IJITKM Volume 7 • Number 1 • December 2013 pp. 143-146 (ISSN 0973-4414)

Effective Information Retrieval Method Based on Hierarchical Agglomerative Clustering

Ravi Bhushan, Mohit Vats
Department of Computer Science & Engg. SKIET, Kurukshetra, Haryana, India
ravibhushan9@gmail.com

Abstract- Now a days searching on the internet is most widely used operation on the World Wide Web. The amount of information is increasing day by day rapidly that creates the challenge for information retrieval. There are so many tools for perform efficient searching. There are differences in the ways various search engines work, but they all perform basic tasks. We proposed the vector based document models and cosine similarity measures are a highly accurate and efficient practical document clustering solution. The Clustering system definitely improves efficiency and effectiveness in information retrieval system. It also reduces the search space. They search the Internet or select pieces of the Internet based on important words. They keep an index of words they find, and where they find them.

I. INTRODUCTION

There is a vast amount of information retrieval in the libraries, Internet etc to which a speedy and accurate access is required. If to some user the desired relevant information is not made available then it may result into the duplication of work and efforts. Information retrieval system informs the whereabouts of the documents related to the request. It does not inform the user on the study of his inquiry. It merely informs on the existence (or non existence) and whereabouts relating to his request and this excludes data retrieval system[1].
Information Retrieval (IR) is the area of computer science concerned with retrieving information about a subject from a collection of data objects. This is not the same as Data Retrieval, which in the context of documents consists mainly in determining which documents of a collection contain the keywords of a user query. The Information Retrieval System extracts keywords from the query, matches the query against the document representations in its collection, and assigns a relevance status value to each document. A number of factors are considered by the Information Retrieval System to judge the relevance of the documents. Unfortunately, the relevance of a document is a very subjective notion and difficult to measure in practice. Many existing search engines use a metric called the "Page Rank" to measure the relevance of a page. The page rank metric consider a page imp if the page is linked to many other pages on the web. Google search engine is also using page rank metric to assign the relevance to the pages although it is considered 100 more factors in determining the final ranking[6]. The whole document collection is then ranked in descending order based on relevance status value and then the documents are returned to the user in the same order. Many models to index and represent documents as well as queries have been proposed in Information Retrieval literature. Among these, the Vector Space Model (VSM) is most widely used, due to its simplicity and intuitive geometric representation[5]. In VSM, each query and document is represented as a vector. The similarity between these two vectors as often measured by the cosine function is commonly used to represent the relevance status value for that document.

II. Literature Review

This section provides a brief description of the concepts which are used in the proposed system and work done under them.

2.1 Ranking

The ordering of the search results becomes a crucial factor to evaluate the effectiveness of a search engine. A query based search engines normally return a large number of relevant Web pages. To be more effective, the returned pages must be adequately ranked according to their importance with respect to the user's information need. In the

past, link-based page ranking has been treated as an eigen system problem. Two most representative link-based page ranking algorithms are Page Rank[4] and HITS (Hypertext Induced Selection) [5].
Page Rank was proposed by Lawrence Page and Sergey Brin, the graduate students of Stanford, in 1998, and has been used as the core ranking algorithm of Google[3], today's most widely used search engine. The basic idea of Page Rank is that a page is considered important if many other important pages link to it. So within the concept of Page Rank, the rank of a page is given by the rank of those pages which link to it. Hence, the Page Rank of a document is always determined recursively by the Page Rank of other documents. Page Rank score of each page is pre-computed for the entire Web graph, which contains more than 50 billion pages today, and must be upgraded periodically and each upgrading needs hundreds of thousand high-end computers and 3 to 5 days to finish.

2.2 Similarity Function

The vector space model[7] is the standard technique for ranking documents according to a query. In a space defined in this way, the similarity of a query to a document is given by a formula that transforms each vector using certain weights and then calculates the cosine of the angle between the two weighted vectors:

$$\text{Sim}(Q,D_i)= \frac{\sum_i W_{Q,j} W_{i,j}}{\sqrt{\sum_j W^2_{Q,j}} \sqrt{\sum_i W^2_{i,j}}}$$

In pure text-based information retrieval systems, documents are shown to the users in decreasing order using this similarity measure. The most used scheme is the TF-IDF weighting scheme, that uses the frequency of the terms in both queries and documents to compute the similarity. g

2.3 Clustering

clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

III. PROBLEM DEFINATION AND PROPOSED SOLUTION

A lot of information has to be managed for the websites, which is increasing exponentially. One way to deal with the extraordinary overflow of data, is cluster analysis. This lead to the need of organizing a large set of documents into categories through clustering. It is used to divide large unstructured documents corpora into groups of more or less closely related documents. We propose a new similarity measure to compute the similarity of text-based documents based on the Term Frequency and Inverse Document Frequency using Vector Space Model. We apply the new similarity measure to the Hierarchical Agglomerative Clustering Algorithm and develop a new Clustering Approach. These models will provide accurate document similarity calculation and also improve the effectiveness of the clustering technique over the traditional methods.

3.1 Proposed algorithm

Proposed algorithm for generating the cluster based on idf (inverse document frequeny) and cosine similarity.

Input: document sets

Output: Similar documents.

Step 1: N -> Number of Documents

Step 2: Idf -> Inverse document frequency.

Step 3: Df -> Document Frequency.

Step 4: di -> Document Identifiers.

Step 5: For each D do

Step 6: For each di do

Step 7: Read the document from left to right over the document D

Step 8: Calculate the term frequency for each term di € D

Step 9: Calculate the inverse document frequency for each term di € D

Step 10: Calculate the cosine similarty

Cos-sim = |dx|.|dy|/|dx|*|dy|

Step 11. If (Cos-sim==1)

The two documents are similar

else

The two documents are not similar.

end if

end for

IV. CONCLUSION

The traditional Vector Space model plays important roles in text-based information retrieval. The Clustering system definitely improves efficiency and effectiveness in information retrieval system. It also reduces the search space. The proposed system, clustering the documents and also give users an overview of the contents of a document collection. If a collection is well clustered, we can search the relevant cluster to find the relevant documents. Searching a smaller collection should improve effectiveness and efficiency.

REFERENCES

[1] Gorden D. Weiguo and Pathak Praveen, "Personalization of Search Engine Services for Effective Retrieval and Knowledge Management", Search Engine Services, PP 20-34, 2001.

[2] Salton, G., Buckley, C. (1987) *Term Weighting Approaches in Automatic Text Retrieval*. Information Processing and Management 24 (5), pp. 513-523.

[3] Porter, M.F. (1980) *An algorithm for suffix stripping*. Program 14 (3) pp. 130–137.

[4] Surgey Brin and Lawrence Page, "*The anatomy of a large-scale hyper textual web search engine* ", in Proceedings of the Seventh International World Wide Web Conference , pages 107-117 in 1998.

[5] Gautam Pant, Padmini Srinivasan, and Filippo Menczer, "*Crawling the Web"* Indiana University, Bloomington, IN 47408, USA

[6] Allan Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the World Wide Web. In *Proceedings of the 10th International World Wide Web Conference*, Hong Kong,(2001).

[7] Google search engine. http://www.google.com/, 2004.

[8] Brian D. Davison. Topical locality in the web. In *Proceedings of the 23rd annual international ACMSIGIR conference on research and development in information retrieval*, pages 272–279. ACM Press, 2000.

[9] Monika Henzinger. Hyperlink analysis for the web. *IEEE Internet Computing*, 5(1):45–50, 2001.