Dealing with Complex Queries on Environment: A Comparison of

The Performances of WWW Search Engines

By

Rajendra Kumar Thaty

Centre for Environment Education

Thaltej Tekra

Ahmedbad – 380054

E-mail: rajendra.thaty@ceeindia.org

Sukhendu Mukherjee

Centre for Environment Education

Thaltej Tekra

Ahmedbad – 380054

E-mail: library@ceeindia.org

ABSTRACT

Performance of three search engines namely Google, Altavista and MSN was evaluated with respect to complex queries drawn from one specific subject ‘Environment’. The main criterion for evaluation was precision. Google, with the highest precision value, was found to be the best of the three search engines considered.

KEYWORDS : Web search engines, performance evaluation, precision, complex queries, Google, MSN, Altavista

0. INTRODUCTION

World Wide Web (WWW or the Web) is perhaps the second most widely used application after electronic mail. The publicity and popularity WWW has gained is so great that many people equate WWW with the Internet. It has been attracting large number of users as well as information providers due to its user-friendly interface and hypermedia features. The size of WWW is not only huge it also has been increasing at a mind-boggling rate making it extremely difficult to retrieve information from it. In order to overcome this difficulty a large number of companies and institutions have developed various search aids such as catalogs, indexes, directories, and search engines for information retrieval. However, the huge number of these search aids and the differences in the search features used by them have confounded users. Are these web search engines equally good or do they differ greatly? Is it a fact that a particular search engine is good for one particular subject field and not so good for another? How do they differ from one another in performance? Is there a single Web search engine that out-performs all others in information retrieval? The current study attempts to seek answers to some of these questions.

1. SEARCH ENGINES, CLASSIFIED DIRECTORIES AND META SEARCH ENGINES

Search Engines are automated tools that employ robots to find and retrieve Web pages. These pages are then analyzed and indexed automatically. Search Engines are usually interrogated by user supplied keywords but some also provide a browsable classified directory. Examples: Google, Altavista, All the Web(Fast), MSN Search etc.

Classified Directories are tools that organize information on the Web into a browsable classification hierarchy. Most Directories also provide a keyword search facility. Information is gathered by user input (i.e. they do not employ a robot) and usually the user specifies the category to which the resource belongs and relevant keywords. These resources tend to be more accurate, because of the human intervention, but less comprehensive because of the lack of automation. Examples: Yahoo, Looksmart, Open Directory Project, Librarian’s Index to the Internet, About.com, Galaxy, Magellan etc.

Meta Search Engines are interfaces that enable to submit query to several Search Engines at once. Some merely provide an interface and others post-process the query to provide one set of results. Examples: SurfWax, ixquick, Vivisimo, Dogpile etc.

2. RELATED STUDIES

Web search engines came into existence around 1994. However, the number of evaluation studies done on Web search engines is very small. In fact, most of these studies were done in and around 1995. The earliest of studies made during 1995 by Shirky [1], Taubes[2] and Wildstrom[3] was descriptive in nature. The next notable study, an evaluation as well as description of various web search engines, was by Notess[4, 5]. Courtois, Baer, and Stark [6] evaluated the performances of about 10 different Web search aids including CUI, Harvest, Lycos, Open Text, World-Wide Web Worm, and Yahoo. Scoville [7] also surveyed a wide range of Web search engines, and suggested that Excite, InfoSeek, and Lycos should be added to one's list of favorites. Chu and Rosenthal [8] compared three Web search engines, namely, Google, Hotbot, and Lycos and evaluated them in terms of their search capabilities (e.g., Boolean logic, truncation, field search, word and phrase search) and retrieval performances (i.e., precision and response time) using sample queries drawn from real reference questions. They also proposed a methodology for evaluating other Web search engines. There are various agencies e.g. CNET (www.cnet.com) , searchenginewatch.com etc. that have been comparing and ranking Web Search Engines from time to time and publishing their results on the web.

It is observed that there is no agreement among the findings of the various studies indicated above and except for the study by Chu and Rosenthal[8], no serious attempt has been made so far to formulate a standard methodology to evaluate the performance of search engines.

3. SCOPE, OBJECTIVES AND METHODOLOGY

It is believed that people now a days depend more on Internet than on libraries to get their required information. Although this statement is highly debatable, it is true that the role of librarians has undergone a sea change with the advent of Internet. It has become imperative for librarians to manage not only printed items of information but also those available in the electronic environment. It is experienced most often that users are not aware of what search engines are best suited for the specialized subject she/he is dealing with. Further, he/she may not have the time to explore all the search features available in a search engine to refine a search query. This is where the librarians can play a highly significant role by orienting users about efficient search methodology for getting their required information from the Internet. The present study aims at providing them with a methodology to compare web search engines.

The study is based on the following assumptions:

1. Most of the time users need to formulate complex queries to get the required information on their specialized subject area
2. Some search engines are more suitable for certain subject areas than others
3. Search engines are more popular among users than catalogs, directories, web databases and indexes
4. Google, Altavista and MSN are the most popular and highest rated search engines

One of the reasons why we chose to compare search engines is that most of the Web search engines are available to users free of charge and that these free services will continue to be available to the Internet community in future. With selected search engines, we compared their search capabilities based on Complex queries, which make use of various Boolean Operators such as AND, OR, Phrases and Parenthesis. Performance was evaluated primarily with respect to precision. Recall is deliberately omitted because it is impossible to determine how many relevant items there are for a particular query in the huge and ever-changing Web system.

The queries were confined only to one specific subject area e.g. Environment. The aim was to compare the performance of the selected search engines with respect to this specific subject area. We have also confined ourselves to three search engines namely Google, Altavista and MSN. The comparison is only elementary and not comprehensive.

4. SELECTED SEARCH ENGINES AND THEIR FEATURES

Some of the useful features of the selected Search Engines are compared in the following table. While ‘y’ indicates that the feature is supported ‘n’ indicates that it is not.

FEATURES		GOOGLE	MSN	ALTAVISTA
	Fully Indexed	y	y	y
Search Options	Boolean	y	y	y
	Natural Language	y	y	y
	Phrase matching	y	y	y
	Assumes AND	y	y	y
	AND , OR, NOT	y	y	y
	Field searching: Title, URL, text, etc.	y	y	y
	Truncation	y	y	n
	Adjacency (Phrase)	y	y	y
	Proximity	n	y	y
	Case Sensitivity	n	partially	y
	Classified Directory	y	y	n
Presentation of results	Abstract	n	y	y
	Relevance Score			

		y	n	n
	Date	y	n	y
Other Sources sounds,images,pdf files etc.		y	y	y

5. SAMPLE QUERIES AND THE TEST ENVIRONMENT

All the five search queries were extracted from real reference questions handled by the librarian at the Centre for Environment Education, Ahmedabad. Complex queries on various environmental issues were selected to test the performance of the selected search engines with respect to areas related to Environment. Simple queries were not considered because the number of results would be unmanageably large.

5.1 Reference Queries

1: Wild Biodiversity: Strategies, Actions for in situ conservation

2: International Guidelines on Environmental Management and Reporting for the Finance Sector - Sustainability Reporting

3: Removal of heavy metals from wastewater of thermal power station by water-hyacinths

4: Potential guidelines for conducting and reporting environmental research: Quantitative methods of inquiry

5: Effect of human genetic engineering on sustainable Society

As can been seen, all queries are complex and require the use of Boolean as well as other operators.

5.2 Search Queries

According to the syntax used in the selected search engines, three separate search queries were constructed for each of the reference questions listed above. The terms and characters listed after the name of each search engine are the queries actually typed in during the searches. As can be seen the queries for Altavista and MSN are identical.

#1. Google ("wild biodiversity" +strategies +action +"in situ conservation") +("protected area" OR community OR institutional OR urban OR threats OR degradation OR conflicts)

Altavista: (wild biodiversity AND strategies AND action AND in situ conservation) AND (protected area OR community OR institutional OR urban OR threats OR degradation OR conflicts)

MSN: ("wild biodiversity" AND strategies AND action AND "in situ conservation") AND ("protected area" OR community OR institutional OR urban OR threats OR degradation OR conflicts)

#2. Google : (“Sustainable Reporting” + “Environmental Management”) +(“International guidelines” OR guidelines OR finance OR sustainability OR Reporting)

Altavista: (Sustainable Reporting AND Environmental Management) AND (International guidelines OR guidelines OR finance OR sustainability OR Reporting)

MSN: ("Sustainable Reporting" AND "Environmental Management") AND ("International guidelines" OR guidelines OR finance OR sustainability OR Reporting)

#3. Google: (wastewater +"heavy metals") +(water-hyacinths OR effluent OR flyash OR absorption OR ashbund OR "Eichho crassipes")

Altavista: (wastewater AND heavy metals) AND (water-hyacinths OR effluent OR flyash OR absorption OR ashbund OR Eichho crassipes)

MSN: (wastewater AND "heavy metals") AND (water-hyacinths OR effluent OR flyash OR absorption OR ashbund OR "Eichho crassipes")

#4. Google ("Environmental Education" AND Research) AND (guidelines OR conducting OR reporting OR Communicating OR "Quantitative methods" OR inquiry)

Altavista: (Environmental Education AND Research) AND (guidelines OR conducting OR reporting OR Communicating OR Quantitative methods OR inquiry)

MSN: ("Environmental Education" AND Research) AND (guidelines OR conducting OR reporting OR Communicating OR "Quantitative methods" OR inquiry)

#5. Google : ("human genetic engineering" +" Sustainable Society") +(biotechnology OR "human genome" OR environmental OR "modified humans" OR "human rights" OR bio-ethics OR diseases OR eugenics OR cloning OR "human biotech industry")

Altavista: (human genetic engineering AND Sustainable Society) AND (biotechnology OR human genome OR environmental OR modified humans OR human rights OR bio-ethics OR diseases OR eugenics OR cloning OR human biotech industry)

MSN: ("human genetic engineering" AND " Sustainable Society") AND (biotechnology OR "human genome" OR environmental OR "modified humans" OR "human rights" OR bio-ethics OR diseases OR eugenics OR cloning OR "human biotech industry")

5.3 The Test Environment

Simple search was used for Google. But advanced search was used for MSN and Altavista as separate options for Boolean search was available there. The most detailed option available was favored since this option would provide more information for evaluation.

Only the first 15 Web records retrieved for each query were considered for evaluation. As all the selected search engines display results in descending order of relevance we assumed that this would not affect the validity of our study.

6. PERFORMANCE EVALUATION

6.1 Evaluation Criteria

Internet is essentially an Information Storage and Retrieval System characterized by its enormous size, hypermedia structure, and distributed architecture. According to Lancaster and Fayen[9] there are six criteria for assessing the performance of information retrieval systems. They are: 1) Coverage, 2) Recall, 3) Precision, 4) Response time, 5) User effort, and 6) Form of output.

Although the above criteria are still valid in spite of remarkable developments in the field of Information Technology, we decided to restrict our study only to ‘precision’. Recall was not considered as it is difficult to determine because nobody knows the total number of websites available in WWW with respect to a particular query. Response time depends on several factors such as the type of hardware used and the time of the day during which search is made. Although no significant difference was observed, in general Altavista seemed slightly better in terms of Response Time than MSN and Google. All the three search engines have user-friendly search interfaces with almost similar features. As mentioned above all the three search engines support almost all types of searches. A little bit of effort on the part of a novice user is required to understand all the search techniques available before he/she is able to formulate complex queries efficiently. Adequate facilities are available in all the three search engines to regulate the presentation of the results. Users are given the choices of viewing 10, 20, 30, or 40 search results a time. In addition, each search result can be displayed using the summary, standard, or detailed format.

6.2 Precision of Search Results

	Google			**Altavista**			**MSN**		
Query	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
1	62	11	0.73	29	4	0.26	8	2	0.25
2	124	13	0.86	44	5	0.33	16	5	0.33
3	24300	15	1.00	11639	8	0.53	6511	4	0.26
4	108000	14	0.93	47124	7	0.46	36151	4	0.26
5	9	7	0.77	6	4	0.66	6	3	0.50
Mean	26499	12	0.85	11768	5.6	0.45	8538	3.6	0.32
%	-	-	85	-	-	45	-	-	32

A: Total number of hits;

B: Number of relevant hits (in the first 15 hits if the total number of hits > 15);

C: Precision Value

As shown in the above table, Google returns fairly higher number of sites than Altavista and MSN. While checking for the relevancy only the first 15 hits were considered. Precision (C) was calculated using the following formula:

C = B/(A or 15 whichever is less)

The mean precision for Google (85%) was observed to be much, much higher than that for Altavista (45%) and MSN (32%). In fact almost all of the first 15 sites returned by Google for each query were found to be relevant.

6.3 Common Hits

Query	Google/Altavista	Google/MSN	Altavista/MSN
1	1	1	2
2	1	2	4
3	0	0	3
4	0	0	1
5	5	3	3
Mean per query	1.4	1.2	2.6

It is evident from the above table that there were more common sites returned by Altavista and MSN

6.4 Other Findings

1. Although Google does not take more than 10 words as query, it was found to be the best search engine for complex queries. Contrary to our assumption basing on the ranking made by various agencies Altavista and MSN fell far behind google in returning relevant sites.
2. Google and Yahoo returned almost the same sites. The format of output was also identical. This was perhaps because Yahoo makes use of the Google search engine since October 2002.
3. Google has the cache function that is very useful. If a link you need no longer exists on the Web, Google shows the last saved version of the page when you click Show matches.
4. Google returns fewest number of broken links.
5. The output of Google are extracted from the original web documents. It only returns some of the broken sentences (sometimes unintelligible) containing the search terms. MSN and Altavista output the first few complete sentences from the original web document.

7. CONCLUDING REMARKS

Web search engines no doubt are different, in various aspects, such as search features, interfaces, documentation, user efforts etc.. Although it is debatable we believe that the performance of search engines varies for different subject areas. We have made a preliminary attempt here to study this taking one subject area namely ‘Environment’ as a sample subject. We have concluded that for this particular case Google is far better than the other two search engines. More studies for other subject areas are needed to verify our claim.

There is a need to include more search engines in the study. Further it is required to have a standard methodology for evaluating search engines in relation to complex search queries.

REFERENCES

[1] Shirky, Clay. “Finding needles in haystacks”. *Netguide*, (October 1995), pp-87-90.

[2] Taubes, Gary. “Indexing the Internet.” *Science*, 269, (8 September 1995). pp-1354-1356.

[3] Wildstrom, Stephen H. "Feeling your web around the Web." *Business Week*, 11 September 1995, pp-22.

[4] Notess, Greg R. "Searching the World-Wide Web: Lycos, WebCrawler and More." *Online*, 19(4), (July/August 1995).pp. 48-53.

[5] Notess, Greg R. "The InfoSeek Databases." *Database*, (August/September 1995). pp-85-87.

[6] Courtois, Martin P., Baer, William M., and Stark, Marcella. "Cool tools for searching the Web: A performance evaluation." *Online*, 19(6), (November/December 1995).pp. 14-32.

[7] Scoville, Richard. "Special report: Find it on the net!" *PC World*, (January 1996). also available at <http://www.lycos.com.>

[8] Chu, Heting and Rosenthal, Marilin. "Search Engine for World Wide Web: A Comparative Study and Evaluation Methodology" *ASIS*, 1998

[9] Lancaster, F.W. and Fayen, E.G. "Information Retrieval On-Line." Los Angeles: Melville Publishing Co., 1973. Chapter 6.

-

BRIEF BIOGRAPHY OF AUTHORS

Rajendra Kumar Thaty currently serves the Centre for Environment Education, Ahmedabad as Librarian. A science graduate in Physics, he received his MLIS degree from Sambalpur University. With over 16 years of working experience in the field of Librarianship and Documentation, his interest areas include Bibliographic Database Management and Training.

Sukhendu Mukherjee currently serves the Centre for Environment Education, Ahmedabad as a Project Associate under the ENVIS programme. He received his MCA degree from IGNOU. With around 3 years of experience in the field of computer application, his interest areas include Web development, Database Management Systems and Networking