Refinement of Feature Terms and Improvement of Classification Accuracy on Multilingual Text Categorization Using Character *N*-gram

Makoto Suzuki[1], **Yi-Ching Tsai**[2], **Takashi Ishida**[3], **Masayuki Goto**[3], **Shigeichi Hirasawa**[3]

[1]Shonan Institute of Technology, 251-8511, Japan
(m-suzuki@info.shonan-it.ac.jp)
[2]Leader University, Tainan City 70901, Taiwan
(tsaiyiching@mail.leader.edu.tw)
[3]Waseda University, 169-8555, Japan
(tishida@fuji.waseda.jp) (masagoto@waseda.jp) (hira@waseda.jp)

ABSTRACT

In our previous paper, we proposed a new classification technique called the Frequency Ratio Accumulation Method (FRAM). This is a simple technique that adds up the ratios of term frequency among categories. However, in FRAM, the use of feature terms is unlimited. Then, we adopt Character *N*-gram as feature terms improving the particularity of FRAM. That is to say, the proposed method is language-independent because it does not depend on grammatical knowledge peculiar to the language by using Character *N*-gram. Therefore, we can classify multi-language into some categories using only one program. In the present paper, we will refine the DB of the feature term set using mutual information and frequency ratio, and improve the classification accuracy. Next, the proposed method is evaluated by performing several experiments. In these experiments, we classify newspaper articles from English Reuters-21578 that provides benchmark data in automatic text categorization, Japanese CD-Mainichi 2002 and Chinese China Times 2005 using FRAM. As a result, we show the effectiveness of refining DB of feature terms. Specifically, the macro-averaged F-measure of the proposed method is 92.3% for Reuters-21578.

Keywords: Text Mining, Classification, *N*-gram, Newspaper

1. INTRODUCTION

This paper discusses automatic text categorization, which is to select an appropriate category from a pre-defined set of categories, given a document [1].

In general, the processing of automatic text categorization involves two important phases. The first phase, namely the training phase, is the extraction of feature terms that will give effective keywords in the test phase; and the second is the actual classification of documents using the feature terms of the test phase. In the present paper, we refer to the former as the feature selection stage and the latter as the document classification stage.

One word is usually considered to be one feature term in the feature selection stage. In the language delimited by a space, in English, for example, there is no need to extract words. However, for Japanese, words should be extracted by morphological analysis. In contrast, a method to generate these feature terms using *N*-gram has been proposed as a language-independent technique [2],[3]. In any case, most of these conventional techniques extract useful feature terms from many words by using mutual information, TFIDF values etc. [4], and these feature terms are used for classification.

On the other hand, the categorization at the document classification stage is a traditional problem of machine learning, and machine learning algorithms are often used, such as, Neural Network [5], Decision Trees [6],[7], Naive Bayes Method [8], k-Nearest Neighbor [9] and Boosting Algorithms [10], as well as Support Vector Machines (SVM) [11].

In our previous paper, we proposed a new classification technique called the Frequency Ratio Accumulation Method (FRAM) [12]. This is a simple technique that adds up the ratios of term frequency among categories. However, in FRAM, the use of feature terms is unlimited. Then, we adopt Character *N*-gram as feature terms improving the above-described particularity of FRAM.

In the present paper, we will refine the DB of the feature term set using mutual information and frequency ratio, and improve the classification accuracy. Next, the proposed method is evaluated by performing several experiments. In these experiments, we classify newspaper articles from English Reuters-21578, Japanese CD-Mainichi 2002 and Chinese China Times 2005 using FRAM. Here, Reuters-21578 provides benchmark data in automatic text categorization. As a result, we show the effectiveness of refining DB of feature terms. Specifically, the macro-averaged F-measure of the proposed method is 92.3% for Reuters-21578.

2. TEXT CATEGORIZATION

2.1 Overview

In this study, the goal of text categorization is to classify the given new documents into a fixed number of pre-defined categories. Figure 1 shows a flow diagram of the text categorization task [13].

The procedure for automatic text categorization is divided into two phases, the training phase and the test phase. In the training phase, the training documents are input along with a category. Next, the feature terms are

—	Operating cost
	Operating gross profit
—	Operating expenditure
	Operating profit (loss)
±	Non-operating income and profit (expense and loss)
	Earning before tax margin (loss)
—	Income tax
	After-tax earning per share

3. EMPIRICAL RESULTS

3.1 Distribution of the Samples

By the end of 2007, there are 1238 listed companies in Taiwan, including 378 listed non-electronic companies; among others, Sunspring Metal Corporation (stock code: 2062) and Taiwan Prosperity Chemical Corporation (4725) do not have complete five-year financial statement and thus they are eliminated. Besides, Fu Sheng Group (1520), TFC (1601) and Norman International Ltd. (9915) announce the suspension of public offering in Jan. 9, 2008, Jan. 22, 2008 and March 3, 2008 and their financial statements in 2007 are not compelte and thus they are eliminated. Upon the selection above, there are **373** qualified listed non-electronic companies.

3.2 Category of Business Models

After statistics, the number of the companies of different types of business models in the listed non-electronic companies in Taiwan is shown in Table 3-1.

Table 3-1: Statistics of the number of listed non-electronic companies according in different types of business models (2003 to 2007)

Type	Type of business models	Number of companies	Percentage of business models
Type 1	International brand marketing	11	2.95%
Type 2	OEM, ODM, EMS	20	5.36%
Type 3	Domestic demand	202	54.16%
Type 4	Mix	119	31.90%
Type 5	Niche innovation	21	5.63%

Source: This study.

The number of the companies in different business models according to industry category is shown in Table 3-2.

Table 3-2: Statistics of business models according to industry category (2003 to 2007)

Type	Number of compan-ies in manufacturing	Number of compan-ies in service industry	Number of compan-ies in finance and banking	Total
Type 1	9	1	1	11
Type 2	20			20
Type 3	95	70	37	202
Type 4	110	9		119
Type 5	15	6		21
Total	249	86	38	373

Source: This study.

As to the percentage of business models of Taiwan listed non-electronic companies, Type 3-- "domestic demand" is the first (54.16%); among 21 industrial sectors, Cement sector (7 out of 7 companies), Electrical and Cable sector (10 out of 13 companies), Construction sector (36 out of 36 companies) and Financial sector (36 out of 37 companies) tend to operate in Taiwan, Penghu, Kinmen and Matsu and the profits are mainly from domestic demand market. They are domestic demand business model; besides, service industry (70 out of 86 companies) and finance and banking (37 out of 38 companies) are also domestic demand business model.

The second is Type 4-- "mix" (31.90%); most of the companies are manufacturing and among 21 industrial sectors, the percentages of Plastic sector (12 out of 21 companies), Textile sector (28 out of 46 companies), Electric Machinery sector (19 out of 34 companies) and Rubber sector (6 out of 9 companies) are more significant.

The third is Type 5-- "niche innovation" (5.63%); most of the companies which meet the five meanings of "niche innovation" are in manufacturing industry (15 out of 21 companies); industrial sectors refer to Food, Plastic, Textile, Electric Machinery, Chemical, Biotechnology and Medicine, Rubber, Shipping and Transportation, Consumers Goods and Others.

The fourth is Type 2-- "OEM, ODM and EMS" (5.36%); among Taiwan listed non-electronic companies, this type of companies are less than those of listed electronic companies and most of them are the small and medium enterprises with final capital stock less than ten billion. Thus, this study combines large-scale and small and medium OEM, ODM and EMS as one type. Type 2 companies are in manufacturing industry and industrial sectors refer to Plastic, Textile, Electric Machinery, Chemical, Biotechnology and Medicine and Others.

The fifth is Type 1-- "international brand marketing" (2.95%); it refers to "top 20 international brands in Taiwan" and well-known Taiwan listed non-electronic companies in the world. Thus, there are only 11 companies; Johnson, Cheng Shin, DEPO, Merida, Giant and Uni-President are among top 20 international brands in Taiwan in 2007; besides, Evergreen Marine is the fourth among AXS-Liner global Shipping and Transportation; Nan Ya Plastic, Cathay Financial extracted via a feature selection process and an indices database is produced, referred to herein as "feature term set (DB)", which is later used for the test phase. In the test phase, new documents to be classified are input one after another, and one category is allocated in these documents with a classifier that uses Naive Bayes Method, SVM, our proposed method, and so on. Finally, the classification results of each technique are evaluated.

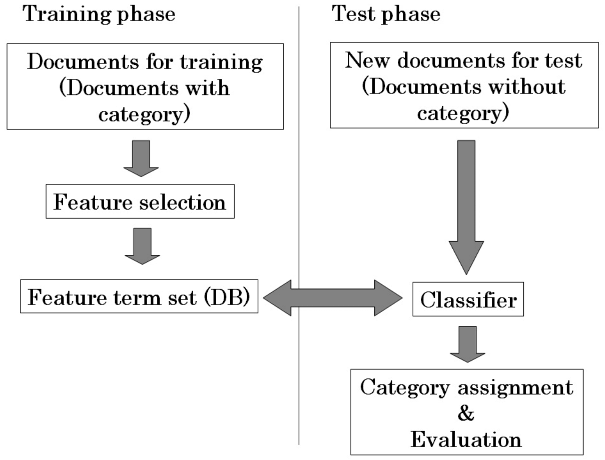
Training phase
Test phase
Documents for training (Documents with category)
New documents for test (Documents without category)
Feature selection
Feature term set (DB)
Classifier
Category assignment & Evaluation

Figure 1: Flow diagram of text categorization.

2.2 Mathematical formulation

In the present paper, the following notations are used.

Definition 1: Document Set

$$D = \{d_i \mid i = 1,2,...,I\} \tag{1}$$

d_i : a document
I : total number of all documents

Definition 2: Word Set

$$W = \{w_j \mid j = 1,2,...,J\} \tag{2}$$

w_j : a word
J : total number of words contained in all documents

Document d_i can be expressed as a sequence of a number of words in the word set W.

Definition 3: Document

$$d_i = \langle w_{i1} w_{i2} ... w_{iL_i} \rangle \tag{3}$$

L_i : total number of words contained in the document (length of document d_i)
w_{il} : a word ($w_{il} \in W, l = 1,2,...,L_i, L_i \leq J$)

In addition, the set of categories to which each document belongs is written as follows.

Definition 4: Category Set

$$C = \{c_k \mid k = 1,2,...,K\} \tag{4}$$

c_k : a category
K : total number of categories

Using the notations mentioned above, the problem of automatic text categorization in the present paper is to classify a new document d_i into a pre-defined category c_k using words w_{il} included therein.

2.3 Feature terms and document vector

Usually, text data stored in the DB of feature terms is composed of character strings of a suitable form for learning and classification. These character strings that perform well in classification are extracted as feature terms in several previous studies. Moreover, documents are expressed using TFIDF values and the like so that they can be processed by the computer. Thus, feature selection plays an important role in achieving good classification performance. In the present paper, a set of *M* extracted feature terms are expressed as follows.

Definition 5: Feature Term Set

$$T = \{t_m \mid m = 1,2,...,M\} \tag{5}$$

M : total number of all feature terms

For example, when a word is used as a feature term, $T \supseteq W$; that is, one feature term t_m corresponds to one word w_j. On the other hand, when the *N*-gram of each character (hereafter, referred to as "Character *N*-gram") is extracted as a feature term t_m, one feature term corresponds to a string of *N* characters. In addition, when the *N*-gram of each word (hereafter referred to as "Word *N*-gram") is extracted as a feature term t_m, one feature term corresponds to a block composed of *N* words, such as $\langle w_j, \cdots, w_{j+N-1} \rangle$.

Here, one type of feature selection methods is adopted. Character *N*-gram is effective as a language-independent method because it does not depend on the meaning of the language. Three examples are shown as follows. That is, the first example concerning English is shown in Table 1, the second example concerning Japanese is shown in Table 2 and the third example concerning Chinese is shown in Table 3[1].

Table 1: Example of English feature terms

Original sentence	**He is sincere.**
Word	**He/is/sincere**
Character N-gram (N=2)	**He/e / i/is/s / s/si/in/nc/ce/er/re**

Table 2: Example of Japanese feature terms

Original sentence	**彼は誠実です.**
Word	**彼/は/誠実/です**
Character N-gram (N=2)	**彼は/は誠/誠実/実で/です**

Table 3: Example of Chinese feature terms

Original sentence	**他是誠實的.**
Word	**他/是/誠實/的**
Character N-gram (N=2)	**他是/是誠/誠實/實的**

Here, each document is expressed as a vector of *M* dimensions using *M* selected feature terms.

Definition 6: Document Vector

$$\vec{d}_i = (t_1, t_2, ..., t_M) \tag{6}$$

[1] The sentence in Table 1, "He is sincere.", the sentence in Table 2, "彼は誠実です", and the sentence in Table 3, "他是誠實的" mean the same thing.

3. PROPOSED METHOD

In this study, a close relationship is assumed between the selection method for feature terms and classification method using these terms in automatic text categorization. Therefore, instead of separately discussing the feature selection and the classifier shown in Figure 1, they must be thought as steps in a single information processing flow while considering the congeniality between the feature selection and the classifier. In our previous paper, extraction of the feature terms and classification using these terms are considered, and the new text categorization technique mentioned below was proposed.

3.1 Classification method using the sum of frequency ratios

In our previous paper, we proposed a classification method using the sum of frequency ratios in each category of an individual feature term [12]. We call the Frequency Ratio Accumulation Method (FRAM). First, we define “frequency ratio” as follows.

Definition 7: Frequency Ratio

$$FR(t_m,c_k) = \frac{R(t_m,c_k)}{\sum_{c_k \in C} R(t_m,c_k)} \tag{7}$$

Where,

$$R(t_m,c_k) = \frac{f_{c_k}(t_m)}{\sum_{t_m \in T} f_{c_k}(t_m)}$$

$f_{c_k}(t_m)$: total frequency of the feature term t_m in a category c_k

In the training phase, the frequency ratios of all feature terms are calculated and maintained for each category. Next, “category score” which indicate the possibility that the target document belongs to the category, are calculated as follows.

Definition 8: Category Score

$$E_{d_i}(c_k) = \sum_{t_m \in d_i} FR(t_m,c_k) \tag{8}$$

Finally, the target document d_i is classified into the category $c_{\hat{k}}$ for which the category score is the maximum as follows.

$$c_{\hat{k}} = \arg\max_{c_k \in C} E_{d_i}(c_k) \tag{9}$$

That is to say, the proposed method maintains M (total number of the feature terms) * K (total number of categories) frequency ratios in the training phase and calculates category scores for each category by adding frequency ratios when a target document includes the feature term in the test phase and classifies the feature term into the category for which the evaluation score is the maximum. The advantage of the proposed method is the ability to maintain feature terms almost without limitation because the calculation is simple.

3.2 Extraction of feature terms using *N*-gram

As shown in equation (8), a document is classified by taking into account both the total frequency of each individual feature term included in the document and its frequency ratio in each category. Thus, the proposed method will reduce computational complexity and enable unlimited use of feature terms. Therefore, *N*-gram is used for feature selection, making the most of the particularity that feature term use can be unlimited. Consequently, we adopted Character *N*-gram as one type of feature selection methods. Character *N*-gram is effective as a language-independent method because it does not depend on the meaning of the language.

4. REFINEMENT OF FEATURE TERM SET

In the present paper, we will refine the DB of feature term set using two method, i.e. mutual information and frequency ratio in document, as follows.

4.1 Case of mutual information

First of all, we use the following mutual information as a criterion to select feature terms. That is, M terms with positive mutual information are extracted as feature terms[1].

Definition 9: Mutual Information

$$I(t_m;C) = \sum_{k=1}^{K} P(t_m,c_k)\log\frac{P(t_m,c_k)}{P(t_m)P(c_k)} \tag{10}$$

t_m : a word

C: set of categories

c_k: a category ($c_k \in C$)

$P(t_m,c_k)$: occurrence probability of documents having a feature term t_m as an element and belonging to category c_k in an entire set of documents

$P(t_m)$: occurrence probability of documents having a feature term t_m as an element in an entire set of documents

$P(c_k)$: occurrence probability of documents belonging to category c_k in an entire set of documents

4.2 Case of frequency ratio in document

Second, we use the following frequency ratio in each document as a criterion to select feature terms.

Definition 10: Frequency Ratio in Document

$$R_{d_i}(t_m,c_k) = \frac{f_{d_i}(t_m,c_k)}{\sum_{c_k \in C} f_{d_i}(t_m,c_k)} \tag{11}$$

[2] That is to say, we drop Character N-gram with $I(t_m;C)=0.$

Here, we compute $P(t_m,c_k)\,log\,\frac{P(t_m,c_k)}{P(t_m)P(c_k)} = 0$

if $P(t_m,c_k)=0.$

$f_{d_i}(t_m, c_k)$: total frequency of the feature term t_m per category c_k in a document d_i

That is, we select terms that meets the following conditions as feature terms[2].

$$\max_{c_k \in C} R_{d_i}(t_m, c_k) > \alpha \qquad (12)$$

5. EXPERIMENT

5.1 Experimental conditions

The present experiment involved three newspapers that contain articles with pre-assigned categories. The first is English Reuters-21578 (Reuters)[3], the second is Japanese CD-Mainichi Newspaper 2002 (Mainichi)[4] and the third is Chinese China Times Newspaper 2005 (China Times)[5].

In the present experiment, for each method, the computer was first made to learn using training data with pre-assigned categories in the training phase. Second, in the test phase, the test data was given to the computer without showing the true categories, requiring the computer to classify it.

We perform experiments with three methods in Tables 4-6 respectively. The first method is FRAM without filtering. This is the same method proposed in our previous paper [12]. The second method is FRAM with filtering based on mutual information in equation (10). The third method is FRAM with filtering based on frequency ratio of document in equation (11). These methods are denoted by “Proposed Method 0-2” in Table 4-6 for English Reuters, Japanese Mainichi and Chinese China Times. In the present paper, an evaluation is made using *recall* and *precision* [14].

i. Experiments using Reuters

Reuters-21578 provides benchmark data in automatic text categorization. *Apte split 10 categories* was used for Reuters-21578. *Apte split 10 categories* is benchmark data that extracts ten categories such as Acquisition, Corn, Crude, Earn, Grain, Interest, Money-fx, Ship, Trade, and Wheat from Reuters-21578 [1].

ii. Experiments using Mainichi

In addition, for each of the seven categories, such as Society, Sports, Entertainment, Home, Economy, International relations and Leaders, included in the CD-Mainichi Newspaper 2002, we randomly selected 1,000 training documents and 500 test documents (7,000 and 3,500 documents in total, respectively).

iii. Experiments using China Times

Equally, for each of the seven categories, such as Society- Domestic administration, Sports, Art-Culture, Education, Economy-Trade, Politics and Science technology- Information, included in the China Times Newspaper 2005, we randomly selected 1,000 training documents and 500 test documents (7,000 and 3,500 documents in total, respectively).

Table 4：Each method for English Reuters

Ex.-No.	*Proposed Method 0 (without filtering)*	*Proposed Method 1 (Mutual Information)*	*Proposed Method 2 (Frequency Ratio)*
1	Character 9-gram	Character 5-gram	Character 5-gram
2	Character 10-gram	Character 6-gram	Character 6-gram
3	Character 11-gram	Character 7-gram	Character 7-gram
4	Character 12-gram	Character 8-gram	Character 8-gram
5	Character 13-gram	Character 9-gram	Character 9-gram
6	Character 14-gram	Character 10-gram	Character 10-gram
7	Character 15-gram	Character 11-gram	Character 11-gram

Table 5: Each method for Japanese Mainichi

Ex.-No.	*Proposed Method 0 (without filtering)*	*Proposed Method 1 (Mutual Information)*	*Proposed Method 2 (Frequency Ratio)*
1	Character 2-gram	Character 2-gram	Character 2-gram
2	Character 3-gram	Character 3-gram	Character 3-gram
3	Character 4-gram	Character 4-gram	Character 4-gram
4	Character 5-gram	Character 5-gram	Character 5-gram
5	Character 6gram	Character 6gram	Character 6gram

Table 6: Each method for Chinese China Times

Ex.-No.	*Proposed Method 0 (without filtering)*	*Proposed Method 1 (Mutual Information)*	*Proposed Method 2 (Frequency Ratio)*
1	Character 1-gram	Character 1-gram	Character 1-gram
2	Character 2-gram	Character 2-gram	Character 2-gram
3	Character 3-gram	Character 3-gram	Character 3-gram
4	Character 4-gram	Character 4-gram	Character 4-gram

5.2 Results

The results are shown in Figures 3-11. Figures 3-5, 6-8, and 9-11 show the results concerning English (Reuters), Japanese (Mainichi), and Chinese (China Times) respectively. Here, the performance was measured as micro-averaged precision (miP), micro-averaged recall (miR), micro-averaged F-measure (miF), macro-averaged precision (maP), macro-averaged recall (maR) and macro-averaged F-measure (maF). The legend of these graphs is shown in Figure 2. In Figure 3-11, the left side bold number is miF, and the right side is maF. Moreover, the left vertical line is accuracy, and the right vertical line is a number of feature terms[6].

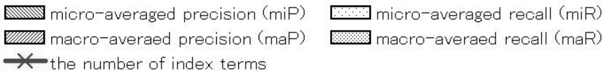
micro-averaged precision (miP)
micro-averaged recall (miR)
macro-averaed precision (maP)
macro-averaed recall (maR)
the number of index terms

Figure 2: Legend of Result Graphs

For example, Figure 5 shows that the highest maF is 92.3% in English, when we use the frequency ratio in document[7]. Let us see Figures 3-5 in a little more detail.

[2] α is an adhoc value in equation (12). We set $\alpha = 0.6$ in the present experiment.

[3] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[4] CD-Mainichi Newspapers 2002 data, Nichigai Associates, Inc., 2003 (Japanese)

[5] http://news.chinatimes.com/mainpage.htm (Chinese)

[6] The unit **k** in the right vertical line means 1000 terms.

[7] Refer to [1] (the table in p.38) for the comparison with other traditional techniques.

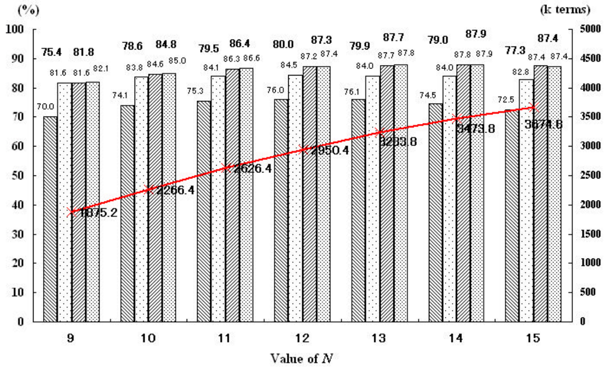

Figure 3: Results for English without filtering

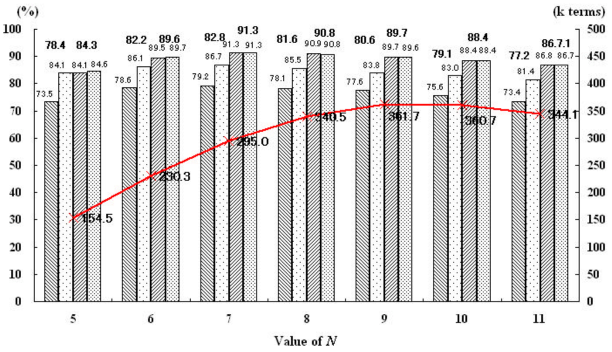

Figure 4: Results for English using mutual information

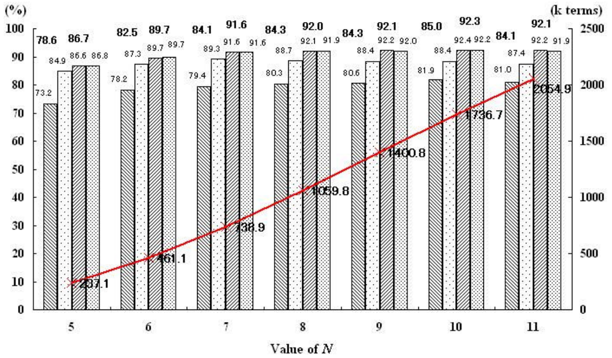

Figure 5: Results for English using frequency ratio

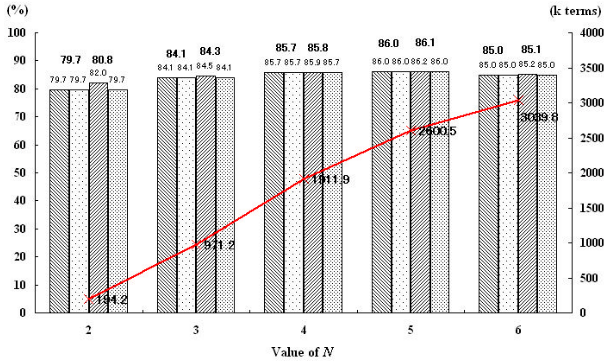

Figure 6: Results for Japanese without filtering

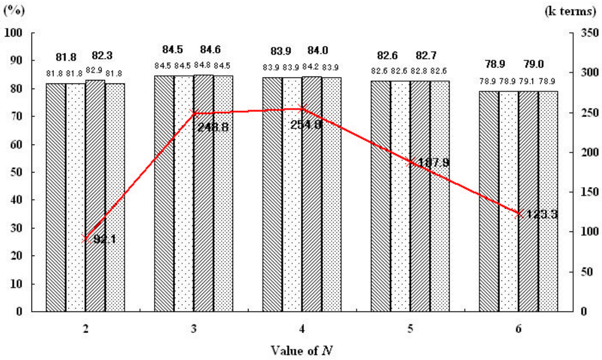

Figure 7: Results for Japanese using mutual information

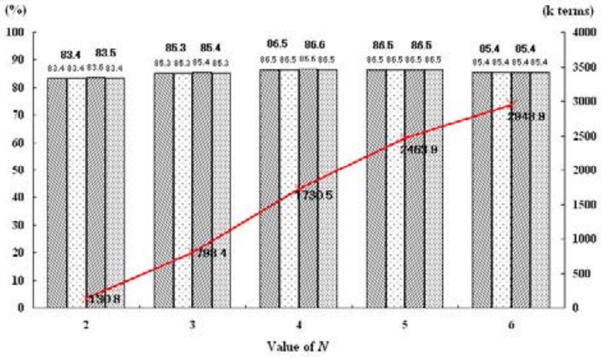

Figure 8: Results for Japanese using frequency ratio

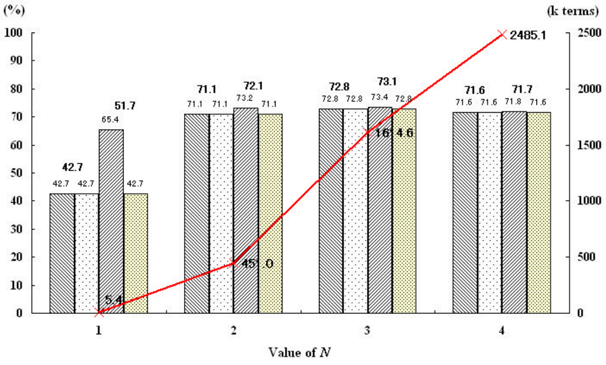

Figure 9: Results for Chinese without filtering

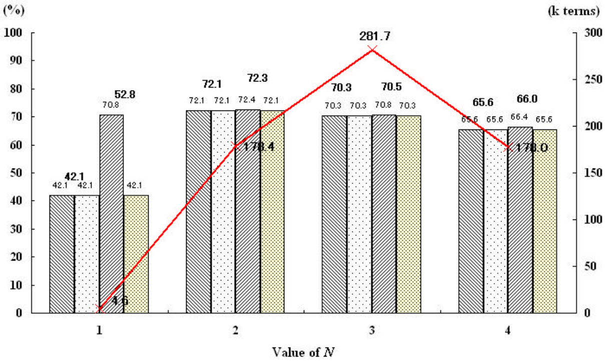

Figure10: Results for Chinese using mutual information

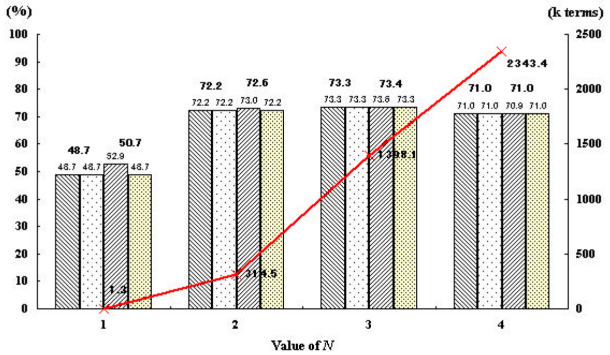

Figure11: Results for Chinese using frequency ratio

As shown in Figure 4, we were able to improve the highest maF up to 91.3% by refining the DB of feature terms using mutual information. On the other hand, Figure 3 shows that the highest maF in our previous paper is 87.9% in case of *N*=14.

Similarly, Figure 8 shows that the highest maF is 86.6% in case of N=4 for Japanese, and Figure 11 shows that the best value is 73.4% in case of N=3 for Chinese.

6. CONCLUSION

In the present paper, we refined the DB of feature terms using mutual information and frequency ratio in documents, and showed the effectiveness of the proposed method by several experiments. These experiments use English Reuters-21578 that is the benchmark newspaper articles, Japanese Mainichi Newspaper 2002, and Chinese China Times Newspaper 2005.

The best results were obtained when we refined DB using the frequency ratio in any case English, Japanese, and Chinese, and we showed the effectiveness of refining DB of feature terms. However, the results of Japanese and Chinese came short of our expectations though those of English were very excellent. As a matter of fact, we performed experiments with Word *N*-gram separately. As a result, it turned out that Word *N*-gram was more effective than Character *N*-gram in case of Japanese and Chinese. However, we did not refer to Word *N*-gram in the present paper, because the language-independence that is the theme of this paper is not filled so that using Word *N*-gram means using grammatical knowledge peculiar to the language. Moreover, our experiment clarified that the results of Chinese were 10% or more low compared with English and Japanese, even if we used Word *N*-gram. It is thought that this cause of the difficulty in Chinese is that the same Character *N*-gram appears in various categories because one Chinese character has various meanings in comparison with other languages. However, we have still not been able to clarify the detail. It is a future subject to clarify this cause.

REFERENCES

[1] F.Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol.34, pp.1-47, 2002

[2] W.Cavnar and J.Trenkle, *N*-Gram-Based Text Categorization, Proc. 3rd Annual Sympo. on Document Analysis and Information Retrieval (SDAIR), pp.161-169, 1994

[3] P.Nather, *N*-gram based Text Categorization, Diploma thesis, Comenius Univ., Faculty of Mathematics, Physics and Informatics, Institute of Informatics, 2005.

[4] Aizawa, The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures, Proc. 23th ACM International Conf. on Research and Development in Information Retrieval, pp.104-111, 2000

[5] E.D. Wiener, J.O. Pedersen, and A.S. Weigend, A neural network approach to topic spotting, Proc. 4th Sympo. on Document Analysis and Information Retrieval (SDAIR), pp.317-332, 1995

[6] *Apte, F.Damerau and S.M.Weiss, Automated Learning of Decision Rules for Text Categorization, ACM Trans. of Information Systems, Vol.12, No.3, pp.223-251, 1994*

[7] R. Rastogi and K. Shim, A decision tree classifier that integrates building and pruning, Proc. 24th International Conf. on Very Large Data Bases, pp.404-415, 1998

[8] D.D. Lewis and M. Ringuette, A comparison of two learning algorithms for text categorization, Proc. 3rd Annual Sympo. on Document Analysis and Information Retrieval (SDAIR), pp.81-93, 1994

[9] Y. Yang, An Evaluation of Statistical Approaches to Text Categorization, Journal of Information Retrieval, Vol.1, No.1, pp.67-88, 1999

[10] T. Joachims, Text categorization with support vector machines: learning with many relevant features, Proc. 10th European Conf. on Machine Learning, No.1398, pp.137-142, 1998

[11] R.E.Schapire and Y.Singer, BoosTexter - A Boosting-based System for Text Categorization, Machine Learning, Vol.39, No.2-3, pp.135-168, 2000

[12] M. Suzuki and S. Hirasawa, Text categorization based on the ratio of word frequency in each category, Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC), pp.3535-3540, 2007

[13] S.M.Namburu, H.Tu, J.Luo and K.R.Pattipati, Experiments on Supervised Learning Algorithms for Text Categorization, Proc. IEEE Aerospace Conf., Big Sky, MT, pp. 1-8,2005.

[14] K. Toutanova, F. Chen, K. Popat, and T. Hofmann, Text classification in a hierarchical mixture model for small training sets, Proc.

ACM Conf. on Information and Knowledge Management (CIKM), pp.105-113, 2001