Optimizing performance of search engines based on user behavior

Dr Jkr Sastry 1 *, M. Sri Harsha Vamsi 1, R. Srinivas 1, G. Yeshwanth 1

1 Professor Department of Electronics and Computer Science Engineering, KLEF
*Corresponding author E-mail: dksastry@kluniversity.in

Abstract

WEB clients use the WEB for searching the content that they are looking for through inputting keywords or snippets as input to the search engines. Search Engines follows a process to collect the content and provide the same as output in terms of URL links. One can observe that only 20% of the outputted URLs are of use to the end user. 80% of output is unnecessarily surfed leading to wastage of money and time. Customers have surfing characteristics which can be collected as the user keep surfing. The search process can be made efficient by including the user characteristics / Behaviors as part and parcel of search process. This paper is aimed at improving the search process through integration of the user behavior in indexing and ranking the web pages.

Keywords: Search Engine; Information; Retrieval; User behavior Search engine.

1. Introduction

Information decimations is being done using WEB. Infor-mation is made available to users in just few seconds through search engines. Many search engines have been introduced quite recently and all of these almost behave in similar fashion. The search engines follow a process, starting from accepting snippets from end user and then looking for the URLs at which the content has the snippet words. A crawl is a moving agent that is made to visit the WEB sites based on the Meta directory maintained for each of the WEB site. The Crawler when finds an URL having the desired content writes the URLs to a database as it moves. The databases of URLs are indexed based on the snippet words using which the URLs are retrieved. The URLs in the indexed database are then ranked based on some criteria as recommended by a ranking algorithm. The URLs are as per the ranking in the as-cending order displayed for the end user. It becomes the responsi-bility of the end user to visit the URLs reported to find where the content for which they are looking is resident. A kind of mini search is required within the list of URLs displayed. At times the users land up to read the content to find if it is of use to them The search process described above leads to 20% of the effec-tive URLs needed by the user and the remaining URLs are of no use. Effort expended for fetching and processing remaining 80% of the URLs is a waste leading to too many wastages including inef-fec-tive use of bandwidth.

Search engine optimization is achieved through adding more aspects to the search process. Many methods to improve the performance of the search engine all of which aim at reducing the URLs to be displayed to the end users are not quite effec-tive.

The process of searching for the content must be optimized considering the user behavior so that the URLs that are exactly required can be fetched and displayed to the user. Search engine optimization will improve the page profile by natural results. It helps in fulfilling the user needs and improves rank position of the URLs, which are fetched by the search engines.

User behavior can be captured in many ways and one of the most important method is to capture search tends / history of the user concerned by capturing the negation characteristics and filter the fetched URLs based on the history either before indexing or after ranking. This paper presents a method of capturing the user behavior and also using the same as an additional process in-built within the search process.

2. Related work

Google Scholar is one of the search engines that is quite frequently used for providing search on publications. The algorithm being used by them has not been published. A reverse engineering technique [1] has been found to determine the kind of indexing and ranking process used by Google to develop Google scholar search engine. It has been found that Citation count of the papers published is being used for ranking the publications. Highly cites papers are more often found when a search is made. It has been found that the snippet words influence more in the search com-pared to number of citations.

With the advent of internet more and more information is being hosted on the WEB. While some information hosted on the WEB is useful and correct, some information is useless. Sometimes incorrect information is being hosted on the WEB. No control as such exists in this case. Search engines generally fetch the information required and also not required as the search is generally undertaken using snippet words. It has been found that considering user behavior profile / behavior as a part of search process will fetch exact information required by the users. Many ideas have been floated [2] that can be considered for building the user be-havior within the search process The traditional algorithms that have been in existence for web searching and caching have not been quite effective especially when the speed of information addition to the WEB is quite rapid and high. Clicking through data analysis which is an inverted file replacement algorithm has been presented [3] which do efficient web caching. A new cache policy has been used
which is based on position arrival model. It has been found that the retrieved information organised as inverted file which enhance the speed of searching quite rapidly.

Mining algorithms are being used for web searching. A deep extraction tools have been presented that uses clustering technique for web searching [4]. The presentation by the authors is limited to information related to researchers and scientists. A group of in- formation is identified as a cluster and the users will make a choice on the cluster using which the content encapsulated into the cluster will be displayed.

Most of the algorithms meant for doing web searching are based on the number of back links that a site has. The algorithm involves in safe building of the links with each of the link weighted with velocity No compromise on page optimisation as such is made. One of the important strategies is to rank a page on a keyword [5]. Every organisation can find the keywords and link the URLs/pages to the keyword and rank the pages bases on the fre- quency of usage of a keyword for searching. The techniques elim- inate the sandboxing of the search into already available search engines.

A tutorial has been made present on the WEB [6] that identifies the partially pages for analyzing the system. It is expressed that keywords must be recognized considering the customer require- ments and the competitiors choices and design the WEB pages that are in- bult with keywords. A search engine uses the keywords for indexing and ranking.

Finding the exact required information from WEB is tough due to existence of extensive information on the WEB. Search engine op- timization has become enormously important for this reason [7]. Search engine optimization involves analyzing the WEB data from the perspective of different users and handling the WEB based on the analysis results. Web logs contain information to certain extent logging the information related to different queries that have been processed. The info log includes the times stamps, URLS pro- cessed, User identification etc. The user’s experiences can be ana- lyzed through analysis of query logs / web logs. The history con- tained within the web logs can be used effectively for understand- ing the user behavior and accordingly uses the same for query opti- mization. The history contained in the WEB log can be grouped dynamically and in an automated manner. The Groups can be then used for optimization through query alteration, re- ranking, query replacement etc. A method [7] has been proposed that links the query groups with URLS that are related over general information needs. It has been proposed to combine word similarity- ty measures with document similarity measures and form into a combined similarity measure. Other measures also are considered that include query reformulation, clicked URLS etc.

Search engines are being used for querying the information re- quired by web sufferers. The users of the WEB require only the top most of the results they require. The web sites designed are more bothered about promoting the WEB site for access by search en- gines. Search engine optimisation has become very important and sometimes the techniques used might break the rules and regula- tions followed by the search engines. A user for instance might be clicking on the same URL several times misleading the rule that the URLS with high clicked count be ranked on the Top. To avoid this Ranking Algorithm that uses the IP address of the users to track the clicking on the URLS has been presented [8].

Trend related queries can be found when users interact with the WEB. A context search engine considers query trends which can be traced from different types of domains. The requirements of the users can be represented as a series of queries based on the search intentions of the user. The context search engines helps in provid- ing search results as required by the end users [9].

Web mining is actually carried for providing the query results. Both web structure mining and web content mining are usually carried for web searching. Page ranking algorithms are used for web structure mining and some algorithms that include HITS and page content ranking are used for mining both WEB structure and WEB con- tent. A new method has been proposed based on the weighted pages and content ranking which uses all web mining techniques (structure, content, usage mining) [10].

Search engines are the only means available to locate and make available the information required. A combination of manually ed- ited directories, automated algorithms and advertisements on the WEB for generating search results. A new algorithm [11] has been presented that implements a modified page rank algorithm that al- locates weights to the in-linked web pages. The weights are distrib- uted to all outbound pages based on the popularity of those pages. This kind of algorithm is called Weightage in-link rank algo- rithm. The algorithm calculates score of every individual web page and the score is used for ranking.

WEB logs provide a source for analyzing the user behavior while transaction on site especially the e-commerce sites. Data mining techniques can be applied to WEB log data for revealing interesting patterns. The user’s behavior as such is modelled using web log in most static way. The sequence of operations carried by the user generally is dynamic and the static data is not good enough to depict dynamic behavior of the users. Capturing the behavior of the users while interacting through the WEB site in terms of the process fol- lowed is more complex and interesting.

A linear-temporal logic model checking approach has been pre- sented for analyzing the e-commerce web logs [12]. The web log records if can be related to event logs dynamic behaviors of the user can be traced.

Driving traffic on the WEB sites has become quite complicated as the completion is increasing to find the data related to an estab- lishment to be right on the top of search results. It is usual that the in- ternet users navigate through the pages which are on the top of the list. Indexing and ranking of the searched pages are usually done, to order the fetched pages in ascending of their ranking. Some of the WEB site owners apply search engine optimisation techniques for optimising the content to ensure that their pages are reflected on top of the search results. Serval methods [13] have been presented in the literature for optimising the search process. Google SEO On- page and Off-page techniques are one such tech- nique that can be used for search engine optimisation. The per- formance of a search engine can be computed by using the SEERP metric.

3. Investigations and findings

The search process that is generally followed includes a web crawler which is made to visit every web site and find the pages if it contains the desired content by the users. The URLs fetched are stored in a database as it crawls various WEB sites. One optimization technique used is to match the key words entered by the users with the meta words stored within a web site. Only when the key words and Meta words matches, the content checking is done fetching the URLS of the documents that has the content matching the keyword entered. All the URLS that are fetched are indexed into the database using the Key words. The URLS are ranked based on the importance of the URLS by following some criteria, for instance, the clicked count on the URLS.

Many algorithms exist in literature for ranking the URLS means the WEB pages. The URLS are weighted based on some criteria. Page rank algorithm has been selected for this work as it weighs the URLS based on the number of clicks made by the user on its hyper- link. Clicked counts on each of the URLS are processed us- ing the WEB log. The URLS and the related clicked count are fetched by the WEB Clawer. The process followed is shown in Figure 1 The searching is done in the lines of above mentioned process using sample key words and just browsing through a Known WEB site.
4. Proposed algorithms

The user behavior is recognized through the navigation paths to the elementary level that the user navigates rather than the number of clicks that the users makes in the normal courses or erratically in the abnormal course. The rank of a navigation path is fixed based on the number of users who navigates to the same path.

The process flow followed for the revised search process is shown in Figure 2.

Fig. 2: Revised Search Flows.

The experimentation on the Known WEB site using the above process has been done and experimental results are shown in the following Tables. Table 4 shows the keywords used and table 5 shows the URLs fetched along with the user counts. Table 6 shows the indexed URLs along with page ranks derived out of user counts on the navigation paths.

Table 2 shows the fetched URLs with clicked counts.

Table 3: Indexed URLs with Clicked Counts and Page Ranks

Table 4: Search Process Outcomes

Table 5: Fetched URLs with Clicked Count

The URLs are displayed to the end user in the ascending order of the page ranks. The user behavior in this case is recognized through the clicked count which sometimes gives wrong results when the users behave in erratic manner clicking on the same link several times without any use or purpose.
From Table 8 it can be seen number of operations to be carried on WEB LOGS get reduced resulting in saving in time of search processing and also the amount of space required to store the in dexed will also get reduced reasonably. More than these the ranks confirm to the usual navigation that is normally followed by the users.

6. Conclusions

Page ranking algorithm uses the clicked counts as the user behavior which does take the reality of the user surfing as there could be some abnormality exhibited by the user in the normal course. User behavior as well be exhibited through number of users clicking on the same path, The process presented follows the normal course that happen when the users starts surfing the WEB sites are initi ates a search process for want of information.

References