Search Engine Indexing & the Page Rank Algorithm

**Introduction**

The World Wide Web contains billions of webpages and so finding the information / resources that we need should be a very difficult job.

Luckily, there are search engines, which are systems with the purpose of finding the webpages and files that we require.

**How search engines work…**

Many people wrongly believe that when you search for some information online, the search engine quickly scans the entire web to find relevant websites. All that actually happens is the search engine will quickly look up your search terms in their database of websites and give you a list of relevant sites.

Search engines call their database an index. A search engine’s index is constantly being updated by ‘software robots’ called spiders (or web crawlers) which follow links from site to site, sending back information such as the websites URL and aspects of their content, so that it can be added to the index.

***What information do spiders return?***

***What information is actually added to the index?***

As the spiders visit each website they will send back huge amounts of information. The following is just a small list of information that the spiders will be recording:

* How many times different words appear on the page
* When the page was last updated
* The words that are placed in <h1> tags (headings)
* The words that are placed in <title> tags (the page title that appears in the bar/tab at the top of a browser)
* The URL (web address)

After a site has been crawled, the site (along with this information) will be indexed (added to the database). This information is then used each time a relevant search is conducted to decide upon which websites should appear on the search results page.

**Deciding where websites should be positioned on the search results page**

It is all very well having a list of webpages that relate to the search terms entered in to the search engine, but considering that there could be millions of related webpages found in a search, the real art of a search engine is to rank the webpages so that the most useful sites appear first.

Google has a search engine that does this very well.

Back in 1995, Larry Page and Sergey Brin developed an algorithm to rank the webpages (that relate to the search terms), which meant that users quickly found the websites / information that they were looking for. This algorithm is the **‘Page Rank Algorithm’**.

**The Page Rank Algorithm**

The algorithm doesn’t just look at the number of times a key word appears or when it was last updated etc, it also considers the number of external links that point to the page. Because of this, it can work out how popular a webpage is and rank it according to its popularity. Furthermore, it also works out how popular the external webpages are that have the links to the page, and so adds weighting to each of the external links. The result is an ordered list of sites which match the search terms AND that are deemed more popular with users of the web…clever stuff!

**The Page Rank Algorithm – An Example**

Let’s say there are two webpages that have equal weighting in terms of its content (key terms, meta data, updates etc).

These of course should both ranked number 1!

…but they can’t be.

This is because the page rank algorithm focuses more on the external webpages which link to the page, than the page itself.

If page one has 100 links pointing towards it, it has 100 popularity votes.

If page two has 200 links pointing towards it, it has 200 popularity votes.

Page 2 therefore would usually rank higher.

Delving deeper however, if the links pointing to page 2 are not popular themselves, this will adversely affect page 2’s rank. So the algorithm is based on the quality of the external links pointing to a webpage, as well as the number.

**The Page Rank Algorithm – How it works…**

In essence, the algorithm works by adding together the page rank score of each external page with links to A…but with a few important extra calculations…



The size of each face (webpage rank) is proportional to the number of arms (links) pointing to it.

**The Page Rank Algorithm – Open to abuse?!**

As the algorithm works on the premise of quality links pointing to a webpage, individuals have attempted to increase the page rank of a webpage by creating hundreds of webpages which each have links pointing to the page.

Search engines are constantly tweaking the algorithm to ensure that situations like this do not wrongly impact the chance of a user finding the websites that they require.