

- Yahoo does not blend hardware and software engineering as Google does. Like Microsoft, getting more storage and more speed from Yahoo's patchwork of servers is expensive. Speed is bounded by the operating system developer and by the cost of branded hardware from IBM, Dell and other vendors. Google uses commodity "white box" servers and its own optimised operating system.
- Yahoo has its data in a state of organization metaphorically identical to the structure of the Balkan states. In a word, fragmented. A lack of data cohesiveness limits Yahoo's ability to know its customers. More importantly, the lack of usage data integration makes it difficult for Yahoo to answer some advertisers' demands for site-wide demographic and usage information.
- Yahoo is neither a technology nor an information company. It is a media company. Yahoo is emerging as a Hollywood mogul's version of a digital media company. It is not that Yahoo's business model is wrong. The situation at Yahoo is that a layer of staff has been added to intermediate between the goals of the executive office and the engineers.

Google has its cart loaded appropriately. Its mule has its four hooves on *terra firma*. The drivers know roughly where to take the load of products and services.

In terms of the broader competitive environment, few companies are poised to beat Google's mule cart over the short term. In the context of business, 20 years is a long time. Google, however, seems to have the mule cart to win in the next 12 to 24 months. After that, all bets are off. Nevertheless, Google's advantages and its lack of legacy software to spend money to support is a plus in comparison with Microsoft. Google's integrated Googleplex gives it an advantage selling advertisements to advertisers who want data about usage patterns. Yahoo has a team in India struggling to make Yahoo's data repositories talk to one another. Google just asks, and the Googleplex spits out data at the same mind-boggling speed Google Search works in dozens of languages.

PageRank

PageRank is explained by Sergey Brin and Larry Page.¹² The abstract, like the paper, is direct:

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at <http://google.stanford.edu/>

To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation,

12. The paper is located at www-db.stanford.edu/~backrub/google.html. The patent for PageRank is held by Stanford University, No. 6,285,999. The annex to this monograph contains a selected list of patents and an abbreviated discussion of Google's intellectual property initiatives.

creating a web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale web search engine – the first such detailed public description we know of to date.

Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

A search of the term *PageRank* on Google on June 12, 2005, yielded about seven million results. To summarize the core technology of one of the dominant search-and-retrieval companies is difficult. Boiling the formulae and millions of words written about PageRank, there are several key points that warrant restating:

- 1 PageRank is a voting system. Votes equal popularity. The more votes a Web page gets, the higher it will be ranked in a Google results list.
- 2 Links from other highly-ranked sites boost a Web page's PageRank significantly.
- 3 PageRank rewards substantive content that is frequently updated with a ranking boost.
- 4 PageRank lowers the relevance score or *downchecks* a Web page for shoddy code, tricks to fool Mother Googlebot, and dead links, among other Web master indiscretions.

Several technologists with a fixation on Google have published detailed analyses about how Google's PageRank algorithm works. Ian Rogers has a lucid description for those with a mathematical bent. An equally detailed explanation comes from the keyboard of Phil Craven. His analysis offers notes, tips, and a more approachable narrative style than some other explanations.¹³

Stepping back from the iterative and recursive calculations that make PageRank tick, what did PageRank deliver that Alta Vista, Excite, Northern Light and Lycos could not provide? The answer is surprisingly simple. Google was the first mover in Web search that said, "If Web pages link to other Web pages, we know that a page is meaningful to people."¹⁴ The pages with the most links are, therefore, the ones that most people looking for a specific topic will find useful.

The secret sauce was figuring out how to use the actions of people who created Web pages as the billion-dollar metric. Over time, Messrs. Brin and Page added other factors to the basic PageRank notion. Looking at semantic vectors was not one of the first technical additions flowing from the dormitory room at Stanford. As Google grew, more sophisticated algorithms were applied to the basic voting notion.

13. See "Google's Patent: Information Retrieval Based on Historical Data" at <http://www.seomoz.org/articles/google-historical-data-patent.php>. Available on June 30, 2005.

14. Google acknowledges the work of Eugene Garfield (founder of the Institute for Scientific Information and the father of online citation analysis) and Oliver McBryan's 1994 work on anchor text, among others. For more precursors of Google, see US Patent 6285999 B1.

As people figured out how to snooker the Google PageRank system, Google added layers of factors to the basic PageRank algorithm. Like a snowball pushed by a group of school children, the Google PageRank algorithm has grown. Mathematics and logic helped protect the value of the Google relevance ranking despite the best efforts of sales people, unscrupulous Web site operators, and search engine optimization gurus. To handle the emergence of tricks that exploit the vulnerabilities of an algorithm, Google has, according to some Google watchers, introduced a penalty. A Web site that uses too many tricks to make a Web page appear with a high PageRank score gets penalized. Some penalties are a downtick in a Google ranking; another is removal from the Google index.

What is emerging is an escalation of PageRank complexity. Whether the driving force is a desire to defeat the cheaters or a pragmatic need to keep PageRank reasonably objective, is not clear. PageRank changes. With every modification that Google makes, Web masters, advertisers, and SEO snake oil vendors howl. Google, like Chrysippus, makes no comment.¹⁵

What makes PageRank go, however, is not the algorithm. Google, to its credit, does little to keep competitors in the dark about what and how it makes its core invention putter along. The competitors seem to be in a state of willing disbelief that Google is implementing processes other than simple link analysis and click tracking.

Imagine that a competitor gets the Google PageRank algorithm in its current incarnation. Assume the competitor's PageRank algorithm runs on a big server with multiple high-speed data lines. Assume the competitor turns on the PageRank algorithm. Immediately the big server slows down and times out due to the iterative nature of the calculations. Frustrated, the competitor orders more high-end, superfast hardware, and the same thing happens. The competitor again hits the computational wall that slows down the competitor's system.

One of Google's competitive barriers is that Google figured out how to build its own big computer out of cheap, commodity hardware. The cost benefit is that Google spends US\$1.00 for a performance boost and competitors must spend US\$7.00 or more for the same performance level. Google also understood that off-the-shelf operating systems from standard Linux distributions, specialists like Sun Microsystems, or Microsoft had built in performance inhibitors.

To removed these bottlenecks, Google re-engineered a standard Linux distribution. The result is that Google uses its own Google Operating System or GOS. Google also developed its own file system, GFS or Google File System. The payoff from these engineering efforts is that Google eliminated traditional back up and restore costs by making three to six copies of every file. By 2003, Google's hardware and software engineers had developed a computational Maserati with the operational costs of a Honda Civic.

The strength of these barriers and Google's success stem from a network effect. One innovation is good. When a series of many incremental innovations combine, the system is greater than any one component.

15. Chrysippus was a Stoic and a pupil of Zeno. Stoics are not chatty Kathies.

Google went critical in 2004. Like a nuclear reaction, no one component is more or less important. A system phase change occurred. Now, Google's competitors must pore through Google's technical documentation, study Google's patents, and reverse engineer the Google system. A big job.

PageRank is the prime mover at Google. Once in motion, PageRank forced the company to find ways to do certain types of computer processing better, faster, and cheaper. Google is benefiting from the confluence of factors that combine to make the Googleplex bigger than the sum of its parts.

The Service Array in 2005

The table below provides a checklist of Google services available in June 2005. This list provides a brief description of the Google service and includes a comment about monetization of the service. The word Google is omitted from the product or service name. To locate these services on Google, run a Google query with quotes about the phrase; for example, "*Google Alerts*". *I'm feeling lucky* hot links to the specific Google service in most cases.¹⁶

Three points warrant comment:

- 1 Google's service array spans more than 50 services with new services appearing frequently. The services include locally-installed software such as Google Earth and Picasa as well as applications such as GMail that offer considerable functionality at no cost to the user because advertisers foot the bill.
- 2 Newer services push into quite different markets such as Video, presumably for consumers, and Scholar, presumably for researchers looking for information from scholarly publications. The unifying element in diverse products and services is that users can search for information.
- 3 Google requires users to create an account. That user name and password allow Google to track user behavior at the level of an individual user as well as using data from these stateful sessions as fodder for more advanced metrics for clustering and analyzing data from stateless sessions.

Product / Service	Description	Comment
AdSense	Process and technology to place context-sensitive ads on participating Web sites' pages.	Per-click model with revenue sharing to site.
AdWords	Google's version of the original Overture service. Google paid Yahoo to avoid litigation prior to its IPO.	Per-click model. Difference from Yahoo is a relatively low-key approach to ads. Otherwise, close enough to describe both as "very similar".

16. An exception is the "I'm feeling lucky" for AdSense. A third-party Web site was more relevant in May 2005 than Google's own description of its service. Google's PageRank algorithm is a script. This error in ranking has been "adjusted" as of June 12, 2005.