Case Study on Google Search Engine

Introduction:

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. 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.

Major Obstacles:

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.

Features of Search Engine :

Google's most important feature is Page Rank, a method that determined the "importance" of a webpage by analyse 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. 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

- 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.

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 .

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).

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).