Google: A Case Study (Web Searching and Crawling)

Krishan Kant Lavania, Sapna Jain, Madhur Kumar Gupta, and Nicy Sharma

***Abstract*—Today search engines are becoming necessity of most of the people in day to day life for navigation on internet or for finding anything. Search engine answer millions of queries every day. Whatever comes in our mind, we just enter the keyword or combination of keywords to trigger the search and get relevant result in seconds without knowing the technology behind it. I searched for "search engine" (I intentionally mis-spelled it) and it returned 68,900 results. In addition with this, the engine returned some sponsored results across the side of the page, as well as some spelling suggestion. All in 0.36 seconds. And for popular queries the engine is even faster. For example, searches for World Cup or dance shows (both recent events) took less than .2 seconds each.**

To engineer a search engine is a challenging task. In this paper, we present Google, most popular search engine, and in-depth description of methods and techniques that the Google uses in searching. Different search engines use different techniques for searching and algorithm to rank the pages. At the end of the paper, We compare major search engines such as Google, Yahoo, Msn etc.

***Index Terms*—Google, yahoo, msn, internet, world wide web.**

I. Introduction

All Google, it is one of the most renowned terms in Internet world. Google's brand has become so universally recognizable that now days; people use it like a verb. For example, if someone asks "Hey what is the meaning of that word? The answer is "I don't know, goggle it".

Google Inc. is an American public corporation specializing in Internet search technology and many products. Google's mission is based on the fundamentals of collaborative teamwork. Its main motive is to organize the world's information and make it universally accessible and useful. Google Company was founded by Larry Page and Sergey Brin while studying PHD at Stanford University in 1998[1].

The main idea behind the Google's search engine is that the web can be represented as a series of interconnected links, and its structure can be portrayed by a giant and complex mathematical graph. Google's innovative search engine technologies connect millions of people around the world with information every second.

The name "Google" derived from the word "googol "which refers to 10^100.

Manuscript received September 15, 2012; revised November 29, 2012.

K. K. Lavania, S. Jain, and M. K. Gupta are with the Department of Information Technology, Arya Institute of Engineering & Technology, Jaipur, India (e-mail: k@lavania.in, sapna_sona1990@yahoo.com, madhur.neelesh@gmail.com).

N. Sharma is with the Department of Computer Science Engineering, Sir Chotu Ram Institute of Engineering and Technology, Meerut, India (e-mail: nicy.sharma1991@gmail.com).

II. Web Search Engine: Google

Developing a search engine that matches even to today's web world presents many challenges before us. Storage technologies must be used optimized to store the documents and the indices. To gather the up to date web documents and information, fast crawling technology (browsing the World Wide Web) is required and it ensures that we can find latest news, blogs and status updates. The indexing system must process hundreds of gigabytes of data efficiently. Speed is the major priority in searching. Queries response time must be very faster.

Google is designed to scale well to keep up with the growth of web. It gives exactly what we want. For fast and efficient access, its data structures are optimized. In addition to smart coding, on the back end it developed distributed computing systems around that globe that ensure fast response times.

A. System Anatomy

First, we will provide a high level discussion of the Google's architecture. Finally, the major methods: crawling, indexing, and searching will be examined in depth.

1) Google architecture overview

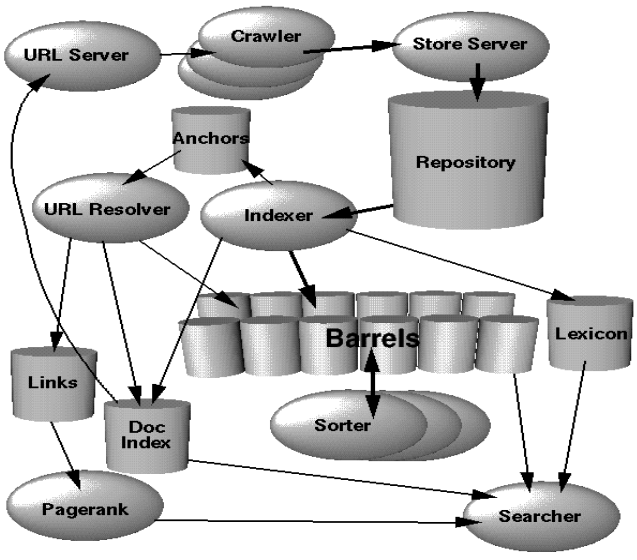

Fig. 1. High level Google architecture.

In Google, the web crawling (downloading of web pages) is done by several distributed crawlers ,which is a computer program that browses the World Wide Web by employing many Computers.URLserver sends lists of URLs(uniform resource locator) 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 docID which is assigned whenever a new URL is parsed out of a web page. The indexer and the sorter perform indexing functions and read the repository, uncompress the documents, and parse 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 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 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 by the searcher. The searcher is run by a web server and uses the lexicon built by DumpLexicon together with the inverted index and the Page Ranks to answer queries[2].

B. System Features

Google's most important feature is Page Rank, a method that determined the "importance" of a webpage by analyze at what other pages link to it, as well as other data.

1) PageRank: Bringing order to the web

Search engine searches for the web pages or documents available on World Wide Web and returns the relevant results. It is not possible for a user to go through all the millions of pages presented as output of search. Thus all the pages should be weighted according to their priority and represented in the order of their weights and importance. PageRank is an excellent way to prioritize the results of web keyword searches. PageRank is basically a numeric value that represents how much a webpage is important on the web.

2) Description of pagerank formula

PageRank is calculated by counting citations or backlinks to a given page. In the paper "The Anatomy of a Large-Scale Hypertextual Web Search Engine" founders of Google, Sergey Brin and Lawrence Page defined PageRank as:

"We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85 …… C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) "

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one[3].

PageRank or *PR(A)* can be calculated using a simple iterative algorithm.

3) Anchor text

The anchor text is defined as the visible, highlighted clickable text that is displayed for a hyperlink in an HTML page. Search engine treat the anchor text in a different way. The anchor text can determine the rank of the page. It provides more accurate descriptions of web pages that are indicated in anchors than the pages themselves. Anchors may exist for documents which cannot be indexed by a text-based search engine, such as images, programs, and databases. Anchor text is ranked or given high weightage in search engine algorithms. The main goal of search engines is to bring highly relevant search results and anchor text can help by providing better quality results[4].

4) Other features

Apart from PageRank calculations and the use of anchor text, Google has several other features.

- First, it has location information for all hits, a set of all word occurrences so it makes extensive use of proximity or probability in searching.
- Second, Google keeps information about some visual presentation details such as font size of words, Words in a larger or bolder font are weighted higher than other words.
- Third, full raw HTML of pages is available in a repository.

C. Crawling

Web crawling or spidering is a process of browsing the World Wide Web in a methodical, automated manner by software program called Web crawler running on a search engine's server .web crawlers are also called indexers, bots, Web spiders, ants, Web robots.

Crawlers start by fetching a few web pages, then they follow all the links contained on those pages and fetch the pages they point to and so on, it is a recursive process and it produce many billions of pages stored across thousands of machines. Running a web crawler is a challenging task because Crawling is the most fragile application since it involves interacting with hundreds of thousands of web servers.

Web crawlers or spiders are mainly used to create a copy of all the visited pages for later processing by a search engine. Search engine will index the downloaded pages to provide fast searches. Other than this work, Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links, validating HTML code etc. Crawlers can also be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for spam)[5].

D. Indexing

The Spider collects the data and stored in "index". When we are doing searching, we are not searching the web, but the cache of the web or index provided by search engine. Rather than searching each page for a query phrase or keywords, Search Engine "inverts" the index to produce a table of the documents containing particular words. For example for the search 'online shopping' the search engine might find the word "online" in documents 13, 28, 49, 54 and 99 and the word "shopping" in documents 13, 23, 49 and 57 as follows.

The Index uses the intersection between different postings lists for different words. The "index" is distributed across many servers to make searching more efficient. Google estimates that for each search, over 500 servers work together to find the best documents. Other than this, Google's Search Engine store many other aspects characterizing a page in the index files for example a document's title, Page Rank, trust or authority, meta description, spam rating, etc . to make the look up more efficient.

E. Ranking or Scoring

The indexing process has produce all the pages that include particular words in a query enter by the searcher, but they are not sorted in terms of importance or relevance. Ranking of the document is measured to provide the most relevant WebPages for the search query entered.

Evaluation of relevance is based factors, they are:

- Page Rank.
- Authority and trust of the pages which refer to a page.
- The number of times the keywords, phrases and synonyms of keywords occur on the page.
- Spamming rate.
- The occurrence of the phrase within the document title, URL (Uniform Resource Locator).

F. Query Request and Result Serving

The Google's Search Engine interface accepts the searchers query. IP address of the user give detail about the user's location and the query is then passed to data centre for searching and processing. This process occurs in real-time and return a sorted list of relevant Web Pages and these Web Pages are then displayed on the Search Results Page. Google refers this approach as Google Universal Search [6].

III. Major Search Engine: A Comparison

Today there are many search engines available to web searchers. What makes one search engine different from another? Following are some important measure.

- The contents of that database are a crucial factor determining whether or not you will succeed in finding the information we need. Because when we are doing searching, we are not actually searching the Web directly. Rather, we are searching the cache of the web or database that contains information about all the Web sites visited by that search engine's spider or crawler.
- Size is also one important measure. How many Web pages has the spider visited, scanned, and stored in the database? Some of the larger Search Engines have databases that are covering over three billion Web pages, while the databases of smaller Search Engines cover half a billion or less.
- Another important measure is how up to date the database is. As we know that the Web is continuously changing and growing. New Websites appear, old sites vanish, and existing sites modify their content. So the information stored in the database will become out of date unless Search Engine's spider keep up with these changes.
- In addition with these, the ranking algorithm used by the Search Engine determines whether the most relevant search results appear or not towards the top of results list.

In this paper I will be highlighting the 'The Big Four'. They are Google, Yahoo!, MSN and Ask Jeeves. Google, Yahoo & MSN account for more than 95% of all searches in the world.

Fig. 2. Google logo.

Google has been in the search game a long time, it has the highest share market of Search Engine (about 81%) [7].

1) Web Crawler-based service provides both comprehensive coverage of the Web along with great relevancy.
2) Google is much better than the other engines at determining whether a link is an artificial link or true editorial link [8].
3) Google gives much importance to Sites which add fresh content on a regular basis. This is why Google likes blogs, especially popular ones [7].
4) Google prefer informational pages to commercial sites.
5) A page on a site or sub domain of a site with significant age or link can rank much better than it should, even with no external citations.
6) It has aggressive duplicate content filters that filter out many pages with similar content.
7) Crawl depth determined not only by link quantity, but also link quality. Excessive low quality links may make your site less likely to be crawled deep or even included in the index [7].
8) In addition we can search for twelve different file formats, cached pages, images, news and Usenet group postings.

Fig. 3. Yahoo logo.

1) It shares the second largest share market of the search engine (about 12%). Yahoo has been in the search game for many years [7].
2) When it comes to counting back lings, Yahoo is the most accurate search engine [7]
3) Yahoo is better than MSN but near as good as Google at determining whether a link is artificial or natural.
4) Crawl rate of the Yahoo's spiders is at least 3 times faster than Google's Spiders [8].
5) Yahoo! tends to prefer commercial pages to informational pages as comparing with Google [7].
6) At Yahoo search engine "exact matching" is given more importance than "concept matching" which makes them slightly more susceptible to spamming.
7) Yahoo! gives more importance to meta keywords and description tags [7].

Fig. 4. MSN logo.

1) MSN has the share of 3% of the total search engine market [7].
2) MSN Search uses its own Web database and also has separate News, Images, and Local databases.
3) Its strengths include: this large unique database, its query building "Search Builder" and Boolean searching, cached copies of Web pages including date cached, and automatic local search options.
4) The spider crawls only the beginning of the pages (as opposed to the other two search engine which crawl the entire content) and also the number of pages found in its index or database is extremely low[8].
5) It is bad at determining if a link is natural or artificial in nature.
6) Due to sucking at link analysis they place too much weight on the page content.
7) New sites that are generally untrusted in other systems can rank quickly in MSN Search. But it also makes them more susceptible to spam [7].
8) Another downside of this search engine is its habit of supplying the results based on geo-targeting, which makes it extremely hard to determine if the results we see are the same ones everybody sees [8].

Fig. 5. ASK Jeeves logo.

1) The Ask search engine has the lowest share (about 1%) out of the total search engine market [7].
2) Ask is a topical search site. It gives more importance to sites that are linked to topical communities [7]
3) Ask is more susceptible to spamming [7].
4) Ask is smaller and more specialized than other search engines, it is wise to approach this engine more from a networking or marketing perspective.

IV. Conclusion

Today the amount of information available on the Web is growing rapidly. Search Engine technology had to scale according to the growth of the Web. Web searching technology has been evolving very rapidly and will continue to evolve. Google is designed to be a scalable search engine. The primary goal of the Google Search Engine is to provide high quality search results over. Google employs a number of techniques or methods to improve search quality including page rank calculation, anchor text, and many other features. Google have some differences than Yahoo! Yahoo! has some differences than Ask. Ask also have differences different than MSN Live. This study quantifies the similarities & the differences among the leading Search Engines.

References

[1] Keyword Density Tool. [Online]. Available: www.google.com/corporate/tech.html
[2] S. Brin and L. Page, "The anatomy of a large-scale hyper textual web search engine," *Proc. Seventh World Wide Web Conf. (WWW7), International World Wide Web Conference Committee (IW3C2)*, 1998.
[3] Seminar Report on Page Ranking Technique in Search Engine by Phapale Gaurav S. [05 IT 6010].
[4] N. Eiron and K. S. McCurley, "Analysis of Anch.or Text for Web Search," IBM Almaden Research Center.
[5] WebCrawler - Wikipedia, the free encyclopedia. [Online]. Available: en.wikipedia.org/wiki/WebCrawler.
[6] Google and Beyond: Finding Information Using Search Engines, and Evaluating Your Results Elizabeth Geesey Holmes University of Georgia School of Law Library.
[7] Yahoo for msn. [Online]. Available: www.sunsigndesigns.com/what-is-the-difference-between-google yahoo-and-msn-how-do-i-rank-well-for-all-three
[8] Different googles yahoo - WEB statsdomain overview for keyword. [Online]. Available: www.domaininform.net/differencesbetween3bigsearchengines.html

Krishan Kant Lavania was born in the city of Taj – Agra on 04th May 1978, completed his M. Tech. (Computer Science) from JNRV University, Udaipur, India. After that he works as programmer at Real Choice Infotech, Kanpur, then with Galgotia Institute of Management &Technology as Lecturer, then with Gyan Vihar School of Engineering &Technology as Reader, then after with Rajasthan College of Engineering for Women as Head of Department (Computer Engineering), presently he is working with Arya Institute of Engineering & Technology. He has published research papers in various international & national journals & conferences .Prof. Lavania is the active member of various national and international organizations like, Indian Society for Technical Education, International Association of Engineers, Universal Association of Computer and Electronics Engineers and World Academy of Science, Engineering and Technology. Prof. Lavania is also in the editorial board of International Journal of Soft Computing and Engineering (IJSCE).

Sapna Jain obtained her Bachelor's of technology degree in Information Technology from Arya Institute of Engg. & Technology,Jaipur (Rajasthan), India in 2011.She is currently working as Programmer Analyst Trainee at Cognizant Technology Solutions India Private Limited and currently working with Mainframe Computers in Insurance Domain in Chennai (India). She is born and brought up in Jaipur, Rajasthan (India).She has 7 month work experience in Java Technologies. Google: A case study (web Searching and crawling) is Author's first research paper.You can contact her on Facebook and drop a mail at Sapna_sona1990@yahoo.com.

Madhur Kumar Gupta did his engineering from Arya Institute of Engg. & Technolgy, Jaipur (Rajasthan) with "Information Technology" stream in 2011. He is presently working with Accenture services pvt. Ltd., Pune(India), as an Associate Software Engineer. Trained in SAP-ABAP with IS utilities. He born and brought up in Jaipur, Rajasthan (India). He is having 10 months of experience in SAP-ABAP. Google: A case study (web Searching and crawling) is Author's Second Paper ; First one was published in CONIAPPS- XIII, held at 4,5 May,2011 in Dehradun, India. If needed, you can reach out to him at madhur.neelesh@gmail.com.

Nicy Sharma was born in the city of Taj – Agra on 02 April 1991, pursuing her bachelor's degree in computer science & engineering from Department of Computer Science Engineering, Sir Chotu Ram Institute Of Engineering and Technology, Meerut, India.