ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 9, September 2014

An Analysis on Search Engines: Techniques and Tools

R. Rubini[1], Dr. R. Manicka Chezian [2]

Research Scholar, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India[1]

Associate Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India[2]

Abstract: Web search is an integral part of our daily lives. Search engines help users locate particular information within large stores of content developed for human consumption. Search engines are developed using standard sets of realistic test cases that allow developers to measure the relative effectiveness of alternative approaches. A search engine is a tool that allows a user to enter keywords and retrieve information on websites contained in its catalog or database. Search engine tools like Google is run by search engine software that allows the database to be searched. The process of searching in the search engine is very efficient and more accurate. This paper surveys about how the search engines reduce the number of unwanted search results in the searching process.

Keywords: **Indexing**, Search Engine, Semantic Engine, Search Engine Optimization, Web Crawling

I. INTRODUCTION

Search engines are one tool used to answer information needs. Search engines are huge databases of web pages as well as software packages for indexing and re-trieving the pages that enable users to find information of interest to them [1]. The search engines directories, portals and indexes are the web's "catalogues" allowing a user to carry out the task of searching the web for information that he/she requires. Search engines typically "crawl" web pages in advance to build local copies and/or indexes of the pages. This local index is then used later to identify relevant pages and answer users' queries quickly.

A. Features of Search Engine

- A true search engine is automated software program that moves around the Web collecting Web Pages to include in its catalog or database.
- It searches when a user requests information from a search engine; not the entire Web.

II. LITERATURE SURVEY

Fu-Ming Hung and Jenn-Hwa Yang et al [11], present an intelligent search engine with semantic technologies. This survey has combine description logic inference system and digital library ontology to complete intelligent search engine.

Inamdar and Shinde et al [12], discussed agent based intelligent search engine system for web mining.

Patrick Lambrix and Nahid Shahmehri and Niclas Wahllof et al [13], presents a search engine is described as one that tackles the problem of enhancing the precision and recall for retrieval of documents. There have been tested the system on small-scale databases with promising results.

Satya Sai Prakash et al [14], present architecture and design specifications for new generation search engines highlighting the need for intelligence and give a knowledge framework to capture intuition.

Dan Meng, Xu Huang et al, discussed an interactive intelligent search engine model based on user information preference [15]. This model can be an effective and useful way to realize the individuation information search for different user information preference.

Xiajiong Shen Yan Xu Junyang Yu Ke Zhang et al, forward an intelligent search engine where Information Retrieval model is found on formal context of FCA (formal concept analysis) and incorporates with a browsing mechanism.FCA is a useful way of supporting the flexible management of documents according to conceptual relation [16].

III. SEARCH ENGINE ARCHITECTURE

Creating a search engine which scales even to today's web presents many challenges. Fast crawling technology is needed to gather the web documents and keep them up to date. Storage space must be used efficiently to store indices and, optionally, the documents themselves. The indexing system must process hundreds of gigabytes of data efficiently. Queries must be handled quickly, at a rate of hundreds to thousands per second.

These tasks are becoming increasingly difficult as the Web grows. However, hardware performance and cost have improved dramatically to partially offset the difficulty. There are, however, several notable exceptions to this progress such as disk seek time and operating system robustness [2]. In designing Google, needs to consider both the rate of growth of the Web and technological changes. Google is designed to scale well to extremely large data sets. It makes efficient use of storage space to store the index. Its data structures are optimized for fast and efficient access. Further, expect that the cost to index and store text or HTML will eventually decline relative to the amount that will be available. This will result in favourable scaling properties for centralized systems like Google.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 9, September 2014

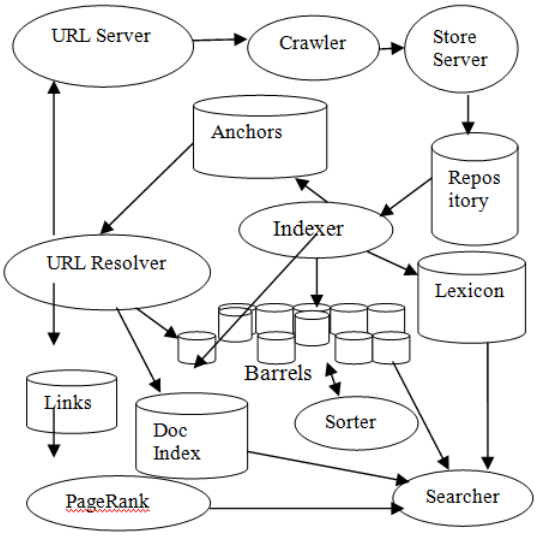

Fig. 1 Search Engine Architecture

The web crawling is done by several distributed crawlers. There is a URL server that sends lists of URLs to be fetched to the crawlers. The web pages that are fetched are then sent to the store server. The store server then compresses and stores the web pages into a repository. Every web page has an associated ID number called a doc ID which is assigned whenever a new URL is parsed out of a web page. The indexing function is performed by the indexer and the sorter. The indexer performs a number of functions.

It reads the repository; uncompressed the documents, and parses them. Each document is converted into a set of word occurrences called hits. The hits record the word, position in document, an approximation of font size, and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It parses out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link.

The URL resolver reads the anchors file and converts relative URLs into absolute URLs and in turn into doc IDs. It puts the anchor text into the forward index, associated with the doc ID that the anchor points to. It also generates a database of links which are pairs of doc IDs. The links database is used to compute Page Rank's for all the documents.

The sorter takes the barrels, which are sorted by doc ID and resorts them by word ID to generate the inverted index. This is done in place so that little temporary space is needed for this operation. The sorter also produces a list of word IDs and offsets into the inverted index. A program called Dump Lexicon takes this list together with the lexicon produced by the indexer and generates a new lexicon to be used by the searcher. The searcher is run by a web server and uses the lexicon built by Dump Lexicon together with the inverted index and the Page Ranks to answer queries.

A. *Types of Search Engines*

The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.
Recommended font sizes are shown in Table 1.

B. *Improvements of Search Engine*

There are different ways to improve the performance of web search engines. Generally speaking, there are three main directions:

- Improving user interface on query input
- Using Filtering towards the query results
- Solving algorithms in web page spying and collecting, indexing, and output.

C. *Basic Types of Search Tools*

1) Crawler Based Search Engines

Crawler based search engines create their listings automatically. Computer programs 'spiders' build them not by human selection. They are not organized by subject categories; a computer algorithm ranks all pages. Such kinds of search engines are huge and often retrieve a lot of information -- for complex searches it allows to search within the results of a previous search and enables you to refine search results. These types of search engines contain full text of the web pages the link to. So one can find pages by matching words in the pages one wants.

2) Human Powered Directories

These are built by human selection i.e. they depend on humans to create listings. They are organized into subject categories and subjects do classification of pages. Human powered directories never contain full text of the web page they link to. They are smaller than most search engines.

IV. GENERAL SEARCH ENGINES

It includes the search engines Google yahoo etc. which provides a number of links when search the user for a query. It became a vast collection of information for these arches. It may not be containing the exact fact but it searches the query related all items which syntactically matches for the searched query. As far as users are concerned they need relevant and precise results.

A. Conventional Searching

Conventional searching helps the user to have the links of the searched query. It gives all the possible urls. In conventional searching it is not considering about the different meanings the words can have infarct it will show all the matches possible. By clicking or going through the links only have the clear picture about the query that what searching for. But it is not the case of the semantic search engines [3]. It is time consuming process if go through each and every links one by one. That may also happen in this type of search engines. While comparing this with the semantic search it is giving a difficult way to the user to get the result of the specified query.

The conventional search engines always provide the links for the user to go through to reach the results. It will also have the shot keys to search in the web, pictures, videos,

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 9, September 2014

news, shopping etc… but the user will not get the answer for the query. In all these searches the search engine will provide the list of links by which the user can reach the destination. In conventional search engines it is not sure that the search thing is the same what is getting or the other possibilities of the same query.

V. SEMANTIC SEARCH ENGINES

Semantic search engines include the searching of the query related to the entered data by the user in the data space. But the thing is that it gives a short review of the commonly used related word description for the convenience of the users, it makes the job easier-the search can be done easily [4]. Semantic search engine differs according to the user. For the ordinary users no need of analysing the data which is in high level.

A. Key requirements of Semantic Search Engine

- **Low barrier to access for ordinary end users**. Our semantic search engine should overcome the problem of knowledge overhead and ensure that ordinary end users are able to use it without having to know about the vocabulary or structure of the ontology or having to master a special query language.

- **Dealing with complex queries**. In contrast with existing semantic-based keyword search engines which only answer simple queries, our semantic search engine should allow end users to ask complex queries and provide comprehensive means to handle them.

- **Precise and self-explanatory results**. Our semantic search engine should be able to produce precise results. Thus, ordinary end users can understand the results (e.g. what they are and why they are there) without having to consult the back-end semantic data repositories or their underlying ontologies.

B.Some of Semantic Engine

1) Hakia

It is a common search semantic search engine in use. It is well organized by the tabs Web results, credible sites, images and news. Credible site includes the sites which are vetted by the librarians and other information professionals. For some of the user queries it produces the resumes. They are the portals which gives all information related to that subject. For each resumes there will be an index of links which refers to the corresponding pages, it helps for a quick reference. According to the query the content of the resume will vary. Resumes are one the most impressive feature of hakia. Hakia will also provide the related queries also, which help the user to reach the target or to get the thing very easy.

2) Sense Bot

It gives a summarized accurate search result according to the query given. The search engine itself tries to understand the concept of the query, actually what it contains and will give an appropriate result. To do this it makes use of the text mining on the web pages which results on the queries to categorize the actual semantic concept. This summarized result helps the user to get the result of the content very fast. The most important fact is that the answer will be relevant and precise.

C.Methodology

1) *Resource Description Framework (RDF)*

It is the foundation of the semantic web. It is a standardized language by W3C.Semantic web consists of web of data that is the data collection is done rather than the document collection.

2) Graph patterns

It is an important concept in semantic search. In semantic search it is used in multiple varying roles. It is used to solve or encode the complex constraint queries given by the user; it is solved by locating the corresponding graph in the RDF network. Graphs patterns also give the idea that where to collect or to fetch the data for particular item.

3) Logics

It is internally very much tied with semantic web. In most of the cases the applications take few entailments for the base and create functionality according to their requirements over that.

VI.SEARCH ENGINE OPTIMIZATION

Search engine optimization (SEO) refers to techniques that help your website rank higher in organic [5].

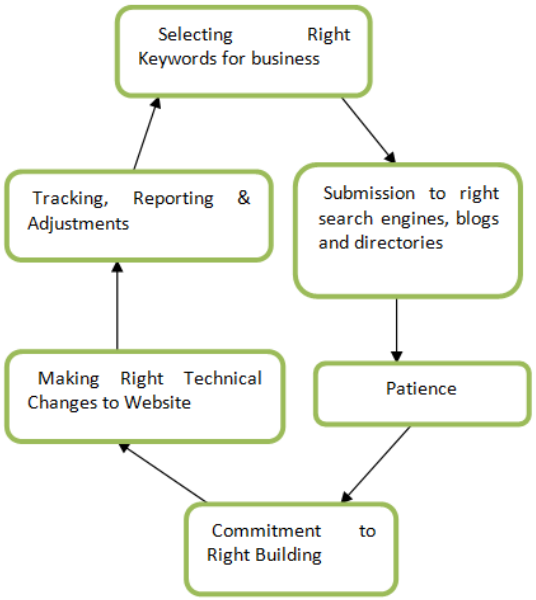

Fig.2 Effective Search Engine Optimization

A. Search Engine Optimization Techniques

1) Directory Submission

Directory submission is one of the important techniques in SEO to create incoming links to a website through related page and category. A website is created and need to be rank to get good business results. Manually submission to directories is the best approach to rank your website. Internet directory is the platform on World Wide Web for

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 9, September 2014

information and links of many websites. Many directories are providing free service to website in directory [6]. To submit website in directories can produce web traffic for your website. This assist you to promote your business needs. The directory submission is used as SEO technique to promote your business.

2) Keyword Generation

Any search engine optimization method used keywords generation process. The keywords are necessary and most important part of SEO. The keywords are must be related to business [7]. Because related keywords boost website in short span of time.

There are many online tools available to generate keywords relevant your needs like: Word tracker, Yahoo keyword selector tool, Google Ad words keyword tool and Thesaurus etc. By using these tools just put one word related your website like gamming. But only keywords are not providing assurance to popularity of website.

3) Link Exchanges

The link exchange is the method in SEO to place link on other websites and other websites place links on your websites means vice versa [8]. There are many types of link exchanges are used like: illustrate interest directly on web pages and other is that send email or discussion forums to show interest for link exchanges.

Only related website but with good page rank websites are required to build reciprocal links.

VII. WEB CRAWLING

Web crawlers are an essential component to search engines. Web crawling speed is governed not only by the speed of one's own Internet connection, but also by the speed of the sites that are to be crawled [9]. Especially if one is a crawling site from multiple servers, the total crawling time can be significantly reduced, if many downloads are done in parallel.

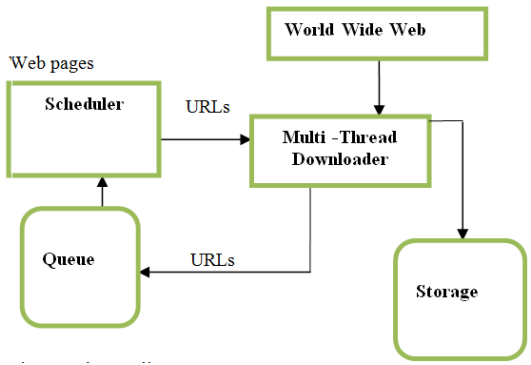

Fig. 3 Web Crawling

The Web crawler can be used for crawling through a whole site on the Inter-/Intranet. It specifies a start-URL and the Crawler follows all links found in that HTML page. a tree-structure, the root is the start-URL; all links in that root-HTML-page are direct sons of the root. Subsequent links are then sons of the previous sons.

A) Crawling Techniques

1) Focused Crawling

A general purpose Web crawler gathers as many pages as it can from a particular set of URL's. Whereas a focused crawler is designed to only gather documents on a specific topic, and a crawler with dynamically reconfigurable priority controls which is governed by the classifier and distiller.

2) Distributed Crawling

Indexing the web is a challenge due to its growing and dynamic nature. A single crawling process even if multithreading is used will be insufficient for large – scale engines that need to fetch large amounts of data rapidly. When a single centralized crawler is used all the fetched data passes through a single physical link.

VIII. INDEXING

Similar to an index of a book, a search engine also extracts and builds a catalog of all the words that appear on each web page and the number of times it appears on that page[10]. The parser can extract the relevant information from a web page by excluding certain common words (such as a, an, the - also known as stop words).Indexes are updated periodically as new content is crawled. Some indexes help create a dictionary (lexicon) of all words that are available for searching.

A. Methods of Indexing

1) Full-Text Indexing

As its name implies, full-text indexing is where every word on the page is put into a database for searching. Full-text indexing will help you find every example of a reference to a specific name or terminology. In this case the websites are indexed my computer software. This software called "spiders" or "robots" automatically seeks out Web sites on the Internet and retrieves information from those sites (which matches the search criteria) using set instructions written into the software.

2) Human Indexing

Yahoo and some of Magellan are two of the few examples of human indexing. In the Keyword indexing, all of the work was done by a computer program called a "spider" or a "robot"[8].

IX. CONCLUSION

As the web and its usage continues to grow, many opportunities to analyse web data and extract all manner of useful knowledge from it. The web presents new challenges to the traditional data mining algorithms that work on flat data. Nobody have time to lose by searching the needed content in this fast life that is the area the semantic search engine gets the chance. Search engine and web crawler describes the functionalities of all the components involved in finding information on the web.

REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hyper textual Web search engine. In Proceedings of the Seventh WWW Conference, Brisbane, Australia, 1998.

[2] Grossan, B. "Search Engines: What they are, how they work, and practical suggestions for getting the most out of them," February1997.
[3] Koyoro Shadeo, Trends in web Based Search Engine 'Journal of emerging trends in computing and information Sciences' Vol 3, No-6, June 2012, ISSN – 2079-8407.
[4] PSSE: Architecture for a Personalized Semantic Search Engine A. M. Riad, Hamdy K. Elminir, Mohamed Abu ElSoud, Sahar. F. Sabbeh. doi: 10.4156/aiss.vol2.issue1.9
[5] Bo Xing, Zhangxi Lin. The Impact of Search Engine Optimization on Online Advertising Market: The Eight International Conferences on Electronic Commerce (ICEC'2006), pp. 519-529, ACM Electronic Commerce, 2006.
[6] Muhammad Akram, Search Engine Optimization Techniques Practiced in Organizations, "A Study of Four Organization", Journal of Computing, Vol-2, Issue-6, June-2010, ISSN- 2151-9617
[7] Dr. S. Sarvankumar, A New methodology for search engine optimization without getting sandboxed 'International journal of Advanced research in computer and communication Engineering Vol 1, issues, Sept 2012, PP- 472-475.
[8] Mike Barus. "Link Exchange and One Way Links Using Web Directories," February 2009.
[9] 'A web crawler design for data mining' Mike Thelwall Journal of information Science, 27 (5) 2001 PP. 321
[10] C. W. Cleverdon. The Cranfield tests on index language devices. In Aslib Proceedings, volume 19, pages 173-192, 1967. (Reprinted in Readings in Information Retrieval, K. Spärck- Jones and P. Willett, editors, Morgan Kaufmann, 1997).
[11] Fu-ing Huang et al. "Intelligent Search Engine with Semantic Technologies"
[12] S. A. Inamdar1 and G. N. Shinde "An Agent Based Intelligent Search Engine System for Web mining" Research, Reflections and Innovations in Integrating ICT in education 2008.
[13] Patrick Lambrix et al, "Dwebic: An Intelligent Search Engine based on Default Description Logics"-1997.
[14] K. Satya Sai Prakash and S. V. Raghavan "Intelligent Search Engine: Simulation to Implementation", In the proceedings of 6th International conference on Information Integration and Web-based Applications and Services (iiWAS2004), pp. 203-212, September 27 - 29, 2004, Jakarta, Indonesia, ISBN 3-85403-183-01.
[15] Dan Meng, Xu Huang "An Interactive Intelligent Search Engine Model Research Based on User Information Preference", 9th International Conference on Computer Science and Informatics, 2006 Proceedings, ISBN 978-90-78677-01-7.
[16] Xiajiong Shen Yan Xu Junyang Yu Ke Zhang "Intelligent Search Engine Based on Formal Concept Analysis" IEEE International Conference on Granular Computing, pp. 669, 2-4 Nov, 2007.

BIOGRAPHIES

R.Rubini is a research scholar in Nallamuthu Gounder Mahalingam College, Pollachi. She received her Master of Computer Application (MCA) in 2013. She has presented papers in International/National Conferences and attended Workshop, Seminar. Her research interest focuses on Data Mining.

Dr. R.Manickachezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published thirty papers in international/national journal and conferences: He is a recipient of many awards like Desha Mithra Award and Best Paper Award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.