Nicholson, S. (2000). Raising reliability of Web search tool research through replication and chaos theory. *Journal of the American Society for Information Science, 51*(8), 724-729.

Raising Reliability of Web Search Tool Research through Replication and Chaos Theory

Scott Nicholson, PhD

scott@scottnicholson.com

Abstract:

Because the World Wide Web is a dynamic collection of information, the Web search tools (or "search engines") that index the Web are dynamic. Traditional information retrieval evaluation techniques may not provide reliable results when applied to the Web search tools. This study is the result of ten replications of the classic 1996 Ding and Marchionini Web search tool research. It explores the effects that replication can have on transforming unreliable results from one iteration into replicable and therefore reliable results after multiple iterations.

Introduction

Web search tools are the only automated retrieval mechanisms available for guidance in the rapidly growing and changing universe of Web documents. In order to better represent this body of literature, the databases underneath these tools are constantly updated. The extremely dynamic nature of Web search tools makes them a very difficult target to study using traditional information retrieval (IR) evaluation methods. In

traditional IR evaluations, a set of queries is entered into a search tool, and the results are judged for relevance. Some type of score is generated for each tool and search, results are established, and the project is ended. This model, however, is not be sufficient for Web search tool research. This article replicates a classic Web search tool study to explore this problem and present a solution.

Literature Review

There has been much discussion of Web search tools in both popular and academic literature. Research works that just compare features of the search tools or are not done in a quantitative, scientific manner (as those in many popular computer magazines) will not be discussed here. Since the present paper focuses on techniques for raising the reliability of evaluative research of Web search tools, this literature review will focus on quantitative evaluative Web search tool research. The recent comprehensive literature review of search tool literature written by Schwartz (1998) in JASIS can guide those wishing to explore the rest of the search tool literature.

One of the earliest published, and therefore heavily cited, quantitative studies of Web search tools is by Courtois, Baer, and Stark (1995). "Cool Tools for Searching the Web: A Performance Evaluation" examined seven different Web search tools. The authors posed three search questions, and chose between one and three Web pages as answer documents for each. They ran searches for each query on each tool, although they do not present the actual queries used for each tool. Results reported were the total number of pages returned from each tool and the number of answer documents found.

There were two studies published on Web search tools in *Global Complexity: Information, Chaos, and Control*, the proceedings of the 1996 ASIS conference. Chu and Rosenthal (1996) examined three search tools: Alta Vista, Excite, and Lycos. They did a quantitative examination of precision and response time of these tools. Queries for each search tool were developed from real-life reference questions. Results presented were response time per query and precision, which was determined by the percentage of relevant hits out of the first ten presented by the tool.

The second study presented that year was by Ding and Marchionini (1996). They also examined three search tools: Lycos, InfoSeek, and OpenText. After the authors qualitatively compared features of these tools, they compared salience, relevance concentration, and three types of precision for each tool. They presented a five-point relevance scale for each question. A five-point answer document was an entire article about the subject, a four-point answer document contained some information about the topic, three points were given if the page contained good links to four or five-point pages, and so on. They then examined the first twenty Web pages, and gave each page a relevance ranking.

Three types of precision values were determined for each query. The first was the percentage of pages, not including duplicates, that scored a three, four, or five on the relevance ranking; the second was the percentage of pages scoring a four or five. The third was figured by comparing the hits returned by each query to a pool of hits that got a three, four, or five retrieved from all three engines. Salience was found by

totaling the relevance scores over all twenty pages for each search (including duplicates). Finally, relevance concentration was found by taking the number of items scoring a four or five in the first ten returned documents and dividing that by the total number of items scoring a four or five found in the first twenty documents.

A similar study using this method of relevance ranking was done by Leighton and Srivastava (1997). They worked to include both natural language and Boolean-based queries (fifteen in total) in their study, as they felt that the exclusion of one or the other was a shortcoming in previous studies. In calculating results, they weighted results higher in the list more heavily, and came to the conclusion that the results of this study changed dramatically when the definition of relevance was changed.

The first large-scale study was done by Tomaiuolo and Packer in 1996. They looked at Lycos, InfoSeek, and Alta Vista and two human-created evaluative search tools, Magellan and Point. They used 200 queries selected to be representative of research needs of undergraduates. They counted the relevant hits returned out of the first ten. When they could not judge the relevance from the surrogate, they visited the site. They listed results from only 32 of the 200 searches and presented charts for the average number of relevant hits out of the first ten for each tool. They did not remove duplicates from the relevance scores, but recognized the influence that duplicates might have had. They found that evaluative search tools presented fewer, but more relevant, results than the non-evaluative tools.

Recall of Alta Vista, Excite, and Lycos was examined by Clark and Willett in 1997. They looked at 30 queries submitted by students and gave each of the top ten returned documents a relevance ranking between 0 and 1. Documents on the topic scored 1, documents with links to useful pages scored .5, and duplicate, non-existent, or off-topic sites received 0. They used these rankings to calculate precision for each tool, and then looked at a pool of all documents returned to determine relative recall.

Peterson (1997) replicated his study three times over ten months. However, he looked only at the total number of documents returned by eight different search tools with no examination of relevancy. Even in this limited study, however, the dramatic differences between the replications are cause for questioning non-replicated evaluative search tool studies.

Another study using replication was conducted by Westera in 1997. Westera looked at a series of searches in eight different tools and replicated the study six months later. It was found that both the number of hits returned and the relevant results returned changed over this time, but the researcher stated that no pattern could be found.

Feldman (1998) conducted an Internet "search-off" by asking professional searchers to use real-world search requests for evaluation. Searchers did the same search for a client in both DIALOG/Dow Jones and in a Web search tool, and had the client rank the top 30 documents returned from each tool. It was found that the traditional services returned more relevant documents in general, but that the Web search tools provided some types of information not available through the traditional services.

Dong and Su (1997) presented a comprehensive literature review of the search tool studies by discussing measures of evaluation used in studies, gathering results from other studies by search tool, and summarizing

the conclusions from other studies. Su (1997) presented a framework for doing user evaluation of these search tools. Her argument, based upon lessons learned from OPAC research, was that user-based evaluations are needed in order to thoroughly understand the search tools and the way they are used. Most of these studies are based off the traditional researcher-based relevance decisions, which may not be generalizable to users.

Theoretical Background

Unlike the printed word, an electronic document can be easily modified. In fact, one commonly cited measure of quality for Web pages is that they have been updated recently. In addition, the author or publisher of the Web page may remove it or change its location without warning. The server holding the Web pages may go down temporarily or for good. The body of literature comprised of Web pages is much more dynamic than the body of print literature.

The search tools work by sending a small mobile agent program (commonly referred to as a Web robot or spider) out to examine a Web page. The robot may visit only sites that are chosen by the administrators of the search tool, it may go to sites when a user requests a visit, or it may just wander the Web looking for sites. Some robots follow links in order to index more of the Web, while other robots are controlled and visit only the top page of a Web site. When the robot visits the site, it may collect the full text or just an extract from the page. That information will be stored, along with a surrogate for the site, in the Web page database. In order to keep this record up to date, most Web robots revisit the indexed sites every few weeks to few months (Sullivan, 1999).

There are thus several different reasons for the dynamic nature of these search tools and the consequent lack of generalizability of traditional search tool research:

- The content of the Web page can be changed at any time (very common with discussion group pages, online serials/news sources), possibly causing different relevancy judgments.

- The page can be moved to a new location or be removed from the Web, causing links once counted as relevant to now be dead.

- The Web page surrogate presented to the user may be changed, causing different search results or different relevance judgments

- New Web pages are added hourly, causing different search results.

- The search tool designers change the underlying searching and indexing algorithms in order to stay competitive and to defeat those trying to cheat the system and get their pages ranked higher. This can cause very different results when doing a search.

With all of these forces of change affecting a search tool, there is cause for concern when using evaluation techniques developed for to static information retrieval tools, because studies using these techniques may not be replicable. The present study thus seeks to determine if results from one of these classical search tool experiments are replicable, and therefore reliable, over time.

Experiment

The Web search tool study by Ding and Marchionini (1996) was selected as research to replicate because of its thorough documentation. The authors listed the searches used, the relevance criteria used, and other critical details of the study; therefore, their study can be replicated and the replications compared to examine the reliability of this technique. Other studies could have been selected for this replication, and future research could look at other studies using the same technique.

Alta Vista, Excite, Infoseek, and Lycos were evaluated with the same queries as in the Ding and Marchioini study:

Query 1. Find some articles about the effects of divorce on children.

Query 2. Find some newsgroups, articles, and home pages on Flamenco dance (but not music).

Query 3. Find movie reviews on The Indian in the Cupboard.

Query 4. Find some articles on how diet can prevent Osteoporosis.

Query 5. Find some articles and sources about video-on-demand (VOD).

The queries used in each search tool, listed in Appendix A, are based upon the original research and modified to take advantage of current features of each search tool.

For each query, the first twenty pages returned by each tool were retrieved and examined. Each page was given a relevance rank from 0 to 5, based upon the criteria published in the previous study. Any Web page surrogate returned in the top twenty that pointed to the exact same URL as an earlier listing in that search tool page received a 0 for each duplication, but surrogates that pointed to copies (or mirrors) of a page received the full score each time. Pages that were down or not responding received a 0.

The measures used to examine these results are four derivatives of traditional precision (as true precision would require an examination of all pages returned). Precision A was the number of items that received a ranking of 3, 4 or 5, and Precision B was the number of items that received a 4 or 5 ranking. These precision counts were found for both the first 10 pages and the first 20 pages; they will be referred to as 10A, 10B, 20A, and 20B. These quantities were selected because in previous research it was found that 58 percent of users looked at no more than the first ten results, and another 19 percent looked at only the first twenty (Jansen, Spink, Bateman & Saracevic, 1998).

This study was replicated ten times throughout the summer of 1998, with at least one week separating each replication. For simplicity of presentation, the number of relevant documents (based on the criteria and precision method used) for each tool were summed over all five queries for that week, and the search tools ranked from 1 to 4. A rank of 1 was awarded to the search tool with the highest number of relevant results and 4 given to the lowest number of relevant results for that week.

Charts 1 - 4 show the variations in the ranking of the tools over the ten-week period. One chart is used for each precision method; the numerical data behind these charts are contained in Appendix B. The fact that the results for each precision method vary throughout the duration of the study can clearly demonstrated through a week-by-week comparison of the rankings.

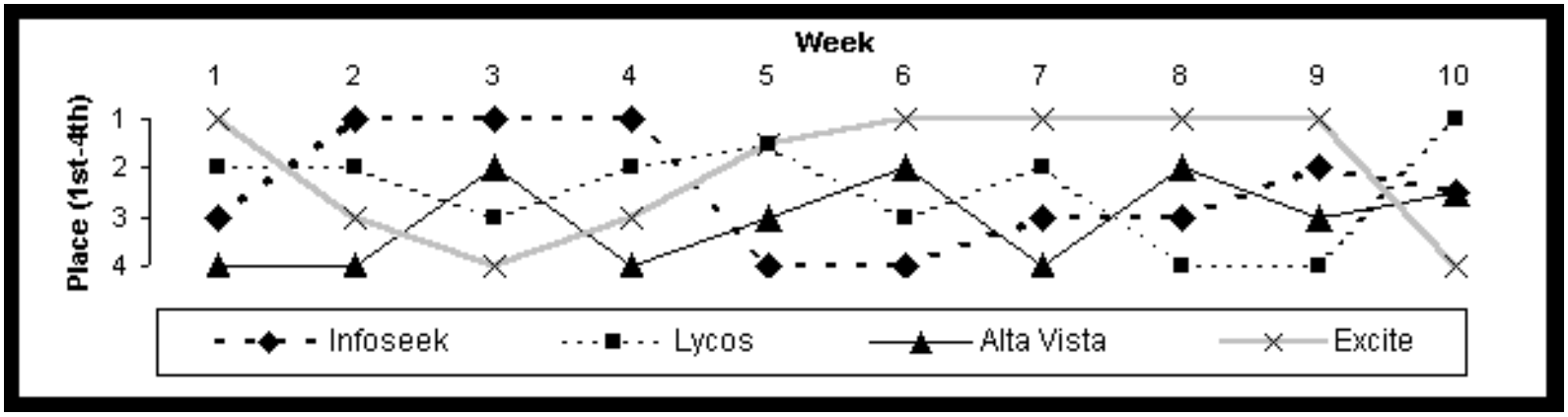

Chart 1. Ranks of tools by week using Precision A for the first 20 hits over all 5 queries

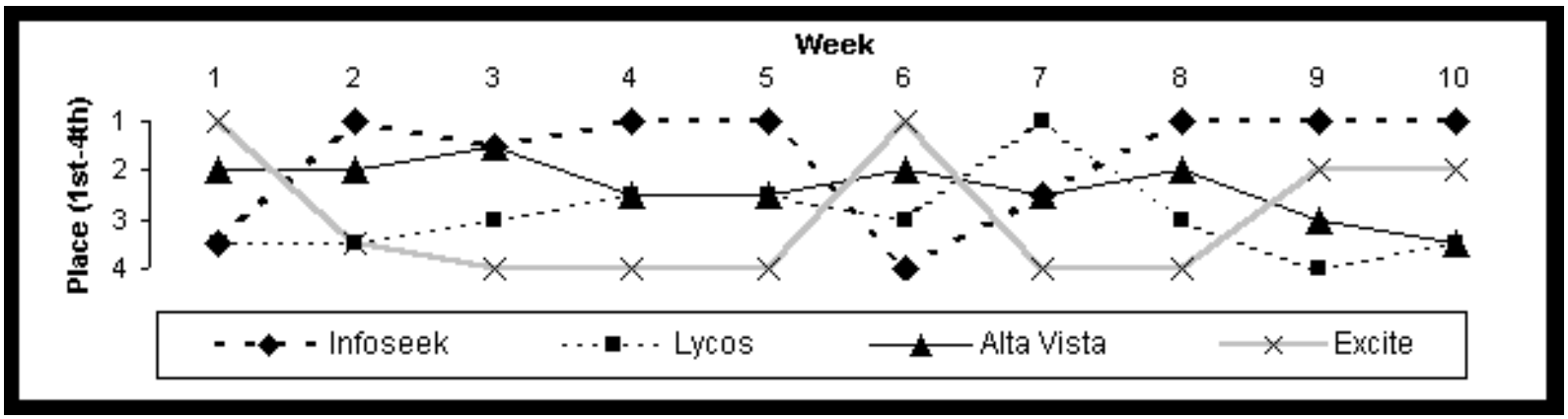

Chart 2. Ranks of tools by week using Precision B for the first 20 hits over all 5 queries

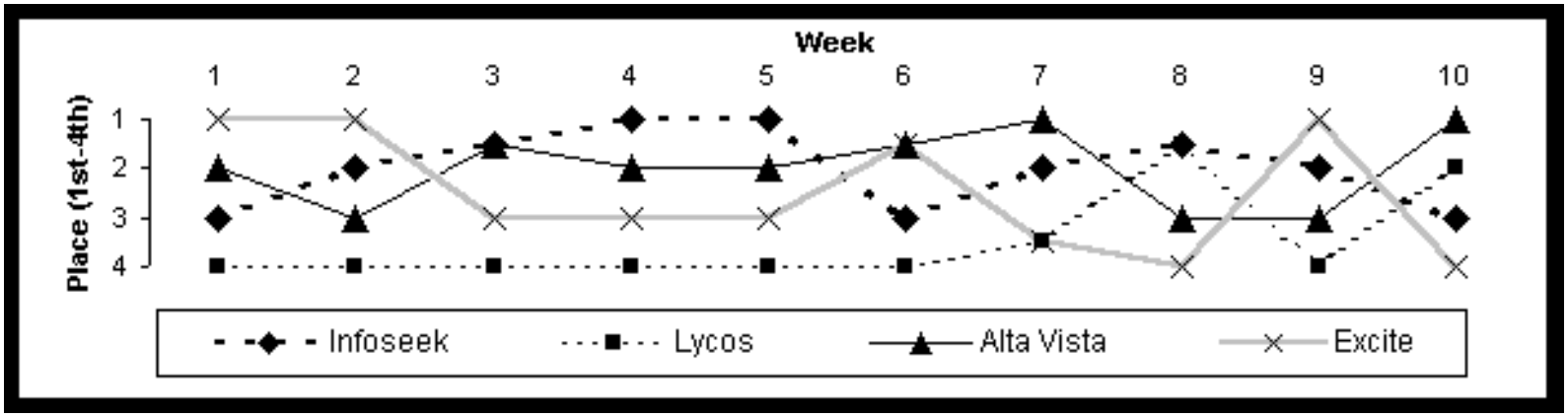

Chart 3. Ranks of tools by week using Precision A for the first 10 hits over all 5 queries

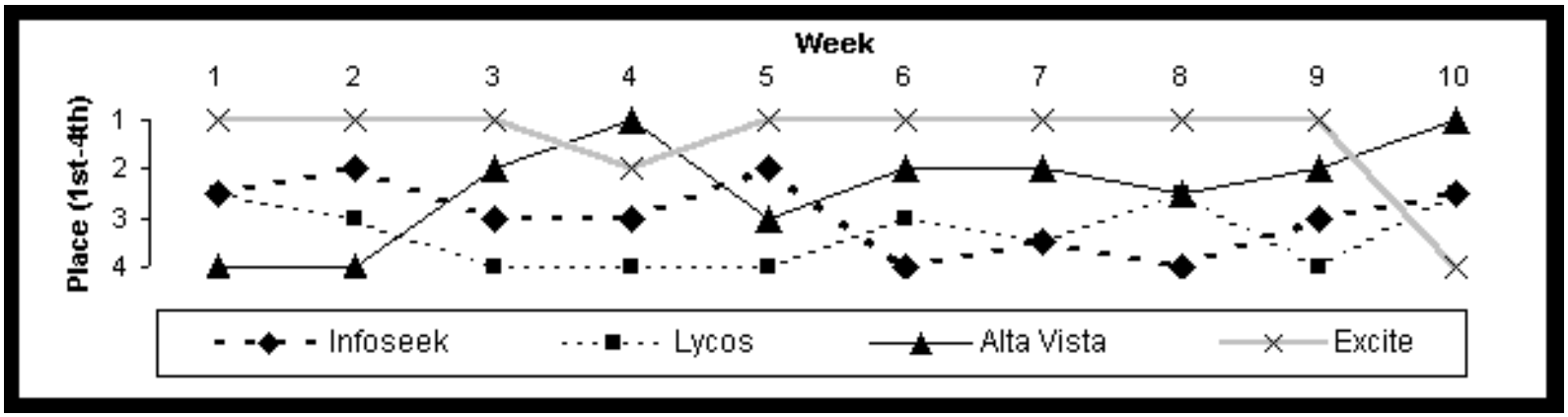

Chart 4. Ranks of tools by week using Precision B for the first 10 hits over all 5 queries

Depending upon the week chosen and the precision method selected, different search tools can be selected as first-ranking. Excite comes out first more often in Chart 3 (Precision 10A), the exceptions being weeks 4 and 10. The Web search tools are so dynamic that a study produces different results depending upon when it is done. The following table shows the Web tool that was ranked first for each week and precision method.

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
Prec. 20A	Excite	Infoseek	Infoseek	Infoseek	L&E tie	Excite	Excite	Excite	Excite	Lycos
Prec. 20B	Excite	Infoseek	I&A tie	Infoseek	Infoseek	Excite	Lycos	Infoseek	Infoseek	Infoseek
Prec. 10A	Excite	Excite	Excite	AV	Excite	Excite	Excite	Excite	Excite	AV
Prec. 10B	Excite	Excite	I&A tie	Infoseek	Infoseek	A&E tie	AV	I&L tie	Excite	AV

Table 1. Search tool ranked **first** over all five queries by week and precision measurement technique

By comparing the two different techniques for selecting relevant documents (A versus B) and comparing the number of returned pages examined (20 versus 10) for each week, it can be seen that these research design decisions can have a dramatic effect on the results of the study.

Discussion

There are two possible explanations for this discrepancy between replications of the study. The first is that the Web page databases are very dynamic and change constantly; the second is that the researcher in this study changed the method of measuring relevance judgments from week to week. While it is possible that the relevance judgments did change some over the course of the study because of learning effects, it is more likely that the Web page databases were very dynamic. Another small experiment was conducted to confirm this hypothesis.

Examination of the Dynamic Nature of Web Page Databases

In this experiment, two queries were run in each of four different tools (Alta Vista, Excite, Infoseek, and Lycos) for five replications with at least a week between each replication. There were no relevance judgments made, as the purpose of this experiment was to see how much the first twenty results changed from week to week.

For this study, the queries were selected to represent two extremes, as it was felt that the

timeliness of the topic would make a difference in the dynamic nature of the Web databases. Therefore, the first (and theoretically more stable) query was for the 1980 census while the second query was for bugs in Windows 98. The queries were modified initially to take advantage of the features of each search tool.

Two counts will be presented here for each topic and each pair of consecutive weeks. The first listing, labeled "New," is how many documents (out of 20) were not on the list from the previous week. The second listing, labeled "Moved," is how many documents changed position from the previous week. Therefore, a score of 0 means that nothing was new from the previous week or nothing moved from the previous week, and a score of 20 means that every page was new or every page moved. For this study, individual pages are determined by unique URLs.

1980 Census	Week 1 to Week 2		Week 2 to Week 3		Week 3 to Week 4		Week 4 to Week 5	
	New	Moved	New	Moved	New	Moved	New	Moved
Alta Vista	3	17	3	17	0	0	0	0
Excite	14	5	19	1	16	4	18	2
Infoseek	4	10	4	12	1	0	6	14
Lycos	3	0	0	12	1	0	0	20

Table 2. Pages (out of 20) that were new or changed positions between weeks with the same query each week about the 1980 Census.

Windows 98 Bugs	Week 1 to Week 2		Week 2 to Week 3		Week 3 to Week 4		Week 4 to Week 5	
	New	Moved	New	Moved	New	Moved	New	Moved
Alta Vista	12	3	13	6	4	6	4	3
Excite	15	4	15	4	18	2	14	4
Infoseek	6	10	9	3	6	12	10	8
Lycos	20	0	0	8	20	0	0	0

Table 3. Pages (out of 20) that were new or changed positions between weeks with the same query each week about Windows 98 bugs.

While the theoretically stable query did provide slightly more stable results, both queries showed considerable change from week to week. Lycos was unusual in its results, as it seemed that new sites appeared only every other week of the study. For the rest of the search tools, most of the time new pages were added and old pages moved during every replication. Change in Excite is noticeably higher than the rest because they rebuilt their entire database weekly at the time of this research. This small experiment shows that the Web page databases are very dynamic.

Looking for Patterns in the Chaos

Is there a way that traditional techniques can be used to reliably evaluate Web search tools? The solution may lie in chaos theory, which was explored by James Gleick in his 1987 work Chaos: Making a New Science. The fundamental idea of chaos theory is that behind every randomly fluctuating set of observations is an overall pattern of some type. Once that pattern is discovered, predictions about future observations can be made.

In searching for a pattern in this data set, the results were aggregated over the entire ten week period. By looking at the frequency of the search tool having the highest number of relevant documents for every precision technique and query, a replicable pattern was discovered. The rank order of the tools in this pattern was Excite, Infoseek, Alta Vista, and Lycos, and the pattern is replicated almost exactly over the odd weeks, over the even weeks, and over all ten weeks of the experiment. It is to be expected that more iterations would produce the same pattern.

	First place over all 10 weeks	First place over all Even Weeks	First place over all Odd Weeks
Excite	19	7.5	11.5
Infoseek	12.5	7.5	5
Alta Vista	5.5	3.5	2
Lycos	3	1.5	1.5

Table 4. Total first place rankings for each search tool over all four Precision methods

Conclusion - Raising the Reliability of Web Search Tool Research

By replicating the study and aggregating the results, reliable results may be found where none existed in a single iteration of the study. By using different techniques of relevance judgments and aggregating those results, the generalizability of the study can be improved as well. Future studies can use more queries in order to raise the reliability of their studies; five queries are not enough to make definitive statements about the quality of individual tools. With this set of results, the difference between the search tools is more clearly defined through the aggregate of ten replications than with either set of five replications. It is hoped that future researchers using traditional techniques to evaluate dynamic Web search tools will replicate their studies in order to produce more stable results.

Acknowledgements

The author would like to thank Dr. Mark Rorvig, Kathleen Murray, and Terry Sullivan for their assistance and encouragement in the development of this research.

References

Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: A comparative study and evaluation methodology. Proceedings of the 59th Annual Meeting of the American Society for Information Science, 33, 127-135.

Clarke, S., & Willett, P. (1997). Estimating the recall performance of Web search engines. Aslib Proceedings, 49(7), 184-189.

Courtois, M., Baer, W., & Stark, M. (1995). Cool tools for searching the Web: A performance evaluation. ONLINE, 19(6), 14-32.

Ding, W., & Marchionini, G. (1996). A comparative study of Web search service performance. Proceedings of the 59th Annual Meeting of the American Society for Information Science, 33, 136-142.

Dong, X. and Su, L. (1997). Search engines on the World Wide Web and information retrieval from the Internet: A review and evaluation. Online & CDROM Review, 21(2), 67-81.
Leighton, H., & Srivastava, J. (1997). Precision among World Wide Web search services (search engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos [Online]. Available: http://www.winona.msus.edu/is-f/library-f/webind2/webind2.htm [1998, April 18].

Feldman, S. (1998). The Internet search-off. Searcher, 6(2) [Online]. Available: http://www.infotoday.com/searcher/feb98/story1.htm [1999, May 15].

Gleick, J. (1987). Chaos: Making a New Science. New York: Penguin Books.

Jansen, B., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life informationretrieval: A study of user queries on the Web. ACM SIGIR Forum, 32(1), 5-17.

Peterson, R. (1997). Eight Internet search engines compared. First Monday, 2(2) [Online]. Available: http://www.firstmonday.dk/issues/issue2_2/peterson/ [1999, May 12].

Schwartz, C. (1998). Web Search Engines. Journal of the American Society for Information Science, 11, 973-982.

Su, L. (1997). Developing a comprehensive and systematic model of user evaluation of Web-based search engines. In M. Williams (Ed.), National Online Meeting: Proceedings - 1997 (pp. 335-345). Medford, NJ: Information Today, Inc.

Sullivan, D. (1999). Search engine EKGs [Online]. Available: http://www.searchenginewatch.com/reports/ekgs/index.html [1999, May 16].

Tomaiuolo, N., & Packer, J. (1996). An analysis of Internet search engines: Assessment of over 200 search queries. Computers in Libraries, 16(6), 58-62.

Westera, G. (1997). Robot-driven Search Engine Evaluation Overview [Online]. Available: http://www.curtin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy /index.htm [1999, May 12].

Appendix A: Searches used in the large replication study

Query 1. Find some articles and facts about the effects of divorce or parental alienation on children

Infoseek: divorce "parental alienation" +effect +child

Lycos: divorce alienation effect child

Alta Vista: divorce "parental alienation" +effect +child

Excite: divorce "parental alienation" +effect +child

Query 2. Find some newsgroups, article, and home pages on Flamenco dance (but not music)

Infoseek: Flamenco dancing -music

Lycos: flamenco dancing -music

Alta Vista: +flamenco dancing -music

Excite: +flamenco dancing -music

Query 3: Find movie reviews on The Indian in the Cupboard

Infoseek: movie review +"Indian in the Cupboard"

Lycos: movie review Indian Cupboard

Alta Vista: +"Indian in the Cupboard" movie review

Excite: +"Indian in the Cupboard" movie review

Query 4: Find some articles on how diet and prevent Osteoporosis

Infoseek: Osteoporosis diet prevent

Lycos: osteoporosis diet prevent

Alta Vista: Osteoporosis diet prevent

Excite: Osteoporosis diet prevent

Query 5: Find some articles and sources about video-on-demand (VOD)

Infoseek: "Video on Demand" VOD

Lycos: video on demand VOD

Alta Vista: "Video on Demand" VOD

Excite: "Video on Demand" VOD

Appendix B: Numerical ranks by Precision type

Week	1	2	3	4	5	6	7	8	9	10
Infoseek	3	1	1	1	4	4	3	3	2	2.5
Lycos	2	2	3	2	1.5	3	2	4	4	1
AV	4	4	2	4	3	2	4	2	3	2.5
Excite	1	3	4	3	1.5	1	1	1	1	4
Best	Excite	Infoseek	Infoseek	Infoseek	L&E tie	Excite	Excite	Excite	Excite	Lycos

Ranks of tools by week using Precision A for the first 20 hits over all 5 queries

Week	1	2	3	4	5	6	7	8	9	10
Infoseek	3.5	1	1.5	1	1	4	2.5	1	1	1
Lycos	3.5	3.5	3	2.5	2.5	3	1	3	4	3.5
AV	2	2	1.5	2.5	2.5	2	2.5	2	3	3.5
Excite	1	3.5	4	4	4	1	4	4	2	2
Best	Excite	Infoseek	I&A tie	Infoseek	Infoseek	Excite	Lycos	Infoseek	Infoseek	Infoseek

Ranks of tools by week using Precision B for the first 20 hits over all 5 queries

Week	1	2	3	4	5	6	7	8	9	10
Infoseek	2.5	2	3	3	2	4	3.5	4	3	2.5
Lycos	2.5	3	4	4	4	3	3.5	2.5	4	2.5
AV	4	4	2	1	3	2	2	2.5	2	1
Excite	1	1	1	2	1	1	1	1	1	4
Best	Excite	Excite	Excite	AV	Excite	Excite	Excite	Excite	Excite	AV

Ranks of tools by week using Precision A for the first 10 hits over all 5 queries

Week	1	2	3	4	5	6	7	8	9	10
Infoseek	3	2	1.5	1	1	3	2	1.5	2	3

Lycos	4	4	4	4	4	4	3.5	1.5	4	2
AV	2	3	1.5	2	2	1.5	1	3	3	1
Excite	1	1	3	3	3	1.5	3.5	4	1	4
Best	Excite	Excite	I&A tie	Infoseek	Infoseek	A&E tie	AV	I&L tie	Excite	AV

Ranks of tools by week using Precision B for the first 10 hits over all 5 queries