IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 12, PP 01-04

Query Mining Based On User Logs

Ameya Ajgaonkar [1], Vijay Gupta [2],Vijay Mistry[3],Virendra Chauhan[4], Prof. Komal Gothwal[5]

[1](IT Department, Atharva College of Engineering/ Mumbai University, India.)
[2](IT Department, Atharva College of Engineering/ Mumbai University, India.)
[3](IT Department, Atharva College of Engineering/ Mumbai University, India.)
[4](IT Department, Atharva College of Engineering/ Mumbai University, India.)
[5](IT Department, Atharva College of Engineering/ Mumbai University, India.)

Abstract:*In General, The Queries Inserted In Search Engines Are Short. They Do Not Provide Sufficient Information For An Effective Selection Of Relevant Documents. For User Requests, There May Be More Unrelated Documents Listed Than Those Listed Once. However, The Most Common Practice Includes Searching For Occurrences Of Query Terms In Documents. In This Study, We Proposed A New Method In Which The Documents Will Be Listed Based On User Records.These User Records Contain The Query Term And The Documents That Were Clicked For The Corresponding Search Term. The Main Idea Is To Extract Correlations Between Query Terms And Document Terms By Analyzing User Records. After Finding The Correlation, This Correlation Between The Query Term And The Document Term Is Used To Expand The User's Queries. With The Help Of This Extended Query, The System Can Now Extract More Relevant Documents*

Keywords*- Correlation ,Extended Query ,Occurrences, Relevant Documents, Sufficient Information.*

I. Introduction

Queries to search engines on the Web are generally short. They do not provide sufficient information for an effective selection of relevant documents. Previous research has proposed the use of query expansion to solve this problem.However, the terms of expansion are determined only in the analysis of the document. In this study, we propose a new method for query expansion based on user interaction information recorded in web query records.

The main idea is to extract correlations between query terms and document terms by analyzing query records. These correlations are then used to select high-quality expansion terms for new queries. Compared to the previous method of expanding the query, our method exploits the opinions of users involved in user registries. Our experimental results will show that the query-based query expansion method can produce much better results than the classic search method and other methods of query expansion.

The People increasingly rely on the Web to meet their diverse information needs. Users often describe their information needs with certain keywords in their queries, which may be different from the terms of the web index.

.As a result, in many cases, the documents returned by search engines are not relevant to the user's information needs. This raises a fundamental problem in terms of mismatching in information retrieval, which is also one of the key factors affecting search engine accuracy.

II. Randomly Directed Exploration

With the rapid growth of e-commerce applications, there is a large accumulation of data in months, not years. Data mining,also known as knowledge. Discovery in database (KDD), to find anomalies, correlations, models and trends to predict the results. The Apriori algorithm is a classic algorithm in data extraction. It is used to extract the frequent elements and the relevant association rules. It is designed to operate in a database that contains many transactions, such as items brought by customers in a store.

It is very important for an effective analysis of market baskets and helps customers to buy their items more easily, increasing market sales .Apriori is an algorithm for the frequent mining of element sets and the learning of association rules on transactional databases. It proceeds by identifying the single frequent elements in the database and extending them to ever larger groups of elements, provided that these sets of elements appear frequently in the database. The sets of frequent elements determined by Apriori can be used to determine association rules that highlight general trends in the database: this has applications in domains such as market basket analysis.

2.2 Algorithm 1

Enter items in Dataset
Check Frequency
Define a Threshold value
IF Frequency is >= Defined Threshold Value
Accept the data items from dataset
Remove the threshold value and again define a new
Threshold value for the remaining items
END IF
Check the final output obtained after applying threshold values
To all the items in the datasets.

2.3 Algorithm Working

Step 1: Create a frequency table of all the items that occur in all the transactions .
Step 2: Assign Threshold Value (Assumption).
Step 3: The next step is to keeping in mind that the order doesn't matter make all the possible pairs of the significant items, i.e. AB is same as BA.
Step 4: We will now count the occurrences of each pair in all the transactions.
Step 5: Again only those item sets are significant which cross the support threshold.
Step 6: Now let's say we would like to look for a set of three items that are purchased together.

2.4 Performance parameters

2.4.1Detection Probability

The algorithm code pseudo is provided below for a T transaction database and a € support threshold. The normal theoretical notation is used, although it is noted that T is a multiple set. C k is the candidate established for level k. At each stage, it is assumed that the algorithm generates the candidate sets from the large element sets of the previous level, based on the closing slogan. count [c] accesses a field in the data structure that represents the candidate set C, which is initially taken as zero. Subsequently, many details are omitted, in general, the most important part of the implementation is the data structure used to store the candidate sets and to count their frequencies

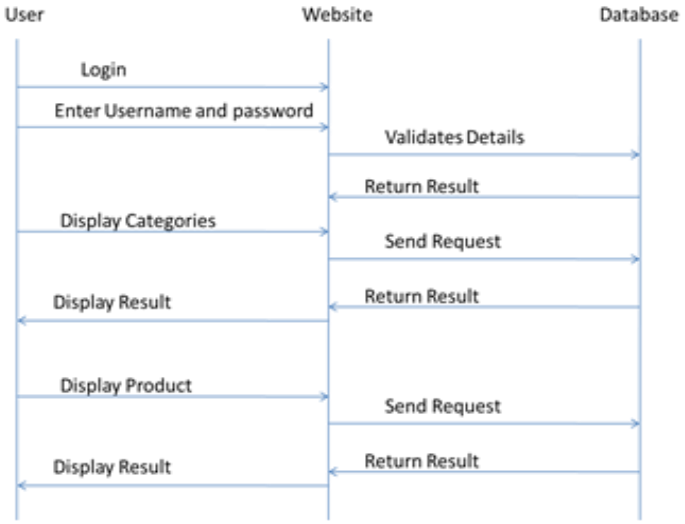

Fig.1 (a) Sequence Diagram

If the sales of a particular product (item) above a certain proportion have a meaningful effect on profits, that proportion can be considered as the support threshold. Furthermore, we can identify item sets that have support values beyond this threshold as significant item sets. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

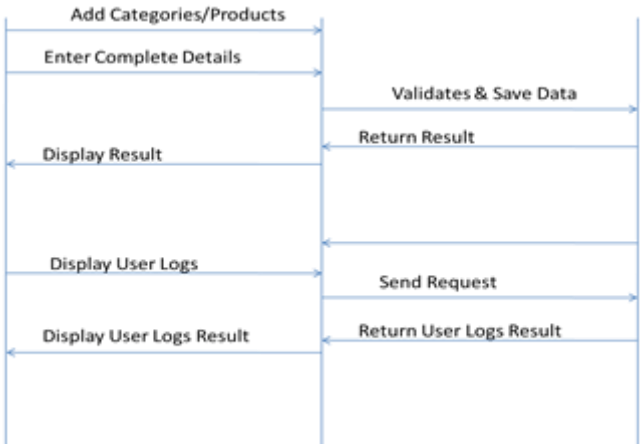

Fig.1 (b) Sequence Diagram

III. Related Works

[1]The Effectiveness of Query Expansion for Distributed Information Retrieval:-

Distributed information retrieval and for single database retrieval where complete collection information is stored query expansion has been effective. One can think that query expansion will then work for distributed information retrieval when complete collection information is not available. However, this does not appear to be the case. The most significant reason query expansion does not work is that merging data of documents retrieved by expanded queries is very tough when using local context analysis for query expansion in distributed retrieval with partial information. It has been found by us that using sampled information for query expansion can give speed up in a single database environment, and that when more information is available, query expansion can also work in distributed environments. For finding good documents, and not from selecting good databases Most of the benefit of Query expansion in distributed retrieval comes.

[2] Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval

Information Retrieval focuses on searching documents whose content matches with a user query from a large document collection. It is know that to formulate well-designed queries might be difficult for most users, it is necessary for us to use query expansion to retrieve the relevant information.For increasing the efficiency of the textual information retrieving systems query expansion techniques are commonly used. We have overcome wording mismatch issues by increasing the original query with an additional related terms and for rewriting the terms in the expanded query special technique are used.

[3]Query Expansion in Information Retrieval Systems using a Bayesian Network-Based Thesaurus

Information Retrieval (IR) is concerned with the identification of documents in a collection that are relevant to a given information need, usually represented as a query containing terms or keywords, which are supposed to be a good description of what the user is looking for. By using a process of query expansion, which automatically adds new terms to the original query posed by an user IR system can improve their effectiveness (i.e., increasing the number of relevant documents retrieved).

IV. Conclusion

1-The proliferation of the World Wide Web prompts the wide application of search engines. However, short queries and the incompatibility between the terms in user queries and documents strongly affect the performance of existing search engines.

2-Automatic query expansion techniques have been introduced, through which we can solve the short query and the term mismatch problem to some extent. They fail to take advantage of the query logs available in various websites, and then uses as a means for query expansion.

3-A method for automatic query expansion based on query logs is introduced by us. This method can establish correlations between document terms and query terms by exploiting the query logs. This is an effective way to narrow the difference between the document space and the query space.

4-High-quality expansion terms can be selected from the document space on the basis of these probability correlations. This method can be tested on a data set that is equally same to the real web environment.

5-Experiments conducted on both long queries and short queries showed that the log-based query expansion method can achieve efficient improvements in performance, not only in the original queries also with respect to

the local context analysis, which has been proved that one of the most effective query expansion methods is in the past. Our method has been proven more useful for short queries than for long queries

References

[1] M.J. Bates, "Search Techniques." Ann. Rev. of Information Science and Technology, M.E. Williams, ed., pp. 139-169, 1981.
[2] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. SIGKDD, pp. 407-416, 2000.
[3] G. Brajnik, S. Mizzaro, and C. Tasso, "Evaluating User Interfaces to Information Retrieval Systems: A Case Study on User Support," Proc. 19th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'96), pp. 128-136, Aug. 1996.
[4] C. Buckley, G. Salton, and J. Allan, "Automatic Retrieval with Locality Information Using Smart," Proc. First Text Retrieval Conf. (TREC-1), pp. 59-72, 1992.
[5] C. Buckley, M. Mitra, J. Walz, and C. Cardie, "Using Clustering and Super concepts within Smart," Proc. Sixth Text Retrieval Conf. (TREC-6), E. Voorhees, ed., pp. 107-124, 1998.