Discussion Report: The Anatomy of a Large scale Hypertextual Web Search Engine

Azbayar Demberel
Department of Computer Science
asic@cs.duke.edu

1 The World Before Google

With the birth of the World Wide Web (WWW), the usage of the Internet has grown dramatically. One of the first search engines, the World Wide Web Worm, had an index of around 110,000 web pages and it received about 1500 queries per day by 1994 - only 4 years after WWW's creation. By November 1997, Altavista, one of the more popular web search engines during that time, claimed to handle roughly 20 million queries per day and indexed over tens of millions of web pages. (As of January 2005, the number of web pages grew to over 11.5 billion and Google claims to process nearly 90 million queries per day.)

Although the number of the users accessing the web has been growing rapidly, the way they access it remained pretty much the same. People tend to surf the web using a graph of links, often starting with search engines or popular web pages like Yahoo.com and continue navigating using their lists of topics or other links.

The automated searches used plain keyword matching or other simple techniques which returned too many low quality results or were very easy to manipulate. For example, some search engines used only the font size of the text to determine the importance of the search result. Hence, advertisers exploited this weakness by creating extra large redundant text messages so that their web page would always appear in the front of search result. For example to gain attention of possible computer buyers, they would insert every possible word connected to computers in a large font like "Dell", "Vaio", "Acer", etc to their web page. In this case if a person searches for Sony Vaio laptops, for instance, he will see the advertiser's web site in the front of the search result instead of Sony Vaio's official web site or other more prominent shopping web sites. Some advertisers would simply pay money to the search engines, so that their web sites could appear on the top of search results.

The human maintained lists, on the other hand, have an advantage of efficiently listing the popular topics but this approach is expensive to build and maintain, slow and difficult to improve and in many cases it is not objective, as a person has to decide which topics to include in the list. These human maintained lists are also not able to scale with the growth of the web, as a person is limited in his/her ability to look to the enormous amount of documents.

S. Brin and L. Page presented a new search engine - Google that addressed many of these problems: it provides high quality search results by using additional information in the hypertext and other mechanisms like page ranking systems, and is able to scale well by using efficient data structures and design.

2 Design Goals

The main goal of Google was to improve the search quality. The ability to find anything on the web is not sufficient anymore. Intuitively, the average number of search results would increase proportionally to the number of the web pages. The users, however, are not interested in the quantity of the results; they are more interested in the quality. As they state in the paper, "People are still only willing to look at the first few tens of results". Hence it was becoming increasingly important to distinguish the most relevant documents from the other thousands of less relevant ones.

It was also important to consider the rate of the growth of the web so that the new search engine could scale well with the future web. As the web is growing hour by hour, the search engine must have a fast crawling technology to gather the documents and keep them up to date. It must make efficient disk usage to store the indices and the documents of billions of web pages. And it is also necessary to have good data structures so that queries can be handled quickly.

Another design goal was to build an architecture that can support research activities. As one of the ways to provide this feature, Google stored all the actual documents when it crawled through the web pages so that other researchers could quickly process and experiment a large amount of information and produce interesting results.

3 Google's new ideas

3.1 Use of Proximity

Google uses several features to assure high quality search results. For example it uses location information for all hits to make extensive use of proximity in search and it uses some visual presentation details like font size, as words in a larger or bolder font tend to be more important and hence are weighted higher than other words. The proximity of the search keywords is also important because for instance when a person searches for “computer science”, he/she is not interested in web sites about ”computers” or ”sciences” or any other person's personal web page containing a sentence like ”Yesterday my computer broke down... I watched a science fiction movie”.

3.2 Page Rank

A page with many inlinks can be considered to be important because many people point to this page. Counting only inlinks, however, can be easily manipulated. For example one can create several small web pages and make a mesh of links, connecting each to one another. So it is important to distinguish the characteristic of the linking web page. For example, a link from Yahoo.com is not the same as a link from an average Joe's home page.

As you can see this notion is similar to academic papers' citations: The more citations a paper has or the more distinguished authors' papers cite it, the more importance or quality the paper probably has.

Google uses this idea of counting inlinks but with the extension of (a) not counting links from all pages equally, and by (b) normalizing by the number of links on a page, to determine the importance of the web page, which is called PageRank.

Page rank can also be seen as a model of user behavior. The probability that a random user, given a random web page as a starting point, and by either clicking on links or by restarting at another random web page, reaches that certain web page, defines the ”Page rank” of that page. So clearly, if many web sites point to that web page, the probability that it will get hit will be higher. Also if the page is pointed from a well-known/popular web site it will also have a higher probability to get hit.

The mathematical interpretation of page rank will be as follows:

If page A has T1 ... Tn pages pointing to it, the Page rank (PR) of page A equals:

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

where ”d” is a damping factor that can be set between 0 and 1, and ”C(x)” is the number of links going out of page x.

3.3 Anchor Text

The designers of Google paid special attention to anchor texts, i.e. the texts of the links. It is possible to get more accurate results because anchor text often have better descriptions of the web pages that they point to. For example, someone would never make a link to michaeljordon.com with an anchor text sumo wrestler or Chevrolet. They are more likely to name the link ”air jordan”, ”#23”, or ”the greatest basketball player ever”, etc.

In addition, anchor texts provide an opportunity to find results which cannot be found by plain text search, like media files, programs, etc. To continue our previous example, if someone posted a video clip of Michael Jordan’s best dunk shots, the only way we can find this file is if the link to the file had an anchor text like ”MJ’s dunk shots” etc.

This idea of propagating anchor text to the page it refers to was first implemented in the World Wide Web Worm. They used it to help search non-text documents and expand the search coverage. Google extended their idea to get better quality results.