ISSN 2047-3338

Data Mining Techniques and Analysis of Concept Based User Profiles from Search Engine Logs

Prof. N. Prasanna Balaji[1], Chenapaga Ravi[2], P. Krishna Prasad[3] and V. Chandra Prakash[4]
[1,3]Department of Information Technology, Gurunank Engineering College, India
[2,4]TKR College of Engineering Hyderabad, India

***Abstract*– Search engine logs are emerging new type of data user profiling component of any personalization interesting opportunities for data mining. Early user profiling work on mining data mostly attempted to discover knowledge at the level of queries based on objects that users are interested in positive preferences but not the objects in negative preferences. In our paper we focus on search engine logs for patterns at the level of terms and develop many concept user profiling methods that are positive preference and negative preference. We shows that our proposed system in data mining techniques how the search engine works in query process, index process and also comparing the existing system with effective mining association search engine logs.**

***Index Terms*– Data Mining, Search Engine Logs, Query Process, Clustering and Association**

I. INTRODUCTION

INFORMATION explosion on the Internet has placed high demands on search engines. People are far from being satisfied with the performance of the existing search engines, which often return thousands of documents in response to a user query. Many of the returned documents are irrelevant to user's need. The precision of current search engines is well under people's expectations. To find more precise answers to a query a new generation of search engines, question answering systems have appeared on the web [11]. Unlike the traditional search engines that only use keywords to match documents, this new generation of system tries to match documents this new generation of systems tries to understand the users question and suggest some similar questions that other people have often asked and for which the system has the correct answers. In fact the correct answers have been prepared or checked by human editors in most cases then guarantee that if one of the suggested questions is truly similar to that of the user, the answers provided by the system will be relevant. The assumption behind such a system is that many people are interested in the same questions – the frequently asked.

Association rule mining is one of the most important and well researched techniques of data mining, [1]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database.

In spite of the recent advances in the Web search engine technologies; there are still many situations in which the user is presented with non-relevant search results. One of the major reasons for this problem is that Web search engines often have difficulties in forming a concise while precise representation of the user's information need. Most web search engine users are not well trained in organizing and formulating their input queries, which the search engine relies on to find the desired search results. Some studies on Web and peer-to-peer (P2P) queries have been conducted (Chau, Fang, & Yang, 2007; Kwok & Yang, 2004; Yang & Kwok, 2005). On one hand, this is due to the ambiguity that arises in the diversity of language itself; no dialog (discourse) context structure is available for search engines. On the other hand, users are often not clear about the exact terms that best represent their specific information needs. In the worst case, users are even not clear of what exactly their specific information need is. For example, in our study of the sample dataset, one frequently submitted query is " " (download) without specifying what exactly the users are seeking to download.

In order to overcome the problems, some web search engines have implemented methods to suggest alternative queries to users. The purpose of these methods is to help users specify alternative related queries in their search process in order either to clarify their information needs or to rephrase their query formulation to retrieve more related search results. The techniques used in these proprietary commercial systems are usually secure however observe that those suggested queries returned from these search engines are rather similar in their terms.

This may imply that those suggested queries are likely to be generated by using simple query expansion techniques. For instance, if the user searches for Yahoo! search engine the following related queries are presented: messenger, *best yahoo mail*. However, as we can imagine, there are a good number of other queries related to mail but presumably not having the term yahoo explicitly in their term vectors.

II. PROBLEM DEFINITION

Searching some kind information in the web became hectic. Search engine logs record the activities of web users, which reflect the actual general user need or interests when conducting a search. Information in Search engine logs have. Text queries that users submitted the time when they searched, and the URLs that they clicked after the queries. Search engine logs are separated by user sessions. A user session includes several queries from the same user for a coherent information need and the clicked URLs for each query in the session.

A. What is Search Engine

The term "search engine" is often used generically to describe both crawler-based search engines and human-powered directories. These two types of search engines gather their listings in radically different ways. Crawler-based search engines, such as Google, create their listings automatically. They crawl the web, and then people search through what they have found. If there is any change on the web pages, crawler-based search engines eventually find these changes, and that can affect on the listing. Some search engines also mine data available in news books, databases, or open directories. Unlike Web directories, which are maintained by human editors, search engines operate algorithmically or are a mixture of algorithmic and human input.

B. Use of Search Engine

Crawler-based search engines [5] have three major elements. First one is spider or crawler. The spider visits a web page, reads it, and then follows links to other pages within the site. This is what it means when someone refers to a site being "spidered" or "crawled." The spider returns to the site on a regular basis, such as every month or two, to look for changes. Everything the spider finds goes into the second part of the search engine, the index The index, sometimes called the catalog, is like a giant book containing a copy of every web page that the spider finds. If a web page changes, then this book is updated with new information. Sometimes it can take a while for new pages or changes that the spider finds to be added to the index. Thus, a web page may have been "spidered" but not yet "indexed." Until it is indexed added to the index -- it is not available to those searching with the search engine. Search engine software is the third part of a search engine. This is the program that searches through the web pages recorded in the index to find matches to a search and then rank them in order of what it believes is most important.

C. Search Engine Priority for Web Pages

To Search for anything search engine use crawler-based search engine. Nearly instantly, the search engine will sort through the millions of pages it knows about and present you with ones that match your topic [4]. The matches will even be ranked, so that the most relevant ones come first. Unfortunately, search engines [6] don't have the ability to focus the search. They also can't rely on judgment and past experience to rank web pages, in the way humans can. So to determine relevancy they follow a set of rules, known as an algorithm. Exactly how a particular search engine's algorithm works is a closely-kept trade secret.

III. SEARCH ENGINE

A. Query Processing

The architecture of generally consist of three layers (Fig. 1) extracting user sessions from search query logs, segmenting extracted user sessions into query transactions and mining related queries from query transactions

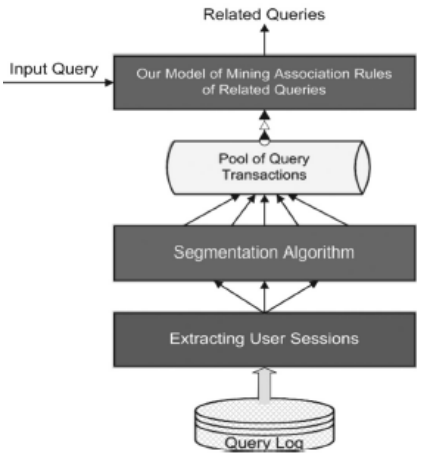

Fig. 1. Search engine log query processing

Extracting User session data from query logs: It extracts user session data from query logs by identifying the query records that belongs to the same user who is organized by a person unique internet protocol address. Segmenting user session data into query transactions: Accepts the extracted user session data generated as inputs and then segments each of them into query transactions properly. The segmented query transactions are then pooled together; user identities and time-stamps of query transactions will no longer be used.

Discovering related queries from query transactions: Initial query submitted by a certain user is directed as the input to the system, if it satisfies some criterion. Query transactions are fetched from the pool and relatedness is calculated between the input query and any other query that satisfies predefined constraints.

B. Index Process

Building data structures that enable searching (Fig. 2):

1) Test Acquisition: Identifying available documents that will be searched using crawling or scanning the web, a corporate intranet. Also the building a document data store containing the test and metadata for all the documents.

Web crawler restricted to a single site supports site search. Types of crawlers; Topic-based crawlers using classification techniques to restrict pages that relevant to specific information; Enterprise document crawler follows links to discover both internal and external pages.

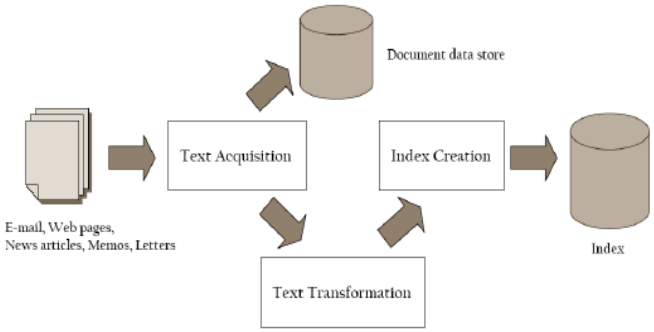

Fig. 2. Process of Search engine in index

2) Text Index Creation: Transforming documents into index terms or features.

Index term is the parts of a document that are stored in the index and used in searching.

Index vocabulary: set of all the terms that are indexed for a document collection.

Gathering and recording statistical information about words features and documents will be used to compute scores of documents e.g., counts of index term occurrences.

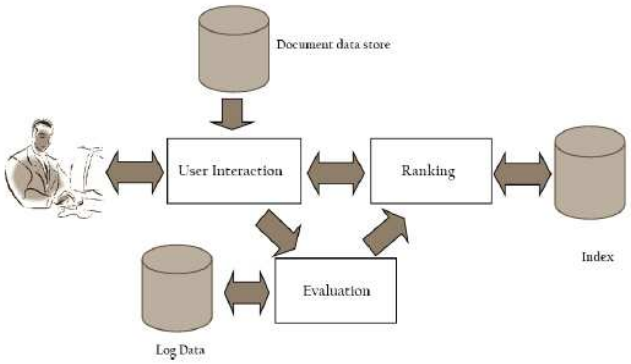

Fig. 3. Process of search engine in query

User Interaction is a providing the interface between the user and search engine, which accepts a user query and transforming it into index terms and also taking the ranked list of documents from the search engine and organizing into results.

Query transforming is to improving the initial query before and after producing a document ranking.

Ranking takes the transformed query from the user interaction component and generating a ranked list of documents using scores based on a retrieval model.

Evaluation monitors the effectiveness and efficiency recording and analyzing user behavior using log data.

IV. DATA MINING FOR SEARCH ENGINE LOGS

Data mining is a stream to apply for the different applications. In our proposed System, we consider that clustering and association rule.

A. Clustering

Clustering is different physical objects are combined similar objects in one cluster and dissimilar objects in other cluster.

1) Search Engine Log Clustering: When we applied cluster technique for search engine consider the Hierarchical clustering.

Hierarchical models [7], [8] propose a better versatility as they do not require any priori definition of the number of clusters to find. It is deterministic in nature and produces the same clustering each time. A perfect visualization of a dendrogram is provided. In hierarchical clustering, the data are not partitioned into a particular cluster in a single step. The rules [4] which govern between which points distances are measured to determine cluster membership include Complete-link (or complete linkage) we merge in each step the two clusters whose merger has the smallest diameter (or the two clusters with the smallest maximum pair-wise distance).

$$dist(C_i, C_j) = \max \{dist(o_i, o_j) \mid o_i \varepsilon C_i , o_j \varepsilon C_j \}$$

Single-link (or single linkage) we merge in each step the two clusters whose two closest members have the smallest distance (or the two clusters with the smallest minimum pairwise distance).

$$dist(C_i, C_j) = \min \{dist(o_i, o_j) \mid o_i \varepsilon C_i , o_j \varepsilon C_j \}$$

Average-link clustering merges in each iteration the pair of clusters with the average distance parameter. $dist(C_i, C_j) = mean\{dist(o_i, o_j) \mid o_i \varepsilon C_i , o_j \varepsilon C_j \}$

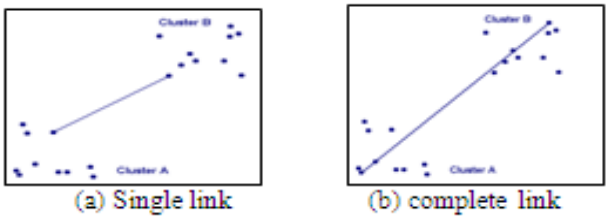

Fig. 4. Rules for membership

Ward's method [8]: This method is distinct from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step. In general, this method is regarded as very efficient; however, it tends to create clusters of small size.

A series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object and the Hierarchical clustering is based on the union between the two nearest clusters. Given a set of N items to be clustered, and an N*N distance (or similarity) matrix. Hierarchical clustering is subdivided into agglomerative methods, which proceed by series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer groupings. For both the hierarchical methods, a hierarchy of tree-like structure is

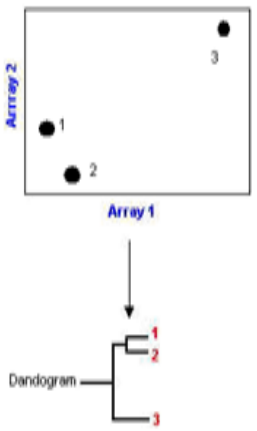

Fig. 5. Dendrogram or tree graph for 2 clusters

constructed which is known as dendrogram or tree graph that is depicted for 2 clusters as shown in Fig. 5:

Hierarchical algorithms like AGNES, DIANA and MONA construct a hierarchy of clusters, with a number of clusters ranging from one to the number of observations. The space complexity for hierarchical algorithm is $O(n^2)$, this is the space required for the adjacency matrix. The space required for the dendrogram is O(Kn), which is much less than $O(n^2)$. The time complexity for hierarchical algorithms is $O(kn^2)$ because there is one iteration for each level in the dendrogram. Depending on the specific algorithm this could actually be O(maxd n^2) where maxd is the maximum distance between the points. Different algorithms may actually merge the closest clusters from the next lowest level or creates new clusters at each level with progressively larger distances.

B. Association

Association model is often used for market basket analysis, which attempts to discover relationships or correlations in a set of items. Market basket analysis is widely used in data analysis for direct marketing, catalog design, and other business decision-making processes.

1) Mining Association Rules: The Concept of mining association rules associates to review the needed information or data Let $I = \{I_1, I_3, \ldots, Im\}$ be a set of binary attributes called items. Let T be a database of transactions. Each transaction can be represented by a binary vector, with $t[k] = 1$ if t bought the item, and $t[k] = 0$ otherwise. Let X be a set of some items in I. A transaction t satisfies X if for all items I_k in X, $t[k] = 1$. By an association rule mean an implication $X \Rightarrow I_j$ where $X \subset I$ and $I_j \notin X$ We define that the association rule has a confidence factor of c if c% of the transactions in T satisfy $\{Ij\}$ given that all transactions in T ($T' \subset T$) satisfy X. We will use the classical notation to specify that the association rule has a confidence factor of c. We also define that the association rule has a *support factor* of s if s% of the transactions in T *satisfy* both X and $\{Ij\}$ at the same time. Note that support should not be confused with confidence while confidence is a measure of the association rule's strength; support corresponds to its statistical significance.

V. COMPARATIVE STUDY

Information retrieval is the technique for searching for the different documents, for the information and metadata about the documents. Early days with the massive growth of the web we assisted to an explosion of information accessible to internet users. Users to explore this huge repository and find needed resources by simply. The biggest problem facing users of web search engines today is the quality of the result get back, one of the important problem is the relationship ranking problem

With so many disadvantages time waste to search, click Expensive process, display unwanted links. Comparing with the early analysis our proposed system work involve a both queries and sessions in order to identify more granular classifications of user intent, web results to the underlying user content need will increase performance of future web search engines. We analyzed a relation-based page rank algorithm to be used in conjunction with semantic web search engine that simply relies on information that could be extracted from user queries, this ranking is called user profile and Google utilizes link to improve search results having more advantages reducing time of search and click expensive process which avoids unwanted links and effectively displays link for best profiles.

VI. CONCLUSION

In this paper, we show the problem of mining search engine logs from the vast amount of data. We defined a term clustering and association patterns and proposed new methods for mining such patterns from search engine logs. Our methods are based on the index process and query process which are mined from a query collection and list out the drawbacks of exiting system with benefits of our system. Few limitations of work First, all the experiments are based on click-through instead of real relevance judgments, so an interesting future work would be to further test the proposed methods with real relevance judgments. Second, building an interactive user interface which can allow a user to modify user queries using our suggested terms can help evaluate our algorithms. Third, search logs have more meaningful click-through information besides queries and sessions.

REFERENCES

[1] Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in large databases. In Proceedings of the ACM Sigmod International Conference Management of Data (SIGMOD'93), pp. 207–216, Washington, DC, May 1993.

[2] Agrawal, R., Imielinski, T., Swami, A. Database Mining: A performance Perspective. *IEEE Trans. Knowledge and Data Engineering*, vol. 5(6), pp. 914-925, 1993.

[3] Beeferman, D., and Berger, A. Agglomerative clustering of a search engine query log. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 407–416, Boston, MA, 2000.

[4] J. Han and M. Kamber, Data Mining: Concepts and Techniques Morgar Kauffmann, pp, 21-25, 2000.

[5] J. Han and M. Kamber, Data Mining: Concepts and Techniques Morgar Kauffmann, pp. 21-25, 2000.

[6] Franklin, Curt. "How Internet Search Engines Work", 2002.
[7] Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan, DEMON-Mining & monitoring evolving data IEEE transactions on knowledge and data engineering, Vol. 13(1), 2001.
[8] Lamia fhalthu Ibrahim, By using of clustering approach for rural network planning, 3rd International Conference on Info Tech, pp. 1-5, 2005.
[9] Brin, S., Motwani, R., & Silverstein, C., Beyond market baskets: Generalizing association rules to correlations. In Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD'97), Tucson, AZ., 1997.
[10] Gillel and, M. Levenshtein. Distance: three flavors. Retrieved from http://www.merriampark.com/ld.htm.
[11] http://www.askgoogle.com/

Prof. N. Prasanna Balaji, and Head IT has done his B.E in Computer science from Bharathidasan University, completed his M.Tech in IT (Part Time) with distinction from Punjab University Patiala, currently pursuing Ph. D in the topic "Enterprise Resource Planning" from Kakatiya University, Warangal.. He has 20+ years of teaching, training and Systems Computerization. Mr. Balaji has worked as Associate Professor in CSE dept at Vignan Institute of technology & Science. At Infosys Campus Connect (two weeks residential December 2006) Programme and was recognized as one of the Best Teacher.

At Institute of Public Enterprise (IPE) he was the ERP-Incharge for Microsoft Business Solutions-Navision, and has organized a National level conference on "e-Customer Relationship Management" and three Management Development Programms in "Recent Trends in Information Technology", two Management Development Programmes in Enterprise Resource Planning-Navission, and one Management Development Programme in Network Security for Public Sector executives. He is the co-editor for the proceedings of National level conference on "e-Customer Relationship Management". He has published and presented papers in National level Seminars and Journals.

His areas of interest are "Enterprise Resource Planning", Relational Database Management Design, Artificial Intelligence, Operating Systems, Mobile Computing, and Customer Relationship Management. He has guided many PG level and engineering students. He is also a member on various professional societies like Life Member of Computer Society of India, Indian Society for Technical Education, and a Member of International Electrical and Electronics Engineers and All India Management Association.

Chenapaga Ravi having 6-years of Academic Experience, currently working as Assoc Prof at TKR College of Engineering Hyderabad, M.Tech from Holy Mary Institute Technology & Science B.Tech from SRTIST College Nalgonda. His areas of research include Wireless Networks, Data Mining and Information Security.

P. Krishna Prasad pursuing M.Tech Information Technology at Gurunank Engineering College, B.Tech Computer Science Engineering from Netaji Institute of Engineering & Technology. His areas of interest include Web Application, Information Retrival System, Wireless Networks, Currently focusing on Data mining.

V. Chandra Prakash pursuing M.Tech Computer Science Engineering at TKR College of Engineering B.Tech from CVR College Computer Science Engineering. His areas of interest include Wireless Networks, Information Security, Web Application, presently focusing on Data mining.