A PROPOSED QUERY OPTIMIZATION METHOD TO BOOST STUDENTS USE OF SEARCH ENGINES

Amine V. Bitar and Antoine M. Melki
Department of Computer Science
University of Balamand
Lebanon

Abstract

Students, as Internet users, expect quality results in terms of relevance, format and response time, and are more concerned about quality than quantity. On the other hand, the main goal of the search engine system is to retrieve information which is supposed to be relevant to the student's request in the shortest time possible. New techniques in the area of information retrieval by search engines are developed. This paper starts by describing the problems faced by students in their searches. A query optimization method is proposed consisting of extracting some information, including IP address, from the server log file to be used in later searches. This enhancement of search results is expected to improve the relevance of information and consequently the reliance of students on the search engine.

Introduction

The Internet, as an international worldwide resource, has millions of information pages with a wide range and nature of materials. Search has become an essential service required by all students looking for information via the Internet. Search can be defined as an information retrieval activity. A search engine must attract the students by finding and returning relevant information in a short response time. The student translates a request into a query of one or more keywords in order to be processed by the search engine. The main goal of the search engine system is to retrieve information which must be relevant to the student's request in the shortest time possible. Successful search is one where results most relevant are listed first. The vast quantity of Internet information sets new challenge to provide students with the results they consider most relevant listed first.

The objective of this paper is to present a method called Search Booster. The Search Booster allows obtaining the usage data specific to each IP address from the web server log file. The aim is to boost the quality of the search engine performance in order to improve the value of the results returned to the student's keyword search in a short response time. It explores and demonstrates how Internet protocol address (IP) information can enhance current log file techniques to supply valuable data.

This paper is organized as follows: information retrieval through search engines is exposed. Search layers structure is presented. The importance of query optimization in search engines is highlighted and a new method for IP information extraction is proposed. The paper is concluded with experimental validation and some future research work is proposed.

Information Retrieval

Information retrieval deals with the representation, storage, organization and access of information (Baeze-Yates & Ribeiro-Neto, 1999). The full description of the desired information can not be directly used in a search engine to request information. A user must first translate a request into a query which can be submitted to the search engine or information retrieval systems.

To be effective in returning relevant results to a query, the information retrieval system must interpret the contents of the documents by matching them to the student need, and rank them according to the relevance to the student query. The difficulty is not only knowing how to interpret and extract this information, but also knowing how to rank returned results in terms of relevance. The relevance of results is a basic issue in the information retrieval development that most researchers are working on to improve results quality. By definition, the primary goal of the information retrieval system is to retrieve all results relevant to the student query, while minimizing the retrieving of non-relevant results.

Some research dealing with searching as an activity of information retrieval relates the issue of relevance to the study approaches of students. Depending on the approach, students tend to focus on different information seeking aspects in addition to shared commonalities. The research finding showed that students with a surface approach prioritized easily available sources, deep students were aware of quality aspects, and strategic students organized and structured their searches (Heinström, 2006).

The findings of a research on the role of semantics in the formulation of a query were related to the issues of the employment of uncertainty and certainty and the topic and comment. Users initiated the need description with uncertainty and then provided certainty to describe the need in detail. Both topic and comment were used in every stage of information seeking interaction, based on which the source person provided information. The study confirmed that the user’s certainty and uncertainty are important for describing the user's information need and that both topic and comment are essential to communicate the need (Yoon, 2007). The issues of certainty and uncertainty can be related to the formulation of the query not to its evaluation. None of published studies relates certainty to the result selection.

Zhang et al. (2005) worked on a study on the effect of domain knowledge on search behavior and search effectiveness. Interesting results showed that the level of domain knowledge affects both query formulation and results evaluation. The study considers that the users evaluate the results based on relevance and reliability. A user with higher domain knowledge has a higher capability of deciding on the result relevance. In the procedure computer logs were used to save the participants' search history and search results with the intent to keep a record of the searches.

Search Engines Problems

Brooks (2004), in discussing what he termed the Google's age, considers that legacy methods of asserting meaning are inappropriate in the meaning space of the open Web. He urged providing a writing guide as a necessary aid for Web authors who must balance enhancing expression versus the use of technologies that limit the aggregation of their work.

Unfortunately, the characterization of the information need is not a simple problem for students. After submitting the requested query, the key goal of an information retrieval system is to retrieve information which must be relevant to the student query. The student may not express his need in a query in an efficient way. In fact, he is concerned more with retrieving information about a subject of his concern rather than retrieving data which satisfies a given query. For an information retrieval system, an inaccurate retrieved data does not mean a failure and may be unnoticed.

In Internet search, students care about quality not quantity. Quantity without quality turns search engines into a source of great dissatisfaction for many students. Quality results must supply greater depth of information in a particular search relevant to the user's query in order to achieve a high level of reliability and accuracy. The wide area of information available on the web, added to the easy way of information access, has attracted tremendous attention from millions of people everywhere since the early beginning. Despite this success and development, the web has introduced new problems in the field of information retrieval. Finding useful information in the web is a difficult task. A student might search through returned results and navigate through web links to satisfy an information need, but the problem gets more complicated. The main obstacle is the absence of a well defined data model for the web. These difficulties have set to develop new techniques in the information retrieval research to provide new solutions to return high quality results.

The model of searching for information on the web as used by many existing search engines does not meet the needs of many students. Students need at least some understanding of basic Internet concepts in order to carry out successful searches. Few students are impressed with the quality of information on the Internet. While doing their search, students expect information to be returned in a clear way. They assume that anything returned in a search less than a perfect match is a failure and that the Internet does not have relevant information, when in fact their queries are answerable. A search engine may return no results because a valid query correctly fails to find relevant results, or because the query was invalid.

A number of studies tackled the issue of query failure (Jansen et al., 1998). In considering students as users, Baeza-Yates and Ribeiro-Neto (1999) relate query failure to many facts:

- Students have difficulties formulating keywords: This is the main problem. Most students have difficulties formulating good searching keywords or can't translate what they know into a successful query even when they had all the information they need. Students may not understand that searching often requires refinement of search queries to have good search results.

- Students fail to specify their requests: Students may use terms of broader or narrower meaning than what they intend. For example, a student looking for "Reebok Trainers" might search for "Sport Shoes" rather than for "Reebok."

- Students have spelling difficulties: Students may misspell a word while searching for a document without any notice. Reasons for spellings errors include unfamiliarity with the technical terms, difficulty in spelling, or simply hitting the wrong key.

- Students don't fully understand the search interface: Not all search interfaces are the same. Pages with multiple search forms or submit bottoms may confuse. Moreover, many sites use a search form with more than one input box. Students may not understand the different functions or roles of the different inputs.

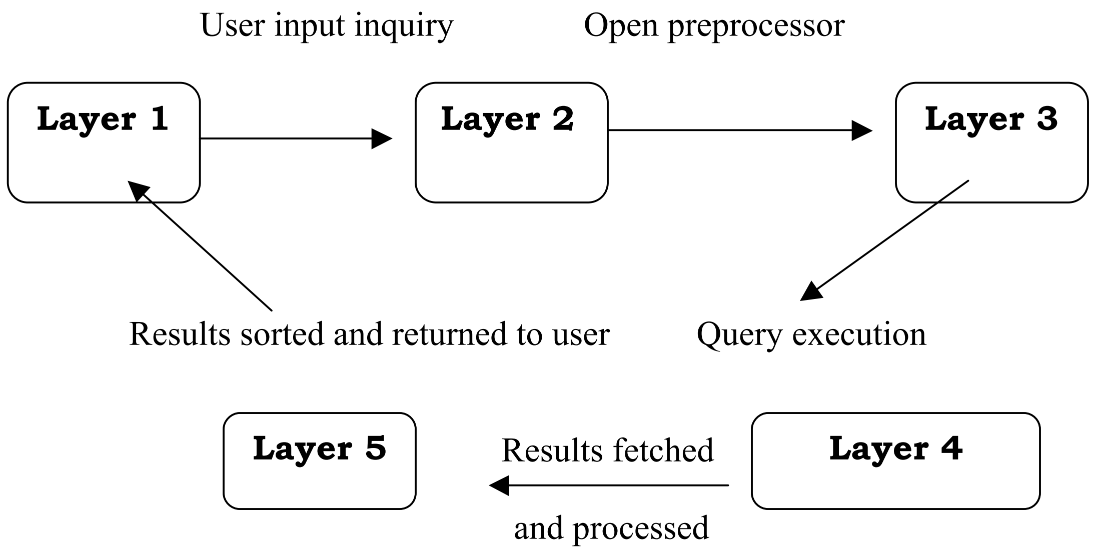
- Most students can't use advanced search or Boolean query syntax: Based on recent study, the mean query length was 2 words. Boolean syntax is a search allowing the inclusion or exclusion of documents containing certain words through the use of operators such as AND, NOT and OR. Students are used to write a single-word query or very

> short multi-word query in the search edit box when they search for information. Boolean search leads them into trouble, and they invariably use it wrongly.

Search Layers Structure

Technically, a multi-layer search structure exists to process any search query. Baeza-Yates and Ribeiro-Neto describe (1999) this structure in the following diagram.

Figure 1: Search layers diagram



Layer 1: Called Search or User Interface. It is the first layer in the search structure. A search form provides an input field, built on the top of an application, used by users to enter query and get results back in return. It is the layer of creation and submission of an appropriate query.

Layer 2: Query Preprocessor. It is the mechanism where modification is performed by the search system before running the query to increase the chance of successful results. The search system may modify, change, or applies Boolean logic (AND, OR) to the query, so the format will be more suited to the collection architecture.

Layer 3: Query Execution. It is the core work of the search. In this layer, the information is fetched and the results are returned. After the query is modified, it runs against an index built from the collection of information in the database. The structure and organization of the collection, and the logic and strategy used to build an index affects how the search query is performed.

Layer 4: Results Processor. After the query is executed, raw results are returned to be processed to the user. Some logic may be applied to the raw results before being processed. This layer involves applying logic to the results to optimize their value for the user.

Layer 5: Results Page. This is the final layer in the search diagram which provides mechanism for re-sorting the results, ways to specify the kind of information shown, a search form that allows the user to re-search from the results page. After the raw results are processed, they are passed and presented to the user, sorted according to alphabetical or numerical order, ordered by a logical or contextual grouping scheme, and ranked according some scale of priority.

Query Optimization

Martzoukou (2004), in a review of Web information seeking research, found that comprehensive studies are limited with many problematic approaches and the absence of a consistent methodological framework. Research has often failed to ensure appropriate samples that ensure both quantitative validity and qualitative consistency because observation has been based on simulated rather than real information needs. Martzoukou sees that the research findings are not so reliable as they tackle the various aspects of cognitive style and ability with variant definitions of expertise and different layers of user experience, and they don’t extensively investigate the effect of social and cultural elements. The existing limitations in method and the plethora of different approaches allow little progress and fewer comparisons across studies. While the review highlights an urgent need for establishing a theoretical framework on which future studies can be based so that information seeking behavior can be more holistically understood, and results can be generalized, Martzoukou falls in the same field on concentrating on the issue of the query formulation and its results evaluation.

Jansen (1990), in a study on the effect of query complexity on searching results, states that his goal is finding the probability of finding relevant information by increasing the complexity of their queries. In his experiment, he used selected fifteen queries from the transaction log of a major Web search service in simple query form with no advanced operators (e.g., Boolean operators, phrase operators, etc.) and submitted to 5 major search engines — Alta Vista, Excite, FAST Search, Infoseek, and Northern Light. The procedure included a series of query modifications using the various search operators supported by each of the five search engines, then an application to the search service. The conclusion reached was that increasing the complexity of the queries had little effect on the search’ results, especially that relevance evaluation was not made concerning the results.

A query, as the formulation of the student information need, is composed of a single word or combination of words which are used to search for documents containing such keywords. Some systems expand the keyword to its synonyms in order to return useful information to student.

Query optimization forms a very large area within the database field that has been studied from different views. It is the process of choosing the most efficient way to execute a statement. In order to improve the performance of queries, it is important to understand the techniques used by the query optimizer to select an access path and the query execution to return the relevant results.

The goal of every website is to increase site visibility and users traffic: “One way to increase site traffic is through search engine optimization” (Patton, 2006). The importance of search engine optimization is to maintain the website at a high technical and content level in order to bring the users qualified results from the search engine: “Search engine optimization is the science of search as it relates to marketing on the web” (Clay, 1996). It is the combination of programming with business, sales and the competition to achieve high rankings in the search engine results page.

Optimization is an important factor for search engines. It increases qualified traffic, leads to more perceptive site navigation, and fast response time to enhance the student experience. For a search engine, it is not always possible to maintain the top ten rankings since results change due to competition. The goal of search engine optimization is to maintain a top ranking using the keywords, monitoring and analyzing ranking results in order to provide the users with updated and valuable results. Search engine optimization involves optimizing the titles, meta tags, the internal and external links of the site, monitoring the results achieved, and keeping track of the user behavior and of the ranking algorithm changes.

A number of studies are going on user modeling techniques based on different issues including relevance estimation (Krömer et al., 2007). The goal of these studies is to provide experimental results in web search framework with evolutionary query optimization. The following proposed solution comes in this direction taking into consideration the issues described in this paper.

Proposed Solution

An interesting study by Huang et al. (2007) on Web users’ behavioral concentration investigates the heterogeneity in Web users’ online information behavior. Their measurement is based on the number of sites visited, the number of page-views per site and the duration per page in order to predict online information behavioral concentration. Despite the interesting finding of this research, relating the issue of concentration to that of relevance needs to be re-examined to find some empirical support, if any exists.

The studies of Spink and Jack (2000) on the Excite search engine users evaluated the relevance feedback. One of those studies showed that there is a variance in the use of relevance feedback and one third of Excite users going beyond the single query, with a smaller group using either query modification or relevance feedback, or viewing more than the first page of results. This study supports our assumption on the effect of relevance on users' utilization of search engines.

Different reports have discussed the ability of students to do successful search and the problems they face. Reports concluded that the majority of students know how to use the technology, but they don't know how to apply it to find what they are looking for (Appel, 2006). Most students do not have the skills to find information (Nielson, 2006). There is a lack of understanding of search and how to use advanced search functions, query reformulations, and choose the top search results without judgment.

Different servers have different log formats. Important information can be derived from the transaction log, such as the IP address of the user, the user name from the DNS lookup, the timestamp as seen by the server, the request made by the user, the visited URL, and the bytes transferred. In some cases, it is impossible for the DNS lookup to get the username, so the log would simply contain the IP address. Therefore, it is better to extract the information based on the IP address rather than the username.

Many studies have been focused on information organizing and locating using Internet protocol address. This method, Search Booster, is proposed to extract the information specific to each IP address from the server log file and usage data aiming to boost the quality of the results returned to the user's keyword search in a short response time. It explores how Internet protocol address (IP) information can enhance current log file techniques by yielding valuable data on information searched by users, especially when couples with other relevance information.

In search engine, a list of related web sites is returned in response to a keyword query. In a new search, the keyword will be checked in the log file for best results previously used by the user. In order to fasten the search, users are categorized according to their IP address range in which their computers are located or associated. The benefit of classifying users in this way is to categorize them according to the subjects related to their interest field of search. The allocation of the computer to a sub-network is fixed. Therefore, the sub-network information provides a very good fix of a client computer. The URL's of the returned information will be associated with the IP address of the user in a log file record.

Conclusion and Future Work

The emphasis in this proposed solution is on increasing students' success in the first attempt so to avoid the typical judgments they make about a website's value based on the quality of the search results. Search results should be returned to users as quickly as possible with the most relevant first.

A run on a preliminary experiment on a list of 10 keywords and a set of 10 students showed a significant improvement when the student uses the same IP. The available data is not enough to jump to any conclusion. Future work will concentrate on the ways to handle the case when the student uses a different IP range. One of the possible solutions is to create a history file, other than the log, in which each keyword searched is stored with the list of all its successful returned hits. When searching using a different IP range for the same keyword, results will be checked first in the history file for possible successful entries. A more elaborated module to handle relevance information also should be provided.

References

Appel, J. (2006). Students struggle with information literacy. Retrieved April 13, 2008, from http://www.eschoolnews.com/news/showStoryts.cfm?ArticleID=6725.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. England: Addison-Wesley.

Brooks, T. A. (2004). The nature of meaning in the Age of Google. *Information Research*, *9*(3), paper 180. Retrieved April 13, 2008, from http://InformationR.net /ir/9-3/paper180.html

Clay, B. (1996). *Search engine optimization*. Retrieved April 13, 2008, from http://www.bruceclay.com/web_rank.htm.

Gillenson, M. (2005). *Fundamentals of database management systems*. Hoboken, NJ: John Wiley & Sons.

Heinström, J. (2006). Fast surfing for availability or deep diving into quality — Motivation and information seeking among middle and high school students. *Information Research, 11*(4), paper 265. Retrieved April 13, 2008, from http://InformationR.net/ir/11-4/paper265.html

Huang, C-Y., Shen, Y-C., Chiang, I-P., &Lin, C-S. (2007). Concentration of Web users' online information behaviour. *Information Research*, *12*(4) paper 324. Retrieved April 13, 2008, from http://InformationR.net/ir/12-4/paper324.html

Jansen, B. J., Spink, A., & Saracevic, A. (1990). Failure analysis in query construction: Data and analysis from a large sample of web queries. *Proceedings of the 3rd ACM International Conference on Digital Libraries*, June 23–26, 1998, Pittsburgh, PA, pp. 289–290.

Krömer, P., Snášel, V., Platoš, J., & Owais, S. (2007). Implicit user modelling for query optimization. *Proceedings of the 18th International Workshop on Database and*

Expert Systems Applications. Retrieved April 13, 2008, from http://ieeexplore.ieee.org/iel5/4312838/4312839/04312872.pdf

Martzoukou, K. (2004). A review of Web information seeking research: Considerations of method and foci of interest. *Information Research*, *10*(2), paper 215. Retrieved April 13, 2008, from http://InformationR.net/ir/10-2/paper215.html

Nielson, J. (2006). College students can't use search engines. Retrieved April 13, 2008, from http://www.sitelogicmarketing.com/blog/11-college-students-cant-use-search-engines.

Patton, T. (2006). Use sitemap standards to help search engines. Retrieved April 13, 2008, from http://articles.techrepublic.com.com.

Spink, A., & Xu, J. L. (2000). Selected results from a large study of Web searching: The Excite study. *Information Research*, *6*(1). Retrieved April 13, 2008, from http://InformationR.net/ir/6-1/paper90.html

Yoon, K. (2007). A study of interpersonal information seeking: The role of topic and comment in the articulation of certainty and uncertainty of information need. *Information Research*, *12*(2), paper 304. Retrieved April 13, 2008, from http://InformationR.net/ir/12-2/paper304.html

Zhang, X., Anghelescu, H. G. B., & Yuan, X. (2005). Domain knowledge, search behaviour, and search effectiveness of engineering and science students: An exploratory study. *Information Research*, *10*(2), paper 217. Retrieved April 13, 2008, from http://InformationR.net/ir/10-2/paper217.html