2011 International Conference on Document Analysis and Recognition

A Study on Automatic Chinese Text Classification

Xi Luo, Wataru Ohyama, Tetsushi Wakabayashi, Fumitaka Kimura
Graduate School of Engineering, Mie University
Tsu-shi, Mie, Japan
e-mail: luoxi@hi.info.mie-u.ac.jp

Abstract—**In this paper, we perform Chinese text classification using *N*-gram (uni-gram, bi-gram and mixed uni-gram/bi-gram) frequency feature instead of word frequency feature to represent documents and propose the use of mixed uni-gram/bi-gram after feature transformation. We further propose a serial approach based on feature transformation and dimension reduction techniques to improve the performance. Experimental results show that our proposed approach is efficient and effective for improving the performance of Chinese text classification. Furthermore, we present several experiments evaluating the selection of features based on part-of-speech analysis and the results show that suitable combination of part-of-speech can lead to better classification performance.**

Keywords-Chinese text classification/categorization; N-gram; part-of-speech; feature selection; dimension reduction; principal component analysis; support vector machines

I. INTRODUCTION

Automatic text classification (ATC) is the task to automatically assign one or more appropriate categories for a document according to its content or topic [1]. Traditionally, text classification is carried out by human experts as it requires a certain level of vocabulary recognition and knowledge processing. With the rapid explosion of texts in digital form and growth of online information, text classification has become an important research area owing to the need to automatically handle and organize text collections.

Many standard machine learning techniques have been applied to automated text classification problems, and *K* Nearest Neighbor system (kNN) and Support Vector Machines (SVM) have been reported as the top performing methods for English text classification [2]. Unfortunately, perfect precision cannot be reached in Chinese text classification and the inherent errors caused by word segmentation always remain as a problem.

In this paper, we perform Chinese text classification using *N*-gram frequency feature instead of word frequency feature to represent documents on TanCorpV1.0 [3] which is a new large corpus special for Chinese text classification. We explain the impact of the different assumptions on *N*-gram (uni-gram and bi-gram) frequency feature and propose to use the combination of different *N*-gram (mixed uni-gram/bi-gram) to battle with artificial assumption. We further propose to conduct the combination of uni-gram and bi-gram after feature transformation (1+2-gram-after FT). Experimental results show that our proposed approach (1+2-gram-after FT) can best represent Chinese documents.

The limitation of using absolute frequency as the feature vector is dependency on text length which usually leads into lower performance. We experimentally evaluate the effectiveness of proposed approach based on feature transformation techniques including normalizing absolute frequency to relative frequency and power transformation. The results show a significant improvement in performance.

N-gram extraction on a large corpus will yield a large number of possible *N*-grams. High dimensionality of the feature space may be problematic in terms of computational time and storage resources. Experiments prove that Principal Component Analysis (PCA) is an efficient and effective way to reduce the dimensionality.

Furthermore, we present several experiments evaluating the selection of features based on part-of-speech analysis. We explored the effects of different part-of-speech (nouns, verbs adjectives, adverbs and pronouns) on Chinese text classification effectiveness. The results show that nouns are the most important features to represent Chinese documents and suitable combination of part-of-speech can lead to better classification performance.

The rest of this paper is organized as follows: In Section II we describe the proposed method. Section III describes part-of-speech analysis. Experiments and results are presented in Section IV and Section V. We summarize our research and point out some future direction in Section VI.

II. THE PROPOSED METHOD

Fig.1 shows the general steps for Chinese text classification based on the proposed approach. In the following subsections, we introduce our approach in more detail.

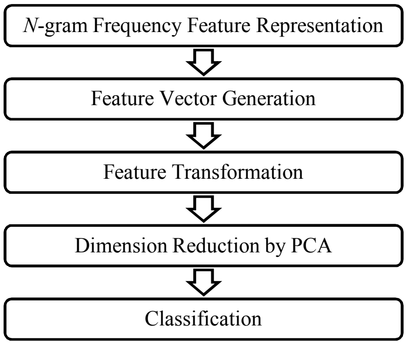
N-gram Frequency Feature Representation
Feature Vector Generation
Feature Transformation
Dimension Reduction by PCA
Classification

Figure 1. General steps for Chinese text classification based on the proposed approach.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.187

A. *N-gram Frequency Feature Representation*

Unlike English and other western languages, there is no natural delimiter between Chinese words and even no uniform smallest semantic units. This means that the word segmentation is necessary before any other preprocessing and the use of a dictionary is required.

In this paper, we use a method independent of languages which represents documents with character *N*-grams [4]. A character *N*-gram is a sequence of *N* consecutive characters. Sequences of one character (*N*=1) are called uni-gram (1-gram). Sequences of two characters (*N*=2) are called bi-gram (2-gram). Table I shows examples of *N*-gram sequences.

TABLE I. EXAMPLES OF *N*-GRAM SEQUENCES

Original Text	明天我们去北京
1-gram	明；天；我；们；去；北；京
2-gram	明天；天我；我们；们去；去北；北京

When we establish *N*-gram frequency feature, we artificially introduce an assumption over the relationship among adjacent words. Uni-gram is based upon the assumption that all words appear in the corpus independently. Bi-gram assumes that only contiguous words correlate with each other. In this sense, single *N*-gram frequency feature could model the language phenomena with some compromise. To make full use of the power of different *N*-gram frequency features, we propose to combine uni-gram and bi-gram to represent documents which called mixed uni-gram/bi-gram (1+2-gram).

Obviously, one way is to combine uni-gram and bi-gram frequency feature before any other processing. Since uni-gram and bi-gram are two relatively independent methods, we further propose to conduct the combination after feature transformation (normalization to relative frequency and power transformation). Fig.2 and Fig.3 show the two different combinations of mixed uni-gram/bi-gram.

In the experiments, the following four methods were used to compare *N*-gram frequency feature representation.

Method 1 (1-gram): Use uni-gram frequency feature.

Method 2 (2-gram): Use bi-gram frequency feature.

Method 3 (1+2-gram-before FT): Use mixed uni-gram/ bi-gram frequency feature and the combination of uni-gram and bi-gram was performed before feature transformation.

Method 4 (1+2-gram-after FT): Same as Method 3 except the combination of uni-gram and bi-gram was performed after feature transformation.

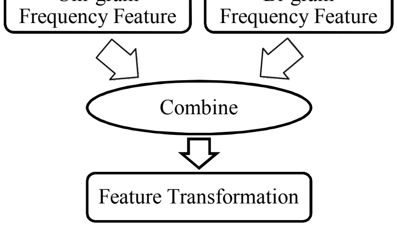

Figure 2. Combination of uni-gram and bi-gram frequency feature before feature transformation (1+2-gram-before FT).

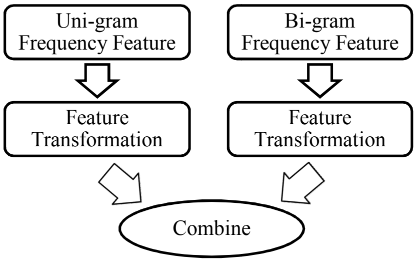

Figure 3. Combination of uni-gram and bi-gram frequency feature after feature transformation (1+2-gram-after FT).

B. *Feature Vector Generation*

In order for a machine learning system to recognize a document there should be a way of representing it. This is usually done by the use of feature vectors. First a lexicon including all different features (*N*-grams) in training data was generated. Then the feature vector represents the frequency of a specific feature in the document. The form of the feature vector *X* can be denoted as:

$$X = [\, x_1 \;\; x_2 \;\; \cdots \;\; x_n \,]^T \tag{1}$$

where *n* is the dimensionality of the feature vector (lexicon size), x_i is the frequency value of i^{th} feature and *T* refers to the transpose of a vector.

Assume that the following two documents of uni-gram sequences represent a text collection:

1. 我；是；一；个；中；国；人

2. 中；国；是；发；展；中；国；家

Figure 4. A text collection composed of two documents.

The lexicon including all different features (*N*-grams) was generated as:

{我 是 一 个 中 国 人 发 展 家}

Figure 5. Lexicon.

Fig.6 shows the feature vector (absolute frequency) obtained for each document from the lexicon.

1. $[1 \;\; 1 \;\; 1 \;\; 1 \;\; 1 \;\; 1 \;\; 1 \;\; 0 \;\; 0 \;\; 0]^T$

2. $[0 \;\; 1 \;\; 0 \;\; 0 \;\; 2 \;\; 2 \;\; 0 \;\; 1 \;\; 1 \;\; 1]^T$

Figure 6. Absolute frequency (AF).

C. *Feature Transformation Techniques*

1) Normalization to Relative Frequency: The feature vector generated by above process is composed of the absolute frequency. In practice, textual data vary in content and length. The limitation of the absolute frequency is dependency on text length which usually leads into lower performance. This is because text length may differ within the same class of documents consequently more complexity of learning. In order to normalize the lengths of documents, absolute frequency is transformed to relative frequency:

$$y_i = \frac{x_i}{\sum_{j=1}^{n} x_j} \tag{2}$$

where x_i is the absolute frequency of feature i and n is the lexicon size. Fig.7 shows the results after transformed to relative frequency.

$$1.\ \begin{bmatrix} \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & 0 & 0 & 0 \end{bmatrix}^T$$

$$2.\ \begin{bmatrix} 0 & \frac{1}{8} & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{bmatrix}^T$$

Figure 7. Relative frequency (RF).

2) Power Transformation: The distribution of absolute/ relative frequency are generally skewed. Therefore in our approach power transformation [5] is applied to improve the symmetry of the distribution:

$$z_i = x_i^v\ (0 < v < 1) \tag{3}$$

This transformation generates Gaussian-like sample distribution. In the experiments, v is set to 0.5. Fig.8 and Fig.9 show the results when power transformation was applied to absolute frequency and relative frequency respectively.

$$1.\ [\ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 0\ \ 0\ \ 0\ ]^T$$

$$2.\ \begin{bmatrix} 0 & 1 & 0 & 0 & \sqrt{2} & \sqrt{2} & 0 & 1 & 1 & 1 \end{bmatrix}^T$$

Figure 8. Absolute frequency with power transformation (AFPT).

$$1.\ \begin{bmatrix} \frac{1}{\sqrt{7}} & \frac{1}{\sqrt{7}} & \frac{1}{\sqrt{7}} & \frac{1}{\sqrt{7}} & \frac{1}{\sqrt{7}} & \frac{1}{\sqrt{7}} & \frac{1}{\sqrt{7}} & 0 & 0 & 0 \end{bmatrix}^T$$

$$2.\ \begin{bmatrix} 0 & \frac{1}{\sqrt{8}} & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} \end{bmatrix}^T$$

Figure 9. Relative frequency with power transformation (RFPT).

D. Dimension Reduction

In ATC, high dimensionality of the feature space may be problematic in terms of computational time and storage resources. In order to solve this problem, the dimensionality is required to be reduced without deterioration of the performance.

1) Dimension Reduction by Feature Selection: N-gram extraction on a large corpus will yield a large number of possible *N*-grams. In fact, only some of them will have significant frequency values in vectors representing the documents and good discriminating power. Yang and Pedersen [6] have shown that it is possible to reduce the dimensionality by a factor of 10 with no loss in effectiveness and a reduction by a factor of 100 bringing just a small loss. Hence in the experiments, features with frequency value of 10 or less in all training data were removed to reduce the high dimensionality.

2) Dimension Reduction by PCA: Then Principal Component Analysis (PCA) was applied to further reduce the high dimensionality. PCA involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

E. Classification

Support Vector Machines (SVM) is a relatively new class of machine learning techniques first introduced by Vapnik [7]. Based on the structural risk minimization principle from the computational learning theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set. In the experiments, we used *SVM^{light}* package [8]. We adopted three different types of SVM kernel functions: Linear Kernel (Linear), Polynomial Kernel (Poly) and Radial Basis Function (RBF).

III. Part-of-Speech Analysis

In this section, we focus on the feature selection process and aim to explore the effects of different part-of-speech (nouns, verbs adjectives, adverbs and pronouns) on Chinese text classification effectiveness.

A. Part-of-Speech Tagging

In order to conduct part-of-speech (POS) analysis, lexical analysis system is required. We used a software called 3GWS (the 3rd Generation Word Segmenter) to perform word segmentation and POS tagging. From all possible tags, we just considered five important POS tags: nouns, verbs, adjectives, adverbs and pronouns.

B. Excluding One Part-of-Speech

Firstly we found out which POS contribute more in classification performance. In this step we removed one group of POS and then evaluate the classification performance to find out the hierarchy of the POS in describing a category's content.

For simplicity let us assume that there are two sets of feature vectors: the feature set generated using only nouns and the feature set generated using only verbs. We denote nouns with a superscript u and verbs with a superscript v. Consequently, we can define the feature vectors as:

$$\boldsymbol{x}^{(u)} = \begin{bmatrix} x_1^{(u)} & x_2^{(u)} & \cdots & x_{n_1}^{(u)} \end{bmatrix}^T \tag{4}$$

for the noun features. The verb features can be expressed as:

$$\boldsymbol{x}^{(v)} = \begin{bmatrix} x_1^{(v)} & x_2^{(v)} & \cdots & x_{n_2}^{(v)} \end{bmatrix}^T \tag{5}$$

Let us denote the original feature set as $\boldsymbol{A}$ and the remaining feature vectors after excluding nouns can be defined as:

$$\boldsymbol{R} = \boldsymbol{A} \ominus \boldsymbol{x}^{(u)} \tag{6}$$

Equation (6) can be used in excluding other POS such as verbs, adjectives, adverbs and pronouns.

C. Combination of Suitable Part-of-Speech

Secondly we found out suitable combination of POS which can lead to better classification performance. The combination of noun and verb feature vectors can be defined as:

$$\boldsymbol{Q} = \boldsymbol{x}^{(u)} \oplus \boldsymbol{x}^{(v)} \quad (7)$$

$$= \left[x_1^{(u)} \ \cdots \ x_{n_1}^{(u)},\ x_1^{(v)} \ \cdots \ x_{n_1}^{(v)} \right]^T \quad (8)$$

Equation (8) can be also used to combine other POS such as adjectives, adverbs and pronouns.

IV. Experiments

A. Data for Experiments

Experimental data were obtained from a Chinese corpus called TanCorpV1.0 [3] which is collected and processed by Songbo Tan. The corpus is categorized in two hierarchies. The first hierarchy contains 12 big categories and the second hierarchy consists of 60 subclasses. It is totally composed of 14,150 texts. This corpus can serve as three categorization datasets: one hierarchical dataset (TanCorpHier) and two flat dataset (TanCorp-12 and TanCorp-60). In our experiments, we use TanCorpHier.

In the experiments, 150 texts were selected randomly from the corpus for each big category, and totally 1800 texts were used. The ratio of training data to test data is set as 2:1.

B. Evaluation

We adopt the most commonly used *F*-measure (*F*) metric introduced by Van Rijsbergen [9], which is the weighted harmonic mean of precision (*P*) and recall (*R*).

For ease of comparison, we summarize the *F*-measure over the different categories using the Micro-averaged *F*-measure which is viewed as a per-document average since it gives equal weight to every document. It is defined as:

$$F(\text{micro-averaged}) = \frac{2RP}{R+P} \quad (9)$$

In micro-averaging, precision and recall are obtained by summing over all individual decisions:

$$P = \frac{TP}{TP{+}FP} = \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M}(TP_i{+}FP_i)} \quad (10)$$

$$R = \frac{TP}{TP{+}FN} = \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M}(TP_i{+}FN_i)} \quad (11)$$

where *M* is the number of categories. *TP*, *FN* and *FP* are the number of true positives, false negatives and false positives, respectively.

V. Results

A. Performance Evalution of the Proposed Method

1) The Effect of Feature Transformation and Principal Component Analysis Techniques.

Fig.10 shows the relationship between the dimensionality and the Micro-averaged *F*-measure for bi-gram with linear kernel. The performance was significantly improved by employing relative frequency instead of absolute frequency. The best Micro-averaged *F*-measure was improved from 79.87% to 85.05%. Power transformation technique further improved the performance, from 79.87% to 83.96% for absolute frequency and from 85.05% to 86.62% for relative frequency. Relative frequency with power transformation (RFPT) gives the best performances throughout all dimensionality.

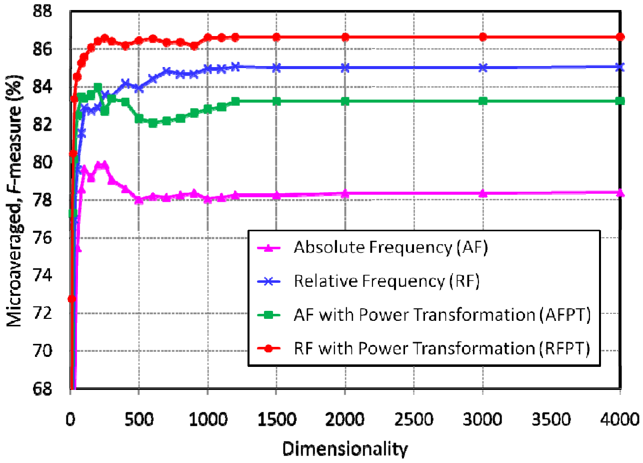

Figure 10. Micro-averaged *F*-measure vs. Dimensionality for bi-gram with linear kernel.

For both AF and AFPT, the best performance was achieved at lower dimensionality (200 dimensionality) and after that the performance decreased slightly. For RF and RFPT, the performance increased before 1000 dimensionality and then became stable. For all kinds of feature vectors, there is nearly no performance improvement after 1100 dimensionality. Principal Component Analysis (PCA) improves the efficiency of the text classification by reducing the dimensionality.

2) Comparing N-gram Frequency Feature Representation

Fig.11 shows the best performance comparison of RFPT in Micro-averaged *F*-measure. The result shows that throughout all dimensionality, 1-gram has the worst results. It also indicates that 1+2-gram-after FT produce the highest effectiveness.

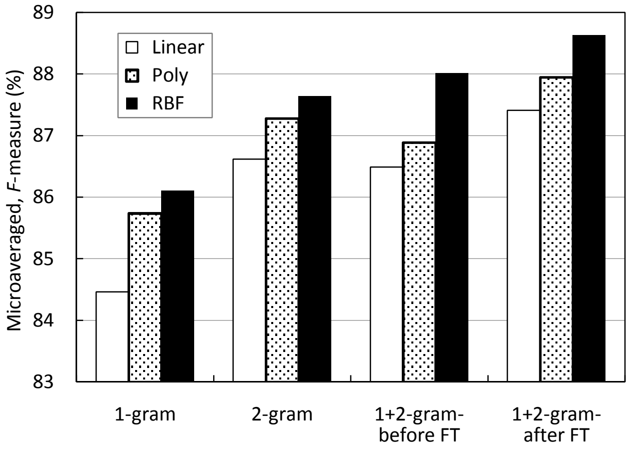

Figure 11. The best Micro-averaged *F*-measure of RFPT.

B. *The Results of Part-of-Speech Analysis*

1) *The Results of Excluding One Part-of-Speech*

Fig.12 shows the best Micro-averaged *F*-measure comparison of three types of SVM kernels of RFPT after excluding one POS.

We can see that when nouns were removed, the results were lower than 79% which achieved the lowest classification performance compared with other POS. This means that nouns are the most important features to represent Chinese documents and contribute most to higher performance. When verbs were removed, the classification performance was only decreased from 87.19% to 85.69% for RBF kernel which means that verbs contribute less in representing Chinese documents. When adjectives were removed, there is no performance decrease but rather an improvement. In other words, adjectives decrease the ability of features to discriminate one category from another. The same case was observed when adverbs or pronouns were removed in the text classification task.

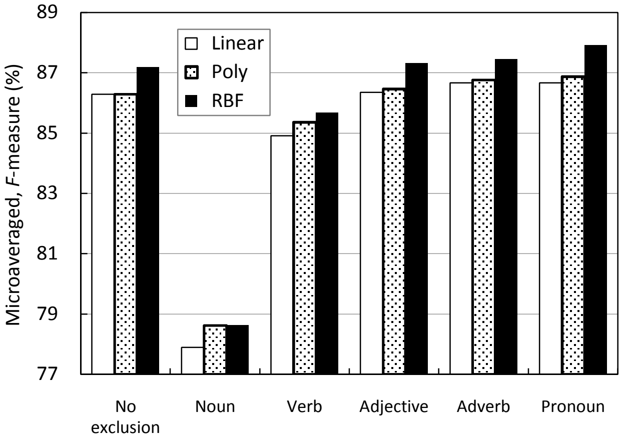

Figure 12. The best Micro-averaged *F*-measure after excluding one POS.

2) *The Results of Suitable Combination of Part-of-Speech*

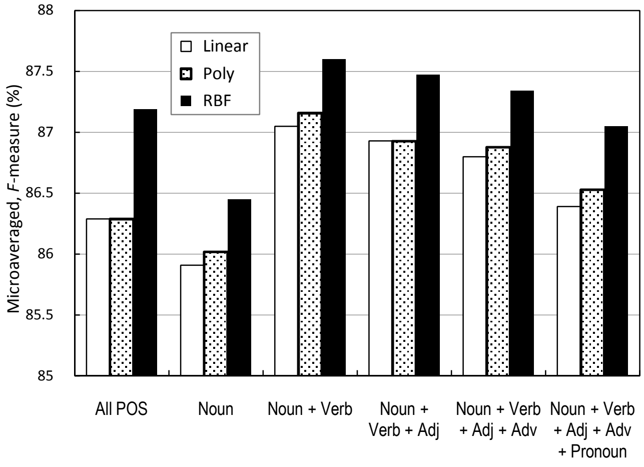

Figure 13. The best Micro-averaged *F*-measure for different combinations of POS.

Fig.13 shows the best Micro-averaged *F*-measure comparison of three types of SVM kernels of RFPT for different combinations of POS. The performance of using only nouns as features is 86.45% for RBF kernel which is only slightly lower than the use of all POS. Then nouns were combined with verbs. The performance was improved from 86.45% to 87.6% which is the best performance as compared to all the other combinations of POS. After combined adjectives with nouns and verbs, the performance was slightly decreased to 87.47%. When adverbs and pronouns were added, the performances were further reduced to 87.34% and 87.05%.

VI. CONCLUSION

In this paper, we perform Chinese text classification using *N*-gram frequency feature representation. We propose to use the combination of uni-gram and bi-gram after feature transformation which proved to be the most efficient method to represent Chinese documents. We further propose a serial approach based on feature transformation and dimension reduction techniques to improve the performance of Chinese text classification. The experimental results show that normalizing absolute frequency to relative frequency followed by power transformation significantly improved the performance. Principal Component Analysis (PCA) effectively reduced the dimensionality without deterioration of the performance. Furthermore, we have explored the roles of the different POS in feature selection and we found that nouns are the most important features to represent Chinese documents and suitable combination of part-of-speech can lead to better classification performance.

Future work includes:

1. Extensive experimental evaluation using more texts on more categories.

2. Text classification on error prone Chinese OCR texts.

REFERENCES

[1] F. Sebastiani, Machine learning in automated text categorization, ACMComputing Surveys, Vol. 34, No. 1, (March 2002), 1-47.

[2] Y. Yang and X. Liu, A Re-examination of text categorization methods. In Proceedings, 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 42-49, 1999.

[3] Songbo Tan and Yuefen Wang, Chinese text categorization corpus-TanCorpV1.0. http://www.searchforum.org.cn/tansongbo/corpus.htm

[4] D. Jurafsky & J.H. Martin, An Introduction to natural language processing, computational linguistics, and speech recognition, Speech and Language Processing, Prentice Hall, 2000.

[5] K. Fukunaga, Introduction to statistical pattern recognition, Academic Press, Inc, (1990), 76-77.

[6] Y. Yang and J. O. Pedersen, A Comparative study on feature selection in text categorization, In Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, TN, 1997), 412-420.

[7] Corinna Cortes and V. Vapnik, Support-Vector Networks, Machine Learning, 20, 1995.

[8] T. Joachims, Learning to classify text using support vector machines: Methods, Theory and Algorithms, Kluwer Academic Publishers Boston Dordrecht London, 2001.

[9] C. van Rijsbergen, Information Retrieval, Butterworths, London, 1979.