

# 带有短语切分的中文文本分类方法

田昕辉, 李成基

(国立庆北大学 计算机工学系, 韩国 大邱 702-701)

**摘要:** Internet 文本信息量极速增加, 在组织和处理这些文本数据时, 文本分类技术显得尤为重要。利用统计学理论, 特征提取和权重计算常常忽略了特征项之间的语法关系。文中提出了一种将短语切分与文本分类相结合的新方法。在经过 TFIDF 计算之后, 在同一个短语中, 特征项之间的关系被计算出来, 然后调整权值向量, 最后可以得到文本分类的正确率。同一般地文本分类方法相比, 加入短语切分的文本分类方法的正确率平均提高了 1.5% 以上。

**关键词:** 特征提取; 文本分类; 短语切分; 权值调整

中图分类号: TP391.1

文献标识码: A

文章编号: 1673-629X(2010)01-0009-05

## Phrase Segmentation for Chinese Text Classification

TIAN Xin-hui, LEE Sung-kee

(Department of Computer Engineering, Kyungpook National University, Daegu 702-701, Korea)

**Abstract:** With the rapid growth of textual information on Internet, text classification has become a more important key technology in organizing and processing large amount of document data. General statistics method of feature extraction and weight calculation ignores the syntax relationship between terms. A new method of how the phrase segmentation is brought into text classification is discussed. After TFIDF is calculated, the relationship between features in a phrase is evaluated. Then coordinate the weight vector with the relationship information. Finally, the accuracy of text classification is evaluated. Compared with the general method of text classification, the improvement of accuracy of text classification with phrase segmentation is increased by 1.5%.

**Key words:** feature extraction; text classification; phrase segmentation; weight coordination

## 1 Introduction

With the development of Internet and information technology, the problem of "Information Explosion" has arisen. It is getting more important to organize and manage electronic documents. Text classification will help people to deal with the huge and disordered documents quickly, accurately, and roundly. The general method of text classification just uses statistical method to extract features and calculate weights associated with features. That method ignores the relationship between word and word.

The paper puts forward a new method of how a phrase is brought into weight calculation to coordinate it in order to improve the accuracy of text classification<sup>[1]</sup> is brought.

## 2 Technology of Text Classification

Text classification is the task of classifying a document under a predefined class. Formally, if  $d_i$  is a document of the entire set of documents  $D$  and  $\{c_1, c_2, \dots, c_n\}$  the set of all the classes, text classification assigns one class  $c_j$  to a document  $d_i$ <sup>[2]</sup>.

Generally, Chinese text classification has the following processes: word segmentation, feature extraction, weight calculation and classification. Finally, we can get the result of text classification.

### 2.1 Word Segmentation

In Chinese language, there is no segmental symbol between two adjacent words, so we need to segment words from the Chinese characters. Mainly, there are 3 kinds of word segmentation method<sup>[3,4]</sup>: string matching method, segmentation based on understanding and the statistic word segmentation method. ICTCLAS is adopted to deal with word segmentation.

ICTCLAS is based on HMM and works at five lev-

收稿日期: 2009-10-14; 修回日期: 2009-11-18

作者简介: 田昕辉(1983-), 男, 硕士, 研究方向为自然语言处理; 李成基, 教授, 研究方向为计算机视觉和个人医疗设备。

els: atom segmentation, simple and recursive unknown word recognition, class - based segmentation and POS tagging<sup>[5,6]</sup>. The speed of segmentation is more than 1000 bytes/s and the recall is over 97% with a precision of over 94%<sup>[5]</sup>.

## 2.2 Feature Extraction

The purpose of feature extraction method is the reduction of the dimensionality of the training corpus by removing the features considered irrelevant to the classification<sup>[2]</sup>. The method for feature extraction for the Chinese text classification task uses an evaluation function applied to a single word. All words are independently evaluated and sorted according to the assigned criterion. The scoring of individual words can be performed using certain measures: for instance, document frequency, term frequency, mutual information, information gain, odds ratio,  $\chi^2$  statistic and term strength<sup>[2]</sup>. 4 methods are mainly introduced here.

### 2.2.1 Document Frequency (DF)

It's the simplest feature extraction method. The value is the document frequency that a certain word occurs in corpus<sup>[3,7]</sup>.

$$DF(f) = \frac{\text{Freq}(f)}{N}$$

Where  $\text{Freq}(f)$  is the number of document that feature  $f$  occurs in corpus and  $N$  the number of document in corpus.

### 2.2.2 Information Gain (IG)

IG is the average information of a text class, in which some features occur in the document in the text class. The formula of IG is as follows:

$$IG(f) =$$

$$\sum \left[ p(c, f) \log \frac{p(c, f)}{p(c)p(f)} + p(c, \bar{f}) \log \frac{p(c, \bar{f})}{p(c)p(\bar{f})} \right]$$

where  $c$  is variable of text classes;  $C$  the set of text classes;  $d$  a document and  $f$  a feature<sup>[8]</sup>.

### 2.2.3 Mutual Information (MI)

MI is the concept of information theory. In the area of feature extraction, class  $c$  and feature  $f$  will measure the interdependence between  $f$  and  $c$  in MI method. The formula is the following<sup>[8,9]</sup>:

$$MI(c, f) = \log \frac{p(c, f)}{p(c)p(f)}$$

### 2.2.4 $\chi^2$ (CHI - square)

$\chi^2$  is developed from the statistical test of the hypothesis. In text classification, given a two - way

contingency table for each feature  $f$  and the class  $c$  as represented in table1,  $\chi^2(c, f)$  is calculated as follows<sup>[3,9~11]</sup>:

Table1. Two - way contingency table<sup>[9]</sup>

	presence of $f$	absence of $f$
labeled as $c$	$A$	$C$
not labeled as $c$	$B$	$D$

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

## 3 Phrase Segmentation and Weight Coordination

Generally, a natural language has its special characteristics. For a sentence, single words are not independent on each other. They play certain roles in the sentence. Furthermore, a sentence consists of the subject, predicate, object, attribute, adverbial modifier, etc. Words are divided into some independent parts. In a sense, words have tight relationship between each other. In other words, they impact each other and express a certain meaning when they combine each other. Thus, we suppose that if two features of a document occur in a certain phrase, they will be more important than others. Because of this, we increase the weight of those two features using a certain method of weight coordination. Then, we design the system of phrase segmentation for the Chinese text classification and implement it. After experiments, the results show that the assumption can improve the accuracy of text classification. In this case, we need to solve the following problems: how to segment phrases and how to coordinate weights.

### 3.1 Phrase Segmentation

In the Chinese language, word is defined clearly; in other words, we can understand each of them easily, because each word have a clear partition. However, phrase has no clear partition and the definition is ambiguous. In this paper, the phrase matching rule is used to obtain phrases. After ICTCLAS word segmentation, we can get the result of word segmentation and part - of - speech (POS). Last, the phrase matching rules are shown as follows:

Figure 1 and 2 point out the phrase matching rules, where the upper characters are phrase tags and lower are POS tags. Figure 1 shows how to generate phrase tags from POS tags directly, and figure 2 shows how to gener-

ate tags from POS tags and phrase tags.

NP->n	NP->r	AVP->dg
NP->an	VP->vg	AVP->vd
NP->ng	VP->v	TP->t
NP->nr	VP->vn	TP->p t
NP->nr	AP->ag	SP->s
NP->ns	AP->a	SP->p s
NP->nt	AP->an	DP->m q
NP->nz	AP->b	
NP->vn	AVP->ad	

Figure1. First set of phrase matching rules<sup>[12]</sup>

NP->AP NP	VP->PP VP	PP->p NP
NP->DP NP	VP->VP u	PP->p NP f
NP->NP NP	VP->TP VP	PP->pp f
VP->AVP VP	AP->AP '的'	DJ->VP '的'
VP->VP NP	AVP->AVP AVP	DJ->NP '的'
VP->VP DP	AVP->p NP	TP->TP TP
VP->VP VP	AVP->AP v	

Figure2. Second set of phrase matching rules<sup>[12]</sup>

Now, the algorithm of matching phrases (figure3 shows phrase matching) is described as follows.

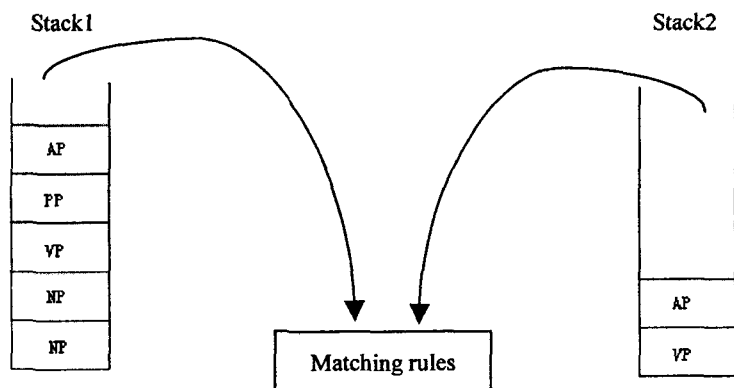


Figure3. Phrase matching

After word segmentation and POS tagging, we get all of the words and their POSs. A phrase is relative to the sentence, so we should segment it. 4 interpunctuations are defined to segment the sentence: “,”, “。”, “:” and “?”. The phrase segmentation algorithm is as follows:

1) Build two stacks: stack1 and stack2.

2) Input a certain sentence and use the first set of phrase matching rules to label the phrase tags. And push the matching result into stack1 sequentially by the sentence order.

3) If stack2 is empty, pop the top of stack1 and push it into stack2.

4) Match the top of stack1 and that of stack2 with the second set of phrase matching rules. If matched, pop

them and combine them with the new phrase and new phrase tag. Then the new phrase and its new tag are pushed into stack2. If not matched, pop the top of stack1 and push it into stack2.

5) Match the rules until stack1 is empty.

6) Pop all of elements in stack2, which are the result of phrase segmentation.

### 3.2 Weight Coordination

In the test processing, after phrase segmentation and weight calculation using TFIDF, weights are coordinated according to phrases.

Assuming a document is  $D$ , its weight vector is  $W = (w_1, w_2, \dots, w_n)$ ,  $n > 0$ , and the phrase of the document is  $P = (p_1, p_2, \dots, p_m)$ ,  $m > 0$ , if two features  $f_i$  and  $f_j$ , associated to  $w_i$  and  $w_j$  ( $w_i, w_j \in W$ ), occur in  $p_k \in P$ ,  $k = 1, \dots, m$ , the two weights  $w_i$  and  $w_j$  will be coordinated, since we believe the two features are more important if the two features occur in the same phrase in the document.

Coordinating method:

1) Calculate the average of weights:  $\bar{w} =$

$$\frac{1}{n} \sum_{i=1}^n w_i, i = 1, \dots, n;$$

2) Calculate the sum of weights:

$$\text{sum}(W) = \sum_{i=1}^n w_i, i = 1, \dots, n;$$

3) If  $(w_i - \bar{w})/\bar{w} < -0.618$ , set the  $w_i$  to 0;

4) If  $(w_i - \bar{w})/\bar{w} \geq -0.618$  and  $(w_j - \bar{w})/\bar{w} \geq -0.618$ , and the features of  $w_i$  and  $w_j$  occur in  $p_k$ , coordinate the weights using

the following formula:

$$w'_i = \text{Smoothing} \cdot w_i / \text{sum}(W) \cdot \bar{w} + w_i$$

where Smoothing is a smoothing value. If we assign different values of Smoothing, the result of text classification will be changed. In this paper, the following smoothing values are chosen to coordinate weights: 0.5, 1.0, 1.5, 2.0, 2.2, 2.5 and 3.0. With different smoothing values, we can get different results of text classification with phrase segmentation.

## 4 Experiment and Result

Evaluation has an important role in automatic text classification. In the paper, the following method is chosen to evaluate the text classification system:

Accuracy = TP / number of sample documents  
where TP is the true false positive documents, which shows the number of documents in the class and the system classifies as belonging. Compared with general process of Chinese text classification, the phrase segmentation and weight coordination are brought into our system.

#### 4.1 Corpus

The corpus of Wuhan University, published on the Internet, is chosen. To do the experiment, 1876 training documents and 931 test documents are selected. These two sets of corpus is divided into ten classes, including art, computer, environment, economy, education, politics, medicine, military, sports and traffic. After the feature extraction with  $\chi^2$ , IG, DF and MI, four feature lists will be calculated from the train corpus. And then the weight vectors are evaluated with TFIDF<sup>[13]</sup>. The test corpus is pre-classified for testing the system only.

#### 4.2 Experiment and Result

In the experiment, four methods of feature extraction, including  $\chi^2$ , IG, DF and MI to extract the features are used and SVM<sup>[14,15]</sup> and KNN<sup>[15~17]</sup> are used to test the system of text classification.

Table 2 shows the result of accuracy of text classification after our system of phrase segmentation for Chinese text classification and non - phrase segmentation are used while figure 4 and figure 5 show the trend with different smoothing values and feature extraction.

With SVM classifier, compared with non - phrase segmentation, the accuracies are improved apparently using different feature extraction methods. The improvement can reach about 2.47% with the smoothing value e-

qual to 2.0, 2.2 and 2.5 using IG. Using other feature extraction methods, the improvement can reach over about 1.5%.

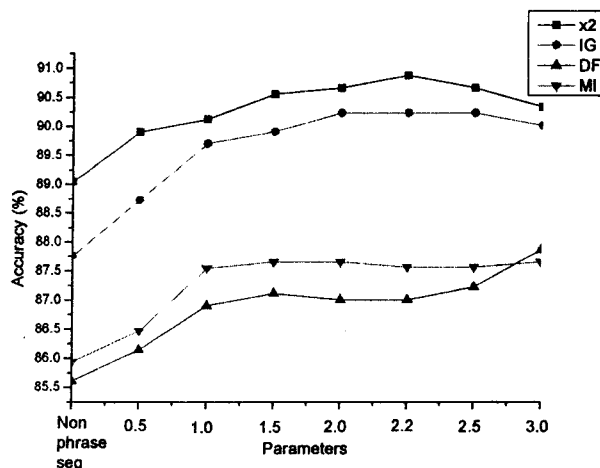


Figure4. Accuracies of different smoothing values with SVM

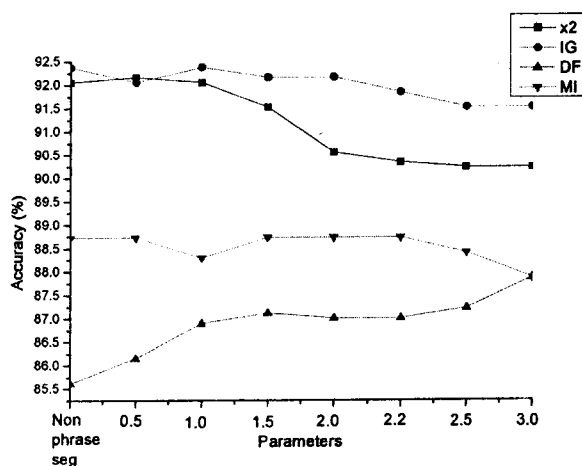


Figure5. Accuracies of different smoothing values with KNN

Table2. Accuracy of text classification

Accuracy (%)

		Without phrase seg	With phrase seg						
			0.5	1.0	1.5	2.0	2.2	2.5	3.0
SVM	$\chi^2$	89.044	89.90	90.12	90.55	90.655	90.87	90.655	90.33
	IG	87.7551	88.7218	89.6993	89.9033	90.2256	90.2256	90.2256	90.0107
	DF	85.6069	86.1439	86.8958	87.1106	87.0032	87.0032	87.218	87.8625
	MI	85.9291	86.4662	87.5403	87.6477	87.6477	87.5625	87.5625	87.6477
KNN	$\chi^2$	92.0516	92.1590	92.0516	91.5145	90.5478	90.333	90.2256	90.2256
	IG	92.3738	92.0516	92.3738	92.159	92.159	91.8367	91.5145	91.5145
	DF	85.6069	86.1439	86.8958	87.1106	87.0032	87.0032	87.218	87.8625
	MI	88.7218	88.7218	88.2922	88.7218	88.7218	88.7218	88.3996	87.8625

However, with a KNN classifier, compared with non - phrase segmentation, the accuracies are not improved greatly compared to that without phrase segmentation, and it even decreases. For example, using  $\chi^2$  feature extraction method, the accuracy is improved by about 0.1% only. And using DF, the improvement increases more than others, reaching about 1.5%.

## 5 Conclusion

Now, we have got the accuracy of text classification with different methods of feature extractions and classifiers. SVM and KNN improve the accuracy of classification, especially for some classes. Based on SVM classifier, using different methods of feature extraction, the accuracies are improved better compared to that without phrase segmentation. Using a KNN classifier, the accuracies are not improved greatly compared to that without phrase segmentation, and it even decreases. It expresses that text classification with phrase segmentation is not suitable for a KNN classifier, since the accuracy does not increase well; it even decreases.

### Reference:

- [1] Tan Songbo. Research on High - Performance Text Categorization[D]. Beijing: Institution of Computing Technology, Chinese Academy of Sciences, 2006.
- [2] Ikonomakis M, Kotsiantis S, Tampakas V. Text Classification: A Recent Overview[C]// In ICCOMP'05: Proceedings of the 9th WSEAS International Conference on Computers. [s.l.]:[s.n.], 2005:1 - 6.
- [3] 丁 琼. 基于向量空间模型的文本自动分类系统的研究与实现[D]. 上海:同济大学, 2007.
- [4] 王 雷. 文本分类相关技术研究[D]. 上海:复旦大学, 2006.
- [5] Zhang Hua - Ping, Liu Qun. Model of Chinese Words Rough Segmentation Based on N - Shortest - Paths Method[J]. Journal of Chinese Information Processing, 2002, 16(5):1 - 7.
- [6] Zhang Hua - Ping, Uu Hong - Kui, Xiong De - Yi, et al. HHMM - based Chinese Lexical Analyzer ICTCLAS[C]// 2nd SIGHAN workshop affiliated with 41th ACL. Sapporo, Japan:[s.n.], 2003:184 - 187.
- [7] Dong Mei, Hu Xue - gang. Text Categorization Based on Multiple Features Selection[J]. COMPUTER TECHNOLOGY AND DEVELOPMENT, 2007, 17(7):117 - 119.
- [8] Zheng Wei, Wang Rui. Comparative Study of Feature Selection in Chinese Text Categorization[J]. Journal of Hebei North University: Natural Science Edition, 2007, 23(6):51 - 54.
- [9] Xu Qinan, Liu Zhijing. Automatic Chinese Text Classification Based on NSVMDT - KNN[C]// Fifth International Conference on Fuzzy Systems and Knowledge Discovery. [s.l.]:[s.n.], 2008:410 - 414.
- [10] Kim Sang - Bum, Han Kyoung - Soo, Rim Hae - Chang, et al. Some Effective Techniques for Naive Bayes Text Classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(2):1457 - 1465.
- [11] 翁金象. 中文文本分类研究[D]. 济南:山东大学, 2007.
- [12] 朱国华. 文本信息处理中汉语句法分析方法研究[D]. Dalian: Dalian University of Technology, 2005.
- [13] SALTON G, BUCKLEY C. TERM - WEIGHTING APPROACHES IN AUTOMATIC TEXT RETRIEVAL[J]. Information Processing & Management, 1988, 24(5):513 - 523.
- [14] Wang Jing, Yao Yong, Liu Zhijing. A New Text Classification Method Based on HMM - SVM[C]// 2007 International Symposium on Communication and Information Technologies. [s.l.]:[s.n.], 2007:1516 - 1519.
- [15] Yuan Pingpeng, Chen Yuqin, Jin Hai, et al. MSVM - kNN: Combining SVM and k - NN for Multi - Class Text Classification[C]// IEEE International Workshop on Semantic Computing and Systems. [s.l.]:[s.n.], 2008:133 - 140.
- [16] Fix E, Hodges J L. Discriminatory analysis - nonparametric discrimination: Consistency properties[R]. Randolph Field, Texas: USAF School of Aviation Medicine, 1951:261 - 279.
- [17] Hao Xiulan, Zhang Chenghong, Tao Xiaopeng, et al. Accurate kNN Chinese Text Classification via Multiple strategies [C]// Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007). [s.l.]:[s.n.], 2007:504 - 508.

### 国立庆北大学简介:

庆北大学是 1946 年由三所大学合并而成的。在 1951 年正式改编为国立综合大学。在过去的 60 年里培养了 15 万各界人才, 现有 1 万 9 千余名本科和 5 千余名硕士、博士在校生。本校由 14 个学院、2 个直属学院(自由专业部和电子电气计算机学部)和 11 个研究生院组成。在 2004 年韩国地区革新展览会上, 首次获得了产学协力最优秀大学的总统奖, 并从 2003 年开始, 连续 3 年在中国上海交通大学公布的世界 500 强名校里被列为地区重点国立大学, 成为世界瞩目的名门大学。