How Web Search Engines Work

Search engines are the key to finding specific information on the vast expanse of the World Wide Web. Without the use of sophisticated search engines, it would be virtually impossible to locate anything on the Web without knowing a specific URL(Uniform Resource Locator), the global address of documents and other resources on the World Wide Web. The first part of the address indicates what protocol to use, and the second part specifies the IP address or the domain name where the resource is located.
There are basically three types of search engines:

Those that are powered by crawlers, or spiders; those that are powered by human submissions; and those that are a combination of the two.

Spider or Crawlers:

Spider is a program that automatically fetches Web pages. Spiders are used to feed pages to search engines. It's called a spider because it crawls over the Web. Another term for these programs is webcrawler.

Because most Web pages contain links to other pages, a spider can start almost anywhere. As soon as it sees a link to another page, it goes off and fetches it. Large search engines, like Alta Vista, have many spiders working in parallel. Because most Web pages contain links to other pages, a spider can start almost anywhere. As soon as it sees a link to another page, it goes off and fetches it. Large search engines, like Alta Vista, have many spiders working in parallel.

Crawler-based engines send crawlers, or spiders, out into cyberspace. These crawlers visit a Web site, read the information on the actual site, read the sites meta tags and also follow the links that the site connects to. Meta Tag is

Meta Tag is a special HTML tag that provides information about a Web page. Unlike normal HTML tags, meta tags do not affect how the page is displayed. Instead, they provide information such as who created the page, how often it is updated, what the page is about, and which keywords represent the page's content. The crawler returns all that information back to a central depository where the data is indexed. The crawler will periodically return to the sites to check for any information that has changed, and the frequency with which this happens is determined by the administrators of the search engine.

Human Powered Search Engines:

Human-powered search engines rely on humans to submit information that is subsequently indexed and catalogued. Only information that is submitted is put into the index. In both cases, when you query a search engine to locate information, you are actually searching through the index that the search engine has created; you are not actually searching the Web. These indices are giant databases of information that is collected and stored and subsequently searched. This explains why sometimes a search on a commercial search engine, such as Yahoo! or Google, will return results that are in fact dead links. Since the search results are based on the index, if the index hasn't been updated since a Web page became invalid the search engine treats the page as still an active link even though it no longer is. It will remain that way until the index is updated.

So why will the same search on different search engines produce different results? Part of the answer to that is because not all indices are going to be exactly the same. It depends on what the spiders find or what the humans submitted. But more important, not every search engine uses the same algorithm to search through the indices. The algorithm is what the search engines use to determine the relevance of the information in the index to what the user is searching for.

One of the elements that a search engine algorithm scans for is the frequency and location of keywords on a Web page. Those with higher frequency are typically considered more relevant. But search engine technology is becoming sophisticated in its attempt to discourage what is known as keyword stuffing, or spamdexing. It is a technique used by Web designers to overload keywords onto a Web page so that search engines will read the page as being relevant in a Web search. Because search engines scan Web pages for the words that are entered into the search criteria by the user, the more times a keyword appears on the Web page the more relevancy the search engine will assign to the page in the search results (this is only one way that search engines determine relevancy, however.) Search engines often penalize a site if the engine discovers keyword stuffing, as this practice is considered poor netiquette, and some search engines will even ban the offending Web pages from their search results.

There are several methods of keyword stuffing. One way is to insert repeating keywords within the input type="hidden" field meta tag or the keyword tag so that the keywords are not seen by the user but are scanned by the search engine. Another

way is to make text in the body of the Web page invisible text by making the text the same color as the page?s background, rendering the text invisible to the user unless the user highlights the text. This method is called invisible keyword stuffing. Keyword stuffing also is referred to as keyword loading and spamdexing.

Another common element that algorithms analyze is the way that pages link to other pages in the Web. By analyzing how pages link to each other, an engine can both determine what a page is about (if the keywords of the linked pages are similar to the keywords on the original page) and whether that page is considered "important" and deserving of a boost in ranking. Just as the technology is becoming increasingly sophisticated to ignore keyword stuffing, it is also becoming savvier to Web masters who build artificial links into their sites in order to build an artificial ranking.

Hybrid Search Engines" Or Mixed Results:

In the web's early days, it used to be that a search engine either presented crawler-based results or human-powered listings. Today, it extremely common for both types of results to be presented. Usually, a hybrid search engine will favor one type of listings over another. For example, MSN Search is more likely to present human-powered listings from LookSmart. However, it does also present crawler-based results (as provided by Inktomi), especially for more obscure queries

Conclusions

Different engines have different strong points; use the engine and feature that best fits the job you need to do. One thing is obvious; the engine with the most pages in the database IS NOT the best. Not surprisingly, you can get the most out of your engine by using your head to select search words, knowing your search engine to avoid mistakes with spelling and truncation, and using the special tools available such as specifiers for titles, images, links, etc. The hardware power for rapid searches and databases covering a large fraction of the net is yesterday's accomplishment. We, as users, are living in a special time when search engines are undergoing a more profound evolution, the refinement of their special tools. I believe that very soon the Web will evolve standards, such as standard categories, ways of automatically classifying information into these categories, and the search tools to take advantage of them, that will really improve searching. I think it's exciting to be on the Web in this era, to be able to watch all the changes, and to evolve along with the Web as we use it.

Source: http://www.go4expert.com/articles/web-search-engines-t324/