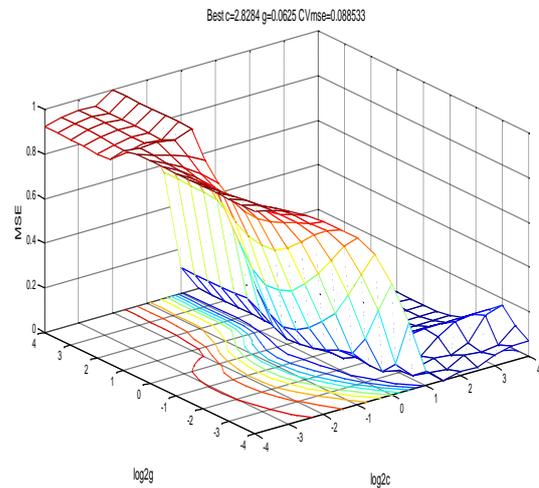


FIGURE 7 The result of SVR parameters (3D view)

When the training model of WLT is finished, the remaining random 5 groups of processed data is applied to test the performance of the model. The comparison result between the original and test values are shown in Figure 8, in which the MSE is 0.14742 the R^2 value is 0.89136.

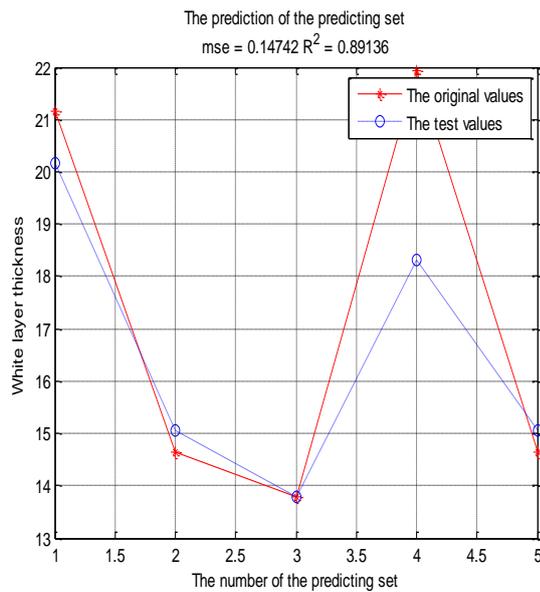


FIGURE 8 The comparison of the original and test values

4 Conclusions

In this study the white layer thickness in the die-sinking EDM process on Ti-6Al-4V was predicted by response surface methodology coupled with least squares support vector machine for the machining parameters. According to the implementation results obtained in the illustrative example, the conclusions are as follows:

References

- [1] Hascalik A, Caydas U 2007 A comparative study of surface integrity of Ti-6Al-4V alloy machined by EDM and AECG *Journal of Materials Processing Technology* **190** 173–80
- [2] Ezugwu E O, Wang Z M 1997 Titanium alloys and their machinability-A review *Journal of Materials Processing Technology* **68** 262–74
- [3] Kao J Y, Tsao C C, Wang S S, Hsu C Y 2010 Optimization of the EDM parameters on Ti-6Al-4V with multiple quality characteristics. *International Journal of Advanced Manufacturing Technology* **47** 395–402
- [4] Haron C H C, Jawaid A 2005 The effect of machining on surface integrity of titanium alloy Ti-6Al-4V *Journal of Materials Processing Technology* **166** 188–92
- [5] Ramasawmy H, Blunt L, Rajurkar K P 2005 Investigation of the relationship between the white layer thickness and 3D surface texture parameters in the die sinking EDM process *Precision Engineering* **29** 479-90
- [6] Ahmet Hascalik, Ulas Caydas 2007 Electrical discharge machining of titanium alloy (Ti-6Al-4V) *Applied Surface Science* **253** 9007–16
- [7] Ulas Caydas, Ahmet Hascalik 2008 Modeling and analysis of electrode wear and white layer thickness in die-sinking EDM process through response surface methodology *Int J Manuf Technol* **38** 1148-56
- [8] Jabbaripour B, Sadeghia M H, Faridvanda Sh, Shabgard M R 2012 Investigating the effects of EDM parameters on surface integrity, MRR and TWR in machining of Ti-6Al-4V *Machining Science and Technology* **16** 419-44
- [9] Lingxuan Zhang, Zhenyuan Jia, Fuji Wang, Wei Liu 2010 A hybrid model using supporting vector machine and multi-objective genetic algorithm for processing parameters optimization in micro-EDM *Int J Adv Manuf Technol* **51** 575-86
- [10] Chong-Jyh Tzeng, Rui-Yang Chen 2013 Optimization of electric discharge machining process using the response surface methodology and genetic algorithm approach *International journal of precision engineering and manufacturing* **14** 709-17
- [11] Somashekhar K P, Ramachandran N, Jose Mathew 2010 Optimization of material removal rate in micro-EDM using artificial neural network and genetic algorithms *Materials and Manufacturing Processes* **25** 467-75
- [12] Vapnik V 1998 *Statistical learning theory* Wiley Interscience, New York.
- [13] Tosun N 2003 The effect of cutting parameters on performance of WEDM *K.S.M.E. Int. J.* **17**(6) 816-24
- [14] Rebelo J C, Dias A M, Mesquita R, Vassalo P, Santos M 2000 An experimental study on electro-discharge machining and polishing of high strength copper-beryllium alloys *J. Mat. Process. Technol.* **103**(2000) 389-97
- [15] Petropoulos G, Vaxevanidis N M, Pandazaras C 2004 Modeling of surface finish in electro-discharge machining based on statistical multi-parameter analysis *J. Mat. Process. Technol.* 155-156 1247-51
- [16] Zorepour H, Tehrani A F, Karimi D, Amini S 2007 Statistical analysis on electrode wear in EDM of tool steel DIN 1.2714 used in forging dies *J. Mat. Process. Technol.* **187-188**(2007) 711-4
- [17] Duric P M 1990 Model selection by cross-validation *IEEE ISCAS* **4** 2760–3
- [18] Lela B, Bajić D, Jozić S 2009 Regression analysis, support vector machines, and Bayesian neural network approaches to modelling surface roughness in face milling *Int J Adv Manuf Technol* **42** 1082–8

1 Response surface methodology is utilized to design the economical experiment and ANOVA indicated that the EDM parameter of discharge current is the most significant factors for white layer thickness. According to the result, the discharge energy increasing lead to the white layer thickness increased. Accompany with the increasing discharge current, the removed material from the machined surface is more when the pulse on time is a constant. The value of the pulse on time increase leading to the WLT increasing because of more discharge energy is transformed to surface of workpiece during a single pulse. The pulse off time is significant and the reason will be that the flushing away by dielectric fluid the less volume of molten particles are re-solidified that leads to induce of WLT when the value of the pulse off time increasing.

2 The LS-SVM is found to give reasonably good prediction accuracy for WLT in the die-sinking EDM machining on the Ti-6Al-4V with the pure cylindrical copper electrode. It gives better prediction results in the experimental runs than just using the SVM method. The LS-SVM model, the predict accuracy is better.

3 According to the LS-SVM model, the result can express the higher discharge current can lead to the higher white layer thickness .However, the longer pulse off time can lead to the lower white layer thickness.

| Authors | |
|---|--|
|  | <p>Jun Li</p> <p>University studies: M.Sc. in Engineering (2008) from southwest University of science and technology. Now he is studying at Sichuan University with a major of Mechatronic Engineering Program for PhD (2014).</p> <p>Scientific interests: on the Electrical discharge machining, Micro Electrical discharge machining and Artificial intelligence used in the mechanical engineering.</p> |
|  | <p>Xiaoyu Liu, born on June 4, 1987, Inner Mongolia, China</p> <p>University studies: Sichuan University, School of Manufacturing Science and Engineering (China), in 2010. She is studying for PHD degree in the same department of Sichuan University.</p> <p>Scientific interests: manufacturing and measurement technology of micro manufacturing.</p> |
|  | <p>Shiping Zhao</p> <p>Current position, grades: full professor of manufacturing science and engineering Department, Sichuan University</p> <p>University studies: PhD in engineering precision instruments and machinery (1991) from Chongqing University.</p> <p>Scientific interests: robot technology and nondestructive test.</p> |

Construction of a computer simulation platform for optical experiments

Hao Wu*, Dewen Seng, Xujian Fang

School of Software Engineering, Hangzhou Dianzi University, 310018, Hangzhou, China

Received 10 July 2014, www.tsi.lv

Abstract

With the rapid development of computer technology, computer-assisted instruction for complex and vulnerable optical experiments has become possible. Computer simulation technique has become an important branch in computer application and a new means in science research and engineering design. People have done a lot of research on optical experiment simulation. But there are still many defects, such as no friendly graphical user interface and the parameters cannot be freely adjusted. We design and develop a well extensible and portable simulation platform for optical experiments including basic optics experiments, information optics experiments and laser experiments and realize flexibility in setting the experimental parameters. Young's double-slit interference experiment, Fraunhofer diffraction experiment and grating diffraction experiment are conducted to show the effectiveness, efficiency and correctness of our simulation platform. The abstract and difficult optical concepts and rules are vividly manifested through the simulation experiments and become easier to understand for the students. The simulation platform will break through the limitation of teaching space, experimental equipment and various other factors and enable students to preview experiment, understand experiment, complete experiment and review experiment much better.

Keywords: optical experiment, simulation platform, interference and diffraction

1 Introduction

The optical experiment instruments are generally sophisticated and expensive. With the expansion of the scale of the university's enrolment students, laboratory equipment is difficult to meet the requirements. Because of the relative shortage of funds, schools are difficult to purchase a large number of laboratory equipment. The students usually carry out experiments in small group. At the same time, there is a certain limit in experimental class, so there are not conducive to students' understanding of the experiment in time and space. Optical experiments have the features of complex operation, difficult understanding of phenomenon, and stringent requirements of experimental data. These factors tend to limit the students to set up the experimental parameters and repeatedly adjust the instrument to observe different phenomena in order to achieve full understanding of the purpose of the experiments [1-4]. The teachers and the students can be freed from the limitations of the experimental apparatus and experimental sites by the use of computer simulation of optical experiments, which will reduce the experimental loss, make accurate simulation of the experiments and help the students to intuitively and easily observe the experiment phenomenon.

In the experimental teaching, the theoretical knowledge involved in the optical experiments is more esoteric than mechanics and electrics. For example, in the experimental teaching of ray diffraction and light

interference, the students cannot see the experimental phenomena, so they will think that the experimental principles are arcane. Some teachers draw the light intensity distribution curve on the blackboard and hand-painted the interference fringe pattern of light and dark. They hope the students will get a better understanding of graphics. However, the teachers' hand drawings are unmatched real phenomena, so the students find it boring, abstract and difficult to be understood, which will even result in the theory and experiment out of touch.

Teachers are also hard-pressed to express the experimental phenomena to the students; therefore, they think that this teaching method is of many deficiencies. In order to enable the students to witness the experimental phenomenon, some teachers personally carry out the experiments to help the students directly to observe the experimental results. But this teaching method is also deficient. It will not only waste the time and will deprive the students to explore their own experiments and the opportunity to exercise experimental ability. This experimental teaching become a repeated mechanical labour of the students from the teachers' experimental procedure and is not conducive to improve the students' ability to experiment.

With the continuous development of computer technology, computer-aided teaching is introduced into the modern education system and becomes an interesting teaching subject from theory to practice [5-14]. Computers have powerful computing, graphics and image processing ability. The use of computer is very useful in

* *Corresponding author* e-mail: hduse@sina.com

the simulation of difficult experiments in the actual operation. The use of computer simulation in the experiments is to break the traditional physical experiments. It will enable the students to better understanding and completion of the experiments and is a very good teaching method for the optical experiments.

Therefore, the use of computer simulation is no doubt an effective experimental teaching adjunct. However, the traditional computer software involves a lot of repetitive and complex programming work. There is a certain degree of difficulty to master. Matlab has powerful symbolic computation and numerical calculation function. The language used in Matlab is intuitive and easy to understand. It will automatically determine the required number of points, the location of the axes, and automatically draw the graphics. The programming is simple, easy to operate and easy to grasp.

2 Computer simulation of optical experiments

In the field of engineering design, through the study of object model, a computer program system run and get the results to find out the optimal solution. The solution will be implemented physically. This is the computer simulation science. In the growing popularity of the computer, the computer simulation technology as a means of virtual experiment has become an important branch of computer applications. It is a new means of understanding the objective world.

The computer simulation is achieved by the running of the simulation programs. Simulation programs firstly set certain parameter values of model to describe the system characteristics. Some of the variables in the model are to be changed in the specified range [8]. The specific circumstances and results of the system movements can be obtained by calculating the variables. The simulation program has a variety of functions.

The computer can display the entire process of the system movements and the various phenomena and status arising in this process. In order to facilitate the observation, the process can be effectively controlled. By changing the external conditions of the system and setting of different parameters of the system, a variety of different characteristics of the system movement can be studied and found. Since computer is of high-speed operation, the simulation system can quickly respond accordingly when the entering experimental conditions and data changed [9]. Using the random number given by the computer system, probabilistic characteristics and the corresponding state of phenomena can be given by certain calculation of the values.

Therefore, the computer simulation has good controllability. The parameters can be adjusted wantonly. The simulation is non-destructive. It will not cause damage to the device or accidents because of unreasonable designs. It is reproducibility, that is, it can exclude a variety of random factor effects, such as temperature, humidity, etc. Also, it is easy to observe the

phenomenon which are difficult to be observed in practical experiments.

The computer simulation of optical experiments includes theoretical experiment and application experiment [15-18]. Theoretical experiment is to move the optical experiments into the computer. The main purpose is to increase students' understanding of the basic theory and to grasp and predict the experimental results. This theory simulation method is very important in scientific research. With the development of computer technology and photovoltaic technology, many of the traditional optical experiments have gradually broken away from the experimental platform and get into the computer simulation environment, such as digital holography [16].

The hologram recording and reproducing can be all digitalized, which makes it easier to measure and analyse quantitatively. In some areas, it has entered the practical use. In addition, the combination of optics and computer is the development trend of optics and optical experiments. Application experiment is to fully develop students' potential and initiative. It starts from a practical engineering problem to discuss the optical applications, so that students will cogitate gradually from the ideal assumptions to the interpretation, analysis and exploration of solutions of practical engineering problems, which will inspire and nurture the students' learning and development capabilities and creativity to solve practical engineering problems. The students can not only master the traditional knowledge but also learn new technology and broaden their horizons. By the optical experiment teaching with computer simulation, the students will really integrate theory with practice and improve their practical ability and innovative ability.

3 The framework of the optical simulation platform

The optical simulation experiment platform includes basic optical experiments, information optics experiment, and the laser experiments, as shown in Figure 1. To improve the scalability of the platform, it used modular design. Each specific experiment is taken as a module for system maintenance, so it is very convenient for adding new experimental simulation.

To allow users to better understand and think about the experiments, the interface in each experiment was added the purpose of the experiment, the experimental principles, experimental steps, thinking questions and other information. Students can enter the corresponding interface for the learning of basic theory. Every window of each experiment was divided into three areas: parameter settings area, the command control area, the simulation results area. Experimental parameters of every experiment are adjustable. The experimental parameters can be set flexibly, and the results can be visualized. The users can observe the different phenomenon under a variety of experimental parameters.

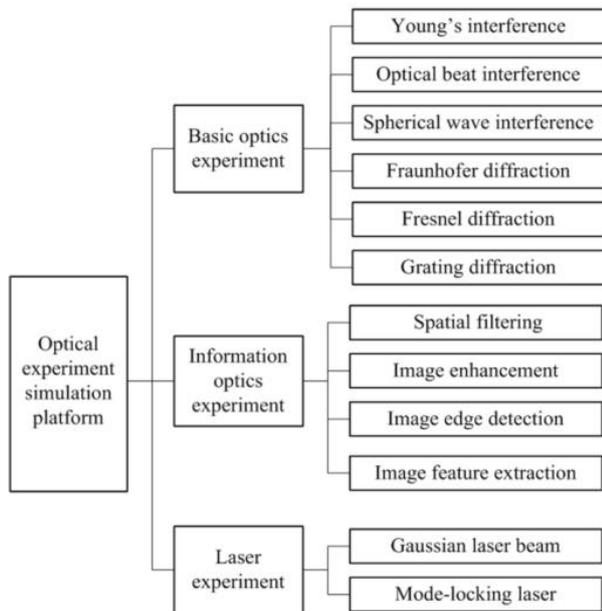


FIGURE 1 Simulation platform of optical experiment

The simulation platform constructed by the above method has friendly interface, simple operation, high portability, scalability and good security. This method makes a breakthrough on teaching space, equipment and various other factors' limitation. The difficult abstract concepts, rules are manifested through vivid simulation experiments and made easier to understand for the students.

MATLAB is a set of computer simulation software which has been widely used in optical experiment simulation. It allows the users to program in mathematical form and has powerful graphics capabilities. It has many graphics library functions, which can be used to draw all kinds of complex two dimensional and multi-dimensional graphics. Users can create a graphical user interface to achieve communication between users and computers. However, the powerful capabilities of MATLAB need the installation of MATLAB system in computer, which gives a big inconvenience to engineering calculations. At the same time, the MATLAB program can be seen directly, which is not conducive to software security and maintenance.

With the development of computer technology, people use MATLAB simulation system for the implementation of optical experiments and conduct a lot of research, but still there are many defects such as no graphical user interface, hence the user must master the language MATLAB and then can adjust the parameters in the experiments program. It is lack of laser experiment simulation. The program cannot run from the MATLAB environment. In response to these shortcomings, we will conduct a comprehensive simulation of the optical experiments and produce a simulation experiment platform including basic optical simulation experiments, information optics simulation experiment and the laser simulation experiments. The platform is of great

scalability, portability of optical experiments. It can really be applied to optical theory and experimental teaching to enhance students' interest in learning. The students can better understand, complete and grasp the experiments.

4 Simulation of Young's interference experiments

In the normal Young's double-slit interference experiment, the interference fringes will change insignificantly by changing the experimental parameter. It is difficult to observe and demonstrate all the characteristics of the experiment. In addition, the experiment requires specific equipment and places which will bring the teaching and research a lot of inconvenience. The experiment can be easily done with our simulation platform based on the theory of Young's double-slit interference. Emulator can display the interference pattern and the light intensity distribution curve whenever the monochromatic or non-monochromatic lights incidence, and will calculate the corresponding fringes spacing and contrast of specific points. We had designed friendly graphical user interface. Users can set different experimental parameters and qualitatively and quantitatively analyse the experimental results of various parameters.

Set the double-slit distance of the experiment apparatus is d , and the distance between the screen and the slit is D . The origin point of the coordinates on the screen is O and is symmetrical with the two slits. The two light source slits meet the interference condition of same vibration direction, same frequency and constant phase difference. When two lights meet in space, interference phenomenon will appear [18]. Bright and dark interference fringes will appear on the screen. Let $OP = y$, then we can know from the geometric relation that the distance between two interference sources and any point of the screen P is:

$$r_1 = \sqrt{D^2 + \left(y - \frac{d}{2}\right)^2}$$

$$r_2 = \sqrt{D^2 + \left(y + \frac{d}{2}\right)^2}$$
(1)

The optical path difference of the two interference lights is $\Delta r = r_2 - r_1$; the phase difference is $\Delta\varphi = \frac{2\pi\Delta r}{\lambda}$. λ is the wavelength of light. Set the amplitudes at point P on the screen of the two light waves are E_1 and E_2 respectively, and the light intensities are I_1 and I_2 . The amplitude of the superimposed waves is:

$$E = \sqrt{E_1^2 + E_2^2 + 2E_1E_2\cos\Delta\varphi}$$
(2)

The light intensity is:

$$I = I_1 + I_2 + 2I_1I_2\cos\Delta\varphi$$
(3)

Let the amplitudes of the two light waves are equal when they meet at the point on the screen, then the light intensity of the point P is:

$$I = 4I_0 \cos^2\left(\frac{\Delta\phi}{2}\right). \tag{4}$$

Interference bright stripes will meet the condition:

$$\Delta\phi = \pm 2k\pi, k = 0,1,2,3, \dots \tag{5}$$

Interference dark stripes will meet the condition:

$$\Delta\phi = \pm(2k + 1)\pi, k = 0,1,2,3, \dots \tag{6}$$

Figure 2 is the interference pattern and the interference intensity distribution curve of monochromatic light and Figure 3 is that of non-monochromatic lights.

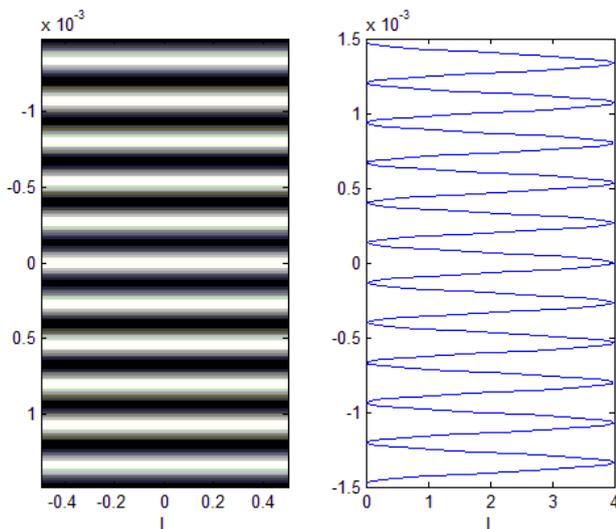


FIGURE 2 Young's double-slit interference pattern and the interference intensity distribution curve of monochromatic light

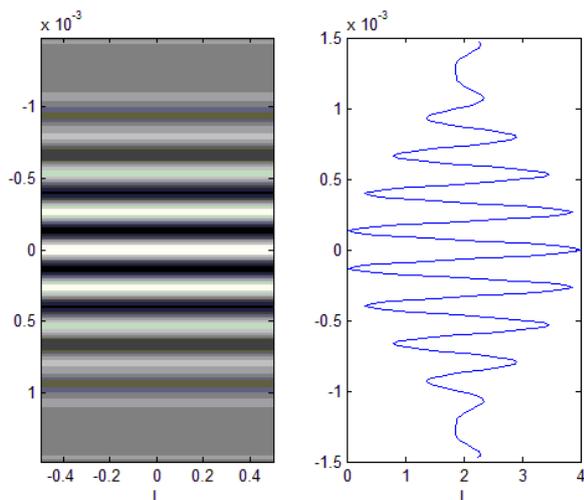


FIGURE 3 Young's double-slit interference pattern and the interference intensity distribution curve of non-monochromatic light

The results show that our simulation platform can draw directly two cases of monochromatic and non-monochromatic light conditions of Young's double-slit interference pattern and the light intensity distribution curve, and can simultaneously calculate the corresponding fringe distance of spacing and degree of contrast. The simulation results and the theoretical calculation are completely consistent. When the incidence light is monochromatic or non-monochromatic, the simulation results and the theoretical derivation are completely consistent. The drawings are delicate, realistic and distinct, which can help to observe and analyse the experiments. In addition, the users can visually analyse the interference results with different parameters. The simulation makes the whole physical process objective and offers a new effective supplementary means for optical theory analysis and experiment teaching.

5 Simulation of light diffraction

Diffraction is a fundamental phenomenon in the light wave propagation. Usually it is divided into two types: Fresnel diffraction and Fraunhofer diffraction. These two diffractions are also called the near-field and far-field diffraction. In optical diffraction experiments, the instruments should be with extremely high precision [19]. During the experiment, it is vulnerable to a number of factors, so the operation is very difficult. People have been conducted a lot of research on the Fraunhofer diffraction simulation experiments, but still there are many defects such as no flexible emulation interface to set parameters and the impact of the changes of wavelength and lens focal length on the diffraction results cannot be easily observed. The use of our simulation platform to do the optical experiments will allow users to intuitively and easily simulate and observe the experimental phenomena.

The Fraunhofer single slit diffraction intensity formula is

$$I = I_0 \left(\frac{\sin(u)}{u}\right)^2, \tag{7}$$

where $u = \frac{\pi a \sin\theta}{\lambda}$, and a is the slit width, θ is the diffraction angle, λ is the wavelength of the incident light.

The centres of the dark stripes meet the condition:

$$a \sin\theta = \pm 2k \frac{\lambda}{2}, k = 1,2,3, \dots \tag{8}$$

The approximate centres of the bright stripes meet the condition:

$$a \sin\theta = \pm(2k + 1) \frac{\lambda}{2}, k = 1,2,3, \dots \tag{9}$$

Set the single-slit width is a . We use the Huygens-Fresnel principle to calculate the diffraction light intensity distribution on the screen. Huygens postulated that every point on a primary wave front acts as a source of spherical secondary wavelets and the sum of these secondary waves determines the form of the wave at any subsequent time. Fresnel developed an equation using the Huygens wavelets together with the principle of superposition of waves, which models these diffraction effects quite well [19]. The wave front is divided equally into N parts, that is, N sub-bands. The amplitude E_0 of every sub-band on the point P of the screen can be regarded approximately equal. As for diffraction angle, the optical path difference of any two adjacent sub-band is $\Delta = \frac{a}{N} \sin\theta$, then the phase difference is:

$$\Delta\varphi = \frac{2\pi a}{\lambda N} \sin\theta. \tag{10}$$

The combination amplitude of the point P on the screen is equivalent to the overlay of N same direction, same frequency simple harmonic vibrations. So the amplitude at point P on the screen is:

$$E = E_0 \frac{\sin(\frac{N\Delta\varphi}{2})}{\sin(\frac{\Delta\varphi}{2})} = E_0 \frac{\sin(\frac{\pi a}{\lambda} \sin\theta)}{\sin(\frac{\pi a}{\lambda N} \sin\theta)}. \tag{11}$$

Because $I = E_0^2$, the light intensity is:

$$I = I_0 \frac{\sin^2(\frac{\pi a}{\lambda} \sin\theta)}{\sin^2(\frac{\pi a}{\lambda N} \sin\theta)}. \tag{12}$$

Figure 4 is the diffraction pattern and light intensity distribution curve with the wavelength close to the slit width, and Figure 5 is the diffraction pattern and light intensity distribution curve with the wavelength greatly less than the slit width.

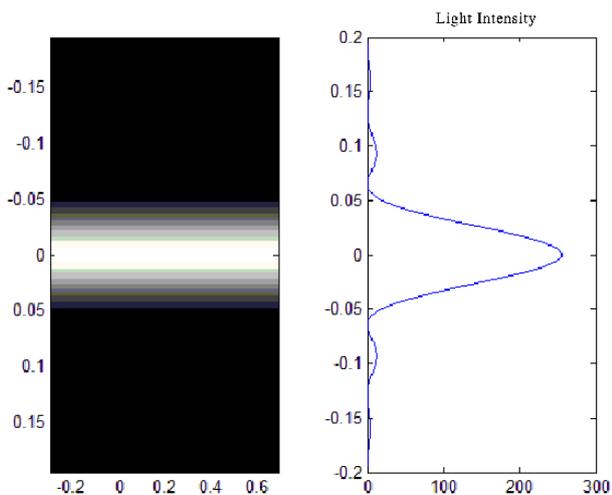


FIGURE 4 Diffraction pattern and light intensity distribution curve with the wavelength close to the slit width

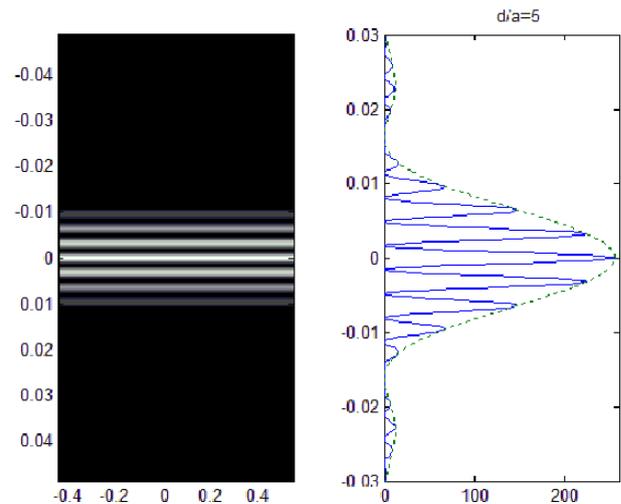


FIGURE 5 Diffraction pattern and light intensity distribution curve with the wavelength greatly less than the slit width

The Fraunhofer grating diffraction intensity formula is:

$$I = I_0 \left(\frac{\sin(u)}{u}\right)^2 \left(\frac{\sin(Nv)}{\sin(v)}\right)^2, \tag{13}$$

where N is the grating number, $u = \frac{\pi a \sin\theta}{\lambda}$, $v = \frac{\pi(a+b) \sin\theta}{\lambda}$, and a is the slit width of the transmittance part; b is the slit width of the opaque part; $d=a+b$ is called the grating constant.

Figure 6 is the grating diffraction pattern and light intensity distribution curve with different slit number and d/a value. The results show that our simulation platform can turn the light intensity distribution to luminance simulation, which will facilitate the computer simulation of relevant diffraction problems, provide multimedia understanding for the students and improve the teaching quality. The parameters can be changed to get different results for the experimental operation, which will save the time and energy for the students and teachers and improve learning efficiency and interest.

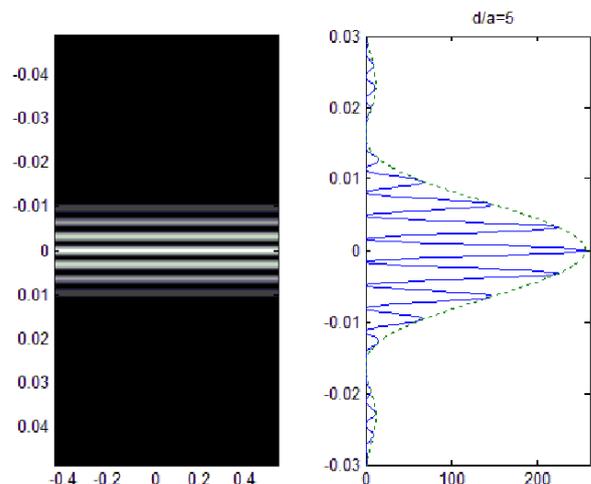


FIGURE 6 Grating diffraction pattern and light intensity distribution curve with 2 slits and $d/a=5$

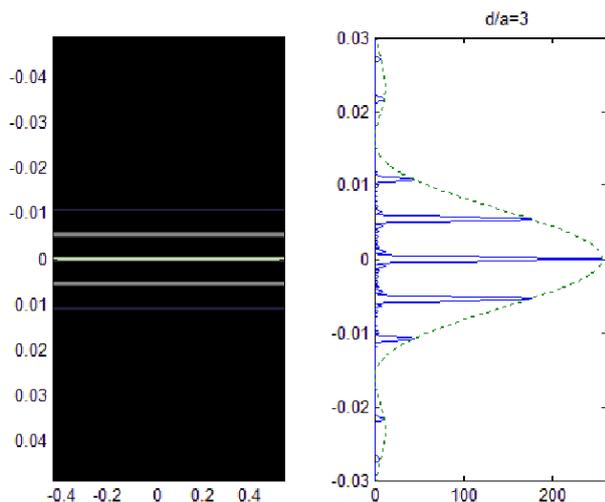


FIGURE 7 Grating diffraction pattern and light intensity distribution curve with 8 slits and $d/a=3$

6 Conclusions

Computer simulation technology is introduced into the optical experiments. According to the basic theory of optical information, we build a computer simulation platform for basic optics experiments, information optics

References

- [1] Lilly T C 2011 Simulated nonresonant pulsed laser manipulation of a nitrogen flow *Applied Physics B: Lasers and Optics* **104**(4) 961-8
- [2] Tan M, Rosenberg P, Yeo J S, etc. 2009 A high-speed optical multi-drop bus for computer interconnections *Applied Physics A: Materials Science and Processing* **95**(4) 945-53
- [3] Seng Dewen 2012 Application of computer in material science and engineering *Applied Mechanics and Materials* **189** 482-5
- [4] Devitt S 2010 Scalable quantum information processing and the optical topological quantum computer *Optics and Spectroscopy* **108**(2) 267-81
- [5] Seng Dewen, Liu Zhongxue 2008 Design and implementation of a 3D simulation system for geological and mining engineering *Journal of Liaoning Technical University* **27**(1) 9-12
- [6] Seng D 2010 Framework and Key Technologies for the Construction of a Virtual Mine *International Conference on Multimedia Technology*
- [7] Seng D, Chen W 2010 3D Visualization and Interaction Technologies Applied in the Modelling of Geological Bodies *2010 International Conference on Computer Application and System Modeling*
- [8] Tornari V, Tsiranidou E, Bernikola Ei 2012 Interference fringe-patterns association to defect-types in artwork conservation: An experiment and research validation review *Applied Physics A: Materials Science and Processing* **106**(2) 397-410
- [9] Seng Dewen 2012 Visualization of composite materials' microstructure with OpenGL *Applied Mechanics and Materials* **189** 478-81
- [10] Seng Dewen 2012 Simulation techniques in the research of structure and performance of nanosized materials *Applied Mechanics and Materials* **189** 457-60
- [11] Zhang J, Wang J Z, Yuan Z, etc. 2011 Computer-aided classification of optical images for diagnosis of osteoarthritis in the finger joints *Journal of X-Ray Science and Technology* **19**(4) 531-44
- [12] Korfiatis D P, Kosmatos A D, Thoma K A T, Vardaxoglou J C 2009 Computer modelling of ultrafast all-optical wavelength conversion in silicon nanophotonic waveguides *Microelectronic Engineering* **86**(4) 1134-7
- [13] Seng Dewen, Shu Yueqing 2013 Framework and construction contents of digital mine *Advances in Intelligent Systems and Computing* 393-400
- [14] Seng Dewen, Shu Yueqing 2013 Research on the application and status quo of digital mine and sensing mine *Advances in Intelligent Systems and Computing* 401-8
- [15] Zhang H, Xie J, Liu J, Wang Y 2009 Optical reconstruction of 3D images by use of pure-phase computer-generated holograms *Chinese Optics Letters* **7**(12) 1101-3
- [16] Shen Y, Hu P, Wang H 2011 The computational complexity of arithmetic based on ternary optical computer *Journal of Information and Computational Science* **8**(5) 850-7
- [17] Gultekin G K, Saranli A 2013 An FPGA based high performance optical flow hardware design for computer vision applications *Microprocessors and Microsystems* **37**(3) 270-86
- [18] Visser T D, Schoonover R W 2008 A cascade of singular field patterns in Young's interference experiment *Optics Communications* **281**(1) 1-6
- [19] Malinka A V, Zege E P 2009 Fraunhofer diffraction by arbitrary-shaped obstacles *Journal of the Optical Society of America A: Optics and Image Science, and Vision* **26**(8) 1763-7

experiments and laser experiments and realize friendly flexibility in setting the experimental parameters. The optics and computer technology are organically integrated in our simulation platform. The optical experiment lab can be moved into the computer. The simulation platform will facilitate the optical theoretical analysis, quantitative measurement and optical experiments. Thus, traditional demonstration experiments are turned into simulation experiments in the computer, which will improve the students' understanding of optical experiments and the learning initiative and enthusiasm. The case study shows the effectiveness, efficiency and correctness of our simulation platform. Meanwhile, it will bring the experiment time a lot of flexibility.

Acknowledgments

The authors wish to thank the National Natural Science Foundation of China for contract 61100194, the foundation of education department of Zhejiang province of China for contract Y201120520 and the research foundation of Hangzhou Dianzi University for contract KYS105612008 and YB1205, under which the present work was possible.

| Authors | |
|---|---|
|  | <p>Hao Wu, born in November, 1978, China</p> <p>Current position, grades: MS Degree, Lecturer University studies: Computer science; software engineering Scientific interest: Cloud Computing and Machine Learning Publications: 8 Experience: H. Wu received the MS degree in computer software from Hangzhou Dianzi University in 2005. He is now mainly engaged in Cloud Computing technology and Machine Learning technology.</p> |
|  | <p>Dewen Seng, born in February, 1977, China</p> <p>Current position, grades: Ph. D. Associate Professor University studies: Computer science; software engineering Scientific interest: Cloud Computing technology and Intelligent Transportation Systems Publications: 21 Experience: D. W. Seng received the Ph. D. degree from University of Science & Technology, Beijing, China in 2005. He is an Associate Professor and the Vice Director of Software Engineering Institute of Hangzhou Dianzi University. He is now mainly engaged in Cloud Computing technology and Intelligent Transportation Systems.</p> |
|  | <p>Xujian Fang, born in September, 1978, China</p> <p>Current position, grades: MS Degree, Lecturer University studies: Computer science; software engineering Scientific interest: Cloud Computing and Data Mining Publications: 6 Experience: X. J. Fang received the MS degree in computer software from Hangzhou Dianzi University in 2005. He is now mainly engaged in Cloud Computing technology and Data Mining technology.</p> |

| Authors' index | | | | | |
|------------------|----------|----------------------|----------|-----------------|----------|
| Bai Lin | 140 | Jiang Ming | 128 | Wang Xiumei | 41 |
| Cai Guoliang | 12 | Jiang Shengqin Jiang | 12 | Wang Yao | 41 |
| Cai Shuiming | 12 | Jun Deng | 7 | Wei Dong | 166 |
| Cao Xining | 356 | Li Dashe | 344 | Wei Jian | 109 |
| Cen Yong | 259 | Li Gongfa | 23 | Wei Jianliang | 236 |
| Chai Xiaodong | 356 | Li Juan | 197 | Wu Hao | 406 |
| Chen Chang | 368 | Li Jun | 398 | Wu Jinzhao | 36 |
| Chen Guohua | 322 | Li Qianmu | 133 | Wu Xian | 116 |
| Chen Guojin | 368, 373 | Li Xiaofeng | 122 | Wu Xiangjun | 65 |
| Chen Weifang | 382 | Li Yan | 54 | Wu Yu | 229 |
| Chen Wentao | 350 | Li Yan | 166 | Xiang Yang | 18 |
| Chen Xingjie | 356 | Li Zhao-Xing | 298 | Xie Junping | 309 |
| Chen Yangyang | 166 | Liang Shuhui | 7 | Xie Yongbin | 80 |
| Chen Yaoting | 175 | Liang Yanxia | 80 | Xing Zongyi | 224 |
| Chen Yizeng | 267 | Lin Xiaowei | 175 | Xiong Kai | 71 |
| Chen Zhongjia | 48 | Liu Haiqiang | 368 | Xu Ming-xing | 159, 244 |
| Cheng Fuwei | 23 | Liu Honghai | 23 | Xu Ning | 304 |
| Cui Jifeng | 388 | Liu Jia | 23 | Xu Shuang | 80 |
| Cui Xiaofeng | 54 | Liu Jianfeng | 337 | Xu Zi-heng | 273 |
| Dai Chenglong | 388 | Liu Jun | 204 | Xue Yun-tao | 273 |
| Dai Feng | 337 | Liu Shue | 344 | Yang Jin | 97 |
| Dau Hoan Manh | 304 | Liu Shunchang | 218 | Yang Min | 133 |
| Deng Zhenrong | 92 | Liu Tang | 97 | Yang Yanfang | 122 |
| Ding Jian-Long | 382 | Liu Xiaohui | 337 | Ye Jizhen | 109 |
| Dong Jian-Gang | 377 | Liu Xiaoyu | 398 | Yong Li | 86 |
| Du Yibo | 253 | Liu Xilong | 267 | You Jianxin | 278 |
| Fan Zheng | 361 | Liu Xinmei | 147 | Yu Jiang | 86 |
| Fang Xujian | 406 | Liu Yezheng | 204 | Yu Rende | 283 |
| Feng Yanli | 344 | Liu Yijun | 41 | Yu Xing | 322 |
| Feng Zhang | 377 | Lu Guangyue | 80 | Yu Yue | 159, 244 |
| Fu Liping | 197 | Lu Hailong | 259 | Yu Yun-jun | 273 |
| Gao Cai-Yun | 212 | Lu Ying | 309 | Yuan Xiangyue | 48 |
| Gao Guangchun | 71 | Ma Peng | 325 | Yuan Yiming | 128 |
| Gao Ji | 382 | Ma Zhong | 65 | Zhang Chuan | 92 |
| Gao Ming | 181 | Mao Lingli | 244 | Zhang Chunying | 103 |
| Gao Ning | 212 | Meng Fei | 236 | Zhang Cui | 71 |
| Gao Wengen | 128 | Peng Jingliang | 109 | Zhang Hua | 218 |
| Han Zhao | 350 | Peng Lingxi | 97 | Zhang Hui | 36 |
| Hao Yang | 36 | Peng Zhangming | 373 | Zhang Jiansheng | 191 |
| He Li-le | 298 | Podolyakina Nataly | 317 | Zhang Jin | 253 |
| He Sheng | 41 | Qi Mingming | 18 | Zhang Shanwen | 60 |
| Hu Jinsong | 86 | Qin Yong | 224 | Zhang Shuiping | 290 |
| Hu X.zhong | 191 | Rao Yuan | 350 | Zhang Xi | 92 |
| Hu Yong-shi | 159, 244 | Ren Yifeng | 330 | Zhang Xiaojie | 147 |
| Huang Junkai | 350 | Seng Dewen | 406 | Zhang Yan | 86 |
| Huang Qing-Huang | 181 | Shang Shanshan | 278 | Zhang Yong-Heng | 377 |
| Huang Wei | 30 | Shen Juan | 283 | Zhang Yunxia | 388 |
| Huang Wenming | 92 | Shi Jihong | 86 | Zhang Zhe | 54 |
| Huang Wenzhun | 60 | Song Yaoliang | 133 | Zhang Zhijie | 330 |
| Huang Xiaodong | 393 | Su Shaohui | 368, 373 | Zhao Jun | 65 |
| Huang Yan | 110, 116 | Sui Peng | 273 | Zhao Junmei | 330 |
| Hui Meng | 140 | Tan Hongyan | 36 | Zhao Shengying | 71 |
| Jia Limin | 122, 224 | Tang Xingxing | 92 | Zhao Shiping | 398 |
| Jia Pengtao | 7 | Ter-Saakova Ilana | 317 | Zhao Yanjun | 103 |
| Jiang Guihong | 54 | Tian Lixin | 12 | Zheng Xianfeng | 361 |
| Jiang Guozhang | 23 | Tong Chao | 273 | Zong Rong | 86 |
| Jiang Jing | 80 | Wang Wenlong | 147 | | |

Cumulative Index

Mathematical and Computer Modelling

Pengtao Jia, Jun Deng, Shuhui Liang A fuzzy combined forecasting model of coal spontaneous combustion
Computer Modelling & New Technologies 2014 18(7) 7-11

This paper focuses on the effective analysis of the coal spontaneous combustion monitoring data, so as to realize the accurate and reliable coal spontaneous combustion limit parameter prediction. Firstly, a weighted multimember fuzzy operation model was constructed. When the additive generator of the model changes, this model can generate new operation clusters. Based on it, a new combined forecasting model of coal spontaneous combustion limit parameter is proposed. The new model can use linear and nonlinear models as its single forecasting models. Its combination is variable and has good generalization ability. Then, the BP neural network model and the support vector machine were used as the single forecasting models of the new model. Finally, for realizing the optimal combination of single models, genetic algorithm and least square method were used to evaluate parameters of new model. The experimental analysis shows that the new model leads to less error and better performance than single models. It can be concluded that the new combined forecasting model is suitable for coal spontaneous combustion.

Keywords: coal spontaneous combustion, limit parameters, combined forecasting, genetic algorithm, least square method

Guoliang Cai, Shengqin Jiang Jiang, Shuiming Cai, Lixin Tian Exponential synchronization of complex networks with non-delayed and delayed coupling via hybrid control
Computer Modelling & New Technologies 2014 18(7) 12-17

In this paper, the different structure synchronization of the two complex chaotic networks with time-varying delay and non-time-varying delay coupling is considered. Based on Lyapunov stability theory, combined with Yong inequality approach, Hybrid control including periodically intermittent control and adaptive control is designed such that the two complex chaotic networks achieves the exponential synchronization. Different numerical simulations are given to illustrate the effectiveness of the proposed method. Moreover through comparing the numerical simulations with the different functions of time delay, we can get how the time delay function impacts the complex chaotic networks synchronization in this model.

Keywords: complex chaotic networks, hybrid controller, time-varying delayed, Lyapunov stability theory

Mingming Qi, Yang Xiang Tensor modular sparsity preserving projections for dimensionality reduction
Computer Modelling & New Technologies 2014 18(7) 18-22

In order to reduce the computational complexity and promote the classification performance of Modular Weighted Global Sparse Representation (MWGSR), Tensor Modular Sparsity Preserving Projections (TMSPP) for dimensionality reduction is proposed. The algorithm firstly partitions an image into several equal-sized modules and constructs these modules into a third-order tensor image; then, the algorithm makes module sparse reconstructions and some modules with less reconstruction errors are selected. These selected modules are recombined into a dataset with fewer dimensions and a new sparse reconstruction weight is gotten on the new dataset, which is denoted as the sparse reconstruction weight of original samples; finally, projection matrices are gotten with steps of tensor sparsity preserving projections on the reconstructed tensor images. The algorithm promotes the computational efficiency and the robust performance of sparse preserving projections on high-dimensional datasets. Experimental results on YaleB and AR face datasets demonstrate effectiveness of proposed algorithm.

Keywords: dimensionality reduction, modular sparsity preserving projections, sparse reconstruction, the third-order tensor

Gongfa Li, Fuwei Cheng, Honghai Liu, Guozhang Jiang, Jia Liu Coke oven production process hybrid intelligent control
Computer Modelling & New Technologies 2014 18(7) 23-29

Coke oven production possesses the characteristics of nonlinear, large inertia, large disturbances, and highly-coupling and so on. According to the characteristics of coke oven production process and control demand of coke oven

production, intelligent control structure and models of coke oven production process was conducted. Firstly the intelligent control structure of coke oven production process was established. Then coal blending intelligent control, gas collector pressure intelligent control and combustion intelligent control of coke oven were discussed simply, while heating intelligent control, the production plan and schedule were discussed in detail. The control principle of combining the intermittent heating control with the heating gas flow adjustment was adopted, and fuzzy hybrid control was proposed to establish heating intelligent control strategy and model of coke oven, which combined feedback control, feed forward control and fuzzy intelligent control. The production plan and schedule of coke oven were optimized by utilizing the dynamic program and genetic algorithm. The practical running indicates that the system can effectively improve quality of coke and decrease energy consumption.

Keywords: Coke oven production process, Hybrid intelligent control, Heating intelligent control, Production plan and schedule

Huang Wei Research on output regulation for saturated systems

Computer Modelling & New Technologies 2014 18(7) 30-35

In this paper, the output regulation problem is investigated, which consists of building a controller to asymptotically steer the output of a saturated linear systems to a given reference signal despite external disturbances. Particularly, for saturated systems subject to periodically time-dependent exosystem, a K -step asymptotically regulatable region was characterized by a set of all the initial states of the plant and the exosystem. Improved internal model principles were constructed on the balance between the state convergence rate and the control of all the initial state. Finally, a state feedback controller was designed to ensure exponential output regulation in the regulatable region with disturbance rejection. Simulation examples were given to illustrate the effectiveness of proposed method. The results show these systems can go into stable rapidly and periodically.

Keywords: saturation constraint, output regulation, internal model principles, feedback controller

Hui Zhang, Jinzhao Wu, Hongyan Tan, Hao Yang Approximate trace equivalence of real-time linear algebraic transition systems

Computer Modelling & New Technologies 2014 18(7) 36-40

In allusion to data error and equivalence relation for software program design, the paper proposes approximate trace equivalence of real-time linear algebraic transition systems. Firstly, it leads real-time algebraic program into transition system and establishes real-time linear algebraic transition system. And then, it uses matrix norm and matrix singular value decomposition to analyze approximation of traces. Afterwards, it obtains approximate trace equivalence of real-time linear algebraic transition systems. Finally, the traffic light control vehicle flow system example shows that approximate trace equivalence of real-time algebraic transition systems can optimize real-time linear algebraic programs and reduce states.

Keywords: transition system, approximate, trace equivalence, algebraic program

Yijun Liu, Sheng He, Yao Wang, Xiumei Wang A comparative study on artificial neural networks for environmental quality assessment

Computer Modelling & New Technologies 2014 18(7) 41-47

The aim of this study is to use neural network tools as an environmental decision support in assessing environmental quality. A three-layer feedforward neural network using three learning approaches of BP, LM and GA-BP has been applied in non-linear modelling for the problem of environmental quality assessment. The case study shows that the well designed and trained neural networks are effective and form a useful tool for the prediction of environmental quality. Furthermore, the LM network has the fastest convergence speed and the GA-BP network outperforms the other two networks in both predictive and final classification accuracies of environmental quality.

Keywords: neural network model, hybrid ga-bp algorithm, environmental quality assessment

Xiangyue Yuan, Zhongjia Chen Design of Q450 pellet molding machine and force analysis of its molding assembly based on Solidworks

Computer Modelling & New Technologies 2014 18(7) 48-53

Energy shortage and environmental pollution is a common serious problem restricting the development of world economy and the society. Biomass energy has become the fourth major energy resource after oil, coal, natural gas energy for the good properties of green, clean, and renewable. So it is important to research biomass energy technology to solve the energy crisis and environmental protection. The technology of biomass densification is a simple solution to make the biomass resource become low cost and high value. In this paper, a new kind of biomass pellet molding mechanism had been deeply studied, and the pellet molding machine, Q450, was designed by the CAD/CAE, Solidworks. The load conditions of three molding assemblies fixed inside the enclosure bodies had also been studied and analysed on the designed machine. Then a method, named FEA (Finite Element Analysis), was conducted to research the mechanical properties of enclosure assembly for the pellet machine in Simulation. Through analysis, the results were obtained that the maximum stress and displacement of enclosure bodies were separately 31.37Mpa and $7.583e^{-2}$ mm, which could provide the reliable strength and stiffness to the enclosure assembly. It convincingly ensured that Q450 pellet molding machine had enough reliability and security.

Keywords: pellet molding mechanism with plunger-roller ring die, Q450 pellet molding machine, enclosure assembly, strength and stiffness with FEA (Finite Element Analysis)

Yan Li, Zhe Zhang, Guihong Jiang, Xiaofeng Cui Model driven testing distributed environment monitoring system

Computer Modelling & New Technologies 2014 18(7) 54-59

Distributed environment monitoring system is more and more widely used, especially the design and verification of embedded system in environmental monitoring is the guarantee of successful use of environmental monitoring. In this paper we demonstrate how test-case prioritization can be performed with the use of model-checkers. For this, different well known prioritization techniques are adapted for model-based use. New property based prioritization techniques are introduced. In addition it is shown that prioritization can be done at test-case generation time, thus removing the need for test-suite post-processing. Several experiments for embedded systems are used to show the validity of these ideas.

Keywords: test case prioritization, software testing, model checking, property testing

Wenzhun Huang, Shanwen Zhang Source enumeration algorithm based on eigenvector: revisit from the perspective of information theory

Computer Modelling & New Technologies 2014 18(7) 60-64

In case of low signal to noise ratio (SNR) and small snapshot condition, it is difficult to separate sources and noises, and the performance of classical eigenvector source estimation algorithm drops quickly. To solve the problem, further research is carried out around the characters of eigenvalue and eigenvector, and a novel eigenvalue algorithm is presented based on the theory of source enumeration. In detail, the eigenvectors of sample covariance matrix are employed as the decision factor, which is insensitive to SNR. And an improved Predictive Description Length (PDL) criterion is adopted to enumerate source number. Theoretical analysis and simulation results demonstrate that the proposed algorithm is available and efficient in case of low SNR and small snapshot condition compared with those of Minimum Description Length (MDL) and PDL.

Keywords: source enumeration, eigenvector, signal-to-noise ratio, information theory, predictive description length

Computer and Information Technologies

Jun Zhao, Zhong Ma, Xiangjun Wu A method for improving real-time communication of switched Ethernet
Computer Modelling & New Technologies 2014 18(7) 65-70

A method has been proposed for improving real-time communication of Switched Ethernet. Based on virtual link ideas, this method offline plans the whole network traffic under traditional Switched Ethernet hardware conditions, improves network terminal TCP/IP protocol by adding real-time communication interface, traffic shaping and priority queuing etc., and uses IEEE802.1p protocol on the switches of communication path. And it employs the network calculus theory to deduce the equation of calculating maximum end-to-end delay of real-time traffic. Meanwhile, it receives a simulation test of OPNET software. Both theoretical calculation and simulation results show that this method can effectively improve real-time communication of Switched Ethernet.

Keywords: Switched Ethernet, virtual link, real-time, traffic shaping, network calculus

Guangchun Gao, Kai Xiong, Shengying Zhao, Cui Zhang Optimal adaptive wavelet transforms without using extra additional information
Computer Modelling & New Technologies 2014 18(7) 71-79

Wavelet transforms via lifting scheme provides a general and an adaptive flexible tool for the construction of wavelet decompositions and perfect reconstruction filter banks. According to the construction of the lifting wavelet transforms, the optimal filter design method for the adaptive update wavelet transform is proposed by the authors. The optimal update filter coefficients can be acquired based on the Minimum Mean Square Error Criteria (MMSE) in the algorithm. In prediction process, take the case of LeGall 5/3 wavelet, we propose an adaptive version of this scheme that it allows perfect reconstruction without any overhead cost for the smooth signals with the jumps. Compare with other wavelet transform scheme, simulation results show that optimal adaptive wavelet transform proposed by this paper can achieve the detail signals being zero (or almost zero) at big probability and the better linear approximation for the piecewise continuous signal.

Keywords: Lifting scheme, adaptive wavelet transform, optimal update filter, MMSE

Jing Jiang, Shuang Xu, Guangyue Lu, Yongbin Xie A large-scale MIMO channel information feedback algorithm based on compressed sensing
Computer Modelling & New Technologies 2014 18(7) 80-85

In order to effectively reduce the feedback overhead of channel state information (CSI), a channel state information feedback algorithm based on compressed sensing was proposed for Large-scale MIMO system. Firstly considering the sparsity of spatial-frequency domain for the large-scale MIMO channel, the channel information was compressed in space domain firstly and in frequency domain subsequently, the receiver acquired the measurement vector based on compressed sensing algorithm; then feedback. CSI observations to the transmitter according to the proposed adaptive feedback protocol, at last the transmitter reconstructed CSI based on the Basis Pursuit (BP) algorithm. It is show in stimulation results that the proposed algorithm can acquire similar BER performance with perfect channel information feedback. The proposed algorithm, which feedbacks the compressed channel information, not only can significantly reduce the feedback overhead, but also ensure that large-scale MIMO performance gain.

Keywords: Large-scale MIMO, Channel State information feedback, Compressed Sensing

Yong Li, Jiang Yu, Rong Zong, Yan Zhang, Jihong Shi, Jinsong Hu Heterogeneous networks model for lower error using concatenated encoding
Computer Modelling & New Technologies 2014 18(7) 86-91

Power line communication of the power distribution network requires higher reliability and better standards. In this paper, the ideas of cooperative communication motivate that we have presented a dual heterogeneous networks model with power line communication network and wireless sensor networks. Concatenated channel encoding with cyclic redundancy check code, convolution code, Reed Solomon code and the interleaver are respectively researched, which

are used to analyse the performance of dual networks' communication scheme. The simulated results reveal that the dual networks have better communication quality than the one of single network. Compared to dual-network with RS code, the probability of frame error transmission and bit error rate for dual-network with convolution code are lower. On the basis of the characteristics of two different kinds of concatenated codes, we put forward an improved model, which is more close to reality. The results verify the feasibility of the design.

Keywords: power distribution network, power line communication, wireless sensor network, dual heterogeneous networks, concatenated channel encoding

Zhenrong Deng, Xingxing Tang, Chuan Zhang, Xi Zhang, Wenming Huang Improvements and implementation of the permission system based on RBAC model

Computer Modelling & New Technologies 2014 18(7) 92-96

Role-based access control as a traditional access control (discretionary access, mandatory access) is a promising place to receive widespread attention. Systematic researches on RBAC models, on one hand, this paper combined with the characteristics of Electronic government affair information management system and added regional filter function to the core RBAC model, besides, the research developed by J2EE framework and this paper presents a high availability and extensibility of RL-RABC competence management system.

Keywords: RBAC, J2EE, permission system

Jin Yang, Lingxi Peng, Tang Liu Anti-spam model based on AIS in cloud computing environments

Computer Modelling & New Technologies 2014 18(7) 97-102

Cloud computing is becoming a hot research topic. However, there is little attention to cloud computing environment work for anti-spam issues. Spam has become a thorny issue facing with many countries. The overflow of spam not only great wastes the network resources, taking up the user's e-mail resources, reducing the network efficiency, affecting the normal use of the Internet, but also violates the user's individual rights. But the traditional spam solutions for anti-spam are mostly static methods, and the means of adaptive and real time analyses the mail are seldom considered. Inspired by the theory of artificial immune systems (AIS), this paper presents an anti-spam system in cloud computing environment. The concepts and formal definitions of immune cells are given, and the hierarchical and distributed management frameworks of the proposed model are built. The results of evaluation indicate that the proposed model has the features of real-time processing and is more efficient than client-server-based solutions, thus providing a promising solution for anti-spam system for heterogeneous cloud environments.

Keywords: cloud computing, artificial immune systems, anti-spam system

YanJun Zhao, Chunying Zhang Research and application on set pair entity similarity model of social network

Computer Modelling & New Technologies 2014 18(7) 103-109

In allusion to the certain and uncertain value which exist in the node and relationship attributes of social network, From the attributes and relation angle of the entity to analyse the similarity degree of the affirmative, negative and uncertain between them, then build the set pair entity similarity model based on the set pair analytical method and apply it to the network association detection. First of all, applying the generalized pair close potential and the generalized set loose potential in social network based on set pair analysis method, and see it as the basis of association detection; secondly, giving the set similarity calculation method based on the entity attribute and relation, from the point of view of node attribute and relations attribute to calculate respectively, by setting the weight to consolidated calculate the set pair similarity of the entity; thirdly, utilizing entity set similarity to divide network association into clustering problem, then give the association partitioning algorithm; finally, integrating with the network instance to verify the effectiveness of the new network association partitioning algorithm.

Keywords: set pair, social network, entity similarity, attribute and relation, association partitioning

Jizhen Ye, Jian Wei, Yan Huang, Jingliang Peng Comparative study of DXT1 texture encoding techniques

Computer Modelling & New Technologies 2014 18(7) 110-115

In this paper, we make a comprehensive survey of many different methods to implement DXT1 (a widely used lossy texture compression algorithm). Besides that, we propose two new methods that aim for computing speed and image quality, respectively to implement DXT1 texture compression algorithm. For computing speed, we propose a new method called Lsq3d fit which achieves a very fast speed to encode texture images while keeping acceptable image quality. For image quality, we propose a new method called kmeans iteration fit and make a combination of it and the cluster fit from libsquish (an open source lib for DXTC). Kmeans iteration fit performs competitively in the quality of compressed texture images compared with the state-of-the-art DXT1 encoders, and we achieve different levels of quality by controlling the times of iteration. Finally, we test all the methods on Kodak Lossless True Color Image Suite, and CSIQ (Computational Perception and Image Quality Lab) image dataset. Our proposed methods have competitive results of speed and quality in both image datasets. The combination of cluster fit and kmeans iteration fit defeats all other methods in the quality of compressed images.

Keywords: Texture compression, DXTC, DXT1, S3TC, k-means clustering

Xian Wu, Yan Huang Real-Time and interactive browsing of massive mesh models

Computer Modelling & New Technologies 2014 18(7) 116-121

We present an efficient method for out-of-core construction and real-time interaction of massive mesh models. Our method uses face clustering on an octree grid to simplify and build a Level-of-Detail (LOD) tree for the model. Each octree node leads to a local LOD tree. All the top layers of the local LOD trees are combined together to make the basis of the global LOD tree. At runtime, the LOD tree is traversed top down to choose appropriate local LOD trees given the current viewpoint parameters. The system performance can be dramatically improved by using hierarchical culling techniques such as view-frustum culling and back-face culling. The efficiency and scalability of the approach is demonstrated with extensive experiments of massive models on current personal computer platforms.

Keywords: massive mesh model, out-of-core, level-of-detail, mesh simplification, culling

Xiaofeng Li, Yanfang Yang, Limin Jia A kernel induced energy based active contour method for image segmentation

Computer Modelling & New Technologies 2014 18(7) 122-127

Active contour model is a promising method in image segmentation. However, existing active contour model and its evolution often suffer from slower convergence rates and easily to be trapped in local optima due to the presence of noise. In this paper, a novel curve evolution model based on kernel mapping method is presented. The method first transforms original image data into a kernel-induced space by a kernel function. In the kernel-induced space, the kernel-induced non-Euclidean distance between the observations and the regions parameters is integrated to formulate a new level set based active contour model. The method proposed in this paper leads to a flexible and effective alternative to complex model the image data. In the end of this paper, detailed experiments are given to show the effectiveness of the method in comparison with conventional active contour model methods.

Keywords: Kernel mapping; Active contour; Chan-Vese model; Level-set; Image segment

Yiming Yuan, Ming Jiang, Wengen Gao Image fusion based on MPCNN and DWT in PCB failure detection

Computer Modelling & New Technologies 2014 18(7) 128-132

The traditional contact-type printed circuit board (PCB) test methods have been unable to meet the needs of the fault detection and maintenance of a variety of increasingly complex electronic equipment. The visible and infrared respectively reflects the background information and the radiation information of PCB, so we can fuse the visible image and infrared image of the board together, and use the new fusion image to locate and identify the abnormal high temperature components or areas of the circuit board. A novel fusion algorithm of multi-sensor image is proposed based on Discrete Wavelet transform (DWT) and pulse coupled neural networks (PCNN) in this paper. Firstly, the IR and visible images are decomposed by DWT, then a fusion rule in the DWT is given based on the PCNN. This algorithm uses the local entropy of wavelet coefficient in each frequency domain as the linking strength, then its value

can be chosen adaptively. After processing PCNN with the adaptive linking strength, new fire mapping images are obtained. According to the fire mapping images, the firing time gradient maps are calculated and the fusion coefficients are decided by the compare-selection operator with firing time gradient maps. Finally, the fusion images are reconstructed by wavelet inverse transform. The proposed algorithm of image fusion using modified pulse coupled neural networks (MPCNN) and DWT results in better quality of fused image with Entropy, Average grads, Cross-Entropy as compared to conventional image fusion Algorithms.

Keywords: PCNN image-fusion, DWT, PCB, failure detection

Min Yang, Yaoliang Song, Qianmu Li Research on virus transmission of online social network

Computer Modelling & New Technologies 2014 18(7) 133-139

Online social networks (OSN) are up-and-coming complex network systems. Experiments indicate that it is difficult for simple complex network theory to describe virus transmission behaviour. Based on comprehensive research into current virus transmission, this paper combines user behaviour with social engineering theory and builds a model of virus transmission on OSN. Key factors affecting virus transmission on OSN are then analysed. Lastly, in light of public opinion transmission theory, this paper refers to social reinforcement factors concepts to describe computer virus transmission on OSN and analyses transmission disciplines in regular and random networks.

Keywords: online social network, virus transmission, epidemic spreading

Lin Bai, Meng Hui SVM classification of hyperspectral images based on wavelet kernel non-negative matrix factorization

Computer Modelling & New Technologies 2014 18(7) 140-146

This paper presents a new kernel framework for hyperspectral images classification. In this paper, a new feature extraction algorithm based on wavelet kernel non-negative matrix factorization (WKNMF) for hyperspectral remote sensing images is proposed. By using the feature of multi-resolution analysis, the new method can improve the nonlinear mapping capability of kernel non-negative matrix factorization. The new classification method of hyperspectral image data combined with the novel kernel non-negative matrix factorization and support vector machine (SVM). The simulations results show that, the method of WKNMF reflect the nonlinear characteristics of the hyperspectral image. Experimental results on Airborne Visible Infrared Imaging Spectrometer 220 bands data in Indian pine test site and HYDICE 210 bands hyperspectral imaging in Washington DC Mall are both show that the proposed method achieved more strong analysis capability than comparative algorithms. Compared with the PCA, non-negative matrix factorization and kernel PCA method, classification accuracy of WKNMF with SVM can be improved over 5%-10%.

Keywords: hyperspectral, non-negative matrix factorization, classification, support vector machine, kernel method

Operation research and decision making

Wenlong Wang, Xinmei Liu, Xiaojie Zhang The optimal promised quality defect model for service guarantees

Computer Modelling & New Technologies 2014 18(7) 147-158

Service quality guarantee is an important tool for firms to boost demands, put up prices, and enhance profits. However, when promised quality defect is too high or low, the impact on the organization and the customer is usually negative. Therefore, determining the level of promised quality defect is of critical strategic and tactical importance in businesses. Yet, systematic quantitative methods aren't found to help managers determine promised quality defect. We propose a simple but powerful model in finding the optimal promised service quality defect. The model makes trade-offs between benefits and costs of service defect guarantees. Firstly, the decision of promised quality defect is analysed when service price is exogenous. We secondly investigated when service price is endogenous, how can a service provider make decisions on service price and promised quality defect simultaneously to maximize its profit. Thirdly, comprehensive analysis of how service providers promise the optimal quality defect from two aspects of

demand and supply is given. Numerical analysis is conducted to illustrate the interactive effect of endogenous service price and affected service supply. In the end, we conclude the paper and suggest areas for future research. With only definitional changes, the model can be applied to other guarantee contexts.

Keywords: quality guarantees; promised quality defect; service providers; affected service supply

Yu Yue, Hu Yong-shi, Xu Ming-xing Research on supply chain competition advantage under repeated games

Computer Modelling & New Technologies 2014 18(7) 159-165

To reveal whether the order of supply chains' competition exerts an effect on their profits and whether the repeated game interferences this effect, the paper builds a Stackelberg game model constructed by two supply chains with each containing a supplier and a retailer based on the previous studies. Through comparing respective profits of the leading and following supply chain represented by 'Copycat', this paper concludes that the following supply chain is more likely to gain more profits than the leading one in this case, and this advantage is determined by the order of decision-making itself. Under repeated games, the possibility of the following supply chain to be more profitable and the approaches to make decisions will be related to the substitutable coefficient.

Keywords: leading-following supply chain, Stackelberg game, repeated games, later-mover advantage

Yan Li, Dong Wei, Yangyang Chen The development and evolution of bridge in Chongqing China

Computer Modelling & New Technologies 2014 18(7) 166-174

In this paper, the development and evolution of bridge in Chongqing and around the world are summarized particularly. Besides, the categories of bridges, the development of bridge design theories as well as the breakthroughs of bridge construction with the passage of time are also introduced systemically. With the introduction of the historical stone-arch bridges, the recent reinforced concrete slab bridges, modern pre-stressed concrete bridge, various arch bridges, suspension bridges and cable-stayed bridges, the prosperity and progress made by human beings in the process of transformation of nature are gradually revealed in this paper, the development and evolution of bridge is also revealed with the introduction of new techniques. The construction of bridge promotes the economic development and strengthens the connection of different areas, brings a booming market. The role of mechanics in bridge design is analysed and the development of Chongqing bridges can also be experienced in this paper. Bridge is not only a construction but also the creator of the soul of a city, showing the fighting spirit and braveness of a generation.

Keywords: Chongqing, Bridges, Development, Evolution model

Yaoting Chen, Xiaowei Lin A comparative study on efficiency of two different circulation modes of agricultural products based on DEA model: wholesale market and logistics distribution centre

Computer Modelling & New Technologies 2014 18(7) 175-180

The purpose of this study is to find out the relatively efficient circulation mode of agricultural products through a comparative analysis on the operating efficiency of two different circulation modes of agricultural products: wholesale market and logistics distribution centre. Based on the input and output data collected from the survey of the main representatives of enterprises in the two modes in Zhangzhou, Fujian, including: fixed assets, number of employees, main business cost, main business net profit, gross margin, the paper uses Data Envelopment Analysis to conduct the analysis. The results show that the third party logistics mode based on logistics distribution centre is relatively more efficient, comparing with the traditional wholesale market mode. Therefore, in order to reduce circulation cost of agricultural products, and to promote the development of agricultural industry, it is necessary to make policies to encourage the development of the third party logistics mode based on logistics distribution centre.

Keywords: data envelopment analysis model, wholesale market mode, logistics distribution centre mode, operation efficiency

Qing-huang Huang, Ming Gao A study on mechanism of environmental protection industry innovation under open innovation - the intermediary effect based on the enterprise network dynamic capability

Computer Modelling & New Technologies 2014 18(7) 181-190

In the dynamically changing external environment, it is the core issue of enterprise innovation strategy that how enterprises maintain a sustained level of innovation by creating their own capabilities. In addition, the open innovation proposed by Chesbrough provides a new way of thought for innovation management. This essay constructs conceptual models of several sets of variables relationships between environmental protection industry innovation performance and external innovation resources, which is based on 85 environmental protection enterprises as the questionnaire objects and a path analysis of the model is conducted. The results show that the cooperation with horizontal and vertical enterprises can significantly affect innovation performance only by virtue of the intermediary effect of the enterprise network dynamic capability, and government-industry-academy-research cooperation can directly improve innovation performance. Mechanistic study not only reveals that the joint action by external innovation resources and network dynamic capabilities can influence the innovation and motivation of environmental protection enterprises, but also reflects that a major source of environmental protection innovation is the internal resources. This provides theoretical guidance for enterprises to effectively implement the open innovation strategy in the innovation practice.

Keywords: open innovation, environmental protection industry, innovation performance, network dynamic capability

Jiansheng Zhang An analysis on the growth and effect factors of TFP under the energy and environment regulation: data from China

Computer Modelling & New Technologies 2014 18(7) 191-196

The paper analyses the growth and effect factors of TFP (Total factor productivity) under the energy and environment regulation with the data of China from 2002 to 2012. The results show that: in the past 10 years, without considering the energy and environmental regulation, the average annual growth rate of TFP is 3.2%, but it is 2.7% when considering them. The technological progress is the major contributor to TFP under the energy and environment regulation. From the comparison of various provinces, the growth difference of TFP was great. The TFP value in eastern coastal region is higher than that in the central and western regions. From the time trend, the average growth rate of TFP is in the lower. After the financial crisis of 2008, the TFP starts to decline and the average annual growth rate is -0.3%. The three variables of the FDI, environmental regulation intensity and industrial structure have a negative impact on TFP growth, but the two variables of R&D investment and energy consumption structure have a positive impact on it.

Keywords: environmental pollution, energy regulation, TFP, effect factors

Liping Fu, Juan Li Analysis of the public satisfaction index of public cultural services based on the Grey Correlation AHP method

Computer Modelling & New Technologies 2014 18(7) 197-203

The public is the service object of the public cultural services while the public satisfaction index is the main indicator in the judgment of the public cultural service effect. The Grey Correlation Method is applied to selecting the main factors which influence the public satisfaction index of public cultural services, and the number of public library, the public cultural activities of organizations, and the number of staff in the public cultural service institutions is the most three important factors. After that, the paper builds public satisfaction model based on grey correlation AHP, applies the method to evaluating the current public satisfaction of public cultural services in China, and proposes the specific measures to improve and promote public cultural services in China on the basis of the evaluation result.

Keywords: public cultural services, public satisfaction index, Grey Correlation AHP method

Yezheng Liu, Jun Liu Asymmetric effects of exchange rate pass-through: an empirical analysis among China, the United States and Japan

Computer Modelling & New Technologies 2014 18(7) 204 – 211

From the perspective of exchange rate direction fluctuations, this paper comparatively studied the asymmetric effect of movements in the nominal exchange rate on consumer prices among China, the United States, and Japan. To this end, the paper used the error correction model (ECM) to conduct an empirical analysis from the first season of 1994 to the last season of 2010 period. The results showed that: (1) the pass-through of exchange rate movements to consumer

prices was incomplete; (2) exchange rates fluctuated in different directions, meaning that when the exchange rate appreciated and depreciated, the pass-through of exchange rate movements to consumer prices was asymmetric. However, the direction varied among the three countries. The influence of depreciation on consumer prices was higher in both China and the United States, while Japan was the opposite; (3) exchange rate pass-through was different in the three countries. The level of exchange rate pass-through in China was higher than the other two countries; (4) when short-term fluctuations deviated from long-term equilibrium, the adjustment was higher in the United States, followed by Japan, and China was relatively lower. These results had important implications for current monetary policies and practices.

Keywords: exchange rate, consumer prices, exchange rate pass-through, asymmetry

Ning Gao, Cai-Yun Gao Deformation forecasting with a novel high precision grey forecasting model based on genetic algorithm

Computer Modelling & New Technologies 2014 18(7) 212-217

The precision of prediction of grey forecasting model depends on the conformation of background value and the selection of the initial condition. Existent literatures optimized grey forecasting model just from one side, respectively. Therefore, a novel model named BIGGM (1,1) is proposed in this paper by integrated optimizing background value and initial condition. In addition, genetic algorithm has also been integrated into the new model to solve the optimal parameter estimation problem. An illustrative example of deformation of Lianzi cliff dangerous rock along the Yangtze River in china is adopted for demonstration. Results show that the BIGGM (1,1) model can increase the prediction accuracy, and it is suitable for use in modelling and forecasting of deformation.

Keywords: GM (1,1) model, background value, initial condition, genetic algorithm, integrated optimization, deformation forecasting

Hua Zhang, Shunchang Liu Effect judgment and effectiveness estimation of anti-dumping duty – an example of the case of canned mushroom

Computer Modelling & New Technologies 2014 18(7) 218-223

The paper aimed to provide a method for accurate estimation of the Anti-Dumping (AD) tax. Having used the case of canned mushrooms, exported from Indonesia to the United States as an example, the paper presented methods for effective judgment and accurate estimation of the AD. By having done so, it provided AD policy makers with a scientific and fair AD tax which, at the same time, would incorporate legislation that will prevent abuse of AD taxation system. The paper further analysed the impact of AD tax on the trend of export; the spread of related indicators between a taxable situation and a non-taxable one through the Chow test and other methods. Final results provided AD users with logical basis and method support.

Keywords: SOFC, discrete sliding mode, control, DC/AC, converter

Zongyi Xing, Lingli Mao, Limin Jia, Yong Qin Identification of key subsystems for urban rail vehicles based on fuzzy comprehensive evaluation

Computer Modelling & New Technologies 2014 18(7) 224-228

Identification of key subsystems for urban rail vehicles is important for the selection of maintenance strategy. The fuzzy comprehensive evaluation technique is applied to determine the key subsystems of urban rail vehicles. Firstly, the vehicle is divided into nine subsystems according to the module partition method. Then, the degrees of occurrence, severity, detection and maintenance cost are chosen as the evaluation factors that are quantified based on fuzzy theory and collected historical data. Finally, the calculation model of critical degree is established based on the fuzzy comprehensive evaluation method. The proposed approaches are applied to Guangzhou Metro Corporation, and five key subsystems are selected. The experiment results, which are consistent with those of most knowledgeable engineers and experts, indicate the validity of the proposed method.

Keywords: key subsystem, urban rail vehicle, fuzzy comprehensive evaluation

Wu Yu Reputation risk contagion and control of rural banks in China based on epidemic model*Computer Modelling & New Technologies 2014 18(7) 229-235*

Rural bank reputation risk is the negative evaluation formed in stakeholders' minds as a result of events which pose both internal and external risks. Regardless of whether or not these risk events have actually occurred, any resulting negative evaluations tend to propagate and accumulate in both the public's mind and within the main financial system. The growing negative opinion can create a herd effect, ultimately creating a reputation crisis. This paper attempts to research the contagion mechanism of rural bank reputation risk based on epidemic model, then explores a simulation study under different situations. The results show that the key to prevent or regulate reputation risk contagion is to reduce the unit available contact rate and the re-entry ratio, as well as the lurker infected rate. Finally, this paper puts forward management and control strategies from the perspective of the entire process. These strategies specifically focus on constructing an early warning mechanism, a dissolving mechanism and a long-term mechanism.

Keywords: reputation risk, reputation risk contagion, reputation risk control, epidemic model, rural bank

Fei Meng, Jianliang Wei Research on the influential factor of consumer model based on online opinion leader*Computer Modelling & New Technologies 2014 18(7) 236-243*

With the development of internet and e-commerce, online opinion leader becomes an important information resource which influences the purchase decision and behaviour model of online consumer, although the influential mechanism is still uncertain. In order to obtain a more precise user model, Grounded Theory is adopted in this paper and an interview table is designed according features of online opinion leader. Then, more than 20 online consumers concerning on opinion leader frequently in internet communities such as Taojianghu and Douban are selected for interview. After open coding, axial coding and selective coding on the interview materials, several findings are obtained: professional knowledge and interactive features of opinion leader influenced the purchase intension of consumer; characters such as visual cues and timeliness of recommended information from opinion leader have impact on consumer intension; consumer perceived value of product recommended by opinion leader influenced their purchase behaviour; and trust is the principle reason for consumer' acceptance on product information recommended by opinion leader.

Keywords: online opinion leader, grounded theory, consumer behaviour model

Ming-xing Xu, Yue Yu, Yong-shi Hu Service and revenue sharing strategies in a dual-channel supply chain with fairness concerns*Computer Modelling & New Technologies 2014 18(7) 244-252*

This paper incorporates the concept of fairness in a dual-channel supply chain to examine the effect of fairness concerns on the supply chain partners' service and revenue-sharing strategies in three different scenarios: only the retailer is concerned about fairness, only the manufacturer is concerned about fairness, and both parties are concerned about fairness. Though applying the equilibrium analysis, the results show that (1) Fairness concerns strongly influence the manufacturer's and the retailer's decision-making and utility. (2) The revenue sharing ratio increases with the strengthening of channel members' fairness concerns. (3) If only the retailer is concerned about fairness, the retailer's service is unaffected by his fairness concerns. (4) There exists a Pareto improvement for channel members' utility when the manufacturer without fairness concern becomes fair-minded.

Keywords: fairness concerns, dual-channel supply chain, service level, revenue sharing

Yibo Du, Jin Zhang Time-varying decision-making for hazardous chemical transportation in a complex transportation network*Computer Modelling & New Technologies 2014 18(7) 253-258*

The transit and storage of hazardous chemicals are harmful. A distributed decision model for hazardous chemicals is developed in this study, with the time window established, to improve the efficiency of transportation and storage. The route, mode, time, and volume of each demand can be determined by this model. The model minimizes the total

transportation risk and cost. The model is divided into two parts, and the corresponding ant colony algorithm is designed and achieved. The feasibility and efficiency of the model are illustrated through a numerical example with eight transfer nodes, six origin–destination (OD) demands, and multiple transportation mode alternatives. The developed model provides an effective approach for hazardous chemical substance transportation.

Keywords: hazardous chemicals, transportation decision, nonlinear mixed integer programming model, complex transportation network, ant colony algorithm

Hailong Lu, Yong Cen Research and implementation on integration information platform in China tobacco industry enterprise

Computer Modelling & New Technologies 2014 18(7) 259-266

For better information technology combined with enterprise management mechanism to solve the information technology into the enterprise production and operation of each link, the integration of information platform for building modern tobacco industry enterprise is proposed. First of all, this study combines the principal business process for building market-driven enterprise, and proposes based on self-assembled dynamic fifth-order business model for enterprise business operation model. According to the actual situation of enterprise information system construction, the integration information platform application architecture and integration architecture designs is developed. This information platform uses SAP XI as the ESB transforms the formats of all data coming from source systems to realize the seamless integration among different systems. With this integration information platform construction it can better support enterprise development strategies, optimizing resource allocation, improve business and management efficiency, and promote scientific enterprise sustainable development.

Keywords: tobacco industry enterprise, business process model, integration information platform

Xilong Liu, Yizeng Chen A fuzzy clustering approach of the customers' demands, which influences the e-banking service quality

Computer Modelling & New Technologies 2014 18(7) 267-272

The interest rate liberalization have a huge influence for commercial banking management in China, the net interest margin (NIM) is more and more low, but the competition is becoming increasingly fierce, so it is a very necessary and urgent work to strength the management of banks by the key financial innovation. Some researches showed that the e-banking service quality plays an important role during competition among banks as well as the core competence of banks' sustainable development. In order to improve the service quality, the first task is to really master and understand the customer demands, which influence the e-banking service quality. The paper proposed a fuzzy clustering method for customer demands and empirical analysis, and the results showed that all the customer demands can be classified into two clusters according to the maximum value of F-statistics, one of which indicated that the new trend of customer demands in e-commerce environment and have great influence on the decisions of users to use the service of e-banking.

Keywords: e-banking, service quality, customers' demands, fuzzy clustering

Yun-jun Yu, Sui Peng, Yun-tao Xue, Chao Tong, Zi-heng Xu An autonomous decision making algorithm applied for the evaluation of power quality

Computer Modelling & New Technologies 2014 18(7) 273-277

An autonomous decision making algorithm applied for the evaluation of the field power quality is proposed. This algorithm can reflect to the characteristics of evaluation objects, develop evaluation objects initiatives, weakens the influence of the subjective weight of index on evaluation results and implements the comparison of different power qualities of the assessed in the area. The paper introduces the implementation steps of autonomous decision making algorithm, analyses the competition scope of the power quality of the assessed with this algorithm. The competition model is established, which output the comprehensive evaluation results of the assessed. The simulation demonstrates the effectiveness and practicability of this method.

Keywords: power quality, autonomous decision, algorithm, evaluation

Shanshan Shang, Jianxin You A simulation model on the formation of knowledge-based collaborative networks*Computer Modelling & New Technologies 2014 18(7) 278-282*

Collaborative network has been a hot topic in the related research field. This paper proposes a simulation model on the formation of knowledge-based collaborative networks mainly based on the Set theory. The paper proposes that formation process as follows: (1) find the key skills and the core members; (2) classify the organizations; (3) establish the relationship between organizations in different classifications.

Keywords: Knowledge-Based, Collaborative Networks, Set theory

Rende Yu, Juan Shen Analysis on road traffic accidents spatial distribution based on the multi-fractal theory*Computer Modelling & New Technologies 2014 18(7) 283-289*

After analysing the characteristics on spatial distribution of road traffic accidents in some areas in China, this paper took road traffic accidents of some provinces/cities in China as an example and thought those provinces/cities as cells. Then the fractal spectrum of road traffic accidents spatial distribution in two-dimensional space and $\ln \varepsilon - \ln \chi_q(\varepsilon)$ curve was obtained by MATLAB programming based on multi-fractal theory. Because of the preferable linear relation $\ln \chi_q(\varepsilon)$ between and $\ln \varepsilon$, the conclusion of which road traffic accidents spatial distribution satisfies power-law form and accord with multi-fractal distribution was obtained. By calculating the relation of related parameters, this paper analysed the characteristics of road traffic accidents spatial distribution further.

Keywords: Road traffic accidents, spatial distribution, multi-fractal spectrum, deaths, injuries

Shuiping Zhang Research on the principal-agent problems in China's low-carbon ecological urban construction*Computer Modelling & New Technologies 2014 18(7) 290-297*

In the game of the interest bodies of low-carbon ecological urban construction, the central government, as a principal, will lose some interests in some ways because of information disadvantages, whereas the local governments, as agents, will make use of their information advantages to make profitable action choices for more interests. As a result, moral risks will appear for the latter. This paper attempts to construct a mathematical model of the game theory for the principal-agent problems in the low-carbon ecological urban construction and analyses the choice actions involved. The conclusion is drawn that for the optimal balance of the game to be realized between the central and local governments, a relevant system must be established. This system is expected to change the information asymmetry by increasing the central government's ability to acquire information while stimulating or restraining the local governments' choice actions so that the external pressure on the local governments will be turned into their internal actions in a low-carbon ecological urban construction.

Keywords: low-carbon ecological cities and towns, agency by agreement, information asymmetry, system

Li Zhao-Xing, He Li-le A multi objective optimization algorithm for recommender system based on PSO*Computer Modelling & New Technologies 2014 18(7) 298-303*

In order to follow the development of Internet information service and improve the accuracy of recommender systems and recommendation algorithm. An optimal selection approach of multi-objective and particle swarm optimization (MOP-PSO) was put forward based on PSO algorithm. Furthermore, through two sets are combined and repeated dynamic adjustments, to achieve a better balance in algorithm efficiency and accuracy. Proposed a weighted cosine similarity method to calculate the user similarity, and then optimizing the weight by the PSO algorithm. Simulation results show that the algorithm has a better effective and can effectively improve the scoring accuracy, effectively improve the quality of the recommendation system.

Keywords: particle swarm optimization, recommended system, multi objective optimization

Hoan Manh Dau, Ning Xu The effectiveness of using methods two-stage for cross-domain sentiment

classification

Computer Modelling & New Technologies 2014 18(7) 304-308

Traditional sentiment classification approaches perform well in sentiment classification but traditional sentiment classification approaches does not perform well with learning across different domains. Therefore, it is necessary to build a system which integrates the sentiment orientations of the documents for every domain. However, this needs much labelled data involving and much human labour as well as time consuming. Thus, the best solution is using labelled data in one existed in source domain for sentiment classification in target domain. In this paper, a two-stage approach for cross-domain sentiment classification is presented. The First Stage is building a bridge between the source domain and the target. The Second Stage is following the structure. The study shows that the mining of intrinsic structure of the target domain brings a considerable effectiveness during the process of sentiment transfer. This is a typical mining approach comparing to previous approaches basing on information from the source domain to address the task of sentiment transfer, which does not depend on intrinsic structure of the target domain. Experimental results on sentiment classification with a two-stage approach indicate that the effectiveness outperforms other traditional methods.

Keywords: cross-domain, sentiment classification, sentiment transfer, opinion mining

Ying Lu, Junping Xie Multi-objective hub location problem in hub-and-spoke network

Computer Modelling & New Technologies 2014 18(7) 309-316

Through observations from the construction of Chinese national emergency material reserve system, we introduce the multi-objective hub location problem. We provide a mathematical model for finding the optimal hub locations to minimize the total transportation cost and maximize the coverage of the hubs simultaneously in the whole network. Then, a procedure for solving this model is proposed. By using a numerical example, we discuss the efficiency of the tabu-search-based algorithm compared to the complete enumeration method and the impact of cost discount factor on the performance of hub-and-spoke network. The results show that the heuristic algorithm based on tabu search may be better than the complete enumeration research method for big size multi-objective hub location problem and as the cost discount factor is increased, the cost savings in the hub-and-spoke network compared to the direct connect network would decrease while the covering rate remains the same unless the cost discount factor is close to 1. Finally, we set future research directions on the multi-objective hub location problem.

Keywords: hub location problem, multi-objective programming, hub-and-spoke network, tabu search

Ilna Ter-Saakova, Nataly Podolyakina Analysis of necessary investments in the production and warranty service of innovative products considering the necessity of their backup

Computer Modelling & New Technologies 2014 18(7) 317-321

Redundancy is one of the commonly used methods to improve the reliability of industrial products and is used in various designs. Another way to increase the reliability is to use more reliable components during the production. This work provides a feasibility study of the redundancy during the manufacturing as well as a comparative analysis of the conditions under, which one or another methods are chosen to improve the reliability of a product as a function of its value.

Keywords: costs structure, warranty service, reliability level, probability of no-failure operation of products

Xing Yu, Guohua Chen The continuous-time optimal portfolio using a multivariate normal inverse Gaussian model

Computer Modelling & New Technologies 2014 18(7) 322-324

This paper develops the continuous-time portfolio model using a multivariate normal inverse Gaussian model. Though the weighted average of lognormal variables is no longer lognormal, it can be approximated by other distributions, such as a multivariate normal inverse Gaussian model. Our method belongs to the analytic approximation class. By comparing to Monte Carlo experiments, it illustrates the computational efficiency and accuracy of our approach.

Keywords: Continuous-time portfolio, Normal inverse Gaussian, Approximation, Monte Carlo, Optimization

Peng Ma Computer information technology and agricultural logistics management system

Computer Modelling & New Technologies 2014 18(7) 325-329

At present, there are kinds of problems on circulation pattern of agricultural products' supply chain, leading to high cost of agricultural logistics system and unreasonable planning. In order to solve the problem of agricultural product circulation pattern, we need to put forward a circulation pattern of agricultural products' supply chain, which takes the agricultural product logistics as a core enterprise. This thesis introduces the idea of combination between computer information technology and logistics management system. It also analyses how to better complete the modules of logistics management system and key points of them, based on the technology of computer, automation, bar code, etc., which focus on analysing the module of farm, customer relations and decision.

Keywords: supply chain; logistics management; circulation pattern; information technology; module

NATURE PHENOMENA AND INNOVATIVE ENGINEERING

Junmei Zhao, Zhijie Zhang, Yifeng Ren Research on speed regulation system for matrix converter fed induction motor

Computer Modelling & New Technologies 2014 18(7) 330-336

This study presents the application of a Matrix Converter (MC) and an active disturbance rejection controller (ADRC) to Direct Torque Control (DTC) system based on an induction motor. Matrix Converter (MC) is applied to Direct Torque Control (DTC) system based on an induction motor in order to reduce power grid harmonic pollution which is caused by AC-DC-AC converter in conventional DTC system. Then a PID controller and an ADR controller are both designed to regulate the speed of the system. Design procedures for ADRC are given in detail. Finally, corresponding results are compared. The simulation results show that the novel DTC system has combined the advantages of both MC and DTC--stable running, strong anti-jamming, good dynamic and static performance.

Keywords: DTC, MC, ADR controller, space vector, PI controller

Xiaohui Liu, Feng Dai, Jianfeng Liu Research on the anisotropy of the coal rock under different bedding direction

Computer Modelling & New Technologies 2014 18(7) 337-343

The Fu Rong mining area was selected to analyse the micro view characteristics of the coal, the ultrasonic acoustic characteristics and the uniaxial compression feature from different directions (the parallel direction and the vertical direction). The results show that: (1) Coal rock has large discreteness with strong anisotropic properties. (2) The impulse wave velocities have obvious anisotropic characteristics. The parallel and vertical wave velocities are different. The parallel bedding velocity is greater than the vertical of coal rock no matter the longitudinal wave or transverse wave of coal rock. (3) The uniaxial compressive strength of parallel bedding coal rock is less than the vertical bedding of coal rock. The uniaxial compressive strength is normally distributed vertical wave velocity, obeying exponential functions or power functions. (4) Failure pattern of the coal rock in the parallel bedding direction is splitting, while the vertical bedding direction of coal rock is shearing. The uniaxial compression strength and deformation parameters in two directions are obviously different. In other words, the anisotropic is apparent.

Keywords: coal rock, bedding, anisotropy, ultrasonic velocity, uniaxial compression test

Yanli Feng, Dashe Li, Shue Liu Research on the laser transmission simulation based on random phase screen in atmospheric turbulent channel

Computer Modelling & New Technologies 2014 18(7) 344-349

On the basis of collimated Gaussian beams, the paper focused on the modelling and simulating of the transmission of laser beams using two-dimension random phase screens in the atmospheric turbulence channel. Firstly, with the analysis of the transmission model of Gaussian beams through the phase screens, the simulation theory of random phase screens and the depth range model of the phase screens were proposed. Then, In accordance with Kolmogorov

atmospheric turbulence theories, a two-dimension random phase screen was built using Fourier transform. Numerical simulation experiments were conducted with low frequency compensation to simulate the propagation of Gaussian collimated beam in Kolmogorov turbulence. Finally, the two-dimension random phase screen was testified by the phase structure function. The results showed that the approach of simulating the random phase screen using Fourier transform was appropriate after compensating the low frequency.

Keywords: random phase screen, atmospheric turbulence, Gaussian beam, Fourier transform, Kolmogorov

Zhao Han, Yuan Rao, Wentao Chen, Junkai Huang The design of a dynamic slope compensation circuit for boost DC-DC converter

Computer Modelling & New Technologies 2014 18(7) 350-355

This paper proposes a structure of peak current mode Boost DC-DC converter with slope compensation circuit, and designs a dynamic slope compensation circuit applied to this converter. With the utilization of the voltage controlled resistance characteristics of MOS transistor and the introduction of a clamp circuit consist of cascade current mirror, a dynamic slope compensation circuit is realized. The circuit is simulated on Cadence Spectre using SMIC 0.18 μ m CMOS technology. Results show that it can provide proper slope compensation following the variation of input and output. The load capacity of DC-DC converter reaches 550mA and the transient response lows to 10 μ s. By eliminating the problem of instability caused by the peak current mode switching power supply of double loop control, the design improves the stability of switching power supply.

Keywords: boost DC-DC, slope compensation, current mirror, voltage controlled resistor

Xingjie Chen, Xiaodong Chai, Xining Cao The time-frequency analysis of the train Axle box acceleration signals using empirical mode decomposition

Computer Modelling & New Technologies 2014 18(7) 356-360

Rail defects usually result in lots of problems such as affecting the comfort of passengers, increasing the wheel-rail forces, exacerbating the train axle boxes vibration and track wear, even threatening the safe operation of trains. In this paper, the characteristic frequency distribution of the changing axle box acceleration caused by defects is analysed by empirical mode decomposition and Hilbert-Huang Transform is used to analyse the time-frequency changes of axle box acceleration. As a result, rail defects can be effectively positioned and the short wave irregularities within a certain degree can be detected. The research provides timely protection for the maintenance of the track.

Keywords: track detection, empirical mode decomposition, time-frequency analysis, Axle box acceleration

Xianfeng Zheng, Zheng Fan Research into voltage sag online detection technology based on wavelet tree

Computer Modelling & New Technologies 2014 18(7) 361-367

A general process model is established using the real-time requirements of data stream processing, and the data is constantly processed with a sliding window. This paper selects the recursion-based complex wavelet as the detecting algorithm for voltage sag, and tries to detect when the voltage sag occurs and ends with amplitude and phase information contained in the wavelet analysis results. Meanwhile, this paper seeks to improve the precision of detection by looking for optimal wavelet scales with information entropy. The shifted wavelet tree-based data flow anomaly detection algorithm and data update method of shifted wavelet tree have been improved to make rapid detection possible. Finally, this paper reports the experimental simulation which proved the instantaneity and accuracy of this method.

Keywords: Data Stream; Voltage Sag; Recursive Complex Wavelet Transform; Shifted Wavelet Tree

Chang Chen, Guojin Chen, Shaohui Su, Haiqiang Liu Modelling and simulation of marine rudder system in a unified M&S platform

Computer Modelling & New Technologies 2014 18(7) 368-372

For modelling and simulating of marine rudder system, there are lots of along with their model libraries, such as AMESim could be used. But the models in these tools lack of flexibility and are not open to the end-user. And these

tools could not model the whole marine rudder system consisted of mechanical, hydraulic and control sub-system in a unified form. In order to solve those problems, a flexible and extensible marine rudder system library was constructed, based on the Modelica, by the object-oriented strategy. It supports the reuse of knowledge on different granularities: physical phenomenon, component model and system model. A conventional model of marine rudder system was built and calculated using the library, and the results shows that the object-oriented modelling strategy is effective; the framework of the library is reasonable.

Keywords: Marine Rudder System, Modelica, M&S Unified Platform, Object-oriented modelling strategy

Zhangming Peng, Guojin Chen, Shaohui Su Study on quantitative diagnosis method of valve clearance based on cylinder head vibration signal of diesel engine

Computer Modelling & New Technologies 2014 18(7) 373-376

The vibration signal of cylinder head contains abundant performance information of diesel engine, and it is inseparable from injection advance angle and valve timing in time domain, so it is easy to separate the response signal of each exciting force from vibration signal. In this paper, the vibration signals of the exhaust valve closing were cut out by extract time interval sampling, and the energy information of feature frequency range was extracted by HHT transform, the corresponding relationship between valve clearance and energy information was established after normalization, so that it is realized to quantitatively diagnose the valve clearance.

Keywords: cylinder head vibration, diesel engine valve, extract time interval sampling, quantitative diagnosis

Jian-Gang Dong, Feng Zhang, Yong-Heng Zhang A water quality changing prediction model for agricultural water-saving irrigation based on PSO-LSSVR

Computer Modelling & New Technologies 2014 18(7) 377-381

In order to improve the prediction of early warning and agriculture information processing level of water quality for agricultural water-saving irrigation, using mathematics and information theory model to predict and estimate the possibility of future changes in water quality based on getting the quality data by using sensor device. The basic process, model for water quality prediction of agricultural water-saving irrigation, forecasting and early warning method of establishing process is designed. Finally, was using the PSO-LSSVR forecasting method to predict the water quality in the agricultural water-saving irrigation of water quality changes prediction. Simulation results show that the parameters of LSSVR were optimized by PSO algorithm, and overcome the cross validation to determine the influence of subjective factors of LSSVR parameters, has better prediction accuracy and generalization ability, its precision can satisfy the need for intensive irrigation production management.

Keywords: prediction model, PSO-LSSVR, water quality, water-saving irrigation, ZigBee

Jian-Long Ding, Weifang Chen, Ji Gao Intelligent data-collaboration mechanism under the distributed application environment

Computer Modelling & New Technologies 2014 18(7) 382-387

Because of the complexity, the dynamic and uncertainty of the distributed applications environment, the data-collaboration crisis caused by isolated information island is serious day by day. Through the establishment of Data Cooperation-based Virtual Organization (DCVO), is conducive to meet the realistic demand of the on-demand dynamic data collaboration, which led to the distributed application to carry out the intelligent data collaboration in effective control, as well as realize the intelligent data retrieval across application domain. Through research of the Distributed Application System-based Data Cooperation Architecture (DASDA), to straighten out the related technology and method of distributed collaborative, from the semantic specification (including the application of domain ontology, relational databases and ontology mapping mechanism, cooperative data transmission standard), rational of data-collaboration (policy representation and configuration), collaborative service personalization, data structure and model of collaborative content level and so on, to provides an important reference to solve the eliminate problem such as semantic fuzzy, dynamic expansion, uncontrollable, cooperative security, recall and precision of conflict which caused in the process of the data-collaboration.

Keywords: ontology, virtual organization, data collaboration, policy configuration

Yunxia Zhang, Chenglong Dai, Jifeng Cui Lifetime forecasting for hemispherical resonator gyroscope with wavelet analysis-based GM(1,1)

Computer Modelling & New Technologies 2014 18(7) 388-392

Because of high cost and small batch of spacecraft like flywheel and gyroscope, how to estimate their reliability and lifetime becomes a tough task. A method to predict the lifetime of hemispherical resonator gyroscope (HRG) is put forward in this paper. This method utilizes grey correlation and mean absolute percent error (MAPE) to estimate the reliability of predictive data sequence. For reducing noise, Daubechies wavelet is used to decompose and reconstruct the test data in the paper as well. After pre-processing, predictive data sequences are gained by using GM(1,1) prediction model and then according to grey correlation and MAPE of each predictive data sequence, the threshold value meets conditions can be gained. Finally, the lifetime of HRG is predicted with using the threshold value. In this paper, the method is applied to the data of one type of HRG provided by a research institute in China and the result shows the gyroscope can normally run 4780 days at least, namely about 13.10 years.

Keywords: hemispherical resonator gyroscope (HRG), lifetime prediction; wavelet analysis, GM(1,1), MAPE, grey correlation

Xiaodong Huang Automatic license plate detection based on colour gradient map

Computer Modelling & New Technologies 2014 18(7) 393-397

License plate detection plays a key role in traffic surveillance, speeding vehicles ticketing and vehicle detecting, and so on. However, most of the previous approaches to detect license plate experience difficulties in handling license plate with the uneven illuminations changes, complex background or tilted alignments. In this paper, we propose a method of license plate detection. License plate regions contain plate characters, frames and screws. First we propose to build the Colour Gradient Map (CGM) based on the colour gradient method. Then we perform the Niblack's method on the Colour Gradient Map (CGM) to retrieve the candidate license plate regions. Finally, we use the template matching to remove most of background noises. Experimental results show that this approach is robust and can be effectively applied to license plate detection.

Keywords: license plate detection, colour gradient, template matching

Jun Li, Xiaoyu Liu, Shiping Zhao Prediction model of recast layer thickness in die-sinking EDM process on Ti-6Al-4V machining through response surface methodology coupled with least squares support vector machine

Computer Modelling & New Technologies 2014 18(7) 398-405

Ti-6Al-4V is widely applied in frontier for its excellent properties such as a high strength-weight ratio, great heat stability and exceptional corrosion resistance. Electrical discharge machining (EDM) is suitable for machining titanium alloys, because it is the technical that removal materials by discharge energy and non-contact in processing progress. The recast layer is formed by the solidification of molten metal on the machined surface during the EDM process. In the present investigation, a hybrid approach using Least squares support vector machines (LS-SVM) and response surface methodology (RSM) for predication the recast layer thickness is proposed. Experimental plan is performed by response surface method with 20 experimental runs. The different machining parameters of pulse current, pulse on-time, and pulse off-time are selected as input factors. The white layer thickness (WLT) is response variable. The LSSVM method is applied to construct the predication model based on the orthogonal experiment swatches. The randomly 15 experimental runs were utilized to train the LS-SVM model to predict the WLT. Finally, support vector machine is used to compare with the proposed method. The proposed model can be good performance in prediction of white layer thickness of the complex EDM process.

Keywords: Electrical discharge machining, Least squares support vector machines, Response surface methodology, Recast layer

Hao Wu, Dewen Seng, Xujian Fang Construction of a computer simulation platform for optical experiments

Computer Modelling & New Technologies 2014 18(7) 406-412

With the rapid development of computer technology, computer-assisted instruction for complex and vulnerable optical experiments has become possible. Computer simulation technique has become an important branch in computer application and a new means in science research and engineering design. People have done a lot of research on optical experiment simulation. But there are still many defects, such as no friendly graphical user interface and the parameters cannot be freely adjusted. We design and develop a well extensible and portable simulation platform for optical experiments including basic optics experiments, information optics experiments and laser experiments and realize flexibility in setting the experimental parameters. Young's double-slit interference experiment, Fraunhofer diffraction experiment and grating diffraction experiment are conducted to show the effectiveness, efficiency and correctness of our simulation platform. The abstract and difficult optical concepts and rules are vividly manifested through the simulation experiments and become easier to understand for the students. The simulation platform will break through the limitation of teaching space, experimental equipment and various other factors and enable students to preview experiment, understand experiment, complete experiment and review experiment much better.

Keywords: optical experiment, simulation platform, interference and diffraction