

Workshop notes: Intermediate SEO

Recommended reading

SEOMoz's Beginners Guide to SEO is an excellent resource:

<http://www.seomoz.org/beginners-guide-to-seo>

How search engines rank sites

Each search engine has a vast database, called an **index**, storing information about all the pages they've found and how the pages are connected to each other by links. The index is created using software known as a **robot, crawler, or spider**. Robots find new pages by following links from existing pages. Some search engines allow you to manually submit new sites to them, but this isn't necessary as long as there are links pointing to the site.

The ranking of a page in the search results is determined by a combination of **relevance** and **importance**. Relevance depends on how much a page relates to the user's search terms. Importance is independent of any particular search query, and depends on factors such as the number and variety of incoming links.

PageRank

In January 1998 Larry Page, one of the co-founders of Google, filed a patent titled *Method for node ranking in a linked database*. Building on earlier work with Google's other co-founder Sergey Brin, this patent describes an algorithm known as **PageRank**, which attempts to determine the importance of web pages based solely on the way they link to each other. Later that year Page and Brin founded Google, and results generated using PageRank were so much better than contemporary search engines that Google quickly became the most popular search engine, remaining so ever since.

The mathematics of the PageRank algorithm are described in detail in the patent, but there are two common non-technical explanations: the **voting model** and the **random surfer model**.

The voting model is more metaphorical than literal, but gives a good intuitive understanding of the algorithm. It says that when a page links to another page, it's like a vote for that page. The PageRank algorithm proceeds over multiple rounds; in the first round every page's voting power is equal, but in each successive round pages that received more votes in the last round have more voting power to give out in the next round. Eventually the results change little from round to round, and the vote a page receives is its PageRank.

The random surfer model is more precise, as it's a direct interpretation of the underlying mathematics. Imagine if a person surfing the web starts at a random page. When they've read the page, they do one of two things: there's a 15% chance they move to another random page, and a 85% chance they click on a random link on the current page. Suppose they continue this process for a large number of steps. The probability that they end up on a particular page is its PageRank.

The end result is that pages with more incoming links tend to rank more highly, especially if the pages that link to them also have many incoming links. As a result, SEO experts sometimes refer to links from higher ranking sites as passing more 'link juice' than links from lower ranking sites.

Link building

As a result of the PageRank algorithm, and similar algorithms used by other search engines, a large part of SEO concerns **link building**: the process of gaining incoming links in order to improve ranking. Some techniques to gain extra incoming links are discouraged by search engines, and links gained in such a way will be ignored if detected.

Natural links are the links that your site gains as a result of people just deciding to link to it, and are the best kind of link. If your content is interesting, entertaining, or controversial, it is likely to gain larger numbers of natural links. Gaining natural links is thus more a process of content creation and copy-writing than one of SEO. It is also a slow process, as to begin with few people will be looking at your site and hence it will take time to gain the first few links. When someone writes an article specifically in order to gain a large number of links, the article is called **link bait**. This term is generally used negatively and reserved for deliberately controversial content; it is less common to use the term for useful or entertaining content. A good way of accelerating the process of gaining natural links is **outreach**: contacting people you think will find your site useful. Outreach may or may not involve actually asking for a link, depending on how you think the recipient will respond; many bloggers, for example, don't like being asked for a link and will be more likely to link if not explicitly asked.

Self-created links are links to your site that you create on other sites. For example, you could link to your site in a comment posted to a blog, or in your profile page on a social network. While it's perfectly normal to have some links to your site of this form, if you have too many links of this form search engines will regard them as suspicious (**comment spam**) and ignore them. In 2005 Google introduced the **nofollow** attribute for links, which is a way for a site to tell search engines to ignore particular links – for example, links from Wikipedia and many blogs use nofollow and so have no effect on your ranking. This was intended to reduce the incentive for comment spam, although the effect has been limited.

Link exchanges and bad neighbourhoods

One common link building technique is the **link exchange**: linking to a site in return for a link back. As with self-created links, it's normal to have some links for this form, but having too many will look suspicious to search engines, especially if the links have been exchanged with irrelevant sites (sites on completely unrelated topics) or sites in **bad neighbourhoods** (groups of sites which search engines already regard as suspicious). Search engines will penalise sites that they detect are engaging in link exchanges. This is a harsher punishment than that for comment spam as search engines realise that comment spam could actually be a malicious rival attempting to frame you, whereas link exchanges involve changes to your own site and hence require your participation.