Journal of Theoretical and Applied Information Technology
15th October 2012. Vol. 44 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NEW METHOD FOR CHINESE TEXT CONTENT IDENTIFICATION

[1]LI WEIWEI, [2]ZHANG TAO, [3]LIN WEIMIN, [4]DENG SONG, [5]SHI JIAN, [6]WANG CHEN
[1-6] China Electric Power Research Institute (Nanjing), Jiangsu Province, 210003, China
E-mail: [1]liweiwei@epri.sgcc.com.cn , [2]zhangtao@epri.sgcc.com.cn ,
[3]linweimin@epri.sgcc.com.cn , [4]dengsong@epri.sgcc.com.cn ,
[5]shijian2@epri.sgcc.com.cn , [6]wangchen2@epri.sgcc.com.cn

ABSTRACT

This article designed and implemented a method to identify the sensitive data in Chinese text. It can be used in data leakage prevention. There are two main innovation of this paper. One innovation is a method of preprocessing sensitive data based on statistical and the other is a method of determining the threshold based on self-learning. Experiments prove that the method is simple and practical.

Keywords: *Sensitive Data, Content Identification, Data Leakage Prevention*

1. INTRODUCTION

There are three main algorithms for text content identification: one is a classification algorithm based on probability and information theory such as Naive Bayes algorithm. Another is based on the TFIDF. The third is based on learning knowledge, such as support vector machine. The method of pretreatment of content identification was complex and lack of flexible method to determine thresholds in the past.

This paper proposed and implemented a method of identification of sensitive data based on Chinese text content. Through learning the sensitive text library and known-classification text library, we determined the threshold of the sensitive data. We analyzed the unknown-classification text and formed a feature vector, then calculated the similarity value between this vector and the feature vector of sensitive data by cosine formula. Finally, we compared the similarity value with the threshold to judge whether the text is sensitive or not. This article described a detailed account of the process and experiments showed that the method was simple, practical and had a high accuracy rate.

The content of this paper is divided into four main sections. The first chapter is the introduction, which analyzed the background and introduced the content of research and the structure of this article. The second chapter is architecture, which introduced the realization of architecture. The third chapter is functional components, which introduced the details of process of pretreatment, text recognition and determination of the threshold. The fourth chapter is the experiment, which can verify the effectiveness of this method.

2. ARCHITECTURE

Generally, the process of text content identification can be divided into the following steps: first, we created the data sets, which including the training set and test set. Then, we established the model to show the text and selected the feature. Then, we established the classifier by learning from training set. Finally, we did the experiment and evaluated the performance.

In this article, the data set contained the training set and test set. The training set included sensitive text library and the known classification text library. The former was used to form classifier by learning. The latter was used to form threshold. The latter library contained two small libraries. One of the small libraries contained a certain number of sensitive texts and the other contained the same number of safe texts.

In this article, the text was mainly expressed through the way of vector space model (VSM) [6], which described the Chinese text content in the form of vector. We proposed a method of identification of sensitive data based on Chinese text content. The main architecture was shown as Figure.1.

Journal of Theoretical and Applied Information Technology
15th October 2012. Vol. 44 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

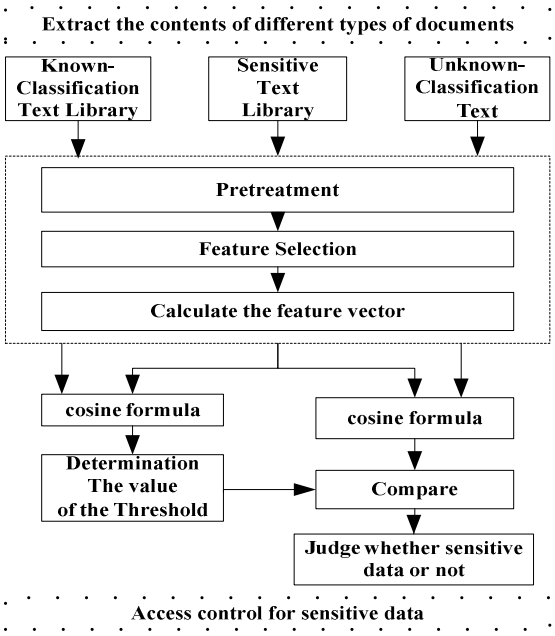

Fig.1 The Architecture Of Method To Identify Sensitive Data Based On Chinese Text Content

The main steps were shown as follow:

Step 1: The sensitive text library formed a feature vector after pretreatment, feature selection and vector space weight calculation.

Step 2: The known-classification text formed a feature vector after pretreatment, feature selection and vector space weight calculation. Then we calculated the similarity between this vector and the feature vector of sensitive data by cosine formula. By gathering statistics, determine the threshold of sensitive data.

Step 3: The unknown-classification text formed a feature vector after pretreatment, feature selection and vector space weight calculation. Then we calculated the similarity between this vector and the feature vector of sensitive data by cosine formula. We compared the value of similarity with the threshold of sensitive data to judge whether the text is sensitive or not.

3. FUNCTIONAL COMPONENTS

3.1 Pretreatment

The first step to identify sensitive data based on Chinese text content is pretreatment. Document collection which expressed as $T\_pre = \{T_1, T_2, \ldots\ldots, T_i\}$ was divided into individual phrases and identified the parts of speech by the interface of ICTCLAS and compiled statistics of word length and word frequency. For example, nouns was shown as (n), verb was shown as (v), adjective was shown as (a) and other.

Text Ti was shown as follow after pretreatment:

$$T_i = \{(a_{i1}, l_{i1}, p_{i1}), (a_{i2}, l_{i2}, p_{i2}), \ldots\ldots, (a_{in}, l_{in}, p_{in})\}$$

Where, T_i is a Chinese text, a_{in} is the phrase which is got by the interface of ICTCLAS, l_{in} is the length of a_{in} and p_{in} is the parts of speech of a_{in}.

3.2 Feature Selection

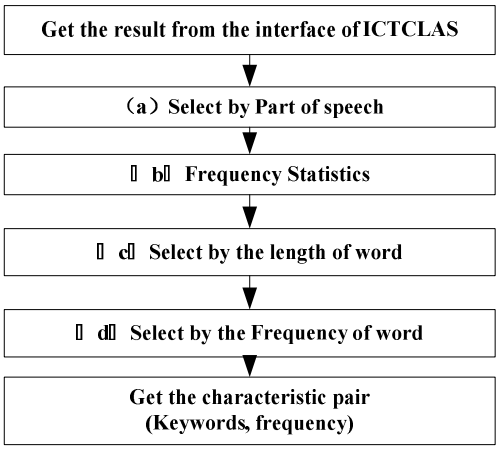

Fig.2 The Process Of Feature Selection

In the process of learning and identifying the Chinese text library, if all the parts of speech of the words were regarded as keyword, this would lead a lot of calculation and too much redundant information. Therefore, we extracted part of the result which get form the interface of ICTCLAS to reduce the dimension of vector space. This made keywords more representative and calculation more simple and effective.

The process of extraction is as follows:

(a) Select by part of speech

In the process of extraction, we only keep the keywords which can most strongly express the content of the article to eliminate redundancy. According to the statistical, the noun can express better than other parts of speech. Therefore, we only keep the noun. Ti was shown as Ta_i as follow after been selected by part of speech:

$$Ta_i = \{(a_{i1}, l_{i1}), (a_{i2}, l_{i2}), \ldots\ldots, (a_{in'}, l_{in'})\}$$

Where, Ta_i is the text which was extracted to keep noun and $(a_{in'}, l_{in'}) \in T_i$ in which $a_{in'}$ is noun.

(b) Frequency statistics

Journal of Theoretical and Applied Information Technology
15th October 2012. Vol. 44 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

We gathered statistics the frequency for each keyword and added the frequency of keyword on the basis of Ta_i and show as Tb_i as follow:

$$Tb_i = ((a_{i1}, l_{i1}, f_{i1}), (a_{i2}, l_{i2}, f_{i2}), \ldots\ldots, (a_{in'}, l_{in'}, f_{in'}))$$

Where, Tb_i is the text after gathering statistics of the frequency of keyword and $f_{in'}$ is the frequency of $a_{in'}$.

(c) Select by the length of word

In the Chinese text, the word has a stronger ability to express than the single word. We calculated the length of the keywords and delete single word, and then further expressed Tb_i as Tc_i as follow:

$$Tc_i = ((a_{i1}, f_{i1}), (a_{i2}, f_{i2}), \ldots\ldots, (a_{in'}, f_{in'}))$$

Where, Tc_i is the text after delete single word and a_{i1} is the keyword which is not a single word.

(d) Select by the frequency of the word

In the Chinese text, the keyword which only appears once is unrepresentative and accidental. Therefore, delete these keywords. The final characteristics was expressed as follow:

$$Td_i = ((a_{i1}, f_{i1}), (a_{i2}, f_{i2}), \ldots\ldots, (a_{in'}, f_{in'}))$$

Where, Td_i is the text after gathering statistics of the frequency of keyword and for which $f_{in'} \geq 1$.

3.3 Calculate The Feature Vector

(1) Calculate the feature vector of sensitive data.

After pretreatment and feature selection, the collection of sensitive text was shown as: $T = \{Td_1, Td_2, \ldots\ldots Td_n\}$. Where, Td_i was shown as follow:

$$Td_i = ((a_{i1}, f_{i1}), (a_{i2}, f_{i2}), \ldots\ldots, (a_{in'}, f_{in'}))$$

Where, Td_i is the text after gathering statistics of the frequency of keyword and for which $f_{in'} \geq 1$.

Calculate the weight of the keyword is the effective way to measure the characteristic value. Currently, the TF-IDF formula is the most widely used which based on statistical methods. This formula proved to be feasible and effective in a large number of actual uses. TF-IDF formula which is commonly used is expressed as follow:

$$d_{ij} = t_{ij} \times \log (N/n_j)$$

Where, t_{ij} is the number of occurrences of a_{ij} in T''_i, N is the total number of documents and n_j is the number of a_{ij} include in the document library.

After calculate the weight of the keyword, feature vector of sensitive data library is expressed as:

$$V = \begin{pmatrix} (a_{11}, d_{11}), (a_{12}, d_{12}), \ldots\ldots, (a_{1m}, d_{1m}), \\ \ldots\ldots, \\ (a_{n1}, d_{n1}), (a_{n2}, d_{n2}), \ldots\ldots, (a_{nm}, d_{nm}) \end{pmatrix}$$

Where, d_{nm} is the weight of a_{nm} calculated by TF-IDF formula for sensitive data library. Simply expressed as:

$$V = (d_{11}, d_{12}, \ldots\ldots, d_{1m}, \ldots\ldots, d_{n1}, d_{n2}, \ldots\ldots, d_{nm})$$

(2) Calculate the feature vector of known-classification text and unknown-classification text in library.

According to the sensitive data feature vector V, we calculated the weight of the keywords a_{ij} respectively and get the feature vector as follows:

$$V' = \begin{pmatrix} (a_{11}, d'_{11}), (a_{12}, d'_{12}), \ldots\ldots, (a_{1m}, d'_{1m}), \\ \ldots\ldots, \\ (a_{n1}, d'_{n1}), (a_{n2}, d'_{n2}), \ldots\ldots, (a_{nm}, d'_{nm}) \end{pmatrix}$$

Where, a_{mn} in V' is equal to a_{nm} in V and d'_{nm} is the weight of a_{nm} calculated by TF-IDF formula for known-classification text. Simply expressed as:

$$V' = \begin{pmatrix} d'_{11}, d'_{12}, \ldots\ldots, d'_{1m}, \\ \ldots\ldots, \\ d'_{n1}, d'_{n2}, \ldots\ldots, d'_{nm} \end{pmatrix}$$

We used the same method to get the feature vector of unknown classification document and expressed as:

$$V'' = \begin{pmatrix} d''_{11}, d''_{12}, \ldots\ldots, d''_{1m}, \\ \ldots\ldots, \\ d''_{n1}, d''_{n2}, \ldots\ldots, d''_{nm} \end{pmatrix}$$

3.4 Calculation Of Cosine

We used the cosine formula to calculate the similarity between two feature vectors, which is shown as follows:

$$\cos\theta = \frac{V_1 \cdot V_2}{\|V_1\| \|V_2\|}.$$

Where, V_1 and V_2 is feature vector of the document; $V_1 \cdot V_2$ dot product of standard vector and define as $\sum_{i=1}^{t} V_{1i} V_{2i}$ and the norm in the denominator $\|V_1\|$ is defined as $\sqrt{V_1 \cdot V_1}$.

Journal of Theoretical and Applied Information Technology
15th October 2012. Vol. 44 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.5 Determination Of The Threshold

The threshold of sensitive data was important to determine whether the document was sensitive or not by comparing it with the results of the cosine similarity calculation.

The process is shown as follows:

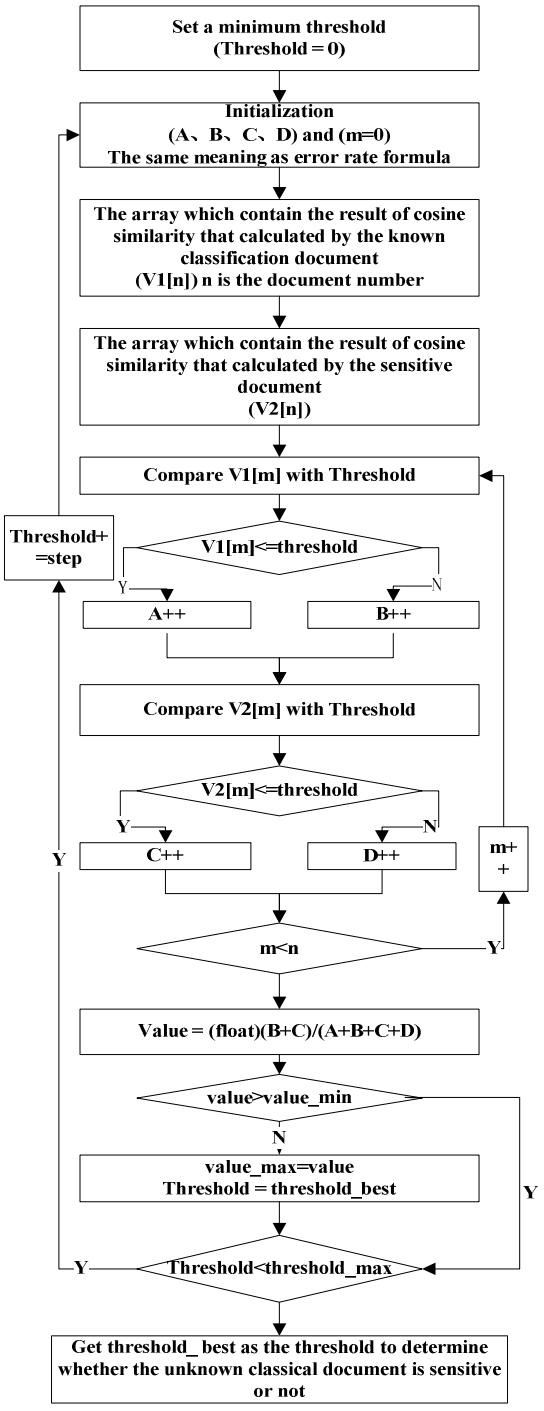

Figure 3 The Code To Determine The Threshold

In this article, we learned from the known classification documents to determine the threshold.

We collected hundred and fifty security documents and the same number of sensitive documents. And then, we calculated cosine similarity with the sensitive feature vector and got the value respectively. By setting the same step interval to threshold, we got the threshold which had the lowest error rate. The threshold would be used to determine whether the unknown classification document was sensitive or not.

The error rate is shown as follows:

$$rate = \frac{(B+C)}{(A+B+C+D)}$$

Where, A is the number of documents that are correctly identified as secure document; B is the number of documents that are mistakenly identified as sensitive document; C is the number of documents that are mistakenly identified as security document and D is the number of documents that are correctly identified as sensitive documents.

4. FUNCTIONAL COMPONENTS

(1) The data sets

This system can achieve identification of sensitive data in different environments by changing the training set.

We created a sensitive data identification system about education and selected 1600 documents about education from SogouC.reduced.20061102 Corpus as sensitive documents library in the experiment. We selected 100 education-related documents and 100 documents about other area randomly from the corpus for the known-classification documents library.

(2) Pretreatment and feature selection

In the process of feature selection, statistics analysis shows that: the proportion of the remaining keyword is about 32% after selecting by the part of speech, about 28% after selecting by the word length, about 11% after selecting by the word frequency. The process of feature selection greatly reduces the redundancy of the keywords and simplifies the calculation.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(3) The process of calculating the feature vector

According to the keywords, we get the feature vector of sensitive data V by the TFIDF algorithm. By the same way, we got the known classification text feature vector V' and the unknown classification text feature vector V".

(4) The process of calculating cosine similarity value

We calculated the feature vector V' of the known classification documents respectively and calculated the similarity between V' and V. They were sorted after statistics in the table below in which the horizontal axis represented the text number and the vertical axis represented the cosine similarity value.

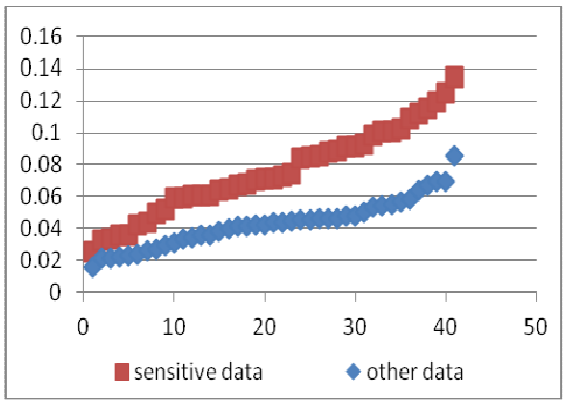

Table 1 Count Up Cosine Similarity Value

(5) The process of determining the threshold

From the bottom of the value to the top of the value, we added an interval length step by step and calculated the error rate of the threshold. Through learning the error rate of threshold, we got the threshold which has the lowest error rate. Through the experimental data, when the value of threshold is 0.060, we got the lowest error rate of 22.1%. The statistical results were illustrated below in which the horizontal axis represented the value of threshold and the vertical axis represented the error rate, recall rate and accuracy.

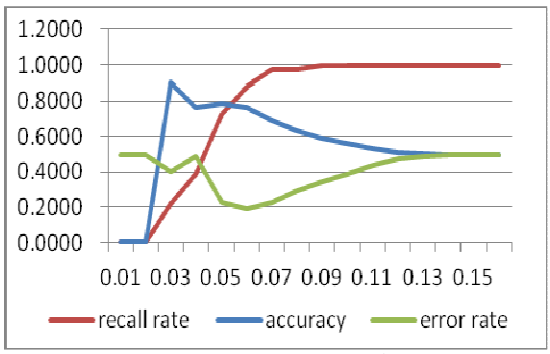

Table 2 Through learning the error rate of threshold get the threshold which has the lowest error rate

(6) The experimentation to identify the unknown classification document

We took forty unknown classified documents from the library and judged whether the document was sensitive or not by comparing cosine similarity value with the threshold. We pre-processed and analyzed the unknown classified documents library. And then, we get feature vectors based on the keywords of sensitive feature vectors. We identified the unknown classified document by comparing the cosine similarity value with threshold 0.061 which was the threshold with lowest error rate 16.52%. The experimental results show that this method is effective and feasible.

5. CONCLUSION

This article designed and implemented a method of preprocessing sensitive data based on statistical and a smart self-learning method to determine the threshold which was used to judge whether the document is sensitive or not. We described the methods and processes to achieve in detail and gave the results of the experiment.

6. ACKNOWLEDGEMENT

This work was supported in part by the information security technology research team of state grid corporation of china. In addition, I would like to offer special thanks to Zhang Tao, Lin Weimin and Deng Song for writing and experiment guidance.

REFRENCES:

[1] Lin Zhenbiao. Research on Key Technologies of File Network Leakage Prevention Based on Data stream Analysis [D]. The PLA Information Engineering University,2009.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[2] Yiming Yang. An Evaluation of Statistical Approaches to Text Categorization[J]. Information Retrieval, 1999,1(1-2) .

[3] Yuan Wensheng,Wang Xiaofeng. Research on Chinese Maritime Muti-class Text Classifier Based on Na ve Bayes[J]. computer and modernization, 2011,(05)

[4] Shi Congying, Xu Chaojun,Yang Xiaojiang. Study of TFIDF a lgor ithm[J] Journal of Computer App lications,2009,29:168-180.

[5] Zhang Xiaoyan, Li Qiang. The Summary of Text Classification Based on Suppor t Vector Machines,2008,28:344-345.

[6] Christopher J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998,2(2) .

[7] Pang Jianfeng , Bu Dongbo , Bai Shuo. Research and Implementation of Text Categorization System Based on VSM[J]. Application Research of Computers , 2001, (9) :23-26