Web Search Engines and Search Strategies

Umesha Naik D Shivalingaiah

Abstract

Information is available in many different forms. Search Engines are very useful tools to find out information available on the net. The competition among Internet search engines is intense, and that's good news for net users. But despite each search service's attempt to become the ultimate online find-it tool, no single service can cater for every need. Library and Information professional use this type of tools and get better result from the site, use popular search engines. Internet is a wide network many type of information can get it. In this paper the Authors suggest that the resources are available in the net somehow know web address of the site without any information of the site find the information. For better use of this one should use the search engine. Search engines have a variety of ways to refine and control your searches. Some of them offer menu systems for this. Others require you to use special commands as part of the query.

Keywords : Search Engine, Information Retrieval, Search Strategies, Search Techniques.

0. Introduction

Imagine Internet as a huge library with no catalogue and no staff to assist. To access a site/source on the Internet, one should know its URL (Uniform Resource Locator) or Web address. It is not possible to remember the addresses of all the required sites/sources. Since information retrieval becomes a major problem a number of search tools or services have been developed. Search engines were originally described as automated programs that compiled and updated databases without human intervention. An Internet novice is often told to "search" when looking for details pertaining to a particular topic. Search Engines are useful because the Internet is made up of literally millions of websites containing various forms of information and products. Search Engines are Internet companies that collect information about all other websites. The information is then listed by category and description to expedite the process of finding available websites in a desired category.

1. What is Search Engine ?

Any arbitrary method of Internet searching is called a "search service" (the most general term). A manually administered database is called a "catalog" or a "directory" (or in simple cases a "list"). An automatic robot, which indexes Internet data mainly by itself, is called a "search engine". Search engine is a program ("robot" or "spider") that indexes other Web pages. The name Search Engine is an intriguing term. Just as a car engine, a search engine has many parts to it. The first element is often called the SPIDER or the CRAWLER, which visits web pages and reads the information found. Next comes the INDEX part, which is also often referred to as the catalog. This is like a giant book, which contains a copy of every page on the web that the SPIDER has found.

A search engine is a collection of software programs that collect information from the Web, index it, and put it in a database so it can be searched. Search Engine is automated keyword searching tools, it use piece of software, usually known as a 'spider' or 'crawler' to gather the information from web and other servers and generate indexes. Search engine crawl the networks continuously to update their databases. It usually indexes the full-text of web page and holds lot of information in the databases. They are quite comprehensive and freely available but not complete.

Search Engine is a specialized program that facilitates information retrieval from large segments of the Internet. Search engines attempt to help a user locate desired information or resources by seeking matches to user-specified key words. The usual method for finding and isolating this information is to compile and maintain an index of Web resources that can be queried for the key words or concepts entered by the user. The indices are often built from specific resource lists, and may also be created from the output of Web crawlers, wanderers, robots, spiders, or worms. The indices are usually compiled during times of minimum network traffic. Different engines are appropriate for different kinds of searches, and most can be optimized for specified results. It allows you to enter words and phrases about what we are looking for and then tries to find the closest match in its database. Some common search engines are AltaVista, Excite, Google, Infoseek and Yahoo.

The search engine performs keyword searches against the database and retrieves a set of Web Pages matching the query. A search engine is a giant database of many Web sites on the Internet. A search engine generally returns the result of a search ranked by 'that search engines' indexing criteria. This formula varies widely between search engines.

1.1 Component of a Search Engine

Internet search engines operate in a manner similar to information retrieval systems. The main components of a search engine are *Gatherer* (A Spider), *Indexer* and the *Search Interface.*

1.1.1 Gatherer

Gatherer or Crawler or Spider gathers content descriptors from the document collection, which continuously traverses the Web and picks up the newly added WebPages/documents.

1. Spiders or Crawlers or Robots travel from site to site, looking for new WWW pages
2. Gatherer or Crawler or Spider gathers content descriptors from the document collection, which continuously traverses the Web and picks up the newly added WebPages/documents.
3. Some spiders only go to the "What's New" or the "What's Hot" pages

1.1.2 Indexer

Indexes the web pages gathered by the spider and build the database.

1. Some search engines do full text indexing. e.g. AltaVista and Open Text
2. Some search engines do keyword indexing, e.g. Lycos and Excite
3. Yahoo and Magellan are two of the few examples of human indexing.
4. While indexing, types of Web sites indexed may be – HTTP, Gopher, Telnet, FTP, and Usenet Groups.
5. Single type of source Indexing. e.g. Dejanews indexes only Usenet group postings
6. Types of files indexing: HTML files, Audio, Video, Images, etc.

1.1.3 Search Interface

1. It is an interface between User and Database
2. Interface has been designed using HTML and CGI
3. It collects query (Keywords) from the user, submit to the Database and display results based on the matching and relevance.

2. Chronological Development of Search Engines

Year	Search Engine name
1990	Archie
1991	Gopher
1992	Veronica
1993	Jughead
1993	World Wide Web Wanderer
1994	Galaxy
1994	Yahoo!
1994	WebCrawler
1994	Lycos
1995	Infoseek
1995	MetaCrawler
1995	Excite
1995	AltaVista
1995	Search Savvy Meta search engine.
1996	Inktomi
1996	HotBot
1996	LookSmart
1997	Ask Jeeves
1997	GoTo
1997	Northern Light
1998	Open Directory Project
1998	Google
1998	MSN
1998	Direct Hit
1999	InfoSeek
1999	NBC
1999	FAST Search
2000	Teoma
2001	Ask Jeeves acquires the Teoma search property
2001	Lycos search Discontinued
2001	AltaVista switches to Yahoo!
2002	LookSmart bought the WiseNut search engine.
2002	Froogle search engine
2003	Google released a contextual based ad program by the name of AdSense
2003	Overture purchased AllTheWeb and AltaVista. Yahoo gobbled up Intomi and Overture.
2004	Yahoo in 2004 dumped Google in favor of its own in house search engine. Yahoo! Slurp is believed to be collecting data to make a new database separate from the Inktomi database. The new Yahoo! database replaced both AltaVista and AllTheWeb in March 2004

3. Categories of Search engine ?

Search engines fall into five categories they are

- Robotic Internet search engines uses a Web robot to retrieve a significant number of documents from the World Wide Web.
- Mega-indexes have links to the robotic search engines
- Simultaneous mega-indexes access the robotic search engines simultaneously
- Subject directories are manually-maintained collections of Web sites organized by topic and
- Robotic specialized search engines focus on a small or specialized segment of the Internet.

4. Types of Search Engine ?

There are five types of search engines

1. Free-Text Search Engines
2. Index or Directory-based Search Engines
3. Multi or Meta-Search Engines
4. Natural-Language Search Engines
5. Resource or site specific Search Engines

4.1. Free-Text Search Engine

Free-text Search Engines are just one way of finding the information that we need on the net. These types of search engine are very easy and useful if we know exactly what we are looking for. Simply search for any single word, a number of word or in some cases a phrase. e.g. AltaVista, Lycos, HotBot and Northern Light etc.

Free-text Search Engines will;

- accept any terms the user wishes to search for,
- can search for terms in any combination,
- can search for phrase as well as single words,
- allow users considerable flexibility in choosing how to search.

4.2 Index or Directory-based Search Engines

Index–based search engines are less useful if we require a broad overview of a subject or unfamiliar with a subject and its technical jargon. e.g. Yahoo, Magellan, Excite

Index–based search engines are;

- arrange data in a structured fashion,
- make use of heading and subheading – general to specific,
- web authors (human made) to submit pages to the engine,
- depend on their category structure for their success,
- generally quite simple to use, and appeal to novice searches,
- useful if we want a broad approach to a subject,
- useful if we are unsure of keywords to use in a search.

4.3 Multi or Meta-Search Engines

An automat that scans search engines in parallel and merges the results is called a "*meta-search engine*". If this automat is running as a client on the user PC it is "*client-based*" meta-search engine. If this automat is running as a server, answering queries of many users, it's a "*server-based*" meta-search engine. They are also called Meta Crawlers or multi search engines and they do not crawl the web compiling their own searchable databases. They search the databases of multiple sets of individual search engines simultaneously from a single site using the same interface. They function as intermediary and present the results of their searches in two ways: - one is single lists (merged and duplicates removed) and Multiple Lists (not collated, displayed specially, duplicates may appear).

4.3.1 Three main factors determine the usefulness of any meta-search engine,

The search engines: the search terms send to (size, content, number of search engines, ability to choose the search engines) all of them search subject directories as well as search engines and intermixed results from all.

How the search terms and search syntax are handled (Boolean operators, phrases, and defaults imposed) How results are displayed (ranking; aggregated into one list, or with each search engine's results reported separately)

4.3.2 What does this Meta Engine do ?

- Translates the search request.
- Queries all relevant search engines intelligently for you.
- Processes the results.
- Presents you with the best possible sites.

Unlike the individual search engines and directories, meta-search engines do not have their own databases; they do not collect web pages; they do not accept URL additions; and they do not classify or review web sites. Instead, they send queries simultaneously to multiple Web search engines and/or Web directories. Many of the meta-search engines integrate search results: duplicate findings are merged into one entry; some rank the results according to various criteria; some allow selection of search engines to be searched. Meta search engines don't crawl the web themselves to build databases. Instead, they allow searches to be sent to several search engines all at once. The results are then blended together onto one page.

Successful use of a meta-search engine depends on the status of each of the individual search engines used. Some may be heavily loaded at the time; some may be unreachable. The added features mentioned above require further resources from the meta-search engines, resulting in slower response time, a serious problem with many of the meta-search engines. Many of them, therefore, have a timeout period, so that attempts to work with a particular search engine can be abandoned if no response comes from it within a set period of time

4.3.3 Seven steps can describe the principle of a meta-search engine

- Accept a user query,
- Convert the query into the correct syntax for every underlying search engine,
- Launch the multiple queries,
- Wait for the results, and in parallel do some searching on a local database (Quick Tips),

- Analyze the results, eliminate duplicates, do a ranking,
- Merge the results,
- Deliver the post-processed results to the user's client.

4.3.4 Limitations of Meta-Search engines

How do you know if your search terms will "work"? Anyone who does Internet searching knows, search protocol (the way you enter search keywords) is far from standardized. Almost all accept " " as causing a *phrase*. A few accept *Boolean* AND, OR, and NOT. Fewer accept ( ) to group terms. Some only accept + or -. Some default to OR, some to AND. Some take * to *truncate.* Other *stem* automatically and so on.

4.4 Natural-Language Search Engines

Some support natural language queries – accepts queries like 'What is the height of Mount Everest' or 'Why is the sky blue'. A recent development in the field of search engine technology is the introduction of what may be termed natural-language search engines that is a search engine which not only understand the request that has been made, but is able to interrupt the question and come up with answers about the subject that are not entirely based on the words or phrases used by the questioner. They can be very useful if you have real problem finding information. e.g. AskJeeves, Albert

4.5 Resource or site specific Search Engines

This type of search engines are perhaps the largest but paradoxically the least used, probably as a result of their diversity. e.g. search a particular resource, such as Bible, a Dictionary, and Encyclopedia.

5. General tips for Searching the Web

5.1. Search Options

- Natural Language Processing
- Boolean Operators
- Vectors
- Fuzzy Matching
- Phrase Searching
- Proximity Matching
- Concept Browsing & Automatic Matching
- Thesaurus
- Query By Example
- Stemming & Substitutions
- Non-English character matching
- Special features (price-range searching, for example)
- Spelling error tolerance.

5.2 Categories of Search Tools

5.2.1 Information

- directories
- search engines
- meta search engines
- subject gateways and virtual libraries

5.2.2 People

- online directory service
- email and phone directories on institutional websites
- mailing list memberships
- directories of people using Usenet news

5.2.3 Resource / Tools for Searching...

Software

- searchable database of software service
- anonymous FTP archives
- Archie

5.3 Boolean Operators

Boolean logic takes its name from British mathematician George Boole (1815-1864), who wrote about a system of logic designed to produce better search results by formulating precise queries.Broad or general terms will return thousands of possible sites. Try to use terms that are more specific to your topic. To narrow your terms, look at sites that you already have found and that are relevant to your topic. Identify possible search terms from those sites. You also can combine terms, using Boolean Operators.

Particular words called Boolean operators: AND, OR, NOT can be used in searching to help specify the information to be located. The following diagrams illustrate how they work.

- **AND** Search will reveal only profiles containing both words.
- **OR** Search will reveal profiles with entire word.
- **NOT** That word will not appear in the results

5.4 Wildcards and other operations

- **NEAR** Result will have the linked words within number of words of each other
- **FAR** The linked words will appear within some number of words apart at least once in result.
- **BEFORE** Words appear in specific order but not necessarily near each other.
- **ADJ** The words will appear next to each other.

✍	**BUT NOT**	That word will not appear in the search results.
✍	**NEAR/#**	Result will have the linked words within the specific number (#) words of ach other.
✍	**O**	Directly before another operation (OADJ) will force result to be in order to specify.
✍	**/#**	Use directly after ADJ, NEAR and FAR to specify the # of words allowed between search words.
✍	**( )**	Group words together so you can search for a couple of different options at a time.
✍	**NEAR#**	Result will have the linked words within the specific number # words of each other.
✍	*****	Substitute for any string of characters.
✍	**?**	Substitute for one letter.
✍	**^**	Substitute for any string of characters
✍	**|**	Narrow by placing between a broad-category search word and a narrow-category search word.
✍	**URL**	Search for link to a URL
✍	**U**	Search word will appear in URL
✍	**T**	Search word will appear in Title
✍	**Site**	Search for pages at a particular website.
✍	**{}**	Search words in brackets will appear within some number words of one another.
✍	**+**	Requires following search word to appear in each search result.
✍	**-**	Requires following search word to be absent in each search result.
✍	**“”**	Place around any number of word you want searched for as a phrase.

5.5 Search Tips

Some Search Engine allows searching of both the web and many Usenet Newsgroups. It allows control of the result lists in a standard, compact, and detailed format. It provides both simple and advanced searches. Advanced searches include all the features of simple ones, and also allow the use of Boolean and proximity operators, grouping of terms by parentheses, and results ranking by keyword.

5.5.1 Case Sensitivity : Search terms entered in lower case letters are case insensitive. The use of capitalized terms (or accented letters) makes the term case sensitive. HotDog finds only the terms spelled exactly with that capitalization; hotdog finds all occurrences of the term, regardless of capitalization. López only finds a word spelled exactly that way.

5.5.2 Phrases : To group search terms into phrases, include them in double quotes. “Abraham Lincoln” finds occurrences of the name Abraham Lincoln, capitalized in just that way.

5.5.3 Required Terms : To require that one of your terms be included in the document being indexed, preface (the formal term is prepend) it with a + symbol

5.5.4 Prohibited Terms : To prohibit the inclusion of a term from a document for which you are searching, prepped it with a – symbol

5.5.5 Wildcards : With simple queries you are allowed to enter a wildcard character at the end of phrases, which will substitute for any combination of letters. The asterisk (*) is AltaVista's wildcard character

5.5.6 Rankings : AltaVista will assign a confidence ranking to the hits it returns based on the following:

- The query terms are found in the first few words of the document (especially the title of web pages).
- The query terms are found in close proximity to one another in the document.
- The document contains more of the search terms than other documents.

6. Search Engines and Information Professionals

6.1 How to use Search Engine

- simply enter relevant words in to a search form
- set the appropriate options
- search terms will then be compared to the index/database and any matching results returned
- results are displayed as per the relevance (ranking)

6.2 When to use Search Engine

- if we need lots of information
- if we have a fairly specific information need
- if we are searching for organizations or people

6.3 Good for simple searches : Meta-Search engines are useful if the user looking for a unique term or phrase (enclose phrases in quotes " "); or simply want to test run a couple of keywords to see if they get. For such straight-forward searches, the unique ranking algorithm used by Google (based on how many other sites link to a site) often finds exactly what they want, better than any meta-search engine.

6.4 For more difficult searches : we recommend a search engine where you can search within results on a term or phrase you specify. We recommend learning Alta Vista Advanced Search and Northern Light Power Search and possibly Info seek whenever you retrieve a huge result and want to focus on some aspect, some other approach. Please consult our recommended search strategy based on what you know and want to know. We will learn when to consult Subject Directories, how to look for expert guides and specialized databases - all of which have a valuable place in the repertoire of searching skills for the experienced searcher.

6.5 Use meta-search engines - but use them 'CAUTIOUSLY' : Most meta-search engines only spend a short time in each database and often retrieve only 10% of any of the results in any of the databases queried. Thi's makes their searches usually "quick and dirty," but often good enough to find what you want.

Most meta-searchers simply pass your search terms along, and if your search contains more than one or two words or very complex logic, most of that will be lost. It will only make sense to the few search engines that support such logic (see table of general search engine features).

Quantity in results does not equal satisfaction. If you get more results than you want, try refining the results by going directly to AltaVista Advanced Search, Northern Light, or Infoseek by clicking on their link in the results. Choose meta-search engines that offer some of these as options.

Look for meta-search engines that also send your terms to selective or odd databases like WebCrawler, Thunderstone, Direct Hit, and WhatUSeek. One of the advantages of a meta-searcher is that you might overlook databases like these which may have sites missed by the big boys.

6.6 A word on Portals

The trend is for many search sites to offer not only searching and links to resources by subject, but also many other services (stock quotes, airline tickets, shopping malls, news links, games, chat rooms, free e-mail, and much more). The goal seems to be to lure as many users to the site and keep them there as long as possible, probably because the site's advertisers may benefit.

7. Advantages

- Best suited for complex/interdisciplinary search topics
- Searches can be limited to a period of time
- Currency of information: Web spiders traverse the Web almost everyday, so the latest additions to the Web can be retrieved
- Exhaustive information is retrieved on a particular topic

8. Disadvantages

- The search is time consuming: The search normally results in too many hits and this contains a lot of irrelevant documents
- The searcher should be familiar with the search techniques
- Search engines vary from each other
- The retrieved set may contain dead links

9. Conclusion

URL guessing can help in finding pages for URLs that no longer work and links that lead to dead ends. Try chopping off parts of the URL starting on the right-hand side and stopping at every. Subject directories select and classify resources into subject categories and subcategories. Some include reviews and/or ratings. Access is by keyword search or by Search Engines attempt to find and index as many sites as possible. Search features vary greatly, as does the actual scope, size, and accuracy of the databases. Unique Keywords, Combinations of unique keywords, Field searching and limiting and Pages buried deep in a web site are the main points. Searching information on the Internet is a complex process and tools that use in locating are directories, search engines, meta-search engines, subject gateways/virtual libraries etc. Keeping up-to-date with the developments in the area is necessary and as information on the net grows, more and more search tools will be designed. There is a greater need for organising these resources using skills of librarianship.

10. References

1. BRADLEY (Phil). The Advanced Internet Searcher's handbook. 2002. Library Association; London
2. CHOWDHURY (GG) and CHOUDHARY (Sudatta). Information Sources and Searching on the World Wide Web. 2001. Library Association; London
3. ERUEST (Ackermann) and KAREN (Harturan). The Information Specialist's Guide to Searching and Researching on the Internet and the World Wide Web. 1999. Fitzroy Bearborn,
4. Iprospect.com: http://www.iprospect.com/search_engine_placement/seo_history.htm Visited: 30/11/2003
5. Librarians' index to the internet: http://www.lii.org Visited: 30/11/2003
6. MINOLI (Daniel). Internet and Intranet Engineering. Technology, Protocols, and applications. 1999. TaTa McGraw hill; New Delhi
7. Positionmaster.com: http://www.positionmasters.com/positioning/definitions.html Visited: 30/11/2003
8. RAJASHEKAR (TB). Internet and Web Search Engines. Library Herald, April-June; 1999; 60-74. Askscott.com: http://www.askscott.com Visited: 30/11/2003
9. Impact web design: http://www3.sk.sympatico.ca/dean030/search_history.htm: Visited on 30/05/2004
10. Internet Complete, 1998. BPB; New Delhi
11. Seoconsultants.com: http://www.seoconsultants.com/search-engines/history.asp: Visited on 30/05/2004

About Authors

Mr. Umesha Naik is currently working as a Lecturer in the Department Library and Information Science, Mangalore University, Mangalore. Prior to this he has worked 8 years at INFLIBNET Centre. He obtained his B.L.I.Sc. degree from Mangalore University and M.L.I.S from IGNOU. His areas of interest are Networking, Internet, Web Design, Digital and Electronic Libraries. He has published four articles in journals and seminar/ conferences.

Dr. D. Shivalingaiah is a Reader in Library and Information Science, Mangalore University, Mangalore. He holds M.A. in Rural Development and M.L.I.Sc. from Bangalore University and Ph.D. from Mangalore University. He has successfully guided a candidate for Ph.D. programme. Presently five candidates are working under him for Ph.D. programme. He has publications in Journals and Conference Proceedings and edited books. He is presently working as Deputy Registrar (Administration) on deputation.