International Journal Of Computer Architecture And Mobility (0012-0187)

Volume 1-Issue 1, November 2012

Study of search engines that travel World Wide Web

Sagar Pandit, Toshi Jain, Rimpal Chugga, Pragya Bagdi

sgrjii37@gmail.com, toshijainsmail@gmail.com, rimpalc123@gmail.com, pragya.bagdi@gmail.com

***Abstract*— All the information present on WWW reaches to the user with a mediator that are search engines. This literature paper is an attempt to extract, an outline the most important software present on internet Search Engine. The paper includes history, types of search engines their working , optimization techniques, common algorithm, comparisons of popular search engines & also sectors for improvement in future.**

***Keywords*— Archie, Crawlers, Spiders, Pigeon Rank's, google Penguin.**

I. INTRODUCTION

A web search engine poised for query fired by user searches through its own databases or open directories, being different from human controlled directories, web search engine's working is controlled by running algorithms on Web Crawlers. Three major steps followed by web search engines are Web Crawling, Indexing, and Retrieving. The difference among various searching tools available on net is due to variations in way of following these three steps.

II. HISTORY

Archie, the very first search tool on internet use to search file names and titles stored in Gopher index systems. Aliweb, Web's second search engine didn't use a web robot, but instead depended on administrator. Jumpstation the first WWW resource-discovery tool viewing similar to present search engines. Later Web Crawler search engine preceded by Lycos, Magellan, Excite, Infoseek, Inktomi, Norhtern Light and Altavista. As I.T industry widens in corporate sense Yahoo became popular in 2000 providing search services based on Inktomi's search engine untill google tied with inktomi and also pigeon ranking used by google promoted it on top. Later yahoo and Microsoft tied [5].

III. TYPES OF SEARCH ENGINES

A. Crawler-based search engines:

Crawling programs are used by these engines to create their listings automatically and form index for future's search base. Web changes can be dynamically caught by crawler-based search engines and will affect how these web pages get listed in search results. Eg are Google, AllTheWeb, AltaVista, etc.

B. Human-powererd directories:

Webmasters submit a short description to the directory for websites, or editors write one for sites they review and these manually edited descriptions will form search base. Therefore, changes made to individual web pages will have no effect on how these pages get listed in search results. Eg are Yahoo directory, OpDirectory and LookSmart.

C. Hybrid search engines:

It is extremely common for crawler-type and human-powered results to be combined when conducting a search. Usually, a hybrid search engines will favour one type of listings over other.

D. Meta-search engines:

User supplied keyword is simultaneously transmitted to several individual search engines to actually carry out search. Search results returned from all the search engines can be integrated, duplicates can be There is another type of search engines that is called meta-search engines.

IV. SEARCH ENGINE ALGORITHM

Basically, a search engine algorithm is used to determine the significance of a web page, and each search engine has its own set of rules [4]. The algorithms, as they are different for each search engine, are also closely guarded secrets, but there are certain things that all search engine algorithms have in common. *1. Relevancy* – search engine algorithm checks for is the relevancy of the page, the algorithm will determine whether the web page has any relevancy at all for the particular keyword. The frequency of the keywords also is important to relevancy. If the keywords appear frequently, but are not the result of keyword stuffing, the website will rank better.

2. Individual Factors – Each search engine has unique algorithms, and the individual factors of these algorithms are why a search query turns up different results on Google than MSN or Yahoo!. One of the most common individual factors is the number of pages a search engine indexes.

3. Off-Page Factors – Off-page factors are such things as click-through measurement and linking. The frequency of click-through rates and linking can be an indicator of how relevant a web page is to actual users and visitors, and this can cause an algorithm to rank the web page highe

V. SEARCH ENGINE ARCHITECTURE

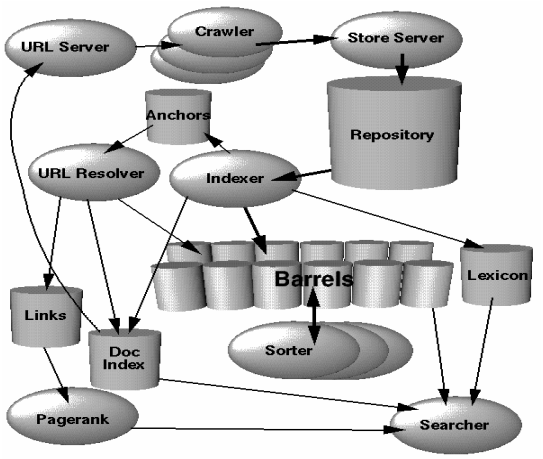

Fig1. Architecture of search engine

TERMS USED IN ARCHITECTURE:

URL SERVER: 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.
CRAWLERS: A *Web Crawler* is a program [6], which automatically traverses the web by downloading documents and following links from page to page. They are mainly used by web search engines to gather data for indexing. Web crawlers are also known as spiders, robots, worms etc. Crawlers are automated programs that follow the links found on the web pages
STORE SERVER: The web pages that are fetched crawled by crawlers are then sent to the store server. The store server then compresses and stores the web pages into a repository
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.
INDEXER: The indexer performs the indexing function. The indexer performs a number of functions. It reads the repository, uncompressed the documents, and parses them.
BARRELS: The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index.
HITS: 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.
SORTER: The sorter takes the barrels, which are sorted by docID and resorts them by wordID to generate the inverted index.
*ANCHOR FILES:*It parses out all the links in every web page and stores important information about them in an anchors file.

VI. BRIEF WORKING OF SEARCH ENGINE

The working start [1] crawler crawls the web pages from the urlserver and stores them into the store server. The store server compresses the web pages and stores them into a repository.Then the indexer reads the repository uncompresses the documents and parses them.The important information about these pages are stored in the anchors file.The urlresolver reads the anchors file. The links database is used to compute PageRanks for all the documents. The sorter takes the barrels, which are sorted by docID and resorts them by wordID 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 wordIDs and offsets into the inverted index. A program called DumpLexicon takes this list together with the lexicon produced by the indexer and generates a new lexicon to be used. The URLresolver reads the anchors file and converts relative URLs into absolute URLs and in turn into docIDs. It puts the anchor text into the forward index, associated with the docID that the anchor points to. It also generates a database of links which are pairs of docIDs. The searcher is run by a web server and uses the lexicon built by DumpLexicon together with the inverted index and the PageRanks to answer queries.

VII. SEARCH ENGINE RELATED TECHNIQUES

A. RANKING TECHNIQUE: PigeonRank's success relies primarily on the superior trainability of the domestic pigeon. The common gray pigeon can easily distinguish among items displaying only the minutest differences, an ability that enables it to select relevant web sites from among thousands of similar pages. When a search query is submitted to search engine, it is routed to a data coop where flash result pages monitors at blazing speeds. When a relevant result is observed by one of the pigeons in the cluster, it strikes a rubber-coated steel bar with its beak, which assigns the page a PigeonRank value of one. For each peck, the PigeonRank increases. Those pages receiving the most pecks, are returned at the top of the user's results page with the other results displayed in pecking order.
B. *SEARCH ENGINE OPTIMIZATION TECHNIQUE*:
Search engine optimization (SEO) is the process of improving the visibility of a website or a web page in a search engine. SEO may target different kinds of search, including image search, local search, video search, academic search, news search and industry-specific vertical search engines. Optimizing a website may involve editing its content and HTML and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing activities of search engines. Promoting a site to increase the number of backlinks, or inbound links, is another SEO tactic. Industry commentators have classified SEO techniques as either white hat SEO, or black hat SEO. White hats tend to

produce results that last a long time, whereas black hats anticipate that their sites may eventually be banned either temporarily or permanently once the search engines discover what they are doing.
White hat SEO ensuring that the content a search engine indexes and subsequently ranks is the same content a user will see. White hat advice is generally summed up as creating content for users, not for search engines, and then making that content easily accessible to the spiders, rather than attempting to trick the algorithm from its intended purpose. Black hat SEO attempts to improve rankings in ways that are disapproved of by the search engines, or involve deception. One black hat technique uses text that is hidden, either as text colored similar to the background, or positioned off screen. Another method gives a different page depending on whether the page is being requested by a human visitor or a search engine, a technique known as cloaking.
.In February 2011, Google announced the "Panda" Technique. which penalizes websites containing content duplicated from other websites and sources.
.In April 2012, Google launched the Google Penguin update the goal of which was to penalize websites that used manipulative techniques to improve their rankings on the search engine.

C. INDEXING TECHNIQUE OF DIFFERENT SEARCH ENGINE:

Following are indexing techniques [3] used by various search engines

GOOGLE: The heart of google indexing technique is "PageRank". PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. Google interprets a link from page A to page B as a vote, by page A, for page B. Google also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."

YAHOO: Yahoo the most popular hierarchically organized search engine uses robots to discover the sites and humans are relied upon for indexing, index includes URL, HTML title tags, very short description of the site.

ALTAVISTA: It uses crawlers to visit every site on the web, indexes all the site they find there. It also uses meta tag for indexing.

LYCOS: Lycos indexes the title, URL, headings and subheadings and the first twenty lines of text < reindexing it's database every two weeks>.

EXCITE: It uses robots to do full text indexing; it also employs "Intelligent concept extraction" which relies on clustering of words to locate the presence of concepts, It indexes more than first level of heading.

INFOSEEK: It uses robot to do full text indexing, meta descriptor tags are also indexed, indexes third and fourth level also.

ALTAVISTA: It uses crawlers to visit every site on the web, indexes all the site they find there. It also uses meta tag for indexing.
LYCOS: Lycos indexes the title, URL, headings and subheadings and the first twenty lines of text < reindexing it's database every two weeks>.
EXCITE: It uses robots to do full text indexing; it also employs "Intelligent concept extraction" which relies on clustering of words to locate the presence of concepts, It indexes more than first level of heading.
INFOSEEK: It uses robot to do full text indexing, meta descriptor tags are also indexed, indexes third and fourth level also.

VIII. COMPARISONS OF PRESENT SEARCH ENGINES

Search engines can be compared on basis of several features and services provided by them, for a user time constraint and relevant results are most important [2]. The below table includes various properties and compares four popular search engines Altavista, Google, Bing, Yahoo. User can also use various search engines optimization tools for judging various properties of a search engine such as domain age, domain range, domain time, domain relevant data mining, etc .

TABLE I
COMAPRION OF SEARCH ENGINES

	Altavista	Google	bing	Yahoo!
WWW	YES	YES	YES	YES
Word in URL	YES	YES	YES	YES
Languages?	All or English	35	38	37
Image search?	YES	YES	YES	YES
News?	YES	YES	YES	YES
Multimedia?	YES	Separate function	NO	YES
Limit by file format?	YES	YES	YES	YES
Boolean	YES	YES	YES	YES
Proximity	NO	YES (Using *)	NO	NO
Wildcards (three * mice)	YES	YES	YES	YES
Search in Domain	YES	YES	YES	YES

Search in Title	YES	YES	YES	YES
Capitalisation?	NO	NO	NO	NO
Sort Results?	NO	NO	YES (date, popularity, exactness)	NO
Search in results	YES	YES	YES	YES
Group results?	NO	NO	NO	NO
Index size given?	YES	YES	YES	NO
Personalised?	NO	YES	NO	YES
Geographic Specific?	NO	YES	YES	YES
Thumbnails?	NO	NO	NO	NO
Limit by date?	YES (excellent)	YES (poor)	NO	YES (poor)

IX. CONCLUSIONS

The search engines are an important part for surfing the net, and making more users to connect to the internet for which several loop holes should be improved. There should be security feature to block the urls having malicious content. Also there should be an age constraint for searching regarding keyword fired by a user that will not only give results in quick time but also will relevant results.

REFERENCES

[1]Monica Peshave, "How Search Engine Work and Web Crawler Application", University of Illinois at Springfield.

[2] www.yuanlei.com/studies/articles/is567-searchengine/page2.htm

[3] http://www.searcheasy.htmlplanet.com/indextech.html

[4] http://www.brickmarketing.com/define-search-engine-algorithm.html

[5] www.wikipedia.org

[6] Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine,
Stanford University, Stanford, CA 94305,US