

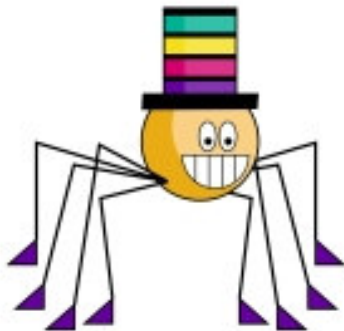
## How does Google work?

We have been asked a number of times how search engines work. Well, actually the question is, almost always, 'How does Google work?'

Now that's a bit like asking 'how does a plane work?' but worse, as Google and other search engines publish their patents but not their algorithms (i.e. the mathematical and other rules they use in their programmes).

However, we thought that we'd try to have a go at giving the basics because without search engines, it would be virtually impossible to locate anything on the Web without knowing a specific web page address.

First let's define what we are talking about. For this article a search engine (and we are not talking about directories here), but programmes that automatically browse the world-wide-web in a methodical manner, database and index the data returned and then allow users to query that data and provide accurate results.



Generally search engines have these components:

- A web crawler: an automated program that accesses a web site like your browser does but 'non visually' and goes through the site following the links or sitemap protocol information and sending data back.
- An indexer that processes crawled web pages into a database and then analyses them. It will look at things such as the page title, headings and sub headings, style (bold, italic), internal links, external links, inbound links and the text on each page itself. In looking at the text it will use techniques such as natural language processing to manipulate, analyse and understand the meaning and mark the page up in a number of ways for storage in the database.

- A database, which is a collection of related electronic records in a standardized format and searchable in a variety of ways.
- A query and results interface into which we put simple or more advanced queries to try to ensure that you get the most relevant result.

So, it's all very simple really. All you need is:

- a few computer programmes that can visit billions of web pages on a regular basis requesting and fetching thousands of different pages simultaneously
- work out each page links internally and externally and who links to them and then go and crawl those web pages too, making sure you don't

duplicate your crawling or visit pages that don't change much too frequently

- collect all the text on each web page that you crawl
- recognise whether it has changed and what has changed
- manipulate and analyse it making sure that 'web spammers' are not manipulating your results



- keeps copies of all the information indexed in your own document servers and all the data available and sorted
- store it all in a database that billions of people can access whenever they like

- answer their queries with great accuracy and in milliseconds even if the information is not that obvious – see <http://googleblog.blogspot.com/2008/07/technologies-behind-google-ranking.html> where Amit Singhal of Google on the official Google blog says "One of the key technologies we have developed to understand pages is associating important concepts to a page even when they are not obvious on the page. We find the official homepage for Sprovieri Gallery in London for the Italian query [galleria sprovieri londra], even though the official page does not have either London or Londra on it" and.....
- simultaneously run an adverts database so that accurate advert matches are also placed on the results page, because that's the main way you make money.

Google themselves say at

<http://www.google.com/corporate/tech.html>: "We use more than 200 signals, including our patented PageRank™\* algorithm, to examine the entire link structure of the web and determine which pages are most important. We then conduct hypertext-matching analysis to determine which pages are relevant to the specific search being conducted. By combining overall importance and query-specific relevance, we're able to put the most relevant and reliable results first".



Piece of cake really!

Well.....

So, with all this going on how can you get your site to the top of Google or other search engines?

Google itself publishes some very clear guidelines at:

<http://www.google.com/support/webmasters/bin/answer.py?answer=35769> and



there are other resources that give their views of what is important, for example <http://www.vaughns-1-pagers.com/internet/google-ranking-factors.htm>

But, the plain fact is that sites need to earn Google's trust before they can rank well for competitive search queries.

\* PageRank™ mainly relies on the 'democratic nature' of the web by using its vast link structure as an indicator of an individual page's value. Important, high-quality sites receive a higher PageRank™. So, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at a lot more than the sheer volume of links a page receives. For example, it also analyzes the page that casts the vote and votes by pages that are important weigh more heavily and help to make other pages important. A site's rate of link acquisition, the longevity of a link, the text used for the link, whether it's a 'deep link' or to the homepage and whether anyone clicks on the link seem also to count.