

CASE HISTORY

How Google works

Sep 16th 2004

From The Economist print edition



AP

Internet searching: With all the fuss over Google's IPO, it is easy to overlook its broader social significance. For many people, Google made the internet truly useful. How did it do it?

ONE thing that distinguishes the online world from the real one is that it is very easy to find things. To find a copy of *The Economist* in print, one has to go to a news-stand, which may or may not carry it. Finding it online, though, is a different proposition. Just go to Google, type in "economist" and you will be instantly directed to economist.com. Though it is difficult to remember now, this was not always the case. Indeed, until Google, now the world's most popular search engine, came on to the scene in September 1998, it was not the case at all. As in the physical world, searching online was a hit-or-miss affair.

Google was vastly better than anything that had come before: so much better, in fact, that it changed the way many people use the web. Almost overnight, it made the web far more useful, particularly for non-specialist users, many of whom now regard Google as the internet's front door. The recent fuss over Google's stockmarket flotation obscures its far wider social significance: few technologies, after all, are so influential that their names become used as verbs.

Google began in 1998 as an academic research project by Sergey Brin and Lawrence Page, who were then graduate students at Stanford University in Palo Alto, California. It was not the first search engine, of course. Existing search engines were able to scan or "crawl" a large portion of the web, build an index, and then find pages that matched particular words. But they were less good at presenting those pages, which might number in the hundreds of thousands, in a useful way.

Mr Brin's and Mr Page's accomplishment was to devise a way to sort the results by determining which pages were likely to be most relevant. They did so using a mathematical recipe, or algorithm, called PageRank. This algorithm is at the heart of Google's success, distinguishing it from all previous search engines and accounting for its apparently magical ability to find the most useful web pages.

Untangling the web

PageRank works by analysing the structure of the web itself. Each of its billions of pages can link to other pages, and can also, in turn, be linked to. Mr Brin and Mr Page reasoned that if a page was linked to many other pages, it was likely to be important. Furthermore, if the pages that linked to a page were important, then that page was even more likely to be important. There is, of course, an inherent circularity to this formula—the importance of one page depends on the importance of pages that link to it, the importance of which depends in turn on the importance of pages that link to them. But using some mathematical tricks, this circularity can be resolved, and each page can be given a score that reflects its importance.

The simplest way to calculate the score for each page is to perform a repeating or “iterative” calculation (see [article](#)). To start with, all pages are given the same score. Then each link from one page to another is counted as a “vote” for the destination page. Each page's score is recalculated by adding up the contribution from each incoming link, which is simply the score of the linking page divided by the number of outgoing links on that page. (Each page's score is thus shared out among the pages it links to.)

Once all the scores have been recalculated, the process is repeated using the new scores, until the scores settle down and stop changing (in mathematical jargon, the calculation “converges”). The final scores can then be used to rank search results: pages that match a particular set of search terms are displayed in order of descending score, so that the page deemed most important appears at the top of the list.

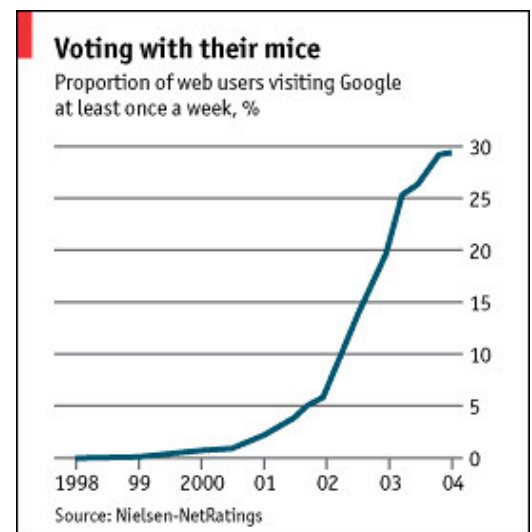
While this is the simplest way to perform the PageRank calculation, however, it is not the fastest. Google actually uses sophisticated techniques from a branch of mathematics known as linear algebra to perform the calculation in a single step. (And the actual PageRank formula, still visible on a [Stanford web page](#) includes an extra “damping factor” to prevent pages' scores increasing indefinitely.)

Furthermore, the PageRank algorithm has been repeatedly modified from its original form to prevent people from gaming the system. Since Google's debut in 1998, the importance of a page's Google ranking, particularly for businesses that rely on search engines to send customers their way, has increased dramatically: Google is now responsible for one in three searches on the web. For this reason, an entire industry of “search-engine optimisers” has sprung up. For a fee, they will try to manipulate your page's ranking on Google and other search engines.

The original PageRank algorithm could be manipulated in a fairly straightforward fashion, by creating a “link farm” of web pages that link to one another and to a target page, and thus give an inflated impression of its importance. So Google's original ranking algorithm has grown considerably more complicated, and is now able to identify and blacklist pages that try to exploit such tricks.

Mr Page and Mr Brin made another important innovation early on. This was to consider the “anchor text”—the bit of text that is traditionally blue and underlined and forms a link from one page to another—as a part of the web page it referred to, as well as part of the page it was actually on. They reasoned that the anchor text served as an extremely succinct, if imprecise, summary of the page it referred to. This further helps to ensure that when searching for the name of a person or company, the appropriate website appears at the top of the list of results.

Ranking the order in which results are returned was the area in which Google made the most improvement, but it is only one element of search—and it is useless unless the rest of the search engine works efficiently. In practice, that means compiling a comprehensive and up-to-date index of the web's ever-changing pages. PageRank sits on top of Google's extremely powerful and efficient



search infrastructure—one that draws on the lessons learned from previous, and now mostly forgotten, search engines.

As the web grew in the early 1990s, a number of search engines, most of them academic research projects, started crawling and indexing its pages. The first of these, the World Wide Web Wanderer and the World Wide Web Worm, used very simple techniques, and did not even index entire web pages, but only their titles, addresses and headers. A number of commercial engines followed, springing out of academic projects (as Google later did). WebCrawler, the first to index entire pages, emerged in 1994 at the University of Washington and was later bought by America Online. It was followed by Lycos and InfoSeek. But the first really capable search engine was AltaVista, unveiled by Louis Monier of Digital Equipment Corporation in December of 1995.

The day before the site opened for business, on December 15th, it already had 200,000 visitors trying to use it. That was because AltaVista successfully met two of the three requirements that later led to Google's success. First, it indexed a much larger portion of the web than anything that had come before. This, says Dr Monier, was because AltaVista used several hundred "spiders" in parallel to index the web, where earlier search engines had used only one. Second, AltaVista was fast, delivering results from its huge index almost instantly. According to Dr Monier, all earlier search engines had been overwhelmed as soon as they became popular. But the AltaVista team had used a modular design right from the start, which enabled them to add computing power as the site's popularity increased. Among some geeks, at least, AltaVista came into use as a verb.

Seek, and Google shall find

Even so, AltaVista still lacked Google's uncanny ability to separate the wheat from the chaff. Experienced users could use its various query options (borrowed from the world of database programming) to find what they were looking for, but most users could not. Although AltaVista's unprecedented reach and speed made it an important step forward, Google's combination of reach, speed and PageRank added up to a giant leap.

When you perform a Google search, you are not actually searching the web, but rather an index of the copy of the web stored on Google's servers. (Google is thought to have several complete copies of the web distributed across servers in California and Virginia.) The index is compiled from all the pages that have been returned by a multitude of spiders that crawl the web, gathering pages, extracting all the links from each page, putting them in a list, sorting the links in the list in order of priority (thus balancing breadth and depth) and then gathering the next page from the list.

When a user types in a query, the search terms are looked up in the index (using a variety of techniques to distribute the work across tens of thousands of computers) and the results are then returned from a separate set of document servers (which provide preview "snippets" of matching pages from Google's copies of the web), along with advertisements, which are returned from yet another set of servers. All of these bits are assembled, with the help of PageRank, into the page of search results. Google manages to do this cheaply, in less than a second, using computers built from cheap, off-the-shelf components and linked together in a reliable and speedy way using Google's own clever software. Together, its thousands of machines form an enormous supercomputer, optimised to do one thing—find, sort and extract web-based information—extremely well.

"Can Google stay on top as search, the activity where it is strongest, moves from centre stage to being just part of a bundle of services?"

Mr Page and Mr Brin created the prototype of Google on Stanford's computer systems. However, as visionaries do, they thought ahead clearly, and from the beginning had sound ideas both for searching and for creating the system of servers capable of handling the millions of queries a day that now pass through Google. It was the clarity of their ideas for scaling the server architecture, and their ability to think big, that made it so easy for them to turn their research project into a business. Andy Bechtolsheim, one of the founders of Sun Microsystems and an early investor in Google, did not even wait to hear all the details: when Mr Page and Mr Brin approached him, he reputedly said, "Why don't I just write you a cheque for \$100,000?" He wrote the cheque to "Google Inc."—a firm which did not yet

exist. So Mr Page and Mr Brin were forced to incorporate a business very quickly, and the company was born.

What was still missing, though it was unfashionable to worry about it in the early days of the dotcom boom, was a way of making money. Initially, Google sold targeted banner advertisements and also made money by providing search services to other websites, including Yahoo! and a number of other, smaller portals. But, says John Battelle, a professor at the University of California, Berkeley, who is writing a book about search engines, Google's revenues did not really take off until 2000, when it launched AdWords—a system for automatically selling and displaying advertisements alongside search results.

Advertisers bid for particular search terms, and those who bid the highest for a particular term—"digital cameras", say—have their text advertisements displayed next to Google's search results when a user searches for that term. Google does not simply put the highest bidder's advertisement at the top of the list, however. It also ranks the advertisements according to their popularity, so that if more people click on an advertisement halfway down the list, it will be moved up, even if other advertisers are paying more. Google's philosophy of ranking results according to their usefulness is thus applied to advertisements too.

The only fly in the ointment, from Google's point of view, was that Overture, a rival firm, claimed to have patented the idea for AdWords-style sponsored links. Overture filed a lawsuit against Google in 2002: it was settled out of court last month when Google agreed to give Yahoo! (which acquired Overture last year) 2.7m shares, worth around \$230m, to resolve the matter. Google was eager to settle the AdWords dispute before its initial public offering, which took place on August 19th.

Google now faces a three-way fight with Yahoo! and Microsoft, which have both vowed to dethrone it as the dominant internet search engine. Yahoo!'s strategy is to interconnect its various online services, from search to dating to maps, in increasingly clever ways, while Microsoft's plan is to integrate desktop and internet searching in a seamless manner, so that search facilities will be embedded in all its software, thus doing away (the company hopes) with the need to use Google. Both firms are also working to improve their basic search technology in order to compete with Google.

Beyond searching?

In response, Google has gradually diversified itself, adding specialist discussion groups, news and shopping-related search services, and a free e-mail service, Gmail, which is currently being tested by thousands of volunteers. It has also developed "toolbar" software that can be permanently installed on a PC, allowing web searches to be performed without having to visit the Google website, and establishing a toe-hold on its users' PCs.

Google's technical credentials are not in doubt. The question is whether it can maintain its position, as search, the activity where it is strongest, moves from centre stage to being just part of a bundle of services. Yet the example of Gmail shows how search can form the foundation of other services: rather than sorting mail into separate folders, Gmail users can simply use Google's lightning-fast search facility to find a specific message. So the technology that made Google great could yet prove to be its greatest asset in the fight ahead. Let battle commence.