Term Selection with Distributional Clustering for Chinese Text Categorization using N-grams

Jyh-Jong Tsay and Jing-Doo Wang
Department of Computer Science and Information Engineering
National Chung Cheng University
Chiayi, Taiwan 62107, ROC.
{tsay, jdwang}@cs.ccu.edu.tw
TEL:886-5-2720411.EXT.6207, FAX:886-5-2720859

Abstract

In this paper we propose an SB-tree approach to extract significant patterns efficiently by scanning the leaves of the SB-tree to decide the boundary of significant patterns for term extraction, and reduce the dimension of term space to an practical level by a combination of term selection and term clustering. Our current experiment uses CNA one year news as training data, which consists of 73,420 articles and is far more than previous related research. In the experiment, we compare the performance four term selection methods, odds ratio, mutual information, information gain and χ^2 statistic, when they are combined with distributional clustering method. Our experiment shows that χ^2 statistic and information gain achieve performance better than odd ratio and mutual information when they are combined with distributional clustering. With the combination of term selection and term clustering, the dimension of term space can be greatly reduced from 60000 to 120 while maintaining similar classification accuracy.

Keywords: Text Categorization, Term Selection, Term Clustering, Naive Bayes Classifier, Information Retrieval.

1 Introduction

Text classification(categorization) is the problem of automatically assigning predefined classes to free text documents, and is gaining more and more importance as the amount of text data available on World Wide Web grows dramatically. A well classified text database will be very helpful for a user to identify interesting data from the huge collection of texts. There are many studies about the text classification as well as web-page classification [17, 1, 9, 10, 27, 32, 33, 23, 24, 7, 38, 15]. While there are a great number of researches on automatic text classification for English texts, text classification for Asian languages such as Chinese, Japanese, Korean and Thai has not been studied seriously until recently [36, 21, 37, 3, 28, 31, 29].

Because text segmentation is not straightforward in Asian languages, 1-grams, 2-grams and n-grams have been used as indexing terms to represent documents. It is reasonable that n-gram is more meaningful and brings more concept than 1-gram or 2-gram. The main obstacle to apply n-grams to Chinese text classification is the huge number of possible n-grams. Notice that many of them are meaningless and non-informative for text categorization. The major challenge is to develop an approach that can reduce the dimension of term space to an acceptable level while maintains similar classification accuracy. There was a related study about term selection in Chinese text classification[29]. A practical problem there is that a news may contain very few or even non of the selected terms, and thus is classified to the default class which is the largest class. On the other hand, a large number of selected terms make Chinese text classification computationally impractical. To overcome the problems, we study the combination of the term(feature) selection and term clustering in this paper. We first use term selection to select a set of significant terms, and then use term clustering to cluster the selected terms into a small number of groups. Our experiment on one year CNA news shows that the dimension of term space can be greatly reduced while maintaining similar classification accuracy.

The remainder of this paper is organized as follows. Section 2 describes the process to remove meaningless and non-informative substrings. Section 3 gives the scoring functions of four term selection methods, and reviews distributional clustering. Section 4 introduces the naive Bayes classifier. Section 5 gives our experimental results. Section 6 gives conclusion. Throughout this paper, we assume $2 \leq n \leq 20$ when n-gram is mentioned.

2 Term Extraction

There are several research[30, 5, 25] on the extraction of meaningful terms from Chinese texts. In [30] Tseng proposed *a multi-linear term-phrasing* technique in which adjacent character sequences are merged pairwisely to form longer character sequences if they satisfy the criteria of the merging rules. This approach is simple but can not run incrementally when new news are added. In [5] Chien proposed *PAT-tree* method to extract keyword. PAT-tree is an incremental method but does not handle the I/O problem when the amount of memory is not large enough to store the whole tree. In this paper, we propose an approach based on SB-trees [13] which use B^+tree to store all the suffix strings[14] of the training documents. Note that SB-tree can grow incrementally, is I/O efficient and is scalable to store large amount of data.

We construct two SB-trees to locate the left and right boundary of terms respectively, and compute the statistics information of extracted term by scanning the leaves of SB-tree. We use *SB-trees* [13, 29] to store all suffix strings [14] of every sentences in the training corpus, and then search for all the repeated strings which appear more than once. To eliminate redundant strings, we gather only the repeated patterns that have, at least, two different kinds of successor Chinese characters. For example, in Figure 1, there are partial sorted suffix strings listed in the SB-tree. The "傳統", "傳統工業" and "傳統工業技術升級"are considered as candidate patterns. Notice that the "傳統工業技", "傳統工業技術" and "傳統工業技術升" are not considered as candidate patterns because they have only one successor Chinese character "術"," 升" and "級" respectively. This process determines the right boundary of terms.

To determine the left boundary of terms, we construct another SB-tree, called *Reverse-SB-tree*, with all suffix strings that come from each reversed sentences in the training corpus. For example, in Figure 2, there are candidate repeated patterns "級升", "級升術技業工" and "級升術技業工統傳". Similarly, the "級升術", "級升術技" and "級升術技業" are

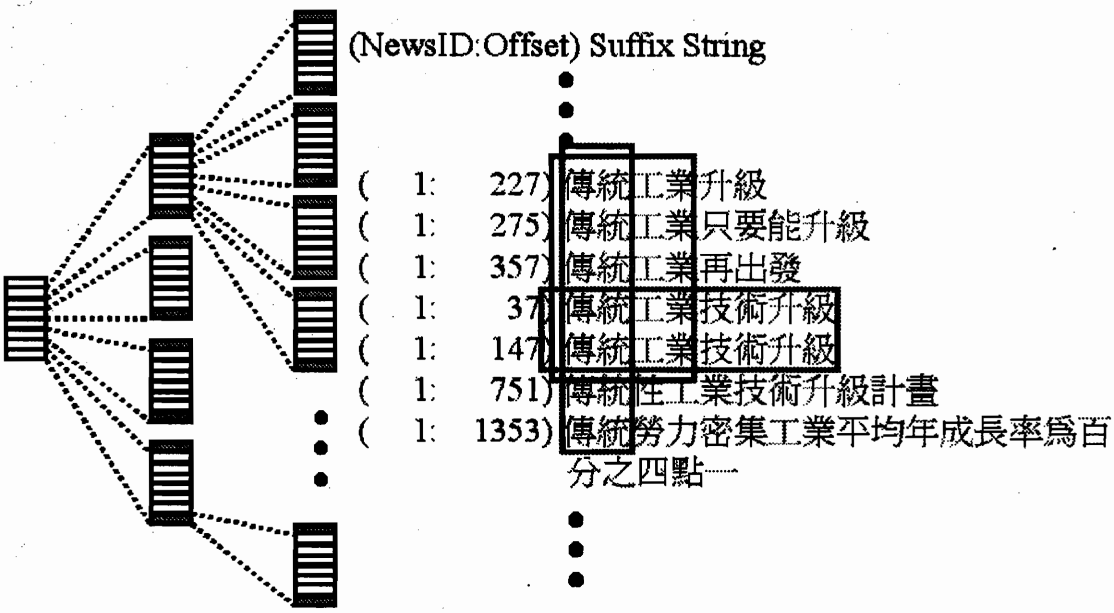

Figure 1: SB-tree

not considered as candidate patterns because they have only one successor Chinese character "技"," 業" and "工" respectively. This process determines the left boundary of terms. Terms identified in above process form an initial set of terms which are used for term selection.

3 Term Selection

After extracting terms from the training corpus as described in section 2, we apply term selection algorithms to select the most representative terms for each class. All terms are given scores by the term selection method, and are choosed according to the scores. There are four term selection methods evaluated individually in this paper. These four term selection methods are odds ratio(OR), information gain(IG), mutual information(MI) and χ^2 statistic(CHI). For a term t and a class c, let A denote the number of times t and c co-occur, B is the number of times t occurs without c, C is the number of times c occurs without t, and N is the total number of documents. The following reviews the term selection methods evaluated in this paper.

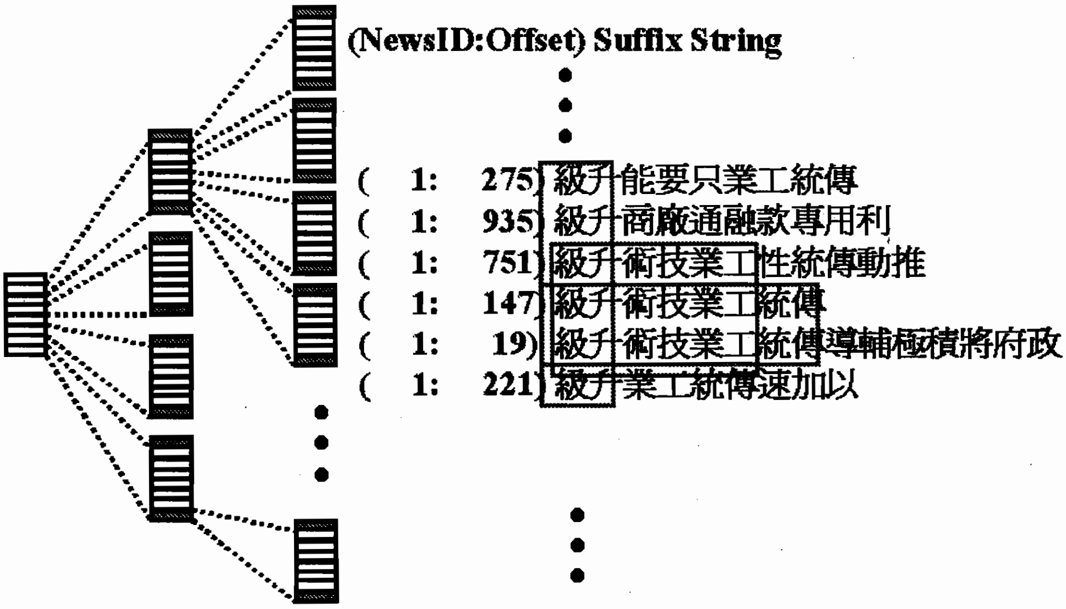

Figure 2: Reverse SB-tree

	c	$\bar{c}$
t	A	B
$\bar{t}$	C	D

Table 1: Two-Way contingency table of a term t and a class c

3.1 Odds Ratio(OR)

The odds ratio value of term t for each class (category) is different. For each term t, the value of odds ratio to class C_k is defined as follows[15].

$$\begin{aligned} OddsRatio(t, C_k) &= \log \frac{Odds(t|C_k)}{Odds(t|C_{neg})} \\ &= \log \frac{P(t|C_k)(1 - P(t|C_{neg}))}{(1 - P(t|C_k)P(t|C_{neg}))}, \end{aligned}$$

where $P(t|C_k)$ is the conditional probability of term t_j occurring given the class value k, $P(t|C_{neg})$ is the conditional probability of term t occurring given the class value $\neq k$. The odds function of X_i is defined as follows.

$$Odds(X_i) = \begin{cases} \frac{\frac{1}{N^2}}{1-\frac{1}{N^2}} & P(X_i) = 0 \\ \frac{1-\frac{1}{N^2}}{\frac{1}{N^2}} & P(X_i) = 1 \\ \frac{P(X_i)}{1-P(X_i)} & P(X_i) \neq 0 \wedge P(X_i) \neq 1 \end{cases}$$

Notice that the value of odds ratio of a term which appears in only one class will be very large even its term frequency is low. It happens that the term selection via the score of odds ratio method might suffer from low hit frequency of selected term when apply to testing documents. This indicates that it is highly possible for a new document to contain very few or even no terms selected by odds ratio method.

3.2 Mutual Information(MI)

The difference between the information uncertainty before adding t and after adding t measures the gain in information due to the Class c. This information is called *mutual information*[35] and is defined as follows.

$$\begin{aligned} MI(t, c) &= \log [\frac{1}{P(c)}] - \log [\frac{1}{P(c|t)}] \\ &= \log [\frac{P(c|t)}{P(c)}] \end{aligned}$$

$$
\begin{aligned}
&= \log[\frac{P(t,c)}{P(t)P(c)}] \\
&= MI(c,t)
\end{aligned}
$$

If the two probabilities $P(t)$ and $P(t|c)$ are the same, then no information is gained and the mutual information is zero. In practice, the score of $MI(t,c)$ is strongly influenced by the marginal probabilities of terms. For terms with an equal conditional probability $P(t|c)$, the term with low term frequency will have a higher score than common terms. The MI can be estimated using

$$MI(t,c) \approx \log \frac{A \times N}{(A+C) \times (A+B)}$$

3.3 Information Gain(IG)

Information Gain is frequently employed as a method of feature scoring in the field of machine learning [26]. Let $|c|$ denote the number of classes. The information gain of term t is defined as follows.

$$
\begin{aligned}
IG(t,C) = E(C) - E(C|t) = & - \sum_{k=1}^{|c|} P(C_k) \log P(C_k) \\
& + P(t=1) \sum_{k=1}^{|c|} P(C_k|t=1) \log P(C_k|t=1) \\
& + P(t=0) \sum_{k=1}^{|c|} P(C_k|t=0) \log P(C_k|t=0)
\end{aligned}
$$

IG is equivalent to the weighted average of the mutual information and is called *average mutual information.* IG makes use of information about term absence, while MI ignores such information. Furthermore, IG normalizes the mutual information scores using the joint probabilities while MI uses the non-normalized scores [35].

3.4 χ^2 statistic (CHI)

The χ^2 statistic measures the lack of independence between t and c, and can be compared to the χ^2 distribution with one degree of freedom to judge extremeness. The χ^2 statistic measure is defined in [20] as follows.

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

3.5 Distributional Clustering

One of the practical problems in term selection is that a document may contain very few or even non of the selected terms(n-grams) if only a small number of significant terms are selected. However, a large number of selected terms will make automatic classification computationally impractical. To overcome the problems, we combine term(feature) selection with term clustering. Notice that term clustering is hard to implement without term selection because the number of extracted terms as described in section 2 is still very large. In this paper we used the distributional clustering[2] to cluster the selected terms. In the following we give a brief description of distributional clustering.

Term clustering algorithms define a similarity measure between terms, and group similar terms into single events that no longer distinguish among their constituent terms. In [2] Baker proposed a weighted average of the parameters of its constituent terms and let, for example, the random variable over classes, C, and its distribution given a particular term, t_i. When term t_i and t_j are clustered together, the new distribution is the weighted average of the individual distributions is as following:

$$\begin{aligned} P(C|t_i \vee t_j) &= \frac{P(t_i)}{P(t_i)+P(t_j)}P(C|t_i) \\ &+ \frac{P(t_j)}{P(t_i)+P(t_j)}P(C|t_j) \end{aligned}$$

The core intuition behind distributional clustering for document classification : the class distributions, $P(C|t_i)$, express how individual terms contribute to classification, and the clustering did preserve the shape of these distributions. Term clustering methods create new, reduced-size event spaces by joining similar terms into groups. The measure of the difference between two probability distributions adapted by [2] is Kullback-Leibler divergence, which is an information-theoretic measure. The KL divergence between the class distributions induced by t_i and t_j is written $D(P(C|t_i)||P(C|t_j))$, and is defined

$$-\sum_{k=1}^{|C|} P(C_k|t_i) \log \frac{P(C_k|t_i)}{P(C_k|t_j)}$$

To avoid the odd properties of KL divergence, such as not symmetric, and it is infinite when an event with non-zero probability in the first distribution has zero probability in the second distribution, they modify the above formula as average KL divergence.

$$\frac{P(t_i)}{P(t_i \vee t_j)} \cdot D(P(C|t_i)||P(C|t_i \vee t_j)) + \frac{P(t_j)}{P(t_i \vee t_j)} \cdot D(P(C|t_j)||P(C|t_i \vee t_j))$$

Instead of comparing the similarity of all possible pairs terms ($O(|V|^2)$ operation), Baker create clusters using a simple, greedy agglomerative approach that consider all pairs of a much smaller subset, of size M, where M is the final number of clusters desired. The clusters are initialized with M terms that have highest score, using information gain(IG) in [2]. The most similar two clusters are joined, the next term is added as a singleton cluster to bring the total number of clusters back up to M. Notice that the number of score for each term measured by IG is just one. Therefore, the M terms as initial cluster may prefer some classes such that result in a biased estimate of term probability distribution to begin with. To avoid a biased estimate of term probability distribution to begin with, we have equal number of selected terms from each class as initial seeds of clusters. Experiment results show that our modification did improve the classification accuracy and smooth the variation of accuracy between each class.

4 Naive Bayes Classifier

There are several well known text classification methods[34] in machine learning or image processing field, such as decision tree method, Neural network method[11], k-nearest-neighbors(KNN)[22], Rocchio algorithm [24] and Naive Bayes classifier [26, 19]. In this research, we implement the naive Bayes classifier for its simplicity and scalability. We are ready to implement other classifiers and measure their performance when they are combined with various term selection methods. The Naive Bayes classifier is one highly practical learning method and is based on the simplifying assumption that the probabilities of terms occurrences are conditionally independent of each other given the class value [26], though this is often not the case. The naive Bayes approach classifys a new document Doc to the most probable class, C_{NB} defined below.

$$C_{NB} = argmax_{C_k \in C} P(C_k|Doc)$$

By Bayes' theorem [18], the $P(C_k|Doc)$ can be represented as

$$P(C_k|Doc) = \frac{P(Doc|C_k)P(C_k)}{\sum_{C_i \in C} P(Doc|C_i)P(C_i)}$$

Where $P(C_k) = |C_k|/\sum_{C_i \in C}|C_i|$ is the probability of the class C_k, and $|C_k|$ is the number of training documents in class C_k.

To estimate $P(Doc|C_k)$ is difficult since it is impossible to collect a sufficiently large number of training examples to estimate this probability without prior knowledge or further assumptions. However, the estimation become possible due to the assumption that a word's(term) occurrence is dependent on the class the document comes from, but that it occurs independently of the other words(terms) in the document. Therefore, the $P(Doc|C_k)$ can be written as follows [19]:

$$P(Doc|C_k) = \prod_{j=1}^{|Doc|} P(t_j|C_k)$$

where $|Doc|$ is the number of words (terms) in document Doc, and $P(t_j|C_k)$ is the conditional probability of t_j given Class C_k. Given the term $T = (t_1, t_2, \ldots, t_n)$ that describe the document Doc, the estimation of $P(Doc|C_k)$ is reduce to estimating each $P(t_j|C_k)$ independently. Notice above equation works well when every term appears in every document; otherwise, the product becomes 0 when some terms do not appear in that document. We use the following to approximate $P(t_j|C_k)$ to avoid the possibility that the product becomes 0, and still keeps the meaning of the equation.

$$P(t_j|C_k) = \frac{1 + TF(t_j, C_k)}{|T| + \sum_j^{|T|} TF(t_j, C_k)}$$

where $TF(t_j, C_k)$ is the frequency of term t_j in documents having class value k, $|T|$ is the number of all distinct terms used in the domain of document representation. The formula used to predict probability of class value C_k for a given document Doc is as the following :

$$P(C_k|Doc) = \frac{P(C_k)\prod_{t_j \in Doc} P(t_j|C_k)^{TF(t_j, Doc)}}{\sum_i P(C_i)\prod_{t_j \in Doc} P(t_j|C_i)^{TF(t_j, Doc)}}$$

5 Experimental Results

Our experiment use one year news, 1991/1/1 to 1991/12/31, which consists of 73,420 news articles, with 23,680,756 characters as training data . We use news from 1992/1/1 to 1992/1/7

	Training : 1991/1/1-1991/12/31 (12 months)		
	Testing : 1992/1/1-/1/7 (7 days)		
		#Train	#Test
	CNA News Group	1/1-12/31	1/1-1/7
1.政治	cna.politics.*	23516	422
2.經濟	cna.economics.*	10160	219
3.交通	cna.transport.*	3423	70
4.文教	cna.edu.*	6064	94
5.體育	cna.1*	4929	73
6.社會	cna.judiciary.*	5679	107
7.股市	cna.stock.*	3313	42
8.軍事	cna.military.*	4646	79
9.農業	cna.argriculture.*	3217	54
10.宗教	cna.religion.*	1315	22
11.財政	cna.finance.*	3622	59
12.社福	cna.health-n-welfare.*	3536	66
	Total	**73420**	1307
23680756 Characters => 322.5 Characters/per News			

Table 2: CNA News : Training&Testing

as testing data. Table 2 summarizes the training and testing data.

We first compare four methods, *OR, IG, CHI and MI* [15, 35] without combining distributional clustering. All methods compute scores to all terms and terms are selected according to their scores. Let the *top k measure* denote the percentage of the correct class is in the first k classes when all the classes are sorted according to their probabilities computed by the naive Bayes classifier. Namely, the top 1 measure denotes the percentage that the news are assigned to their pre-defined classes. Notice that the top k measure will be very meaningful in a semi-automatic system when the number of classes is large as it can quickly identify the most possible k classes. Let the *HitAvg* denote the average number of the selected terms been found in testing news and use to see the popularity of selected terms. Let the *Macro Accuracy* denote the average of the accuracy of each class, and the *Variance of Accuracy* denote the variance of the accuracy of each class. Notice that Macro Accuracy and Variance of Accuracy are used to inspect the variation of accuracy between each class. The less value of Variance of Accuracy is, the less difference of classification accuracy between each class is.

Table 3 shows that the accuracy of top 1 measure of the CHI method changes from 69.17% to 77.35% as the number of selected terms from each class increases from 100 to 5000. The performance of the IG method is similar to the performance of the CHI method. The HitAvg of IG and CHI are 39.02 and 25.35 respectively when the number of selected terms from each class is 1000. This indicates that IG prefers terms with high term frequency. Notice that the accuracy of top 2 measure of CHI is about 90% and is very meaningful in a semi-automatic system. In Table 3 CHI performs the best and achieves 77.35% accuracy in top 1 measure when the number of selected terms from each class is 5000. Both the performance of OR and MI are worse than CHI because both of them prefer to select terms whose term frequencies are low. This can be observed from their low HitAvg, and is consistent with previous theoretic assumption in section 3.1 and 3.2.

Term clustering can reduce the dimension of term space by clustering similar terms into the same group. In addition, redundant substrings and their original strings will be clustered into the same group. This compensates the weakness of term extraction methods which do not remove all redundant substrings. In Table 4, substrings "二屆國", "二屆國代" and "二屆國代選舉" are clustered into group 12; "交易所", "券交易所" and "證券交易所" are clustered into group 300. Furthermore, performance may be increased by clustering when training data is sparse because averaging statistics for similar words together can result in more robust estimates. In Table 4, similar terms, "旅行業"(a travel agent) and "旅行協會"(travel agency association) are clustered together into group 100;" 交響樂團"(a philharmonic orchestra), "巡迴演出"(a show on tour) and "演奏"(to perform) are clustered in group 207;" 犯案"(to commit a crime), "刑事警察"(penal police), "看守所"(a jailer's room) and 槍枝(firearms) are clustered into group 225.

Table 5 shows the difference among different number of selected terms when the number of term groups is fixed at 120. In Table 5, the accuracy of top 1 measure increases as the number of selected terms increases for all term selection methods. When the number of

			Micro Accuracy					
The number of selected terms from each class	The number of total selected terms	Feature Selection Method	Top1	Top2	Top3	HitAvg	Macro Accuracy	Variance of Accuracy
100	1200	OR	50.73	64.50	70.08	1.02	39.21	718.23
100	1200	IG	67.64	87.45	92.81	13.01	68.82	346.01
100	1195	CHI	69.17	86.92	91.58	9.49	68.09	329.17
100	1200	MI	37.49	54.25	61.29	0.24	18.60	616.76
500	6000	OR	62.43	74.75	79.57	2.76	56.73	470.21
500	6000	IG	72.53	89.21	94.41	28.74	74.15	214.42
500	5939	CHI	74.22	91.58	95.10	18.97	73.52	231.13
500	6000	MI	47.28	66.11	72.07	1.13	38.61	432.53
1000	12000	OR	66.03	77.43	82.17	4.04	61.23	370.12
1000	12000	IG	74.22	89.82	94.19	39.02	74.89	207.25
1000	11821	CHI	74.45	91.20	95.26	25.35	75.13	170.24
1000	12000	MI	57.23	74.98	80.49	2.30	54.89	443.64
2000	24000	OR	69.01	79.72	85.77	6.32	66.04	253.29
2000	24000	IG	73.83	90.13	95.26	49.43	75.44	163.70
2000	23513	CHI	75.82	91.51	95.26	32.31	76.81	126.21
2000	24000	MI	64.04	79.19	84.77	4.38	64.39	313.37
5000	59921	OR	74.60	86.46	91.66	16.23	74.44	166.95
5000	60000	IG	75.06	90.36	94.95	62.73	76.10	130.04
5000	57482	CHI	77.35	91.43	95.10	44.01	77.74	123.11
5000	59914	MI	73.53	85.54	91.58	14.59	73.57	214.06

Table 3: Feature Selection Comparison : Testing News(1992/1/1-1992/1/7)

	Group ID				
	12	100	207	225	300
1	二屆國	公路和	交響	犯案	今天在東京
2	二屆國代	在交通	交響樂團	刑事警察	交易所
3	二屆國代選舉	的快樂	巡迴演出	在逃	券交易所
4	的候選人	的班	的音樂	收押	證券交易所
5	候選人	旅行業	的舞	判處死刑	
6	候選人的	旅客的	奏會	官認為	
7	國大代表	旅遊協會	國立藝	押回	
8	國代候選人	泰航	國樂	前科	
9	國代選舉	機票	演奏	看守所	
10			演奏會	書指出	
11			舞蹈	處死刑	
12			樂家	被告	
13			樂團	槍枝	
14			鋼琴	辦案	
15			藝術學	警方在	

Table 4: Term clustering Examples

terms selected from each class is 5000, the accuracy of top 1 measure of IG and CHI are 77.51% and 76.89% respectively. Compared with the accuracy of top 1 measure in Table 3, we find that we can reduce the dimension of term space from 60000 to 120 while the loss of accuracy is less than 1%.

Table 6 shows the difference among different number of term groups when the number of the selected terms from each class is fixed at 1000. The accuracy of top 1 measure of CHI ranges from 74.06% to 75.29% when the number of term groups changes from 60 to 1200. From this observation, we believe that the accuracy is not influenced significantly by the dimension of term space unless the number of term groups is very small(say,12).

				Micro Accuracy					
The number of selected terms from each class	The number of total selected terms	The number of groups	Feature Selection Method	Top1	Top2	Top3	HitAvg	Macro Accuracy	Variance of Accuracy
100	1200	120	OR	50.73	64.04	70.01	1.02	39.21	718.23
100	1200	120	IG	66.41	87.22	92.12	13.01	68.48	377.49
100	1195	120	CHI	69.55	86.76	91.43	9.49	67.68	351.75
100	1200	120	MI	37.34	54.25	61.51	0.24	18.67	603.88
500	6000	120	OR	62.36	73.60	78.96	2.76	56.61	471.53
500	6000	120	IG	72.07	88.60	92.96	28.79	74.46	183.69
500	5939	120	CHI	74.22	90.51	94.03	18.97	73.31	225.94
500	6000	120	MI	46.67	65.42	71.92	1.13	38.64	419.26
1000	12000	120	OR	66.64	77.35	82.25	4.04	61.52	354.31
1000	12000	120	IG	73.64	89.36	93.19	39.02	75.18	149.71
1000	11821	120	CHI	74.22	90.51	94.57	25.35	74.54	186.58
1000	12000	120	MI	56.47	74.52	80.72	2.30	54.47	435.64
2000	24000	120	OR	68.78	80.59	85.77	6.32	65.49	261.91
2000	24000	120	IG	75.06	89.98	94.19	49.43	76.45	124.64
2000	23513	120	CHI	75.44	91.35	95.26	32.31	75.81	129.89
2000	24000	120	MI	64.19	78.50	84.24	4.38	65.31	269.52
5000	59921	120	OR	74.98	88.14	92.12	16.23	71.02	314.07
5000	60000	120	IG	77.51	90.82	94.72	62.73	76.47	132.35
5000	57482	120	CHI	76.89	91.43	94.95	44.01	76.43	126.65
5000	59914	120	MI	66.72	81.71	89.82	14.59	72.14	130.21

Table 5: Term clustering comparison : 120 groups

			Micro Accuracy					
The number of total selected terms	The number of groups	Feature Selection Method	Top1	Top2	Top3	HitAvg	Macro Accuracy	Variance of Accuracy
12000	12	OR	62.51	75.36	81.41	4.04	58.48	506.62
12000	60	OR	66.41	77.43	82.25	4.04	61.40	352.57
12000	120	OR	66.64	77.35	82.25	4.04	61.52	354.31
12000	600	OR	66.49	77.20	81.94	4.04	61.30	358.29
12000	1200	OR	66.11	77.28	81.87	4.04	61.22	363.81
12000	12	IG	70.39	85.00	91.20	39.02	69.99	267.81
12000	60	IG	71.46	88.60	93.27	39.02	73.64	146.79
12000	120	IG	73.64	89.36	93.19	39.02	75.18	149.71
12000	600	IG	73.91	89.82	93.88	39.02	74.89	172.34
12000	1200	IG	74.37	89.90	94.03	39.02	74.44	181.35
11821	12	CHI	70.54	87.15	92.58	25.35	69.53	374.38
11821	60	CHI	74.06	89.90	94.34	25.35	74.00	164.21
11821	120	CHI	74.22	90.51	94.57	25.35	74.54	186.58
11821	600	CHI	74.06	91.20	95.03	25.35	74.38	191.07
11821	1200	CHI	75.29	91.20	95.64	25.35	75.72	166.63
12000	12	MI	53.25	68.86	75.98	2.30	49.15	713.99
12000	60	MI	56.54	73.68	80.18	2.30	55.24	423.26
12000	120	MI	56.47	74.52	80.72	2.30	54.47	435.64
12000	600	MI	56.31	74.45	80.57	2.30	54.29	446.19
12000	1200	MI	56.08	74.29	80.49	2.30	54.16	453.86

Table 6: Term clustering comparison : 1000 Terms selected from each class

6 Conclusions

In this paper, we sketch an implementation of approaches that can handle large amount of training data such as several years of news articles, and automatically assign predefined class to Chinese free text documents. We implement a SB-tree-based approach to extract terms from the original text data, and develop a simple approach to remove redundant subtrings. We also compare four term selection methods combined with distributional clustering and use the naive Bayes classifier to evaluate their performance. In our experiments Information Gain(IG) and χ^2 statistic(CHI) achieved better performance than Odd Ratio(OR) and Mutual Information(MI). With proper term selection and clustering methods, the dimension of term space can be reduced from 60000 to 120 while the loss of classification accuracy is less than 1%.

Acknowledgment. We would like to thank Dr.Chien Lee-Feng and Prof.Tseng Yuen-Hsien for many valuable discussions and comments during this research, and Mr. Lee Min-Jer for kindly help to gather the CNA news.

References

[1] Chidanand Apte, Fred Damerau, and Sholom M. Weiss. Towards language independent automated learning of text categorization methods. In *SIGIR 94*, 1994.

[2] L.Douglas Baker and Andrew Kachites McCallum. Distributional clustering of words for text classification. In *SIGIR 98*, 1998.

[3] Aitao Chen, Jianzhang He, and Liangjie Xu. Chinese text retrieval without using a dictionary. In *SIGIR 97*, 1997.

[4] Chun-Liang Chen, Bo-Ren Bai, Lee-Feng Chien, and Lin-Shan Lee. Cpat-tree-based language models with an application for text verification in chinese. In *Research on Computational Linguistics Conference(ROCLING XI)*, 1998.

[5] Lee-Feng Chien. Pat-tree-based keyword extraction for chinese information retrieval. In *SIGIR 97*, 1997.

[6] Lee-Feng Chien, Min-Jer Lee, and Hsiao-Tieh Pu. Improvements of natural language modeling approaches with information retrieval techniques and internet resources. In *Information Retrieval with Asian Languages(IRAL 1997)*, 1997.

[7] Willian W. Cohen and Yoram Singer. Context-sensitive learning methods for text categorization. In *SIGIR 96*, 1996.

[8] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. McGraw-Hill Book Company, 1990.

[9] David D.Lewis and William A. Gale. A sequential algorithm for training text classifier. In *SIGIR 94*, 1994.

[10] David D.Lewis and Marc Ringuette. A comparison of two learning algorithm for text categorization. In *3rd Annual Symosium on Document Analysis and Information Retrieval*, 1994.

[11] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesain classifier under zero-one loss. In *Machine Learning*, 1997.

[12] Brian S Everitt. *Cluster Analysis*. Halsted Press, New York, third edition edition, 1993.

[13] Paolo Ferragina and Roberto Grossi. An experimental study of sb-trees. In *ACM-SIAM symposium on Discrete Algorithms*, 1996.

[14] William B. Frakes and Rick Kazman. *Information Retrieval Data Structures & Algorithm*. Prentice Hall, Englewood Cliffs, New Jersey 07632, 1992.

[15] Marko Grobelink and Dunja Mladenic. Feature selection for classification based on text hierarchy. In *Conference on Automated Learning and Discovery CONALD-98*, 1998.

[16] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences : computer science and computational biology*. Cambridge University Press, 1997.

[17] Rainer Hoch. Using IR techniques for text classification in document analysis. In *SIGIR 94*, 1994.

[18] M. James. *Classification Algorithms.* Wiley, 1985.

[19] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, 1997.

[20] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data Analysis : An Introduction to Cluster Analysis.* John Wiley and Sons,Inc., New York, 1990.

[21] K.L Kwok. Comparing representations in chinese information retrieval. In *SIGIR 97*, 1997.

[22] Wai Lam. Using a generalized instance set for automatic text categorization. In *SIGIR 98*, 1998.

[23] Leah S. Larkey and W. Bruce Croft. Combining classifiers in text categorization. In *SIGIR 96*, 1996.

[24] David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. Training algorithms for linear text classifiers. In *SIGIR 96*, 1996.

[25] Yih-Jeng Lin, Ming-Shing Yu, Shyh-Yang Hwang, and Ming-Jer Wu. A way to extract unknown words without dictionary from chinese corpus and its applications. In *Research on Computational Linguistics Conference(ROCLING XI)*, 1998.

[26] Tom M. Mitchell. *Machine Learning.* The McGraw-Hill Companies, Inc, 1997.

[27] Mehran Sahami, Marti Hearst, and Eric Saund. Applying the multiple cause mixture model to text categorization. In *Machine Learning: Proc. of the 13th International Conference*, 1996.

[28] Von-Wun Soo, Pey-Ching Yang, Shih-Huang Wu, and Shih-Yao Yang. A character-bases hierarchical information filtering scheme for chinese news filtering agents. In *Information Retrieval with Asian Languages(IRAL 1997)*, 1997.

[29] Jyh-Jong Tsay, Jing-Doo Wang, Chun-Fu Pai, and Ming-Kuen Tsay. Implementation and evaluation of scalable approaches for automatic chinese text categorization. In *The 13th Pacific Asia Conference on Language, Information and Computation*, 1999.

[30] Yuen-Hsien Tseng. Fast keyword extraction of chinese document in a web environment. In *Information Retrieval with Asian Languages(IRAL 1997)*, 1997.

[31] Shih-Hung Wu, Pey-Ching Yang, and Von-Wun Soo. An assessment of character-based chinese news filtering using latent semantic indexing. *Computational Linguistics and Chinese Language Processing*, 3(2):61–78, 1998.

[32] Yiming Yang. Effective and efficient learning from human decisions in text categorization and retrieval. In *SIGIR 94*, 1994.

[33] Yiming Yang. Noise reduction in a statistical approach to text categorization. In *SIGIR 95*, 1995.

[34] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *SIGIR 99*, 1999.

[35] Yiming Yang and Jan O. Pedersen. A comparative study on feature in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, 1998.

[36] Yun-Yan Yang. A study of document auto-classification in mandarin chinese. In *Research on Computational Linguistics Conference(ROCLING VI)*, 1993.

[37] Ogawa Yasushi and Matsuda Toru. Overlapping statistical word indexing : A new indexing method for japaness text. In *SIGIR 97*, 1997.

[38] Bo-Hyun Yun, Min-Jeung Cho, and Hae-Chang Rim. Korean information retrieval model based on the principle of word formation. In *Information Retrieval with Asian Languages(IRAL 1997)*, 1997.