SEARCHING ON THE WORLD WIDE WEB

by

Dr. D. K. Singh* and
Dr. B. K. Singh**

ABSTRACT

Lot of Information is available on the World Wide Web. But Searching on the World Wide Web can be confusing. A myriad of search engines exist, often with little or no documentation, and many of these search engines work differently from the standard commercial search engines user normally use [(1)]. There are many directories, which attempt to organize the Internet by subject. Today there are many search engines that combine directory and keyword search capability. This paper defines search engine, directories and covers basics of searching, provides criteria for choosing search engines, as well as provides a comparison of some of the search engines available.

*** Assistant Librarian, Banasthali Vidyapith, Rajasthan**
**** Assistant Librarian & Officer In-charge, Kota Open University, Kota**

0 Introduction

The plentiful content of the World Wide Web is useful to millions [(2)]. Some simply browse the web through entry points such as Yahoo. But many information seekers use a search engine to begin their Web activity. In this case, users submit a query, typically a list of keywords, and receive a list of Web pages that may be relevant, typically pages that contain the keywords.

Many of the search engines use Well-known information retrieval (IR) algorithms and techniques [(3,4)]. However, IR algorithms were developed for relatively small and coherent collections such as newspaper articles or book catalogs in a (Physical) library. The Web, on the other hand, is massive, much less coherent, changes more rapidly, and is spread over geographically distributed computers. This requires new techniques or extension to the old ones to deal with gathering information, making index structures scalable and efficiently updateable, and improving the ability of search engines to discriminate. For the last item, the ability to discriminate, it is possible to exploit the linkage among web pages to better identify the relevant page s.

Web is huge and challenging. Several studies have estimated the size of the web [(5,6,7,8)], and while they report slightly different numbers, most agrees that over a billion pages are available. Given that the average size of a web page is around 5 to 10k byte, just the textual data amounts to at least tens of terabytes. The growth rate of the web is even more dramatic. According to Lawrence and Giles [(9)], size of the web has doubled in less than two years, and this growth rate is projected to continue for the next two years.

1 What are Search Engines and Directories?

Search engines in use on the Internet use automated programs, called robots, to search the web. These automated programs are also known as spiders, crawlers, wanderers and worms. The robots crawl about the web indexing web sites. Some of them index web sites by title, some by uniform resource locators (URLs), some by words in each document in a web site, and some by combinations of these. Because the Internet is always growing and because these search engines search in different ways and search different parts of Internet, doing the same search using different search engines will often give on widely differing results.

Many directories on the Internet were created by humans tired of stumbling about the Internet looking for topics of interest. These personal lists grew in size and complexity, and eventually the humans started to use the available search engines to assist them in their quest to bring order to the mess. Yahoo is perhaps the best known of the directories. It was started by a couple of students at Stanford and now employs a variety of people, including librarians, who review and categorize, web sites.

2. How to Search?

Browsing a directory is a simple matter of following the links for the topic of interest. Searching either a directory or the portion of the web that a search engine covers works very much the same in almost all search engines. The basic format is that of a dialogue box, pane, or line where search terms can be entered followed by options to either submit or clear the search. Once the search request is received, the search engine searches its own indexed database first, then, based on design, sends out spiders or other robots to add to the database. Results are sent back to the searcher, some annotated extensively, with links to the sources retrieved.

Full-featured search engines also have options to expand or limit searches in a variety of ways. For example, in Lycos, the basic search assumes a Boolean "or", which means those two or more terms will return results if any of the terms occur in documents indexed by Lycos. To obtain documents containing all the terms in a search, the Enhance user Search option must be chosen
and adjustments made to the default options.

3. Choosing a Search Engine

Choosing a search engine depends on the results user's looking for, though there are some criteria that may be useful. These criteria include:
Browsability – how easy is it to understand the results? Do user receive enough information from the retrieved results to make a decision about the usefulness of the results?
Customizability -- can user construct a sufficiently detailed search so as to eliminate or greatly reduce irrelevant results?
Relevance – no matter how browsable or customizable, are the results returned relevant to user search?
The following table describes various search engines for different type of information.

Fields & File Types	
If user wants to search for:	Choose:
Audio/Music	akoo.com, AllThe Web, AltaVista, CNETMP3 Search, FindSounds.com, FtpFind, Genie Knows, HotBot SuperSearch, Ithaki, Ixquick Matasearch, Lycos MP3 Search, Lycos RichMedia Search, MSN Search Advanced Search, MetaMission, Researchville, SavvySearch MP3 Search Turtle.com, Singingfish
Data last Modified	Alta Vista Advanced Search, AltaVista Search Assistant, Hotbot, MSN Search Advanced Search, Northern Light Power Search
Domain/Site/URL	AllThe Web Advanced Search, Alta Vista Search Assistant, Clickey, Direct Hit Advanced, Excite Advanced Search, HotBot SuperSearch, MSN Search Advanced Search, Namedroppers.com, Northern Light, SearchEdu.com
Geographic location	Alta Vista Search Assistant, Excite Advanced Search, Fossick, Northern Light, GeoSearch, HotBot SuperSearch, MetaCrawler Power Search, MSN Search, Advanced Search, Northern Light Geosearch, Northern Light Power Search
Images	AllTheWeb, AltaVista, The Amazing Picture Machine, Corbis, FtpFind, Genie Knows, Google Image Search, HotBot, Ithaki, Ixaquick Metasearch, Lycos RichMedia Search, MSN Search Advanced Search, Researchville, Savvy Search, Scour, SearchTurtle.com, Yahoo!News Image Gallery
Language	AllTheWeb, AltaVista, Excite Advanced Search, Google, HotBot, Lycos Advanced Search, MSN Search Advanced Search, Northern Lig ht Power Search
Multimedia & video	akoo.com, All The Web, AltaVista, FtpFind, Genie Knows, HotBot, Lycos RichMedia Search, MetaMission, Researchville, Scour, SearchTurtle.com, Singingfish, StreamSearch
Programming language	HotBot SuperSearch, MSN Search, Advanced Search
Proper name	AltaVista, Excite, HotBot, Yahoo, Search Options
Title	AllTheWeb Advanced Search, AltaVista, AltaVista Search Assistant, Direct Hit Advanced, Fossick, Google, HotBot, iLOR, Ixquick Metasearch, Lycos Pro, Mamma, Northern Light, Virtual Learning Resources Center

4. Search Engines: A Comparison

We can divide search engines into four categories: (I) Classics, (II) Leaders, (III) Newer Kids on the Block, and (IV) Search Engines for Search Engines. By Classics we mean search engines that have been around for awhile, that are well known and well used. Leaders are search engines that may or may not have been around for awhile, but are well known, have high use and return relevant results. Newer Kids on the Block is our designation for more recent arrivals on the search engine scene. And Search Engines for Search Engines covers two meta-search tools that give user a single interface for searching multiple searches engines at the same time.

The information given for each search engine is the name, the URL, how big the database is (if available), what it searches, general information on how to search, and why you might want to use it. Also included are characteristics specific to a given search engine.

5. Classics

World Wide Web Worm: http://www.goto.com/
A keyword oriented search engine good for general topic searching in a database of around 3 million sites. Has limited customization capability because it is forms based. Searches only http:// sites (no gopher, ftp sites). Once user make a connection to the server, the searching is very fast; of all the search engines we tried, though, this one took the longest to connect. The spiders in this engine seek out only URLs and web page titles for its index, so it's not the ideal place to find in-depth information on specific pages. Users can search with "and" and "or" Boolean operators, and they can retrieve sound/graphics files.

Why search with the Worm? It is good for simple, one or two-word topic searching, as well as generating lists of URLs in a certain area: Lists of business pages, organizations, etc.

WebCrawler: http://www.webcrawler.com
WebCrawler was begun in 1994 at the Department of Computer Science and Engineering at the University of Washington. In 1995 WebCrawler was punched by America Online, has spiders that crawl over the entire web looking for popular sites. They index the contents of the documents as well as the URLs and titles, and claim to update their entire database of around 500,000 web pages on a monthly basis. There are no descriptions of the sites with the results, which makes gauging relevance difficult. However, in many simple, broad topic searches, relevant home pages appear at the top of the results list, allowing user to avoid scanning long lists of less relevant sites.

This engine searches for ftp and gopher sites, not just http's. It searches words, not strings. For example, a search for "Colorado River" will turn up hits for those two words anywhere on the page. User can also search using the Boolean "and" and "or."

Why search with the WebCrawler? It's good for simple searches, has some customization capability--user can specify the number of words to search in their query and the number of desired results in blocks of 10, 25 or 100. They can also bookmark the results, making going back to specific sites very easy.

Yahoo: http://yahoo.com.

Yahoo is the web's most popular search service and has a well-deserved reputation for helping people find information easily. The secret to Yahoo's success is human beings. It is the largest human-compiled guide to the web, employing about 150 editors in on efforts to categorize the web. Yahoo has well over 1 million sites listed. Yahoo has a GUI (graphical user interface) that makes searching and browsing a piece of cake. It offers hourly news summaries from Reuters. Open Text search results are clearly marked, showing all URLs and the size of each. Results are scored by relevancy.

However, all these wonderful features of Open Text, including three types of searching, don't always work for simple queries. This is because the engine searches strings, not words. All words in a query must be present in the order given. However, the Boolean search capability is strong, and user can create their own weighted search. Yahoo! mixed with Open Text is a study in searching contrasts: On the one hand, the directory search does the work for user, on the other, user, the searcher, must do most of the work if they want the best results from the Open Text "power search."

Why search with the Yahoo? It's probably the best place to start any search of the Internet. It helps novices (and we're all novices in something) become acquainted with what the Internet has to offer.

EINet Galaxy: http://galaxy.einet.net/

The Galaxy is another hierarchical, topically organized search engine. Each topic has its own page in the Galaxy, and each page is organized into many lists. For example, the Topic List page provides links to other Galaxy pages containing specific information about your topic. Consists of a series of indexes from which to choose. For example, user can search an index of pages only found on the Galaxy itself, the web, gophers (to improve quality of gophers found, only those also referenced in Gopher Jewels appear in the index), Hytelnet--for access to thousands of telnet sites, and Galaxy Entries. This last index contains only information references in the Galaxy itself. Let's say user wants to know if there are any references to the American Association of Retired People, or AARP. One can search on the full word or on the acronym to find out if they should continue their search further. Boolean "and," "or," and "not" can be used to refine the search process.

The Galaxy has a link "One can add information to this page!" Clicking on it will bring up a form, which can be used to add references to an existing page, or send comments to Galaxy staff.

Why search with the Galaxy? It allows the option of searching areas of the Internet not found on the web. It has a convenient browse page with preformatted searches on approximately 100 commonly chosen topics to save user time. Has topic lists and document lists relating to users topic.

Leaders

InfoSeek Guide: http://infoseek.go.com/

InfoSeek Guide is the free directory and keyword searchable service of InfoSeek. Use the Guide to direct their browsing of the Internet or to look for specific information. InfoSeek Guide indexes over 1 million web pages. It also indexes Usenet newsgroups, FTP and Gopher sites, e-mail addresses, and Frequently Asked Questions lists. Search features are many, and complex. But even with the complexity InfoSeek Guide offers great search customizability and includes features such as: indexing of all words on a page, case sensitivity so that user can get a precise match on proper names, proximity searching, the "not" operator, symbol searching, and phrase searching. Results are ranked by relevancy and include that ranking, a link to the site of the information, the URL of the site, the size of the document, some description of the document, and a link to similar pages. Users can bookmark their results too, making return visits to the sites much easier.

Why search with the InfoSeek Guide? It's convenient (as of this writing it is the first search engine listed on Netscape's Net Search page) and offers many useful search features. Internet World tests also show it to provide the most relevant results [10].

Lycos: http://www.lycos.com/

It was one of the first engines developed for the Net. Back in December of 1995, Lycos claimed to have indexed 92% of the web. Now it claims to be the only complete guide to the Internet. Hype aside, they do have a huge database. They, too, have gone from being simply a keyword searchable index to adding a directory, which goes by the name of A2Z. Lycos also provides a service called Point, which provides reviews and ratings of the top 5% of all the Internet sites they index. Lycos searches every word in a web site and defaults, for some unfathomable reason, to an "or" search. To get the full range of search options user need to go into "Enhance your search". Once there, user can choose variations on "and" to match all user search terms, only two of their search terms or as many as seven search terms. User can also choose the level of relevancy of their search.

Why search with the Lycos? It covers a lot of the web, it is easy to use and the results are not only easy to read but user also get enough information in the standard display to determine how relevant the results really are. User can also bookmark their results, making return visits much easier.

OpenText: http://index.opentext.net/

OpenText provides little documentation on what or how it searches until user does a search, but it is popular because they do get results. The search form looks a bit intimidating at first, but is actually simple to use. User enter a word or phrase on each search line, indicate where they want to search (anywhere, summary, title, first heading, URL) and how they want to search (and, or, but not, near, followed by). Results include a link to the document, the relevancy ranking, and the size, the URL, an excerpt describing the document, links to similar pages and an option to see the matches on the page. This option lets user see the key words in the context of the document.

Why search with the OpenText? It offers a variety of sophisticated search options with a clear display of the results and extras such as links to similar pages and keywords in context.

Newer Kids on the Block

Magellan: http://magellan.excite.com/

Magellan offers added value to user searching by providing sites that have been evaluated by a staff of reviewers on the basis of depth, ease of use, and innovation. It also rates newsgroups, listservs and mailing lists.

User can search a directory mode: Explore Topics, or a keyword searchable mode: Search Magellan. Searches default to "or" if no other connectors are specified, and instructions are provided for Expanded Search utilizing more complex syntax.

Magellan provides a feature called Green Light which appears next to reviewed sites that, at the time of review, have no material "apparently intended for mature audiences." This feature pertains only to http sites, and applies only to the homepage itself, not to its links.

Why search with the Magellan? Its spider uses natural language processing software to hunt down sites for the database. Although it's a small database, it's growing at a steady rate. Thousands of users submit their sites for review, and there are over 1.6 million unrated sites found by the Magellan robot awaiting review. Its value lies in the refereed sites and the ease of searching–both of which will improve with time.

Inktomi: http://inktomi.berkeley.edu/

Full-text search engine for the web that claims to be the fastest (1-2 second response time), and is named for a Trickster Spider of Plains Indian mythology that brought culture to the people. The Trickster also represents the weak vs. the strong, the triumph of the underdog. Inktomi will accept upto 20 words in a query, and ranks documents by how many of the search terms are found in it. The searcher is offered the option to display results with or without full graphics (dispensing with graphics could be a real time-saver). It also searches for same word roots instead of endings (e.g. watch, not watch-ing, or watch-ed). Using a + (plus) before a word indicates that it must be included in the results. A - (minus) indicates it must be excluded from the results.

Why search with the Inktomi? Because it represents the future of web searching. It may now provide too many irrelevant results, but the technology is improving and a new iteration is imminent.

Alta Vista: http://www.altavista.com/

The Alta Vista is one of the most powerful and flexible of the major global WWW search engines on the Net today. Alta Vista searches for words on web pages. It allows user to perform simple or complex searches and has speedy retrieval times and well-developed robot technology (spiders, etc.). If no connector is used in the search the default is "or." Truncation is possible, as are field searches in text, URLs, title and links. The link search retrieves pages where at least one link represented on that page matches their search query. Advanced searching is also available by using Boolean operators and adjacency symbols. The near symbol ~ can be used as can parentheses for nesting.

Web pages are evaluated for relevance--its ranking system is not as effective as that of other search engines because it indexes any and all references to a search term, no matter

How far off it may be from the query's intent. Its search engine doesn't allow "stemming" as others do which means that search are performed only on the exact phrase–plurals and other forms of words are left out. However, if a document is found in user search, they can be sure their search terms are somewhere in it. Alta Vista also provides dates in its result list. Although user can refine their search by using the Power Search option, Alta Vista doesn't have as much on-screen help as other search engines. In terms of sheer scope, however, User will know the Internet universe was scoured once their query is sent out. They can bookmark their results, making future site visits much easier.

Why search with the Alta Vista? Because it searches for the obscure and hard-to-find subjects and performs its searches with speed. If User wants to find as much as they can about a certain topic, this is the search engine for user. Digital's Alpha architecture, and claims to have 21 million, fully indexed pages in its database power its spider technology.

Excite: http://www.excite.com
This search engine offers two ways of searching Concept or keyword. Many times there are no significant differences between the results of these searches. There is no Boolean searching, so trying to find specific information on a topic can be frustrating. The pluses of this engine however lie in its service offerings: User can do a directory search, much like that of Yahoo, or a keyword search. They can search for reviews, cartoons, news summaries, newsgroup texts and public ads. Unlike Alta Vista, its aim is not to build a comprehensive database, but one that is popular and current. The entire database is checked and updated weekly by spiders that are sent out on specific missions: One is sent to what's new sites to compile a database of new URLs. Another is then sent out to bring back the page contents to the Excite database.

Why search with the Excite? Because it incorporates the technology of the future: Concept searching, using natural language processing, needs to be further refined in this engine, but it's being utilized. Excite also provides a complete search service, with news, subject searching and classified ads.

Search Engines for Search Engines a.k.a. Meta Search Engines

MetaCrawler: http://www.go2net.com/search.html
MetaCrawler is a search service that has no internal databases. It simply acts as a front end for 9 different search engines: OpenText, WebCrawler, Inktomi, Alta Vista, InfoSeek, Yahoo, Lycos, Excite, and EINet Galaxy. MetaCrawler sends user query to the search engines then puts them into a uniform format for display. The search screen gives user a number of options. There is the usual search line but beneath it are 3 search

options: search as a phrase (~3 min), search all these words (~ 1 min), search any of these words (~ 1 min).

Why search with MetaCrawler? It provides a single interface for 9 popular search engines, allows user to use some fairly sophisticated search options and will check the document URLs to make sure the link is valid.

SavvySearch: http://www.savvysearch.com/
SavvySearch is a search tool that provides a common interface for searching a variety of search engines. One enters their search on the Query line and it sends their query to multiple search engines. It ranks search engines by a number of factors, including how appropriate they might be and how fast the response time is currently. By requesting that the results be integrated, it will remove duplicate results! To search, enter the search words, choose the "and", "or", or "adjacency" operators from the query options, choose the number of results to be returned from each search engine, choose the display format, tell it to integrate the results if user want, and wait. Since it is searching more than one search engine, the wait may be longer than that when using a single search engine. The normal display will give user most of the standard display for the specific search engine providing the results. If the results are coming from WebCrawler, user gets the URL, if they are coming from OpenText, you will get the usual OpenText display. SavvySearch lists the name of the search engine providing the results. Another nice feature is that SavvySearch is currently available in 18 different languages.

Why search with the SavvySearch? It's one stop shopping and it searches a lot of different search engines. In one search it reviewed 17 search engines as having possibly relevant information and searched 3 of them.

6 Conclusion

There are many similarities and differences in the way the search engines work. Think about what you want to get out of your search, try out a number of the search engines, and understand that the Internet and the search engines are changing daily. Yesterday's favorite search engine may be completely different today, and, most certainly, yesterday's search will provide completely different results today. The concept of an expert as someone who knows almost everything about a subject is no longer valid. A better definition may be that an expert is someone who adapts to new information, digests it more quickly, and soon is hungry for more.

For getting the best results form the search engines, use the "advanced features" and options that are available. When getting to know a new search engine, take a minute to read the "helpful hints" for searching that many of them provide on their main pages.

Reference

GAN (ANN EA) and BENDER (LAURA). Spiders and Worms and Crawlers, Oh My: Searching on the World Wide Web.
http://www.library.ucsb.edu./untangle/eagan.htm

ARASU (ARVIND) et al. Searching the Web. ACM Transactions on Internet Technology. I,1; 2001. p 2-43.

SATON (G) (ed). Automatic Tent Processing. 1998. Addison-Wesley Series in Computer Science. Addison-Wesley, Inc., Reading, MA.

FALOUTSOS (C). Access methods for text. ACM Comput. Surv. 17,1;1985. p 49-47.

BAR-YOSSRF et al. Approrcimating Aggregate Queries about Web Pages via Random Walks. In Proceeding of the 26th International Conference on Very Large Data Bases. 2000.

LAWRENCE (S) and GILES (C). Accessibility of information on the Web. Nature 400. 1999. p 107-109.

LAWRENCE (S) and GILES (C). Searching the World Wide Web. Science. 280. 1998. p 98-100.

BHARAT (K) and A.BRODER. Mirror, mirror on the web: A study of host pairs with replicated content. In proceeding of the Eighth International Conference on The World Wide Web. Ref. No. 6. 1999.

VENDITTO (GUS). Search Engine Showdown. Internet World. 7,5; 1996. p 79-86.

Bibliography

Comparing Search Engines.
http://www.hamline.edu/library/links/comparisons.html

DECY (DON E). All Aboard the Internet: Searching the World-Wide Web. Techtrends. 40,4; 1995. p 7-8.

GRALLA (PRESTON). Underground Internet. PC Computing. 8,11; 1995. p 195-200+.

MOELLER (MICHAEL). Open Text, Yahoo Meld Search Engines. PC Week. 12,38; 1995. p 135.

NOTESS (GREG R). Searching the World-Wide Web: Lycos, WebCrawler and More. Online, 19; 1995. p 48-50+.

JANE SCALES (B) and CAUFIELD FELT (ELIZABETH). Diversity of the World Wide Web: Using Robots to Search the Web. Library Software Review. 14,3;1995. p 132-136.

UDELL (JON). Web Search. Byte. 20,9;1995. p 223-224+.