Software Engineering 4C03

Research Paper: Google™ Servers

Researcher: Nathan D. Jory
Last Revised: March 29, 2004
Instructor: Kartik Krishnan
Introduction

The Google™ search engine is a powerful and highly accurate search engine tool used by many individuals and corporations alike. The name Google™ is derived from the word googol which is a prefix denoting the number one with one hundred zeros\(^1\). This research paper will attempt to highlight some of the technologies used to make Google™ such an effective search engine. A brief background and history of Google™ will be provided followed by an analysis of the hardware and software technology used by Google™. A conclusion and sources will follow.

History

The Google™ search engine, originally known as BackRub, began by Larry Page and Sergey Brin in an attempt to alleviate the difficulty of retrieving and prioritizing information from an immense set of data. Naturally, the Internet was an excellent target for their research. After many hardships in attempting to sell their technology, Page and Brin decided to create Google™ Incorporated in 1998 with initial investments of almost one million dollars. Eventually, after a wildfire of investments in their technology from major firms across the world the Googleplex, the current home to the massive search engine, was founded and in September 1999 the beta label was removed from Google™ and its successes have only increased since then.\(^2\)

Technology

Server Technology

After using Google™ and discovering its incredible speed and accuracy, one might think that the technology behind the Google™ search engine must be tremendous. One might imagine a super computer crammed full of the most up-to-date server technology performing high speed calculations to satisfy the public’s searching needs. However, much to one’s surprise, this is certainly not the case. In fact it is very much the opposite. It has been discovered in the last ten years that many technologies need not be state-of-

---

1  Google History: [http://www.google.com/corporate/history.html](http://www.google.com/corporate/history.html)
2  Ibid.
the-art and, as a repercussion, very expensive. This is certainly true of the server technology used to power the Google™ search engine. As a goal of the Google™ project, Page and Brin set out to find a technology solution that suited their needs as students but was still highly effective. Rather than purchasing a few very powerful machines they began to network and coordinate a large number of low-end technology machines.\(^3\) This network layout requires incredible software coordination. Often, because the technology is lower-end, a great deal of redundancy is required on the servers in order to prevent loss of performance in the event of a server failure. It has been recently reported that the number of servers within the Google™ empire is now over one hundred thousand machines and is considered by some to be the largest computing facility to date.\(^4\)

**Software Technology**

In order to develop a better understanding of how Google™ and many search engines develop their databases of cached links, a commonality between all search engines has developed. The primary software element in a search engine is known as the 'spider.' These software spiders 'crawl' accessing and following the web of links that are throughout the Internet. As they trace their way from the Google™ server network, they develop a database of information about the pages found and the links that they contain.\(^5\) It is this information that is the base for the search engine results. It is also important to note that if a page has no external pages that link to it, it is impossible for it to be discovered by a search engine spider and these pages must be manually inserted into the search engine databases to be discovered.

A number of unique software technologies are used in the Google™ search engine that set it apart.. The technology that distinguishes Google™ from other search engines is its PageRank™ algorithms. Google™'s servers develop, maintain, and update a database in almost an artificial intelligence sort of manner which maintains a ranking system of pages based on the number of links to that page. It also evaluates the ranking of the pages that link to a particular page and this evaluation also contributes to the ranking

---

3 Google History: [http://www.google.com/corporate/history.html](http://www.google.com/corporate/history.html)
5 Search Engines: [http://www.lib.berkeley.edu/TeachingLibGuides/Internet/SearchEngines.html](http://www.lib.berkeley.edu/TeachingLibGuides/Internet/SearchEngines.html)
of the linked page.\textsuperscript{6}

The Google\textsuperscript{TM} algorithms used to crawl the web are sophisticated and a general formula was developed by Page and Brin to rank the importance of a page given a particular search criteria.

“PageRank is defined as follows:

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. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

\[
PR(A) = (1-d) + d \left( \frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(Tn)}{C(Tn)} \right)
\]

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.” \textsuperscript{7}

It is this iterative formula that provides the basis to the PageRank\textsuperscript{TM} system provided by Google\textsuperscript{TM}.

\textbf{Conclusion}

Google\textsuperscript{TM} is one of the fastest and most effective search engine technologies on the Internet at present. It runs on a massive collection of thousands of servers running low end technologies to reduce cost and provide similar performance to a supercomputer server. It is the incredible combination of the coordination of these servers through inter networking and the sophisticated software algorithms developed by Page and Brin that will help Google\textsuperscript{TM} to maintain its champion status in the realm of the search engine wars.

\textsuperscript{6} Google Information: \url{http://infolink.schoolnet.org.za/google.htm}
\textsuperscript{7} The Anatomy of a Search Engine: \url{http://www-db.stanford.edu/~backrub/google.html}
Sources


