IJCSC Volume 4 • Number 2 September 2013 pp.184-190 ISSN-0973-7391

A Framework to Evaluate Search Engines

Deepak Rathee, Jaideep Atri, Jagdeep Rathee, Rajender Nath
Department of Computer Science & Applications, Kurukshetra University Kurukshetra, Haryana, India
deepakrathee5@gmail.com, jds094@gmail.com, rathee.jagdeep@gmail.com, rnath2k3@gmail.com

Abstract- A search engine is designed to search information on World Wide Web and FTP server. URLs are presented as search engine result. URLs consist of web pages, images, information and other type of file. The main goal of this document is to analyse the efficiency of different search engines and ascertain the best one and the options that can be used to enhance search results. Twenty search engines are analysed experimentally based on various parameter like relevant search, irrelevant search, total resultant document, time taken for search, current status of search engine, registration required, advertisement availability, efficiency, language support, support to semantic based query variance and the redundancy in search results.
Keywords: Relevant search, Search Engines, Semantic based query, variance and redundancy.

1. INTRODUCTION

The World Wide Web (WWW) now a day is a store house of huge amount of information but none of the search engine can explore more than 16 % of available web data. Though million of web pages are available but the result of each search is only few web pages. For a search engine it's very difficult to decide which web pages are most important for user or in other words it is very difficult for search engine to decide which web page is relevant to the intent of the user. Some times for a search engine it is very important to maintain semantic relation between web pages so that user can get information that is relevant [1] [2]. Here different search engines are discussed and then analysis is done based on some parameters like relevant search, irrelevant search, total resultant document, time taken for search, current status of search engine, registration required advertisement availability, efficiency, language support, support to semantic based query variance and the redundancy in the results of search. Relevancy parameter defines how much user's requirement is being fulfilled with search results. Total resultant URLs refers to the total number of web pages being shown by the search engine as a result of the query entered by the user and the time span to produce the resultant URLs is named time taken.

Advertisement availability shows whether the search results carry advertisement as well. Language support provides a mode to a user to opt desired language Semantic based queries refer to Boolean queries and thus are important for search engine to optimize the results [1] [3]. Variance factor is the deviation from the similar results that are being produced by other search engines or in term of mathematics it is the mean of total number of results by different search engines. Redundancy is an important factor as the duplicate search results mislead the user and responsible for extra time and efforts, which is defined as

$$\text{Redundancy} = \frac{\text{Total number of results shown}}{\text{Total number of distinct results}}$$

Based on all above factors, the performance analysis of different search engines can be done. Rest of the paper is organized as follows section II discusses the related work, section III presents various popular search engine, section IV presents the analysis and results of the experiment conducted in order to evaluate the search engines based on their performance, Section V discusses the results of the experiment conducted in the section IV. Section V concludes the paper.

2. RELATED WORK

The explosive growth of the Internet has rendered the World Wide Web as the primary tool for information retrieval today. For information retrieval purposes on the Web, most people use search engines like Google, Yahoo and MSN. Sullivan and Griesbaum assert that Google is the dominant and most successful search engine among its contemporaries. For assessing the performance of these search engines, there are various measures such as database coverage, query response time, user effort and retrieval effectiveness, that can be used [9]. The dynamic nature of the Web also brings some more performance measure concerns regarding index freshness and availability of the Web pages as time passes [10]. The most common effectiveness measures are precision (ratio of retrieved relevant documents to the total number of retrieved documents) and recall (ratio of retrieved relevant documents to the total number of relevant documents in the database) [10] Measuring the search engine effectiveness is expensive due to

the human labour involved in judging relevancy. This has been reported in several studies. Evaluation of search engines may need to be done often due to changing needs of users or the dynamic nature of search engines (for example, their changing Web coverage and ranking technology) and therefore it needs to be efficient [9].Per Thousand Precision approach was formulated and used for computing the precision of the search engines. Since we based our human judgment on the first 1000 top documents retrieved for the relevant documents retrieved by the search engines, we therefore adopted this approach, although the first top 20 documents are commonly used [14]. There are two types of search engine evaluation approaches in the literature: testimonial and shootout Testimonials are casual studies and state the general impression obtained after executing a few queries. Shootouts are rigorous studies and follow the information retrieval measures for evaluation purposes. The Jansen and Pooch [13] study provides an extensive review and analysis of current Web searching results. It provides a rich reference list and proposes a framework for future research on Web searching. In addition, the Gordon and Pathak [11] study, which measures the performance of eight search engines using 33 information needs also recommends seven features to maximize the accuracy and informative content of such studies (see Table 1 below). Hawkins et al., [12] also proposes some more features in addition to the seven items specified in Gordon and Pathak [11]. In another study, A set of URLs were collected ("X" topics, "Y" searches, "Z" search engines, "R" results). For each search, duplicates (where the same URL was returned by two different search engines) were removed, leaving with unique search-URL pairs. The search-URL pairs corresponding to each topic were then given to the student concerned, in an Excel file in which the search terms and the URL appeared in consecutive columns (with the URL appearing as a clickable link to the corresponding site). The student had to evaluate the document indicated by the URL, without knowing which search engine it came from, and report the following information in supplementary columns:
Dead link (1 if the site does not respond, otherwise 0), Pornographic link (1 if the link points to a pornographic site, otherwise 0), Topic (regardless of the quality of the information, 1 if the document is on-topic, otherwise 0)Commercial site (1 if the link points to an e-commerce site, otherwise 0) Relevance (grade from 0 to 5, with 0 being a that is totally useless or off-topic, and 5 being a document that provides a perfect answer to the question posed) Conclusion reached to the point stating proportion of commercial links in resultant URLs is high, varying between 7 and 16% depending on the search engine. In itself, the presence of commercial links does not necessarily have a negative impact on quality: for a search like "Harry Potter", returning the page on Amazon where the book can be purchased may be relevant. However, as things stand, we can see a clear degradation of the results in terms of perceived relevance for commercial links, for all search engines alike.
Literature shows that several studies have been carried out by researchers in line with all these features with some modifications and / or enhancements For instance, Griesbaum [15] compares the performances of three German search engines – Altavista, Google and Lycos – in terms of relevance and precision using 50 queries. Results from his study shows that Google reached the highest values, followed by Lycos and then AltaVista. Hananzita and Kiran [16] also compare the performances of four Malaysian web search engines using Google as the benchmark search engine. Their results also show that Google outperforms the four engines considered.

3. EVALUATION OF MAJOR SEARCH ENGINES

In this paper, the most commonly used 20 search engines have been analyzed based on queries from different domains. There are many search engines developed so far, but the twenty most widely used have been selected for this study. Each of these search engines have been studied based on different parameters like relevant search, irrelevant search, total resultant document, time taken for search, current status of search engine, registration required advertisement availability, efficiency, language support, support to semantic based query variance and the redundancy in search results. In the ensuring paragraphs each of these twenty search engine is described in brief:
Google: The order of search results on Google's search-results pages is based on a priority rank called a "PageRank". Google Search provides many options for customized search, using Boolean operators such as: exclusion ("-xx"), alternatives ("xx OR yy"), and wildcards ("x * x"). The main purpose of Google Search is to hunt for text in publicly accessible documents offered by web servers (in formats such as HTML, PDF, etc.), as opposed to other data, such as with Google Image Search [4],[8].
Alta Vista: AltaVista was a web search engine owned by Yahoo! AltaVista was once one of the most popular search engines but lost its ground due to the rise of Google [5] [7]. The distinguishing feature of AltaVista was its minimalistic interface compared with other search engines of the time; a feature which was lost when it became a portal, but was regained when it refocused its efforts on its search function.
Lycos. Lycos is one of the oldest Search Engines on the web, It ceased crawl the web for its own listings in April 1999 and instead provides access to human-powered results from Look Smart for popular queries and crawler-based results from Yahoo or others [7].

Bing: When you search on Bing, in addition to providing relevant search results, the search engine also shows a list of related searches on the left hand side of the Search Engine Results Page (SERP). Bing is a new search engine market but it has prospered a lot. Bing is a new search engine from Microsoft that was launched on May 28, 2009. Microsoft calls it a "decision Engine", because it is designed to return search results in a format that organizes answers to address one's need.

Cuil: Cuil was a search engine that organized web pages by content and displayed relatively long entries along with thumbnail pictures for many results. Cuil said it had a larger index than any other search engine, with about 120 billion web pages. A user could log into their Face book account via Cuil, which would then search friend updates for topics, with search links. A user could also send messages to their friends through Cuil. Cuil worked on an automated encyclopedia called Cpedia, built by algorithmically summarizing and clustering ideas on the web to create encyclopedia-like reports. Instead of displaying search results, Cuil would show Cpedia articles matching the searched terms. This was meant to reduce duplication by combining information into one document. Cuil is available in many languages.

Gigablast: Gigablast is a search engine that does real-time indexing. Gigablast provides large-scale, high-performance, real-time information retrieval technology for partner sites. It offers a variety of features including topic generation and the ability to index multiple document formats. This search delivery mechanism gives a partner "turn key" search capability and the capacity to instantly offer search at maximum scalability with minimum cost. In addition, the Gigablast website provides unique "Gigabits" of information, enabling visitors to easily refine their search based upon related topics from search results. Clients range from NASDAQ 100 listed corporations to boutique companies.

Mamma (http://www.mamma.com): Mamma is used to search the web, news, stock company names MP3 files pictures and sounds. Again it is easy to use. It provides good search results as it gets easily connected to all top Search Engines. It was one of the web's first tier 2 Meta Search Engine

MSN: MSN is a collection of Internet sites and services provided by Microsoft. The range of services offered by MSN has changed since its initial release in 1995. MSN was once a simple online service for Windows 95, an early experiment at interactive multimedia content on the Internet, and one of the most popular dial-up Internet service providers. Today, MSN is primarily a popular Internet portal.

Netscape : Netscape's web browser is once dominant in terms of usage share, Netscape is credited with developing the Secure Sockets Layer Protocol (SSL) for securing online communication, which is still widely used, as well as JavaScript, the most widely used language for client-side scripting of web pages.

Yahoo: Yahoo Search interface send queries to searchable index of pages supplemented with directory of sites. Yahoo search combines the capabilities of all the Search Engine companies they have acquired, with its existing research and put them into a single Search Engine. The new Search Engine results were included in all of yahoo's sites that had a web search function. Yahoo! Search can be accessed using URL in.search.yahoo.com in browser. The yahoo also support selection-based search feature which enable users to invoke search using only their mouse and receive search suggestions in floating windows while remaining on Yahoo! properties such as Yahoo! Mail

Topsy : Topsy is a social analytics company that gives instant answers to critical business questions through real-time analysis. Topsy introduced four Domains of Search that search engines must follow that is : query speed, relevance (ranking results), frequency of index updates and recall (indexing historical data). It is observed that traditional web search engines compromise on update frequency; real-time search engines typically compromise on recall and relevance.

Board Reader: Boardreader is search engine for Forums and Boards. Board Reader has aggregated link data from Message Board Posts, and Forum Threads, which have linked TO and FROM wikipedia.org both internally and externally. This page reports by time period the link matrix for the data

Regator: Regator the human-curated blog directory and news aggregator just relaunched with a vastly improved and easier to use design, an improved search engine, and tight integration with Facebook Connect and Delicious. Regator's mission is to aggregate the best content from blogs across over 500 categories. To do so, Regator's editors created a vast directory of the best blogs on the Internet, with topics ranging from tech news and politics to tourism and beekeeping. The service's algorithms then create front pages for every topic that includes the most popular and interesting articles from these blogs, as well as an index of related posts and lists of trending topics.

Webopedia: An online computer dictionary and Internet search engine for Internet terms and technical support. Webopedia's Quick Reference section help to understand how technology works. Webopedia is complicated technology where Quick Reference articles break it down to just the facts and information needs to start learning more about common computer, Internet and technology topics.

Yippy: A research-oriented search engine, perfect for education. Yippy is a company built on Web search technology developed by Carnegie Mellon University researchers, much like Lycos was a decade earlier. Clusty

added new features and a new interface to the previous clustering web Meta search. Different tabs also offer Meta searches for news, jobs. government info and blogs.

Deeper Web: DeeperWeb is an innovative search engine plug-in and an essential Firefox addon for Google. The Deep Web (also called the Deepnet, the Invisible Web, the Undernet or the hidden Web) is World Wide Web content that is not part of the Surface Web, which is indexed by standard engines. Most of the Web's information is buried far down on dynamically generated sites, and standard search engines do not find it. Traditional search engines cannot retrieve content in the deep Web—those pages do not exist until they are created dynamically as the result of a specific search. The deep Web is several orders of magnitude larger than the surface Web

Zuula: Customizable tabbed interface for the major search engines. Zuula is a metasearch engine that provides search results from a number of different search engines. Zuula can be used to carry out standard web searches, image searches, video searches, news searches, blog searches, and job searches. Results are available from major search engines, such as Google, Yahoo, and Bing, and smaller engines, such as Gigablast and Mojeek. Zuula does not combine the results from its source search engines. Instead, tabs are used to organize the results from source engines. When a user carries out a search, the first results that are displayed are those from the search engine assigned to the first tab. The user can then click on other tabs to see the results from other source engines. Users can change the order of the tabs for each search type by moving them into the desired order.

Swoogle : Swoogle -- the semantic web search engine and metadata service provider. Swoogle provides services to human users through a browser interface and to software agents via Restful web services. Several techniques are used to rank query results inspired by the Page Rank algorithm developed at Google but adapted to the semantics and use patterns found in semantic web documents

Mojeek: Mojeek is a search engine with its own crawler, index and search results. Mojeek - a new independent crawler based search engine offering unbiased, fast and relevant search results, combined with a clean user interface and minimal on-screen clutter. All results are gathered and indexed by crawling the web. It does not contain other engine's. One of the main aims at Mojeek is to implement and allow full customization on a personal, search by search basis. To do this Mojeek's ranking and weightings have been designed and are fully configurable at the time of each search. This feature would also benefit site owners and webmasters wishing to implement site searches and multiple domain searches, using their own customized ranking parameters, a feature no other engine offers. Integrated with free xml feed, webmasters can offer their visitors a site search, ranked and weighted as they feel appropriate.

PCH search engine. It is a web search engine that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages. The information may be a specialist in web pages, images, information and other types of files Publishers Clearing House (PCH).

4. PROPOSED FRAMEWORK AND EXPERIMENTAL EVALUATION OF SEARCH ENGINES

To compare search engines, the following parameters have been identified. The first parameter presented here is relevant search which represents the number of relevant results returned by search engine in response to any query put by user. The relevant search parameter shows that google is best among the entire twenty searches with a percentage of 95 percent. The next parameter which is being used here is irrelevant search which shows the number of irrelevant document shown in response to a query. The third parameter is shows the total no. of document returned in response to a query which is followed by time taken to return all these result. The next parameter shows whether the search engine is active or not. There are few other factors which effect the performance of search engine such as advertisements are available or not, whether it supports the semantic based query or not and whether registration is required or not one more important factor which effects the performance of search engine is redundancy score of search engine. The redundancy score here refers to the number of redundant results shown in response to any query. The results corresponding to these parameters are shown below in form of tables.

Following are the five tables to show the different parameter analysis of these search engines. Table 1 shows Google, Altavista, Lycos and Bing engines. Table2 shows Cuil, Gigablast, Mamma and MSN. Table 3 shows Netscape, Yahoo, Topsy amd BoardReader. Table 4 shows Regator, Webopedia, Yippy and Deeperweb. Lastly, Table 5 shows Zulla, Swoogle, PCH and Mozeek search engines comparison on 10 different parameters.

Table 1: Comparison of Search Engines

Sno.	Search Engine / Parameter	Google	Alta vista	Lycos	Bing
1	Relevant search %	95	90	88	87
2	Irrelevant search %	5	10	12	13
3	Total resultant document	897000000	383000000	392000000	317000000
4	Time taken for search /sec	0.21	0.27	0.26	0.29
5	Current status of search engine	Active	Active	Active	Active
6	Advt. Availability	No	NO	NO	NO
7	Efficient to any search	YES	YES	YES	NO
8	Support to semantic based query	YES	NO	NO	NO
9	Redundancy in search results	1.09	1.62	1.92	1.78
10	Registration required	Optional	Optional	Optional	Optional

Table 2: Comparison of Search Engines

S No.	Search Engine / Parameter	Cuil	Gigablast	Mamma	MSN Search
1	Relevant search%	88	86	90	88
2	Irrelevant search %	12	14	10	12
3	Total resultant document	963000000	15029632	685000000	276000000
4	Time taken for search /sec	0.17	0.27	0.25	0.27
5	Current status of search engine	Active	Active	Active	Active
6	Advt. Availability	No	NO	NO	YES
7	Efficient to any search	YES	YES	YES	YES
8	Support to semantic based query	YES	NO	NO	YES
9	Redundancy in search results	1.19	1.70	1.56	1.26
10	Registration required	Optional	Optional	Optional	optional

Table 3: Comparison of Search Engines

S No.	Search Engine / Parameter	Netscape	Yahoo	Topsy	Board reader
1	Relevant search %	84	92	92	84
2	Irrelevant search %	16	8	8	16
3	Total resultant document	228000000	392000000	3042	1000
4	Time taken for search /sec	0.22	0.28	0.27	0.22
5	Current status of search engine	Active	Active	Active	Active
6	Advt. Availability	No	YES	NO	NO
7	Efficient to any search	YES	YES	YES	NO
8	Support to semantic based query	YES	YES	NO	NO
9	Redundancy in search results	1.49	1.32	1.49	1.58
10	Registration required	Optional	Optional	Optional	optional

Table 4: Comparison of Search Engines

S No.	Search Engine / Parameter	Regator	Webopedia	Yippy	Deeper web
1	Relevant search %	90	87	89	89
2	Irrelevant search %	10	13	11	11
3	Total resultant document	10000+	2400	472000000	96300000
4	Time taken for search /sec	0.21	0.12	0.26	0.21
5	Current status of search engine	Active	Active	Active	Active
6	Advt. Availability	No	YES	NO	NO
7	Efficient to any search	YES	YES	YES	NO
8	Support to semantic based query	YES	YES	NO	NO
9	Redundancy in search results	1.39	1.52	1.59	1.53
10	Registration required	Optional	Optional	Optional	Optional

Table 5: Comparison of Search Engines

S No.	Search Engine / Parameter	Zulla	Swoogle	PCH search	Mojeek
1	Relevant search %	90	88	85	88
2	Irrelevant search %	10	12	15	12
3	Total resultant document	89700000	476	4620000	5007208
4	Time taken for search /sec	0.27	2.1	0.26	0.81
5	Current status of search engine	Active	Active	Active	Active
6	Advt. Availability	No	YES	NO	YES
7	Efficient to any search	YES	YES	YES	NO
8	Support to semantic based query	YES	NO	NO	NO
9	Redundancy in search results	1.38	2.52	1.49	1.43
10	Registration required	Optional	Optional	Optional	optional

5. DISCUSSION

In a way to find the best search engine as per the need of user an experiment is conducted by using 20 popular search engines .In this experiment the queries from different fields are given to these search engines and then the results have been analysed based on different parameters. The Table1-5 show comparison of Google, Alta Vista, Lycos, Bing, Cuil, Gigablast, Mamma, MSN Search, Netscape,Yahoo, Topsy, Board Reader, Regator, Webopedia, Yippy, Depperweb, Julla, Swoogle, PCH Search,Mojeek based on the above mentioned paremeters. In order to analyze the results based on these parameters a number of queries related to different fields are given to the above mentioned search engines. The percent of relevant results corresponding to different search engines have been shown in the table 1-5, It can be seen clearly that the relevant search percentage is maximum in case of google with 95 and it is least for boardreader. However in case of total number of results cuil is ahead of google but it really does not matter because user normally does not looks beyond a specific limit. The total time taken for each search is a parameter which really matters a lot .the table 1-5 show that the webopedia takes least amount of time which is .12 second while it is .21 second in case of google. But due to large number of relevant results google may still be considered ahead to webopedia. There are other factors like advertisement availability which also have significant impact on the performance of availability. The Google again proves better as it does not contain any advertisement. Semantic based queries are those which are interpreted through there meaning not on the bases of keywords only. The tables also show the search engines with semantic support.

6. CONCLUSION

This paper has presented a framework to evaluate search engines. This framework is evaluated on twenty search engines. In market a lot of search engine are available. But it depends upon user and type of query that which they want to use. A user must use a search engine which provides secure, relevant and quick search results to the queries. On evaluation ,it has been formed that based on the various factors such as variance, redundancy, time taken and percentage of relevant and irrelevant search for each search google proves to be better than others.

REFERENCES

1) D.Tumer, M.A. Shah and Y.Bitirim, "An Empirical Evaluation on semantic Search Performance of Keyword based and Semantic search engines google, Yahoo,Msn and Hakia", 4th International Conference on internet Monitoring and Protection, 2009.
2) Diana Botluk "update to search Engines compared", LLRX, 2000
3) Faizan Shaikh, Usman A. Siddiqui,Iram Shahzadi, Syed I.Jami and Zubair A. Shaikh, " SWISE : Semantic Web Based Intelligent Search Engine" ICIET IEEE,2010.
4) B.T Sampath Kumar,J.N Prakash," Precision and relative recall of search engines:A comparative study of google and yahoo" Singapore Journal of Library & Information Management , Vol.38 , 2009 .
5) Jean Veronis "A Comparative study of Six Search Engines" 2006
6) Jiandong Cao, Yang Tang, Binbin Lou "Social Search Engine Research" IEEE,2010
7) Judit Bar-Ilan, "Search Engine Results over Time-A Case Study on Search Engine Stability", International Journal of Scientometrics, Informetrics and Bibliometrics, 2/3(1),1998.
8) Bernard J. Jansena, Amanda Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs", Information Processing & Management, Vol. 42, No. 1, Pages 248–263,January 2006
9) Enid Burn, "US search Engine Ranking, April 2007" http://www.clickz.com/clickz/news/1692204/us-search-engine-rankings-april-2007, may 2007.
10) Akinola, S. Olalekan, Yara, P. Olatunde and Osunade O. Oluwaseyitanfunmi . "Internet Information Seeking Behaviour of Nigerian Students: University of Ibadan as a Case Study", Proceedings of the Eighth International Working Conference of International Federation for Information Processing, (IFIP WG 9.4), Abuja Nigeria, 26 – 28 May, 2005.
11) Fazli, Can, Rabia, Nuray, Ayisigi, B. Sevdik (2003). Automatic Performance Evaluation of Web Search Engines, Information Processing and Management Vol. 40, No. 3 ,Pg 495–514, 2004,
12) Hananzita Halim and Kiran Kaur, " Malaysian Web Search Engines: A Critical Analysis" , Malaysian Journal of Library and Information Science, Vol.11, No. 1, pg. 103 -122 ,2006
13) Hong Yu and David Kaufman "A cognitive evaluation of four online search engines for answering definitional questions posed by Physicians", Pacific Symposium on Biocomputing, Vol. 12: Pg. 328-339. 2007
14) Jansen, B. J. and Pooch, U. " A review of Web searching studies and a framework for future research". Journal of the American Society for Information Science and Technology, Vol.52 No.3, Pg.235 – 246. 2001
15) Gordon, M. and Pathak, P., "Finding information on the World Wide Web: the retrieval Effectiveness of search engines", Information Processing and Management, Vol. 35 No.2 ,Pg 141 – 180. 1999
16) Griesbaum Joachim , "Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de", Information Reearch, Vol 9 No.4., 2004