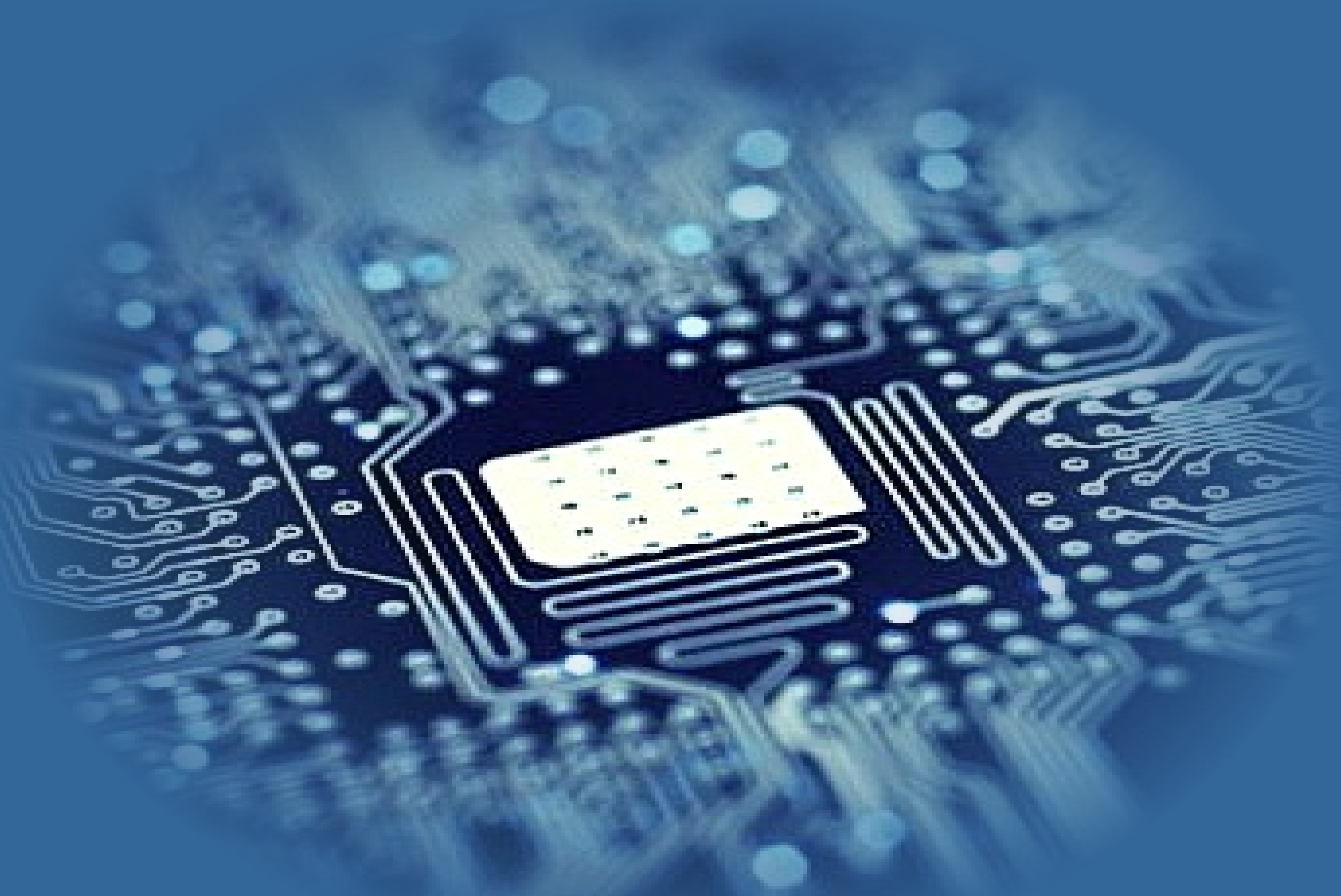
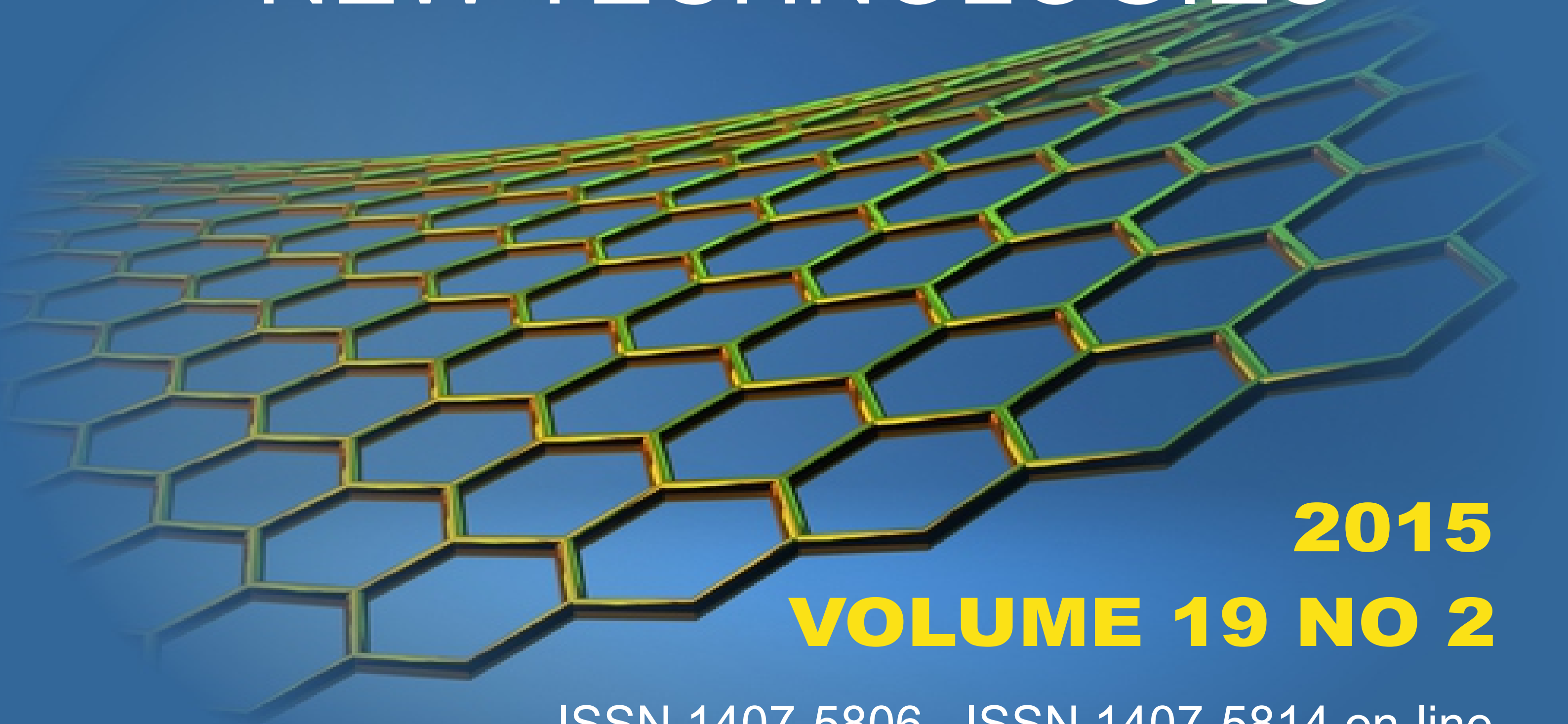


COMPUTER MODELLING AND NEW TECHNOLOGIES



2015

VOLUME 19 NO 2

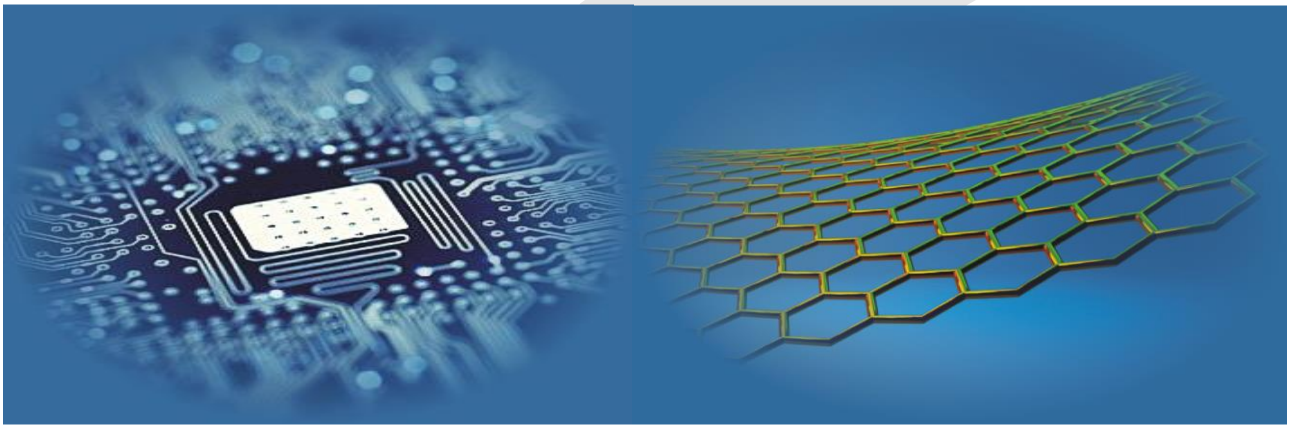
ISSN 1407-5806 ISSN 1407-5814 on-line

Latvian Transport Development and Education Association

Computer Modelling and New Technologies

2015 Volume 19 No 2

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



Riga – 2015

Journal topics:	**Publisher**	**Supporting Organizations**
• mathematical and computer modelling • computer and information technologies • natural and engineering sciences • operation research and decision making • nanoscience and nanotechnologies • innovative education	Latvian Transport Development and Education Association	Latvian Academy of Sciences Latvian Operations Research Society Transport and Telecommunication Institute, Latvia Fraunhofer Institute for Factory Operation and Automation IFF, Germany International IT University, Kazakhstan

Articles should be submitted in **English**. All articles are reviewed.

EDITORIAL CORRESPONDENCE	**COMPUTER MODELLING AND NEW TECHNOLOGIES, 2015, Vol. 19, No.2** **ISSN 1407-5806, ISSN 1407-5814 (on-line: www.cmnt.lv)**
Latvian Transport Development and Education Association	**Scientific and research journal** **The journal is being published since 1996**
68 Graudu, office C105, LV-1058 Riga, Latvia **Phone: (+371) 29411640** E-mail: yu_shunin@inbox.lv http://www.cmnt.lv	The papers published in Journal 'Computer Modelling and New Technologies' are included in: **INSPEC,** www.theiet.org/resources/inspec/ **VINITI,** http://www2.viniti.ru/ **CAS Database** http://www.cas.org/ **EI Compendex**

Editors' Remarks

****

A Moments Indulgence

by Rabindranath Tagore

I ask for a moment's indulgence to sit by
thy side.

The works that I have in hand I will finish
afterwards.

Away from the sight of thy face my heart
knows no rest nor respite,

and my work becomes an endless toil in a
shoreless sea of toil.

Today the summer has come at my
window with its sighs and murmurs;

And the bees are plying their minstrelsy
at the court of the flowering grove.

Now it is time to sit quite, face to face
with thee, and to sing

dedication of life in this silent and
overflowing leisure.

Rabindranath Tagore (1861-1941)[♣]

****

This 19th volume No.2 consists of four topical parts, namely, **Part A: Mathematical and Computer Modelling, Part B: Computer and Information Technologies, Part C: Operation Research and Decision Making and Part D: Nature Phenomena and Innovative Engineering.** These parts have a particular page numbering. References should include the symbols belonging to the part of the journal issue (A, B, C or D) and the pages of the paper quoted. (e.g.: ... **19**(2C) 77-89) We are planning to expand CMNT topics within the scope of its scientific interests.

Our journal policy is directed to fundamental and applied scientific researches, innovative technologies and industry, which is the fundamentals of the full-scale multi-disciplinary modelling and simulation. This edition is the continuation of our publishing activities. We hope our journal will be of interest for research community and professionals. We are open for collaboration both in the research field and publishing. We hope that the journal's contributors will consider collaboration with the Editorial Board as useful and constructive.

EDITORS

Yuri Shunin

Igor Kabashkin

[♣] **Rabindranath Tagore (7 May 1861 – 7 August 1941),** was a Bengali poet, novelist, musician, painter and playwright who reshaped Bengali literature and music. As author of Gitanjali with its "profoundly sensitive, fresh and beautiful verse", he was the first non-European and the only Indian to be awarded the Nobel Prize for Literature in 1913. His poetry in translation was viewed as spiritual, and this together with his mesmerizing persona gave him a prophet-like aura in the west. His "elegant prose and magical poetry" still remain largely unknown outside the confines of Bengal.

Content A

PART A **Mathematical and Computer Modelling**		
Yonglong Tang	Fuzzy knowledge searching on the basis of the traditional and-or graph search algorithm	**7**
Yuwan Gu, Dongyu Zhang, Yan Chen, Bixia Chao, Yuqiang Sun	Parallel computation of matrix norm based on MapReduce	**13**
Authors' Index		**17**
Cumulative Index		**18**

Fuzzy knowledge searching on the basis of the traditional and-or graph search algorithm

Yonglong Tang

School of mathematics and Statistics Jishou University, China

Corresponding author's e-mail: 11755600050@qq.com

Received 1 March 2014, www.cmnt.lv

Abstract

Based on the fuzzy propositional logic FLCOM and fuzzy set FSCOM, we research the formal denotation, inference and computation of fuzzy knowledge. We extend the fuzzy and-or graph, turn the propositional formulas as state nodes, express the logical rules as the search space, construct and-or graph of the fuzzy propositional formula. We modify heuristic function on the basis of the traditional and-or graph search algorithm, and give out a method to process negation information in the process of reasoning, transforming the fuzzy knowledge reasoning into the state space searching problem, and using the state space searching to solve the problem of fuzzy knowledge reasoning.

Keywords: fuzzy propositional logic FLCOM, fuzzy set FSCOM, fuzzy propositional formula, negation information, state space searching

1 Introduction

With the development of information science and the further research of non classical logic, in 80's and 90's of twentieth Century, research on treatment of domestic and foreign scholars on the negative information in the field of information processing begins with new ideas and methods. Among them, the domestic famous scholar Zhu Wujia and Xiao Xian are the representative, the work of the famous scholar Gerd Wager is the most representative [1-3].

Zhu Wujia and Xiao Xian founded the formal logic system of medium logic theory in 1985 [4-5]. The negative relations in concept are divided into two types by the medium logic: contradiction and opposition, and certainly the transition state intermediate between the opposite concept, namely intermediary state [6].

In 1991, Wagner G first proposed that the database needs two kinds of negation to deal with partial information [7], and in 1994 two kinds of partial logics distinguished methods were given, and studied the two negative knowledge reasonings. Local logic mainly distinguish denotative negative information of the knowledge information: strong negation and weak negation. Local logical negation introduced two types of theories from the part of the model to represent the deletion, and processing information explicitly rejected and pseudo information [8].

In 2005, Kaneiwa K gave an extensive description logic $ALC_{\_}$, and $ALC_{\_}$ with classical negation and strong negation. Kaneiwa K believed that contradiction, opposition parties and subcontrary available predicate negation (e.g. Not Healthy) negation and the verb (such as Unhealthy) to distinguish from a possible sentence type [9]. He would improve the semantic predicate negation and negation of predicates, properly explain the combination of various classical negation and strong negation, and proved that the opposition and contradiction of $ALC_{\_}$ semantic concept were improved, and the characteristic was not all types of description logic owning. For example, building description logic ($CALC_{\_}$) with the construction of Heyting negation and strong negation would not be able to maintain this property [10-12].

In 2006, Ferré S presented a logic transformation based on modal logic AIK, and the transformation was in the concept of logic analysis LCA. This essential feature of the logic transformation is that LCA will not lose general, and distinguishes three kinds of relations in the only formal: negation, opposition and possibility [13].

In 1987, Pan Zhenghua gave a semantic interpretation of three value of medium logic [18], and proved in the interpretation that medium propositional logic (MP), medium predicate logic (MF) as well as the completeness and reliability of the extended medium propositional logic (MP^*) [14]. In 2003, Pan Zhenghua proposed and insisted that the medium logic is a kind of infinite valued logic [15], and in the infinite value semantic model of medium logic he proved the reliability and completeness of the medium proposition logic [16]. In 2007, Pan Zhenghua distinguished five different negative relationships in the conceptual level of clear knowledge and fuzzy knowledge [17]. In 2010, Pan Zhenghua gave the fuzzy set FSCOM with opposite negation, intermediary negation and contradiction negation, and different negative relation of fuzzy concepts from the set point of view characterizations was portrayed [18]. In 2012, Pan Zhenghua put forward a fuzzy propositional logic system FLCOM to differentiate contradiction, intermediary negative and opposite negation.

We distinguish three kinds of negative fuzzy logic FLCOM based on fuzzy knowledge representation to represent the state space through the reasonable extension of fuzzy and-or graph, and a three negative searching method of fuzzy knowledge reasoning method was given in the case of the FSCOM to represent and process knowledge [19-21].

2 Fuzzy knowledge representation based on FLCOM

The knowledge representation of the fuzzy logic adopts the fuzzy proposition of the fuzzy logic and 'and ($\wedge$)', 'or ($\vee$)', 'non ($\neg$)', 'implication ($\rightarrow$)', 'equivalence ($\leftrightarrow$)' logical connectives showing [22]. FLCOM on the basis of the introduction of intermediary negative word "~" and opposite negation word "϶", and " $\neg$ " represents the contradiction to expand the traditional logic expressions, and they will be a combination of wff more complex, in order to express the concept of facts more complex.

A fuzzy production rule with the general P ← Q, CF, τ representation. Conclusion and the premise of P, Q denote respectively, the truth degree is expressed in a fuzzy way, and CF (0<CF ≤ 1) is the confidence of the rule, and τ (0< τ ≤ 1) is a threshold [23].

3 And / or graph representation of extended fuzzy propositional formula

Definition 1 Representation of fuzzy rules:

(1) If a plurality has the same conclusion rules, and it can also be activated to perform, that is:

$$P \leftarrow Q_1, \quad CF_1$$

$$P \leftarrow Q_2, \quad CF_2$$

$$\cdots$$

$$P \leftarrow Q_k, \quad CF_k$$

Use and-or graph representation, and Q_1, Q_2,... Q_k can be regarded as node P or its parent node as shown in Figure 1.

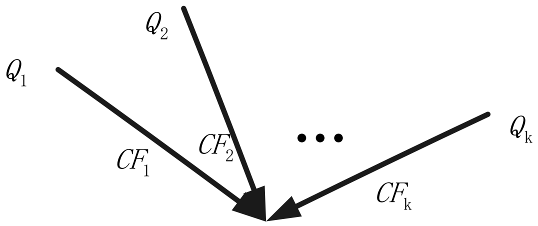


FIGURE 1 And-or graph of node P and its parent node set

(2) A condition to represent the rules can be different degrees of activation of several results, that is:

$$P_1 \leftarrow Q, \quad CF_1$$

$$P_2 \leftarrow Q, \quad CF_2$$

$$\cdots$$

$$P_k \leftarrow Q_k, \quad CF_k$$

Use and-or graph representation, and P_1, P_2,... P_k is Q or sub node set as shown in Figure 2.

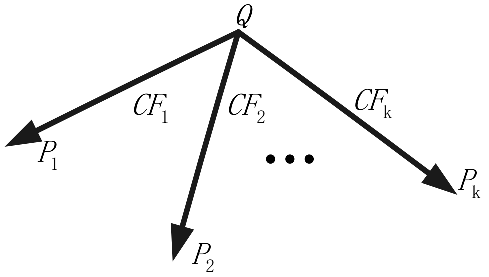


FIGURE 2 And-or graph of node Q and sub node set

(3) For the multidimensional fuzzy such as P ← Q_1, Q_2,... Q_k, CF, and-or graph representation, namely Q_1, Q_2,... Q_k, P and the father node set as shown in Figure 3.

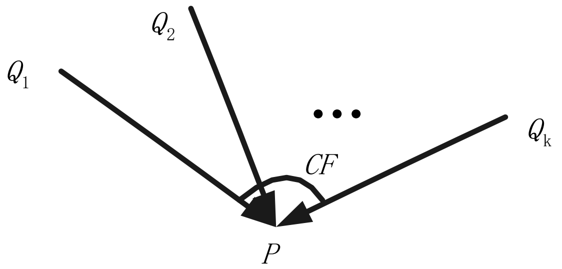


FIGURE 3 And-or graph of node P and the parent node set

(4) If the condition to represent the rules can be the same degree of activation of multiple objective, namely P_1, P_2,... P_k ← Q, CF, and-or graph representation, then P_1, P_2,... P_k are Q and the sub node set as shown in Figure 4.

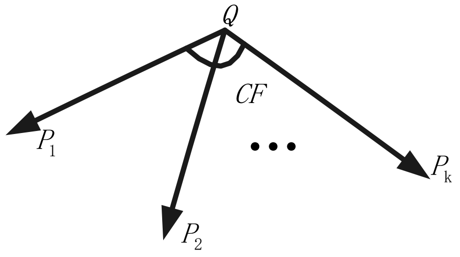


FIGURE 4 And-or graph of node Q and son sets

As shown in Figure 1, Figure 2, Figure 3 and Figure 4, in describing the fuzzy knowledge, with a correlation of arc to represent the relationship between parent and child nodes, and this relationship is the confidence of the rule. Figure 3 and Figure 4 from node P, Q are issued by the arc by a curve together and represents and node. P, Q nodes in Figure 1 and Figure 2 express or node [24].

4 State space searching of fuzzy knowledge reasoning

4.1 ALGORITHM ANALYSIS

To improve the search efficiency of the problem space needs a lot of support strategies, and these strategies are to solve problems and solutions related control knowledge, and the evaluation function reflects the control information. The general form of evaluation function is: $f(n) = g(n) + h(n)$, and $h(n)$ is the estimating cost from n to the target node, and $g(n)$ is the actual cost from the initial node to n [25]. After given a problem function definition method can have many kinds of means according to the characteristics and problems. In the fuzzy knowledge reasoning, what we want is the most likely out come, but in the traditional state space search process is looking for the shortest path to the target node. Therefore, in the search process of fuzzy knowledge reasoning can be to find the shortest path problem to find comprehensive credibility of the highest path, and it can be integrated with the reliability function to define the evaluation function.

Zadeh operators $a \wedge b = \min(a,b)$ in t- van of fuzzy 'and' operators in the fuzzy reasoning define evaluation function:

Definition 2 [45] $f(n) = g(n) \wedge h(n) = \min(g(n), h(n))$, of which, $h(n)$ is to estimate the credible degree from node n to the target, and $g(n)$ is the credibility from the initial node to node n. $h(n)$ depends on the confidence of the target node, and

$g(n)$ depends on the relational degree of node n and its parent node (set) of the arc and the confidence of the parent node (set).

In the actual search, the confidence of the initial node (set) is expressed degree of membership with initial conditions, and the correlation of the arc is the confidence of rules of inference. In fuzzy reasoning, in addition to the confidence of the inference rules, fuzzy reasoning algorithm determines the degree of confidence of the conclusion, and this paper uses CRI algorithm of Zadeh to determine the confidence degree of sub nodes in the search process [26].

If P, Q respectively are the child nodes and parent nodes, and the correlation degree of the arc of P, Q is CF (0<CF ⩽ 1) (i.e. rule confidence of P ← Q), then $\Psi(P) = \Psi(Q) \wedge CF$, $\Psi(P)$ and $\Psi(Q)$ respectively are the confidence of node P and node Q . In particular, for or parent node point set $Q = \{Q_1, Q_2, \cdots, Q_k\}$, $\psi(P) = \overset{k}{\underset{i=1}{\vee}} (\Psi(Q_i) \wedge CF_i)$; and the father node set $Q = \{Q_1, Q_2, \cdots, Q_k\}$, $\psi(P) = \overset{k}{\underset{i=1}{\vee}} (\Psi(Q_i) \wedge CF_i)$

Here are the fuzzy search process of reasoning.

Procedure fuzzy search

Begin

Let S represent the initial state of the node INITIAL, mark begins with an arc of S, and calculate $f(S)$.

Repeat

(1) Track this arc, and choose one on the path on the future expansion of node expansion, when expansion, if you need one or several nodes participate at the same time(i.e. with the parent node set), these nodes are added to the diagram, and the selection of the new node (node set) called NS, successor node generate NS [27].

If NS has a successor node labeled succeed, successor node for each not NS ancestor

Do Begin

Succeed is added to the chart.

If succeed has no successor node, then labeled SOLVED, the node of the confidence value is f (succeed) value.

Else calculate f (succeed).

Else Begin

f (NS) values are for the confidence of NS, and NS is labeled SOLVED.

End

The latest information is returned, node set D (initial contains only the node S), D include the marker SOLVED node, and the value of f has been the need to change back to the ancestor node.

(2) Repeat

Choose a node CS from D removed from D, and their descendants are not in D. Calculate the confidence each sub node CS, and the maximum degree of confidence in all of the sub node calculated, namely f (CS) value. Arc to the previous step confidence largest sub nodes from the start, marked as the best path begins with CS.

The best path If labeled and CS connected nodes with SOLVED markers.

Then CS is labeled SOLVED.

Else CS is marked as SOLVED or f (CS) has changed

Then return to its new status. All the ancestor of CS are added to D.

Untie D is empty;

Until S marked SOLVED succeed, else, f (S) is less than μ , failure;

End

4.2 TREATMENT OF NEGATION KNOWLEDGE

To solve the negative contains nodes in the search process. The treatment includes contrasting negation, intermediary negation and contradiction negation [28].

Each fuzzy set has its membership, we set each node of the membership degree as the node degree of confidence, in reality, the negative situation, do the following treatment: let A be the domain U of FSCOM set, for the PSI lambda $\psi_\lambda(A(x))$ for the membership of A, abbreviated $A_\lambda(x), \lambda \in \left(\frac{1}{2}, 1 \right]$

(1) A opposition $A^{\ni}$ membership psi lambda $\psi_\lambda(A^{\ni}(x)) = 1 - A_\lambda(x)$, abbreviated $A^{\ni}(x)$

(2) A mediated negative membership psi lambda set $A^{\sim}$ $\psi_\lambda(A^{\sim}(x))$, abbreviated as $A_\lambda^{\sim}(x)$

$$A_\lambda^{\sim}(x) = \begin{cases} \dfrac{2\lambda - 1}{\lambda - 1} A_\lambda(x) + \lambda & A_\lambda(x) \in [0, 1-\lambda) \\ \dfrac{2 - 2\lambda}{1 - 2\lambda} (1 - \lambda - A_\lambda(x)) + \lambda & A_\lambda(x) \in [1-\lambda, \frac{1}{2}) \\ \dfrac{2 - 2\lambda}{1 - 2\lambda} (A_\lambda(x) - \lambda) + \lambda & A_\lambda(x) \in [\frac{1}{2}, \lambda) \\ \dfrac{2\lambda - 1}{\lambda - 1} A_\lambda(x) + \lambda & A_\lambda(x) \in [\lambda, 1) \\ \dfrac{1}{2} & A_\lambda(x) = \dfrac{1}{2} \end{cases}$$

(3) $A^{\neg}$ contradiction membership psi lambda set $\psi_\lambda(A^{\neg}(x))$, abbreviated as $A_\lambda^{\neg}(x) = Max(A_\lambda^{\ni})(x), A_\lambda^{\sim}(x))$

The membership is determined according to improved FLCOM infinite value semantic interpretation. Discuss the fuzzy decision-making in the lambda value in [29] according to the value of significance, and λ value is as the threshold, and in practice the general threshold λ is greater than 0.5.

(a) when $A_\lambda(x) \in [0, 1-\lambda)$, x does not belong to A, and $A_\lambda^{\ni}(x) \in (\lambda, 1]$, x fully belongs to $A^{\ni}$, $A_\lambda^{\sim}(x) \in (1-\lambda, \lambda)$, $A_\lambda^{\sim}(x)$ decreases with $A_\lambda(x)$ increasing, x partly belongs to $A^{\sim}$.

(b) When $A_\lambda(x) \in \left[1-\lambda, \frac{1}{2}\right)$, $A_\lambda^{\ni}(x) \in \left(\frac{1}{2}, \lambda\right]$, x partly belongs to A and $A^{\ni}$, $A_\lambda^{\sim}(x) \in (\lambda, 1)$,

$A_\lambda^{\sim}(x)$ increases with $A_\lambda(x)$ increasing, x partly belongs to $A^{\sim}$.

(c) When $A_\lambda(x) \in \left(\frac{1}{2}, \lambda\right]$, $A_\lambda^{\ni}(x) \in \left[1-\lambda, \frac{1}{2}\right)$, x partly belongs to A and $A^{\ni}$, $A_\lambda^{\sim}(x) \in (0, 1-\lambda)$, $A_\lambda^{\sim}(x)$ increases with $A_\lambda(x)$ increases, x partly belongs to $A^{\sim}$.

(d) When $A_\lambda(x) \in (\lambda, 1]$, x fully belongs to A, $A_\lambda^{\ni}(x) \in [0, 1-\lambda)$, x does not belong to $A^{\ni}$, $A_\lambda^{\sim}(x) \in (1-\lambda, \lambda)$, $A_\lambda^{\sim}(x)$ decreases with $A_\lambda(x)$

increasing, x partly belongs to $A^{\sim}$.

(f) When $A_\lambda(x)=\frac{1}{2}$, $A_\lambda^{\ni}(x)=\frac{1}{2}$, $A_\lambda^{\sim}(x)=\frac{1}{2}$, x partly belongs to $A^{\sim}$, $A^{\ni}$ and $A^{\sim}$.

The above method is the membership of objects in practical application for opposition set, intermediary negative set and contradiction set, and the set is selected as the basis, sums up the membership degree and the threshold according to a domain object, by the above calculation method obtained membership this object belongs to its different negative set, the the method of negative information processing in the search process [30-31].

5 Application examples

Below a group of information composed of a number of rules:

(1) The middle-aged and elderly people like foods high in sodium leading to arteriosclerosis, and it has the high reliability.

(2) Eat too much sodium food leading to arteriosclerosis, and overweight people have high rates of hypertension, and it has the high reliability.

(3) Young people without family history of hypertension and the proper weight is not prone to high blood pressure.

(4) Ms. Liu is 40 years old with no family history of hypertension, it is absolutely reliable.

(5) Ms. Liu is about 1.64m and her weight is about 70kg.

5.1 FSCOM INTRODUCTION

In practical reasoning we represent the fuzzy knowledge by fuzzy set FSCOM.

The domain belongs to all people for any x belonging to the domain. Obviously 'the elderly', 'youth', 'obese' and 'moderate weight' are the fuzzy set, and the relationship among 'the elderly' and 'young people' in the concept of 'adult' 'contradiction, therefore, if the fuzzy set 'youth' expresses with YOUNG, then fuzzy set 'the elderly' is represented by $\mathrm{YOUNG}^{\neg}$, and $\mathrm{YOUNG}^{\sim}$ expresses fuzzy set 'middle-aged people', and $\mathrm{YOUNG}^{\ni}$ expresses fuzzy set 'the elderly'. $\mathrm{YOUNG}^{\sim}(x)$, $\mathrm{YOUNG}^{\ni}(x)$ and $\mathrm{YOUNG}^{\neg}(x)$ respectively express the membership for the corresponding fuzzy set x. Similarly, if the fuzzy set 'obese' is expressed using FAT, then the fuzzy set 'moderate weight' is expressed using $FAT^{\sim}$. FAT (x) and $FAT^{\sim}(x)$ respectively express the membership for the corresponding fuzzy set x. MUCHNa expresses eating foods high in sodium. ARTERIOSCLEROSIS(x) expresses x arteriosclerosis. HYPERTENSION (x) will have high blood pressure x, $\mathrm{HYPERTENSI\ ON}^{\ni}(x)$ said that it would not have high blood pressure [32].

Confidence of the above rules according to the actual situation can be given that credibility language assignment. Such as 'high reliability' can give the confidence value 0.7, and 'very big credibility' can gives the confidence value 0.85, and 'easy' gives the confidence value 0.65, and 'absolute confidence' gives the confidence value 1.

Based on fuzzy set FSCOM and the fuzzy production rules, the above rules are represented as:

(1) MUCHNa(x) ← $\mathrm{YOUNG}^{\neg}(x)$, CF=0.7;

ARTERIOSCLEROSIS(x) ← $\mathrm{YOUNG}^{\neg}(x)$, CF=0.7.

HYPERTENSION(x) ← MUCHNa(x) ∧ ARTERIOSCLEROSIS(x) ∧ FAT(x),

CF=0.85

HYPERTENSION $^{\ni}$ (x) ← YOUNG(x) ∧ FAMILYHISTORY $^{\ni}$ (x) ∧ FAT $^{\neg}$ (x),

CF=0.65。

(4) AGE(Liu, 50),FAMILYHISTORY $^{\ni}$ (Liu),CF=1

(5) HEIGHT(Liu, 164),WEIGHT(Liu, 65),CF=1.

5.2 STATE SPACE REPRESENTATION AND SEARCH

In the example 40-year-old Ms Liu belongs to 'the elderly' involves determining the degree of membership, for example, and 'the youth' generally refers to the year from 18 years old to 30 years old, and 'the old' generally refers to the year after the age of 60, then Ms Liu belongs to the membership of 'young people': $YOUNG(Liu)=\frac{d(40,60)}{d(30,60)}\approx 0.67$ belongs to the membership of 'the old' $YOUNG^{\ni}(Liu)=1-YOUNG(Liu)\approx 0.33$. And Ms Liu belonging to the membership of 'the middle-aged' refers to thee stablishment of a specific threshold, see article [33], here the assumption that λ =0.8, get

$$YOUNG^{\sim}(Liu)=\frac{2-2\lambda}{1-2\lambda}(YOUNG(Liu)-\lambda)+\lambda\approx 0.887$$

by using $\frac{2-2\lambda}{1-2\lambda}(A_\lambda(x)-\lambda)+\lambda \quad (A_\lambda(x)\in\left(1/2,\lambda\right])$, so Ms Liu belongs to the membership $\psi^{\neg}(YOUNG(Liu))=Max(YOUNG^{\sim}(Liu))$, $YOUNG^{\in}(Liu)=0.887$ of 'the elderly'. Ms. Liu belongs to 'overweight' or 'moderate weight' according to Ms. Liu's body mass index: body weight (kg) / height (m^2) to establish. Body mass index into Ms. Liu calculation is about 24.17, and it is generally believed that the 'weight' of the body mass index of less than 18, and 'obese' refers to the body of prime number greater than 28. Therefore, Ms. Liu belongs to the membership $FAT(Liu)=\frac{d(18,24.17)}{d(18,28)}=0.617$ of 'fat obese', still assume that $\lambda=0.8$, using the formula $\frac{2-2\lambda}{1-2\lambda}(A_\lambda(x)-\lambda)+\lambda \quad (A_\lambda(x)\in\left(1/2,\lambda\right])$ to obtain

$$FAT^{\sim}(Liu)=\frac{2-2\lambda}{1-2\lambda}(FAT(Liu)-\lambda)+\lambda=0.922.$$

The search graph of the state space is shown in Figure 5:

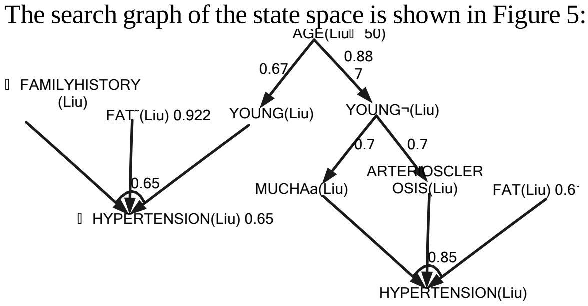


FIGURE 5 Search graph of state space about the instances

According to the algorithm and combined with Figure 5, look for the maximum path of general reliability, and the search path is as followings:

(1) AGE (Liu, 50) is the only node, and the end of the path has the highest credibility at present, and the confidence level is 1.

(2) Extended AGE (Liu, 50) gets or sub node YOUNG (Liu) and YOUNG (Liu), and the confidence is respectively 0.67 (1 ^ 0.67) and 0.887 (1 ^ 0.887), and YOUNG¬ (Liu) node has the max confidence, so it is the most authentic path. At this time, AGE (Liu, 50) is estimated at 0.887.

(3) Extended node YOUNG¬ (Liu) gets or sub node MUCHNa(Liu), ARTERIOSCLEROSIS (Liu), and the confidence is 0.7, and the estimated value of YOUNG¬ (Liu) is 0.7, and the estimated value of AGE (Liu, 50) is 0.7, too. This path is still the most reliable path.

(4) Select the node MUCHNa (Liu) extension to get sub node HYPERTENSION (Liu) obtained by three and father nodes: MUCHNa (Liu), ARTERIOSCLEROSIS (Liu) and FAT (Liu), and the confidences respectively are 0.7, 0.7 and 0.617, and the related degree of the arc is 0.85, so the confidence of HYPERTENSION (Liu) is 0.617 (0.7 ∧ 0.7 ∧ 0.617 ∧ 0.85), so the estimation of YOUNG¬ (Liu) is 0.617, and this path is no longer the most reliable path.

(5) Therefore, the estimated value of YOUNG¬ (Liu) is amended as 0.7, and select the extension node ARTERIOSCLEROSIS (Liu), at the same time the confidence degree of HYPERTENSION (Liu) is also is 0.617, then the estimated value of YOUNG¬ (Liu) is 0.617 to get the estimated value of AGE (Liu,50) for 0.617, and this path is not the most reliable path.

(6) Therefore, the estimated value of AGE (Liu, 50) is amended as 0.67, and extended node is YOUNG (Liu), and the path is the most authentic path at present. The obtained child node ∋ HYPERTENSION(Liu) has three and the parent nodes: ¬ FAMILYHISTORY(Liu), FAT˜ (Liu) and YOUNG (Liu), and the confidences respectively 1, 0.922 and 0.67, ant the the association degree of the arc is 0.65, so the obtained confidence ∋HYPERTENSION (Liu) is 0.65 (1 ^ 0.922 ^ 0.67 ^0.65). This path is the most trusted path, so Ms. Liu will not have high blood pressure [34].

6 Conclusion

This paper describes the fuzzy knowledge based on fuzzy logic FLCOM, and the fuzzy propositional formula is looked as the state nodes to expand the fuzzy and-or graph, and the propositional calculus for reasoning about state space are described to give the method of negative information processing in a fuzzy information. Use the state space search to solve the problem of fuzzy knowledge reasoning.

References

[1] Zadeh L A 1965 Fuzzy Sets *Information and Contral* **8**(3) 338-53

[2] Zadeh L A 1975 The Concept of Linguistic variable and it's Apllication to Approximate reasoning *Information Sciences* **8** 199-249

[3] Atanassov K 1986 Intuitionistic fuzzy sets *Fuzzy Set s and Systems* **20**(1) 87-96

[4] Elkan C 1994 The paradoxical success of fuzzy logic *IEEE Expert* **9**(4) 1-14

[5] Elkan C 1994 The paradoxical controversy over fuzzy logic *IEEE Expert* **9**(4) 47-9

[6] Watkin F A 1995 False Controversy: Fuzzy and non-fuzzy faux pas *IEEE Expert* **10**(4) 4-5

[7] Wagner G 2003 Web Rules Need Two Kinds of Negation *In: F Bry, N Henze, and J Maluszynski,eds. Proc of the 1stinternational workshop on Practice of Semantic Web Reasoning. Heidelberg: Springer Verlag* 33-50

[8] Wujia Zhu, Xi'an Xiao 1988 On the Naive Mathematical Models of Mathematical System *J Math Res& Exposition* **8**(1) 139-47

[9] Wagner G 1991 A Database Needs Two Kinds of Negation *In: B.Thalbeim, J Demetrovics, H-DGerhardt,eds. Proc of 3rd Symposium on Mathematical Fundamentals of Database and Knowledge Bases Systems (MFDBS). Heidelberg: Springer-Verlag* 357-71

[10] Kaneiwa K 2005 Negations in Description Logic-Contraries, Contradictories, and Subcontraries *In: F. Dau, M.-L. Mugnier, and G. Stumme, eds. Contributions to ICCS2005. Kassel: Kassel University Press* 66–79

[11] Kaneiwa K 2007 Description Logics with Contraries, Contradictories and Subcontraries *New Generation Computing* **25**(4) 443-68

[12] Ferré S 2006 Negation, Opposition, and Possibility in Logical Concept Analysis *In: B Ganter and L Kwuida, eds. ICFCA2006,LNAI 3874. Heidelberg: Springer* 130-145

[13] Zhenghua Pan 1987 On the completeness of medium proposition logical calculus system (ML) *Proc of Int Symposium on Fuzzy Systems and knowledge Engineering. Guangzhou, China* 858-60

[14] Zhenghua Pan 1988 Construction of a model of medium logical calculus system (ML) *Proc of Workshop on knowledge-Based Systems and Models of Logical Reasoning,Cairo* 165-75

[15] Zhenghua Pan, Wujia Zhu 2003 A finite and infinite-valued model of the medium proposition logic *Proc of the Second Asian Workshop on Foundations of Software. Nanjing: Southeast University Press* 103-6

[16] Zhenghua Pan, Wujia Zhu 2004 An interpretation of infinite valued for medium proposition logical *Proc of IEEE-Third International Conference on Machine Learning and Cybernetics. Washington: IEEE Computer Society Press* 2495-9

[17] Zhenghua Pan 2013 *Three Kinds of Negation of Fuzzy Knowledge and their Base of Logic* Lecture Notes in Artificial Intelligence7996. Heidelberg: Springer Verlag 83-93

[18] Zhenghua Pan 2013 *Fuzzy Decision Making Based on Fuzzy Propositional Logic with Three Kinds of Negation* Lecture Notes in Computer Science 7995, Heidelberg: Springer Verlag 128-40

[19] Zadeh L A 1965 Fuzzy sets *Information and Control* **8**(3) 338-53

[20] Atanassov K 1986 Intutionistic fuzzy sets *Fuzzy Sets and Systems* **20**(1) 87-96

[21] Pawlak Z 1982 Rough sets *International Journal of Computer and Information Sciences* **11** 341-56

[22] Hájek P 1998 *Metamathematics of Fuzzy Logic* Kluwer Academic Publishers. Dordrech

[23] Kitainik L 1993 *Fuzzy decision procedures with binary relations* Kluwer, Dordrecht

[24] Cordon O, Herrera F, Peregrin A 1995 T-norms wersus Implication Functions as Implication Operators in Fuzzy Control *Proc.6th IFSA Congr* 501-4

[25] Cordon O, Herrera F, Peregrin A 1997 Applicability of the Fuzzy Operators in the Design of Fuzzy Logic Controllers *Fuzzy Sets and Systems* **86** 15-41

[26] Dujet C, Vincent N 1995 Force Implication: A new approach to human reasoning *Fuzzy Sets and Systems* **69** 53-63

[27] Gupta M, Qi J 1991 Theory of T-norms and Fuzzy Inference Methods *Fuzzy Sets and Systems* **40** 431-50

[28] Trillas E, Valverde L 1985 On implication and indistinguishability in the setting of fuzzy logic *In: K.Janusz and Yager R.R,eds. Management Decision Support Systems using Fuzzy Sets and Possibility Theory* 198-212

[29] Trillas E 1997 On a Mathematical Model for Indicative *FUZZIEE'97-Sixth IEEE International Conference or Fuzzy Systems* **1** 3-10

[30] Dubois D, Prade H 1991 Fuzzy Sets in approximate reasoning. Part 1:

Inference with possibility distributions *Fuzzy Sets and Systems* **4** 143-202

[31] Smets P, Magrez P 1987 Inplication in fuzzy logic *Internat. J. Approx. Reason* **1** 327-47

[32] Fodor J C, Roubens M 1994 *Fuzzy preference modeling and multicriteria decision support* Kluwer, Dordrecht 31-6

[33] Gottwal S 2001 *A trestise on many-valued logics* Research Studies Press, Baldock 53-6

[34] Shanshan Wang, Zhenghua Pan, Lei Yang 2012 *Fuzzy Decision Making Based on Fuzzy Logic with Contradictory Negation, Opposite Negation and Medium Negation* Lecture Notes in Computer Science,Springer-Verlag Berlin Heidelberg **7530** 200-8

Authors

Tang yonglong, born on May 20, 1961, in Zhangjiajie of Hunan province

Current position, grades: associate professor in college of mathematics and statistics in Jishou University
University studies: graduated from Huan Normal University on July, 1982 with a mathematics bachelor degree. And he acted as a visiting fellow in Zhongshan University in 2007.
Scientific interests: applied mathematics and cryptology.

Parallel computation of matrix norm based on MapReduce

Yuqiang Sun, Dongyu Zhang, Yan Chen, Bixia Chao, Yuwan Gu*

School of Mathematics and Physics, ChangZhou University, Jiangsu, Changzhou213164, China

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

Received 24 November 2014, www.cmnt.lv

Abstract

A kind of parallel programming method based on MapReduce model is proposed, in allusion to data characteristic of having specific data partitioning requirement, parallel computation of matrix norm is implemented on the platform of high-performance MapReduce. Comparing with the traditional parallel programming model, MapReduce model parallel program can satisfy to requirement of high performance numerical calculations well, its programming for simplicity and readability can improve parallel programming efficiency in effect.

Keywords: MapReduce, Numerical computation, Matrix norm, Parallel computation, Data partitioning, High-performance

1 The definition and property of matrix norm

Definition 1: given $A \in C^{m \times n}$, prescribed a real-valued function of A on $C^{m \times n}$ according to a certain rule, marked $\|A\|$, it satisfies to the following 4 conditions:

(1) Non negative: if $A \neq 0$, then $\|A\| \succ 0$; if $A = 0$, then $\|A\| = 0$.

(2) Homogeneity: for arbitrary $k \in C$, $\|kA\| = |k|\|A\|$.

(3) Triangle inequality: for arbitrary $A, B \in C^{m \times n}$, $\|A+B\| \leq \|A\| + \|B\|$.

(4) Compatibility: when the matrix product AB has meaning, if $\|AB\| \leq \|A\|\|B\|$, then $\|A\|$ is called matrix norm.

Given $A = (a_{ij}) \in C^{n \times n}$, the real-valued function of the following provisions

$$\|A\|_{m_1} = \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}| \quad , \quad \|A\|_{m_\infty} = n \bullet \max_{i,j}|a_{ij}| \quad ,$$

$\|A\|_{m_2} = (\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2)^{\frac{1}{2}}$, they are all norm of matrix A.

Theorem 1: given $A = (a_{ij}) \in C^{m \times n}$, $x = (x_1, x_2, \cdots, x_n)^T \in C^n$, operator norm of three kinds of norms $\|x\|_1$, $\|x\|_2$, $\|x\|_\infty$ that belongs to the vector x is $\|A\|_1 = \max_j \sum_{i=1}^{m}|a_{ij}|$ (known as a column norm); $\|A\|_2 = \sqrt{\lambda_{\max}(A^H A)}$ (known as the spectral norm), where $\lambda_{\max}(A^H A)$ is the maximum of the absolute value of matrix $A^H A$ characteristic value; $\|A\|_\infty = \max_i \sum_{j=1}^{n}|a_{ij}|$ (called line norm) in turn.

2 Outline of MapReduce

MapReduce can be implemented in many ways, and indeed it has various implementations [1-4]. Here, we will outline MapReduce as described in [1]. In a nutshell, MapReduce computations consist in processing input data sets by creating a set of intermediate (key, value) pairs, and then reducing them to yet another list of (key, value) pairs. The computations are performed in parallel.

More precisely, MapReduce applications are divided into two steps. In the first step a Map function processes the input dataset (e.g. a text/HTML file), and a set of intermediate (key1, value1) pairs is generated. In the second step the intermediate values are sorted by key1, and a Reduce function merges the intermediate pairs with equal values of key1, to produce a list of pairs (key1, value2). Thus, the input dataset is transformed into a list of key/value pairs. Let us consider examples given in [1]. Counting occurrences of words in a big set of documents can be organized in the following way. Map function emits intermediate pair (word,1) for each word in the input file(s). The intermediate pairs are reduced by sorting them by word, summing 1s, and producing pairs (word, count). In the inverted index computation all documents comprising certain words must be identified. The Map function emits pairs (word, docID), where docID is a document identifier (e.g. a URL of a web page). In the Reduce function all (word, docID) pairs are sorted, and pairs (word, list_docIDs) are emitted, where list_docIDs is a sorted list of docIDs. There are many types of practical applications which can be expressed in the MapReduce model. More detailed and advanced examples are given in [1, 2, 3, 5].

Both map and reduce operations are performed in parallel in a distributed computer system. Processing a MapReduce application starts with splitting the input files into load units (in [1] called splits). Many copies of the program start on a cluster of machines. One of the machines, called the master, assigns work to the other computers (workers). There are m map tasks and r reduce tasks to assign. In the further discussion the map tasks will be called mappers, and the reduce tasks reducers. A worker which received a mapper reads the corresponding input load unit and processes the data using the Map function. The output of this function is divided into r parts by the partitioning function and written to r files on the local disk. Each of the r files corresponds to one of the reducers. Usually the partitioning function is something like hash(key1) mod r. The information about local file locations is sent back to the master, which forwards it to the reduce workers.

When a reduce worker receives this information, it reads the buffered data from the local disks of the map workers.

After reading all intermediate data, the reduce worker sorts it by the intermediate keys in order to group together with all occurrences of the same intermediate key. Each key and the corresponding set of values are then processed by the Reduce function. Its output is appended to a final output file for a given reducer. Thus, the output of MapReduce is available in r output files. The execution of MapReduce is completed when all reducers finish their work [6, 7].

Basic framework of MapReduce is shown in figure 1:

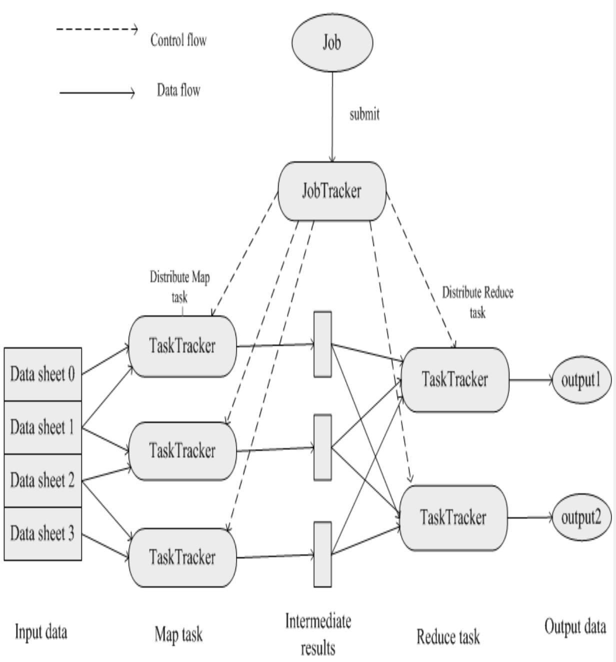


FIGURE 1 Basic framework of MapReduce

3 Parallel computation of matrix norm based on MapReduce

Given $A=(a_{ij})\in C^{n\times n}$, serial algorithm of real-valued function $\|A\|_{m_1}=\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|$ is as follows:

```
Begin
  A=0
  For i=1 to n do
    For j=1 to n do
    A=A+ | $a_{ij}$ |
    End for
  End for
  $\|A\|_{m_1}$ =A
End
```

Because of MapReduce Still belongs to the explicit data partition and parallel computing model, it shall be that the matrix will be divided into several blocks of data in accordance with the specific data partition strategy, the using partition method is according to line by continuous partition or according to column by continuous partition etc. Then according to serial computing semantic and partition method of matrix norm $\|A\|_{m_1}=\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|$, so it can easily write treatment process of Map and Reduce, description is as follows:

(1) Map: Key=ij,Value=| a_{ij} |, if according to line by continuous partition, then the intermediate process is adding Value of the same i; if according to column by continuous partition, then the intermediate process is adding Value of the same j.

(2) Reduce: if according to line by continuous partition, key=i, value= the value corresponding to i obtained in the intermediate process; if according to column by continuous partition, key=j, value= the value corresponding to j obtained in the intermediate process. Finally, all value is added together, namely to obtain the value of matrix norm $\|A\|_{m_1}$.

Given $A=(a_{ij})\in C^{n\times n}$, serial algorithm of real-valued function $\|A\|_{m_\infty}=n\bullet\max_{i,j}|a_{ij}|$ is as follows:

```
Begin
  A=| $a_{11}$ |
  For i=1 to n do
    For j=1 to n do
    If  A<| $a_{ij}$ | then A=| $a_{ij}$ |
    Else A=A
    End for
  End for
  $\|A\|_{m_\infty}=n*A$
End
```

Because of MapReduce Still belongs to the explicit data partition and parallel computing model, it shall be that the matrix will be divided into several blocks of data in accordance with the specific data partition strategy, the using partition method is according to line by continuous partition or according to column by continuous partition etc. Then according to serial computing semantic and partition method of matrix norm $\|A\|_{m_\infty}=n\bullet\max_{i,j}|a_{ij}|$, so it can easily write treatment process of Map and Reduce, description is as follows:

(1) Map: Key=ij,Value=| a_{ij} |, if according to line by continuous partition, then the intermediate process is getting the maximum Value of the same i; if according to column by continuous partition, then the intermediate process is getting the maximum Value of the same j.

(2) Reduce: if according to line by continuous partition, key=i, value= the value corresponding to i obtained in the intermediate process; if according to column by continuous partition, key=j, value= the value corresponding to j obtained in the intermediate process. Finally, obtaining the maximum Value, this value is multiplied by n times, then namely matrix norm $\|A\|_{m_\infty}$.

Given $A=(a_{ij})\in C^{n\times n}$, serial algorithm of real-valued function $\|A\|_{m_2}=(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2)^{\frac{1}{2}}$ is as follows:

```
Begin
  A=0
  For i=1 to n do
    For j=1 to n do
    A=A+ $|a_{ij}|^2$
```

```
      End for
    End for
    $\|A\|_{m_2} = \sqrt{A}$
End
```

Because of MapReduce Still belongs to the explicit data partition and parallel computing model, it shall be that the matrix will be divided into several blocks of data in accordance with the specific data partition strategy, the using partition method is according to line by continuous partition or according to column by continuous partition etc. Then according to serial computing semantic and partition method of matrix norm $\|A\|_{m_2} = (\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2)^{\frac{1}{2}}$, so it can easily write treatment process of Map and Reduce, description is as follows:

(1) Map: Key=ij,Value= $|a_{ij}|^2$,if according to line by continuous partition, then the intermediate process is adding Value of the same i; if according to column by continuous partition, then the intermediate process is adding Value of the same j.

(2) Reduce: if according to line by continuous partition, key=i, value= the value corresponding to i obtained in the intermediate process; if according to column by continuous partition, key=j, value= the value corresponding to j obtained in the intermediate process. Finally, all value is added together and the result is squared root, namely to obtain the value of matrix norm $\|A\|_{m_2}$.

Given $A=(a_{ij})\in C^{m\times n}$, serial algorithm of Column norm $\|A\|_1 = \max_j \sum_{i=1}^{m}|a_{ij}|$ is as follows:

```
Begin
  For j=1 to n do
    A[j]=0
    For i=1 to m do
      A[j]=A[j]+ | $a_{ij}$ |
    End for
  End for
  A=A[1]
For j=1 to n do
  If A<A[j] then A=A[j]
End for
```

Because of MapReduce Still belongs to the explicit data partition and parallel computing model, it shall be that the matrix will be divided into several blocks of data in accordance with the specific data partition strategy, the using partition method is according to column by continuous partition etc. Then according to serial computing semantic and partition method of matrix norm $\|A\|_1 = \max_j \sum_{i=1}^{m}|a_{ij}|$, so it can easily write treatment process of Map and Reduce, description is as follows:

(1) Map: Key=ij, Value=| a_{ij} |, the intermediate process is adding the Value of the same j.

(2) Reduce: key=j, value= the value corresponding to j obtained in the intermediate process. Finally, obtaining the maximum Value, namely to obtain the value of matrix norm $\|A\|_1$.

Given $A=(a_{ij})\in C^{m\times n}$, serial algorithm of line norm $\|A\|_\infty = \max_i \sum_{j=1}^{n}|a_{ij}|$ is as follows:

```
Begin
  For i=1 to m do
    A[i]=0
    For j=1 to n do
      A[i]=A[i]+ | $a_{ij}$ |
    End for
  End for
  A=A[1]
For i=1 to m do
  If A<A[i] then A=A[i]
End for
```

Because of MapReduce Still belongs to the explicit data partition and parallel computing model, it shall be that the matrix will be divided into several blocks of data in accordance with the specific data partition strategy, the using partition method is according to line by continuous partition etc. Then according to serial computing semantic and partition method of matrix norm $\|A\|_\infty = \max_i \sum_{j=1}^{n}|a_{ij}|$, so it can easily write treatment process of Map and Reduce, description is as follows:

(1) Map: Key=ij, Value=| a_{ij} |, the intermediate process is adding the Value of the same i.

(2) Reduce: key=i, value= the value corresponding to i obtained in the intermediate process. Finally, obtaining the maximum Value, namely to obtain the value of matrix norm $\|A\|_\infty$.

Given $A=(a_{ij})\in C^{m\times n}$, parallel computation process based on MapReduce of spectral norm $\|A\|_2 = \sqrt{\lambda_{\max}(A^H A)}$ is as follows:

There are two Mapreduce processes, $A^H A$ is implemented in the first, λ value of $A^H A$ is solved in the second.

The Map and Reduce treatment process of implementing $A^H A$ is as follows:

A^H is partitioned by line, A is partitioned by column.

(1) Map: $\mathrm{Key}_1 = \mathrm{i}$, $\mathrm{Value}_1 = (\mathrm{a}_{i1}, \mathrm{a}_{i2}, \cdots, \mathrm{a}_{in})$, $\mathrm{Key}_2 = j$, $\mathrm{Value}_2 = (\mathrm{a}_{1j}, \mathrm{a}_{2j}, \cdots, \mathrm{a}_{mj})$, then the intermediate process is that the vector corresponding i and the vector corresponding j multiplied by two.

(2) Reduce: key=ij, value= a_{ij}, namely to obtain matrix $A^H A$.

The Map and Reduce process of treatment process of implementing λ value of matrix $A^H A$ according to a lower triangular in the parallel algorithm of LU decomposition in reference[8], then the diagonal is multiplied, namely to obtain all λ value, getting the maximal λ value, and the maximal λ value is squared root, namely to obtain the value of spectral norm $\|A\|_2$.

4 Conclusion

A parallel computation method of matrix norm based on the MapReduce model is proposed in the paper, in some areas related to computation of the matrix norm, parallel computation method in the paper brings convenient. As a new type of parallel and distributed programming model, MapReduce model has a high parallel representation abstract [9, 10], can effectively reduce the difficulty of parallel programming, and upgrades the parallel programming productivity. The next step for the research work is that the model is introduced to high performance computing area of more numerical value / non numerical value.

Acknowledgment

Supported by The National Natural Science Fund (11271057, 51176016) and the project of general office of Broadcasting and Television (GD10101) and Natural Science Fund in JiangSu (BK2009535) and Natural Science Fund in ZheJiang (Y1100314) and Jiangsu Province ordinary university innovative research project (SCZ1412800004).

References

[1] Dean J, Ghemawat S 2004 MapReduce: simplified data processing on large clusters *in: Proc. OSDI'04: Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA 137–50 http://labs.google. com/papers/mapreduce.html

[2] Lin J, Dyer C 2010 Data-Intensive Text Processing with MapReduce *Morgan &Claypool* 2010

[3] Pike R, Dorward S, Griesemer R, Quinlan S 2005 Interpreting the data: parallel analysis with Sawzall *Scientific Programming* **13**(2005) 277–98

[4] Wikipedia, MapReduce, *http://en.wikipedia.org/wiki/MapReduce* [Online; 17-February-2010]

[5] Ranger C, Raghuraman R, Penmetsa A, Bradski G, Kozyrakis C 2007 Evaluating MapReduce for multi-core and multiprocessor systems *in: Proceedings of International Symposium on High Performance Computer Architecture* HPCA 13–24

[6] Wang W, Li J H, Ding R F 2011 Maximum likelihood parameter estimation algorithm for controlled autoregressive autoregressive models *International journal of computer mathematics* **88**(16) 3458-67

[7] Bonettini S, Landi G, Piccolomini E L, Zanni L 2013 Scaling techniques for gradient projection-type methods in astronomical image deblurring *International journal of computer mathematics* **90**(1) 9-29

[8] Zheng Qi-long etc. 2010 Application of HPMR in Parallel Matrix Computation *Computer Engineering* **36**(8) 49-51

[9] Zhiyuan Shi, Volker Gruhn, Yuwan Gu, Yuqiang Sun 2013 The Study of Reuse Mashup Technology Based on Using Frequency *Information Technology Journal* **12**(14) 2669-72

[10] Yihong Cao, Yuwan Gu, Huanhuan Cai, Yuqiang Sun 2013 An Improved Decision Tree Algorithm Based on The Attribute Set Dependency *Information Technology Journal* **12**(22) 6641-5

Authors' index

Chao Bixia	13
Chen Yan	13
Gu Yuwan	13
Sun Yuqiang	13
Tang Yonglong	7
Zhang Dongyu	13

Mathematical and Computer Modelling

Fuzzy knowledge searching on the basis of the traditional and-or graph search algorithm

Yonglong Tang

Computer Modelling & New Technologies 2015 ***19****(2A) 7-12*

Based on the fuzzy propositional logic FLCOM and fuzzy set FSCOM, we research the formal denotation, inference and computation of fuzzy knowledge. We extend the fuzzy and-or graph, turn the propositional formulas as state nodes, express the logical rules as the search space, construct and-or graph of the fuzzy propositional formula. We modify heuristic function on the basis of the traditional and-or graph search algorithm, and give out a method to process negation information in the process of reasoning, transforming the fuzzy knowledge reasoning into the state space searching problem, and using the state space searching to solve the problem of fuzzy knowledge reasoning.

Keywords: fuzzy propositional logic FLCOM, fuzzy set FSCOM, fuzzy propositional formula, negation information, state space searching

Parallel computation of matrix norm based on MapReduce

Yuwan Gu, Dongyu Zhang, Yan Chen, Bixia Chao, Yuqiang Sun

Computer Modelling & New Technologies 2015 ***19****(2A) 13-16*

A kind of parallel programming method based on MapReduce model is proposed, in allusion to data characteristic of having specific data partitioning requirement, parallel computation of matrix norm is implemented on the platform of high-performance MapReduce. Comparing with the traditional parallel programming model, MapReduce model parallel program can satisfy to requirement of high performance numerical calculations well, its programming for simplicity and readability can improve parallel programming efficiency in effect.

Keywords: MapReduce, Numerical computation, Matrix norm, Parallel computation, Data partitioning, High-performance

Content B

Part B Information and Computer Technologies		
Wu Yang, Wang Tao, Li Jin-Dong	Research on characteristic parameters mining and clustering of unknown protocols bitstreams	7
Li Yongbiao	Cache Pre-fetching System based on data mining on Web	16
Li Buyi, Deng Chongjing, Li Shuang	High Resolution Remote Sensing Image Classification based on Particle Swarm Optimization and Support Vector Machine	22
Amirgaliyev Y, Yunussov R	Pattern recognition systems in the problems of automatic person identification using the passport data	27
Authors' Index		31
Cumulative Index		32

Research on characteristic parameters mining and clustering of unknown protocols bitstreams

Yang Wu*, Tao Wang, Jin-dong Li

Dept. of Information Engineering, Shijiazhuang Mechanical Engineering College, Shijiazhuang, 050003, P.R. China

**Corresponding author's e-mail: baiyanwy@163.com*

Received 25 November 2014, www.cmnt.lv

Abstract

Characteristic parameters mining of unknown protocol bitstreams and parameters optimizing of clustering algorithm are the foundations of unknown protocol bitstreams analyzing. The parameters such as the bit frequency, runs and bit frequency within a block are defined according to the frequency of zero and one, frequency of sequential zero and one, bit frequency within a block. As the parameter of bit frequency within a block is sensitive to the block length, an optimal block length selection algorithm is proposed based on the principle of variance. In order to select effective initial clustering centers for division clustering algorithms such as the k-means algorithm, an initial clustering centers selection algorithm is proposed based on the peak value of sample density for each dimension. In order to select the optimal clustering number, a function of clustering quality evaluation is given by the sample density in cluster and cluster density. Taking the bitstreams of HTTP, DNS, ICMP, TELNET and UDP datasets as the unknown protocols bitstreams, the experimental results not only verified the effectiveness of the proposed algorithms but also point out the necessity of mining more effective parameters.

Keywords: Unknown protocol, bitstreams, clustering, characteristic parameter, bit frequency within a block

1 Introduction

Generally Speaking, the main task of unknown protocol identification is to find out the format information of the protocol from its bitstreams based on frequent sequences mining and the established association rules. Unknown protocol identification can provide supports for further unknown protocol analysis and utilization. Dividing the unknown protocol bitstreams with similar characteristics into corresponding clusters is the foundation of frequent sequences mining and unknown protocol identification. As the known protocol data is the main research object of protocol identification currently, the protocols of network data are distinguished mainly based on pattern matching [1], machine learning [2] and some other known protocol identification methods. The network data capture and analysis tools such as the Snifter and Ethereal are all based on above methods. The main challenges of unknown protocol identification are to identify the protocol fields of unknown protocol and user data accurately in the absence of any prior knowledge of the unknown protocol. However, most of the existing protocol identification technologies are based on the characteristics of known protocols; they cannot be effectively used to analyze the unknown protocol data.

2 Related works

It is easy to know that the key processes of unknown protocol bitstreams clustering mainly include characteristic parameters mining of bitstreams, initial clustering centers selection, optimal clustering number selection, clustering results evaluation and other key issues.

2.1 CHARACTERISTIC PARAMETERS MINING OF BITSTREAMS

The traditional characteristic parameter mining contents of the bitstreams include protocol type, ports, bitstream length, bitstream direction, characteristic fields and some other characteristic parameters. In 1999, the MIT Lincoln Labs provided the 41-dimensional real network traffic data for KDD competition, which is the acknowledged DARPA data for intrusion technologies testing [3]. But to unknown protocol characteristic parameter mining, there are only a few relevant researches in encrypted traffic identification. Charles [4] proposed a method to identify the application protocol of encrypted traffic according to the bytes number of data packets, durations, interactive processes and flow directions. Based on the interactive processes of SSL/TLS traffic, Sun [5] proposed a hybrid multi-level encrypted SSL/TLS traffic classification method, which identifies the specific application protocol of the encrypted traffic by statistical analysis. From the perspective of protocol independent, literature [6] provided a protocol independent online identification scheme for encrypted traffic by extracting the different statistical information of the encrypted and non-encrypted bitstreams. Literature [7] also proposed an encrypted bitstream identification scheme based on the statistical distributions of the zero and one in random and un-random bitstreams.

2.2 CLUSTERING ALGORITHMS AND THEIR PARAMETERS SELECTION

(1) Clustering algorithms selection

Clustering is one of the most important data analysis methods; it divides the samples with similar attributes into corresponding clusters according to a certain similarity measure rule. However, all of the clustering algorithms cannot be widely used to reveal the structures of multidimensional data [8]. The traditional clustering methods mainly include division clustering, hierarchical clustering, grid-based clustering, density-based clustering and model-based clustering. Most of the clustering algorithms are sensitive to their parameters; different parameters may bring completely different clustering results. As division clustering methods have lower implementation complexity, they are widely used in large-scale data clustering; many researchers naturally pay their attentions to the research of parameters selection for division clustering. There are many typical division clustering algorithms such as the *k*-means, PAM (Partitioning Around Medoid), *k*-modes and EM (Expectation Maximization) algorithm [9]. The pivotal problems of the division clustering algorithms mainly include initial clustering centers selection, optimal clustering number selection and clustering results evaluation.

(2) Initial clustering centers selection

The initial clustering centers selection methods of division clustering algorithm mainly include the RS (Random Selection) method, MMD (Maximum and Minimum Distance) algorithm and other improved algorithms.

(a) RS method: If the number of clusters is *k*, the RS method randomly chooses *k* samples as the initial clustering centers. Although the process of the RS method is very simple, its clustering results are usually inconsistent. Different initial clustering centers could inevitably result in different clustering results.

(b) MMD algorithm: The basic idea of the MMD algorithm is to select the samples with maximal distance as the initial clustering centers.

To avoid clustering algorithm converging to a local minimum, Likas [10] proposed a global *k*-means algorithm, in which the initial clustering centers are more and more close to the real clustering centers during the iterative processes. In order to increase the likelihood of obtaining the globally optimal solution, literature [11] provided an initial clustering centers selection algorithm based on selecting the dispersed samples as the initial clustering centers. Based on the MMD algorithm, literature [12] proposed a scheme to select the high-density points farthest from the initial clustering centers as the new centers. Literature [13] proposed a fuzzy clustering algorithm based on large density region to avoid the clustering algorithm converging to a local minimum, but the algorithm needs to calculate the density values of all samples, it is not suitable for large-scale data clustering. Literature [14] proposed a method using recursive calls to find the initial clustering centers with farthest distance for the *k*-means algorithm.

(3) Optimal clustering number selection

The optimal clustering number has important significance for getting high accuracy clustering results. Many classical indices are proposed such as the CH (Calinski-Harabasz) index [15], DB (Davies-Bouldin) index [16], KL (Krzanowski-Lai) index [17], Wint (Weighted inter-intra) index [18], IGP (In-Group Proportion) [19] and so on. But all of these indices are often unable to obtain the correct clustering number when the clustering structures are difficult to determine. Literature [3] proposed a clustering results evaluation method called COPS (Clusters Optimization on Preprocessing Stage) based on hierarchical division, which effectively improves the accuracy of clustering number selection. Furth more, literature [9] proposed the BWP (Between-Within Proportion) index for the *k*-means algorithm. All of the above indices are based on Euclidean distances of the samples or clusters, with the increase of sample dimension, the distance approaching phenomenon will be more obvious and the above methods will become invalid.

3 Unknown protocol bitstreams clustering scheme

3.1 CHARACTERISTIC PARAMETERS MINING FOR UNKNOWN PROTOCOL BITSTREAMS

(1) Bit frequency statistics parameter mining

Firstly, the bitstreams are $DB=(X_1,X_2,...,X_N)$, where $X_i=\left(x_1^i,x_2^i,...,x_{l_i}^i\right)$ is the bitstream *i*, l_i is the length of X_i. The bit frequency statistics mainly checks the bit frequency distribution of zero and one in a bitstream. Taking the bit frequency statistics parameter calculating process of X_i as an example, based on the $y_j^i=2x_j^i-1$ transformation, we change X_i to be a new sequence $Y_i=\left(y_1^i,y_2^i,...,y_{l_i}^i\right)$ composed of -1 and 1, and then get the binomial sum of the sequences as shown in formula (1).

$$S_i=y_1^i+y_2^i+...+y_{l_i}^i\,. \tag{1}$$

Further normalize the binomial sum of sequence as shown in formula (2).

$$F_{X_i}=\frac{|S_i|}{l_i}\,. \tag{2}$$

Then F_{X_i} is the bit frequency statistical parameter of X_i. From the definition of F_{X_i}, we can know that if the bits of X_i are all zero or one, the maximum value of F_{X_i} is one. Generally, F_{X_i} is normal distributed.

(2) Runs statistical parameter mining

Run is composed by successive zero or one bit; there are zero runs and one runs respectively with different lengths in X_i. We set z_{ij} as the frequency of zero run, e_{ij} as the frequency of the one run in X_i, where *j* is the length of the run, γ_0 as the longest lengths of zero run, γ_1 as the longest lengths of one run. On above definitions, we define the run statistical parameter as in formula (3).

$$R_{X_i}=\frac{\left|Var\left(e_{ij}\right)-Var\left(z_{ij}\right)\right|}{Var\left(e_{ij}\right)+Var\left(z_{ij}\right)}, \tag{3}$$

where

$$Var\left(e_{ij}\right)=\frac{1}{\gamma_1}\sum_{j=1}^{\gamma_1}\left(e_{ij}-\tilde{e}_i\right)^2, \tag{4}$$

$$Var\left(z_{ij}\right)=\frac{1}{\gamma_0}\sum_{j=1}^{\gamma_0}\left(z_{ij}-\tilde{z}_i\right)^2. \tag{5}$$

According to the definitions of z_{ij}, e_{ij}, γ_0 and γ_1, the binomial sum of sequence Y_i can be expressed as:

$$S_i=\sum_{j=1}^{\gamma_1}je_{ij}-\sum_{j=1}^{\gamma_0}jz_{ij} \tag{6}$$

and then F_{X_i} can be expressed as:

$$F_{X_i} = \frac{\left| \sum_{j=1}^{\gamma_1} je_{ij} - \sum_{j=1}^{\gamma_0} jz_{ij} \right|}{\sum_{j=1}^{\gamma_1} je_{ij} + \sum_{j=1}^{\gamma_0} jz_{ij}} . \tag{7}$$

Formula (2) and (7) show that there is no simple linear relationship between F_{X_i} and R_{X_i} .

(3) Bit frequency within a block statistics parameter mining

As described above, bit frequency within a block mainly focuses on the frequency distribution of zero and one in a block with a certain block length. In this situation, m is the block length, the bitstream X_i can be divided into $H_{im} = \left\lfloor \frac{l_i}{m} \right\rfloor$ blocks. π_{ij} is the bit one frequency of the block *j*.

$$\pi_{ij} = \sum_{k=1}^{m} x^{i}_{(j-1)m+k} . \tag{8}$$

When the block length is *m*, define B_{X_i} as the bit frequency within a block statistical parameter of X_i as shown in formula (9).

$$B_{X_i} = \frac{\sum_{j=1}^{H_{im}} (j\pi_{ij} - \tilde{\pi}_i)^2}{H_{im}\Phi_i} , \tag{9}$$

where $\tilde{\pi}_i = \frac{1}{H_{im}} \sum_{j=1}^{H_m} j\pi_{ij}$ and $\Phi_i = \max_{1 \le j \le H_{im}} \left((j\pi_{ij} - \tilde{\pi}_i)^2 \right)$.

(4) Optimal block length selection

Before we give the optimal block length selection algorithm, we firstly give the following definitions.

Definition 1: σ_k is the variance of π_{ji} for cluster C_k .

$$\sigma_k = \frac{1}{H_m N_k} \sum_{i=1}^{H_m} \sum_{j=1}^{N_k} \left(\pi_{ji} - \tilde{\pi}_{ki} \right)^2 , \tag{10}$$

where $\tilde{\pi}_{ki} = \frac{1}{N_k} \sum_{j=1}^{N_k} \pi_{ji}$ is the average value of π_{ji} for cluster C_k , N_k is the number of bitstreams included in C_k .

Definition 2: $\tilde{\sigma}$ is the average value of σ_k for the bitstreams sets $C = (C_1, C_2, ..., C_p)$.

$$\tilde{\sigma} = \frac{1}{p} \sum_{k=1}^{p} \frac{1}{H_m N_k} \sum_{i=1}^{H_m} \sum_{j=1}^{N_k} \left(\pi_{ji} - \tilde{\pi}_{ki} \right)^2 . \tag{11}$$

When $\tilde{\sigma}$ obtains the minimum value, we can confirm that the frequencies of bitstreams in each cluster have least differences as the block length is *m*.

Definition 3: σ is the variance of all $\tilde{\pi}_{ki}$ for the bitstreams sets $C = (C_1, C_2, ..., C_p)$.

$$\sigma = \frac{1}{H_m p} \sum_{i=1}^{H_m} \sum_{k=1}^{p} \left(\tilde{\pi}_{ki} - \frac{1}{p} \sum_{k=1}^{p} \tilde{\pi}_{ki} \right)^2 . \tag{12}$$

When σ obtains the maximum value, we can confirm that the frequencies of bitstreams in different clusters have greatest differences as the block length is *m*.

Definition 4: Q_m is the difference of σ and $\tilde{\sigma}$ for optimal block length selection.

$$Q_m = \sigma - \tilde{\sigma} . \tag{13}$$

The purpose of Q_m definition is to balance $\tilde{\sigma}$ and σ . The optimal block length should ensure $\tilde{\sigma}$ is as small as possible, but σ is as large as possible. So when we get the maximum Q_m , we take m as the optimal block length. Based on above definitions, the main steps of the optimal block length selection algorithm are as follows:

Step 1: Calculate the bit frequency statistical parameters, runs statistical parameters and bit frequency within a block statistical parameters for all the bitstreams respectively, m_0 is the initial block length, H_{km_0} is the minimum block number defined in formula (14).

$$H_{km_0} = \left\lfloor \frac{\min(l_1, l_2, ..., l_N)}{m_0} \right\rfloor . \tag{14}$$

Step 2: Using the *k*-means algorithm cluster the bitstreams into *p* clusters as $C = (C_1, C_2, ..., C_p)$

Step 3: Set $m = m_0$, confirm Q_{m_0} according to formula (10), (11), (12) and (13).

Step 4: Set $m = m + 1$, get the new block number H_{km} , and then get the new Q_m according to formula (10) (11) (12) and (13).

Step 5: if $m < m_{\max}$, repeat Step(4), get corresponding Q_m for different block length.

Step 6: Select the optimal block length m_{opt} according to formula (15).

$$m_{opt} = \arg\max_{m_0 \le m \le m_{\max}} \{Q_m\} . \tag{15}$$

3.2 UNKNOWN PROTOCOL BITSTREAMS CLUSTERING BASED ON THE *K*-MEANS ALGORITHM

Once we get the *F*, *R* and *B* characteristic parameters of bitstreams, the bitstreams will be clustered by the *k*-means algorithm. The initial clustering centers selection and optimal clustering number selection algorithms for the *k*-means algorithm are as follows:

(1) Initial clustering centers selection algorithm

(a) Confirm the range of characteristic parameters for each dimension as $[u_{j_{\min}}, u_{j_{\max}}]$, where $1 \le j \le h$ and h is the maximum dimension number.

(b) Set λ_1 as the number of sections for sample density statistics of the first dimension, $\varphi_1(i)$ is the sample density of section *i*.

$$\phi_1(i) = \frac{\Delta N_1(i)}{\Delta u_1}, 1 \le i \le \lambda_1 . \tag{16}$$

(c) If $\varphi_1(m)$ is a peak value, the sample density statistics sections between $\varphi_1(m)$ and the previous candidate is not less than η_1, the average value of section *m* is the candidate for clustering center.

(d) Adjust corresponding parameters, until *h* is equal to the maximum dimension, then return the candidates for initial clustering centers.

(e) Initial clustering centers selection

Set up the relationship tree of the candidates for initial clustering centers according to their mapping relationships of each dimension. The initial clustering centers selection process is based on the MMD algorithm.

(2) λ_j and η_j selection

As the average sample density of each dimension may be different, the values of λ_j as defined in formula (17) should be also different. If $W_j > W_{j+1}$, there will be more sample density statistical sections in dimension *j* than dimension *j*+1. W_j is the difference of the maximum value and minimum value of dimension *j*. k , N and h usually satisfy $k << N$, $h << N$ and $\sqrt[h]{N} \ge k$. The parameter $\frac{W_j}{\sqrt[h]{\prod_{i=1}^{h} W_i}}$ is the section number inching parameter for the dimension *j*.

$$\lambda_j = \frac{W_j}{\sqrt[h]{\prod_{i=1}^{h} W_i}} \sqrt[h]{N} . \tag{17}$$

There may be many density peak values in the overlap sections among clusters. So when we check the peak values, the peak values in the η_j sections radius will be ignored. The parameter η_j is defined in formula (18) and the corresponding conclusions are as follows.

$$\eta_j = \frac{\lambda_j - k}{2k} . \tag{18}$$

Conclusion 1: If there is no overlap structure between any two clusters in the dimension *j*, η_j can make sure that the selected initial clustering centers are all included in their clusters.

In order to prove the conclusion 1, we give an example of clusters distributions as shown in Figure 1. To cluster 1, the maximum and minimum value in the direction of x are x_{12} and x_{11}, the maximum and minimum value in the direction of y are y_{12} and y_{11} . To cluster 2, the maximum and minimum value in the direction of x are x_{22} and x_{21}, the maximum and minimum value in the direction of y are y_{22} and y_{21} . Where $y_{12} > y_{22}$ and $y_{11} > y_{21}$, the proof of conclusion 1 is as follows.

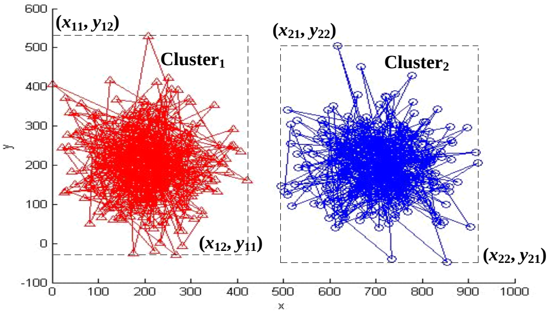


FIGURE 1 Example of clusters distributions

Proof: Based on the above definitions and according to formula (17) λ_x can be expressed as:

$$\lambda_x = (x_{22} - x_{11})\sqrt{\frac{N}{(x_{22} - x_{11})(y_{12} - y_{21})}} . \tag{19}$$

After we confirmed the λ_x, the length of singe sample density statistic section in the direction of x is:

$$\Delta u_x = \frac{W_x}{\lambda_x} = \sqrt{\frac{(x_{22} - x_{11})(y_{12} - y_{21})}{N}} . \tag{20}$$

According to formula (18), η_x can be obtained as follows:

$$\eta_x = \frac{\lambda_x - 2}{4} = \frac{(x_{22} - x_{11})\sqrt{\frac{N}{(x_{22} - x_{11})(y_{12} - y_{21})}} - 2}{4} . \tag{21}$$

The length of the sections which the parameter η_x corresponding is:

$$\Delta x = \eta_x \Delta u_x = \frac{(x_{22} - x_{11}) - 2\Delta u_x}{4} . \tag{22}$$

We can assume that the coordinates of the density peak values of Cluster1 and Cluster2 in the direction of x are O_{x1} and O_{x2}, where

$$\begin{cases} O_{x1} = \dfrac{x_{12} + x_{11}}{2} \\ O_{x2} = \dfrac{x_{22} + x_{21}}{2} \end{cases} . \tag{23}$$

$\bar{W}_x$ is the distance of O_{x1} and O_{x2} in the direction of x as shown in formula (24).

$$\bar{W}_x = O_{x2} - O_{x1} . \tag{24}$$

S_x is the difference of $\bar{W}_x$ and Δx as shown in formula (25).

$$S_x = \bar{W}_x - \Delta x . \tag{25}$$

If there is no overlap structure between Cluster1 and Cluster2 in the direction of x , where $x_{22} > x_{11}$ and $x_{21} > x_{12}$, S_x satisfies $S_x > 0$. Conclusion 1 holds.

Conclusion 2: When there are some overlap structures between any two clusters, η_j can also make sure that the selected initial clustering centers are all included in their clusters.

Proof: As shown in Figure 2, we can assume that the clustering centers of cluster1 and cluster2 are $\frac{x_{11}+x_{12}}{2}$ and $\frac{x_{21}+x_{22}}{2}$ in the direction of *x*, the distance of them is Δx. The length of overlap structure is $\frac{\overline{W}_x+2\Delta u_x}{2}$. As the samples in a cluster are normal distributed around their cluster center, the density peaks generally appear in $\left[\frac{x_{11}+x_{12}}{2},\frac{x_{21}+x_{22}}{2}\right]$. We will respectively analyze the distributions of clustering centers according to the relationships of $\left|\frac{x_{21}+x_{22}}{2}-\frac{x_{11}+x_{12}}{2}\right|$ and Δx.

(a) If $\left|\frac{x_{21}+x_{22}}{2}-\frac{x_{11}+x_{12}}{2}\right|\le \Delta x$, there will be only an initial clustering center. Conclusion 2 holds.

(b) If $\left|\frac{x_{21}+x_{22}}{2}-\frac{x_{11}+x_{12}}{2}\right|> \Delta x$, the will be only a clustering center, two clustering centers or three initial clustering centers candidates. In summary, when there are some overlap structures between any two clusters in dimension *j*, the initial clustering centers are all included in their clusters. Conclusion 2 is proved.

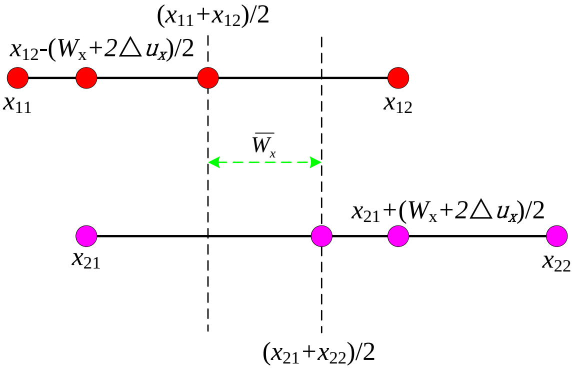


FIGURE 2 Example of clusters overlap structure

(3) The optimal clustering number selection

Once we selected the clustering algorithm, it is very important to establish an effective function $V\left(C^*\right)$ to evaluate the quality of clustering. As most of the current clustering validity functions are complex, based on sample density and clustering density we give a new clustering validity index called CVED (Clustering Validity Evaluation based on Density). When the division of the bitstreams is confirmed as $C^k=\left(C_1,C_2,...,C_k\right)$, $\tilde{\xi}$ is the sample density distribution of all samples of all dimensions, $\tilde{\rho}$ is the clustering density of all dimensions.

$$\tilde{\xi}=\frac{1}{hk}\sum_{j=1}^{h}\sum_{i=1}^{k}\frac{|C_i|}{W_{ij}}, \tag{26}$$

$$\tilde{\rho}=\frac{k}{h}\sum_{j=1}^{h}\frac{1}{\overline{W}_j}, \tag{27}$$

where $|C_i|$ is the number of the samples included in cluster C_i, W_{ij} is the difference of the maximum value and minimum value of dimension *j* for cluster C_i. $\overline{W}_j$ is the difference of the maximum and minimum clustering centers of dimension *j*.

Definition 6: $V\left(C^*\right)$ is the clustering validity index as shown in formula (28):

$$V\left(C^k\right)=\frac{\tilde{\xi}-\tilde{\rho}}{\tilde{\xi}+\tilde{\rho}}. \tag{28}$$

The optimal clustering number k_{opt} is confirmed according to formula (29):

$$k_{opt}=\underset{k_{\min}\le k\le k_{\max}}{\arg\max}\left\{V\left(C^k\right)\right\}. \tag{29}$$

3.3 ALGORITHM COMPLEXITY ANALYSIS

To facilitate the analysis, assuming the number of density sections in each dimension is λ. The operation times of density statistics for the first dimension are $2N\lambda$. The average ratio of effective density peak values is α, when we confirm the parameters of the dimension *j*+1 form the parameters of dimension *j*, the operation times are $\lambda\alpha\left(N\lambda+\lambda\right)$. If h is the number of the dimensions, the all operation times are $2N\lambda+\lambda\alpha\left(N\lambda+\lambda\right)+\cdots+(\lambda\alpha)^{h-1}\left(N\lambda+\lambda\right)$. When confirm the initial clustering centers by the MMD algorithm, the main complexity of the algorithm is to calculate the distances of the initial clustering centers, but the operation times of distances calculating can be ignored as the number of initial clustering centers is much smaller than the number of samples. So the complexity of our initial clustering centers selection algorithm is $O\left(N(\lambda\alpha)^{h-1}\right)$. In extreme cases, every sample density statistics section only contains a sample, where $\lambda^h\le N$ is satisfied according to formula (17), the actual complexity of the algorithm is far less than $O\left(N^2\right)$.

The main complexity of the proposed algorithm is mainly due to the process of clustering. When $k=k_{\max}$, the operations times for $\tilde{\xi}$ and $\tilde{\rho}$ are respectively $hk_{\max}$ and h. The complexity of the proposed algorithm is $O\left(hk_{\max}\right)$. The complexity of the CH, DB, KL and COPS indices is $O\left(N\right)$. The complexity of the Wint, IGP and BWP indices is $O\left(N^2\right)$.

4 Experimental results and analysis

4.1 EXPERIMENTAL SUBJECTS AND SETTINGS

In our experiment, the system of the computer is Windows XP, all bitstreams are from the internet including the HTTP, DNS, ICMP, TELNET and generic UDP bitstreams, the number of each dataset is different with each other. The detail information of HTTP, DNS, ICMP, TELNET and UDP bitstreams is shown in Table 1. In our experiments, we took the bitstreams of HTTP, DNS, ICMP, TELNET and UDP datasets as the bitstreams of unknown protocols.

TABLE 1 Data sets information

Datasets	Sample numbers	Dimensions
HTTP	285	3
DNS	47	3
ICMP	270	3
TELNET	102	3
UDP	1000	3

4.2 RESULTS AND ANALYSIS OF BITSTREAMS CHARACTERISTIC PARAMETERS SELECTION

To verify the affects of different block lengths to the character value of bit frequency within a block, we initially cluster the bitstreams of the HTTP, DNS, ICMP, TELNET and UDP datasets based on the *k*-means algorithm. When we calculate the characteristic value of bit frequency within a block, randomly choose 20 as the block length. In our experiment, the shortest length of the bitstreams is 320; we choose 160 as the longest block length, so the block length is ranging from 2 to 160, the values of Q_m for different block lengths are shown in Figure 3. As shown in Figure 3, the values of Q_m is ranging from -7.729 to 23.866, the maximum value of Q_m is 23.866 when the block length is 88. On the other hand, the minimum value of Q_m is -7.729 when $m \in [107,121]$. So when we calculate the characteristic value of bit frequency within a block, $m = 88$ is the optimal block length.

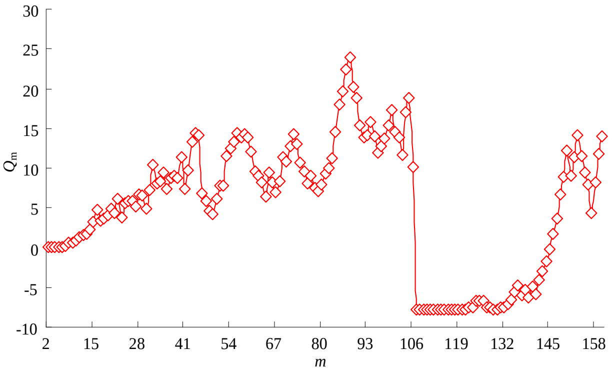
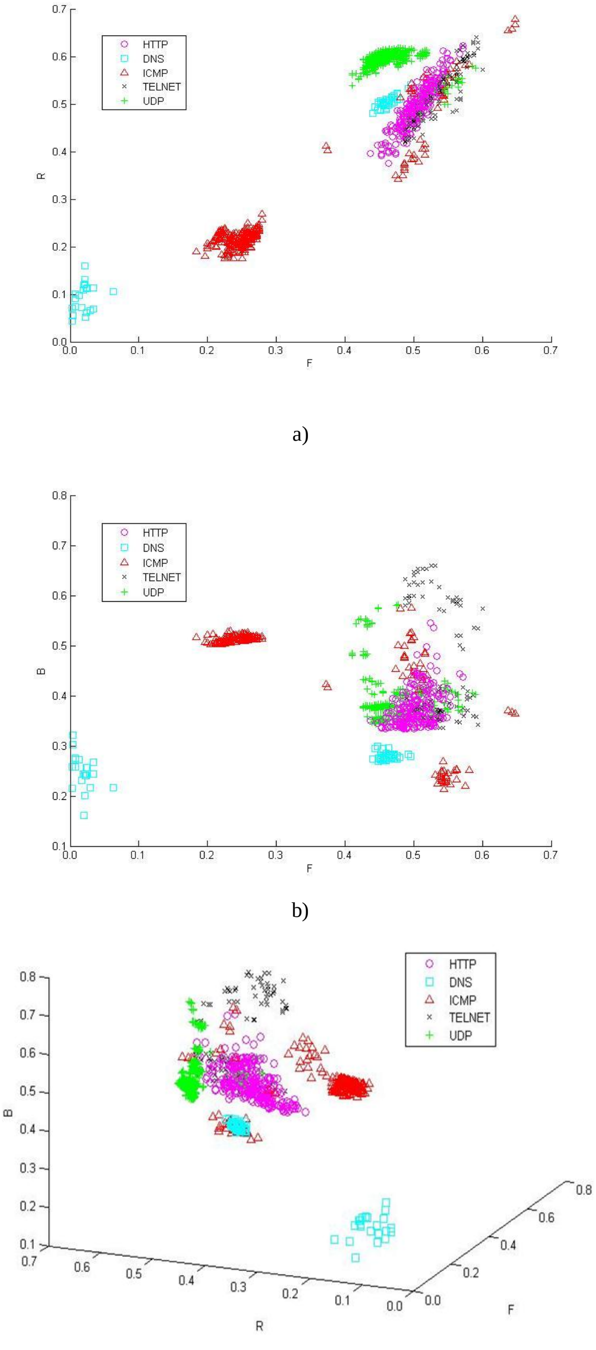


FIGURE 3 Q_m for different block length

When the block length is 88, the distributions of the *F* and *R* values are shown in Figure 4a), the maximum and minimum values of *F* are 0.6479 and 0.0026, the maximum and minimum values of *R* are 0.6788 and 0.0435. Meanwhile, the distributions of *F* and *B* are also shown in Figure 4b), the maximum and minimum values of *B* are 0.6601 and 0.1616. The three-dimensional distributions of *F*, *B* and *R* are shown in Figure 4c). Although there are some overlap structures among the characteristic parameters of the bitstreams, but most of the bitstreams have presented effective clustering characteristic, we can cluster them into corresponding bitstream datasets.



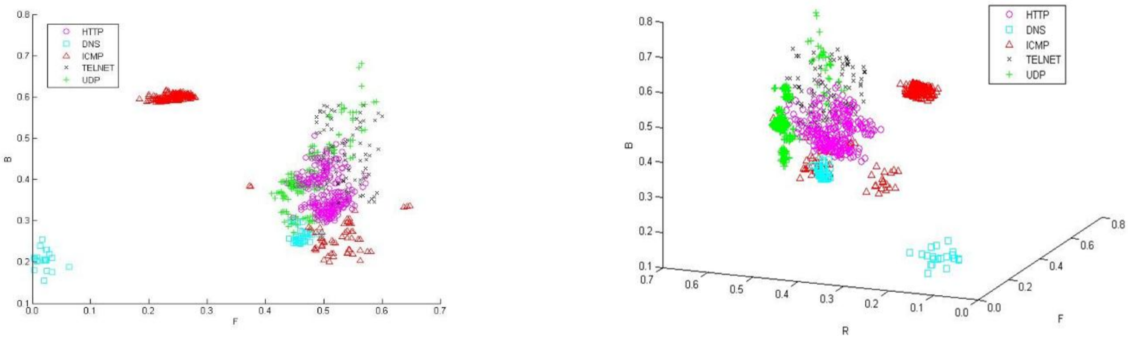
a)

b)

c)

FIGURE 4 Distributions of *F*, *R* and *B* for maximal Q_m: a) Distributions of *F* and *R*, b) Distributions of *F* and *B*, c) Distributions of *F*, R and *B*

To further illustrate the importance of selection optimal block length, under the conditions of $m = 120$ (the value of Q_m is minimum), we recalculate the *B* values for the bitstreams. The distributions of *F* and *B* are shown in Figure 5a). The distributions of *F*, *R* and *B* are shown in Figure 5b). In Figure 5a) and Figure 5b), there are more overlap structures of the *B* values. The clustering characteristic of the *B* values in Figure 5a) are absolutely more indistinctive than the *B* values in Figure 4b). The experimental results demonstrate the validity of the proposed optimal block length selection algorithm.



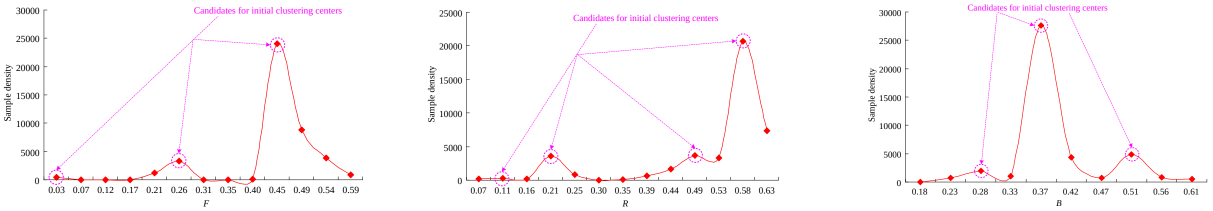
a) b)

FIGURE 5 Distributions of *F*, *R* and *B* for minimal Q_m : a) Distributions of *F* and *B*, b) Distributions of *F*, *R* and *B*

4.3 BITSTREAMS CLUSTERING RESULTS AND ANALYSIS

(1) Initial clustering centers selection

With the proposed algorithm, we can get the sample density distribution characteristics of the HTTP, DNS, ICMP, TELNET, and UDP bitstreams as shown in Figure 6a), Figure 6b) and Figure 6c). According to formula (17), we can respectively calculate the section numbers in the direction of *F*, *R* and *B*; they are 13, 10 and 13. The values of η_F , η_R and η_B can also be confirmed by formula (18), they are 0.8, 0.8 and 0.5. As η_F , η_R and η_B are all less than 1, so all of the peak values in the directions of *F*, *R* and *B* should all be taken as the candidates for initial clustering center. There are three candidates for initial clustering center both in the direction of *F* and *B*, their coordinates are F_1=0.03, F_2=0.26, F_3=0.45, B_1=0.28, B_2=0.37 and B_3=0.51. In the direction of *R*, there are four candidates for initial clustering center, their coordinates are R_1=0.11, R_2=0.21, R_3=0.49 and R_4=0.58. Based on the MMD algorithm and the relationship tree of the candidates for initial clustering center, we obtained five initial clustering centers, they are (0.03,0.11,0.28), (0.26,0.21,0.51), (0.45,0.49,0.28), (0.45,0.58,0.37), (0.51,0.49,0.37).



a) b) c)

FIGURE 6 Distributions of sample density: a) Sample density in the direction of *F*, b) Sample density in the direction of *R*, c) Sample density in the direction of *B*

(2) Similarity analysis of clustering centers and impacts on the iteration times

To illustrate the effectiveness of the proposed initial clustering centers selection algorithm, the similarity value of the initial clustering centers $U'=(u'_1,u'_2,...,u'_k)$ and the final clustering centers $U=(u_1,u_2,...,u_k)$ is defined in formula (30).

$$\tau_i = \frac{4(u_i, u'_i)}{\left(|u_i|+|u'_i|\right)^2}. \qquad (30)$$

When the *k*-means algorithm is converged, we get the final clustering centers, they are (0.02, 0.09, 0.24), (0.25, 0.22, 0.51), (0.50, 0.50, 0.36) and (0.46, 0.59, 0.39). The similarity values of the initial clustering centers and final clustering centers are 99.47 %, 99.97%, 99.40%, 99.98% and 97.36%. Furth more, we also get the average similarity values of the initial clustering centers and the final clustering centers by respectively running the RS, MMD and our initial clustering centers selection algorithm. The results are shown in Figure 7. During 100 repeated experiments, the constant average similarity value of our algorithm is 99.24%. As shown in Figure 7, the average similarity values of the RS method are unstable due to the randomness of the clustering centers; its value is ranging from 86.25% to 99.80%. On the other hand, the average similarity values of the MMD algorithm are less unstable than the RS method as there is only one random clustering center; its value is ranging from 91.88% to 98.63%.

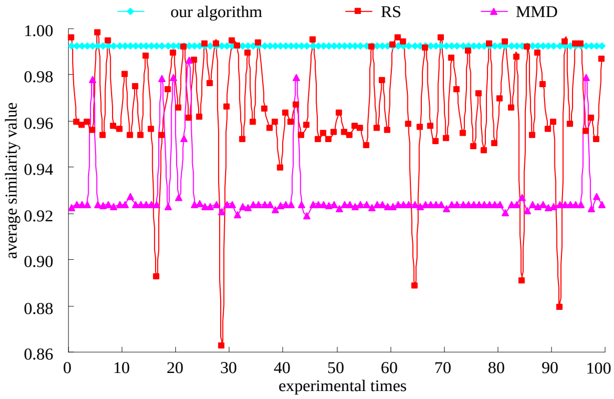


FIGURE 7 Average similarity values

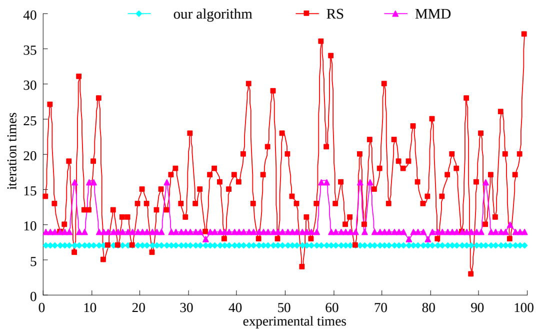
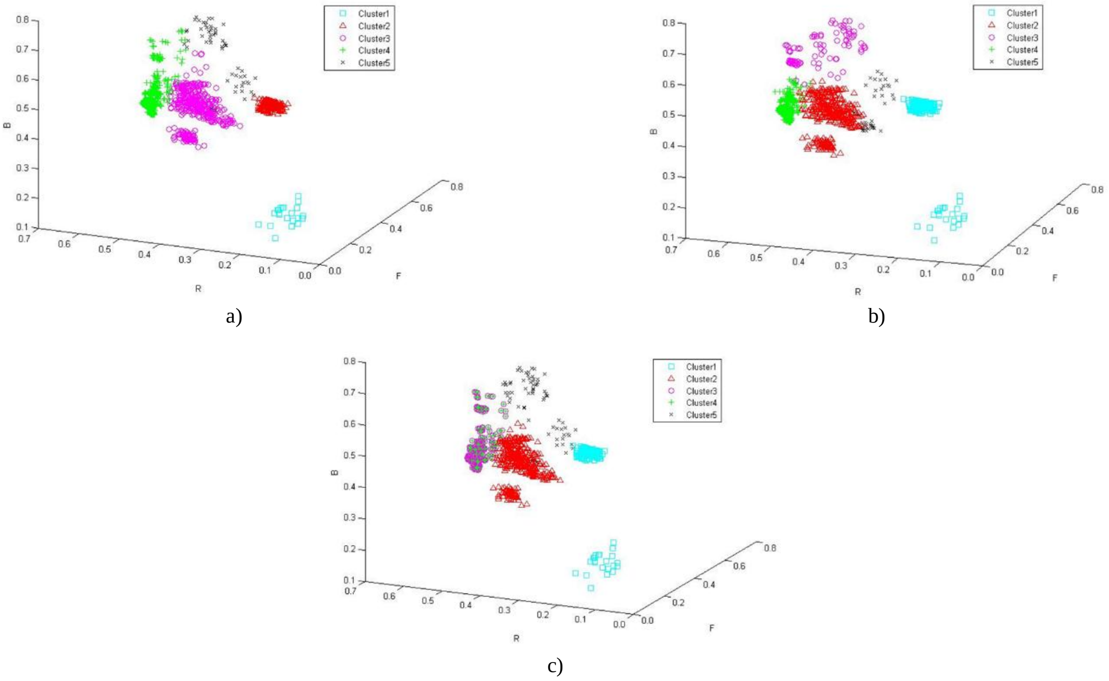


FIGURE 8 Iteration times

To verify the effect of the initial clustering centers to the iteration times of the *k*-means algorithm, we run the RS method, MMD algorithm, our initial clustering centers selection algorithm and the *k*-means algorithm for 100 times. The iteration times of the *k*-means algorithm are shown in Figure 8. As shown in Figure 8, when using our algorithm, the iteration times of the *k*-means algorithm is 7, but to the RS method and MMD algorithm the iteration times of the *k* -means algorithm are respectively ranging from 3 to 37, 8 to 16. Although, the iteration times 3 from the RS method is less than 7 from our algorithm, the clustering results of our algorithm are steadier than the RS method and MMD algorithm.

(4) Impacts on the clustering results

To verify the effect of the initial clustering centers to cluster results, we set 5 as the number of the initial clustering centers, the clustering results of the *k*-means algorithm for our algorithm, RS method and MMD algorithm are respectively shown in Figure 9a), Figure 9b) and Figure 9c). The results of our algorithm are more close to the original clustering characteristics of bitstreams in Figure 4c).



a)

b)

c)

FIGURE 9 Affects of initial clustering centers to clustering results: a) Clustering results of our algorithm, b) Clustering results of the RS method, c) Clustering results of the MMD algorithm

(5) Optimal clustering number selection

In order to verify the effectiveness of the CVED index, we have calculated the values of the KL, Wint, IGP, COPS, BWP and CVED indices and given the clustering number of these indices referring to as shown in Table 2. The experimental numbers of clusters from the KL, Wint, IGP and COPS indices are larger than the actual number of clusters due to the dispersive distributions of the bitstreams. On the other hand, the experimental numbers of clusters from the BWP and CVED indices are closer to the actual number of clusters.

TABLE 2 Numbers of clusters for different indices

Indices	Actual values	Experimental values
KL	5	9
Wint	5	10
IGP	5	7
COPS	5	8
BWP	5	6
CVED	5	6

5 Conclusions

In order to get the characteristic parameters of the bitstreams from the aspect of independent protocol, we defined the characteristic parameters of bit frequency, runs and bit frequency within a block for bitstream respectively. As the characteristic parameter of bit frequency within a block is sensitive to the block length, we proposed an algorithm based on the principle of the variance to obtain the optimal block length. As the sample density in each cluster is generally higher than the average sample density, we firstly calculated the sample density in each sample density calculating section for every dimension, the average sample value of section with the density peak value is taken as the candidate for initial clustering center. The relationship tree of candidates for initial clustering centers is set up based on the mapping relationships of dimensions. With the combination of the MMD algorithm, the initial clustering centers are selected from the relationship tree.

Furthermore, we also defined the function of clustering quality evaluation based on the definitions of sample density in cluster and cluster density. Taken the bitstreams of HTTP, DNS, ICMP, TELNET and UDP datasets as unknown protocol bitstreams, the experimental results demonstrate that our proposed algorithms can effectively mine the characters of protocol bitstreams and divide the bitstreams into the corresponding clusters. However, with the considerations of multi-value property of protocol field, there are some overlap structures among F, R and B values respectively which have some affects to the bitstreams clustering. Our next research work is to mine more effective parameters for unknown protocol bitstreams.

Acknowledgments

This work is supported by the National Natural Science Foundation of China. The authors gratefully acknowledge assistance and encouragement from Prof. Yuri Shunin. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] Zheng T M, Wang T, Guo S Z 2012 Improved space protocol identification algorithm *Journal on Communications* **33**(5) 183-90

[2] Tan J, Chen X S, Du M 2012 A novel real-time p2p identification algorithm based on BPSO and neural networks *Journal of Central South University (Science and Technology)* **43**(6) 2190-97

[3] Chen L F 2008 Research on clustering methods for high dimensional data and their applications *PhD dissertation of Xia Men University*.

[4] Charles V W, Fabian M, Gerald M M 2006 On inferring application protocol behaviors in encrypted network traffic *Journal of Machine Learning Research* **7**(12) 2745-69

[5] Sun G L, Xue Y B, Dong Y F 2010 A novel hybrid method for effectively classifying encrypted traffic *GLOBECOM 2010*: *Proc. Communications and Systems Security (Miami, Florida, USA, 6-10 December 2010) IEEE* 2010, pp 1-5

[6] Bo Z 2012 Research on protocol independent online identification of encrypted traffic *PhD dissertation of PLA Information Engineering University*

[7] Zhao B, Gong H, Liu Q R 2013 Protocol independent identification of encrypted traffic based on weighted cumulative sum test *Journal of Software* **24**(6) 1334-45

[8] Xu R 2005 Survey of clustering algorithm *IEEE Tran on Neural Networks* **16**(3) 645-78

[9] Zhou S B 2011 Research and application on determining optimal number of clusters in cluster analysis *PhD dissertation of Jiang Nan University*

[10] Likas A, Ulassis M, Uerbeek J 2003 The global k-means clustering algorithm *Pattern Recognition* **36**(2) 451-61

[11] Liu Y M, Zhang H X 2011 Approach to selection initial centers for *k*-means with variable threshold *Computer engineering and applications* **47**(32) 56-8

[12] Xiong Z Y, Chen R T, Zhang Y F 2011 Effctive method for cluster' initialization in K-means clustering. *Application Research of Computers* **28**(11) 4188-90

[13] Li X, Zhang J F, Cai J H 2012 A fuzzy clustering algorithm based on large density region *Journal of Chinese Computer Systems* **33**(6) 1310-5

[14] Chen G P, Wang W P, Huang J 2012 Improved initial clustering center selection method for *k*-means algorithm *Journal of Chinese Computer Systems* **33**(6) 1320-3

[15] Calinski T, Harabasz J 1974 A dendrite method for cluster analysis *Communications in Statistics* **3**(1) 1-27

[16] Davies D L, Bouldin D W 1979 A cluster separation measure *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**(2) 224-7

[17] Dudoit S, Fridlyand J 2002 A prediction-based resampling method for estimating the number of clusters in a dataset *Genome Biology* **3**(7) 1-21

[18] Dimitriadou E, Dolnicar S, Weingessel A 2002 An examination of indexes for determining the number of cluster in binary data sets *Psychometrika* **67**(1) 137-60

[19] Kapp A V, Tibshirani R 2007 Are clusters found in one dataset present in another dataset? *Biostatistics* **8**(1) 9-31

Authors	
	Wu Yang, 1985, Cheng Du, China **Current position, grades:** PhD. Student **University studies:** Shijiazhuang Mechanical Engineering College **Scientific interest:** Network security and network data analysis **Experience:** From 2005 to 2009, studied in UESTC and got the Bachelor Degree in 2009; From 2009 to 2012, studied in Shijiazhuang Mechanical Engineering College and got the Master Degree in 2012; From 2012 to now, studies in Shijiazhuang Mecha nical Engineering College for Doctor Degree.
	Wang Tao, 1964, Shi Jia-zhuang, China **Current position, grades:** Professor **University works:** Shijiazhuang Mechanical Engineering College **Scientific interest:** network security and cryptology **Publications:** Principles and Methodologies of Side-Channel Analysis in Cryptography, published by the Science Press, 2014. **Experience:** From 1990 to now, works in the Department of Information Engineering of Shijiazhuang Mechanical Engineering College.
	Li Jin-dong, 1990, Shi He-zi, China **Current position, grades:** Graduate Student **University studies:** Shijiazhuang Mechanical Engineering College **Scientific interest:** Network security and network data analysis **Experience:** From 2008 to 2012, studied in Xinjiang University and got the Bachelor Degree in 2012; From 2009 to now, studies in Shijiazhuang Mechanical Engineering College for the Master Degree.

Cache Pre-fetching system based on data mining on Web

Yongbiao Li*

Human resource management Department, Jinan University, Huangpu Street 601# Tianhe Distric, Guangzhou City,Guangdong Province, 510632, P.R.C

**Corresponding author's e-mail: gql.jnu@gmail.com*

Received 1 March 2014, www.cmnt.lv

Abstract

From 20th century 90's to now on, Internet and data mining techniques had developed rapidly and became mature, kinds of application on Web data mining had been proposed to the market. In this paper, we would first introduce the development of cache Pre-fetching technique, and then present a cache pre-fetching System model based on Web data mining, details of each implementation would follow. Our aim was to enhance caching effectiveness, and network accessing speed. Such technique could be applied in personnel, educational, and official information managing system in database of educational scope. Accessing speed of educational information system for numerous teachers and students, benefit high school personnel management, and also the effective scientific structuralize educational management.

Keywords: Web data mining, sequential mining, cache pre-fetching system

1 Introduction

While WWW had become the most popular service in Internet, Web work load grew explosively, the problem came along was the simultaneous growth of network ping. How to raise the responding speed had become the most urgent problem. Caching technique was an effective solution and widely used in Web clients and servers, but most of them implemented using traditional memory paging strategy, for example LRU (Least Recently Used). Thomas MK et al. [1] stated that: As the ratio of dynamic content and personal service on WWW kept rising, performance improvement brought by caching was not obvious. Meanwhile, Chen Xin et al. [2] stated that: Hit rate of web cache would usually float from 24% to 45%. Pre-fetching technique was a further extension of cache technique, enhancement of Web system could implemented in different ways:

Pre-fetching mechanism could further enhance hit rate into around 60% to 80%, reducing network accessing latency to improve quality of service (QoS). Compared to cache technique, pre-fetching mechanism was pertinence; hence the memory space usage was relatively small, which was able to fulfil personalization need, showing users' personal interest. Pre-fetching mechanism could also smooth work load of network, the usage of limited network resource would be more effective. Therefore, the study of Web pre-fetching technique was important to reduce web access latency and raise QoS.

In this paper, we would combine pattern features of Web proxy server service flow and cache mechanism; then studied about the design of Web pre-fetching strategy. By studying the relation between cache size, replacement algorithms and cache hit rate, explored a kind of Web proxy server side Web cache replacement algorithms with high hit rate. With the use of Web log and data mining skills, linked up the internal attributes of Web page with cache size and hit rate, in order to optimize the usage of proxy server.

2 Related Work

2.1 WEB CACHE TECHNIQUE

Cache technique was well developed in many fields, like operating system, distributed file system, but Web cache was different from those traditional cache systems.

Firstly, "file" would be the unit for Web cache system operations like save, edit, and replace, and its size for request, save, and transmission operations on the Internet various. Therefore, replacement strategy had to consider not just frequency and recency, but file size also.

Secondly, the cost for traditional cache to maintain different cache object was basically equal; but in Web cache, the cost of Web cache was related to the cost of getting Web cache object, and path and server for data transmission, thus the time for downloading different object various.

Moreover, there was usually a small number of accessing program in traditional cache; but for Web cache, other than client side, there might be a great number of client connections, and this kind of connections were usually come from decades to thousands clients. Besides, cache consistency need maintenance in Web cache.

Many countries were using Web cache, for example, the JANET from Britain, DFN from Germany, FREEnet of Russia, SingNet of Singapore, ThaiSARN from Thailand, and etc. All of them were national Web cache system, providing high speed cache service with low price. CERNET from China introduced a level-structural Web cache project: built a L1 cache system in nation center, then built L2 cache system in each connect school network, forming a cache hierarchy through cache interacting protocol in CERNET scope.

There were large number of Web cache replacement algorithms, GDSF (Greedy Dual Size Frequency), GDSize (Greedy Dual Size), LFU (Least Frequently Used), LRU (Least Recently Used) were those representatives.

Performance of cache replacement strategy relied on the practical properties of Web access; there was none of recent strategy performed well under different access condition. Ways to make the replacement strategy adaptive to different Web access properties had been a great concern.

2.2 WEB PRE-FETCHING TECHNIQUE

Pre-fetching was aimed to hide communication latency, and could be classified into the following models:

(1) Pre-fetching algorithm based on access probability

Numbers of literature implemented Web pre-fetching based on the pre-fetching algorithm of Markov process. Traditional Markov chain model was a simple and effective prediction model, but the prediction accuracy was relatively low. XING Yong-Kang et al. [3] stated and built a multi-Markov chain user browsing prediction model based on classification of user. The works [4, 5] used hidden Markov process to raise prediction accuracy. Su Zhong et al. [6] used N-gram prediction model to predict the Web access request might be occurred in the future.

(2) Pre-fetching algorithm based on data mining

According to historical and recent access data, with the use of data mining technique, predicted possible future behavior of user, in order to prefetch the related Web page. Data in user's data buffer could be used as historical data for data mining. The works [7, 8] mined interest relation rules using data mining, applied those rules as pre-fetching foundation to predict pages. Pre-fetching based on data mining was more suitable for user personalize recommendation.

(3) Pre-fetching algorithm based on Web semantics

Zhu et al. [9] proposed to extract features of user session, then classified semantically. While responding user's request, server would calculate user accessing path, and the distances between user and each category center, in order to confirm the type of session. According common features of session category, predicted access-possible documents, pre-transmit them to client side. T.I.brahim et al. [10] introduced semantic web page pre-fetching with the use of neural network. By extracting hyperlinks in web page, using keywords described in hyperlink text as input of neural network; output of neural network would be the basis for pre-fetching. Browsing path of user would be the training sample for the learning of neural network

(4) Pre-fetching algorithm based on network performance

JIN et al. [11] studied Web intelligent boost technique based on RTT (round trip time) and other network performance index. Proposed an intelligent pre-fetching control technique and new cache replacement method based on the service analysis on web proxy server and measurement of network RTT. R.P .Klemn et al. [12] designed and implemented pre-fetching agent WebCompanion in Java. The pre-fetching algorithm was based on an estimated RTT, only applied pre-fetching for those Web object with relative long respond time and low resource usage.

(5) Pre-fetching algorithm based on popularity

E.P. Mareatos et al. [13] proposed a classical Top-10 method, basic concept was to find out the Top-10 popular document (named it asTop-10) on server periodically; when clients sent request, server would send the Top-10 to them. X.Chen et al. [14] implemented Web pre-fetching using the popularity-based PPM model. Declared four levels for popularity of URLs' access mode, then constructed a prediction tree using these models for pre-fetching. Through post-processing, reconstruction and other methods, reduced space usage of PPM algorithm, and raised the accuracy of prediction.The recent study of cache and pre-fetching development was to discover a unified cache-and-pre-fetching model according to Web object browsing features, raise adaptability of cache strategy and pre-fetching algorithms, achieved a better performance in a reasonable time and space usage.

(6) Pre-fetching system

Griffioen et al. [15] had studied the pre-fetching and cache model of file system, assuming that cache and pre-fetching shared the same cache space. Result had shown that cache pre-fetching-unifying model could improve performance of cache system. Z.Jiang et al. [16] cooperated the use of server sides with client sides to implement pre-fetching, studied pre-fetching mechanism based on network work load and waiting time of users. The paper studied the pre-fetching control problem also, but not yet considered the competition of cache space between pre-fetching mechanism and cache mechanism. Pei cao et al [17] studied the combination of cache and pre-fetching in file system. Proposed and analyzed two combination strategy: proactive strategy and conservative strategy; through simulation test, these two method could reduce application latency for more than 50%.

N.J.Tuah et al. [18] combined pre-fetching and cache to raise the utilization rate of memory. Deduct the calculation formula of pre-fetching threshold under two interacting models of cache and pre-fetching. Then it made a conclution that limitation for the number of pre-fetching object was no longer needed once the access probability met the close value. In the two used models, improvement of access time had been considered, while the whole system resource usage had not.

Yang et al. [19] retained a fixed cache space for pre-fetching object in unified system of cache and pre-fetching while pre-fetching Web object. pre-fetching model in the system was constructed from the mined visit path from log file. Real data declared that cache performance in cache pre-fetching unified system was better than cache-only system.

3 Our Pre-fetching System

While update frequency of network resources kept raising, performance improvement by cache was no longer obvious. Numbers of studies had shown that pre-fetching would gain great benefit only when cooperating with suitable cache algorithms.

3.1 PRE-FETCHING SYSTEM MODEL

As shown in the figure 1, pre-fetching system model concludes several components: log file processing, relation (sequence) pattern mining, relation rules, buffer management based on prediction, cache buffer, pre-fech queue, log file.

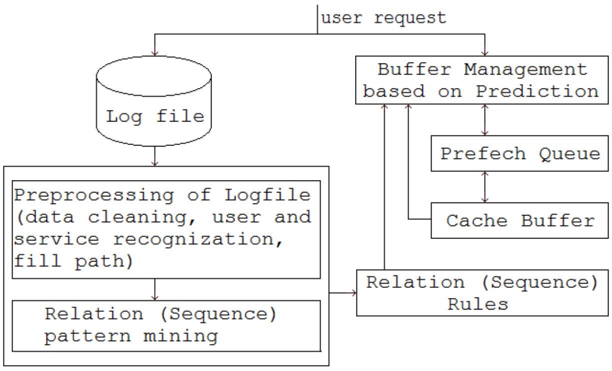

user request
Log file
Buffer Management based on Prediction
Prefech Queue
Cache Buffer
Preprocessing of Logfile (data cleaning, user and service recognization, fill path)
Relation (Sequence) pattern mining
Relation (Sequence) Rules

FIGURE 1 Pattern Mining Model in Web

4 Log File Processing

Quality of data preprocessing was closely related to the efficiency and result of modeling mining. Data preprocessing included data cleaning, path filling, and users, session, and services recognition.

(1) Data cleaning

The main task for data cleaning was to remove redundant data that was not related to pattern mining in data source. At server side, the page requested by user presented as numbers of .html (or .asp/.jsp) files, numbers of image file or script files. What we were going to study was the relation between pages and pages, which was independent from images or scripts in them. And log file would record all transmission of files, hence, files could be removed in order to reduce total size of data.

(2) Path filling

Part of the user accessing information in Web log might be incomplete. For instance, while there was no direct chain relation between the recent accessing page and previous requested page, and the requesting page was a recently requested page, so we could assume that the user was using the back button of browser; which invoking the local cached page, this should not be recorded in server side log. Heuristic rules were be used for the situation, filling the user path with inference of network topology structure.

(3) Users recognition

That was hard to identify a user by only user access log. With respect to practical experimental environment, network environment under CERNET was the only consideration, using fixed IP as a sign to specify user; and assumed that a user was corresponding to one and only one IP, thus recognized different users.

(4) Session recognition

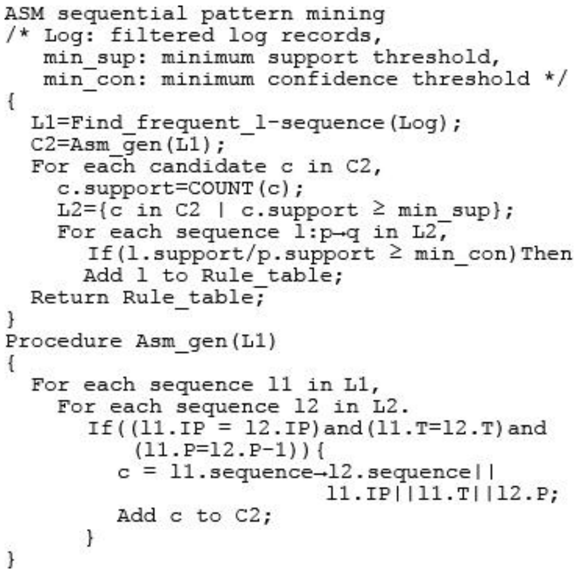
Session referred to consecutively requested pages by a user, pages accessed by different users belonged to different sessions. Session recognition was to separate user accessing record into plural independent session records.

(5) Service recognition

Service recognition was to transfer user session into smaller, more accurate, and relatively semantic user accessing service, which was the page sequential for user to visit specific information. While service recognition, filtration of part of the service should be applied; for example, some users just visited the web page but was not interested about web content, services liked this could be filtered, avoided too much relations generated.

5 Relation (Sequence) Pattern Mining

After the filtration of log file, we might start using the ASM sequential mining algorithm to generate rule table, then decided cache-prefetching strategy based on those rules generated. Figure 2 is the ASM sequential pattern mining algorithms:



```
ASM sequential pattern mining
/* Log: filtered log records,
   min_sup: minimum support threshold,
   min_con: minimum confidence threshold */
{
  L1=Find_frequent_1-sequence(Log);
  C2=Asm_gen(L1);
  For each candidate c in C2,
    c.support=COUNT(c);
    L2={c in C2 | c.support ≥ min_sup};
    For each sequence l:p→q in L2,
      If(l.support/p.support ≥ min_con)Then
      Add l to Rule_table;
  Return Rule_table;
}
Procedure Asm_gen(L1)
{
  For each sequence l1 in L1,
    For each sequence l2 in L2.
      If((l1.IP = l2.IP)and(l1.T=l2.T)and
          (l1.P=l2.P-1)){
        c = l1.sequence→l2.sequence||
                             l1.IP||l1.T||l2.P;
        Add c to C2;
      }
}
```

FIGURE 2 ASM sequential pattern mining

We would use a practical example to explain the algorithm. After data cleaning at the first stage, and then index each accessed pages of user, we could obtain an initial log as table 1 shown.

First of all, scanned through these logs, calculated browsing count (support threshold, 7th, 8th lines in the algorithm) of each page, and then recorded them by a 3-tuple group (IP, T, P). As shown in table 2, (202.116.32.46, 1, 3) of sequence 3 meant that user with IP 202.116.32.46 accessed the page with sequence number of 3 at the 3rd position in 1st service.

After table 2 was obtained, filtered those sequence with support larger than the minimum support (here assumed the minimum support be 2), obtained sequence 1, 2, 3, 5, 7, 8, 9, 10. After that, compared the 3-tuple group of each sequence pair (19th to 21st lines in the algorithm); if the pair met requirement, then put it into candidate set with sequence length of 2, as shown in table 3.

The ASM pattern mining algorithm would stop after the generation of candidate set with sequence length of 2 finished. After that, calculated the global confidence for each rule p→q, divided by confidence of p for each global confidence equal to p→q (11th line in algorithm). Assumed the minimum confidence be 1/2, removed the rule of global confidence (as the rule had been fulfilled), obtaining rule table 4 at last.

Notably, here we used local confidence for each user IP; for the same rule p→q, sum of each local confidence equal to the global confidence.

TABLE 1 Filtrated Log

IP	Visited Page Sequence
202.116.32.46	(1,2,3)(4,5)
202.116.32.47	(6,5)(1,2,3)
202.116.32.48	(7,8)(9,10)
202.116.32.49	(1,2)(9,10)
202.116.32.50	(5,1,3)
202.116.32.51	(7,8)(11,9,10)

TABLE 2 Candidate Set with Sequence Length 1

Sequence number	Support	(IP,T,P)
1	4	(202.116.32.46,1,1) (202.116.32.47,2,1) (202.116.32.49,1,1) (202.116.32.50,1,2)
2	3	(202.116.32.46,1,2) (202.116.32.47,2,2) (202.116.32.49,1,2)
3	3	(202.116.32.46,1,3) (202.116.32.47,2,3) (202.116.32.50,1,3)
4	1	(202.116.32.46,2,1)
5	3	(202.116.32.46,2,2) (202.116.32.47,1,2) (202.116.32.50,1,1)
6	1	(202.116.32.47,1,1)
7	2	(202.116.32.48,1,1) (202.116.32.51,1,1)
8	2	(202.116.32.48,1,2) (202.116.32.51,1,2)
9	3	(202.116.32.48,2,1) (202.116.32.49,2,1) (202.116.32.51,2,2)
10	3	(202.116.32.48,2,2) (202.116.32.49,2,2) (202.116.32.51,2,3)
11	1	(202.116.32.51,2,1)

TABLE 3 Candidate Set with Sequence Length 2

Sequence	Support	(IP,T,P)
1→2	3	(202.116.32.46,1,2) (202.116.32.47,2,2) (202.116.32.49,1,2)
1→3	1	(202.116.32.50,1,3)
2→3	2	(202.116.32.46,1,3) (202.116.32.47,2,3)
4→5	1	(202.116.32.46,2,2)
5→1	1	(202.116.32.50,1,2)
6→5	1	(202.116.32.47,1,2)
7→8	2	(202.116.32.48,1,2) (202.116.32.51,1,2)
9→10	3	(202.116.32.48,2,2) (202.116.32.49,2,2) (202.116.32.51,2,3)
11→9	1	(202.116.32.51,2,2)

TABLE 4 Rule Table

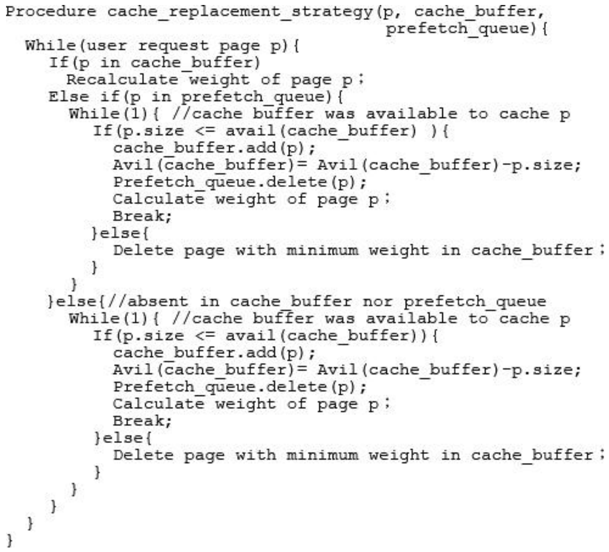
IP	Access rule	Local confidence	Global confidence
202.116.32.46	1→2	1/4	3/4
202.116.32.47	1→2	1/4	
202.116.32.49	1→2	1/4	
202.116.32.46	2→3	1/3	2/3
202.116.32.47	2→3	1/3	
202.116.32.48	7→8	1/2	1
202.116.32.51	7→8	1/2	
202.116.32.48	9→10	1/3	1
202.116.32.49	9→10	1/3	
202.116.32.51	9→10	1/3	

5.1 BUFFER MANAGEMENT BASED ON PREDICTION

There were two main processes for buffer management based on prediction:

1. Handled requested page in cache buffer from recent user
2. Handled pages which were going to be visited in pre-fetching Queue

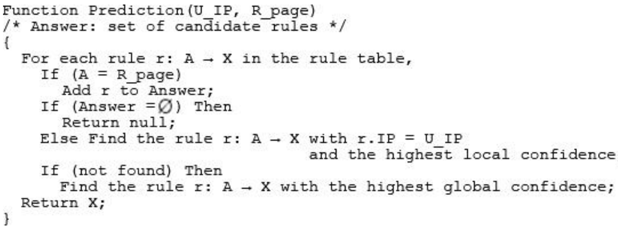
Handling requested page in cache buffer from recent user, practically was a kind of common cache buffer function, only once the user sent a page request, proxy server would first check the existence of requesting page in cache buffer, returned it to user if exists, sent page request to remote server if no existence. The algorithm is as follow as Figure 3:



```
Procedure cache_replacement_strategy(p, cache_buffer,
                                           prefetch_queue){
  While(user request page p){
    If(p in cache_buffer)
      Recalculate weight of page p;
    Else if(p in prefetch_queue){
      While(1){ //cache buffer was available to cache p
        If(p.size <= avail(cache_buffer) ){
          cache_buffer.add(p);
          Avil(cache_buffer)= Avil(cache_buffer)-p.size;
          Prefetch_queue.delete(p);
          Calculate weight of page p;
          Break;
        }else{
          Delete page with minimum weight in cache_buffer;
        }
      }
    }else{//absent in cache_buffer nor prefetch_queue
      While(1){ //cache buffer was available to cache p
        If(p.size <= avail(cache_buffer)){
          cache_buffer.add(p);
          Avil(cache_buffer)= Avil(cache_buffer)-p.size;
          Prefetch_queue.delete(p);
          Calculate weight of page p;
          Break;
        }else{
          Delete page with minimum weight in cache_buffer;
        }
      }
    }
  }
}
```

FIGURE 3 Cache replacement strategy

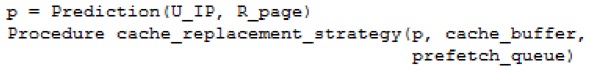
After the sequence pattern mining, a rule table had been obtained. Then the predicted buffer management would match according to the rule table specified by user IP; matching principle was to choose page with same IP and highest confidence in rule table (if that page was not in cache), if the one with equal IP could not be found, then chose the page with highest global confidence, as the page might be interested by other users [14], and the prefetching algorithm is shown as Figure 4:



```
Function Prediction(U_IP, R_page)
/* Answer: set of candidate rules */
{
  For each rule r: A → X in the rule table,
    If (A = R_page)
      Add r to Answer;
    If (Answer =∅) Then
      Return null;
    Else Find the rule r: A → X with r.IP = U_IP
                                and the highest local confidence;
    If (not found) Then
      Find the rule r: A → X with the highest global confidence;
  Return X;
}
```

FIGURE 4 Prediction algorithm

As the concept of the process was consistence to cache replacement strategy, therefore we could invoke the previous cache_replacement_strategies, difference in between was that the requesting page p was not the user clicked page, but the page predicted according to Rule Table. The algorithm for the whole process is Figure 5 as follow:



```
p = Prediction(U_IP, R_page)
Procedure cache_replacement_strategy(p, cache_buffer,
                                     prefetch_queue)
```

FIGURE 5 Invoking cache_replacement_strategy for prediction

In the cache replacement algorithm, the main concept was to remove pages with lowest weight when cache buffer was not big enough; and the weight setting should be related to visit count, size, staying time for cache and plural factors of the page, thus the formula for hypothetical weight as follow:

In the cache:

Wn(p) = L+(Wn-1(p)*T_stay/(T_cur-T_ref))/size(p)

In the above formula:

Wn(p) was the weight of page p,

L was accommodation coefficient, purpose was to avoid cache pollution

F (p) was the usage frequency for page p,

Wn-1(p) was the original weight of p,

T_stay was staying time of p,

T_cur was current time referencing p,

T_ref was the time of last reference of p,

size(p) was size of p

In prefetch queue:

Wn(p) = Pro(p)/size(p)+Wn-1(p)*1/(T_cur-T_pre)

In the above formula:

Wn(p) was the weight of page p,

Pro(P) was confidence of p

size(p) was size of p

Wn-1(p) was the original weight of p,

T_cur was current time referencing p,

T_ref was the time of last reference of p,

Formulas above followed one principle: new weight of page was related to original weight, also, shorter staying time, shorter user access interval, greater access probability; smaller page, greater weight.

5 Conclusions

We mainly proposed a Cache-Prefetching System model based on Web Data Mining, and made detail introduction of implementation of each part, included the way to generate rule table from server logs, page replacement strategy of cache buffer manager, and weight calculation of each pages. Prospect was to implement those concepts, and compared the efficiency to each cache scheduling algorithms like LRU and LFU. With the use of practical result, raise caching effectiveness by kept adjusting page weight formula and replacement strategy of cache. Such technique could be applied in personnel, educational, and official information managing system in database of educational scope. Accessing speed of educational information system for numerous teachers and students, benefit high school personnel management, and also the effective scientific structuralize educational management.

Acknowledgments

This work was supported by Major Program of National Natural Science Foundation of China (61133014, 61272415, 61272413), the CEEUSRO project of Guangdong province, China (2012A080102007, 2011J430004, 2011B090400324, 2010A011200038, 2012B040305008), Science and Technology Planning Project of Guangzhou city (2013Y2-00071), The Project for Engineering Research Center of Guangdong Province (GCZX-A1103), JNU innovation project (JNU-J-CXGCWK2014006) and the Fundamental Research Funds for the Central Universities (21613336).

References

[1] Kroeger T M, Long D D E, Mogul J C 1997 Exploring the bounds of Web latency reduction form caching and perfecting *Proceedings of the USENIX Symposium on Internet Technologies and Systems California USENIX Association* 13-22

[2] Chen Xin, Zhang Xiao dong 2003 Accurately modeling workload interactions for deploying prefetching in web sevrers *Proceeedings 2003 International Conference on Parallel Processing* 427-35

[3] Xing Yong-Kang, Shao-Ping M A 2003 Modeling user navigation sequences based on multi-markov chains *Chinese Journal of Computers* **26**(11) 1510-17

[4] Wang Shi, Gao Wen, Li Jin-Tao, Huang Tie-Jun 2001 Mining Interest Navigation Patterns Based on Hidden Markov Model *Chinese Journal of Computers* **24**(2) 152-57

[5] Xu Huan-Qing, Wang Yong-Cheng A Web Pre-fetching Model Based on Analyzing User Access Pattern *Journal of Software* **14**(6) 1142-47

[6] Su Zhong, Shao-Ping M A, Yang Qiang, Zhang Hong-Jiang 2002 An N-Gram Prediction Model Based on Web-Log Mining *Journal of Software* **13**(l) 136-41

[7] Xu Bao-Wen, Zhang Wei-Feng 2001 Applying Data Mining to Web Pre-Fetching *Chinese Journal of Computers* **24**(4) l-7

[8] Yang Q, Zhang H H 2003 Web-log mining for predictive Web caching *IEEE Transactions on knowledge and Data Engineering* **15**(4) 1050-53

[9] Zhu Pei-Dong, Lu Xi-Chengg 1999 Traffic Smoothing of WWW Presending *Chinese Journal of Computer* **22**(6) 668-71

[10] Xu C, Ibrahim T I 2004 A Keyword-Based Semantic Prefetching Approach in Internet News Services *IEEE Transactions on knowledge and Data Engineering* **16**(5) 601-11

[11] Jin Zhi-Gang, Zhang Gang, Shu Yan-Tai 2001 Intelligent Prefetch and Cache Techniques based on Network Performance *Journal of Computer Research & Development* **38**(8) 1000-4

[12] Klemn R P 1999 Web Companion a friendly client-side Web prefetching agent *IEEE Transactions on knowledge and Data Engineering* **11**(4) 577-94

[13] Mareatos E P, Chronaki C E 1998 A top-10 approach to prefetching the Web *Proceedings of the Eighth Annual Conference of the Internet Society* Geneva

[14] Chen X, Zhang X 2003 A popularity-based prediction model for Web prefetching *Computer* **36**(3) 63-70

[15] Griffioen J, Appleton R 1994 Reducing file system latency using a predictive approach *Proc of USENIX Summer Conefrenee* 197-207

[16] Jiang Z, Kleinrock L1998 Web prefrtching in a mobile environment *IEEE 1nt Conference on Communications* **5**(5) 25-34

[17] Pei Cao, Edward W, Feltn Anna R 1995 A Study of Integrated Prefetching and Caching Strategies *Proc of the ACM SIGMETRICS Conference on Measurement and Modeling of ComputerSystems*

[18] Tuah N J, Kumar M, Venkatesh S 2003 Resource aware Speculative prefetching in Wireless Networks *Wireless Networks* **9**(1) 61-72

[19] Yang Qiang, Zhang Henry Hanning 2001 Integrating Web Prefetching and Caching Using Prediction Models *World Wide Web* **4**(4) 299-321

[20] Padmanabhan V N, Mogul J C 1996 Using predictive prefetching to improve World Wide Web latency *Proceedings of the ACM SIGCOMM’96 Conference* 22-36

[21] Jiang Z, Kleinrock L 1998 An adaptive network prefetch seheme *IEEE Jounral on Selected Areas in Comm-unieations* **16**(3) 358-68

[22] Jiang Z, Kleinrock L 1998 Web prefetching in a mobile environment *IEEE Int Conference on Communications* **5**(5) 25-34

[23] Palpanas T, Mendelzon A 1998 Web Prefetching using Partial match Prediction *Technical Report CSRG-376 Dept.of CS,Univ.of Toronto*

[24] Mahanti A, Eager D, Williamson C 2000 Temporal locality and its impact on Web proxy Cache Performance *Performance Evaluation* **42**(2-3) 187-203

[25] Chen X, Zhang X 2003 A Popularity-based Prediction model for Web Prefetching *Computer* **36**(3) 63-70

[26] Nanopoulos A, Katsaros D, Manolopoulos Y 2003 A data mining algorithm for generalized web Prefetching *IEEE Transactions on Knowledge and Data Engineering* **5**(5) 1155-69

Authors

Li Yongbiao, China

University studies: Jinan University
Scientific interest: Information systems

High resolution remote sensing image classification based on particle swarm optimization and support vector machine

Buyi Li, Chongjing Deng, Shuang Li*

International School of Software, Wuhan University, Luoyu Road 37#, Wuhan, China, 430079

**Corresponding author e-mail: sli@whu.edu.cn*

Received 1 November 2014, www.cmnt.lv

Abstract

Many algorithms have been developed for image classification and support vector machine (SVM) is a kind of supervised classification that has been widely used recently. However, the accuracy of a SVM classifier heavily depends on the selection of a right kernel model and appropriate parameter. In this paper, a comparative analysis of the impact of four kernels (linear kernel, polynomial kernel, radial basis function kernel and sigmoid kernel) on the accuracy of SVM classifiers is conducted. Moreover, the Particle Swarm Optimization (PSO) is used to search for the optimum parameters for each kernel function in order to improve the classification accuracy of SVM classifiers. Our experiments for optimizing the kernel function parameters and assessing the robustness of SVM classifiers were carried out with classifications of QuickBird-2 images over Wuhan, China for monitoring urban land cover/land use information. The experimental results indicate that the polynomial kernel outperforms the other kernels in classifying high resolution remote sensing image. The sigmoid kernel performs worse than any other kernels. Our findings also suggest that selected parameter by PSO will improve the classification accuracy, especially for radial basis function kernel.

Keywords: high resolution remote sensing image, support vector machine classification, parameter optimization, particle swarm optimization

1 Introduction

The classification of land use and land cover (LULC) from remotely sensed imagery is a challenging topic due to the complexity of landscapes. Numerous classification algorithms have been proposed especially since more and more remote sensing images with various spatial and spectral resolutions are sent back to the earth. Among the most popular algorithms, Support Vector Machine (SVM) is a new machine learning method based on statistical learning theory, which can solve the classification problem with small sampling, non-linear and high dimensions [1].

It is well-known that the performance of SVM depends on the training features, kernel type and its corresponding parameters [2-4]. The kernel function in SVM is used to convert non-linear separating boundaries into linear ones by mapping the input data into a high-dimensional space. Thus, determine the kernel type and kernel parameters are important for image classification accuracy. There are many kinds of support vector kernels such as the linear kernel, the polynomial kernel, the radial basis function kernel, etc. For the kernel type selection, Pal (2002) suggested that the radial basis function kernel achieved higher accuracy than linear kernel, polynomial kernel and the sigmoid kernel [5]. Villa et al. (2008) concluded that polynomial kernel outperformed the Gaussian Kernel in remote sensing image classification [3]. Kavzoglu and Colkesen (2009) indicated that radial basis function kernel performed better than polynomial kernel in land cover classification [4]. However, they also indicated that further research should be conducted on the effects of kernel type and their parameters on classification accuracy.

For kernel parameter optimization techniques, the traditional way is grid search with cross validation. However, the grid search is time consuming as the model needs to be evaluated at many grid points for each parameter set. In recent years, the artificial intelligent algorithms are employed in SVM parameter optimization, i.e. genetic algorithm (GA), simulated annealing (SA), particle swarm optimization (PSO). The GA updates the population by crossover and mutation operations to generate optimal parameters. The SA technique can also be applied to ensure that the global optimum of parameter combinations. However, these methods obtain the optimal parameters from the population evolution iteratively, which require much training time in SVM classifier. Inspired by social behavior of bird flocking or fish schooling, PSO is proposed by Kennedy and Eberhart in 1995 [6]. Through the competition and collaboration among the population, each particle in the swarm can dynamically adjust its velocity according to its own and its companion’s experience and finally can find the best position to land. Compared with other intelligent algorithms, PSO demonstrates its high efficiency, easy implement and powerful both global and local exploration abilities in parameter optimization in support vector machine [7-12]. However, the ACO algorithm is only used to optimize the RBF kernel. According to the above analysis, in this research, the proposed PSO-SVM model is applied for classification of remote sensing image from Quickbird-2 sensor, in which PSO is used to determine optimized parameters of support vector machine with different kernels. The remainder of this paper is organized as follows. Section 2 describes the basic idea of support vector machine and Section 3 introduces the recommended PSO and the optimization procedure for SVM kernel parameters. Section 4 testifies the performance of the proposed method and presents the analysis for the experimental results. Finally, conclusions are made in Section 5.

2 Support vector machines and its kernels

Consider data set

$$\{(x_1,y_1),...,(x_i,y_i),...,(x_N,y_N),\ y_i\in(1,-1)\},$$

where N is the number of samples, x_i is the training sample, y_i is the class label of x_i. Optimum hyper plane is used to maximize the margin between classes.

The hyper plane is defined as

$$w\cdot x+b=0 \tag{1}$$

where x is a point lying on the hyper plane, w determines the orientation of the hyper plane, b is the bias that indicates the distance between hyper plane and the origin. For the linearly separable case, the hyper plane is defined as

$$y_i\left(w\cdot x_i+b\right)\geq 1 \tag{2}$$

As the margin width between both bounding hyperplanes equals to $2/\left(\|w\|^2\right)$, the constraint optimization model of soft margin based SVM is as follows:

$$\min_{w,b,\xi}\frac{1}{2}w^2+c\sum_{i=1}^{l}\xi_i \\ s.t.\ \ y_i\left(\left(w\cdot x_i\right)+b\right)\geq 1-\xi_i;\ \xi_i\geq 0,\ i=1,2,...,l \tag{3}$$

where c is the penalty parameter which allows striking a balance between two competing criteria of margin maximization and error minimization, whereas ξ_i is the slack variable which indicate the distance of the incorrectly classified points from the optimal hyper plane. The larger the c value, the higher the penalty associated to misclassified samples.

To solve non-linear classification tasks, a nonlinear function $\phi(x)$ is usually employed to map the input space to a higher dimensional feature space. Thus, the input point x can be represented by $\phi(x)$ in high-dimensional space. The time-consuming computation of $\phi(x)\cdot\phi(x_i)$ is reduced by using a kernel function $K\left(x_i,x_j\right)=\phi\left(x_i\right)^T\phi\left(x_j\right)$. Thus, the classification decision function is defined as:

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{N}\alpha_i y_i K\left(x_i\cdot x\right)+b\right) \tag{4}$$

where $\operatorname{sgn}(\cdot)$ is the sign function, $K(\cdot)$ is the kernel function and the magniude of α_i is Lagrange multiplier. A multiplier exits for each training data instance and data instances corresponding to non-zero α_i are support vectors.

The typical SVM kernels include linear kernel function, polynomial kernel function, radial basis kernel function and sigmoid kernel function. They are defined as follows:

Linear kernel function

$$K\left(x,x'\right)=\left(x,x'\right)$$

Polynomial kernel function

$$K\left(x,x'\right)=\left(\gamma\left(x,x'\right)+r\right)^d,\gamma>0$$

Radial basis function

$$K\left(x,x'\right)=e^{\gamma\|x-x'\|^d},\gamma>0$$

Sigmoid kernel function

$$K\left(x,x'\right)=\tanh\left(\gamma\left(x,x'\right)+r\right),\gamma>0$$

Generally d is set to be 2 since the kernel value is related to the Euclidean distance between the two samples [13]. r is set to be 0 [14]. For the linear kernel function, only the penalty parameter c in SVM is needed for optimization. For the polynomial kernel function, radial basis function and sigmoid kernel function, the parameters (c,γ) should be set properly. c is the penalty parameter and γ is related to the kernel width.

3 Parameter optimization by particle swarm intelligent

In standard PSO algorithm, the particle swarm starts with the random initialization of a population, and each particle in the search space is characterized by two factors: its velocity and position. The velocity and position vectors of the particle $i\left(i=1,2,...,n\right)$ in d-dimensional space can be represented as $v_i=\left(v_{i1},v_{i2},...,v_{id}\right)$ and $x_i=\left(x_{i1},x_{i2},...,x_{id}\right)$, respectively. Then, the new velocity and position of particle i for the next generation in d-dimensional subspace is calculated as follows:

$$v_i(t+1)=\omega v_i(t)+c_1r_1\left(p_{best}(t)-x_i(t)\right)+c_2r_2\left(g_{best}(t)-x_i(t)\right) \tag{5}$$

$$x_i(t+1)=x_i(t)+v_i(t) \tag{6}$$

where $v_i(t)$ represents the previous velocity and its value is limited in the range of $\left[-V_{\max},V_{\max}\right]$. $p_{best}(t)$ is the particle's personal best position obtained so far at t *-th* generation and this part encourages the particles to move toward their own best position found so far. $g_{best}(t)$ is the global best position obtained so far by all particles and this part always pulls the particles toward the global best particle. c_1 and c_2 are constants known as acceleration coefficients which determine the relative influence of the social and cognition components. r_1 and r_2 are two independent random number uniformly distributed in the range of (0, 1). ω is the inertia weight that controls the impact of particle's previous velocity on its current generation.

The fitness function is used to guide the direction of search. As the classification accuracy is the object of our study, the recognition rate (RR) is used for fitness function, which is defined as follows:

$$RR=\frac{n_{correct}}{n_{total}}\times 100\% \tag{7}$$

where n_{corect} is the number of corrected classified samples, n_{total} is the total number of samples.

4 Experiments and results

The original image is shown in Figure 1. The image size is 400*400. The image is classified into seven classes, i.e. water, grass, bare land, blue roof, red roof, road, trees.

In the standard SVM, c=2 and g=0.125, which is according to the experience. Only the spectral features are used in SVM classification, i.e. Blue, Green, Red and Near-Infrared band. The results of different kernels are shown in figure 2 (a)-(c). The result of the sigmoid kernel is not shown here as it describes only one class the grassland.

From Figure 2(a)-(c), it can be seen that polynomial kernel performs better than linear and RBF kernels. Many bare lands are misclassified into roads in the result of linear kernel SVM classifier. The result of RBF is very bad as most of study area is misclassified into blue roof.

In the PSO-SVM method, the results of different kernels are shown in figure 2 (d)-(f). From figure 2 (d)-(f), it can be seen that less bare soil is misclassified into roads for polynomial kernel. The result of RBF kernel is improved greatly by PSO. The improvement of linear kernel result is not obvious.

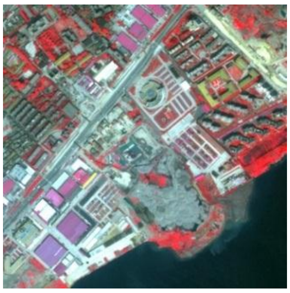
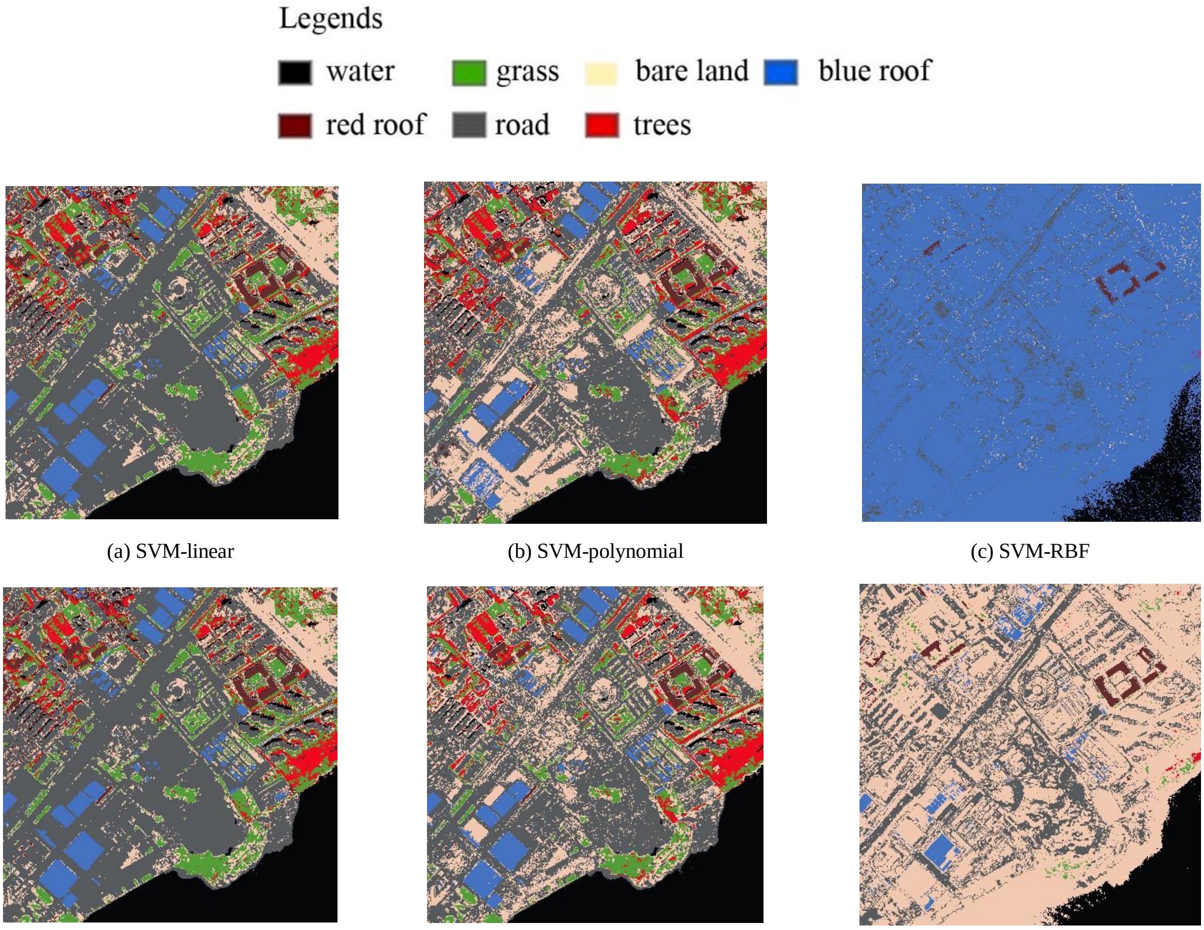


FIGURE 1 The original image



(a) SVM-linear

(b) SVM-polynomial

(c) SVM-RBF

(d) PSO-SVM-linear

(e) PSO-SVM-polynomial

(f) PSO-SVM-RBF

FIGURE 2 The classification results of different SVM methods

To further compare the results of different kernels, we compute the Producer's accuracy, user's accuracy, overall accuracy and kappa coefficient for the classified images. The producer's accuracy refers to the probability that a certain land-cover of an area on the ground is classified as such, which is the complementary of omission error. The user's accuracy refers to the probability that a pixel labeled as a certain land-cover class in the map is really this class, which is the complementary of commission error. The producer's accuracy and user's accuracy for any given class typically are not the same. From table1 and table 2, for SVM-Linear, an estimate for the producer's accuracy of bare land is 73%, while the user's accuracy is 79%. As a producer of classification, only 73% of all the bare land as such. As a user, roughly 79% of all the pixels classified as bare land are indeed bare land on the ground. As the producer's and user's accuracy are computed based on the diagonal of confusion matrix, the Kappa coefficient, which is calculated by all the values in the confusion matrix, is used for accuracy assessment. From table 3, it can be seen that, polynomial kernel outperforms other kernels. The PSO-SVM improves the results of SVM. For RBF kernel, the improvement of PSO-SVM is most obvious. The accuracy assessment further demonstrate the proposed PSO-SVM improve the results of classification. The improvements of different kernels by PSO-SVM are different.

TABLE 1 Producer's accuracy of classified image

	SVM			PSO-SVM		
	Linear	**Poly**	**RBF**	**Linear**	**Poly**	**RBF**
Water	90%	92%	87%	94%	95%	92%
Grassland	66%	69%	14%	73%	82%	19%
Bare land	73%	74%	9%	84%	92%	99%
Blue roof	62%	65%	97%	75%	90%	45%
Red roof	56%	59%	18%	68%	82%	79%
Road	92%	90%	19%	90%	91%	71%
Trees	68%	73%	11%	66%	92%	23%

TABLE 2 User's accuracy of classified image

	SVM			APSO-SVM		
	Linear	**Poly**	**RBF**	**Linear**	**Poly**	**RBF**
Water	94%	94%	98%	98%	100%	100%
Grassland	83%	84%	95%	96%	91%	89%
Bare land	79%	84%	49%	67%	73%	28%
Blue roof	97%	98%	19%	100%	100%	93%
Red roof	97%	97%	100%	97%	98%	100%
Road	41%	42%	58%	48%	78%	86%
Trees	77%	81%	88%	92%	94%	88%

TABLE 3 Overall accuracy (OA) and kappa coefficient (KC) of classified image

	SVM			SVM-PSO-MF		
	Linear	**Poly**	**RBF**	**Linear**	**Poly**	**RBF**
OA	72%	75%	36%	75%	89%	61%
KC	0.68	0.70	0.26	0.71	0.87	0.55

5 Conclusions

Support vector machines (SVM) are receiving increasing attention in remote sensing applications, such as image classification, land cover/land use change detection and so on. However, SVM is very sensitive to the parameters setting. In this study, a comparative analysis of the impact of four kernels (linear kernel, polynomial kernel, radial basis function kernel and sigmoid kernel) on the accuracy of SVM classifiers is conducted. Moreover, the Particle Swarm Optimization (PSO) is used to search for the optimum parameters for each kernel function in order to improve the classification accuracy of SVM classifiers. The experimental results show that the result of SVM with polynomial kernel is best while the result of sigmoid kernel is worst. The PSO improves the classification accuracy of RBF kernel most significantly, while the accuracy of linear kernel is not obvious. The PSO-SVM-Poly outperforms other methods. Of course, the experiment is limited. More experiment would be conducted in our future study, especially for feature selection in SVM classifiers.

Acknowledgments

This research was supported by the Natural Science Foundation of Hubei Province (2013CFB285), the Fundamental Research Funds for the Central Universities, Wuhan University (2042014KF0054) and the State 973 project (2012CB719901) in China.

References

[1] Buciu I, Kotropoulos C, Pitas I 2006 Demonstrating the stability of support vector machines for classification *Signal Processing* **86** 2364-80

[2] Huang C, Davis L, Townshend J 2002 An assessment of support vector machines for land cover classification *International Journal of Remote Sensing* **23**(4) 725-49

[3] Villa A, Fauvel M, Chanussot J, Gamba P, Benediktsson J A, Gradient optimization for multiple kernel's parameters in support vector machines classification *Geoscience and Remote Sensing Symposium 7-11 July 2008* IGARSS 2008 IEEE International **4** 224-7

[4] Kavzoglu T, Colkesen I 2009 A kernel functions analysis for support vector machines for land cover classification *International Journal of Applied Earth Observation and Geoinformation* **11**(5) 352-9

[5] Pal M 2002 Factors influencing the accuracy of remote sensing classifications: a comparative study *University of Nottingham*

[6] Kennedy J, Eberhart R 1995 Particle swarm optimization *Proceedings of IEEE international conference on neural networks* 1942-48

[7] Wu Q, Wu S, Liu J 2010 Hybrid model based on SVM with Gaussian loss function and adaptive Gaussian PSO *Engineering Applications of Artificial Intelligence* **23**(4) 487-94

[8] Bazi Y, Melgani F 2007 Semisupervised PSO-SVM regression for biophysical parameter estimation *Geoscience and Remote Sensing, IEEE Transactions on* **45**(6) 1887-95

[9] Mao Y, Xia Z, Yin Z, Sun Y X, Wan Z 2007 Fault diagnosis based on fuzzy support vector machine with parameter tuning and feature selection *Chinese Journal of Chemical Engineering* **15**(2) 233-9

[10] Huang C L, Dun J F 2008 A distributed PSO–SVM hybrid system with feature selection and parameter optimization *Applied Soft Computing* **8**(4) 1381-91

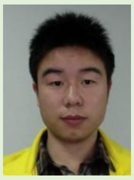
[11] Lin S W, Ying K C, Chen S C, Lee Z J 2008 Particle swarm optimization for parameter determination and feature selection of support vector machines *Expert Systems with applications* **35**(4) 1817-24

[12] Melgani F, Bazi Y 2008 Classification of electrocardiogram signals with support vector machines and particle swarm optimization *Information Technology in Biomedicine IEEE Transactions on* **12**(5) 667-77

[13] Wu K P, Wang S D 2009 Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space *Pattern Recognition* **42**(5) 710-17

[14] Chang C C, Lin C J 2011 LIBSVM a library for support vector machines *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(3) 27

Authors

	**Li Buyi, 1991, Xinyang, China** **Current position, grades:** Graduate Stundet **University studies:** Wuhan University **Scientific interest:** image processing and big data mining **Experience**: From 2010 to 2014, studied in Wuhan University and got the Bachelor Degree in 2014, from 2014 to now, studies in Wuhan University for the Master Degree, honorable mention in Mathematical Contest In Modeling (MCM) in 2013, took part in the developing of digital library of China from 2012 to 2014.
	Chongjing Deng, 1992, Changde, China **Current position, grades:** Graduate Student **University studies:** Wuhan University **Scientific interest:** remote sensing image processing **Experience:** From 2010 to 2014, studied in Wuhan University and got the Bachelor Degree in 2014, from 2014 to now, studies in Wuhan University for the Master Degree.
	Li Shuang, 1982, Wuhan, China **Current position, grades:** Associate professor **University works:** Wuhan University **Scientific interest:** image processing, remote sensings image classification and interpretation **Experience:** in 2010 got Doctoral degree from State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing in Wuhan Univerity.

Pattern recognition systems in the problems of automatic person identification using the passport data

Y Amirgaliyev, R Yunussov*

Suleiman Demirel University, Information technologies and computations Instutute, Kazakhstan

**Corresponding author's e-mail: yunussov@gmail.com*

Received 12 November 2014, www.cmnt.lv

Abstract

The work describes the implementation of modern technology for remote sensing and data processing in the area of human activities concerned to the security provision, based on usage of pattern recognition algorithms and neural networks. The Republic of Kazakhstan State Identities and Passports were used as the basis; The ICAO 9303 MRZ Standard was used. Obtained stable recognition model for identification of known passport types, and MRZ section decoding.

Keywords: neural networks, pattern recognition, raster data, image processing

1 Introduction

Security systems, based on usage of modern hardware and software solutions have had a huge spread abroad as well as in our country. It is necessary to note the state program "Secured City" [1], which realized by establishing video cameras in large cities of RoK. Passengers registration systems in airports [2], based on face recognition algorithms. Multiple penetration of such systems in different areas of human activity shows the huge potential of researches and development in this knowledge area. And growing problems in the area of security systems, that the governments and corporate sectors face, necessitated further development of more effective method of problem solutions. Proposed model in this paper allowsto automate process of person registration by automatically recognizing it's passport data and extracting the meta information from passport using OCR methods based on neural networks.

Application of similar models was proposed by Young-Bin Kwon and Jeong-Hoon Kim [3]

2 Overview of the study area

The problem of automated passport recognition is solved under different tasks and business processes, where it is necessary to improve the work throughput of person identification process, like – migration processes in the border control area, the logging of visitors for the secured area and so on. Currently almost all countries have accepted ICAO 9303 standard of passport template, that should have the Machine Readable Zone(MRZ), to allow for automated recognition processes to be implemented. And this standards simplifies the algorithms creation of extracting the meta information from documents acquired by scanners on other optical sensors.

Under scope of this work the problems of automated national identities and passports of Republic of Kazakhstan recognition and extraction are surveyed. Currently there are 4 types of national ID presented in RoK, that have MRZ:

1 National ID Type A

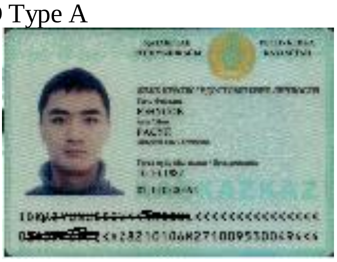


FIGURE 1 A type document

2 National ID Type B

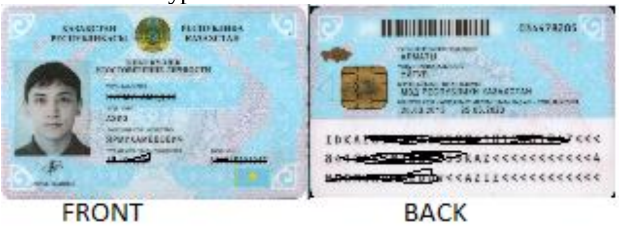


FIGURE 2 B type document

3 National ID Type C

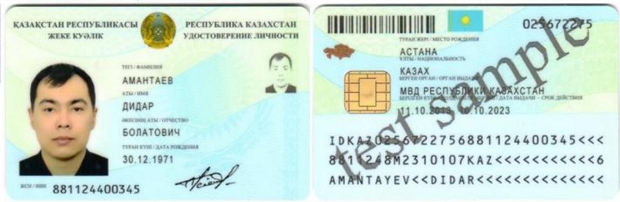


FIGURE 3 C type document

4 National Passport of the RoK citizen

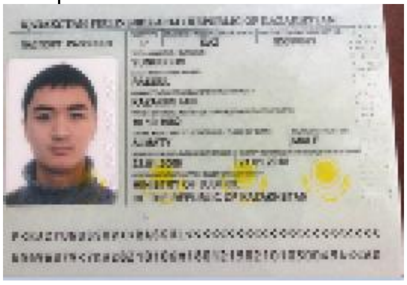


FIGURE 4 Passport

Depending of the document type the image can contain person face picture and MRZ, and also the type of the MRZ (3 lines or 2 lines, described by ICAO 9303).

Thus this work surveys next problems:

1. Identification of document type by analysis of image;
2. Determining the possible rotation angle of document against horizontal scanning plane and implementing the de-skew process;
3. Information extraction from the document – identification of human face and MRZ for further processing.

Under scope of problem the scanner Fujitsu 65fi (format A6) was used. The scanner area is bigger than the possible different passport sizes, and it imposes additional conditions of variable environment of scanned area, where the passport borders should be found.

Number of algorithms were used to solve the problem:

1. The borders extraction algorithm based on the gradient direction analysis [4];
2. Strict lines detection algorithm [5];
3. Template matching algorithm on the basis of matrices correlation [6];
4. Geometry topology comparison;
5. Face detection algorithm on the image [7];
6. MRZ recognition on the basis of neural network OCR system.

3 Adopting relevant technology

This paper surveys the problems of information extraction from images and the ways of improving the quality by the implementation of artificial neural networks.

Among the existing approaches for the OCR there are two base methods – the template matching and the invariant topology of character extraction. Current work uses the first approach due to the simplicity of implementation of ANN classifier for the one type of font used under the ICAO 9303 standard.

The image analysis for solving the problems surveyed under this paper consists of 5 stages, depicted on Figure 5.

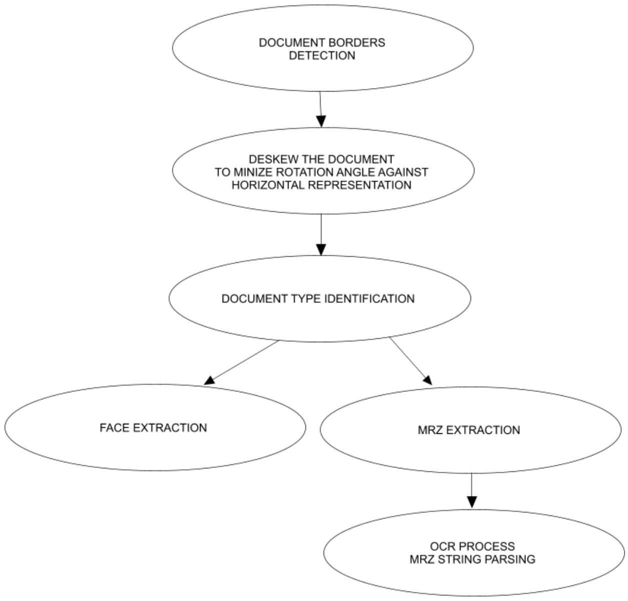

DOCUMENT BORDERS DETECTION
DESKEW THE DOCUMENT TO MINIZE ROTATION ANGLE AGAINST HORIZONTAL REPRESENTATION
DOCUMENT TYPE IDENTIFICATION
FACE EXTRACTION
MRZ EXTRACTION
OCR PROCESS MRZ STRING PARSING

FIGURE 5 Passport recognition process

3.1 DOCUMENT BORDERS IDENTIFICATION

The document borders on the scanned image search problem is common, because there is a variance in sizes of different document types and also it is inevitable that human puts the document in the scanner with possible shifts and rotations. Also the task becomes more difficult when the scanning process can take with closed or opened flat, that leads to different environment on the image where it is necessary to detect the document.

For borders identification there were used algorithms of gradient change detection on the basis of Canny Edge detector [4].

The application of such algorithm and its different coefficients gives results, that are dependent on optical resolution of scanner that produces the image (300DPI is used under scope of work) and also the type of document template.

To reduce the number of not meaningful elements, that are obtained as a result of applying this algorithm, the image was resized and blurred in advance.

Obtained edges (also borders of documents) allow to step into the next stage of processing – searching of angle degree against the horizontal scanning plane and de skew the image to compensate the rotation till getting the zero degree offset.

3.2 DE SKEW THE DOCUMENT

The result obtained under previous stage can be processed for the line detection analysis of connected pixels. For this process we have used the Hough Line Transform algorithm [5]. And the lines, that we are looking at should be not shorter than the smallest document size divided by 2.

It is possible that we will get the lines, that do not represent the real borders of document. But such lines can be easily filtered by the statistics analysis because of their casual appearance.

In the scope of work, we have used the algorithm that takes into account the statistics of the lines, that should not exceed the deviation from the horizontal plane of 10 degrees.

After obtaining straight lines, that represents boundaries of the document it is possible to determine the slope relative to the horizontal plane, and compensate the tilt by rotating the entire scanned image by the inverse value. This procedure allows to solve several problems:

Preparation of the horizontal position of the text information;

1. Determine type of document template;
2. Extraction (if present) of a human face.

3.3 TEMPLATE MATCHING ALGORITHM FOR DOCUMENT TYPE DETERMINATION.

For all document types and for each side of the document we have selected the unique areas that are not repeated in position, size and textural features on other types of documents.

For each document has been selected for at least 5 characteristic features. Characteristic features have been saved as templates for future use of the search algorithm on the image [6].

The Threshold of correlation value for patterns was chosen sufficiently low valued by 60%, which allows the algorithm to work in different environments noise. But making the low threshold gives us the possibility of false positive occurrences. The To reduce the possibility of incorrect identification of the type of documents have been applied topological matching algorithms mutual arrangement patterns. Ie all distances and offset relative to each other templates. When no matching geometric topology found fragment is taken as a false positive and is excluded from the sample.

At each document type and for each side of the document highlighted the unique areas are not repeated in position, size and textural features on other types of documents.

For each document has been selected for at least 5 characteristic features. Characteristic features have been saved as templates for future use of the search algorithm of the image [6].

Threshold when searching for patterns was chosen sufficiently low value of 60%, which allows the algorithm to work in different environments noise. But making the low threshold creates higher false positive rate. The topological matching algorithm of mutual patterns arrangement is applied to reduce the possibility of incorrect identification of the type of document. All distances and offset relative to each other templates are compared to found on the current image. When no matching geometric topology found for particular fragment, this fragment is marked as a false positive and is excluded from the sample.

3.4 FACE DETECTION ON THE DOCUMENT

If the document that is scanned under the Republic of Kazakhstan standard and we have determined it by template matching algorithm it becomes very easy to mark the area with the presence of photos of the human face. However, in this work we assume the use of passports, the form of which is not known (templates, which were not collected, such as a US passport). Then such a passport is treated as a standard ICAO 9303 template with two lines of rows in the MRZ. However, the position of the human face can vary depending on the issuer country of passport. To solve this problem we have used search algorithms of human faces [7]. This algorithm is invariant to the size of a human face depicted on image, because it uses a pyramidal descent for matching purposes.

3.5 MRZ EXTRACTION

To extract MRZ we solve the problem of selecting the proper threshold value for image binarization. The machine readable zone has the following characteristics - the symbols have significant pixel intensity values range from 0 to 100, and a background - a monotonic textured surface, white noise or some light pattern generally in the range of pixel intensity 120 and 255. As seen in the boundary between significant and insignificant pixels is not great. In this case, there are noises of different nature - worn, creased, partial lack of character, the presence of stains. Particularly in the case of document types A document radiographic phenomenon appears when the flat is opened, which increases the complexity of determining the threshold value for binarization. Thus, the combination of adaptive threshold and a linear threshold binarization algorithms have been applied [8].

3.5 MRZ PARSING

The figure 6 demonstrates the sample MRZ image obtained from the scanner. We can easily determine the little slope of the text direction. Such slope can persist even after second stage of de-skew process during the passport processing algorithm. The text direction slope is defined as the relative position of the leftmost character and the rightmost character in each line. The de skew method applied using affine transformation by rotating around the central axis of the entire document and repeating the whole algorithm from the first step.

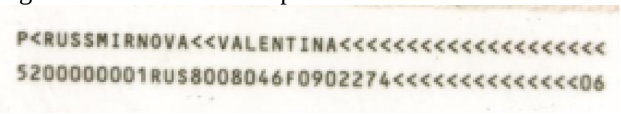

P<RUSSMIRNOVA<<VALENTINA<<<<<<<<<<<<<<<<<<<<
5200000001RUS8008046F0902274<<<<<<<<<<<<<<06

FIGURE 6 Machine readable zone

After rotation of the text and obtaining the desired effect we repeat the binarization process for characters selection. We take into consideration the noise and the false characters that can appear along the lines and out of the lines.

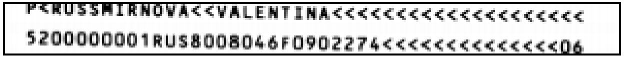

P<RUSSMIRNOVA<<VALENTINA<<<<<<<<<<<<<<<<<<<<
5200000001RUS8008046F0902274<<<<<<<<<<<<<<06

FIGURE 7 Binarized machine readable zone

We calculate the confidence number of recognition of each letter separately and then summarize total confidence of MRZ string recognition process.

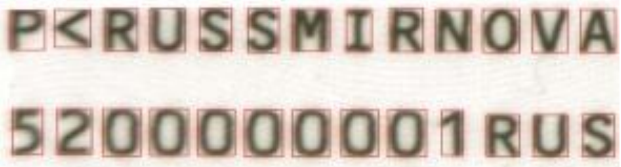

P<RUSSMIRNOVA
5200000001RUS

FIGURE 8 Bounding box of each symbol calculation

And on the final stage we calculate the control sums provided by ICAO 9303 of meaningful MRZ string elements, that can give us the information – if the recognition passed right or if the document has compliance to the ICAO 9303 standard.

4 Suggestions

Commercial application of OCR have been used since 1955. Among the solutions for automated recognition of text information we can note these main directions:

1. Adaptive OCR, that covers problems of recognition different languages ant text styles, recognition of mono text styles, document segmentation and mathematic models of text recognition [9].
2. Hand write text recognition system, which is still in active research and covers many problems of text direction, text position and style, recognition of hand write text, free writes text recognition systems and specialized devices for hand writing [10].
3. Image preprocessing [11], which covers problems of filters application for getting the clear input for classifiers.

4. Post processing intellectual systems [12], which covers problems of intensive input noises.
5. Text recognition systems in multimedia [13], which cover problems of text recognition on simple photo and have deal with edge and contour selecting, projective and non linear distortion.

It is important to note, that absolute validation in text recognition systems is still cannot be made without human correction. That is why the active researches still continue in this area.

5 Conclusions

In the scope of the work, there were solved such tasks as:

1. Redundant algorithms of face detection on images;
2. Redundant OCR algorithms using neural network;
3. Pattern matching algorithm using correlation computation;
4. Application of geometry topology check algorithm.

High quality rate of passport recognition and information extraction have been obtained under conditions of 300 DPI scan resolution and little values of passport rotation degree (less than 10 slope). The neural network (for OCR) was trained to recognize characters on only one image of each character without pan and tilt. This affects the quality of recognition and solving ambiguities in recognizing dissimilar characters in high noise environments. It is therefore critical to minimize the slope of the passport to the horizontal scan axis. The text recognition can be improved after first iteration and determination of text slope by analyzing the first letter and last letter in each line and their vertical offset. If the offset is greater than 0.1 degree we rotate whole image again to compensate this offset and repeat OCR process again which in all cases gives better confidence result than previous.

There are still opened questions of characters identification ambiguity using algorithms based on neural networks. The most striking example is the symbols "0" - "Zero" and "O" - «The letter of the alphabet". In most cases, the distance between the two presentations is extremely low and the percentage of occurrence of false-positive identification symbol becomes high enough. Therefore, more research is needed in the area of how to properly recognize the pattern of character depending on the context position, and researches in the construction of context dependent neural network.

Acknowledgement

The work was funded by JSC NATD Republic of Kazakstan (Agreement No. 163, 2014).

References

[1] «Безопасный город» в Казахстане – краткий исторический обзор *Журнал Rибеж* **1** Апрель 2013, Главный редактор Михаил Динеев, Издатель и учредитель ООО «Компания Р-Медиа» http://www.aips.kz/ru/otchety/otchety-o-vystavke-ot-partnerov/bezopasnyj-gorod-v-kazakhstane

[2] *Company bets on airport of the future passing security with an iris scan* Ministry of innovation/Business of Technology ArsTechnica Sept 2012 by Cyrus Farivar http://arstechnica.com/business/2012/-09/company-bets-on-airport-of-the-future-passing-security-with-an-iris-scan/

[3] Young-Bin Kwon, Jeong-Hoon Kim Recognition based Verification for the Machine *Readable Travel Documents* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.9373 &rep=rep1&type=pdf

[4] Canny J 1986 A computational approach to edge detection *IEEE Tans. Pattern Analysis and Machine Intelligence* **8**(6) 679-98

[5] Bhattacharya P, Rosenfeld A, Weiss I 2002 Point-to-line mappings as Hough transforms *Pattern Recognition Letters* **23**(14) 1705–10

[6] Tempate Matching http://docs.opencv.org/doc/tutorials/imgproc/-histograms/template_matching/template_matching.html

[7] Cascade Classifier http://docs.opencv.org/doc/tutorials/objdetect/-cascade_classifier/cascade_classifier.html

[8] Thresholding Operators http://docs.opencv.org/doc/tutorials/-imgproc/threshold/threshold.html

[9] Nielson H E, Barrett W A Consensus-based table form recognition. *In Proceedings of the International Conference on Document Analysis and Recognition* **II**

[10] Ching Y Suen, Shunji Mori, Soo H Kim, Cheung H Leung 2003 *Analysis and recognition of Asianscripts - the state of the art In Proceedings of the International Conference on Document Analysis and Recognition* **II**

[11] Summers K 2003 Document image improvement for OCR as a classification problem *In Document Recognition and Retrieval X* **5010**

[12] Yefeng Zheng, Huiping Li, and David Doermann 2003 Text identification in noisy document images using markov random field *In Proceedings of the International Conference on Document Analysis and Recognition* **I**

[13] Clark P, Mirmehdi M 2000 Finding text regions using localised measures *In Majid Mirmehdiand Barry Thomas, editors, Proceedings of the 11th British Machine Vision Conference* 675-84

Authors

	Yedilkhan Amirgaliyev **Current position, grades**: Head of Engineering Faculty, Corresponding member of National Academy of Engineering, Republic of Kazakhstan **University studies**: Ph. D, Professor, Kazakh National Technical University **Scientific interest**: Pattern recognition and classification, information theory, remote sensing, neural networks, system analysis and decision making **Publications**: More than 140 science papers
	Rassul Yunussov, 1982, Almaty, Kazakhstan **Current position, grades**: Instructor, Kazakhstan, Kaskelen, Suleman Demirel University **University studies**: PhD, Information Technologies and Computations Institute 2010, Almaty, Kazakhstan. **Scientific interest**: neural networks, data analysis. **Publications:** 10 papers**.**

Authors' index

Amirgaliyev Y	27
Deng Chongjing	22
Li Buyi	22
Li Jin-Dong	7
Li Shuang	22
Li Yongbiao	16
Wang Tao	7
Wu Yang	7
Yunussov R	27

Information and Computer Technologies

Research on characteristic parameters mining and clustering of unknown protocols bitstreams

Yang Wu, Tao Wang, Jin-dong Li

Computer Modelling & New Technologies 2015 ***19****(2B) 7-15*

Characteristic parameters mining of unknown protocol bitstreams and parameters optimizing of clustering algorithm are the foundations of unknown protocol bitstreams analyzing. The parameters such as the bit frequency, runs and bit frequency within a block are defined according to the frequency of zero and one, frequency of sequential zero and one, bit frequency within a block. As the parameter of bit frequency within a block is sensitive to the block length, an optimal block length selection algorithm is proposed based on the principle of variance. In order to select effective initial clustering centers for division clustering algorithms such as the k-means algorithm, an initial clustering centers selection algorithm is proposed based on the peak value of sample density for each dimension. In order to select the optimal clustering number, a function of clustering quality evaluation is given by the sample density in cluster and cluster density. Taking the bitstreams of HTTP, DNS, ICMP, TELNET and UDP datasets as the unknown protocols bitstreams, the experimental results not only verified the effectiveness of the proposed algorithms but also point out the necessity of mining more effective parameters.

Keywords: Unknown protocol, bitstreams, clustering, characteristic parameter, bit frequency within a block

Cache Pre-fetching system based on data mining on Web

Yongbiao Li

Computer Modelling & New Technologies 2015 ***19****(2B) 16-21*

From 20th century 90’s to now on, Internet and data mining techniques had developed rapidly and became mature, kinds of application on Web data mining had been proposed to the market. In this paper, we would first introduce the development of cache Pre-fetching technique, and then present a cache pre-fetching System model based on Web data mining, details of each implementation would follow. Our aim was to enhance caching effectiveness, and network accessing speed. Such technique could be applied in personnel, educational, and official information managing system in database of educational scope. Accessing speed of educational information system for numerous teachers and students, benefit high school personnel management, and also the effective scientific structuralize educational management.

Keywords: Web data mining, sequential mining, cache pre-fetching system

High resolution remote sensing image classification based on particle swarm optimization and support vector machine

Buyi Li, Chongjing Deng, Shuang Li

Computer Modelling & New Technologies 2015 ***19****(2B) 22-26*

Many algorithms have been developed for image classification and support vector machine (SVM) is a kind of supervised classification that has been widely used recently. However, the accuracy of a SVM classifier heavily depends on the selection of a right kernel model and appropriate parameter. In this paper, a comparative analysis of the impact of four kernels (linear kernel, polynomial kernel, radial basis function kernel and sigmoid kernel) on the accuracy of SVM classifiers is conducted. Moreover, the Particle Swarm Optimization (PSO) is used to search for the optimum parameters for each kernel function in order to improve the classification accuracy of SVM classifiers. Our experiments for optimizing the kernel function parameters and assessing the robustness of SVM classifiers were carried out with classifications of QuickBird-2 images over Wuhan, China for monitoring urban land cover/land use information. The experimental results indicate that the polynomial kernel outperforms the other kernels in classifying high resolution remote sensing image. The sigmoid kernel performs worse than any other kernels. Our findings also suggest that selected parameter by PSO will improve the classification accuracy, especially for radial basis function kernel.

Keywords: high resolution remote sensing image, support vector machine classification, parameter optimization, particle swarm optimization

Pattern recognition systems in the problems of automatic person identification using the passport data

Y Amirgaliyev, R Yunussov

Computer Modelling & New Technologies 2015 ***19****(2B) 27-30*

The work describes the implementation of modern technology for remote sensing and data processing in the area of human activities concerned to the security provision, based on usage of pattern recognition algorithms and neural networks. The Republic of Kazakhstan State Identities and Passports were used as the basis; The ICAO 9303 MRZ Standard was used. Obtained stable recognition model for identification of known passport types, and MRZ section decoding.

Keywords: neural networks, pattern recognition, raster data, image processing

Content C

PART C Operation Research and Decision Making

Shubing Qiu	The real estate enterprise performance evaluation model study empirical research on the real estate enterprise statistics in China: 2009-2013	**7**
Shuaihui Tian, Lan Chang	Reliability research on dynamic logistics alliance based on GO methodology	**11**
Qingyong Xu, Shunliang Jiang, Wei Huang, Longzhen Duan, Shaoping Xu	Multi-feature fusion based spatial pyramid deep neural networks image classification	**16**
Hong-xia Liu	Difference processing on financial accounting of institutions and enterprises	**22**
Author's Index		**25**
Cumulative Index		**26**

The real estate enterprise performance evaluation model study empirical research on the real estate enterprise statistics in China: 2009-2013

Shubing Qiu

School of Management Engineering, Anhui Polytechnic University, Wuhu, China

Corresponding auhtor's e-mail: 44528934@qq.com

Received 1 November 2013, www.cmnt.lv

Abstract

Based on our real estate business development, for the shortcomings of traditional performance evaluation methods, combined with hierarchical fuzzy neural network evaluation method, using BP neural network training corporate financial indicators, and fuzzy neural network training non-financial indicators, and then to build a fuzzy neural network evaluation model integratedly, so the value of enterprise performance evaluation results can be calculated. The results show: the model is of high accuracy, which can more accurately reflect the performance of the real estate development business.

Keywords: fuzzy neural network, BP neural network, real estate business, business performance

1 Introduction

Real estate business is engaged in real estate development and management activities of a comprehensive industry, which has a pilot, basic, driven and risk characteristics, on the one hand, the real estate industry as a new growth point of macro-economic development for the domestic economy rapid growth, driven by the joint development of other related industries made a significant contribution. On the other hand, due to the lack of current evaluation system and the imperfect overall evaluation of the performance of real estate development, these factors result in the development of the domestic real estate disorder. Therefore, it is necessary to select the scientific method to construct performance evaluation model to reflect the status of the real estate business development, which is an important content.in the study of sustainable development in the real estate industry.

2 Problem statement and preliminaries

In the previous literatures, the research on the real estate enterprise performance evaluation model are focused only on selected indicators for economic analysis, research methods, mainly using the traditional balanced scorecard method, principal component analysis, composite index, efficacy coefficient method, factor analysis, economic Value Added (EVA), etc. [1], but most of these methods are of "subjective factors affecting large, poor accuracy", In addition, in the selection of indicators, only the financial indicators (quantitative indicators) are considered, while ignoring some of the non-financial indicators (qualitative indicators) and lacking accuracy and implement on the research results. Therefore, to establish a comprehensive and objective performance evaluation system can provide a strong basis for the adjustment of corporate strategy, then the implementation effect of the enterprise development strategy can be evaluated timely.

Based on this, now to select the fuzzy neural network method to evaluate the development status of the real estate enterprise performance in China. And because the fuzzy neural network, fuzzy logic and neural networks combined, can form a better system than a single fuzzy neural network system or a separate system, its operation process is not black-box operation [2, 3], so it can more efficiently and accurately reveal the development of China's real estate enterprise performance.

3 A fuzzy neural network structure selection

In the selection of neural network model based on Takagi-Sugeno (TS model for short) is blurred by the first member of the network structure of the network. And the network is consisted of two parts, of which the first network element is to match the fuzzy rules the antecedent of the network and the last to generate fuzzy rules. Its fuzzy neural network structure is simpliy shown in Figure 1.

As is shown in Figure 1, layer 1 is the input layer, and this layer is expressed as the number of nodes $N_1 = n$.

Layer 2 is the fuzzy layer, and the layer number N_2= $\sum_{i=1}^{n} m_i$ (Now the membership function can be calculate by the Gaussian Gauss function method).

Layer 3 is the rule based reasoning layer, the total number of nodes in this layer N_3=m.

Layer 4 is the normalized layer and the number of nodes in this layer N_4=N_3=m, which implements the normalization calculation and it is calculated as shown in Equation (1).

$$\bar{a}_i = a_j / \sum_{i=1}^{m} a_i ,\ j = 1,2,\ldots,m \,. \tag{1}$$

Layer 5 is the output layer. Take the enterprise performance value as the output value of fuzzy neural network, which is calculated as shown in Equation (2).

$$y_1 = \sum_{j=1}^{m} y_{1j} \bar{a}_j \,. \tag{2}$$

Among them:

$$y_{1j} = p_{j0}^1 + p_{j1}^1 x_1 + p_{j2}^1 x_2 + ... + p_{jn}^1 x_n,\ j = 1, 2, ...m \,,$$

which p_{ji}^1 is the connection weights, the total number of nodes in this layer, N$_5$=1.

4 Learning algorithm selection

As is shown in Figure 1, the choice of fuzzy neural network is essentially a multilayer before-feed-forward networks, and the error back propagation algorithm can be used to adjust the parameters, which is mainly to adjust p_{ji}^1 (the connected weight of Layer 5 and the membership parameters value c_{ij} and the width σ_{ij} (i=1,2,....,n, j=1,2,....,m$_i$) of Layer 2 .

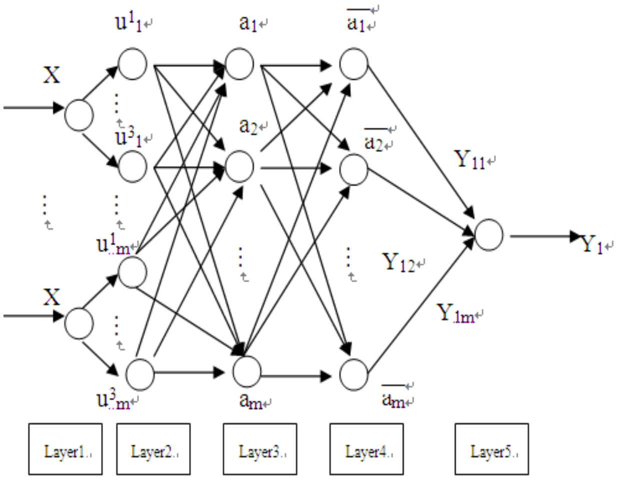


FIGURE 1 The fuzzy neural network structure based on T-S model

Assuming the input of the neurons node j in the fuzzy neural network layer q is

$$f^{(q)}(x_1^{(q-1)}, x_2^{(q-1)}, ...y^{(q-1)}{}_{nq-1}, y_{j1}^{(q)}, y_{j2}^{(q)}, ..., y_{jnq-1}^{(q)})$$

the output is $x_j^{(q)} = g^{(q)}(f^{(q)})$, then the function of each node can be expressed as:

Layer 1:

$$f_i^{(1)} = x_i^{(0)} = x_i,\ x_i = g_i^{(1)} = f_i^1, \quad i = 1,2,..n \,.$$

Layer 2:

$$f_{ij}^{(2)} = -(x_1^{(1)} - c_{ij})^2 / \sigma_{ij}^2 x_{ij}^{(2)} = u_i^j$$

$$= g_{ij}^{(2)} = e^{f_{ij}^{(2)}} e^{-(x_i - c_{ij})^2 / \sigma_{ij}^2} \,,$$

$$i = 1, 2, ...n,\ \ j = 1, 2, ...m_i$$

Layer 3:

$$f_i^{(3)} = x_{1i_1}^{(2)} x_{2i_2}^{(2)} ... x_{ni_n}^{(2)} = u_1^{i_1} u_2^{i_2} ... u_n^{i_n} x_j^{(3)} = a_j = g_j^{(3)} = f_j^{(3)}$$

$$j = 1,2,...,m$$

Among them, $m = \prod_{i=1}^{n} m_i$.

Layer 4: $\bar{a}_j$

$$f_j^{(4)} = x_j^{(3)} / \sum_{i=1}^{m} x_i^{(3)} = a_i / \sum_{i=1}^{m} a_j x_j^{(4)} = \bar{a}_j = g_j^{(4)} = f_j^4, j = 1,2,...,m$$

Layer 5:

$$f_1^{(5)} = \sum_{j=1}^{m} y_{1j} x_j^{(4)} = \sum_{j=1}^{m} y_{1j} \bar{a}_j x_1^{(5)} = y_1 = g_1^{(5)} = f_1^{(5)}$$

Assuming the square error function is

$$c_{ij}(k+1) = c_{ij}(k) - \beta \frac{\partial E}{\partial c_{ij}} \,,$$

$$i = 1, 2, ...n,\ \ j = 1, 2, ...m_i$$

(in which t_1 represents the expected output and y_1 represents the actual output), to adjust p_{ij}^1 , c_{ij} and σ_{ij} according to the BP algorithm (error back-propagation algorithm), and then according to the learning algorithm of parameters p_{ij}^1 (as is shown in Equation (3) and Equation (4)), to adjust c_{ij} and σ_{ij} fixing the parameters p_{ij}^1 , then can obtain:

$$c_{ij}(k+1) = c_{ij}{}^{(k)} - \beta \frac{\partial E}{\partial c_{ij}}$$

$$\sigma_{ij}(k+1) = \sigma_{ij}^{(k)} - \beta \frac{\partial E}{\partial \sigma_{ij}}, i = 1, 2, ..., n,\ j = 1, 2, ..., m_i$$

(β is the learning rate， and β >0)

$$\frac{\partial E}{\partial p_{ji}^1} = \frac{\partial E}{\partial y_1} \frac{\partial y_1}{\partial y_{1j}} \frac{\partial y_{1j}}{\partial p_{ji}^1} = -(t_1 - y_1) \mathrm{a}_j x_i \tag{3}$$

$$p_{ji}^1(k+1) = p_{ji}^1(k) - \beta \frac{\partial E}{\partial p_{ji}^1}, i = 1, 2, ..., n; \qquad (4)$$

$$j = 1, 2, ..., m$$

5 Build performance evaluation model of the real estate enterprises in China

5.1 SELECT THE INDEX SYSTEM

Performance evaluation is currently used by the majority of the real estate business for financial indicators (quantitative indicators), which select a single indicator, lack industry-specific and can not fully and objectively reflect the performance level of the real estate enterprise development

[4]. For example, the current domestic real estate companies on the Top 50 rankings, are most ranking in index selection process based on sales revenue index focused solely on the growth of "quantity", which does not combine our real estate with the ubiquitous "development disorder, housing vacancy rate increases, "and so on, and ignore the Chinese real estate industry's increase of " quality ", as is shown:

1) Lack indicators of intangible assets and intellectual capital. The current development of the individual real estate businesses are of great differences, which develop extensively and disorderly. So in the current era of knowledge and economy, we must increase the index reflects the value of intangible assets and intellectual capital, in order to measure the lasting vitality of enterprises effectively.

2) Lack indicators of reflecting environmental costs. Real estate development is taken as a modern city in the largest project land use area, and also as the city's main construction contents, which consume large amounts of energy and resources [5] and also affect the ecological quality of life of residents and the city. Thus, environmental indicators and environmental cost accounting must be increased for the real estate development performance.

5.2 IDENTIFICATION AND ANALYSIS ON THE INDICATOR OF THE REAL ESTATE ENTERPRISE PERFORMANCE EVALUATION

Enterprise performance is a very complex system, which involves a number of factors, with the multiplicity and complexity of other features, this combination of the characteristics of the real estate index design, which includes both financial indicators (quantitative indicators), but also contains the non-financial indicators (qualitative indicators). Basing on the real estate enterprise statistics in China from 2009 to 2013, to integrate select three areas of "social, economic and environmental" comprehensive estimates, expectations of the business as a whole in order to compare the performance of accurate evaluation, the specific contents of each indicator, as shown in Table 1.

TABLE 1 Indicators of the real estate enterprise performance evaluation

Financial Indicators				**Non-financial indicators**
Social Indicators	Financial Indicators	Environmental Indicators	Development capacity indicators	
X1: Per capita living space	X5: Real estate added value	X9: The growth rate of agricultural land expropriated area	X13: Capital accumulation rate	X16: Enterprise infrastructure management level
X2: Complete housing rate residential	X6: The growth rate of real estate practitioners	X10: Real estate development of residental green coverage	X14: Total asset growth	X17: Value of identity in the post
X3: The residents' average living area	X7:Ratio of real estate value added and GDP	X11: Construction waste and sewage emissions	X15: Operating profit growth	X18: Consumer satisfaction
X4: Ratio of housing price and income	X8: Real estate price index	X12: Effective utilization of land	____	X19: Real estate business brand influence

5.3 SYSTEM TEST AND SIMULATION TRAINING

Since the number of fuzzy rules with fuzzy neural network input dimension increases exponentially，so when the input level increase in the number of large, the network-structural will inevitably be more complex and the training learning time will be longer [6], in order to solve this problem, and to minimize the influence of subjective factors on the evaluation results, there to adopt the layered-in-order-evaluation method based on the use of hierarchical fuzzy neural network, and firstly BP neural network can be used in the financial indicators simulated training, and the fuzzy neural network can be used in non-financial indicators simulated training. And then to use the fuzzy neural network to train on their results again, and come up with a final evaluation results. Set for the final evaluation result as Y→ Z={excellent, good, fair, poor}. Among them, Y is the final output of the fuzzy neural network, Z is the corresponding grade of the enterprise performance evaluation.

This paper uses batch-training methods and LM learning algorithm to calculate, and the financial indicators of BP neural network is constructed with three-tier network structure, in which the input layer neurons is 15, the output layer neuron number is 1 (the output is of the level of the financial grade of the enterprise), the hidden layer neurons number is 10 (determined according to Kolmogorov theorem), and at this time BP neural network training error is of the minimum value and shortest training time (its training error curve shown in Figure 2, where performance is 0.000895863, goal is 0.001).

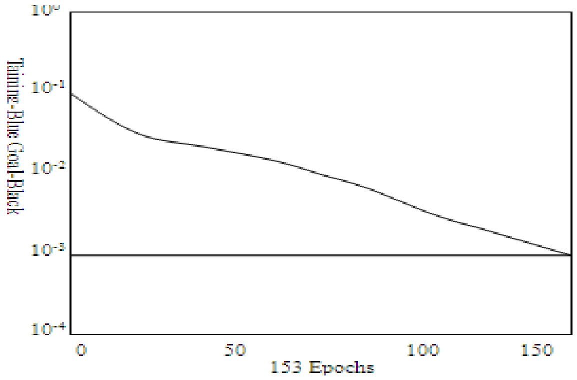


FIGURE 2 RMSE change curve in training process

The non-financial indicators of fuzzy neural network model structure is a made by 4 non-financial indicators as input, 1 output, and fuzzy layer has 16 neurons, the number of fuzzy rules is 256; the final fuzzy neural network model structure includes 2 inputs (Level results of the financial and non-financial status), 1 output (the final results of the enterprise performance evaluation), and the fuzzy layer contains 8 neurons, and the fuzzy rules is 16 (its training data mean square error curve shown in Figure 3). From Figure 3, we can see that the mean square error curve of the training data is relatively smooth, the network training is valid.

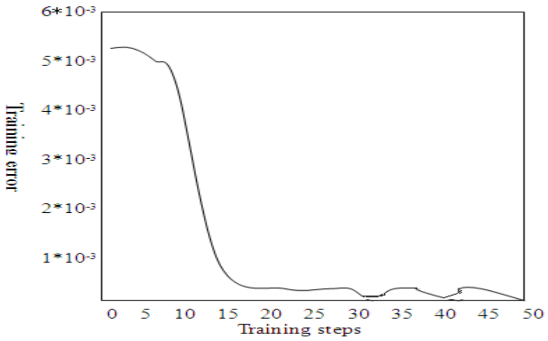


FIGURE 3 Mean square error of the training data curve

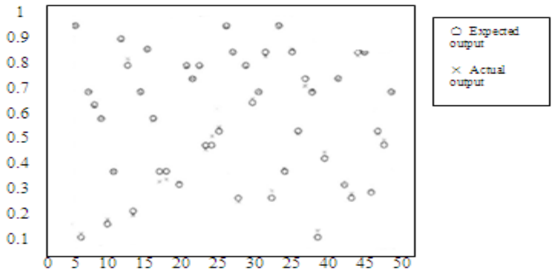


FIGURE 4 The predicted and expected output of the models

Finally, to put the test sample in the well-trained fuzzy neural network to carry out the results of the enterprise performance evaluation when the fuzzy neural network training is completed. The test results expected output and the actual output of the network are shown in Figure 4. From Figure 4, there can seen that the fuzzy neural network model we have structured in this article, which can build a completed better business performance evaluation, and the error of its network between the predicted output and the expected output is tiny, and the degree of match is up to 90%, so it is of higher accuracy.

6 Conclusions

Above all, fuzzy neural network evaluation model constructed in this paper can better solve the unshaped, nonlinear and other issues, which has parallel computing, distributed information storage, fault-tolerant capability, adaptive learning function and other advantages. And we can draw a conclusion that it is suitable for the actual developing stastus of the real estate in China. It also can provide reliable performance information for the business owners, creditors, small investors and other stakeholders who are involved in real estate, and it can effectually reduce the investment risks of the real estate market in China.

Acknowledgements

This work is partially supported by the National Foundation of Social Science of China (NO. 14BGL163) and the National Foundation of Natural Science of China (NO. 71171002). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] Peng Z 2007 A Fuzzy Neural Network Evolved by Particle Swarm Optimization *Journal of Harbin Engineering University* **3** 316-21

[2] Cao Y 2011 An Approach of C2C E-commerce Consumer Community Classification Based a Modified Fuzzy Neural Network *Journal of Computational Information Systems* **16** 5738-45

[3] Zhou X 2013 Attern Recognition Based on Fuzzy Neural Network with Improved Structure and Learning Algorithm *Journal of Computational Information Systems* **3** 1103-11

[4] Jyh J, Roger S 1993 ANF IS: Adaptive Network based Fuzzy Inference System *IEEE Transactions on Systems, Man and Cybernetics* **3** 665-85

[5] Liu K 2013 Energy Efficient Heterogeneous Network Selection Method Based on Fuzzy Neural Network*Journal 5of Information and Computational Science* **10** 3373 - 85

[6] Wei J J 2010 Research on a Fuzzy Neural Network Based Enterprise Performance Evaluation Approach *Shandong Science* **1** 36-41

Author

Shubing Qiu, born on October 18, 1980, Shandong Province of China

University studies: M.S. in 2006 and currently in College of Management Engineering at Anhui Polytechnic University (Wuhu).
Scientific interest: enterprise application integration and distributed systems.

Reliability research on dynamic logistics alliance based on GO methodology

Shuaihui Tian[1]*, Lan Chang[2]

[1]*College of Economics and Management, Chongqing University of Posts and Telecommunication, Chongqing 400065, China*

[2]*College of Economic Management, Chongqing Electric Power College, Chongqing 400053, China*

**Corresponding author's e-mail: tianshuaihui2007@163.com*

Received 1 October 2013, www.cmnt.lv

Abstract

In order to calculate the reliability accurately and dig out the dominant influencing factors of dynamic logistics alliance, GO methodology is applied in reliability research on dynamic logistics alliance. Through building the structure model of dynamic logistics alliance, failure factor of each subsystem is diagnosed. With the GO methodology, the dynamic logistics alliance is transformed into the GO chart, the system reliability is calculated in detail, its failure mode diagnosis and importance calculation are quantitatively studied and then a case of automobile dynamic logistics alliance is employed to verify GO methodology for effectiveness and validity.

Keywords: dynamic logistics alliance, system reliability, GO methodology, influencing factor, failure mode

1 Introduction

Currently, Dynamic Logistics Alliance (DLA for short) has always been one of the hot fields of logistics. As the most competitive logistics operation model in the 21st century, it is a complex organization, which is involving multiple logistics cooperation, dynamic, uncertainty. In this paper, the research is from perspective of logistics service integrator. Logistics service integrator can not only provide a comprehensive logistics operation solution for the customer, also have a strong ability to integrate logistics resources [1]. Due to complexity and uncertainty in the integration process of the DLA, it has exposed some problems, such as low alliance' stability, high integration cost, poor information and other issues. In order to improving the operational efficiency and digging out the dominant influencing factors of DLA, we need study the reliability of DLA. Reliability is used to measure the probability that the DLA integrate the logistics resources to complete the logistics tasks in accordance with customer needs within a specified period of time. Research scholars about the reliability of the logistics system has focused on logistics network systems and specific logistics (including military logistics and emergency logistics), these documents always is created complex mathematical models and solved very difficult. Chen A [2] studied the reliability of Military Logistics in the war, ÁrniHalldórsson [3], Wang N [4], Yu X C [5] studied the reliability of the logistics network, or system. Chen J [6] designed regional emergency logistics network based on reliability analysis. But reliability problem about DLA, which include reliability' calculation and factors, are studied very little, and it is very necessary to find a method to study the reliability problem about DLA.

GO methodology is a system probability analysis technique, which is success-oriented, its' main steps is to step up the GO chart, and calculate the system reliability. The two factors of GO chart are Operator and Signal flow. Shen Zupei [7] has described the 17 Operators and its Algorithms. Operator represents the logical relation between unit functions and input, output signals. Signals flow represent the association between system unit and its input, output signals, the attribution of Signals flow is state value and probability. Currently, the application of GO methodology is mainly focused on reliability analysis of complex repairable machine, equipment and other hardware systems [8-11].And the GO methodology is also used in reliability of supply chain systems, emergency management systems, etc [12-15]. DLA is a system, which is more complex, dynamic than general logistics system. Each subject and logistics resources in the process of operation are likely in a variety of state, so the reliability study of dynamic logistics alliance using the GO method is feasible.

In view of this, we will use the GO method to study the reliability of DLA. Through comprehensively analyzing the subsystem in the DLA, such as logistics services integrator, strategic logistics resource provider, common logistics resource provider and client, some important factors are analyzed which leads to their failure. The GO model, which is used to diagnosis the reliability of the DLA, is established with GO method, and it provides a new idea for the reliability analysis of DLA.

2 Application of GO method in the DLA reliability

2.1 STRUCTURE MODEL OF DLA

For researching simply, first of all, the structure model of DLA must be assumed. DLA, which take logistics services integrator as the core, include logistics services integrator, functional logistics resource providers, clients and other units. Logistics service integrator integrate logistics resources of various functional logistics resource provider and provide all kinds of high level, low cost, on time logistics service to customers. The structure model of DLA has been constructed based on core competitiveness of the enterprise and resource advantage, which is shown in Figure 1.

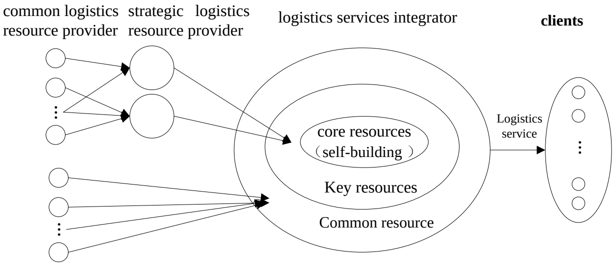

common logistics resource provider
strategic logistics resource provider
logistics services integrator
clients
core resources (self-building)
Key resources
Common resource
Logistics service

FIGURE 1 The structure model of DLA

In Figure 1, according to the own ability and the strategy need of logistics service integrators, we divide all kinds of logistics resources into core resources, key resources, and common resources. In order to maintain the core competitiveness of enterprises, logistics services integrator usually build the core resources by themselves. Because of the great influence of key resources, logistics services integrator is generally associated with some important partners (strategy resource providers) to get strategic cooperation mode. Common resources is the basic resources to achieve customer demand, and logistics service integrator is usually in the form of outsourcing to establish common trading partnership with other common resource provider. Strategic resource providers also have capability to integrate logistics resources.

2.2 FACTORS ANALYSIS OF FAILURE DIAGNOSIS ABOUT DLA

From the perspective of system operation, DLA includes subsystem, such as logistics services integrator, client (Group), strategic resource provider, general resource provider. To ensure the successful operation of dynamic logistics alliance, each subsystem must keep normal state. The dominant factors of supply chain cooperation from two aspects of hardware and software is given in the literature [16]. From two aspects of soft and hard environment, important factors affecting the successful operation of DLA is summed up, which is in the Table 1.

2.3 THE GO CHART ESTABLISHMENT OF DLA

Based on the structure model of DLA and the factor of each subsystems, we have established a GO chart about DLA, as shown in Figure 2.There are two operator symbols: the fifth (signal generator) and the tenth (And gate), which make the realization of computer programming and automation calculation conveniently.

The illustration of operator symbols:

$Y_{ij}(i=1,2\cdots m, j=1,2,\cdots 8)$: the influencing factor j of common resource supplier i of strategic resource provider;

$X_{pq}(p=1,2\cdots e, q=1,2,\cdots 8)$: the influencing factor q of common resource supplier of logistics services integrator;

$Z_{bt}(b=1,2\cdots n, t=1,2,\cdots 8)$: the influencing factor t of strategic resource provider b ;

$J_k(k=1,2,\cdots 8)$: the influencing factor k of logistics resource integrator;

$C_{ar}(a=1,2\cdots d, r=1,2)$: the influencing factor r of customer a .

In the signal flow, Y_{ij}, X_{pg}, Z_{bc}, J_k, C_a respectively represent their signal flow number. The signal issued by input operator is according to the respective operation symbol stream, and other signal flows are in numerical order from small to large order.

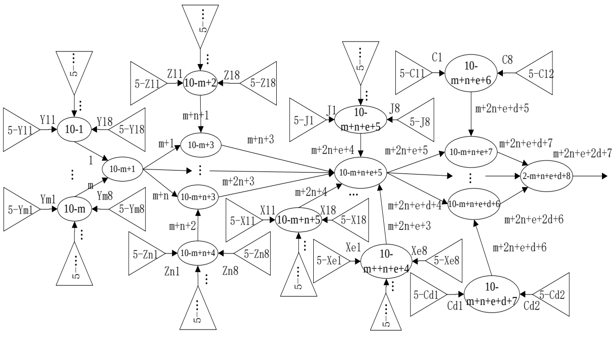
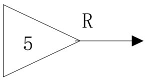


FIGURE 2 The GO chart of DLA

2.4 GO METHODOLOGY CALCULATION OF DLA RELIABILITY

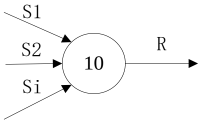
DLA is a two-state system, there are only two states in the input operator: success and failure. Meanwhile the operator symbol has two types: type 5 and type 10.The intermediate operator symbols of GO chart about DLA (Figure 2) are all logical operator symbols, so the quantitative calculation are more convenient. Suppose that the successful state of input operator is 1, the failure state is 0, so the signal flow only has successful state 1 and failure state 0. Suppose the probability of success for the operator symbol as $P(S=1)$, the probability of failure for the operator symbol as $P(S=0)$, the probability of the output signals as $P(R)(R=1\,or\,0)$. The algorithm of 2 types of operator symbols in the GO chart about DLA is shown as follows:

1) Single signal generator (type 5)



State probability of output signal in single signal generator is the operator of state probability.

2) And gate (type 10)



When there is one failure of N independent input signals, the output signal is failure, the algorithm is as follows: the success probability of the output signal is

$$P(R=1)=P(S1=1)P(S2=)...P(Si=1)$$

Failure probability:

$$P(R=0)=1-P(S1=1)P(S2=)...P(Si=1)$$

According to the above algorithm, starting from the input signal of the input operator in the GO chart about DLA, the calculation is operated according to the operation rules of an operator, and get the state probability of the output signal. In accordance with the signal flow sequence in the GO chart, the state probability of each output signal are gradually calculated, we can conveniently calculate the state probability of the final output signal, namely the reliability of DLA.

2.5 DETERMINATION AND IMPORTANCE DEG-REE CALCULATION OF DLA FAILURE MODE

The determination of dynamic logistics alliance failure mode is mainly to determine the minimum cut sets of DLA. For DLA, the two states of the system, can be directly applied to the quantitative calculation method for qualitative analysis to get the minimal cut sets. The methods are as follows: for the 1 order cut set, as long as suppose that one operator in M operators is in a failure state and the probability of success is 0, the other operator symbol keep the same, and the system success probability is directly calculated. If the probability of success is 0, the fault state of the operator symbol is a one-order cut set of a system. M operators are calculated successively and one-order cut sets can obtained. Then we take 2 operator symbols out of M operators except the one- order cut set, and use the same method to get all of the two-order minimum cut sets. And so on, each order cut set can be got. For higher order cut sets, if high order combination already contains low order cut set, it will not need to calculate system success probability, then cut set is the minimal cut sets [8]. This method is easy to programming and it can also realize the automatic calculation of complex system.

Combination of operator symbol failure states in the minimal cut set represents the combination of the system functional subsystem failure event, the product of failure probability of these subsystems represents a minimum cut sets of probability. The probability of the minimum cut set can be used to evaluate the important degree of minimal cut sets, so as to system optimization.

In the DLA, including logistics services integrator, customers, strategic resource provider, common resource providers and other subsystems, so the DLA failure mode has 1 order, d order, n order, m+e order minimum cut sets. One-order cut set is mainly determined by the influencing factors of logistics resources integrator. D-order cut set is the main collection, which is composed by one factor of each client's customers in d customers, n order cut set is mainly consisted of one factors of each strategic logistics resources provider in providers, m+e order cut set is mainly consisted of one factors of each common resource provider in m+e providers.

3 Application case

3.1 THE BASIC SITUATION

There is a automobile dynamic logistics alliance, whose core is a third party logistics company, which consists of 4 common resource providers (Y_1, Y_2, X_1, X_2), 2 strategic resource providers (Z_1, Z_2), 1 logistics resource service integrator (J), 3 customers (C_1, C_2, C_3). Through the cooperation with the quality department of the company, we statistics particular data of operation situation of the system in July 2012 to October 2013 during the implementation of logistics task 120 times, and use the fish bone diagram method to analyze the abnormal problem, and various reasons are classified and analyzed to obtain quantitative failure probability as shown in Table 1, Table 2, Table 3.

TABLE 1 Failure probability of logistics services integrator

Influencing factors	***J***
The poor ability of anti-risk inside and outside	0
adjustment difficulty about enterprise management mode, organization structure, business process	0
The high cost of logistics alliance cooperation	0.07
The unreasonable operation plan of DLA	0
Low coordination ability to enterprise resource provider	0
Unsmooth information communication	0.04
weak control of the operation process of DLA	0
Imperfect Logistics network/equipment	0

TABLE 2 Failure probability of strategic resource provider

Influence factors	Z_1	Z_2
The poor ability of anti-risk inside and outside	0	0
adjustment difficulty about enterprise management mode, organization structure, business process	0	0
The unreasonable operation plan of logistics alliance	0	0
The low alliance income about participation in logistics cooperation	0.03	0.07
poor convergence of information	0.09	0.04
Poor execution force of logistics personnel	0	0
Low level of Logistics service technology	0.06	0
Imperfect Logistics network / equipment	0	0

TABLE 3 Failure probability of common resource providers and Clients

	Influence factors	Y_1	Y_2	X_1
Common resource provider	The poor ability of anti-risk inside and outside	0.06	0	0
	adjustment difficulty about enterprise management mode, organization structure, business process	0	0	0
	Low profits of participation in the integration of logistics resources	0.07	0.06	0.08
	Unreasonable logistics task operation plan	0	0	0
	Unsmooth information communication	0.05	0.07	0
	Pool executive force of logistics management and operating personnel	0	0	0
	Low level of Logistics service technology	0	0.08	0.04
	Imperfect Logistics network / equipment	0	0	0
	Influence factors	C_1	C_2	
Clients	The change degree of Customer demand plan	0.04	0	
	The smooth degree of Customer information communication	0	0.06	

3.2 RELIABILITY ANALYSIS OF DLA BASED ON GO METHODOLOGY

3.2.1 The establishment of GO chart

Combining each subsystem structure of the dynamic logistics alliance and each subsystem failure situation, DLA GO chart is established (Figure 3). The factor whose probability of failure is 0 has no influence on the system reliability calculation, so it will not be put in the GO chart.

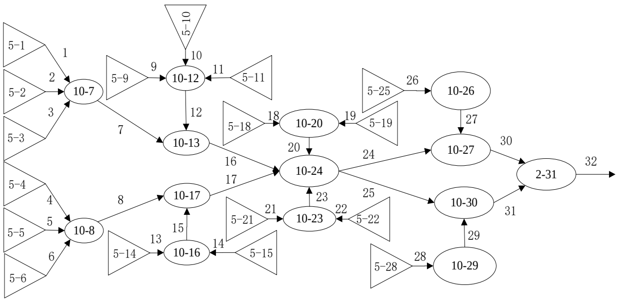


FIGURE 3 Automobile logistics alliance GO diagram

3.2.2 Probability Calculation of each signal flow

In Figure 3, according to the operation rules of GO chart and signal flow order in the GO chart, the reliability degree of output signal in dynamic logistics alliance are calculated, the ultimate output signal 32 represents the reliability of dynamic logistics alliance. The calculated results are shown in Table 3, in which the calculation of signal flow 32 involves a total signal flow, which is shown in reference [8].

In Table 4, we can find the reliability of dynamic logistics alliance is not high, only 0.4823, which is obviously related to multi-system structure of the dynamic logistics alliance and the complex operation process, so it is necessary to find out the key factors restricting the dynamic logistics alliance reliability, so as to improve and have better control it.

TABLE 4 The calculation results of DLA reliability

Signal flow	P	Signal flow	P	Signal flow	P
1	0.9400	12	0.8297	23	0.8832
2	0.9300	13	0.9300	24	0.3923
3	0.9500	14	0.9600	25	0.3923
4	0.9400	15	0.8928	26	0.9600
5	0.9300	16	0.6928	27	0.9600
6	0.9200	17	0.7181	28	0.9400
7	0.8305	18	0.9300	29	0.9400
8	0.8043	19	0.9600	30	0.3766
9	0.9700	20	0.8928	31	0.3688
10	0.9100	21	0.9200	32	0.4823
11	0.9400	22	0.9600		

3.2.3 Determination of the minimal cut set and important degree calculation

The composition of the DLA of minimal cut sets is shown in Table 5, probability of cut set is obtained by the product of the failure event probability, the important degree represents the percentage of the cut set failure probability for total system failure probability. The larger importance degree means the larger impact on the system for the factor, and then it should be the focus of attention.

Table 5 shows that the importance degree in the first place is cut set composed of No. 18 operator, namely logistics alliance cooperation cost of logistics service integrator is high, up to 51.19%, which is the most important factors of logistics alliance operation, it is in accordance with the current logistics services integrator's concern (logistics cost) and cost of resources integration. The important degree in second place is cut set composed of No.19 operator, namely the bad information communication of logistics service integrator, reached to 29.25%. The important degree in third place, fourth place are composed of No. 10, No. 14 operators cut sets and No. 11, No. 14 operators cut sets, namely strategic resource provider's 'poor information convergence', 'low income from participation in the logistics alliance cooperation', 'low level of logistics services technology', in turn, we can determine the factors which we should pay more attention on the operation process of DLA. And it provides ideas and basis for the improvement and optimization of DLA.

TABLE 5 The minimum cut sets and the important degree

Order number	Operator number in Cut set	Cut set probability (10^{-2})	Important degree
1	18	7	51.19%
1	19	4	29.25%
2	9,14	0.21	1.54%
2	9,15	0.12	0.88%
2	10,14	0.63	4.61%
2	10,15	0.36	2.63%
2	11,14	0.42	3.07%
2	11,15	0.24	1.76%
2	27,28	0.24	1.76%
3	1,4,21	0.0288	0.21%
3	1,4,22	0.0144	0.11%
3	1,5,21	0.0336	0.25%
3	1,5,22	0.0168	0.12%
3	1,6,21	0.0384	0.28%
3	1,6,22	0.0192	0.14%
3	2,4,21	0.0336	0.25%
3	2,4,22	0.0168	0.12%
3	2,5,21	0.0392	0.29%
3	2,5.22	0.0196	0.14%
3	2,6,21	0.0448	0.33%
3	2,6,22	0.0224	0.16%
3	3,4,21	0.024	0.18%
3	3,4,22	0.012	0.09%
3	3,5,21	0.028	0.20%
3	3,5,22	0.014	0.10%
3	3,6,21	0.032	0.23%
3	3,6,22	0.016	0.12%
Total			100%

5 Conclusions

Due to the characteristics and advantages of GO method, more and more attention and application in the enterprise and scholars are gradually done in recent years. In this paper, through the establishment of DLA structure model, and the diagnostic analysis of failure factors, we established the GO chart model of DLA reliability using the GO method. This model not only can accurately calculated the reliability of system, but also can diagnose and calculate the important degree of failure mode of DLA, at the same time the calculation of the method is simple, easy to programming.

The reliability analysis of DLA through GO method, can not only calculate the reliability of DLA and the reliability of each subsystem, but also can find the factors that restricts the successful operation of the DLA, which provides important basis for the improvement of the reliability of dynamic logistics.

Acknowledgments

This paper is supported by the Science and Technology Research Project of Chongqing Municipal Education Commission (Grant No. KJ1400415). The youth scientific research project of Chongqing University of Posts and Telecommunications (Grant No. A2013-34), the doctoral start-up fund of Chongqing University of Posts and Telecommunications (Grant No. A2013-20), Chongqing science and technology project (Grant No. cstc2012gg-yyjs80024) and the open Fund of Chongqing E-commerce and Modern Logistics Key Laboratory (NO. ECML201404).

References

[1] Wang X L, Ma S H 2010 Logistics resource integration of a multi-stage supply chain with service time windows *Systems Engineering* **28**(12) 1-5

[2] Chen A, Yang H, Tang W 1999 H.A capacity related reliability for transportation networks *Journal of Advanced Transportation* **33**(2) 183-200

[3] Halldórsson Á, Aastrup J 2003 Quality criteria for qualitative inquiries in logistics *European Journal of Operational Research* **144** 321-32

[4] Wang N, Lu J C, Kvam P 2006 *IEEE Transactions on Reliability* **55**(3) 525-34

[5] Yu X C, Ji J H 2007 The Researches on the optimization reliability of Logistics System *Journal of Industrial Engineering Manage- ment* **21**(1) 67-70

[6] Chen J, Yan Q P, Huo Y M 2011 Regional emergency logistics network design Based on reliability analysis *Journal of Southwest Jiaotong University* **46**(6) 1025-31

[7] Shen Z P, Huang X R 2004 Principle and application of GO methodology *Beijing: Tsinghua University Press*

[8] Cai G Q, Zhou L M, Li X,et al. 2011 Reliability Analysis of Urban Rail Transit Vehicle's Door System Based on GO Method *Journal of Southwest Jiaotong University* **46**(2) 264-70

[9] Wang C, Zhang X S, Xu Z 2009 Reliability analysis of ultra high voltage direct current system based on GO methodology *Journal of Zhejiang University(Engineering Science)* **43**(1) 159-65

[10] Wang Z, Bao C Y 2007 Application of GO methodology for reliability analysis o f YAG laser system *Journal of Tsinghua University: Natural Science Edition* **47**(3) 377-80

[11] Wang S J, Wang G L, Zhai G F 2006 Application and research of GO methodology in reliability analysis of synthetical simulated experiment system of railway electronic antiskid devices *Systems engineering theory & practice* (10) 95-101

[12] Jia Z K 2011 Reliability Analysis of emergency management system based on GO Method *Systems Engineering* **29**(10) 123-6

[13] Zhang G B, Chen G H, Pang J H 2010 Application of GO methodology in reliability analysis of supply chain *Journal of Chongqing University* **33**(12) 40-6

[14] Cai J M, Zeng F 2007 Reliability analysis of the supply chain based on the GO Methodology *Journal of Highway and Transportation Research and Development* **24**(3) 141-4

[15] Tian SH, Wang X, Chang L 2013 Reliability of whole circulation process for porkproduct based on GO methodology *Transaction of the Chinese Society for agricultural machiner* **44**(6) 168-74

[16] Yao J M, Liu L W 2007 A decision analysis on supply chain resource integration in 4PL mode *Systems Engineering* **25**(4) 1-7

Authors

	Shuaihui Tian, born in November, 1984, HeBei Province, China **Current position, grades**: Dr and Lecturer at Chongqing University of Posts and Telecommunication, China. **University studies**: Doctor's degree in Chongqing University. **Scientific interest**: logistics and E-commerce. **Publications number or main**: 10 research papers.
	Lan Chang, born in May, 1985, Chong Qing city, China **Current position, grades**: Master, lecturer at Chongqing Electric Power College, China. **University studies**: Master's degree in Chongqing University. **Scientific interest**: logistics and supply chain management. **Publications number or main**: 4 patents, 6 research papers.

Multi-feature fusion based spatial pyramid deep neural networks image classification

Qingyong Xu[1, 2]*, Shunliang Jiang[2], Wei Huang[2], Longzhen Duan[2], Shaoping Xu[2]

[1]*School of Economic and Management, Nanchang University, Nanchang 330031, China*

[2]*School of Information Engineering, Nanchang University, Nanchang 330031, China*

**Corresponding author e-mail: xyongle@ncu.edu.cn*

Received 1 October 2013, www.cmnt.lv

Abstract

The scalable and efficient multi-class classification algorithm is now a well-known hard problem. Traditional methods of computer vision and machine learning cannot match human performance on images classification tasks. This paper proposes a novel semi-supervised classifier called Spatial Pyramid Deep Neural Networks (SPDNN). SPDNN utilizes a new deep architecture to integrate the ability of neural networks and spatial pyramid model because deep neural networks do not considerable the spatial information. Feature fusion has been more and more important for image and video retrieval, indexing and annotation because of the lack of single feature. We use multiple feature fusion over any single feature instead of pixels of images. The features include color feature, shape feature and texture feature. The performance of experiment shows that the algorithm improved the state-of-the-art image classification.

Keywords: multi-feature fusion, spatial pyramid deep neural networks, image classification

1 Introduction

In the last decades, the availability of digital images produced by scientific, educational, medical, industrial and other applications has increased dramatically. Images have become one of the main sources of information representation in human life. Thus, images retrieval and images classification has become a challenging task. In order to reach the goals, some pattern recognition techniques have been proposed and become a research hotshot. Deep learning methods as one of pattern recognition techniques have become the focus of the study in image processing and computer vision. Recent advances in deep learning methods have led to a widespread enthusiasm among pattern recognition and machine learning researchers [1, 2]. Deep learning move machine learning towards the discovery of multiply levels of representation.

Deep learning is a set of algorithms in machine learning that attempt to model high-level abstractions in data by using deep model architectures composed of multiple non-linear transformations. For deep model, it can extract more sophisticated and invariant feature from original raw input signals. Lower layers aim at extracting simple features, which are clamped into higher layers [7]. Generally speaking, deep architectures can be exponentially more efficient than shallow ones. For shallow architectures, it need more nodes in order to increase performance and leads to more time to train. For deep architectures, it increases deeper layers other than number of nodes. Those make it more efficient than shallow architecture. So the depth of architecture may be more important from the point of view of statistical efficiency [7].

The concept of neural networks started in the late-1800s as an effort to describe how the human mind performed. These ideas started being applied to computational models with Turing's B-type machines and the perceptron. A deep neural network (DNN) as one of deep learning is defined [4, 5] to be an artificial neural network with multiple hidden layers of units between the input and output layers. DNNs can model complex non-linear relationships. The extra layers enable composition of features from lower layers, giving the potential of modeling complex data with fewer units than a similarly performing shallow network [4, 8]. The main purpose of DNNs is to extract generally useful features from unlabeled data, to detect and remove input redundancies, and to preserve only essential aspects of the data in robust and discriminative representations [9]. DNNs fully unfold their potential when they are big and deep [10].

In recent years, spatial pyramid model has been extremely popular in image classification. Spatial Pyramid is a widely used method for embedding both global and local spatial information into a feature, and it shows good performance in terms of generic image recognition and classification [11]. For spatial pyramid model, the image is divided into a sequence of increasingly finer grids on each pyramid level. Then the features are extracted from every grid cell and are concatenated to form one huge feature vector. Spatial information is usually embedded in the feature extraction process.

Since the emergence of extensive multimedia data, because of the lack of single feature, feature fusion has been more and more important for image and video retrieval, indexing and annotation. Existing feature fusion techniques simply concatenate a pair of different features or use canonical correlation analysis based methods for joint dimensionality reduction in the feature space.

In this paper we propose a novel semi-supervised classifier called spatial pyramid deep neural networks (SPDNN). The SPDNN utilizes a new deep architecture to integrate the abstraction ability of deep neural nets (DNN) and discriminative ability of spatial pyramid model. For input data, we use multiple feature fusion over any single feature instead of pixels of images because of the lack of single feature and pixels.

The paper is structured as follows: In section 2 we will detail feature fusion. Sections 3 present the framework of our proposed spatial pyramid deep neural networks model. Section 4 asserts the validity of our method by the experiment using COREL 1000 data-set and section 5 draws the conclusions and points out future work.

2 Features fusion

In our lives, there are more and more technology and feature extracting methods to be proposed for image retrieval and image classification in order to increase the precise. In the beginning, the researchers mainly focus on the text based image retrieval(TBIR) using simple feature, then more and more researchers start to research the content based image retrieval(CBIR) because of the limitation of TBIR. If the features come from the entire image, we called it as global based image retrieval (GBIR) and otherwise we called region based image retrieval (RBIR). The features of GBIR are often low-level features from images. RBIR focuses on contents from regions of images not form entire image. Generally speaking, RBIR have better performance than GBIR and TBIR. The performance of the methods including CBIR, GBIR and RBID is depended on features extracting.

The features consist of colour feature, texture feature and shape feature.

The color feature is one of the most widely used visual features and is invariant to image size and orientation. But the color feature does not contain the spatial information. Some CBIR systems employ color to retrieve images such as QBIC system and Visual SEEK. Colour features consist of color moment, color histogram, the edge histogram, Gabor wavelet transform, partial binary image, GIST etc. Color moment is often used for color representation. It contains mean, variance and skewness.

Texture is another important characteristics for image retrieval. Image texture refers to surface patterns which show granular details of an image. Image texture gives us information about the spatial arrangement of color or intensities in an image or selected region of an image, and it can be used for image segmentation or classification. Texture includes edge detection, gray co-occurrence matrices (GLCM), autocorrelation features, Tamara [12] feature etc. Co-occurrence matrix is constructed based on the distance and orientation between pixels. GLCM is one of the most well-known and widely used texture features and is defined based on different combinations of pixel brightness values (i.e., grey levels). It considers the spatial relationship among pixels. Tamura is another important texture feature and consists of coarseness, contrast, directionality, linelikeness, regularity and roughness.

Shape feature is different from color feature and texture feature. Popular shape feature consists of edge histogram, Fourier descriptors, polygonal approximation, invariant moments, curvature scale space, etc. Invariant moments is proposed by Hu [13].

Obviously we cannot cover all features that have been proposed in our experiment. We select some features that can represent images. For color feature, we select the color moment in RGB and HSV color space because it is simple to adopt and effective for retrieval. It mainly describes the image color distribution. For texture feature, we use the GLCM and Tamara. For Shape feature, we use the invariant moments.

3 Spatial pyramid deep neural networks

In order to increase the performance of image retrieval, the machine learning methods are applied. Deep learning which is one of machine learning is proposed in recent years and is a set of algorithms in machine learning that attempt to model high-level abstractions in data by using architectures composed of multiple non-linear transformations. And it is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text.

A deep neural network is one of the most important deep learning methods. For labeled training examples ($x^{(i)}$; $y^{(i)}$). Neural networks give a way of defining a complex, non-linear form of hypotheses $h_{W;b}(x)$, with parameters W; b that we can fit to our data. A neural network is put together by hooking together many of our simple neurons, so that the output of a neuron can be the input of another. For example, a small neural network as Figure 1.

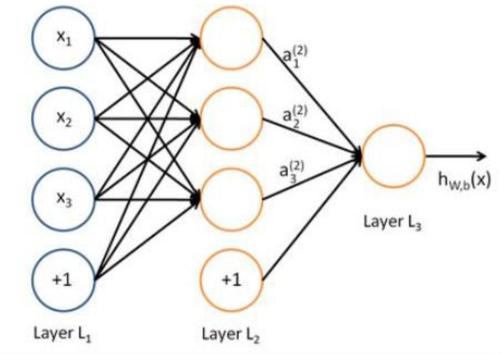


FIGURE 1 The Graph of an NN with hidden

A deep neural network (DNN) is defined [4, 5] to be an artificial neural network with multiple hidden layers of units between the input and output layers. Similar to shallow ANNs, DNNs can model complex non-linear relationships. The extra layers enable composition of features from lower layers, giving the potential of modeling complex data with fewer units than a similarly performing shallow network. DNNs are typically designed as feed forward networks, but recent research has successfully applied the deep learning architecture to recurrent neural networks for applications such as language modelling [14].

A DNN can be discriminatively trained with the standard back propagation algorithm. The weight updates can be done via stochastic gradient descent using the following Equation (1).

$$\triangle W_{ij}(t+1) = \triangle W_{ij}(t) + \mu \frac{\partial C}{\partial W_{ij}} , \qquad (1)$$

where μ is the learning rate and C is the cost function. He choice of the cost function depends on factors such as the learning type (supervised, unsupervised, reinforcement, etc.)

and the activation function. But the deep neural networks do not considerable the spatial information.

A SPDNN is composed of one or more spatial pyramid layers with fully connected layers on top. It also uses tied weights and pooling layers. This architecture allows SPDNN to take advantage of the 2D structure of input data. In comparison with other deep architectures, convolutional neural networks are starting to show superior results in both image and speech applications. They can also be trained with standard back propagation.

4 Experiments and analysis

4.1 EXPERIMENTAL SETUP

In the following we give a detailed description of all the experiments we performed. We evaluate our architecture on various commonly used object recognition benchmarks and improve the state-of-the-art on all of them. The architecture of SPDNN is three layers used for the experiment. The description of the SPDNN is given as following: architecture of the pyramid is 1-4-16; architecture of the deep neural networks is 48-500-500-500-500-500-10(48 is the size of input data and 10 is the output of SPDNN). The architecture has five hidden layers with 500 hidden units and a fully connected output layer. All SPDNN are trained using on-line gradient descent. Initial weights are drawn from a uniform random distribution in the range [-0.05; 0.05] [15].

4.2 DATA-SET

The Corel image database contains a large amount of images containing various contents, ranging from animals and outdoor sports to natural scenes. It is often used for image retrieval system and image classification. There are two subsets. One is the min Corel image set with 1000 images and another is a bigger images set with 10000 images. In our experiment, the Corel 1000 images set is used in order to compare with other results of anthers. The Corel 1000 is a data-set have 1000 labeled high-resolution images with JPEG format belonging to 10 categories with 100 images each. The 10 categories are Africa, Beach, Buildings, Buses, Dinosaurs, Flowers, Elephants, Horses, Food and Mountains. The size of every image is 256×348 or 348×256. The images of every category are shown as Figure 2.

FIGURE 2 Examples of 10 class images

4.3 FEATURE EXTRACTING AND NORMALIZE

The feature extracted is one of the most important for image classification. Obviously we cannot cover all features that have been proposed in our experiment. However, we have tried to make the selection of features as representative and at the state-of-the-art as possible. Generally speaking, the features can be classified into three groups: colour features, texture features and shape features. Some good features is crucial for obtaining competitive performance in classify-cation. For color features, we select color moment proposed by stricker [16]. The colour moment include mean, variance and skewness. It does not need color space quantization and the dimension of feature vectors is low. It can be extracted from RGB and HSV space. For texture features, we select tamura, entropy and gray-level co-occurrence matrices (GLCM). Tamura consist of six texture features (coarseness, contrast, directionality, linelikeness, regularity, and roughness) corresponding to human visual perception: Image entropy is a quantity which is used to entropy measures the randomness of the distribution of intensity levels in bins. Co-occurrence matrix is constructed based on the distance and orientation between pixels. GLCM is one of the most well-known and widely used texture features and is defined different combinations of pixel brightness values (grey levels). It considers the spatial relationship among pixels. For shape features, Hu invariant moment is used. Hu derived these expressions from algebraic invariants applied to the moment generating function under a rotation transformation. They consist of groups of nonlinear centralised moment expressions. The result is a set of absolute orthogonal moment invariants, which can be used for scale, position, and rotation invariant pattern identification. Thus we can obtain 31 features (color:9, tamura:6, entropy:1, GLCM:8, Hu invariant moment:7) for each images. The data-set is 1000×31 array and every row represent an image. By this way, every row represented features of an images which contained the color feature, texture feature and shape feature. Then we used SPDNN to train the train data-set and test the testing data-set using the training results.

After feature extracting, the features are normalized in order to keep as the unified scale because the scale of different feature is not same, so the normalization is needed. Normalization of the feature refers to adjusting values measured on different scales to a notionally common scale and brings the indicators into the same unit. The intention is that these normalized values allow the comparison of corresponding normalized values for different data-set in a way that eliminates the effects of certain gross influences. In our work, 0-1 normalization is use. This is also called unity-based normalization. The formulation as follows:

$$X_{\mathrm{i}} = \frac{x_{\max} - x_i}{x_{\max} - x_{\min}}. \qquad (2)$$

By this way, the feature scaling used to bring all values into the rang [0, 1].

4.4 PARAMETERS SETTING

After pre-processing, the parameters must be set for SPDNN. The architecture of SPDNN is three layers used for the experiment. The description of the SPDNN is given as following: architecture of the pyramid is 1-4-16; architecture of the deep neural networks is 31-500-500-500-500-500-10(31 is the size of input data and 10 is the output of SPDNN). The networks have one visible layer, five hidden layer and an out layer. Each hidden lever has 500 hidden units and the output layer has 10 units. The sign function as feature mapping function is used. The sign functions as follows:

$$f(x) = g(Wx+b) = \frac{1}{1+e^{-(wx+b)}} \,. \qquad (3)$$

We do the experiment using the learning rate from 0.01 to 1, and the moment from 0.1 to 1. The experiment as Figure 3 shows that the learning rate has great effect on the results, and other elements have little influence on the result.

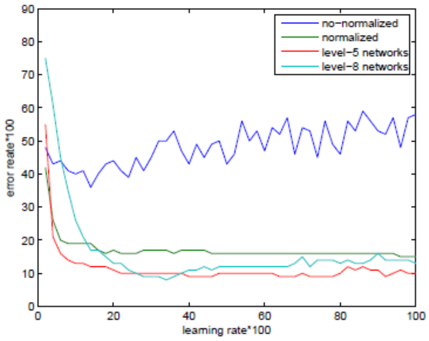


FIGURE 3 Learning rate figure

There are many hyper-parameters involved in training this deep learning method. The parameters in Table 1 which is determined by lots of experiments using different deep learning methods is used to test the effectiveness of our dataset and our experience.

TABLE 1 Parameters lists

Parameters	Value
number of hidden unit	90
learning rate	0.4
zero Masked Fraction	0.5
momentum	0.5
alpha	0.5
weight-decay	0.0001
number epochs	500
number epochs function	sigma

4.5 ANALYSIS

4.5.1 Data Grouping

In our experiment, there are 1000 images. The images was randomly into 10 groups, and each group has 100 samples. The results of groups is showed in Table 2.

TABLE 2 Data-set groups

Groups	Africa	Beach	Building	Buses	Dinosaurs	Flowers	Elephants	Horses	Food	Mountains	Total
1	10	5	10	11	12	9	10	14	9	10	100
2	11	14	6	7	14	9	8	12	10	9	100
3	8	8	10	12	10	8	5	17	11	11	100
4	10	14	11	16	5	12	9	7	4	12	100
5	11	10	16	12	4	11	7	10	7	12	100
6	9	11	11	7	7	11	10	11	7	16	100
7	13	10	9	9	6	10	14	11	10	8	100
8	8	12	11	11	10	10	14	7	10	7	100
9	12	6	9	10	18	11	10	4	12	8	100
10	8	10	7	5	14	9	13	7	20	7	100
total	100	100	100	100	100	100	100	100	100	100	1000

4.5.2 Experiment results

For our groups, we select 9 groups as training set and 1 group as testing set. We do 10 experiments, each with different test sets. 10 results are obtained. The average result of classification as Figure 4.

From the correct rate of every groups, we know that the best is 91% for tenth group. The worst is 78% for seventh group. The average correct rate is 84.2%. It is better than the-state-of-the-art.

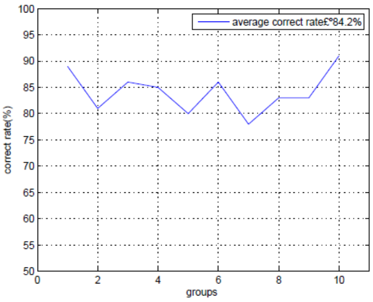


FIGURE 4 Correct rate for every group

4.5.3 Compared with single feature

Table 3 lists the classification results of several common features, including histogram, color straight direction, gray level co-occurrence matrix, color co-occurrence matrix and the results in this paper. It is seen that the average correct classification rate of single features are not more than 70%, and the results of multi feature fusion is achieved 84.2%. It has better performance than single feature.

4.5.4 Compared with different methods

Table 4 lists the common image classification and related scholars are the classification result is at the COREL 1K gallery. It is seen that both in the average correct rate and maximum/minimum rate of correct classification, the SPDNN algorithm has better performance.

TABLE 3 Compared with Single Feature

Feature	Average correct rate (%)
Gray level histogram(size:16)	69.9
RGB histogram(size:16)	67.4
Index histogram(size:16)	57.2
Dominant colour descriptor(size:16)	48.6
Dominant Codebook(size:16)	38.1
Grey Level Co-occurrence Matrix	67.4
Color Co-occurrence Matrix	58.4
Gabor Wavelets	58.8
Scan pattern co-occurrence matrix	50.2
This paper	84.2

TABLE 4 Compared with different methods

Methods	Best	Worst	Average（%）
SIMPLIcity(2013) [26]	98.1/Dinosaurs	33.0/Building	46.7
Edge based(2013) [26]	95.0/Dinosaurs	25.0/Elephants	51.0
Fuzzy Club(2013) [26]	95.0/Dinosaurs	30.0/Elephants	55.9
DD-SVM(2004) [25]	99:7/Dinosaurs	----	81.5
CS_LBP(2012) [25]	96.2/Dinosaurs	31.4/ Mountains	59.1
LEPSEG(2012) [25]	96.0/Dinosaurs	37.2/ Mountains	65.2
LEPINV(2012) [25]	95. 5/Dinosaurs	34.9/ Beach	60.8
Hiremath's method (2007)[25]	95.0/Dinosaurs	30.4/Beach	54.9
Wang-yu-yang method (2010)[28]	95.0/Dinosaurs	30.0/Mountains	59.2
M.Babu Rao, Ch.Kavitha etc method(2013) [26]	99.0/Dinosaurs	55.0/Beach	75.13
Fazal Malik Baharum (2013) [27]	100.0/Dinosaurs	60.0/Food	82.0
This paper	100.0/Dinosaurs	64.0/Building	84.2

Due to the characteristics of the image itself, such as the difference between the differences of different objects, different in the image size, the foreground and background color of the size and not the same image, image classification correct rate of different category has certain difference. From the experimental results, ten types of images, classification of each class is the correct rate of each are not identical, the dinosaur a set of correct classification rate is highest, for 100%, all classified correctly; secondly is the flower and the automobile, the correct rate of classification was 99% and 98%; the correct ratio of less than 80% of the building, Africa, elephant and beach four class.

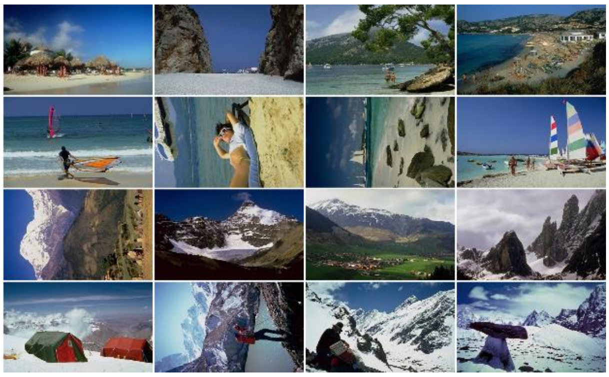


FIGURE 5 Some misclassification images

In real images, images belong to the same category sometimes have the obvious difference, and the images which belong to different categories and sometimes very similar. This is mainly because the channel between the image low-level features and high-level semantic. The semantics for the same class, both in the form are quite different, image semantic belong to different categories, may form is very similar, this will cause great difficulty for image classification. For example, the beach has 8 images is divided into the mountains of this group, mountains has 8 images is divided into beach in this group, the 16 images in Figure 5.

In Figure 5, the images of first two rows belong to beach but they are misclassified to mountains. The images of last two rows belong to mountains but they are misclassified to beach. As show in Figure 5, the image itself is not much difference and they are very similar.

5 Conclusion

Deep neural networks (DNN) as one of deep learning methods and Spatial pyramid are an active research topic in image processing and computer vision Based on DNN and spatial pyramid, we proposed a new methods called multi-future fusion spatial pyramid deep neural networks (SPDNN). SPDNN utilizes a new deep architecture to integrate the advantage of deep neural networks (DNN) and overcome the disadvantages of DNN without considering the spatial structure of image. Then it is successfully applied to visual data classification. For input data-set, images pixels are replaced by the based features for input of SPDNN. By this way, the size of input data vector can be reduce and keep the information of images. In our experiment, we stochastically classify the images database into 10 groups. The results show that SPDNN has better performance than the-state-of-the-art.

The further work will be explored from two aspects. Firstly, we will study how to determine the scale of deep architecture for various applications and the parameters are decide. Secondly, we will consider that how to improve the performance of region based image retrieval and classification use deep learning method in a large scale data-set.

Acknowledgments

This work was supported by Grants 61363046 & 41261091 & 61163203 approved by the National Natural Science Foundation of China.

References

[1] Markoff J 2012 Giant steps in teaching computers to think like us: neural nets mimic the ways human minds listen,see and execute *International Herald Tribune 24-25 (November)(2012)* 1-8

[2] Larochelle H 2007 An empirical evaluation of deep architectures on problems with many factors of variation *Proceedings of the 24th international conference on Machine learning ACM* 2007 473-80

[3] Lopes N, Ribeiro B 2013 Towards adaptive learning with improved convergence of deep belief networks on graphics processing units *Pattern Recognition* (47) 114-27

[4] Bengio Y 2009 Learning Deep Architectures for AI. Foundations and Trends in Machine Learning 1-127

[5] Schmidhuber J 2013 Deep Learning in Neural Networks: An Overview *arXiv preprint arXiv* 1404.7828

[6] Liou C-Y. 2013 Auto encoder for words *Neurocomputing* **139** 84-96

[7] Bengio Y, Courville A, Vincent P 2013 Unsupervised feature learning and deep learning: A review and new perspectives *IEEE Transaction Pattern Analysis and Machine Intelligence (PAMI)*

[8] Masci J 2011 Stacked convolutional auto-encoders for hierarchical

feature extraction.Artificial Neural Networks and Machine LearningCICANN *Springer Berlin Heidelberg* 52-9

[9] Ciresan D C 2011 Flexible, high performance convolutional neural networks for image classification. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (22)

[10] Harada T, Ushiku Y, Yamashita Y 2011 Discriminative spatial pyramid *Computer Vision and Pattern Recognition (CVPR) 2011 IEEE Conference on IEEE* 1617-24

[11] Tamura H, Shunji M, Takashi Y 1978 Textural features corresponding to visual perception *Systems, Man and Cybernetics IEEE Transactions on* 460-73

[12] Hu M-K 1962 Visual pattern recognition by moment invariants *Information Theory IRE Transactions on* 179-87

[13] Mikolov T 2010 Recurrent neural network based language model *Interspeech*

[14] Ciresan D, Meier U, Schmidhuber J 2012 Multi-column deep neural networks for image classification *Computer Vision and Pattern Recognition (CVPR)*

[15] Stricker M A, Orengo M 1995 Similarity of color images. IST/SPIE Symposium on Electronic Imaging: Science and Technology *International Society for Optics and Photonics*

[16] Rao M Babu, et al. 2013 A new feature set for content based image retrieval *Information Communication and Embedded Systems (ICICES), 2013 International Conference on IEEE*

[17] Hinton G E 2002 Training products of experts by minimizing contrastive divergence *Neural Computation* **14**(8) 1771-800

[18] Subrahmanyam Murala R P Maheshwari R Balasubramanian 2012 Directional local extrema patterns a new descriptor for content based image retrieval *Int J Multimed Info Retr* 191-203

[19] Hiremath P S, Pujari J 2007 Content based image retrieval using color, texture and shape features *Advanced Computing and Communications International Conference on IEEE*

[20] Wang X-Y, Yu Y-J, Yong H-Y 2010 An effective image retrieval scheme using color, texture and shape features *Journal of computer stabdards and interfaces*

[21] Fazal Malik, Baharum Baharudin 2013 Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain *Journal of King Saud University - Computer and Information Sciences* **25**(2) 207-18

[22] Zheng L, Wang S, Qi Tian 2013 Coupled Binary Embedding for Large-scale Image Retrieval *IEEE Transactions on Image Processing (TIP)* (8) 3368-80

[23] Zheng L, Wabg S, Liu Z, Qi Tian 2013 Packing and Padding: Coupled Multi-Index for Accurate Image Retrieval *In: CVPR*

Difference processing on financial accounting of institutions and enterprises

Hong-xia Liu*

Nanyang Normal University, Nanyang City, Zip Code: 473061, Henan Province, China

Corresponding author's e-mail: zmeditlqx@sina.com

Received 21 November 2014, www.cmnt.lv

Abstracts

In modern society, both enterprises and institutions would face a common problem, which was how the accounting staff handle the finances. Due to the different properties, this two types of units could lead to the difference how to handle this units. This paper was to research and analyze problems from this aspect, and based on the analysis and research, and to compare the specific difference of institutions and enterprises, then put forward the corresponding point of view for research better understanding.

Keywords*:* Institutions, Enterprises, Accounting Staff, Financial Processes, Difference

1 Introduction

The development of society had promoted the constant progress of economic system. The most obvious reflection was in the accounting field. Our country had implemented a new accounting standard for business enterprises, and made our country's accounting standards gradually bridge to the world. And we all knew that institution was the national department, which was the non-profit organization, the main purpose was to provide public service for the society. As the country further reforms had penetrated in a wide range of industries in society, there was no exception for institutions. The state had reformed the management mode of institutions, and made the management pattern of institution was similar to enterprises operation and management mode. In the end of 2012, the institution accounting system had been issued, this system had exceeded the previous accounting system, and also showed the characteristics of accounting management system reform after institution. Although the enterprises and public institutions were quite different in accounting management, the two organization had certain similarities between them. From the accounting system of the two organization, this study analyzed the two systems were different, this could assist institutions learn from enterprises management mode, and offer assistant to deepen the reform of institutions.

2 Basic classification of accounting

There was a process for the forming of accounting system in our country. The forming of this system was mainly to convenience people daily financial management. In general, according to the accounting object, it was divided into two categories: enterprise accounting and budget accounting. Enterprise accounting was mainly suitable for society organization, including agriculture, industry, business, enterprise unit. Its purpose was to supervise and manage each combination and process of social production in the field of enterprise capital operation situation. While, the budget accounting applied to the government department, institutions and administrative unit of the society. Its purpose was to supervise and manage the process of social production, material distribution and social welfare institutions in funds management field. Typically, government departments and institutions did not directly provide material products, instead, they mainly provided some public services for the society, played a significant role in social production activities. The main difference between the enterprises and institutions was social functions. When institutions held various public activities, most were paid by individuals or units, and the investment was free and voluntary. Most of the money was appropriation by national finance department. Therefore, the accounting management of institutions was based on the social benefits, such as measurement, recording, reporting, etc.

3 Reason analysis of institutions and enterprises for accountant processing

In financial accounting treatment, institutions and enterprises had a very big difference, here listed was the reason why the difference exists and the analysis of the three aspects.

The fundamental reason for the difference between institutions and enterprises was because the unit was a big difference in nature between. The differences of course was only part of the reason, and enterprises and institutions internal financial management aspects of the difference is more obvious. Institution was to provide public service for the society, it was a non-profit organization, their work was unpaid, and basically no profit. Institution was mainly to ensure the normal development of society and people's daily activities. Enterprise unit was focused on the maximization of self-interest, so many of their daily activities and accounting work was based on profit for the purpose. In addition, the institution of the profit and loss had nothing to do with the unit itself, while the company unit of profit was closed related to its loss.

Institutions and enterprises were the main reason for the difference, was because they both had significant difference on the internal capital form. Under the normal circumstances, the institution of the funds was very fixed, his internal capital form is divided into two kinds: liquidity and conserve cash. Institution's liquidity referred to the government's funding, these funds in the process of circulation was basically not possible money back. Regular money was fixed assets within the unit. And enterprises and institutions were quite different in this respect, in addition to some fixed building enterprise unit, the rest of the money had been in circulation in order to be able to earn more money, the differences made them in financial accounting treatment be a very big difference.

For institutions and enterprises, there were a difference in terms of money, this was what they both in financial accounting treatment of different direct cause. Capital operation was to do something with money. To some extent, the institution of capital operation was a form of social services and establishing a good security system for the society. So, enterprises in the use of funds, was considering how to legalize the funds. The enterprise units were more to take the money out to invest other business and expect in return for greater economic benefits, realized the enterprise capital flow, they were considered when using money more and how to make these funds into maximize interests.

4 Financial accounting treatment differences of institutions and enterprises

4.1 THE DIFFERENCE IN THE FORM OF ACCOUNTING ENTRIES

In daily work, the basic work of accounting staff was resorting the accounting work. The capital of the work on the establishment of the financial records had a vital role, the difference of capital entry work was enterprises and institutions in the financial accounting treatment of a concrete embodiment. For institutions and enterprises to take loans, we illustrated it would be different results even the loan subject was same, due to the two units was different in area of financial system. Therefore, accounting staff would sort the accounts based on the account name. In addition, for accounts these lend funds, according to the duration of loan record the payments, the enterprises would divided it into long-term borrowing and short-term borrowing. Institution was on the contrary, institution loan would like to put the money into the unit interior projects, these two approaches was different, the two units in the overall financial management had a certain difference, also brought the certain effect accounting of financial records management work, the formation of the differences and the nature of the unit had a certain relationship between.

4.2 THE DIFFERENCE ON THE TAX PAYMENT

For taxes payment, a unit must fulfill the responsibility, at the same time, it was also an important aspect for constituent units for financial problems. Both institutions and enterprises needed to pay taxes to the state. While, these two units had a certain differences on the tax issues in processing. The two units would purchase a certain amount of production data, for example, under the normal circumstances, the institution tax payment method was more complicated, they would launch a distinction and made to taxpayers. In dealing with these tax, generally, there were two ways, the first was the institution purchase was the price without tax; The second was the material production of institution for its own use which was include tax. Instead, enterprises in the treatment of the tax issue was relatively simple, they divided taxpayers into two kinds: general taxpayer and small unit taxpayer. For the general taxpayer, there was no need for them to pay tax for the raw material, for the post one, when dealing with matter, it became relatively easy. It mainly due to dividing the taxpayers into tow type: the general taxpayer, and small-scale taxpayers. There was no tax when the former purposed raw material, the latter in purchasing production material always had certain tax price.

4.3 DIFFERENCE IN EXPENDITURE ACCOUNTING

In daily business activities, any enterprises needed process the expenditure. In the treatment of the cost of the accounting. Firstly, the enterprise unit main purpose was profit, so what they thought more was cost and profit accounting, based on this, they implemented a series of measures to manage the production of the enterprise profit, their ultimate goal was the cost accounting and accounting profits. Instead, Institutions were different, when check the cost accounting, the profits was not considered. In daily operation process, when needed to check the cost accounting, they could count it into the unit itself spending account. Through the comparison and analysis, institutions and enterprises had a certain difference on cost calculation, the cause of the difference was the purpose of the unit content of cash flow.

4.4 DIFFERENCES IN ACCOUNTING AND BUSINESS ACTIVITIES

No matter institutions or enterprises, in the process of production and business operation, they would have a business activity, on the accounting performance of both thing was not the same. Due to the business unit's business activities mainly for business class and professional class, its spending and income included the business activities of the whole business activities of the expenditure and revenue. These accounts were in the business activities between accounting difference. Took pending as the research object, this study set the activities in the relevant administrative department office expenses, there was a difference between, enterprises would take the activity cost accounting in the records of financial management, and professional activities associated with the business entity would be as spending.

4.5 THE DIFFERENCE IN FINANCIAL STATEMENTS

Institutions and enterprises also had certain difference on accounting statements, institution accounting statements, such as the daily revenue expenditure tables, sheets, government subsidies, etc. This unit report mainly reflected business entity for a specified period of operating conditions and relevant information, these statements played a certain reference value, and was reference to the relevant units and daily inspection unit. And could be the instruction of operation be accordance with the specification. While the enterprise unit was not like this, their report was mainly for some enterprise profit statement, balance sheet and cash flow table, etc. These statements could react systematically and comprehensive enterprise information in the normal course of business activities, and also could reflect enterprises in financial management and operation for the enterprise decision makers to provide some beneficial information, convenient they work out scientific and reasonable scheme.

4.6 DIFFERENCE ON ASSETS PROCUREMENT PROCESSING

Production materials could be required for enterprises and institutions in the daily procurement internal needs. The accounting staff shall make records of these information.

Institutions and enterprises had a certain differences between, the purchased assets of enterprises had a category of “fixed assets”, recycling had a “bank deposit” category. While, institutions unit was different, accounting staff would had two records, one was the category” fixed assets” and “non-current assets funds”; another was the “services” and “operating expenditure”.

4.7 DIFFERENCE ON ACCOUNTING OF FINAL TRANSFERRING

There were many differences for the final transferring of institutions and enterprises. For institution, their final accounting transaction, the incomes and expenditure accounts were shifted to company annual profits accounts. The purpose was to convenient the proportional distribution of the profits. While, the institution was different, accounting staff transferred incomes accounts to finance allowance accounts, according to the different properties and USES funds, and expenditure accounts were shifted into non-fiscal allowance accounts. The funds were transferred into fiscal spending allowance accounts in the end of annual year.

4.8 DIFFERENCES IN REPORTING

There were certain difference of enterprises and institutions in preparation of accounting statements. Under the normal circumstances, there were main four reasons for institutions: the profit statement, balance sheet, funds recycling statement and shareholders’ equity table. These statements, report, or table could fully reflect the whole enterprise daily financial situation, operating conditions and capital structure, and it was very beneficial to facilitate the enterprise to provide information to the enterprise users. The purpose was to help them make decisions more scientific and effective. For institution reports, there were main three types: the balance sheet, income and expenditure tables and fiscal subsidy. For institutions, these tables were mainly embodies the business entity in a specific period of time’s financial situation and business activities. These tables could provide advantaged information for institutions in economic and financial management, and could also help the business unit leadership analysis unit interior operation specification and improve the efficiency of financial management.

4.9 THE DIFFERENCE ON GOODS INVENTORY PROCESSING

Institutions and enterprises had certain difference on goods inventory. General storage could be less at some time, and full at some time. This condition was consistent with the loss and gain of profits. The checking of goods inventory of enterprises was according to the batches of examination and approval in a time order, and could be divided into pending property loss accounts and different credit management cost respectively. While, the profit and loss of enterprise was in the same situation, which could be divided into property accounts loss and different situation to be processed respectively and saving management charges. While, institutions was on the contrary, the goods storage and inventory was not according to the examination and approval, and would not base on the record of the income, the profit and loss of these goods was directly recording of expense.

5 Conclusion

As is known to all, enterprises and public institutions were the two main forms of our country unit, this two units had significant difference at unit properties and funds operation conditions. These may lead to both had many differences in accounting and financial management. Through the series of analysis of studies, this paper had confirmed the differences of financial management, accounting records, business activities, the accounting and reporting. The purpose was to give the accounting management personnel assist to these unit in order to improve their efficiency in the following work.

References

[1] Zhang Yun 2012 Difference analysis of institutions and enterprises accounting financial processing *Accounting and Finance* **8** 56

[2] Wei Pei-shu 2013 Similarities and differences of institutions and enterprises in the accounting treatment *Modern economic information* **8** 115

[3] Xu Ai-mei 2014 Comparison of institutions and enterprises accounting financial processing *Finance Managing* **2** 55-8

[4] Tang An-xiang 2009 Influence of the limitations of financial accounting in the new period to accounting information quality *Commercial economy* **2009**(20)

[5] Xie Lan-lan 2008 Treatment of human resource accounting from reliability and relevance *Caikuai Monthly publication (integrated)* **2008**(06)

[6] Wang Li-yan, Wu Li-na, Luo Zheng-ying 2012 Accounting principle *Peking University Press* **2012**(06)

Author

Hong-xia Liu, June, 1991, Xinyang, Henan Province

Current position, grades: Certified public accountant;
University studies: Financial Accounting;
Scientific interest: The internal financial management;
Publications: 1 - Problem and suggestion on official business card payment and settlement, China Economist, July, 2014.
Experience: From 1999 to 2003, I had formally studied accounting in School of Management Harbin Institute of Technology. During the internship, I had worked as an assistant in a certified public accountants audit of Guangdong for two months, and joined in the annual audit work involved in Daya Bay hydropower station. After graduation, I worked in the bursar’s office of Nanyang Normal College in Henan Province, and engaged in college staff wage accountings, accumulation funds and unit taxes. During this period, i had joined in the accounting and cashier jobs in bank, and made the financial budget in the beginning of year, and involved in financial management statements. I am professionalized in the routine of institute accounting, the figure calculation and financial matters.

Authors' index

Chang Lan	11
Duan Longzhen	16
Huang Wei	16
Jiang Shunliang	16
Liu Hong-xia	22
Qiu Shubing	7
Tian Shuaihui	11
Xu Qingyong	16
Xu Shaoping	16

Operation Research and Decision Making

The real estate enterprise performance evaluation model study empirical research on the real estate enterprise statistics in China: 2009-2013

Shubing Qiu

Computer Modelling & New Technologies 2015 ***19****(2C) 7-10*

Based on our real estate business development, for the shortcomings of traditional performance evaluation methods, combined with hierarchical fuzzy neural network evaluation method, using BP neural network training corporate financial indicators, and fuzzy neural network training non-financial indicators, and then to build a fuzzy neural network evaluation model integratedly, so the value of enterprise performance evaluation results can be calculated. The results show: the model is of high accuracy, which can more accurately reflect the performance of the real estate development business.

Keywords: fuzzy neural network, BP neural network, real estate business, business performance

Reliability research on dynamic logistics alliance based on GO methodology

Shuaihui Tian, Lan Chang

Computer Modelling & New Technologies 2015 ***19****(2C) 11-15*

In order to calculate the reliability accurately and dig out the dominant influencing factors of dynamic logistics alliance, GO methodology is applied in reliability research on dynamic logistics alliance. Through building the structure model of dynamic logistics alliance, failure factor of each subsystem is diagnosed. With the GO methodology, the dynamic logistics alliance is transformed into the GO chart, the system reliability is calculated in detail, its failure mode diagnosis and importance calculation are quantitatively studied and then a case of automobile dynamic logistics alliance is employed to verify GO methodology for effectiveness and validity.

Keywords: dynamic logistics alliance, system reliability, GO methodology, influencing factor, failure mode

Multi-feature fusion based spatial pyramid deep neural networks image classification

Qingyong Xu, Shunliang Jiang, Wei Huang, Longzhen Duan, Shaoping Xu

Computer Modelling & New Technologies 2015 ***19****(2C) 16-21*

The scalable and efficient multi-class classification algorithm is now a well-known hard problem. Traditional methods of computer vision and machine learning cannot match human performance on images classification tasks. This paper proposes a novel semi-supervised classifier called Spatial Pyramid Deep Neural Networks (SPDNN). SPDNN utilizes a new deep architecture to integrate the ability of neural networks and spatial pyramid model because deep neural networks do not considerable the spatial information. Feature fusion has been more and more important for image and video retrieval, indexing and annotation because of the lack of single feature. We use multiple feature fusion over any single feature instead of pixels of images. The features include color feature, shape feature and texture feature. The performance of experiment shows that the algorithm improved the state-of-the-art image classification.

Keywords: multi-feature fusion, spatial pyramid deep neural networks, image classification

Difference processing on financial accounting of institutions and enterprises

Hong-xia Liu

Computer Modelling & New Technologies 2015 ***19****(2C) 22-24*

In modern society, both enterprises and institutions would face a common problem, which was how the accounting staff handle the finances. Due to the different properties, this two types of units could lead to the difference how to handle this units. This paper was to research and analyze problems from this aspect, and based on the analysis and research, and to compare the specific difference of institutions and enterprises, then put forward the corresponding point of view for research better understanding.

Keywords: Institutions, Enterprises, Accounting Staff, Financial Processes, Difference

Content D

Part D Nature Phenomena and Innovative Engineering		
Shi Li	Energy consuming control of building based on fussy temperature control	7
Xiaoshuai Xin, Jinxi Chen	Using cubature Kalman filter to estimate the vehicle state	12
Xiaolu Li, Wenfeng Zheng, Dan Wang, Lirong Yin, Zhengtong Yin	The application of R/S analysis for the earthquake prediction in Sichuan, China	18
Qiuxian Shen, Xu Wang	Effect of 3-S-isothiuronium propyl sulfonate on bottom-up filling in copper electroplating	23
Junjie Yang, Zhihe Fu, Yibiao Fan, Wenxiang Chen, Zhiping Xie, Wei Wu, Xiaoyu Shan	Study on signal processing technology based on the reflective intensity modulated fiber optic sensor	26
Authors' Index		31
Cumulative Index		32

Energy consuming control of building based on fussy temperature control

Shi Li*

Department of Architectural Engineering, Yulin University, Yulin 719000

**Corresponding author's e-mail: shili9988@126.com*

Received 1 March 2014, www.cmnt.lv

Abstract

Energy saving is an hot topic recently as the energy crisis is more and more serious. Among the energy consumer of the world, building is often ignored by many people. Nowadays many researchers noticed that research on the energy saving of building is meaningful, especially the research on the energy saving of air-conditioning. As the energy consuming of air-conditioning is very significant. Traditional control method of air-conditioning is based on PID. In the paper, fussy control is introduced and applied in the air-conditioning control, the result shows that response speed and accuracy of fussy controller are significantly better than PID controller.

Keywords: Energy consuming, Public building, Temperature adjustmen, Air conditioner

1 Introduction

Energy is vital important in the society, as nowadays large amount of energy are consumed everyday. It makes energy become the focus of the world recently. A large number of investigations carried out on energy. One of investigation statistics the energy consumption in different areas. The statics shows that three of the largest energy consumers are industry, transport and agriculture [1]. Table 1 shows the energy index of the world in 1973 and 2012.

TABLE 1 Energy index of world

Parameter	1973	2012	Ratio/%
Population(million)	3938	6352	64.3
GDP(G$ year)	14451	35025	142.4
Per capita income	3670	5514	50.2
Primary energy(Mtoe)	6034	11059	83.3
Final energy(Mtoe)	4606	7644	66.0
Electrical energy(Mtoe)	525	1374	161.8
Per capita Primary energy(toe)	1.53	1.77	15.7
Per capita CO_2 emitions(ton)	3.98	4.18	5.0
Primary energy intensity	418	316	-24.4
Primary energy intensity	319	218	-31.5

With the growth of population and development of the economic, the rate of consumption of fossil fuels increase more and more fast. With the depletion of fossil fuels, energy crisis has become a hot topic recently. So energy saving becoming an important goal of the world now. Among the energy consumer, buildings often is ignored by many people. Actually buildings are one of the largest energy consumers in the world [3].

Due to the increasing use of unitary air-conditioners, there has been a substantial increase in electricity consumption during summer. Therefore, the use of better control techniques for steady temperature control and energy saving has become a major topic in the study of air-conditioning systems [4]. Most research on power savings regarding air conditioners are focused on large/medium-sized chillers as the subject. The control units include; cooling tower, compressor, and fan coils. The amounts of energy savings range from 6% to 13%. While there is little researches focused on unitary systems for energy saving controls. Small-sized shops, offices, laboratories and classrooms generally use 2–3 unitary systems, such as window or split type air-conditioners, for their main air-conditioning devices. Unitary systems are mass-produced by manufacturers, therefore, low production cost, steady performance, low installation fees, and low operational cost, with proper control settings, are the reason unitary systems are now widely used [5]. Unitary systems mainly use the ON/OFF method as temperature control, which causes unstable room temperatures. The changes in room temperatures, from various unitary systems working at the same time, create large surges in energy consumption; therefore, energy crisis related to the buildings is defined with regard to occupant thermal comfort, energy savings and temperature control.

Daily maximum load chillers and pumps of air conditioning systems are selected according to the size of the maximum design load of pipes currently. But the air conditioning system running at full capacity in a relatively short time actually, and most of the time it is at part load. Due to the strong nonlinear characteristics air conditioning systems such as time-varying, large inertia, large lag, strong interference, etc. The classic means of control or PID control can not meet the control demand of air conditioning systems. Considering that the fuzzy control technology is fit for all types of non-linear, strong coupling, uncertainty, variable time-varying complex system. And it has been widely used in various control areas, and achieved good control effect. An air conditioning temperature control system based on fuzzy control is studied in the paper, the result shows that the result is very good.

2 Classic AIR-Conditioner Temperature Adjustment Methods

Air conditioning is a complex system more than just cool down the rooms of building. Actually it includes dehumidifying, cleaning (filtering), and circulating the air. A complete air conditioning system perform all of the functions above, not oily cool down the room. As the air conditioning system is so big, there are a lot of dynamical variables and nonlinear variables [6]. So it very hard to find a simple and fit mathematical model to describe air conditioning, which make design a good control system for air conditioning becomes a challenging work.

Two main classical controllers used in air conditioning control are two-position control (on/off) and PID (proportional-integral-derivative) control.

On/off control is one of the oldest techniques that is practiced in buildings for the purpose of energy saving and occupant thermal comfort. It's diagram is shown in Figure 1. As is shown in Figure 1. On/off control is a simple, fast and inexpensive feedback controller that accepts only binary inputs which is also known as bang-bang control and hysteretic control [7]. This control technique is still being using in domestic and commercial buildings effectively, as the well known thermostat, humidity and pressure switch. It is based on cutting of the power supply. The method is very simple and the cost of controller is very cheap. While its performance is so bad that it can't meet the demand of air conditioning now.

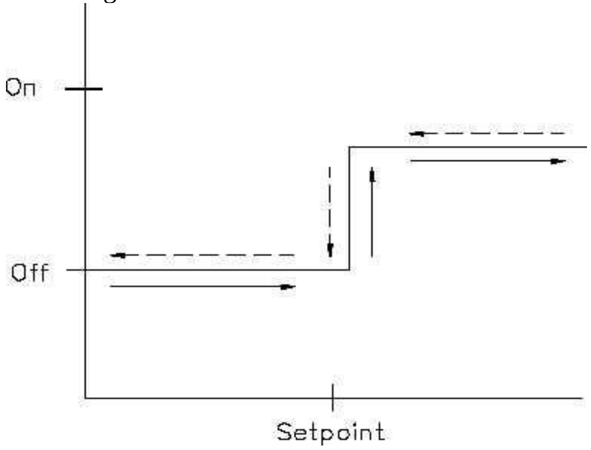


FIGURE 1 Diagram of two-position control

The most important and popular controller in industrial process is Proportional Integral Derivative (PID) as it is easy to understand and to be used as a controller. So Proportional integral derivative control (PID control) is another control method that use in the air conditioning control. Its diagram is shown in Figure 2.

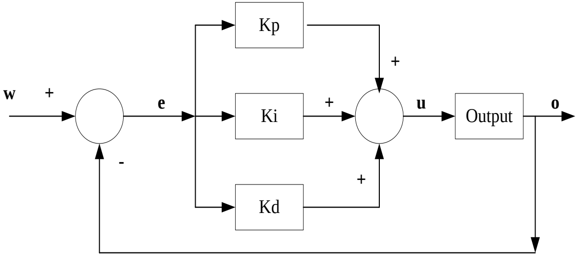


FIGURE 2 Diagram of PID control

But there are major problems that occur when using the PID controller which cause disturbance and environmental condition on the structural of the system. However when it compared to other controller, the PID are better and simple structure. To the controlled object in air condition system, the traditional PID control can be applied, but it has some disadvantages such as inconvenient tuning parameters, faint anti-interference and large overshoot. The traditional PID control method has the characteristics of simple construction, good stability, and mature theories. But the PID method excessively depends on the model parameters, and the robustness is poor. From a mathematical viewpoint, the PID control works to push the error e to zero, where

$$\mathrm{e} = o - w, \tag{1}$$

where w is the target of the PID control system and o is the output of the PID control system.

The change of output can be expressed as:

$$\Delta o = K_p e(t) + K_i \int e(t)dt + K_d \frac{\partial e(t)}{\partial t}, \tag{2}$$

where K_p , K_i and K_d are the scale factors for the proportional, integral and differential terms respectively.

There are three separate control techniques used in the PID control algorithm:

1. proportional term relates to the present offset;
2. integral term depends on the accumulation of past errors;
3. derivative term predicts the future offsets based on the current rate of change of the process.

A control signal is delivered based on a weighted sum of these three actions. The distinct effect of these three terms causes the most important stimulus for the survival of the PID control mechanism, and it also committed to the evolution of modern control approaches. It could be beneficial for certain applications to apply only one or two actions out of the three by setting the other parameters to zero. P control and PI control are two mostly used control algorithms. Thermal process dynamics in a building is usually a slow responding process. Therefore, proportional control can be engaged in building temperature control with a good stability and a reasonable small offset. Also, it is good in building humidity control. Derivative term also contributes to combat the sudden load changes encountered in the system. Still, small amounts of measurement and process noise can cause large variations in the output due to the derivative term present in the PID control.

Even though there are a number of advantages in using PID control such as simplicity of implementation, it may not be the most suitable controller for building control due to several reasons. It requires three parameters to be trained for each building zone after the installation. This is quite a time consuming task and re-tuning after the commissioning may be inconvenient. They are unable to handle random disturbances, and therefore large deviations from the set point can occur. In buildings, thermal interaction between the zones leads to multi-variable behavior. However, standard PID controller assumes a single input single-output (SISO) system during the analysis which may cause unacceptable deviations. Since these controllers operate at low energy deficiencies they may not be suitable in the long run. Smart temperature control technique for energy.

3 Consuming control

3.1 ALGORITHM OF FUSSY CONTROL

With the development of fussy mathematic, a new control algorithm - Fussy Logic Control is developed. Fussy Logic Control is based on the fussy logic [9]. Fussy Logic Control is a mathematical method that ake on continuous values between 0 and 1. The fussy logic simulate fuzziness of human information processing.

Fuzzy Logic Control is one of the intelligent control systems that are a successful solution to many control problems. The fuzzy models can represent the highly nonlinear processes and can smoothly integrate a prior knowledge with information obtained from process data. Many control solution need the mathematical model of the system to be controlled, but the Fussy Logic Control only need the measurement of input and output signals of the system to be controlled.

This controller consists of fuzzy membership function, fuzzy rules and defuzzification. Fuzzy membership rules are used to set the input and output range in several level such as low, medium and high. The fuzzy rules are used to relate and combine the input and output of Fussy Logic Control. Commonly, the relation of input and output are using “OR” and “AND” logic. Defuzzification is used to convert the rules output to appropriate value which is to be used by plant. This controller is widely used in air conditioner [10].

Fuzzy Logic Controller has three successive blocks through which the control signal is generated in Figure 3.The first block fuzzification the input, this fuzzification input is sent through an inference block where decisions are made by firing certain rules. The Fussy Logic Control system is based on the theory of fuzzy sets and fuzzy logic. Previously a large number of fuzzy inference systems and defuzzification techniques were reported.The output of the inference engine is a set of fuzzification knowledge which is converted to a crisp control signal through a technique of defuzzification. This crisp output is applied to the plant to be controlled.

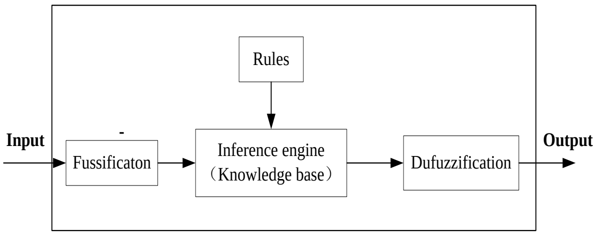


FIGURE 3 Structure of Fussy Block

The Fussy Logic Control can overcome some shortcomings of traditional PID. The fuzzy controller is a language controller. The algorithm of Fussy Logic Control can be obtained from experience and optimized from the operation, which has advantages such as powerful anti-interference, faster response and strong robust. Fussy control is appied in the control of air conditioning in the paper, so as to improve the temperature control of the air conditioning.

3.2 DIAGRAM OF THE FUSSY CONTROL SYSTEM

Air conditioning Fussy Logic Control system is shown in Figure 4. Temperature is measured by the indoor temperature sensor. Then A / D converted the measured temperature value into the digital value T. The set temperature S is compared with T. The digital temperature deviation and temperature change rate δT. And e as a controlled amount of input look-up table obtained after the fuzzy control output U_1 and U_2 is the system of cooling air volume increment increment. Changing the cooling capacity is adjusted by the electric proportional valve regulating water flow rate is achieved, then the fan flow by adjusting the fan speed regulator circuit.

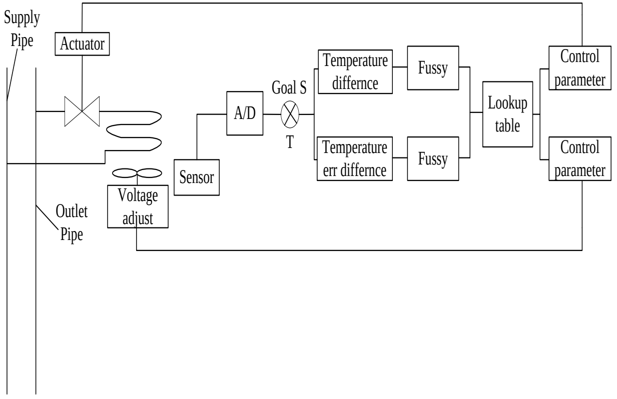


FIGURE 4 Diagram of the fussy control system of air conditioning system

3.3 FUSSY OF THE INPUT AND OUTPUT OF THE CONTROL SYSTEM

Basic domain of temperature change ΔT is [-2,2] ºC in the paper. And the corresponding linguistic variables E of ΔT is divided into 8 files: Negative Big change(NB), Negative Middle change(NM), Negative Little change(NL), Negative Zero(NZ), Positive Zero(PZ), Positive Small change(PS), Positive Middle change(PL) and Positive Big change(PB)[8]. And divide ΔT into 12 levels: -5,-4,-3,-2,-1, -0,+0,1,2,3,4,5. The quantization factor K_t of ΔT is:

$$K_t = 2 \tag{1}$$

Temperature change ΔT is transfer into a number ranges from -2 to 2 by appropriate variation.

Basic domain of temperature change rate δT is [-0.25,0.25] ºC the paper. And the corresponding linguistic variables δT is divided into 7 files: NB, NM, NL, Z, PS, PM, and PZ. And divide δT into 12 levels: -5,-4,-3,-2,-1,0,1,2,3,4,5. The quantization factor K_t of ΔT is:

$$K_T = 5/0.25 = 20 \tag{2}$$

Temperature change rate δT is transfer into a number ranges from -0.25 to 0.25 with the same variation above.

Output of the control system is defined as U_1. Basic domain of U_1 is [U_{1min},U_{1max}]. U_1 is divided into 7 files: NB, NM, NL, Z, PS, PM, and PZ. And divide U_1 into 7 levels: -3,-2,-1,0,1,2,3. The quantization factors of U_1 and U_2 are shown below:

$$\mathrm{K}_{u1} = \frac{3}{U_{1\max}} \tag{3}$$

Also, U_1 is transfer into numbers ranges from -3 to 3.

3.4 DESIGN AND SIMULATION OF FUZZY SYNTHESIS ALGORITHM

After summarizing the above-mentioned principles, the fussy control structure of air-conditioning can be concluded as below:

$IF \quad E$

$THEN \quad U_{1k} \quad and \quad U_{2l}$

in which $k = 1, 2, 3; \; l = 1, 2, 3.$

And the fussy rule of the air-condition control system is:

$$\begin{cases} R_{U_k} = R_{U_{1k}} \cup R_{U_{2k}} \cup \ldots \cup R_{U_{52k}} \\ R_{U_l} = R_{U_{1l}} \cup R_{U_{2l}} \cup \ldots \cup R_{U_{52l}} \end{cases} \tag{5}$$

After the iterative calculation of equation 5, R_{n1} and R_{n2} can be obtained. Depending on the fuzzy subset affiliations of temperature ΔT and the temperature change rate deviation δT, the corresponding U₁ and U₂ is calculated in accordance with the rules of control fuzzy decision. But it is a blur amount can not directly control the controlled object. A reasonable approach need to taken the amount of blur into a precise amount. In order to play the best decisions effect the fuzzy inference result, the principle of maximum membership degree of actual control in the paper, as long as the sampled values to calculate the temperature deviation ΔT and temperature change rate δT.

After conclude, there are 56 fussy rules in the fussy control system. The rules are shown in Table 2.

TABLE 2 Rule table of fussy control

δT	NB	NM	NL	Z	PS	PM	PB
ΔT	*Output of the fussy control system*						
NB	PB	PS	Z	NM	NM	NB	NB
NM	PB	PM	PS	NS	NM	NB	NB
NL	PB	PM	PM	NS	NS	NB	NB
NZ	PB	PB	PB	Z	NS	NM	NB
PZ	PB	PB	PB	Z	NS	NM	NB
PS	PB	PB	PB	PM	NM	NS	NB
PM	PB	PB	PB	PB	NM	Z	NB
PB	PB	PB	PB	PB	NS	PS	NB

According to the purpose of the corresponding affiliation, precise control of the amount charged is realized. At the end of the fussy control system, the control for freezing water to establish the simulation model shown in Figure 5. In which, Gain, Gain1 and Gain2 are constant gains. Derivative is the differential link, MinMax is seeking the best value. Fuzzy Logic Control is fuzzy logic controller block. Matlab Function is Matlab functions. Switch and Switch1 is to convert the block. Transport Delay is the input signal at a given time to do the delay. The selected scale factors of the fuzzy controller are $K_E = 3.2$, $K_F = 0.8$, $K_{U_{1k}} = 0.85$.

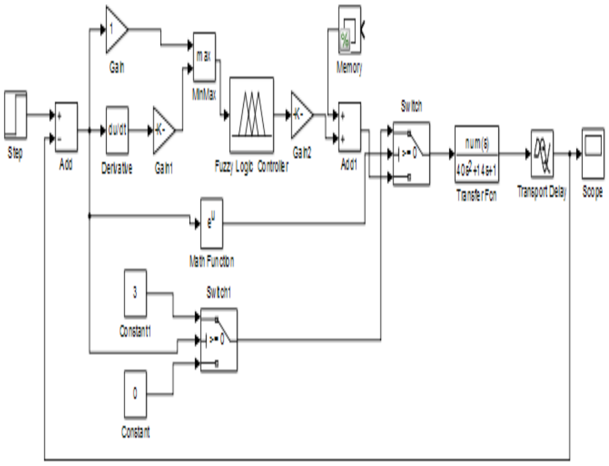


FIGURE 5 Simulation model of fussy control system for air conditioning

Assumed that indoor temperature variation is 1.4ºC, the simulation curve of fussy logic controller is shown in Figure 6.

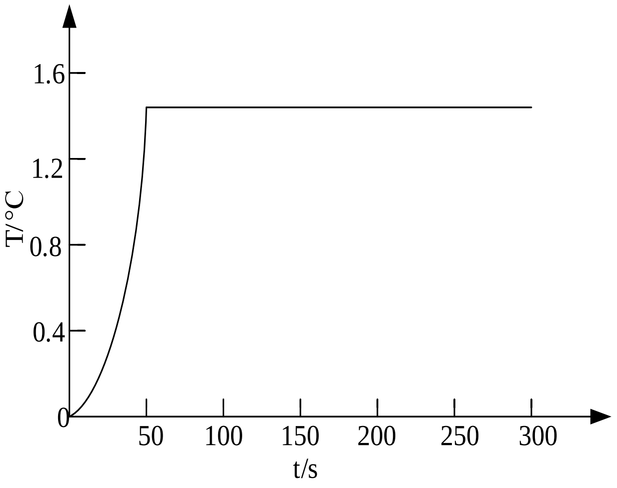


FIGURE 6 Temperature adjustment performance of fussy controller

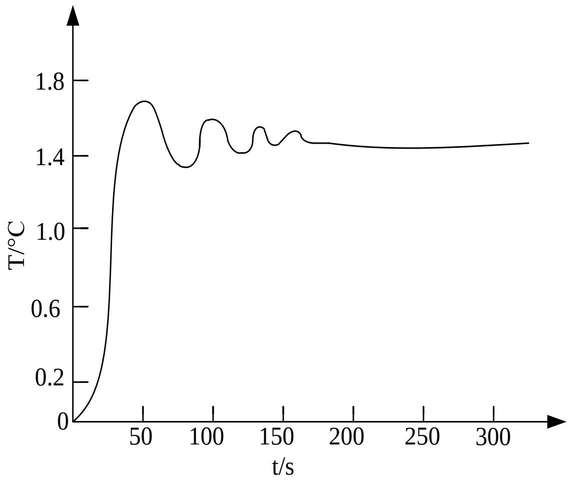


FIGURE 7 Temperature adjustment performance of PID controller

From the simulation curves, it can be seen when the temperature change of the stable room is less than 0.1, the response time of the fuzzy controller is 60 s. Its response speed and accuracy are significantly better than PID controller. Thus the temperature adjustment performance of air-conditioning with fuzzy controller better than that with PID controller. Good temperature regulation will be able to significantly reduce the energy consumption of air conditioning. So the fuzzy control system developed in the paper is significant for building energy saving.

4 Conclusion

An air-conditioning controller based on fussy control for air-conditioning is developed in the paper. In order to evaluation the performance of the fussy controller, a simulation model is set up with the help of Matlab. After simulation, the temperature adjustment performance of fussy controller and PID controller are obtained. The simulation results shows that response speed and accuracy of fussy controller are significantly better than PID controller. And the fuzzy control system developed in the paper is significant for building energy saving.

References

[1] Patil R M, Keeli A, Shi D, et al Zone Based Control of Building Air Conditioning Systems Through Temperature Monitoring and Control *ASME 2014 Power Conference American Society of Mechanical Engineers*

[2] Agarwal S, Rengarajan S, Sing T F Nudges of School Children and Electricity Conservation: Evidence from the “Project Carbon Zero” Campaign in Singapore *Available at SSRN* 2014

[3] Fong K F, Lee C K 2015 Investigation of separate or integrated provision of solar cooling and heating for use in typical low-rise residential building in subtropical Hong Kong *Renewable Energy* **75** 847-55

[4] Wu Z, Xia X, Wang B 2014 Improving Building Energy Efficiency by Multiobjective Neighborhood Field Optimization *Energy and Buildings*

[5] Chajaei S, Habib F, Hossini A M Building Design in Hot and Dry Climate with a Climatic Approach to Reduce Energy Consumption by Using Software *HEED*

[6] Bedwell B, Leygue C, Goulden M, et al 2014 Apportioning energy consumption in the workplace: a review of issues in using metering data to motivate staff to save energy *Technology Analysis & Strategic Management* (ahead-of-print) 1-16

[7] Zhao Y, Hu G, Zhou Y, et al 2014 *HVAC Control System For Household Central Air Conditioning* U.S. Patent 20,140,324,230 2014-10-30

[8] Zhang F 2014 Design of Control Scheme for Variable Air Volume Air-Conditioning System *Applied Mechanics and Materials* **686** 113-20

[9] Bai J B, Li Y, Wang M 2014 Fuzzy Adaptive PID Control of Indoor Temperature in VAV System *Applied Mechanics and Materials* **638** 2092-96

[10] Costa H R N, La Neve A 2014 Study on application of a neuro-fuzzy models in air conditioning systems *Soft Computing* 1-9

[11] Mu A, Santos M, López V *Design of intelligent control for hvac system using fuzzy logic*

[12] Hua X, Duan P, Lv H, et al 2014 Design of fuzzy controller for air-conditioning systems based-on semi-tensor product *Control and Decision Conference (2014 CCDC) The 26th Chinese IEEE* 3507-12

Authors

Shi Li, 1983, Shaanxi, China

Current position, grades: The lecturer
University studies: Yulin University
Scientific interest: Engineering management and Building energy
Publications: four

Using cubature Kalman filter to estimate the vehicle state

Xiaoshuai Xin*, Jinxi Chen

School of Automation Engineering, University of Electronic Science and Technology of China. No.2006, Xiyuan Ave, Chengdu, China

Corresponding author's e-mail: xinxiaoshuai@gmail.com

Received 1 March 2014, www.cmnt.lv

Abstract

The vehicle state is of significant to examine and control vehicle performance. But some vehicle states such as vehicle velocity and side slip angle which are vital to active safety application of vehicle can not be measured directly and must be estimated instead. In this paper, a Cubature Kalman Filter (CKF) based algorithm for estimation vehicle velocity, yaw rate and side slip angle using steering wheel angle, longitudinal acceleration and lateral sensors is proposed. The estimator is designed based on a three-degree-of-freedom (3DOF) vehicle model. Effectiveness of the estimation is examined by comparing the outputs of the estimator with the responses of the vehicle model in CarSim under double lane change and slalom conditions.

Keywords: cubature Kalman filter, vehicle state, 3DOF, CarSim

1 Introduction

A variety of active vehicle safety applications are being developed in modern cars to reduce driver burden and road accidents. Traction control system (TCS) and electronic stability program (ESP) are two popular active safety applications in vehicles. TCS concerned with controlling longitudinal motion of the vehicle and ESP concerned with controlling lateral motion of the vehicle. Traction control system works by controlling slip ratio of the four vehicle wheels. Although vehicle speed is required to calculate the slip ratio of the wheel in TCS, the absolute vehicle speed can not be accurately measured by wheel speed because of wheel slip. ESP works by controlling yaw rate and side slip angle of the vehicle. Side slip angle can not be measured directly. Due to these factors, vehicle speed and side slip angle are not directly measured on production cars and must be estimated instead.

Although there are other nonlinear observer [1-5] based study about vehicle state estimation, the main research activities in the field concentrate on the application of Kalman filter theory, which is the most powerful tool for multi-sensor data fusion problems [6]. In [7-9], Kalman filter is used to estimate yaw rate, lateral acceleration and tire slip angle with linear vehicle model. Since Kalman filter is based on linear stochastic differential equations, it can only be used in the linear system estimation. As a nonlinear filter, extended Kalman filter (EKF) extend the use of Kalman filtering through a linearisation procedure. Ray proposes an extended Kalman filter (EKF) based method for estimating vehicle speed, braking forces, wheel slip and side-slip angle [10]. A nonlinear extended adaptive.

Kalman filter is proposed for the estimation of vehicle handling dynamic states in [11]. In [12-13], dual EKF is used for vehicle state and parameter estimation. The EKF works well in many application, but may suffer from large estimate errors when system have strong nonlinearities, and also suffer from the computation burden of the Jacobians [14]. Unscented Kalman filter (UKF) is used to vehicle state estimation because it overcomes these hurdles [15]. The UKF reduces computational costs compared to EKF and needn't linearize the system and measurement equations as required by the EKF.

Recently, a cubature Kalman filter is proposed by Arasaratnam and Haykin, which improves the performance over UKF [16]. Since nonlinear filtering can be reducing to a problem of how to compute integral, cubature Kalman Filter introduce a third-degree spherical-radial cubature rule to achieve the cubature points which are used to approximate the multi-dimensional integral [14]. CKF has been proposed and used in many application, such as positioning [17-18] and attitude estimation [19]. For nonlinear system with additive Gaussian noise, cubature Kalman filer (CKF) can achieve more accurately than the UKF with similar computational complexity [20].

In this paper, we propose a CKF based estimator with a 3DOF vehicle is to estimate vehicle velocity, yaw rate and side slip angle. The inputs of the estimator are steering wheel angle, longitudinal acceleration and lateral acceleration with additive noise. Effectiveness of the estimation is examined by co-simulation between the software CarSim and Matlab-Simulink under double lane change and slalom conditions.

The rest of paper is structured as follow: The 3DOF vehicle model are described in Section 2. CKF based estimator is presented in Section 3. Our experiments and results are introduced in Section 4. Finally the main conclusion and future works are summarized in Section 5.

2 Vehicle model

The proposed method is based on a nonlinear 3DOF vehicle model, which is shown in Figure 1.

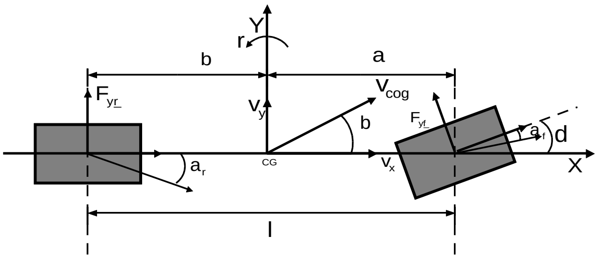


FIGURE 1 Nonlinear 3DOF vehicle model

Segel present a vehicle model with 3DOF in order to describe lateral movements including both roll motion and yaw motion [21]. By reducing the roll motion, a two-degrees of freedom linear bicycle model is obtained [22]. The linear two degrees of freedom related to vehicle body are yaw rate (r) and side slip (β). The motion of yaw rate is described as:

$$\dot{r} = \frac{a^2C_f + b^2C_r}{I_z v_x} r + \frac{aC_f - bC_r}{I_z}\beta - \frac{aC_f}{I_z}\delta, \tag{1}$$

where a is the distance from the front axel to the centre of gravity (CG), b is the distance from the rear axel to CG, C_f is the effective cornering stiffness of the front axel, C_r is the effective cornering stiffness of the rear axel, I_z is the vehicle moment of inertia about Z axis and δ is the steering wheel angle.

The motion of side slip is described as:

$$\dot{\beta} = (\frac{aC_f - bC_r}{mv_x^2} - 1)r + \frac{C_f + C_r}{mv_x}\beta - \frac{C_f}{mv_x}\delta, \tag{2}$$

where m is vehicle mass and v_x is the longitudinal velocity. In order to estimate the longitudinal velocity of the vehicle, longitudinal motion is required.

The longitudinal motion is described as:

$$\dot{v}_x = r\beta v_x + a_x \cdot \tag{3}$$

Equations (1), (2) and (3) form the nonlinear three degrees of freedom of vehicle model. In this paper, $[\gamma \quad \beta \quad v_x \quad]^T$ is the state vector of the proposed estimator, and a_y is the measurement. The measurement equation is written as:

$$a_y = (\frac{aC_f - bC_r}{mv_x} - 1)r + \frac{C_f + C_r}{m}\beta - \frac{C_f}{m}\delta \cdot \tag{4}$$

3 CKF state estimation

3.1 CUBATURE KALMAN FILTER

Kalman filter is a special case of the Bayesian filter, which assuming that the dynamic system is linear and both the dynamic noise and measurement noise are statistically independent processes [23]. Considering a nonlinear discrete-time system of the form

$$x(k) = f(x(k-1), u(k)) + w(k-1)), \tag{5}$$

$$y(k) = h(x(k-1), u(k-1), v(k)) \cdot \tag{6}$$

where $x(k)$ is a N-dimensional state vector, the output $y(k)$ is a M-dimensional vector, $u(k)$is the known control input, $w(k-1)$and $v(k)$ are independent process and measurement Gaussian noise sequences with zero means and covariance Q and R respectively. The heart of the Bayesian filter is to compute multi-dimensional weighted integral of the form

$$I(f) = \int_{R^n} f(X)\omega(X)dX. \tag{7}$$

Since it's difficult to obtain the solution of the above integral, the challenge is to compute the integral numerically by finding a set of cubature point ω_i and ξ_i that approximates the integral $I(f)$ by a weight sum of function evaluations

$$I(f) \approx \sum_{i=1}^{m} \omega_i f(\xi_i) \,. \tag{8}$$

Cubature Kalman Filter introduce a third-degree spherical-radial cubature rule to achieve the cubature point as:

$$\xi_i = \sqrt{n}\,[1]_i, \tag{9}$$

$$\omega_i = \frac{1}{2n} \qquad i = 1, 2, ..., 2n. \tag{10}$$

The entire algorithm is presented as follows:
1. Time update
Evaluate the cubature points

$$S(k-1) = chol\{P(k-1)\}, \tag{11}$$

$$\hat{X}_i(k-1) = S(k-1)\xi_i + X(k-1) \cdot \tag{12}$$

where $P(k-1)$ is associated covariance matrix, $chol\{\}$ denotes a Cholesky decomposition of a matrix.

Evaluate the propagated cubature points

$$X_i^*(k-1) = f(\hat{X}_i(k-1), U(k)) \cdot \tag{13}$$

Estimate the predicated state

$$\hat{X}(k) = \sum_{i=1}^{2n} \omega_i X_i^*(k-1) \cdot \tag{14}$$

Estimate the predicated error covariance

$$\hat{P}(k) = \sum_{i=1}^{2n} \omega_i X_i^*(k-1) X_i^*(k-1)^T - \hat{X}(k)\hat{X}(k)^T + Q \,. \tag{15}$$

2 Measurement update
Evaluate the cubature points

$$\hat{S}(k) = chol\{\hat{P}(k)\}, \tag{16}$$

$$\hat{X}_i(k) = \hat{S}(k)\xi_i + \hat{X}(k). \tag{17}$$

Evaluate the propagated cubature points

$$Y_i^*(k) = h(\hat{X}_i(k), U(k)) \cdot \tag{18}$$

Estimate the predicated measurement

$$\hat{Y}(k) = \sum_{i=1}^{m} \omega_i Y_i^*(k) \cdot \tag{19}$$

Estimate the innovation covariance matrix

$$P_{yy}(k) = \sum_{i=1}^{m} \omega_i Y_i^*(k) Y_i^*(k)^T - y(k)y(k)^T + R \cdot \tag{20}$$

Estimate the cross-covariance matrix

$$P_{xy}(k) = \sum_{i=1}^{m} \omega_i \hat{X}_i(k) Y_i^*(k)^T - \hat{X}(k)\hat{Y}(k)^T \cdot \tag{21}$$

Estimate the Kalman gain

$$K(k) = P_{xy}(k)P_{yy}^{-1}(k)\cdot \tag{22}$$

Estimate the updated state

$$X(k) = \hat{X}(k) + K(k)(Y(k) - \hat{Y}(k))\cdot \tag{23}$$

Estimate the corresponding error covariance

$$P(k) = \hat{P}(k) + K(k)P_{yy}(k)K(k)^T\cdot \tag{24}$$

The proposed sate estimation method is based on CKF, the block diagram of the method is shown in Figure 2. As can be seen in Figure 2, the state estimator is designed to estimate the vehicle state by using steering wheel angle, lateral and longitudinal acceleration signals.

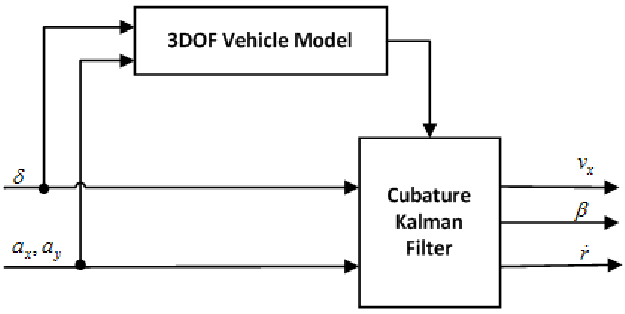


FIGURE 2 Estimation block diagram

For the state estimator, the state vector is written as

$$x(k) = \begin{bmatrix} \gamma & \beta & v_x \end{bmatrix}^T. \tag{25}$$

The measurement is written as

$$y(k) = a_y. \tag{26}$$

The known input is written as

$$u = \begin{bmatrix} \delta & a_x \end{bmatrix}^T. \tag{27}$$

The state vector equation of the proposed estimator can be written as:

$$\begin{cases} r(k) = (\frac{a^2k_f+b^2k_r}{I_zv_x}r(k-1) + \frac{ak_1-bk_2}{I_z}\beta(k-1) \\ \qquad -\frac{ak_1}{I_z}\delta(k-1))\Delta t + r(k-1) \\ \beta(k) = (\frac{ak_f-bk_r}{mv_x^2}r(k-1) + \frac{k_f+k_r}{mv_x}\beta(k-1) \\ \qquad -\frac{k_f}{mv_x}\delta(k-1))\Delta t + \beta(k-1) \\ v_x(k) = (r(k-1)\beta(k-1)v_x(k-1) + a_x(k))\Delta t \\ \qquad +v_x(k-1) \end{cases}, \tag{28}$$

where $\Delta t = t_{k+1} - t_k$ is the sampling interval.

The measurement matrix is described as:

$$H = \begin{bmatrix} (ak_f - bk_r)/(mv_x) \\ (k_f + k_r)/m \\ -(ak_f - bk_r)/(mv_x^2) \end{bmatrix}. \tag{29}$$

4 Experiments

Two simulation cases under double lane change and slalom conditions are conducted based on Matlab/Simulink and CarSim. CarSim is a multi-DOF nonlinear simulation software for vehicle dynamics control and integration, and detailed mathematical models for simulating automotive vehicle dynamics have been in use for decades [24]. Since CarSim can work with Simulink, we build estimation model in Simulink and test it with the full nonlinear CarSim vehicle model. The Simulink Model for the proposed estimation is shown as Figure 3. The known parameters of the vehicle model are listed in Table 1.

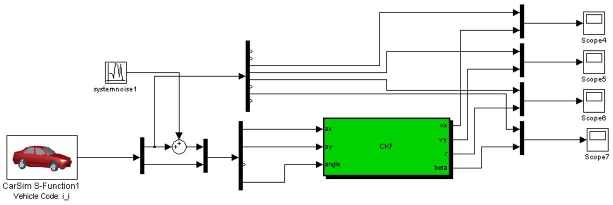


FIGURE 3 The Simulink model

TABLE 1 Specification of the vehicle model

Parameter	Symbol	Unit	Value
Vehicle mass	m	kg	1650
Vehicle moment of inertia about Z axis	I_z	kgm^2	3234
Distance from front axel to CG	a	m	1.4
Distance from rear axel to CG	b	m	1.65
Effective cornering stiffness of the front axel	C_f	N/rad	- 97000
Effective cornering stiffness of the rear axel	C_r	N/rad	- 120000

The process noise covariance of CKF is $Q = I_{3\times3}$, and measurement noise covariance is $R = [10000]$. The sampling interval is $\Delta t = 0.001s$.

4.1 DOUBLE LANE CHANGE TEST

The initialization of the State vector of the double lane change simulation case is $x(0) = [0, 0, 80]^T$. Simulation results are shown in Figure 4, Figure 5, Figure 6, Figure 7, Figure 8 and Figure 9. For the double lane change test, Figure 4, Figure 5 and Figure 6 are respectively the vehicle sensor signal of steering wheel angle, longitudinal acceleration and lateral acceleration.

As can can be seen from Figure 4, Figure 5 and Figure 6, all simulated sensor signals for CKF contain white noise which simulates the sensor noise in the real world. Figure 7, Figure 8 and Figure 9 are respectively the estimation of longitudinal velocity, side slip angle and yaw rate. As can be seen from Figure 7, Figure 8 and Figure 9, the estimated value of longitudinal velocity, side slip angle and yaw rate capture the trends in the data from CarSim. The additive noise of the sensor signal is filter by the CKF well.

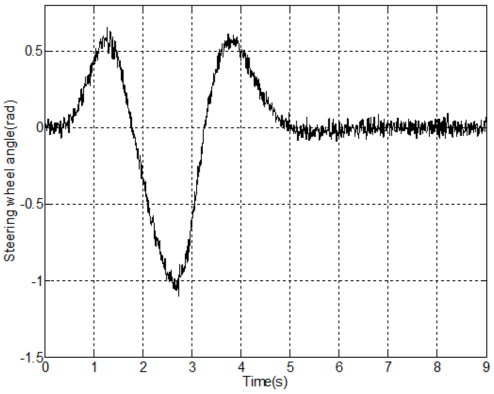


FIGURE 4 Steering angle with noise

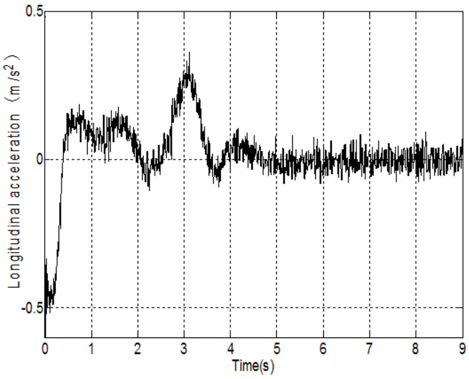


FIGURE 5 Longitudinal acceleration with noise

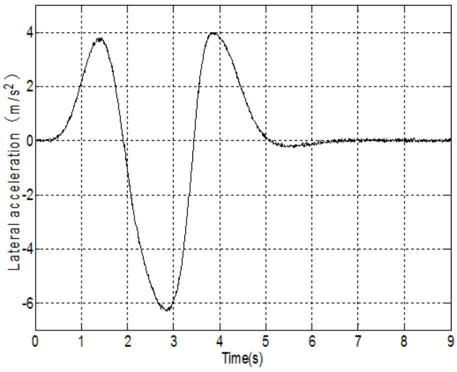


FIGURE 6 Lateral acceleration with noise

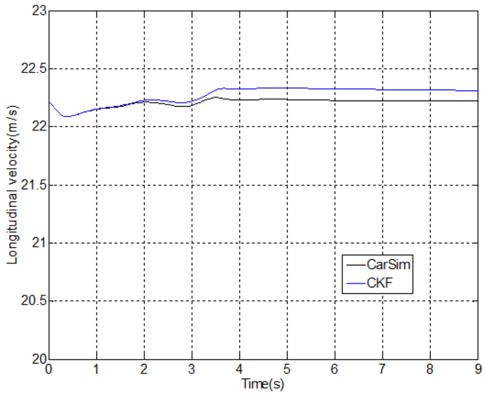


FIGURE 7 Estimation of longitudinal velocity

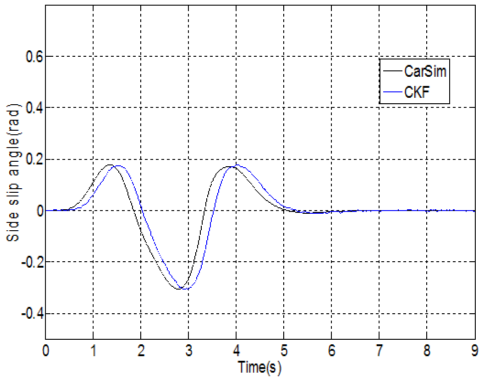


FIGURE 8 Estimation of side slip angle

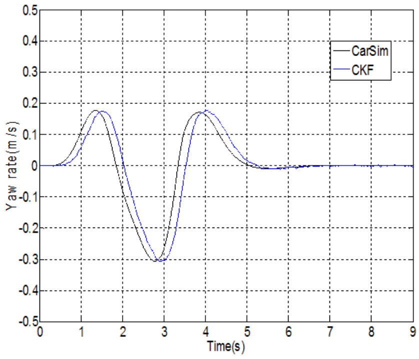


FIGURE 9 Estimation of yaw rate

4.2 SLALOM TEST

The initialisation of the State vector of the slalom simulation case is $x(0) = [0, 0, 50]^T$. Simulation results are shown in Figure 10, Figure 11, Figure 12, Figure 13, Figure 14 and Figure 15.

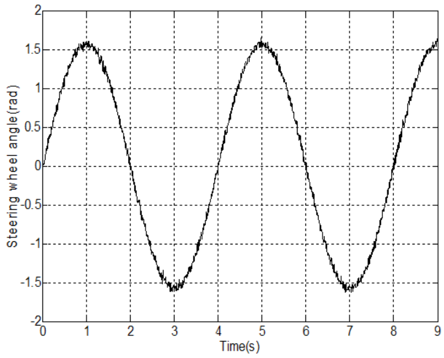


FIGURE 10 Steering angle with noise

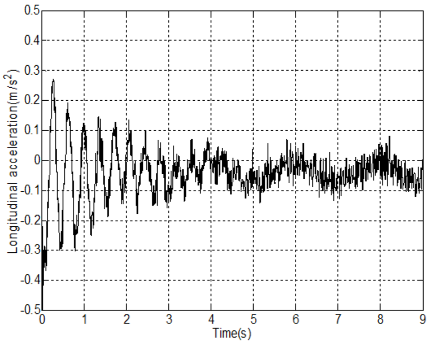


FIGURE 11 Longitudinal acceleration with noise

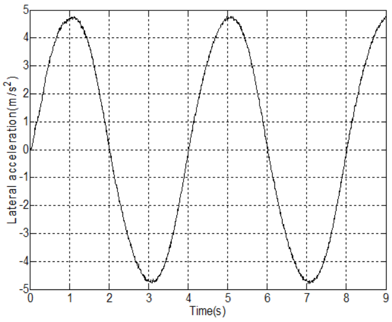


FIGURE 12 Lateral acceleration with noise

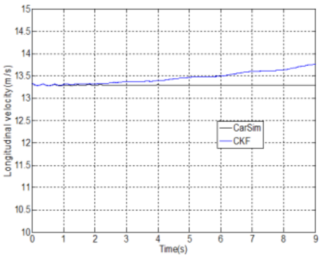


FIGURE 13 Estimation of longitudinal velocity

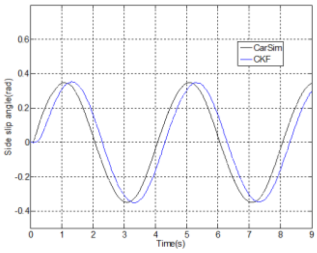


FIGURE 14 Estimation of side slip angle

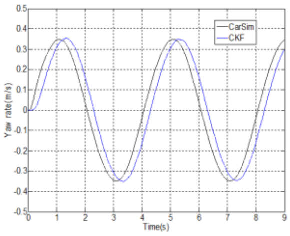


FIGURE 15 Estimation of yaw rate

For the slalom test, Figure 10, Figure 11 and Figure 12 are respectively the vehicle sensor signal of steering wheel angle, longitudinal acceleration and lateral acceleration. Figure 13, Figure 14 and Figure 15 are respectively the estimation of longitudinal velocity, side slip angle and yaw rate. As can can be seen from Figure 13, Figure 14 and Figure 15, the estimated value of longitudinal velocity, side slip angle and yaw rate capture the trends in the data from CarSim. The additive noise of the sensor signal is filter by the CKF well.

5 Conclusions

In this paper, some works are proposed to estimate vehicle speed, side slip angle and yaw rate of the vehicle. Firstly, a nonlinear 3DOF vehicle model is presented. Secondly, the estimator based on CKF is designed. Finally, the estimation is examined by comparing the outputs of the estimator with the responses of the vehicle model in CarSim under double lane change and slalom conditions. Experimental results of the simulation show the effectiveness of the proposed method.

Acknowledgements

This research has been supported by the National Nature Science Foundation of China under Grant No. 61201010.

References

[1] Hac A, Simpson M D 2000 Estimation of vehicle side slip angle and yaw rate *SAE Technical Paper*

[2] Imsland L, Johansen T A, Fossen T I, H Aa Vard Fj Ae R Grip, Kalkkuhl J C, Suissa A 2006 Vehicle velocity estimation using nonlinear observers *Automatica* **42**(12) 2091-103

[3] Shraim H, Ananou B, Fridman L, Noura H, Ouladsine M 2006 Sliding mode observers for the estimation of vehicle parameters *Forces and states of the center of gravity* 1635-40

[4] Baffet G, Charara A, Lechner D 2009 Estimation of vehicle sideslip, tire force and wheel cornering stiffness *Control Engineering Practice* **17**(11) 1255-64

[5] Jaballah B, M'Sirdi N K, Naamane A 2014 Partial vehicle state estimation using Hight Order Sliding *Mode Observers* 1-7

[6] Antonov S, Fehn A, Kugi A 2011 Unscented Kalman filter for vehicle state estimation *Vehicle System Dynamics* **49**(9) 1497-520

[7] Venhovens P J, Naab K 1999 Vehicle dynamics estimation using Kalman filters *Vehicle System Dynamics* **32**(2-3) 171-84

[8] Ryu J, Moshchuk N K, Shih-Ken Chen 2007 Vehicle State Estimation for Roll Control System *2007 American Control Conference* 1618-23

[9] King Tin Leung, Whidborne J F, Purdy D, Barber P 2011 Road vehicle state estimation using low-cost GPS/INS *Mechanical Systems and Signal Processing* **25**(6) 1988-2004

[10] Ray L R 1992 Nonlinear estimation of vehicle state and tire forces 526-30

[11] Best M C, Gordon T J, Dixon P J 2000 An extended adaptive Kalman filter for real-time state estimation of vehicle handling dynamics *Vehicle System Dynamics* **34**(1) 57-75

[12] Wenzel T A, Burnham K J, Blundell M V, Williams R A 2006 Dual extended Kalman filter for vehicle state and parameter estimation *Vehicle System Dynamics* **44**(2) 153-71

[13] Changfu Zong, Dan Hu, Hongyu Zheng 2013 Dual extended Kalman filter for combined estimation of vehicle state and road friction *Chinese Journal of Mechanical Engineering* **26**(2) 313-24

[14] Dai Hong-de, Dai Shao-wu, Cong Yuan-cai, Wu Guang-bin 2012 Performance comparison of EKF/UKF/CKF for the tracking of ballistic target *TELKOMNIKA Indonesian Journal of Electrical Engineering* **10**(7) 1692-99

[15] Tianjun Zhu and Hongyan Zheng 2008 *Vehicle State Estimation Based on Unscented Kalman State Estimation* **1**(42-46)

[16] Arasaratnam I, Haykin S 2009 Cubature Kalman Filters *IEEE Transactions on Automatic Control* **54**(6) 1254-69

[17] Pesonen H, Piche R 2010 Cubature-based Kalman filters for positioning 45-9

[18] Fernandez-Prades C, Vil A Valls J 2010 *Bayesian nonlinear filtering using quadrature and cubature rules applied to sensor data fusion for positioning* 1-5

[19] Chao Li, Quan-Bo Ge 2011 *SCKF for MAV attitude estimation* **3**(1313-18)

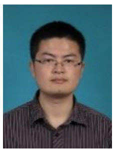
[20] Huimin Chen 2012 *Adaptive cubature Kalman filter for nonlinear state and parameter estimation* 1413-20

[21] Segel L 1956 Theoretical prediction and experimental substantiation of the response of the automobile to steering control *Proceedings of the Institution of Mechanical Engineers: Automobile Division* **10**(1) 310-30

[22] Stephant J, Charara A, Meizel D 2004 Virtual sensor: application to vehicle sideslip angle and transversal forces *Industrial Electronics, IEEE Transactions on* **51**(2) 278-89

[23] Kalman R E 1960 A new approach to linear filtering and prediction problems *Journal of Fluids Engineering* **82**(1) 35-45

[24] CarSim User Manual 2009 Mechanical Simulation Corp *Ann Arbor, MI*

Authors	
	Xiaoshuai Xin, 1982, Henan, China **Current position, grades:** The lecturer, PhD, student of School of Automation Engineering, University of Electronic Science and Technology of China **University studies:** University of Electronic Science and Technology of China, in the School of Automation Engineering (2006-2011) **Scientific interest:** Robust control and optimal control, advanced control of vehicle **Publications:** 11 papers **Experience:** He has completed 6 scientific research projects
	**Jinxi Chen, 1988, Fujian, China** **Current position, grades:** Engineer of FAW-VALKSWAGEN, Master of Science **University studies:** University of Electronic Science and Technology of China, in the School of Automation Engineering (2001-2005) **Scientific interest:** Signal processing, Vehicle control **Publications:** 3 papers **Experience:** He has completed 2 scientific research projects

The application of R/S analysis for the earthquake prediction in Sichuan, China

Xiaolu Li[1], Wenfeng Zheng[1]*, Dan Wang[1], Lirong Yin[2], Zhengtong Yin[3]

[1] *School of Automation, University of Electronic Science and Technology of China, Chengdu, 610054, China*

[2] *Geographical & Sustainability Sciences Department, University of Iowa, Iowa City, IA, 52242, USA*

[3] *School of Resources and Environment, Guizhou University, Guiyang, 550025, China*

**Corresponding author's e-mail: wenfeng.zheng.cn@gmail.com*

Received 1 March 2014, www.cmnt.lv

Abstract

Fractal is one of the powerful analysis for the study of complex natural phenomena. This paper employed fractal analysis in seismology based on the Statistical fractal concept and gave a simple overview to fractal characteristics of seismic activity in the spatio-temporal distribution. Analyzed by the R/S scale invariance of seismic time sequence and time interval sequence, this paper explored the self-shot fractal characteristics in the seismic activity.

Keywords: statistical fractal, earthquake, spatio-temporal distribution, R/S analysis method

1 Introduction

The research on the Seismic activity means the research on the overall time, space and the strength distribution feature of groups of Seismic activity. Using the catalog of earthquake to describe the statistical characteristics of regional seismic activity is one of the primary ways of today's research on seismic activity [1]. However, as a kind of instability phenomenon of the Earth's lithosphere which is nonlinear dynamical system, the mechanism of earthquake is very complicated, the strength of earthquake had a significant fluctuation over time and showed active and quiet alternating changes [2]. Because of its complexity, so we get a lot and Different aspects of the parameters those are using to describe the Earthquake activity, time change of seismic activity, the cluster of earthquake on time and space and Earthquake spatial concentration. Such as Kagan which uses the coefficient of variation CV to describe the statistical characteristics of the intervals of adjacent earthquake and describe the uniform or non-uniform of the process of earthquake quantitatively and so on [3-4].

Most of these studies are based on statistics, describe the feature such as the overall situation and fluctuations of the earthquake in a region with statistical features, such as average value, mean square error and so on [5]. The question now is whether we can make estimates to the situations may occur in the future of variables by the observation in the past or in a short time. Or on the contrary, speculate the situations may occur in a short time by the statistical characteristics of the long-time variable [6]. And this is significant in the analysis of seismic activity and seismic hazard.

R/S is a method of time series analysis which is derived by the self-affine fractal. Essentially, the Hurst exponent is the fractal dimension of standard deviation or basis structure[7-8]. Hurst analyzed many complex phenomena in nature, most of their time series conform to the self-affine fractal, and H>0.5, which means that complex phenomena in nature not only have randomness, but also a certain regularity [9]. They are long-range correlations in regularity of time. In other words, what happened later is not random, there is a certain relationship with what happened before, and reflects a certain regularity [10]. The value of H has an important physical meaning. The greater it declined from 0.5, the more regularity of the sequence, conversely the less [11]. Changing the length of time, and studying its regularity based on the relatively short time observation, and extrapolate the situation in the future, make a relatively conservative estimate for the future events. It can reflect the inner regularity of complex time series [12].

Based on the theory above, this paper proves that the scale invariance characteristics and the long-range correlation of seismic activity, to show that seismic activity time is not independent Poisson process, but the seismic activity later is affected by which in a previous time, so it is meaningful that we use R/S to study the regularity of the seismic time sequence [13]. Based on this, it is analyzed earthquakes in Sichuan since 1970 in a full time and systematically. At the same time, did an R/S analysis to groups of earthquake these are greater than or equal to magnitude 7 in Sichuan, China, to extract the variation characteristics of the value of H which is Hurst's dimension in the medium-strong earthquake [13-14].

This paper used R/S method to study the question that Hurst index and H values of seismic interval sequence in Sichuan change with time, to seek in the abnormal variation of the value of H before a medium-strong earthquake.

2 The principle and calculation steps of R/S method

R/S method was put forward in 1965 by Hurst. It is a way to analyze the time series. Its main principle is as follows [15].

Consider a time increment $\{\xi(\mathrm{t})\}$, here $\{\xi(\mathrm{t})\} = B(\mathrm{t}) - B(\mathrm{t}-1)$, $B(\mathrm{t})$ is the observed value of time t(t = 1,2,……). To any positive integer τ , define the mean sequence:

$$\langle \xi \rangle_t = \frac{1}{\tau}\sum_{t=1}^{\tau}\xi(t) . \tag{1}$$

In the formula $\tau = 1,2,...$, which means the lag time. $X(t)$ means the cumulative deviation:

$$X(\mathrm{t},\tau)=\sum_{t=1}^{\tau}\left(\xi(t)-\langle\xi\rangle_\tau\right),1\le t\le\tau . \tag{2}$$

So define range $R(\tau)$:

$$R(\tau)=\max_{1\le t\le\tau}[X(t,\tau)]-\min_{1\le t\le\tau}[X(t,\tau)],\tau=1,2... \tag{3}$$

Define standard deviation $S(\tau)$:

$$S(\tau)=\left\{\frac{1}{\tau}\sum_{t=1}^{\tau}(\varepsilon(t)-\langle\varepsilon\rangle_\tau)^2\right\}^{\frac{1}{2}},1\le T\le\tau . \tag{4}$$

The ratio of range and standard deviation is $R(\tau)/\mathrm{S}(\tau)$, which can be thought of as R/S . After analyzed the statistical law of R/S , Hurst found the following relationship:

$$R/S \propto (\tau/2)^H . \tag{5}$$

Taking the logarithm of formula (5), we can get:

$$Lg(R/S) \propto HLg(\tau/2) . \tag{6}$$

We can see that $Lg(R/S)$ is directly proportional to $Lg(\tau/2)$. According to the relationship, there are $n(n>2)$ values of $Lg(R/S)$ and $Lg(\tau/2)$ could be used to curve fitting. The slope of the linear which was got by curve fitting is Hurst index H.

3 Seismic data selection and processing

3.1 THE RANGE OF SEISMIC DATA

This paper used the earthquake directory which is offered by these two earthquake site CENC (The China earthquake networks center) [16] and USGS (The U.S. geological survey) [17] to proofread the earthquake catalogue of mainland China, and delete duplicates, supplement the missing item, to study the earthquake what the magnitude of earthquake is huge in Sichuan from 1900 to 2013. Using the K - K method to remove the handle of aftershocks for the data of seismic network and the USGS data.

3.2 THE INTEGRITY ANALYSIS OF THE SEISMIC DATA

The Regional seismic network in China was built in 1950s, the regional seismic network which has the small earthquake monitoring function was established in 1970s. So we have a lot of seismic data now. But in fact, the earthquake monitoring ability is different in different regions of China, After 1970 the earthquake is edited by the seismological bureau of Municipalities directly under the central government, autonomous regions and provincial, establish the earthquake catalog database. Earthquake catalog is the basal data of analysis of the seismic activity, is the precondition of the research on the rules of seismic activity, and is the indispensable data to study the dynamics of lithosphere. Because of the different level of development in various areas, there are many differences in earthquake monitoring ability, and earthquake catalogue which was written is limited, so the seismic record in the data of middle-strong earthquakes might be incomplete. So we need to consider its impacts to our study. In order to avoid the effect to the results by the lack of seismic data, we must attach great importance to the integrity of seismic data. The integrity of Seismic data refers to begin with a certain level of the earthquake, the earthquake can be observed above and be recorded completely.

MC is the important parameters, which is used to represent the smallest integrity of historical earthquake catalog. There are a lot of sources of seismic data, and those need to be screened. Because if we choose a higher minimum magnitude, that would miss a lot of useful data for the experiment. However, if we choose a low minimum magnitude, historical earthquake catalog may not be complete any more, and that would affect the result of the experiment. So in order to make full use of historical earthquake data, we need to give the distribution of time domain, the airspace to ensure the earthquake catalogue is complete and keep the useful information do not be missed. Especially for the research of different time scale and different spatial distribution.

Because with the development of social economy, the number of seismic stations are increasing and the monitor ability improved constantly. Which makes the minimum magnitude (MC) of historical earthquake catalog is in constant change with time. Usually the later the time, the smaller the value of MC is, that is to say, the more complete the seismic data is. At the same time, because of unbalanced development of various regions and the uneven of the base station facilities, which makes MC have some differences in spatial distribution, for example the difference at the edge of the seismic network and seismic network covering area is very obvious. All in all, MC not only changes with time, but also exist differences in the spatial distribution.

This paper selected the earthquake catalogues from 1900 to 2013 in Sichuan as the research object, so analyzed the minimum magnitude of the earthquake catalogue in time domain in the study of the region at the following content and gives the main earthquake activity area and the integrity in time and space distribution.

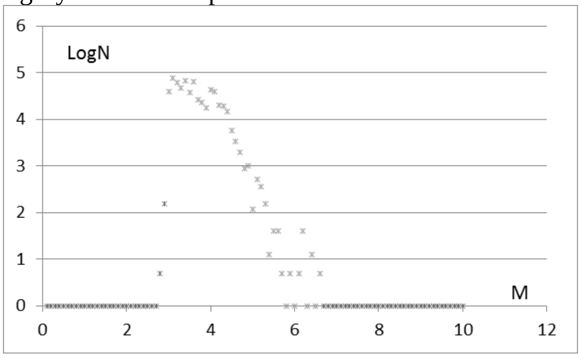


FIGURE 1 The magnitude-frequency diagram from 1900-2013

The method this paper used to analyze the earthquake catalogue integrity is the G R relationship (FMD), calculate the minimum integrity magnitude as shown in Formula 7:

$$LogN = a - bm, \tag{7}$$

where N is a number of earthquake magnitude file or the cumulative sum of the earthquakes which above a certain magnitude m, a & b are constant. In the result diagram and the magnitude – frequency diagram which was calculated by the method, the number of small earthquake magnitude is in decline, we think that this is due to the earthquake catalogue is not complete.

The process of using magnitude - frequency relation to estimate the integrity minimum magnitude of seismic data as follows: first will begin the earthquake from 0, Level 1 as the step length increasing, until 10.0, get each file number of earthquake magnitude, make earthquake - frequency chart, as shown in figure 1; And then take out different lower limit Mi to fit by the formula (7), as shown in figure 2 to figure 4, the results show that the fitting residual error R is generally decreasing with the increase of Mi, it reduced to the minimum until Mi = Mc, and then increase along with the increase of Mi. Finally fitting results statistical figure as shown in figure 5, abscissa means the minimum magnitude, ordinate means the residual error R, the results showed that from 1900 to 2013 the minimum integrity is $Ms = 2.8$. Choosing the earthquake catalogue at the lower magnitude limit $Ms = 2.8$ or above the limit from 1900 to 2013 is reasonable.

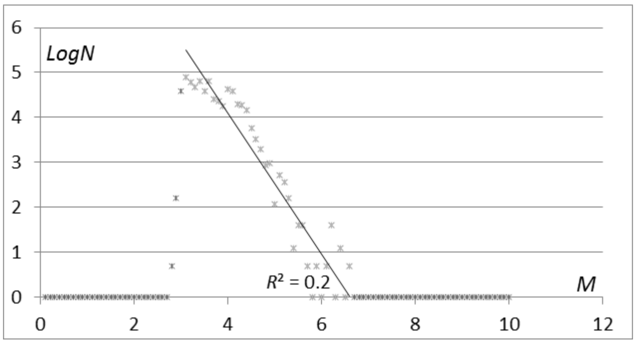


FIGURE 2 The magnitude-frequency chart from 1900-2013 (Ms=2.5)

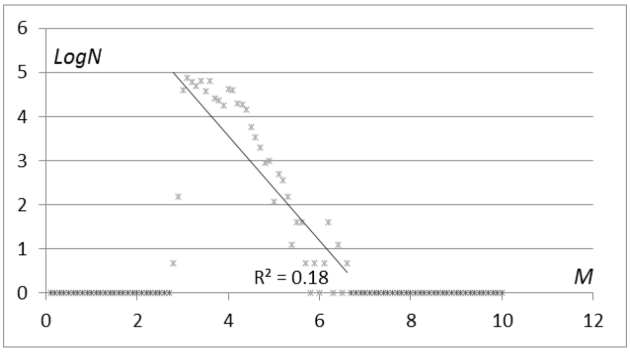


FIGURE 3 the magnitude-frequency chart from 1900-2013 (Ms=2.8)

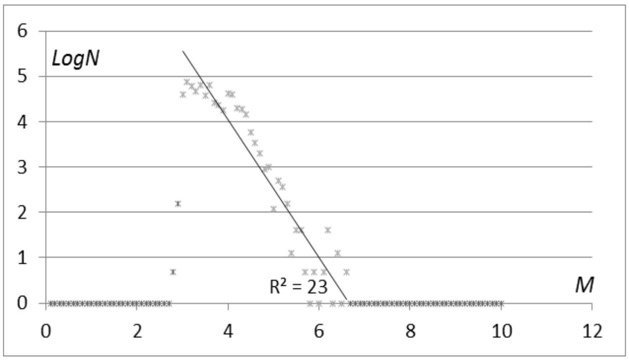


FIGURE 4 The magnitude-frequency chart from 1900-2013 (Ms=3.9)

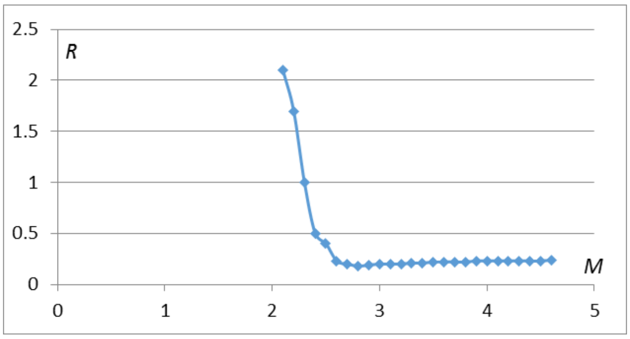


FIGURE 5 The sample of FMD method to estimate minimum integrity magnitude

4 The calculation results and analysis

In the R/S study of Earthquake sequence, such as Liu, C. used the method of non-integer order differential and integral calculus considering the integrity of earthquake catalogue and determination of the minimum magnitude [18], we research the earthquake which is more than 3 and from 1990 to 2013. We chose the frequency parameters values fitting.

The main factors influencing the value calculation precision and credibility are the size of the sample size and the reasonable selection of window. When maximum window size and the sample size is large, then get a high precision on H and the abnormal of H is highly reliable. When maximum window size and the sample size is small, then the credibility of abnormal of H's value is relatively low. So we improve accuracy by long zoom window in practical work. Considering the calculation accuracy of H, and to the considerations of precursor recognition even the substantiality of earthquake prediction, in computing, choose the window size as two years in the R/S analysis of the frequency is suitable.

Do the R/S analysis to the data of Sichuan region (latitude 26.0661-34.3203°N, longitude 97.3661-108.5329° E) from January 1990 to June 2013 a total of 20 years. Using two years as the length of time window to count years and setting a month as sliding step. Processing the data by taking 12 groups as a sliding step. The result of the calculation are shown in table 1, the time process curve of H is shown in figure 6, among them, the dotted line is the actual value of the slide. Because the value of H is affected by the man-made factors of the study area and time window, we set H's sliding step length as 5 and take its average, to reduce the influence of noise. Which used solid line to express.

TABLE 1 time series values of the earthquake in Sichuan province

Date	Time	Latitude	Longitude	Depth	MS	Time sliding H
2013/04/20	02:47.5	30.3	102.99	17	7	0.3553
2013/01/18	42:50.1	30.95	99.4	15	5.5	0.3737
2011/10/31	58:15.0	32.6	105.3	6	5.2	0.4353
2009/06/29	03:51.5	31.46	103.96	24	5.5	0.5109
2008/05/12	27:59.5	31.01	103.42	14	8.0	0.3852
2001/02/23	09:21.8	29.55	101.14	24	6	0.4049
1998/11/19	38:13.8	27.27	101.03	33	6.1	0.2963
1996/12/21	39:39.9	30.56	99.51	10	5.6	0.3054
1996/02/28	22:01.6	29.13	104.73	32	5.4	0.3807
1994/12/29	58:29.7	29.11	103.83	32	5.6	0.4148

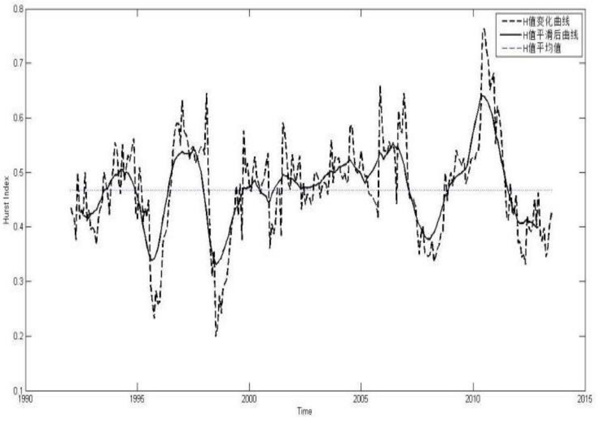


FIGURE 6 The time process curve of the H of earthquake in Sichuan area

The result shows that 90% of H are lower than the average value which is 0.4353, the H of 77.78% of big earthquake are in the process of recovery of down - low-recovery stage. H isn't on decline phase except the December 1994 and October 2011. The rest 8 groups of earthquake are all had an earthquake in 36 months. Abnormal interval sequence is concentrated in the 5 to 12 months. Which is in the process of recovery of down - low- recovery stage in the future, these are more likely happened in the middle-huge earthquake.

5 Conclusion

The study of the seismicity means that first analyze the earthquake of a certain magnitude interval in time and space distribution characteristics, then discuss its physical meaning, and then make a scientific summary for the law of earthquake. As a kind of nonlinear time series analysis method, seismic activity R/S method is applied to sliding Hurst index H value, mainly based on the earthquake catalog data for research. Through this paper we can get the following conclusions.

(1) Under the smooth background, in the steady background, took sliding R/S analysis of frequency time series of the earthquakes in the area above magnitude 3 since 1990.80% of earthquakes showed prominent or morphological characteristics is uniform precursory anomaly. Abnormal morphology is value of H which was begin with normal, but then down-low-recovery. The earthquakes occurred in the value of H on the decline and the process from low basis to rise.

(2) The unusual duration and the lowest value, the decrease and the low value to the occurrence time and total time have no relationship. Although the earthquake in the study area size is different, but the abnormal form before the earthquake almost show a same order of magnitude. The abnormal process of the 8 magnitude earthquake is clear. Such as the abnormal of H in Wenchuan 8.0 magnitude earthquake in 2008. However there are no abnormal low value in two 5.0 magnitude earthquakes in Sichuan area. These might related to tectonic characteristics, characteristic of medium and the earthquake location and other factors. Which also reflects the complexity of the occurrence of earthquakes. And these may also be related to the limitations of the method.

(3) According to the time interval value H, you can see that seismically active areas where is the reliability of H is high, the value of H stable value in the normal background relatively, so that is extremely clear and reliable.

(4) The research of time interval and frequency sequence in Sichuan. For the relatively concentrated area of earthquakes: the earthquake which is greater than 5.0, are in abnormal down-low-recovery, but the earthquake occurred in the recovery stage were more than 70%. And according to the results of the research region by the lowest earthquake occurred in the process of abnormal, H began to rise within three years. Among them the advantage of frequency sequence time rallied for a minimum of 4 to 7 months and 4 to 8 months, the time interval sequence is concentrated in two closed period: 4-12 months or 11-12 months. The quickly picked up of H from low value is a short-term precursor worth noting. If H value appears this change, there would be a risk of earthquake in the coming months.

(5) To Sichuan, the possibility of the occurrence of earthquakes is relatively large.

References

[1] Li J, Chen Y 2001 Rescaled range (R/S) analysis on seismic activity parameters *ACTA Seismologica Sinica* **14**(2) 148-55

[2] Kawamura M, Wu Yh, Kudo T, Chen Cc 2014 A statistical feature of anomalous seismic activity prior to large shallow earthquakes in Japan revealed by the pattern informatics method *Natural Hazards and Earth System Science* **14**(4) 849-59

[3] Chen C, Lee Y T, Chang Y F 2008 A relationship between Hurst exponents of slip and waiting time data of earthquakes *Physica A Statistical Mechanics and its Applications* **387**(18) 4643-8

[4] Xu Y, Burton P W. 2006 Time varying seismicity in Greece: Hurst's analysis and Monte Carlo simulation applied to a new earthquake catalogue for Greece *Tectonophysics* **423**(1-4) 125-36

[5] Wheeler R L, Mueller C S 2001 Central US earthquake catalog for hazard maps of Memphis Tennessee *Engineering Geology* **62** 19-29

[6] Yao Q L 2003 A fuzzy method for evaluating the influences of some geological factors on earthquake disaster risk *Seismology and Geology* **2** 245-59

[7] Enescu B, Ito K, Radulian M, Popescu E, Bazacliu O 2005 Multifractal and Chaotic Analysis of Vrancea (Romania) Intermediate-depth Earthquakes: Investigation of the Temporal Distribution of Events *Pure and Applied Geophysics* **162**(2) 249-71

[8] Rong Y M, Wang Q, Ding X, Huang Q H 2012 Non-uniform scaling behavior in Ultra-Low-Frequency (ULF) geomagnetic signals possibly associated with the 2011 M9 Tohoku earthquake *Chinese Journal Geophysics* **55**(11) 3709-17

[9] Silva P G, Goy J L, Zazo C, Bardaji T 2003 Fault-generated mountain fronts in southeast Spain: geomorphologic assessment of tectonic and seismic activity *Geomorphology* **50** 203-25

[10] Gao H, Zhu Y, Han M, Dou M, Li J 2009 Research on deformation characteristics of Weihe basin *Journal of Geodesy and Geodynamics* **29**(3) 60-6

[11] Han W, Jiang G 2004 Study on distribution characteristics of strong earthquakes in Sichuan-Yunnan area and their geological tectonic background *ACTA Seismologica Sinica* **26**(2) 211-22

[12] Bassingthwaighte J B, Raymond G M. 1994 Evaluating rescaled range analysis for time series *Annals of Biomedical Engineering* **22** 432-44

[13] Shao H, Du C, Liu Z, Sun Y, Xia C 2004 Multi-scale analysis of earthquake activity in Chinese mainland *ACTA Seismologica Sinica* **26**(1) 102-5

[14] Omori S, Komabayashi T, Maruyama S 2004 Dehydration and earthquakes in the subducting slab: empirical link in intermediate and deep seismic zones *Physics of the Earth and Planetary Interiors* **146**(1-2) 297-311

[15] Meng X, Zhao P 1991 Fractal method for statistical analysis geological data *Journal of China University of Geosciences* **2**(1) 111-6

[16] CENC. Available from: http://www.csndmc.ac.cn/newweb/data.htm

[17] USGS. Available from: http://earthquake.usgs.gov/earthquakes/

[18] Liu C, Zhang J, Liu Y 1995 Time-space scanning of seismic time interval fractals for the large north China region by R/S analytical method *Earthquake* **4** 372-8

Authors

Xiaolu Li, 1983, Hubei, China

Current position, grades: PhD candidate
University works: University of Electronic Science and Technology of China, in the School of Automation Engineering
Scientific interest: Geographical information system, Spatial analysis

Wenfeng Zheng, 1969, China

Current position, grades: Associate professor
University works: University of Electronic Science and Technology of China, in the School of Automation Engineering (2001-2005)
Scientific interest: Geographical information system, Spatial analysis

Effect of 3-S-isothiuronium propyl sulfonate on bottom-up filling in copper electroplating

Qiuxian Shen, Xu Wang*

College of Science, Lishui University, Lishui 323000, China

**Corresponding author's email: lsxywx@sina.com*

Received 1 March 2014, www.cmnt.lv

Abstract

The effect of 3-S-isothiuronium propyl sulfonate (UPS) upon the microholes filling by Cu electrodeposition was investigated by cross-sectional images using optical microscopy. The bottom-up filling of the electroplating bath was achieved with an addition of UPS. The electrochemical study indicated that the polarisation on the cathode was decreased with an addition of UPS. Furthermore, X-ray diffraction analyses showed the crystallography and the peak intensity ratio I(111)/I(200) of plated Cu film were decreased with addition of UPS. The results present UPS as an accelerator which is beneficial for microholes filling for high density interconnections printed circuit board.

Keywords: Damascene copper plating, accelerator, Microhole filling

1 Introduction

Depolarizers (accelerators, anti-suppressors) belong to one important class of copper plating additives that are used in the integrated circuit (IC) industry for the on-chip metallization of holes and trenches [1]. It is the non-uniform distribution of such accelerators and suppressor additives across those trenches and holes that allow their superfilling with copper. Origin of these non-uniformities in the additive surface coverage is the conjunction of purely geometric shape evolution effects upon fill with the distinct transportation and adsorption kinetics of the depolarizer and suppressor additives involved [2–5].

Most common suppressor used for the Damascene process is polyethyleneglycol (PEG) that are known to form barriers for cupric ions on the copper surface when combined with chloride [6–12]. Sulfur-containing, organic additives typically serve as depolarizers. Bis (3-sulfopropyl) disulfide (SPS) is the most widely used depolarizer for Damascene applications [3,12]. SPS shows such a mild depolarizing effect on the copper deposition, but only when chloride is present as a co-additive [8, 13]. Such intrinsic acceleration therefore needs to be considered as a synergistic effect of the SPS and the chloride.

It was reported that other good accelerator was 3-N, N-dimethylaminodithiocarbamoyl-1-propanesulfonic acid (DPS) [14], 3,3-thiobis-1-propanesulfonate (TBPS) [15], 3-mercapto-1-propanesulfonic acid (MPS) [16-18], as potential substitutes of the SPS.

UPS had been used as brightener for Cu electrodeposition [19] and as stabilizer for electroless nickel deposition [20], but the copper filling of UPS as accelerator of a three-additive system has not been reported.

In this article, we address the copper filling of 50 μm microholes using UPS as accelerator. And the effect of UPS on the crystallography was studied.

2 Experimental

PCB fragments with many microholes formed by CO2 laser ablation were used as plating samples. The dimensions of the PCB fragment were 45 mm × 60 mm. The diameters of the microholes were 50 μm. The depth of the microholes was 50 μm. Before metallization, the microholes were conducted through a so-called desmear process in order to remove the smear that was formed by laser ablation at the microhole bottom. The desmear process could thoroughly clean the via bottom to make sure of its conductivity. Following the treatment of desmear process, electroless copper plating was used for sidewall metallization of the microhole. Following that, an electroplated copper layer with a thickness of 2–3 μm was deposited on the sidewall in order to increase the thickness of electroless copper layer for prevention of electroless copper oxidation.

The PCB fragment was plated at a current density of 1.5A dm−2 for 120 min. Two phosphorus-containing copper plates were used as anodes and placed directly in the plating bath with a working volume of 700 mL. The plating solution was constantly agitated by continuously flowing air bubbles at a flow rate of 2.5Lh−1 during the electroplating to ensure good convection.

The electroplating solution used for microhole filling experiments was composed of 220 g L^{-1} $CuSO_4{\cdot}5H_2O$, 55 g L^{-1} H_2SO_4, 12 ppm 3-S-isothiuronium propyl sulfonate (UPS), 4 ppm Janus Green B (JGB) aswell as Cl^{-} (added as NaCl), 300ppm PEG-8000. Temperature of the plating solution was controlled at 25℃. The filling performance of the plating bath was evaluated by cross-sectional views of the microholes using optical microscopy (DFC290, Leica) at a magnification of 200×.

To investigate the effect of UPS concentration on microhole filling characteristics, the cross-sectional images of the holes were observed by an optical microscope (OM). As shown in Fig. 1, the filling capability of a microhole is expressed as a filling performance. Height from the bottom of a via-hole to the deposited Cu surface and Cu film thickness at the centre of the microholes are expressed as H_1 and H_2 respectively; the filling performance calculated by a proportion of H_2 to H_1. Linear sweep voltammetry was performed to analyse the effect of UPS concentration on cathodic polarisation of the electrolyte for Cu deposition. In the electrochemical analyses, a φ3.0 mm pure Cu electrode was used as the working electrode, and a 10×10 mm^2 platinum sheet and a commercial electrode of Ag/AgCl saturated with KCl were used as the counter and reference electrodes, respectively. Linear sweep voltammetry experiments were carried out at 25℃ and at a scan rate of 10 mV s−1 in the range from 0 to −0.6 V.

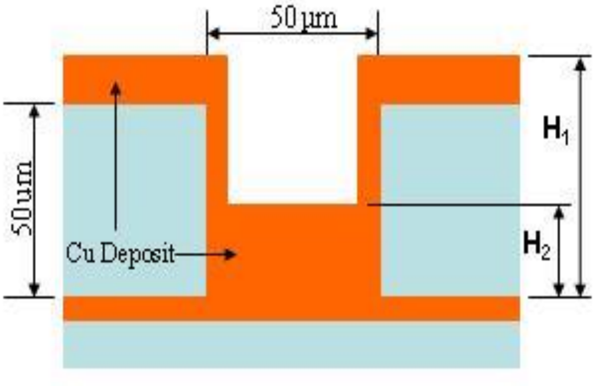


FIGURE 1 Filling Power of microhole

The crystalline structures of plated Cu films were measured by an X-ray diffractometer (Dmax3C Rigaku) using θ–2θ scan with a Cu Kα source working at 40 kV and 40 mA.

3 Results and discussion

To investigate the effect of UPS concentration on microholes filling characteristics, the cross-sectional images of the microholes were observed by an optical microscope (OM). Figure 2A shows the cross-sectional OM image of the hole before electrodeposition. After plating for 30 min (Fig. 2B) a somewhat conformal filling was obtained, and 60 min (Fig. 2C), a significant bottom-up filling of electroplated Cu in the microhole was observed. Further to that, void-free filling of Cu was completed after electroplating for 80 min (Fig. 2D), and a bottom-up filling was obtained.

It was noted that thickness of Cu on the top surface changes significantly over plating time. The deposition rate of Cu on the top surface was very small before microhole had been filled by plated Cu. After the microhole were almost filled with Cu, the thickness of Cu on the top surface started to increase noticeably. According to N. T. M. Hai [15], this was attributed to PEG-Cl^- suppressing mainly Cu deposition on the surface of the substrate, and the acceleration of UPS-Cl^- acting at the bottom of microhole before they had been filled in plating of Cu. When the microhole were almost filled with Cu, the combined effects of UPS-Cl^- and PEG-Cl− cause the deposition rate of plating of Cu on the surface of the substrate to increase.

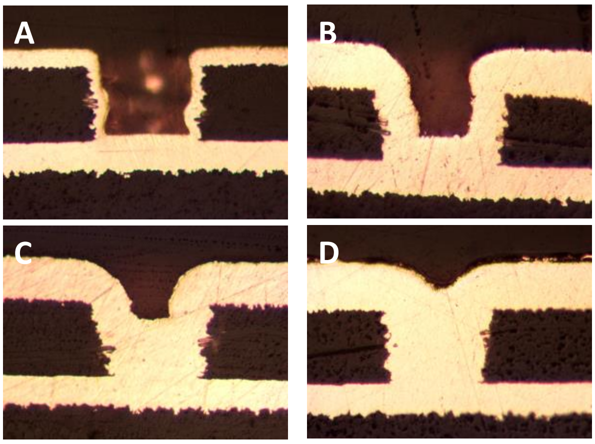


FIGURE 2 Cross-sectional OM images of holes with different plating times. Plating times: (A) 0 min, (B) 30 min, (C) 60 min, (D)80 min

Effects of UPS concentration on the cathodic polarisation of the electroplating bath were investigated by linear sweep voltammetry, and the results are shown in Figure 3.

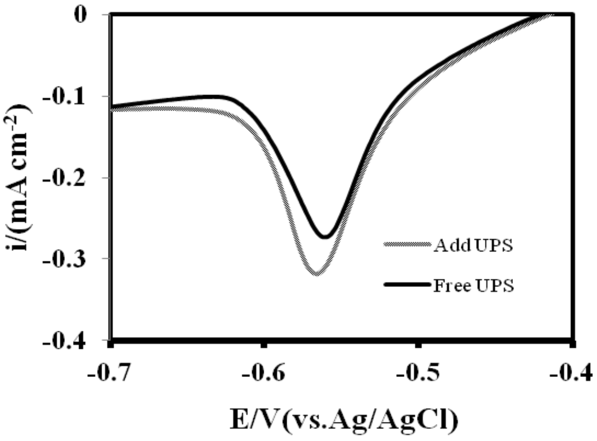


FIGURE 3 Effect of EPE-8000 on the cathodic polarization behavior of electrolyte for copper plating

From Fig. 3, it was found that copper reduction current changed with addition of UPS. The reduction peak current was about -0.269 Am cm^{-2} without addition of UPS, it shifted to -0.319 Am cm^{-2} with 12 mg L^{-1} UPS addition, but copper reduction potential did not change obviously with addition of the UPS. As mentioned above, the change of the reduction peak current of copper with an addition of UPS was significant agreement with the tendency that of copper deposition with the tendency that copper deposition with an addition of UPS, which in-depth indicated that copper reduction reaction was accelerated by addition of UPS.

The crystallography of superfilling plated Cu films, which deposited from the plating solution in the absence and presence of UPS, were characterized by XRD. When UPS was added in bath, the peak intensity ratio I(111)/I(200) was 4.6, and the full-width at half-maximum (FWHM) of (111) for 3.0 μm thick Cu film was 0.21°. For plated Cu film without additives, the peak intensity ratio I(111)/I(200) was 2.4 and FWHM of (111) for 3.2 μm thick Cu film was 0.27°, which indicated that the crystallinity and peak intensity ratio I(111)/I(200) were increased with addition of UPS. According to Scherrer formula [21], the results indicated the crystallinity of plated Cu film was reduced by addition of UPS, but peak intensity ratio I(111)/I(200) was decreased upon addition of UPS. It is well known that the copper film with a strong (111) texture can enhance electromigration resistivity performance because of the reduced degree of anisotropy in grain boundary transport. Consequently, the performance of the plated Cu films was improved by an addition of UPS.

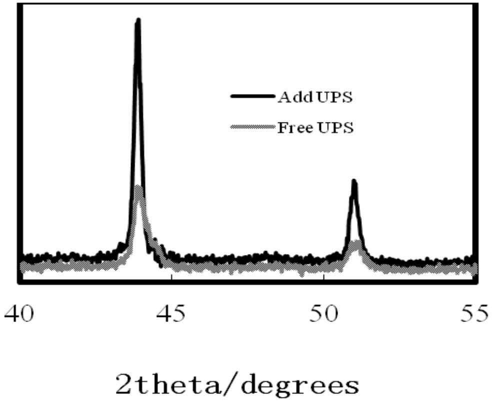


FIGURE 4 XRD patterns of plated Cu films

4 Conclusions

The effect of UPS as an additive on bottom-up filling characteristics in copper deposition was investigated. Bottom-up fillings with microholes of 50 µm and aspect ratio of one were obtained in the electroplating Cu bath with addition of UPS. Linear sweep voltammetry measurement indicated that UPS accelerated on Cu deposition, which is in agreement with the results of cross-sectional OM observation. The crystallinity and the peak intensity ratio I(111)/I(200) of plated Cu films were increased with an addition of UPS; the surface of electroplated Cu films become more smooth with an addition of UPS. From the results obtained in this study, it is concluded that UPS was highly effective for void-free filling of trenches.

Acknowledgments

This work was supported by the Commonweal Technology Project of Lishui (2012JYZB35) and Provincial Natural Science Foundation of Zhejiang (Y14E010005).

References

[1] Andricacos P C, Uzoh C, Dukovic J O, Horkans J, Deligianni H 1998 *IBM Journal of Research and Development* **42** 567
[2] Moffat T P, Bonevich J E, Huber W H, Stanishevsky A, Kelly D R, Stafford G R, Josell D 2000 *Journal of Electrochemical Society* **147** 4524
[3] Moffat T P, Wheeler D, Edelstein M D, Josell D 2005 *IBM Journal of Research and Development* **49** 19
[4] Moffat T P, Wheeler D, Kim S K, Josell D 2006 *Journal of Eletrochemical Society* **153** C127
[5] Shen Q X, Wang X, Zhu H P 2014 *International Journal of Electrochemical Society* **9** 367
[6] Kelly J J, West A C 1998 *Journal of Electrochemical Society* **145** 3472
[7] Hebert K R 2005 *Journal of Electrochemical Society* **152** C283
[8] Dow W P, Huang H-S, Yen M-Y, Chen H-H 2005 *Journal of Electrochemical Society* **152** C77
[9] Moffat T P, Wheeler D, Josell D 2004 *Journal of Electrochemical Society* **151** C262
[10] Yokoi M, Konishi S, Hayashi T 1984 *Denki Kagaku* **52** 218
[11] Healy J P, Pletcher D, Goodenough M 1992 *Journal of Electrochemical Society* **338** 155
[12] Vereecken P M, Binstead R A, Deligianni H, Andricacos P C 2005 *IBM Journal of Research and Development* **49** 3
[13] Broekmann P, Fluegel A, Emnet C, Arnold M, Roeger-Goepfert C, Wagner A, Hai N T M, Mayer D 2011 *Electrochimica Acta* **56** 4724
[14] Cho S K, Kim S-K, Kim J J 2005 *Journal of Electrochemical Society* **152** C330
[15] Hai N T M, Furrer J, Gjuroski I, Bircher M P, Cascella M, Broekmann P 2013 *Journal of Electrochemical Society* **160** D3158
[16] Lee C H, Kim A R, Koo H-C, Kim J J 2009 *Journal of Electrochemical Society* **156** D207
[17] Hai N T M, Odermatt J, Grimaudo V, Kramer K W, Fluegel A, Arnold M, Mayer D, Broekmann P 2012 *Journal of Physical Chemistry C* **116** 6913
[18] Hai N T M, Kramer K W, Fluegel A, Arnold M, Mayer D, Broekmann P 2012 *Electrochimica. Acta* **83** 367
[19] Xu X L, Webb E 2012 *High speed copper plating bath* US Patent No. 8262894
[20] Liu H P, Bi S F, Li N 2010 *Russian Journal of Electrochemistry* **46** 383
[21] Norkus E, Vaskelis A 1987 *Russian journal* of *inorganic chemistry* **32** 130

Authors

	Qiuxian Shen, 1964, Zhejiang, China **Current position, grades:** Master degree, Senior Engineer and A-class Expert on Lishui University **University studies:** Zhejiang University of Technology **Scientific interest:** metal interconnection in ultra large scale integration
	Xu Wang, China **Current position, grades:** assistant professor of ecology college, in Lishui University **University studies:** Henan Normal University **Scientific interest:** Scientific interest: metal interconnection in ultra large scale integration

Study on signal processing technology based on the reflective intensity modulated fiber optic sensor

Junjie Yang[1]*, Zhihe Fu[2], Yibiao Fan[2], Wenxiang Chen[1], Zhiping Xie[1], Wei Wu[1], Xiaoyu Shan[1]

Department of Mechanical and Electrical Engineering, Xiamen University, Xiamen, Fujian 361005, China

Department of Mechanical and Electrical Engineering, Longyan College, Longyan, Fujian 364012, China

**Corresponding author's email: junjieyang677@126.com*

Received 1 March 2014, www.cmnt.lv

Abstract

Sensor technology is one of the most representative of the emerging technology. At present, the sensor has been widely used in national defence, industry, agricultural production, environmental protection, biological science, measurement, transportation, each field of automatic control and household appliances, etc.. Optical fiber sensing technology is accompanied by the development of optical communication technology gradually formed, compared all kinds of optical fiber sensor and the traditional sensor has a series of unique advantages, such as high sensitivity, anti electromagnetic interference, corrosion resistance, electrical insulation, explosion-proof, light path with the flexible, convenient for connecting with a computer, the structure is simple, small volume, light weight, low power. In this paper, the intensity modulation type reflective optical fiber displacement sensor, studied the basic principle, in fact, is the displacement measurement in particular, on the assumption that the condition of uniform distribution, the emergent light field is analyzed in detail, the expression intensity modulation function under various conditions were obtained.

Keywords: Algorithm, Fiber optic displacement sensor, weak signal processing, band-pass filter

1 Introduction

In recent years, optical fiber sensors have been widely used in various industries [1]. In particular, with the rapid development of communication industry, the application of optical fiber is also increased. The research of optical fiber sensor has been more in-depth. Optical fiber sensor has the advantages of some other traditional electromagnetic sensor can match, especially suitable for some worse condition, therefore, it is of great significance to study the optical fiber sensor. With the rapid development of communication industry, optical fiber sensors have been widely used in all aspects of the industry [2]. Compared with the traditional electromagnetic sensor, optical fiber sensor has the very big difference in the detection principle [3-5].

Detection is the basis of industry, displacement detection is one of the most common means of detection in mechanical industry, and to realize the displacement detection using optical fiber sensor has great potential for development [6]. For example [7], optical fiber displacement sensor can be used in many kinds of occasions detection by detecting the surface morphology, surface morphology to finally realize the reconstruction of 3D topography, which compared with traditional electromagnetism sensor has great advantages, detection accuracy is guaranteed, but also in the abominable environment continue to use without failure [8].

The main purpose of this paper is to study the working principle of reflective optical fiber displacement sensor based on intensity modulation, concrete implementation scheme is proposed for detection of weak photoelectric signal, according to the detection scheme, a reasonable choice of devices, design the corresponding optical fiber displacement sensor signal processing circuit, finally completes the circuit debugging, which can accurately detect the light signal passing through and the modulation filter noise, that the measured displacement [9, 10].

At present in the industrial field in common are: mechanical displacement measuring instrument, often with a mechanical transmission mechanism (lever, gear, rack etc.) measurement of the displacement magnification and some with optical reading device corresponding; displacement of the sensor structure is changed, the displacement is converted into electricity, such as a potentiometer type sensor (displacement of the sliding contact mobile), capacitance sensor (variable distance, variable area type), eddy current sensor, Holzer sensor can realize the displacement measurement; use effect of some functional materials, such as piezoelectric sensor, metal strain plate and semiconductor strain resistance, the displacement transformation into small pressure to the piezoelectric sensor the crystal surface charge sheet or the strain resistance changes to achieve the displacement measurement; magnetoelectric: magnetic grid (linear disc) and inductosyn (linear, circular) [11].

With the development of optical detection element and the precision manufacturing process improvement and electronic components, with the development of computer automatic control technology and the upgrading of the industry, combined with the method of using photoelectric is an effective way to solve the above problems, such as grating, encoder, triangulation, spot scattering method, its measurement precision is high, the reaction speed is fast, easy to realize digital measurement, but the back-end circuit and digital processing device is complex, expensive. In recent years, high precision, high speed object displacement measurement has attracted more and more attention, especially the measurement of small displacement of the narrow space of the doing a lot of research and discussion. Thus, displacement measurement by optical method more and more, especially the displacement measurement method using optical fiber sensing technology has the unique advantage of the much attention. Displacement detection system is often affect the control performance of the system, although the displacement detection in different conditions detection requirements focus is different, but the basic requirement of similar, all want to fast, accurate and reliable and economical realization of displacement measurement.

2 Related theory

2.1 THE BASIC COMPONENTS OF OPTICAL FIBER SENSOR

Optical fiber sensor, as a kind of detection device, in the external environment of various physical, chemical content, biomass effect, will make some specific optical properties of light transmission in fiber to change, a process known as modulation also, corresponding to the modulator part of optical fiber sensor, detecting part through changes optical properties of light detection, can be detected by measuring, this process is also called the demodulation [12-14].

(1) Sensor light source

Fiber optic light source in the system is determined by the character of the design can achieve the expected indicators measuring system for light source of different needs, in accordance with the light source coherence can be divided into coherent light source and non coherent light source, common non coherent light source is composed of light-emitting diode (LED) and incandescent light source, coherent light source is the main variety of semiconductor laser, gas laser.

(2) The photoelectric converter

The photoelectric converter is a variety of photoelectric detection device, is converting the optical signal into electrical signal special device, optical receiving system is the front-end device, its sensitivity or bandwidth directly affects the performance of the whole optical fiber sensing system. Photoelectric detectors commonly used with PIN photodiode, charge coupled device (CCD), photomultiplier tube.

(3) Optical fiber connector and a fixed connector

Optical fiber and other optical devices interconnect, the inevitable power losses occur, such as connecting the optical fiber and optical fiber, light source and optical fiber connection, the loss is not negligible, directly influence the detection results, so it needs to consider the attenuation of the signal interconnection.

(4) Optical fiber coupler

The optical coupler is a device used for optical signal transmission and distribution. The general form of the optical coupler is light enters from one end of the coupler, and from another or several output ports, mainly all kinds of beam splitters, wave division multiplexer, isolators and circulators and so on, these devices are basically passive devices.

2.2 THE SHORTCOMINGS OF THE EXISTING RESEARCH METHODS

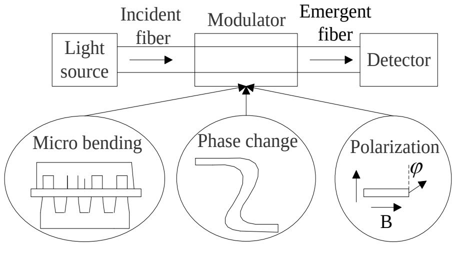
The advantages of the traditional sensor optical fiber sensor has the incomparable, but due to late start time, although the broad application prospects, but also has many insufficiencies, a lot of technology is not mature, large-scale commercial level there is still a lot of difficulty, it is difficult to quickly replace the electromagnetic sensor, the future still need further exploration.

In the development of fiber industry in China is also very quickly, from the end of twentieth Century began launched the corresponding research work, also to be included in the seven five plan, the research continues to expand the scale of. Development throughout the entire optical fiber sensor industry, a lot of research in developed countries in the area of optical fiber sensors in China lags behind the foreign countries, for example, there are many kinds of other optical fiber sensor is still in the experimental stage of the laboratory, the accuracy is not high enough, and foreign gap is not small, a large number of industrial products still exist many problems. Now the level of development of optical fiber sensor industry in China is still not formed a large scale, large scale commercial within short term is difficult to achieve.

3 The basic principle and the classification of optical fiber sensors

3.1 THE WORKING PRINCIPLE OF REFLECTIVE OPTICAL FIBER DISPLACEMENT SENSOR

Light emitted from the light source passes through the optical fiber transmission into the modulator, was measured under the action of some optical properties of signal light will change, such as light intensity, phase, frequency, wavelength and polarization. As shown in Figure 1, the light through the modulator becomes modulated signal light, through the photoelectric detector and signal processing circuit of follow-up, we can demodulate is measured, this is the basic principle of optical fiber sensor.



Fundamental principle of function (light transmission) fiber sensor

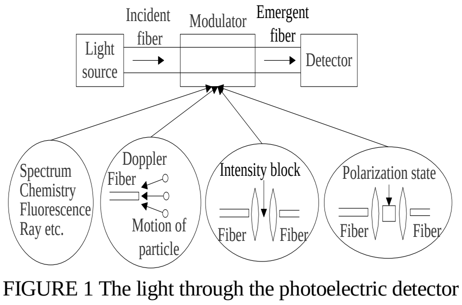


FIGURE 1 The light through the photoelectric detector

Signal transmission in optical fiber, light can use the following formula:

$$E = E_0 \cos(\omega t + \phi) . \quad (1)$$

The above formulas in E_0 wave amplitude, ω is the angular frequency, which is a phase angle, from this equation, after transformation, five parameters can be obtained by the light signal. Are the strength of E_0^2 (amplitude squared), angular frequency, wavelength $\lambda_0 = \frac{2\pi c}{\omega}$ (c is the speed of light), phase (ωt+ψ) and polarization.

3.2 ANALYSIS OF THE GEOMETRY OF REFLECTIVE OPTICAL FIBER DISPLACEMENT SENSOR

The chart can be 2 basic structure of optical fiber sensors, due to the light out of the optical fiber field range restriction, so the sensor can detect displacement range is limited, the receive optical fiber must be within the light cone reflecting light transmitting fiber in form or part in the cone of light in, can receive the reflected light, the optical fiber displacement sensor optical field analysis of geometry as shown in Figure 2.

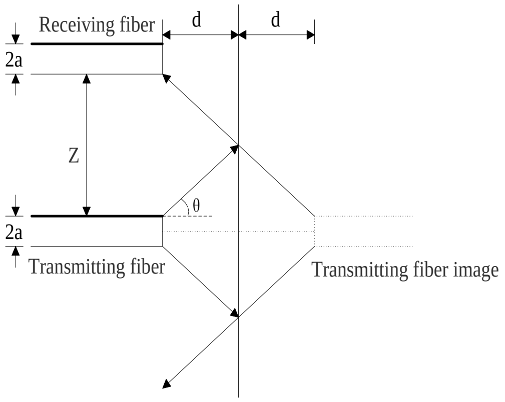


FIGURE 2 Sensor optical field analysis of geometry

When the receiving optical fiber located outside the light cone, will not be able to receive the optical signal, this model will fail, above can be

$$\tan\theta = \frac{Z}{2d} . \quad (2)$$

According to the definition of numerical aperture of NA, numerical aperture is a parameter to measure fiber light gathering ability. From improving the coupling efficiency between light source and optical fiber perspective requirements for fiber with large NA, but the greater light mode dispersion of optical fiber is also more serious, the information transmission capacity is small.

$$\theta = \arcsin NA . \quad (3)$$

Then

$$d = \frac{Z}{2\tan(\arcsin NA)} . \quad (4)$$

3.3 ANALYSIS OF THE PRINCIPLE OF REFLECTIVE INTENSITY MODULATION

Receiving optical power Pr and transmitting optical fiber Pt optical power ratio:

$$M = \frac{P_r}{P_t} . \quad (5)$$

The intensity modulation function is influenced by many factors of a function, typical of a receiving fiber core diameter is Rr, a receiving optical fiber numerical aperture NAr, Rr sends the fiber core diameter, numerical aperture of NAt transmitting optical fiber, optical fiber and optical fiber transmitting receiving axial spacing p, reflector reflectivity δ, and a receiving fiber and reflector between the distance d. Taken together, as functions of the form can be expressed as:

$$M = f(r_r, r_t, NA_r, NA_t, p, \delta, d) . \quad (6)$$

Visible light intensity modulation function is more than the combined effects of the physical quantity results, these related physical quantities and structure parameters of the optical fiber, selection and combination of different physical quantities, will have different characteristic curves, it is suitable for different measurement range, basis of the next section will in certain assumptions on the research of intensity modulation function.

4 Result and discussion

4.1 THE IMPROVED PRINCIPLE OF REFLECTIVE OPTICAL FIBER DISPLACEMENT SENSOR

In this paper, the requirements of displacement sensor in the detection sensitivity are high and the volume is particularly demanding. Sensor or front belong because volume is too large to meet the requirements, or because of insufficient accuracy can not be adopted. Therefore, this paper uses the reflection type sensors improved, as shown in Figure 3, increase a way of receiving fiber as the reference light path, in order to eliminate same-sex interference effects such as the external environment.

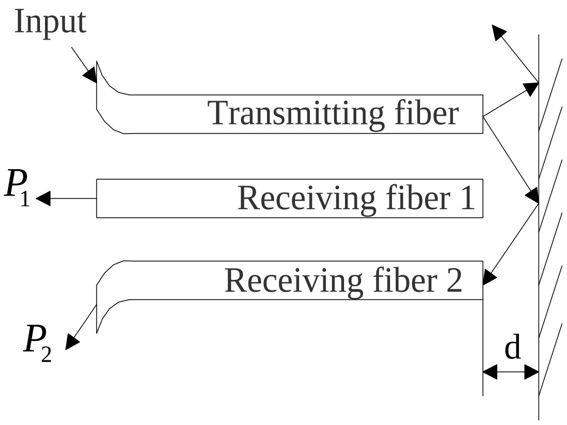


FIGURE 3 The improved reflection type sensors

4.2 STRUCTURE DESIGN OF INTENSITY MODULATION REFLECTION TYPE OPTICAL FIBER

Using different combinations of optical fiber, optical structure is different, will have different response curves, the measurement and the sensitivity and dynamic range is suitable for different conditions. Over the past thirty years, in order to improve the performance of a variety of sensors, many scholars put forward different fiber structure, the main common single fiber type, fiber of type, three optical fiber type, double beam, stochastic, coaxial and semi-circular, coaxial type I, coaxial II type, double ring beam type, double beam type, coaxial random type, icircle with models and so on. Intensity modulation characteristics such as shown in Figure 4 curves corresponding to several typical structure.

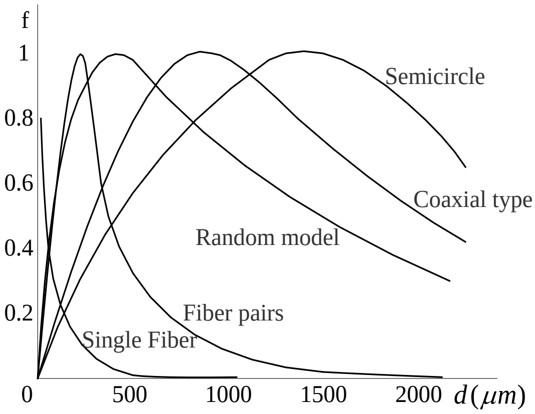


FIGURE 4 Intensity modulation characteristics in curves corresponding to several typical structure.

Different fiber arrangement has been light intensity modulation characteristic curve of different. From the figure of the curve, we can see that different combinations, characteristic curves is dead zone length, different linear range, before the slope after slope length is different. Single fiber and fiber on the front slope of high sensitivity but linear range is small, random type, front slope sensitivity small coaxial and semicircle, but the linear range is bigger, need to choose the fiber arrangement according to the requirements of measurement.

4.3 WEAK SIGNAL PROCESSING

In order to describe the degree of signal quality, introduces two concepts, signal-to-noise ratio and improves signal-to-noise ratio.

(1) Signal to noise ratio

Signal to noise ratio refers to the useful signal components in the S effective value and the noise component of the effective value of the ratio of N.

$$SNR = \frac{Signal}{Noise} = \frac{S}{N} . \tag{7}$$

Measurement uncertainty or error can be expressed as:

$$r = \frac{1}{SNR} . \tag{8}$$

The higher the SNR, the measurement error or uncertainty of measurement is smaller.

If the signal-to-noise ratio is improved, can improve the signal-to-noise ratio to measure of the number of relations, to improve the signal-to-noise ratio is defined as

$$SNIR = \frac{Output\ SNR}{Input\ SNR} = \frac{S_\theta / N_\theta}{S_i / N_i} . \tag{9}$$

The output noise bandwidth of the system is to improve the signal-to-noise ratio and better, so, in ensuring the pass-band useful signal under the condition of the bandwidth of the system, the more narrow the better.

Mainly based on the implementation of correlation detection technology is the randomness of the measured signal periodicity and noise, the measured signal generally contains the periodic component, and the noise is generally do not contain periodic components, through autocorrelation or cross-correlation operation, can effectively achieve the purpose of filtering noise. Measures of correlation is the use of correlation function, autocorrelation function and cross correlation function, detection method respectively correspond to the autocorrelation detection and correlation detection.

The self correlation function is usually used to measure associated with a random process, the autocorrelation function:

$$R(\tau) = R_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)x(t-\tau)dt . \tag{10}$$

Here, put forward a kind of weak signal detection scheme, displacement sensor to realize the difference compensation, is adopted in the structure of a received two, its signal processing block diagram was shown in Figure 5.

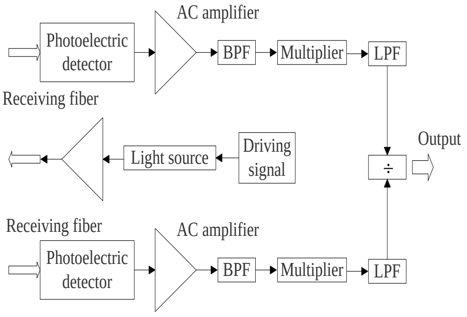


FIGURE 5 The signal processing block diagram

The principle of weak signal detection theory to achieve the above diagram: drive circuit issued a certain frequencies of light, into the transmitting fiber communication to the measured object surface, the modulated signal is received by the receiving fiber, the signal is the signal being measured requires demodulation, the measured signal through band-pass filter, filter out of band noise, after correlation detection to detect weak signal, correlation detection is implemented through a multiplier and low-pass filter, the final result of the DC control signal, the two signal phase, the results obtained shows that the size of the measured displacement.

4.4 ANALYSIS OF EXPERIMENTAL RESULTS

The micrometer displacement measurement platform in a rotating to change the distance optical fiber probe to the reflecting surface, so that it can be measured by displacement of a different point, get the voltage displacement curve of the sensor, as shown in Figure 6. In the figure, the solid line shows the displacement voltage curve, in order to verify the accuracy of the micro controller ADUC812 work, we also measured in into the micro controller before the two analog signal after signal processing, which are marked with an asterisk curve is measuring coaxial reflection optical fiber beam signal obtained with the plus sign, the curve is measurement of random type of reflective optical fiber bundle is obtained.

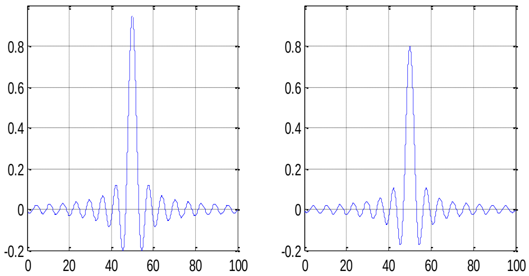


FIGURE 6 The voltage displacement curve of the sensor

According to the design of the signal processing circuit, design of analog signal processing circuit, and the results of the signal processing circuit. The optical fiber sensor probe is fixed on the 2D jogging platform, because of the particularity of reflective light displacement sensor, here need to ensure that the probe perpendicular to the test surface, otherwise it will affect the output results. The test pieces is made from aluminium block surface smooth, read the output voltage value using the oscilloscope, by adjusting the two-dimensional platform, can change the probe and the distance between the test pieces.

5 Conclusions

In the information age, along with the human understanding, expand the scope of activities to the infinite, extreme and new field in space and time, Admiral, development to develop sensing various strong, high, weak, sensor and edge effect has become a starting point for a variety of emerging areas, key projects and breakthrough. These emerging areas and key engineering process the break will bring immeasurable progress to human science and technology, produce the enormous economic benefits. Therefore, the sensor has become a pioneer in development of modern science and technology. This paper is about the design and development of reflective intensity modulated fiber optic displacement sensor for measuring micro displacement.

In order to weak signal detection of optical fiber displacement sensor, puts forward a method of weak signal detection, and completed the design of the circuit part, focuses on the design of filter circuit, the amplitude detection circuit, the circuit simulation software has also been a good result, finally completed the debugging of the signal processing circuit board, and with the completion of the displacement measurement with optical fiber probe, the corresponding results obtained, the results may make analysis.

The measured curve and graph into two signals after dividing the match, indicating that the micro design of peripheral circuit and software of the controller is correct. Take the measurement range is 0.2mm1.2mm are linearized, the range can be obtained for the voltage displacement curve of 1mm, as shown in Figure 7. The sensitivity is 5mV/1um, linearity is less than 1%. Some experimental results show that this sensor can meet the basic technical indexes design.

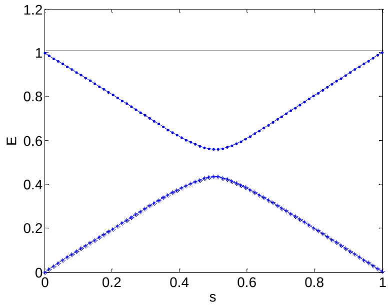


FIGURE 7 The range obtained for the voltage displacement curve of 1mm

References

[1] Yuan Chang Su, Chih Chung Chang, Jia Lin Wang 2008 Construction of an automated gas chromatography/mass spectrometry system for the analysis of ambient volatile organic compounds with on-line internal standard calibration *Journal of Chromatography A* **1201**(2) 134-40

[2] Zhen Li, Endalkachew S-D, Ashraf A H, Sorial G A 2011 Transport and deposition of CeO2 nanoparticles in water-saturated porous media *Water Research* **45**(15) 4409-18

[3] Goldovsky N, Goldovsky V 2003 Correlational gas analyzer *Measurement* **33**(3) 273-9

[4] Leonhardt J W 2003 A new ppb-gas analyzer by means of GC–ion mobility spectrometry (GC-IMS) *Journal of Radioanalytical and Nuclear Chemistry* **257**(1) 133-9

[5] Jue Wang, Ren Wang, Duo Qian Miao 2010 Data enriching based on rough set theory *Journal of Environmental Sciences* **29**(3) 63-9

[6] Changjun Hou, Jiang Jie Li, Danqun Huo, Xiao gang Luo, Jia le Dong, Mei Yang, Xiao Jie Shi 2008 Design of an embedded gas detector based on spectral analysis *Chinese Journal of Scientific Instrument* **29**(4) 471-5

[7] Fort A, Mugnaini M, Rocchi S, Vignoli V, Comini E, Ponzoni A 2010 Metal-oxide nanowire sensors for CO detection: characterization and modeling *Sensors and Actuators B: Chemical* **148**(1) 283-91

[8] Breysse M, Claudel B, Faure L, Guenin M, Williams R J 1976 Chemiluminescence during the catalysis of carbon monoxide oxidation on athoria surface *Catal* **45**(2) 137-44

[9] Xian An Cao, Ying Tao, Liling Li, Yong Hui Liu, Yan Peng, Jin Wen Li 2011 An ethyl acetate sensor utilizing cataluminescence on Y2O3 nanoparticles *Luminescence* **26**(1) 5-9

[10] Li Tang, Ya Ming Li, Kai Lai Xu, Xian Deng Hou Yi Lv 2008 Sensitive and selective acetone sensor based on its cataluminescence from nano-La2O3 surface *Sensors and Actuators B: Chemical* **132**(1) 243-9

[11] Xiao An Cao, Zhen Yu Zhang, Xin Rong Zhang 2004 Sensitive a A novel gaseous acetaldehyde sensor utilizing cataluminescence on nanosized BaCO3 *Sensors and Actuators B: Chemical* **99**(2) 30-5

[12] Zhi Ming Rao, Lin Jie Liu, Jing Yi Xie, Yu Yun Zeng 2008 Development of a benzene vapour sensor utilizing chemiluminescence on Y2O3 *Luminescence* **23**(3) 163-8

[13] Huimin Cao, Youping Chen, Zude Zhou, Gang Zhang 2005 General models of optical-fiber-bundle displacement sensors *Microwave and Optical Technology Letters* **47**(5) 494-7

[14] Zhang Gang, Chen Youping, Cao Huimin 2006 Design of an embedded optical fiber micro-displacement measurement system in *Proceedings of SPIE* **6150**(1)

Authors' index

Chen Jinxi	12
Chen Wenxiang	26
Fan Yibiao	26
Fu Zhihe	26
Li Shi	7
Li Xiaolu	18
Shan Xiaoyu	26
Shen Qiuxian	23
Wang Dan	18
Wang Xu	23
Wu Wei	26
Xie Zhiping	26
Xin Xiaoshuai	12
Yang Junjie	26
Yin Lirong	18
Yin Zhengtong	18
Zheng Wenfeng	18

NATURE PHENOMENA AND INNOVATIVE ENGINEERING

Energy consuming control of building based on fussy temperature control

Shi Li

Computer Modelling & New Technologies 2015 ***19****(2D) 7-11*

Energy saving is an hot topic recently as the energy crisis is more and more serious. Among the energy consumer of the world, building is often ignored by many people. Nowadays many researchers noticed that research on the energy saving of building is meaningful, especially the research on the energy saving of air-conditioning. As the energy consuming of air-conditioning is very significant. Traditional control method of air-conditioning is based on PID. In the paper, fussy control is introduced and applied in the air-conditioning control, the result shows that response speed and accuracy of fussy controller are significantly better than PID controller.

Keywords: Energy consuming, Public building, Temperature adjustment, Air conditioner

Using cubature Kalman filter to estimate the vehicle state

Xiaoshuai Xin, Jinxi Chen

Computer Modelling & New Technologies 2015 ***19****(2D) 12-17*

The vehicle state is of significant to examine and control vehicle performance. But some vehicle states such as vehicle velocity and side slip angle which are vital to active safety application of vehicle can not be measured directly and must be estimated instead. In this paper, a Cubature Kalman Filter (CKF) based algorithm for estimation vehicle velocity, yaw rate and side slip angle using steering wheel angle, longitudinal acceleration and lateral sensors is proposed. The estimator is designed based on a three-degree-of-freedom (3DOF) vehicle model. Effectiveness of the estimation is examined by comparing the outputs of the estimator with the responses of the vehicle model in CarSim under double lane change and slalom conditions.

Keywords: cubature Kalman filter, vehicle state, 3DOF, CarSim

The application of R/S analysis for the earthquake prediction in Sichuan, China

Xiaolu Li, Wenfeng Zheng, Dan Wang, Lirong Yin, Zhengtong Yin

Computer Modelling & New Technologies 2015 ***19****(2D) 18-22*

Fractal is one of the powerful analysis for the study of complex natural phenomena. This paper employed fractal analysis in seismology based on the Statistical fractal concept and gave a simple overview to fractal characteristics of seismic activity in the spatio-temporal distribution. Analyzed by the R/S scale invariance of seismic time sequence and time interval sequence, this paper explored the self-shot fractal characteristics in the seismic activity.

Keywords: statistical fractal, earthquake, spatio-temporal distribution, R/S analysis method

Effect of 3-S-isothiuronium propyl sulfonate on bottom-up filling in copper electroplating

Qiuxian Shen, Xu Wang

Computer Modelling & New Technologies 2015 ***19****(2D) 23-25*

The effect of 3-S-isothiuronium propyl sulfonate (UPS) upon the microholes filling by Cu electrodeposition was investigated by cross-sectional images using optical microscopy. The bottom-up filling of the electroplating bath was achieved with an addition of UPS. The electrochemical study indicated that the polarisation on the cathode was decreased with an addition of UPS. Furthermore, X-ray diffraction analyses showed the crystallography and the peak intensity ratio I(111)/I(200) of plated Cu film were decreased with addition of UPS. The results present UPS as an accelerator, which is beneficial for microholes filling for high density interconnections printed circuit board.

Keywords: Damascene copper plating, accelerator, Microhole filling

Study on signal processing technology based on the reflective intensity modulated fiber optic sensor

Junjie Yang, Zhihe Fu, Yibiao Fan, Wenxiang Chen, Zhiping Xie, Wei Wu, Xiaoyu Shan

Computer Modelling & New Technologies 2015 ***19****(2D) 26-30*

Sensor technology is one of the most representative of the emerging technology. At present, the sensor has been widely used in national defence, industry, agricultural production, environmental protection, biological science, measurement, transportation, each field of automatic control and household appliances, etc.. Optical fiber sensing technology is accompanied by the development of optical communication technology gradually formed, compared all kinds of optical fiber sensor and the traditional sensor has a series of unique advantages, such as high sensitivity, anti electromagnetic interference, corrosion resistance, electrical insulation, explosion-proof, light path with the flexible, convenient for connecting with a computer, the structure is simple, small volume, light weight, low power. In this paper, the intensity modulation type reflective optical fiber displacement sensor, studied the basic principle, in fact, is the displacement measurement in particular, on the assumption that the condition of uniform distribution, the emergent light field is analyzed in detail, the expression intensity modulation function under various conditions were obtained.

Keywords: Algorithm, Fiber optic displacement sensor, weak signal processing, band-pass filter