

Searching the Web Effectively

UK
UNIVERSITY
OF KENTUCKY
Human Resource
Development



Participant Guide

HRD Technology Training

For technical assistance, please call 257-1300

Copyright 2003

Searching the Web

Course Objectives

After completing this class, you should:

- Understand how a search engine works.
- Understand the difference between an index search engine and a directory search engine.
- Select and use the appropriate search engine for your needs.
- Formulate your query by defining your search terms and key words.
- Evaluate the results of your search and refine your search terms if necessary.
- Evaluate the quality and credibility of the sites produced by your search.

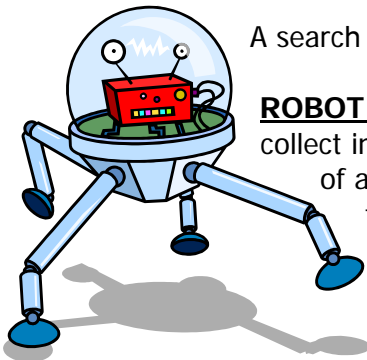
Five Steps to Searching the Web:

1. Select a search engine
2. Formulate your query (define your search terms and select keywords)
3. Examine the results and select the sites you want to explore
4. Refine or change your search terms if necessary
5. Evaluate the quality of the sites you have selected

STEP 1: Select a Search Engine

A search engine is a program that searches web documents for specified keywords and returns a list of sites that contain the keywords. Although "search engine" is really a general class of programs, the term is often used to describe specific systems like Google and Hotbot.

In a nutshell, a search engine sends out a "robot" or "spider" to fetch as many web documents as possible. Then another program called an indexer reads these documents and creates a database based on the words contained in each document. Each search engine uses a proprietary algorithm (a step-by-step problem-solving procedure) to create its database so that, ideally, only meaningful results are returned when a search is performed.



A search engine consists of several parts:

ROBOT, SPIDER, or CRAWLER – This software visits web pages to collect information to put in the database. It often starts at the main page of a site, and then follows links to other pages and returns periodically to the site to look for changes. The program is called a spider or crawler because it crawls over the web.

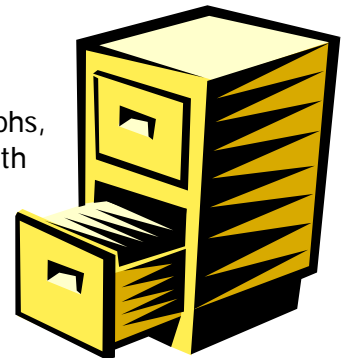
How does a robot decide where to visit?

Each robot uses a different strategy. Generally a robot starts from a historical list of URLs, especially of documents with many links elsewhere, such as server lists, "What's New" pages, and the most popular sites on the Web. Most indexing services also allow web designers to submit URLs manually, which will then be queued and visited by the robot/crawler.

Sometimes other sources for URLs are used, such as news group postings and published mailing list archives. Given these starting points, a robot can select URLs to visit and index, and to parse (break down and examine) and use as a source for new URLs.

How does an indexing robot decide what to index?

Some robots index the HTML titles, some index the first few paragraphs, and some parse the entire HTML source code and index all words, with weightings depending on HTML constructs. Some parse the MetaTag, which is a special HTML command that provides information about the web page, such as what the page is about or what keywords represent the page's content.



DATABASE or INDEX – This is produced when a robot or spider crawls across the web, collecting and indexing web pages. The search engine's information is indexed, stored, and updated here. Some time may elapse between the collecting and the indexing, depending on the search engine.

RETRIEVER – This is software that sorts through pages stored in the database, to find matches to a search and rank them in order of relevance.



INTERFACE – This is what you use to query the database. It usually consists of a search box in which you type a query and a button you click to launch the search. Sometimes there are additional menus and fields to further refine your query. This is the part you see – the other functions occur behind the scenes.

Each search engine collects information differently and offers different search options, search speed, and

interface design. Factors that affect results include the size of the database, how often the database is updated, and the engine's search capabilities

How do search engines index their databases?

Indexing can include every document on a site, or only the homepages.

Within a document, a search engine might index every word (**full-text indexing**) or only important words or phrases (**keyword indexing**). Some engines have a person look at each page (**human indexing**) to decide which keywords or phrases to include, according to specific criteria.

Ranking

Search engines use a ranking algorithm to determine the order in which documents are listed on the result page. Documents can be ordered according to how many of the search terms are included, how often the terms appear, where the words are located in the document, and how close they are to each other.



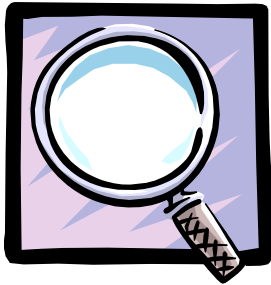
Generally the results are listed in order of importance, so **the first 20 to 50 hits are likely to be of most relevance to your request**. The highest-ranking documents will include all of the keywords that had to appear, and then as many as possible of the other search terms. If all the terms appear, the engine will consider that site most relevant and will put it first in the results.

Other things that enhance relevance and result in higher ranking include:

- Unique or unusual keywords
- Keywords that appear several times in a document
- Keywords that appear in the title/header or at the top of the page

Some search engines also look at how **popular** a site is:

- How many links there are to the site from other pages
- How popular pages are with users, by looking at what they click from search results
- Whether the site is listed in the search engine's directories.



Types of Search Engines

INDEX search engines are usually based on *software* (*robots or spiders*) that crawls on the web and collects a great deal of information but does not evaluate it. As a result, these contain more information than other search engines. Examples: Google.com, Teoma.com

DIRECTORY search engines (subject guides) are compiled by *people* and use *human indexing*. When you search a directory, the search engine looks at the descriptions in its own database rather than at the site itself. Directory databases are organized by category or subject and are generally smaller than index engines. Examples: Yahoo.com, About.com

When you are looking for SPECIFIC information, start with an index search engine.

When you are looking for GENERAL information, use a directory.

TIP: Some directories will give you results from a search engine if they cannot find what you are looking for in their own database, and some index search engines also offer a directory.

REGIONAL SEARCH ENGINES focus on one particular language or region. Most are directories and quite limited. Example: italiamia.com

METASEARCHERS use a single uniform interface to search using several engines simultaneously. You type your query once and the results are produced quickly. The downside is, the uniform interface prevents you from using advanced features of individual search engines. Also, metasearchers do not conduct exhaustive searches, but instead only return some pages from each search engine (generally the top 10-100 hits). Metasearchers are useful for searching obscure terms. Examples: Metacrawler.com, Webcrawler.com

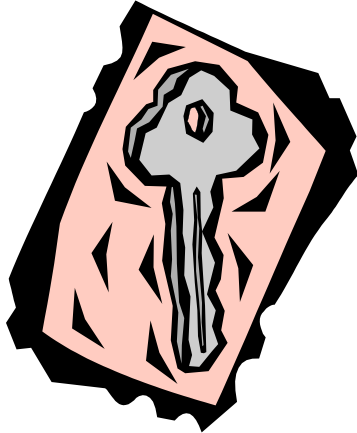
SPECIALIZED DATABASES provide specific information, and sometimes the contents can only be accessed through the search function provided with that specialized database. Examples: imdb.com (Internet Movie Database), lawcrawler.findlaw.com (LawCrawler)

TIP: Try different search engines, select two or three that you like, and learn to use their features such as advanced search and image search.

STEP 2:

Formulate your query (define your search terms and select keywords)

No matter which search tool you use, the following strategy should be effective:



1. Formulate your question
Example: What do my cholesterol test results mean?
2. Identify the important concepts within the question
Example: what are normal cholesterol levels and how do my numbers compare?
3. Identify search terms to describe those concepts
Example: cholesterol, normal, levels
4. Consider synonyms and variations of those terms
Example: levels, test, measurement
5. Prepare your search logic
Example: cholesterol +test + "normal levels"

Search Tips:

Use nouns and objects as query keywords. Verbs, modifiers, and conjunctions are often ignored by search engines or are too general to be useful.

Use six to eight keywords in your query. More keywords can reduce the total documents returned and help ensure that the ones you get are relevant.

Use the OR operator to include synonyms. This helps cover the different ways a concept may be described. *Example: cholesterol +high OR elevated*

Combine keywords into phrases when possible (use quotation marks to indicate phrases). Phrases restrict results to EXACT matches and can better narrow and target results.

Combine two or three concepts in a query (use quotation marks to indicate concepts). Multiple query concepts narrow and target results. *Example: "high cholesterol" AND "treatment options"*

Order concepts with the subject first. Some search engines rank documents more highly that match the first terms or phrases provided in the query.

Use the AND operator to link concepts. The resulting query is not overly complicated and correct left-to-right evaluation order is ensured.

Some search engines allow you to truncate words to pick up singular and plural versions (use the asterisk wildcard). The wildcard tells the search engine to match all characters around it, preserving keyword slots and increasing coverage.



STEP 3: Examine the results and select the sites you want to explore



Relevancy Ranking – The first few pages of results should yield the most productive information.

Ctrl-F (Find command) – If it is not obvious why the search engine has produced a document, use the Find command to look for your keyword(s).

Bookmark your results – This will allow you to go back to the search without resubmitting it.

Right-truncate URLs – To go to the main page of a site, work backward and delete everything in the URL until you reach the domain.

Guess URLs – For example, typing a company or university name plus the appropriate domain can take you to the site you want.

STEP 4: Refine or change your search terms if necessary

You can further refine your search by using Boolean logic in your query. If you know how to order a pizza, you can use Boolean operators.

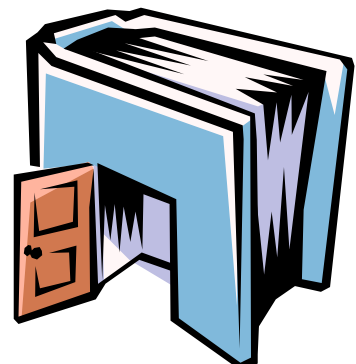
AND – think of as “only if also.” Using AND will narrow your search. Example: Find pizza with mushrooms AND olives (you want *both*). Some search engines will also let you use the plus sign (+) to represent AND, but do *not* put a space after the plus sign.

OR – think of as “either or.” Using OR will broaden your search to include items that include either term. Example: Find pizza with mushrooms OR olives (sites that include *either* word)

NOT – think of as subtracting a concept from the search. Example: Find pizza AND mushrooms NOT pepperoni (do **not** include pages that mention pepperoni). Some search engines will also let you use the minus sign (-) to represent NOT, but do *not* put a space after the minus sign. Avoid using NOT as the first entry in your search string.

You can also specify **PHRASES** within quotation marks. Example: Find pan pizza with green olives by typing: “pan pizza” AND “green olives”

Each search engine has its own interpretation of search strings. For example, if you are using google.com and type:
pizza olives
that search engine will automatically search for:
pizza AND olives.



STEP 5:

Evaluate the quality of the sites you have selected



Is the information fact or opinion? Does the site contain original information or simply links to other sites? Does the resource stand alone, or has it been abstracted from another source, perhaps losing meaning or links in the process? Are there political, ideological, or other biases?

Accuracy - How reliable and error-free is the information? Almost anyone can publish on the web, and many web resources are not verified by editors or fact checkers. Web standards to ensure accuracy are not fully developed.

Authority - What are the author's qualifications for writing on the subject? How reputable is the publisher? Often it is difficult to determine authorship of web resources. If the author's name is listed, his/her qualifications might be absent. Publisher responsibility often is not indicated.

Objectivity - Is the information presented with a minimum of bias? To what extent is the information trying to sway the opinion of the audience (the web often functions as a virtual soapbox). On the other hand, the goals and aims of persons or groups presenting material might not be clearly stated.

Timeliness - Is the content of the work up-to-date? Is the publication date clearly indicated? Dates are not always included on web pages, and if included, a date may have various meanings such as date first created, date placed on web, or date last revised.

Coverage -- What topics are included in the work? To what depth are topics explored?

Challenges Presented by Web Resources

- Web evaluation techniques are only beginning to be developed
- Technology is outpacing the ability to create standards and guidelines
- Establishing evaluation procedures will be an ongoing evolutionary process

Use of Hypertext Links - quality of web pages linked to the original web page may vary.

Search Engines Can Retrieve Pages Out of Context - When in doubt, try to return to the home page to determine the source of information.

Marketing-Oriented Web Pages - On the web, distinctions between advertising and information can become blurred. Try to determine if advertising and informational content are supplied by the same person or organization.

Instability of Web Pages - Full access may require additional software. Web pages may move or disappear without notice, and you might not be able to refer back to a page. Browsers may alter the appearance of web pages.

Susceptibility of Web Pages to Alteration - Try to verify information using other sources.

Places to Start

All about search engines: searchengineguide.com

Google

Search engine: google.com (notice the tabs: **Web, Images, Groups, Directory**)

US Government: google.com/unclesam

Zeitgeist (patterns and trends): google.com/press/zeitgeist.html

Electronic Resources: On the University of Kentucky web site (www.uky.edu), go to Libraries and click on "All Electronic Resources" to access an alphabetical list of brief, annotated guides to stand-alone and web accessible resources, including many full-text resources as well as indexes, abstract services, and data sets. Of special note: **Academic Universe** from LEXIS-NEXIS & CIS provides nearly 5,000 publications (most full-text, the rest abstracted) covering news, financial, medical, and legal information. The service covers newspapers, magazines, wire services, federal and state court opinions, federal and state statutes, federal regulations, and more. News information is updated daily and wire services several times daily.

Great Reference Links: www.nytimes.com/learning/general/navigator/index.html

The newsroom of the New York Times uses this site. It was created to give editors and reporters a solid starting point in terms of useful web sites.



Maps: mapquest.com or expedia.com

Resources include detailed maps and destination guides.

States of the Union information: 50states.com

Library sites: libraryspot.com, lii.org

The Internet Archive: archive.org

This is a public nonprofit organization founded to build an Internet library offering access for researchers, historians, and scholars to historical collections in digital format. It includes a fun feature called the Wayback Machine, which makes possible to surf pages stored in the Internet Archive's web archive.

People (includes reverse phone lookup): anywho.com/index.html

Dictionary, thesaurus, and more: dictionary.com

How things work (all kinds of information about everything, from car engines to quicksand to CDs to hypnosis and much more): howstuffworks.com