Looking for the Eye of the Needle in a Haystack: Retrieving Information from the Web

Jamshid Beheshti, Ph.D.
Director

Graduate School of Library & Information Studies
McGill University
3459 – McTavish Street
Montreal, Quebec, H3A 1Y1
Canada

beheshti@gslis.lan.mcgill.ca
www.gslis.mcgill.ca
voice: (514) 398-4204 ext. 3366#
fax: (514) 398-7193

I. ABSTRACT

This paper surveys the major search engines and examines their retrieval capabilities to investigate information recall versus precision. The main objective is to rank the search engines for precision. The study is conducted under the assumption that since more than one billion pages are now available on the Web, retrieving specific information in a particular context, such as e-commerce, is more problematic than retrieving broad and general information. One example is used as a case study to determine the effectiveness of the search engines. Information about a commercial multimedia CD-ROM product is searched using various features of the engines and the results are compared.

Keywords: Web, Information Retrieval, Ranking, Relevance

I. INTRODUCTION

It is estimated that as of January 2000, one billion Web pages are available online through the Internet. It is also estimated that every day an additional one million pages are added to the Web. [2] This huge depository of information, regardless of the quality, authenticity, and authority, poses an enormous challenge to users who are seeking specific information. It is equally formidable for corporations and other organizations to post their products, which should easily be located by potential consumers. A varied array of search engines and directories are now available for expert and novice searchers alike to assist them to find the needed information and products. These range from comprehensive directories such as Open Directory Project [13] to advanced search engines such as Altavista.

Despite all the sophisticated search tools, bringing the user and the information together (or the product and the consumer) remains a difficult and serendipitous task. A typical search on any engine or directory, which have indexed millions of pages, produces a very high number of hits (recall) with a very low precision rate. While in the traditional retrieval situations, such as searching large bibliographic databases, experts have attempted to create a balance between recall and precision, the information overload on the Web and the nature of search tools prohibits any controlled approach to information retrieval.

The purpose of this paper is to demonstrate the inconstancies in searching for an specific product on the Web. A case study methodology is used to investigate the results of searching for a product on various search engines.

II. RECENT RESEARCH

The most formidable task for a search engine is locating relevant pages among the millions of sites, based solely on a term or a phrase entered by the user. As one search engine designer states; "Imagine walking up to a librarian and saying, 'travel.' They're going to look at you with a blank face." (How Search Engines Rank Web Pages) The search engines, unlike the librarian, do not have the ability to ask questions from the users to determine the context and the nature of their enquiries. The best solution that search engine designers can provide is based on relevance ranking. To determine relevancy, designers rely on algorithms that include factors such as number, location and frequency of terms matched, metadata, and URLs. These algorithms, however, are closely guarded secrets and do not reveal how each company is actually ranking the outcome of a query. [8]

Courtois and Berry tested five major search engines to determine the effectiveness of their relevancy and ranking techniques. [1] They used twelve multi-term search statements, which were derived from actual reference questions in various libraries or previous studies. Different search techniques ranging from simple to advance were employed to test the engines. They concluded that search results vary significantly depending on the topic of enquiry. They also observed that "ranking was consistently more reliable within the first 20 hits than within the first 100 hits. In a few cases, however, the first 100 hits produced better scores than the first 20 hits."

Sirapyan in a PC Magazine test of search sites used more than 50 queries ranging from single words to natural-language questions. [17] The default settings of each site were used to conduct the searches and the first ten hits were evaluated. She concluded that "none of the search engines out there today are perfect, but using the right one at the right time can make all the difference."

A more rigours study was conducted in 1998 by Naisos, et al, the results of which was reported in a sixty-page document. [6] The purpose of the research, part of a project for the European Commission, was to examine the problems associated with information retrieval and the criteria by which search engine performance should be evaluated. Twenty queries were searched on six engines and the first ten hits were evaluated. The questions were divided into three categories: exact phrases using quotations, natural

language without any operators or quotations, and terms with embedded search features such as Boolean and proximity operators. One general conclusion of the report was that "search engines basically differentiate in efficiency, with respect to different types of queries." The investigators recommended that the choice of search engine should depend on the type and the nature of the query.

Several other recent studies[4][5][16], which perhaps have employed less vigorous methodologies, but are nevertheless very useful, demonstrate the strengths and weakness of search engines. Their ratings of the search engines illustrate the general and broad conclusions reached by previous studies that the type of query should determine the choice of the search engine.

In this study, unlike previous research, general or encyclopaedic queries have not been utilized to rank the search engines. The methodology is based on using a specific product and formulating queries, which might be posed by consumers who are potentially interested in finding more information about that product. Searching for products has gained prominence in the recent years, since 83% of sites contain commercial content [3] and 34% of popular searches are for products as oppose to 16% for places and 6% for recent news. [15]

III. METHODOLOGY

A. a) A case study

Treasures of Islam: Art and Design in Islamic Manuscripts was conceived to present the rich Islamic collections at McGill University to other institutions, researchers and the general public. Islamic bookmaking has been a source of fascination for many centuries. Authors, calligraphers, miniature painters, illuminators, and bookbinders collaborate to produce what are often regarded as works of arts. The materials included in the CD-ROM were collected from the Rare and Special Collections Division of McGill Libraries under the supervision of the curator and the Islamic Studies Librarian. They were arranged in four separate categories: Calligraphy, Miniature, Bindings, and Manuscripts. Calligraphy includes a collection of 40 samples of the works of different Islamic scribes from the 9th to the 19th centuries. Miniature section consists of 42 miniature paintings representing a variety of styles and sources, selected for their beauty and elegance. Many are taken from different Persian manuscripts dated from the 13th to the early 20th centuries. Sixteen lacquer bindings are included in the CD-ROM, chosen on the basis of their rich and impressive decorations, originating in Persia or Kashmir. They date from the 16th to the 20th centuries. The final section, Manuscripts, is a collection of eight Arabic manuscripts, included in their entirety, representing prayer books, poems and a manuscript about calligraphy.

The CD-ROM was packaged and prepared for marketing in December 1999. It was available for sale in January 2000 through McGill Systems Inc. (MSI), and was registered with Amazon.com in early March 2000. A Web site was created in December on the Graduate School of Library and Information Studies' Web server (GSLIS), (www.gslis.mcgill.ca/Treasures), which consisted of three pages, one for each language on the CD: English, French, and Arabic. The pages were identical and each contained four visual samples and a more extensive version of the above description. Each image is also available on a separate page. The site contains meta tags to describe the CD and was initially registered with Yahoo!, Infoseek, and Altavista.

MSI also created a Web site for the CD-ROM independently from GSLIS, which contained two pages, neither of which used meta tags or were registered with any search engines or directories. (http://musicm.mcgill.ca/msi/http/toi.html). In April 2000, MSI completely overhauled its main gateway with a new address (http://www.cs.mcgill.ca/~gwen/), using Java scripts and meta tags. The tags, however, do not have any information regarding *Treasures of Islam.*

B. b) Procedures

A total of nineteen search engines were chosen for this study. The choices were based on reports from Nielsen [7], Search Engine Showdown [11], PC Data Online [14], and Nua [12]. Although the reports were written at different times with dissimilar ranking criteria, together they provided a relatively comprehensive list of engines for the purposes of this study. Meta-search engines and directories were excluded from this study.

During April 2000, four searches were performed on each engine, consisting of the following phrases:

1. "Treasures of Islam" - the user knows the exact title and simply needs to find more information about the product.
2. treasures of Islam - the user does not know the exact title or is unaware of the use of quotations or capitalization.
3. Islamic manuscript - the user is interested or needs to locate products related to the subject.
4. Islam art - A relatively more sophisticated searcher is interested or needs to find products related to the subject.

The rational behind choosing these particular phrases was based on two factors. First, all the terms appear in the title of the CD-ROM. The first, second and third search terms are exact phrases embedded in the title, while the fourth search phrase appears both as a phrase, as well as two independent words in the title. Second, the words and phrases represent the potential terms, for which the majority of users might search the major sites.

Several assumptions had to be made in conducting this study. First, it is assumed that the users are novice searchers who are unfamiliar with advanced features of the search engines, or are not inclined to use them. Hence, they will only use the default settings of each engine. Second, while attempts were made to register the GSLIS Web site with several engines and directories, the number of engines, which actually accepted the site, is unknown.

IV. RESULTS

Table 1 summarises the results of all the searches. The first number represents the position or ranking of the hit in the results set display. The lower this number is the more relevant the page has been ranked by the search engine and the more visible it is to the consumer. The second number is the total number of hits (recall) for the specific phrase. If the number of hits exceeded 200, the first one hundred was examined manually to locate the *Treasures of Islam* CD-ROM. The assumption is that users do not browse beyond the first fifty to one hundred hits for finding the desired information. Some search engines do not report the total number of hits, in which case a '+' sign is used to designate more hits. All the hits are from the GSLIS site, unless otherwise noted.

The table is organized by the popularity of the sites based on the *PC Data Online reports*, which is monthly and current. Many other evaluative reports are either dated (1997 to 1999) [19] or have weekly reporting frequency [7], which are not suitable for the purposes of this study. In addition, PC Data Online tracks the Internet activity of more than 100,000 home users and reports on a range of parameters. Although many problems have been cited with using any rating mechanism for search engine usage [20], it is assumed that the PC Data Online reports represent the most suitable ratings for our purposes.

As the results in Table 1 indicate, while twelve search engines returned at least one hit, seven did not find any of the *Treasures of Islam* pages. Those users who know the exact title of the CD-ROM can easily locate the Web pages for this product. With one exception, all the remaining eleven search engines rank the CD pages among the top ten hits and on the first page of the result set. In the case of Fast search engine, the first eight hits out of 58 are *Treasure* pages. Fast Search and Transfer company has moved

swiftly to capture a sizeable market. It first appeared in May 1999 and by January 2000 had already indexed 300 million pages. [10]

When the quotation marks are not used, as in the case of the second query, again Fast ranks *Treasures* among the top seven hits out of a total of 7914 pages. Three other engines Go, Altavista, and Northern light also rank the CD pages higher than any other pages they locate. Go is particularly useful in finding the *Treasures* pages as it reports a total of just under three million hits for this query. Disney's Go Network portal acquired Infoseek search engine recently. Infoseek ranks the pages based on location and frequency of the words within the document, as well as the rarity of the search terms. [9]

Another scenario presented with the third query assumes that users are interested in finding products dealing with Islamic manuscripts. Only three search engines can locate the *Treasures* pages under this heading, and only Go ranks one page among the top ten. Northern light also locates one page and ranks it twelfth among more than eleven thousand hits. After Fast, Northern light has the second largest index among the engines and uses ranking to sort the search results.

If users enter two words, *Islam* and *art* in the hope of finding the *Treasures of Islam*, only two search engines, Go and Northern Light shall return a single page each. Go retrieves the page and ranks it ninth among more than 3.5 million pages.

Netscape, ranked tenth in the PC Data Online reports, relies on its partners, Google and Open Directory, to provide searching capability to its site. It did not find any of the *Treasures* pages for any of the queries. Similarly, newcomers LookSmart, and AskJeeves did not display any of the relevant pages.

Table 1. Search results.

	"Treasures of Islam"	treasures of Islam	Islamic manuscript	Islam art
Yahoo! http://www.yahoo.com	1 : 15	6 : 2406	0 : 3	0 : 65
AOL http://search.aol.com	1 : 1	9 : 204	0 : 202	0 : 1416
MSN http://search.msn.com	2 : 13	7 : 1889	0 : 2336	0 : 25180
Go http://www.go.com	1 : 18 2 : 18*	1 : 28622994 2 : 28622994*	7 : 147226* 23 : 147226	9 : 3550110* 0 : 3550110
Excite http://www.excite.com	0 : 21	4 : 200 +	82 : 200+	0:200+
Altavista http://www.altavista.com	1 : 25 2 : 25 3 : 25*	1 : 845680 2 : 845680	0 : 631405	0 : 782315
Iwon http://www.iwon.com	1 : 19 3 : 19**	7 : 562200 9 : 562200**	0 : 338354	0 : 5960694
Snap http://www.snap.com	1 : 14 3 : 14**	6 : 66 10 : 66**	0 : 200+	0 : 38
Goto http://www.goto.com	1 : 7**	0 : 200+	0 : 200+	0 : 200+
Northern Light http://www.northernlight.com	1 : 1487 2 : 1487	1 : 10383 2 : 10383	12 : 11527*	17 : 79721*
Hotbot http://hotbot.lycos.com/	1 : 13**	6 : 1000+**	0:1000+	0:10000+
FAST http://www.ussc.alltheweb.com/	1: 58 2 3* 4 5 6 7* 8	1 : 7914* 2 3 4 5 6 7*	0:8181	0:91671

*MSI old pages
**MSI new site

Netscape (http://www.netscape.com), Lycos (http://www.lycos.com), About (http://www.about.com), AskJeeves (http://www.ask.com), Direct Hit (http://affiliate.directhit.com/), LookSmart (http://www.looksmart.com), and InfoSpace (http://www.infospace.com) yielded zero hits in all categories.

V. DISCUSSION AND CONCLUSION

Two independent Web sites were created to publicize and promote a multimedia CD-ROM product. Another site was developed at a later date, which contained information about the CD-ROM, but was not exclusively about the product. Each site had different characteristics and only one was manually registered with three search engines. Four queries were formulated to simulate potential consumers' search patterns for the CD-ROM and were posed to nineteen search engines. The results indicate that while the majority of search engines are able to locate one or more pages for the product based on an exact title search, only a few find the relevant pages based on queries that contain disparate words from the title.

The results of this study show that it is highly probable (58%) to find Web sites and pages for a product, if the exact name or title is known and if several search engines are used. There is an equal chance of finding the appropriate pages if the first words of the title are used without quotations. Based on these results, Fast search engine may be ranked as the most appropriate for known item searching, particularly in light of the fact that it indexes the most number of Web pages. For both queries, it ranked the highest number of pages related to the *Treasures* than any other engine.

The search results change dramatically when a subject matter is entered, even though the terms used appear in the title or the name of the product. In this case, the probability decreases to only 16% if all the search engines are considered and 25% if the results are confined to the twelve search engines. The probability is reduced even further if only disparate terms from the title are utilized in locating a product. In both cases, Go should be considered as the most appropriate search engine for retrieving and ranking relevant pages.

The results also demonstrate that many factors influence the indexing of a site or a page by the search engines. It has been suggested that the main factor affecting the indexing and ranking is location and frequency of keywords on the page. [18] Three sites for the CD-ROM were created with very different characteristics. Since the title of the CD-ROM appears as the title of the Web pages on two of the sites as well as in the heading sections, nine out of the twelve engines rank the pages very high in the results set for query one. Three engines, however, behave differently than this norm. Hotbot and Goto only locate the new MSI site, which uses meta tags but does not include any keywords from the CD-ROM, nor does it contain any terms related to the topic in its title or major headings. Both search engines, however, utilize Inktomi databases, which are also used by Yahoo!, AOL, MSN, and Snap. Hence, the differences among these engines cannot be attributed to the databases alone.

Excite, on the other hand, does not find any of the pages for the first query, but does find one page for the second query and one for the third one. This is a puzzling result considering that all the other engines find pages for the exact title match before they locate pages for other types of queries.

It has been suggested that linking is becoming an important factor in ranking the relevance of the search results. [2] Google uses this feature extensively and has become popular due to its precision for finding information. The *Treasures* pages have few links to them and therefore this factor could not be verified. In addition, Google along with several other search engines mentioned above have not indexed the CD-ROM pages and could not be tested for relevance ranking.

Another important factor suggested in the literature is registration of the site with the search engines. [3] While many tools are now in the market to assist Web masters in registering their products (for example *Web site Traffic Builde*r and *Web Position Gold*), manual submission is preferred. The results of this study show that while registering the GSLIS site improved the search outcome for at least three search engines, it did not affect the other engines, where the site was not formally registered. Noteworthy is also the speed with which Hotbot, Snap, and Iwon indexed the new MSI site, and the delays in indexing the new site by other search engines.

The inconsistencies of search engines in indexing and relevance ranking have been acknowledged in the literature. Little is known about how each engine finds and ranks a desired page. A similar conclusion was reached by Lawrence and Giles who stated that the search engines "do not index sites equally. The current state of search engines can be compared to a phone book which is updated irregularly, is biased toward listing more popular information, and has most of the pages ripped out." [3] The results of this study confirm these findings.

II. REFERENCES

[1] M.P. Courtois and M.W. Berry, “Results Ranking in Web Search Engines,” *Online*, May 1999. http://www.onlineinc.com/onlinemag/OL1999/courtois5.html [April 2000].

[2] “Hypersearching the Web,” Members of the Clever Project, IBM. *Scientific American*, June 1999. . www.sciam.com/1999/0699issue/0699raghavan.html [April 2000]

[3] S. Lawrence and L. Giles, "Accessibility of information on the web," *Nature*, Vol. 400, pp. 107-109, 1999.

[4] Jennifer Mack, "Want to win $1M? Don't Ask Jeeves," *ZDNet*, January, 22, 2000. http://www.zdnet.com/zdnn/stories/news/0,4586,2425612,00.html [April 2000]

[5] Michael J. Miller, "Best Search Sites on the Web," *PC Magazine*, September 20, 1999. http://www.zdnet.com/anchordesk/story/story_3871.html [April 2000]

[6] Naisos, Y., et al. *Evaluation of search engines*. Athens: National Technical University of Athens, July 1998. http://piper.ntua.gr/reports/searcheng/ [April 2000]

[7] The Nielsen/Netratings Reporter. http://www.nielsen-netratings.com/weekly.htm[April 2000]

[8] G. Notess, "On the Net: Rising Relevance in Search Engines," Online, May 1999. http://www.onlineinc.com/onlinemag/OL1999/net5.html [April 2000]

[9] G. Notess, “Review of Go (Infoseek),” Search Engine Showdown, March 3, 2000. http://www.searchengineshowdown.com/features/infoseek/review.html [April 2000]

[10] G. Notess, “Review of Fast Search,” *Search Engine Showdown*, March 3, 2000. http://www.searchengineshowdown.com/features/fast/review.html [April 2000]

[11] G. Notess, “Search Engine Features Chart,” *Search Engine Showdown*, March 3, 2000. http://www.searchengineshowdown.com/features/ [April 2000]

[12] Nua. Nua Internet Surveys. http://www.nua.ie/surveys/analysis/index.html [April 2000]

[13] Open Directory Project. http://dmoz.org/

[14] PC Data Online reports. http://www.pcdataonline.com/reports/tmSitesSingleFree.asp [April 2000]

[15] http://searchenginez.com/ratings.html

[16] C. Sherman, "Want to be a Millionaire?" *About.com Web Search Guide*, Feb. 18, 2000. http://websearch.about.com/internet/websearch/library/weekly/aa021500a.htm[April 2000]

[17] N. Sirapyan, “Search Sites,” PC Magazine, September 15, 1999. http://www.zdnet.com/pcmag/stories/reviews/0,6755,2327803,00.html [April 2000]

[18] D. Sullivan, “How Search Engines Rank Web Pages,” *Search Engine Watch*. http://www.searchenginewatch.com/webmasters/rank.html [April 2000]

[19] D. Sullivan, “Media Metrix Search Engine Ratings,” *Search Engine Watch*, March 24, 2000. http://www.searchenginewatch.com/reports/mediametrix.html [April 2000]

[20] D. Sullivan, “The Problems With Rating Services,” The Search Engine Report, April 4, 2000. http://www.searchenginewatch.com/sereport/00/04-ratings.html [April 2000]