International Journal of Computer Theory and Engineering, Vol. 1, No. 4, October, 2009
1793-8201

# Research on Correlative Techniques of Building Specific Topic Lexicon

Shouning QU, Jian Lu and Jing Li

***Abstract*—for information extraction and topic classification, this paper extracts topic words from the classified documents, and builds the topic lexicon according to the topic. Topic words are extracted from each document by pretreating the document and using the TF-IDF weight formula. The topic words are extracted by the size of weight proportionally. After processing each document uninterruptedly, the topic lexicon is built according to the topic. Experiments prove that it has good accuracy to extract topic words. The topic lexicon is easy to build, and it satisfies the needs of word segmentation of all kinds of documents. It is a new method in information extraction and text classification.**

***Index Terms*—text classification, TF-IDF, topic lexicon, topic words**

## I. INTRODUCTION

Topic words [1] extraction is the foundation of text classification, and Chinese word segmentation [2] is the first step, if we want to acquire the topic words of the documents. By the complexity of Chinese, the mechanical word segmentation techniques that base on lexicon are used widely. Therefore, the lexicon good or not is the key that whether can we extract the topic words effectively.

Currently, topic words extraction mainly uses existing source of knowledge, such as Wordnet [3] in English and Hownet [4] in Chinese. All sorts of word segmentation systems have been adopted to meet the need of word segmentation, such as Tsing Hua' s SEG [5]system and Beijing University of Aeronautics and Astronautics's CDWS system. They perform well in deal with words that usually use in common documents, but perform badly in deal with technical terms [6,7]. The paper just extracts the words [8] that relate to the topic rather than all the ones. So, it becomes a significant research projects to build the lexicon that meets the need discussed above and divided by topic. The methods that build topic lexicon mainly use weight calculation that bases on documents frequency now. Its maximum defect is that it needs huge training documents which waste plenty of time. In this paper, on the basis of weight calculation, an improved using method is adopted, which reduces the number of training documents largely. At the same time, semantic comprehension is added to make the topic lexicon more reasonable. Building topic lexicon is not only a great help for the information extraction and text classification but also a great help for enhancement the efficiency of topic words extraction obviously.

## II. RESEARCH OF RELEVANT TECHNIQUES

### *A. Text feature representation and Vector Space Model*

The most common method to represent the feature of text is called VSM[9](Vector Space Model), which was proposed by professor Salton. In this model, documents are mapped into term vector space[10]. Each text is represented feature vector, $V(d)=\{t_1,w_1(d)\ldots\ldots;t_i,w_i(d);\ldots\ldots t_n,w_n(d)\}$, $t_i$ is the term , and every term $t_i$ has a weight $w_i(d)$,which is assigned a value according to the importance. Thus, the document is represented as N-dimensional vector.

### *B. Relativity of the words*

Word is the minimum unit of text topic expression. The meaning of a sentence is expressed by forward-backward correlation of words. Mutual information [11] usually measure the relativity between words, it represents the size of information, when the two things co-occurrence. The formula of mutual information of words a and b is shown as: (1)

$$MI(a,b)=\log_2\frac{p(a,b)}{p(a)p(b)} \quad (1)$$

In the formula, P(a) is the frequency of a in the document. P(a,b) is the frequency of a and b co-occurrence in the document. The bigger the MI(a,b) is, the stronger the relativity between a and b, and they should be combined together.

### *C. Extracting topic words by the method of TF-IDF*

The text is composed of huge number of words, and the dimensions of vector of the text are enormous too. But the possibility to be the topic words is different greatly. The topic words are that they are the most representative feature of the text. In order to get them, the topic words should be sorted by the size of weight that is calculated by the weight calculation formula. Then they are extracted proportionally.

The classic methods to calculate the weight is the TF-IDF [12] formula. TF (term frequency) is the frequency of the words in a document. It reflects the distribution of a feature in a document. IDF (inverse document frequency) is the reciprocal that the words occur in the documents. It reflects the distribution of a feature in all the documents. The thoughts [13] of the methods are that the topic words of a document should have a higher word frequency in this document and have a lower word frequency in other documents. The formula of weight calculation [14] is shown as (2)

$$weight(t_i,d)=\frac{tf(t_i,d)\times\log(N/n_{ti}+a)}{\sqrt{\sum_{t_i\in d}[tf(t_i,d)\times\log(N/n_{ti}+a)]^2}} \quad (2)$$

In the formula, $tf(t_i, d)$ is the frequency of the key words $t_i$ in document $d$, $n_{ti}$ is the number of documents that has key words $t_i$, N is the number of all documents. The letter a is a constant, usually taking the value 0.01. Denominator is the normalization factor.

## III. Establishment of Topic Lexicon

Traditional methods to build topic lexicon analyze the word frequency and the inverse documents frequency in huge number of documents, getting the topic words. Then the topic words are added to the topic lexicon, and the topic lexicon is built. In this paper, each document that extracts topic words is regarded as a whole. Text frequency in the document and inverse document frequency in all documents is calculated firstly, and then the weight order is calculated. In the end, topic words are extracted proportionally. On the basis of that, to solve the defects of the TF-IDF weight formula, we add semantic comprehension to enhance the accuracy of topic words extraction. Repeating the above steps constantly, we can build a topic lexicon. Using these methods is easy to build the topic lexicon in a smaller set of training documents, and has better accuracy. At the same time, it is a process that perfects the lexicon constantly, and can perform well in word segmentation.

### *A. Extraction of topic words*

a. The processing of foreground document

Firstly, the foreground document that the paper extracts the topic words is processed [15]. Word segmentation is done, and the stop words that can not express the topic of a document are discarded [16], such as the words "的","了","啊", and so on. Secondly, the paper represents the text by VSM model. Finally, the word frequency [17] is calculated in the document.

b. The processing of background documents

The background documents are that they belong to different domains comparing with foreground documents. Word segmentation is done, and the stop words are discarded. Then the documents are represented by VSM model. Taking each document as a unit, counts the word frequency that occurs in all documents.

In the end, calculating the weight uses TF-IDF formula, and extracts the topic words proportionally. The faults of traditional method of TF-IDF are that they only calculate the word frequency in different domains without calculating the relationship among the words and the degree of importance that composes a document of the different parts. So, the paper goes on analyzing the foreground document by adding to semantic comprehension to make up the faults.

### *B. Word co-ccurrence of a sentence*

In Chinese language, the meanings of a sentence are expressed by the words that compose a sentence and the relationship among the words. The most direct relationship among the words is the word co-occurrence [18]. It shows the times that two words co-occurrence in a sentence. The bigger the word co-occurrence is, the more important the word. Because it is a word that often modifies other words or is modified by other words.

In addition, in Chinese scientific literature, it is usually composed of title, abstract, keywords, text and paragraph title. The importance of them to express the topic of a document is different largely. The title, the abstract and the paragraph title is the summary of the article, so they are more important than others. We take them which are pretreated firstly and the keywords to calculate the word co-occurrence, and the words that satisfy the threshold value are added to the topic lexicon. The formula of word co-occurrence is shown as (3).

$$occurrence_{ij} = \frac{f_{ij}}{f_i + f_j - f_{ij}} \quad (3)$$

In the formula, $f_{ij}$ is the times of the word $t_i$ and the word $t_j$ co-occurrence in a sentence. $f_i$ is the times of the word $t_i$ occurrence. $f_j$ is the times of the word $t_j$ occurrence.

### *C. The connection of compound word*

As the documents are processed by word segmentation that bases on lexicon, some compound words that have special meanings may be segmented to simple one. For example, the word"世界贸易组织" may be segmented by"世界", "贸易", "组织". They affect the accuracy of extracting the topic words. The compound word is the one that is composed of m ($2 \le m \le x$)(X is the maximum number of building the compound word)simple words that are adjacent[19], and it has lower word frequency but stronger ability to distribute the topic of the text.

The paper uses the deformation formula of mutual information to calculate the degree of connection of the m adjacent simple words. If the result is larger than the threshold value, It can be confirmed as a compound word[20].The formula of compound word connection is shown as (4).

$$MI(c_1, c_2, \cdots, c_x) = \log_2 \frac{p(c_1, c_2, \cdots, c_x)}{p(c_1)p(c_2)\cdots P(c_x)} = \log_2 \frac{freq(c_1, c_2, \cdots, c_x) * N^{x-1}}{freq(c_1)\, freq(c_2)\cdots freq(c_x)} \quad (4)$$

In the formula, $freq(c_w)/N$ is the relative word frequency of word $c_w$ in the foreground document. N is the number of words in the foreground document.

### *D. The processing of synonym*

Firstly, a table of synonym is built, which records the corresponding relationship of the synonym, and the acronym that used commonly. After the topic lexicon is built, the table of synonym is searched to add their synonym and acronym to the topic lexicon. The flow chart of building specific topic lexicon is shown as Fig. Ⅰ.

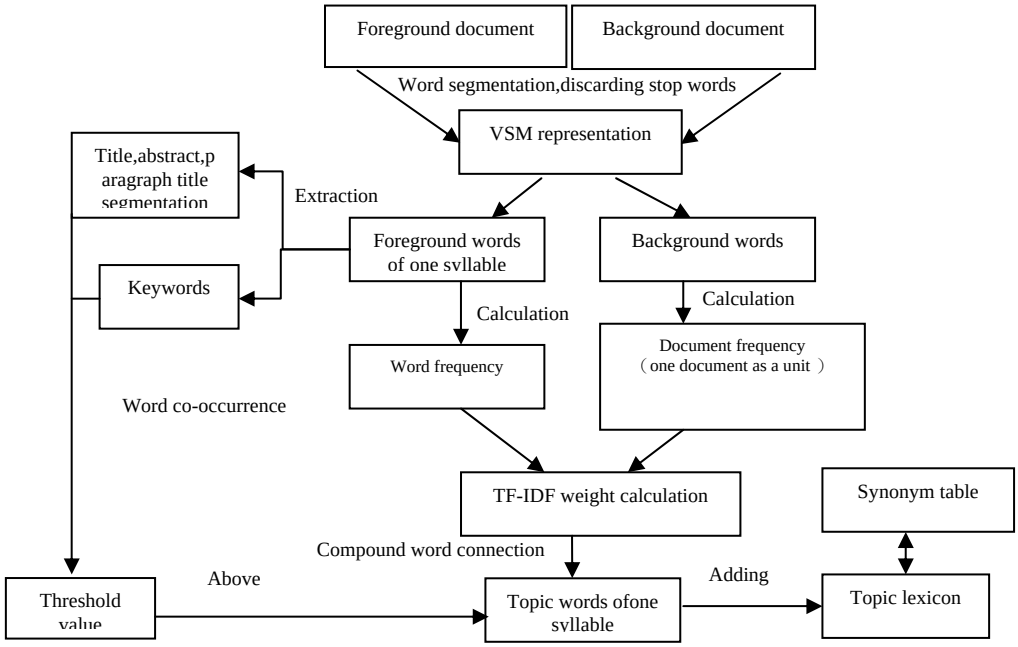


Fig. 1 flow chart of building

## IV. The Experiment and Result of Building Topic Lexicon

The experiment is done on the platform of windows xp sp2 operating system, and adopts NetBeans 6.1 integrated development environment which is based on java techniques to program the software.

### *A. The methods of the experiment*

Take to build topic lexicon of sports as an example. One hundred economic articles, one hundred educational articles, one hundred historical articles, one hundred martial articles and one hundred political articles are chosen from Fudan text classification corpus as the background documents. The average length is one hundred and ten words. Four hundred sports articles are chosen as the foreground documents. Each sports article is processed to extract topic words. Then the topic words are added to the sports lexicon. After finishing the work, a sports lexicon is built.

### *B. The experiment and analysis of topic words extraction*

Choosing an article named *The brief history of Olympic football* from the foreground documents. There are two hundred and fourteen words in all, after the pretreating, five percentage words are chosen as the topic words of this article. At the same time, the words that have the same weight as the dividing point of threshold value are added into lexicon. Then word co-occurrence calculation and compound word connection is done. Finally the topic words are shown as (6)

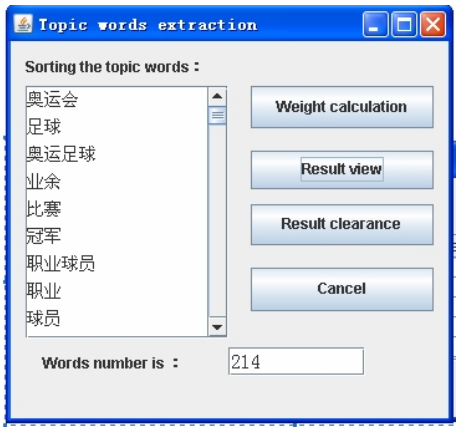


Fig.2 Topic words generation

While there are no objective evaluation methods in topic words classification, the paper uses the methods of artificial judgment. In this experiments, only the words "业余", "职业", don't belong to the sports words in the twelve topic words that are extracted.

### *C. Evaluation standard*

The words are divided into topic words and non-topic words. The formula of accuracy is: Accuracy=topic words/all the words. The result is gotten as following in Table Ⅰ and Table Ⅱ.

TABLE Ⅰ 300 BACKGROUND DOCUMENTS

| | | | | |
|---|---|---|---|---|
| Sports articles numbers | 100 | 200 | 300 | 400 |
| Topic words number | 1149 | 2327 | 3478 | 4631 |
| Accuracy | 83.72% | 83.71% | 84.12% | 83.74% |

TABLE Ⅱ 500 background documents

| | | | | |
|---|---|---|---|---|
| Sports articles numbers | 100 | 200 | 300 | 400 |
| Topic words number | 1149 | 2327 | 3478 | 4631 |
| Accuracy | 84.16% | 84.21% | 84.17% | 84.19% |

### D. *Results*

In 300 background documents, the accuracy of extraction words reposeful, though the number of topic words is different. In 500 background documents, the accuracy of extraction words increased a little. As with the increment of background documents, the document frequency of common words increases too. When calculating the weight, the technical terms will be chosen easily.

## V. THE STRUCTURE OF THE TOPIC LEXICON


Fig. 3 Structure of topic lexicon

In the table, the attribute of wordID is the exclusive number of the topic words. The attribute of words is the topic words. |The attribute of useFrequency is the frequency of this topic words in the word segmentation in future. The attribute of addFrequency is the frequency of the topic words that is added from the training documents. We can clear up the topic lexicon according to the attribute of addFrequency and useFrequency in future to make it more reasonable.

## VI. CONCLUSION

On the basis of traditional TF-IDF methods, the paper improves the using methods. Semantic comprehension is added to solve its defects too. The methods that build topic lexicon are easy and have good accuracy. The topic lexicon will be an effective tool in information extraction and topic words extraction of professional documents in future.

## REFERENCES

[1] Kai Kang, Kunhui LIN, Changle Zhou. "New text categorization method based on the frequency of topic words". Journal, Computer Applications, 2006,Vol. 26,NO.8, pp.1993-1995.
[2] Shouning Qu, Qin Wang. "Intelligent Question Answering System Based on DataMining". Journal, J.of ZhengzhouUniv. (Nat.Sci.Ed.), 2007, Vol.6, NO.1, pp.50-54.
[3] Jian Zhang, Chunping Li. "WordNet-based Concept Vector Space Model for Text Classification". Journal, Computer Engineering and Application, 2006,Vol.4, NO.5, pp.174-178.
[4] Xuexia Gao."Research of Word similarity Model Based on HowNet". Journal, Software Guide,2008,Vol.7, NO.4,pp.30-32.
[5] Haiyan Zhang. Research and implementation of Chinese automatic classification based on word segmentation[D]. University of Hu Nan,2002.
[6] Jiaheng Zheng, Jiaoli Lu."Study of an Improved Keywords Distillation Method". Journal,Computer Engineering, 2005,Vol.31,NO.18, pp. 194-196.
[7] Fei Liu, Xuanjing Huang, Lide Wu. "Approach for Extracting Thematic Terms Based on Association Rules". Journal, Computer Engineering, 2008,Vol.34,NO.7, pp.81-83.
[8] Jiaheng Zheng, Yongping Du,Changyu Liu." The Initial Research on the Method of Dynamically Acquiring Domain-specific Lexicals Based on Corpus". Journal, Computer Engineering ,2002,Vol.28, NO.5, pp.64-66.
[9] Salton G,Lesk M E. "Computer evaluation of indexing and text processing". Journal, Association for Computing Machinery,1968, Vol.15, NO.1, pp.8-36.
[10] Shouning Qu, Sujuan Wang, YanZou. Research and Design of Intelligent Question Answering System[C]. 2008 International Conference on Computer and Electrical Engineering. 2008, pp.711-714.
[11] Li Liu, Zhongshi He." Term selection and weighting approach based on key words in text categorization". Journal, Computer Engineering and Design, 2006,Vol.27, NO.6, pp.934-936.
[12] Salton G, Buckley C. "Term-weighting approaches in automatic text retrieval", Journal, Information Processing & Management, 1988, Vol.24, NO.5, pp: 513-523 .
[13] Jie Luo,Li Chen, Delin Xia,et. "Research on Fast Text Classifier Based on New Keywords Extraction Method". Journal, Application Research Of Computers.2006, Vol.32, NO.4, pp.32-34.
[14] Gang Yu, Yuehua Chen, Zhengyu Zhu, et. "Text feature description based on word co-occurrence". Journal, Computer Engineering and Design,2005, Vol.26,NO.8, pp.2180-2182.
[15] Jiong Chen, Yongkui Zhang. "Novel Chinese text subject extraction method based on word clustering". Journal, Computer Applications, 2005,Vol.25, NO.4, pp.754-756.
[16] CHANG H-C,HSU C-C, "Using topic keyword clusters for automatic document clustering". Journal, IEEE Transactions on Information and Systems, 2005,Vol.88, NO.8, pp.1852-1860.
[17] Qiwen Zhang, Ming Li. "study and implementation of approach to automatic extraction of text topic" Journal, Computer Engineering and Design,2006,Vol.27,NO.15, pp.2744-2766.
[18] Yinghua Ma, Yongcheng Wang,Guiyang Su,et. "Novel Chinese Text Subject Extraction Method Based on Character Co-occurrence" Journal, Journal of Computer Research and Development,2003,Vol.40, NO.6, pp.874-878.
[19] Song Lu,Shuo Bai." Quantitative Analysis of Context Field in Natural Language Processing". Journal, China Journal of Computers, 2001, Vol.24, NO.7, pp.742-747.
[20] Liao Hao, Zhishu Li, Qiuye,Wang,et. "Text feature word selection based on relationship between words". Journal, Computer Applications, 2007, Vol.27, NO.12, pp.3009-3012.

**Shouning Qu** (1962-), male, borned in Yantai of Shandong province of China. Professor of school of Information Science and Engineering University of Jinan of China.
**Jian Lu**(1983-) male, borned in Jinan of Shandong province of China. The graduate student of University of Jinan of China.
**Jian Lu**(1984-) female, borned in Jinan of Shandong province of China. The graduate student of University of Jinan of China.