

How Google Works

If you aren't interested in learning how Google creates the index and the database of documents that it accesses when processing a query, skip this description. I adapted the following overview from Chris Sherman and Gary Price's wonderful description of How Search Engines Work in Chapter 2 of The Invisible Web (CyberAge Books, 2001).

Google runs on a distributed network of thousands of low-cost computers and can therefore carry out fast parallel processing. Parallel processing is a method of computation in which many calculations can be performed simultaneously, significantly speeding up data processing. Google has three distinct parts:

- Googlebot, a web crawler that finds and fetches web pages.
- The indexer that sorts every word on every page and stores the resulting index of words in a huge database.
- The query processor, which compares your search query to the index and recommends the documents that it considers most relevant.

Let's take a closer look at each part. 1. Googlebot, Google's Web Crawler

Googlebot is Google's web crawling robot, which finds and retrieves pages on the web and hands them off to the Google indexer. It's easy to imagine Googlebot as a little spider scurrying across the strands of cyberspace, but in reality Googlebot doesn't traverse the web at all. It functions much like your web browser, by sending a request to a web server for a web page, downloading the entire page, then handing it off to Google's indexer.

Googlebot consists of many computers requesting and fetching pages much more quickly than you can with your web browser. In fact, Googlebot can request thousands of different pages simultaneously. To avoid overwhelming web servers, or crowding out requests from human users, Googlebot deliberately makes requests of each individual web server more slowly than it's capable of doing.

Googlebot finds pages in two ways: through an add URL form, www.google.com/addurl.html, and through finding links by crawling the web.

Unfortunately, spammers figured out how to create automated bots that bombarded the add URL form with millions of URLs pointing to commercial propaganda. Google rejects those URLs submitted through its Add URL form that it suspects are trying to deceive users by employing tactics such as including hidden text or links on a page, stuffing a page with irrelevant words, cloaking (aka bait and switch), using sneaky redirects, creating doorways, domains, or sub-domains with substantially similar content, sending automated queries to Google, and linking to bad neighbors. So now the Add URL form also has a test: it displays some squiggly letters designed to fool automated "letter-guessers"; it asks you to enter the letters you see — something like an eye-chart test to stop spambots.

When Googlebot fetches a page, it culls all the links appearing on the page and adds them to a queue for subsequent crawling. Googlebot tends to encounter little spam because most web authors link only to what they believe are high-quality pages. By harvesting links from every page it encounters, Googlebot can quickly build a list of links that can cover broad reaches of the web. This technique, known as deep crawling, also allows Googlebot to probe deep within individual sites. Because of their massive scale, deep crawls can reach almost every page in the web. Because the web is vast, this can take some time, so some pages may be crawled only once a month.

Although its function is simple, Googlebot must be programmed to handle several challenges. First, since Googlebot sends out simultaneous requests for thousands of pages, the queue of "visit soon" URLs must be constantly examined and compared with URLs already in Google's index. Duplicates in the queue must be eliminated to prevent Googlebot from fetching the same page again. Googlebot must determine how often to revisit a page. On the one hand, it's a waste of resources to re-index an unchanged page. On the other hand, Google wants to re-index changed pages to deliver up-to-date results.

To keep the index current, Google continuously recrawls popular frequently changing web pages at a rate roughly proportional to how often the pages change. Such crawls keep an index current and are known as fresh crawls. Newspaper pages are downloaded daily, pages with stock quotes are downloaded much more frequently. Of course, fresh crawls return fewer pages than the deep crawl. The combination of the two types of crawls allows Google to both make efficient use of its resources and keep its index reasonably current. 2. Google's Indexer

Googlebot gives the indexer the full text of the pages it finds. These pages are stored in Google's index database. This index is sorted alphabetically by search term, with each index entry storing a list of documents in which the term appears and the location within the text where it occurs. This data structure allows rapid access to documents that contain user query terms.

To improve search performance, Google ignores (doesn't index) common words called stop words (such as the,

is, on, or, of, how, why, as well as certain single digits and single letters). Stop words are so common that they do little to narrow a search, and therefore they can safely be discarded. The indexer also ignores some punctuation and multiple spaces, as well as converting all letters to lowercase, to improve Google's performance. 3. Google's Query Processor

The query processor has several parts, including the user interface (search box), the "engine" that evaluates queries and matches them to relevant documents, and the results formatter.

PageRank is Google's system for ranking web pages. A page with a higher PageRank is deemed more important and is more likely to be listed above a page with a lower PageRank.

Google considers over a hundred factors in computing a PageRank and determining which documents are most relevant to a query, including the popularity of the page, the position and size of the search terms within the page, and the proximity of the search terms to one another on the page. A patent application discusses other factors that Google considers when ranking a page. Visit SEOMoz.org's report for an interpretation of the concepts and the practical applications contained in Google's patent application.

Google also applies machine-learning techniques to improve its performance automatically by learning relationships and associations within the stored data. For example, the spelling-correcting system uses such techniques to figure out likely alternative spellings. Google closely guards the formulas it uses to calculate relevance; they're tweaked to improve quality and performance, and to outwit the latest devious techniques used by spammers.

Indexing the full text of the web allows Google to go beyond simply matching single search terms. Google gives more priority to pages that have search terms near each other and in the same order as the query. Google can also match multi-word phrases and sentences. Since Google indexes HTML code in addition to the text on the page, users can restrict searches on the basis of where query words appear, e.g., in the title, in the URL, in the body, and in links to the page, options offered by Google's Advanced Search Form and Using Search Operators (Advanced Operators).

Let's see how Google processes a query. Copyright © 2003 Google Inc. Used with permission.

For more information on how Google works, take a look at the following articles.

- Google's page on Google's Technology, www.google.com/technology/.
- How does Google collect and rank results?, www.google.com/newsletter/librarian/librarian_2005_12/article1.html.
- Google's PageRank Algorithm and How it Works, www.iprcom.com/papers/pagerank/
- Google's PageRank Explained and How to Make the Most of It, www.webworkshop.net/pagerank.html

Source: googleguide.com