# Research on Feature Selection Method in Chinese Text Automatic Classification

Ying Hong [a], Xiwen Shao

Computer Information Center, Beijing Institute of Fashion Technology, Beijing, China

[a]jsjsxw@bift.edu.cn

**Keywords:** text classification, feature selection, Bayes algorithm, chi-square statistic.

**Abstract.** This paper introduced the importance and workflow of Chinese web page classification. It studied the defects of chi-square statistic algorithm and improved it. At last, it verified the improved chi-square statistic algorithm combined with Bayes algorithm through a series of experiments. The experimental results show that the improved algorithm improved the accuracy of Chinese web page classification.

## 1. Introduction

Chinese web page automatic categorization analyses characteristics of web page unclassified and divided it into corresponding category [1]. Web page automatic classification has the advantage of rapid classification without human intervention. It has been widely used in analysis of user behavior, personalized recommendation service, precision marketing and other fields. It is a very practical technology.

This paper focuses on the study of Chinese web page classification techniques. It introduced the related technologies of Chinese web page classification. It studies on feature selection algorithm deeply and puts forward the corresponding improvement ideas. Finally, it has verified the validity of improvement ideas through experiments.

## 2. Workflow of Chinese text classification

The workflow of Chinese web page classification is shown in Fig. 1. As we can see from Fig.1, the workflow of Chinese web page classification is divided into two parts: training process and classification process [2]. In the training process, we will select feature in training set and get a collection of feature items after Chinese segmentation.

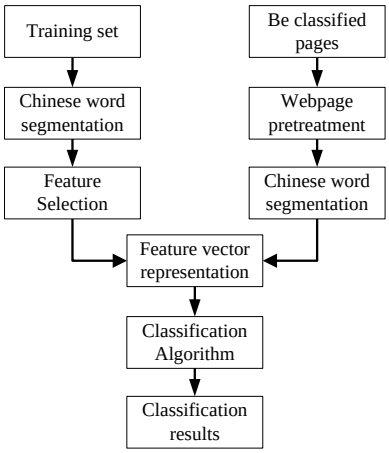


Fig.1 The workflow of Chinese web page classification

Then, we will create a feature vector space and all instances are represented as vectors. At last, we use classification algorithm to construct classifier and train the classifier using the training set. In the

 

classification process, the web page to be classified will be expressed in vector form after a Chinese word processing. Then, we use the trained classifier to classify them.

## 3. Chinese word segmentation

Chinese word segmentation is a unique concept in Chinese text classification [3]. There are no obvious segmentation signs between words in Chinese text. The computer needs to find out automatically the dividing line between words using the Chinese word segmentation. Therefore, the Chinese word segmentation is the foundation of Chinese web page classification technology [4].

The Chinese word segmentation tool used in this paper is Pangu word components. The version number is V2.1.0.0, It is an open source word components. The main features are:

Pangu word components can automatically identify some of the unknown word which is not in the dictionary.

Pangu word components can solve the problem of segmentation ambiguity according to frequency.

Pangu word components can effectively identify Chinese place names and names.

Punctuation, conjunctions, auxiliary word and other needs to be filtered out in Chinese word segmentation. Pangu word components offer a file named StopWord.txt. If users want to filter the words, they just need to add the words to the file.

Pangu segmentation provides a dictionary management tool named DictManag. Users can add, modify and delete dictionary words through this tool.

As used herein, the Chinese stop word list is released by Harbin Institute of Technology Social Computing and Information Retrieval Research Center, which contains 767 Chinese stop words.

## 4. Feature selection algorithm

Feature selection means to select the representative words from all words contained in a document to constitute the set of feature items [5]. The feature set must meet the following three requirements:

Feature items can accurately describe the document content.

Feature items can distinguish the target document from other documents.

Make the dimension of feature vectors as small as possible.

The feature selection algorithms commonly used are document frequency (DF), information gain (IG), mutual information (MI), chi-square (CHI), etc. The results show that: CHI algorithm is best, followed by the IG algorithm, MI algorithm and DF algorithm is the worst [6].

### 4.1 The chi-square statistic algorithm (CHI)

The chi-square statistic algorithm illustrates the importance of features by measuring the degree of correlation between feature items $t$ and category $c$. The prerequisite is assuming that the relationship between $c$ and $t$ meets $\chi^2$ distribution. It is shown in equations (1):

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

In the formula, $A$ is document frequency included in category $c$ and has the feature item $t$. $B$ is document frequency which has the feature item $t$ but is not included in category $c$. $C$ is document frequency included in category $c$ and does not include the feature item $t$. $D$ is document frequency does not be included in category $c$ and does not has the feature item $t$. $N$ is the total number of training document.

In the equations (1), $\frac{N}{(A + C)(B + D)}$ can be removed. We can use simplified formula when calculating:

$$\chi^2(t,c) = \frac{(AD - CB)^2}{(A + B)(C + D)} \tag{2}$$

High frequency characteristics that appear in any of the categories are determined having a contribution for judgment on the category during feature extraction using the chi-square statistic algorithm [7]. When $t$ and $c$ are independent, the value of $\chi^2$ is 0.The larger the value of $\chi^2$, the higher the degree of correlation between characteristics $t$ and category $c$.

For author/s of more than two affiliations: To change the default, adjust the template as follows.

Selection: Highlight all author and affiliation lines.

Change number of columns: Select the "Columns" icon from columns.

**4.2 Identify the Headings**

After analysis of the chi-square statistic algorithm, we found two flaws:

It ignores the influence of frequency of feature.

It tends to the features which observed value of document frequency is less than the theoretical value.

For these two shortcomings, this paper proposes improvement ideas and get improved chi-square statistic algorithm, referred to as NCHI algorithm.

This paper considers selecting feature items which positive correlation with the category can get better classification results. The chi-square statistic algorithm ignores the impact of word frequency. So in this paper, we has been improved the calculation of $\chi^2(t, c)$. The improved algorithm we called NCHI is shown in equations (3):

$$\chi^2(t, c) = \frac{(AD - CB)^2}{(A + B)(C + D)} \log fq(t, c) \tag{3}$$

Among them, $fq(t,c)$ is frequency of feature words $t$ in category $c$.The purpose of multiplied by the frequency in the formula is to improve the chi-square value of high-frequency words in the same category.

For the feature words in category $c$ which observation value of document frequency is less than the theoretical value, the punishment should be implemented. It aims to reduce the chi-square value of this type of feature words. As for the feature words which observation value of document frequency is greater than the theoretical value that we do not impose penalties. We assume that the document frequency of feature $t$ is equal to the observed values $A$ and the theoretical value of document frequency is $E = (A + C)\frac{A + B}{N}$, then the penalty function is shown in equations (4):

$$PH(t, c) = \begin{cases} 1 & , A \geq E \\ \dfrac{1}{|A - E| + 1} & , A < E \end{cases} \tag{4}$$

We get the equation (5) by introducing penalty function $PH(t, c)$ into the equation (3):

$$\chi^2(t, c) = \begin{cases} \dfrac{(AD - BC)^2}{(A + B)(C + D)} \log(t, c) & , A \geq E \\ \dfrac{(AD - BC)^2}{(A + B)(C + D)(|A - E| + 1)} \log(t, c) & , A < E \end{cases} \tag{5}$$

## 5. Naive Bayes text classification algorithm

Naive Bayes algorithm is a simple and effective classification algorithm [8]. It can predict the probability of a given sample belongs to a category. Naive Bayesian classification algorithm assumes that the characteristics and the impact of a given class independent of other characteristics [9].

That is feature independence assumption. Naive Bayes text classification method is more widely used because it is simple, computationally efficient, better classification results [10].

The text to be classified through word segmentation is divided into a set of word $A(a_1, a_2, \ldots, a_n)$, and then, it will compare the probability which belonging to each classification of $B(B_1, B_2, \ldots, B_m)$ and determine the category which $A(a_1, a_2, \ldots, a_n)$ belongs to. It is shown in equations (6):

$$P(B \mid a_1, a_2, \ldots, a_n) = \frac{P(a_1, a_2, \ldots, a_n \mid B_i) P(B_i)}{P(a_1, a_2, \ldots, a_n)} \tag{6}$$

Obviously, in the formula (6), the denominator of the right side is a constant. We only need to compare the probability of belonging to each category, so the equation (7) can be turned into the formula for solving the maximum:

$$C = \max_{1}^{m} P(a_1, a_2, \ldots, a_n \mid B_i) P(B_i) \tag{7}$$

Wherein $P(B_i)$ is the prior probability which is determined by the training text. Naive Bayes assumes that the distribution of characteristic properties $a_1, a_2, \ldots, a_n$ of the text is not relevant. We get equations (8) by introducing the joint probability distribution:

$$C = \max_{1}^{m} \left( \prod_{j} P(a_J \mid B_i) P(B_i) \right) \tag{8}$$

Here when the equation (8) was substituted into equation (6) we get equation (9) which is shown as follows:

$$P(a_1, a_2, \ldots, a_n \mid B_i) = \prod_{j} P(a_J \mid B_i) \tag{9}$$

The equation (9) can be used to determine the text to be classified is belong to that category which it is in.

## 6. Experimental Analysis

### 6.1 Training and test sets preparations

In experiments, we use Bayes algorithm and conducted two experiments to verify the advantage of improved text feature extraction algorithm. In this paper, we use news dataset provided by Sogou laboratory as Experimental corpus. We select six categories amount to 4500 corpus including 2756 training examples and 1744 test examples from the data set.

### 6.2 Evaluating index

In this paper, the three indicators are used to evaluate the classifier on a single class of classification performance. They are precision ( $p$ ), recall ( $r$ ), and $F1$ value [11].

$$p = \frac{a}{a + c} \tag{10}$$

$$r = \frac{a}{a + b} \tag{11}$$

Among them, parameter $a$ is the number of relevant records retrieved. Parameter $b$ is number of relevant records not retrieved. Parameter $c$ is the number of irrelevant records retrieved. $F1$ is weighted average of precision and recall. It is shown in equations (12):

$$F1 = \frac{2 \times r \times p}{r + p} \tag{12}$$

### 6.3 Experimental results and analysis

Bayes algorithms and the original chi-square statistic algorithm are used in the first set of experiments. Bayes algorithm and the improved chi-square statistic algorithm are used in the second set of experiments. The experimental results are shown in Table I.

The distribution of text category and the number of samples set. As we can see from Table I, the classification accuracy has been increased obviously when we used the chi-square statistic improved algorithm.

Table I

| No. | Feature selection algorithm | Average precision (%) | Average recall (%) | *F*1 Value (%) |
|---|---|---|---|---|
| 1 | CHI | 73.32 | 75.41 | 71.83 |
| 2 | NCHI | 82.26 | 81.31 | 82.12 |

## 7. Conclusions

This paper introduced the Chinese word segmentation process and studied the feature selection algorithm. It improved Chi-square statistic algorithm and designed two set of experiments to demonstrate the effect of the improved algorithm. The results show that the improved algorithm improves the accuracy of classification.


## Acknowledgment

The research work was financially supported by the Foundation for Beijing teacher team construction-youth with outstanding ability project (No.YETP1414), the twelfth five-year plan important subject of Beijing education science research (No. AJA11174) and the scientific research program of Beijing institute of fashion technology (No.2012A-17).


## References


[1] Kousu Ling: Research on feature selection in Chinese text classification. Computer Simulation, 2007, 24(3): 289-291(In Chinese)

[2] Jianming Cui, Jianming Liu, Zhouyu Liao: Research of text categorization based on support vector machine. Computer Simulation, 2013, 30(2): 299-302(In Chinese)

[3] Kangxin Fan: Design of NB combination text classifier based on various feature selection. Computer Engineering, 2009, 35(24): 191-193(In Chinese)

[4] Xiaoying Su, Yanpeng Hu, Junhui Yang, Ming Li: A new probabilistic classifier design for text categorization. Computer Technology and Development, 2014, 24(3): 46-48(In Chinese)

[5] Peng Di, Liguo Duan: New Naive Bayes text classification algorithm. Journal Data Acquisition and Processing, 2014, 29(1): 71-75 (In Chinese)

[6] Haifeng Liu, Zhan Su, Shousheng Liu: Improved CHI text feature selection based on word frequency information. Computer Engineering and Applications, 2013, 49(22): 110-114(In Chinese)

[7] Zhe Zhao, Yang Xiang, Jisheng Wang: Text classification based on parallel computing. Journal of Computer Applications, 2013, 33(S2): 60-62, 66(In Chinese)

[8] Juan Lai: Simulation research of text categorization based on data mining. Computer Simulation, 2011, 28(12): 195-198(In Chinese)

[9] Lin Mu: Performance comparison based on support vector machine algorithm and other algorithms in text categorization. Journal of Inner Mongolia University (Natural Science Edition), 2011, 42(6): 703-707(In Chinese)

[10] Zhe Wang: Study and comparation on feature selection method in Chinese text categorization. Journal of Computer Applications, 2011, 19: 18-20(In Chinese)

[11] Lin Chen, Jian Wang: Comparison and research on algorithms of three Chinese text classification. Computer and Modernization, 2012, 2: 1-4(In Chinese)