Web Usage Mining and Web Content Mining – A Combine Approach for Enhancing Search Result Delivery

Miss. Shital C. Patil, Prof. R. R. Keole
Department of Computer Science and Engineering.
H.V.P.M’s College of Engg. & Tech
Amravati University, India

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 two important issues: improving search-engine performance through dynamic caching of search results, and helping users to find interesting web pages.

Keywords- Search engine result, Query, Web usage mining, web content mining, Reranking.

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. Extraction of interesting information from Web data has become more popular and as a result of that web mining has attracted lot of attention in recent time. Web mining is the process of discovering knowledge, such as patterns and relations, from Web data. Web mining generally has been divided into three main areas [7]: content mining, structure mining and usage mining. Each one of these areas are associated mostly, but not exclusively, to these three predominant types of data found in the Web:

• Content: The real data that the document was designed to give to its users. In general this data consists mainly of text and multimedia.

• Structure: This data describes the organization of the content within the Web. This includes the organization inside a Web page, internal and external links and the website hierarchy.

• Usage: This data describes the use of a website or search engine, reflected in the Web server’s access logs, as well as in logs for specific applications.

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 [9]. 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 [6], 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:

• Restructure web search results [4], [11], [15] according to user search goals by grouping the search results with the same search goal.

• User search goals represented by some keywords can be utilized in query recommendation [1], [3]; thus, the suggested queries can help users to form their queries more precisely.
• The distributions of user search goals can also be useful in applications such as reranking web search results that contain different user search goals.

II. LITERATURE REVIEW and RELATED WORK

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 [9]. 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. [5] 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 [4]. 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 [14] 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 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) 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 [11], 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 [16], improving web search [6], recommending queries [2], suggesting authoritative web sites [12], and enhancing performance of search engines [13]. A good survey of web usage mining can be found in [8].

III. ANALYSIS of PROBLEM

World Wide Web plays a starring role for retrieving user requested information from the web resources. 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.

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 many 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.

### IV. ENHANCING SEARCH RESULT DELIVERY

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 proposed approach is to organize search results by aspect learned from user click through logs. Given an input query the general procedure of the approach is:

1. User Query will be pre-processed to identify the root words.
2. All the feedback sessions (The feedback session is defined as the series of both clicked and unclicked URLs) of the query will be extracted from user click- through logs.
3. Resulted feedback sessions will be mapped to pseudo documents.
4. User search goals will be inferred by clustering these pseudo documents and depicted with some keywords.
5. When any query will be entered for the first time and no matching urls will there in user click- through logs the search results will be displaced by weighted ranking approach in the web content mining.
6. Evaluate the performance of restructuring search results.

### V. 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**


