Comparative Study of Search Engines

RK SHUKLA* & JATA SHANKAR GUPTA**

ABSTRACT

The internet can be said to be most exhaustive, important and useful source of information. Data on almost all the topics on earth are hosted in million of servers connected to the internet around the world. There is no set policy for hosting the data of number of centralized database for organization and searching the desired information. With the availability of excessive information, it has become very difficult for the LIS professionals to find the precise information required by the user. However, there are a variety of search tools for resource discovery on the web.

OBJECTIVES

Internet is growing day by day, making a jungle of information where the information is highly unorganized and finding relevant information in such a jungle is not an easy task. Although various search engines are available to solve this problem. But again finding information through search engines is not an easy task, until the users have sufficient knowledge about different search techniques available on a particular search engine. Thus the main objective of the present study is:

1 To acquainted LIS professionals with search engines.
2 To suggest various ways to get the desired information on the web.
3 To make acquainted with the difference among the search engines, directories, and meta-search engines.
4 To make understand functioning of the search engines.
5 To identify behaviour that contributes to success in searching when success is identified by standard measures such as recall value and precision.

PURPOSE

The purpose of the present study is to determine and analyse various search engines and strategies on different search engines , which will be helpful for LIS professionals in searching information and get required information with greater recall and precision value.

LIMITATION

The present study aims to understand various techniques and strategies available on different search engines. Since the number of search engines available on internet is large, the study has been restricted to some highly used search engines i.e. GOOGLE, HOTBOT, ALL THEWEB, YAHOO, MSN, ALTAVISTA.

METHODOLOGY

For the problem under the study, the data has been collected using following techniques:

Literature survey has been conducted through various primary and secondary sources to have a clear understanding of the topics, aspects to be covered on internet and web searching.

In order to get background information on search engines, systematic and exhaustive search

*Delhi College of Engineering email ramakant.shukla@gmail.com **IIT Delhi bholajata@gmail.com

for published literature has been done to collect available information from various categories of sources.

In the pursuit, internet has been surfed to compare the features as well as the retrieval performance of the above mentioned search engines. The features were obtained from home pages of search engines and articles given on different websites. As regards to the retrieval performance, the performance of search engines were tested by a set of queries. Thus experimental study has been conducted

The term search engine more properly refers to any software used to search any database, on the internet the phrase usually refers to a very large database of web sites that are automatically built by robots. These internet search engines use a software robot (or spider) that seeks out and index the words on web pages. In common language search engines might be called a search engines service or a search service. Following are some of the definitions of search engines.

Search engines are the keys to finding specific information on the vast expanse of the World Wide Web.

Search engines are system on the internet dedicated to finding, classifying and storing information and providing this information to users. When you visit a search engine, you can look for sites, which have information of reference on a particular topic by supplying some keywords to the engine.

A program that search documents for specified keywords and returns a list of the documents where the keywords were found. Although search engine is really a general class of programs, the term is often used to specifically describe systems like Alta Vista and Excite that enable users to search for documents on the World Wide Web and USENET newsgroup.

A generic term for the software that “searches” the web for pages relating to a specify query.

TYPES OF SEARCH ENGINES

There are basically four types of search engines:

1. Free text search engines
2. Index or directory based search engines
3. Multi or meta search engines
4. Resources or site specific search engines

FREE TEXT SEARCH ENGINES

Free text search engines are those software which helps simply search for any single word, a number of words or in some cases a phrase. Free text search engines are useful if one knows exactly what one is looking for. They are less useful if one want a broad overview of a subject or are searching in an area that one don't know very well and consequently have no idea as to the best terms to use.

Example: Google (http:/www.google.com)

INDEX OR DIRECTORY BASED SEARCH ENGINES

Index or directory based search engines classify information under a series of major subject headings, and then subdivide these into a tree structure of more specific headings, and sites are listed as appropriate in this directory structure.

Example: Yahoo (http://www.yahoo.com)

MULTI OR META SEARCH ENGINES

Multi or Meta search engines do not compile their own searchable database; instead, they search the databases of multiple sets of individual search engines simultaneously from a single site and using the same interface.

Example: Metacrawler (http://www.metacrawler.com)

RESOURCE OR SITE SPECIFIC SEARCH ENGINES

This type of search Engines are created for one particular resource type or sites such as Dictionary, Bible etc. Example: (http://www.reference.com)

WORKING OF SEARCH ENGINES

Any search engine performs the following three basic tasks:

- They search the internet based on a set of criteria.
- They maintain an index of the words / phrases with specific information such as where they found them how many times they found them.
- They allow users to search for words / phrases or combinations of for words / phrases available in their index.

There are three main components of a search engine: “the spider ", "the index”, and "search engine software and interface".

THE SPIDER

To extract information out of the millions of web pages that exist, a search engine employs a special program, called a spider, to build lists of the terms found on websites. It is called a spider because it crawls over the web to automatically fetch web-pages at regular intervals such as every week or month, to look for changes. Web pages contain links to other pages and spider uses these links. As soon as it finds a link to another page, it visits the pages, reads it, and then follows links to other pages within the site.

Spiders fetch terms occurring in the title, subtitles, meta-tags and other positions of relative importance of selected web pages. The selection criteria vary from one search engines to another, and these different approaches usually result in the differences in the results.

THE INDEX

The term that a spider finds matching the criteria go to index. If the contents of a web page changes, then the spider notices it, updates the index with the new information. However, it may take sometime before the changes in a web page are noticed and indexed by a search engine, depending on several factors, like how frequently the spider visits the same pages, how frequently the index is updated etc.

THE SEARCH ENGINE SOFTWARE AND THE INTERFACE

The search engine software is the information retrieval program that performs two major tasks:

- It searches through the millions of terms recorded in the index to find matches to a search.
- It ranks the retrieved records in order of what it believes is the most relevant.

The criteria for selection of search terms and assigning weight to them depend on the policy of the search engine concerned, as does the specific information that is stored along with each keyword such as where in a given web page it occurred, how many times it occurred, the attached weight and so on.

COMPARISIONS AMONG SEARCH ENGINES

CRITERIA FOR COMPARISION

The study revealed various aspects of the search engines. A few features are not common but majority are common to all search engines.

Comparison of search engines have been done on two aspects namely Interface capability, retrieval performance.

INTERFACE CAPABILITY

	Boolean	Phrase searching	Field searching	Special features	Help
Google	Default	“”	Link:,	Offers a similar pages	Help

*Delhi College of Engineering email ramakant.shukla@gmail.com   **IIT Delhi bholajata@gmail.com

	AND between words;(place immediately in front of term to exclude); can use OR (use capitals); in advanced search can select “all”, “without”		allintitle :, Intitle:, Allinurl:, Inurl:, site:, Related:, info:, (details at advance search operators-be careful in the use of these multiple field searching may provide differing results). Advanced mode limits include: language, format, date and domain	option within each result; offer translation of pages; can do a general search e.g. Austrailia and get news information at the top of your results; can use as part of a phrase search e.g. three mice will provide a phrase result of three [anyword] mice; if searching on a broad topic, will have option of going to a category to help guide your search. Can use the title (-) in front of a word to search for synonyms (not: not reliable!); use Google Proximity Search (GPS) to do easy proximity searching: can limit by file type. Can set up alerts via www.googleallert.com (note: this service is not affiliated with google but works very well!)	
MSN Search	Default AND between words;(place immediately in front of term to exclude); in search builder select: all of these terms/any of	“”	In search builder can limit by site/domain. Find. pages that link to a site you input; limit by country and language	Offers a nifty search builder (similar to advance search in other search engines); allow you to influence results ranking by popularity, recent updates and approximate/exact matches.	Help

	these terms/this exact phrase/none of these terms				
Yahoo search	Default AND between words;(place immediately in front of term to exclude)	“”	Site:, hostname:, link:, url:, inurl:, intitle:	Provide some nifty search shortcuts; incorporates some directory features which can be useful (see the Help info); provides related search terms; allow to replace a word e.g. three mice Try out Ujiko which uses Yahoo search techn0ology in a fun way- and also provides grouped results to help you narrow down your search	Help
AltaVista	Supports full Boolean searching with the operators AND, OR, and NOT… Also, symbols can be used: & for AND,/ for OR, !for AND NOT. Be sure to use AND NOT rather than just NOT.	“”	Anchor, applet, host, image, link, text, title, url, like:-	First to introduce a free translation service this currently offers translations between English and Chinese, French, German, Italian, Japanese, Korean, Spanish, and Portuguese. It also goes from Russian to English, German to French, and French to German.	Help
ALLTHEWEB	AlltheWeb uses ‘and’,		url:, link: or link.all,	AlltheWeb’s index (provided by Yahoo!)	Help

	‘or’, and ‘andnot’, allows the uses of a + for AND, - for NOT, and multiple words in parentheses		title: or normal.title, site	includes billions of web pages, as well as tens of millions of PDF and MS Word@ files, offers a variety of specialized search tools and advanced search features	
HOTBOT	Use the operators AND, OR, and NOT. Searching can be nested using parentheses. Operators must be in upper case.	“”	Title, domain, depth, linkdomain	Advanced searching capabilities, Quick check of three major database, Advanced search help	Help

RETRIEVAL PERFORMANCE

Estimation of Precision and Recall

Precision is the fraction of the search output that is relevant for a particular query. Its calculation, hence, requires knowledge of the relevant and non-relevant hits in the evaluated set of documents. Thus it is possible to calculate absolute precision of search engines which provide an indication of the relevance of the system. In the context of the present study precision is defined as:

Precision	Sum of these scores of scholarly documents retrieved by a search engine/Total no. of results evaluated

To determine the relevance of each page, a four point scale was used which enabled us to calculate precision. The criteria employed for the purpose is as under:

1 A page representing full text of research paper, seminar/conference proceedings or a patent given a score of three.
2 A page corresponding to an abstract of a research paper, seminar/conference proceedings or a patent is given a score of two.
3 A page corresponding to a book or a database is given a score of one.
4 A page representing other than the above (i.e. company web pages, dictionaries, encyclopaedia, organization etc.) is given a score of zero.
5 A page occurring more than once under different URL is assigned a score of zero.
6 A non-response of the server for subsequent thee searches is assigned a score of zero.

The recall on the other hand is the ability of a retrieval system to obtain all or most of the relevant documents in the collection. Thus it requires knowledge not just of the relevant and retrieved but also those not retrieved. There is no proper method of calculating absolute recall of search engines as it is impossible to know the total number of relevant in huge databases. This study also followed the method used by Clark and Willett by pooling the relevant results (corresponding here to scholarly documents) of individual searches to form the denominator of the calculations. The relative recall value is thus defined as:

Relative Recall=	Total number of scholarly documents retrieved by a search engine/ Sum of scholarly documents retrieved by all five search engines

However, in the case of overlapping between search engines results, only the overlapped results are included for the pooling by taking five search engines(say a, b, c, d, and e) into consideration which retrieve a1, b1, c1, d1, and e1 scholarly documents respectively. Further, where there is no overlap between search engines (i. e. a ∩ b, a ∩ c, a ∩ d, a ∩ e is zero) then the relative recall of search engine ‘a’ is calculated as a1/(a1+b1+c1+d1+e1). Again if overlapping exists between search engines i.e. a ∩ b=b2, a ∩ c=c2, a ∩ d=d2, a ∩ e=e2 then the relative recall of search engine ‘a’ is a1/(a1+b2+c2+d2+e2). The relative recall is more in case of overlapping between search engines. The mean value for precision and relative recall is obtain by micro-averaging i.e. average score for each engine against a query is summed over all the twenty queries and mean value calculated from these totals for single, compound and complex terms separately.

Results and Discussion

The mean precision and relative recall of select search engines for retrieving scholarly information are presented in Table 1.

Table 1: Precision and Relative Recall of search engines

	AltaVista	Google	HotBot	Yahoo	MSN	Alltheweb
Precision	0.27	0.29	0.28	0.25	0.24	0.23
Recall	0.18	0.20	0.29	0.17	0.15	0.14

The results depict better performance. Google is the best alternative for getting web-based scholarly documents and its recent introduction of ‘Google Scholar’ for accessing scholarly information offers better dividends for researchers. HotBot offers a good combination of recall and precision built has a larger overlap with other search engines which enhance its relative recall over Google search engine. AltaVista once prominent on the web has lagged behind and the Alltheweb is the weakest among the select search engines in all respects. Further, the results reveal that structured queries (i. e. phrased and Boolean) contribute in achieving better precision and recall. The findings also establish the case that precision is inversely proportional to recall i.e. if precision increases recall decreases and vice versa.

FINDINGS

Table 1:- Database

Search Engines	Google	Yahoo	AltaVista	HotBot	MSN	Alltheweb
Directory	Available	Available	Available	Not available	Not available	Not available

Table 2:- Content Size of search engines

Search Engines	Google	Yahoo	AltaVista	HotBot	MSN	Alltheweb
Content Size	1.25 billion sites	Not available	250M pages	110M pages	Not available	Not available

Table 3:- Indexes of search engines

Search Engines	Google	Yahoo	AltaVista	HotBot	MSN	Alltheweb
Indexes	First 101KB of webpage & 120 KB of PDFs	First 500 KB of web pages	First 110K of web pages & 750 K of PDFs	Not available	Not available	Full web page & PDF files

Table 4:- Cached archive of web pages of search engines

Search Engines	Google	Yahoo	AltaVista	HotBot	MSN	Alltheweb
Cached archive of	Available	Available	Not available	Not available	Available	Not available

web pages						

Table 5:- Search strategy of search engines

Search Engines	Google	Yahoo	AltaVista	HotBot	MSN	Alltheweb
Default word	AND	AND	OR	AND	AND	AND
Boolean connectors	Limits including & excluding words	AND, OR, NOT/AND NOT	AND, AND NOT, NEAR	OR, AND NOT in UPPER CASE	AND, OR, NOT/AND NOT	AND, OR, AND NOT
Phase search	Double quotes	Double quotes	Double quotes	Double quotes	Double quotes	Double quotes
Truncation	No, Use *	No	No, Use *	No	No	No
Case sensitive	No	No	Yes	Yes	No	No
Special features	Limit by files, language, family filter, Domain	Limit by File, language, Safe search filter, date, country, Domain	Limit by language, region, file, date, domain	Limit by domain. Date, language, Page type, Page content, region	Limit by domain, location, file, language	Not available

The study has revealed that all search engines are having same search techniques with the exception of one or two. For finding specific information, search engines and Boolean logic are seems to be appropriate, most of the times. And all the search engines have the options for Boolean searching.

The Boolean operators AND is for narrowing the search. It is same as looking for information from both keywords, provided. The Boolean OR operator is valuable when the searches want to make sure that he or she doesn't miss any important information. If one, find a term that occurs in most of the documents that is not wanted, and then in such condition NOT can be used. This is equivalent of request everything containing one keyword WITHOUT the misleading wor

RECOMMENDATIONS

The study recommends that while formulating a search strategy, the searcher should keep in mind the following things:

Be Specific: This will get, the searcher, closer to what is really wanted. Typing "search engine" and "search engine comparison" will give different set of results.

Add Quotation Marks: The searcher can avoid confusing many search engines by keeping phrase intact with quotation marks. It is particularly helpful for searches that include proper names like “NISCAIR”.

Be Advanced: It is being advised that not to be shy away from the advanced search tool offered by many search engines. In one easy-to-use form, one can focus results in numerous ways.

Think outside the web: Much information is buried in database and not in web pages scoured by search engines. Access such hidden information through invisible web.net.

CONCLUSION:

The study reveals that while searching information on the internet, first of all it is important to know, what type of information is to be found out, as number of searching tools are available on internet. Search engines are one of them, and will be only beneficial if a searcher is looking for specific information or locations like WWW sites, Usenet discussion groups, FTP sites, other databases etc. after choosing a search engine it is important to learn how to use the tool, as each search engine has specific rules and syntax to follow.

Comparative Study of Search Engines

RK SHUKLA* & JATA SHANKAR GUPTA**

REFERENCES:

1.Guide to effective searching of the internet (http://www.brightplanet.com) visited on 1.11.2008

2. Search engines features chart (http://www.searchengineshowdown.com) visited on 1.11.2008

3.Chu, H,& Rosenthal, M(1996) Search engines for world wide web: acomperative stdy and evaluation methodology. In Proceedings of the ASIS 1996 Annual conference, October, 127-35 (http://www.asis.org/annual-96/electronicproceedings/chu.html) visited on 1.11.2008

4.Ding, W & Marchinonini, G(1996) A comperative study of the web search service performance. In Proceedings of the ASIS 1996 Annual conference, October, 127-35 (http://www.asis.org/annual-96/electronicproceedings/ding.html) visited on 1.11.2008

5.Modi, G.(1996)Searching theweb for Gigabucks.New Scientist, 150(2024),36-40

6.Oppenheiem, C.,Moris, AMcknight, C., Lowley, S.(2000)The evaluation of www search engines.Journal of documentation 56(2), 190-211

7. http://www.webology.ir/2005/v2n2/a12.html
8. http://www.google.com
9. http://www.msn.com
10. http://www.yahoo.com
11. http://www.alltheweb.com
12. http://www.altavista.com
13. http://www.hotbot.com

*Delhi College of Engineering email ramakant.shukla@gmail.com **IIT Delhi bholajata@gmail.com