A New Approach to Feature Selection for Text Categorization

Shoushan LI and Chengqing ZONG

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing 100080,China
{sshanli,cqzong}@nlpr.ia.ac.cn

***Abstract*-Text categorization (TC) is a problem of assigning a document into predefined classes. One of the most important issues in TC is feature selection. In this paper, we propose a new approach in feature selection called Strong Class Information Words (SCIW). Different from many existing feature selection methods, our method takes many kinds of information into account. Moreover, the method can easily use some implicit regularities of natural language. Our extensive experiments resulted in a good performance on precision by a linear classifier using SCIW feature selection method. The most attractive aspect of the classifier as a combining part in the categorization system is shown in our experiments and the combining system outperforms performances in comparison with conventional classifiers.**

***Keywords*-feature selection, linear classifier, text categorization, classifier combination**

I. INTRODUCTION

The task of TC is to automatically assign natural language texts with thematic categories from a predefined category set [8]. Because of the growing need for automatic processing of large amounts of information from text, the research on text categorization has become more and more popular.

A TC process possesses several characteristics different from other pattern recognition problems. The TC problems normally involve an extremely high dimensional feature space [10]. A standard procedure to reduce feature dimensionality is feature selection (FS). Many selection methods such as term strength (TS), document frequency (DF), mutual information (MI) and information gain (IG) have been applied to TC [13].

This work has been partially funded by the Natural Science Foundation of China under the grant 60121302, the China-France PRA project under the grant PRA SI02-05, and the Outstanding Overseas Chinese Scholars Fund of the Chinese Academy of Sciences (No.2003-1-1) as well.

Another open question for TC research is what classifier is suitable for a TC task. As we know, many standard machine learning techniques have applied to TC, such as Bayes classifiers, support vector machines (SVM), linear classifiers and K-nearest neighbor classifiers (KNN) [8] [12].

In our paper, we propose a new approach to feature selection, SCIW, mainly according to the features' distribution in classes. The experiments have shown that a linear classifier using this approach can achieve higher performance on precision.

The remainder of this paper is organized as follows. Section 2 briefly introduces the related work. Section 3 proposes the motivation of our work. Section 4 describes the new feature selection method in detail and the classifiers used in our experiments. Experimental results are presented and analyzed in Section 5. Finally, section 6 draws some conclusions and outlines the future work.

II. RELATED WORK

Feature selection is an important step in TC. Many methods have been proposed to solve this problem, such as document frequency (DF), Information Gain (IG), Mutual Information (MI), Category Term Descriptor (CTD), Class Discriminating Words (CDW), Term Strength (TS), odd-ratio, Term Strength (TS) [13] [15] [16] [2] [9] [3]. Among these measures, DF is the simplest method but ignores much other information such as the category information present in the training set. MI is also simple but biases towards low frequency features because of its ignorance of features' frequency [13]. The performance of CHI and IG is good, but their computation is expensive. CDW takes the features' distributions into account and achieves a good performance, but it has the same weakness as MI in that the document frequency information is not considered.

According to our analysis, some problems in feature selection may be summarized as the following aspects:

1) How to take full advantage of the statistic information such as category information, document frequency information?
2) How to find other kinds of new statistic information for better performance on TC?

3) How can the implicit regularities of natural language be used? By treating TC as a classical classification problem, standard feature selection measures mentioned above ignore the fact that texts are written in natural language, which means that they have many implicit regularities of natural language itself.

The recent work dealing with the FS problem has focused on the first problem and has significantly improved the performance. The method CTD utilizes the document frequency information in *IDF* as well as the category information in *ICF* [2]. However, simply multiplying the two types of information, $IDF * ICF$, is not sufficient for handing the problems of TC because the improvements in performance are not explicit in some training data. Based on multi-criteria, the methods in [9] and [3] select the combination features. In these two methods, some of the statistic information is used in parallel. But they do not separately consider the characteristic of each kind of information.

As to the second problem, the method CDW takes the feature discrimination into account and gains the best selecting effect among many selection methods in experiment [16]. The method ignores many other kinds of important information such as document frequency, and definitely cannot deal with some kind of TC task.

There are few papers reported on the third problem. One related work appears in [5], where only an agricultural term dictionary is used but the accuracy was clearly improved. One obvious weakness is that the work to obtain the term dictionary demands a lot of manual work.

III. MOTIVATIONS

The problem of feature selection can be examined in many perspectives. The two major problems are (1) How to search for the "best" features? (2) What should be used to determine best features, or what are the criteria for evaluation? [7] In order to answer these questions, we need a classifier that has learned from the training data and can be tested on the test data. An optimal classifier is the equivalent of direct table lookup, which can, in turn, be formulated in precise mathematical terms – the probabilistic theory of Bayesian analysis [11]. Adopting the Bayesian approach, the true class of a given object is considered as a random variable c taking values in the set $\{c_1, c_2, ..., c_m\}$. The initial uncertainty regarding the true class is expressed by the prior probabilities $P(c_i)$. Mathematically speaking, the table lookup criterion can be stated as selecting the class c_i with the greatest posterior probability for a given pattern of evidence x where c_i is chosen such that

$$P(c_i \mid x) > P(c_j \mid x) \quad \text{for all } i \neq j.$$

When it is assumed that features $t_1,....,t_n$, with conditional probabilities, are probabilistically independent.

$$P(c_i \mid x) = P(c_i \mid t_1, t_2, ..., t_n) = \frac{P(t_1, t_2, ..., t_n \mid c) P(c_i)}{P(t_1, t_2, ..., t_n)}$$

$$= \frac{\prod_{j=1}^{n} P(t_j \mid c_i) P(c_i)}{\prod_{j=1}^{n} P(t_j)} = \frac{1}{P(c_i)^{n-1}} \prod_{j=1}^{n} P(c_i \mid t_j).$$

This formula shows that the $P(c_i \mid x)$ is in direct proportion to $P(c_i \mid t_j)$. So, $P(c_i \mid t_j)$ plays an important role in classification. In our approach, we will use the features with high value of $P(c_i \mid t_j)$.

When the feature t_j has a high value of $P(c_i \mid t_j)$, we think the feature word have strong class information in TC. For example, the word "football" usually appears in the class "sports", it must have a high vale of $P(c_i \mid t_j)$, so it must be a word with strong class information. We propose a new approach to select the words with strong class information. We call this method Strong Class Information Words (SCIW). In fact, there are a lot of such words and many documents can easily be classified by only using the SCIW in TC.

IV. FEATURE SELECTION AND CLASSIFIERS

A. SCIW method

According to the motivations, our SCIW method consists of the following steps:

First, compute the distribution of the feature t in class C_i:

$$Distribute(t) = (P(C_1 \mid t), P(C_2 \mid t), ..., P(C_m \mid t)).$$

According to the Bayes theorem, $P(C_i \mid t) = \frac{P(t \mid C_i) P(C_i)}{P(t)}$, $P(t) = \sum_{i=1}^{m} P(C_i) P(t \mid C_i)$, and $P(t \mid C_i) = \frac{\sum_{k=1}^{d_i} tf(t_k)}{\sum_{j=1}^{|v_i|} \sum_{k=1}^{d_i} tf(t_{jk})}$, where $tf(t_{jk})$ represents that the word t's frequency in the document D_k in class C_i, d_i is total number of files in class C_i, and $|v_i|$ is the total number of words in class C_i

Then, the selection criterion is defined according to the contributions of the features. We define the SCIW value, $SCIW(t) = \max\{P(C_i \mid t)\}$, and two thresholds, $TS (0 \le TS \le 1)$ which stands for the threshold of the SCIW value, and $DS (0 \le DS \le 1)$ which stands for the threshold of the document frequency.

The features which satisfy the conditions bellow will be selected:

$$\begin{cases} SCIW(t) \geq TS & (1) \\ \dfrac{d_{class(t)}(t)}{d_{class(t)}} \geq DS & (2) \end{cases}.$$

Features that satisfy the first condition indicate that they usually appear only in one single class when the TS is assigned a high value. The second condition is used for filtering the words that have only a low frequency in documents. These words can be the noisy.

B. Classifiers

In our algorithm, linear classifier is used. Linear classifier is a simple approach to classification [6]. The main idea of the linear classifier is to construct a feature vector as one representative for each class. For each class C_i, a prototype vector $G_I = (g_{i,1}, g_{i,2}, ..., g_{i,n})$, is computed, where $g_{i,j}$ corresponds to the weight of the j_{th} feature, which is usually trained by the training corpus. Then, the similarity between D_i and G_i is computed to determine which class the document belongs. The similarity measurement can be cosine or inner product measure. After obtaining the features selected by our FS method, vector G_i is easily expressed as $(g_{i,1}, g_{i,2}, ..., g_{i,n})$ where the weight $g_{i,j} = P(C_i \mid t_{ji})$. When comes a test document $D_i(w_{i,1}, ..., w_{i,j}, ..., w_{i,n})$ where $w_{i,j}$ is the frequency of feature t_j in document D_i, the similarity will be computed with the following formula:

$$Simi(D_i, C_j) = \sum_{k=1}^{|s_i|} w_{i,k} \bullet g_{i,k}.$$

The document D_i will be classified in the class I, such that

$$I = \arg\max_j \{Simi(D_i, C_j\},$$

and the maximum value expresses as $MaxSimi$:

$$MaxSimi = Simi(D_i, C_I).$$

To achieve a high precision, some control conditions on categorization results are proposed as follows:

$$\begin{cases} MaxSimi > 3 \\ MaxSimi > \sum_{j=1}^{m} Simi(D_i, C_j) - 0.5 * MaxSimi \end{cases}.$$

We use another classifier k-nearest-neighbor (KNN) as the combining classifier and the baseline. The KNN algorithm is quite simple: given a test document, the system finds the k nearest neighbors among the training documents, and uses the categories of the k neighbors to weight the category candidates [12]. KNN classifier outperforms many other classifiers [4] [12].

V. EXPERIMENTS

Two different corpra are used to test the method. All the documents in the corpus are obtained form the Internet. Corpus 1 is composed of 1960 documents with about 10000 words and Corpus 2 is a large scale corpus of 23387 documents with about 50000 words. The distributions of the two samples are shown in TABLE 1:

TABLE 1
THE DISTRIBUTION OF EACH CORPUS

Categories in Corpus 1	Number of documents in training set	Number of documents in test set
economy	250	120
politics	175	82
computer	130	55
sports	300	282
education	150	64
law	200	152
Total	1205	755

Categories in Corpus 2	Number of documents in training set	Number of documents in test set
entertainment	3770	644
finance	3170	429
health	3360	392
news	2580	356
science	3530	378
sports	4350	428
Total	20760	2627

We tested our approach through the following experiments. In experiment 1, different threshold TS values of the SCIW are tested. In experiment 2, a special control condition is tested. In experiment 3, our method is used to combine the traditional classification method KNN, and a compare study is done to the method KNN.

Classification effectiveness has been evaluated in terms of the standard precision, recall, F1 and miroF1 measure, which are defined as:

$$R_i = \frac{\alpha_i}{\beta_i}, P_i = \frac{\alpha_i}{\gamma_i}, F1 = \frac{2P_iR_i}{P_i + R_i},$$

$$miroR = \frac{\sum_1^m \alpha_i}{\sum_1^m \beta_i}, microP = \frac{\sum_1^m \alpha_i}{\sum_1^m \gamma_i},$$

$$miroF1 = \frac{2 \times miroR \times microP}{miroR + miroP}.$$

Where α_i is the number of documents correctly classified by system, and β_i is the number of documents classified by system to category C_i, and γ_i is the number of documents form category $C_i (i = 1,2,.....,m)$.

A. Experiment 1

We first compare the impact of different TS values on classification performance. The experimental results indicate that the performance is best when the TS value is within the range, 0.5-0.6. The experiment results are showed in TABLE 2, where the TS values are 0.6 in corpus 1 and 0.5 in corpus 2:

TABLE 2
THE RESULT OF THE TEST OF THE TS VALUE

Categories in Corpus1	precision	recall	Number of features
economy	0.990	0.858	243
politics	0.952	0.487	97
computer	1.000	0.418	58
sports	1.000	0.943	236
education	0.971	0.516	69
law	1.000	0.592	150
Total	0.992	0.735	853

Categories in Corpus2	precision	recall	Number of features
entertainment	0.987	0.572	234
finance	0.980	0.562	281
health	0.977	0.853	412
news	0.790	0.244	78
science	0.905	0.429	168
sports	0.994	0.804	239
Total	0.962	0.585	1412

As shown in TABLE 2, the precision is especially high. This confirms our assumption that many documents can easily classified by SCIW. The low recall of some categories in Table 2 is not strange because they only use a small portion of the features. Moreover, the recall of one certain category is relative to the number of features selected in the category. For example, the number of the features selected in category "news" is comparatively small and the recall of this category is very low.

B. Experiment 2

According to our analysis of the features selected by our method, we find that there are many noisy words in the features which do not help to provide useful information for classification. Good control conditions on feature selection and thresholding can filter some of these words to enhance the precision or the recall. For example, in the features selected by our method, we find that the single character word such as "男 (men)" contains smaller class information than two character word "男子（men's）", so we use a more strict control condition to filter the single character, so as to improve the classifier's performance. For experiment 2, the single character words are assigned a DS value of 0.25. This resulted in a reducing the total number of features from 853 to 409 in corpus 1 and from 1412 to 1082 in corpus 2. The overall precision values remained similar but the total recall value improved form 0.735 to 0.818 in corpus 1 and from 0.585 to 0.619 in corpus 2.

The result confirms that using the implicit regularities of nature language will indeed make sense.

C. Experiment 3

Only using our algorithm usually can not accomplish the classification task because the value of recall is not satisfying. So we combine our method with other methods. In this experiment, the KNN method is combined with our linear classifier. The process can be simply described. We use the KNN(k=10) classifier with the feature selection method, IG, to classify the documents that can not classified by our method. The results are showed in TABLE 3:

TABLE 3
THE COMPARE RESULT BETWEEN THE COMBING CLASSIFIER AND OTHER METHODS

	KNN	SCIW+KNN
miroF1(corpus1)	0.904636	**0.952318**
miroF1(corpus2)	0.888466	**0.935668**

The results indicate that our method is effective for improving the performance of the whole classification system. It is proved to be a promising method.

VI. CONCLUSIONS AND FUTURE WORK

We have developed a novel textual document categorization method, utilizing the words with strong class information. Different from the previous work, our feature selection method has some advantages below:

1) Many kinds of statistic information are used in our method such as term frequency information, document frequency information, and term distributing information. Moreover, we consider the characteristic of each kind of information separately and these kinds of information are used in cascading.
2) Some implicit regularities of language can be considered to filter the noisy words, which is one contribution expressed in experiment 2.

Future work extending in our research includes several aspects. One aspect is to add control additions in feature

selection using more information such as text length information or the language knowledge. The second aspect is to combine more other classifiers in the state-of-the-art methods, such as SVM, to enhance the performance.

REFERENCES

[1] Anil,K. Jain, Robert P.W.Duin and Jianchang Mao. 2000. Statistical Pattern Recognition: A Review. IEEE Transaction on Pattern Analysis and Machine Intelligence. vol.22, no.1, pp.4-37..

[2] Bong,Chih How,Narayanan K. 2004. An Empirical Study of Feature Selection for Text Categorization based on Term Weightage. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence(WI'04).

[3] Dai, Liuling, Heyan Huang and Zhaoxiong Chen. 2004. A Comparative Study on Feature Selection in Chinese Text Categorization. Journal of Chinese Information Processing. Vol.18,no.1, pp.26-32.

[4] He, Ji, Ah-Hwee Tan, and Chew-Lim Tan. 2000. A Comparative Study on Chinese Text Categorization Methods. PRICAI 2000 Workshop on Text and Web Mining. pp.24-35.

[5] Horyu, D, Kiura, T and Ninomiya, S. 2004. Effect of Agricultural Term Dictionary in Text Categorization of Japanese Agricultural Documents. AFITF/WCCA joint Congress in Agriculture 2004. pp.326-329.

[6] Lewis, D.D., Schapire, R.E., Callan, J.P. and Papka, R. 1996. Training algorithms for linear text classifiers. In Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval(SIGIR'96). pp.298-306.

[7] Liu, Huan and Hiroshi Motoda. 1998. Feature selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers.London.

[8] Sebastiani，F. 2002. Machine Learning in Automated Text Categorization. ACM Computing Surveys. 34(1),pp.1-47.

[9] Son, Doan and Horiguchi Susumu.2004. An Efficient Feature Selection Using Multi-Criteria in Text Categorization. Proceedings of the Fourth International Conference on Hybrid intelligent Systems(HIS'04).

[10] Wai, Lam, and Yiqiu Han. 2003. Automatic Textual Document Categorization Based on Generalized Instance Sets and a Metamodle. IEEE Transaction on Pattern Analysis and Machine Intelligence.vol.25,no.5,pp.628-633.

[11] Weiss, S.M. and Kulikowski, C.A. 1991. Computer Systems That Learn. Morgan Kaufmann Publishers, San Mateo, California.

[12] Yang, Yiming, Liu X. 1999. A re-examination of text categorization methods. In Proceeding of SIGIR-99,22nd ACM International Conference on Research and Development in Information Retrieval 1999. pp.42-49

[13] Yang, Yiming. 1997. A Comparative Study on Feature Selection in Text Categorization. In Proceeding of the Fourteenth International Conference on Machine Learning(ICML'97),pp.412-420.

[14] Yang, Yiming. 1999. An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval Journal, 1/2.

[15] Zheng, Z., and R. Srihai. 2003. Optimally Combining Positive and Negative Features for Text Categorization. ICML 2003 Workshop.

[16] Zhou, Qian, Mingsheng Zhao and Hu Min. 2004. Study in Feature Selection in Chinese Text Categorization. Journal of Chinese Information Processing. vol.18,no.3,pp.17-23.