The Chinese Text Categorization System with Category Priorities

Huan-Chao Keh
Department of Information Engineering, Tamkang University, Taipei, Taiwan
Email: keh@cs.tku.edu.tw

Ding-An Chiang, Chih-Cheng Hsu and Hui-Hua Huang
Department of Information Engineering, Tamkang University, Taipei, Taiwan
Email: chiang@cs.tku.edu.tw, 894190130@s94.tku.edu.tw, 893190040@s93.tku.edu.tw

***Abstract*—The process of text categorization involves some understanding of the content of the documents and/or some previous knowledge of the categories. For the content of the documents, we use a filtering measure for feature selection in our Chinese text categorization system. We modify the formula of Term Frequency-Inverse Document Frequency (TF-IDF) to strengthen important keywords' weights and weaken unimportant keywords' weights. For the knowledge of the categories, we use category priority to represent the relationship between two different categories. Consequently, the experimental results show that our method can effectively not only decrease noise text but also increase the accuracy rate and recall rate of text categorization.**

***Index Terms*—text categorization, feature selection, filtering measure, text mining**

I. INTRODUCTION

English is taken as the main language family in many recent text mining studies [1][2][3]. Chinese language family studies are not common. Therefore, we have a fact study of Chinese text categorization system. In Chinese text, there are no obvious spaces between Chinese words and English words, numbers, and symbols are often included, so the feature extraction needs punctuation. Chinese punctuation is to divide particular text into some words of uncertain lengths. Since a single Chinese character has different meanings when combined with different characters, Chinese punctuation has to rely on a large word library and context comparison in order to acquire the most appropriate words. As this system only categorizes Chinese articles, in the preprocess phase, we remove all characters except Chinese words and use the Chinese Punctuation System [4], developed by Library Team of Central Research Academy, to make Chinese punctuation. We find that some features may be missed or divided into different features with different meanings; for example "大腸桿菌" (colon bacillus) is cut into "大腸"(colon) and "桿菌" (bacillus), though "大腸桿菌"(colon bacillus) should be regarded as a single feature. Although some features may not be cut out from one feature, the correlation between these features exists. If this type of combining feature is regarded as a special feature, it will be helpful in classification processing. The association rule can be used to find terms which may have relations to each other. Therefore, we utilize the associative classification technique to deal with such subject in ref. [5].

In this paper, we use feature terms longer than two characters to compute weights of these terms relative to categories. After being punctuated, the document can be represented by the bag of words [1]. The document D can be converted to $d= ((f_1,w_1),(f_2,w_2)\ldots(f_i,w_i))$, where each f_i is a document word, and w_i denotes its frequency. Since the number of different words appearing in the collection may be very large and contain many irrelevant words for the classification, in addition to eliminating stop words and auxiliary words such as「的」(of),「而且」(but also),「和」(and),「因爲」(because), feature reduction is usually performed. As pointed out by ref. [3], filtering and wrapping are two main approaches to feature reduction. Since wrapper approaches are time-consuming and sometimes unfeasible to use, in this paper, we use filtering measure, TF-IDF, for feature selection in our text categorization system. Although the TF-IDF is performed well in many situations [6], this formula still has some problems. To solve these problems, we will modify this formula to improve the classification recall rate and accuracy rate. We introduce this improved formula in section 3.

The data sources of this study are from the thesis abstracts of universities in Taiwan. These thesis abstracts are extracted from "National Dissertations and Theses Information Web" [7]. Generally, documents are classified into different categories by the content of the document directly. However, documents are thesis abstracts in this research and some theses of a department may cross different fields. If we do not consider previous knowledge of the categories, theses should be categorized by departments which release the theses, and only consider the content of these theses, it may cause classification errors. For example, some theses of the chemistry department may apply related chemistry knowledge to the field of biology, thus in document classification, it may be wrongly classified into the "Biology" category because of many biological keywords in these abstracts. In this paper, we use category priority to solve this kind of classification

doi:10.4304/jsw.5.10.1137-1143

mistake. The category priority will be introduced in section 3.

The rest of this paper is organized as follows: In Section 2, related work is summarized. In Section 3, improved TF-IDF method and category priorities are presented. Section 4 introduces experiment results. In section 5, conclusions and further research are described.

II. Relative Work

A. Document Classification Process

The different text categorization systems have been proposed recently [8][9][10][11], in this paper, we refer Aas and Eikvil [12] and sum up the systemic process for document classification as shown in Fig. 2.1. The system classified documents as training documents and test documents, pretreated training documents with known categories and extracted all kinds of feature terms through phrase distribution statistics, retained meaningful phrases and built the classification estimation model, then categorized the test documents.

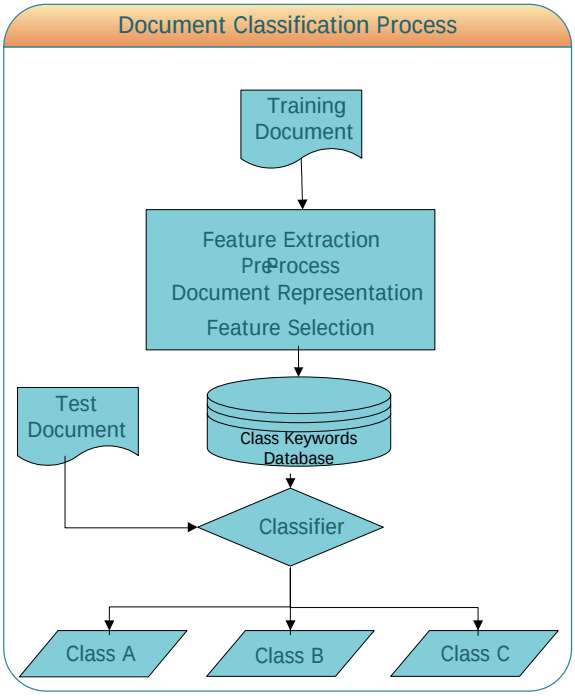

Figure 2.1 The process of text categorization system

Before document classification, we have to select and determine a document profile. Titles, abstracts or specified chapters can be taken as document profiles. If all document content is selected as the document profile, there would be a huge amount of text information and a lot of meaningless information. The document profile shall therefore be selected according to category feature and shall be the most representative content. The proper document profile can reduce redundant text information so as to enhance classification performance. For example, Maron [10] took Transaction on Electronic Computer periodicals as a document source, and selected abstracts as document profiles. Hamill [8] took Chemical Abstracts Titles as document profiles.

Feature extraction is further divided into pre-process, document representation and feature selection. In this part, most of the related literature are English documents as test data source. If Chinese documents are to be tested; in the feature extraction part, a Chinese punctuation treatment must be additionally done. After being punctuated, in addition to eliminating stop words and auxiliary words, we have to perform feature selection to reduce the affect of irrelevant words on classification.

B. Weight Computation

Feature extracted documents are often expressed as vector patterns (weight, keyword). The weight can be computed by different methods, such as information gain [10][13], mutual information [10], etc. As pointed out by ref. [14], the TF-IDF method performs well in many situations. In this paper, we use this filtering measure method in our system and it is introduced as follows.

The TF-IDF method uses the term frequency and document frequency to compute the weight of a word in a document. The term frequency *TF(t,d)* is the frequency of a word *t* in the document *d*. The document frequency *DF(t)* is the number of documents that contain a word *t*. The inverse document frequency of a word *t*, *IDF(t)*, can be computed by the following formula:

$$IDF(t) = \log\left[\frac{D}{DF(t)}\right] \quad (1)$$

In the above formula, *D* is the number of documents and *IDF(t)* is the discretion degree of a word *t* over the integral document. Since the importance of a word *t* in a document *d* is proportional to the frequency of a word occurring in a document and inverse document frequency, the weight of a word *t* in a document *d*, *W(t,d)*, can be computed by the following formula:

$$W(t,d) = TF(t,d) \times IDF(t) \quad (2)$$

In the above formula, a larger value of *W(t,d)* indicates a higher frequency of a word *t* occurring in a document *d*, but a lower frequency of *t* occurring in all documents.

The problem with this formula is that it easily results in constant *IDF(t)* of term *t* at each category, or very similar weights of term *t* with respect to various categories. When this situation occurs, the weight of *t* in different document *d* totally depends on the term frequency *TF(t,d)*. When the frequency of a noise term, *TF(t,d)*, is higher, the weight *W(t,d)* with respect to each category will be bigger; therefore, the classification error probability will increase. To improve the classification accuracy, the discretion should take account of category distribution of each word. This article primarily adopted TF-IDF weight method and made some improvements. The modified TF-IDF will be introduced in the next section.

C. Classification

Many document classification algorithms, such as Rocchio classification [15][16], decision tree classification [17], SVM(support vector machine)[18], KNN nearest neighbor rule [19], and Naïve-Bayes [13][20], have seen proposed recently. Various classification algorithms have their own advantages and different classification models. In this paper, we use Naïve-Bayes classification method to classify documents; therefore, only this method is introduced in this section.

Naïve-Bayes classification method is designed on the basis of Bayesian analysis theory. Bayesian Analysis was proposed by Thomas Bayes in the early 18th century. Its basic principle is to modify (or improve) boundary probability of a certain incident according to some additional information. It predicts the probability of an object being the member of a certain category so as to complete classification. This study adopted Bayes probability [13][20] as the classification criterion. For the Naïve-Bayes classifier, this paper calculates the weights of all characteristic terms according to TF-IDF formula. For test documents, Bayes performs as the basis of the classification rules. It makes the occurrence number of characteristic terms$<f_1,f_2,f_3\ldots,f_i>$ of a document(j) and multiplies its matched *IDF*. The obtained weight is the cumulative integral of that class provided that the number of the classes is known. And then it sums the cumulative integral of one class in the document to get the matched integral for that class. Based on the obtained matched integral, it can infer the class of this document through the following formula:

$$Value(j) = \sum term(fi) \times w(i \mid j) \tag{3}$$

where *Value(j)* is the cumulative integral of each document. Based on this probability data, we classify this uncategorized data to the category with highest the probability.

III. System

A. Process of the system

As shown in figure 3.1, the proposed system firstly summarizes the occurrence of each term in training documents and uses improved TF-IDF to build weighting table. Thereby, the system computes the sum of the weight of each test document relative to each category. In addition, as the document has cross-field property, the highest weight or second highest weight will be selected as the final classification result, on the basis of category priority.

B. TF-IDF Improvement

The traditional TF-IDF method did not consider the distribution of feature terms over different categories; therefore, it may not only discriminate important words since they occur less times in the document, but may also be less useful in reducing noise terms because *IDF(t)* is close to zero. To differentiate meaningless noise terms and important feature terms, the category

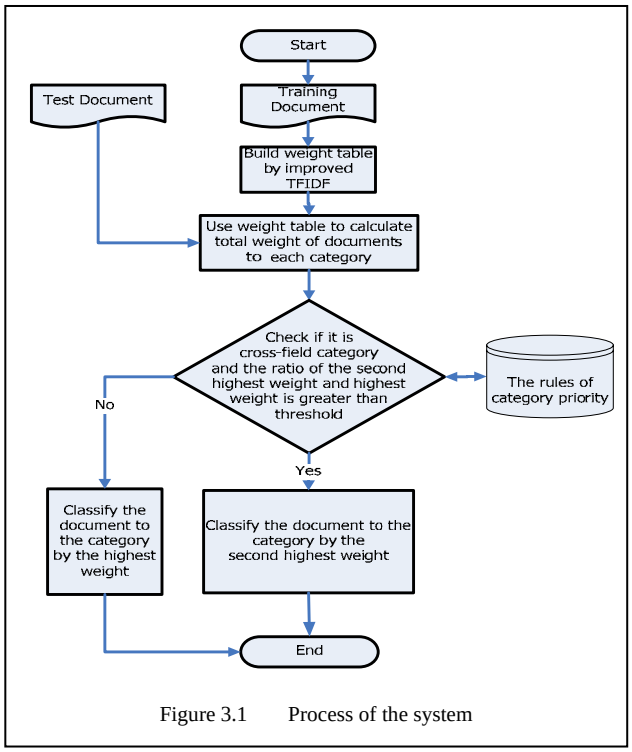

Figure 3.1 Process of the system

distribution of feature terms should be taken into consideration. Accordingly, the improved TF-IDF formulas are given as bellows:

$$W(t,c)=TF(t,c) \times IDF(t,c) \tag{4}$$

$$IDF(t)=log\left[\frac{TF(t)}{TF(t)\text{-}TF(t,c)+1}\right]\times log\left[\frac{D}{DF(t)}\right]\times log\left[\frac{\alpha}{\beta}\right]^2 \tag{5}$$

where ，*TF(t)*：*frequency of term t*
TF(t,c)：*frequency of term t at category c*
D：*total number of documents*
DF(t)：*document frequency of term t*
α：*total number of categories*
β：*number of category where term t appears*

The improved *IDF* is composed of three items: the first item is to determine discretion of term *t* by its concentration at one category, that is, the higher the occurrence of term *t* is at the category, the smaller the denominator is (near 1). The lower the occurrence of term *t* at the category, the bigger the denominator (near numerator) so that *IDF* gets lower; the second item is traditional *IDF*; the third item is to determine the distribution of term *t* over categories by quantity of categories where term *t* is distributed. If term *t* appears at only a few categories, then *IDF* will rise; however, if term *t* appears at multiple categories, then *IDF* will fall.

For example, as shown in Table 3.1, the traditional TF-IDF weighting method fails to filter the unimportant words “研究” (study) and “結果”(result). Unlike the traditional TF-IDF weighting method, the improved TF-IDF method is able to use category distribution parameter to reduce the influence of noise terms. As shown in Table 3.1, weights of “研究” (study) and “結果”(result) are reduced to zero.

TABLE 3.1
AFTER IMPROVED, THE WEIGHT OF USELESS KEYWORD IS WEAKENED

Class	Feature	Class Freq.	Total Freq.	Class Num.	Traditional TF-IDF	Improved TF-IDF
大氣 Atmosphere	研究 Study	298	5394	6	34.65	0.0
大氣 Atmosphere	結果 Result	370	2199	6	90.25	0.0
音樂 Music	研究 Study	492	5394	6	57.21	0.0
音樂 Music	結果 Result	68	2199	6	16.59	0.0
化學 Chemistry	研究 Study	455	5394	6	52.91	0.0
化學 Chemistry	結果 Result	332	2199	6	80.98	0.0
教育 Education	研究 Study	2446	5394	6	284.43	0.0
教育 Education	結果 Result	344	2199	6	83.91	0.0
土木 Civil	研究 Study	1309	5394	6	152.22	0.0
土木 Civil	結果 Result	728	2199	6	177.57	0.0
生物 Biology	研究 Study	394	5394	6	45.82	0.0
生物 Biology	結果 Result	357	2199	6	87.08	0.0

Another example shows that, according to distribution at various categories, the improved TF-IDF method can strengthen or weaken the importance of the same keyword with respect to different categories at the same time. For example, as shown in table 3.2, "颱風" (typhoon) has weight at both "Civil" category and "Atmosphere" category. Comparing with the weights of "颱風" (typhoon) of the improved TF-IDF and that of the traditional TF-IDF weighting method, obviously, the weight of "颱風" (typhoon) at category "Civil" is weakened; and another example, "鋼琴" (piano) has weight at both "Music" and " Education" categories. After being improved, the weight at category "Music" was strengthened and at " Education" is reduced to zero. Therefore, the improved TF-IDF method can strengthen keyword differentiation at various categories, so as to increase classification accuracy.

C. Category Priority

As mentioned in the introduction, the data sources are thesis abstracts of universities in Taiwan and are classified by departments which release the theses in this study. Since some theses may cross two different departments, this situation may cause classification error. For example, a document originally belonging in the "Chemistry" category with an abstract that contains biological keywords, such as "bacteria", "infection", etc, will be classified in the "Biology" category because the weight relative to "Biology" is greater than to "Chemistry". In order to identify the major department

TABLE 3.2
AFTER IMPROVED, VARIATION OF KEYWORD AT DIFFERENT CATEGORIES ARE STRENGTHENED

Class	Feature	Class Freq.	Total Freq.	Class Num.	Traditional TF-IDF	Improved TF-IDF
音樂 Music	作品 Work	518	539	2	618.66	820.09
教育 Education	作品 Work	21	539	2	25.08	0.39
教育 Education	音樂 Music	45	1162	2	46.56	0.74
音樂 Music	音樂 Music	1117	1162	2	1155.67	1546.59
大氣 Atmosphere	氣流 Air flow	199	201	2	290.97	507.03
土木 Civil	氣流 Air flow	2	201	2	2.92	6.04
土木 Civil	混凝土 Concrete	582	587	2	718.18	1364.12
化學 Chemistry	混凝土 Concrete	5	587	2	6.17	1.75
生物 Biology	蛋白 Protein	997	1018	2	1053.47	1674.09
化學 Chemistry	蛋白 Protein	21	1018	2	22.19	0.18
土木 Civil	颱風 Typhoon	33	368	2	52.95	1.99
大氣 Atmosphere	颱風 Typhoon	335	368	2	537.50	530.54
音樂 Music	鋼琴 Piano	317	318	2	485.41	1019.68
教育 Education	鋼琴 Piano	1	318	2	1.53	0.0

of these theses, different priorities should be given to these two different categories. In this paper, we propose a simple approach, category priority, to solve this problem. Since we found that, the chemistry department often produces theses similar to biology, but the biology department seldom produces theses similar to chemistry, we can use category priority to describe the above relationship of this kind of cross-field theses. In this case, the priority of the "Chemistry" category is higher than that of the "Biology" category and this relationship can be represented as follow:

$$\text{Chemistry} \rightarrow \text{Biology}. \quad (6)$$

Moreover, since these theses are crossing biology and chemistry fields, weights of these theses relative to the "Chemistry" category and the "Biology" category should be higher than those to other departments. Accordingly, we can define the following algorithm, as shown in the following algorithm, to classify cross-field theses.

The Algorithm Classify Cross-Field Theses.

/* Let the document *D* be a thesis crossing "A" and "B" fields, and the category priority of "A" be higher than that of "B"*/

{ If the weight of thesis to "A" is highest,
Then this thesis belongs to "A" category,
Else if the ratio of weight between "A" and "B" reaches a certain threshold,
Then this thesis belongs to "A" category,
Else this thesis belongs to "B" category. }

To classify these theses which are crossing biology and chemistry fields by the above algorithm, we can select the documents where the highest weight is the "Biology" category and second highest weight is the "Chemistry" category. If the ratio of weight between the "Chemistry" category and the "Biology" category reaches a certain threshold, we can classify this document to the "Chemistry" category according to its second highest weight. This study sets the threshold as 0.6 by the experimental experience. For example, comparing the experimental results of using the improved TF-IDF method without and with category priorities, as shown in Table 3.3, when category priority Chemistry → Biology is used, there are 15 theses are classified from "Biology" category into "Chemistry" category and 14 theses are classified correctly.

TABLE 3.3
NUMBER OF DOCUMENTS ARE CLASSIFIED FROM ORIGINAL CATEGORY TO NEW CATEGORY BY THE IMPROVED TF-IDF WITH CATEGORY PRIORITIES

Highest Class	Second Class	Effected Doc.	Wrong→ Correct	Correct→ Wrong	Wrong→ Wrong
生物 Biology	化學 Chemistry	15	14	1	0
土木 Civil	化學 Chemistry	10	8	1	1
教育 Education	土木 Civil	13	9	2	2

Moreover, some theses of the chemistry department may also apply related chemistry knowledge to the field of the civil department, thus in document classification, it may be wrongly classified into the "Civil" category. For example, a document originally belonging in the "Chemistry" category which has an abstract that contains keywords of a "Civil engineering" nature, such as "concrete", "cement", etc, will be classified in the "Civil" category because the weight relative to "Civil" is greater than to "Chemistry". For the same reason between the "Chemistry" category and the "Biology" category, we define that the priority of the "Chemistry" category is higher than that of the "Civil" category. This relationship can be represented as follow:

Chemistry → Civil. (7)

As shown in Table 3.4, when the traditional TF-IDF method and category priority Chemistry → Civil are used, there are 65 theses which are classified from the "Civil" category into the "Chemistry" category and 52 theses are correctly classified. Moreover, as shown in Table 3.3, when improved TF-IDF method and category priority Chemistry → Civil are used, there are 10 theses classified from the "Civil" category into the "Chemistry" category and 8 theses are classified correctly. Clearly, the accuracy rate and recall rate are improved when category priorities are used.

TABLE 3.4
NUMBER OF DOCUMENTS ARE CLASSIFIED FROM ORIGINAL CATEGORY TO NEW CATEGORY BY THE TRADITIONAL TF-IDF WITH CATEGORY PRIORITIES

Highest Class	Second Class	Effected Doc.	Wrong→ Correct	Correct→ Wrong	Wrong→ Wrong
生物 Biology	化學 Chemistry	31	15	15	1
土木 Civil	化學 Chemistry	65	52	8	5
教育 Education	土木 Civil	26	16	4	6

IV. EXPERIMENTAL RESULTS

This article selected 6065 thesis abstracts from the "National Dissertations and Theses Information Web" as the document profile, and categorized them on the basis of six departments [7]. The thesis distribution is shown in Table 4.1. We select 10% of the documents as training data to build a classification model, and 30% of the documents as testing data.

TABLE 4.1
NUMBER OF DOCUMENTS SELECTED FROM EACH DEPARTMENT

Dept. Name	Document Num.
土木 Civil	1794
生物 Biology	1004
化學 Chemistry	1003
大氣 Atmosphere	670
音樂 Music	658
教育 Education	936
Total	6065

After analysis, it was found that Chinese and English words are mixed in thesis abstracts, documents of the chemistry department contain abbreviated chemistry formulae and compound names, and a minority of articles has no abstracts. Therefore, we have to remove characters except Chinese words and use the Chinese Punctuation System [4] to make Chinese punctuation. Moreover, since theses may cross two different fields, we have to define the category priorities before classification. The category priorities with respect to these six departments are:

Chemistry → Biology. (8)
Chemistry →Civil. (9)
Civil → Education. (10)

After that, we use traditional TF-IDF method and improved TF-IDF method with and without category priority to classify documents. The effect of document classification can be evaluated by Recall rate and Accuracy rate, while Recall rate and Accuracy rate for a category "A" are defined as follows:

$$\text{Recall rate} = \alpha / (\beta+\alpha) \quad (11)$$

$$\text{Accuracy rate} = \alpha / (\gamma+\alpha) \quad (12)$$

Where α is the number of documents which belong to "A" category and they are also classified into "A" category,

β is the number of documents which belong to "A" category, but they are not classified into "A" category,

γ is the number of documents which do not belong to "A" category, but they are classified into "A" category.

Comparing the traditional TF-IDF method and improved TF-IDF method, as shown in Table 4.2 and 4.3, we find that classification number, recall rate and accuracy rate are improved by our method. Without considering category priority, the recall rate and accuracy rate are increased from 85.95% in traditional TF-IDF to 91.03% in improved TF-IDF. When traditional TF-IDF is used, the recall rate and accuracy rate are increased from 85.95% without category priority to 89.12% with category priorities. When category priorities are used, the recall rate and accuracy rate are increased from 89.12% in traditional TF-IDF method to 92.56% in improved TF-IDF method.

Apart from improving TF-IDF weight algorithm, the biggest breakthrough of this study is to exploit the special relationship between categories to pay different priorities to different categories. It was found from the experiment result, as shown in Table 4.2 and Table 4.3, when using category priority the performance of classification is better than when not using it. When improved TF-IDF and category priorities are used, the recall rate and accuracy rate of each category perform well; most of the data reach above 90%, except that recall of "Chemistry" category appears lower than other data. However, the "Chemistry" category has the biggest improvement. Its accurate classification number is increased from 158 in to 201. The recall is increased from 58.74% to 74.72%. Although the accuracy rate drops a little, on the whole, the "Chemistry" category classification is significantly improved. Also, the other categories have relative improvement.

TABLE 4.2
COMPARISON OF CORRECT CLASSIFICATION AND PREDICTION NUMBER

Class	Total	Traditional TF-IDF Correct	Traditional TF-IDF+ Priority Correct	Improved TF-IDF Correct	Improved TF-IDF+ Priority Correct	Traditional TF-IDF Predict	Traditional TF-IDF+ Priority Predict	Improved TF-IDF Predict	Improved TF-IDF+ Priority Predict
化學 Chemistry	269	158	225	179	201	169	265	191	216
音樂 Music	197	168	168	191	191	170	170	195	195
教育 Education	276	272	268	271	269	332	306	306	293
土木 Civil	535	513	521	499	507	653	614	533	536
大氣 Atmosphere	201	128	128	186	186	128	128	203	203
生物 Biology	295	285	270	288	287	321	290	345	330
Total	1773	1524	1580	1614	1641	1773	1773	1773	1773

TABLE 4.3
COMPARISON OF ACCURACY AND RECALL RATE

Class	Traditional TF-IDF Accuracy Rate	Traditional TF-IDF+ Priority Accuracy Rate	Improved TF-IDF Accuracy Rate	Improved TF-IDF+ Priority Accuracy Rate	Traditional TF-IDF Recall Rate	Traditional TF-IDF+ Priority Recall Rate	Improved TF-IDF Recall Rate	Improved TF-IDF+ Priority Recall Rate
化學 Chemistry	93.49	84.91	93.71	93.06	58.74	83.64	66.54	74.72
音樂 Music	98.82	98.82	97.94	97.95	85.28	85.28	96.95	96.95
教育 Education	81.92	87.58	88.56	91.81	98.55	97.10	98.18	97.46
土木 Civil	78.56	84.85	93.62	94.59	95.89	97.38	93.27	94.77
大氣 Atmosphere	100.0	100.0	91.62	91.63	63.68	63.68	92.53	92.54
生物 Biology	88.78	93.10	83.47	86.97	96.61	91.53	97.62	97.29
Total	85.95	89.12	91.03	92.56	85.96	89.12	91.03	92.56

V. CONCLUSION

In this paper, we improve traditional TF-IDF to compute terms' weights. Besides, in order to cope with interdisciplinary research case, we introduce category priorities to solve cross fields problem. It does not select the highest weighted category, but chooses the second highest weighted category in some cases. The experiment results verified that it can get better classification results than when not using category priority.

This study is to categorize Chinese documents, and the results may also be achieved when applying it to process English documents. We also plan to use weighted classification method coupled with data mining approach to find some useful rules so as to increase classification accuracy. These shall serve as the direction of our further study.

REFERENCES

[1] V. Vapnik , S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal Processing," *Neural Information Processing Systems 9*, pp. 281-287, 1997

[2] Y. Huang, J. Tan, and L. Zhang, "A context analytical method basing on text structure," *Journal of Software*, vol. 4, no. 1, pp.3-10, February 2009.

[3] J. Myung, J.-Y. Yang, and S.-G. Lee, "Picachoo: A text analysis tool for customizable feature selection with dynamic composition of primitive methods," *Journal of Software*, vol. 5, no. 2, pp.179-186, February 2010.

[4] Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica, "Chinese Word Segmentation System," http://ckipsvr.iis.sinica.edu.tw/.

[5] D. A. Chiang, H. C. Keh, H. H. Huang, and D. Chyr, "The chinese text categorization system with association rule and category priority," *Expert Systems with Applications*, vol. 35, no. 1-2, pp. 102-110, 2008.

[6] I. Díaz, J. Ranilla, N. Elena Monta, J. Fernández, and E. F. Combarro, "Improving performance of text categorization by combining filtering and support vector machines: Research articles," *J. Am. Soc. Information Science and Technology*, vol. 55, no. 7, pp. 579-592, May 2004.

[7] National Central Library, "Electronic Theses and Dissertations System," http://etds.ncl.edu.tw/theabs/index.jsp.

[8] K. A. Hamill and A. Zamora, "The use of titles for automatic document classification," *Journal of the American Society for Information Science*, vol. 31, no. 6, pp. 396-402, 1980.

[9] K. L. Kwok, "The Use of Title and Cited Titles as Document Representation for Automatic Classification," *Journal of Information and Management*, Vol. 11, pp. 201-206, 1975.

[10] M.E. Maron, "Automatic Indexing : an Experimental Inquiry," *J. of the ACM*, Vol. 8, pp. 404-417, 1961.

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

[12] K. Aas and L. Eikvil. "Text categorisation: A survey," *Technical report, Norwegian Computing Center*, 1999.

[13] Mingyu Lu, Keyun Hu, Yi Wu, Yuchang Lu, and Lizhu Zho, "SECTCS: towards improving VSM and Naive Bayesian classifier:Systems, Man and Cybernetics," *IEEE International Conference on 2002* , Vol. 5, pp. 6-9, Oct. 2002.

[14] Elı´as F. Combarro, Elena Montan˜e´ s, Irene Dı´az, Jose´ Ranilla, and Ricardo Mones, "Introducing a Family of Linear Measures for Feature Selection in Text Categorization," *IEEE Transactions on knowledge and data engineering*, vol. 17, no. 9, pp. 1223-1232, September 2005.

[15] D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka, "Training algorithms for linear text classifiers," *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pp. 298-306, 1996.

[16] Thorsten Joachims, "A probabilistic analysis of the Rocchio Algorithm with TF-IDF for text categorization," *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pp. 143-151, 1997.

[17] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, March 1986.

[18] K. R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," *in ICANN '97: Proceedings of the 7th International Conference on Artificial Neural Networks*. London, UK: Springer-Verlag, pp. 999-1004, 1997.

[19] P. Soucy, and G.-W. Mineau, "A simple KNN algorithm for text categorization,"*Proceedings IEEE International Conference on Data Mining(ICDM 2001)*, pp. 64-68, 29 Nov.-2 Dec. 2001.

[20] K.-M. Schneider, "Techniques for improving the performance of naive bayes for text classification," *in Computational Linguistics and Intelligent Text Processing*, pp. 682-693, 2005.