Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 3, March 2014, pg.7 – 14

RESEARCH ARTICLE

The Role of Web Content Mining and Web Usage Mining in Improving Search Result Delivery

Ms. Shital C. Patil[1], Prof. R. R. Keole[2]

[1]H.V.P.M's College of Engg. & Tech,
Amravati University, India
shitalcpatil@gmail.com

[2]Department of Computer Science and Engineering,
H.V.P.M's College of Engg. & Tech,
Amravati University, India
ranjitkeole@gmail.com

Abstract - In today's e-world search engines play a vital role in retrieving and organizing relevant data for various purposes. However, in the real ground relevance of results produced by search engines are still debatable because it returns enormous amount of irrelevant and redundant results. Providing relevant information to user is the primary goal of the website owner. Web mining is ample and powerful research area in which retrieval of relevant information from the web resources in a faster and better manner. Web content mining improves the searching process and provides relevant information by eliminating the redundant and irrelevant contents. However for a broad-topic and ambiguous query, different users may have different search goals when they submit it to a search engine. Web usage mining plays an important role in inferring user search goals as they can be very useful in improving search engine relevance and user experience. The paper focuses on combine approach of web usage mining and web content mining.

Keywords- Search *Engine Result, Information Retrieval, Web Usage Mining, Web Content Mining, Re-ranking*

I. INTRODUCTION

World Wide Web (WWW) is very popular and interactive. It has become an important source of information and services. The web is huge, diverse and dynamic. As the web is growing very rapidly, the users get easily lost in the hyper structure. The primary goal of search engines is to provide relevant information to the users to cater to their needs. Therefore, finding the content if the Web and retrieving the users' interest and needs have become increasingly important.

Web Mining is the application of data mining and information extraction techniques aimed at discovering patterns and knowledge from the Web. This may also be the data related to the Web activity. Web data can be:

- Content of web pages like text and images.
- Intra page structure, which includes the HTML tags or XML tags.
- Inter page structure, which is the linkage structure between Web pages.
- Usage data, which describes how web pages are accessed by various visitors on the internet.

Web mining can be divided into three main subareas:

1.1 Web Content Mining – Web content mining is used to examine the content of Web pages as well as results of Web searching. The content may include text as well as graphics data. Web content mining is further divided into Web page content mining and search results mining. Web page content mining is traditional searching of Web pages with the help of content while search results mining is a further search of pages found from a previous search.

1.2 Web Structure Mining - Web structure mining is done at the hyper link level. This kind of mining tries to discover the model underlying the link structure of the web.

1.3 Web Usage Mining - Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

There is not a clear-cut distinction among these categories, and all three mining tasks can be combined [1].

The paper presents weighted technique to mine the web content catering to the user needs. In the proposed work a new approach is introduced to rank the relevant pages based on the content and keywords rather than keyword and page ranking provided by search engines. Based on the user query, search engine results are retrieved.

Every result is individually analyzed based on keywords and content. If a match is found then particular weight is awarded to each word. Finally, the total relevancy of the particular link against user request is computed by summarizing all the weights of the keyword and content words.

When user clicks the URL out of the re-rank search results list, the contents are extracted from that particular page using NLP technique. These extracted contents, user query and the clicked url are then stored in the server log.
When next time user enters the query, the results are retrieved from search engines and compared with the data saved in the server log and rank the search results accordingly so that users can reach effortlessly what they are looking for.

II. WEB SEARCH ENGINE

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as Search Engine Result Pages (SERPs). The information may be a specialist in web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain Real Time Computing information by running an algorithm on a web crawler.

Figure.1 Various Web Search Engines

A. Traditional Search Engine Architecture:

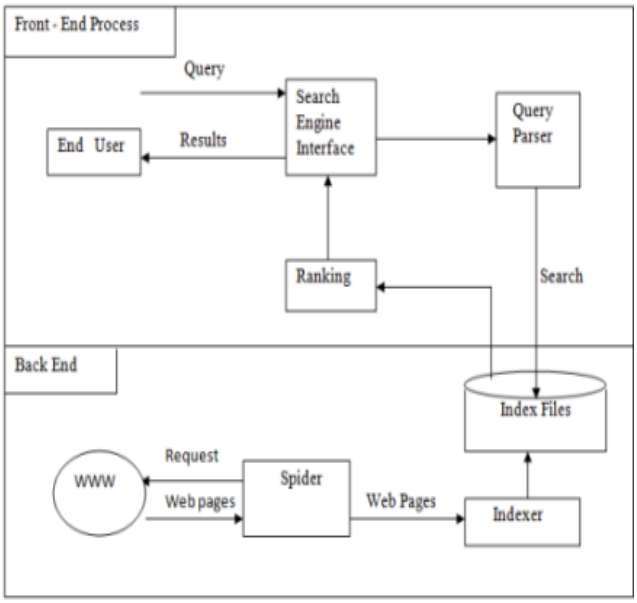

Figure 2. Search Engine Architecture

B. Document Features Which Make a Good Match to a Query

The key features helping to retrieve a good representation of documents/pages are as follows.

• **Term frequency**: How frequently a query term appears in a document is one of the most obvious ways of determining a document's relevance to a query. While most often true, several situations can undermine this premise. First, many words have multiple meanings — they are polysemous. Think of words like "pool" or "fire." Many of the non-relevant documents presented to users result from matching the right word, but with the wrong meaning.

• **Location of terms:** Many search engines give preference to words found in the title or lead paragraph or in the metadata of a document. Some studies show that the location — in which a term occurs in a document or on a page — indicates its significance to the document. Terms occurring in the title of a document or page that match a query term are therefore frequently weighted more heavily than terms occurring in the body of the document. Similarly, query terms occurring in section headings or the first paragraph of a document may be more likely to be relevant.

• **Link analysis:** Web-based search engines have introduced one dramatically different feature for weighting and ranking pages. Link analysis works somewhat like bibliographic citation practices, such as those used by Science Citation Index. Link analysis is based on how well-connected each page is, as defined by Hubs and Authorities, where Hub documents link to large numbers of other pages (out-links), and Authority documents are those referred to by many other pages, or have a high number of "in-links".

• **Popularity:** Google and several other search engines add popularity to link analysis to help determine the relevance or value of pages. Popularity utilizes data on the frequency with which a page is chosen by all users as a means of predicting relevance. While popularity is a good indicator at times, it assumes that the underlying information need remains the same.

• **Date of Publication:** Some search engines assume that the more recent the information is, the more likely that it will be useful or relevant to the user. The engines therefore present results beginning with the most recent to the less current.

• **Length:** While length per se does not necessarily predict relevance, it is a factor when used to compute the relative merit of similar pages. So, in a choice between two documents both containing the same query terms, the document that contains a proportionately higher occurrence of the term relative to the length of the document is assumed more likely to be relevant.

• **Proximity of query terms:** When the terms in a query occur near to each other within a document, it is more likely that the document is relevant to the query than if the terms occur at greater distance. While some search engines do not recognize phrases per se in queries, some search engines clearly rank documents in results higher if the query terms occur adjacent to one another or in closer proximity, as compared to documents in which the terms occur at a distance.

• **Proper nouns:** sometimes have higher weights, since so many searches are performed on people, places, or things. While this may be useful, if the search engine assumes that you are searching for a name instead of the same word as a normal everyday term, then the search results may be peculiarly skewed.

III. LITERATURE REVIEW

Due to the heterogeneity of network resources and the lack of structure of web data, automated discovery of targeted knowledge retrieval mechanism is still facing many research challenges. Moreover, the semi-structured and unstructured nature of web data creates the need for web content mining. In paper [4] the author differentiates web content mining from two different points of view. Information retrieval view and database view. In paper [5] research area of web mining and different categories of web mining are discussed briefly. They also summarized the research works done for unstructured data and semi structured data from information retrieval view.

Effective organization of search results is critical for improving the utility of any search engine. The utility of a search engine is affected by multiple factors. While the primary factor is the soundness of the underlying retrieval model and ranking

function, how to organize and present search results is also a very important factor that can affect the utility of a search engine significantly. Compared with the vast amount of literature on retrieval models, however, there is relatively little research on how to improve the effectiveness of search result organization. The most common strategy of presenting search results is a simple ranked list [2]. Intuitively, such a presentation strategy is reasonable for non-ambiguous, homogeneous search results; in general, it would work well when the search results are good and a user can easily and many relevant documents in the top ranked results. However, when the search results are diverse (e.g., due to ambiguity or multiple aspects of a topic) as is often the case in Web search, the ranked list presentation would not be effective; in such a case, it would be better to group the search results into clusters so that a user can easily navigate into a particular interesting group.

People attempt to infer user goals and intents by predefining some specific classes and performing query classification accordingly. Lee et al. [6] consider user goals as "Navigational" and "Informational" and categorize queries into these two classes. Other works focus on tagging queries with some predefined concepts to improve feature representation of queries. However, since what users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical.

Methods of organizing search results based on text categorization are studied in [7]. In this work, a text classifier is trained using a Web directory and search results are then classified into the predefined categories. The authors designed and studied different category interfaces and they found that category interfaces are more effective than list interfaces. However predefined categories are often too general to reflect the finer granularity aspects of a query.

Clustering search results [8] is an effective way to organize search results, which allows a user to navigate into relevant documents quickly. As a primary alternative strategy for presenting search results, clustering search results has been studied relatively extensively. The general idea in virtually all the existing work is to perform clustering on a set of top- ranked search results to partition the results into natural clusters, which often correspond to different subtopics of the general query topic. A label will be generated to indicate what each cluster is about. A user can then view the labels to decide which cluster to look into. Such a Such a strategy has been shown to be more useful than the simple ranked list presentation in several studies. However, this clustering strategy has two deficiencies which make it not always work well:

i) The clusters discovered in this way do not necessarily correspond to the interesting aspects of a topic from the user's perspective. For example, users are often interested in finding either "phone codes" or "zip codes" when entering the query "area codes." But the clusters discovered by the current methods may partition the results into "local codes" and "international codes." Such clusters would not be very useful for users; even the best cluster would still have a low precision.

ii) The cluster labels generated are not informative enough to allow a user to identify the right cluster.

iii) Since feedback is not considered, many noisy search results that are not clicked by the users may be analysed as well.

Wang and Zhai clustered queries and learned aspects of these similar queries, which solves the problem in part. However, their method does not work if we try to discover user search goals of one single query in the query cluster rather than a cluster of similar queries. For example, in [9], the query "car" is clustered with some other queries, such as "car rental," "used car," "car crash," and "car audio." Thus, the different aspects of the query "car" are able to be learned through their method. However, the query "used car" in the cluster can also have different aspects, which are difficult to be learned by their method.

Some works take user feedback into account and analyze the different clicked URLs of a query in user click-through logs directly. However the number of different clicked URLs of a query may be not big enough to get ideal results.

Web usage mining aims to capture, model, and analyze the behavioral patterns and profiles of users interacting with the Web. Data stored in usage logs can be used for solving navigational problem [10], improving web search [3], recommending queries [11], suggesting authoritative web sites [12], and enhancing performance of search engines [13]. A good survey of web usage mining can be found in [14].

IV. ANALYSIS of PROBLEM

In order to retrieve user requested information, search engine plays a major role for crawling web content on different node and organizing them into result pages so that user can easily select the required information by navigating through the result pages link. This strategy worked well in earlier because, number of resources available for user request is limited. Also, it is feasible to

identify the relevant information directly by the user from the search engine results. When the Internet era increases, sharing of resource also increases and this leads to develop an automated technique to rank each web content resource. Different search engine uses different techniques to rank search results for the user query. Web content mining improves the searching process and provides relevant information by eliminating the redundant and irrelevant contents according to user queries [2] **.** However, sometimes queries may not exactly represent users' specific information needs since many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query. For example [3], when the query "the sun" is submitted to a search engine, some users want to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun. Therefore, it is necessary and potential to capture different user search goals in information retrieval. The inference and analysis of user search goals can have a lot of advantages in improving search engine relevance and user experience.

As the Web's contents grow, it becomes increasingly difficult to manage and classify its information. The high level of competition in the Web makes it necessary for websites to improve their organization in a way that is both automatic and effective, so users can reach effortlessly what they are looking for.

The problems are:

- **Incomplete or Limited Information Problem**: A number of heuristic assumptions are typically made before applying any data mining algorithm; as a result some patterns generated may not be proper or even correct.

- **Incorrect Information problem:** When a web site visitor is lost, the clicks made by this visitor are recorded in the log, and may mislead future recommendations. This becomes more problematic when a website is badly designed and more people end up visiting unsolicited pages, making them seem popular.

- **Persistence Problem:** When a new pages are added to a web site, because they are not visited yet, the recommender system may not recommend them, even though they could be relevant Moreover, the more a page is recommended, the more it may be visited, thus making it look popular and boost its candidacy for future recommendation.

- **Incorrect recommendation:** Since what user cares about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical.

V. PROPOSED METHODOLOGY

Optimization of search-engine performance is obviously of paramount importance, given that a typical search engine receives a huge number of queries every second, and users expect very low response times. The inference and analysis of user search goals can have a lot of advantages in improving search engine relevance and user experience which can be achieved by web usage mining while web content mining removes persistence problem. The paper focuses on combine approach of web usage mining and web content mining.

The paper presents weighted technique to mine the web content catering to the user needs. In the proposed work a new approach is introduced to rank the relevant pages based on the content and keywords rather than keyword and page ranking provided by search engines. Based on the user query, search engine results are retrieved.

Every result is individually analyzed based on keywords and content. If a match is found then particular weight is awarded to each word. Finally, the total relevancy of the particular link against user request is computed by summarizing all the weights of the keyword and content words.

4.1 Mining contents

(1) User Request- User request is processed for search engine to obtain the results.
(2) Top n Results Extraction – Top n results are extracted from search engine based on the user query.

(3) Content Mining - Statistical parameters such as a term frequency (TF) are calculated. For this every result is individually analyzed based on keywords and content. The calculations depend on the user query. Every result of the keywords and content

words are compared by full word matching. If a match is found then particular weight is awarded to each word. Likewise each link is given the final matching score.

(4) Page Reranking- At last, the normalized value of each result is sorted in descending order to get the most relevant content for the user query. Re-ordered results are sent back to the user so that the top most page is more relevant for the user query.

4.2 Log Data

Web usage mining is the application of data mining techniques to the data generated by the interactions of users with web servers. This kind of data, stored in server logs, represents a valuable source of information, which can be exploited to optimize the document-retrieval task, or to better understand, and thus, satisfy user needs.

(1) When user clicks the url out of the re-rank search results list, the contents are extracted from that particular page using NLP technique.

(2) These extracted contents, user query and the clicked URL are then stored in the server log.

(3) When next time user enters the query, the results are retrieved from search engines and compared with the data saved in the server log and rank the search results accordingly so that users can reach effortlessly what they are looking for.

VI. CONCLUSION

In web search applications, queries are submitted to search engines to represent the information needs of users. However, sometimes queries may not exactly represent users' specific information needs since many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query. The proposed system improves the search engine results by inferring user search goals, removing incorrect or limited information problems.

REFERENCES

[1] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.

[2] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-ow graph: model and applications. In CIKM, 2008.

[3] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click-Through," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875-883, 2008.

[4] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.

[5] U. Lee, Z. Liu, and J. Cho, "Automatic Identification of User Goals in Web Search," Proc. 14th Int'l Conf. World Wide Web (WWW '05), pp. 391-400, 2005.

[6] Z. Lu, H. Zha, X. Yang, W. Lin, Z. Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions," Proc. IEEE Transactions on Knowledge and Data Engineering, pp. 502-513, 2013

[7] I. Mele, " Web Usage Mining for Enhancing Search –Result Delivery and Helping Users to Find Interesting Web Content," ACM SIGIR Conf. Research and Development in Information Retrival (SIGIR '13), pp. 765-769, 2013.

[8] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor. Newsl., 1(2):12{23, 2000.

[9] P. Sudhakar, G. Poonkuzhali, R. Kishor Kumar, " Content Based Ranking for Search Engines," Proc. International Multi Conference of Engineers and Computer Scientists (IMECS 12), 2012.

[10] N. Tyagi, A. Solanki and M. Wadhwa, "Analysis of Server Log by Web Usage Mining for Website Improvement," International Journal of Computer Science Issues, Vol.7, issue 4, No 8, pp. 17-20, July 2010

[11] X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.

[12] R. W. White, M. Bilenko, and S. Cucerzan. "Studying the use of popular destinations to enhance web search interaction,". In SIGIR, 2007.

[13] Y. Xie and D. O'Hallaron, "Locality in search engine queries and its implications for caching," In IEEE Infocom 2002, pages 1238{1247, 2002.

[14] O. Zamir and O. Etzioni, " Web Document Clustering: A Feasibility demonstration," ACM (SIGIR, 99) , pp. 46-54.

