Journal of Computer Applications (JCA)
ISSN: 0974-1925, Volume IV, Issue 1, 2011

Search Engines:A Study

[1] Mr.K. Tarakeswar , [2] Ms. D. Kavitha

Abstract— **The Internet is a huge collection of data. To get the appropriate information from it, using a search engine is the most effective way. Many Search Engines were introduced since 1990. In this paper we present a brief study on search engines. First, we present the definition of search engine, types of search engines and the general working process of a search engine. Then we give an example for the working process with a description of the Google search engine architecture. Later, we present a short description of the next generation search engines. Then we present comparisons among some major search engines.**

Index Terms— **Internet; Search Engine; working; architecture.**

I. INTRODUCTION

Access to various types of information is necessary these days. The World Wide Web (WWW) contains a lot of web pages. To search for the information necessary for us from that huge collection of web pages, using a Search Engine will provide with efficient results. Many web pages in the WWW contain inappropriate information. This is due to the inappropriate naming and unnecessary highlighting of the content of web pages by their web masters. This raises the need of a search engine. Using a good Search Engine will filter out the necessary and relevant information needed by the user.

This paper presents an overview on the search engines. In the second section of this paper, we present the definition of search engine and we describe the types of search engines in the third section. We describe the general working of a search engine in the fourth section and present an example for it in the fifth section by explaining the Google search engine architecture. In the sixth section we present a brief explanation on the next generation search engines and in the seventh section we present comparisons among some major search engines.

II. DEFINITION OF SEARCH ENGINE

Definition 1: Search Engine is a program which searches the database, gathers and reports the information which contains the specified or related terms.

Definition 2: The term Search Engine [11] refers to the process of searching files using the key words specified. The key words found are returned and collated into the user information.

III.TYPES OF SEARCH ENGINES

Search Engines are of four types[6]. They are

A.Crawler based search engines.
B.Human powered directories.
C.Hybrid search engines.
D.Meta search engines.

A. Crawler based Search Engines

Crawler based search engines contain three parts. The first part is the 'Crawler' (bot or robot or spider). It is used to wander the web and create listings of web pages. The second part is the 'Index', which is a huge collection of copies of web pages and the third part is the 'Search Engine Software' which ranks the results. Because the crawler in this engine searches the web constantly, it provides updated information. Google, Live Search, Ask and most other search engines are crawler based.

B. Human Powered Directories

Human powered directories are search engines which depend on humans for their web page listings. These types of search engines get their listings of web pages from the submissions made by the respective web page masters. The submission contains the address, title and a brief description of the site. Later, the submission is reviewed by editors. A directory searches for results only from the page descriptions submitted to it. This is an advantage because, as the pages are submitted manually, the quality of the content will be better and more appropriate compared to the results retrieved by a crawler based search engine. But, the disadvantage is, any change made to an already submitted web page will not be updated until it is submitted again. Also, the ranking of pages can't be changed once ranking is done. Yahoo, dmoz and Galaxy are some examples.

C. Hybrid Search Engines

Hybrid search engines include the features of crawler based search engines and human powered directories. Currently, some search engines are using both features to provide effective results. MSN, Google and Yahoo are some examples.

D. Meta Search Engines

Meta search engines fetch results from other search engines. The fetched results are combined and ranked again according to their relevancy. These search engines were useful when each search engine had a significantly unique index and search engines were less savvy. Because the search has improved a lot, the need for these has reduced. MetaCrawler and MSN Search are some examples.

IV. WORKING OF SEARCH ENGINE

The working [3], [4], [5] of Search Engine involves three basic tasks. They are,

A.Searching the WWW and collecting the pages.
B.Keeping the index of the words they find and where they were found.
owing users to search for words or a combination of rom the index by using efficient software.

Manuscript received Apr 18, 2011.

Mr.K. Tarakeswar, Department of Computer Science and Engineering, G. Pulla Reddy Engineering College, Kurnool-518002, Andhra Pradesh, India. (e-mail : eshwartarak158@gmail.com)
D. Kavitha, Department of Computer Science and Engineering,G. Pulla Reddy Engineering College, Kurnool-518002, Andhra Pradesh, India. (e-mail : dwaramkavithareddy@gmail.com)

These tasks are performed by the three parts of a search engine. They are,
1) Crawler
2) Index
3) Search Engine Software.
The working of a search engine is shown in the Fig.1.

A. Searching theWWW and collecting the pages

Definition of Computer Robot, Spider or Crawler:
Computer Robots [10] are programs, which automate repetitive tasks at speeds impossible to be done by humans. The term 'bot' on the internet implies anything which interfaces with the user or collects data.
To present the result pages for a query a search engine must search and collect it. To find the web page from the millions of web pages present, search engines use the software robots called 'Crawlers or Spiders'. They build lists of words found in the web pages. This process of building lists is called 'Web Crawling'. A lot of pages must be traced to collect a useful list of words.
A spider chooses a list of heavily used servers and popular web pages as its starting point. It then begins with a popular web site, indexes the words present in it and also follows every other link present in that page. In this way, it quickly starts to travel spreading across widely used parts of the Internet.
The crawler carefully chooses at each step about which page to index. Some policies were introduced to guide the crawler. They are

1. Selection policy: Selection policy states which pages to download.
2. Revisit policy: Revisit policy states when to check for changes in web pages.
3. Politeness policy: Politeness policy states how to avoid overloading of web sites.
4. Parallelization policy: Parallelization policy states how to coordinate the different web crawlers distributed.

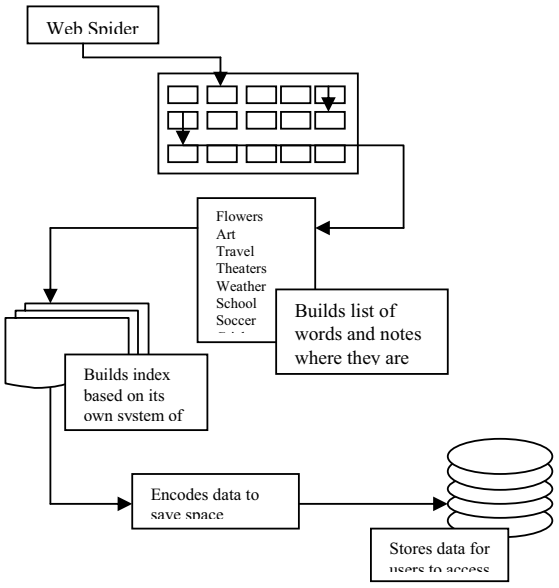

Figure1. Working of a Search Engine

B. Keeping an index of the words they find and where they were found.Before describing this task, we give a brief explanation on Meta Tags.

Definition of Meta Tags: Meta Tags[7] allow a web page owner to mention key words and concepts under which hi page will be indexed. They guide a search engine in choosing appropriate meaning for a word from the several possible.
But, over reliance on meta tags leads to pages with popular topics, but which have very less or irrelevant content. To compensate this, the crawlers correlate the meta tags with the page content. They reject the meta tags which don't match with the words in the page.
After collecting the information, it must be stored in a way useful to the user. The stored data is encoded by the search engines to save the storage space. Two important components are present in making the collected data accessible to the user. They are mentioned below.

1. The information stored with the data

A search engine can store only the word and URL (Universal Resource Locator). This is a simple way of storage. In this case, the results cannot be ranked for their relevancy. To provide relevant results, weights can be assigned to the words based on their locations in the page.

2. The method by which the information is indexed

Indexing of words is made to allow the information to be accessed as fast as possible. An effective way is to use a Hash table for indexing. In the Hash table indexing we apply a formula for attaching a numerical value to the words. The formula used must evenly distribute all the entries.
In a dictionary, more pages will be present for the words starting with the letter's' than for the letter 'z'. So, the time to search for a word starting with's' is more, compared with the time taken to search for a word starting with'z'. Hashing evens out such differences. It also reduces the average search time for an entry. The hash table will contain the hashed values and pointers to the actual data. Hence, using efficient indexing and effective storage methods provide quick and better results for complicated queries also.

C C. Providing results by using efficient search engine software

The third task is performed using search engine software. This software sifts through the results and ranks them according to their relevancy. Some basic principles are followed by all search engines to determine the relevancy of results. They are,

- Principle 1: The location of key words in a web page is a factor for determining the relevancy. The pages containing the search term in its HTML (Hyper Text Markup Language) tag, at the beginning of the page, in the links or subheadings and meta tags are more relevant.
- Principle 2: Frequency of key words in the page is another factor for determining the relevancy. The page with more occurrences of a search term is said to be more relevant.

Each search engine has its own method for assigning weights. Because of this, for the same query, different search engines provide differently ordered results.

1. Off Page Factors

Off Page Factors[10] are also used to rank web pages. They do not depend on the content of the page. They are

- Factor 1: Look of the web page.

Search engines infer a lot about the content of a page with a look of the page. Sophisticated techniques exist to find [illegible]cial, fake and useless links and remove them.

- Factor 2: Click Through Measurement.

This determines the behavior of the user in relation to what results they choose while searching.

V. The Google Search Engine Architecture

In the Google search engine[1],[2], the three tasks of a search engine are performed as follows. The Google Architecture is shown in Fig. 2.

A. Searching the WWW and collecting the pages

The first task is performed using several distributed crawlers. The URL Server will send the lists of URLs to be fetched to the crawlers. The fetched web pages are sent to the Store Server. The Store Server compresses the web pages and stores them in a repository. Each web page is assigned a 'docID'. It is assigned each time a new url is parsed out of a page.

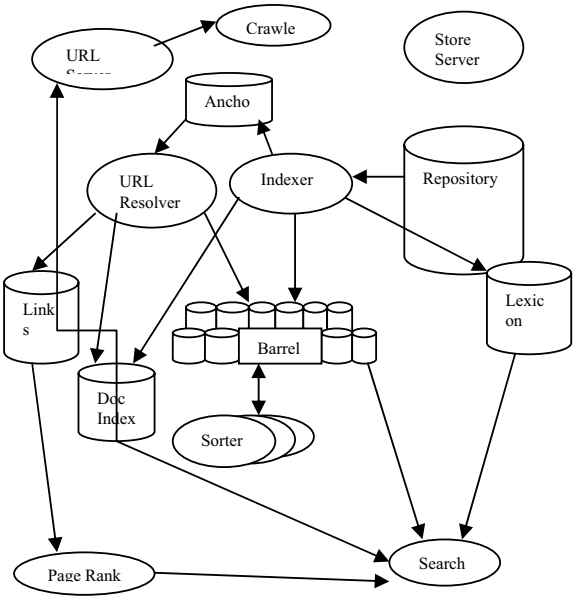

Figure.2 The High Level Google Architecture

B. Keeping an index of the words they find and where they were found.

In Google, Indexing is done by the Indexer and Sorter. The Indexer reads the repository to uncompress the documents. It then parses the documents. Every document is converted into a set of word occurrences which are referred to as 'Hits'. The Hits contains the words, their position in the document, their font size and capitalization. These hits are distributed by the Indexer into a set of Barrels, thus creating a partially sorted forward index. The Indexer also parses out the links in all web pages and stores the key information about them in 'Anchors' file.

The URL Resolver reads the Anchors file, converts relative URLs into absolute URLs which are then converted into docIDs. It also puts the anchor text into the forward index according to their docIDs. It generates a database of links which are used to compute page ranks of all documents. The Sorter takes the Barrels which are sorted by docIDs. These are resorted according to their wordIDs to create the inverted index. The Sorter produces a list of wordIDs and also offsets into the inverted index.

C. Providing results by using efficient search engine software.

The Dump Lexicon program takes the list generated by Sorter along with the lexicon generated by the Indexer. It then produces the lexicon which is used by the Searcher. The Searcher is run by a web server. It uses the lexicon built by the Dump Lexicon program, inverted index and page ranks to efficiently answer the queries.

VI. Next Generation Search Engines

The next generation search engines are referred to as Peer-to-Peer Search engines. They employ major types of discovery methods which are mentioned below.

- Selective forwarding systems.
- Flooding broadcast of queries.
- Centralized indexes and repositories.
- Decentralized hash table networks.
- Distributed indexes and repositories.
- Relevance driven network crawlers.

The Peer-to-Peer search implementation has two models. They are

A. Centralized server-client model.

B. Decentralized model.

A. Centralized server-client model.

The Centralized server-client model[9] contains a single, centralized server. It contains a directory of the shared files which are stored on the computers of users in the network. When a user searches for some file, the central server creates a list of files from its database of files which belong to users currently connected to the network. The server displays that list of files to the user. After the user chooses the file, a direct connection is setup with individual computers which contain that file at that moment. Opennap, kazaa and eDonkey are examples of Centralized server-client models.

Advantages

- The single, centralized index locates files quickly and efficiently.
- The search requests are sent to all clients who have logged in to the network. So, the search will be as through as possible.

Disadvantages

- The centralized server results in a single point of failure.
- As the centralized index is updated only periodically, the client may receive outdated information.

B. Decentralized model.

Decentralization of the network is made so that each peer can communicate as an equal to all the other peers. The Decentralized model[8] will not be having a single, central server. This model can be explained as follows.

Let there be some peers a, b, c, d, e, f etc., Whenever a peer 'a' enters the decentralized network, it connects to another peer 'b' to announce that it is alive. The peer 'b' announces to all other peers to which it is connected about the peer 'a'

being alive. The other peers c, d, e, f etc., repeat this pattern. After 'a' announces that it is alive, it can send search requests to 'b'. 'b' will pass this request to c, d, e, f etc., If 'c' has a copy of the file requested by 'a', 'c' sends a reply to 'b'. 'b' passes this reply back to 'a'. 'a' then opens a direct connection to 'c' and downloads the file. This scenario allows for an infinite network. In practice, a time to live (TTL) is used to limit the number of nodes reached by a request. Gnutella, mnet, freenet and gnunet are examples of Decentralized model.

Advantages

- The problem of a single point of failure is eliminated.
- The network is harder to shutdown.

Disadvantages

- Searching is slower in a decentralized network.
- Because of the TTL, the request for a file can't reach the node which will be having the file needed.

VII Comparison of Search Engines

In this section, we present comparisons among some major search engines based on some factors [12], which make a search engine provide satisfactory results. The results of the comparisons are presented in the below table, Table I.

TABLE I COMPARISONS OF MAJOR SEARCH ENGINES

	AltaVista	**Yahoo**	**Google**	**Ask**	**Teoma**	**MSN Search**	**Bing**
Links to a URL	No	Yes	Yes	No	No	No	No
Languages provided	All or English	41 languages	44 languages	6 languages	10 Languages	38 Languages	41 languages
Similar pages	No	No	Yes	No	No	No	No
Boolean and Phrase search	Yes	Yes	Yes	Yes	Yes	Yes	Yes
News and Multimedia search	Yes	Yes	Yes	Yes	No	Yes	Yes
Stemming	No	Yes	Yes	No	No	Yes	No
Other databases provided	Yes	Yes	Yes	Yes	No	Yes	Yes
Word in URL	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Search by File Type	Yes	Yes	Yes	Yes	No	Yes	Yes
Truncation	Yes	Yes	No	Yes	No	Yes	Yes
Grouping and Sorting Results	No	Yes	Yes	No	Only Grouping of results	Yes	Yes
Domain Search	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Thumbnails of results	No	No	Yes	Yes	No	No	Yes
Personalize	No	Yes	Yes	Yes	No	No	Yes
Date Limit Search	Yes	Yes	Yes	Yes	Yes	No	No

VIII CONCLUSIONS

Concluding this paper, using a Search Engine is obviously good to gather the necessary information. Many search engines have been developed to provide the best results for users. The present day search engines provide a variety of results like geographic search, domain search, personalization etc., but, they are unable to present satisfactory results for scientists, analysts, research students etc. To compensate this, the Peer-to-Peer search engines are being developed, which are referred to as the next generation search engines. The Peer-to-Peer search engines use the major searching techniques like flooding broadcast of queries, selective forwarding of queries, relevance driven network crawlers etc., They also use scalable and self-organizing algorithms and data structures and the results provided by them will be more quick and efficient compared to the present day search engines.

REFERENCES

[1] Sanjay Ghemawat, Howard Gobioff & Shun-Tak Leung, "The Google File System", Proc. The Nineteenth ACM Symposium on Operating Systems Principles, pp. 29-43, 2003.

[2] William Yip & Dr. Liz Quiroga, "Google Page Rank Algorithm", LIS 678 Personalized Information Delivery, Oct 11, 2008.

[3] Mark Levene, "An Introduction to Search Engines and Web Navigation", John Wiley & Sons, Inc., 2010.

[4] Fidel Cacheda, Diego Fernandez & Rafael Lopez, "Experiences on a Practical Course of Web Information Retrieval:Developing a Search Engine", Proc. Second International Workshop on Teaching and Learning of Information Retrieval, 2008.

[5] Curt Franklin, "How Internet Search Engines Work", [online] Available at: http://computer.howstuffworks.com/internet/basics/search-engine.htm

[6] J. M. Kassim & M. Rahmany, "Introduction to Semantic Search Engine", Proc. International Conference on Electrical Engineering and Informatics, 2009.

[7] Pegah Pishva & Mousa Majidi, "Study of HTML Meta-Tags Utilization in Web-based Open-Access Journals", Journal of Information Sciences and Technology, Vol. 22 Number 3(4-2007), 2007.

[8] Gabor Vincze, Zoltan Pap & Robert Horvath, "Peer-to-Peer based distributed file systems", International Journal of Internet Protocol Technology, Vol. 2, Number 2/2007, pp. 117-123, 2007.

[9] L. Plissonneau, J. L. Costeux & P. Brown, "Detailed Analysis of eDonkey transfers on ADSL", Proc. Second Conference on Next Generation Internet Design and Engineering, 2006.

[10] S. Brin & L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine", Proc. the Seventh World Wide Web Conference. Brisbane, Australia.

[11] Wen-Jen Yu, Shrane Koung Chou, "A Bibliometric Study of Search Engine Literature in the SSCI Database", Journal of Software, Vol5, No 12 (2010), 1317-1322, Dec

[12] Ran Hock, (2010), "Major Search Engines – Features Guide", [online] Available at: http://extremesearcher.com/sechart.pdf

BIOGRAPHY

K. Tarakeswar obtained his B.Tech degree from Sri Krishna Devaraya University, Anantapur in the year 2009. He is pursuing his M.Tech in Computer Science and Engineering from Sri Krishna Devaraya University, Anantapur, India. He presented a survey paper at a national level conference.

D. Kavitha obtained her B.Tech degree from Sri Krishna Devaraya University, Anantapur and M.Tech degree from Jawaharlal Nehru Technological University, Anantapur, in the years 2001 and 2005 respectively. She is pursuing her Ph.D. from Sri Krishna Devaraya University, Anantapur, India. She is working as an Associate Professor in the Department of Computer Science and Engineering at G. Pulla Reddy Engineering College, Kurnool, Andhra Pradesh, India. She presented six research papers in international journals and five in national and international conferences so far. Her research areas include Computer Networks and Network Security.