

# Research on the internal influence factors of the text multi-classification problem

Mingqiang Wu<sup>1</sup>, Furong Chang<sup>1</sup> Kui Zhang<sup>2</sup>

<sup>1</sup>College of computerscience and technology, Kashi University, No.29 Xueyuan Road, Kashi, China

<sup>2</sup>College of computerscience and technology, Kashi University, No.29 Xueyuan Road, Kashi, China

**Abstract:** This paper mainly deals with the classification of text type data. The statistics show that more than 8000 articles have been reached in all kinds of documents retrieved by the optical network. However, there are few papers on the factors that affect the classification of text. The text classification method used is important, but the internal factors sometimes play a great role, and even affect the success or failure of the whole text classification. In order to make up for this deficiency, this paper selects the Rocchio algorithm as the classification method, mainly from the category clustering density, class complexity, category definition, stop words and document's length five internal factors, we tested their influences on text classification by the experiment. Experiment shows that the clustering density is higher and the complexity of the lower class, class definition is higher, the higher the accuracy of text classification, text classification effect is better, and better effect to text stop words, the length of the text does not directly affect the effect of text classification, but according to the text classification algorithm is more suitable to choose the length of the document.

## 1 INTRODUCTION

With the increasing popularity of computer applications, the Internet, as the greatest technological achievement of twentieth Century, it is changing the world with great strength, changing the human society and changing the way of life for everyone. The massive data information on the Internet, including text information, picture, audio and video, is constantly flooding into our minds, and these information is growing exponentially. How to classify them quickly and accurately will continue to become a hot topic in the field of information processing for a long time.

There are two kinds of text categorization: if a text belongs to only one category, it is called 1-of-N classification; if a text can belong to multiple or even all categories at the same time, or it doesn't belong to any category, it is called m-of-n classification<sup>1</sup>. The feature selection of multi class classification can be to select a feature subset for a single class or to select a common feature subset for all classes<sup>2</sup>.

The commonly used text classification techniques include Rocchio, Bias, clustering, neural network, support vector machine and so on. No matter what algorithm is used, we need to preprocess texts before

classifying text, such as stopword removal, selecting appropriate corpus, selecting suitable training samples, etc. If the preparatory work is not well done before classification, then a good text classification algorithm can't play its due role, so the pretreatment of text is particularly important.

This paper will analyze the internal factors of the text multi classification problem, and try to provide a reference for the study of the text classification problem.

## 2 Analysis the internal factors of multi text classification PROBLEM

### 2.1 Multi Text Classification Process

#### 2.1.1 Classification Method In This Paper

There are many methods of text categorization, such as naive Bayes algorithm (NB), K- nearest neighbor algorithm (KNN), Rocchio algorithm, support vector machine (SVM) and so on<sup>3</sup>. For the convenience of research, the improved Rocchio algorithm is used in this paper. The flow of the training set classification algorithm is as follows:

\* Corresponding author: {986449600,514286392}@qq.com,zhangkui319201@126.com

(1) participle

In English sentences, words and words are separated by spaces. The information processing of English texts is much more convenient than Chinese. Many operations, such as statistical word frequency, word analysis and so on, can be directly carried out in existing articles.

The text is clearly to Chinese words as text interval to store, we are in accordance with Chinese grammar to understand every word meaning, such as the basic "SVO" structure, the meaning of this and we just want to express the oral agreement, in fact, every word I say in everyday communication implicit separated into different parts, but the write is a complete sentence. In order to make a better study of each sentence, a participle tool is usually used. There are many common word segmentation tools, which are based on the statistical word segmentation tool Ictclas. By using the Ictclas word segmentation tool, words and its part of speech can be obtained at the same time.

(2)stopword removal

Some words, such as "de", "in it", "are", "but" and so on, have played a certain role in the connection of sentences<sup>4</sup>. But these words tend to be of little significance to the topic of expressing articles, so relatively speaking, they are more likely to appear in articles. In order to better mine their influence on text classification, some experiments will be done later.

(3)selection of characteristic words

In this paper, the chi square based feature selection, the chi 2 statistic, is used to detect the independence of two events. It is a measure of the degree of deviation between the expectation and the observation. If two events depend on each other, then the emergence of an event will make the occurrence of another event more likely or impossible, so it is chosen as a feature<sup>5</sup>. Let N denote the total number of documents, t said the feature words, C said a category, then agreed to A: the training set number, category contains features of T documents in C B: the total number of training set, the document contains features of t other categories except category C in the C: the training set number, category feature does not contain t the document in C D: the total number of training set document does not contain features of other t categories except category C in the. The value of the  $\chi^2$  is shown in Eq.1 as follows:

$$\chi^2(t,c) = \frac{N*(AD-BC)^2}{(A+C)(A+B)(B+D)(C+D)} \quad (1)$$

(4)feature weighting

The selected feature words may appear in every class, but the selected feature words should generally satisfy the principle of "compact within class and loosely between classes", and assign certain weights to the corresponding feature words. This can better reflect the role of the feature words in the text. Here we mainly use the supervised feature weighting method  $iqf*qf*icf$  to carry out the feature weighting. This method shows good

performance in text classification and achieves ideal results on multiple data sets in two fields. The calculation method is shown in Eq.2:

$$iqf*qf*icf = \frac{1}{\log(N/(a+c))} * \log(a+1) * \log(|C|/cf+1) \quad (2)$$

In the upper form, IQF means using a global weighted item. QF represents the number of documents that contain the current features in the positive class document. ICF represents the ability of feature to distinguish all kinds of data in the data set. In addition, a indicates that the number of documents containing characteristic t is in the positive class. C indicates the number of documents that contain the feature t in the negative class. N represents the total number of documents in the dataset, C is the total number of categories in the dataset, and CF represents the number of categories that contain the least number of current features.

(5)compute the classification vector for a training set

The use of the word vector dimension reduction, words of each text vector is trained in the category of text said, then these feature vector sum and normalized, essentially constitutes a document theme matrix, each row of the matrix represents category category document vector current.

The training set algorithm flowchart is shown in Figure 1:

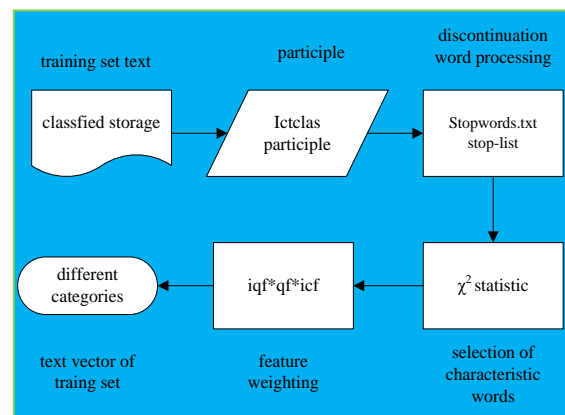


Figure 1: flow chart of training set algorithm.

The test set algorithm as follows:

(1)Participle

The test set also uses the Ictclas word segmentation tool to get the text after the word segmentation.

(2)discontinuation of words

Disuse word processing according to more than 500 words provided in the disable word list stopwords.txt.

(3)first feature selection

The statistical test focuses on the DF value and extracts the maximum n values from it.

(4)second feature selection

Use the TF-IDF formula to calculate the TF-IDF value of the largest n words of the above TF, from which the maximum value of m TF-IDF is taken out<sup>6</sup>. The

$$\text{sim} = \frac{\mathbf{V}_t \cdot \mathbf{V}_s}{|\mathbf{V}_t| |\mathbf{V}_s|} \quad (5)$$

for  
mul  
as  
for  
the

calculation of TF and IDF are as follows:

$$tf(w, d) = \frac{f(w, d)}{\sum_i f(w_i, d)} \quad (3)$$

$$idf_w = \log\left(\frac{N}{1 + df_w}\right) \quad (4)(2)$$

In formula two and formula three,  $f(w, d)$  represents the number of words  $w$  appears in document  $D$ ,  $DFW$  represents the number of documents contained in the corpus, and  $N$  represents the total number of documents in the corpus. The weighting of word  $w$  is  $tfidf_w = tf(w, d) * idf_w$ .

(5) compute the text vector of the test text set

After calculating the TF-IDF value of each text, the same test text set is made up of document topic matrix, and finally, every word's vector expression can be obtained through the theme feature word matrix.

(6) textual vector de centralization

The first test text into different groups, each group contains 5 text, then start to the center of the processing, the specific methods are as follows: first calculate the average of each line of text vector feature theme - word matrix values, a vector is obtained, then each line of the TF-IDF value minus the vector, then for each column, and the standard deviation, then each column of data divided by the standard deviation, so that they get to the de-centralization matrix<sup>7</sup>.

The test set algorithm flowchart is shown in Figure 2:

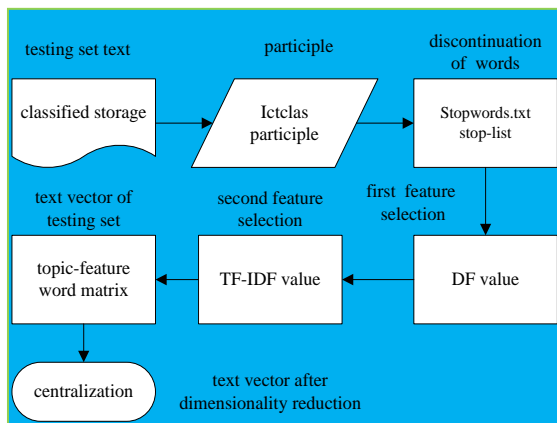


Figure 2: flow chart of test set algorithm.

Finally, the test text set is classified according to the generated training model:

Here using the cosine theorem to calculate the similarity between the text vector and the class vector.

If the text vector sets as  $V_t$ , the class vector sets as  $V_s$ , the similarity between the text and the class text is set to  $\text{sim}$ , as shown in Eq.5:

According to the similarity values between test texts and different classes of training texts, we lookup the maximum value from these values, the test text belongs to the category of the training text with the maximum similarity value.

## 2.2 The Internal Factors Of The Multi Text Classification In This Paper

The following are a brief introduction to several factors that affect text classification in this article.

①Category clustering density

Class clustering (CCD) is used to measure the degree of application about document features in this category when the category is represented<sup>8</sup>. Set the category clustering density to be  $cd$ , as shown in Eq.6.

$$cd = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j \neq i}^N \text{sim}(d_i, d_j) \quad (6)$$

In which  $\text{sim}(d_i, d_j)$  represents the similarity between document  $d_i$  and document  $d_j$ ,  $N$  represents the total number of documents in the corpus, and  $i$  and  $j$  represent the subscript of the document.

②Category complexity

Category complexity (COC) measures the degree of difference between document features within a category, which is independent of the value of the average similarity of documents, and relate to that number of value taken and the number of documents under different value taken<sup>8</sup>. The complexity of a class sets as  $co$  and the calculation method is shown in Eq.7.

$$co = N * \frac{1}{N-1} \sum_{j \neq i}^{N-1} \text{sim}(d_m, d_n) \quad (7)$$

Among them,  $d_m$  represents a document under the current category.  $d_n$  represents other documents in the corpus.  $\text{sim}(d_m, d_n)$  represents the similarity of the two documents, and  $N$  represents the total number of corpus documents.

③Category definition

First, we use the average value of all document vectors in a category to represent the centroid of the class. Then we calculate the similarity between the centroid of the class and the average of all documents in another category as the similarity between the two categories. By calculating the similarity between two categories, the degree of distinction between categories can be measured<sup>8</sup>. The category definition sets as  $doc$ , and the calculation method is shown in Eq.8.

$$\text{doc} = 1 - 1/2N * \left( \sum_{i=1}^N \text{sim}(d_i, d_o) + \sum_{j=1}^N \text{sim}(d_j, d_o) \right) \quad (8)$$

Among them, variable “ $d_i$ ” represents the  $i$ th document vector of a category, variable “ $d_o$ ” represents the centroid vector of this class, and it can directly get the classification clarity of the two categories of documents to be calculated by using the similarity of 1- two classes.

④stop words

Mentioned before, stop words are in the sentence, mainly used in conjunction a word, generally no meaning of the stopword on the expression of the article’s theme, in the end they have no effect on text classification, play a positive role or a negative role? How much is the impact? These issues will be discussed later.

⑤Document length

After word segmentation, the corresponding word is stored in a list collection, in order to facilitate, here the elements of the collection, namely the number of words or the length of the text, in a randomly selected text test set, whether the length of text will be affected in the division of categories?

### 3 Experimental results and conclusions

#### 3.1 Corpus Selection

This paper selects the part of the corpus in the Sogou corpus as experimental data, which includes three types of text sets: education, internet and recruit. Each category of the training set was randomly selected 200 articles, a total of 600, and each category of the test set randomly selected 100 articles, a total of 300. The samples selected from the training set are shown in Table 1.

**Table 1:** training set sample.

trainset		
education	internet	recruit
200	200	200

The sample selected by the test set is shown in Table 2.

**Table 2:**test set sample.

trainset		
education	internet	recruit
100	100	100

#### 3.2 Classification Result

This paper uses Java as a development language, MyEclipse as development tool, first calculate the class vector in training set include education category, internet category, recruit category, then compute the similarity between the various categories of text vectors in the test set and the corresponding three class vectors. Category of maximum similarity value as the current document category. The positive samples number and accuracy of the class in the three categories, as shown in Table 3, are:

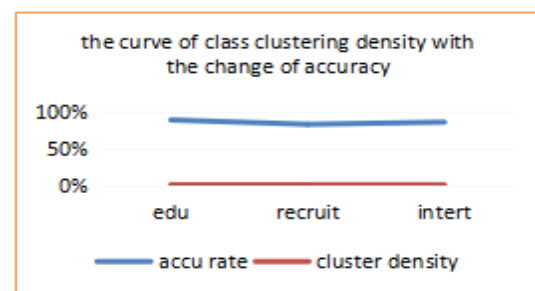
**Table 3:** classification result.

	education	internet	recruit
accuracy rate	89%	86%	83%

### 3.3 Analysis Of The Influencing Factors

#### 3.3.1 Category Clustering Density

After calculating the clustering density of each category, and get a variation curve between the category clustering density and the accuracy of each category, as shown in Figure 3.

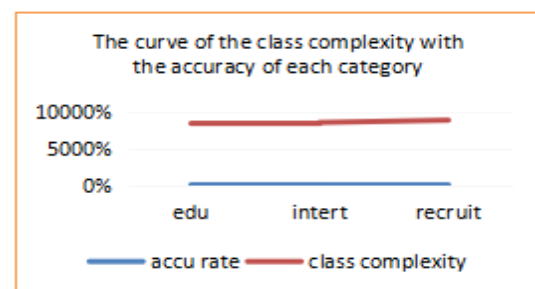


**Figure 3:** Class clustering density and accuracy rate relation curve.

According to the above curve, it can be seen that the larger the classification density of the document, the higher the accuracy of the document, the better the effect of the classification.

#### 3.3.2 Class Complexity

Similarly, after calculating the class complexity of the three categories, the change curve between the class complexity and the accuracy of each category as shown in Figure 4.

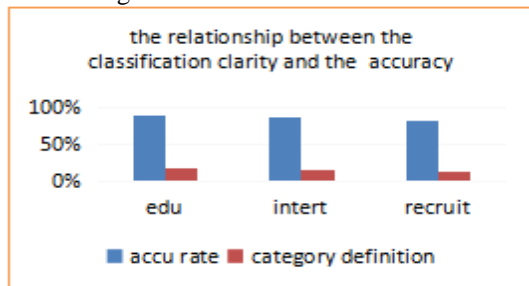


**Figure 4:**class complexity and accuracy rate relation curve.

As can be seen from the above picture, the higher the class complexity is, the lower the classification accuracy is, that is, the worse the classification effect is.

#### 3.3.3 Category Definition

After calculating the center of mass vector of each class, we calculate the value of category definition with the corresponding formula, and get the relationship between the classification clarity and the classification accuracy. It is shown in Figure 5:

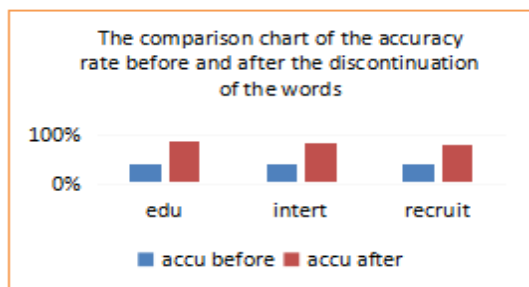


**Figure 5:** Relation diagram of category definition and accuracy.

As can be seen from the figure, the categories definition is higher, the higher of classification accuracy is.

### 3.3.4 Stop Word

In the earlier classification methods, both the training set and the test set all go through the step of remove stop word. Here we try to delete the stopword list, and then classify the test set according to the same method. The accuracy rate and the accuracy rate before classification are compared. The result is shown in Figure 6.

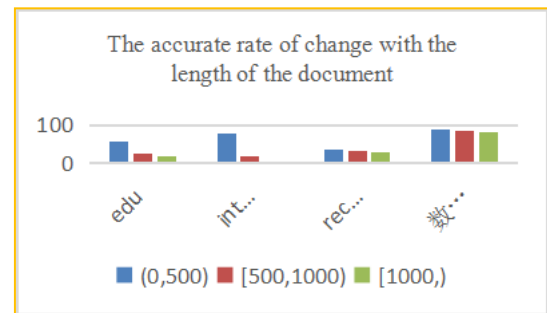


**Figure 6:** diagram of the contrast before and after removal stopword.

From the above, we can see that the accuracy rate of text classification is quite different before and after removal stopword. The accuracy of text classification is much lower than that removal stopwords before. Therefore, treatment the stopwords can improve the effect of text classification.

### 3.3.5 Document Length

The number of words in the list set is added to the length of a document as a standard for measuring the length of a document. Here select three type documents: the length is less than 500, the length is greater than or equal to 500 and less than 1000, the length is greater than or equal to 1000, different length types documents as the research object. the relationship between accuracy of text classification and length change as shown in Figure 7:



**Figure 7:** the relationship between text classification accuracy and document length.

It can be seen from the figure, within the same category, the document length is shorter, the higher the accuracy of text classification, but in different types of documents, with short document number, the accuracy of text classification reduces. Similarly, when the text category from the education class to internet class, the accuracy declines as the number of documents reduce, when the text category from the internet class to recruit class, with the increasing number of long documents, the accuracy is still in decline. This indicates that there is no necessary relationship between the length of documents and the accuracy of documents. The different text classification methods selected will have a certain influence on the classification effect of different length of documents. This is the commonly called short text classification problem and the long text classification problem, and there are special classification methods to solve the two classification problems<sup>9</sup>.

## 4 CONCLUSIONS

From the perspective of training set and test set, we extract the class vectors of training set and the text vector of the test set. We select three kinds of language materials in Sogou corpus as training set and test set, and randomly extract a certain number of texts from each category. Through continuous calculation, the accuracy of text similarity in test sets in each category is obtained. and then calculate the category clustering density, category complexity, category definition, by drawing the line map and cluster shaped column diagram clearly depicts the relationship between the three factors and the accuracy of the similarity. Because the text classification algorithm adopted in this paper filters the stopwords by default, after calculating the accuracy of similarity, by annotating the stopword list, the similarity test is carried out in the same way as the original method. After processing with participle and discontinuation words, the list added to the number of elements in the collection as the basic unit to measure the length of the text, in order to facilitate the analysis, the length of the document is divided into three sections, after the length of the statistical documents, through the histogram simple analysis of the relation between them and the text classification accuracy.

Finally draw a conclusion: the similarity accuracy increases with increasing the density of clusters, with the increase of the complexity of the categories decreased, increased with the increase of category definition. The accuracy of text classification after removal stopwords is better than that without removal stopwords, there is no necessary relation between the length and the classification accuracy of the document. The selection of text classification algorithms plays decisive role in the classification effects of different length documents.

Through the analysis of the five internal factors that affect the text classification, we can see that in the future when doing text classification experiments, it is necessary to select the text with high density, low class complexity and high definition as experimental corpus. According to the different classification algorithms plays attention to an antidote against the disease, selection of different classification algorithms based on the length of the text, if it is short text, choose a better algorithm for the classification of short text, if long text, choose a wider application algorithm on long text. Not only to filtering stopword, but also take a variety of ways select more representative words to add to the stopword list. This will be of great benefit to improving the efficiency of text classification.

## REFERENCES

1. De Wu, Sanyang Liu, Jinjin Liang. (2016), Multi-class text classification algorithm Computer Science, Vol. 43 August, pp190-191, Chongqing City: China Academic Journal Electronic Publishing House
2. Bo Wang et al. (2010), Study of feature selection in text multiclassification Computer engineering and Science, Vol. 32 August, pp90-92, Changsha City: China Academic Journal Electronic Publishing House
3. Wan C H, Lee L H, Rajkumar, R et al. (2012), A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine Expert System with Application, Vol. 39 December, pp11880-11888, America: 2018 Elsevier B.V.
4. Jiao Xu. (2015), Research on text feature selection algorithm based on Hadoop, MD thesis. Lanzhou: Lanzhou University
5. Guangqiang Pan, Jun Zhou, Yang He. (2014), The study of text feature selection based on the simple Bias classification model Computer knowledge and technology, Vol. 10 January, pp133-134, Hefei City: China Academic Journal Electronic Publishing House
6. Yongliang Wu et al. (2017), Text classification method based on TF-IDF and cosine similarity Chinese information journal, Vol. 31 September, pp139-140, Beijing: Academic Journal Electronic Publishing House
7. Tiantian Wang, Yu Kang. (2016), Research on the use of variance and word vector in text reduction Application of computer system, Vol. 25 November, pp2-5, Beijing: Academic Journal Electronic Publishing House
8. Xiangdong Li, Zhichao Ba, Li Huang. (2014), Research on the impact of text classification performance based on corpus information measurement Journal Of Intelligence, Vol. 33 September, pp158-159, Xi'an City: China Academic Journal Electronic Publishing House
9. Erjing Chen, Enbo Jiang. (2017), A review of the method of text similarity calculation Data analysis and knowledge discovery, Vol. 6 June, pp6-10, Beijing: Academic Journal Electronic Publishing House