Search Engines : Comparative View

By

Mukesh Sakia
Asstt. Librarian
Tezpur University
Assam
Email : *mukesh@tezu.ernet.in*

A S Chandel
Professor and Head
Department of Library & Information Science
NEHU, Shillong: 793014
E-mail : *chandel46@yahoo.co.uk*

ABSTRACT

The paper discusses the main features of four widely used search engines namely AltaVista, Google, Yahoo and Excite. Also makes a comparative study of searching features available in these Search Engines along with their limitations.

KEYWORDS : Search Engines, Evaluation, Performance, Excite, yahoo, AltaVista, Google, Information retrieval, Excite

0. INTRODUCTION

In the beginning of 1990's the world wide web started to grow more due to introduction of new features of transferring and receiving of high quality colour images, video and sound files in addition to plain text. Hence, Internet became more common and attractive to everybody. People started recognising the importance of Internet for business, entertainment, education, research and recreation. Thousands of companies, scientific laboratories, universities, private homes and individuals started placing their materials on the Internet. Everyday, more and more information are placed in the net and turned the Internet into Cyber jungle (Ding and Marchionini,1996). The constant increase of information on the net and frequent changes which are being brought in the SE are making it difficult for anyone to monitor such changes. To support the navigation in this vast virtual universe, the development of effective retrieval tool became a necessity. To cope with this situation, the search mechanisms known as Search Engines (SE) was developed. Today, there are several hundred search engines with different features. These Search Engines deal with the problem of indexing and retrieval of digital information on the web. However, every SE provides users with an interface that enables them to locate documents containing information that matches their interest.

1. SEARCH ENGINES

A search engine is a database creation, manipulation and search programme which can browse information on Internet. Search engines use software programmes, which create automatically

their own databases consisting of list of web pages. Search engines have following three different parts :

- Program called a spider (or Robot / crawler)
- A database with index
- A search software

The indexing software agents called robots/ spiders wanders through the web. These agents are programmed to constantly "crawl" the web in search of new or updated pages. The different search engines use different types of spiders whereas some spiders visit every possible site and some spiders visit only popular sites based an selective principles. When visiting a Web site, agent will record the full text of every page within the site. It will then continue to visit external links. Following these external links is how search engines are able to find site. Robot revisits periodically to refresh the recorded information. Every page found by the spiders in the database is indexed. It is built by extracting automatically words from the web pages and ranking them alphabetically using same principles as is used by any other index file. The index is therefore a list of every word found (except irrelevant words stop words – a, an, the etc.) with a pointer to its location on the database. The different approaches that search engines use in crawling the web, finding the new pages and indexing them will produce different result.

2. TYPES OF SEARCH ENGINES

There are four different types of search engines:

- Robot driven search engines : *e.g.* AltaVista, Excite, Lycos.
- Directory based search engines : *e.g.* Yahoo
- Meta Search Engines : *e.g.* Metacrawler, Dogpile
- Software tools : e.g. Web search

Features

Many studies on evaluation of most popular search engines have been attempted which has generated a lot of literature on the subject. No doubt, search engines are fast, robust, scalable, and sustainable and use a variety of techniques. However, considerable variations exist among them in the techniques used for indexing, ranking, search features and display of retrieved results. In the present study, main features of the following most popular SE has been discussed to develop familiarity in their use:

i) Google ii) Yahoo iii) AltaVista iv) Excite

The reason of this choice is due to the fact that the above search engines are the most commonly used ones and each having some unique features. On the basis of the following features these SE have discussed.

- **Search capability,** *Sorting, Stop- words, Fields Limit, Case sensitivity, Stemming Truncation, Proximity, searching Proximity, Search capability, Documentation and Display.* These features are discussed below:

- **Search capability:** Search capability measure the performance of the search engines based upon Boolean logic, phrase searching, truncation, *etc.*

- **Proximity Searching:** Phrase search requires terms to be in the exact order specified within the phrase markings. The difficult standard for identifying phrase is to use double quotes to surround the phrase.

- **Truncation:** The symbol asterisk is used to represent the rest of a term (S).

- **Stemming:** Relate to truncation, usually refers to the ability of a search engine to find word variants such as plural, singular forms, past tense, present tense *etc.*, Some stemming only covers plural and singular forms.

- **Case Sensitivity:** Test upper, lower, mix case of search term.

- **Fields:** Fields searching allows to designate where a specific search term will appear, rather than searching for words anywhere on a Web page, fields define specific structural units of a document. The title, the URL, an image tag on a hypertext link is common fields on a Web page.

- **Limit:** Commonly available limits are the data limit, and language limit, and many more are being added for the purpose of filtering information.

- **Stop-words:** Frequently occurring words are not searchable .

- **Sorting:** Typically, Internet search engines sorts the results by 'relevance' determined by their proprietary relevance ranking algorithms.

- **Sorting:** Typically, Internet search engines sort the results by 'relevance' determined by their proprietary relevance ranking algorithms. Other options are to sort by date, alphabetically by title a by root URL a host name.

- **Display:** No. of record displayed, style of display etc.

- **Documentation:** Search link to any online help files, FAQ, tutorial, explanation, manuals, press release or other selected documentation.

Some of the important features of these four SE are discussed as under:

2.1 Google

Google has become the pre-eminent Internet search engines. Google innovates search technologies and connects millions of people around the world with information everyday. Google was officially launched on 21 September, 2003 (Greg R, 2003).

Database: Google offers its own database of indexed web pages along with another collection of URL that it has not indexed. Google has an image database and a News database known as Google groups. Google also has a Page Rank version of open directory. Google database is used by AOL at Netscape's search site, as the backend search engine at Yahoo.

Strength: It is one of the largest search engines which include *pdf* documents and many other files.

Weakness: Limited search feature, no truncation, no nesting, does not support full Boolean logic. Site clustering is difficult to turn off.

Default operator: AND is default operator.

Boolean searching: Google uses an automatic Boolean but does not support AND, NOT operator, AND is default in Google. But Google supports the logical operator "OR". "+" and "_" sign are used to include and exclude common words or characters for improving search results.

Synonyms: Google can search not only a particular keyword but also for its Synonyms by placing the tilde sign ("~") immediately in front of both of the keyword.

***Proximity searching*:** Google enables phrase searching by endorsing the phrase in double quotes (" ").

Stemming/ Wildcard or Truncation: No truncation is available nor there is any automatic plural searching, word stemming.

Phrase search: Phrase search by enclosing the search term in double quotation marks exactly retrieves the phrase.

***Case sensitivity*:** Google searches are not case sensitive.

Field: Google offers several fields searches like *In title, In URL, All in title, all in url site, all in anchor.* This is valuable feature in Google.

Limits: Google has languages, domains date, file type and adult content limits.

Stop words: Google ignores frequent words such as '*the*',' '*of*' , '*and*' and '*or*'. These can be searched by putting '+' in front of them if required.

Sorting: Google results are sorted by relevance which is determined by Google's page rank analysis, determined by links from other pages with a greater weight given to authoritative sites.

Display: Display includes title, URL, a brief extract short text near the search terms, the file size and for many hits, a link to a cached copy of the page. The default output is 10 hits per screen but searcher can also choose 20, 30, 50 or 100.

Uniqueness: Google provides access to pages at the time they were indexed designated as 'cached' pages.

Documentation: Help page and press releases, Google Zeitgeist (search pattern and trends).

2.2 Yahoo

Yahoo is one of the best Internet subject directories. It can be searched directly or browsed by category. Yahoo is known as directories rather than search engines, (Greg R, 2003). But it serves both purposes quite satisfactorily.

Databases: Yahoo directory, Google for web pages, News, stocks, shopping and others. If no items are found in Yahoo itself, the search statement is automatically sent to Google data base.

Strength: Popular and larger directory databases.

Weakness: Heavy commercial bias.

Boolean searching: Yahoo does not support Boolean operators or nested searching. Boolean operators "AND", "NOT" and "OR" are not necessary to use. However, it accepts AND OR operators in addition to the use of. (+) to include a term and (_) to exclude a term. AND operator is also by default. For two or more alternative OR in capital letters is used in Yahoo search.

Proximity searching: Exact phrase searching can be done by using double quotes around a phrase. Yahoo also retrieves related searches at the top of web results page.

Truncation: Single search terms in Yahoo of five or more characters are automatically truncated. To truncate a search term of less than five characters asterisk is used. No truncation is available in phrase searching.

Case sensitivity: No case sensitivity.

Field searching: Two field searches are available namely title and URL.

Limit: In advanced search, limit is Yahoo directory sites or directory categories. Limit to entries within specified period of time viz. 1 day, 3 day, 1week, 1 month, 3 months, 6 months, 4 years are available

Stop word: Common words are ignored.

Sorting: Yahoo provides results in various categories like- websites, web pages, Yellow pages, images, news and events.

Display: Yahoo displays the category name and hierarchy, the site title, URL and sometimes a brief description.

Documentation: Search guide, about the yahoo Projects (includes press releases)

2.3 AltaVista

AltaVista is also one of the largest search engines. It has two distinct search modes like many other SE- Basic search and Advanced search.

Databases: AltaVista has a variety of data bases in addition to their regular web page database.

Strength: Powerful search features with some unique features.

Weakness: Database is not large. Slow in updating database.

Default Operator: The default operator for multiple terms has changed many times and depends on which search form is being used. In both search modes, symbol, the punctuation marks are removed and replaced with a space and the string is searched as a phrase.

Simple Search: The default is usually an AND. However, some automatic phrase recognition is still in force and certain phrases will automatically be searched as a phrase.

Advance Search: If no operators are used in the Boolean expression box, AltaVista interprets the search as a phrase search even if it does not match a phrase in their database of phrases. If multiple terms are entered in the ranking keywords box are processed as an OR operators.

Boolean searching

Simple search: Allows only the use of + for AND and - for NOT.

Advanced Search: Supports full Boolean searching with AND, OR and AND NOT. Searching can be nested using parentheses. Operators can be in lower or upper case. Also these symbols can be used: & for AND, | for OR ! for AND NOT. Altavista does not accept operator NOT.

Proximity searching

Simple: Phrase search is done by using double quotes around a phrase. Phrase searching can also be designated by putting punctuation marks between words.

Advanced: Phrase search is done by using double quotes around a phrase. Phrase search can also be done by putting punctuation mark between words. The operator NEAR (N) can also be used in Boolean expression box to designate that the search terms must be within 10 words of each other and in any order, NEAR can be entered in upper a lower case and can be nested.

Wild card or Truncation: The truncation symbol – an asterik* can be used for truncation of 0 – 5 extra characters. To get more than 5 extra characters, double asterik** is used. Truncation can also be used in phrase searches by * mark which must be typed after three letters.
Case sensitivity

Simple: Searches are not case sensitive.

Advanced search: Searches are case sensitive. If search terms are entered completely in lower case, all mixtures of upper and lower cases are searched. If a search term contains one or more upper case letters, the search is limited to only records that exactly match the specified case. This feature may disappear any time having no utility.
Field Search: Simple and Advanced Field searching is available by using the field name followed by a colon (:) followed by the field query.

Limit

Simple: Language, Region, File type and other limits can be created by the use of field searches.

Advanced: Limit by date specifying a start date and or an end date, File type, Domain limit, Region, language limit and other limits created by use of Boolean operators and field searches are also available.

Stop words

Simple search: Simple search will search some stop words but some others are discarded. Stop words within a phrase are now searched. In case of quotation marks around a stop words, SW are searched. Advanced search: No stop words are searched in advanced search.

Sorting: The actual web database results are ranked by Altavista's relevance ranking formula.

Display: AltaVista displays title, URL, file size, language and a two line extract for each hit.

Documentation: Help file, press releases, search through all sites.

Special features: AltaVista provides free translation service between English and Chinese, French, German, Italian, Japanese, Korean, Spanish and Portuguese. It also provides from Russia to English, German to French, and French to German.

2.4 Excite

Excite is one of the smaller search engines. It provides sophisticated personalization, excellent relevant results from popular queries. It's News search provides important access to web versions of newspapers, magazines and news wires, (Greg R, 2003).

Database: Excite offers its own directory database and channels along with a current news database and several reference databases. It also has a small database of customized links for commonly requested topics.

Strength: Excite provides personalization features and high relevance on popular topics.

Weakness: Excite is a smaller database There is no truncation. Field searching is also not available.

Default Operator: If multiple terms are searched, they are processed as an OR operator.

Boolean searching: Excite use + for AND – for NOT. It also supports full Boolean searching with the operators AND, OR and NOT. OR search is default when no operators are used. Excite uses either AND NOT or NOT. Searching can be nested using parentheses. Boolean operators must be in upper case. In Advanced searches neither the +, nor - Boolean operators can be used.

Proximity searching: Phrase searching is available by using double quotes around a phrase.

Truncation: Excite does not support this facility.

Case sensitivity: Excite is not case sensitive.

Field searching: This facility is not available in Excite.

Stop words search: Excite does not have a Stop word list. However, many common words and numbers will not be searched.

Sorting: Excite sorted the results by relevance with groupings by site available at the end of each brief records.

Limits: Excite automatically limits its search to English language only by default. After running a search from the main page, a follow up search box also has the language limit options. Currently one language can be searched at a time. No other limits are available on regular search. In more search or Advanced web search, options for limiting the search to nine languages are available. Display: Excite's display includes relevance score. title, URL, and brief summary.

Documentation: Help index, General search help, advanced search help, press releases.

Special features: Excite's personalization capabilities is one of the most advanced features. It searches in other important Search Engines like Google, Looksmart, Inktomi and some other.

3. INFORMATION RETRIEVAL PROBLEMS

Some of the important features of these above described SE have been discussed. It is evident from the present study that every SE has its own features. Different strategies and searching techniques are applicable. Users familiar with one SE may not be able to retrieve right information from the other. It is also not easy for the common user of WWW to ascertain and establish as to what SE is suitable for what kind of search. But paradox is that most of the Internet users feel confident that they can search the information on net without any difficulty and most of them also feel that Internet is substitute for libraries.

Let us take a few examples where these SE don't provide solution to the searching problems. As a common searcher on Net, we find certain intriguing problems in finding information we look for such as:

First query: Searching on Apple or Apples which should exclude either apple as fruit or apple as computer, similarly Penguin as publisher not as bird. No SE has a solution to differentiae between the two mutually exclusive terms. AltaVista provided cross reference to differentiate between the Apple as fruit and apple as computer. But while searching Apple as computer, literature on apple fruits were also being retrieved in addition of many other unrelated output of concepts like: Apple vacation, Newton's Apple and Apple card game and many other such like most irrelevant concepts not associated with either of the two.

Second query: We wanted to find winners of Templeton Award in chronological order which we could not find mainly due inadequate provision of truncation. There are innumerable such other queries which are not being correctly responded. We found different results searching in different modes in the same SE (Basic search and advanced search). There are inconsistencies in searching in two modes.

Third query: Believing that searching through directory will yield better results with more precision, therefore certain queries were processed first going to the relevant directory. But surprisingly, much less records were retrieved as compared to searching directly without invoking relevant directory. The obvious reason was that the directory was not updated.

The database connection, Index and user system features have impact on performance of SE.Indexing strategies are adopted differently by different SE. The majority will index every word on the page, other index only frequently occurring words or words occurring within certain markup tags or only the first X number of words or lines of HTML files. Stop words may, may not be applied. It may include words of very high frequency such as "web". The use of meta tags, traditionally used to improve a search by providing a common ground indexing terminology is seemingly discarded by search engines. Most of the SE are using software for automatic indexing without any role of human indexer. Some of the inquisitive questions likely to arise are:

- Has the role of human indexer been completely and successfully taken over by automatic indexing?
- Has the role of controlled vocabulary been fully and successfully taken over by free text indexing through software?
- Can concept indexing be replaced by term indexing? There are many more such questions to be answered by the information professionals.

We must seek answers to such queries. We also observe that free text indexing in certain cases scores over the controlled indexing languages. But how far the varied features of different SE are going to help the users of different disciplines. Different SE adopts different principles and techniques of indexing. Some depend entirely on software whereas some take the help of human indexer also. Therefore, due to such different approaches, variation in research output is bound to occur.

4. CONCLUSION

There had been continuous refinement in searching technique in almost all the search engines over the years. Some of the search engines also have been responding to the feedback from users. The changes are quite frequent. But problem in extracting right information still continues. There had been continued professional contribution for effective retrieval of information yet it is becoming beyond control day by day. With the increase in new SE, problem is becoming more complex. Introduction of field search by some SE is welcome step. More filtering devices are required to further narrow down the search. Adoption of concept of classification in developing directories and subdirectories may give good guidelines for better retrieval. The efforts are on to make all the search engines user friendly. But the complexity of retrieval of information can not be solved so easily. Richness of the languages particularly due to synonyms and homonyms is complicating the effectiveness of retrieval process in the absence of control vocabulary. Due to such problem some of the search engines are requisitioning the services of human indexer. But modern trend of free indexing is resulting decline in use of controlled vocabulary

This is a fact that searching information through the search engines is not easy as every user feels that he is competent enough to form query and get what ever is available. There would be a small number of web browsers who go into the details of advance features of search engines. Majority

of them depend upon basic search where retrieval of exact information is not always possible. While comparing the various features of these search engines, it is quite apparent that every search engine has its own features, scope and coverage including display format. These formats are yet to be standardised.

REFERENCES

1. Greg R, Review of Yahoo (2003) " http://www.searchengineshowdown.com/dir/yahoo/index.html" (visited 23 September,2003).
2. Greg R. Review of Altavista (2003) " http://www.searchengineshowdown.com/features/av/index.html" (visited 23 September, 2003)
3. Greg R. Review of Excite (2003)" http://www.searchengineshowdown.com/features/excite/index.html" (visited 23 September, 2003).
4. Ljosland. Mildrid. Evaluation of web search engines and the search for better ranking algorithms: http://www.aitel.hist.no/~mildrid/dring/paper/sigir.html (visited 8 August, 2003).
5. Li, Longzhaung and Shang, Yi, A new method for automatic performance comparison of search engine. World Wide Web.3:241-247, 2000, Netherlands.
6. Johnson,F.C., Griffiths, J R and Hartley, R, DEVISE: a frame work for evaluation of Internet search engines.
7. Li, Longzhuang and Shang, Yi. A new statistical methods for performance evaluation of search engines: 12th IEEE International conference on tools with artificial intelligence(ICTA'00)November 13-15,2000: http://csdl.computer.org/comp/proceedings/ictai/2000/0909/00/0909208abs.html" (visited 9 September, 2003).
8. Search engine watch (online)." http://www.searchenginewatch.com "
9. Search engine showdown (online)." http://www.searchengineshowdown.com".
10. University libraries:university at Albany, searching internet:recommended sites and search techniques.http://library.albany.edu/interest/search.html (visited 26 September).
11. Information centre & central library:Ferdowsi university of Mashhad, What are search engines " http://c-library.un.ac.ir/what_se.html "(visited 26 September, 2003).
12. PIPER :" http://piper.ntua.gr/reports/searcheng/toc.html "(visited 20 September, 2003).
13. Wei Ding and Gary Marchionini, Comparative study of web search service performance .In: proceedings of the ASIS 1996.Annual conference, October, 1996.

BRIEF BIOGRAPHY OF AUTHOR

Dr. A S Chandel is presently Professor and Head in Library and Information Science, North Eastern Hill University, Shilong. He held the position of University Librarian at Dr. Y S Parmer University of Horticulture and Forestry for about 10 years. He also served earlier at Isabella Thoburn Colege, Lucknow and Lucknow University as lecturer before joining NEHU. He had also served as Deputy Director (Information) at SAARC Agricultural Information Centre, Dhaka from 1994-1997. Has published about 30 papers including 5 books.