How does Google work?

Three aspects:

1. How does Google know what is on the Web?
2. How does Google search the Web for your query phrase?
3. How does Google decide which results you may be interested in?

Google is not the first Search Engine, earlier were, for example, Altavista and Yahoo.

Google introduced a ‘page ranking’ to help with item (3).

How does Google know what is on the Web?

Answer is spiders!

These are processes (sometimes called web crawlers or bots) that traverse the Web by going along links, and for each page encountered storing the text in a compact, easy to access, form.

How does Google search the Web for your query phrase?

The problem:

How many proper webpages are there? Most estimates are between 10 and 50 billion ($1 - 5 \times 10^{10}$) webpages.

The average webpage has approx. 500 words of text, so approx. 3,000 characters.

Therefore, the simple matching of a search phrase of 10 characters to the entirely of webpage texts to find all occurrences requires approx.

10^{15} equality tests on characters.

And Google has approx. $30,000$ searches per second!

There are 3×10^{16} nanoseconds in a year, and the age of the universe is approximately 4×10^{26} nanoseconds!

How does Google search the Web for your query phrase?

The solution:

Google does not search for words in text, like a text editor (eg MS Word or Emacs).

Instead it stores the pages using indexing.

In its simplest form, the index links a word with all webpages that contain the word. A search, consisting of a sequences of words, results in the intersection of the indexed sets of pages.

Of course, we don't use words, sets and intersection - this would be impossibly inefficient. We use hashing techniques.

How does Google decide which results you may be interested in?

The results of a search are presented to you in some order. What order?

It is determined by:

1. "Page Rank"
2. High preference sites (eg Wikipedia)
3. The importance of the phrase on the page
4. Your browsing history
5. Other things...

It combines these order relations and sorts the results using the combined relation into a final arrangement to present the 'hits'.

Page Rank

Due to: Larry Page and Sergey Brin (1996)

Idea: The importance of a page is determined by the importance of pages that link directly to it.

This is a recursive definition and leads to an iterative algorithm for calculating pagerank as the probability that someone clinking a link at random arrives at a particular page.

See e.g. `http://www.sirgroane.net/google-page-rank`.