# Automatic Identification of Stop Words in Chinese Text Classification

Lili Hao
Institute of Mathematics
Jilin University
Lsystem@163.com

Lizhu Hao
School of Mathematics and Statistics
Northeast Normal University
Hsystem@163.com

## Abstract

*Text classification is an active research area in information retrieval and natural language processing. A fundamental tool in text classification is a list of 'stop' words(stop word list) that is used to identify frequent words that are unlikely to assist in classification and hence are deleted during pre-processing. Till now, many stop word lists have been developed for English language. However, there is no standard stop word list which has been constructed for Chinese text classification yet. In this paper, we give a refined definition for stop words in Chinese text classification from a perspective of statistical correlation, then propose an automatic approach to extracting the stop word list in text classification based on the weighted Chi-squared statistic on 2*p contingency table. We evaluate the stop word lists using accuracies obtained from text classification experiments in the real-world Chinese corpus. The results show that the proposed approach is effective. The stop word lists derived by the approach can speed up the calculation and increase the accuracy of classification at the same time.*

## 1. Introduction

A fundamental tool for improving the accuracy of classification techniques is the employment of stop word lists. In traditional view, words in documents that frequently occur but unlikely assist in classification are called stop words, such as "and", "the", and "of" in English documents. Stop words almost can be found virtually in every sentence. These words account for a very significant fraction of all text's size. We usually refer to this set of stop words as stop word list [1]. The use of stop word list is justified by research suggesting that over 50% of all words in a small typical English passage are contained within a list of about 135 common words [2]. These can be considered to be noise words as suggested by Van Rijsbergen [3], and should be removed as part of any pre-processing in text analysis experiments. Stop words make very minimum information for text classification task. Elimination of stop words could contribute to reduce the size of the text feature space considerably and help to speed up the calculation and increase the accuracy of text classification [4].

Till now many stop word lists have been developed for English language. The most commonly referred stop word lists are the Van Rijsbergen stop word list and the Brown corpus stop word list [3, 5]. They are widely used as English standard stop word lists. These stop word lists are traditionally extracted by frequency analysis of all the words in large corpus [6]. Stop words disparity exists for different document corpora and different text processing tasks. So many other methods for generating stop word lists automatically were proposed by researchers as well. For example, Lo et al. [7] proposed the term-based random sampling approach, which was introduced based on the Kullback-Leibler divergence measure. Sinka et al. [8] produced a new stop word list based on word-entropy. Different from English language, up to now, no commonly acceptable stop word list has been constructed for Chinese language. Some research work on Chinese information retrieval even makes use of manual stop word lists [9, 10], and other may automatically generate stop word list [11]. In Chinese text classification, the approach that automatically identifies the stop words is few. Many related works on Chinese text classification use a single fixed manual stop word list which can only remove a small portion of the stop words.

In this paper, we give a refined definition of stop word from the perspective of statistical correlation in terms of characteristic of text classification tasks. In order to save time and release the burden of manual stop words selection, we propose a new automatic approach to identify stop words, which is based on the weighted statistic on 2*p contingency table. We empirically demonstrate that stop word lists derived by our method are effective by comparing the performances of classification with and without stop word lists from a real-world Chinese corpus of the Mayor's Public Access Line Project texts.

978-0-7695-3336-0/08 $25.00 © 2008 IEEE
DOI 10.1109/CSSE.2008.829

The remainder of this paper is organized as follows: Section 2 gives a refined definition of the stop words, and proposes an automatic approach to identify the stop words, which is based on the weighted statistic on 2*p contingency table; Section 3 gives the experimental data set and the evaluation approach, then discusses and analysis the results. Finally, Section 4 provides our conclusions.

## 2. Identification of stop words based on the weighted Chi-squared statistic

With traditional definition of stop word, it is difficult to extract these words from documents using statistical methods. According to characteristics of given categories in text classification, we propose a new definition of stop word from the perspective of the correlation theory in statistics. A word that satisfies the following two conditions is called stop word:

(1) It has a high document frequency (DF);

(2) It has small statistical correlations with all the classification categories.

From the above definition of stop word, we can extract these words from documents using statistical methods.

### 2.1. Chi-squared statistic

In order to satisfy the second condition of the above definition of stop word, we use a statistic to measure statistical correlation between a word and classification categories.

Let $C = \{C_1, C_2, \cdots, C_p\}$ denotes classification categories. After word segmentation using the Backwards Maximum Matching algorithm, we use the Boolean vector space model to express each document vector. Obviously, word frequency is the same as document frequency. Assume that the appearance of the *r*th word $W_r$ is independent of any category, so the correlation between the word $W_r$ and classification categories could be researched by a 2*p contingency table, as shown in Table 1 below.

**Table 1. The 2*p contingency table of word $W_r$**

| | $C_1$ | $C_2$ | $\cdots$ | $C_p$ | $+$ |
|---|---|---|---|---|---|
| $W_r$ appears | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1p}$ | $n_{1+}$ |
| $W_r$ not appears | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2p}$ | $n_{2+}$ |
| $+$ | $n_{+1}$ | $n_{+2}$ | $\cdots$ | $n_{+p}$ | $N$ |

where

$$n_{+j} = \sum_{i=1}^{2} n_{ij},\ n_{i+} = \sum_{j=1}^{p} n_{ij}, \tag{1}$$

$$N = \sum_{i=1}^{2}\sum_{j=1}^{p} n_{ij} = \sum_{i=1}^{2} n_{i+} = \sum_{j=1}^{p} n_{+j}. \tag{2}$$

In Table 1, $n_{1j}$ denotes the number of documents containing word $W_r$ in the category $C_j$; $n_{2j}$ denotes the number of documents which don't contain word $W_r$ in the category $C_j$ ; $n_{+j}$ denotes the total number of documents in the category $C_j$; $n_{1+}$ denotes the number of documents in corpus containing the word $W_r$ ; $N$ denotes the total number of documents in the training set. To test whether a word is independent of all the classification categories, we can make use of the following Chi-squared statistic [12]

$$\chi^2_{2*p} = \sum_{i=1}^{2}\sum_{j=1}^{p} \frac{(Nn_{ij} - n_{i+}n_{+j})^2}{Nn_{i+}n_{+j}}. \tag{3}$$

We can sort the words by the $\chi^2_{2*p}$ values in increasing order. The smaller the $\chi^2_{2*p}$ value is, the weaker the correlation relationship with all the categories is. These words satisfy the second condition of the definition of stop word. Nevertheless, there are many low-frequency words and a few high-frequency words among the top ranked words in the list. For the data which is severely skewed and contains many short documents, the low-frequency words sometimes are the feature words. Deleting these low-frequency words will debase the accuracy of classification. Therefore, in order to satisfy the first condition of the definition of stop word, we add document frequency as weight factors of $\chi^2_{2*p}$ values to our model and extract stop word list based on the weighted $\chi^2_{2*p}$ values.

### 2.2. Weighted Chi-squared statistic

As the smaller the $\chi^2_{2*p}$ value is, the stronger the independence is. We use the reciprocal of these $\chi^2_{2*p}$ values and consider a weight factor of every $\frac{1}{\chi^2_{2*p}}$ value related to the document frequency of each word. The higher the document frequency is, the higher the weight is. Let $DF_r$ denote the document frequency of the *r*th word, we definite the weight as follows

$$\frac{DF_r}{\sum DF_r}, \tag{4}$$

where $\sum DF_r$ is the unitary factor, then we give the following formula

$$\frac{DF_r}{\sum DF_r}\frac{1}{\chi^2_{2*p}}, \tag{5}$$

which is equivalent to

$$\chi^2_{weighted} = \frac{\chi^2_{2*p}}{DF_r}. \tag{6}$$

$\chi^2_{weighted}$ is called weighted Chi-squared statistic. This statistic balances the strength of the dependent relationship between a word and all categories and document frequency

of a word. Subsequently, all words are increasingly ordered according to the value of weighted Chi-squared statistic. The first word in the ordered list has the minimum value of weighted Chi-squared statistic, i.e. it has a higher document frequency and lesser correlations with all the categories. Therefore it is a stop word. In order to ascertain size of stop word list in the ordered list, a threshold needs to be determined. The threshold can be determined by cross-validation experiments [13].

## 3. Experiment Results

With the above theoretical foundations, we resort to a Naive Bayes classifiers to classify Chinese corpus of the Mayor's Public Access Line Project texts for the purpose of evaluating above provided approach.

### 3.1. Naive Bayes classifiers

To evaluate the performance of the stop word lists constructed by our method, we use the Naive Bayes classifier [14] to classify the data. Let D denote the training text data set, and $d = (W_1, \cdots, W_n)$ denotes a test document, where $W_j (1 \leq j \leq n)$ the *j*th word in the document and n the number of different words in the training set. Let C denote the document categories set, where $C = \bigcup_{k=1}^{p} C_k$, and p is the total number of categories. By the assumption of the conditional independence of the Naive Bayes model, to determine the category of the test document, the Naive Bayes classifier needs to satisfy

$$C_{NB} = \arg \max_{C_i \in C} P(C_i) * \prod_{j=1}^{n} P(W_j|C_i), \qquad (7)$$

where $P(C_i)$ is the a priori probability of category $C_i$ and $P(W_j|C_i)$ is the conditional probability of word $W_j$ given category $C_i$.

### 3.2. Data Set

This research focuses on the real-world Chinese corpus of the Mayor's Public Access Line Project texts. The telephone is a hot line set by the city government with a special number of 12345, via which citizens may express their complaints, suggestions or praises etc.to the receivers, who in turn record these as texts and send them to relevant departments for handling. In the corpus, there are two kinds of classifications, i.e., assigned departments laterally and assigned professions vertically. In this paper, the samples are classified by professions, including totally 41 professions such as the appearance of the city, environmental protection, water supplement, industry and commerce, and so on. The data set includes 98465 documents, which have been labelled manually. The data is severely skewed on professions. For example, the samples corresponding to the appearance of the city occupies 1/10 of the total samples, while the least one occupies less than 1/10000. The average length of text document in corpus is 65 word, which belongs to short document.

### 3.3. Performance measures

To evaluate a text classification system, we use the several measures introduced by Van Rijsbergen [3]. For the problem of two class classification, the measures are Precision, Recall and F1. Their definitions are as follows:

$$\text{Recall}=\frac{\text{number of correct positive predictions}}{\text{number of positive examples}}, \qquad (8)$$

$$\text{Precision}=\frac{\text{number of correct positive predictions}}{\text{number of positive predictions}}, \qquad (9)$$

$$\text{F1}=\frac{2 \times \text{Recall} \times \text{Precision}}{(\text{ Precision + Recall })}. \qquad (10)$$

To evaluate the performance of multi-class classification algorithm, there are two conventional methods, namely macro-averaging and micro-averaging [15]. Micro-average is considered as a per-document average (more precisely, an average over all the document/category pairs). Likewise, macro-average is considered as a per-category average. The most often used performance evaluations are micro-averaging precision, micro-averaging F1, macro-average precision, macro-average recall and so on.

### 3.4. Experimental results

We use the approach of identifying the stop words to Chinese corpus of the Mayor's Public Access Line Project texts. This corpus consists of 87540 labelled training samples and 10925 labelled testing samples. We extract the stop word lists (see Table 2) by the weighted Chi-squared statistics in the training data set, and we make some analysis and comparison for the testing data set based on the lists obtained, considering deleting/not deleting the stop words respectively (see Tables 5 and 7).

**Table 2. The stop word list in profession (weighted Chi-squared order).**

| Word | H01 | $\cdots$ | DF | Chi | WChi |
|---|---|---|---|---|---|
| of | 367 | $\cdots$ | 4031 | 339.183 | 0.08414 |
| relevant department | 82 | $\cdots$ | 1445 | 184.119 | 0.12742 |
| handling | 68 | $\cdots$ | 1159 | 218.247 | 0.18831 |
| reflect | 350 | $\cdots$ | 5646 | 1070.189 | 0.18955 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

In Table 2, H01, $\cdots$ are profession codes; DF denotes the total document frequency in training texts; Chi denotes the Chi-squared value; WChi denotes the weighted Chi-squared value obtained by the weighted Chi-squared statistic. From Table 2, we can see that the results sorted by the values of the weighted Chi-squared statistic are more intuitional. The high-frequency words with small correlation relationship among all the categories are sorted in front, which are the complimentary words that often appear in the complaint document, such as "the relevant department","‘ handling" and so on. We use different numbers of stop words to construct the stop word lists, and test the performance based on the Naive Bayes classifier (see Table 3 ).

**Table 3. Performance comparison in Naive Bayes classifier of profession.**

| | The numbers of stop words deleted | | | | |
|---|---|---|---|---|---|
| Number | 0 | 50 | 400 | 500 | 800 |
| $R_{micro}$ | 83.48% | 87.51% | 88.93% | 88.98% | 88.02% |
| Percentage | 0% | 20.95% | 41.06% | 43.71% | 50.01% |

In Table 3, $R_{micro}$ is micro-average recall of the Naive Bayes classifier after deleting stop words, which is the percentage of the total stop words among all the words( include repeat words). From the testing results, we can see that the classifying accuracy is improved greatly after deleting 50 stop words, and the effect is most excellent when deleting 500, but it will fall if deleting more.

The reasons of improving the classification accuracy after deleting 500 stop words are: the 500 stop words deleted appear in the training text for 83674 times, and occupy 43.71% of 191433 which is the frequency of all the words. Although the stop words are only a small fraction in all the 13909 words (not repeating words), deleting them can decrease the number of all words in the texts by 43.71%, and if we delete 800 stop words, it will delete 50.01% of those. So it is obvious that the high-frequency words have small effect on classifying which seriously affect the classifier's performance. By deleting or not deleting the stop words, the effect of each category by the Naive Bayes classifier is also different (see Table 4 and Table 6).

**Table 4. Results of Naive Bayes classifier of profession with stop words not deleted.**

| No | IY | TS | CS | CI | P(%) | R(%) | F1(%) |
|---|---|---|---|---|---|---|---|
| 1 | LC | 1128 | 1529 | 1109 | 72.53 | 98.32 | 83.48 |
| 2 | EP | 1122 | 1245 | 1094 | 87.87 | 97.50 | 92.43 |
| 3 | WS | 917 | 1138 | 909 | 79.88 | 99.13 | 88.47 |
| 4 | HS | 748 | 820 | 734 | 89.51 | 98.13 | 93.62 |
| 5 | PL | 606 | 676 | 554 | 81.95 | 91.42 | 86.43 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | |
| 40 | BS | 6 | 1 | 0 | 0 | 0 | 0 |
| 41 | WP | 4 | 0 | 0 | 0 | 0 | 0 |

In Table 4, IY denotes the profession code; TS denotes the testing samples number in a certain profession ; CS denotes the sample numbers assigned by the classifier; CI denotes the number of the testing samples being assigned correctly in a certain profession ; P: Precision; R: Recall; F1: F1 measure; LC: the appearance of the city; EP: environmental protection; WS: water supply; HS: heat supply; PL: peasants life; BS: banks; WP: work proposal.

**Table 5. Synthetical evaluation of Naive Bayes classifier with stop words not being deleted.**

| Micro-average | | | Macro-average | | |
|---|---|---|---|---|---|
| P | R | F1 | P | R | F1 |
| 84.73% | 83.48% | 81.39% | 72.31% | 51.61% | 54.44% |

Obviously, the micro-average is the average of documents, which can be easily affected by large classes; while the macro-average is the average of classes, which can be easily affected by small clusters.

**Table 6. Naive Bayesian classifier of profession (500 stop words deleted).**

| No | IY | TS | CS | CI | P(%) | R(%) | F1(%) |
|---|---|---|---|---|---|---|---|
| 1 | LC | 1128 | 1093 | 1106 | 92.04 | 89.18 | 90.59 |
| 2 | EP | 1122 | 1105 | 1064 | 96.29 | 94.83 | 95.55 |
| 3 | WS | 917 | 921 | 882 | 95.77 | 96.18 | 95.97 |
| 4 | HS | 748 | 749 | 727 | 97.06 | 97.19 | 97.12 |
| 5 | PL | 606 | 638 | 550 | 86.21 | 90.76 | 88.43 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | |
| 40 | BS | 6 | 3 | 0 | 0 | 0 | 0 |
| 41 | WP | 4 | 2 | 1 | 50 | 25 | 33.33 |

**Table 7. Synthetical evaluation of Naive Bayes classifier of profession two (500 stop words deleted).**

| Micro-average | | | Macro-average | | |
|---|---|---|---|---|---|
| P | R | F1 | P | R | F1 |
| 88.93% | 88.98% | 88.76% | 72.34% | 67.51% | 68.83% |

From Table 4 and Table 6, we can see that the classification accuracy is greatly influenced by the deleting of the stop words. For example, when we do not delete the stop words, the sample number assigned by the classifier of 'the appearance of the city' is 1529. And it becomes 1093 after deleting the stop words, with the precision changing from 72.53% to 92.04%. It shows that there is skew in the data. The samples from 'the appearance of the city' take up 1/10 of the total samples, so by the effect of stop words, more documents are classified falsely into 'the appearance of the city'. After deleting the stop words, the noise of the data decreases, and the precision increases greatly at the same time. From Table 5 and Table 7, we can see that after deleting 500 stop words, the micro-average precision improves nearly 4% from 84.73% to 88.93%. Micro-average F1 is the joint consideration of the micro-average precision and the micro-average recall, whose values are improved nearly 7% from 81.39% to 88.76% after deleting. Moreover, the

relatively low values of the macro-average show that more categories and the skewed data put badly negative influence on the classifier.

## 4. Conclusion

In this paper, we propose a new approach to automatically generate a stop word list for a given Chinese corpus, under the definition of stop word from a perspective of statistical correlation. We have employed a weighted Chi-squared statistic based on 2*p contingency table measure in order to identify potential stop words for the Chinese corpus, which we would expect to be able to identify high frequency words which have small statistical correlations with all the categories as the stop words, and we regard these words as the noise words of the classifier.

We evaluate stop word lists by using accuracies obtained from text classification experiments in Chinese corpus of the Mayor's Public Access Line Project texts. Our experiments compare and analyze the results of classifiers constructed by deleting and retaining the stop words. Thus come to the conclusion that the stop word list involving 500 words constructed by the foregoing experiment can reduce the words in profession data by 43% of all the words in corpus, and the micro-average F1 improves nearly 7% from 81.39% to 88.76%. The results show that methods given in this paper save time, release the burden of manual stop words selection, and also speed up the calculation and increase the accuracy of classification at the same time.

## References


[1] P. S. Mark, and W. C. David. Evolving better stoplists for document clustering and web intelligence. *Proceedings of the 7th WSEAS Int'l Conf. on Artificial Intelligence*, 1015–1023, 2003.

[2] G. W. Hart. To decode short cryptograms. *Communications of the ACM*, New York: Association for Computing Machinery, 37(9):102–108, 1994.

[3] C. J. Van Rijsbergen. *Information retrieval*. London: Butterworths, 1979.

[4] C. Silva, and B. Ribeiro. The importance of stop word removal on recall values in text categorization. *Neural Networks*, 320–24, 2003

[5] C. Fox. Lexical analysis and stoplists. *Information Retrieval: Data Structures and Algorithms*, Upper Saddle River, New Jersey: Prentice Hall, 102–130, 1992.

[6] Y. M. Yang. Noise Reduction in a Statistical Approach to Text Categorization. *In Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval* (SIGIR'95), 256–263, 1995.

[7] R. Lo, B. He, and I. Ounis. Automatically Building a Stopword List for an Information Retrieval System. *Proceedings of the 5th Dutch-belgian Information Retrieval Workshop*, Utrecht, the Netherlands, 141–148, 2005.

[8] M. P. Sinka, and D. W. Corne. Web intelligence WI 2003. *Proceedings IEEE/WIC International Conference on Soc*. Los Alamitos: IEEE Comput, 396–402, 2003.

[9] K. H. Chen, and H. H. Chen. Cross-Language Chinese Text Retrieval in NTCIR Workshop: towards Cross-Language multilingual Text Retrieval. *ACM SIGIR Forum*, 35(2): 12–19, 2001.

[10] H. Nakagawa, H. A. Kojima, and Maeda. Chinese term extraction from web pages based on compound word productivity. *IJCNLP*, 269–279, 2005.

[11] F. Zou, F. L. Wang, X. S. Deng, Han, and L. Wang. Stop Word List Construction and Application in Chinese Language Processing. *WSEAS Transanction on Information Science and Applications*, 3(6): 1036–1045, 2006.

[12] Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2002.

[13] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3): 103–134, 2000.

[14] P. Langley, W. Iba, and K. Thompson. An Analysis of Bayesian Classifiers. *Proc. of the 10th National Conf. on Artificial Intelligence*, Menlo Park, AAAI Press, 223-228, 1992.

[15] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1): 76–88, 1999.