Factors Affecting Information Retrieval
Rajesh Kumar Goutam\textsuperscript{1}, Dr. Sanjay K. Dwivedi\textsuperscript{2}
\textsuperscript{1} BBAU, Lucknow, India
Email: rajeshgoutam82@gmail.com
\textsuperscript{2} BBAU, Lucknow, India
Email: skd200@yahoo.com

Abstract—Web is a largest information repository containing web pages, continuously expanding with time thus effectively and efficiently searching relevant information on the web is being a challenge. In recent years large numbers of search engines have been introduced which are still dealing with few problems of indexing and retrieval of relevant information. Therefore it becomes essential to evaluate the performance of available search engines. It is mandatory to have a prior knowledge of factors affecting the search engine performance before its evaluation. In this paper, we attempted to identify those factors by taking a sample set of 25 TREC queries. We also explored that the HTML code has an important role in changing the web page ranking which affects the performance of the search engines.

Index Terms—Information retrieval, Search engine performance, Search Length.

I. INTRODUCTION

Search engines are essential tools for information retrieval on the web. However, it is observed that majority of them do not give up to the mark performance. Due to the rapid expansion of web, lack of structure and the uncountable number of users, search engines frequently face formidable performance related challenges. Many different techniques for evaluation of search engines performance have been proposed but nobody attained satisfactory results. The available measurement techniques and algorithms are not efficient in highly interactive end user system. It is observed that lots of web search engines have been introduced by the various companies like Google, AltaVista, Yahoo etc. struggling with performance related issues and companies also tried to search out for better solutions but still satisfactory results are not achieved.

The basic aim of this paper is to identify the factors that affect the performance of the search engines. We had taken a sample set of 25 TREC queries. That are posed practically to five search engines and 25 results of each query were studied and evaluated carefully. The affect of various factors on the performance of the search engines is studied.

II. EVALUATION PARAMETERS

The search engine performance revolves around the Response time, Total number of results and its ordering, Quality of Results and User efforts. These are the main criteria to differentiate the performance of one search engine from others.

Three levels of the evaluation parameters hierarchy have been shown in Figure 1.

A. Response Time

This is the time period that starts from the point of query submission till user gets a huge list of responded results. Response time is directly related to activeness of search engines and kept minimum as much as possible.

B. Total number of results and its ordering

To find the relevant result it is essential to cover the whole area of web and select the best among them. User does not prefer for huge list of results. Rather his interest is to get most relevant results on the first page of result set retrieved. The results that completely fulfill users requirements are called ‘relevant results’. Sometimes search engines response for duplicate links, such types of problems affect the search engine performance in direct way.

C. Quality of Result

Result Quality is important parameter to justify the search engines quality. Sometimes search engine retrieves results with lots of junk information, which often generates the confusion for end users during results selection. Every search engine is expected to provide only relevant results and it should not include junk material.

D. User Effort

The search engine which minimizes the “user efforts” in order to perform search, is considered to be the better one. End users are interested in getting the results on a single click. Nobody is interested in waiting for long time for their results. User wants to pay minimum efforts to query formulation process. A better search engine should be capable of understanding the user’s interests. In this highly competitive scenario, search engines are searching new ways to automate the ranking system for the documents because everyday huge numbers of documents are getting uploaded. It is not always possible to sort out all the documents with the help of statistical methods. Automatically judging the effectiveness of information...
retrieval systems, based on previous users behavior hold guarantee for making retrieval evaluation quicker, cheaper and more user focused. The advantages are clear unlike human evaluators judgments, usage data can be captured essentially at zero cost, it is existing in real time, and reflects the decisions of the users, not those of judges holds the experience in their area.

III. FACTORS AFFECTING SEARCH ENGINE PERFORMANCE

Web is continuously changing the way people seek information. One should be aware of searching trends that are being used. We identified and examined following factors which affect the information retrieval process directly or indirectly.

1. Query formulation
2. User feedback to a webpage
3. W3 rules
4. Web Developer’s Fake Techniques

A. Query formulation

Query formulation is an essential part of successful information retrieval [3]. A ‘query’ is collection of one or more searching terms which normally includes logical operators and modifiers. “Term” is basically string of characters without any space. Information searchers seldom employ advance searching techniques, such as query phrase searching and Boolean operators [1]. Keily [6] published a study utilizing queries from two Web search engines, “WebCrawler” and “Magellan”, and concluded that only 12% from the 2000 queries contained Boolean operators. Hoelscher [4] presented data and analysis from a German Web Information retrieval system “Fireball” with approximately 16 million queries and declared approximately 3% queries consisting Boolean operators, and 8% contained phrase searching. Jansen, et al. [5] conducted a study concerning searching with “Excite” search engine. In this study approximately 8.5% queries contained Boolean operators. Hoelscher and Strube [2] examined web searchers and reported information on sessions, terms and queries, emphasized that experts adopt different searching patterns than novices.

B. User feedback to a webpage

Implicit user feedback enables a search engine to judge the relevancy level of available web pages on the web efficiently, at zero cost. This technique collects real time data and promises to do evaluation faster. The potential advantages are clear unlike expert judgments, usage data can be collected at zero cost, it is available in real time, and it reflects the action of the users and their intention [7, 8]. This technique utilizes the users behavior within single webpage such as save, copy, print, add to favorite, hyperlinks clicked, the amount of scrolling and mouse activity, to judge the relevance level. But in explicit user feedback method, expert judgments are required to fix the relevancy status to a webpage that is challenging task because to hire experts is too expensive to collect judgments for search engine evaluation process. Implicit users feedback technique is given preferences because of its advantages over several other techniques.

Search Length

Search length is a measurement by which user efforts can be measured in searching process. It can be defined as the number of links the user has to go through to obtain relevant documents. In our experiment we select five search engines and 25 TREC queries. The mean for search length measure is displayed in Table 1. When this study was conducted, Google significantly outperformed among five search engines. It is observed that user has to go through the few pages to find the relevant documents.

Broken Links

A web link that doesn't response is called a broken link or dead link. A link may become dead for several reasons: The result of a dead link is actually 404 error, which indicates that the web server is alive and responding, but the page could not be found on the web. ‘Broken-link’ could be an indication towards time in which a search engine checks the links in its own database for currency thoroughly. In an experiment we select 25 TREC queries and check out top 65 returned links for each query, the percentage of broken links are shown in Table 2.

C. W3 Rules

There are millions of live websites running on, to advertise business and organization around the world. To compete with competitors, it is essential to improve the traffic potential of your website to be on-line marketer. Web designer is important person that is partially responsible to hike ranking of any website by applying W3 rules during the website design that have its own importance in web environment. These rules also play a major role because all search engines takes W3 rules of designing in consideration during ranking /indexing the documents.

Some vital factors that play important role during searching to calculate the relevancy of web document are listed below:
• Powerful content
• Excellent use of ‘Cascading Style Sheet’
• Validated HTML coding
• Appropriate use of hyperlinks
• Meta Tags (Title, Keywords, Alt, Description )
• Use of site map or RSS feed
• Domain name selection and its age.
• Avoid broken links

We utilized 25 different TREC queries and Submitted to Google, yahoo, AltaVista, MSN, Excite. To carry out experiment we use ‘Dreamviwer’ software to validate the pages. Findings show that all the five search engines results into the websites where pages incorporate Title Tag, Meta Tag, Abstract, Description, and Keywords matching with the query terms. This indicates towards the fact that HTML coding plays important role in information retrieval process of search engines. We investigated first 25 results with reference to each query and examined their source code. We evaluated that almost all top ranked documents are built with strong HTML coding part, powerful content and contain all the properly arranged tags as mentioned above. The top ranked documents are also without broken links, URL submission and domain age is also important. The results of real time testing for HTML coding are summarized in Table 3.

D. Web Developer’s Fake Techniques

To earn money web developer use following tactics to get a place in top 10 places.
• Use of Fake keywords while page content does not match with keywords at all.
• Huge advertisement of any particular website to www so that number of people can hit in order to make that page popular.
• Fake title
• Use of PPC (Pay Per Click)
• Words on black back ground
• Frequently submission of particular URL

It is always better for a good page to be higher in the ranked list because most of searchers visit only top ranked pages, unfortunately some web developers interfere with their fake techniques in ranking system of search engines and get success to promote their website in top ranking.

IV. CONCLUSION

The paper discusses about the three level evaluation parameters hierarchy and also describes the importance of each one during performance evaluation of search engines. We emphasized on various factors that directly or indirectly affect the performance of search engines. Implicit user behavior technique is considered most effective because of its cost factor and real time data collection. We highlighted HTML page related issues and its importance in w3 environment. We also discussed the fake techniques that are being used by web developers to hike ranking of their websites so that large numbers of people may visit the website.

ACKNOWLEDGMENT

The authors wish to thank to Prof. B. Hanumaiah, Vice-Chancellor, Babasaheb Bimrao Ambedkar University (A Central University) Lucknow for providing the excellent facilities in the computing lab of B. B. Ambedkar University, Lucknow, India.

REFERENCES