A Rapid Method for Text Tendency Classification

LI Yan-ling[1,2], DAI Guan-zhong[1], QIN Sen[1]

(1. College of Automation, Northwestern Polytechnical University Xi'an 710072; 2. Xi'an Research Institute of Hi-Technology Xi'an 710025)

Abstract A rapid method for text tendency classification is proposed in this paper. By means of class space model to display the tendency of the words to the categories, the method realizes the classification based on the statistic characteristics of words. In this method, through the studies of the complexity of text tendency categorization, three statistic characteristics of word such as frequency, document frequency and the distribution of words are comprehensively taken into account, and a new method of twice feature selection is proposed：In the first characteristic selection process, using combination characteristic selection method, the words that those distributions are uniform in each category and the low-frequency words are deleted. Then in the second process, the words that those category tendencies are not obvious are deleted. The experimental results show that the algorithm is running-fast, and has high performance.

Key words category weight; class space model; text tendency categorization; twice feature selection

快速的文本倾向性分类方法

李艳玲[1,2]，戴冠中[1]，覃 森[1]

(1. 西北工业大学自动化学院 西安 710072; 2. 西安高技术研究所 西安 710025)

【摘要】提出了一种快速的文本倾向性分类方法，即采用类别空间模型描述词语对类别的倾向性，基于词的统计特征实现分类；针对倾向性分类的复杂性，在综合考虑词频、词的文本频、词的分布三种统计特征的基础上，提出一种新的二次特征提取方法：第一次特征提取，采用组合特征提取方法，除去低频词以及在各类中均匀分布的噪音词；第二次特征提取，去除类别倾向性不明显的词。实验表明该分类方法不仅具有较高的分类性能，而且运行速度快，在信息检索、信息过滤、内容安全管理等方面具有一定的实用价值。

关 键 词 类别权重; 类别空间模型; 文本倾向性分类; 二次特征提取
中图分类号 TP391 文献标识码 A

Text classification refers to the processes that under the given classification system, the texts of the unknown category are processed automatically to determine their categories according to the text characteristics[1]. With the rapid development of the Internet, the information of the Internet is great enriched and becomes the main sources to obtain and diffuse the information for people. The applications of the Internet, such as effective information retrieval, information filtering, content management and online public opinion analysis, become increasingly important. And text classification is an effective solution. It has become a key practical technology.

Text classification can be divided into two categories according to classification granularity, the theme classification and the tendency classification. Theme classification is based on the theme discussed in the articles. Tendency classification is based on the views or attitudes of the articles in terms of a particular theme. Tendency classification has very important applications in the sensitive information identification and filtering out the particular theme, and in public opinion extraction, analysis of public opinion tendency and content security management. So it has received

Received date: 2007 – 08 – 20
收稿时间：2007 – 08 – 20
Foundation item: Supported by the National 863 Projects (2005AA147030)
基金项目：国家863计划项目(2005AA147030)
Biography: LI Yan-ling was born in 1970. She is a PHD candidate and associate professor, Her research interests are in the data mining and information security.
作者简介：李艳玲(1970 –)，女，在职博士生，副教授，主要从事数据挖掘、网络信息安全等方面的研究.

much attention in domestic and abroad in the past two years.

A lot of text classification methods[2-5] are proposed in recent years and there are numerous methods have achieved a better classification effect in the theme classification. But this paper found that there would be a greater degree of lower classification accuracy directly using some methods to the tendency classification. We used an experimental method in Ref. [2], and the average rate of correct classification was 58.9%. There are two main reasons by our analyses. On the one hand, in the tendency classification process, lots of the same words appear in various categories since the articles belong to the same theme. On the other hand, when people express their views, they often carry emotional or subjective awareness. And there are a number of ironies or metaphors to deliver their tendency. Therefore, although it is the same word, different meaning of the expression is displayed in different articles.

In this paper, aiming at the two main reasons mentioned above, a method of text tendency classification based on statistics is proposed. In the method, class space model is adopted to depict the word tendency to the categories[2]. And the texts are classified according to the category tendency of words.

1 Class space model

1.1 The concept of class space model

The class space model is constituted by the categories of types[2]. And in the model, each sort C_j is regarded as an axis X_j of the space. Each word t_i is regarded as a node v_i of the space. The weight that v_i is mapped in the axis X_j is the weight W_{ij} that the word t_i in the category C_j. In a M-dimensional category space, if the weight of the word t_i in each category are $W_{i1}, W_{i2}, \cdots, W_{im}$, the mapping coordinate of the point v_i is $W_{i1}, W_{i2}, \cdots, W_{im}$.

1.2 The weight of word

In class space model, the weight of words indicates the tendency of the words to the categories. That is to say, it measures the abilities of the words embodying the categories. According to previous studies, the higher the frequency of the characteristic t_i appears in the sort C_j, the stronger the tendency of the word t_i to the class is. Meanwhile, in order to highlight the role of important words and inhibit the role of secondary words, the weight W_{ij} that the *i*th word in the *j*th category can be calculated by：

$$W_{ij} = \exp\left(\frac{f_{ij}}{\sqrt{f_{i1}^2 + f_{i2}^2 + \cdots + f_{im}^2}}\right) \quad j = 1, 2, \cdots, m \quad (1)$$

where f_{ij} is the frequency of the word t_i appears in the sort C_j. In order to remove the influence of the document length and the number of each type of the training documents, the formula of the frequency f_{ij} is as follows:

$$f_{ij} = \frac{\sum_{k=1}^{N_j} \frac{\mathrm{Num}(i,k)}{\mathrm{Num}(k)}}{N_j} \quad (2)$$

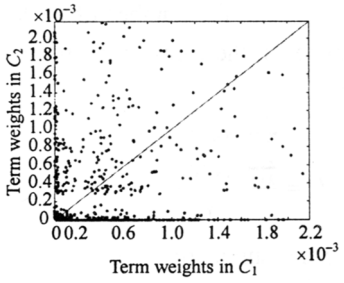
where N_j is the summation number of the jth class document, and $\mathrm{Num}(i,k)$ is the times that the ith characteristic word appears in the kth document. $\mathrm{Num}(k)$ is the summation number of the characteristic words in the *k*th document.

However, in the follow-up studies, we have found when using Eq. (1) to calculate the category weight of words there are some inadequacies: Eq. (1) only measures the category representation by means of the frequency of the words, but does not consider the distribution of words. Generally, the frequency of the majority of the 1 000-character Chinese texts is ten to the third power of magnitude[6]. This makes that the calculation of the weight difference in the type of words be very small, thus reduces the accuracy of the classification. In fact, there is a great influence that the word contains the information of categories in terms of its distribution in the all categories[7]. The word appears more in the same category, while it appears less in other types, the word should reflect the nature of the category to greater. The calculation Eq. (1) of category weight of word is modified as follows:

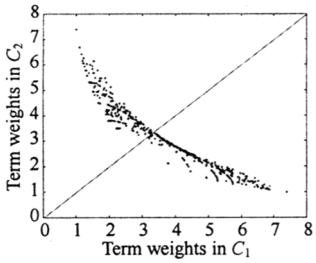
$$w_{ij} = \exp\left(\frac{f_{ij}}{\sqrt{f_{i1}^2 + f_{i2}^2 + \cdots + f_{im}^2}}\right)\exp\left(\frac{N_{ti,j}}{N_{ti}}\right) \quad j = 1, 2, \cdots, m \quad (3)$$

where $N_{ti,j}$ is the document number that the document of the *j*th category contains the *i*th characteristic t_i, N_{ti} is the all document number containing the *i*th characteristic t_i.

The data of the experiment 1 is used as a sample, and is divided into two categories according to their different viewpoints tend to create a type of two-dimensional space, which are noted C_1 (X axis) and C_2 (Y axis). And the characteristic word t_i is mapped into a node in the space. Its coordinate in the space can be expressed as (W_{i1}, W_{i2}). The distribution of all characteristic words in the two- dimensional space is shown in Fig. 1.



a. Based on the Eq. (1)



b. Based on the Eq. (3)

Fig.1 The distribution of the characteristics in the two-dimensional space

So it is obvious that if we calculate the category weight of words based on the Eq.(3), it is very effective to distinguish the category tendencies of the majority of the words.

2 Twice feature selection

2.1 The first feature selection

Feature selection is a necessary prerequisite for text classification. In the tendency classification, the same words appear in many different categories and disrupt the effect of the classification. However, their importance is not the same to the different categories. Therefore, they can not be generalized to the removal.

According to previous studies, the word is a noise word when the word appears in various categories randomly and uniformly.

Then, the distribution coefficient[7] concept is adopted in this paper. If the text is classified into m-categories, the distribution coefficient α_i of the characteristic word t_i can be calculated by the following formula since there are C_m^2 kinds of combinations:

$$\alpha_i = \exp\left(\frac{1}{2C_m^2} \sum_{\substack{j,k=1 \\ j \neq k}}^{m} \left| \frac{N_{t_i,j}}{N_j} - \frac{N_{t_i,k}}{N_k} \right| \right) \tag{4}$$

where N_j is the summation number of the jth class documents.

The larger the value α_i is, the larger the category information of the word t_i is, so as to the contribution of the word to classify the categories is very important.

On the other hand, reference Ref.[8], the majority of low-frequency words are the noise words and do not become the feature words.

Given the above analysis, for the first feature selection, we use the combined feature selection method; using the Ref.[8] method get rid of those low-frequency words for the little significance in the classification firstly. Then, the words of those α_i value below a certain threshold will be removed.

2.2 The Second feature selection

From Fig.1b, we see that，although most feature words have obviously category tendentious, there are a little of words that those tendencies are hazy. They centralize the strip district in the figure, which district nears a close angle bisector region.

Suppose that the width of the strip district be ε, the words dropping the region $|x-y|<\varepsilon$ will disrupt the effect of the classification. So the twice feature selection is proposed in the paper to solve the problem. Select a certain ε and to delete the words dropping the region $|x-y|<\varepsilon$, which is shown as Fig.2 (ε=0.002).

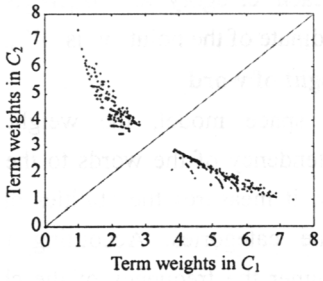


Fig. 2 Feature distribution after the second feature selection

We can see from the figure that the category tendency of feature words has been more obvious distinction.

3 The algorithm of text tendency classification

The algorithm of text categorization proposed in the paper can be classified into two stages, training stage and classification stage[7]. The main steps of the algorithm are as follows.

Training stage:

(1) Pretreatment of text;

(2) Feature selection for the first time;

(3) Calculate the category tendency coefficient W_{ij} of each word according to the Eq. (3);

(4) Set a given value of ε, then the second feature selection is processed and a general feature words table is obtained.

Classification stage:

(1) Pretreatment of the not-categorized documents;

(2) Calculate the class weights D_j of the not-categorized documents. The calculation formula is as follows:

$$D_j = \sum_{i=1}^{T} \lambda_i W_{ij} \quad j = 1, 2, \cdots, m \tag{5}$$

$$\lambda_i = \begin{cases} 0 & \text{Feature } t_i \text{ not appeared in the document} \\ 1 & \text{Feature } t_i \text{ appeared in the document} \end{cases} \tag{6}$$

where j is the serial number of the categories and T is the sum number of the words in the general feature word table; D_j is the class weight of the document for the jth category.

(3) Seek the corresponding index of the maximum of D_j. If D_k is the maximum, the category of the document is k and the algorithm finishes.

4 Experiment and result analysis

In this paper, the open test of the algorithm is processed using the three common methods to assess the performance of classification: Precision(P)、Recall (R) and F_1, whose calculation formulas of the three performances can be found in Ref. [9-10].

Experiment 1　The data of this experiment are downloaded from the website: http://bbs. people. com. cn. They are divided into two categories: C_1、C_2. The results of this experiment are shown as the Table 1.

Table 1　Results of classification

Category	training set	test set	P/(%)	R/(%)	F_1/(%)
C_1	120	42	87.23	97.62	92.13
C_2	111	39	97.06	84.62	91.43

From the table, we can find that the effect of the classification algorithm is good and the operating speed is quick because of a simple calculation processing (polynomial time complexity level).

Experiment 2　The other data about releasing in a month is downloaded and classified six groups. The measures are: Macro-average precision(Macro$P$)、Macro-average recall(Macro$R$) and Macro-average F_1(MacroF_1)[1]. Second feature selection to the classification results of a test is shown as Fig.3.

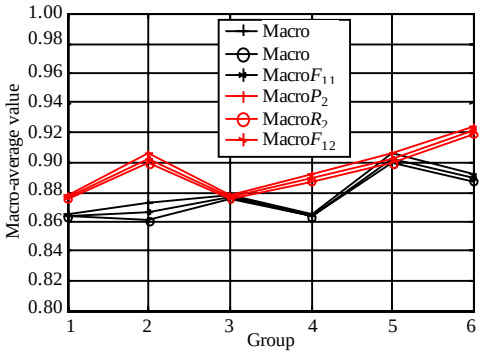


Fig. 3　Second feature selection results

In the figure, MacroP_i、MacroR_i、MacroF_{1i}(i=1,2) are the corresponding classification results before and after the second feature extraction, respectively.

We can found from the figure that the stability of the algorithm is very high and the performance of the classification after the second feature extraction has some enhancement.

Experiment 3　For test the influence of the classification results with α_i, we set that the value DF be 3 in the experiment. And the distribution coefficients α_i of all feature words are sorted by increasing order and connected a curve, which is shown in Fig.4. The classification results in the different values of α_i are shown in Fig.5. From the two figures, we can see when α_i=1.002, the inflexion of the curve appears, and it was found that the effect of the classification is the best.

Therefore, if two words have vague classification tendency, the values of their distribution coefficients are very close. When α_i is given a certain threshold, a majority of words with vague classification tendency are wiped off and the precision of classification is advanced evidently. However, when α_i is excessively large, the number of characteristic words is reduced obviously. On the contrary, the effect of classification is not good, which is similar to the results of Ref.[8, 10-11].

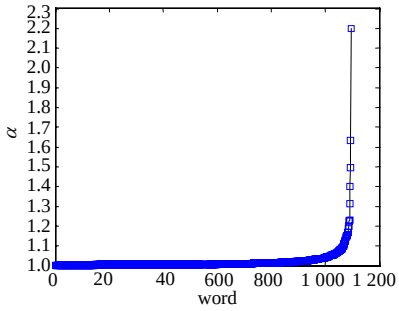


Fig. 4 The distribution coefficient of the feature words

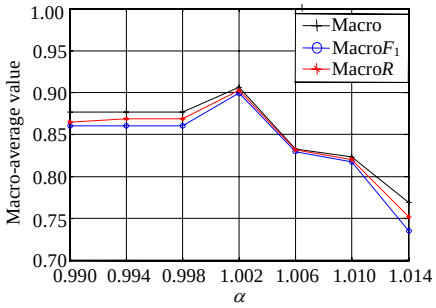


Fig. 5 The classification results in the different values of α_i

5 Conclusion

A method of text tendency classification is proposed in this paper. The class space model is adopted to depict the tendency of words to the categories in the method. And the texts are classified according to the category tendency of words. Aimed at the condition that the number of the same words is very much in the tendency classification, twice feature selection method is introduced. In the first characteristic selection process, using DF and distribution coefficient combination characteristic selection method, the words that those distributions are uniform in each category and the low-frequency words that they have no sense to classification are deleted. Then in the second process, the words that those category tendencies are not obvious are deleted. The experiment results show that the algorithm is simple and running-fast, and has high classification accuracy to deal with the cosmically text set timely.

References

[1] ZHAO Shi-qi, ZHANG Yu, LIU Ting, et al. A feature selection method based on class feature domains for text categorization[J]. Journal of Chinese Information Processing, 2005, 19(6): 21-27.

[2] HUANG Ran, GUO Song-shan. Research and implementation of text categorization system based on class space model[J]. Application Research of Computers, 2005, 22(8): 60-63.

[3] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[4] LI Y H, JAIN A K. Classification of text documents[J]. The Computer Journal, 1998, 41(8): 537-546.

[5] LIU B, HSU W, MA Y. Mining association rules with multiple minimum supports[C]//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-99). New York, USA: ACM Press, 1999: 337-341.

[6] LUO Xin, XIA De-lin, YAN Pu-liu. Improved feature selection method and tf-idf formula based on word frequency differentia[J]. Computer Applications, 2005, 25(9): 2031-2033.

[7] LI Yan-ling, DAI Guan-zhong, ZHU Ye-hang, et al. A high performance extraction method for public opinion on internet[J]. Wuhan University Journal of Natural Sciences, 2006, 12(5): 902-906.

[8] DAI Liu-ling, HUANG He-yan, CHEN Zhao-xiong. A comparative study on feature selection in chinese text categorization[J]. Journal of Chinese Information Processing, 2004, 18(1): 26-32.

[9] CHEN Zhi-gang, HE Pi-lian, SUN Yue-heng, et al. Research and implementation of text classification system based on VSP[J]. Journal of Chinese Information Processing, 2005, 19(1): 36-41.

[10] FAN Xing-hua, SUN Mao-song. A high performance two-class chinese text categorization method[J]. Chinese Journal of Computers, 2006, 29(1): 124-130.

[11] KIM H, HOWLAND P, PARK H. Dimension reduction in text classification with support vector machines[J]. Journal of Machine Learning Research, 2005, 6(1): 37-53.

编 辑 黄 莘