Iranian Int. J. Sci. ***4****(1), 2003, p. 87-96*

Intelligent Document Categorization for Information Retrieval based on Inter-connected Neurons

D. Sridharan[1], P. Sakthivel[1] and S.K.Srivatsa[2]
[1] Ramanujan Computing Centre, Faculty of Electrical Engineering
Anna University, 600025, Chennai, India
mail:sridhar@annauniv.edu
[2] School of Electronics and Communication Eng., Faculty of Electrical Engineering
Anna University, 600025, Chennai, India
E-mail: profsks@annauniv.edu

(received: 10/19/2001 ; accepted: 8/3/2002)

Abstract

The World Wide Web contains huge amount of dynamic, heterogeneous, and hyperlinked distributed documents. Many modern information retrieval systems are developed based on matching process, which is automated, but classification and query formulations are manual process. The approach taken in this paper is to add intelligence to Information Retrieval by way of Document Classification based on Vector Space Model using Inter-connected neurons with relevance feed back from the retrieved documents to the intelligent classifier as well as to the user query. The basic idea is to integrate three existing techniques: classification, query expansion and relevance feedback both to classified documents and user query to achieve a concept-based information search for the Web.

Keywords: *Information Retrieval, Document Classification, Categorization and Inter-connected Neurons*

Introduction

With the advent of the World Wide Web in mid 90s, Internet became a household word and everyone was searching the relevant information in the web. A number of search engines became available on the Web and most of these engines are developed from University research projects. In the early days (1996-1997), the search engines competed on the number of pages indexed and the speed of retrieval. Now, most searchers realize that it is completely irrelevant if the search engine retrieves 100,000 or 1 million records and also it is very difficult to find the relevant document among large collection of retrieved documents. Now researchers recognized the limits of the massive quantity and lack of quality of information on the Web and employed a new idea based on concept searching, which automatically expanding the search to look for similar ideas and synonyms and alternatives of the search words. This type concept searching or classification retrieves relevant pages that would have been missed in a standard Boolean search. The important feature of categorization is to identify the key concepts within a document and group together according to their concepts on specific topics and displaying it in a way that the user understands. Researchers have also focused on automatic query expansion to help the user formulate what information is really needed. Another research topic is on relevance feedback from the user, which gives the relevance of documents to clarify the ambiguity. In fact, these three techniques complement each other. However, the mechanisms of relevance feedback based on words or documents in the past research both have their own deficiencies. Word feedback has its upper bound performance in lexical-semantic expansion (Harman, 1992) and document feedback is sometimes too tiring for the users. An automatic query reformulation can be used to overcome these problems. The proposed system can be accommodated easily and inexpensively in future generations of web-based retrieval systems and technologies

The basic process of Information Retrieval using relevance feedback both to query and to the classifier is visualized in figure 1. The remainder of this paper is organized as follows. Section 2 and 3 describes Vector Space Model followed by query reformulation via relevance feedback based on Rocchio algorithm. And the next section 4 describes the taxonomy of designs and techniques used in this work. Section 6 outlines various related work on intelligent document retrieval system. The approach used in this paper is restricted to a relatively simple but effective relevance feedback for retraining the network and refining the user query to improve the performance of retrieval.

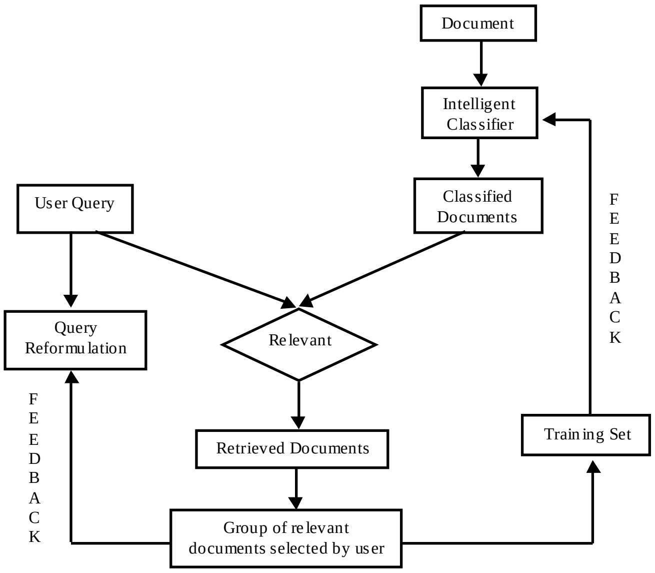

Figure 1 – Basic Process of Information Retrieval.

Vector Space Model

The Vector Space Model was invented by Gerard Salton in the 1960's, is a mathematical model that represents documents and queries as term sets and computes the similarities between queries and documents. When combined into connectionist method with relevance feed back from the retrieved document gives intelligence for classifying the web documents and also for refining the query. The required criterion is that the queries and document use the same term set. In the vector space model, both queries and documents are represented as term vectors of the form Di = (di1, di2, ...,dit) and Q = (q1, q2, ...,qt). A document collection is then represented as a term-document (TD) matrix A:

$$sim(D_iD_j) = \sum_{1\leq l\leq n} d_{li}d_{lj}$$

$$A = \begin{array}{cc} & \begin{array}{cccc} T_1 & T_2 & & T_t \end{array} \\ \begin{array}{c} D_1 \\ D_2 \\ D_3 \end{array} & \left(\begin{array}{cccc} a_{11} & a_{12} & \ldots. & a_{1t} \\ a_{21} & a_{22} & \ldots. & a_{2t} \\ a_{i1} & a_{i2} & \ldots. & a_{it} \end{array}\right) \end{array}$$

The rows of the above TD matrix represent the individual documents, the columns represent unique words and each entry in the matrix represents the number of occurrence of that word in that document. The similarity between a query vector Q and a document term vector D can also be computed. This is particularly advantageous because it allows one to sort all documents in decreasing order of similarity to a particular query. The required knowledge base for the connectionist method is developed from this Term-document matrix and remembered by a network of Inter-connected neurons, weighted synapses and threshold logic units. Learning algorithm can be applied to adjust connection weights, so that the network can predict or categorize the documents correctly.

Query refinement via relevance feedback

Relevance feedback is the reformulation of a search query in response to feedback provided by the user for the results of previous versions of the query and has long been suggested as a solution for query modification. Past research has applied Robertson and Sparck Jones' term weighting to query expansion (Robertson et al., 1988) given the top 10–30 documents as relevant and all other documents in the corpus as non-relevant. Rocchio describes an elegant approach and shows how the optimal vector space query can be derived using vector addition and subtraction given the relevant and non-relevant documents (Rocchio *et al.,* 1971). The goal of relevance feedback is to retrieve and rank highly those documents that are similar to the document(s) the user found to be relevant

Now, given the group of documents selected by user as feedback unit, and then each document in this group has been digested as a set of relevant document vector (R) and a set of non-relevant document vector (S), the original query can be modified by Rocchio's algorithm.

The new query is obtained by Q’ = a Q + b sum (R) – c Sum (S). Where Q is original query vector and a,b,c are Rocchio weights.

Design of Neural Network for Classification

The proposed model for Intelligent Document Classifier is a feed forward Network with two layers of units fully connected between them as shown in Figure 2.

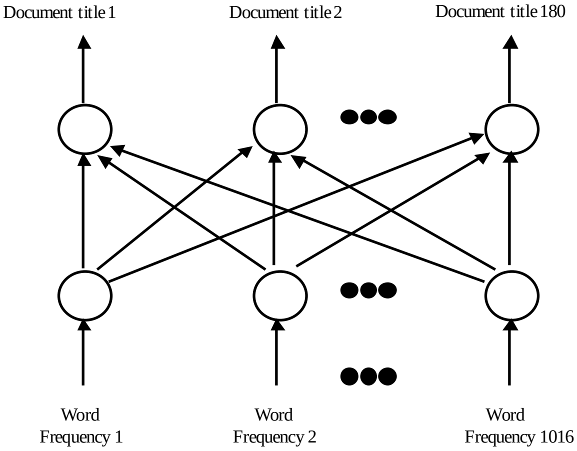

Figure 2 – Topology of the network

The first layer has the input unit and second layer has the output units. Each input unit in the first layer is associated with each word-stem and weight of the input unit is the word frequency in the document. Each output unit is associated with title of the research area and threshold logic units, defining, if the particular title characterizes the input document.

The data set used for training and testing the Intelligent Document Classifier consists of 2,344 research document and each one contains the title, authors name, citation information, abstract. This collection of documents was processed in such a way that tokenizing words from title and abstract, eliminating common words using a stop list, and the remaining words are passed through the stemmer. After this process, a total of 12,292 stemmed-words and 4,049 different titles were obtained from the entire collection. Finally the document frequencies are computed for each word-stem and document titles and it is found to be varying from 1 to 1,511 for word-stem and from 1 to 2,102 for document titles.

The number of stemmed-words in the above collection decides the number of nodes in the input layer of the inter-connected neurons. Hence it is decided to reduce the number of units in the input layer by setting a minimum Inverse Document Frequency (IDF) threshold as 75 on the entire collection. This is based on the assumption that a rare term that appears in only a few documents will have a high IDF score and at the same time the term that appears in most document will have a low IDF score. The terms that appear frequently in one document, but rarely on outside, are more relevant to the topic of categorization. A threshold of 75 for Inverse Document Frequency offers a reasonable reduction in the number of terms while retaining terms that are neither too specific nor too general. Using this criterion, the stemmed words are reduced to 1,016 for IDF threshold of 75. So, the number of units in the input layer was also reduced to 1,016. Using the same criterion for research titles, it is possible to have 180 neurons in the output layer. Finally, connections between all input units and all output units were created: 1016 X 180 = 182880. A Stuttgart Neural Network Simulator (SNNS, 1998) is used for developing this intelligent classifier

Experiments

In the first experiment, the entire collection of 2,344 documents was divided into two sets: one is training set and another one is validation set. For training the network, 586 documents were used from the entire collection and this set is called as training set. Training documents were selected randomly among the training set. For validating the network, 1,758 documents were used and called this set as validation set and used to evaluate the network's ability to recognize different categorization of documents.

For training the network, a criterion was set that the average distance between patterns was less than a predefined value. An initial run was trained for a maximum of 25 cycles. The average distance per pattern in the last iteration was 0.75 and the error per pattern in the last cycle was 19.15. Attempts to train the network allowing more number of iterations resulted in higher values of error. Finally the network is converged to the value of an acceptable error (0.2) in 19 cycles and took about 6 hours (using a HP-700 workstation).

As a result of validation it is revealed that 90% of the proposed documents were correct. It is possible that the 10% of the documents are not completely correct and they may be coming under different categorization. This is because the network is trained by training set with the minimum Inverse Document Frequency threshold of 75.

In the second experiment, the user viewed an average of 62 full documents during the first search. Of the full documents seen, on average 18 of them were declared relevant. After a few tuning experiments, the constant are fixed as follows: a=0.4, b=0.4 c=0.2. This resulted in a great reduction in the number of terms without significant drop in performance and the new query is calculated. The performance of this system can be measured by evaluating precision and recall.

During the search the users used a number of search criteria other than thesaurus concepts such as free-text, author searching, limiting by publication types, although thesaurus searching accounted for 70 % of all the search criteria selected with free-text accounting for 12.5 % of all search criteria. This gives an average total of 27.7 relevant documents per query. Then the Relevance feedback average recall is 17.7/27.7=0.64 and Relevance feedback average Precision is 17.7/61.6=0.29(for titles)

When looking at the rankings a maximum of 22 relevant records were found in any one set of thirty documents produced in the ranking with an average of 5.2 documents per ranking using feedback only to query. Then the Relevance Feedback average recall is 5.2/27.7=0.19 and Relevance Feedback average precision = 5.2/30 = 0.17. The results seem to indicate that this method performs far better than the simple relevance feedback process in which only user query is modified.

Related work

Many interesting knowledge-based systems have been developed in the past few decades for applications such as medical diagnosis, engineering troubleshooting, and business decision making (Hayes - Roth *et al.,*1994). Most of these systems have been developed based on the manual knowledge acquisition process, a significant bottleneck for knowledge-based systems development. A recent approach to knowledge elicitation is referred to as ``knowledge mining'' or ``knowledge discovery'' (Hayes-Roth *et al.,* 1994 and Frawley *et al.*, 1991).

PLEXUS is an expert system that helps users find information about gardening (Vickery, *et al.*, 1987). The system has a knowledge base of search strategies and term classifications similar to a thesaurus. EP-X is a prototype knowledge-based system that assists in searching environmental pollution literature (Smith, *et al.*, 1989). The system interacts with users to suggest broadening or narrowing operations. GRANT is an expert system for finding sources of funding for given research proposals (Cohen, *et al.*, 1987). An intelligent system(AWPC) is developed for Automatic Web Page Categorization for Information Retrieval System (Hisao Mase, *et al.,* 2001) . This system classifies the web page according to the content available in the home page

The Yahoo server developed at the Stanford University represents one attempt to partition the Internet information space and provide meaningful subject categories (e.g., science, entertainment, engineering, etc.). However, the subject categories are limited in their granularity and the process of creating such categories is a manual effort. The demand to create up-to-date and subject specific categories and the requirement that an owner place a homepage under a proper subject category has significantly hampered Yahoo's success and popularity.

Conclusion

This article integrates the three mechanism viz. classification, query expansion and relevance feedback to retrieve a research article from the free text, which contains authors name, citation information abstract and set of manually assigned research key words. The results suggest that this method performs better than the models based on relevance feedback in which only the user queries are reformulated from the retrieved documents. The employment of conceptual feedback with query expansion based on the two models is a new approach in information retrieval. By expanding a query, one could not only increase the number of relevant documents retrieved but also classify the candidate documents. Thus, Web Information Retrieval can be carried out by generating new queries and filtering through the existed evidence. The issue of scalability has to be tested by using a larger collection of documents. The knowledge obtained by the network during the training phase can also be used as fuzzy rule for categorizing the documents.

References

Cohen, P.R., and Kjeldsen, R. (1987) *Information retrieval by constrained spreading activation in semantic networks.* Journal of Information Processing and Management, **23(4)**, 255-268.

Frawley, W.J., Pietetsky-Shapiro, G., and Matheus, C.J. eds. (1991) *Knowledge discovery in databases*: an overview. The MIT Press, Cambridge, MA.

Harman, D. (1992) *Relevance feedback revisited.* In Proceedings of ACM SIGIR , International Conference on Research and Development in Information Retrieval, 1–10.

Hayes-Roth, F., and Jacobstein, N. (1994) *The state of knowledge-based systems.* Communica-tions of the ACM, **37(3),** 27-39.

Hisao Mase., and Hiroshi Tsuji. (2001) *Experiments on Automatic Web Page Categorization for Information Retrieval System.* Journal of Informa tion. Processing Society of Japan, **42(2),**

Robertson, S.E., and Sparck Jones, K. (1988) *Relevance weighting of search terms.* Journal of the American Society for Information Science, **27(3),** 129–146.

Rocchio, J.J., and Salton, G., eds. (1971) *Relevance feedback in information retrieval.* The SMART Retrieval System. Prentice-Hall, Inc., Englewood Cliffs, NJ, pp. 313–323.

SNNS (1998) *Stuttgart Neural Network Simulator ver 4.2 user Manual,* Institute for Parallel and Distributed High Performance Systems, University of Stuttgart

Smith, P.J., Shute, S.J., Galdes, D., and Chignell, M.H. (1989) *Knowledge - based search tactics for an intelligent intermediary system*. ACM Trans. on Information Systems, **(7),** 246-270.

Vickery, A., and Brooks, H. M. (1987) *PLEXUS – The expert system for referral*. Journal of Information Processing and Management, **23(2),** 99-117.