IMPLEMENTATION AND EVALUATION OF SCALABLE APPROACHES FOR AUTOMATIC CHINESE TEXT CATEGORIZATION

Jyh-Jong Tsay , Jing-Doo Wang , Chun-Fu Pai and Ming-Kuen Tsay
Department of Computer Science and Information Engineering
National Chung Cheng University
Chiayi, Taiwan 62107, ROC.
{tsay, jdwang, pcf84, tmc86}@cs.ccu.edu.tw

Abstract

The purpose of this research is to identify scalable approaches that can handle large amount of training data such as several years of news articles, and automatically assign predefined category to Chinese free text documents. Our approach consists of the following processes: (i) term extraction, (ii) term selection, and (iii) document classification. The approach first builds a recently developed SB-tree to identify all repeated substrings, called *patterns*, from the text. We then proceed to identify possible boundary of terms appearing in the identified patterns. After terms are extracted from the training articles, we run term selection algorithms to select the most significant terms and to reduce the number of terms to an acceptable level. The selected terms are used by the classifier to assign a predefined category to each text document. 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 implement and compare four term selection methods, the odds ratio method, the mutual information method, the information gain method and the χ^2-test method, when they are combined with the naive Bayes classifier.

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

1 Introduction

Text categorization is the problem of automatically assigning predefined categories 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 categorization as well as web-page classification [11, 3, 7, 8, 21, 25, 26, 18, 6, 5, 2, 10]. While there are a great number of researches on automatic text categorization for English texts, text categorization for Asian languages such as Chinese, Japanese, Korean and Thai has not been studied seriously until recently [17, 29, 1].

It is well known that written Asian language consists of strings of ideograph separated by punctuation signs. An ideograph (or character) can function as a word with meaning(s), or it can act as an alphabet to form a "word" with one or more adjacent characters. Determining the boundaries of single or multi-character words in a string, a process called *segmentation* [4], is very difficult because no delimiter or while space is used in the text and one has to rely on the context

contents. Because text segmentation is not straightforward, 1-grams, 2-grams and n-grams have been used as indexing terms to represent documents in Asian languages. Among them, 1-gram-based approaches is the simplest one that uses single characters as indexing terms, and should be good for recall in information retrieval(IR) because it guarantees that if there are correct word matches between queries and documents, there will be 1-gram matches. However, single characters (1-grams) are ambiguous in meaning, which results in low precision in IR. A number of research have proposed to use n-grams, instead of 1-grams, as indexing terms. An n-gram is a sequence of n contiguous characters in the text. The 1-gram-based approaches [23] simply use every single character as a single term, and the 2-gram-based approaches use every 2 contiguous characters as indexing terms, and the general n-gram-based approaches use all 1-grams, 2-grams, 3-grams,..., n-grams as indexing terms. Although 2-gram and n-gram perform similarly well as indicated in our experiment, in this research, we take n-grams, $1 \leq n \leq 10$, as indexing terms because n-grams can catch the concept of a document. Notice that the possible number of n-grams in Chinese is dramatically huge, and furthermore many of them are meaningless and non-informative for text categorization. The major challenge is to develop approach that can reduce the number of n-grams to an acceptable level, while at the same time maintains similar categorization accuracy.

The purpose of this research is to identify scalable approaches that can handle large amount of training data such as several years of news articles, and automatically assign predefined category to Chinese free text documents. Our approach consists of the following processes: (i) term extraction, (ii) term selection, and (iii) document classification. Identifying terms, or so-called word segmentation, from text documents is one of the most difficult problems in processing Chinese texts. In this research, we develop a scalable approach to identify terms from large amount of text data, which does not use a dictionary. The approach first builds a recently developed SB-tree [9, 4, 19] to identify all repeated substrings, called *patterns*, from the texts. We believe important terms will appear repeatedly in the articles. The SB-tree also gives the information such as the frequency of a pattern, the documents and the locations where a pattern appears which are then used to identify possible boundary of terms appearing in the same pattern, and to remove meaningless patterns which are substrings of some terms. Term boundaries are used to partition patterns into terms. After terms are extracted from the training articles, we run term selection algorithms to select the most representative terms and to reduce the number of terms to an acceptable level. The selected terms are used by the classifier to assign a predefined category to each text document.

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 which use either one month news or sampled articles from the whole year news. Notice that although sampling methods are very interesting research issues, most of the commercial systems prefer to extract information from the original whole-set data as done in the recent data mining applications. We believe the whole year training data can make conclusions from our experiment more reliable than previous research. We implement and compare four term selection methods, the odds ratio method, the mutual information method, the information gain method and the χ^2-test method, when they are combined with the naive Bayes classifier [22]. Our experiment shows that χ^2-test achieve the best performance.

The remainder of this paper is organized as follows. Section 2 describes the process to remove meaningless and non-informative substrings, and to select the most representative terms. Section 3 introduces the naive Bayes classifier. Section 4 gives our experimental results. Section 5 gives conclusion and further remarks. Throughout this paper, we assume $2 \leq n \leq 10$ when n-gram is mentioned.

2 Term Selection

To avoid the segmentation problem and extract meaningful terms efficiently, we use n-gram-based approach which is based on simple statistics rather than complex syntax and semantic analysis. It is very important to reduce the number of n-grams generated from the original data. In the section, we describe how to reduce the number of n-grams generated from the original data. The process consists of two main steps: substring removal and term selection. Substring removal is to remove patterns that are substrings of other identified terms, and term selection is to select the most representative terms. Two common term selection methods, odd ratio method and information gain method, are implemented and compared in this research.

2.1 Substring Removal

For Chinese, it is very important to remove the redundant substrings because there are $n(n+1)/2$ substrings derived from each n-gram and, furthermore, most of the substrings are meaningless and non-informative. For example, the substrings of 股票市場(stock market) are listed in the table below. The substrings 票市(?), 股票市(?), 票市場(?) derived from "stock market" (股票市場) are not meaningful "words" in Chinese, and should be removed from the term set.

1-gram	股, 票, 市, 場
2-gram	股票 (stock), 票市 (?), 市場 (market)
3-gram	股票市 (?), 票市場 (?)
4-gram	股票市場

The method that removes the meaningless substrings is motivated by the method developed by Chein [4]. Let T denote the total set of n-grams, and $T = \{t_1, t_2, \ldots, t_k\}$. Our observation is that if the string t_j is a substring of t_i, and the term frequency ratio of t_i and t_j is almost equal to 1.0, say ≥ 0.9, then we can assume t_j is a redundant substring generated from t_i, and remove t_j from the term set. In this experiment, we remove the substring t_j when the ratio of term frequency of t_j over the term frequency of t_i is greater than or equal to 0.9. The original number of n-grams ($n \leq 10$), whose term frequency ≥ 5, generated our training data is 935734. After the substring removal the number is reduced to 425903.

2.2 Term selection methods

Substring removal is just to remove redundant substrings. The number of remained n-grams is still very large. Most of them are not significant for the purpose of categorizing text documents. Term selection, or so-called *feature selection*, is the process to select most significant terms, and to reduce the number of terms to an acceptable level as the time and space required by current classifiers greatly depend on the size of the term set. In addition, the noise, i.e. the non-significant terms, can reduce the precision achieved by a classifier. Several term selection methods have been proposed for occidental languages [20, 16, 15, 10, 27]. In this experiment, we implement the odds ratio method, the mutual information method, the information gain method and the χ^2-test method, and compare their performance when they are combined with the naive Bayes classifier. We next review them.

For convenience of the definition of feature selection, we claims that the two-way contingency table of a term t and a category c, where A is the number of times t and c co-occur, B is the number of time the t occurs without c, C is the number of times c occurs without t, and N is the total number of documents. We summarized above statements as

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

2.2.1 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[10].

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

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 $\neq k$, and the odds function of X_i is defined as follows.

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

Where N is the number of training documents. Notice that the value of odds ratio of one term which just appear in only one class would be very large no matter term frequency is low or high. It happens that the term selection via the score of odds ratio maybe suffer from low hit frequency of selected term when apply testing documents.

2.2.2 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*** and is naturally defined as[27]

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

If the two probabilities $p(t)$ and $P(t|c)$ are the same, then we have gained no information and the mutual information is zero. In practical, 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

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

2.2.3 Information Gain(IG)

Information Gain is frequently employed as a method of feature scoring in the field of machine learning [22]. Let $|c|$ denote the number of categories. The information gain of term t is defined as

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

IG can be proven equivalent to the weighted average of the mutual information and is called *average mutual information.* IG makes a use of information about the 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 [27]. Notice that the number of score for each term measured by IG is just one.

2.2.4 χ^2-test(CHI)

The χ^2-test measures the lack of independence between t and c, and can be computed to the χ^2 distribution with one degree of freedom to judge extremeness. The χ^2-test measure is defined as [14]

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

3 Naive Bayes Classifier

There are several well known classification methods in machine learning or image processing field, such as decision tree method, k-nearest-neighbors(KNN), Neural network method, Rocchio algorithm and Naive Bayes classifier [22, 13]. 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 [22], 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 [12], 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 [13]:

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

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 :

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

4 Experimental Results

The amount of training&testing data of previous related experiments [28, 4, 23] are thousands of news articles which were just within one month or sampling from several months. In order to close the reality of the term distribution of Chinese corpus, we select 12 Central News Agency (CNA) news group from 1991/1/1 to 1991/12/31, which contains 73420 news articles and 23680756 Chinese characters, and chose 21 days out of the next month (January 1992) as testing news. The statistics of training&testing news are listed in Table 1.

4.1 Comparison : 1-gram, 2-gram, 3-gram and n-gram

There are discussions to chose 1-gram, 2-gram(bigram) or n-gram to be basic indexing unit of Chinese texts [17, 24, 1] in Information Retrieval. The character-based approach (1-gram) is good for recall in IR, but not for precision. In [23], they developed a Chinese news filtering agents using character-based approach, and got the result of filtering news efficiently, but the precision of the filtering news is quite low. There is the limitations of character-based approach. For example, considering the order of the Chinese character, the two words 國 中 (junior high school) and 中 國 (China) make no difference via character-based approach. In [17], some reference states that the major of the modern Chinese words are bisyllable. Therefore, they take a lot of experiments and conclude that 2-gram indexing is effective and performs as well as short-word indexing in IR. Notice that the number of 3-grams is more than the number of 2-grams no matter before or after the

	Training : 1991/1/1-1991/12/31 (12 months)				
	Testing : 1992/1/1-/1/7,1/11-1/17,1/21-1/27 (3*7=21 days)				
		#Train		#Test	
	CNA News Group	1/1-12/31	1/1-1/7	1/11-1/17	1/21-1/27
1.政治	cna.politics.*	23516	422	395	437
2.經濟	cna.economics.*	10160	219	211	330
3.交通	cna.transport.*	3423	70	84	78
4.文教	cna.edu.*	6064	94	119	140
5.體育	cna.l*	4929	73	84	75
6.社會	cna.judiciary.*	5679	107	148	183
7.股市	cna.stock.*	3313	42	76	51
8.軍事	cna.military.*	4646	79	64	63
9.農業	cna.argriculture.*	3217	54	82	60
10.宗教	cna.religion.*	1315	22	22	41
11.財政	cna.finance.*	3622	59	86	49
12.社福	cna.health-n-welfare.*	3536	66	71	92
	Total	73420	1307	1442	1599

Table 1: CNA News : Training&Testing

substring removal for terms with frequency ≥ 5. This observation is different from the statements in [17] (see table 2). Notice that the number of the n-grams is the number after substring removal.

Let the *top k measure* denote the percentage of the correct category is in the first k categories when all the categories 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 categories. Notice that the top k measure will be very meaningful in a semi-automatic system when the number of categories is large as it can quickly identify the most possible k categories. We choose the n-gram, $2 \leq n \leq 10$, as the basic indexing unit. Table 3 gives the accuracy achieved when 1-gram, 2-gram, and n-gram are used as term unit and no term selection is performed for n-grams. In the top 1 measure, the gap between the 1-gram-based and n-gram-based approach is about 8%, about 68% and 76%, respectively. The gap decreases as k increases. In top 3 measure, the gap becomes about 3%. The 1-gram-based approach uses only 3089 distinct characters; however, the n-gram-based approach uses 299386 terms. Although the 2-gram-based and the n-gram-based approaches achieve similar accuracy, we use n-grams to measure the difference performed of odds ratio and information gain methods because n-grams can catch the concept of an article and can be assigned as keyword. This can be important in other area of information retrieval.

4.2 Term Selection Comparison : OR, MI, IG and CHI

In this experiment we implement and compare four methods , *OR, IG, CHI and MI* [10, 27], which require much less computation time and are more scalable. All methods compute scores to all terms. Terms are selected according to their scores. Let the *top k measure* denote the percentage of the correct category is in the first k categories when all the categories are sorted according to their

Term Length	tf>=5	Percent	tf>=5+(sub90)	Percent
1	4134	0.4%	3628	0.9%
2	129938	13.9%	84441	19.8%
3	212784	22.7%	113249	26.6%
4	172807	18.5%	89472	21.0%
5	122745	13.1%	51296	12.0%
6	88384	9.4%	33573	7.9%
7	66787	7.1%	20759	4.9%
8	53640	5.7%	14388	3.4%
9	45101	4.8%	9694	2.3%
10	39414	4.2%	5403	1.3%
	935734		425903	

Table 2: Term Length Distribution

			Testing News(3*7=21days)		
		#Term	1/1-1/7	1/11-1/17	1/21-1/27
Top1	1-gram	3089	**68.94**	**68.72**	**65.23**
	2-gram	103072	76.36	75.66	72.23
	3-gram	153558	76.21	75.56	72.73
	(2+...+10)-gram	295910	**77.12**	**76.63**	**72.67**
Top2	1-gram	3089	86.61	84.67	82.74
	2-gram	103072	91.43	89.04	87.49
	3-gram	153558	90.97	88.49	87.37
	(2+...+10)-gram	295910	**92.04**	**88.7**	**87.99**
Top3	1-gram	3089	92.35	90.78	90.06
	2-gram	103072	95.72	93.62	92.87
	3-gram	153558	94.72	92.86	91.81
	(2+...+10)-gram	295910	**95.56**	**92.93**	**92.62**

Table 3: Accuracy Comparison : 1-gram, 2-gram, 3-gram and n-gram

probabilities computed by the naive Bayes classifier. 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. As in Table 4 shows, the accuracy of top 1 measure of the CHI method is from 71.61% to 77.43% as the number of selected term from each class increases from 100 to 10000. The performance of the IG method is similar to the performance of the CHI method while IG prefer the terms whose term frequency are high. The HitAvg of IG and CHI are 43.77 and 26.36 respectively when the number of selected terms from each class is 500. Notice that the accuracy of top 2 measure of CHI is about 90% and is very meaningful in a semi-automatic system. In Table 4 CHI is the best and achieves 75.90% accuracy in top 1 measure when the number of selected terms from each class is 500. That is, the number of n-gram can be reduced from 295910 to 5918 while the accuracy only lose less 2% accuracy as compared with Table 3. Both the performance of OR and MI are worse than CHI's because both of them prefer to select term whose term frequency is low such that their HitAvg are 4.43 and 2.68 respectively. This observation is consistent with previous theoretic assumption in section 2.2.1&2.2.2. Notice that OR achieve 78.04 in top 1 measure, better than CHI's 77.43% when the number of selected terms from each class is 10000, but the number of total selected term by OR is 91745, larger than CHI's 79202.

To illustrate the effectiveness of selected term by term selection of CHI method, for example, we have 20 top score terms selected from four classes respectively in the table 5. To state the characteristic of n-gram, there are significant terms such as "行政院經濟建設委員會" (Council for Economic Planning and Development of Executive for R.O.C), "台灣鐵路管理局" (Taiwan Railway Administration) and "台灣省教育廳" (The Department of Education , Taiwan Provincial Goverment) chose from 經濟 *(cna.economics.*)*, 交通 *(cna.transport.*)* and 教育 *(cna.edu.*)* classes respectively. Notice that using n-gram are more meaningful and informative than using 1-gram or 2-gram(bigram).

5 Conclusions and Further Remarks

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 category 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, the odds ratio method, the mutual information method, the information gain method and the χ^2-test method, and use the naive Bayes classifier to evaluate their performance. Among four feature selection method, χ^2-test achieve the best performance. 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. We believe the whole year training data can make conclusions from our experiment more reliable than previous research. The experiment shows that the character-based approach performs poorly in the top 1 measure; however is quite competitive in the top 3 measure. Notice that the top k measure will be very meaningful in a semi-automatic system when the number of categories is large as it can quickly identify the most possible k categories. This paper present an initial experimental study of Chinese text categorization. There are a lot of work to be proceeded in the future. The naive Bayes classifier is a basic approach in the probability model. There are many other classifiers in the vector model such as KNN and Rocchio algorithm.

			Average Accuracy			
The number of selected terms from each class	The number of total selected terms	Feature Selection Method	Top1	Top2	Top3	HitAvg
100	1200	OR	56.85	69.32	74.29	1.68
100	1200	IG	69.85	87.22	91.97	19.41
100	1200	CHI	**71.61**	86.92	92.43	12.66
100	1200	MI	42.08	57.77	65.42	0.56
500	6000	OR	67.25	76.97	82.40	**4.43**
500	6000	IG	72.69	89.36	93.57	**43.77**
500	**5918**	CHI	**75.90**	90.90	94.49	**26.36**
500	6000	MI	55.62	72.61	78.73	**2.68**
1000	12000	OR	69.24	79.27	84.93	6.84
1000	12000	IG	73.37	89.14	94.26	58.60
1000	11770	CHI	76.28	91.28	95.03	35.79
1000	12000	MI	61.82	77.81	83.09	4.99
2000	23991	OR	73.14	82.71	87.83	12.61
2000	24000	IG	74.67	89.82	93.96	71.24
2000	23075	CHI	76.28	90.67	95.03	47.95
2000	23990	MI	68.32	81.64	86.76	11.08
5000	57726	OR	76.13	88.14	93.34	46.47
5000	60000	IG	76.36	89.44	93.50	81.51
5000	52399	CHI	76.97	90.74	94.95	70.25
5000	57161	MI	74.60	87.22	92.58	41.55
10000	91745	OR	**78.04**	90.36	93.57	71.79
10000	120000	IG	76.66	89.44	93.88	84.64
10000	79202	CHI	**77.43**	90.21	93.73	80.30
10000	91024	MI	77.58	90.21	94.34	69.58

Table 4: Feature Selection Comparison : Testing News(1992/1/1-1992/1/7)

	政治(cna.politics.*)	經濟(cna.economics.*)	交通(cna.transport.*)	教育(cna.edu.*)
1	政黨	經濟部	交通部	教育
2	選舉	出口	航空公司	教育部
3	執政	進口	航空	學校
4	民主	經貿	航線	學生
5	執政黨	行政院經濟建設委員會	華航空	教育廳
6	屆國	院經濟	台灣鐵路	國立
7	總理	經濟建設	民航局	校長
8	民黨	行政院經	中華航空公司	教學
9	領袖	國貿	台灣鐵路管理局	藝術
10	二屆國	外貿	鐵路局	省教育
11	候選人	國經	旅客	省教育廳
12	黨中	經濟部國	交通處	科學
13	國民	景氣	班次	學年度
14	日國	中油	飛航	作品
15	的政	油公司	民航	灣省教育廳
16	外長	江丙坤	交通部長	教師
17	秘書長	建設委員會	民用航空	文化
18	國代	生產	省交通處	台灣省教育廳
19	人民	國經濟	民用航空局	高中
20	黨中央	億美元	交通部民	招生

Table 5: 20 top score terms using CHI method

We will do experiment to understand their performance in Chinese text categorization. In addition, the category structure in this experiment is flat. However, most of the information search engines provides hierarchical structures. We will do experiment on web information, and study approaches that takes advantages of the hierarchical structures provided in search engines.

Acknowledgment. We would like to thank Dr.Chein 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] Liangjie Xu Aitao Chen, Jianzhang He. Chinese text retrieval without using a dictionary. In *ACM SIGIR*, 1997.

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

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

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

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

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

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

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

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

[10] Marko Grobelink and Dunja Mladenic. Efficient text categorization. 1998.

[11] Rainer Hoch. Using IR techniques for text classification in document ahalysis. In *ACM SIGIR*, 1994.

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

[13] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *ICML-97*, 1997.

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

[15] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *ICML-97*, 1997.

[16] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *Machine Learning : 13th International Conference*, 1996.

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

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

[19] Min-Jer Lee Lee-Feng Chien and Hsiao-Tieh Pu. Improvements of natural language modeling approaches with information retrieval techniques and internet resources. In *IRAL'97*, 1997.

[20] David D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of Speech and Natural Language*, 1992.

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

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

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

[24] Chi-Yin Wong Wai Lam and Kam-Fai Wong. Performance evaluation of character-, word- and n-gram-based indexing for chinese text retrieval. In *Information Retrieval with Asian Languages*, 1997.

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

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

[27] Yiming Yang and Jan O. Pedersen. A comparative study on feature in text categorization. 1998.

[28] Yun-Yan Yang. A study of document auto-classification in mandarin chinese. In *ROCLING*, 1993.

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