97-556 C

Point & Click:
Internet Searching Techniques

May 12, 1997

Rita Tehan
Information Research Specialist
Congressional Reference Division

Point & Click:
Internet Searching Techniques

SUMMARY

Finding information on the Internet can be challenging for even the most experienced searchers. Since the most popular means of accessing the Internet is through the World Wide Web, this report focuses on search strategies that locate Web information through the use of subject guides, search engines, Usenet news groups, e-mail discussion lists, and other resources.

CONTENTS

Point & Click:
Internet Searching Techniques

Problems with Internet Searching

Finding information on the Internet can be challenging for even the most experienced searchers. Since the most popular means of accessing the Internet is through the World Wide Web, this report focuses on search strategies that locate Web information. Some search engines index gopher[1] and FTP (file transfer protocol)[2] sites as well as Web sites and Usenet newsgroups. When the most comprehensive search is needed, it might be necessary to search gopher and FTP sites using Archie and Veronica.[3]

If a searcher enters a simple query, such as *African elephant* into any of the top World Wide Web search engines, the resulting set ranges from 30,000 hits in AltaVista, to 253,783 in Excite, to 277 in OpenText, to 218,730 in InfoSeek, and 30,217 in Lycos. A quick review of the results *does* show some relevant hits near the top of each list, but the large number of items retrieved is not productive. Since there is no centralized catalog of Internet resources, a searcher must find other ways to retrieve more precise, relevant, and useful information. This report will suggest a number of strategies, tips, and techniques to use.

The tools that are available today are going to change, and there will be new and different ones a month or a week from now, or tomorrow. Ultimately you will find a handful of useful sites by trial and error. Bookmark[4] these and return to them for future reference. Internet sites may change their URL[5] addresses slightly, but usually only to move files from one directory to another. World Wide Web sites seldom disappear

[1] See Glossary for definition, p. 9.

[2] See Glossary for definition.

[3] Archie helps find files available at File Transfer Protocol (FTP) hosts. When searching for a particular term, Archie searches the database and displays the name of each FTP host that has that file or directory and the exact path to that directory. See *Archie Services*, a gateway to Archie servers on the Web at:
http://www.nexor.co.uk/archie.html/

Veronica is an indexer that can query every gopher on the gopher system to search for a keyword or phrase in a menu title and give the address of all menus with those key words.
gopher://munin.ub2.lu.se/11/resources/veronica

[4] See Glossary for definition.

[5] See Glossary for definition.

completely. If it is a valuable resource, the organization that created the Web page has a stake in maintaining it. If the page moves, a responsible organization will provide a pointer URL to the new location.

Standards for Determining Information Quality

Almost anyone with an Internet connection can "publish" on the Web. Some criteria to consider when judging an Internet site's quality are:

Content. Is the site a provider of original content or merely a pointer site? What is the purpose of the site? Is it stated? Sites containing durable, timely, fresh, attributable information are usually more useful.

Comprehensiveness. What is the scope of the information? How deep and broad is the information coverage? If the site links to other resources, the links should be up to date and to appropriate resources.

Balance. Is the content accurate? (You may have to check other Internet or print resources.) Is it objective? If there are biases in the information, they should be noted at the site. The organization's motivation for placing the information on the Web should be clear (is it an advertisement? does it support a particular viewpoint?). An organization's Web page will provide information it wants to release and nothing more.

Currency. Is the site kept up to date? If it is a pointer site, what percentage of the links work when you click on them? Dates of updates should be stated and correspond to the information listed in the resource.

Authority. Does the resource have a reputable organization or expert behind it? Who is the author? What is the author's authority? Does the author or institution have credibility in the field? Can the author be contacted for clarification or to be informed of new information? There is nothing wrong with amateur, club, or fan sites. In fact, they may deliver more passion and enthusiasm than professional sites. The researcher must remember, however, that many amateur sites have no standards for accuracy, no fact checkers, and no peer review board.

Where to Start

The first thing to decide is what type of resource is needed. One resource is information from the World Wide Web; another is information posted to special interest e-mail lists or Usenet newsgroups. Some search engines concentrate on the Web, others focus on Usenet, and others, such as AltaVista and InfoSeek, let you search both. Many search engines scan for gopher and FTP sites as well.

If you are looking for general information on a subject, start with subject guides, which are compiled and categorized by human indexers (discussed below). These are organized hierarchically, so you can "drill down" through their links. Once you find the correct terminology for your subject, you can use search engines to locate additional information. A rule of thumb for a comprehensive search would be to check three subject indexes and three search engines.

You will retrieve more information from a search engine than a subject index, because software robots[6] visit many more sites than human indexers. However, human indexers add structure and organization to their indexes.

Subject Guides

Subject guides typically present an organized hierarchy of categories for information browsing by subject. Under each category or sub-category, links to appropriate Web pages are listed. Some sites (for example, the Argus Clearinghouse) include subject guides that function as bibliographies for Internet resources and are authored by specialists.

The lack of a controlled vocabulary within and between different subject trees increases the difficulty of browsing them effectively. Some subject guides allow keyword searching, which is useful. Examples of well-organized and comprehensive subject guides are:

- *Argus Clearinghouse* (formerly the Subject-Oriented Clearinghouse Guide to Internet Resources): http://www.clearinghouse.net/
- *Galaxy* (formerly EINet Galaxy): http://www.einet.net
- *Internet Public Library*: http://ipl.sils.umich.edu/ref/
- *Librarians' Index to the Internet*: http://sunsite.berkeley.edu/InternetIndex/
- *LookSmart*: http://www.looksmart.com/
- *World Wide Web Virtual Library*: http://www.w3.org/pub/DataSources/bySubject/Overview.html
- *Yahoo*: http://www.yahoo.com/

Search Engines: Spiders, Crawlers, Robots

Search engines are automated software robots which typically begin at a known page and follow links from it to others, downloading pages and indexing them as they go.[7]

[6] See Glossary for definition.

[7] For more information comparing the features of different search engines, spiders, robots, and crawlers, see: *Top Keyword Resources of the Web*, March 19, 1997, from December Communications at:

http://www.december.com/web/top/keyword.html

Search Engines Get Faster and Faster, But Not Always Better - September 1996 issue of *PC World* at:

http://www.pcworld.com/workstyles/online/articles/sep96/1409_engine.html

(continued...)

Search engines vary according to the size of the index, the frequency of updating the index, the search options, the speed of returning a result, the relevancy of the results, and the overall ease of use. Unfortunately, no two search engines work the same way.

To decide on which search engine to use, it helps to understand which parts of a Web page the search engines index. For example, AltaVista, InfoSeek, and OpenText index every word of a Web page, while Lycos indexes the title, heading, and the most significant 200 words. These differences contribute to the different results returned by different search engines for the same query. Search engines are not in any way comprehensive maps of the Internet. The World Wide Web is simply too vast for even the most advanced search engine to cover adequately.

All search engines don't use the same syntax. Many search engines ignore words of three or fewer letters, or will not search numbers (i.e., a date). Lycos offers relatively simple search options, but because it indexes the largest number of Internet resources, it is often successful where others fail.

Meta-search engines (MetaCrawler, SavvySearch, etc.) scan several search engines sequentially and eliminate duplicates, though not always reliably. The parallel (or meta-search) search engines are good for uncomplicated searches of very general concepts or very narrow searches of unique words or concepts, because you can't use advanced search techniques with them.

Examples of some of the most useful search engines are:[8]

- *Altavista*: http://altavista.digital.com/
- *Excite:* http://www.excite.com/
- *Hotbot:* http://www.hotbot.com/
- *Infoseek*: http://www.infoseek.com/
- *Lycos*: http://www.lycos.com/
- *Metacrawler*: http://metacrawler.com/
- *Opentext*: http://www.opentext.com/
- *SavvySearch*: http://guaraldi.cs.colostate.edu:2000/

There is no "best" search engine, and one search engine is not better than another at finding different types of documents (for example, government reports, or corporate press releases, or movie reviews). Search engines look for keywords, not concepts, so to find information on a particular topic, you need to create a precise search. That is why

[7] (...continued)
A Higher Signal-to-Noise Ratio: Effective Use of Web Search Engines, October 9, 1996, from the Wisconsin Educational Technology Conference, Green Bay, WI at:
http://www.state.wi.us/agencies/dpi/www/search.html

[8] Some sites with compilations of multiple search engines are:
All-in-One Search Page at: http://www.albany.net/allinone/
Scout Toolkit: Searching the Internet at:
http://wwwscout.cs.wisc.edu/scout/toolkit/search/index.html
WebCrawler: Database of Web Robots, Overview at:
http://info.webcrawler.com/mak/projects/robots/active/html/index.html

it is important to learn the advanced search syntax for a few different search engines, in order to refine and narrow a query when the number of items retrieved is too large.

Search Engine Features

- Most allow for phrase searching, usually by enclosing the phrase in quotation marks, for example, *"aurora borealis."*

- Most are case-insensitive, so you can enter a keyword in lower case, and the search engine will find both upper and lower case matches. Other search engines allow an exact match, which means you can retrieve words that are capitalized, such as "AIDS," or all lower case, such as "e.e. cummings."

- Most search for word variations. Some search engines support the asterisk (*) symbol (known as a wildcard) to find word variations. For example, if you enter sing*, you will retrieve pages on singers, singing, and Sing Sing.

- Most allow for advanced searching. All of the top sites use Boolean search operators to help limit the set if a large number of results is retrieved. The most important of these is "AND." When you use "AND" in a search, for example, *travel* AND *Antarctica*, the search engine will find Web pages where both those words appear. Another useful Boolean operator is "NOT" (or "AND NOT" in Alta Vista). For example, if the search is for *beetle* NOT *volkswagen*, the search engine will find information on the insect and not the automobile. Some search engines allow you to use the Boolean operator "NEAR." For example, *vaccine* NEAR *HIV*. In this case, both words will be in the document, and within a few words of each other.

Search Tips

- Read the help pages of the search engines you use regularly. These explain how to search, what is and is not covered by the database, and special syntax or retrieval rules. Take advantage of advanced searching features, such as narrowing the results by document title, date, or domain (i.e., .gov, .edu, .com, etc.)

- To increase your chance of precision searching, try to use unique or uncommon words or acronyms, especially when using a parallel search engine (such as Metacrawler or SavvySearch). If there is a synonym or less common word, this will reduce the number of items retrieved. Also remember to vary your spelling to account for differences in British or other spelling (for example colour or *labour*.)

- Think of which organizations are interested in the subject and visit those Web sites to see if they provide position papers or link to material on it. For example, if you wanted to find information on handgun control issues, check the Web pages for the National Rifle Association and the Center to Prevent Handgun Violence.

- Use specialized search engines or indexes. These focus on collecting relevant sites for a particular subject. Some examples are:

- *FindLaw* - legal resources - http://www.findlaw.com
- *Health Finder* - health resources - http://www.healthfinder.gov
- *Govbot* - government information
  http://www.business.gov/Search_Online.html

- If you don't find anything useful with one search engine, try another. There is surprisingly little overlap when using the same query in more than one Web search engine.

Some Common Problems

- *The search engine did not find a Web page you know is available*. If the page is new, it's possible the Web robot hasn't found it yet. Your search phrase or term is checked against an index of documents that the robot has scanned on a previous indexing run. While some robots search the Web continuously, others go out only once a week or every 2 or 3 weeks.[9] Some dynamic sites, by their very nature, are impossible to index correctly. News sites such as CNN or the *New York Times* are updated daily. No search engine can find very recent material.

  It's possible the desired document may be on a server that is not within a robot's scope. For example, files on a gopher or FTP server are missed by Web search tools that index only Web pages (only HTML files on Web servers).

  Information can vanish for other reasons. Webmasters move pages or entire sites without notifying search engines. Pages are deleted when customers' accounts are terminated. The job of keeping search engine indexes up to date is unrelenting.

- *The Web robot found the document but was not permitted to access it.* If the page you want is on a server protected by a firewall,[10] access will be denied. Most search engines skip sites that demand a password or registration for entrance, even those like the *New York Times*, which offer passwords free of charge. Additionally, some Web servers install software to specifically prohibit Web robots from entering.

- *The Web robot could not access the document, at least for the moment.* This problem is related to the vagaries of Internet traffic and connectivity. The Internet is most congested during the afternoon hours. If you see a

[9] *Search Engine Tutorial for Web Designers*, from Northern Webs. The "Search Engine Summary" is a chart comparing database update times for seven major search engines.
http://www.digital-cafe.com/~webmaster/set01.html#summary

[10] See Glossary for definition.

message such as "no DNS entry found," this is an indication that the host server is busy or unavailable. Frequently, an immediate attempt to reconnect will be successful.

Usenet News Groups and Email Discussion Lists

Usenet is a discussion system distributed worldwide. It consists of a set of "newsgroups" with names that are classified hierarchically by subject. "Articles" or "messages" are "posted" to these newsgroups by people on computers with the appropriate software; these articles are then broadcast to other interconnected computer systems via a wide variety of networks. Some newsgroups are "moderated;" in these newsgroups, the articles are first sent to a moderator for approval before appearing in the newsgroup.[11]

Human expertise is very accessible on the Web. A researcher can find information from other people via Usenet newsgroups, listservs, or an e-mail link on a Web page.

The Web site *Reference.com* (http://www.reference.com/) allows you to find, browse, search, and participate in more than 150,000 newsgroups, mailing lists, and Web forums.

Before posting to a Usenet group, read its *Frequently Asked Questions* (FAQ) guide. Chances are good that your question will be answered there. The FAQ is often compiled by the experts who moderate a particular newsgroup. Two good sources of Usenet FAQs are the *FAQ Archive* at:

http://www.cis.ohio-state.edu/hypertext/faq/usenet/FAQ-List.html

and *The FAQ Finder* at:

http://ps.superb.net/FAQ/

Another good practice is to read a few discussion threads before posting a question to a newsgroup. You will get a feeling for the group's style and attitudes and will reduce your chance of getting "flamed"[12] for posting an inappropriate query.

When you send a message to the Usenet group, your question may be sent out globally. People who take the time to answer are likely to feel strongly about the issue or have information that you need. Such direct personal communication is one of the Usenet's strengths. Some of its weaknesses, however, are that some Usenet groups are unmoderated, and there is no way to verify that a poster is who he/she claims to be, and that what they say is true.

If you see that a particular person frequently posts to a certain Usenet group or seems to be informative on a particular subject, you can search for the poster's name in *DejaNews* (http://www.dejanews.com/) to see what else he/she has written on that (or any other) topic.

[11] For more information on Usenet, see "What is Usenet" at the *FAQ Archive* at: http://www.cis.ohio-state.edu/hypertext/faq/usenet/usenet/what-is/part1/faq.html

[12] See Glossary for definition.

An e-mail discussion list is a computerized mailing list in which a group of people are sent messages pertaining to a particular topic. The messages can be articles, comments, or whatever is appropriate to that topic. There are more than 70,000 electronic mailing lists on every imaginable topic.

Email lists have been used for more than a decade to distribute information efficiently to research and academic communities. Scholarly lists/newsgroups are still more common than scholarly Web sites. To find listservs on various topics, check the *Publicly Accessible Mailing Lists* at:
http://www.neosoft.com/internet/paml/ or *Liszt* at: http://www.liszt.com

Gopher vs. Web

The probability of finding something current, valuable, important, and unique on a gopher diminishes as the Web becomes more popular and gophers less so. Gophers are becoming less well-maintained. However, gophers cannot be ignored because a lot of static (but still useful) information is conveyed via gopher. Most search engines also index gophers.

A catalog of many of the best Gopher sites by category is *Gopher Jewels* at:
http://galaxy.einet.net/GJ/

Miscellaneous Sources

Additional information on Internet searching is available at the Library of Congress Home Page. See *Internet Search Tools* at:
http://lcweb.loc.gov/global/search.html

Internet News

The sites listed below provide annotated evaluations of new Internet resources within a few days of their availability. A user can also subscribe to them via e-mail, if desired. Most of the sites archive their previous issues, so it is not necessary to keep copies of postings.

- CNet Digital Dispatch: http://www.cnet.com/
- Edupage: http://www.educom.edu/
- Net Happenings: http://www.mid.net/NET/
- Netsurfer Digest: http://www.netsurf.com/nsd/
- Scout Report: http://www.scout.cs.wisc.edu/scout/report/bimonth/

Glossary

Bookmark -- Using a World Wide Web browser, a bookmark is a saved link to a Web site. Like bookmarks for paper books, Web bookmarks are markers that permit you to quickly return to a Web page. Netscape and some other browsers use the term "bookmark." Microsoft's Internet Explorer uses the term "favorite."
Firewall -- A dedicated gateway machine with special security precautions on it, used to service outside network connections and dial-in lines. The firewall protects a cluster of more loosely administered machines hidden behind it from individuals attempting to gain unauthorized access.

Flame -- An electronic mail or Usenet news message intended to insult, provoke or rebuke, or the act of sending such a message.

FTP -- The FTP command allows an Internet-connected computer to contact another computer, log-on anonymously, retrieve texts, graphics, audio, or computer program files, and transfer desired files back to itself.

Gopher -- The gopher software program, developed at the University of Minnesota, organizes information into a series of menus. Using gopher is like browsing a table of contents: a user crawls through a set of "nested" menus to zero in on a specific subject.

Robot -- A program that automatically explores the World Wide Web by retrieving a document and retrieving some or all the documents that are referenced in it. This is in contrast with normal Web browsers that are operated by a human and do not automatically follow links other than graphic images and redirections (pointers to new URLs).

Search Engine -- A remotely accessible program that lets you do keyword searches for information on the Internet. There are several types of search engines; the search may cover titles of documents, URLs, headers, or the full text.

URL -- Uniform Resource Locator. It is the unique Internet address which begins with "http://" This address is used to specify a WWW server and home page. For example, the House of Representatives URL is: http://www.house.gov and the Senate URL is http://www.senate.gov