Research on Food Complains Document Classification Based-on Topic

Xiquan Yang*
Northeast Normal University, Changchun, 130117, China
* Corresponding author: yangxq375@nenu.edu.cn

Caifeng Zou*, Lin Yue and Rui Gao
Northeast Normal University, Changchun, 130117, China
* Corresponding author: zoucf306@nenu.edu.cn

***Abstract*—In this paper, we design a classifier based-on topic for food complain documents, and take a series of measures to the implementation process. In order to accomplish feature reduction, the filter method named term filtering for independent topic features is proposed to compress each topic feature vector. We introduce the created food ontology as background knowledge and to expand the semantic of complaint documents with the aid of HowNet. So we can supplement effective information and improve the effect of text classification. In addition, we take account of different importance between title and body in the text, considering that title can stand out the topic of text better than the textual body. Consequently, we separately calculate the word frequency which words are in textual title and body. The experiments show that it is necessary to consider the different importance between textual title and body, and we can achieve good feature reduction effect using the proposed filter method, and the classification performance get obvious improvement after the process of term expanding.**

***Index Terms*—ontology, food complain, text classification, topic**

I. INTRODUCTION

Food is the material base that human survived and developed. In recent years, due to the vicious food security incident happened frequently, the tremendous amount of food complaint documents have appeared on the web. However, the issued food complaint documents on the Internet uploaded by the customer are out of order, which are not effectively organized. And different information users have different require for complaint information. Therefore, aiming at the trait of food complain documents, our paper proposes a classifier based-on topic. It can effectively organize and manage these complaint documents, and classify them according to the topic. Thus, people can find the information which they want rapidly and accurately.

The food complaint documents on the web mostly are short texts. Their main trait is that the document features which are used to express the topic are quite sparse. Thus, feature selection for short text is very important. At present, there are many researches about short text classification at home and abroad. Zelikovitz, S proposed applying LSI (Latent Semantic Indexing) for the term selection of short text classification [1]. LSI is a commonly used feature extraction algorithm. It is often used for dimension reduction of high-dimensional sparse text data. However, it cannot perfectly solve the problem of polsemantic words. The existing text classification methods mostly have studied on long text. We cannot use them for short text classification very well, especially for such complaint documents. In each complaint document, there are only a few feature words expressing the document topic, and documents with the same topic may have been expressed by synonyms and nearsynonyms. So the existing traditional classification methods cannot suit our problem. We have many studies on exploiting the semantic knowledge of ontology to supervise the classifier learning. For example, many researchers exploit some general ontologies such as WordNet[2], Euro WordNet, CoreNet and HowNet[3], to give an intuitive way to simulate human's document labeling. Many kinds of semantic relations in these general ontologies are used as a bridge to provide the linkage between the category and the unlabeled documents for realizing automatic text classification [4].

This work enhances the term weighting model based on domain knowledge in ontology and with the help of HowNet to improve the classification process. The instances in food domain ontology are used to help each topic get its feature words. At the same time, we exploit word similarity computation method provided by HowNet to calculate the similarity between the words in topic term vector and the instance words in food ontology, and then the words whose similarity value is greater than a threshold are extracted as the related words of topic term vectors. Finally, according to the proposed rectifying feature weights method, we use these related words to rectify the weight of each topic term vector and each topic inverse document frequency vector. For each topic, its

[1] Manuscript received Sept. 8, 2011; revised Oct. 18, 2011; accepted Oct. 28, 2011.
[2] project number: 20090303

doi:10.4304/jsw.7.8.1687-1693

inverse document frequency vector has the same vector space with its term vector.

The rest of this paper is organized as: We briefly generalize related work in section II; In section III, section IV and section V, we describe our proposed food complain document classification method, and show the experiments and evaluation of classification in section VI, in conclusion we summarize our results and mention future research work.

II. RELATED WORK

Document classification is the process of classifying documents into pre-defined categories. In the text mining process, ontology can be used to provide background knowledge of a domain. Some recent researches show that domain ontology is importance in the text classification process [5-8]. In [9] the authors present a novel automatic text categorization method based on ontological knowledge that does not require learning and can implement real-time classification. The [10] presents a novel ontology-based automatic classification and ranking method, in which web documents are characterized by a set of weighted terms, categories are represented by ontology. In [11] the authors present an approach towards mining ontology from natural language, in which they considered a domain-specific dictionary for telecommunications documents. The [4] presents exploiting the semantic of the category name for supervising the classifier learning. The semantic of category name is represented as a set of keywords for automatically document labeling and then for fully automatic text categorization.

Some of the recent literatures show that works are in progress for the efficient feature selection to optimize the classification process. A novel term weighting method is presented in [12], which estimate statistically the importance of a word to the classification problem. The [13] proposed a new feature scaling method, called class–dependent–feature–weighting (CDFW) using Naive Bayes (NB) classifier. Furthermore an algorithm exploiting the extracted associated frequent sentences and co-occurring terms is presented in [14], which reduce the feature space and then using domain ontology to convert these terms to a concept and update the VSM with new feature weights. Rossitza[15] et.al presented a method to create relationships between keywords and concepts in WordNet, generate a concept set representing each text, and classify these texts using concept sets. Jesús Oliva et.al[16] computed the similarity between texts according to the sentence structure and semantic information from the ontology in short text classification.

In our work, we exploit the built food domain ontology, and have the aid of word similarity computation method provided by HowNet. And then we combine the domain knowledge from ontology and traditional statistical information, to expand the topic term in semantic and rectify the weight of topic term vectors, finally to realize the semantic expanding of text.

III. THE PROPOSED CLASSFICATION METHOD

A. Methodology

The purpose of classification is to discover the common characteristics existed in the same kind of data objects to build classifier, and then determine the class or topic of test text. The text classification method proposed in our paper realizes classification based on the document's topic. According to the nature distribution of each topic's dataset, we use traditional statistical method to form its topic term vector from its training set. Since in each formed topic term vector, there are many irrelevant terms, even most of them are noises, so we make a further term filtering process. The term filtering method we adopt here is called as term filtering for independent topic features. Its purpose is to make the term in each topic term vector and the term in other topic term vectors completely dissimilar and non-overlapping. And then we can reduce noise and achieve feature reduction. Because mostly complaint documents are short texts, it has the shortcomings that the number of words in each document is small, and text features in each document are not enough. For solving the problem and making up the defect that the important features are lost after the process of term filtering for independent topic features, we exploit the domain knowledge of food ontology and together with the aid of word similarity computation method provided by HowNet to expand topic term vectors, and then improve the effect of classification. The schema of our proposed text classification method is shown in Fig. 1.

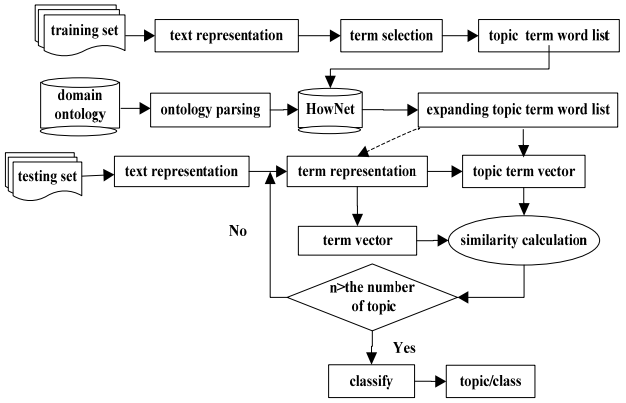


Figure 1. An overview of the method.

B. Weight Calculation and Text Representation

First of all, we should preprocess all food complaint documents. The preprocessing of Chinese text includes Chinese word splitter and removing stop words. In order to facilitate the following computer processing for text, after preprocessing, we implement text representation for documents. We represent the words in each document as mathematical vector. Therefore, we should calculate the weight of words in document. We have many term weighting computation methods now. Eq. (1) is the classical formula of TF*IDF used for term weighting.

$$w_{ij} = tf_{ij} * \log(N / df_i) \quad (1)$$

Where w_{ij} is the weight of the term i in document j, N is the number of documents in the collection, tf_{ij} is the

term frequency of the term i in document j and df_i is the document frequency of the term i in the collection. Eq. (1) shows that the higher of tf_{ij}, the more improtant the term to document. The higher of df_i, the effect is lower when use the term to measure the similarity of document. Considering that the title of complaint doucuments can stand out the textual topic better than the textual body, when compute the word frequency, we calculate the weight of textual title and body separately. In the experiment, we adopt the following two ways to compute the tf_{ij} value.

Way 1: Without consideration of the importance difference between textual title and body, their word frequency is normally computed according to the times that the word appears in its document.

Way 2: Distinguish the importance difference between textual title and body, we think the word in title is more important than words in the body, considering, which can stand out the topic of document better. The computation of their tf_{ij} is according to the proportion λ, where λ>1. The tf_{ij} of words in textual body is computed with the times that the word appears in the document, but in textual title, the tf_{ij} of word is computed with the λ times of its appearing times.

IV. TERM SELECTION

A. Forming the Initial Topic Term Vector (named operation A)

We assume that the dimension of each topic term vector and testing text term vector is enough high when we adopt VSM in our work, having enough features used for classifying. But in practical application, the length of document is far less than the dimension of term vector. In the model of VSM, term vector space is completely built based on primitive training text set. That is to say, all terms in each topic term vector are from primitive documents. Then we compute the similarity between the testing text vector and topic term vector to classify testing text.

Our paper uses a simple term selection method TF-IDF to produce initial topic term vectors. In order to construct topic term vector C_i for topic i, at first we should compute the sum of tfidf value of each word in training documents of topic i, and then sort them according to its sum in ascending order. Finally we take words whose values are big as terms. And the average of each word's tfidf is used as its term weight, and the length of each topic term vector is set as a stated percentage of its original word vector. That is to say, if the length of its original word vector is L, the length of topic term vector is α*λ, where α ranges from 0 to 1. So the lengths of topic term vectors may not be the same. The topic term vector C_i is represented as $C_i = \{t_{i1}, w_{i1}; t_{i2}, w_{i2}; ...; t_{in}, w_{in}\}$, where tf_{ij} is the term or word in topic term vector, w_{ij} is the term weight of that word. The w_{ij} is computed by Eq. (2).

$$w_{ij} = \sum_{k=1}^{m} w_{ij} / m \quad (2)$$

Where w_{ij} is the average of m word weights of term tf_{ij} in topic i, m is the number of documents containing the same term tf_{ij} in topic i.

B. Term filtering for independent topic features (named operation B)

Because there are many irrelevant feature terms in each topic term vector formed in operation A, and most of them are noises. To address this problem, we do the further term filtering process. Here we exploit one term filtering method named term filtering for independent topic features.

The basic idea of term filtering for independent topic features method is described in detail as follows. We deal with each topic term vector in operation A, throwing away all terms which not only exist in its topic term vector, but also exist in other topic term vectors. Thus in the new forming topic term vectors, the term in each topic term vector is completely different from other topic term vectors. That ensures the independence among the topic term vectors, reducing the interference of classification.

The experiments demonstrate that it can reduce the dimension of each topic term vector after term filtering, and achieve the purpose of feature reduction. In our work, we can reduce half of the dimension for our dataset.

C. Expanding Topic Term (named operation C)

Complaint documents are mostly short, which have the inherent shortcomings that features are not enough and the information is sparse. In addition, it loses some important features after term filtering for independent topic features. Thus, we need supplement the feature of topic term vectors to improve the performance of classification. So based on operation B, we do the further term expanding process. For getting some appropriate features to expand the topic term vectors semantically, our paper introduce the domain knowledge of ontology and have the aid of word similarity computation method provided by HowNet to expand each topic term word list, and then achieve the purpose of semantic expanding. The detail procedure describes as follows.

Step 1: We parse the instances of abnormal subclasses in food domain ontology by adopting Jena2.5.3 API. Take the milk product as example, we build food ontology. In the ontology, we set the subclass of "normal" and "abnormal" for the class of "ingredient", "expiration date", "sterilizing method", "moisture", "volume", "weight", and so on. Combining with information extraction method, we extract the abnormal food information from the documents about the milk product, and add the information as instances into the related subclass of "abnormal".

Step 2: Selecting a topic, we perform a match between the words of instances parsed by Step 1 and the term from the selected topic term vector. If succeed, we will enlarge

the weight of the term, here we weight the term according the Eq. (3).

$$newtfidf = oldtfidf + \beta * oldtfidf \quad (3)$$

Where β is an adjustable parameter according to the effect of experiment, ranging from 0 to 1.

Step 3: Exploiting word similarity computation method provided by HowNet, all the unsuccessful instance words are calculated word similarity with the words in topic term vector. The instance words having the biggest similarity value are selected and compared with the threshold. If they are bigger than the threshold, the instances are taken as the related feature words and are added into the topic term vector containing the word which is most similar with the instance word, and their term weights are set as the term weight of its most similar term.

Here we give an example to explain. There are two Chinese words in the "food quality" topic: 苦涩 and 涩.One of them exists in a topic term vector, and the other belongs to the instance word of domain ontology. We use the word parser of HowNet to calculate the similarly of the two words. 苦涩={ADJ aValue|属性值,taste|味道,bitter|苦,undesired|莠}, 涩={ADJ aValue|属性值,taste|味道,bad|坏,undesired|莠}, we compute their similarity using the word similarity computation method of HowNet. Their similarity is 0.86842, so the instance word can be regarded as related feature term. And then it was added into the term vector of "food quality" topic, and was given the same weight with the term in topic term vector.

Step 4: Using the related feature terms produced by Step 3 to filter the topic term vector, that is to say, deleting the same terms in other topic term vectors. So we can guarantee the term is completely different between each two topic term vectors. Up to now, we get the new topic term vector which is expanded.

We implement Step 1 to 4 for all topic term vectors which produced by operation B, and end it until we get new topic term vectors of all topics.

V. SIMILARITY COMPUTATION

After part C, we get the final topic term vectors and topic word lists. Next, we need represent the test text by text representation method. Here we still use TF-IDF method for representing. After the text representation of test text, we get test text term vectors with the same space of topic term vectors separately to different topics. So we can get the topic similarities of test text by computing the cosine value of their vectors. Before similarity computing, we compute the word frequency tf of each word in test text. The word frequency computation method we adopt here is the same as methods in training text. We still exploit two ways to compute its word frequency. Eq. (4) is the formula of Cosine similarity computation.

$$Sim(C_i, T_j) = \cos(C_i, T_j) = \frac{\sum_{k=1}^{n} w_{ik} * v_{ik}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2} \times \sqrt{\sum_{k=1}^{n} v_{jk}^2}} \quad (4)$$

Where C_i represents the topic term vector of topic i, and T_j represents the test text vector of document j in test text set, which formed by the process of term representation. Here, $C_i = \{t_{i1}, w_{i1}; t_{i2}, w_{i2}; ...; t_{in}, w_{in}\}$, $T_j = \{t_{j1}, w_{j1}; t_{j2}, w_{j2}; ...; t_{jn}, w_{jn}\}$. C_i and T_j have the same vector space. The similarity computation algorithm works as follows:

Algorithm 1:

Input: $tf_j \leftarrow$ the word frequency vector of test text j;
$df_j \leftarrow$ the document frequency vector of test text j;
$C \leftarrow$ a set of topic term vectors;
$Tt \leftarrow$ a set of topic term word lists;
$n \leftarrow$ the number of topics
Output: $t \leftarrow$ the topic of document j in test text set
1: k=1;
2: for each topic i
3: $Tt_i \leftarrow$ select the topic word list of topic i ;
4: forming the topic inverse document frequency vector idf_i with Tt_i and d
5: $T_j \leftarrow$ do term representation for test text j with tf_j and idf_i ;
6: $C_i \leftarrow$ select the topic term vector of topic i ;
7: $Sim_k \leftarrow Sim(C_i, T_j)$;
8: k++;
9: if($k > n$)*break*;
10: end for
11: comparing n similarity value obtained from aboved ;
12: $t \leftarrow$ the topic that has the highest similarity value ;

For each document in test text set or any new complain document, we use Algorithm 1 to determine what topic it belongs to.

VI. EXPERIMENT

A. Experiment evaluation

To examine the efficiency of the constructed method, our paper adopted the most commonly used measures in data mining, namely, precision, recall, and F1, for the general assessment [14]. Their equation is given in the following.

$$\Pr ecision = \frac{TC}{TC + FC} \times 100\% \quad (5)$$

$$\mathrm{Re}\, call = \frac{TC}{TC + MC} \times 100\% \quad (6)$$

$$F - measure\,(F1) = \frac{2 * \Pr ecision * \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \times 100\% \quad (7)$$

Where TC is the number of documents, and the documents are those belonging to the topic or class actually and are also divided into the topic or class; FC is

the number of documents which do not belong to the topic or class, but divided into the topic or class wrongly; MC is the number of documents which belongs to the topic, but not divided into the topic correctly.

B. Expeimental Results and Analysis

We collected document sets related to milk products complaint from "315 electronic consumption complaint" website. There are 997 documents in total, where 197 documents belong to "food hazard" topic, 206 documents belong to "food hygiene" topic, 365 documents belong to "food quality" topic, and 219 documents belong to "sale service" topic. We organize training documents and testing documents by the ratio of 2 to 1. That is to say, for each topic we take about 2/3 of its topic's documents as training set, and the rest is taken as testing set.

To demonstrate the effective of our proposed method, we do two groups of experiments aiming at two kinds of word frequency computation method. And in each group, we have three pairs of experiments for three operating associations of operation A, B and C in section IV. The experimental results are shown in table Ⅰ and table II, where tf_2 represents that its word frequency computation method is the way distinguish the importance of textual title and body, and tf_1 adopt the normal word frequency computation method. According to our experimental effect, we set the ratio λ as 1.5, that is to say, if the word appears in textual title 1 time, we record it as 1.5 times.

Comparing table Ⅰ and table II, we can discover that using the same operation, the experimental effect with tf_2 is better than the experimental effect with tf_1. So it is necessary to concern the different importance between textual title and body. Furthermore, we can see that the classification result which is processed by the operation A and B is better than the result which is only processed by operation A. This demonstrates that the term filtering for independent topic features method can draw out disturbing noise, and reduce the classification disturbance.

In addition, from table III and table IV, we can also see after operation B, we can reduce about half of dimension of vectors in our experiment, achieving good dimensionality reduction effect. From the item of operation A+B+C in table I and table II, we can see no matter what tf_1 or tf_2 we use, after the term expanding process the classification effect we achieved is obviously better than before. Its precision, recall and F1 values of classification are higher than other two combinational operations. Thus the term expanding method we proposed effectively improve the performance of classification.

TABLE I.
THE EXPERIMENTAL RESULT OF TF1

operation association	topic	Precision(%)	Recall(%)	F1(%)
A	*sale service*	76.71	62.92	69.14
	food hygiene	54.37	59.57	56.85
	food hazard	47.54	75.33	58.29
	food quality	74.42	51.61	60.95
	average	63.26	62.36	61.31
A+B	*sale service*	80.23	77.53	78.86
	food hygiene	68.54	64.89	66.67
	food hazard	65.12	72.73	68.71
	food quality	68.07	65.32	66.67
	average	70.49	70.12	70.23
A+B+C	*sale service*	89.47	76.40	82.42
	food hygiene	82.98	82.98	82.98
	food hazard	68.63	90.91	78.21
	food quality	79.46	71.77	75.42
	average	80.14	80.52	79.76

TABLE II.
THE EXPERIMENTAL RESULT OF TF2

operation association	topic	Precision(%)	Recall(%)	F1(%)
A	*sale service*	80.00	62.92	70.44
	food hygiene	57.43	61.70	59.49
	food hazard	47.93	75.33	58.59
	food quality	73.91	54.84	62.96
	average	64.82	63.70	62.87
A+B	*sale service*	79.07	76.40	77.71
	food hygiene	69.41	62.77	65.92
	food hazard	67.86	74.03	70.81
	food quality	66.93	68.55	67.73
	average	70.82	70.44	70.54
A+B+C	*sale service*	88.16	75.28	81.21
	food hygiene	87.36	82.61	84.92
	food hazard	70.59	93.51	80.45
	food quality	78.99	74.60	76.74
	average	81.27	81.50	80.83

To illustrate our method, we do experiment based on mutual information and support vector machine (SVM) in the same dataset. The experimental result shows that our proposed method is superior in the complaint documents. The experimental result is shown in table V. We compare F1 value of all the experimental results and show them in Fig. 2. Frown Fig. 2 we can see obviously, the tf_2+(A+B+C) method our paper proposed has the highest F1 value.

TABLE III.
THE DIMENSION COMPARISON OF VECTORS IN TF1

	food hazard	food hygiene	food quality	sale service
initial vector	2420	2448	3490	2647
A (α =0.8)	1936	1958	2792	2117
B (term filtering)	802	859	1359	931
C (term expanding)	806	859	1369	933

TABLE IV.
THE DIMENSION COMPARISON OF VECTORS IN TF2

	food hazard	food hygiene	food quality	sale service
initial vector	2420	2448	3490	2647
A (α =0.8)	1936	1958	2792	2117
B (term filtering)	792	852	1348	932
C (term expanding)	805	856	1358	935

TABLE V.
MUTUAL INFORMATION AND SVM

	Precision(%)	Recall（%）	F1 (%)
sale service	87.67	71.91	79.01
food hygiene	71.79	60.87	65.88
food hazard	78.21	79.22	78.71
food quality	62.73	80.16	70.38
average	75.10	73.04	74.06

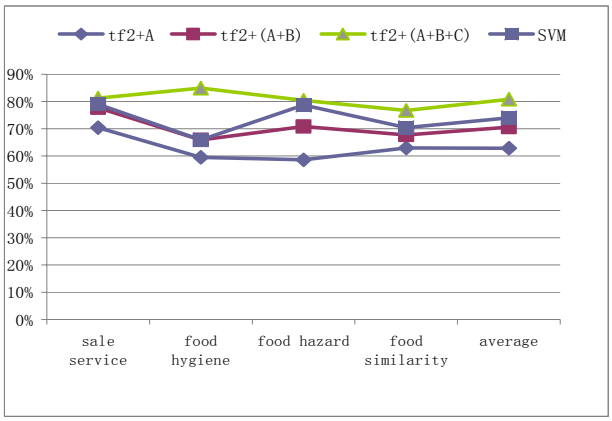


Figure 2. The comparison of F1 value for classification results

VII. CONCLUSION AND FUTURE WORK

In our paper, we introduce how to combine the food ontology and the word similarity computation method provided by HowNet. We also introduce how to expand topic term vectors and how to rectify the term weight for topic term vectors. Finally we realize semantic expansion of text. Furthermore, before term expansion, we use term filtering for independent topic features method to realize term reduction and drew out disturbing noise of classification. At the same time, we consider the importance difference between textual title and body, so we separately compute the term weights of textual title and body. The experiments show that our proposed method can achieve good classification effect.

In the future, we will further study how to perfect our ontology. We will expand the instances belong to classes of "abnormal" in ontology by ontology learning. In that case, after term expanding process, we will get more features closely related to topics. Then we will improve the precision and recall of classification, achieving better classification effect.

ACKNOWLEDGMENT

This work is supported by Jilin Provincial Science & Technology Department (No.20090303). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] ZELIKOVITZ S, MARQUEZ F. Transductive learning for short text classification problems using latent semantic indexing[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2005, 19(2): 143-163.

[2] de Buenaga Rodriguez, M.; Gomez-Hidalgo, J.; and Diaz-Agudo, B. 1997. Using WordNet to complement training information in text categorization. In Proc. of RANLP.

[3] LI Dun, MA Yong-tao, Guo Jian-li. Words semantic orientation classification based on HowNet. The Journal of China Universities of Posts and Telecommunications, 2009, 16(1): 106-110.

[4] Jianqiang Li , Yu Zhao , Bo Liu. Fully Automatic Text categorization by Exploiting WordNet[C].Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology, 2009, 1-12.

[5] D.Fensel, "Ontologies: Silver Bullet for Knowledge Management and e-Commerce", Springer Verlag, Berlin, 2000.

[6] B. Omelayenko., "learning og ontologies for the Web: the analysis of existent approaches", in the proceeding of the International Workshop on Web Dynamics, 2001.

[7] OWL Web Ontology Language, viewed March 2008 http://www.w3.org/TR/owl-features.

[8] Zheng Hai-Tao, Borchert Charles and Kim Hong-Gee. GOClonto: An ontological clustering approach for conceptualizing PubMed abstracts, Journal of Biomedical Informatics 2010;43(1), 31-40.

[9] Song Mu-Hee, Lim Soo-Yeon, Kang Dong-Jin, et a1. Automatic Classification of Web Pages based on the concept of Domain Ontology[C]. In: Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05), IEEE Computer Science, 2005. 645-651.

[10] Jun Fang, Lei Guo, XiaoDong Wang and Ning Yang "Ontology-Based Automatic Classification and Ranking for Web Documents" Fourth International Conference on Fuzzy Systems and Knowledge Discovery -FSKD -2007.

[11] Alexander Maedche and Ste_en Staab "Mining Ontologies from Text" LNAI 1937, pp. 189-202, 2000. Springer-Verlag Berlin Heidelberg, 2000.

[12] P. Sccuy, G.W.Mineanu "Beyoned TFIDF weighting for text Categorization in the Vector Space Model", 2003.

[13] E. Youn, M. K. Jeong , "Class dependent feature scaling ethod using naive Bayes classifier for text datamining", Pattern ecognition Letters , 2009.

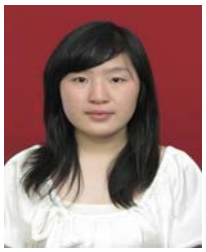
[14] Aurangzeb khan, Baharum Baharudin, Khairullah khan. Efficient Feature Selection and Domain Relevance Term Weighting Method for Document Classification[C]. In: 2010 Second International Conference on Computer Engineering and Applications, 2010. 398-403.

[15] Rossitza Setchi, Qian Tang, Ivan Stankov. Semantic-based information retrieval in support of concept design. Advanced Engineering Informatics. vol.25, pp.131-146, 2011.

[16] Jesús Oliva, José Ignacio Serrano, María Dolores del Castillo, ángel Iglesias. SyMSS: A syntax-based measure for short-text semantic similarity. Data&Knowledge Engineering.vol.70, pp.390-405,2011

XiQuan Yang JiLin Province, China. Birthdate: December, 1963. is Communication and Communication System Ph.M., graduated from Dept. Communication and Communication System JiLin University. And reasearch interests on Semantic Web and Ontology, Data Mining, Machine Learning.

He is a associate professor of Dept. Computer Science and Information Technology Northeast Normal University.



CaiFeng Zou JiLin Province, China. Birthdate: February, 1987. is Computer Software and Theory Ph.M., graduated from Dept. Computer Science and Information Technology Northeast Normal University. And reasearch interests on Semantic Web and Ontology, Data Mining.

She is a student of Dept. Computer Science and Information Technology Northeast Normal University.

Lin Yue JiLin Province, China. Birthdate: August, 1985. is Computer Application Techonology Ph.M., graduated from Dept. Computer Science and Information Technology Northeast Normal University. And reasearch interests on Semantic Web and Ontology, Data Mining.

She is a student of Dept. Computer Science and Information Technology Northeast Normal University.