Text Categorization Based on the Ratio of Word Frequency in Each Categories

Makoto SUZUKI, *Member, IEEE,* and Shigeichi HIRASAWA, *Fellow, IEEE*

Abstract— **In the present paper, we consider the automatic text categorization as a series of information processing and propose a new classification technique called the Frequency Ratio Accumulation Method (FRAM). This is a simple technique that calculates the sum of ratios of word frequency in each category. However, in FRAM, feature terms can be used without limit. Therefore, we propose the use of the character N-gram and the word N-gram as feature terms using the above-described property of FRAM. Next, we evaluate the proposed technique through a number of experiments. In these experiments, we classify newspaper articles from Japanese CD-Mainichi 2002 and English Reuters-21578 using the Naive Bayes method (baseline method) and the proposed method. As a result, we show that the classification accuracy of the proposed method is far better than that of the baseline method. Specifically, the classification accuracy of the proposed method is 87.3% for Japanese CD-Mainichi 2002 and 86.1% for English Reuters-21578. Thus, the proposed method has very high performance. Although the proposed method is a simple technique, it provides a new perspective and has a high potential and is language-independent. Thus, the proposed method can be expected to be developed further in the future.**

I. Introduction

With the spread of computers, the amount of accumulated electronic text is increasing rapidly. Recently, automatic text categorization has receiving a great deal of attention because it is becoming impossible to manually classify the enormous amount of text for the purpose of, for example, later category-based retrieval.

Text categorization is the problem of selecting 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 problems. The first is the extraction of feature terms that become effective keywords in the training phase, and the second is the actual classification of the document using these feature terms in the test phase. In the present paper, we refer to the former as a feature selection stage and the latter as a document classification stage.

One word is usually considered to be one feature term in the feature selection stage. In the language delimited in space, in English, for example, we need not extract words. However, for Japanese, we should extract words by morphological analysis. In contrast, a method to generate these feature terms with N-grams has been proposed as a language-independent technique [2]-[3]. In any case, many previously-proposed techniques extract useful feature terms from several words by using mutual information, TFIDF values, and so on [4]. These extracted feature terms are then used for classification.

On the other hand, the categorization in the document classification stage is a traditional problem of machine learning, and we often use machine learning algorithms, such as the neural network [5], the decision trees [6]-[7], Naive Bayes [8], the k-nearest neighbor [9], and support vector machines [10], as well as boosting algorithms [11].

Many of these previous researches tended to deal with the feature selection and the classification respectively as independent problems in automatic text categorization.

In this research, we consider the automatic text categorization as a series of information processing. In the present paper, we propose a new framework that classifies documents without extracting feature terms in the feature selection stage. The proposed procedure is as follows. First, we define the frequency ratio (FR) in each category for an individual feature term, and propose a classification technique using these summations. Second, we propose an extraction method of suitable feature terms for the proposed classification technique. Finally, we perform experiments using newspaper articles from Japanese CD-Mainichi 2002[1] and English Reuters-21578 [2]. We then show that the classification accuracy of the proposed method is better than that of the Naive Bayes method, which is one of the most famous techniques that are available at present.

II. Text Categorization

2.1 Overview

In the present paper, the goal of text categorization is to classify the given new documents into a fixed number of pre-defined categories. Fig.1 shows a flow diagram of the text categorization task [12].

The procedure for automatic text categorization is divided into two phases, the training phase and the test phase, as shown in Fig.1. In the training phase, we input the training documents along with a category. Next, we extract the feature term via a feature selection process and produce an indices database, referred to herein as DB, which is later used for the test phase. In the test phase, several new documents to be classified are input one after another, and one category is allocated in these documents with a classifier that uses the

Makoto Suzuki is with the Department of Information Science, Shonan Institute of Technology, 1-1-25 Tsujido Nishikaigan, Fujisawa-shi, Kanagawa, 251-8511, Japan (e-mail: m-suzuki@info.shonan-it.ac.jp).

Shigeichi Hirasawa is with the School of Science and Engineering, Waseda University, 3-4-1 Ohkubo, Shinjyuku-Ku, Tokyo, 169-8555, Japan(e-mail: hirasawa@hirasa.mgmt.waseda.ac.jp).

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

[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/ This provides benchmark data in automatic text categorization

Naive Bayes, the proposed method, and so on. Finally, we evaluate the classification results of each technique.

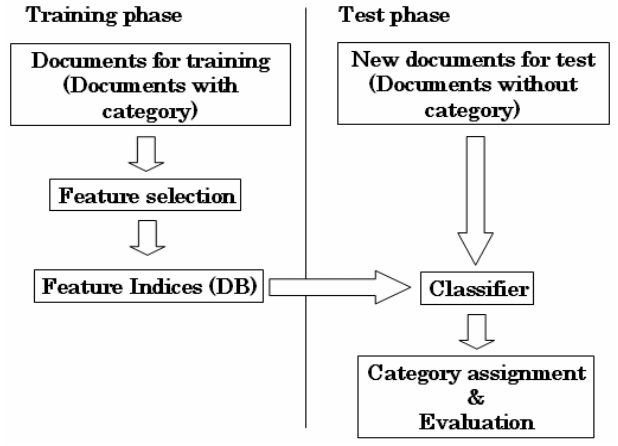

Fig.1. Flow diagram of text categorization

2.2 Mathematical Formulation

In the present paper, we use the following notation:

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.

Definition3: 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)

In addition, we write the set of categories to which each document belongs 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 notation 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.

III. Previous Method

3.1 Selection of Feature Terms

3.1.1 Feature Terms

Usually, text data stored in the DB of feature terms are composed of character strings of a suitable form for learning and classification. These character strings that function well in classification were extracted as feature terms in several previous studies. Moreover, documents are represented 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, $T \subseteq W$. That is, one feature term t_m corresponds to one word w_j. On the other hand, when the N-gram of each character (henceforth, referred to as the 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 (hereinafter referred to as the word N-gram) is extracted as a feature term t_m, one feature term corresponds to a block composed of N words, such as $\langle w_j, ..., w_{j+N-1} \rangle$.

3.1.2 Mutual Information

In many previous studies, the feature space was reduced by stemming, using the information gain, mutual information criteria, and so on, in the training phase. Here, we use the following mutual information as a selection criterion of feature terms. That is, we sequentially extract *M* terms with large mutual information as feature terms.

Definition 6: Mutual information

$$I(t_m; C) = \sum_{k=1}^{K} P(t_m, c_k) \log \frac{P(t_m, c_k)}{P(t_m) P(c_k)} \tag{6}$$

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

3.2 Naive Bayes

The method based on Naive Bayes is a probabilistic classifier using joint probabilities of words and categories to calculate the category of a given document. The systems based on Naive Bayes are probably the most frequently used systems in text classification. Therefore, the present paper considers the Naive Bayes as a method for comparison (baseline method).

First, we express each document as a vector of *M* dimensions using *M* selected feature terms.

Definition 7: Document Vector

$$\vec{d}_i = (t_1, t_2, ..., t_M) \tag{7}$$

The goal of using Naive Bayes is to classify a new document $\vec{d}_i$ into category c_k, the probability of which is highest when document $\vec{d}_i$ is given. Therefore, we can write the following. Here, Eq.8 uses Bayes' theorem.

$$\begin{aligned} c_{\hat{k}} &= \arg\max_{c_k \in C} P(c_k \mid \vec{d}_i) \\ &= \arg\max_{c_k \in C} \frac{P(\vec{d}_i \mid c_k) P(c_k)}{P(\vec{d}_i)} \\ &= \arg\max_{c_k \in C} P(\vec{d}_i \mid c_k) P(c_k) \end{aligned} \tag{8}$$

In general, it is difficult to calculate the conditional probability $P(\vec{d}_i \mid c_k)$ in Eq.8. Then, in the Naive Bayes based classification, we assume that, given a category, each word w_{il} occurs independently in document $\vec{d}_i$. We are able to calculate the probability efficiently using this assumption.

Assumption 1: Conditional Independence

$$\begin{aligned} P(\vec{d}_i \mid c_k) &= P(t_1, ..., t_M \mid c_k) \\ &= \prod_{m=1}^{M} P(t_m \mid c_k) \end{aligned} \tag{9}$$

$\vec{d}_i$: a document
c_k : a category
t_m : a feature term

The above conditional probability becomes 0 when no feature term appears, because the probability has the sparseness problem. Then, we calculate this conditional probability using the Laplace method, which is a well known smoothing method in the present paper.

3.3 Problems with the previous method

The method based on Naive Bayes requires unrealistic assumptions such as Assumption 1. Moreover, Naive Bayes decreases the classification accuracy when it maintains too many feature terms with low frequency (henceforth, referred to as low-frequency terms). Therefore, the number of feature terms is usually limited[3]. Thus, it is thought that there are limitations of Naive Bayes with respect to classification accuracy.

IV. Proposed Method

In the present study, we assume a close relationship between the selection method of feature terms and the classification method using these terms in automatic text categorization. Therefore, instead of discussing the selection and the classification separately, we need to think of them as steps in a single information processing flow while considering the compatibility between them. In the present paper, we consider the extraction of the feature terms and the classification using these terms and propose a new text categorization technique as mentioned above.

4.1 Classification method using the sum of frequency ratios

Here, we propose a classification method using the sum of frequency ratios in each category of an individual feature term. We call the Frequency Ratio Accumulation Method (FRAM). First, we will define the frequency ratio as follows.

Definition 8: Frequency Ratio

$$FR(t_m, c_k) = \frac{R(t_m, c_k)}{\sum_{c_k \in C} R(t_m, c_k)} \tag{10}$$

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 evaluation values, which indicate the possibility that the target document belongs to the category, are calculated as follows.

[3] Refer to Fig.3 and Fig.5 of the following experiment results for details.

Definition 9: Category Score

$$E_{d_i}(c_k) = \sum_{t_m \in d_i} FR(t_m, c_k) \quad (11)$$

Finally, the target document d_i is classified into the category $c_{\hat{k}}$ for which the category score is the maximum, as in Eq.12.

$$c_{\hat{k}} = \arg\max_{c_k \in C} E_{d_i}(c_k) \quad (12)$$

Namely, 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 the frequency ratio 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.

4.2 Extraction of feature terms using the N-gram

Here, we propose three types of feature selection methods, including the words extracted by morphological analysis, the character N-gram, and the word N-gram, as suitable feature terms for the classification method proposed above. In particular, the N-gram is effective as a language-independent method because it does not depend on the meaning of the language. Moreover, the word N-gram is the N-gram of each word after morphological analysis in Japanese. An example for the Japanese language is shown in Table I and an example for the English language is shown in Table II.

Table I. Example of Japanese feature terms

Original sentence	**これは論文です.**
Word	**これ/は/論文/です**
Character N-gram (N=2)	**これ/れは/は論/論文/文で/です**
Word N-gram (N=2)	**これは/は論文/論文です**

Table II. Example of English feature terms

Original sentence	**This is a paper.**
Word	**This/is/a/paper**
Character N-gram (N=2)	**Th/hi/is/s / i/is/s / a/a / p/pa/ap/pe/er**
Word N-gram (N=2)	**Thisis/isa/apaper**

V. EXPERIMENT

5.1 Experimental conditions

The present experiment involved two newspapers that contain articles with pre-assigned categories. The first is the Japanese CD-Mainichi Newspaper 2002 (Mainichi), and the second is the English Reuters-21578 (Reuters).

We classified the two types of newspaper articles mentioned above using six methods shown in Table III.

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, we gave the test data to the computer without showing them their true categories, and made the computer classify them.

Moreover, we performed several experiments by limiting the number of feature terms in order to confirm the differences by the number of feature terms in experiment 0. We report the results of these experiments. On the other hand, in experiments 1 through 5, we did not limit the number of feature terms and used all of the feature terms to achieve the best performance by FRAM.

Table III. Each method in the present experiment

No	Form of feature terms	Categorization method	Notation
0	Word	Naïve Bayes	Baseline
1	Word	FRAM	Prop.1
2	Character 2-gram	FRAM	Prop.2
3	Character 3-gram	FRAM	Prop.3
4	Character 4-gram	FRAM	Prop.4
5	Word 2-gram	FRAM	Prop.5

5.1.1 CD-Mainichi Newspaper 2002

For each of the seven categories, such as Economy, International, Home, Culture, Entertainments, Sports, and Leader, included in the CD-Mainichi, we randomly selected 1,000 training documents and 500 test documents (7,000 and 3,500 documents in total, respectively). Moreover, we performed morphological analysis when we perform extraction from articles. Here, we used MeCab [4] as a morphological analysis tool[5].

5.1.2 Reuters-21578

In addition, we used Apte split 10 categories of 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.

5.2 Measuring classification performance

In the present paper, we perform simple evaluation with only *recall* and refer to the evaluation results as classification accuracy, although general criteria include *recall*, *precision*, and *F-measure* [13].

Here, accuracy is defined as the number of correctly categorized documents divided by the total number of categorized documents.

[4] http://mecab.sourceforge.net/
[5] We did not remove affixes.

5.3 Results

These results are shown in Fig.2 through Fig.5. Fig.2 shows that the classification accuracy for Mainichi by Prop.1 (word × FRAM) is 83.7% and that by the baseline method (word × Naive Bayes) is 76.7%. That is, a difference of 7% was observed between the performance of the baseline method and that by Prop.1. This difference is a result of the use of FRAM because the same morphological analysis is performed in both the baseline method and Prop.1. Here, for the baseline method, 76.7% is the highest accuracy, as shown in Fig.3. That is, Fig.3 shows the accuracy for each case in which the baseline classifies Mainichi by limiting the number of feature terms using mutual information. Similarly, the accuracy of the baseline is 75.6%, while that of Prop.1 is 86.1% for Reuters. That is, a difference of 10.5% was observed between the performance by the baseline and that by Prop.1. Thus, we were able to confirm the effect of Prop.1 to be greater.

On the other hand, Prop.2 – Prop.5 are methods that use FRAM as a classifier and use the character N-gram in Prop.2 – Prop.4 and the word N-gram in Prop.5 as feature terms. For both Mainichi and Reuters, the classification accuracy of Prop.5 (word 2-gram×FRAM) was the highest.

Moreover, for Mainichi, Fig.2 shows that the highest classification accuracy of Prop.5 is 87.3% and the improvement compared to Prop.1 is 3.6%. This difference is an effect of using word N-grams as feature terms in the proposed method. The use of character N-grams as feature terms had some effects for the cases of N = 3 and 4, whereas the value of N was too small and was ineffectual in the case of N = 2.

On the other hand, the highest classification accuracy for Reuters is 86.1%, which was improved by only 0.1%, as compared with 86.0% of Prop.1. The classification accuracy of character N-grams is lower than Prop.1, although we only have results until N=4. Thus, the effect of feature terms based on N-grams is not expected to be large for English.

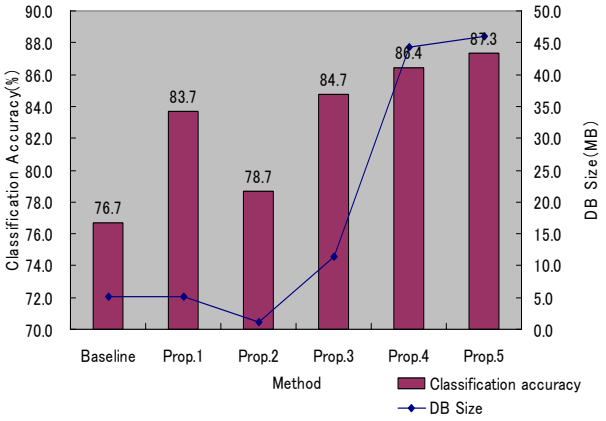

Fig.2. Results of Proposal in the case of Mainichi

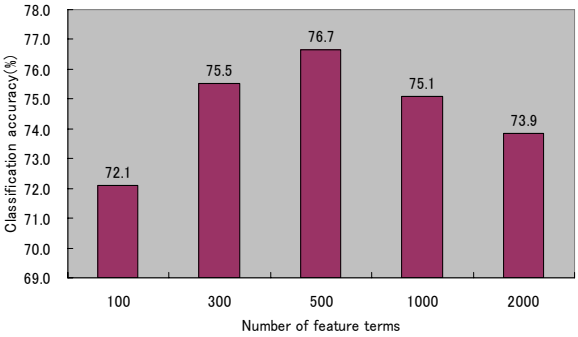

Fig.3. Results of Baseline in the case of Mainichi

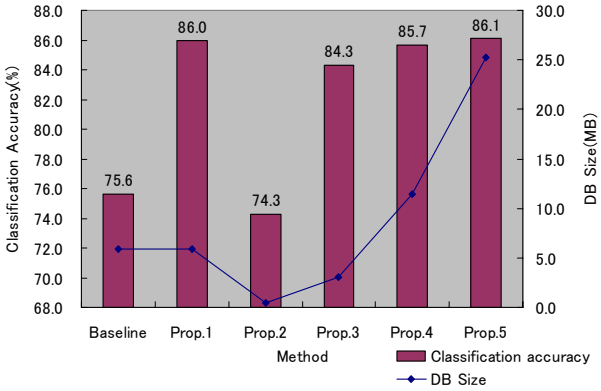

Fig.4: Results of Proposal in the case of Reuters

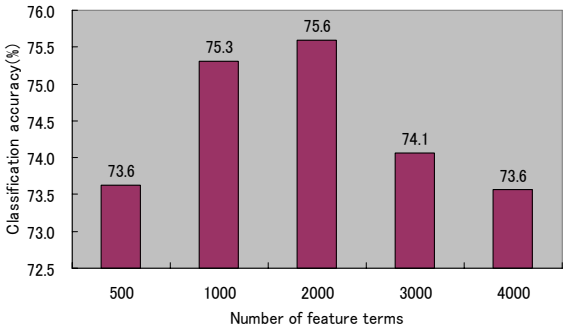

Fig.5: Results of Baseline in the case of Reuters

VI. EVALUATION AND DISCUSSION

When we deal with the mass of texts, we are interested in the memory capacity of data and the computational complexity. Therefore, several studies have examined the feature selection and the classification in automatic text categorization.

In general, feature space reduction in feature selection improves the performance of the learning algorithm, decreases the data size, controls the classification time, and avoids over-fitting of the data.

However, as shown in Fig.3 and Fig.5, the Naive Bayes decreases the classification accuracy if the number of feature terms is increased too much. This reason is the great influence of low-frequency terms. For example, if, by chance, an extremely-low-frequency term for a certain specific category is included in the target document, the possibility exists that the document will be classified into that category becomes very low. Therefore, the situation mentioned above happens frequently if the ratio of low-frequent terms in the set of feature terms increases, and the classification accuracy of Naive Bayes decreases.

On the other hand, the proposed method used words, character N-grams, and word N-grams as feature terms, and all feature terms were adopted without limiting their numbers. Therefore, the DB of the feature terms generated by the proposed method contains useless feature terms. For example, in the case of word N-gram, the feature term selected depends on whether “is” or “was” appears after “This”. The former feature term becomes "This is", and the latter feature term becomes "This was". That is to say, each feature term is distinguished completely. Thus, the proposed method can reduce the computational complexity, although the set of feature terms is redundant because it uses a simple classification technique of calculating the sum of the frequency ratio in the classification process and improves the classification accuracy.

Next, we consider the relationship between the size of the feature terms and the classification accuracy in the proposed method. As shown in Fig.2 and Fig.4, we can confirm that the classification accuracy has increased in proportion to the logarithmic value of the size of feature terms. However, in the case of the character N-gram, it is not expected that the classification accuracy will increase remarkably as the size of the feature terms increases exponentially with N. In this sense, it is necessary to select an appropriate value of N.

VII. Conclusion

A number of previous studies examined feature selection and classification, respectively, as independent problems in automatic text categorization. We also searched an approach to extract unnecessary information and maintain only necessary information as feature terms compactly in feature selection from the viewpoint of the information theory.

However, we proposed a new classification technique called FRAM from a different angle in the present paper, whereby feature terms can be used unlimitedly, although this is a simple classification method of summing frequency ratios. We then proposed the use of the character N-gram and the word N-gram as feature terms in feature selection and maintained terms that are considered to be redundant in the set of feature terms. Thus, we consider automatic text categorization as a series of information processes and propose a technique that combines a selection method of feature terms and a simple classification method.

As a result, we show that the classification accuracy (*recall*) of the proposed method improves greatly compared with the baseline. That is, the classification accuracy is 87.3% for Mainichi and 86.1% for Reuters. Thus, the proposed method has a very high performance [1]. Moreover, the proposed method based on N-grams is basically language-independent, although the word N-gram requires morphological analysis. Consequently, the proposed method provides a new perspective and has a high potential. Thus, it can be expected to be developed further in the future.

Acknowledgment

We would like to express my sincere gratitude and appreciation to Dr. Tetsuya Sakai with the NewsWatch, Inc. and Toru Nakata with the eSOL Co.,Ltd, for their cooperation and help.

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] A. 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] C.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] 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.

[13] 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.