

On the Instability of Web Search Engines

Erik Selberg * Oren Etzioni *

Go2Net, Inc.

999 Third Ave. Suite 4700

Seattle, WA 98104

{erik, oren}@go2net.com

<http://www.cs.washington.edu/homes/{selberg, etzioni}>

<http://www.cs.washington.edu/research/metacrawler>

Abstract

The output of major WWW search engines was analyzed and the results led to some surprising observations about their stability. Twenty-five queries were issued repeatedly to the engines and the results were compared. After one month, the top ten results returned by eight out of nine engines had changed by more than fifty percent. Furthermore, five out of the nine engines returned over a third of their URLs intermittently during the month.

1 Introduction

While the World Wide Web began as a convenient method for scientists to disseminate information to one another, the Web has rapidly become a key medium for information dissemination for everyone. Indeed, the Web has been likened to a “digital library,” and to a searchable 15-billion word encyclopedia (Barrie and Presti, 1996). As pointed out by Lawrence and Giles, “Web search engines have made the large and growing body of scientific literature and other information resources accessible within seconds” (Lawrence and Giles, 1998b). Web search services such as Excite or Yahoo! are essential for locating information on the Web, and are queried millions of times daily. In the 10th Web Survey done by Georgia Tech’s Graphics, Visualization, and Usability Center, 84.8% out of 16,891 respondents found web pages via search engines (Kehoe and Pitkow, 1999). Considering that in December, 1999, Excite reported an average of 123 million page-views per day with 51 million registered users, and Yahoo! reported 465 million page-views per day and more than 100 million registered users (Yahoo, Inc., 2000; Excite@Home, Inc., 2000), it is clear that these services have a substantial impact on millions of people daily.

*Work completed while at the University of Washington.

Since a relatively small set of search engines is responsible for the majority of online searching, the performance of these key search engines merits detailed scrutiny. Experiments by Selberg and Etzioni in 1995 and 1999 demonstrated that each of the major search engines returned only a fraction of the URLs of interest to users, and that the overlap of the results returned by different search engines was surprisingly small (Selberg and Etzioni, 1995; Selberg, 1999). Lawrence and Giles refined and extended that work in two studies in a number of important ways (Lawrence and Giles, 1998b; Lawrence and Giles, 1999). Lawrence and Giles estimated the relative size of the indices used by various major search engines to the size of the “indexable web” or web pages reachable by some search engine. In the process, they found that all had a small coverage of available data, the search engine with the largest index (Northern Light) covering 16% of the indexable web, estimated at 800 million documents in July, 1999. Similar experiments by Bharat and Broder (Bharat and Broder, 1998) reported that the size of the Web to be roughly 200 million documents in Nov. 1997, and subsequently found that the search engine with the largest index (AltaVista) covered 50% of the web.

Another key question for users is how the sundry search engines behave. If we view the Web as a digital library, then the search engines could be likened to an online card catalog. All available information is indexed somewhere in the search engines, and for a given search, while the results may change slightly over time due to the addition of new titles and the removal of other titles, they will remain largely unchanged. In essence, a search engine should be reasonably *stable* in the way it returns information to given queries.

However, the stability of search engine results has not been deeply investigated: how do the results of a query change over time? and how do the results of a query change when a different query syntax is used? Recently, Notess observed that search engines are often very inconsistent in the way they report the number of URLs that match a particular query (Notess, 2000). Stability is closely linked to the issue of search engine reliability: Is the user receiving the results most relevant to his query? Will minute changes to query syntax result in massive changes to query results? Can a user retrace his steps by re-issuing a query a week later?

2 An Experiment Measuring the Degree of Engine Instability

We define the *stability* of a search engine as the difference between two sets of results that were calculated from the same query submitted at two different times. Because the Web is a dynamic environment, there will naturally be some degree of instability as URLs are added and removed from an index. However, a question remains as to whether or not a search engine exhibits instability that is in line with the rate of Web change and growth.

As a means of evaluating the behavior of the engines, we conducted an experiment to measure the change in the search engines’ output and compared that change with what was expected.

We issued a set of twenty-five queries to nine major search engines: AltaVista, Excite, Lycos, HotBot, InfoSeek, Lycos, Northern Light, PlanetSearch, WebCrawler, and Yahoo!, and analyzed the documents that were returned. The queries are an independently generated set that were used as part of the Lawrence and Giles study on measuring search engine coverage (Lawrence and Giles, 1998b). These queries were originally issued by scientists at NEC using the Inquiris meta-search engine (Lawrence and Giles, 1998a). To be included in the Lawrence and Giles study, each query had to return a combined total of 600 or fewer URLs. For our study, we modified the queries by removing any syntactic or logical modifiers in the query and treated them as pure keyword queries. All queries except one consisted of two or more words. While there is some bias introduced by the queries as they come from a scientific community, we opted for an independently generated set rather than use either random or popular queries taken from a search engine in order to keep the number of results returned by each query as close to a constant as possible. In this manner, we attempted to remove potential bias from queries that returned either a very small or very large number of results. The queries are presented in Appendix A.

The twenty-five queries were issued twenty-five times over a one month period, between December 7, 1998 and January 8, 1999. Some search engines have reported using a form of query result caching, mostly to improve performance for users that have their browser reload the page, causing a reissue of the query (Brewer, 1997). Because we had no knowledge of how long search results might remain in a cache, we attempted to avoid cached results by increasing the interval between issuing the set of queries exponentially, with the initial interval being 15 minutes, then increasing to 30, 60, and so forth.

The URLs in the output from the search engines were extracted into a URL set. Only URL sets that came from the same engine and same query were compared. As each service is consistent in the way it reports a particular URL, simple string comparison was used to detect duplicates. The difference between the two URL sets A and B was measured using the bi-directional set difference:

$$|(A - B) \cup (B - A)|$$

This measure does not consider changes in the position of URLs because while a change in position can make it more or less convenient to locate information, that information can still be found via the query. The percentage of URL set difference is measured by simply dividing the set difference by the number of elements in the union, or:

$$\frac{|(A - B) \cup (B - A)|}{|A \cup B|}$$

For example, if URL set A contained URLs $\alpha, \beta, \gamma, \delta$ and URL set B contained URLs α, β, ϵ , then the percentage difference between A and B would be:

$$\frac{|(A - B) \cup (B - A)|}{|A \cup B|} = \frac{|(\{\alpha, \beta, \gamma, \delta\} - \{\alpha, \beta, \epsilon\}) \cup (\{\alpha, \beta, \epsilon\} - \{\alpha, \beta, \gamma, \delta\})|}{|\{\alpha, \beta, \gamma, \delta\} \cup \{\alpha, \beta, \epsilon\}|}$$

$$\begin{aligned}
&= \frac{|\{\gamma, \delta\} \cup \{\epsilon\}|}{|\{\alpha, \beta, \gamma, \delta, \epsilon\}|} \\
&= \frac{|\{\gamma, \delta, \epsilon\}|}{|\{\alpha, \beta, \gamma, \delta, \epsilon\}|} \\
&= \frac{3}{5} \\
&= 60\%.
\end{aligned}$$

2.1 Experimental Results

A reasonable assumption to make is that the results of a search engine do not change substantially over a month. In the initial query submission run 25,386 URLs were returned from the nine engines. A run done a month later contained 25,756 URLs. However, 8,222, or 32.39%, of those URLs were not present in the initial submission's results. Clearly substantial change had occurred. Further analysis broken down by the individual search engines as shown in Figure 1 shows that change is not isolated to a few engines. The output from eight of the nine engines changed by over 40%, with the only exception being AltaVista which changed by slightly over 20%. If the results were restricted to the Top 10 results, most engines report more change, with InfoSeek, the highest, at 64%.

These initial results were created using the Default search syntax and options of each engine, which typically involves some kind of “best match” ranking. “Best match” ranking uses a function that ranks documents based on how many of the query words they contain. Two common alternatives were also examined: submitting the entire query as a phrase, and “AllPlus” syntax, which is to preface each term with a “+” sign. Prefacing a term with a “+” has become standard on Web search engines to designate a term that is required to be present in the referenced document. One might assume that the default search syntax produces general results that would change substantially over time, and that more specific syntax would produce results that are more stable. Table 1 summarizes the average difference across all engines for Default, AllPlus, and Phrase syntax. Averaged across all engines, the change for the Top 10 was 54.38% and for the Top 200 45.60%. Although there is less change when using different query syntax, the change is still substantial. Furthermore, numerous studies have shown that most users use the default syntax (Selberg and Etzioni, 1995; Silverstein et al., 1998). Thus, our observations regarding queries using the default syntax are likely more applicable to average Web users.

One might argue that this high rate of change in search engine output is due to growth and change in the Web. To determine if this is the case, we plotted the average rate of change for the search engine output throughout the month for the Default, AllPlus, and Phrase syntax, as shown in Figure 2. In addition, we plotted a flat 25%/mo. growth and change rate for the Web. This rate is an upper bound that combines the growth estimate of Bharat and Broder (Bharat and Broder,

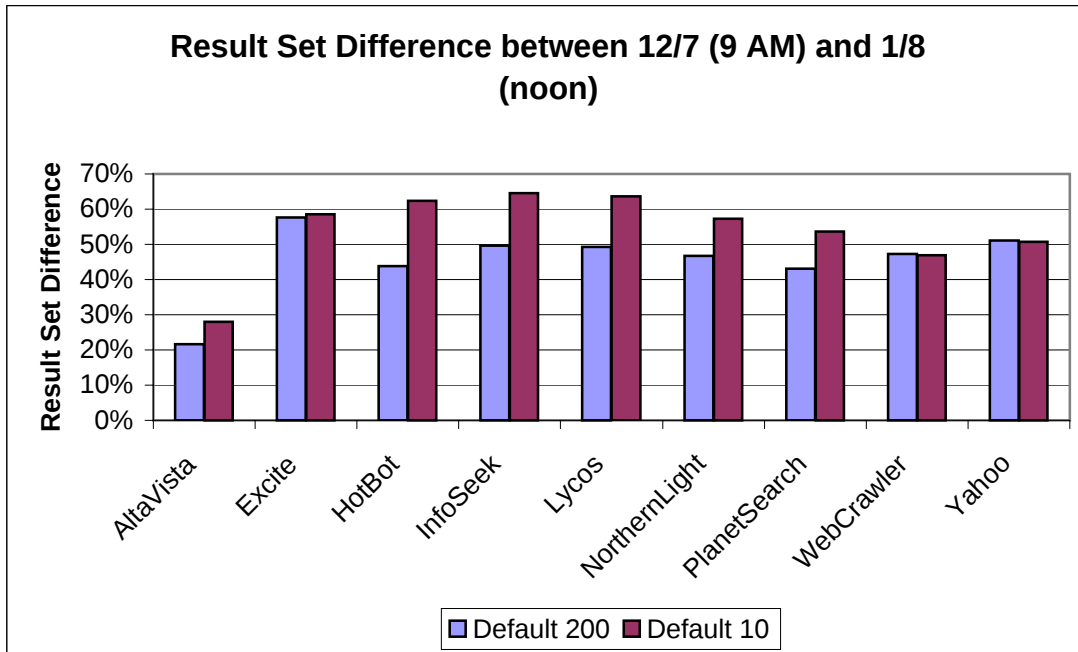


Figure 1: Top 200 and Top 10 results using default options.

As shown, all the Top 200 results except AltaVista changed by over 40% difference over the course of 1 month. If only the Top 10 are examined, then most differ between 50% and 60%, indicating that some URLs are being re-ordered. These figures compare the URLs from the first run to the second to last run.

	Default	AllPlus	Phrase
Top 10	54.38%	33.34%	30.77%
Top 200	45.60%	22.92%	19.84%

Table 1: Change over one month for Default, Phrase, and AllPlus options.

This table shows the percent the original results changed over one month for Default, Phrase, and AllPlus, averaged across all engines for Top 10 and Top 200. Note that the “Default” syntax and Top 10, which is the common case for most queries (Silverstein et al., 1998), changes by over 50%. The best case, using the Phrase option and getting 200 results, is still substantial at a little under 20% difference.

1998) and change estimate by Douglis *et al.* (Douglis et al., 1997). The output changes at a roughly linear 40% rate for Default, 30% for the AllPlus and Phrase options. Clearly, while Web change and growth may explain some of the change in the results, it does not explain all of the change.

Another argument for the high rate of change in search engine output is that because the Web is much larger than any one search engine, search engines are growing at a faster rate than the Web in order to stay up-to-date and competitive with one another. However, Sullivan has reported that most engines did not report any substantial growth in the testing time period except for Northern Light, although PlanetSearch was not part of his study (Sullivan, 1999). While slightly dated, a 1997 article by Brake gives a good discussion as to why this is the case (Brake, 1997), indicating that search engine companies are focusing on improving finding quality matches within their index rather than expanding their index.

Even though the queries used in this experiment were independently generated, one concern may be they were simply susceptible to returning results that had a high degree of change due to new documents being added and obsolete documents being removed. If a significant number of new documents were added, then previous documents may still be present in the results, but pushed outside of the Top 10 or Top 200 window, thus showing a misleadingly high degree of change. To determine if this was in fact happening, we examined how many URLs were removed in subsequent queries only to reappear later. In order to remove the effect of URLs falling outside either the Top 10 or Top 200 window, we looked at which of the Top 10 did not appear in the Top 200 of a subsequent query, only to reappear in the Top 10 of later query results. Figure 3 shows our results. As shown, in five out of the nine engines over a third of the URLs returned were temporarily removed from the search results. While not as pronounced, three of the remaining four also exhibit some of this phenomenon. WebCrawler, which has the smallest reported index of the nine, was the only engine that did not display this phenomenon on at all.

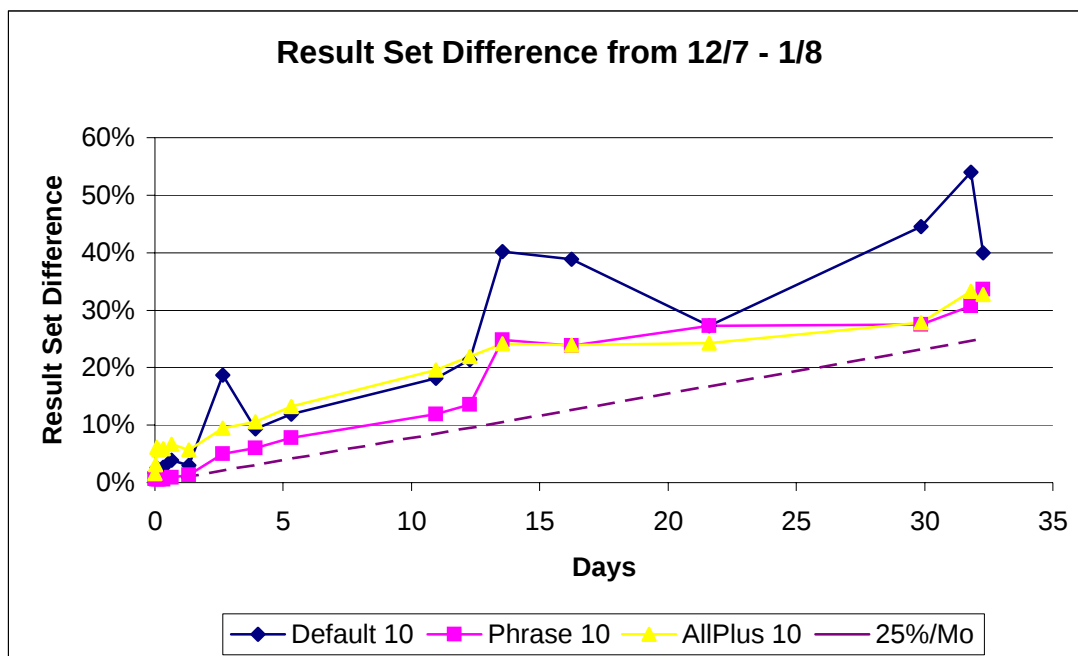


Figure 2: Average change over time for Top 10 URLs.

The AllPlus results grow roughly linearly. Phrase results are also fairly linear, with a notable jump near the Day 13. Default results are much more errant, with a corresponding jump to the phrase graph as well as two others. However, all results grew faster than the estimate for Web growth and change of 25% per month. The interval was reset to an hour after a week and a half in order to obtain more data points, and at the termination of the experiment the query sets were run two more times manually, once at noon and once at midnight, to ensure that there were no errors in the automatic submission script.

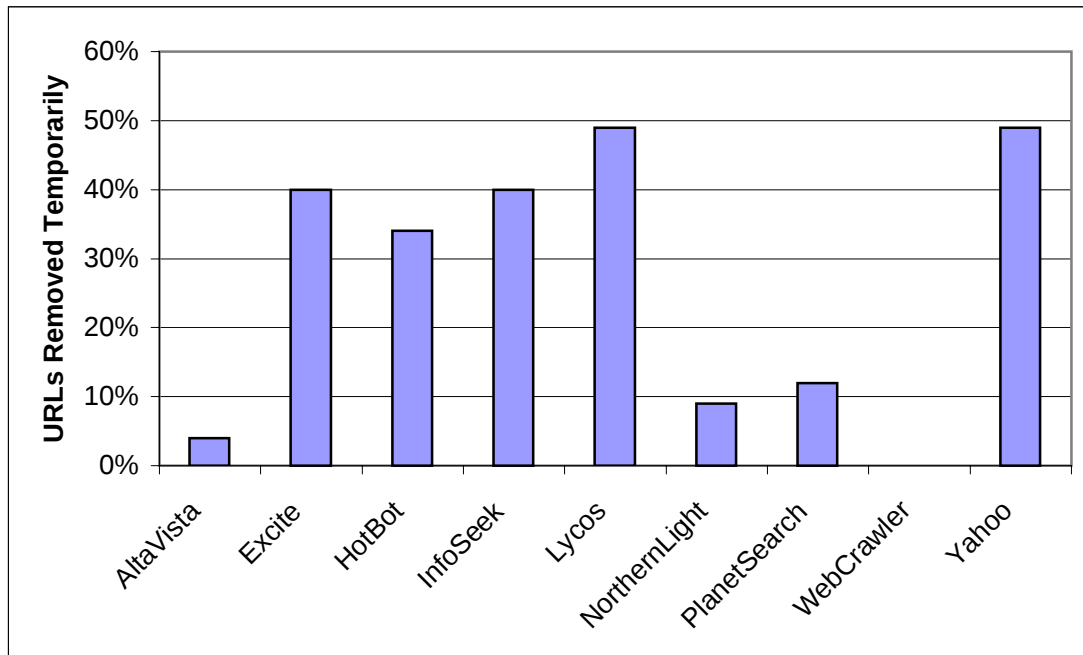


Figure 3: Percentage of URLs removed temporarily.

This chart shows for each engine the percentage of URLs returned in the Top 10 that were not present in the Top 200 of a subsequent result set, only to reappear in the Top 10 of another subsequent result set. These results used the Default query settings; results using AllPlus and Phrase were extremely similar except for InfoSeek, where the results lowered by roughly 10% and 20% respectively.

Rank Range	% Covered
1-10	52.20%
1-50	88.49%
1-100	96.97%
1-200	99.69%

Table 2: Percentage of user clicks in various height ranges.

This table shows shows four ranges of URL ranks in a ranked relevancy list, where a rank of 1 is the top of list, along with the percentage of URLs that were viewed were contained within that range. These numbers are calculated from 38,849 URLs that were viewed during the month of April, 1998 on the HuskySearch parallel web search engine.

Since most engines show a higher degree of stability when 200 documents are retrieved compared to 10, and most engines reported that they have thousands of matches for most of the twenty-five queries, it is a valid conjecture that by retrieving all available results, there might be substantially less instability in the results over a month's time. This may in fact be true. Unfortunately, some engines have a hard limit on how many documents are retrievable regardless of how many are advertised. Therefore we were unable to retrieve all available documents. At the time of these experiments, AltaVista had a hard limit of 200, which was the smallest limit of all nine engines. In addition, a histogram of the documents viewed during April 1998 on the HuskySearch engine (Selberg and Etzioni, 1997) shown in Table 2 demonstrates that 99.69% of the documents viewed are ranked 200 or higher. A similar study by Silverstein *et al.* on a million queries submitted to AltaVista showed that 95.7% of all users did not look beyond 30 results (Silverstein et al., 1998). These findings may be exaggerated because AltaVista presents ten documents to the user at a time, as compared to HuskySearch which presents all available documents on a single page. For all practical purposes, the top 200 documents are sufficient to compare instability.

2.2 Analysis

This experiment demonstrates that there is substantial change in Web search service results, and we show that there is substantial change that is unexplained by the Web's dynamic nature. Our hypothesis for this high rate of change in search service results is that the search services trade off quality for speed, and we saw the impact of those tradeoffs. One common technique among search services is to use a limit on the resources available for each query. Thus, if a system is heavily loaded, the results may not be of as high a quality as when the system is lightly loaded.

For our experiment, “lower quality” translates to a greater difference between those URLs and the originals. For example, the spikes in Figure 2 are consistent with this technique. The large positive jump near the 12 day mark began with a run that started on Saturday at 11 PM PST, which is a light load time for the search services. The apex of the jump was a run submitted on a Monday at 7:30 AM PST, which is a heavy load time, according to the search services. The negative jump at the end of the experiment started with a run submitted on a Friday around noon, still peak use hours, and ended with a run submitted on a Friday near midnight, which is decidedly off-peak. Both of these jumps occurred when one run was done during “peak” traffic time, and the other during “off-peak” time. The spike that occurred near the beginning of the experiment was the only other occurrence of two consecutive points where one run was during “peak” time and the other “off-peak” traffic times.

Another common technique is to “threshold” results. This technique involves finding a set number of results that match the query within a certain threshold, rather than finding all matches and placing them in an absolute order. This technique is typically used on a portion of the search service’s index that is placed in main memory. Searching a memory-based database is significantly faster than searching a disk-based one; therefore, if an appropriate number of “good” results can be found in main memory, those will be returned under the thresholding scheme. If the requisite number of results are not found in main memory, then the search service will swap a portion of the memory index with a portion of the disk index that does have the appropriate results. Either way these results may not be the best results that could be obtained by searching through the entire index contained on disk. Furthermore, depending on when the query was issued, URLs may be removed and then reappear depending on the contents of the memory-based index. The high rate of change, especially the type of change exhibited by Figure 3, is consistent with the application of this technique.

2.3 Specific comments on various engines

There were some interesting notes concerning individual engines. HotBot’s ranking implied that many pages were ranked equivalently and sorting among equivalent ranks was random. For each query option, it had a high rate of change at the Top 10, but much lower rate of change for Top 200. AltaVista did exceptionally well with phrases, with change under 5% for both Top 10 and Top 200. WebCrawler was poor with Default syntax, however it had near-zero change for AllPlus and Phrase syntax. Finally, Yahoo! was exceptionally high in every category. At the time of these experiments, Yahoo! had recently switched its backup search engine for queries that contained terms not found in its hand-created directory from AltaVista to an engine by Inktomi Inc., which provides the same underlying technology for HotBot. We believe this is attributed more towards Yahoo! changing its backup search engine and bringing it up to speed recently rather than an intrinsic behavior of that engine.

3 Conclusions

In conclusion, the output of search engines changes at a surprisingly high rate over time, as high as 64% depending on the search engine used and the time between query submissions. Furthermore, as high as 49% of the URLs that appear in the Top 10 of some result set disappear in subsequent results, only to reappear again later. We hypothesize that the reason for the observed instability in search results is quality-for-speed tradeoffs made by the search engines, not the addition of new documents that push older ones outside the 200 document window. Other studies have indicated that the most search engines are not increasing the size of their indices dramatically indicating that not only are the search engines unable to return information contained within their own indices, but are covering a decreasing portion of the Web (Sullivan, 1999; Brake, 1997). Unstable search engine results are counter-intuitive for the average user leading to potential confusion and frustration when trying to reproduce the results of previous searches. Scientists that use the Web to locate up-to-date research information run into the same problems. Scientists using search engine results as part of their experimental research (e.g., (Lawrence and Giles, 1998b)) need to consider whether instability affects the results of their experiments. Educators that make use of search engines in assignments may find those assignments to be unfair because their results cannot be replicated. And people that use the Web for dissemination of information may find that even though a search engine has indexed their information, it may still not be retrievable. In addition, unstable searching does make large-scale, albeit slower search resources more attractive. For example the Internet Archive (Internet Archive, Inc., 1998) is attempting to archive the entire Web. While it will unlikely be as fast as modern search engines, it may be both stable and comprehensive.

4 Acknowledgments

We would like to thank Mary Kaye Rodgers, for her assistance and support during the writing of this paper. Steve Lawrence, Susan Dumais, Oren Zamir, Steve Tanimoto, and Efthimis Efthimiadis also contributed to early drafts of this paper. This paper marks the final chapter in Drs. Selberg and Etzioni's joint research on metasearch while at the University of Washington, and both would like to extend hearty thanks to the faculty, staff, and students of the University of Washington for their support for the past six years.

This research was funded in part by Office of Naval Research grant 98-1-0177, and by National Science Foundation grants IRI-9357772 and DL-9874759.

References

- Barrie, J. M. and Presti, D. E. (1996). The World Wide Web as an Instructional Tool. *Science*, 274(5286):371.
- Bharat, K. and Broder, A. (1998). A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. In *Proceedings of the 7th World Wide Web Conference*, Brisbane, Australia.
- Brake, D. (1997). Lost in cyberspace. *New Scientist*.
<http://www.newscientist.com/keysites/networld/lost.html>.
- Brewer, E. (1997). The HotBot Search Engine [talk only]. In *Proceedings of the American Library Association 1997 Annual Conference*, San Francisco, CA.
- Douglis, F., Feldmann, A., Krishnamurthy, B., and Mogul, J. (1997). Rate of Change and other Metrics: A Live Study of the World Wide Web. In *Proceedings of the 1997 Usenix Symposium on Internet Technologies and Systems*, Monterey, CA.
- Excite@Home, Inc. (2000). Excite@Home Reports Fourth Quarter And Fiscal Year Results . Press Release.
http://www.excitehome.net/news/pr_000120_01.html.
- Internet Archive, Inc. (1998). The Internet Archive.
<http://www.archive.org>.
- Kehoe, C. and Pitkow, J. (1999). *GVU's Tenth WWW User Survey Report*. Office of Technology Licensing, Georgia Tech Research Corporation.
http://www.gvu.gatech.edu/user_surveys/survey-1998-10/.
- Lawrence, S. and Giles, C. L. (1998a). Inquirus, the NECI meta search engine. In *Proceedings of the 7th World Wide Web Conference*, Brisbane, Australia.
- Lawrence, S. and Giles, C. L. (1998b). Searching the World Wide Web. *Science*, 280:98–100.
- Lawrence, S. and Giles, C. L. (1999). Accessibility of information on the web. *Nature*, 400:107–109.
- Notess, G. (2000). Search engine inconsistencies. *Online*, 24(2).
- Selberg, E. (1999). *Towards Comprehensive Web Search*. PhD thesis, University of Washington.
- Selberg, E. and Etzioni, O. (1995). Multi-Service Search and Comparison Using the MetaCrawler. In *Proceedings of the 4th World Wide Web Conference*, pages 195–208, Boston, MA USA.
<http://huskysearch.cs.washington.edu/papers/www4/html/Overview.html>.

Selberg, E. and Etzioni, O. (1997). HuskySearch Home Page.

<http://huskysearch.cs.washington.edu>.

Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. (1998). Analysis of a Very Large AltaVista Query Log. Technical Report 1998-014, Compaq Systems Research Center, Palo Alto, CA.

<http://gatekeeper.dec.com/pub/DEC/SRC/technical-notes/abstracts/src-tn-1998-014.html>.

Sullivan, D. (1999). Search Engine Watch.

<http://www.searchenginewatch.com>.

Yahoo, Inc. (2000). Yahoo! reports fourth quarter and 1999 fiscal year end financial results. Press Release.

<http://docs.yahoo.com/docs/pr/4q99pr.html>.

A Queries from Lawrence and Giles study

1. adaptive access control
2. neighborhood preservation topographic
3. hamiltonian structures
4. right linear grammar
5. pulse width modulation neural
6. unbalanced prior probabilities
7. ranked assignment method
8. internet explorer favourites importing
9. karvel thornber
10. zili liu
11. softmax activation function
12. bose multidimensional system theory
13. gamma mlp
14. dvi2pdf
15. john oliensis
16. rieke spikes exploring neural
17. video watermarking
18. counterpropagation network
19. fat shattering dimension
20. abelson amorphous computing
21. histogram equalization algorithm
22. mixture distance
23. selective attention memory task sequential
24. universal approximation bounds
25. bayesian interpolation