Comparative study of Metasearch Engines performance in retrieving Library and information Science documents on the web

Fatemeh Nabavi

Abstract

Internet searching tools, Search engines and Metasearch engines are among. Metasearch engines do not have their own databases. They send user's queries simultaneously to multiple web Search engines and/or web Directories. Some of researchers and compaines that conduct the Metasearch engines, believe that these tools retrieve more relevant hits. This research was conducted to determine this idea in searching and retrieving Library and information science documents. 12 major Metaseach engins that introduced in www.searchengine watch.com at September 11, 2000 were selected. On the other hand, keywords of articles titles which selected were published in library and information science periodicals during one year (1378 = march 21, 1999 – march 20, 2000).

Keywords were classified acording to LISA classification of library and information science subjects. Then, these keywords were searched in 12 Metasearch engines. The 10 first records were considered to be evaluated. The results showed that inspite of using the best Metaseach engines, Just approximately %30 of sources are relevant. Also results indicate that C4 Metasearch engine in comparison with others, retrieves more sources in various subjects, but more relevant sources retrieve form C/ Net Metasearch engine.

Introduction

If you know where the document is placed on the web, you should only give the address to the browser. In this case, the browser communicates with server by following the address, and transfers the information to the client computer. But what will you do if you want to find the document and you don't know the address? (Han,1376,44). To find the proverbial needle in this immense haystack (or tiny fly in the web), you may use two basic approaches: a search engine or a subject guide. Subject guides are fine for browsing general topics, but for specific information use a search engine. (Introduction to search engines, 1999). Each of search engines has its unique content and presents a unique interface, requires a unique set of rules for searching and displays search results differently. To exhaust a search a search, one often has to use several of them and has to

be familiar with the different interfaces and searching rules.

It would be highly desirable to have a central place with a uniform interface, where a query can be entered and the search can be conducted simultaneously in as many search engines and directories as necessary, and search results can be brought back and displayed in a consistent format. Tools with these features have come to be called Meta – search engines. (Liu, 1999)

At the first, it seems that using metasearch engines provide all the information that user needs. But should one use a metasearch engine instead of an indivisual search engine? There is no definitive answer to this question. Much depends on what one is seeking. For a specific, obscure search term, I would recommend starting with a metasearch engine, as it will search many sites at the same time, thus saving you a lot of time and making your search less tendious. On the other hand, if you are reasonably confident that any major search engine will return the page you are looking for, starting with an individual search engine would be recommended. (Liu, 1999). The study of searching tools performance on the internet and introducing the best, was always considered by the researchers. Also about metasearch engine some studies was conducted. Robert kiley believes that: “supporters of metasearch engines argue that by utilising multiple search engines the potential for finding information is much greater. Although this is indeed true, the downside of this question is that for too many resources are identified and, as you start to sift the results, you naturally find numerous duplicates”. He continued: “The only occasion I can recommend using them is if you are searching for something that is very obscure or rare, and you don’t want to manually visit a number of search tools and rekey your search.”(Kiley, 2000, 30). Ripman and karlson at 1999 did research about 16 metasearch engines and introduced 5 of them as the majors. They were: Bytesearch, Mamma, MetaCrawler, Profusion, and Savvy Search. (Asadi, 1379, 61). Tomaiuolo at 1999 in the article “Are metasearches better searches?”, compared 4 metasearch engines (Dogpile, Cyber411, Internet Sleuth, MetaCrawler) with 2 search engines (Altavista, and HotBot). He said “Although some metasearch engines certainly interface better with certain individual engines as compared to others, this investigation illustrates that metasearch engines work relatively”. (Tomaiuolo, 1999, 32). Metasearch engines may find the same information that the single search engines find, But because no engine has precisely the same coverage as another, the searcher using a metasearch engine efficiently maximizes the potential for locating relevant information. (Tomaiuolo,1999,34). This research was conducted using 12 leading metasearch engines to determine the suitable of them for finding library and information scenice (LIS) documents on the web.

The purposes of study

Comparative study of metasearch engines in order to finding a metasearch engine which retrieves the most relevant documents about library and information science (LIS)

Hypothesises

1. There is significant difference between total retrieval hits from metasearch engines in various subjects.
2. There is significant difference between mean ranks of relevancy in retrieval hits from each metasearch engines in various subjects.

Questioning

1. In which classified subjects of LIS, each of metasearch engines retrieve more hits on the web?
2. In which classified subjects of LIS, metasearch engines have more dealer of relevant hits on the web?
3. In which classified subjects of LIS, metasearch engines have the best performance (by calculating mean ranks)?

Methodology of research and data collecting

This research was conducted as a survey – analytic research. 12 major metasearch engines that introduced in www.searcheginewatch.com at September 11, 2000 were selected. 6 selected LIS periodicals were published in IRAN during the year (1378 = March 21, 1999 – March 20, 2000). They were surveyed to determine the articles that were originally in English and translated. 47 titles were selected. Then some keywords were appointed in natural language. They were 49 keywords. After that, these keywords were classified acording to LISA classification of library and Information science subjects. Total metasearch engines were 12. But finally decreased to 9. Because one of them (Terespondo) was in spanish, the second was shut down (InferenceFind) and the third also was not accessible and the reason isn't clear to researchers. Of course we did some correspondences with owners but we couldn't get consequence. Nine metasearch engines being studied are listed as follow:

1- C4 (www.c4.com)
2- C/Net Search (www.savvysearch.com)
3- Dogpile (www.dopile.com)
4- Ixquick (www.ixquick.com)
5- Go2Net (www.go2net.com)
6- Mamma (www.mamma.com)
7- Profusion (www.profusion.com)
8- QuickBrowse (www.quickbrowse.com)
9- Surfwax (www.surfwax.com)

Finally, searches wre done in 9 metasearch engines using keuwords. To do this, default settings of metasearch engines were considered. Simultaneously the sources were evaluated. Evaluation was as will be mentioned. The first 10 retrieval hits were considered in every metasearch engines, and we surveyed one by one. In the 2 of metasearches (Dogpile, and Quick browse) retrieval hits were ordered and displayed acording to search engines and directories. So, by using statistical methods, we made random chosen from all

retrieval hits in this 2 metasearches. Because we couldn't judge correctly in the result pages, and sometimes there is no explanations about the retrieval hits (i.e. surfwax), so we clicked on the retrieval hits titles and after loading the main pages, evaluated them. In the ckeck list 6 columns were determined: completely relevant, relevant, nearly relevant, Irrelevant, errors and duplicated links.

Data Analysis Method:

Datas were analyzed by using SPSS. The application tests were chi-square and kruskal–wallis. when kruskal wallis was significant, sheffe test was used for finding differences between groups (subjects). It's necessary to mention that kruskal-wallis was used to calculate mean ranks in every table. Because of method that was used in data entry every group that scored minimum mean rank, showed the better performance in comparison.

Research Results

diag. 1 shows the number of articles that were selected from LIS periodicals during one year. As you see, the payam-e-ketabkaneh has the largest number of articles (18 titles) and ketabdari has just 1.

1. Etela Resani (Information Science)
2. Pzhouheshname-e-Etala Resani (Information Science Research Bulletin)
3. Payam-e-Ketabkhane (Library Messenger)
4. Faslname-e-Ketab (Book Quarterly)
5. Ketabdari (Library Science)
6. Ketabdari va Etela Resani (Library & Information Science)

As you can see table 1. "*Comunication and information technology*" subject has the largest percentage of articles (%25.6). In 3 subjects no article was found. They are "*Bibliographic controls*", "*Bibliographic records*", and "*Reading*". The smallest percentage of translated articles belongs to "*knowledge & learning*" group (%2.1). In the other words, about half of total articles were about technology and computer.

TABLE 1

Subjects Codes	Subjects	Number of Titles	Percentage
1	Library and Information Science	3	6.3
2	Library Profession	2	4.3
3	Libraries & Documents Centers	3	6.3
4	Using Libraries & Users	2	4.3
5	Materials & Sources	3	6.3
6	Library Organization	4	8.5
8	Library Technology	8	17
9	Technical Services	2	4.3
10	Information Communications	2	4.3
11	Bibliographic Controls	0	0
12	Bibliographic Records	0	0
13	Saving & Retriving Computer Information	3	6.3
14	Comunication and Information Technology	12	25.6
15	Reading	0	0
16	Medias	2	4.3
17	knowledge & Learning	1	2.1
	Total	47	100

For testing the first hypothesis chi-square test was used. It's considered =%5 and df=100. It illustrates that there is significant difference between total retrieval hits from metasearch engines in various subjects (P>124.3). So the first hypothesis was confirmed. Totally, C4 metasearch engine has the largest percentage of retrieval hits in various subject (%24.4) and Go2Net metasearch engine has the smallest (%2.6).

For testing the 2nd hypothesis, data analysis was done by using SPSS, and kruskal–wallis test. Except in 3 of metasearches (Go2Net, QuickBrowse, and surfwax), there is significant difference between mean ranks of relevancy in retrievaling hits from metasearches in various subjects.

Fig.2 illustrates that all of metasearches had largest percentage of retrieval hits in "*Communication and information technology*" subject. and this is the answer of 1st question.

Fig. 3 answers to 2nd question. It shows that 8 of metasearches had the largest

percentage of relevant retrieval hits in "*Communication and information technology*" subject. Just QuickBrowse had the largest percentage in "*Library technology*" subject.

Table 2 indicates that except of 2 metasearches (Mamma, and QuickBrowse) the others had the best performance in "*technical services*" subject. Mamma in *"Libraries and documents centers*" and QuickBrowse in *"information communications*" had the best performance.

TABLE 2.

Metasearch Engines	Subject
C4	technical services
C/Net	technical services
Dogpile	technical services
Go2Net	technical services
Ixquick	technical services
Mamma	Libraries and documents centers
Profusion	technical services
Quick Browse	information communications
Surfax	technical services

When all the retrieval hits from metasearch engines in various subjects were surveyed, it emerged that totally 24157 hits were retrieved from them. The largest percentage of hits were in "*communication and information technology*" subject (%30.3). This finding has coordination with population of study. So, it expected that the smallest percentage would be in "*knowledge & Learning*". But that isn't that is in "*using Libraries and users*" subject (% 1.3).

Among 730 relevant hits which were retrieved from metasearches in various subjects, the largest percentage were retrieved from Ixquick (%15.6), and the smallest from QuickBrowse (%3.3).

If we imagine that the retrieval hits which classified in completely relevant and relevant groups are useful for users, we can say except of QuickBrowse, in the other metasearches, the largest percentage of these hits beloged to *"technical services*" subject. They are beloged to C/Net (%27.2). This study has emerged that, by using the best metasearch engines, only about %30 of the retrieval hits would be relevant. Totally all of metasearch engines had the largest percentage in irrelevant hits.

The mean ranks of relevancy in retrieval hits were determined. Only in 3 of metasearches there wasn't significant difference between mean ranks in various subjects ($P>0.05$). In the other 6, for determining significant difference between subjects, sheffe test

was used. In comparison of mean ranks of relevancy retrieval hits, these results were obtained. In C4 there is significant difference between "*Materials & sources*" with 6 other subjects. They are *"Library and Information science*", "*Library technology*", "*Technical services*", "*Saving & retrieving computer information*", "*communication and information technology*" and "*Medias*". Also there is significant difference between "*Library organizations*" with "*Communication and information technology*".

In C/Net: "*Library & Information science*" with "*Material & Sources*", and "*Library technology*". Also "*technical services*" with 9 subjects. In the other words the mean rank of "*technical services*" didn't have significant difference with only 3 subjects: "*Library & Information science*", "*Using Libraries and users*", "*knowledge & Learning*"

In Dogpile: "*Technical services*" with 5 subjects, "*Library profession*", *"Libraries and document centers*", "*Materials & sources*" , "*Library technology*".

In Ixquick: "*Using Libraries and user*" with 5 subjects, "*Library & Information science*", "*Library profession*", "*Library technology*", "*Technical services*", and "*Communication and information technology*". Also "*Technical services*" with 5 subjects, *"Libraries & document centers*", "*Library organizations*", "*Library technology*", "*Information communication*", and "*Medias*".

In Mamma: "*Materials & Sources*" with 3 subjects, "*Libraries & document centers*", "*Technical services*", "*Saving & retrieving computer information*".

In Profusion: "*Libraries & document centers*" with "*Library technology*". Also "*Technical services*" with 5 subjects, "*Library & Information science*", "*Libraries & document centers*", "*Materials & sources*", "*Communication & Information technology*", and "*knowledge & Learning*".

Totally, all of metasearch engines that significant difference was found by kruskel–wallis in them, the mean rank of "*Technical services*" had significant difference with other subjects at least in 1 of them. The next subject was "*Library technology*" that only in Mamma didn't have significant difference with other subjects. In the other hand, "*knowledge & Learning*" had only in 1 case (Profusion) significant difference with "*Technical services*".

Apart from being significant or no, performance of metasearch engines in various subjects were very different. In 4 subjects: "*Library & Information science*", "*Using Libraries & users*", "*Technical services*", and "*Knowledge & Learning*", The C/Net had the best performance. C/Net was the only metasearch that didn't have the average of minimum performance in any subjects. The other metasearches had the best performance in one or several subjects and the worst in other(s).

Conclusion

Metasearch engines have very different performance in retrieving Library and Information science documents on the web. 2 of them (Dogpile, and QuickBrowse) display the results acording to search engines and/or directories. The others in result page, mention the search engines and/or directories that they have retrieved hits from them. 3 of

metasearches (C/Net, Ixquick, and Profusion) refer in retrieval hits to 2 or 3 search engines and/or directories. So, users expect at least these metasearches don't retrieve duplicate link. But this study determined that they do it. Each of mentioned metasearch engines retrieved orderly %2.5, %3.2 and %2.2 duplicate links. As mentioned Robert kiley metasearches find numerous duplicates.

C4 retrieves the largest percentage of hits in various subjects, but the largest percentage of relevant hits in various subjects belongs to C/ Net.

Recommendations

Acording to this study, QuickBrowse is not recommended for searching LIS documents on the web in various subjects. In the others, these items deserve notice:

1. For "*Library & Information science*", "*Using Libraries & users*", "*technical services*", and "*knowledge & Learning*": C/Net
2. For "*Library profession*": Go 2 Net
3. For "*Library & document centers*" and "*saving & retrieving computer information*": Mamma
4. For "*Materials & sources*" and *"Library technology*": Ixquick
5. For *"Library organization*": Dogpile
6. For "*Information communication*": Profusion
7. For "*Communication & Information technology*" and "*Medias*": C4

Persian Resources

ÇÓÏí ÑßÇäí¡ ÝÇØãå. (1379). *ÌÓÊÌæ ÏÑ æÈ: ãæÊæÑåÇí ÌÓÊÌæ æ ÇÈÑãæÊæÑåÇí ÌÓÊÌæ*. ***æÈ.*** - ***.ãÇåäÇãå ÂãæÒÔí ŽæåÔí æ ÇØáÇÚýÑÓÇäí ãÌÊãÚ Ýäí ÊåÑÇä***. ÔãÇÑÉ 8: 59-61

ÕÏíÞ ÈåÒÇÏí¡ ãÇäÏÇäÇ. (1379). * ßíÏåýäÇãÉ ÇíÇäýäÇãåýåÇí ßÊÇÈÏÇÑí æ ÇØáÇÚýÑÓÇäí.* - .ÊåÑÇä: ßÊÇÈÎÇäÉ ãáí Ìãåæ Ñí ÇÓáÇãí ÇíÑÇäº ãÑßÒ ÇØáÇÚýÑÓÇäí æ ÎÏãÇÊ Úáãí ÌåÇÏ ÓÇÒäÏ í

áíæ¡ ÌíýíÇä. (1378). *ÑÇääãÇí ÇÈÑãæÊæÑåÇí ßÇæÔ*. (ÊÑÌãå ßíæÇä ßæÔÇ). *** ŽæåÔäÇãÉ*** - ***.ÇØáÇÚýÑÓÇäí***. ÏæÑÉ 3¡ ÔãÇÑÉ 5¡ 1378: 6-7

ÊÑÌãå ãÍãÏÑÖÇ ÂíÊýÇáåå ÒÇÏå ÔíÑÇÒí). ÊåÑÇä:) **.** åÇä¡ åÇÑáí. (1376). *ÑÇääãÇí ÌÇãÚInternet* - ßÇæä äÔÑ Úáæã.

English Resources

-*Introduction to Search Engines.* (1999).

Internet: "http://www.kepl.lib.mo.us/search/searchengine.htm". Viewed January 10, 2000.

-Kiley, Robert. (2000). *Medical information on the internet: a guide to health professionals.* Edinburgh: Churchill Livingstone.

-Tomaiuolo, Niclas. (1999). *Are metasearch engines better search?*. ***Searcher***. Vol.7. pp: 30- 34.