Journal of Korea Multimedia Society Vol. 11, No. 6, JUNE 2008(pp. 875-882)

Table based Matching Algorithm for Soft Categorization of News Articles in Reuter 21578

Taeho Jo[†]

ABSTRACT

This research proposes an alternative approach to machine learning based ones for text categorization. For using machine learning based approaches for any task of text mining, documents should be encoded into numerical vectors; it causes two problems: huge dimensionality and sparse distribution. Although there are various tasks of text mining such as text categorization, text clustering, and text summarization, the scope of this research is restricted to text categorization. The idea of this research is to avoid the two problems by encoding a document or documents into a table, instead of numerical vectors. Therefore, the goal of this research is to improve the performance of text categorization by proposing approaches, which are free from the two problems.

Key words: Text Categorization, Table Matching Algorithm

1. INTRODUCTION

Text categorization is the process of assigning one or some among predefined categories to each document. The task belongs to pattern classification where texts or documents are given as patterns. Note that almost information in any system is given as textual formats dominantly over numerical one. For managing efficiently the kind of information given as the textual format, techniques of text categorization are necessary; 'text categorization' became a very interesting research topic in both academic and industrial worlds. However, as the preprocessing, documents or texts should be encoded into numerical vectors for using traditional techniques for the task.

Encoding documents so causes the two main problems. The first problem is huge dimensionality where documents must be encoded into very large dimensional numerical vectors. In general, documents must be encoded at least into several hundreds dimensional numerical vectors in previous literatures. This problem causes very high costs for processing each numerical vector representing a document in terms of time and system resources. Much more training examples are required proportionally to the dimension for avoiding over-fitting.

The second problem is sparse distribution where each numerical vector has zero values dominantly. In other words, more than 90% of its elements are zero values in each numerical vector. This problem degrades the discrimination among numerical vectors. This causes poor performance of text categorization. In order to improve performance of text categorization, the two problems should be solved.

The idea of this research is to avoid the two problems by encoding documents into tables instead of numerical vectors. The proposed approach to text categorization is called index based class of approaches in this research. Each table is a collection of entries consisting of words and their weights indicating the importance of words in a given document or a corpus. Category by category,

※ Corresponding Author : Taeho Jo, Address : 1212 Hitech Inha University, Yonghyundong Namgu Incheon 402-751, Korea, TEL : +82-32-860-8984, FAX : N/A, E-mail : tjo018@inha.ac.kr

Receipt date : Dec. 29, 2007, Approval date : Feb. 11, 2008

[†] School of Computer and Information Engineering Inha University

we can sum weights of matched words between a table given as the surrogate of a document and a table as a categorical profile which will be explained later. Therefore, an unseen document is classified as the category corresponding to the maximum summed weight.

The performance of the proposed approach will be validated through the three sets of experiments in section 5. For doing that, the test beds used for the experiments are three collections of news articles: Retuer21578. The proposed approach is compared with the three machine learning based approaches: KNN (K Nearest Neighbor), NB (Naïve Bayes), and SVM (Support Vector Machine). F1 measure where recall and precision are combined with their equal proportion is adopted as the evaluation measure. In section 5, it is shown that once the optimal option is given, the proposed approach is better than any machine learning based approach.

This paper consists of six sections including this section. In section 2, we will survey previous cases of applying one of the machine learning based approaches to text categorization. In section 3 and 4, we will describe the process of encoding documents into tables and the proposed text categorization system, respectively. In section 5, the performance of the proposed approach is validated by comparing the approach with the three machine learning based ones on the test bed: Reuter21578. In section 6, we will mention the significance of this research and further research as the conclusion.

2. PREVIOUS WORKS

This section concerns the exploration for previous research on text categorization. In 2002, Sebastiani mentioned two kinds of approaches to text categorization in his research paper [Sebastiani 2002]. One is rule based class of approaches and the other is machine learning based one of approaches. He count the former out since the class of approaches is very naïve, and he surveyed only machine learning based class. Among approaches belonging to the machine learning based class, we will survey representative four approaches: KNN (K Nearest Neighbor), NB (Naïve Bayes), SVM (Support Vector Machine), and Neural Networks in this section, because of their popularity.

The first representative approach to text categorization is KNN. In 1992, KNN was initially applied to classification of news articles by Massand et al [1]. In 1999, Yang compared 12 approaches to text categorization with each other, and observed through her experiments that KNN is one of recommendable approaches [2]. In 2002, Sebatiani evaluated KNN as a simple and competitive algorithm with SVM evaluated as the best algorithm [3]. Its disadvantage is that KNN costs very much time for classifying objects, given a large number of training examples because it must computes similarities of each unseen example with all individual training examples for selecting some of them.

Another popular approach to text categorization is NB. This approach is a variant of the Bayes Classifier based on the Bayesian Rule which assumes the independence of attributes [4]. In 1997, Mitchell mentioned NB as a typical approach to text categorization in his text book [4]. In terms of a supervised learning algorithm, its advantage is that it learns training examples with its higher speed than neural networks. However, its disadvantage is that an almost zero value of probability influences on the entire posteriori probability; a smoothing scheme was proposed for solving the problem [4].

The third representative approach to text categorization is SVM. In 1998, it was initially applied to text categorization by Joachims [5]. He validated the better classification performance of SVM in text categorization by comparing it with KNN and NB. Drucker et al adopted SVM for implementing a spam mail filtering system and compared it with

NB in implementing the system in 1999 [6]. In 2000, Cristianini and Shawe-Taylor presented a case of applying SVM to text categorization in their textbook [7].

The last representative approach to text categorization is Neural Networks. Among models of neural networks, MLP (Multi Layers Perceptron) with the back propagation algorithm is most popular model. The model of neural networks was initially applied to text categorization in 1995 by Wiener [8]. In 2002, successively to Wiener, Ruiz and Srinivasan applied several MLPs to text categorization by combining them hierarchically [9]. The combined model of neural networks in their research was called HME (Hierarchical Mixture of Experts).

In order to apply one of traditional machine learning based approaches including the four representative approaches, documents must be encoded into numerical vectors. Encoding so causes the two main problems: huge dimensionality and sparse distribution as mentioned in section 1. There was a previous attempt to solve the two problems without encoding documents so. In 2002, Lodhi et al proposed a string kernel for applying Support Vector Machine to text categorization [10], and in their research, documents are used as their raw form. However, their proposed version of SVM failed to be better than the traditional version of SVM [10].

3. DOCUMENT ENCODING

This section concerns the process of encoding a document or documents into a table. Figure 2 illustrates the process with three steps. A document or documents is given as input of the process, and a list of words and their frequencies is generated from the process. The three steps illustrated in figure 2 will be explained. After that, the three schemes of weighting words will be also mentioned.

As illustrated in figure 1, a document or documents may be given as input of this stage. If more than two documents are given as the input, their full texts are concatenated into an integrated full text. The integrated full text becomes the target for the tokenization. The full text is tokenized into tokens by a white space or a punctuation mark. Therefore, the output of this step is a list of tokens.

The next step to the concatenation & tokenization is the stemming & exception handling, as illustrated in figure 2. In this step, each token is converted into its root form. Before doing that, rules of stemming and exception handling are saved into a file. When the program encoding documents is

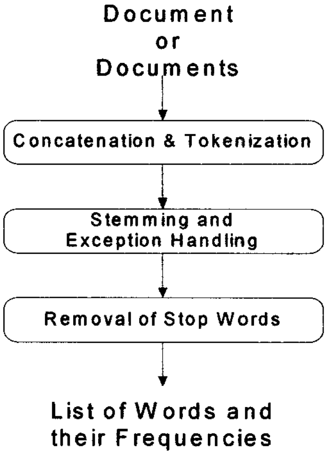

Fig. 1. The Process of Mapping Document or Documents into a Table

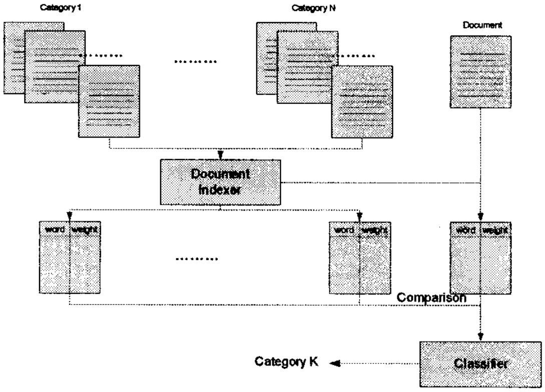

Fig. 2. The Modules involved in Implementing the Proposed Text Clustering System

activated, all rules are loaded into memory and the corresponding one of them is applied to each token. The output of this step is a list of tokens converted into their root forms.

The last step of extracting feature candidates from a corpus is to remove stop words as illustrated in figure 1. Here, stop words are defined as words which function only grammatically without their relevance to content of their document; articles (a an, or the), prepositions (in, on, into, or at), pronoun (he, she, I, or me), and conjunctions (and, or, but, and so on) belong to this kind of words. It is necessary to remove this kind of words for more efficient processing. After removing stop words, frequencies of remaining words are counted. Therefore, a list of the remaining words and their frequencies is generated as the final output from the stage illustrated in figure 1.

Although there are other schemes of weighting words, we will mention only three schemes as representative ones. For first, we can assign frequencies themselves to words as their weights. For second, we may assign normalized frequencies generated from dividing their frequency by the maximum frequency. For third, we can weights words using equation by equation (1),

$$weight_i(w_k) = tf_i(w_k)(\log_2 D - \log_2 df(w_k) + 1) \qquad (1)$$

where $weight_i(w_k)$ indicates a weight of the word, w_k, which indicates its content based importance in the document, i, $tf_i(w_k)$ indicates the frequency of the word, w_k in the document, i, $df(w_k)$ is the number of documents including the word, w_k, and D is the total number of documents in a given corpus. Among the three schemes, we adopt the third for weighting words in this research.

4. PROPOSED TEXT CLUSTERING SYSTEM

This section concerns the proposed text categorization in terms of its architecture and flow. Figure 2 illustrates the architecture of the proposed text categorization. As shown in figure 2, there are two modules involved in the system. The first module is named as 'document indexer', and encodes a document or documents into a list of words and their frequencies. The second module is named as 'classifier', and categorizes directly an unseen document.

The left part of figure 2 shows the process of building category profiles using labeled sample documents. In the view of machine learning, the process may be called 'learning' [Mitchell 1997]. Tables are generated from this process and become references for categorizing unseen documents. Each table corresponds to each category. The weights of the table indicate the relevancy of words to the given category, and they are called categorical weights in this research.

The right part of figure 3 shows the process of categorizing an unseen document based on categorical weights. In this process, a particular unseen document is given as the input. The process corresponds to generalization in view of machine learning [Mitchell 1997]. The unseen document is converted into a table by the module, 'document indexer'. The weights of the table given as the surrogate of the unseen document indicate the relevancy of

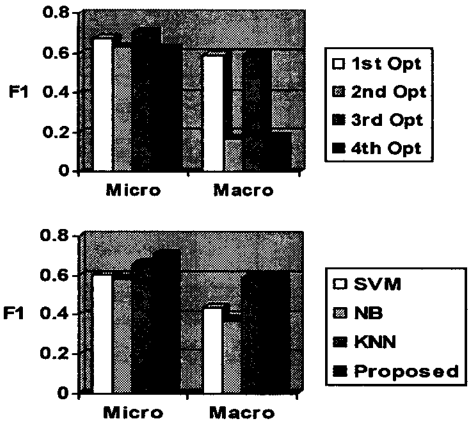

Fig. 3. Results of Four Options of the Proposed Approach and ML based Approaches in Reuter21576

words to its content; the weights are called substantial weights in this research.

We already above mentioned the two kinds of weights of words involved in the proposed text categorization system. One kind is categorical weights given in the categorical profiles. In this research, we may define the categorical weights as the real value indicating how much words are relevant to the given category. The other kind is substantial weights given in a surrogate of an unseen document. In this research, we define substantial weights as real values how much words concern the content of the document.

In the proposed system, unseen documents are categorized based on matched words between two tables. One is given as a surrogate of an unseen document, and the other i^ given as a categorical profile. We can get matched words from the two tables. We can compute a categorical score by summing the weights of the matched words. Wewillassignthecategorycorrespondingtothemaximumcategoricalscoretotheunseendocument. We use categorical weights, substantial weights, or products of both depending on the given option of the proposed system. The detail explanation is included in section 5.

5. EMPIRICAL RESULTS

These experiments concern the comparisons of the proposed and machine learning based class of approaches to text categorization. We used three collections of news articles as the test beds of the experiments: NewsPage.com, 20NewsGroups, and Reuter21578. In using machine learning based class of approaches, documents are encoded into 250 dimensional numerical vectors, while in using the proposed approach, documents are encoded into table with 25 entries. In the experiments, the proposed approach is compared with the three machine learning based approaches: KNN, NB, and SVM. The goal of the experiments is to validate the performance of the proposed approach to text categorization by comparing it with the three traditional ones.

The test bed is Reuter21578, which is a typical standard test bed for evaluating approaches to text categorization. Each news article is given as a document in SGML format, while each news article is given as a plain text in the previous test beds. In the experiment, we selected the most frequent ten categories and table 1 shows the ten categories and a number of documents in each category. The partition of the test bed into training set and test set follows the version, ModApte, which is the standard partition of Reuter 21578 for evaluating text classifiers [Sebastiani 2002]. The difference of this test bed from 20NewsGroups is that in this test bed each news article is labeled with more than one category, while in the previous test bed each news article is labeled with only one category.

We defined the four options in using the proposed approach to text categorization, as illustrated in table 2. In the first option, categorical scores are computed based on the number of matching words as the base option. In the second option and the third option, categorical scores are computed by summing substantial weights and categorical weights of matching words, respectively. In the forth option, categorical scores are computed by

Table 1. Training Set and Test Set of Reuter21578

Category Name	Training Set	Test Set	#Document
Acq	1452	672	2124
Corn	152	57	209
Crude	328	203	531
Earn	2536	954	3490
Grain	361	162	523
Interest	296	135	431
Money-Fx	553	246	799
Ship	176	87	263
Trade	335	160	495
Wheat	173	76	249

Table 2. Four Options in the Proposed Approach

First Option	Number of Matched Words
Second Option	Substantial Weights of Matching Words
Third Option	Categorical Weights of Matching Words
Forth Option	Substantial * Categorical Weights of Matching Words

summing products of both weights of matching words.

We can implement a text categorization system without or with decomposing a text categorization task into several binary classification tasks. In the two previous sets of experiments, text categorization was not decomposed into binary classification tasks, since each news article is labeled exclusively with only one category. In other words, each classifier generates one of predefined categories as its output. However, in this set of experiments, since each document is labeled softly with more than one categories, the text categorization task should be decomposed into binary classification tasks. Therefore, each classifier answers whether an unseen document belongs to its corresponding category or not; F1 measure which has been traditionally used for information retrieval systems is used as the evaluation measure, instead of accuracy.

F1 measure is the combination of precision and recall with their identical portions. Each classifier corresponds to a category one to one; ten classifiers are given according to the number of predefined categories. There are two ways for averaging F1 measures of the ten classifiers: micro averaged F1 and macro-averaged F1.

Figure 3 shows the results of evaluating the four options within the proposed approach and those of comparing the proposed one under its optimal option with the three approaches in the test bed, 20NewsGroups. In this set of experiment, SVM was added to the two machine learning based approaches. The third option is optimal in the results of evaluating the four options within the proposed one. This set of experiments also shows that the proposed approach with its optimal option also outperforms the three machine learning based ones.

Let's consider the four options within the proposed one. The three sets of experiments show that the third option is optimal. Especially in the second option, the performance of the proposed one is very poor, as illustrated in figure 3. When the fourth option where both kinds of weights are considered is selected, the performance is not good as the proposed one with the third option. Through the three sets of experiments, we can conclude that the substantial weights of words are very harmful for categorizing documents; we should improve the process of computing the substantial weights of words.

According to the results of the experiments, we can conclude that if the third option is adopted in the proposed approach, it outperforms the three machine learning approaches: SVM, NB, and KNN. Especially in the tasks of text categorization without the decomposition, the proposed scheme is more recommendable than the machine learning based approaches. The three machine learning algorithms were comparable to the proposed one, as shown in the right of figure 3.

6. Conclusions

This research proposes an alternative approach to machine learning based ones to text categorization. In the proposed approach, a document or documents are encoded into a table, instead of a numerical vector or numerical vectors. In other words, we can avoid the two main problems in encoding documents into numerical vectors: huge dimensionality and sparse distribution. The performance of the proposed approach was validated in the previous section using the three test beds: NewsPage.com, 20NewsGroups, and Reuter21578. Since the two problems are solved, the proposed approach are shown to work better than machine learning based ones for text categorization.

There may be many ways of computing weights of words. In this research, we computed weights of words using equation (1), because of the popularity in the information retrieval. Note that the weights do not reflect exactly the relevancy of words to a given category or a content of a document. We need to develop several state of the art schemes for computing weights. In further research, we will compute weights of words using by combining multiple schemes with each other.

If we could develop various schemes for computing weights of words, we may define multiple tables to a document or corpus. There are two ways for treating multiple tables. The first way is to integrate multiple tables corresponding to a document or a corpus into a table. The second way is to treat the multiple tables as a committee. In further research, we will evolve the proposed approach by encoding a document or corpus into multiple tables.

In this version of the proposed text categorization system, the number of entries of tables is fixed constantly. The proposed one is called static index based approach. However, the optimal number of entries is very dependent on the given document or corpus. The size of each table should be optimized in terms of two factors: reliability and efficiency. In the further research, we will propose dynamic index based approach where the size of table may be changed automatically for satisfying the both factors.

The weights of words may be automatically adjusted to improve the performance of text categorization in implementing the proposed approach. We need an additional set of labeled documents, called validation set. The set is built by separating some of a given training set. The weights of words are updated to minimize misclassification rate of the examples in the validation set. The modified version may be regarded as a fusion of the proposed approach and the machine learning based one.

REFERENCES

[1] B. Massand, G. Linoff, G, and D. Waltz, D, "Classifying News Stories using Memory based Reasoning," *The Proceedings of 15th ACM International Conference on Research and Development in Information Retrieval*, pp. 59-65, 1992.

[2] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information Retrieval*, Vol.1, No.1-2, pp. 67-88, 1999.

[3] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Survey*, Vol.34, No.1, pp. 1-47, 2002.

[4] T. M. Mitchell, Machine Learning, McGraw-Hill, 1997.

[5] T. Joachims, "Text Categorization with Support Vector Machines: Learning with many Relevant Features," *The Proceedings of 10th European Conference on Machine Learning*, pp. 143-151, 1998.

[6] H. Drucker, D. Wu, and V. N. Vapnik, "Support Vector Machines for Spam Categorization," *IEEE Transaction on Neural Networks*, Vol.10, No.5, pp. 1048-1054, 1999.

[7] N. Cristianini and J. Shawe-Taylor, Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

[8] E. D. Wiener, "A Neural Network Approach to Topic Spotting in Text," The Thesis of Master of University of Colorado, 1995.

[9] M. E. Ruiz and P. Srinivasan, "Hierarchical Text Categorization Using Neural Networks," *Information Retrieval*, Vol.5, No.1, pp. 87-118, 2002.

[10] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C., 2002, Text Classification with String Kernels, Journal of Machine Learning Research, Vol.2, No.2, pp. 419-444.

Taeho Jo

Taeho Jo received PhD degree from University of Ottawa in 2006. Currently, he works for Inha University as a professor. He has submitted and published more than 100 research papers to journals and proceedings since 1996. Previously he has ever worked for industrial organizations: Samsung, ETRI, KISTI, and KAIST Institute for IT Convergence. His research interests are text mining, neural networks, machine learning, and information retrieval.