The Technology Research of The Semantic Text Classification

Guixian Xu[1*], Lirong Qiu[1]

[1]*College of Information Engineering, Minzu University of China*

Received 30 October 2014, www.cmnt.lv

Abstract

Semantic text classification is to classify the text according to the concepts of semantic relation. It can improve the performance of classification. This paper provides an efficient and accurate method of semantic text classification. First, the classification ontology is constructed by using the concepts extracted from Hownet. Second, Text is represented by semantic vector and general vector space. Then the semantic similarity calculation method is proposed among concepts. The similarity of concepts is calculated based on it. At last, semantic text classification is conducted based on KNN. The comparison of semantic classification and traditional classification is studied. Experiments show that the text classification method based on semantic relation can improve the classification accuracy effectively. The research is meaningful in the application of text clustering, information retrieval, natural language processing and construction of high-quality Tibetan corpus.

Keywords: Semantic Similarity Calculation, Text classification, Ontology, KNN

1 Introduction

Text classification refers to determine a category for each document in the collection of documents according to the predefined subject categories. The text classification technology is widely used in the fields of information filtering, information retrieval, search engine, and digital library. Text classification research began in the 1960s. From the early knowledge engineering classification to machine learning classification, the accuracy of text classification has achieved a lot of improvement. Many classical classification algorithms are produced, such as: naive bayesian model, KNN classification, support vector machine (SVM) method, and artificial neural network. The classification method improved the accuracy of classification much more, compared with the knowledge engineering method, and had good portability.

These classification methods, however, regarded the text as a vector space composed of keyword weight, and ignored the semantic relations among the feature vector. It reduced the accuracy of classification. So some scholars proposed text classification algorithm based on semantic. During the process of feature extraction, the semantic factors are joined. The words of feature vector became concepts. It eliminated ambiguity among words, and solved the problem of high dimension. Text classification based on semantic mainly improved the classifier performance effectively. In the terms of obtaining semantic, many scholars used the hownet [7] and wordnet to get the relationship among the different concepts and obtained knowledge represented by the different concepts. Hownet and wordnet became the bridge of communication of natural language. In the text classification process based on semantics, Some researchers used ontology [5,6], some used semantic similarity algorithm [7] [8] and so on. At present, these techniques have been preliminarily mature, and have brought Gospel for text classification.

In this paper, the semantic classification technology based on the text uses the training set to form the original space vector table, and extracts the concept of ontology from hownet to form a semantic vector space and a normal vector space. The text as semantic vector and normal vector is expressed. From this, it counts semantic similarity of two texts. At last, it classifies the test text according to the calculation method of semantic KNN. The study can be convenient for people to do text classification management, and reduce the burden on people's work. Experiment proves that the study in the classification results are good.

2 Background

Text categorization refers to define a class for each document of document collections, according to predefined subject category. The early knowledge engineering method required higher for sorters, which was difficult extending and was low accuracy. In the 1990 s, with the development of computer technology and Internet technology, many effective text classification algorithms sprang up. For example, the references [1-2] proposed the expansion of the naive Bayesian algorithm based on probability and statistics and information theory. The Naive bayes algorithm used the Naive Bayesian theorem to predict category attributes of the text. The principle of the formula is as formula (1):

$$P(C \mid F_1,...,F_n) = \frac{P(C)P(F_1,...,F_n \mid C)}{P(F_1,...,F_n)} \quad (1)$$

* ***Corresponding author's*** e-mail: xuguixian2000@sohu.com

The algorithm was simple and effective in theory, but there was an assumption that the classification properties were independent among samples, and this assumption was not easy to achieve in reality. The reference [3] proposed the algorithm of supporting vector machine (SVM) based on knowledge study. Through the sample space mapped to a high-dimensional space, support vector machine (SVM) algorithm, which used kernel function to solve the problem of high-dimensional space dimension calculation, could convert nonlinear problem to Linear low dimensional problems, so that it could achieve the purpose of text classification. Kernel function is as formula (2):

Let x, z∈X, and X belong to R (n) space, and nonlinear function Φ realize the mapping of input space X to feature space F. Where the R belonged to F (m),and n < < m.

$$K(x,z) = <\phi(x), \phi(z)> \tag{2}$$

Among them, <, > expressed inner product, K (x, z) was the kernel function. KNN algorithm based on the weight of TF*IDF was put forward in the reference [4]. According to the calculation of the similarity between unspecified samples and each sample of the training sample collections, find the nearest training samples whose total is k(k≥1). Then according to some decision rules, finding unspecified samples belonged to which class.

However ,with the increasing of unspecified dates, the high dimensional and sparse properties of feature vector were more obvious. Only improving the traditional classification algorithm to solve the problem wasn't enough. So the text classification based on semantic was put forward and got attention. The initial algorithm only stayed in the theoretical level, however, as the function of wordnet and hownet improved and the technology of ontology was mature, the text classification based on semantic has become the main way to solve the problem of classification bottlenecks. Text categorization was developed rapidly, so many algorithms were proposed.

In the aspect of ontology application, the characteristics extension method based on the theme of the ontology was put forward in the reference [5], which considered the semantic association among the features, so it obtained a better performance of classification. The literature [6] used ontology to abstract classification, and used abstract words ontology to create an abstract summary of the text. And through these files, it used the abstract analysis to perform text organization. In the aspect of semantic similarity computation, The literature [7] proposed the improved semantic similarity algorithm based on the "HowNet". At the same time, it considered the contribution of "righteousness element" and symbolic. The semantic similarity algorithm is as formula (3). For the concept in the "HowNet", it used "righteousness element" and symbolic to describe something, whose righteousness element was the direct description, and symbols was indirect description. In this formula, sim1 was the basic "righteousness element" similarity, and sim3 was the symbol similarity, and sim2 was "righteousness element" direct description (neither a symbol, nor the basic "righteousness element") similarity. For α+β+γ=1, the value of α was maximum. The literature [8] was that there were two different correlations among different concepts, one of which was Lexical similarity, and the other was Hyperlinks among articles based on encyclopedia. Based on these two kinds of correlations, an algorithm was proposed to calculate the distance of different nodes. Due to the introduction of the real world's knowledge, the algorithm's accuracy was higher than the traditional similarity calculation. The literature [9] combined the ontology with semantic similarity calculation, and used the adjustable parameters to balance the influence of semantic distance, node level and node density, so that it could adapt to the application of different situations. The algorithm is as formula (4).

$$sim(c_1,c_2) = \alpha \times sim_1(c_1,c_2) + \beta \times sim_1(c_1,c_2) + \gamma \times sim_3(c_1,c_2) \tag{3}$$

$$sim(A,B) = \{\frac{a}{distant(A,B)+a}\}^{\alpha} \cdot \{\frac{depth(A)+depth(B)}{|depth(A)-depth(B)|+1}\}^{\beta} \cdot \{\frac{1}{density(A,B)}\}^{\gamma} \tag{4}$$

In order to adapt to the development of the technology, the literature [10] presented distributed Chinese adaptive text classification algorithm based on semantic keywords extraction in the cloud computing environment, which in the agent did distributed extraction based on semantics for Chinese text keywords. Then according to the key words, classify the text. It improved the performance of the algorithm in the cloud computing environment.

3 The proposed method

3.1 FORM THE ORIGINAL SPACE VECTOR TABLE

Vector space model transmutes the given text into high dimensions vector, and transmutes text processing problems into vector arithmetic problems. In the original text in the library, the text is defined as d, and each text uses d={w1,w2,…,wn} to express it, where n is the text vector space dimension. Characteristics of each component have a weight vector v, which represents the feature in the important degree of a text.

There are a lot of feature weights for the calculation of the method, such as TF algorithm, the IDF algorithm, TF*IDF algorithm, and so on. The study of our research is the TF*IDF algorithm. The algorithm is as formula (5) (6):

$$TFIDF(t,d) = TF \times IDF \tag{5}$$

$$IDF = \log(|D|/DF) \tag{6}$$

TF is characteristic frequency, which stands for the frequency of characteristic t appearing in text d. DF is the frequency that characteristic t appears in the training text set. IDF is the inverse document frequency. |D| is the total number of training document. The greater the characteristic frequency of TF is, the higher frequency feature vector appears in the text, and the more important it is to classification. The greater The IDF inverse document frequency is, the less the feature vector appears in the other documents, and the stronger the ability distinguishes the document category.

Currently, TFIDF weight calculation method is the most wide research and application of a weighting method, which shows better performance in text processing.

3.2 THE FEATURE PROCESSING

The characteristics of the processing is the most important part in text categorization. This study uses training set to form the original space scale, and picks the concept of ontology up from hownet to form a semantic vector space, and then it uses semantic vector and normal vector to express text, so that it can do it easily.

3.2.1 The concept of ontology extraction

Ontology is a clear and formal specification of the shared conceptual mode [12] [13], and it has the characteristics of conceptualization, explicit, formalization, and sharing. Semantic information can be got through ontology, so we can get text categorization based on semantics. In this paper we will study extract ontology concepts based on HowNet [7].

Extract the words in collections that express categories. Such as: MaterialOf collection, RelateTo collection, Part - Position collection, belong collection, whole collection, domain collection and so on. Then construct MaterialOf, RelateTo PartPosition, belong, whole and domain files. W_C vocabularies of Each file are taken with nouns, and there is only a collection of categories of these nouns. If there are multiple collections, we do complex processing. In other words, it will set up file to contain all of the complex information. If there contain two words expressing category, the first one is taken. If there contain three words, the first two are taken. If there are many words, the first one is taken.
Hownet structure is as shown in figure 1:

NO.=111567
W_C=南天竹
G_C=noun [nan2 tian1 zhu2]
S_C=
E_C=
W_E=nandina
G_E=noun [3 nandina┴noun┴-0┴static┴种]
S_E=
E_E=
DEF={FlowerGrass|花草:MaterialOf={medicine|药物}}

FIGURE 1 Example Of Hownet

3.2.2 Create the classification ontology

It is defined as 16 categories, which were respectively, the universe, biological, medical, social, diet, emotional, agriculture, science and technology, information, transportation, economy, politics, education, religion, physical, and mathematical. The categories with hyponymy relations extracted from the hownet is defined as the middle class, and each middle class is classified into a large class by domain experts, so they accomplished a category data collection. Encoding all kinds of concepts in large classes to construct ontology tree of large classes. According to the hyponymy relation among concepts, constructing the hierarchy relationship among codes. If the leaf nodes and internal nodes are repeated, internal nodes are marked. If the internal nodes and internal nodes are repeated, both of them are marked. Duplicate nodes in semantic computation are deleted. Then all the category ontology tree are gathered together to form a domain ontology classification.Ontology coding structure is as shown in figure 2:

A 类别
A_01 :二级节点
A_01_001: 三级节点

FIGURE 2 Example Of The Ontology Coding

3.2.3 Concept matching

Match the vector space form V and the concept of semantic ontology O, and change concept that appeared in the vector space form and ontology at the same time into semantic vector space. $O_1=\{S_1,S_2,S_3,\ldots,S_n\}$ Changing characteristic that appeared in the vector space form but doesn't appeared in ontology into the common feature vector space $V_1=\{T_1,T_2,T_3,\ldots,T_n\}$.

According to the semantic vector space and the common feature vector space, each text d of after word segmentation is shown as semantic vector O(d) and common feature vector V (d), where $O(d)=\{w_1,w_2,\ldots,w_m\}$, and $V(d)=\{w_1,w_2,\ldots,w_p\}$. wi=TF*IDF, where TF is the characteristic word frequency of T or S, and DF is the document frequency of T or S, and IDF is the reverse document frequency.

3.3 THE IMPOVED KNN CLASSIFICATION BASED ON SEMANTICS

KNN algorithm has become a very important method of classification in the 1960s. KNN algorithm is a simple and effective method, which has the advantages of low cost of re-training, but the KNN algorithm is a kind of lazy classification algorithm. The training is faster, but the classification is slower. The computational complexity will increase linearly with the increase of training set size. A lot of people put forward the improvements on KNN based on this [14]. In this study, first we calculate the text similarity according to weight, and then we start to calculate.

3.3.1 Calculation of text similarity

Calculate similarity $S(d_1,d_2)$ of two texts. Text similarity is composed of two parts, where sim_1 is a semantic similarity, sim_2 is a non semantic part similarity. The algorithm is as formula(7) (8) (9) (10).

$$S(d_1,d_2)=sim_1(d_1,d_2)+sim_2(d_1,d_2) \tag{7}$$

$$sim_1(d_1,d_2)=\sum_{i=1}^{m}\sum_{i=1}^{m}\frac{sim(S_i,S_j)}{a\times b} \tag{8}$$

$$sim(S_i, S_j) = w_i \times w_j \times \gamma \quad (9)$$

$$sim_2(d_1, d_2) = \frac{x_1 \times u_1 + x_2 \times u_2 + ... + x_p \times u_p}{\sqrt{x_1^2 + x_2^2 + ... + x_p^2} \times \sqrt{u_1^2 + u_2^2 + ... + u_p^2}} \quad (10)$$

w_i is the concept of d_1 document in the weight of S_i, and w_j is the concept of d_2 document in the weight of S_j, and γ is the similarity parameters. When S_i and S_j are equal, γ=1; and when S_i and S_j are brothers nodes of classification of semantic ontology , γ = 0.9; and when Si and S_j are cousins nodes of classification of semantic ontology, γ = 0.5; and when S_i and S_j are without genetic relationship, γ=0. a is non zero concept of text d_1 in semantic space, and b is non zero concept of text d_2 in semantic space.

3.3.2 Based on the weighted similarity classification decision

The text adopts the method of weighted similarity of making decisions, Its function of decision rules is shown in the following formula (11) (12) (13):

$$S(d) = \arg \max_{i=1}^{|C|} score(d, c_i) \quad (11)$$

$$score(d, c_i) = \sum_{d_j \in KNN(d)} sim(d, d_j)\delta(d_j, c_i) \quad (12)$$

$$\delta(d_j, c_i) = \begin{cases} 1 & d_j \in c_i \\ 0 & d_j \notin c_i \end{cases} \quad (13)$$

Where S (d) stands for category of the highest scores to make the decision category of document D. Score (d, ci) is the scores of d belonging to ci category, and Sim (d, d_j) is similarity score of d and d_j, and KNN (d) is K neighboring document of d. δ(d_j, c_i) indicates whether the d_j is as part of the c_i.

3.4 EVALUATION METHOD

The major source of the accuracy of classification model estimation is confusion matrix or classification matrix and contingency table. The following figure shows a binary classification problem of confusion matrix. The confusion matrix binary classification problem as shown in table 1.

TABLE 1 The Confusion Matrix Of Binary Classification

		The actual classification	
		Positive example	Negative examples
The predicted Category	Positive examples	The number of the correct Positive cases(TP)	The number of the false Negative cases (FP)
	Negative examples	The number of the false Negative cases (FN)	The number of the correct Positive cases

Based on the origin of the correctly classification model. we use standard called to evaluate the correctness of classification model.

Evaluation formula is as shown in formula (14) :

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

The percentage of TP for the correct classification of the document, TN is the document number of the correct classification, The percentage of FP for the misclassification of the document, FN is the document number of the misclassification of the document.

4 Experiments and result analysis

This study selects 16 kinds of ontology from HowNet , The ontology category statistics is shown in table 2 and table 3, ontology coding signal is shown in figure 3.

TABLE 2 The Total Number Of Ontology

The number of big classes	The number of all middle classes	The number of leaves
16	92	9096

TABLE 3 Ontology Category Table

Name of big classes	The number of every middle class	The leaf number of every middle class
1. Universe	8	187,70,24,76,67,108,88
2. Biology	9	176,70,154,140,43,33,11,32
3. Medicine	7	218,184,142,233,44,122,257
4. Society	7	102,233,114,40,232,128
5. Food	9	143,187,73,64,77,59,62,96
6. Emotion	1	340
7. Information	5	58，204,29,82
8. Agriculture	7	106,158,96,34,35,26
9. Technology	3	59,56
10. Traffic	5	100,84,27,94
11. Economy	6	208,119,389,31,389
12. Politics	9	357,83,88,197,101,75,162,77
13. Education	10	23,148,63,147,57,102,76,12,139
14. Religious	1	28
15. Substance	4	41,31,28
16. Math	1	221

A政治类
A-1外交
A-1-1外交使节
A-1-2外交家
A-1-3一等秘书
A-1-4批准书
A-1-5教廷公使
A-1-6公约
A-1-7外交使团
A-1-8代办
A-1-9不受欢迎的人
A-1-10使臣
A-1-11使团
A-1-12受欢迎的人
A-1-13特使
A-1-14对外政策
A-1-15公使
A-1-16外相
A-1-17来使
A-1-18领事条例
A-1-19使节身分
A-1-20教廷大使
A-1-21密使
A-1-22外交豁免权
A-1-23代理领事
A-1-24单边主义

C经济类
C-1证券
C-1-1证券
C-1-2创业板
C-1-3大盘
C-1-4股票交易所
C-1-5股本
C-1-6上市公司
C-1-7重获股份
C-1-8股市
C-1-9猴市
C-1-10券商
C-1-11证券市场
C-1-12二板
C-1-13次级证券
C-1-14证券业
C-1-15一板市场
C-1-16混合证券
C-1-17债券市场
C-1-18期货市场
C-1-19不通过交易所直接售给顾客的证券
C-1-20开盘价格
C-1-21流动性证券
C-1-22证券承销商
C-1-23股东
C-1-24创业板市场

FIGURE 3. Example Of Ontology

We use 10 classes a total of 300 corpus in the process of classification. 80% of the corpus are used for training, 20% are used for testing. Semantic text classification software screenshots is shown in figure 4.

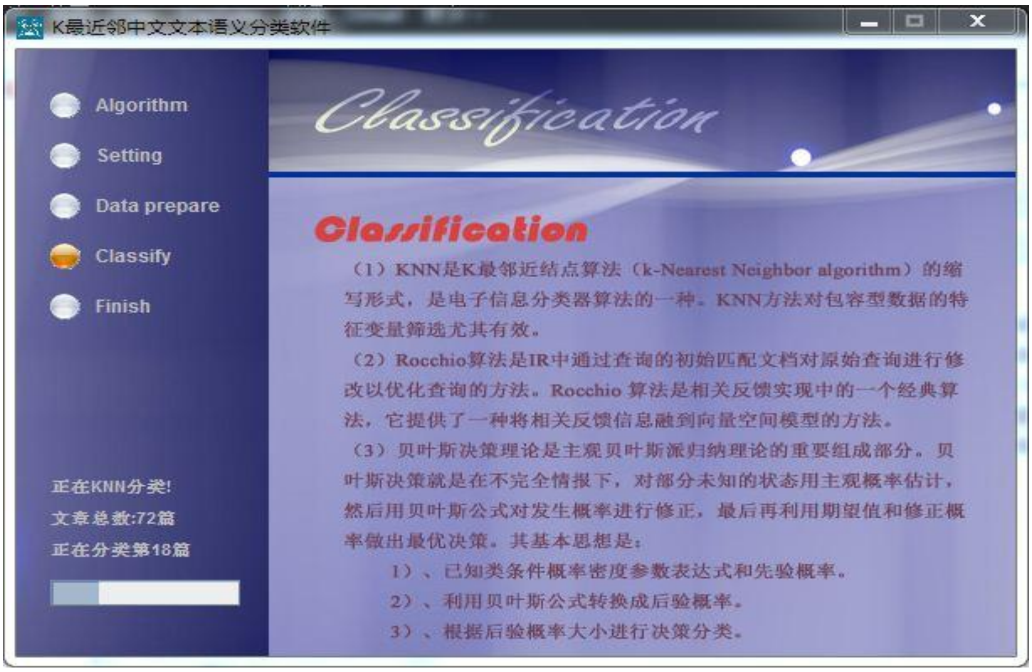

FIGURE 4. Semantic Classification Software

In order to illustrate the effect of this research better, we select the detailed results of a category to observe specifically text classification performance. Experiment randomly selects the biology class and economy class, and there are five texts selected randomly in the two types of testing texts. We show the similarity of five biological testing texts and biology, agriculture, technology, information, economic training sets, which are shown in table 4. We show the texts similarity between five testing texts of economic class and economy, information, technology, transportation, social training sets, which are shown in table 5 (Where the out9 prefix is the economic article, and the out8 prefix is biological article).

TABLE 4. Biology class five articles in the class of similarity and the degree of similarity in the other four classes

	out8-14	out8-15	out8-17	out8-18	out8-2
Biology class	0.844	0.809	0.766	0.652	0.610
Agricultural class	0.012	0.023	0.414	0.033	0.025
Tech	0.011	0.015	0.626	0.023	0.024
Information class	0.003	0.027	0.023	0.006	0.014
Economy class	0.361	0.004	0.008	0.013	0.017

TABLE 5 Economy class five articles in the class of similarity and the degree of similarity in the other four classes

	out9-9	out9-12	out9-30	out9-1	out9-11
Economy class	0.823	0.801	0.798	0.786	0.77
Information class	0.004	0.013	0.019	0.005	0.018
Tech class	0.023	0.037	0.010	0.022	0.271
Traffic classes	0.048	0.025	0.029	0.031	0.016
Society class	0.034	0.039	0.036	0.025	0.062

According to table 4, in the five testing texts of biology class, the similarity between each text and biological class is highest. According to the highest similarity, it is marked biology. In table 4, out8-14 belonging to the biology class has the highest similarity, and similarity value is up to 0.844. The similarity of Out8-2 belonging to the biology class is 0.610, and the similarity belonging to other four classes is less than 0.025. It is observed that the similarity belonging to the biological class has very strong class discrimination. According to table 5, in the five testing texts of economy class, every text belonging to economy class has the highest similarity. And according to similarity, it is marked economy class. In table 5, the similarity of out 9-9 belonging to economy class is as high as 0.823, and the similarity of out 9-11 belonging to economy class is 0.77. In table 4 and table 5, because there contains the concept of semantic ontology in the testing texts, the text similarity of it and the training set can be greatly improved, thus the performance of the classification is improve.

We conduct classification experiments on these data five times, and we chose the "accuracy" as the evaluation standard. The accuracy of using semantics and not using semantics are as shown in table 6.

TABLE 6. The experimental evaluation results

Number of test	First	Second	third	Fourth	fifth
correctness（Don't use the semantic）	0.826	0.827	0.917	0.875	0.83
correctness（use the semantic）	0.900	0.917	0.983	0.942	0.908

According to table 6 we can know, in the five experiments, after introducing ontology semantic, the classification correct rate is obviously improved than before. In the third experiment, the improvement of the correct rate is 0.066, which is minimum. In the second experiment, the correct rate is up to 0.09, which increases most. Five experiments are improved by more than 0.066. This shows that the classification algorithm of semantic played a great role on improving classification performance, which joined the semantic factors. The classification algorithm based on semantic mapped traditional features vector into semantic vector and normal vector. For semantic vector, eliminating isolation among words by using the concept of the hyponymy, so it greatly improves the feature similarity. Compared with traditional classification algorithms, the classification algorithm based on semantic has obvious advantages.

5 Conclusion

This study uses the training set to form the original space vector table, and extracts the concepts from hownet. Then classification ontology is constructed. Then semantic vector and normal vector are used to express text, and then the text is shown as semantic vector and normal vector. Then semantic similarity of two texts is computed. Finally text classification according to semantic KNN classification is executed. Through the analysis of experimental results, it can be shown that the semantic classification is better than the common classification. It has a good application prospect in the information filtering, information retrieval, search engine, and text databases.

Acknowledgement

This work is also supported by “the National Natural Science Foundation of China (No.61309012)”. This paper is also supported by the Fundamental Research Funds for the Central Universities. The project title is "The Research on Hot Topic Discovery and Tracking Technology of Tibetan Network".

References

[1] Ferreira-da-Silva, AR. 2009 Bayesian mixture models of variable dimension for image segmentation. Computer *Methods and Programs in Biomedicine* **94**(1), 1-14
[2] Iribarren, I, &Chacon E, &De-Miguel, E.2009 A Bayesian approach to probabilistic risk assessment in municipal playgrounds. *Archives of Environmental Contamination and Toxicology* **56**(1), 165-72
[3] Adankon, M. M, &Cheriet, M., &Biem, A.2011 Semisupervised learning using Bayesian interpretation: application to LS-SVM. *IEEE Trabs Neural Netw* **22**(4), 513-24
[4] Cover T M, &H art P E. 1967 Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13**(1), 21 -7.
[5] Yan Zhan, &Hao Chen. 2014 Short text classification based on extensible characteristic theme. *Journal of Hebei University (NATURAL SCIENCE EDITION)* **34**(3):307-11.DOI:10.3969 /j.issn.1000-1565.2014.03.017
[6] McAllister, R.A., &Angryk, R.A. 2013 Abstracting for dimensionality reduction in text classification. *International journal of intelligent systems* **28**(2), 115-38.
[7] Liyong Cao, &Cheng Zheng. 2014 Improved algorithm of semantic similarity based on HowNet. *Electronic technology* **47**(5),1-3.DOI:10.3969/j.issn.1000-0755.2010.05.001.
[8] Yazdani, M., &Popescu-Belis, 2013 A. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia.. *Artificial Intelligence*, **194**, 176-202.
[9] Shenyan Chen, &Junhua Wu. 2008 Concept semantic similarity computation based on ontology and its application. *Microelectronics and computer* **15**(12),96-9.
[10] Jiajie Shen, &Hong Jiang, &Su Wang, and so on. 2014 Computing semantic text keyword cloud based on adaptive classification. *Computer engineering*, (7),247-253. DOI:10.3969/j.issn. 1000-3428.2014.07.051.
[11] Kim H, &Howland P, &Park H. 2005 Dimension Reduction in Text Classification with Support Vector Machines. *The Journal of Machine Leaming Research* **6**(1), 37-53
[12] Hua, Hu 2014. Research on ontology construction and information extraction technology based on wordnet. *Journal of Digital Information Management*, **12**(2), 114-119.
[13] Zhang, S. 2012 Nearest neighbor selection for iteratively kNN imputation. *The Journal of Systems and Software* **85**(11), 2541-52.
[14] Shengyi Jiang, &Guansong Pang, &Meiling Wu et al. 2012 An improved K-nearest-neighbor algorithm for text categorization.. *Expert Systems with Application* **39**(1), 1503-9.

Authors

< Guixian Xu >, <1974.12>,< Beijing, P.R. China >

Current position, grades: Vice Professor
University studies: Beijing Institue of Technology
Scientific interest: Data Mining
Publications <number or main>: published more than 10 academic papers
Experience: works at Minzu University of China, engages in teaching of data structure and operating system, conducts the research of data mining and text mining.

< Lirong Qiu >, <1978.8>,< Beijing, P.R. China >

Current position, grades: Vice Professor
University studies: Chinese Academy of Sciences
Scientific interest: Knowledge Engineering
Publications <number or main>: published more than 20 academic papers
Experience: works at Minzu University of China, engages in teaching of computer network, conducts the research of natural language processing, artificial intelligence and distributed systems.