ISSN: 2349-7300
ISO 9001:2008 Certified
International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences
(IJIRMPS)
Volume 1, Issue 2, December 2013

Text Categorization Using Soft Computing Method

P Tiwari, M.S.Rajput

Abstract: ***Text categorization is the task of deciding whether a document belongs to a set of pre- specified classes of documents. Automatic classification schemes can greatly facilitate the process of categorization. Categorization of documents is challenging, as the number of discriminating words can be very large. The traditional method of text categorization like KNN has a defect that the time of similarity computing is huge. In this paper, neural network technique back propagation is proposed. Comparative study of back propagation technique was done with traditional technique of KNN and it was found that the time of similarity computing is decreased largely in back propagation. Following are the objectives of this study -***

- ***Investigate approaches to analyzing large sets of data, including representation, feature selection and automatic classification.***
- ***Build an automatic document classifier for the content on the newsgroup dataset.***
- ***Compare the performance of the suggested approach of KNN and ANN.***

I. INTRODUCTION

In the last few years, there has been enormous growth in the amount of text documents available in digital form. Part of the driving force behind this is the advance in document imaging technologies, which facilitate the conversion of paper documents into digital forms, and the popularity of word processing software enabling the creating of digital document. The popularity of the internet recently has only accelerated this growth. Internet services such as electronic mail, USENET Nes and the World Wide Web are still mainly text based. The explosive growth of unstructured information on the internet and in particular the World Wide Web has greatly increased the need for information retrieval system. This is evident from the rapid growth of the number of World Wide Web indexing services which has become available recently. Some of the most popular World Wide Web indexing service include Alta vista, infoseek, Lycos, Inktomi and Yahoo.

A text document database containing a large number of text documents, a text retrieval engine has to select a subset of the documents to he returned or displayed to the user. Selection is usually based on an input query which represents the information needs of the user? In keyword based text retrieval, the query can be a set of keywords selected by the user as representative to his information needs. The selection criterion can he binary, by which a document is either selected or not selected, or it may involve sorting the documents according to some scoring scheme based on the estimated relevance of each documents to the query , and selecting only a subset of the documents with relatively higher scores than other . The process of sorting the selected documents based on their relevance scored is commonly known as documents ranking. Text categorization [2] (also known as text classification or topic spotting) is the task of automatically sorting a set of documents into categories from a predefined set. This task has several applications, including automated indexing of scientific articles according to predefined thesauri of technical terms, filing patents into patent directories, selective dissemination of information to information consumers, automated population of hierarchical catalogues of Web resources, spam filtering, identification of document genre, authorship attribution, survey coding, and even automated essay grading.

II. RELATED WORK

The neural network model accepts the structure of conceptual, linguistic oriented model, where the problem of document database creation and document indexing for keyword determining is solved. Query entering uses the same mechanisms as document formalization for example document database creating method. Proposed information retrieval system serves [5,7] for information extraction from text documents in natural ovak language. These documents are stored in document base. Each of them is marked with its index, which expresses the document content and the document relevance. User enters the question for that system and system returns him a document subset relevant to his query.

Because of modular structure complexity the information retrieval system can be divided into three subsystems: administrator subsystem, indexation subsystem and user subsystem. Administrator subsystem guaranties document set operations. Administrator determines document set due to creation of document base from them. Document base manager then provides the system representation of documents. He also determines suitable model of document storing and creates document base of system representation. Indexation subsystem solves two tasks. Firstly it is creation of index and secondly it is creation of question representation that is comparable with document index.

III. PERFORMANCE EVALUATIONS AND EXPERIMENTAL RESULTS

Due to the large amount of research efforts spent on the text retrieval task, standard procedures are available for evaluating the performance of text retrieval systems, which are widely used by the information retrieval community.

As many of the concepts in evaluating text categorization systems[2,3] are adapted from the evaluation of text Retrieval systems, an over view of these procedures will be given before we proceed to discuss the techniques for performance evaluation of text categorization systems Evaluating Text Retrieval Systems As is the case with many other systems such as database management systems, the performance of text retrieval systems and information retrieval systems in general can be measured based on a number of different criteria, such as the execution time needed to perform a certain task, or the storage and memory overhead required by the system for execution. While these performance measures [8] are common to many types of systems, most of the performance studies in information retrieval have been focused on measuring the retrieval effectiveness of text retrieval systems. In this section, we concern ourselves only with the evaluation of retrieval electiveness for text retrieval systems.

By this Comparison, the retrieval effectiveness can be computed according to two commonly used measures, namely precision and recall

Precision: Precision measures the accuracy of the retrieval result as indicated by the pro portion of retrieval document that are relevant. it is defined as

$$\text{Precision} = \frac{\text{Number of test set category members assigned to category}}{\text{Total number of test set members assigned to category}}$$

Recall: Recall measure how extensive the retrieval result is ,as indicated by proportion of relevant documents retrieved . it is defined as

$$\text{Recall} = \frac{\text{Number of test set category members assigned to category}}{\text{Number of category members in test set}}$$

For the set of retrieval results corresponding to the set of test queries, a set of precision and recall values can be computed. The overall performance can them be found by averaging, giving the average precision and recall for the system being tested

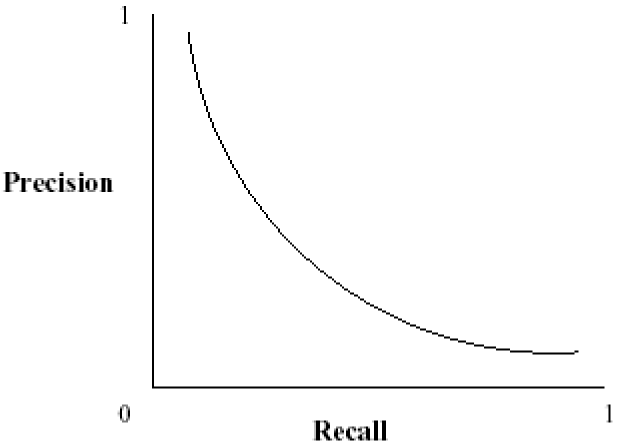

Fig: 1 a typical precision recall graph

For text retrieval systems having a parameter which can affect precision and recall, different pairs of precision and recall values can be calculated by varying the parameter. A common approach is then to plot a precision-recall graph [6, 8] based on the set of pairs of precision and recall values. Generally, there is a trade such that if recall is raised by varying the parameter, precision will drop advice-versa. Figure 1 shows the typical shape of a precision-recall graph for text retrieval systems. .Over the years, a number of test collections have been developed and used in the evaluation of text retrieval systems.

Like text retrieval systems, performance of text categorization systems can be evaluated based on their categorization Effectiveness. Just as retrieval ejectiveness indicates the ability of a text retrieval system in retrieving relevant documents based on a user input query, categorization electiveness indicates the ability of a text categorization system in providing accurate classification of text documents based on a set of pre-defined categories. To evaluate the performance of text categorization systems, we need to measures for the categorization electiveness. One obvious approach is to re den precision and recall in the context of text categorization. We can arrive at the dentitions by looking at an analogy between the text retrieval task and the text categorization task. In text retrieval, the retrieval system has to make the decision of whether to retrieve a document or not, based on whether the document is relevant to a given query. This decision has to be made for each document and each different query. Similarly, a text categorization system has to make the decision of whether to assign a category to a document or not, based on whether the document is relevant to the topics represented by the category. This decision is made for each document and each pre-defined category.

When there are more than one pre-defined category, there are two different ways for computing the precision and recall of the categorization system, referred to as macro averaging and micro averaging in .In macro averaging, separate precision and recall values are calculated for each category, and then averaged over all categories to get the overall precision and recall. On the other hand, micro averaging calculates the precision and recall by considering assignment de-cessions for all categories at once. Similar to the case of text retrieval, categorization electiveness of text categorization systems can be evaluated by testing the system using a test collection designed for the categorization task. A test collection for text categorization systems should at minimal consists of these three components: a set of text documents, a set of pre-defined categories, and a speciation of which of the pre-defined categories each document belongs to. The speciation of category membership for each document is analogous to the standard judgments provided in test collections designed for the text retrieval task. To evaluate the categorization electiveness, the categorization system under test will be used to categorize the set of documents based on the set of pre-defined categories provided in the test collection. Then the categorization result is compared with the speciation given. By this comparison, the precision and recall can be computed.

Experiment-1

The performance of text categorization systems can be evaluated based on their categorization effectiveness The effectiveness measures of recall, precision and F-measure are defined as

$$\text{Precision} = \frac{\text{Number of test set category members assigned to category}}{\text{Total number of test set members assigned to category}}$$

$$\text{Recall} = \frac{\text{Number of test set category members assigned to category}}{\text{Number of category members in test set}}$$

We used the macro-average method to obtain the average value of the precision and recall. The F-measure is based on the micro-average value. The performance results are given in table1.

Category	BPNN pecision	BPNN Recall	KNN Precision	KNN Recall
Money-supply	0.938	0..946	0.832	0.913
Coffee	0.929	0.933	0.847	0.901
Gold	0.955	1.000	0.943	0.914
Sugar	0.824	0.895	0.952	0.884
Trade	1.000	0.932	0.725	0.786
Crude	0.948	0.924	0.944	0.896
Grain	0.948	0.924	0.937	0.928

Table (1) Compare result between KNN and Back propagation algorithm

Following Figure show the replots of the precision and recall values

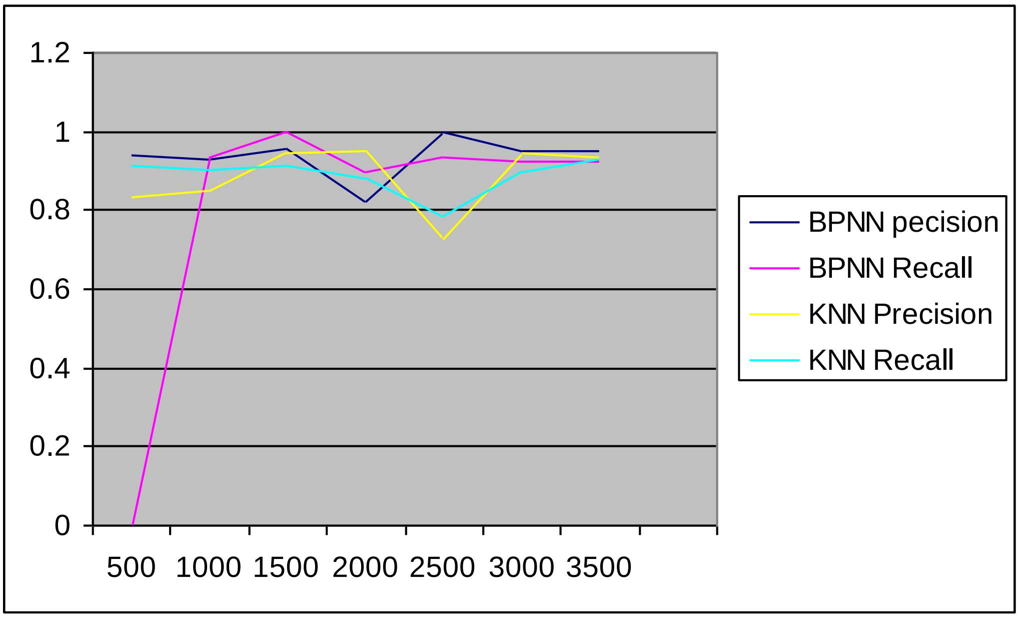

Fig (2) Compare Result between BPNN and KNN Algorithm

The size of the network and some parameters used in our experiments are given in table.

#Input Node	#Hidden Nodes	#output nodes	Learning Rate	Momentum
1000	15	10	.01	0.8

Table (2) Size of the network and parameter

IV. CONCLUSION

With the rapid growth in the amount of electronically stored textual data, information overloading is becoming a serious problem. The popularity of text intensive Internet services such as the World Wide Web, USENET news and electronic mails has resulted in a rapid increase in the number of online information retrieval systems trying to help the users and the relevant information they need. Without such systems, finding the information that one need is often difficult and time consuming given the huge amount of information being put online.

Future Work There are a number of directions that this research can be pursued further. Here we summarize these future research directions: To further test the effectiveness of the proposed model and to increase the generality of the empirical study, more extensive experiments should be conducted by using larger training and test sets. he categorization model proposed is general enough to accommodate most neural network models with the supervised learning paradigm. Possible models include linear neural networks such as the Adaline and newer variants to the original gradient descend based Back propagation learning such as the conjugate gradient [40] and Rprop [45] algorithms. Further comparison study can be conducted to investigate the performance of each of these models. In this thesis, we used the term occurrence frequencies (TF) as the document features. One disadvantage of this approach is that longer documents tend to have higher TF and are favored in the categorization process. A possible solution to this problem is to normalize TF by the document length. Experiments should be carried out to and out whether categorization performance will be increased by this variation

REFERENCES

[1] Chen et al., Generating, integrating, and activating thesauri for concept based document retrieval,"IEEE Expert vol. 8, no. 2, pp. 25-34, 1998.

[2] H. Chenet al. Automatic concept classification of text from electronic meetings," Communications of the ACM, vol. 37, no. 10, pp. 56-73, 1999.

[3] H. Chen and J. Kim, Gannet: a machine learning approach to document retrieval," Journal of Management Information Systems, vol.11, no.3, pp.7-41, 1994.

[4] H. Chen and T. Ng, An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound search vs. connectionist hope field net activation," Journal of the American Society for Information Science , vol.46, no. 5, pp. 348-369,1995.

[5] F.Crestani,Learningstrategiesforanadaptiveinformationretrievalsystemusing neural networks," in Proceedings of the IEEE International Conference on Neural Networks , vol. 1, pp. 244-249, 1993.

[6] F. Crestani, An adaptive information retrieval system based on neural networks," in Proceedings of the International Workshop on Artificial Neural Networks, pp. 732-737, 1993.

[7] F.Crestani,Learningstrategiesforanadaptiveinformationretrievalsystemusing neural networks," in Proceedings of the IEEE International Conference on Neural Networks , vol. 1, pp. 244-249, 1993.

[8] S. Deerwester et al. Indexing by latent semantic analysis," Journal of the American Society for Information Science , vol. 41, no. 6, pp. 391-407, 1990.12. W. B. Frakes and R. Baeza-Yates, Information Retrieval Data Structures and Algorithms. Englewood Clis, New Jersey: P T R Prentice Hall, 1992.