Measuring the Effectiveness of the WWW Search Engines

By

William K. McHenry

School of Business
Georgetown University
Washington DC 20057

Phone: 202 687-3808
Fax: 202 687-4031

mchenryw@gunet.georgetown.edu
http://www.gsb.georgetown.edu/dept/facserv/faculty/mchenryw

September 7, 1997

Published in the Proceedings of the Eighth Annual Conference of the International Information Management Association, Toronto, Dec. 10-12, 1997, pp. 64-69.

Abstract: Eight World Wide Web search engines are compared on the basis of traditional Information Retrieval measurements Recall and Precision. It is found that the most significant factor that is correlated to precision is the topic of the query. Significant groupings at the level of the search engines themselves are found for both recall and precision. The paper also categorizes search engines by information retrieval features, and examines the appropriateness of recall and precision for evaluating WWW search engines.

I. Introduction

Over the past few years, the amount of information stored on the World Wide Web (WWW) has grown enormously. At the beginning of 1997, the creators of the HotBot search engine claimed to have indexed more than 50,000,000 web pages. In theory, well-known traditional information retrieval (IR) techniques should permit these search engines to find any needed information; in practice these techniques threaten to be overwhelmed by magnitudes of this size (Carmel et al., 1989). The problem of language ambiguity, which makes using the indexes so haphazard, looms ever larger as the amount of information stored on the Web grows at unprecedented rates. Meanwhile the question of the practical performance of the search engines has received little attention in the literature (Lesk, 1997).

Over the forty-year history of the IR field, two measurements (precision and recall) have traditionally been used to evaluate whether or not a particular IR algorithm performs better or worse

than another. With millions of inquiries to dozens of search engines every day, the performance of any given algorithm takes on even greater importance. The goal of this paper is to examine eight of the most popular search engines using these two measurements, and to thereby answer the following questions: 1) which search engine is better? 2) are these measurements useful in answering the first question, and 3) if not, what other measurements should we use? If we abandon traditional measurements, how can we relate current experiences with the accumulated knowledge in the IR field?

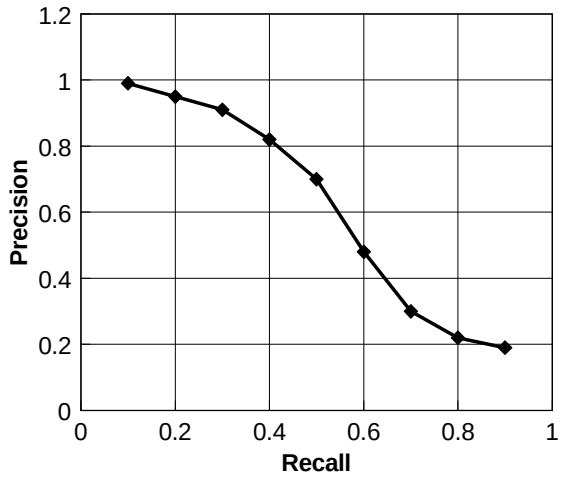

Figure 1: Typical Recall vs. Precision Curve
(Source: Salton & McGill, 1983)

Precision is defined as the percentage of documents retrieved that are relevant to the given query. If, for example, 100 documents are retrieved, and 34 are deemed relevant, precision is 34%. Recall is defined as the proportion of relevant documents retrieved to the entire set of relevant documents in the collection. If the database contains 100 relevant documents, but our query and algorithm combination only retrieved 13, then recall is 13%. The use of very specific terms in indexes and queries ensure a high level of precision, i.e. rejecting useless documents, but may cause us to miss related documents that do not use those specific terms. Conversely an index which is broad enough to encompass all the representations of an idea and therefore to facilitate finding all useful documents (high recall), will permit useless documents to be retrieved because of language ambiguity (Salton & McGill, 1983, p. 160). There are several immediate problems with applying recall and precision measurements to the WWW search engines.

First, what is a document? A document might be considered the smallest unit that is retrieved when a found web address is clicked, i.e. one file that is being retrieved through the WWW. In this case we might think of the web page as being like an article in a journal. If the web page itself does not contain information relevant to the query, then the query failed, even if relevant information is just one or more hyperlinks away. Traditional IR experiments did not allow browsing from one article to another once an article was retrieved, so neither may our measurements of the WWW.

Alternatively, the document might be considered a web site, i.e. all of the pages that branch out from one "home page" within the confines of one server domain. In this case we may consider the retrieved file to be like a chapter in a book. Traditional IR experiments have dealt with databases that are on-line, consisting primarily of abstracts or full texts of articles, and not with sources as massive as books. One can argue that by excluding linked pages from a measurement of relevance, we are ignoring the very property of the WWW that makes it useful. This may be so, and we shall take up this question again, but clearly, to be consistent with past IR research, we must not allow links to be followed.

Second, how can we measure recall when the WWW itself is so big? Finding every relevant page about a subject can require a prohibitive level of effort without any final assurance that all pages have indeed been found. Manual examination of documents, as in the early IR experiments, cannot be applied to millions of web pages. The best we can do is to use every available search technique, and

to pool the relevant documents together to form one large set. This technique has been used successfully in the recent TREC experiments (Harman, 1993), in which a large textual database was used for experiments involving dozens of algorithms processing the same queries.

A third question is why do we care about recall in the first place? Some users are perfectly happy to stop as soon as they find one or two hits that are relevant, but others want to know that they have looked at every available source. More substantively, high recall gives the assurance that the most important and most useful sources will not be overlooked. Since the veracity of any individual page may be questionable, especially in the world of "spin" and "counter-spin" in which we live, the assurance of finding close to everything is important. Otherwise we risk a new form of information tyranny that is based on the arbitrary or willfully misleading performance of a search engine algorithm that is completely hidden from the ordinary user (c.f. Corn, 1996).

Despite the numerous difficulties involved, recall and precision in fact capture what we are interested in: did we get everything we wanted without getting anything we didn't want? If they can be applied to the WWW search engines, then we can make comparisons to previous research.[1] They remain the only generally accepted measurements used in the IR community (Harmon, 1993). In the following sections we examine differences in the search engine algorithms, the experiment that was carried out, its results, and some conclusions. This research is based on experiments done at the end of November, 1996. It describes the status of the search engines at that time; numerous new features have now been added to some of the search engines, while some that were still popular at that time have faded in popularity.

2. Differences in the Search Engines

Over the past forty years, researchers in the IR field have attempted to find automated ways to create indexes by which humans could easily find relevant documents in large document collections. It is not always clear how these generic solution types map into the specific algorithms used in the WWW search engines, especially since those algorithms have not necessarily been described in full on the

[1]See, for example, the comparison of the search engines performed by Internet World Labs (Venditto, 1996). In this test, only the first 25 pages retrieved were examined. Only keywords were used; there was no formulation of a query subject *per se.* No recall and precision statistics are reported, and therefore the comparisons of the engines tend to be quite impressionistic. Westera (1996) examined the relevancy of a number of query results, looking at the first and last five pages retrieved, or if more than 100 were retrieved, pages 100-104. Westera also provided some longitudinal data. Chu and Rosenthal (1996) only examined the summaries of the first ten pages retrieved. Tomaiuolo (1995) also examined just the first ten pages retrieved. Feldman (1997) examined at least the first ten pages retrieved, but did not report any formal statistics. She did use a variety of queries and attempted to ground her analysis in search engine features. Lake (1997) reports on the PC Computing search engine "Shoot Out" as if it were a sporting event; apparently the competing teams tried to find answers to questions, and the only criterion of success was whether or not they found the answer in the given time frame. Haskin (1997) provides only an impressionistic view of some searches.

Web pages or in the literature. Table 2 attempts to perform this mapping, but may be incomplete or incorrect in a few cases. For each engine, the table reflects the most advanced features available in the engine, even if those features had to be accessed by going to another page or selecting "advanced query."

AltaVista:	http://www.altavista.digital.com/
Excite:	http://www.excite.com/
HotBot:	http://www.hotbot.com/
InfoSeek Ultra:	http://ultra.infoseek.com/
Lycos:	http://www.lycos.com/
Magellan:	http://www.mckinley.com/
OpenText:	http://www.opentext.com/
WebCrawler:	http://www.webcrawler.com/

Table 1: Search Engine Addresses used in Nov.- Dec., 1996

These search engines were chosen because they are the most popular and the most written about. The table is based on Morgan (1996) and the descriptions of the engines provided on the web pages themselves. Appendix One contains an explanation of each of the search engine techniques listed in the table. The rows in the table are sorted by the frequency of the use of each technique. The table shows that the approaches adopted by the search engine designers vary considerably. Only Boolean operators appear in almost all the engines. Proximity, stop words out, and tagged fields are the next most popular features, appearing in five search engines each.

Technique/Search Engine (Number of search engines using the technique)	AltaVista	Excite	Hotbot	Infoseek	Lycos	Magellan	Open Text	Web Crawler
Boolean Operators (7)	YES	YES	YES	YES	NO	YES	YES	YES
Proximity (5)	YES	NO	YES	YES	NO	NO	YES	YES
Search Tagged Fields (5)	YES	NO	YES	YES	NO	NO	YES	YES
Stop Words Out (5)	YES	NO	YES	YES	NO	NO	YES	YES
Query Expansion (4)	NO	YES	NO	YES	NO	YES	NO	YES
Partial Match (4)	YES	NO	YES	YES	YES	NO	NO	NO
Relevance Feedback (3)	NO	YES	NO	NO	NO	NO	YES	YES
Term Frequency and Inverse Document Frequency (3)	NO	NO	YES	YES	NO	NO	NO	YES
Vector Match (3)	NO	NO	YES	NO	YES	YES	NO	NO
Term Frequency Alone (2)	YES	NO	NO	NO	YES	NO	NO	NO

Table 2: Search Engine Attributes as of December, 1996

On the basis of these various features, we may expect the performance of the search engines to vary. Since this research was exploratory, no particular hypotheses were made in advance regarding which engine might be better and which technique might perform better than another.[2]

3. The Experiment

Team	TOPIC AREA	QUESTION	Mean Precision in KWS-assigned grouping
3	SCIENCE	Can the meteorite discovered in Antarctica be considered as evidence of life on Mars?	Mean= 43.4% STD = 30.1%
5	FOREIGN AFFAIRS	What are the possible causes of Gulf War Syndrome?	Mean = 35.9% STD= 25.3%
6	SPORTS	What steps are being taken to prevent franchise/free agency in the NFL?	
7	BUSINESS	How can we make money from Jai Alai?	
8	POLITICS	What are the current proposals for U.S.Federal campaign finance reform?	
9	COMPUTERS	Information processing (OS) companies and how they've tackled the issue of Japanese language processing	Mean = 14.9% STD = 15.8%
10	MEDICINE	Who will pay for telemedicine?	
4	ARTS & ENTERTAIN-MENT	What Washington-area museum exhibitions will be showing the week of December 23-27?	Mean = 11.2% STD = 10.2%
1	HISTORY	How did innovations in medieval weaponry change/influence castle-building?	Mean = 4% STD = 4.6%
2	CRIME	Who are the major non-American serial killers?	

Table 3: Query Topics Grouped by Precision Performance

A class of 43 masters students was divided into ten teams of four or five people. Each team came up with a query (Table 3) and submitted the query to each of eight search engines at roughly the same time at the end of November, 1996. The queries were selected from ten different subject matter areas in order to reflect the diversity of the information on the WWW. Using each search engine, students were told to retrieve as close to 100 working WWW pages as possible, and to record whether or not each retrieved page was relevant. In order to ensure as much consistency as possible, all team members had to agree on what made a page relevant before

[2]Sullivan (1997) provides extensive information about the evolving search engine field, including some comparative information. Webber (1997) provides links to a number of papers about the search engines. Notess (1997) provides a useful update of new search engine features. Basch (1997) also reviews search engines features. Westera (1996) provides a bibliography on this subject and also tables of search engine features.

starting the queries. Students were not allowed to follow links. In addition, students were also not allowed to refine their queries, taking whichever produced the best results. Instead, students were encouraged to think in advance of all useful search terms and how to use the apparatus of each individual search engine to get the best results on the first try. If we had allowed multiple tries (known as "routing" in the IR literature), some students undoubtedly would have taken it much further than others. Students were encouraged to use as many features of each engine as possible in order to take advantage of the unique aspects of the engine by which its authors claimed efficacy.

Precision for each search engine was calculated by dividing the number of relevant pages found by the total number of retrievable pages actually retrieved by the students (usually 100 or less).[3] As in the TREC-1 experiments, the teams pooled the relevant pages from all eight search engine queries, eliminated the duplicates, and counted the total number of relevant pages found (Harmon, 1993). Thus ten measurements of recall and precision were made for each search engine.

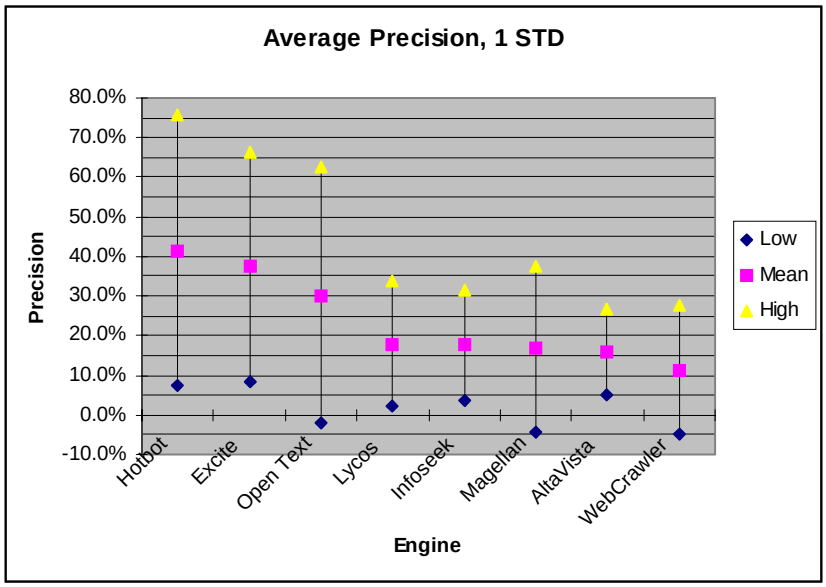

Figure 2: Average Precision showing one standard deviation

4. Results

The average recall across all eight search engines was 13.9% (STD=11.6%). Assuming that the pooling method accurately reflects the total number of pages that are relevant to these queries on the WWW, then the typical search engine found less than one in five on the first 100 pages. The mean precision was 23.6% (STD=24.5%).

In order to rank the search engines, it is necessary to compare the mean precision for each engine, and to compare the mean recall for each engine. A visual inspection of the graphs (Figures 2 and 3) leads one to doubt that there are significant differences across many of the means. To statistically examine the data, we used a data mining tool called KnowledgeSEEKER IV, Version 4.2.2 (KWS), which tests many possible clusters and finds clusters that are statistically significant, taking into account the requisite Bonferroni adjustments (KnowledgeSEEKER IV, 1997).[4] Two tests were performed; in the first the

[3]This definition differs from the traditional one. The reasons for this are explained below.

[4]KnowledgeSeeker uses the cluster technique. The reference manual explains: “When the Cluster method is applied to continuous dependent variables, it is similar to the CART technique (see L.Breiman et al, *Classification and Regression Trees*, Wadsworth, 1984). The Exhaustive method (see Biggs et al, "A Method of Choosing Multiway Partitions for Classification and Decision Trees", *Applied Statistics*, 18, 1, 1991, pp. 49-62) was developed to address shortcomings in the Cluster method.... [T]he disadvantage of the Cluster method is that it is overly conservative.” For this research, the cluster method was used in all cases except the split for partial matches; in this case the exhaustive method was used.

dependent variable was precision, and in the second it was recall. In both tests the independent variables were search engine, team, and all of the features listed in Table One with the exception of "Boolean," "relevance feedback," and "search tagged fields." Using multiple iterations and hence relevance feedback was not permitted, nor was it likely that using a tagged field such as URL would help perform these queries. Boolean was omitted because it is present in virtually all search engines.

The precision-based analysis indicates that the most statistically significant grouping is the query topic as represented by the team number (P=0.01). These topic areas are shown in Table 3, ordered by decreasing precision, in the groups found by KWS. It is interesting to note that the most successful query involved an on-going current event (Team 3). The next four queries (Teams 5-8) all included at least one unambiguous concept that could be described in a fairly unusual way. Teams 9 and 10 used queries that were somewhat more abstract, yet still had identifiable hooks. The query of Team 4 was difficult because it included the concept of time. Finally, there may have been little information on the WWW about the subjects of the last two queries (Teams 1 and 2).

The next most significant grouping is search engine. Excite, Hotbot, and OpenText have a mean precision of 36.4% (STD=31.1%), while the other five have a mean of 15.9% (STD=15.5%) (P=0.03). Both of these groups may be split one level deeper. For the Excite-Hotbot-OpenText group, the split is based on the query, and is identical to the split described above and shown in Table 3 (P=0.03). For example, Team 3 achieved an average precision of 71% using these three search engines. For the other five search engines, there is also a split based on team, but the teams are fewer (P=0.02). Teams 1 and 2 still show the lowest precision, teams 3 and 4 are next, and the rest of the teams form one large group. Thus these results confirm the often reported result that certain techniques work better with certain queries (c.f. Harman, 1993, pg. 41). However, the fact that there were no statistically significant splits based on the specific techniques listed in Table 2 may indicate that the sample size was too low, or that the students did not make use of particular features in ways that would have revealed their effectiveness or lack thereof.

No statistically significant further groupings were found. Thus a rough ordering of the search engines based on precision would be:

First Tier: Excite, Hotbot, OpenText
Second Tier: AltaVista, InfoSeek, Lycos, Magellan, WebCrawler

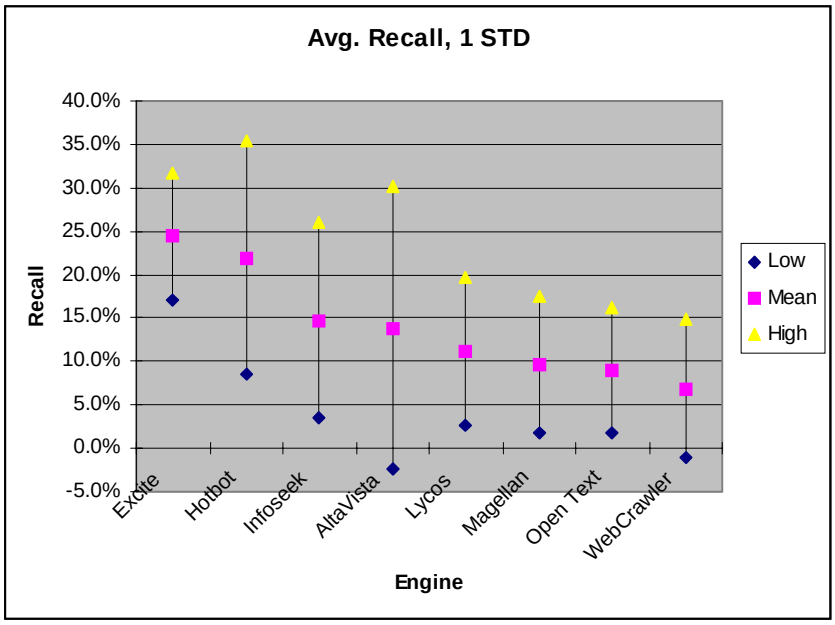

Figure 3: Average Recall Showing One Standard Deviation

For recall, the most statistically significant grouping is search engine (P=0.01). Hotbot and Excite have a mean recall of 23.2% (STD=10.6%), while the other six have a mean recall of 10.8% (STD=10.2%). No other variables proved to be significant in clustering the Hotbot/Excite group. However, the other six search engine results may be distinguished further by whether or not they allow wild cards (partial matches) in the search (P=0.06). For those that do not (OpenText, Magellan, WebCrawler) , the average recall was just 8.3% (STD=7.2%), while for those that do (AltaVista, InfoSeek, Lycos), the average recall was 13.3% (STD=12.1%).

Therefore, a rough ranking of the search engines on the basis of recall would be:

Top Tier: Excite, Hotbot
Second Tier: AltaVista, InfoSeek, Lycos
Third Tier: OpenText, Magellan, WebCrawler

As noted above, precision was calculated as the number of relevant pages retrieved divided by the total number of retrievable pages in the sample taken by the students. Dividing by the total number reported as "retrieved" by the search engine is not practical. Consider the results of entering the terms "campaign finance reform" into the eight search engines (Table 4). If 30 relevant pages were found in the first 100, and the rest are assumed to be irrelevant, then even OpenText has a precision of just 1.6%. AltaVista has a precision of 0.0066%. AltaVista used a default Boolean operator of OR and did not employ Inverse Document Frequency; hence it reported any document containing any of the query terms as potentially relevant. To use this number found in the denominator would be to render all AltaVista results meaningless for the sake of comparison.

Search Engine	Number Retrieved
Opentext	1,882
Lycos	2,709
HotBot	15,983
Magellan	53,618
Webcrawler	61,214
AltaVista	450,330
Infoseek	562,871
Excite	1,021,928

Table 4: Number of pages retrieved for query "campaign finance reform" on 9/7/97

In addition, in traditional IR experiments, the same document collection is tested against

various algorithms. However, the search engines are working from different universes of documents. The spiders that collect pages from the web move at different speeds, the indexes are updated at different rates, and old materials are expunged at different times. The sizes of the collections indexed may vary substantially. Traditional IR collections were usually limited to a single subject area, whereas the range of information on the Web potentially encompasses all human knowledge.

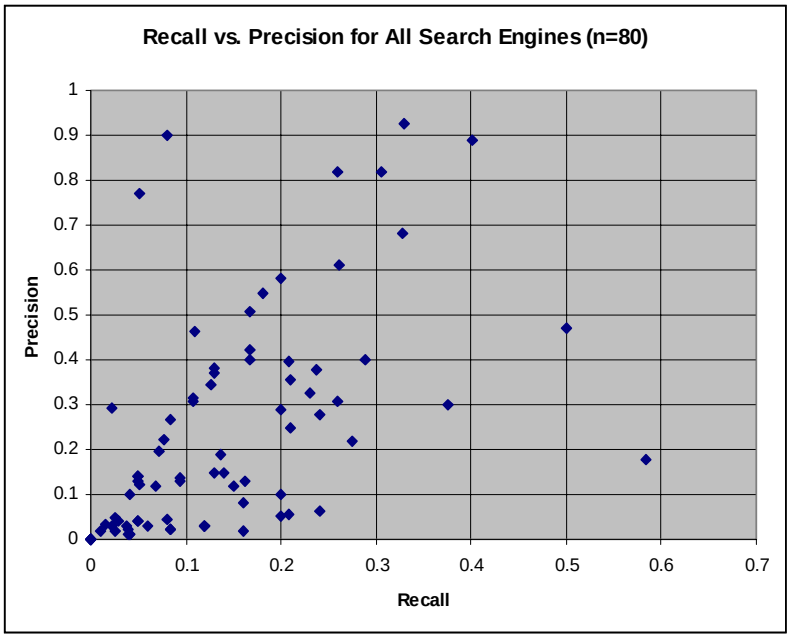

Figure 4: Recall vs. Precision for all 80 Observations

Once we have agreed that we cannot use the total number found in the denominator, using the sample size instead is the only other measure that makes sense. It is intuitively appealing to compare how the search engines do on a number that is close to the number that a typical user would likely review. However, this leads to another difficulty.

The Recall vs. Precision curve (Figure 4) does not at all resemble the curve in Figure 1. This result can be attributed to the specific measures used in this experiment. If the number of pages found by a query is held to be a constant, then recall and precision are correlated.[5]

Given the use of the 100 page cutoff (i.e., a constant number found), how is the variability in Figure 4 to be explained? We do not believe that there is a hidden typical recall-precision curve in the data. Rather, the number found varied from case to case for several reasons. First, some search

---

[5]The explanation for this result can be found in the definitions of recall and precision used in this experiment. Let R be Recall, P be precision, NF be the number found by the query, and NR be the number found that are relevant. Then for the i^{th} search engine, $R_i = \frac{NR_i}{\sum_{k=1,10} NR_k}$ and $P_i = \frac{NR_i}{NF_i}$. Note that the term in the denominator for recall, the sum of all relevant pages found (minus duplicates), is a constant. If the term NF (number found) is held to be a constant, then we may solve for the relationship between recall and precision and find that: $R_i = \frac{P_i \ NF_i}{\sum_{k=1,10} NR_k}$. As long as NF is a constant, there is a simple linear relationship between recall and precision. Furthermore, the slope of the line formed by this equation is never negative.

Team	Number Relevant Found
8	288
3	222
5	194
6	190
10	139
7	119
4	80
9	50
2	25
1	24

Table 5: Number of Unique Relevant Pages Found by Each Team

engines and queries produced less than 100 pages (Table 5). Second, some teams retrieved 100 pages including links that could not be followed (expired, protected by security, etc.), while other teams retrieved 100 pages that could be followed, even if that meant retrieving more than 100 total pages. In about 48% of the cases, the teams retrieved 100 or more useable documents. In the next 33% of the cases, they retrieved 80-99 documents.

Thus we must conclude that the traditional measure of precision may not always be applicable when studying the effectiveness of the WWW search engines.

Another measure that does not show up in traditional IR experiments is the number of documents that could not be retrieved. Unfortunately the WWW search engines' indexes are often out of date. Table 6 shows the average percentage of unreachable links found for each search engine.

Engine	Average Number of Unreachable Links
Excite	7.9%
InfoSeek	10.1%
Lycos	11.4%
Hotbot	14.8%
OpenText	16.2%
AltaVista	16.7%
Webcrawler	17.6%
Magellan	26.7%

Table 6: Unreachable Links

The students were also asked to record the relative value of each retrieved page as they evaluated the pages in the sequence given by the search engine. For example, after seeing pages 1-4, the level of information satisfaction may be 0%, after page 5 it may have jumped to 30%, and by page 10, it may be at 100%. At this point the searchers may have stopped. The level of satisfaction does not measure whether the searcher retrieved everything he or she should have retrieved in order to find out everything that is necessary, but does represent a practical measure of when the searcher may stop. Some students did not fulfill this part of the task, some felt they never reached this point, and some found that it came fairly quickly. This suggests the possibility of creating an evaluation method based on the perceived incremental value of the next page retrieved. Figure 5 shows the satisfaction data recorded by Team 3. Note that in one case the level of satisfaction actually drops; this reflects the fact that the team realized, upon seeing a new web page, that the previous number assigned was too high.

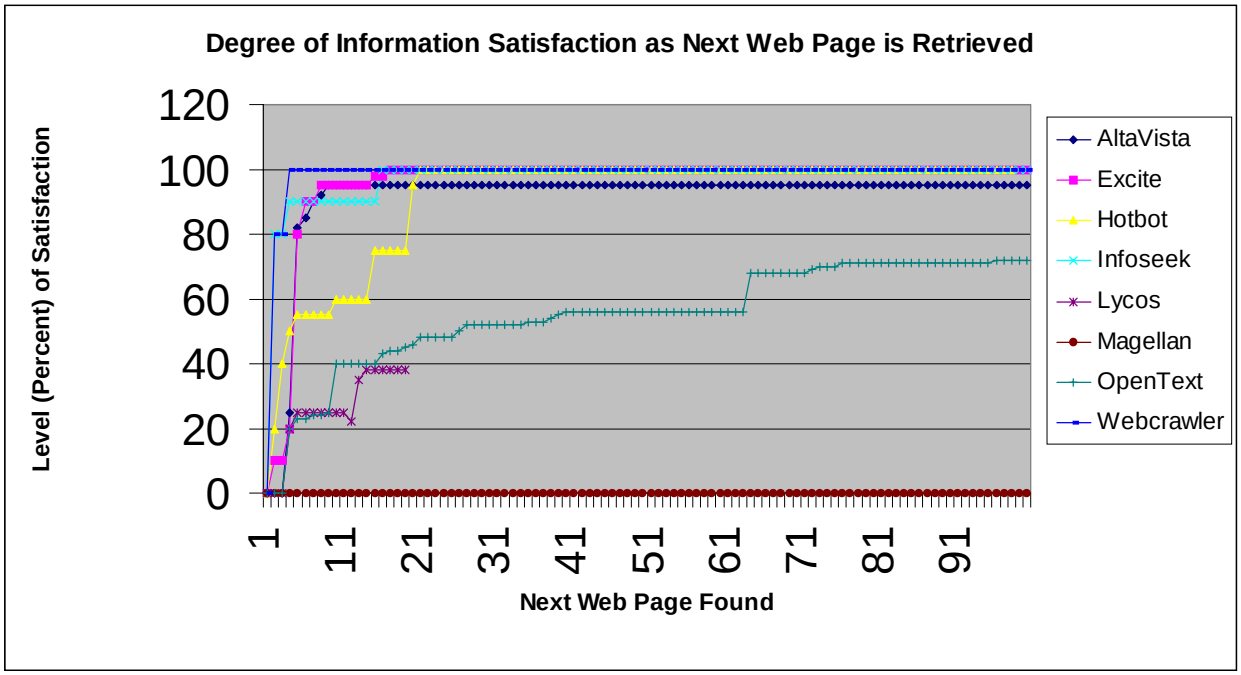

Figure 5: Team 3 Search Satisfaction Data

5. Conclusions

The traditional measures of recall and precision are both problematical when applied to the WWW. However, by using the pooling technique to address the problems with recall, and by measuring precision as a function of the pages actually accessed, it is possible to make a rough estimate about which engines work better across a fairly wide range of queries. The most important determinant of search engine performance was the topic of the queries themselves. Whether or not the search engine allowed partial matches influenced the recall in some cases.

Developing a methodology that tracks the user's changing perceptions of information satisfaction as he or she examines the retrieved web pages would be useful. This kind of measure would address the order in which the pages are served by the search engine, which only enters the recall and precision measures above at whatever evaluation cut off point is chosen.

Another possible measurement could relax the constraint of not following hyperlinks. It could be centered around the number of clicks it takes to reach information satisfaction. In this scenario a search engine is only a jumping off point. If the search engine suggests one page that then has "golden" links to many other pages, one could argue that it has done its job, even if the other 1,000,000 pages retrieved are totally irrelevant. To measure this type of behavior on the Web would require observation of more general search strategies: does the user start from a bookmark, a search engine, a recommendation from a friend, a printed guide to the web, a link that is found by chance while doing another task, or by another means? Until machines can "understand" language in the same way that humans do, there will always be problems with algorithms for finding needed information, but the WWW has opened up myriad possibilities that go well beyond the traditional

means supplied by the Information Retrieval field. Measuring the efficacy of various approaches will continue to be a daunting task.

Acknowledgments: I am very grateful to the 1996-1997 first year students in the Georgetown CCT program (http://www.georgetown.edu/grad/cct) who put forth the great effort to evaluate all those web pages. Also many thanks to Professor Charles Iacovou, Georgetown University School of Business, who helped me to refine my ideas and tighten up the manuscript. Any errors of course remain my sole responsibility.

Appendix 1: Search Engine Techniques Used in the WWW Search Engines

For overviews of search engine techniques and models, see Kantor (1994), Salton and McGill (1983), and Paijmans (1992).

- **Vector Match** - The query is formulated as a vector of keywords which are then matched against the vectors of key words for each document in the collection. Those which are "close"enough are selected as matching the query. Sometimes the words are weighted to reflect better the specific nature of the query.

- **Term Frequency Alone** - Words which appear more frequently in the document are inferred to have more significance and are weighted more heavily. In assuming some search engines work this way, web page designers would repeat words on their pages to get them to come up towards the top of the list of retrieved pages.

- **Term Frequency and Inverse Document Frequency** - This technique not only counts the number of words in the document but finds how rare they are in the whole collection. Terms which appear frequently in one document but rarely in any others are deemed most significant

- **Query Expansion** - Query terms are expanded by adding related terms from a thesaurus.

- **Boolean Operators** - The most common technique in which the user specifies certain combinations of terms that may or must be present. For example: "democratic **and** national **and** convention."

- **Relevance Feedback** - The user selects the most relevant documents from the first iteration and the terms in those documents are used to refine the query for the next iteration.

- **Proximity** - The search engine is able to take into account the relative position of the words in a sentence or paragraph.

- **Partial Match** - The search engine is able to match parts of words, e.g. democr* matches both democracy and democratic. Included in this category is the idea of stemming, in which the stems or roots of words are stored in the index instead of the whole word.

- **Stop Words Out** - Common words such as “a,” “the,” and “and” are excluded.

- **Search Tagged Fields** - The search engine distinguishes between HTML fields such a <head>, <body>, <meta> and so forth, and is also able to handle searches for URLs, etc.

References Cited

Basch, Reva (1997). Find Anything Online. *ComputerLife Online.* http://www.zdnet.com/complife/fea/9708/findny10.html.

Carmel, E., McHenry, W., & Cohen, Y. (1989). Building Large, Dynamic Hypertexts: How Do We Link Intelligently? *Journal of Management Information Systems* 6, 2, 33-51.

Chu, Heting and Rosenthal, Marilyn (1996). Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. *ASIS 1996 Annual Conference Proceedings, Oct. 19-24, 1996.* http://www.asis.org/annual-96/ElectronicProceedings/chu.html

Corn, David (1996, July 7). Anatomy of a Netscam. *The Washington Post*, C5.

Feldman, Susan (1997, May). "Just the answers, please": choosing a Web search service. *Searcher* 5, 5, 44.

Harman, Donna (1993). Overview of the first TREC Conference. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Korfhage, Robert, Rasmussen, Edie, & Willett, Peter, eds., New York: ACM Press, 36-47.

Haskin, David (1997, Sept) . The Right Search Engine: IW Labs Test. *Internet World Online*. http://www.iw.com/1997/09/report.html.

Kantor, Paul (1994). Information Retrieval Techniques. *Annual Review of Information Science and Technology*, 29, 53-90.

KnowledgeSEEKER IV (1997). Version 4.2.2 Help Files, Build Date March 13, 1997. Toronto: Angoss International Limited (see http://www.angoss.com to download demo).

Lake, Matthew (1997, Sept). 2nd Annual Search Engine Shoot-Out. *PC Computing.* Http://www4.zdnet.com/pccomp/features/excl0997/sear/sear.html.

Lesk, Michael (1997). Real Life Information Retrieval: Commercial Search Engines (Panel Session Abstract). *Proceedings of the 20th Annual International ACM SIGIR Conference on Research*

and Development in Information Retrieval, Philadelphia, Pa., July 27-July 31, 1997. Belkin, Nicholas, Narasimhalu, Desai, & Willett, Peter, eds. New York: ACM Press, 333.

Morgan, Cynthia (1996, Nov.) The Search Is On -- Finding the right tools and using them properly can shed light on your Web search efforts. *Windows Magazine*, 218-230.

Notess, Greg (1997, Sept/Oct). New features of the Web indexes. *Online* 21, 5, 52-55.

Paijmans, Hans (1992). An Inventory of Models in Information Retrieval. *ITK-workshop - AIIR* - May 7, 1992, Tilburg. http://pi0959.kub.nl:2080/Paai/Ai_ir/ai_ir.html.

Salton, Gerard & McGill, Michael (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Sullivan, Danny (1997, Oct 6). Search Engine Watch. http://searchenginewatch.com.

Tomaiuolo, Nicholas G. (1995). Quantitative Analysis of Five WWW "Search Engines": Results of 200 Subject Searches performed Oct. to Dec. 1995. http://neal.ctstateu.edu:2001/htdocs/websearch.html.

Venditto, Gus (1996, May). Search Engine Showdown: IW Labs Tests Seven Internet Search Tools. *Internet World*, 78-86. See also: http://www.iw.com/1996/05/showdown.html.

Webber, Sheila (1997, July 23). Business information sources on the Internet. Http://www.dis.strath.ac.uk/business/searchindepth.html.

Westera, Gillian (1996, Oct). Robot-Driven Search Engine Evaluation Overview. http://www.curtin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy/.