PRINCIPLES OF WEB SEARCHING

http://www.mannlib.cornell.edu/reference/tutorials/search/index.html

---

Tutorial developed by Tom Turner, Philip Davis , and Jim Morris-Knower

Introduction

This tutorial discusses ways of finding information on the World-Wide Web. It assumes that you have some familiarity with the World-Wide Web and with the operation of Web browsers, such as Netscape. If you are not familiar with these technologies, you might want to look at Surfing the Internet on the World Wide Web, before taking this tutorial.

What are the best tools to find information on the Internet?

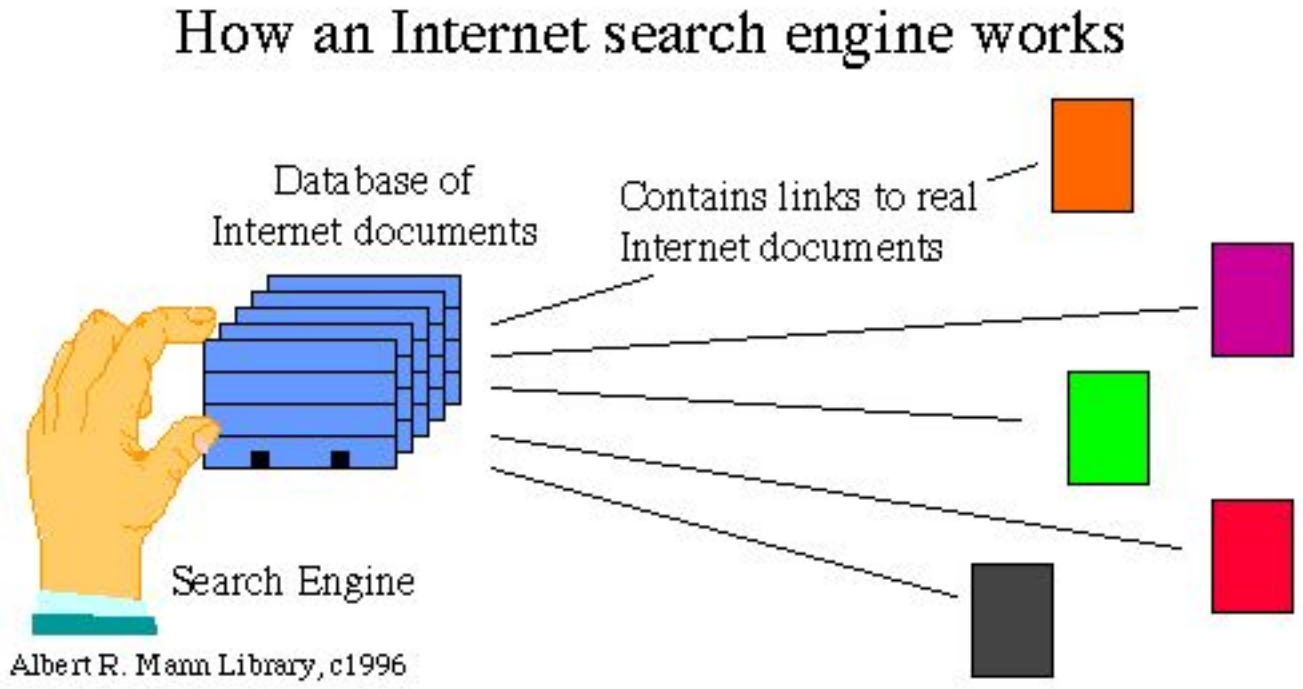
Searching for information on the Internet is more challenging than searching for information in a well-organized library. Browsing the Web can be extremely time consuming and often results in little gain; unless of course, you set out just to have fun. When you want to do research, you need a more structured approach to finding materials.

In this tutorial, we will discuss **Internet search engines**. We will talk about how these search engines organize (or don't organize) information on the Internet. Once we've set that groundwork, we will move on to the uses of particular search engines.

What Is an Internet Search Engine?

Search engines are the software used to search databases. **An Internet search engine is software used to search a database of Internet documents.** The following diagram illustrates how an Internet search engine works. Each time you use one of these search engines, you are querying a database of Internet documents. Each of these records contains a link to a real document on the Internet.

This is contrary to the popular belief that you search the entire Internet every time you use an Internet search engine. This also explains why sometimes a search result will display a "file not found" message when you try to link to that website.



The biggest differences among search engines are how the databases are created. A database can be created either by hand or by automated means. Automatically-created databases will be referred to in this tutorial as ***robot-generated databases*** and human created resources as ***human-selected databases.*** There are other notable differences as well. Some Internet search engines focus on particular subjects or geographic locations. Some also have browsable subject lists. It's these differences that make it difficult to know which one to use.

Some important points to keep in mind:

- **No search engine is perfect.** They are too new to be flawless. Many lack important search features. Many contain links to web sites that have moved or no longer exist. Many may not contain the site you need.
- **No database is complete.** No one search engine will help you find everything. They each have sites that the others do not. Each also displays results in a different way.

- **No database is up to date.** Since the content of the Internet changes on a daily basis, Internet databases always lag behind what is really out there. Some haven't been updated for nearly six months.
- **Search engines often change.** Internet technology changes rapidly. As a result, Internet search engines often differ each time you use them. Certain features usually remain the same. Sometimes, however, Internet search engines disappear entirely.

Basic Web Searching Tips

Before diving into the different types of search engines, we thought it would be helpful to offer some basic searching tips that apply to almost every search engine.

While of course you can simply type your search terms into the search engine's search box and get results, there are a few basic rules to follow to help focus your search and increase the relevancy of what you retrieve.

Here are 4 to get you started:

Rule #1--Use + and - signs
Say you are searching for information on growing Pinot Noir grapes in New York state, and you type in:

growing Pinot Noir grapes New York State

Certain search engines will assume that you want results with *any* of those words. Some of the major search engines that perform a default "or" search are AltaVista, Direct Hit, Excite, LookSmart, and Yahoo's Web Page search.

With these search engines, you will need to put a "+" sign before every word you absolutely, positively must have show up. Other search engines assume that if you type in words you want them to be in your results (a default "and" search).

Almost every search engine will allow you to exclude words by putting a "-" sign before them. If you were interested in apple growing in New York, you might want to exclude words like "computer" or "Macintosh" to screen out info on Apple computers. You might type:

apple growing new york -computer

Rule #2--Use quotation marks for phrases
To further refine your search results, put phrases in quotation marks. For the wine search, you might type:

growing "pinot noir" grapes "new york state"

In Google, this phrase search resulted 287 results, down from 1060 when no quotation marks were used.

Rule #3--Use truncation symbol when available
The truncation or wild card symbol is placed at the end of a word to tell the search engine to retrieve any and all forms of that word. Common truncation symbols are + and *; thus, grow+ would look for the

words grow, growing, growers and growth (among others). Not all search engines will allow truncated searches, and the symbol used varies among those that will allow them.

Here's a brief list of common truncation symbols and the search engines that use them:

*	AltaVista, Inktomi (iWon), Northern Light
?	OL Search, Inktomi (iWon)
%	Northern Light

Rule #4--Use site searching to specify where you search

Most search engines will allow you to specify where your search results come from. These sites can be broad, like all .gov sites, or specific, like all information from www.cornell.edu. As with truncation, different search engiens have differernt commands for site searches. Here's a list:

host:	AltaVista
site:	Excite, Google (Netscape, Yahoo)
url.host:	AllTheWeb, Lycos (for AllTheWeb results only)
domain:	Inktomi (HotBot, iWon, LookSmart)

So if you wanted to search via HotBot all educational (.edu) sites with information on genetically modifed organisms, you might type:

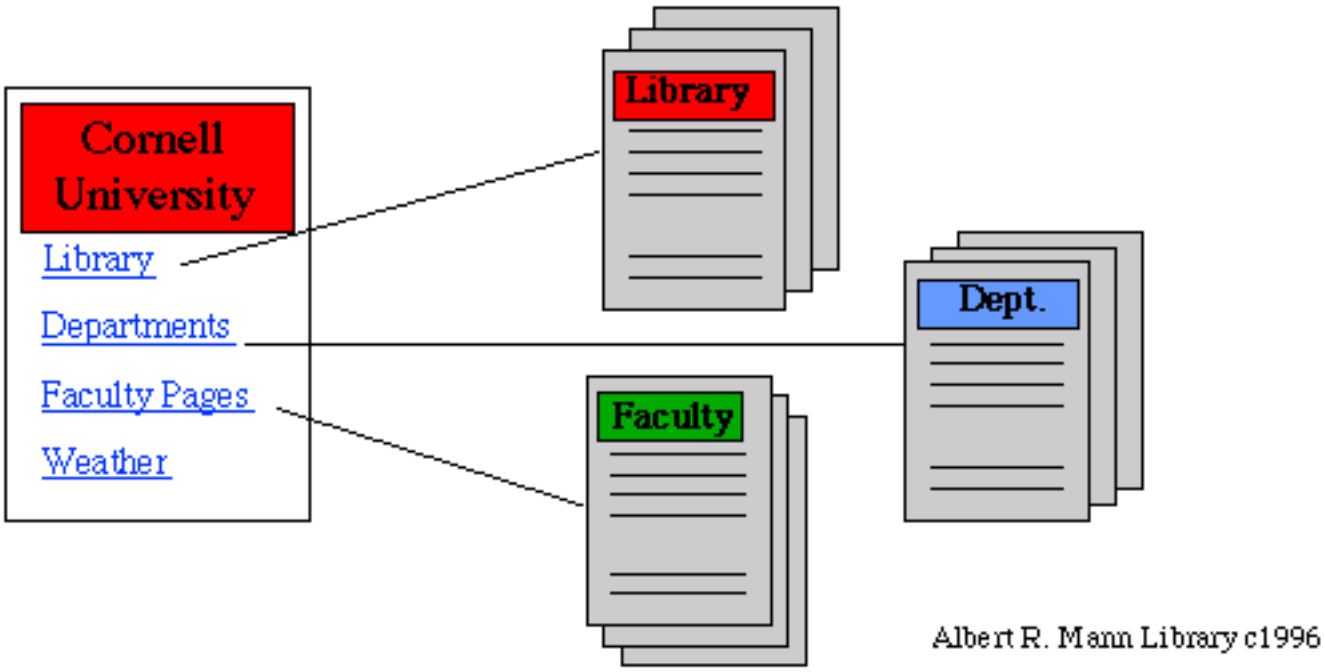
"genetically modifed organisms" domain:.edu

Or if you were searching Google for admission information from Stanford, you might type:

admission site:www.stanford.edu

Robot-Assembled Databases

Robot-assembled databases are created by computer software, often called **robots**, crawlers, spiders, or worms. These programs differ in how they go about indexing materials, but the databases they build are usually quite large and many of them are well-known.



The robot begins at a document (let's say the Cornell University Home Page), and records relevant text into its own database. It then connects to all of the pages linked to the first page, and adds them to its database. In this case, it looks at the library, faculty and departmental home pages. It then accesses all the sites linked to by these pages. The robot just keeps going.

Ranking Results

1. Food Dye Instead of Ink in a HP deskjet 540 cartridge
   91% - Directories & Lists: Posted by on December 02, 2001 at 06:36:01: Hello,My question is a rather odd one, however any help would be appriciated. I
   Commercial sites: http:/ / stroudco.com/ wwwboard/ messages/ 209.html
2. Research on Food Dye
   90% - Articles & General info: Some Research on Food Dyes Last update 12/20/2000. what is allowed and where, A study on the reproductive toxicity of
   Non-profit site: http:/ / feingold.org/ research dye.html
3. FOOD DYE
   89% - Directories & Lists: Posted by on February 28, 2000 at 15:40:01: I have a 4-year old grandson who is alergic to food dye.Would like ... 04/24/2000
   Commercial site: http:/ / www.thefunplace.com/ fence/ techboard/ messages/

Most robot-assembled databases also provide a way of **ranking results**. Ranking is based on how relevant the Web document is to your search. Results are based on a mathematical formula that takes into account how the Web page is marked up by HTML.

The most important factor is location--where your search terms show up in a web page strongly determines how relevant it is. For example, a Web document would receive a higher ranking if the search words are found in the title or header than at the bottom of the page. Another important factor is how many times your search terms who up on any particular page.

Money is another factor. Increasingly, search engine companies are "selling" their rankings, with "sponsored" sites geting relevancy boosts by paying for them. For more on this, see the June 2001 San Francisco *Chronicle* article on the subject.

Robot Generated Databases

Advantages	**Disadvantages**
Very large. Good for locating hard-to find documents	No human quality control done for the content of documents
Results are often ranked in order of relevancy to the search terms	Relevancy to search topic is often questionable
	Information contained in images is not indexed

Some examples of robot assembled databases:

- Alta Vista
- Excite
- HotBot
- Northern Light
- Lycos
- Google
- WISEnut

MetaSearch Engines

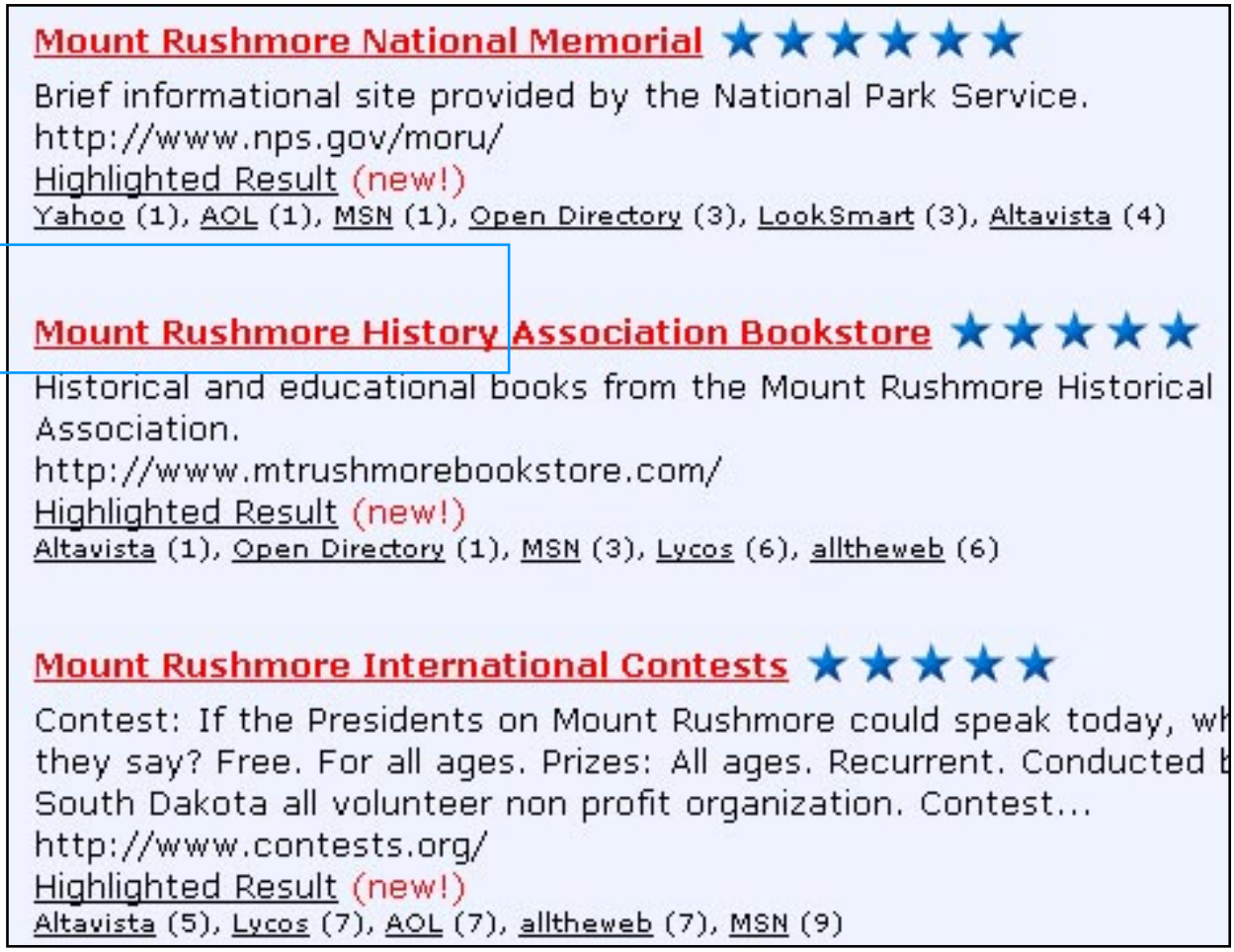
Metasearch engines represent a newer trend in Internet access technology. Rather than compiling a database of their own, they rely on other Internet search engines to do the work for them. Each takes a query, searches it in several different databases and retrieves or points to the results that are found. It can often be confusing to understand how these search tools are getting their results and exactly how the searches are being conducted.

These tools are very useful for doing searches for materials that you suspect exist on the World-Wide Web. They are intended to save some time by enabling you to search a variety of search engines without having to connect sequentially to each.

Here is a list of all the search engines that one metasearch engine, Vivisimo, searches. While each meta search engine searches a different group, this list gives you a good idea of the scope of these all-in-one finding aids.

In this search of Ixquick for information on Mount Rushmore, the first result shows the official Mount Rushmore page as retrieved by six search engines.

Ixquick awards one star for each search engine that placed a site in its top ten, and then lists where in the top ten the page was.



Strengths	Weaknesses
Searches multiple search engines at once	Often lowest common denominator in terms of search capabilities
Many metasearch engines group results in clusters	Some redundancies, often no detailed entries
Some allow users to set search time limits	Limited number of results from each search engine

Some examples of metasearch engines:

- Dogpile
- Ixquick
- Metacrawler
- Profusion
- Vivisimo

For names of more metasearch engines, see searchenginewatch.com's Metacrawlers page

Geographically-Focussed Search Engines

Some databases only contain Web sites from particular locations. Some of these databases are created automatically using a robot while others are generated via human selection. In many cases, the search engines themselves are in the language of the geographical area. This type of database is useful when looking for information that is related to a particular region. Its important to remember that materials about a particular area are not always produced or housed in that country. Some of these databases are, in fact, subsets of larger databases (ie. Lycos and Yahoo).

An example of a geographically-focussed search engine--*Yahoo China*

Some other examples of geographically focussed search engines:

- Yahoo China
- Ananzi South Africa
- Iran Index
- MexicoGlobal
- Lycos France
- Excite Australia

For names of more geographically focussed search engines, see searchenginewatch.com's Regional Search Engines page

Subject-Specific Databases

Like their geographically-focused counterparts, **subject-specific Internet databases** can be created either by robots or by human effort. Although the catalogs are not usually large, they do reflect a tighter control of subject matter than many of the general catalogs. As a result, when you search with a particular word, you may get a meaning that you expect. For example, an agricultural database will give you different results with the search word *apple* than a general database, where you might get results that include Apple Computers

A few examples are:

- FirstGov
- Search4Science
- MedHunt
- LawCrawler

Evaluating Content on the Internet

Is it Scholarly, Substantive, Popular or Sensational?

Scholarly information is written by academics for the academic community. Often these journals are online supplements to existing printed journals. The information normally contains experiment data, graphs, and citations. The language used reflects the discipline covered. Examples: Journal of Biological Chemistry | Psycholoquy | Nature

Substantive information generally covers news and other events. Substantive articles will often report scholarly information in a way that is understood by a larger group of educated people. They are usually more attractive than scholarly articles, and often contain illustrations that make the information clearer for the layperson. They are typically written by an editor or reporter of a news service. Examples: New York Times | Washington Post | Scientific American

Popular information contains lots of illustrations, and are written at a level that most of the population can understand. The main purpose of popular information is to entertain, sell products, or promote a viewpoint. Examples: Reader's Digest | PC Magazine

Sensational information is intended to create disbelief, or make inflammatory remarks about politicians and celebrities. It can be viewed as entertainment (or truth for those of us who are that gullable!). Examples: The Weekly World News | National Enquirer

Check the Source

On the Internet, everyone can be his or her own publisher. Many have argued that this has democratized the Internet, allowing voices that were not previously heard in the publishing world of books and magazines to be heard for the first time. On the other hand, editors (who work for publishers) act like filters, and become a quality check for what gets printed.

If the Internet document is not coming from a well-known publisher, check the credentials of the author. Is the author even listed? Is there a way of contacting the author? Is the author credible in the printed world. Is the author representing a larger group of people? The saying "on the Internet, no one knows that you're a dog", may have some truth to it.

Check the Domain

The domain name is the last part of the Uniform Resource Locator (or URL). The most common domains include: **edu** for the educational domain, **gov** for government, **com** for commerce and **org** for organization. Countries outside of the United States tend to use country codes, ie. **ca** for

Canada, and **fr** for France.

The domain name may help you evaluate the reputability of the information. Companies won't necessarily lie to you, but they may report findings in a way as to sway you to purchase their product or service. Public policy and statistical information is generally more reliable coming from a government sites or an educational site.

Is it Current?

Check to see when the file was created, and the last time it had been updated. You can usually find this information at the bottom of each Web page. Currency is typically not an issue for Internet documents, unless however, they contain time-sensitive information (like stock quotes).

Further Reading on Search Engines

Websites

Search Engine Showdown: The Users' Guide to Web Searching

Hosted by Greg Notess, librarian at Montana State University and a columnist for Online and Econtent magazines. Features reviews of and statistics on search engines, as well as links to Greg's columns.

Search Engine Watch

Hosted by Danny Sullivan, an Internet consultant and journalist. Site includes extensive tips on using search engines, as well as reviews and tests.

About: Web Search

Hosted by Kevin Elliott, president of a Web consulting firm based in Alberta, Canada, this is About.com's guide to web searching. Includes archive of Elliott's articles on the subject, as well as subject guides in categories like New Sites, Search News, How To Search, & The Big Ten.

Articles

Verne Kopytoff, **Searching for Profits Amid Tech Slump, More Portals Sell Search Engine Results to Highest Bidder**, *San Francisco Chronicle, Monday, June 18, 2001*.

ABSTRACT: Once relatively objective, search engines are increasingly becoming commercial. In an effort to survive the online industry's financial struggles, they are providing links based not just on relevancy, but on who pays for top billing.

Mark L Van Name & Bill Catchings, **Searching for the Truth,** *PC Magazine; New York; Oct 16, 2001; Vol. 20, Iss. 17; pg. 133-.*

ABSTRACT: Van Name and Bill Catchings evaluate the quality of of five popular Internet search engine sites, including Google, Northern Light, HotBot, Direct Hit, and Oingo, using a new Web site evaluation tool. The speed of each search engine, the relevance of its results, and the general impressions of the testers are discussed.

Gary Price, **Web Search Engines FAQS: Questions, Answers, and Issues**, *Information Today, Oct. 2001*.

ABSTRACT: Reviews the latest goings on in the search world and tries to provide some suggestions and tools to make you more knowledgeable and save you some time.

Randolph Hock,: **Revisiting Web Search Engines**, *Online; Sep/Oct 2001; Vol. 25, Iss. 5; pg. 18-*.

ABSTRACT: A chart listing Web search engines' features and commands is presented. The goal of the chart is to present in a clear fashion those items of information about search engines that will aid the searcher in making the most effective use of what the search engine has to offer. Every major search engine now has at least one
advanced version.

Lisa Guernsey, **Mining the 'Deep Web' With Specialized Drills**, *New York Times, Jan. 25, 2001*.

ABSTRACT: Traditional search engines have access to only a fraction of 1 percent of what exists on the Web.
As many as 500 billion pieces of content are hidden from the view of those search engines, according to BrightPlanet.com, a search company that has tried to tally them. To many search experts, this is the "invisible Web."

Robert Berkman, Searching for the Right Search Engine, *The Chronicle of Higher Education, January 11, 2000, p. B6.*

ABSTRACT: One important lesson is to understand the range of search tools now available. Many researchers don't realize that they can use hierarchical indexes, standard search engines, alternative search engines, meta search engines, and databases--and that those tools are not all the same.