TELKOMNIKA, Vol.12, No.4, December 2014, pp. 905~910
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v12i4.536

Feature Selection Method Based on Improved Document Frequency

Wei Zheng[1*], Guohe Feng[2]
[1] College of Science, Hebei North University, Zhangjiakou 075000, Hebei, China
[2] School of Economics & Management, South China Normal University, Guangzhou 510006, Guangdong, China
*Corresponding author, e-mail: 43399078@qq.com

Abstract

Feature selection is an important part of the process of text classification, there is a direct impact on the quality of feature selection because of the evaluation function. Document frequency (DF) is one of several commonly methods used feature selection, its shortcomings is the lack of theoretical basis on function construction, itwill tend to select high-frequency words in selecting. To solve the problem, we put forward a improved algorithm named DFMcombined withclass distribution of characteristics and realize the algorithm with programming, DFM were compared with some feature selection method commonly used with experimental using support vector machine, as text classification .The results show that, when feature selection, the DFM methods performance is stable at work andis better than other methodsin classification results.

Keywords*: feature selection, document frequency, text classification*

1. Introduction

The classification technology is to assign automatically a new document into one or more predefined classes based on its contents. With the development of WWW, in recent years, text categorization (TC) has become one of the key techniques for handing and organizing text data, and the technology has got extensive use in rubbish mail filtering, classification for Web page and document. Therefore, it is very necessary and meaningful to study the key technology of text categorization for improving the speed and accuracy of categorization.Some classication algorithms used in TC are:Support Vector Machines,k-NearestNeighbor (kNN) and naive Bayes [1],[2]. A major difficulty of text categorization is the high dimensionality of the original feature space.Feature selection is an import method to reduce the amount of feature in text categorization, and its goals is improving classification effectiveness and computational efficiency. Currently, the feature selection method's principle of operation is that it will compute and score for each feature word using statistical knowledge, according to sort the feature words, then it select some feature whose score is higher to act final document feature. Some well-know methods are document frequency(DF), information gain (IG),expected cross entropy(ECE),the weight of evidence of text (WET), x^2 statistic (CHI) and so on [3],[4],[5], and it is highly desirable to reduce the feature space without the loss of classification accuracy.

The document frequency (DF) thresholding, the simplestmethod with the lowest cost in compution, has shown to behave well when compared to more methods, it can be reliably used instead of IG or CHI when the computation of these measures are too expensive. Aexperimentin literature [6] is carried out on the performance of feature extraction among DF,IG,WET and CHI, and the experiments result show that the DF method has its own advantages such as easy to realize, widely to use in Chinese text categorization and English text categorization, and the feature select performance is good while it compared with others feature selection methods.However, this method overlook the usefullow-frequency feature words in the category and no considering the contributions to each category ,so it is usually considered an empirical method to improve efficiency, not a principled criterion for selecting predictive features.

In this paper we propose a feature selection method based on improved document frequency (DF),named DFM, derived from the DF original definition. The DMF overcome the shortcomings of DF such as overlook the usefullow-frequency feature words in the category and

Received August 23, 2014; Revised October 28, 2014; Accepted November 15, 2014

no considering the contributions to each category. Experiments on Chinese text data collection collected by the Fudan University show the performance of MDF method .

The rest of this paper is organized as follows. Section 2 describes the term selection methods commonly used, and gives a improved document frequency method MDF, Section 3 discusses the classifier using in experiment to compare MDF with other text feature selection methods, and presents the experiment's results and analysis. In the last section, we give the conclusion and future work.

2. Research Method

In this section we summarize and reexamine the feature selection methods DF,IG,ECE and CHI which are commonly used in feature selection for text categorization, and we implement a new feature selection method of DFM (Document Frequency Modified), which it's evaluation function based on Document Frequency (DF) method.

2.1. Feature Selection Methods

The following definitions of DF, IG, ECE and CHI are taken from [7],[8], and They will be simply introduced.

1. Document Frequency (DF)

Document frequency is the number of documents in which a term occurs. In text categorization, according to the setting threshold, the term is retained or removed. DF is the simplest technique for feature reduction. It scales easily to very large corpora with an approximately linear computational complexity in the number of training documents [8].

2. Information gain (IG)

Information gain is used to measures the amount of information obtained for category by knowing the presence or absence of a term in category documents. Let $C = \{c_1, c_2, ...c_m\}$ be the train set of categories. The information gain of term t is defined as following:

$$IG(t) = -\sum_{i=1}^{n} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{n} p(c_i \mid t)\log p(c_i \mid t) p++ p(\bar{t})\sum_{i=1}^{n} p(c_i \mid \bar{t})\log p(c_i \mid \bar{t}) \quad (1)$$

All the feature terms are computed according to formula IG, whose information gain is less than some predetermined threshold are removed from the feature space.

3. Expected cross entropy (ECE)

Cross entropy, also known as the KL distance. It reflects the probability distribution of text topic class and in can computer the distance between specific term with text topic class under the condition of probability distribution .If the cross entropy of termis bigger, the effect on distribution of text topic class is bigger. The difference to information gain is consider the relation of word occurrence and categories, only calculating theterm appear in the text.

$$ECE(t) = P(t)\sum_{i=1}^{n} P(c_i \mid t)\log\frac{p(c_i \mid t)}{p(c_i)} \quad (2)$$

4. Chi statistic（CHI）

The chi statistic method measures the lack of independence between the term and the category. If term t and category c_i are independent, then CHI is 0.

$$CHI(t, c_i) = \frac{[P(t, c_i)P(\bar{t}, \bar{c}_i) - P(t, \bar{c}_i)P(\bar{t}, c_i)]^2}{P(t)p(\bar{t})P(c_i)P(\bar{c}_i)} \quad (3)$$

If there are n classes ,then each term value will have n correlation value, the average value calculation for a category as follows:

$$CHI\ (t) = \sum_{i=1}^{n} P(c_i) \log CHI\ (t, c_i) \tag{4}$$

The above 4 method is the most common methods in the experiment andthe different points of ECE and IG is that ECE only considers the effects to category while that words appear in the documents. the DF method is simple, and complexity is lower. CHI method shows that the CHI statistic value is greater, the correlation between features and categories is more strong. The literature [3] experiments show that IG and CHI is most effectivein the English text classification DF followed .Experiments prove that the DF can apply to large scale calculation, instead of CHI whose complex is larger. Literature [7],[9] points out that IG, ECE and CHI methods have same effect on feature extraction in Chinese text classification, followed by DF.

DF is one of the most simple feature term extraction methods. Because of the extraction performance and corpus into a linear relationship, we can see that, when a term belong to more than one class ,the evaluate function will make high score to it; however, if the term belong to a single category, lower frequency of occurrence lead to a lower score. DF evaluation function theory based on a hypothesis that rare term does not contain useful information, or it's information is litter as so to exert useful influence on classification. However, there are few conflicts between this assumption and general information view,In information theory, there are point of view that some rare term with a greater amount of information can reflect the category information than those of high frequency words, and therefore those termsshould not be arbitrarily removed, so the choice only using DF methodwill lose some valuable features. Document frequency method is easy to implement and simple, and it' effect is similar to other methods in the Chinese and English text classification. Aiming at the shortcoming of DFmethod, we present an improved feature selection method based on the DF.

2.2. A Feature Selection Method Based on Document Frequency improved

From the research of literature [10],[11], this paper summed up, to meet the following 3 points of entry is helpful for classification, these requirements are:

1. Concentration degree: in a corpus of many categories, if a feature term appear in one or a few categories, but not in other category text, the term 'srepresentation ability is strong and it is helpful for text classification.
2. Disperse degree: if a term appear in a category,it has strong correlation with the category. That is, a feature termis more helpful to classification while it is dispersedin a large of text of a category.
3. Contributiondegree: if a feature term's correlation with a certain category is more strong , the amount of information is greater andit is value of classification.

This article uses document frequency DF and adopts the following method to quantitatively describe the above three principles:

(1) Concentration degree：Using following formula to expression, the ratio of formula is biggerthat the term is the more concentrated in the class.

$$DF(t, c_i) / (1 + \sum_{j=1 且 j \neq i}^{n} DF(t, c_j)) \tag{5}$$

(2) Disperse degree: There m different classes, $C = \{c_1, c_2, ... c_m\}$,Using $DF\ (t, c_i) / N(c_i)$ to expression, $N(c_i)$ is the text amount of c_i , the ratio is more bigger ,the more the term's disperse degree is bigger.

(3) Contributiondegree：Expectation crossing entropy (ECE) considering the relationship between the feature appearance and categories, so through the calculation information that feature appearing to category. This article uses a simplified formula of ECE to a simplified formula to examine contribution of category. The simplified formula is:

$$P(c_i,t)\log\frac{p(c_i\mid t)}{p(c_i)} \tag{6}$$

In this paper, we implement a new feature selection method of DFM (Document Frequency improved), which its evaluation function based on Document Frequency (DF) method, and the Concentration degree, Disperse degree, Contribution degree are introduced in DFM .The DFM evaluation function is as follows:

$$DFM(t,c_i)=DF(t,c_i)/(1+\sum_{j=1且j\neq i}^{n}DF(t,c_j))++p(t\mid c_i)+P(c_i,t)\log\frac{p(c_i\mid t)}{p(c_i)} \tag{7}$$

2.3. Data Collections

The experimental data used in this paper is from Chinese natural language processing group in Department of Computing and Information technology in Fudan university.The training corpus is " train.rar " which has 20 categories includes about 9804 documents and "test.rar" includes about 9833 documents is used for test. We just choose some of the documents for our experiments because of considering the efficiency of the algorithm. Table 1 shows the specific quantity of samples in each category we chose

Table 1. Experimental Data

	Quantity of training documents	Quantity of testing documents
C1-computer	134	66
C2-Enviornment	134	67
C3-Education	147	73
C4-Medical	136	68
C5-Traffic	143	71
In all	694	345

Experiment environment:CPU is Intel Pentium Dual Core Processor, Intel G2020;Memory is 2G DDR3; Windows XP+Microsoft Visual C++ .

2.4 performance measure

To evaluate the performance of a text classifier, we use F1 measure put forward byrijsbergen (1979) [12]. This measure combines recall and precision as follows:

$$\text{Recall}=\frac{number\ of\ \ correct\ \ positive\ predictions}{number\ of\ \ positive\ examples} \tag{8}$$

$$\text{Precision}=\frac{number\ of\ \ correct\ \ positive\ predictions}{number\ of\ \ positive\ predictions} \tag{9}$$

$$\text{F1}=\frac{2\times \text{Re}\,call\times \text{Pr}\,ecision}{(recall+\text{Pr}\,ecision)} \tag{10}$$

3. Results and Analysis

Our objective is to compare the DF, IG, ECE and CHI methods with theDFM methods. A number of statistical classification and machine leaning techniques have been applied to text categorization, we use SVM classifier. We choose SVM because evaluations have show that it outperforms all the other classifier.

Figure 1 show the selecting performance used SVM on Fudan corpus after feature selection using DF, IG, ECE, Chi, and DFM. It can be seen in Figure 1 that the DFM method outperforms the DF method.

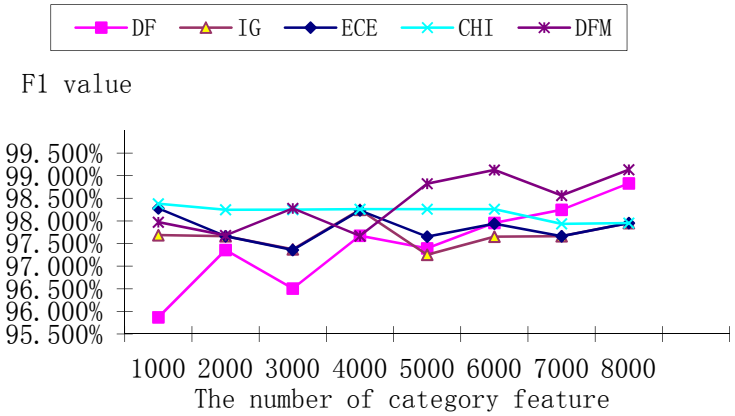

Figure 1. The performance of five feature selection methods

A further observation emerges from the categorization result of SVM. Through the selection of Figure 1, we find that DF F1'rising along with the increase of the characteristic dimension, IG and ECE produce similar performance of classification, because the ECE is a simplified version of IG ,and it only takes into account the condition of feature terms appeared in corresponding category. CHI and DFM are the most effective in our experiment, and CHI classification F1 value has been very stable in the process of the classification. The feature selection method DFM curve are significantly higher than the others methods while the characteristic dimension between 5000 and 8000 ,especially when characteristic dimension is 8000 , Figure 1 appear a maximum points and F1 valueis 99.133%. The classification effect of DFM is better than other four feature selection methods. The extreme value of five kinds of feature selection methods are show in Table 2 with Precision and Recall and F. From Table 2, we can notice that five feature selection methods show better performance all , and DFM gets the best categorization performance that the F1 value is 99.133%.

Table 2. The bestperformance of five feature selection methods

	Recall	Precision	F1
DF	95.868%	95.875%	95.871%
IG	98.226%	98.269%	98.233%
ECE	98.284%	98.287%	98.280%
CHI	98.233%	98.289%	98.261%
DFM	99.121%	99.146%	99.133%

From Figure 1 and Table 2, we can see that DFM can extract category characteristics from Chinese text classification and improve the classification accuracy, and it has the stability in feature extraction.

4. Conclusion

This paper has proposed an improved feature selection method based on DF, named DFM. DFM implemented three principles which are Concentration degree, Disperse degree and Contributiondegree. The experiment has shows that DFM is an effective method to extract category characteristics for feature selection, and it can effectively improve the performance of

text categorization. In the future, we will continue to work on the study of contribution of categories characterization .

Acknowledgements

This work was supported by National Foundation of Social Science (No:08CTQ003),andFoundation ofHebei North University Natural Science for Young (2010).

References

[1] Yiming Yang, XinLiu. *A Re-Examination of Text Categorization Methods*. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). 1999: 42-49.
[2] Mccallum, A, Nigam, K. *A Comparison of Event Models for Naive bayes Text Classification*. Proceedings of the AAAI-98 Workshop on learning for Text Categorization. 1998: 41-48.
[3] Yang Y, Pederson J O. *A Comparative Study on Feature Selection in Text Categorization*. Proceedings of the 14th International Conference on Machine learning.1997: 412-420.
[4] Mladenic, D. Grobelnik, M. *Feature Selection for Unbalanced Class Distribution and Naive Bayes*. Proceedings of the Sixteenth International Conference on Machine Learning.1999: 258-267.
[5] Forman, Guan. Experimental Study of Feature Selection Metrics for Text Categorization. *Journal of Machine Learning Research*. 2003:1289-1305.
[6] KaiFeng Yang, YiKun Zhang, Yan Li. A feature selection method based on document frequency. *Computer Engineer*. 2010; 26(17): 33-35.
[7] Mladenic, D, Grobelnik, M. *Feature Selection for Unbalanced Class Distribution and Naive Bayes*. Proceedings of the Sixteenth International Conference on machine Learning. 1999: 258-267.
[8] Forman, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*. 2003; 3(1): 1289-1305.
[9] Yan Qiu Chen, Nixon, M.S, Damper, R.I., I*mplementing the k - nearest neighbour rule via a neural network*. Proceedings IEEE International Conference on Neural Networks.1995: 136-1401.
[10] DDeDi Tai, Jun Wang. Improved Feature Weighting Algorithm for Text Categorization. *Computer Engineer*. 2010; 36(17): 33-35.
[11] YYong Lou. A Improved Study on Feature Selection of Text categorization Based on Mutual Information. *FuJian Computer*. 2009; 12(4): 82-83.
[12] YYang YM. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*. 1999; 1: 67−88.