February 12, 2009

Search Engine Optimization:

Searching? What are you REALLY searching?

Finding the Web documents (a.k.a. Web "pages" or "sites") you want can be easy or seem impossibly difficult. This is in part due to the sheer size of the WWW, currently estimated to contain 1 billion documents. It is also because the WWW is not indexed in any standard vocabulary (unlike a library's catalogs which assign Library of Congress or some other widely accepted standardized subject descriptors to their documents, or a journal-article index which uses its standardized subject descriptors).

When you do what is called "searching the Web," you are NOT searching it directly. It is not possible to search the WWW directly. The Web is the totality of the many web pages which reside on computers (called "servers") all over the world. Your computer cannot find or go to them all directly. What you are able to do through your computer is access one of several intermediate databases and/or web-pages which contains selections of other web pages organized to allow you to find other web pages and sometimes other databases. You search these intermediate "search tools," and they can provide you with hypertext links (URLs) to other pages. You click on these links, and retrieve documents, images, sound, and more from individual servers around the world.

One last thing… a listing on a search engine is controlled by many factors but one thing to keep in mind is that "You cannot BUY a higher listing". A ranking on a search engine is not like buying a larger ad size or buying a TV ad during the evening newscast. You can make sure that your rankings will be better than average by selecting the appropriate and proper keywords and using the talents and services of a professional Site Marketing firm. (Yet even this is not always true…Yahoo allows only one *perhaps* two keywords.)

Additionally, then maintaining that service year after year, think of it as "making sure your phone number for your business is at least listed in the Yellow pages".

Some of the engines mentioned below are no longer available. But the information is presented to give you an overview of the types of Search Engine techniques and styles of searching, the names may have changed but the services are the same.

Introduction to Search Engines (Historical Review)

The Internet has tens of millions of sites at this point; growth is exponential and bibliographic control does not exist. To find the proverbial needle in this immense haystack (or tiny fly in the Web), you may use two basic approaches: a search engine or a subject guide such as Yahoo, AOL, LookSmart, and Google.

Subject guides are fine for browsing general topics, but for specific information use a search engine.

All search engines do keyword searches against a database, but various factors influence the results from each. Size of the database, frequency of update, search capability and design, and speed may lead to amazingly different results.

Recent addition of new content, redesign and partnership changes have fulfilled the mission of the new name: portals. This name implies a starting point and central location for all uses of the Web. We prepared this document to help you understand the differences between some of the most popular search engines.

There are also metasearch sites or metacrawlers that send searches to several search engines. Since metasearch engines do not allow for input of many search variables, their best use is to find hits on obscure items or to see if something is on the Internet. Some of the best-known ones are or were Dogpile, Inference Find, Mamma, Metacrawler, and SavvySearch.

Search Engines are often casually lumped together but there are some real differences: Yahoo, AOL, Looksmart-MSN are 'DIRECTORIES'. These are search tools that are managed and updated by 'humans'. Typically they are starting to charge for a listing ($250.00-350.00).

Excite, Alta-Vista, Google are 'SEARCH ENGINES'. These tools are managed by humans but most of the updating is done by software programs that 'crawl' your web site indexing the pages as they go along. These software crawlers are often referred to as 'Robots or Spiders'.

Feature Comparison Chart
(Some listed below no longer exist or were purchased)

AltaVista - Summary: one of the largest and most comprehensive search engines available, it searches the *entire* HTML file and also gives the option for searching Usenet.

Excite - The feature "More Like This" can be helpful in narrowing down the search; "Channels" have news and reviews of other sites.

HotBot - Summary: has some unique search features that make this a very useful search engine, including sorting results by date or media type.

Infoseek - Summary: search the Web, Infoseek Select sites, Usenet newsgroups, company directory, and timely news.

Lycos - Summary: one of the oldest search tools on the Internet, but keeping up-to-date with a variety of result options available.

Northern Light - Summary: the newest one with good, extensive content and a feature for classification of hits by division into custom search folders.

WebCrawler - Summary: best for identifying major WWW resources or most relevant sites for mainstream resources. It may be a good place to start for those new to keyword searching with full-text indexing.

SNAP – Summary: directory similar to Yahoo. Bought by NBC in 1999. Heavy promotion campaign on TV and is ranked 8th or 9th in size.

MSN – Summary: Default Home Page when you connect through MSN or do not change the default Home Page for IE. (Microsoft's Browser - Internet Explorer)

AOL NetFind – Summary: Similar to Excite, portal entry page for AOL members (20 Million strong).

For more information on any of these search engines, see the help files located on their sites.

How to Search a Search Engine Matrix

Information need	Characteristics of the search engine
I have a general idea of a **broad topic** and want to narrow it.	(Yahoo) organizes information as "subject trees" from general to specific topics. (Northern Lights) dynamically groups results into folders of related hits to help you narrow your query. Currently the largest engine indexing 16% of the web.
I want a small number of **relevant hits** and an idea of what's in each document before I go to each page.	(Google) returns important, relevant hits quickly, with your terms matched in close proximity and bolded in the results. (Excite) hits have excellent summaries. When you find a relevant hit, you can submit a "more documents like this one" to locate similar pages.
I want to **overview multiple engines together** to get an idea of what's available for my topic. This "metasearch" will also help me with harder-to-find or more recent information.	(Inference Find) performs parallel searches, merges results, removes redundancies, clusters the results (Dogpile) can be customized to exclude particular search engines, or search them in a certain order
I want **quality, evaluated sites** that have reviews and ratings because I have limited time.	(Magellan) smaller database containing descriptive reviews of sites (Lycos TOP 5% Home) reviews and rates the "Top 5%" (WebCrawler) smaller database with relevancy ranking
I need to do a **pinpoint search** because my terms are narrow.	(Alta Vista) massive and fast indexer of full text, good for very specific searches
Is there an **encyclopedia** of information on the Internet?	(Information Please) searches *Information Please Almanac*, *Entertainment Almanac* and *Sports Almanac*, the *Columbia Encyclopedia*, and the *Random House College Dictionary.* (Free Internet Encyclopedia) The MacroReference of the contains references to large areas of knowledge and FAQs; The MicroReference contains short bits of information and references to specific subjects.
I have **common keywords** that probably appear in many documents and should make my search specific.	(HotBot) fast, powerful, with ranked results and many options for defining a search

I have **common keywords in a phrase** like *New* in New Orleans or *A* in Vitamin A that cannot be ignored.	(Infoseek) includes words in a phrase that other engines ignore
I know the **date of an event** and am looking for more information.	(Hot Bot Super Search) limits by date
I need **programming language** (e.g. JavaScript, Shockwave) web pages or information.	(HotBot Super Search) limits by programming language
I need balanced information from **verified sources** for a **school research** project to **take home**.	(Nueva's Library catalog) has resources collected for the school's curriculum ProQuest Direct has full text of over 450 magazines and newspapers
I need **scientific information** to back up the research for my science fair project.	(Alta Vista) rated best for finding scientific information on the Internet in research study.
I need **mathematics** or **statistics** information.	(MathSearch) locates material concerned with research-level and university mathematics
I can describe my topic in a sentence (**natural language**).	(Infoseek) (Ask Jeeves!) allow natural language searches
I need information on a **proper name** (a place, person, or object).	(Alta Vista) (Infoseek) use capital letters to force an exact case match on the entire word (HotBot) person search will retrieve the name in both reversed and normal order (e.g. Picasso, Pablo and Pablo Picasso)
I want web pages from a **geographic region**	(MetaCrawler) can search by continent
I want web pages from an **Internet domain** (e.g.schools)	(HotBot Super Search) can search by Internet domain (e.g. edu, com)
I want to search for **images** and **sounds** (photos, art, designs, logos, videos, music, noises), **media types** (Java, VRML) or **file extensions** (.gif)	(Columbia University's WebSEEk) (AltaVista Photo Finder) (Lycos Media) (HotBot: SuperSearch) (The Amazing Picture Machine) (Yahoo! Computers and Internet:Multimedia:Pictures) (Arriba Vista)
I need a **quotation.**	(The Quotations Page) (Bartlett) (Land of Quotes)
I want to get **advice and opinions** from others.	(Reference.COM) (Deja News) searches archives of Usenet posts (strong in computers, the Internet, science, recreational activities, sociology, psychology, k12 education) Join a discussion group on a topic (Liszt)
I bet this search has been **done before**.	(Ask Jeeves!) has prepared answers to common questions.
I want to **browse**.	(A2Z)(Yahoo) are subject tree directories with short descriptions of sites
I want to see sites **just for kids**.	Yahooligans!

General Search Tips

What is the best search tool? It depends on your premises and why you need the information. If you are just browsing, start at Yahoo, or use the directory services of WebCrawler (GNN) or one of the other subject catalogs. If you are looking for best of web--and your interests are "pop,"--use Magellan. If you need a specialized database, try Internet Sleuth first. If you are doing "serious" research, start with AltaVista, but be prepared to use the other good search engines too, and follow these general rules of thumb:

- Enter as many precise search terms or phrases (if allowed) as possible in order to limit the search. The biggest problem is noise. That is, irrelevant or inconsequential sites among the jewels. Use of the required/prohibited term operator (prepending +/-) helps in reducing noise: +radio* -radiology

- Enter singular terms. Most search engines will find the substring and return rivers for river. To generalize a subject, use wildcards where allowed (surg* for surgery, surgeries, surgical).

- Do not use common, generic search terms, or if you must, include them in a phrase with more specific terms. The term book would be far too generic unless it were part of a phrase like "book binding."

- Enter multiple spellings where appropriate: Khaddafi Quadafy Kaddafi Qadaffi... If you know the correct spelling, using synonyms will broaden a search that is returning too little.

- Use booleans and especially proximity operators to increase the relevancy of your hits. Where allowed, (AltaVista, for example) you may control relevancy based on search terms. Including words in phrases with some search engines (Infoseek) is the same as using proximity operators. Use the adjacency operator where word order is important. WebCrawler has the best implementation of proximity and adjacency operators.

Most of all, be persistent and creative. It's a big web out there. The search tools are wonderful but far from complete. Be prepared to supply the ingenuity to make the most of their features.

Search Engines – *The Details*

AltaVista
AltaVista is the premier search engine on the web. It has the largest, most inclusive indices. That does not mean it is the only one you need, or in all situations the best one to use. Different robot and indexing strategies have resulted in different results when using the various search engines. AltaVista, however, returns consistently useful information, but since no editorial decisions have been made regarding content, it also has the largest "noise to signal" ratio.

AltaVista allows searching of both the web and many Usenet Newsgroups. It allows control of the result lists in a standard, compact, and detailed format. It provides both **simple** and **advanced** searches. Advanced searches include all the features of simple ones, and also allow the use of boolean and proximity operators, grouping of terms by parentheses, and results ranking by keyword.

Excite
Excite uses a combination of text and subject indices to search either by **keyword** or by **concept**. Concept searches, according to the Excite authors, find documents related to the idea of your

search, and not just documents explicitly containing the search terms you enter. From the initial screen you choose which way you would like to search, by clicking the **keyword** or **concept** radio button.
Concept is the default. You may search web documents, reviews, usenet newsgroups or classifieds. Simple or more advanced features are entered in the same search box.

There are not separate entry screens for either type of search, but advanced features like boolean searching and logical grouping are supported. You may not control the appearance of the hit list into standard/summary/detailed formats as you can with some other search engines.

WebCrawler

WebCrawler, now sponsored by America On-Line, is an outstanding search engine very much in the mold of AltaVista. In fact, it has more power than AltaVista in implementing advanced features such as the proximity operators NEAR and ADJ. It also includes a catalog of pre-classified subjects (directory services) by editors at GNN. It implements a feature of further searching based on pre-set search terms from the subject catalog, very much like Excite. (This feature hides behind the Spidey button. [Sometimes I feel silly writing this stuff.]). Finally, like AltaVista, it is so good in its own right, and associated with such a large company, that it can afford to be less gaudily commercial than Excite or Lycos.

WebCrawler touts "natural language searching," so you can enter a search like "highest mountain in the world." It throws out the noise words, and does a fuzzy AND search on the others, weighting pages with occurrences of all search terms highest, but including pages that contain only one of the search terms. This is the common strategy among the best search engines. WebCrawler is different in that its definition of "noise" words is rather broad. The term "web" for example, is not indexed.

Lycos

Many of us who have used the Internet for a while have a fond spot for Lycos from its Carnegie Mellon days when it was truly a Godsend. Since the explosion of the web, better search engines have appeared, but Lycos is still good and fast, if not as sophisticated as some of the others. It offers both keyword and subject searching (the subject searches are called directory services), as well as a Point rating system which rates web pages. Its strong points are its speed, ease of use, and the large size of its indices, which often produce usable results by sheer brute force. Its weakest point is that it does not support boolean searching or any of the more sophisticated searches that can be made with AltaVista, WebCrawler or Excite.

Infoseek

Infoseek was once the only Netscape default search engine. It is not the best available. Its virtues are speed and ease of use. Its defects are a lack of sophistication (booleans are not supported) and a 'teaser' approach to showing the first 100 hits and offering to show more for pay. It is both a search engine, and a searchable subject catalog, with options to search Usenet newsgroups, email addresses and web FAQs.
Searches are quasi-case sensitive. Capitalized words are taken as proper nouns and the search is limited. Searching for Babe will find the famous hitter and the famous pig, searching for **babe**, will also find the Sonny and Cher lyrics. Adjacent capitalized words links them into a phrase. Capitalized phrases must be separated with commas: **The Great Bambino, Baseball Hall Of Fame**. Phrases may be formed by enclosing the words in double quotes: **"I've got you babe"**. Yet a third way to link words into phrases is to place hyphens between them: **wonderful-life**.

Yahoo!

Yahoo is not a search engine, but strictly a hierarchically arranged subject index. It has developed over a long time, with lots of editorial care, so the quality is very high. Browsing Yahoo

is the best way to surf for good sites when you don't know (or perhaps care) where exactly you are going. It is also the best way to find good 'starter' sites, from which you can branch out to more specialized ones.

Using Yahoo is simple. Just enter your search term(s) in the search window and click SEARCH. Yahoo will return three types of information: 1) Yahoo categories that match the search term (so you can explore them for cross referencing); 2) Actual matching end-sites; and 3) The Yahoo categories from which the various pages are indexed--sort of a 'much broader term' cross reference.

NlightN
NlightN is more along the lines of the classical information/document delivery service, like Ebsco. The difference is you can use NlightN's Universal Index free, and only pay if you order a document. It indexes not only the web, but reference works, news wires, books, dissertations, and many public and private databases. NlightN bills itself as 'the world's largest table of contents,' but this is the sort of hype one gets used to using the Internet. AltaVista is the largest, but it depends on what you mean by 'table of contents.' Remember, this is a for profit organization, but you cannot spend money by accident, and are free to use the databases for research. They ask that you sign up for a free NlightN account, but this is not required. If you do, you gain some searching power, but you are free to make the call. If you are serious about using the service for pay, get the FAQ available from the help pages.

The Internet Sleuth
This is a very useful tool, but not as inclusive as you might imagine. Its concept is somewhat different from the tools considered here. It indexes a large number of databases, and provides a front end from which they may be searched. Therefore, in the opening search box it is best to put as broad a single term as possible, and then from the resulting search window(s) be more specific. For example, if I were searching for the lyrics to that timeless classic by Sonny Bono called 'I Got You Babe,' I would search initially on "music." This search would result in a list of 29 searchable databases, such as the CD-ROM Database, Music Colleges, Chicago Concerts, Smithsonian Folkways, and so on. Among the databases (for each of which the Sleuth presents you with a search window) there is one called Lyrics Server. In the search window titled Artist or Title the terms "sonny and cher" result in a list of two songs: 'I Got You Babe', and 'The Beat Goes On'. (Where are all the other great hits, one wonders). Clicking on the appropriate title yields the actual lyrics--reading which, in the light of subsequent history, is somewhat amusing.

Where the database being searched allows for booleans or wildcards, the Sleuth gives you search hints next to the appropriate search window. Even the Yahoo index can be searched from within the Sleuth.

Magellan
Magellan is not actually a search engine, but rather an on-line guide to the Internet that contains a directory of rated and reviewed sites, along with an index to lots of unreviewed sites. It is like Yahoo, only less inclusive with a more thorough rating system. (One to four stars, rather than Yahoo's shades to indicate a cool site). Magellan's strength is its system of reviews.

It is not a good starting place to do a search, but is rather more useful when looking for sites which are tried and true. The emphasis at Magellan is on pop sites (UFOs are one of the main categories on the front page), but if that is what you are looking for the site is great. The only drawback is the inevitable advertising.

(Yet Another)

Search Engine Comparison Chart

Database	AltaVista	Excite	HotBot	Infoseek	Lycos	Northern Light	WebCrawler
Scope							
Content Size	150M pages	55M pages	110M sites	75M pages	50M pages	160M sites	2M sites
Newsgroups	Yes	Yes	Yes	Yes	Yes	No	Yes
Full-text	Yes	Yes	Yes	Yes	No	Yes	Yes
Logic							
Default word	Or	OR	AND	Or	And	And	Or
Boolean connectors	And, not, near	AND, AND NOT	OR, NOT	And, not	Not, adj., near, before, far	Or, not	And, not, adj.
Phrase search	Quotation marks	Quotation marks	Quotation marks	Quotation marks	Quotation marks	Quotation marks	Quotation marks
Truncation	No, use *	Automatic plurals	No, use*	Automatic	No	Automatic	No
Case sensitive	Yes	No	Yes	Yes	No	No	No
Words included	Use +	Use +	Use +	Use +	Use +	Use +	Use +
Word elimination	Use -	Use -	Use -	Use -	Use -	Use -	Use -
Duplicate detection	Grouped under one title	Yes	Grouped under one title	Yes	Yes	Yes	
Special features	Limit by date, language, or format and field followed by:	Concept searching suggests terms	Limit by date, media type, file type, by suffix location		Search for image and sound files	Custom folders	
	AltaVista	**Excite**	**HotBot**	**Infoseek**	**Lycos**	**Northern Light**	**WebCrawler**

Using the various search tools on the web is enhanced by knowing how they were actually designed, and especially by knowing the specific rules--all too often quite different--for each tool. The search engines and catalogs are listed in order of usefulness, and then spell out, in short form, the rules for using them. Provided are some short examples, but for detailed examples consult the help documentation available at the site.