Anatomy of a large-scale hyper-textual web search engines

By: Diana Zaidenshtein
18575013
Anatomy of a large-scale hyper textual web search engines

With the huge amounts of information being created on the internet every minute by millions of people from all around the world, searching for a particular piece of information an individual needs is hard if not impossible. To help assist everyone in this task, systems known as “Search engines”, such as Google have been created. However, despite the importance of search engines, there has not been a heavy amount of academic research done into the area.

Google uses a relatively complex method to index the huge web efficiently. This method is known as “crawling”, where the system will constantly search and index the internet all the time therefore maintaining the most up-to-date references. “Because humans can only type or speak a finite amount, and as computers continue improving, text indexing will scale even better than it does now” (2000; Brin, Page) For many, Google has a reputation as the best search engine, and is the most commonly used internet search engine available, being used by most people. Tens of millions of queries are asked and answered every day.

Handling vast amounts of searches being performed all the time by millions of users, the technology behind Google must be state of the art. The three most important things to Google are availability, response time, and quality results. In order to meet these goals, Google is a system full of original thoughts and creative solutions to meet some tricky problems.

Inside a search engine as large as Google, things are very complex and sophisticated. Within Google, three major things are always being done: Crawling the Internet for up-to-date information, processing the data that is found from this process, and returning the search results back to the users. The task of ‘Crawling the Internet’ is constantly being performed, and the results to be handed back to the users are being retrieved from an extremely large database.

Web crawling is done by a “…fast distributed crawling system” (2000; Brin, Page). A URL server is in charge of sending lists of URL’s to be fetched by the crawlers. The
web pages that are found by the crawlers are then sent to the ‘Store server’. This server is then given the task of compressing and storing these Web pages in a repository in a process known as ‘Indexing’. Google then uses some sophisticated algorithms to process the knowledge, then compare and deduce what is good information, and what is bad information on the web today. The database that is created during this whole process is then used to generate the search results for the user. Around the end of November 1997, only one of the top four commercial search engines would find itself. This was a rather strange aspect of these major search engines, but has now been fixed.

When measuring the effectiveness of a search engine, the most important measure is the quality of the search results. The “Perfect search engine," defined by co-founder Larry Page (2004, Google Technology Overview) as something that, "understands exactly what you mean and gives you back exactly what you want”. Google has shown itself to produce better search results than the major commercial search engines, for most types of searches. Another important aspect of a search engine is the ability for it to scale cost effectively to the size of the web as it grows. Engineers working at Google maintain their server base using low cost, redundant PC hardware.

For all search engines, crawling the web and indexing pages efficiently is very important. This means that information can be kept up to date easily, and any major changes to the system can be tested relatively fast. In the beginnings of Google, it took about “nine days to download the 26 million pages (including errors) being indexed. However, after some minor changes, the system started to run more smoothly and more efficiently. The last 11 million pages to download were obtained in just 63 hours” (2000; Page, Brin). The current version of Google answers most user queries in between one and ten seconds.

Google uses a rather complex algorithm to rank pages. “Page Rank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value”(2004, Our Search; Google Technology) This method is not as easy to fool as the method other common search engines use. When a search is performed by Google, a list of pages is retrieved by the system. The results are then ranked, based on the number of other websites linking to each page. The page with
the most links from other sites will be ranked the highest. Many of the other major
search engines base their ranking system on the number of the search words that occur
in each page. The higher the word occurs, the higher the rank. This method of ranking
is easy to fool.

Nowadays, it is much harder to create a web search engines from scratch than it was
three years ago. To cope with the billions of pages on the web, and the many hundreds
and thousands being created every day, fast web crawling technology is needed to
gather the documents and keep them up to date. With so much information, there is an
issue with the amount of storage space available. Therefore the space must be used
efficiently to store indices of the pages, and sometimes the actual documents
themselves. “When Google was first designed, the rate of growth and technological
changes were considered” (2000; Page, Brin) and allowed Google to scale and handle
the extremely large data sets required.

There are automated search engines, such as vivisimo.com that rely on keyword
matching, and therefore return too many low quality matches. Google goes far beyond
guessing which pages are relevant to the searches. Google’s algorithms weigh in a
tremendous amount of factors, which are capable of producing excellent results. No
one knows exactly how many factors are used, as Google is constantly updating and
tweaking its search algorithms.

What's more interesting, is that Google constantly updates and tweaks its search
algorithms. Spam is pushing Google engineers towards perfection of their search
algorithms. It is amazing how precise Google is, and how close it to reading your
thoughts sometimes. Sometimes you can see really unusual things in the search
results; you're probably lucky to peek at an obscure feature that will become
extremely popular in two to three months.
References:

Google corporate information: Technology overview (2004),


Anatomy of a large-scale hypertextual search engine (2000); Page, Larry; Brin, Sergey