

Good Google - Writing for the most powerful robot in the world

Wednesday, December 1st 2004

By Jim Hedger, Contributing Writer

Google "...is big. Really Big. You just won't believe how vastly hugely mind-bogglingly big it is." (excerpt from The Hitchhiker's Guide to the Galaxy)

Google is the most powerful information resource humans have ever constructed. The power of any major search tool boggles the mind but considering the vastness of Google's complex simplicity can truly hurt one's brain. With over 8-billion references in its rapidly growing, organically generated index, Google sets the standards other search engines follow. Benefiting from a three year reign as the undisputed leader of search, Google has had a very good year and looks poised to make 2005 an even better year.

In 2004, Google introduced more new and improved applications for its users than any other tech company, posted one of the most successful IPO's in business history in a most unorthodox Dutch-Auction format, and met or exceeded any challenges its rivals threw at.

Google is no longer just a search engine, it is an advertising machine. Drawing about 90% of its revenues from paid advertising and contextual ad-delivery, Google has had two major focuses this quarter. The first is increasing the number of places paid-advertising might show up. The second is to develop new products and features that will retain current user loyalty and win new users from the other search firms. Both initiatives rely heavily on Google's reputation for delivering fast, free and relevant search results. Google has the world's largest database of indexed websites and it acquires site information through its spider GoogleBot.

GoogleBot is probably the most well-known spider working the web today. It is also likely among the most analyzed applications ever written. On one level, GoogleBot is quite simple and can be depended on to act in a very specific manner. GoogleBot lives to follow links. GoogleBot will often chase down a link-path until it can no longer work its way deeper into a site. It will also work its way through any site linked to from any other site. Google finds the majority of new sites in its index by following links from established sites. If a link exists, Google will (A) find it, (B) follow it, (C), record every bit of information it can possibly record, and (D) weigh that information against a fairly rigid algorithm to determine the perceived topic or theme of a site for future reference. If a site in Google's index is modified or changes, Google will re-spider the site as quickly as it possibly can.

GoogleBot's mission is to create a snap-shot of the World Wide Web and store it across Google's network of data centers around the world. When you reference information from Google, the results you see reflect Google's most recent snap-shot of the web. Parts of that snap-shot might be hours or even weeks old but overall the index is updating itself every minute of every day, 24/7. The fastest way to see exactly what Google views as the most recent version of your site is to click on the "Cached" link generally below the main link-reference Google displays for your site.

How GoogleBot behaves as it acquires sites is one thing. What Google does with the information its bot gathers is another thing. Google's method of ranking websites is extremely (and increasingly) complex. To understand how Google works today, a brief (and over simplified) explanation of the principle of PageRank is in order.

Google was originally developed as a means of finding information in research documents at Stanford University where its inventors Larry Page and Sergey Brin met as grad students. PageRank was developed as the basic sorting algorithm for their search tool (then known as Backrub) and was based on a very simple concept, trust.

Page and Brin understood that documents on the Internet could be linked together. They speculated that if someone took the time to code a link (by hand in those days) to another document there was likely a relevance between the two documents. Why else would one researcher link to another researcher's work? Simply put, the more incoming links a particular document has, the better it would rank when sorted by PageRank. Given the environment in which it was developed, Google's genesis proved to be the perfect tool for intelligent users. Transferring that simplicity from a dorm room at Stanford to practically every living room and office space on Earth has been a great challenge for Google's engineers. While it is still somewhat based on the original, "democratic" nature of PageRank, Google's sorting algorithm has become infinitely more complicated.

Google continues to weigh the number of links directed towards a site as positive indicators that there is relevant information to be found there. Since links are the veins and arteries of the web, links continue to be the most important factor influencing Google's perception of the relevance of a website. As the Google index has grown so rapidly over the past six years, and search engine marketers have learned how to use Google's behaviours to influence rankings, Google weighs several other factors when considering the relevance of a site but the core of the algorithm remains rooted in PageRank.

Not all Links are Created Equal

Back in the good old days, seven or eight years ago at Stanford, one link could represent one positive vote. As marketers learned to manipulate links, Google learned to apply different standards and measures when looking at those links and the content of sites in its index. Today, Google considers different links in different ways. As a matter of interest, our recent studies show that Google displays less back-links for sites than any other search engine, leading us to conclude that Google has become much stricter about how it views and values incoming links.

Google looks at a number of factors when determining the value of a link. Where the link originates from is as important as where the link is directed in Google's eyes. Google, like its rivals, is trying to find relationships between documents aside from obvious keywords. Google has the ability to fundamentally understand documents in its index and determine the topic, theme or context of those documents. This is an important measure as Google is becoming increasingly strict about link-relevance. To receive a highly positive response from Google, the pages or sites linked together must somehow relate to each other in topic as well as by sharing similar keywords. An excellent example would be in regional tourism.

A local tourism bureau will almost certainly have a website. That site will link to the sites of member-clients in its region. Each of those sites represent businesses dependent on regional tourism, thus establishing relevance between the sites. The tourism bureau becomes the "hub" from which Google follows links to other, topically related websites. In this way, the Hub site becomes a highly positive link-reference in Google's eyes.

The very best links, in Google's eyes, come from "authority sites". An authority site is one that is very well established and respected such as mainstream news sites (CNN, TIME, NYTimes, etc...) other search directories, industrial leaders (Macromedia, HP, Pitney Bowes, Nike, etc...), and other highly credible sources such as the regional tourism bureau mentioned above. While a website doesn't necessarily have to represent a large corporation to be considered an authority site, the sheer number of pages and references, combined with high visitor numbers generally associated with large corporate sites helps. Some personal Blogs, smaller companies and alternative news sources/blogs have also enjoyed "authority" status. This status is, in some ways, flexible and situational. A link from the tourism bureau mentioned above will not tend to help a business outside of its region unless a tangible relevancy factor is somehow introduced.

In practical terms, the "authority" status of a website is irrelevant for SEOs as the vast majority of sites in Google's index are just regular, run of the mill websites run by regular, run of the mill folks like us. Small businesses, researchers, governments, NGOs, musicians, artists, families, hobbyists and others write websites to offer the world access to their information. 99.999999% of these sites contain links of some sort or another and the vast majority of those links lead to topically relevant documents. While not "authority" sites, Google still considers these links

extremely important when sorting and ranking sites. Again, the stress is on topical relevancy as Google places enormous value in good, solid links.

Google does not live on links alone

Much as been written in this article and thousands of others about Google and links. If links were the only factor Google looks at, the SEO business would not exist and Google's index would be as off-kilter as a Batman set. As stated in previous paragraphs, Google has the ability to read sites and understand what it is reading. Google is able to reference a world of information when figuring out the context of text used in Titles, Meta Tags, Body Text and Anchor Links. Since we know that Google is actually reading and comprehending content, we need to place specific content in places we know GoogleBot likes to look for it. Writing and placing this information is where SEO becomes an artful science that stems from simple common sense. Think about what Google knows about your website before it even visits.

It finds your site by following links. Therefore it "assumes" your site is topically relevant to the site it acquired the link to your site from. Google knows the address of the site, the URL. It also knows what anchor text the original linking site used when phrasing the link to your website. Keyword enrichment of both elements is beneficial with Google. In other words, if you can, use a target keyword phrase in the URL of your site, and request that others linking to your site use your target keyword phrases as the anchor text of links directed to your site.

Once Google hits your site, it learns a lot more very quickly. It sees the title, tags, text and links, and records these elements as it moves through the site. These are the basic elements SEOs examine and modify when working on your site.

The first thing GoogleBot sees is the title of the site. Keyword enriched titles are very useful but webmasters are cautioned to be very conservative in the number of keywords or phrases they place in the title of a page. We generally use two or three keyword phrases when writing titles. Page titles should be page specific with keywords focused on the topic of the page. The second (or third) keyword set in the title is used to provide an overall context to the site. For example, <title="Blue Widgets :: Preformed Blocks and Spacers :: Construction Materials"> Overloading the title with keywords is useless and may be considered spam in extreme cases.

Next, Google looks at the meta tags. Unless you wish to exclude Google from sections of your site, there are only two really important meta tags, the description and the keywords tags. Of these two, the description is the most important. Google uses the description tag as a topical reference and may draw from the description tag when generating the two to three sentence site description shown under links in the SERPs. As with titles, each page should have a page specific description tag that outlines the topic of that page and the theme of the overall site. The keywords tag is of much lesser importance but is still considered to carry minor weight. Mentioning keywords that might be associated with your website, including common misspellings doesn't hurt. Packing the keywords tag with dozens of mentions of the same word, or using keywords that do not relate to your website might. We still use the keyword tag on client sites and still use page-specific keyword tags.

After the meta tags, Google looks at page content or body text. Again, relevance is extremely important. The Internet is a very big place and Google's index is pretty big itself. Finding documents in an 8-billion page universe requires precision. Webmasters can help themselves by simply addressing one topic or issue per page. Google is extremely intelligent and intuitive, but even the smartest robots get confused. Keeping it simple for GoogleBot makes good ranking much simpler to achieve for your site. As Google reads information from left to right in columns, like we read a newspaper, placing your keyword phrases early in the body text of pages in your site is very beneficial. Well written sentences that are topically focused are the best spider food for Google as it has become wary of words that "float" on a page without supporting words to provide context.

Lastly, GoogleBot comes back to links. GoogleBot moves through your website following links you place there. It reads the text that phrases the links to determine what it might find when it gets to the next page. For example, the

second page in most websites is the "About Us" page. Billions of websites use "About Us" as the anchor text linking the index page to the about us page. A better link would read About "Blue Widgets Inc." as the keyword phrase Blue Widgets is used as the anchor text from one page to the next. Keyword enrichment of anchor text also effects Google's perception of external links . Going back to our tourism bureau example, a link to a local bed and breakfast might read "Humboldt House" Bed and Breakfast or it might read Humboldt House "Victoria – Bed and Breakfast". The anchor text used in the second example would be far more beneficial than the first.

Remember, links provide the pathway for GoogleBot and other spiders. A final element that should be included on all pages is a text-based sitemap that links to all pages in the site and is linked to from the Home or INDEX page.

In a nutshell, that's how GoogleBot examines a site. Here is a quick rundown of which elements GoogleBot is looking for:

- Relevant Incoming Links
- Good URLs that are not too spammy
- Easy to follow link paths including a sitemap
- Keyword enriched titles
- Well written Description Meta Tag
- Well written Keywords Meta Tag (less important than Description)
- Topically focused Body Text
- Keyword Enriched Anchor Text
- NO SPAM
- Relevance, relevance, relevance

Next week, we'll look at MSNBot.