World-Wide Web indexes: a utility for marine sciences

Peter Brueggeman
Scripps Institution of Oceanography Library, University of California San Diego

In: Tradition & Innovation: planning our future. James W Markham and Andrea L Duda, eds. Proceedings of the 22nd annual conference of International Association of Aquatic and Marine Science Libraries and Information Centers held at Monterey, California, 13-18 October 1996. International Association of Aquatic and Marine Science Libraries and Information Centers, Fort Pierce, Florida, 1997. pp.123-128.

ABSTRACT: Full text Web indexes have varying degrees of utility in finding Web resources in marine science. The major full text Web indexes are reviewed and their basic search features discussed. Some sample searches illustrate these search features.

The most comprehensive Web indexes utilize a spider robot program that automatically wanders the Web by following links between Web pages. The full text of the discovered Web pages is compiled into an index database. When the user searches that Web index database, the Web index outputs the search results as the URL, Web page title, and usually some text. The more comprehensive full text Web indexes, Alta Vista, Excite, HotBot, InfoSeek and OpenText, are covered in this review. One full text Web index, WebCrawler, is very limited in the number of Web pages indexed and is not reviewed. Web indexes like Lycos index the partial text of Web pages; for comparison, Lycos is reviewed. Basic search concepts familiar from online and CD-ROM databases are presented. Look at each system's help documents for additional information on advanced search features. Search words on these Web indexes are typed into input boxes; square brackets are used in this review to represent these input boxes.

Each Web index claims its excellence and it is difficult to sift through the hyperbole. I have repeated some of the hyperbole in each review. A Web page discussing Web index metrics is at http://inktomi.berkeley.edu/counting.html. However your experience is the best guide; I advise trying Alta Vista, Excite, and HotBot. As a gross indicator, search results from October 1996 for two simple one-word searches on several full text Web indexes and one partial text Web index, Lycos, are given. Of course, speed and search features should be considered as well as results.

kuroshio: Alta Vista (600); Excite (467); HotBot (1018); Infoseek (22); Open Text (61); Lycos (148); WebCrawler (18).
oceanography: Alta Vista (50,000); Excite (33,771); HotBot (58551); Infoseek (3449); Open Text (6065); Lycos (6001); WebCrawler (1800).

ALTA VISTA: Alta Vista at http://altavista.digital.com/ has two search modes: simple and advanced. Alta Vista claims to be the largest Web index and it

indexes the full text of Web pages. In October 1996, it claimed 30 million pages on 275,600 servers, and four million articles from 14,000 Usenet news groups. Alta Vista seems very comprehensive and is in the top tier of full text Web indexes with regard to comprehensive retrieval.

ALTA VISTA SIMPLE SEARCH: An OR search is implied between words; search phrases within quote marks. Any text within Web pages are searchable including titles, URLs, image file names. Searching a word without its diacritical marks will retrieve its usage with and without diacritical marks. [kuroshio current] is an OR search retrieving Web pages containing either of these words ranked so that pages with the most matches are presented first. Thus pages with either "kuroshio" or "current" are retrieved; that's too unfocused. Do a phrase search using quotes ["kuroshio current"]. Punctuation within phrases is indexed as spaces e.g. "http://scilib.ucsd.edu/sio/" is parsed as five words ignoring truncation. Search synonym phrases using quotes and with the implied OR ["enhydra lutris" "sea otter"]. Typing letters in lower case will find upper or lower case words; typing letters in upper case will force searching on those upper case letters. ["el nino"] retrieves "el nino" or "El Nino" or other variants whereas ["El Nino"] retrieves only "El Nino". Word truncation or wildcard is with asterisk (*) e.g. pollut* for pollutant, pollutants, pollution, etc. Internal truncation is allowed e.g. ["ocean colo*r"] retrieves "ocean color" and "ocean colour". Use a plus sign (+) or minus sign (-) to require or prohibit words or phrases in a search. To search for Southern California Bight and exclude a subset of CalCOFI reports that one sees, use ["southern california bight" -"california cooperative oceanic fisheries"]. To search for sea otters in Monterey, search ["sea otter*" "enhydra lutris" +monterey]. To search for "El Nino" and exclude Web pages already seen at the Pacific Marine Environmental Laboratory's extensive El Nino site, search ["el nino" -url:http://www.pmel.noaa.gov/]. Since institutions may have more than one Web server, this prohibits retrieval of Web pages mentioning only that one PMEL Web server. If PMEL has several Web servers with El Nino pages, the non-specified Web servers would be included in the search results. An alternative is to use "host:" instead of "url:". To see everything mentioning your Web site and exclude your own institutional pages, use ["scripps institution of oceanography library" -host:ucsd.edu].

ALTA VISTA ADVANCED SEARCH: Advanced searching adds on several more features and the searcher is given a larger input box in which to type it all. Operators AND, OR, NEAR, and AND NOT are used to combine words and phrases. The NEAR operator is used to specify that both words are within ten words of each other in the Web pages. Parentheses are used to group search searches and operators. Examples: ["ozone hole" and plankton]; ["weddell sea] near penguin*]; [("enhydra lutris" or "sea otter*") and (monterey or "big sur" or carmel)]

EXCITE: Excite at http://www.excite.com/ has one search mode. Excite indexes the full text of Web pages. In July 1996, Excite claimed to have the largest database: 50 million fully indexed URLs; it claimed to index more Web pages

than Infoseek, Alta Vista, and Lycos. It seems to be almost as comprehensive as Alta Vista and is in the top tier of full text Web indexes with regard to comprehensive retrieval.

EXCITE SEARCH: An OR search is implied between words. Phrases can be searched within quote marks but the results are fuzzy. Word truncation is not available. Excite uses a plus sign (+) or minus sign (-) to require or prohibit words in a search. Excite also uses the operators AND, OR, and AND NOT and allows nesting of words within parentheses. The operators AND, OR, and AND NOT must be in ALL CAPITALS to be recognized as such e.g. [kuroshio AND current]. If you capitalize the first letter of several words, Excite assumes you are looking for a proper name, e.g. [Peter Brueggeman] or [Intergovernmental Oceanographic Commission]. Excite will find pages where these words are NEXT to each other as a proper name. When put into a search with operators, proper names need to be set off with quotes or parentheses, e.g. ["Intergovernmental Oceanographic Commission" AND gemim]

HOTBOT: HotBot at http://www.hotbot.com/ has two search modes: regular and expert. HotBot indexes the full text of Web pages. In October 1996, HotBot claimed that it indexed 54 million Web pages; this would make it number one. HotBot seems the most comprehensive and is in the top tier of full text Web indexes with regard to comprehensive retrieval.

HOTBOT REGULAR SEARCH: Through a pull-down menu, the user can specify conditions on the words being searched: all words (AND operator), any words (OR operator), phrase, person, or URL. Phrases can also be enclosed within quote marks ("). The default for the pull-down menu is a search for all words, an implied AND search. Words with mixed upper and lower case letters are searched as a case-sensitive search. Word truncation is not available. Operators or parentheses for nesting words are not available by typing operators and parentheses; instead, these tools are utilized via the pull-down menu and the Modify/Revise search option. The text in the initial input box can be modified two more times by text in two additional input boxes. This can allow for nesting of words among three search concepts. Modify operators are MUST, SHOULD, and MUST NOT. MUST is AND and MUST NOT is AND NOT. SHOULD is a relevance ranking tool like those available on some other Web indexes. Text being modified can be words, phrase, person, or URL. A person search uses a limited proximity search. A person search on [Paul Dayton] will find Web pages with Paul Dayton and Dayton, Paul. [Paul K. Dayton] to find Paul K Dayton, Paul Dayton, and Dayton, Paul. Searches cannot be limited on the middle initial.

HOTBOT EXPERT SEARCH: In expert search, additional search criteria can be added to the regular search via radio buttons and pull down menus. Users can specify before and after dates for Web pages, media type (e.g. Java, Acrobat, Shockwave, Audio, Image), and location by domain name or geographical regions.

INFOSEEK: Infoseek at http://guide.Infoseek.com/ has one search mode.

Infoseek indexes the full text of Web pages. In April 1996, Infoseek claimed to be faster than other full text Web indexes and more current. Infoseek seems to be in a third tier of full text Web indexes with regard to comprehensive retrieval.

INFOSEEK SEARCH: An AND search is implied between words. Phrases can be searched within quote marks ("). Word truncation is automatic and there are no operators. Infoseek uses a plus sign (+) or minus sign (-) to require or prohibit words in a search. Use a hyphen between words to specify that one word is within one word of the other e.g. [santa-island] retrieves "santa catalina island", "santa cruz island", etc. Another word proximity option is to use square brackets to find words that appear within 100 words of each other, e.g. [[California ocean]]. If you capitalize the first letter of several words, Infoseek assumes you are looking for a proper name, e.g. [Peter Brueggeman] or [Intergovernmental Oceanographic Commission]. Infoseek will find pages where these words are NEXT to each other as a proper name. Several proper names can be searched by placing commas between them, e.g. [San Clemente Island, Santa Cruz Island].

OPEN TEXT: Open Text at http://index.opentext.net/ has two search modes: simple and power. Open Text is a full text Web index. In their undated FAQ, Open Text claimed to index about the same number of Web pages as Lycos; since Open Text does full text indexing, it will find more Web pages. Open Text seems to be in a second tier of full text Web indexes with regard to comprehensive retrieval.

OPEN TEXT SIMPLE SEARCH: An AND search is implied when searching several words. Through a pull-down menu, the user can specify searching for words or a phrase. There are no OR or NOT operators, no word truncation, and no nesting of search words within parentheses.

OPEN TEXT POWER SEARCH: Word searching is exact; Open Text does not automatically truncate words. Word truncation or wildcard is not available. You can specify where you want the text searched: anywhere, Open Text's summary for a page, title, first heading, URL. The default operator is AND for multiple words. Operators are AND, OR, BUT NOT, NEAR, or FOLLOWED BY and are applied in the order they appear. The first three operators are obvious. Word proximity has two options. NEAR finds all Web pages in which the word or phrase entered in the second input box occurs within 80 characters either before or after the word or phrase entered in the first input box. FOLLOWED BY finds all Web pages in which the word or phrase entered in the second input box follows within 80 characters of the entry in the first input box. Search words are nested using multiple input boxes. No parentheses can be typed for nesting words and operators.

LYCOS: Lycos at http://www.lycos.com/ has two search modes: regular and custom. Lycos doesn't index the full text of Web pages. In October 1996, Lycos claimed over 66 million URLs indexed; however a URL can be a Web page

and also a Web page's inline and external images. Alta Vista claims the most Web pages (though it appears HotBot is now ahead); Lycos claims the most URLs. Lycos claims that its spidering technology allows it now to catalog more URLs than Alta Vista. Lycos indexes key areas like titles, headers, and links and a certain amount of text but not the full text. Though Lycos is very fast, you will find much more with a full text Web index like Alta Vista or Excite. Though indexing only partial text, Lycos seems to be in a second tier of Web indexes with regard to comprehensive retrieval

LYCOS REGULAR SEARCH: Lycos has automatic truncation though it specifies to use a dollar sign ($) for truncation. Lycos uses a period (.) after a keyword to force an exact match and stop automatic truncation. Enter "ocean." to find ocean, but not oceanic, oceanographic, etc. Searching several words e.g. [sea otter] is an implied OR search retrieving Web pages containing either of these words ranked so that pages with the most matches are presented first. Thus pages with either "sea" or "otter?" are retrieved; that's too unfocused. Phrase searching and nesting of words and operators within parentheses is not available. The AND operator is initially unavailable. At the bottom of a search results page, one can edit and rerun the search so that Lycos will match all the words -- an AND search. Therefore an AND search is allowed only as a second step. Lycos uses the minus (-) symbol in front of a word to decrease its relevancy in the search but not eliminate its appearance. This is not an adequate substitute for NOT; enter "otter -sea" to find "otter" with the word "sea" being ranked lower in relevance but not excluded from the search results.

LYCOS CUSTOM SEARCH: In Lycos' custom search, the words typed in the input box can be specified so that all words are present (an AND search), any word is present (an OR search), or two through seven words are present. If you type in three words and specify that two words be present, you cannot specify which two words. You can also specify the relevance used for search retrieval by specifying loose-fair-close-good-strong matches. Truncation is automatic and can be forced off by using a period at the end of a word. There is no phrase searching or nesting of words and operators within parentheses.

REFERENCES

Web indexers and their search engines are evolving rapidly. A reviewer's comments are relevant at the time of the review and may not be as relevant today since more sites may be covered and search engine speed improved. Following are a selection of the many reviews of Web indexes.

Web-based reviews and some sites linking or listing Web reviews:
http://www.lib.berkeley.edu/Web4Lib/archive/9604/0103.html;
http://www.hamline.edu/library/links/comparisons.html;
http://www.dis.strath.ac.uk/business/search.html;
http://www.library.ucsb.edu/untangle/eagan.html;
http://www.library.ucsb.edu/untangle/lager.html;

http://neal.ctstateu.edu:2001/htdocs/websearch.html;
http://www.unn.ac.uk/features.htm;
http://www.state.wi.us/agencies/dpi/www/search.html;
http://www.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/Searching_the_Web/Comparing_Search_Engines/

Print reviews: Annotations rate Web indexes "useful", "good", and "great" and exclude directory services like Yahoo, Point, Magellan, etc.

Clyman, J. 1996. "Finding your needle in the Web's haystack" PC Magazine 15(13):39-44. Good review. Reviews Alta Vista (most comprehensive and powerful), Excite (found fewer sites than other indexes, quirky), Infoseek (less comprehensive than Alta Vista), Lycos (less comprehensive than Alta Vista, search feature limitations), and WebCrawler (not comprehensive).

Conte, R. 1996. "Searching on the Web, Guiding lights" Internet World 7(5):41-44. Useful review. Reviews briefly Alta Vista, Excite, Lycos, Infoseek, and Open Text.

Courtois, M.P. 1996. "Cool tools for Web search, an update" Online 20(3):29-36. Great review. Reviews Excite, Inktomi (HotBot's underlying technology), and Alta Vista (the largest). Updates previous review of Infoseek, Lycos, Open Text, and WebCrawler.

Morris, J. 1996. "Refine your serach" PC Magazine 15(18):48. Good review of Excite and Hotbot.

Munson, K.I. 1996. "World Wide Web indexes and hierarchical lists: finding tools for the Internet" Computers in Libraries 16(6):54-57. Useful review but doesn't cover enough indexes. Reviews Lycos and Open Text.

Notess, G.R. 1996. "Searching the Web with Alta Vista" Database 19(3):86-88. Good review of one of the most comprehensive Web index.

Prosise, J. 1996. "Researching the Web" PC Magazine 15(11):235,238. Good review. Reviews Alta Vista (most comprehensive), Excite (less comprehensive than Lycos or Alta Vista, slower than Lycos and Alta Vista), Lycos (speedy, thorough), and WebCrawler (not comprehensive).

Tomaiuolo, N.G. & Packer, J.G. 1996. "An analysis of Internet search engines: assessment of over 200 search queries" Computers in Libraries 16(6):58-62. Good review; doesn't cover Open Text or Excite. Reviews Alta Vista (performed well), Lycos (retrieved duplicates), and Infoseek (performed well).

Venditto, G. 1996. "Search engine showdown" Internet World 7(5):79-86. Great detailed review of the major indexes. Reviews Alta Vista (speedy, most comprehensive), Excite (not comprehensive, equal to Infoseek and Open Text in quality and quantity of search results), Infoseek (nice search features but

not comprehensive), Lycos (since it indexes partial text, it cannot find all occurrences of text buried within Web pages), Open Text (best designed, great search features, less comprehensive than Alta Vista), WebCrawler (fast, not comprehensive), WWWorm (anachronism).

Zorn, P. et al. 1996. "Advanced Web searching, tricks of the trade" Online 20(3):15-28. Great review. Reviews Alta Vista (most comprehensive, highly relevant and accurate retrieval), Infoseek (highly relevant and accurate retrieval), Lycos (most comprehensive), and Open Text (highly relevant and accurate retrieval, most advanced search features).