A Survey on Effective Query Processing Techniques in Web Search Engines

B. Senthil Kumar
Associate Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India

Asha Unnikrishnan
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India

Abstract – A large information repository employ the effective search techniques to retrieve the required data based on user query. There are several popular search engines like Google, Yahoo, ASK and AOL etc providing such information retrieval in different ways from the large repositories. Data retrieval based on the user query. Query processing is an important task of every search engine, where it need many process to retrieve optimal results, which includes query extraction, preprocessing, crawling, indexing, ranking and so on. This paper gives an overview of such process, challenges, issues and different domain specific search engines. And finally the paper describes the problem of search engine process in terms of delay, energy, and optimization etc.

Index Terms – Web Mining, Web Structure Mining, Web Search Engine, Query Processing, Data Indexing, Crawler

1. INTRODUCTION

Search engines are too mandatory in the current scenario, which is an information retrieval tool. The search engine performs the query processing, data retrieval process from the large data repositories. It returns the results and list of pages which matches for the user queries [1]. Web search engines are continuously crawling and index a numerous count of web documents and gives fresh results to the user. The process of search engine has different types of issues in terms of quality, energy and time. In this paper, we analyzed the search engine process and reviewed various techniques and methods used to make effective search engine. The information gathered by crawlers is used to create a searchable index of the Net. In general, search engines maintain very large databases that contain information about the numerous web pages. They are automatically updated by crawlers that search the WWW for new content and then report their findings to the database. The search engines can be categorized into three types like crawler based, directory based, which is known by human powered, hybrid search engines [2], the table 1.0 show the basic comparison of the three types of search engines.

Search engine Type	Search engine	Advantages	Disadvantages
Crawler -Based	Google	• They contain a huge amount of pages. • Ease of use.	• Information overload problem. • Data manipulation issues • Page rank manipulation
Directory/ human empowered	Open Directory Project and the Internet Public Library	• Each page is reviewed for relevance and content before being included. This means no more surprise porn sites. • Quick retrieval	• Unfamiliar design and format. • Delay in creation of a website and its inclusion in the directory. • May have trouble with more obscure searches.
Hybrid	MSN	• Quick retrieval	• Human results are usually listed first. • Quality of the result is not good

Table 1.0 different type of search engines and its comparison

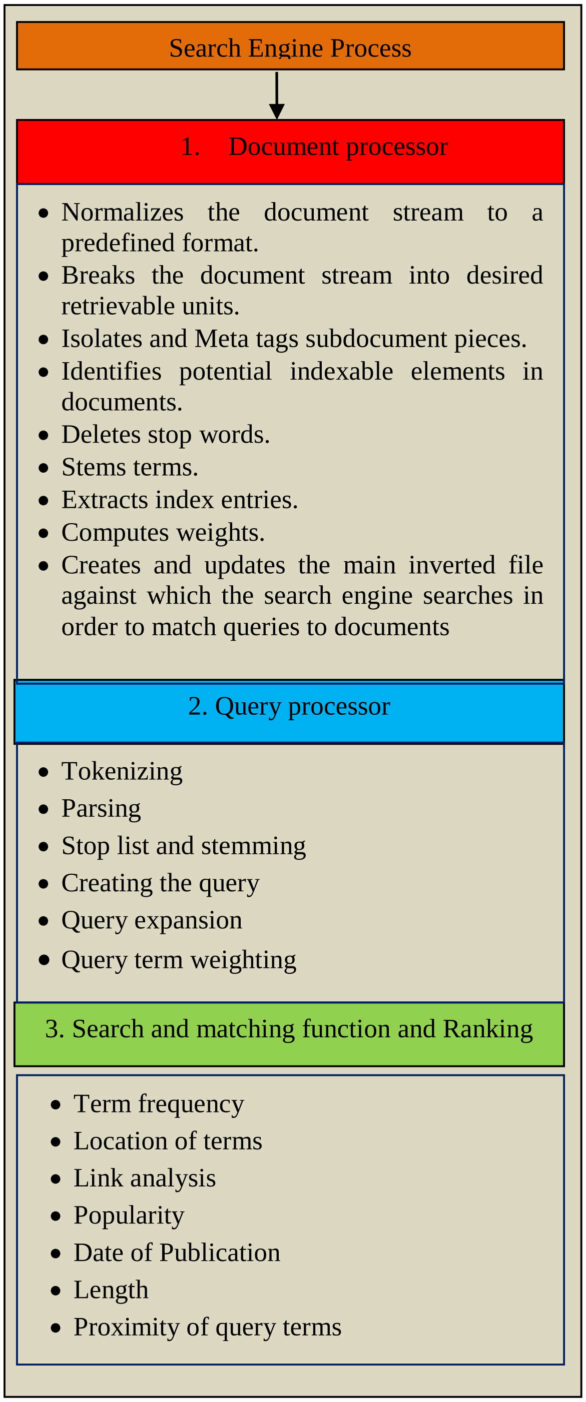
Search Engine Process
1. Document processor
Normalizes the document stream to a predefined format.
Breaks the document stream into desired retrievable units.
Isolates and Meta tags subdocument pieces.
Identifies potential indexable elements in documents.
Deletes stop words.
Stems terms.
Extracts index entries.
Computes weights.
Creates and updates the main inverted file against which the search engine searches in order to match queries to documents
2. Query processor
Tokenizing
Parsing
Stop list and stemming
Creating the query
Query expansion
Query term weighting
3. Search and matching function and Ranking
Term frequency
Location of terms
Link analysis
Popularity
Date of Publication
Length
Proximity of query terms

Fig 2.0 different phases of search engine

A. Crawler-Based Search Engines:

Crawler-based search engines, such as Google, create their listings automatically. They “crawl” or "spider" the web, then people search through what they have found. If web pages are changed, crawler-based search engines eventually find these changes, and that can affect how they are listed. Page titles, body copy and other elements all play a role. The crawler based search engines are easy to use and it contains huge sized data. The disadvantages of this kind of search engines are easy to manipulate the search result and page rank [3].

B. Directories / Human-Powered search engines:

A human-powered directory, such as the Open Directory, depends on humans for its listings. A short description is submitted to the directory for entire site, or editors write one for the sites they review. A search looks for matches only in the descriptions submitted. Changing the web pages has no effect on the listing. Things that are useful for improving a listing with a search engine have nothing to do with improving a listing in a directory.

C. Hybrid Search Engines:

In the initial decade of search engines many users used a search engine either crawler based results or human powered listings, but later its extremely general for both types of results to be presented, so it is known as hybrid search engines. Usually, a hybrid search engine will favor one type of listings over another. The MSN Search is the best example, where it is more likely to present human powered listings from LookSmart And it does also present crawler-based results especially for more obscure queries.

Search engines match queries against an index that they create. The index consists of the words in each document and pointers to their locations within the documents. This is called an inverted file. A search engine or IR system comprises four essential phases shown in fig 2.0. While users focus on "search," the search and matching function is only one of the three phases. Each of these three phases may cause the expected or unexpected results that consumers get when they use a search engine.

2. LITERATURE SURVEY

Many search engines have been developed and commercially implemented, however, only few researches has been done on the domain specific search engines, and very few determined and considered the other factors of search engines like performance, and retrieval efficiency in terms of time and accuracy of web search engines. This section provides the methods and techniques proposed for effective search engine process.

In paper [5] authors in his study proposed a new frontier of search entitled and mentioned that in a traditional search engine interaction scenario, a user begins with a certain concept and finds documents that are similar to their concept. However, the user may wish to compare alternatives and a search capability should compare concepts and present the best alternatives. This task can be difficult without proper decision aids. They propose a concept comparison engine as a decision support tool that may be used to compare attributes of different alternatives and aid in making an informed selection. Authors describe architecture, an interaction scenario and implemented a prototype. They also elaborated a number of evaluation metrics for measuring the viability of different terms for the purpose of comparing concepts. Finally they concluded that in scripted experiments, orderings for candidate terms from the prototype are compared to gold standard ranking lists from structured external sources. Our results indicate that a Rank or analysis may be promising as a measure of the differentiating power of candidate terms a user might choose to support concept comparison.

Authors in [6] analyzed the performance of Natural Language (NL) in search engines in retrieving exact answers to the NL queries differs from that of keyword searching search engines. natural language queries were posed to Google and three NL search engines: Ask.com, Hakia and Bing. The first results pages were compared in terms of retrieving exact answer documents and whether they were at the top of the retrieved results, and the precision of exact answer and relevant documents. Ask.com retrieved exact answer document descriptions at the top of the results list in 60 percent of searches, which was better than the other search engines, but the mean value of the number of exact answer top list documents for three NL search engines was a little less than Google’s . There was no significant difference between the precision for Google and three NL search engines in retrieving exact answer documents for NL queries.

Authors in [7] made an attempt to study students' use of search engines for information retrieval on the web in Adeyemi College of Education, Ondo, the authors found out that majority of the respondents (63.12%) had no specific place for their online search; they used their mobile phones / laptop everywhere to search the internet. Only a very few of respondents (3.55%) used virtual library for their online search, many of the respondents (39.01%) used the search engine occasionally and majority of students (71.63%) used just one or two search engines on regular basis. The authors finally concluded that students should be enlightened on the importance of online resource for their academic success to propel them to use search engines.

Authors in [8] proposed a conceptual model and research issues’ proposed an evaluation method for search engines by developing a conceptual model based on the literature. Authors mentioned that model identifies the key factors that influence user evaluation of search engines, effective and efficient criteria for evaluation by considering user satisfaction and usage as the search engine success variables. They also elaborated that the model attempts to identify the attributes that determine a good search engine, why users repeatedly visit their favourite search engines, and why users switch between different search engines. The authors finally concluded that the relevance of the results with utility plays a crucial role in revisiting the search engine by the users. The research issues are evolved out the conceptual model and the implications for searchers and search engine providers are given.

In paper [9] authors made a study on search engines and published it with the title ‘Search engines: Left side quality versus right side profits’. Authors stated that Search engines face an interesting trade off in choosing the way to display their results. While providing high quality unpaid, or “left side” results attracts users, doing so can also cannibalize the revenue that comes from paid ads on the “right side”. The present paper examines this tradeoff, focusing, in particular, on the role of users' post-search interaction with the websites whose links are displayed. Authors also elaborated on the model, high quality left side results boost demand from users, causing them to tolerate a search engine on which advertisers do not offer the lowest possible prices for the goods that they sell. Finally Authors concluded that websites appearing on the left side still have an incentive to compete in the same market as advertisers, an increase in quality on the left side may reduce advertisers' equilibrium prices. Author analyzed the circumstances under which this will occur and discuss the model's potential implications for antitrust policy.

Authors in [10] in their paper entitled, ‘The Role of Search Engine Optimization on Keeping the User on the Site’ mentioned that 93% of internet traffic is managed by search engines, exploring the potential of search engines is crucial, it shows the critical role of search engines on routing users to the right websites. Due to the important effects of search engines, search results are getting more crucial for websites to compete with other rivals. They also mentioned that the most important part of defeating other rivals is optimization of search engines, after this optimization, website owners expect that the search engine results display their website first, before other websites. The authors finally concluded that the study is to scientifically justify the importance of search engines and search engine optimization (SEO). The author reveals some surprising results that the main focus was to measure the significance of time, speed, reduced bounce rate, page views, and page layout in keeping the user on the site.

Authors in [11] compared the competition between an inferior search engine and a superior search engine with the option to introduce a knowledge-sharing service. The author focuses on the pure strategy, Nash equilibrium of the competition between inferior and superior search engines attempting to maximize their either profits or market shares. If one search engine introduces a knowledge-sharing service, it decides whether to make its answer database accessible by the other competing search engine. They also elaborated on the compatibility decision of each search engine is shown to be significantly influenced by whether it maximizes its profit or market share. The superior search engine should keep its answer database closed to maximize its market share, but may make its answer database open to maximize its profit unless the amount of information available on the Internet is small. The inferior search engine should keep its answer database open to maximize its market share if its search technology is far behind that of the superior search engine. Both the inferior and superior search engines should make their answer databases open to maximize their profits if the amount of information available on the Internet is large [12]. Finally authors revealed some surprising results that equilibrium strategies for inferior and superior search engines depend primarily on the amount of information available on the Internet, the degree of searchers' patience to wait for answers, and the search quality difference.

A. Energy reduction Techniques in Query processing:

While Web search engines can consume huge electric power to operate the query and to get the results, there is only a limited body of research that aims to reduce the energy expenditure of Web search engines. These works can be divided in three categories which focus on different level of a Web search engine architecture: 1) geographically distributed datacenters, 2) processing clusters within a datacenter, and 3) a single query processing node. The works in the literature focused on multi-site Web search engines, where the search engines composed by multiple and geographically distant datacenters. These studies propose to use *query forwarding*, the query forwarding techniques are to shift the query workload between datacenters.

Authors in [13] consider a scenario where datacenters hold the same replica of the inverted index. Authors propose to use query forwarding to exploit the difference in energy price at different websites due to the different datacenter locations and time zones. Using this technique authors aim to minimize the energy expenditure of the search engine. However, the approach ensures that the remote sites can process forwarded queries without exceeding their processing capacity.

Blanco et al. [14] extend this idea by forwarding queries towards datacenters that can use *renewable energy sources* that are both environmentally friendly and economically convenient. Teymorian et al. [15], instead, consider a scenario where each site holds a different inverted index. In their approach, the authors use query forwarding to maximize the quality of search results, collecting relevant document from the different sites, while satisfying energy cost budget constraints. Query forwarding techniques may be applied in conjunction with PESOS to deploy more energy-efficient architectures.

3. PROBLEM DEFINITION

Getting desired results from the search engines is the common expectation of every user in the internet. Due to the massive growth of web contents, data search is not fully optimized. This created many internal issues and challenges. Internet users use text based queries as a request to seek information using any search engine. Search engine then tries to infer and retrieve the relevant documents by performing the matching of query to the surrogates of documents and present the likely relevant documents to users in the form of hits list. Search engine performs query processing process by performing document indexing methods. So effective indexing, query processing with energy minimization is an important attribute of the future research. And moreover, the search engines are common and very few developed with domain specific features. So this will be interesting to have a search engine for biomedical and healthcare domains. Because the query processing may always creates some problem when it is domain specific. User may not have proper knowledge about the health related queries. So this will be an interesting idea for the further research.

4. CONCLUSION

This paper presented several Related Literature dealt with the search engine concepts. The survey on Web Searches shows every author performed different types of searching algorithms to improve the efficiency. Google appears to be dominating the other search engines with its advanced features, performance at query processing and retrieval efficiency. The recent studies also show that Google is superior for its coverage and accessibility. From this survey we concluded that there is a need to develop a domain specific search engine with reliable query suggestion, expansion and optimization features.

REFERENCES

[1] Catena, Matteo, and Nicola Tonellotto. "Energy-Efficient Query Processing in Web Search Engines." *IEEE Transactions on Knowledge and Data Engineering* 29.7 (2017): 1412-1425.

[2] NainB, S., and H. Lall. "Deep Web Data Scraper: Search Engine." (2014): 52-56.

[3] Song, Yang, et al. "Context-aware web search abandonment prediction." *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014.

[4] Zhang, Zhiyong, and Olfa Nasraoui. "Mining search engine query logs for query recommendation." *Proceedings of the 15th international conference on World Wide Web*. ACM, 2006.

[5] Abrahams, Alan S., and Reza Barkhi. "Concept comparison engines: A new frontier of search." *Decision Support Systems* 54, no. 2 (2013): 904-918.

[6] Habernal, Ivan, and Miloslav KonopíK. "SWSNL: semantic web search using natural language." *Expert Systems with Applications* 40, no. 9 (2013): 3649-3664.

[7] Jato, Michael, and Jamogha Oresiri. "Students' Use of Search Engines for Information Retrieval on the Web: A Case Study of Adeyemi College of Education, Ondo." *Greener Journal of Internet, Information and Communication Systems* 1.2 (2013): 55-60.

[8] Palanisamy, Ramaraj. "Evaluation of search engines: a conceptual model and research issues." *International Journal of Business and Management* 8, no. 6 (2013): 1.

[9] White, Alexander. "Search engines: Left side quality versus right side profits." *International Journal of Industrial Organization* 31, no. 6 (2013): 690-701.

[10] Egri, Gokhan, and Coskun Bayrak. "The role of search engine optimization on keeping the user on the site." *Procedia Computer Science* 36 (2014): 335-342.

[11] Kim, Kihoon, and Edison Tse. "Search engine competition with a knowledge-sharing service." *Decision support systems* 66 (2014): 180-195.

[12] Bifet Figuerol, Albert Carles, Carlos Castillo, Paul-Alexandru Chirita, and Ingmar Weber. "An analysis of factors used in search engine ranking." (2005).

[13] E. Kayaaslan, B. B. Cambazoglu, R. Blanco, F. P. Junqueira, and C. Aykanat, “Energy-price-driven query processing in multicenter web search engines,” in *Proc. SIGIR*, 2011, pp. 983–992.

[14] R. Blanco, M. Catena, and N. Tonellotto, “Exploiting green energy to reduce the operational costs of multi-center web search engines,” in *Proc. WWW*, 2016, pp. 1237–1247.

[15] Teymorian, O. Frieder, and M. A. Maloof, “Rank-energy selective query forwarding for distributed search systems,” in *Proc. CIKM*, 2013, pp. 389–398.