How Search Engines Work

By Danny Sullivan, Editor
October 14, 2002

The term "search engine" is often used generically to describe both crawler-based search engines and human-powered directories. These two types of search engines gather their listings in radically different ways.

Crawler-Based Search Engines

Crawler-based search engines, such as Google, create their listings automatically. They "crawl" or "spider" the web, then people search through what they have found.

If you change your web pages, crawler-based search engines eventually find these changes, and that can affect how you are listed. Page titles, body copy and other elements all play a role.

Human-Powered Directories

A human-powered directory, such as the Open Directory, depends on humans for its listings. You submit a short description to the directory for your entire site, or editors write one for sites they review. A search looks for matches only in the descriptions submitted.

Changing your web pages has no effect on your listing. Things that are useful for improving a listing with a search engine have nothing to do with improving a listing in a directory. The only exception is that a good site, with good content, might be more likely to get reviewed for free than a poor site.

"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.

The Parts Of A Crawler-Based Search Engine

Crawler-based search engines have three major elements. First is the spider, also called the crawler. The spider visits a web page, reads it, and then follows links to other pages within the site. This is what it means when someone refers to a site being "spidered" or "crawled." The spider returns to the site on a regular basis, such as every month or two, to look for changes.

Everything the spider finds goes into the second part of the search engine, the index. The index, sometimes called the catalog, is like a giant book containing a copy of every web page that the spider finds. If a web page changes, then this book is updated with new information.

Sometimes it can take a while for new pages or changes that the spider finds to be added to the index. Thus, a web page may have been "spidered" but not yet "indexed." Until it is indexed -- added to the index -- it is not available to those searching with the search engine.

Search engine software is the third part of a search engine. This is the program that sifts through the millions of pages recorded in the index to find matches to a search and rank them in order of what it believes is most relevant. You can learn more about how search engine software ranks web pages on the aptly-named How Search Engines Rank Web Pages page.

Major Search Engines: The Same, But Different

All crawler-based search engines have the basic parts described above, but there are differences in how these parts are tuned. That is why the same search on different search engines often produces different results. Some of the significant differences between the major crawler-based search engines are summarized on the Search Engine Features Page. Information on this page has been drawn from the help pages of each search engine, along with knowledge gained from articles, reviews, books, independent research, tips from others and additional information received directly from the various search engines.

Now let's look more about how crawler-based search engine rank the listings that they gather

How Search Engines Rank Web Pages

By Danny Sullivan, Editor
July 31, 2003

Search for anything using your favorite crawler-based search engine. Nearly instantly, the search engine will sort through the millions of pages it knows about and present you with ones that match your topic. The matches will even be ranked, so that the most relevant ones come first.

Of course, the search engines don't always get it right. Non-relevant pages make it through, and sometimes it may take a little more digging to find what you are looking for. But, by and large, search engines do an amazing job.

As WebCrawler founder Brian Pinkerton puts it, "Imagine walking up to a librarian and saying, 'travel.' They’re going to look at you with a blank face."

OK -- a librarian's not really going to stare at you with a vacant expression. Instead, they're going to ask you questions to better understand what you are looking for.

Unfortunately, search engines don't have the ability to ask a few questions to focus your search, as a librarian can. They also can't rely on judgment and past experience to rank web pages, in the way humans can.

So, how do crawler-based search engines go about determining relevancy, when confronted with hundreds of millions of web pages to sort through? They follow a set of rules, known as an algorithm. Exactly how a particular search engine's algorithm works is a closely-kept trade secret. However, all major search engines follow the general rules below.

Location, Location, Location...and Frequency

One of the main rules in a ranking algorithm involves the location and frequency of keywords on a web page. Call it the location/frequency method, for short.

Remember the librarian mentioned above? They need to find books to match your request of "travel," so it makes sense that they first look at books with travel in the title. Search engines operate the same way. Pages with the search terms appearing in the HTML title tag are often assumed to be more relevant than others to the topic.

Search engines will also check to see if the search keywords appear near the top of a web page, such as in the headline or in the first few paragraphs of text. They assume that any page relevant to the topic will mention those words right from the beginning.

Frequency is the other major factor in how search engines determine relevancy. A search engine will analyze how often keywords appear in relation to other words in a web page. Those with a higher frequency are often deemed more relevant than other web pages.

Spice In The Recipe

Now it's time to qualify the location/frequency method described above. All the major search engines follow it to some degree, in the same way cooks may follow a standard chili recipe. But cooks like to add their own secret ingredients. In the same way, search engines add spice to the location/frequency method. Nobody does it exactly the same, which is one reason why the same search on different search engines produces different results.

To begin with, some search engines index more web pages than others. Some search engines also index web pages more often than others. The result is that no search engine has the exact same collection of web pages to search through. That naturally produces differences, when comparing their results.

Search engines may also penalize pages or exclude them from the index, if they detect search engine "spamming." An example is when a word is repeated hundreds of times on a page, to increase the frequency and propel the page higher in the listings. Search engines watch for common spamming methods in a variety of ways, including following up on complaints from their users.

Off The Page Factors

Crawler-based search engines have plenty of experience now with webmasters who constantly rewrite their web pages in an attempt to gain better rankings. Some sophisticated webmasters may even go to great lengths to "reverse engineer" the location/frequency systems used by a particular search engine. Because of this, all major search engines now also make use of "off the page" ranking criteria.

Off the page factors are those that a webmasters cannot easily influence. Chief among these is link analysis. By analyzing how pages link to each other, a search engine can both determine what a page is about and whether that page is deemed to be "important" and thus deserving of a ranking boost. In addition, sophisticated techniques are used to screen out attempts by webmasters to build "artificial" links designed to boost their rankings.

Another off the page factor is clickthrough measurement. In short, this means that a search engine may watch what results someone selects for a particular search, then eventually drop high-ranking pages that aren't attracting clicks, while promoting lower-ranking pages that do pull in visitors. As with link analysis, systems are used to compensate for artificial links generated by eager webmasters.

Learning More

The Search Engine Features Chart has a section that summarizes key areas of how crawler-based search engines rank web pages. The Search Engine Placement Tips page also summarizes key tips that will help you improve the relevancy of your pages with crawler-based search engines.

Search Engine Features For Webmasters

By Danny Sullivan, Editor
December 5, 2002

The search engine features chart below is designed primarily for webmasters who care about how crawler-based search engines index their sites. It provides a summary of important factors and features that can affect how sites are indexed and ranked. Full explanations of items can be found immediately below the comparison chart.

Human-powered search engines like the Open Directory are not listed on this chart because they do not crawl the web to create their listings. See the How Search Engines Work page for an explanation of the differences between crawler-based and human-powered services.

See the Search Engine Features For Searchers page for a summary of how search engines display their results and other information that may be of interest to searchers, rather than search engine marketers and site promoters.

This chart covers the crawler of AllTheWeb, AltaVista, Google, Inktomi and Teoma. Some of these crawlers power other search engines, and the relationships are shown on the Search Engine Results page.

Crawling	Yes	No	Notes
Deep Crawl	AllTheWeb, Google, Inktomi	AltaVista, Teoma	
Frames Support	All	n/a	
robots.txt	All	n/a	
Meta Robots Tag	All	n/a	
Paid Inclusion	All but...	Google	
Full Body Text	All	n/a	Some stop words may not be indexed
Stop Words	AltaVista, Inktomi, Google	FAST	Teoma unknown
Meta Description	All provide some support, but AltaVista, AllTheWeb and Teoma make most use of the tag		
Meta Keywords	Inktomi, Teoma	AllTheWeb, AltaVista, Google	Teoma support is "unofficial"
ALT text	AltaVista, Google, Teoma	AllTheWeb, Inktomi	
Comments	Inktomi	Others	

Deep Crawl

All crawlers will find pages to add to their web page indexes, even if those pages have never been submitted to them. However, some crawlers are better than others. This section of the chart shows which search engines are likely to do a "deep crawl" and gather many pages from your web site, even if these pages were never submitted. In general, the larger a search engine's index is, the more likely it will list many pages per site. See the Search Engine Sizes page for the latest index sizes at the major search engines.

Frames Support

This shows which search engines can follow frame links. Those that can't will probably miss listing much of your site. However, even for those that do, having individual frame links indexed can pose problem. Be sure to read the Search Engines And Frames page for tips on overcoming the problems with frames and search engines.

robots.txt

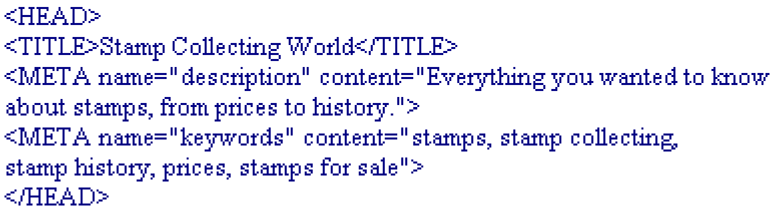
The robots.txt file is a means for webmasters to keep search engines out of their sites.

Meta Robots Tag

This is a special meta tag that allows site owners to specify that a page shouldn't be indexed. It is explained more on the How HTML Meta Tags Work page. The Web Robots Pages: The Robots META tag page also provides official information about robots.txt.

What are meta tags? They are information inserted into the "head" area of your web pages. Other than the title tag (explained below), information in the head area of your web pages is not seen by those viewing your pages in browsers. Instead, meta information in this area is used to communicate information that a human visitor may not be concerned with. Meta tags, for example, can tell a browser what "character set" to use or whether a web page has self-rated itself in terms of adult content.

Let's see two common types of meta tags, then we'll discuss exactly how they are used in more depth:



```
<HEAD>
<TITLE>Stamp Collecting World</TITLE>
<META name="description" content="Everything you wanted to know
about stamps, from prices to history.">
<META name="keywords" content="stamps, stamp collecting,
stamp history, prices, stamps for sale">
</HEAD>
```

In the example above, you can see the beginning of the page's "head" area as noted by the HEAD tag -- it ends at the portion shown as /HEAD.

Meta tags go in between the "opening" and "closing" HEAD tags. Shown in the example is a TITLE tag, then a META DESCRIPTION tag, then a META KEYWORDS tag. Let's talk about what these do.

The meta keywords tag allows you to provide additional text for crawler-based search engines to index along with your body copy. How does this help you? Well, for most major crawlers, it doesn't. That's because most crawlers now ignore the tag. The few supporting it can be found on the Search Engine Features page.

Paid Inclusion

Shows whether a search engine offers a program where you can pay to be guaranteed that your pages will be included in its index. This is NOT the same as paid placement, which guarantees a particular position in relation to a particular search term. The Submitting To Crawlers page provides links to various paid inclusion programs.

Full Body Text

All of the major search engines say they index the full visible body text of a page, though some will not index stop words or exclude copy deemed to be spam (explained further below). Google generally does not index past the first 101K of long HTML pages.

Stop Words

Some search engines either leave out words when they index a page or may not search for these words during a query. These stop words are excluded as a way to save storage space or to speed searches.

Meta Description

All the major crawlers support the meta description tag, to some degree. The ones actually named on the chart are very consistent. If you have a meta description tag on your pages, you'll most likely see the content used in some way.

The How HTML Meta Tags Work page explains how to use the meta description tag.

Meta Keywords

Shows which search engines support the meta keywords tags, as explained on the How HTML Meta Tags Work page.

ALT Text / Comments

This shows which search engines index ALT text associated with images or text in comment tags.

http://www.sebi.com.ar/projects/php/ir/ search engine open source

Search Features Chart

By Danny Sullivan, Editor
October 26, 2001

The search engine features chart below is designed primarily for users of search engines. It summarizes key search commands and search assistance features. These are described more fully on the Search Engine Math, Power Searching For Anyone and Search Assistance Features pages. You may also find information on the Search Engine Features For Webmasters page to be of interest. See the Major Search Engines page for links to the services listed below.

Search Engine Math Commands

Updated: March 11, 2003
(See Search Engine Math and Power Searching For Anyone for more details)
Covers: AllTheWeb, AltaVista, AOL Search, Ask Jeeves, Google, HotBot, Lycos, MSN Search, Teoma and Yahoo. HotBot references are only for its Inktomi-powered results.

Command	How	Supported By
Must Include Term	**+**	**All**
Must Exclude Term	**-**	**All**
Must Include Phrase	**" "**	**All**
Match All Terms	**Automatic at**	**All**
Match Any Terms	**Via Advanced Search**	**AllTheWeb, AltaVista, Google, Lycos, MSN Search, Teoma, Yahoo** *(HotBot offers but failed to work when tested)*
	OR	**AltaVista, AOL Search, Ask Jeeves, Google, HotBot, MSN Search, Teoma, Yahoo** *(must be done in ALL CAPS)* **AllTheWeb, Lycos** *(only works for two words)*

NOTE: By default, all the major search engines named above will match ALL of the terms you enter into a search box. This means that it is not necessary to use the + symbol in front of a particular word, though it won't hurt if you do so.

Power Searching Commands

(See Power Searching For Anyone for more details)

This section is being updated to cover the major crawler-based search engines of AllTheWeb, AltaVista, Google, Inktomi and Teoma. These crawlers also provide results to other search engines, so you may find commands on them work with their partners. Major partnerships are as follows: AllTheWeb (Lycos), Google (AOL Search, Yahoo), Inktomi (HotBot), MSN Search (Inktomi), Teoma (Ask Jeeves)

Command	How	Supported By
Title Search (Updated March 11, 2003)	title:	AltaVista, AllTheWeb, Inktomi
	intitle:	Google Teoma
	allintitle:	Google
Site Search	host:	AltaVista
	site:	Excite, Google (Netscape, Yahoo)
	url.host:	AllTheWeb, Lycos (for AllTheWeb results only)
	domain:	Inktomi (HotBot, iWon, LookSmart)
	none	AOL, Direct Hit, HotBot, LookSmart, Lycos, MSN, Netscape, Northern Light, Open Directory, Yahoo
URL Search	url:	AltaVista, Excite, Northern Light
	url.all:	AllTheWeb, Lycos (for AllTheWeb results only)
	allinurl: inurl:	Google
	originurl:	Inktomi (AOL, GoTo, HotBot)
	u:	Yahoo
	none	AOL, Direct Hit, HotBot, LookSmart, MSN **Not yet updated, but may be still correct:** Open Directory
Link Search	link:	AltaVista, Google, Northern Light

	linkdomain:	Inktomi (AOL, HotBot, iWon, MSN) (NOTE: measures links to entire domains)
	link.all:	AllTheWeb, Lycos (for AllTheWeb results only)
	none	AOL, Direct Hit, Excite, HotBot, LookSmart, Northern Light **Not yet updated, but may be still correct:** Netscape, Yahoo (n/a)
Wildcard	*	AltaVista, Inktomi (iWon), Northern Light **Not yet updated, but may be still correct:** Yahoo
	?	AOL Search, Inktomi (iWon)
	%	Northern Light
	none	AllTheWeb, Direct Hit, Excite, Google, HotBot, LookSmart, Lycos, MSN (MSN's help says it offers wildcard, but it failed to during testing)
Anchor Search	anchor:	AltaVista
	None	AllTheWeb, AOL Search, Direct Hit, Excite, Google, Inktomi, HotBot, Lycos

NOTE: The commands above are primarily useful when dealing with crawler-based search engines. "None" indicates any crawler-based or human-powered search engine that creates its own listings but which does not provide a particular command for searching within those listings. It may also indicate a portal that that outsources for its listings and which lacks a single command to work across the multiple datasets it uses.

Search Assistance Features

(See the Search Assistance Features page for more details)

Feature	Offered By
Related Searches	AltaVista, AllTheWeb, Excite, HotBot, Lycos, MSN, Yahoo **Not yet updated, but may be still correct:** iWon
Clustering	AltaVista, AllTheWeb, Excite, Google, HotBot, MSN, Northern Light
Find Similar	AltaVista, AOL Search, Google
Stemming	AOL Search, Direct Hit, HotBot, Inktomi (HotBot, MSN)
Search Within	AltaVista, Google, HotBot, Lycos
Spidered Version	Google
Search By Language	AltaVista, AllTheWeb, Excite, Google, HotBot, Lycos, MSN, Northern Light
Page Translation	AltaVista, Google, Lycos
Porn Filter	AltaVista, AllTheWeb, Google
Porn Warning	HotBot, MSN, Northern Light

Customization & Display Features

(See Search Assistance Features page for more details)

Feature	Supported By
Number Of Listings Shown (10 unless noted)	AltaVista, AllTheWeb, AOL Search (5), Direct Hit, Excite, Google, HotBot, LookSmart (15), Lycos, MSN (15), Northern Light **Not yet updated, but may be still correct:** iWon, Netscape, Yahoo (20)
Ability To Increase Number Of Listings?	AltaVista, AllTheWeb, Excite, Google, HotBot, MSN **Not yet updated, but may be still correct:** Yahoo
See 20 Results	AltaVista, AllTheWeb, Excite, Google, HotBot, MSN **Not yet updated, but may be still correct:** Yahoo
See 50 Results	AltaVista, AllTheWeb, Excite, Google, HotBot, MSN **Not yet updated, but may be still correct:** Yahoo
See 100 Results	AllTheWeb, Google, HotBot, **Not yet updated, but may be still correct:** Yahoo
Sort By Date	MSN Search, Northern Light
Date Range	AltaVista, Google, HotBot, MSN, Northern Light **Not yet updated, but may be still correct:** iWon, Yahoo
Date Displayed?	AltaVista, HotBot (for Inktomi results), Northern Light
Display Titles Only?	AltaVista, Excite, HotBot (URLs only option), MSN
Other Major Customize Options	AltaVista, AllTheWeb, Google

Boolean Commands

(See Boolean Searching page for more details)

Command	How	Supported By
Or	OR	AltaVista, AOL Search, Excite, Google, Inktomi (HotBot, MSN), Lycos, Northern Light
	None	AllTheWeb, Direct Hit, LookSmart, **Not yet updated, but may be still correct:** Yahoo
And	AND	AltaVista, AOL Search, Excite, Inktomi (HotBot, MSN) Lycos, Northern Light
	None	AllTheWeb, Direct Hit, Google, LookSmart **Not yet updated, but may be still correct:** Yahoo
Not	NOT	AOL Search, Excite, Inktomi (HotBot), Lycos, Northern Light
	AND NOT	AltaVista, Inktomi (MSN) **Not yet updated, but may be still correct:** Netscape
	None	AllTheWeb, Direct Hit, Google, LookSmart, **Not yet updated, but may be still correct:** Yahoo
Nesting	( )	AltaVista, AOL Search, Excite, Inktomi (MSN), Northern Light
	None	AllTheWeb, Direct Hit, Google, Inktomi (HotBot), LookSmart, Lycos **Not yet updated, but may be still correct:** Yahoo
Near	NEAR	AltaVista (10 words), AOL Search (specify number), Lycos (25 words)
	None	AllTheWeb, Direct Hit, Google, Inktomi (HotBot, MSN), LookSmart
Notes At AltaVista, Boolean only works on advanced search page. At Excite, Google & MSN, Boolean commands must be in UPPERCASE At Inktomi-powered services, set menu to "Boolean"		