A Proposal for Categorization and Nomenclature for Web Search Tools

Nicholson, S. (2000). A proposal for categorization and nomenclature for Web search tools. Internet Searching and Indexing: The Subject Approach, ed. Thomas, A. and Shearer, J. 9-28. (also published in Journal of Internet Cataloging, 2(3/4), 9-28, 2000).

A Proposal for Categorization and Nomenclature for Web Search Tools

Scott Nicholson, Ph.D. Syracuse University School of Information Studies.

Abstract:

Ambiguities in Web search tool (more commonly known as "search engine") terminology are problematic when conductin precise, replicable research or when teaching others to use search tools. Standardized terminology would enable Web searchers to be aware of subtle differences between Web search tools and the implications of these for searching. A categorization and nomenclature for standardized classifications of different aspects of Web search tools is proposed, and advantages and disadvantages of using tools in each category are discussed.

Keywords: Web Search Tools, Search Engines, Web searching, Search tool comparisons, Web indexing, Web searching guides, Web searching tutorials

Introduction and Inspiration

Search engines, search tools, robot-generated databases, Web search services, and Web keyword indexes are all names given to tools that allow searching for and retrieval of World Wide Web sites [1, 2, 3, 4, 5]. As any field develops, the terminology becomes more and more precise. In current Web search tool research, however, there is much imprecision. Not only can this lead to confusion on the part of the reader of this research, but assumptions based on this imprecision can cause researchers to conduct biased or inconsistent studies.

For example, in a recent Science article, Lawrence and Giles discuss a group of "full-text search engines"[6]. They go on to state that "the Web can be viewed as a searchable 15-billion-word encyclopedia"[7]. What problems would it cause if some of the pages of this encyclopedia had every word listed in the index while other pages had only a few selected words listed? In order to successfully find a page, one would have to know what type of indexing was used for it. Several of the "full-text" search tools (e.g. Lycos and HotBot) examined in this research[6] actually do not index the full text of each page. Proper categorization of the search tools would allow researchers to select groups of similar tools for study.

Clark and Willett[8] ran the same searches on three different search tools: Alta Vista, Excite, and Lycos. They reported that there were "significant differences between Alta Vista and Excite, and between Alta Vista and Lycos, but not between Excite and Lycos"[9] and then pointed out the "high level of performance of the Alta Vista system"[9]. At the time of that research, Excite and Lycos were in one category of search tool (extract search tools), and Alta Vista was in a different category (full-text search tools). However, the majority of their queries involved phrases which only the full-text search tools could handle properly. Their research design therefore favors any full-text search tool, such as Alta Vista. Clearly, without precise categorization of Web search tools, biases can be unknowingly introduced into the research.

Standardized terminology and categorization also can assist in teaching others how to use search tools effectively. Most books on Internet searching teach how to use Alta Vista, Excite, and other specific tools. This method is like teaching users how to use the Oxford English Dictionary, then showing them how to use the Merriam-Webster Dictionary, without telling them what is assumed when one uses that particular dictionary. Categorization allows trainers to teach about a class of search tools and use examples from that group during the lesson, instead of just teaching about individual tools. More importantly, categorization also allows users to understand new or changed tools and modify their searching techniques to fit the tool in question.

Sans "Search Engine"

The first problem is what to call the whole group of tools used for Web searching. In the popular culture, "search engine" is the term generally used to describe these tools. However, "search engine" means something different to most information retrieval researchers as has traditionally referred to the programs that do the actual matching of query terms to a database in an IR system. Therefore, many are uncomfortable applying "search engine" to Web directories like Yahoo and other search tools. Using this term can result in unclear communication, and it is suggested that the term "search engine" be avoided in the discussion of these tools.

Instead, to describe the comprehensive class of tools used to search the World Wide Web, the author proposes the use of Web search tools or search tools for short. This is clear, recognizable, and unambiguous; it also allows the inclusion of tools not yet invented that may not use any type of search engine. Most search tools are made up of several components; these may include the Web robot, a Web page database, a search interface, and Web page surrogates.

Proposed Classifcations for Aspects of Web Search Tools

Web Robots

The Web robot is a program that traverses the World Wide Web, gathering candidates for pages to be indexed in the search tool. Some Web robots visit a page and all linked pages, some visit pages submitted by Web page authors, and others start with "What's New" or "What's Cool" Web pages. These programs are also called spiders, crawlers, and harvesters, among other things[2], but in order to aid in successful, standardized communication, Web robot is the proposed terminology. Some search tools do not use Web robots; a human finds the pages for inclusion in the database, so the candidate pages will be human-gathered as compared to robot-gathered. A few tools provide access to pages both robot-gathered and human-gathered.

The method of gathering affects the quality of pages available in the database. If humans have gathered the pages, there can be a level of quality control that does not exist in the robot-gathered Web search tools. However, the robot-gathered Web search tools may be more methodological in gathering all of the pages in a Web site (if programmed that way) and will gather more pages overall than a human. Therefore, human-gathered Web search tools can provide higher-quality Web sites for searching, and robot-gathered Web search tools will most likely provide a larger quantity of Web sites to search.

Web Page Databases

Information retrieved by the Web robot is then put into a Web page database (the word "page" is used in order to distinguish this from databases of material that are Web-accessible). Most Web page databases consist of indexes and related surrogates. Knowing certain aspects of the Web page database index is the key to selection and proper use of the appropriate search tool for a specific search request[10]; however, most articles on Web search tools in the popular media compare only the search interfaces.

Topicality of Web Page Database

Some databases contain records only on a specific subject such as health or business. These come in many different forms but will all be called subject-specificsearch tools, while tools that collect data on a variety of topics will be called generalsearch tools. Most of the commonly discussed search tools are general search tools (such as Alta Vista, Excite, and Hotbot); however, the rewards of finding and using a subject-specific search tool will be discussed later in this article.

Method of Indexing Used

One aspect of Web page databases is the method used to index retrieved Web pages. The information collected by a Web robot may be presented to a person for review, or the human that gathers the page may also index it. In either case it will be termed a manually indexed Web page database. These databases may consist of keywords and/or controlled vocabulary selected from the Web page and short abstracts or reviews of the page. If the Web robot works with another computer program to process the information without human intervention, it is an automatically-indexed Web page database.

In automatically-indexed Web page databases, there is an important distinction to make. Some Web page databases contain every word from every page indexed, and are called full-text search tools. Other tools index every word on the page except for a list of stopwords, and these are called full-text without stopwords search tools. A third class of tools contain Web page databases made up of automatically selected extracts (the first n words, the most frequently used words, or words from the title or headers) from each Web page. These are called extracting search tools. The distinction between indexing methods is very important. Knowing the form of indexing used helps the searcher determine what topics will be easily searchable and what types of search strategies may be effective[10].

Full-text Search Tools (Alta Vista, Infoseek)

These tools attempt to be the most comprehensive Web search tools by indexing the full text of every page included in the Web page database. Because of the full-text indexing, these search tools can be challenging to use for the novice searcher; they tend to overwhelm the user with results. The relevance ranking algorithm used in these tools can bring some useful pages to the top of the list; however, unscrupulous Web designers search for ways to design their pages so that they rise to the top, regardless of the usefulness of their content.

While a user might imagine that the full-text search tools provide a comprehensive listing of Web pages, this is not the case. Not only is each one linked to a different Web page database, each indexes a different part of the Web. In fact, even the largest tools index no more than (and probably considerably less than) 35% of the Web[6]. Therefore, if a high-recall, comprehensive search is desired, a user must search using all tools available and be prepared to follow links from pages to discover pages not indexed in any Web searching tool.

Full-text without Stopwords Search Tools (Excite, Hotbot)

These tools index every word on every page in their database with the exception of stopwords, such as and, or, not, the, and other common words. To the unsuspecting searcher, these tools act like the full-text search tools. However, if a searcher tries to find a phrase containing stopwords, such as "To be or not to be", they will have difficulties. For this reason, it is important to distinguish between these two types of full-text search tool.

Extracting Search Tools (Lycos)

These tools will usually provide fewer pages on a topic than either type of the full-text search tools. As these tools do not index every word on the page, many of the techniques used to limit the output described above will not work. For example, the lack of phrase searching with these tools can make it difficult to achieve a high level of precision. However, if the search topic is general, if there are not any applicable phrases that might appear on all the target pages, or if the search terms tend to be used out-of-context, these tools may provide better results than a full-text tool.

When reading older research, it is useful to know that Excite used to be an extracting search tool, and now is a full-text without stopwords search tool. However, this change was not announced; it just happened one day. Thus, as a researcher, it is worth checking the help screens on search tools every so often to catch significant changes like this.

Search Interfaces

The search interface is the method that allows the user to extract information from the Web page database. Although most Web page databases have only one interface attached to them, a few search tools allow different methods of accessing the same database. The two types of interfaces currently available are query boxes and subject trees.

Query Box

A query box accepts free-text from the user and extracts records from the Web page database via a matching and ranking algorithm. Query boxes can link into automatically-indexed or manually-indexed Web page databases. A user will type insearch terms (words representing the different facets of the search) and can useoperators. The operators available may be Boolean connectors, term weighting, grouping, or some of each. Boolean operators may include AND, OR, NOT, AND NOT, and truncation. One caveat with Boolean is that these operators do not always mean what they have meant in traditional types of searching. The AND, for example, may be a "fuzzy AND," where first all items containing both search terms are returned, and then items containing one or the other are returned. The only place to learn the meanings of the Boolean terms used is the help screens attached to the search tool.

Term weighting operators are symbols like + or - entered directly before a search term. The plus sign usually means that items must have that search term to be returned (i.e., very high term weight), while the minus sign means that items with that search term will not be returned or be returned at the bottom of the list and thus have a very low term weight.

Grouping operators are symbols like "". The quotes are used for phrase searching and are the most powerful tool in the searcher's arsenal. Apart from obvious cases, such as proper nouns or lines from poems or songs, phrases can be used to find factual answers easily when a full-text index is accessed through the interface. In the future, search tools may offer other grouping modifiers like those found in DIALOG (NEAR, WITH).

Another issue to consider is the default operator. When a searcher does not enter any operator, the search tool uses some type of default. This is important to consider in Web search tool research and controlling variables. Some search tools default to AND, some default to OR, some default to a "fuzzy AND" (where they do an AND and then an OR), and some just have the defaults built into their relevance ranking algorithms.

When entering terms without using any connectors in many search tools, the user can gain more precise results by entering more terms. However, if more synonyms are entered for one search facet than another facet, pages matching more of the synonyms for that facet may be placed higher on the list. This may produce a desired effect or may create a failed search, depending upon the situation. Thus, the user needs to be aware of this quirk/feature.

Subject Tree

A subject tree allows the user to select from a hierarchical menu of categories. These lists are almost always created by humans, and thus have all of the problems with human indexing that have been faced in libraries for years. These interfaces are usually linked into manually-indexed Web page databases.

Given that these search tools are usually smaller and manually indexed, there is a level of quality control that does not exist in the automatically-indexed search tools. These are the only search tools good for browsing, as the subject tree interface allows users to look around the Web without a specific topic in mind. The subject tree allows users to wander from topic to topic, finding things they would never have known to look for. It is the closest there is on the Internet to simulating the "shelf browsing" experience in a library.

If the topic is general and there is a heading for that topic, a subject tree is the best place to start research. One can begin to see the type of information existing on the Web for that topic, and may be able learn how pages in that field are organized. Compared to results from a general topic search in a automatically-indexed tool, the resulting list of Web sites will be short (perhaps too short); however, users generally will not feel inundated with hits.

Even if there is not a heading for the general topic in question, a tool based on a subject tree (such as Yahoo) may be used to find a few pages on the topic if the tool has a query box. Many times, once a user finds one good page on a topic, links on that page can be followed for further information on that topic. The chances of finding a relevant starting point with a subject tree are better because of the human intervention.

It is important to specify the type of search interfaces and modifiers available with the search tool so that a search can be planned. For example, a user who to find a song lyric will have much better luck using a full-text search tool that has a query box that allows grouping. If the user wants to be able to jump to a short listing of pages on a general topic, then the user needs a Web search tool with both a subject tree and a query box. In Web tool research, it is essential to acknowledge what type of search interface and modifiers are used in order for the study to be replicable.

After accepting the search, the search interface uses a matching and ranking algorithm which will find pages that match the user's search terms under the constraints of the connectors and will produce a list of references to Web pages. This list, called aresult set, will usually be ordered by a descending relevance score, which rates the similarity between the search and the Web page assigned to each Web page database entry by the algorithm. These algorithms are guarded by the companies, as unscrupulous site designers will attempt to figure out how best to manipulate their page to end up near the top of the list.

Because the matching and ranking algorithms are held secret, Web search tool research usually requires a black box approach. Queries are dropped into the black box, and results come out the other side. Other aspects of the Web search tools can be learned and thus controlled. While research that compares Web search tools can control for many aspects of the search; however, the effectiveness of the matching and ranking algorithms will always be a point of variance.

Web Page Surrogates

Just as a card in the card catalog represents a book in the library, the Web page surrogate is a representation of a Web page. Web page surrogates are presented to the user in response to a query, and may or may not be the same as the Web page database entry used for matching a query. This can be frustrating for the user who does not know this, as they can not determine why a page was returned from the Web page surrogate.

These surrogates are made up of a citation and an abstract. The citation usually is the relevance score, the title of the page, the URL, and other information about the page. One important piece of information in many surrogate citations is the refresh date. This date is the last time the Web search tool visited the Web page. These, along with dead links, can be an indicator of a Web search tool that is not kept up-to-date.

There are different types of abstracts available. If the Web page database is full-text, the abstract may be a context abstract, which presents the parts of the Web page with the words in context, or a standard abstract, which will usually consist of the first few lines of the Web page. However, if the Web page databases are indexed only by keywords, then there will be only a standard abstract available. A brief abstract contains the title and URL of the page, and possibly a few words about the content of the page (created automatically or by humans).

Here are examples of each type of abstract for the same Web page:

Brief abstract (from www.go.com's Topics):

AskScott: Your guide to finding it on the Internet
Virtual reference librarian recommends search tools for specific types of searches.
http://www.askscott.com/

Standard Abstract (from www.altavista.com):

AskScott - Your guide to finding it on the Internet.

Need Help? I'm Scott, the Virtual Reference Librarian. As you are answering questions on the right, look at this column to get advice, commentary, or see..
URL: askscott.com/
Last modified 20-May-99 - page size 4K - in English [Translate]

Context Abstract (from www.google.com):

AskScott - Your guide to finding it on the Internet.
...other options: * Orientation to **AskScott** * Searching...
...stocks > Something else What is **AskScott**? Just as the reference...
www.askscott.com/ Cached (4k)

In some search tools, the user can choose the type of information provided in the Web page surrogate. If this is possible, it is important to note what types of surrogates are examined in a Web search tool study. If the study uses selection of pages from a result set, the type of abstract presented may change the selections made by the subject. If relevance judgments are going to be made directly from the search results screen, then the type of surrogate examined will have great effect on the study. Being aware of the different types of surrogates allows the researcher to control for bias that might be introduced by comparing different types of surrogates.

Other Parts

Most Web search tools have all or most of these four parts. However, there are some parts not listed here which exist only in a few tools. One of these is Excite's "More Like This" feature; it allows a user to re-define a search based on characteristics (chosen by Excite) of a seed document. A similar tool is Lycos's WiseWire addition, which watches a search and makes adjustments based on searcher feedback. These feedbackmechanisms, if developed and properly advertised, may become a popular and useful feature. It is important to recognize if these are being used in a Web tool study, as they can bias the results if they are used but not reported.

Meta-Search tools (Inference Find, Dogpile, Metacrawler)

The most problematic yet very popular tool is the meta-search tool. These search tools have a search interface, which accepts a query from a user, may or may not translate it, and submits it to several other search tools. The returned surrogates are collected and organized for display. The problem in defining meta-search tools is that they combine the results from searching different types of Web page databases through different interfaces, and the surrogates will come from a variety of ranking and matching algorithms. Because of this chaotic variety, it is difficult to perform precise, controlled searches with these tools.

Summary of Classification

In summary, here are the questions to ask when examining a Web search tool:

- Are the pages gathered from the Web by robots, humans, or both?
- Does the Web page database contain pages on a specific subject or is it more general?
- Are the entries in the Web page database automatically indexed or manually indexed?
- Is the full text or only portions from each page indexed for the Web page database? Is there a list of stopwords that the index does not include?
- Does the Web search tool have a subject tree or query box interface?
    - If it has a query box interface, what operators does it allow?
- Are the returned abstracts based on the context in which the terms appear, or are they standardized or brief abstracts?
- Is it a Meta-search tool?

Implications for Classification

Once the aspects of a Web search tool have been classified, the searcher can have a better idea of what type of search will be successful there. Conversely, if the searcher has a query in mind, s/he can select the most appropriate search tool for that query. This is the concept behind AskScott (http://www.askscott.com), where the user is asked questions about a query and directed to the most appropriate search tool.

Categories for Web Search Tools

Five categories are suggested based upon the above terminology, although new categories can be created as needed. Each of these categories involves one or more assumptions about aspects of search tools. When an assumption is not met, then the area of difference should be specified when discussing the search tool. For example, a directory-based search tool that is on a specific topic would be referred to as a subject-specific directory-based search tool, but a directory-based search tool that is general would need no additional qualifiers.

Directory-based Search tools (Yahoo, Looksmart, About.Com)

These are tools based on a subject tree search interface. The pages may be gathered by Web robots or humans, but the pages are indexed by humans. The surrogates, usually brief abstracts, are organized in a subject tree, although some tools have a query box as well. When there is a query box and a subject tree that search the same directory, any controls placed upon use need to be stated in Web tool research.

Full-text Search tools (Alta Vista, Excite, Infoseek)

Search tools that automatically index every word on every Web page are called full-text search tools. These have pages gathered by robots, although users can suggest pages for the robots to visit. The search interface is a query box that allows a combination of Boolean and term weighting connectors, including the phrase connector. The abstracts are usually standardized, although the user may have a choice of abstract format.

Also included in this category are the full-text without stopwords search tools. In most respects, they can be used in the same way as the full-text search tools. However, when phrase searching, it is important to remember that full-text without stopwords search tools will produce poor results if stopwords are in the phrase used for searching.

Extracting Search tools (Lycos, Webcrawler)

In comparison to the full-text search tools, extracting search tools automatically index only parts of the Web pages. In some cases, they index only the title, headings and first 100 words. In other cases, they index the most commonly used words. These tools tend to be targets for dishonest Web-page designers who learn their indexing methods and exploit them. Other than the difference in indexing, these tools created are just like full-text search tools. Some even allow limited phrase searching, although it will not be as useful in an extracting search tool as it is with a full-text search tool as any phrase containing a term not indexed from the Web page will fail to find that page.

Subject-Specific Search tools (Achoo, Mutual Funds Online, HomeArts Network)

The only thing these tools have in common is that the Web page database contains records on a given topic. Each one of these tools is different, and very few of them offer the variety of search options that the larger general tools offer. However, if there is a subject-specific search tool available, searching on the subject will be much less frustrating and faster than searching with a general search tool. It can be difficult to locate these subject-specific search tools but there are resources such as http://www.search.com, http://www.ipl.org/ref/, or http://www.clearinghouse.net that are lists of these tools.

Meta-Search tools (Inference Find, Dogpile, Metacrawler)

The most problematic yet very popular tool is the meta-search tool. These search tools have a search interface, which accepts a query from a user, may or may not translate it, and submits it to several other search tools. The returned surrogates are collected and organized for display. The problem in defining meta-search tools is that they combine the results from searching different types of Web page databases through different interfaces, and the surrogates will come from a variety of ranking and matching algorithms. Because of this chaotic variety, it is difficult to perform precise, controlled searches with these tools.

Table 1. Categories of Web Search Tools

Type of search tool	**Assumptions made**
Directory-based Search tool	Manually indexed Web page database General Web page database Subject tree search interface Brief abstracts
Full-text Search tool	Robot-gathered Web page database General Web page database Automatically-indexed Web page database Full-text index (with or without stopwords) Phrase searching allowed Query box search interface Standard Abstracts

Extract Search tool	Robot-gathered Web page database General Web page database Automatically-indexed Web page database No phrase searching Keyword index Query box search interface Standard Abstracts
Subject-specific Search tool	All pages in the Web page database are related to a specific subject Human-gathered Web page database
Meta-Search tool	Searches other Web search tools and combines the results General Web page database Query box search interface

-->

Advantages and Disadvantages of Tools from Different Categories

Different categories of Web search tools are useful for different tasks. Just as one would not use an encyclopedia to look up the definition of a word, one would not use a directory search tool to find the source of a quotation. By knowing the advantages and disadvantages of each category of search tool, the user can select the best search tool for a particular query. As new search tools become available, users can also categorize them and predict what type of searches will be successful.

Directory-based Search Tools

Given that directory-based search tools are usually smaller and manually indexed, there is a level of quality control that does not exist in the automatically-indexed search tools. These are the only search tools good for browsing, as the subject tree interface allows users to look around the Web without having a specific topic in mind. The directory-based search tools allow users to wander from topic to topic, finding things they wold not otherwise have known to look for. It is the closest there is on the Internet to simulating the "shelf browsing" experience in a library.

If the topic is general and there is a directory listing for that topic, a directory-based search tool is the best place to start research. One can begin to see the type of information existing on the Web for that topic, and may be able learn how pages in that field are organized. Compared to results from a general topic search in other categories of search tools, the resulting list of Web sites will be short (perhaps too short); however, users generally will not feel inundated with too many hits.

Even if there is not a heading for the general topic in question, a directory-based tool may be used to find a few pages on the topic if the tool has a query box. Many times, once a user finds one good page on a topic, links on that page can be followed for further information on that topic. The chances of finding a relevant starting point with a directory-based search tool are better because of the human intervention.

The main disadvantage to a directory-based search tool is its size. For example, as of this writing, Yahoo has about 1 million entries while Alta Vista has over 140 million[11,12]. They are many times smaller than search tools created automatically, and thus if they do not adequately answer the user's questions, the user will have to either use another search tool or find some pages with links to other pages.

Full-text Search Tools

The key to using a full-text search tool is the phrase connector. Usually indicated by quotation marks, a phrase search helps users narrow their results and raise their searching precision. For example, the phrases allow users to find all instances of a single form of a person's name or a company in the indexed Web pages. If the person or company is well-known, a directory-based search tool may give more accurate results. But for most identity searches, the full-text search tools will give good results. If there is more than one entity with the same name, placing the term weighting connector to require a term (usually "+") in front of the name and adding a place or topic related to the person can help bring the desired results to the top.

Quotations, song lyrics, and poetry can also be found with the phrase connector. For example, only tools that index every word on the page (including common stop words like the, be, to, or, and, not, etc.) can find the phrase "To be or not to be." As with most Web searching, careful attention should be paid to the authority of the source. Many people, for example, will attach a favorite line from a poem or quotation to a Web page. While personal home pages may provide some information about a quotation, printed resources may need to be consulted if an authoritative answer is needed.

Specific topics, acronyms, and any unique phrases will aid in the precision of a search with the full-text search tool. If there is some specific item that appears somewhere in the body of the Web page, it can be included in quotes when searching. One example of this is in answering factual questions. If the user can conceive of a phrase that would exist on the perfect Web page that would contain an answer to the question, then the user can enter that phrase in quotes. For example, if the question is, "What is the capital of Zimbabwe," the perfect Web page to answer the question might say, "The capital of Zimbabwe is . . .". Therefore, the search to enter is "capital of Zimbabwe is" in a full-text search tool. Again, the authority of these answers should be examined closely.

It is important to be aware when using a full-text search tool that doesn't index stopwords. As discussed earlier, these tools can be used like the full-text search tools with one exception - phrase searching with a stopword will fail. If users are not aware of this, they may get frustrated when they cannot find "Gone with the Wind," for example. DIALOG makes up for this with the NEAR connector, but so far, Web search tools do not allow the use of this connector. Other than that, searching in these tools is identical to searching the full-text tools.

Extracting Search Tools

These tools are a good second choice after trying either directory-based search tools or full-text search tools. They will provide more results than directory-based tools, although they may still overwhelm the user with results. They will usually provide fewer results than a full-text search tool because only selected parts of the target pages are indexed. The lack of phrase searching with these tools makes it difficult to achieve a high level of precision. However, if the search topic is general, if there are not any applicable phrases that might appear on all the target pages, or if the search terms tend to be used out-of-context, these tools may provide better results than a full-text tool.

ome extracting search tools index the text of the page excluding stop words and therefore any phrase searching allowed will be limited. Many of the comments regarding phrase searching from the full-text search tools section will be applicable here. However, a search including any of the stop words will fail to retrieve pertinent pages.

When entering terms without using any connectors in extract search tools, the user can gain more precise results by entering more terms. However, if more synonyms are entered for one search facet than another facet, pages matching more of the synonyms for that facet may be placed higher on the list. This may produce a desired effect or may lower the ranking of desired pages, depending upon the situation. Thus, the user needs to be aware that entering more synonyms will weight the search toward that facet.

Subject-Specific Search Tools

The advantage to subject-specific search tools is that searching done with them can be very precise, assuming the search tool covers the desired topic. There are full-text and extract subject-specific search tools, as well as subject-specific directory-based search tools. If comprehensive searching is desired, the searcher may want to see if a subject-specific search tool exists before using the general search tools.

Meta-Search Tools

Many people prefer the meta-search tools, as they allow searching in a number of Web page databases at once. The problem with these tools is a lack of control. It is difficult to use connectors properly when searching directory-based tools, extract search tools, and full-text search tools simultaneously, as different tools do different things with connectors. Many of these tools present only the first few pages returned by each search tool; therefore, the performance with these tools varies considerably.

These tools can be good for a "quick and dirty" search, in order to learn what types of pages are on the Internet. Since different search tools tend to index different types of pages, by examining the first few returned pages from a number of search tools, a searcher can use a meta-search tool to decide on which search tools to pursue in-depth searching.

Some of the meta-search tools such as Inference Find, http://www.infind.com, group the returned results by topic instead of by type. This is useful in quickly identifying aspects of a search term that create false drops. After examining the topic categories returned in a search, the searcher can use NOT or "-" to remove groups of pages from the returned list.

Table 2. Advantages and Disadvantages of Categories of Web Search Tools

Type of tool	Advantages	Sample tools
Directory-based search tool	Browsing Starting point for searching Not as intimidating for novices	http://www.yahoo.com http:// www.about.com http://www.looksmart.com
Full-text search tool	Specific companies or people Quotes, poems, or lyrics Specific topic searching "Ready reference" fact searching	http://www.altavista.digital.com http://www.infoseek.com http://www.excite.com
Extracting search tool	Good secondary tool General topic searching Search terms commonly used out-of-context	http://www.lycos.com http://www.webcrawler.com http://www.hotbot.com
Subject-specific search tool	More precise when applicable	http://www.search.com http://www.clearinghouse.net http://www.ipl.org/ref/
Meta-search tool	"Quick and dirty" searching Overview of topic area	http://www.infind.com http://www.dogpile.com http://www.metacrawler.com

Conclusion

Why go to all this trouble? Why not just label anything that helps people search the Internet a "search engine" and be done with it? There are differences in the Web page databases that cause problems for the unknowing searcher. Successful queries with one tool may be unsuccessful with another type of tool. This categorization structure brings to light these hidden differences.

When doing research on Web search tools, care should be taken to compare tools from the same category using the same queries. Using search tools from different categories with the same queries may introduce bias toward certain categories of tools into the research results.

Different types of Web search tools perform better in different circumstances. Regardless of how the users enter the query, if the wrong search tool has been selected, users will get poor results. If users are familiar with the five broad categories of tools and when to use them, they will be better searchers. Thus, when training users (in person or through educational Web sites), The use of this categorization will help to search the Web more successfully.

Finally, by moving away from the ambiguous "search engine" and moving toward more precise terminology, discussion and research in this field will become more precise, replicable, and less confusing.

References

1. Xiaoying Dong and Louise T. Su, "Search Engines on the World Wide Web and Information Retrieval from the Internet: A Review and Evaluation", Online & CDROM Review 21:2 (1997): 67-81.

2. Venkat N. Gudivada, Vijay V. Raghavan, William I. Grosky, and Rajesh Kasanagottu, "Information Retrieval on the World Wide Web", IEEE Internet Computing1:5 (1997): 58-68.

3. Stacy Kimmel, "Robot-Generated Databases on the World Wide Web",Database 19:1 (1996): 40-49.

4. Wei Ding and Gary Marchionini, "A Comparative Study of Web Search Service Performance", in Global Complexity: Information, Chaos, and Control, Proceedings of the 59th Annual Meeting of the American Society for Information Science, Steve Hardin, ed. (Medford, NJ: Information Today, 1996): 136-140.

5. Greg R. Notess, "Searching the World-Wide Web: Lycos, WebCrawler and More", Online 19:4 (1995): 48-53.

6. Steve Lawrence and C. Lee Giles, "Searching the World Wide Web", Science280 (1998): 98-100.

7. as 6. p. 98.

8. Sarah J. Clarke and Peter Willett, "Estimating the recall performance of Web search engines", Aslib Proceedings 49:7 (1997): 184-189.

9. as 8. p. 187.

10. Scott Nicholson, "Indexing and Abstracting on the World Wide Web: An Examination of Six Web page databases" Information Technology and Libraries 16:2 (1997): 73-81.

11. Danny Sullivan, "Search Engine Sizes" <http://www.searchenginewatch.com/reports/sizes.html> (January 1999) Seen 8 January 1999.

12. Danny Sullivan, "Directory Sizes" <http://www.searchenginewatch.com/reports/directories.html> (November 1998) Seen 8 January 1999.