First 20 Precision among World Wide Web Search Services (Search Engines)

H. Vernon Leighton*
Winona State University Library, Winona, MN 55987-5838.
E-mail: vleighton@winona.msus.edu

Jaideep Srivastava
Computer Science, University of Minnesota, 4-192 EECSci, Minneapolis, MN 55455.
E-mail: srivasta@cs.umn.edu

Five search engines, Alta Vista, Excite, HotBot, Infoseek, and Lycos, are compared for precision on the first 20 results returned for 15 queries, adding weight for ranking effectiveness. All searching was done from January 31 to March 12, 1997. In the study, steps are taken to ensure that bias has not unduly influenced the evaluation. Friedmann's randomized block design is used to perform multiple comparisons for significance. Analysis shows that Alta Vista, Excite and Infoseek are the top three services, with their relative rank changing depending on how one operationally defines the concept of relevance. Correspondence analysis shows that Lycos performed better on short, unstructured queries, whereas HotBot performed better on structured queries.

Introduction

As we state at the outset, the data for this study were collected in the first part of 1997. An earlier version of this study was published on the Web promptly in June of 1997, when the data was still fresh. The data are now of historical value. The point of having this paper appear here is to have our methods reviewed by peers in the information science profession and to give the study a permanent location in the published literature.

Since this study was done, every search service in it has undergone major change in its features, including its ranking and retrieval strategy, as anyone who follows news reports on search engines knows (Sullivan, 1998a,b,c,d,e). To give some examples, Lycos, which had the lowest median scores here, has added Lycos Pro, where users can customize each search's ranking algorithm. Lycos has bought Wired Digital and with it, HotBot. There has been a trend toward raising the rank of more popular or more heavily cited pages. One- and two-word queries are now often rerouted to Web directories so that users can enjoy the topical coverage that comes with a classification scheme. Clustered pages are now grouped under a link that allows users to view all pages retrieved from the same site in Infoseek, Lycos, and HotBot. Infoseek and Excite have raised the ranking of reviewed sites, and Alta Vista has incorporated the Ask Jeeves service into their results. The results in this study are not useful in helping the user choose a search engine today.

The most important features of this study are the blinding procedure to lessen evaluator bias, the use of medians and statistical techniques for non-normally distributed data, the use of confidence intervals for testing the least significant difference, and the use of correspondence analysis to compare queries to services. Aspects to be concerned with are whether the definitions of relevance are adequate, whether the relevance categories as defined match with the judgments of real users, and whether the metric formula is recommended for future studies.

The Present Study

In this study, we compare five commercial World Wide Web search services, which are also commonly called *search engines*: Alta Vista, Excite, HotBot, Infoseek, and Lycos. Our test suite is a set of 15 questions that are submitted to all of the services. The measurement that we use, *first 20 precision*, rates the services based on the percentage of results within the first 20 returned that were relevant. We use a variant of first 20 precision that adds weight for ranking effectiveness. We then analyze the sta-

Received July 9, 1998; revised February 22, 1999; accepted March 17, 1999.

* To whom all correspondence should be addressed.

tistics and compare the services. Searching for this study was done from January 31 to March 12, 1997.

In the realm of search engine studies, many studies comparing relevance have been conducted. We began this study because many previous studies have arrived at conflicting conclusions as to which services are better at delivering superior precision and because most of those studies have either had small test suites or have not reported how the study was conducted. This study compares all search services that had been recommended in 1996 for their performance at delivering relevant results for the user, and uses a carefully controlled and documented experimental design to ensure valid results. It is our opinion that evaluating pages for relevance without blinding the sources to the evaluator is a major problem for search service studies.

The results of the various experiments indicate that there are three services, alphabetically Alta Vista, Excite, and Infoseek, that are superior to the other two studied, HotBot and Lycos, in terms of first 20 precision. How these top three rank compared to each other depends on how one defines relevant. The chief problem with HotBot is the frequent occurrence of duplicate links; the chief problems with Lycos's performance are irrelevant and inactive links. Lycos's irrelevant links are often caused by the lack of a *required* operator (often denoted by a + prefix). Neither HotBot nor Lycos seems to implement case sensitivity.

In the next two sections of this report, we define the problem, critique related work in the field and summarize the known features of the search services under consideration. In part three, we explain the methodology used to develop the test suite, to conduct the searches, and to evaluate the results from the search services. Next, we explain the evaluation criteria, discuss how we defined the concept of relevance, and explain the formula by which these evaluated results are used to calculate first 20 precision. In part five, the experimental design for the primary statistical tests is described. Finally, we give results, discuss their meaning, and offer conclusions. For an earlier, unreviewed version of this paper, including the specific search expressions that were used and other data, see Leighton and Srivastava (1997).

Problem Statement and Related Work

This work attempts to compare the precision of major Web index services in an objective and fair manner with regard to general subject queries that may be posed in the undergraduate academic setting. For such a study one must develop an unbiased suite of queries to test those services. Then one must design fair methods for searching the services and evaluating the results of the searches. At many points in the design, it is possible to subtly favor one service over another. Conscious or unconscious bias must be guarded against. Finally, one must arrive at some metric for measuring precision and use that metric to analyze performance.

For this study, we chose Web search services that attempt to provide access to a large portion of the pages freely available on hypertext transport protocol (HTTP) servers worldwide. There are other types of services that we did not include. Selective Web databases—C|Net, Excite Reviews, Magellan, and Yahoo—are certainly helpful at cutting through the user's information overload, and may even be better at satisfying user information needs than the major search services, but they are much smaller and are to some degree manually selected. Traditional vendors of databases, such as Dialog (now Knight–Ridder Information Inc.), are also different, in that their databases are often commercially developed and professionally edited records of commercially published texts in academic, scientific and technical areas. Although all of these other services may be able to answer many of the same questions, and may therefore be comparable in terms of satisfying some level of information need, they really are different from the major search services in scope and purpose and were not included.

This study has a test suite large enough for valid statistical analyses. The queries have been obtained in such a way that the researcher did not personally pick the queries in the test suite. The searching has been done to minimize the possibility of favoring the service queried first, or the one queried last. The resulting pages have been blinded so that the evaluator would not know from which service they came. These steps have been taken to prevent the various types of bias that can unfairly influence the results.

Most published precision studies have had test suites that were too small for statistical usefulness. Ding and Marchionini (1996) studied first 20 precision, but used only five queries. Leighton's 1995 study only had eight queries, and there have been a host of minor reports that purport to judge precision based on three, two, or even one query.

Chu and Rosenthal (1996) tested first 10 precision, had 10 queries (enough for statistical comparisons), recorded crucial information about how the searching was conducted, and performed some statistical tests. However, they only studied three search services, and they did not subject their mean precision figures to any test for significance. Gauch and Wang (1996) had twelve queries, studied almost all of the major search services (and even the major metasearch services) and reported first 20 precision, but did not test for significance in the differences reported. Tomaiuolo and Packer (1996) studied first 10 precision on 200 queries. They did list the query topics that they searched, but they used structured search expressions (using operators) and did not list the exact expression entered for each service. They reported the mean precision, but again did not test for significance in the differences. They reported that they often did not visit the link to see if it was in fact active. Nor did they define the criteria for relevance.

After our study was conducted, Clarke and Willett (1997) conducted a study of both first 10 precision and approximate recall using a pooling method on three search engines (Alta Vista, Excite, and Lycos). They searched for 30 queries dealing with topics in library and information

science using when possible unstructured search expressions. Thirty queries allows one to treat the data as normally distributed. They used the Friedmann test for significance and found Alta Vista significantly more precise then Lycos.

Studies reported in popular journals were often vague about how many or exactly what queries were searched. Venditto (1996) used first 25 precision, but did not report how many queries were used nor what the exact statistics were. Scoville (1996) used first 10 precision and gave exact mean scores, but explained neither how many queries were used, nor whether the differences in mean were significant. Munro and Lidsky (1996) also used first 10 precision in a large 50 query by 10 search engine study, but did not list the queries or the statistical analysis. From their description, it is clear that their range of types of queries was much wider than that used in this study. They reported their results as a scale of four stars, indicating that more exact numbers would be easily misleading (perhaps because they tested for statistical significance). Their study was part of a larger survey of search engines by Singh and Lidsky (1996).

None of the studies in the related literature indicate that an attempt was made blind the service of origin of the links evaluated. Unless this step is taken, there must always be the question of bias, conscious or unconscious, on the part of the evaluator. All of the studies used straight first X precision, without adding weight for ranking effectiveness.

The Search Services

What techniques do these search services use to retrieve and rank Web pages? For a thorough investigation of what their features were at the time that we collected our data, see Maze, Moxley, and Smith (1997). As they note, while the exact algorithms for indexing and ranking are commercial secrets, the companies will publicize some general information about these techniques, and testing can reveal other details. We will briefly summarize important details of each service below as described in Maze et al.'s work.

Alta Vista stored information on the position and tag of every word on every page, using no stopwords, and was one-way case sensitive (where capitals match capitals, and lower case match either). In preparing for our study, we understood the Alta Vista documentation to say that in its advanced search, if the relevancy ranking box were left blank, it would use all of the search terms in its ranking algorithm. Maze, Moxley, and Smith (1997, p. 104) claim that if the ranking box were left blank, the ranking algorithm would not be engaged at all. In simple searches, the ranking algorithm was engaged on all words all the time. Alta Vista's algorithm ranked pages higher if the terms were in the title, if they were more frequent and if they were proximate to one another.

Excite indexed the full text of each Web page minus a predefined list of stopwords. Excite was the only service claiming to index by latent semantics, grouping pages with terms on similar topics, so it should have retrieved pages that used synonyms of the terms in the search. Ranking did not weigh terms relative to their fields, but was otherwise not explained. HotBot indexed most of the words in each page, ignoring predefined stopwords. HotBot claimed to be one-way case sensitive, but did not appear to be in our tests. Maze, Moxley, and Smith observe that the case-sensitivity worked only for commercial names such as NeXT and HotBot, which have unusual case. HotBot added weight to terms using "a ratio of the query words' frequencies in the document to the overall document length," (Maze et al., 1997, p. 123) which may have biased it toward shorter documents. Title and meta tag words were give more weight.

Infoseek indexed the full text of all pages, using no stopwords. It used Xerox Linguistic Technology to match word variants and was one-way case sensitive. In ranking documents, it used the location of the term in the document, the frequency of the term in the document, and the rarity of the term on the Web. It added weight for title and heading words, but not for meta tag words. Lycos did not index all of the words in a Web page, but would use information such as word placement and frequency to choose the 100 weightiest words from the page. It would index the title, HTML heading fields, the first 20 lines and the 100 weightiest words. Lycos was two-case insensitive, and used the search terms as stems, finding words that began with the string the user keyed in.

Methodology

Development of the Test Suite

The development of the test suite requires two steps: One must first choose which information needs will be searched for and second choose exactly what search expression will be submitted to each service. Biases, both conscious and unconscious, can enter the process with either step, as, for example, one may select general subject areas that one knows a given search engine is stronger in than others, or one may choose a form of expression that exploits a feature in one search service that is not available in other services.[1]

Because the type of query is the general subject inquiry in an undergraduate academic setting, the queries are ones actually asked at a University Library Reference Desk. During February 1997, the test suite developer recorded the verbal request of every reference question that he was asked in which the patron specifically requested that the Internet be used as a source of information. We stopped after 10

[1] As an example of subject areas, we noticed that Lycos was particularly weak with pages from commercial firms and on business topics, whereas Excite and HotBot were much stronger in that area. As far as search expression, a good example is Infoseek (1997), where it appeared to us that in some cases the forms of the queries were picked to exploit search features that only Infoseek had. But the Infoseek study made no claims to being unbiased, and the forms in question could be defended as fair in some sense if it could be demonstrated that actual users often perform searches using those forms.

queries were obtained. These verbal requests were neither invented nor selected by someone who knew the abilities of the various services. These topics were supplemented by selecting five queries from another study.[2]

The selection of exactly what search expression to enter is perhaps the single weakest point in the design of this study. Other studies have suffered from bias here.[3] First, one must choose how many words to enter in the search, then choose which words. These queries are usually narrowly defined academic topics and used multiple words, as was done in other studies (Chu & Rosenthal, 1996; Ding & Marchionini, 1996; Tomaiuolo & Packer, 1996; Clarke & Willett, 1997).[4]

In addition to choices about what words to use to search, one must decide whether to use an unstructured collection of words (sometimes referred to as natural language) or to use text structured by operators. If the text is structured, one must choose the appropriate operators (proximity and Boolean) and constraints. Our decision as to how many and which queries would be structured or unstructured was our most difficult one. When conducting preliminary queries, we became uneasy about our own ability to know for each of the five services and for each query exactly what expression would be optimal. Furthermore, as Magellan's Voyeur (Magellan Internet Guide, 1997) indicates, most users do not use operators in their searching. Finally, unstructured queries force the search service to do more of the work, ranking results by its own algorithm rather than the constraints specified by the operators. Because of all of these factors, we chose the unstructured—or as we refer to it, simple—text as the preferred expression, and only chose the structured text when, without it, the topic was too easily open to multiple interpretations. Whether we made optimal, or even adequate, choices, is an issue for criticism. For example, Alta Vista advanced allows one to weigh some terms with more relevance. We did not use that feature. Alta Vista's results might improve with the correct application of that ranking function. To allow for that criticism, we have recorded the exact settings for the forms submitted to the services in Appendix A of Leighton and Srivastava (1997). We categorized the queries as *structured*, *simple*,or *personal name*.[5] There are seven simple queries, seven structured queries, and one personal name in the final test suite.

Search Method

When studying search services, it is important to perform an operation on one service as soon as possible after it was performed in the other services, so as to reduce the possibility that the Web's state has changed between operations. In the present study, a given query was searched on all of the search engines on the same day. For most queries, the engines were all searched within half an hour of each other. The first two pages of results were saved into files. That same day or the next day, an automatic Web browser retrieved all of the pages from the results lists within two hours of each other. The stored pages were then evaluated by the researcher over a period lasting from a day to a week.

Evaluation Method

For each query, we wrote down a draft set of criteria for categorizing the links based on the general criteria discussed in Relevancy Categories, below. This was done before evaluating any links. Then, as the links were evaluated, the criteria were adjusted as necessary to take into account the nature of the subject involved.

To prevent biases from clouding our judgment in categorizing individual pages returned by the services, we developed a method of blinding the pages so that for any given page, we would not know ahead of time which service had returned it as a result. A PERL program run by a research assistant was used to strip the URLs of the search service results and load them into batch files. The automated Web browser then retrieved the documents listed in the batch files and saved them with blinding labels so that the source of each page was hidden from the evaluator. The automated Web browser was the *get* program from the PERL suite of web utilities called libwww-perl-0.40.[6]

The evaluator then called up each page using the text editor EMACS, inspected the HTML code and assigned the page to one of these categories (see Relevance Categories below for how these categories are defined): Inactive, zero, one, two or three (duplicates were not discovered until the pages could be matched to search engines). Some unique feature of each page (part of the title, etc.) was jotted down so that if a later page in the evaluation looked similar, a match could be discovered. In this way, even if the evaluation was not evenly or fairly done in other respects, at least the same page would receive the same score throughout the

[2] The suite collection was done at the Maxwell Library at Winona State University, Winona, MN. From the Tomaiuolo and Packer (1996) study, we selected every twentieth query for the first 100 queries, obtaining five queries. Tomaiuolo and Packer also reported recording their topics by getting them in part from real reference questions, and we felt that the addition of their queries should help overcome the location bias of Winona State.

[3] For example, see the forward to Leighton's 1995 study.

[4] If one observes Magellan's Voyeur (Magellan Internet Guide, 1997), or has access to another database of search expressions actually entered by users, one will notice that the typical popular query is only one or two words long.

[5] In the case of personal names, we capitalized the name, and in HotBot, we did indicate that the expression was a person. In our preliminary test queries, we tried personal names in HotBot with and without this specification, and it performed significantly worse without the personal name specification than with it.

[6] Located at http://www.ics.uci.edu/pub/websoft/libwww-perl/ and maintained by Roy Fielding at the University of California at Irvine.

evaluation of a query. For some pages, the evaluation was reassessed when a version of the page could be viewed in which graphics were present.

For each query, after the hundred retrieved pages were inspected in this way, the evaluations were then mapped back to the results pages generated by the search services. The results pages were evaluated to detect duplicates and mirror sites. For sites that gave a 603 (server not responding) or a Forbidden error, those links were checked several more times over a several week period.

It happened that on three occasions the blinding method failed to work correctly, and the results of one of the search engines had to be retrieved later, up to a week later. The results were only used in this case if, when we searched the query a second time in the same search service, all of the links that were inactive in the earlier results were still present in the first 20 hits of the later results. We then evaluated those pages knowing the service they came from. We were able to use the pages on all three occasions.

In both the method the queries were chosen and the method the resulting pages were evaluated, we attempted to prevent our own natural biases from effecting the study. The blinding process for page evaluation only worked for the initial inspection of the page, because later checking and updating was done with an awareness of the source of the page. Nevertheless, that initial blinding was important to establishing relevancy precedents within a process that by its nature is very subjective and open to subtle bias.

Evaluation Criteria

For a study of relevance to a user, the measures of relevancy must be defined. In this study, general criteria for categories of abstract, topical relevance have been devised. Results are then categorized. Once the data is available by category, a formula is used to characterize the basic measure—first 20 precision—using those categories to create a single metric for each query for each service.

Relevancy Categories

In order to describe how the concept of relevance is used in this study, we will summarize Mizzaro's (1997) framework for relevance. Relevance relates two entities: One being some form of information resource, the other being some form of the need for the information. Information resources include the document itself, a surrogate for the document such as its abstract, and the information obtained from the document by the user. Information needs include the abstract problem the user is trying to solve, the information need in the user's mind that the user believes will address the problem, the user's verbal request describing the information need, which we will call the *query*, and finally the search expression based on the verbal request that is keyed into the information retrieval system.

The relevance relationship can be broken into three components—topic, task and context. The topic component relates the resource to the subject area of the need. The task component relates the resource to what the user wants to do with the information. Context relates to everything else: What the user already knows, what reading level the resource is at, how much time and money the resource will cost, etc.

For this study, the information resource is the HTML document itself. The need is studied at the level of the query or the search expression. The primary component of the relationship is the topic, along with anticipated possible tasks. The context is the universe of materials that are published on free Web servers in HTML and that can be located using an Internet search service.

The original user who posed the information request did not participate in judging the relevance of the results from each search service to his or her actual information need.[7] Instead, the query was recorded, and the researchers then defined criteria for relevance categories based on a speculated range of information needs that would be represented by that request on that topic. While having the evaluator be the person with the actual information need is desirable (to allow for a richer concept of relevance), it is not feasible to have a user wait for the query to be prepared, evaluate all 100 Web pages for relevance, and then wait again for the results to be remapped to services to resolve any final discrepancies. None of the other scholarly studies reviewed above utilized evaluators with actual information needs for all of the queries evaluated, and most had no actual users evaluate results at all.

The categories are named as follows: Duplicate links, inactive links, category zero, category one, category two and category three.

- Duplicate links: The same basic URL as a link earlier in the return list, regardless of its other qualities (being inactive, valid, or relevant). Mirror sites are not counted as duplicates.
- Inactive links: File not found (404), forbidden, and server is not responding (603) errors. For forbidden and 603 errors, the links are checked again several times over a several week period.
- Category zero: The page is irrelevant because it does not satisfy an important aspect of the search expression.
- Category one: The page technically satisfies the search expression (structured) or contains all of the search terms or synonyms of them (unstructured), but it is not relevant to the user's query, either because is not related to the topic indicated or because it was too brief to be useful.
- Category two: Relevant to the request and relevant to at least some narrow range of information needs described by the request. These pages are at least potentially relevant to some users. Also pages that have links to category three pages.
- Category three: Relevant to a wide range of possible

[7] In the actual library context, the librarian constructed queries for one or two search services, and then let the patron explore the results on his or her own.

information needs described by the request, such as a clearinghouse of links, or a particularly thorough treatment of the subject.

All pages retrieved by the search services are placed into one of these categories. Within these categories, the relevance judgment is binary—a document is either in the category or it is not in the category. The categories are not a linear scale of relevance. Other studies have used a scale of relevance. Chu and Rosenthal (1996) and Clarke and Willett (1997) used a three value scale, while Ding and Marchionini (1996) used a five value scale modeled after the categories in Leighton (1995).

The Basic Measurement

Once relevancy judgments have been performed, it remains to be established what measure shall be used to compare the performance of the search services. Our model is based on our perceptions about undergraduate information seeking behavior; we did not try to correlate our model with the observed behavior of users. This study did not evaluate search services for their limiting features, screen layout or their presentation of results, which are also important for user satisfaction. The results of the above evaluation process are measured for their ability to put relevant pages within the first 20 links returned by a query, what we call first 20 precision. We chose to study precision because we believe that in the undergraduate context precision is usually more important to the user than recall: Searches tend to be more exploratory than comprehensive.[8] Five different experiments are conducted for first 20 precision, to show how the search services perform using various different versions of what counts as relevant.

Recall can be measured directly or can be approximated. True recall, the ratio of the total number of relevant elements in the space to the total of relevant results returned by the search, cannot be calculated for most searches in Web space, because the total number of links returned by the search services is too great. Lawrence and Giles (1998) conducted a large study for recall and coverage, limiting themselves to searches that retrieved only a modest total number of hits. Recall can be approximated by the pooled method pioneered by the TREC experiments (Harman 1995, 1996). Clarke and Willett (1997) used the TREC method to measure recall for Internet search services. We chose not to attempt to measure recall, whether approximate or complete.

[8] This belief appears contrary to the findings of Su, where she stated, "[P]recision, one of the most important traditional measures of effectiveness, is not significantly correlated with user's judgment of success [in online searching]" (1994, 207). In her study, most users were Ph.D. students with a great deal of subject expertise. Naturally, that demographic group *is* more concerned with recall than precision. The users who provided information requests for this study are from an entirely different demographic group, one less concerned with exhaustive research.

True precision, the ratio of relevant elements returned to the total number of elements returned, is also too arduous to calculate for most searches, again because it would mean examining all of the links returned by a service, which may number in the thousands or millions. The *first X precision* is designed to reflect the quality of service that the typical user would experience: How good is the relevancy within the first few pages of results? A persistent user might well examine the first 50, or 100, returned links, but in practice: "Only about 7% of users really go beyond the first three pages of results," according to Gary Culliss, founder of Direct Hit (Sullivan 1998c). In calculating the precision, we have added a weighing factor, to increase value for ranking effectiveness.

The formula for calculating the metric desired qualities.

The qualities that we decided to measure were first X precision, ranking effectiveness, lack of redundancy and the active status of the links retrieved. We combined precision and ranking effectiveness into one metric. For some of our experiments, we penalized truly duplicate links, for others we did not. We always penalized links returned that were inactive. Spool, Scanlon, Schroeder, Snyder, and DeAngelo (1997) studied actual user behavior in Web sites, and confirmed that users of search engines are discouraged by both poorly ordered results and links with redundant content.

When one attempts to measure both initial precision and ranking effectiveness, one could use two separate metrics, one for precision and one for ranking. As we stated above, all of the other studies have used a straight first X precision, the ratio of relevance scores to the total number of items retrieved within the first X returned records. For ranking, one could use the Coefficient of Ranking Effectiveness (Noreault, Koll, & McGill, 1977), which measures the actual ranking on a linear scale between a best possible ranking and a worst possible ranking.

We chose to create a metric that measures precision with weights for ranking effectiveness. There are several qualities that we made sure the final metric incorporated. First, for each different analysis, a link either meets the criteria under examination, or it does not. We have chosen a binary scale of relevance, because the categories are not defined as intervals on a scale, but are different definitions of relevance. Because the evaluator in our study is not a user with an actual information need, we do not feel it appropriate to use an interval scale for relevance. Second, we want to give more weight to effective ranking of relevant items. Third, the statistic should reflect the fact that if the search service returns fewer hits with the same number of good hits (i.e. if it has better true precision), the relevant links are easier for the user to find. However, this third goal of rewarding true precision should be moderated, so as not to reward a service for being too cautious and returning no hits or only very few.

Finally, one must decide how to treat inactive and duplicate links. In Leighton's 1995 study and in Chu and

Rosenthal's study (1996), duplicates were eliminated from both the numerator and denominator of the precision ratio. That is, a search service was not penalized for not eliminating them. The present study has five tests: In the first three tests, duplicates are only eliminated from the numerator, and services are penalized; in the final two tests, duplicates are not penalized. Services that return inactive links are penalized in all five experiments.

Experiments four and five were designed for any critics who want to see how the services rate when duplicates are not penalized, because duplicates are usually easy enough to detect by the page's description. One should be sensitive to the issue of duplicates because our study did not penalize two other types of pages: Mirror pages—the same page on different servers—and clusters of pages from the same directory or server. Mirror pages and clusters also load the top ranks with somewhat redundant content. Clustering was common for Alta Vista, Excite, and Infoseek, and mirror pages were most common for HotBot. [9]

The actual formula.

In order to analyze and compare the first 20 precision of the five search services, we have devised a formula that gives the performance of the service on a query as a single number between zero and one. The results, up to the first 20, from each service for each query have been categorized already by their status (duplicate or not, active or not) and type of relevance. The categorized results are then used to produce this metric.

The formula for our metric begins by converting the categories into binary values of zero or one. For example, if the test is for pages that minimally satisfied the search expression, then pages in categories one, two and three are assigned a one in the formula, all others are given a zero.

One way to weigh ranking would be to assign each position a value on a 20 position linear scale of value, with first position being the most valuable and the twentieth position being the least valuable. The first ranked item would be assigned a weight of 20, the second item would be assigned a weight of nineteen, and so forth until the last item would be given a weight of one. The same set of relevant items near the top would score higher than it would if it were near the bottom of the 20 item list.

We decided instead to divide the first 20 links returned into three groups: The first three links, the next seven links and the last 10 links. These groups were chosen because the first three usually constitute the user's first screen, the next seven round out the user's first results page and the last 10 make up the second page of results. (The links beyond the twentieth could be said to be in a fourth group—one that is given zero weight.) In each group, the links are assigned an equal weight, the weight that the first item in the group would have had in a 20 item linear scale of value. The first three links are assigned a weight of 20; the next seven are assigned a weight of seventeen; and the last 10 are assigned a weight of 10. These weighted values are then summed to generate the numerator of the metric. For example, if for a given test, a service returned five good links, it would score $(3 \times 20) + (2 \times 17) = 60 + 34 = 94$ if they were the first five links, but only $(5 \times 10) = 50$ if they were all between ranks 11 and 20.

The denominator is calculated by the number, up to 20, of links returned by the search service. If the service returned 20 or more links, then the sum of all weights to 20 is used, $(3 \times 20) + (7 \times 17) + (10 \times 10) = 279$. The denominator is adjusted if there are fewer than 20 links returned, in order to give some benefit to true precision. If the denominator were not adjusted, it would always be 279.

One could calculate the denominator just by summing the weights up to the number of links returned and using that number. The problem with using that number, is that the service could receive high scores by returning only a very few links, and benefit greatly from an overly conservative algorithm. Indeed, if no links are returned, the denominator would be zero, with the metric undefined. Because of this boundary condition, the denominator for this metric is generated by summing all of the weights to 20, 279, and then subtracting 10 for each link less than 20 returned. For instance, if a service returns fifteen links, the denominator is $279 - (5 \times 10) = 229$, but if it returns one link, the denominator is $279 - (19 \times 10) = 89$.

The numerator is divided by the denominator to calculate the final metric. For a search service that returns over 20 links, and the first fifteen of them are good, the metric is 229/279. For the search service that returns only 15 links, and all of them are good, the metric is 229/229, or a perfect 1. For the search service that returns one hit, and it is good, it scores 20/89. A search engine that returns no hits gets 0/79, or zero. This function is a Rube Goldberg machine, but it seems to capture our attempt to credit the search engine for not padding out the rest of the first 20 hits with bad hits, yet without pathological boundary conditions. Below is the complete formula:

$$\frac{(\text{Links } 1\text{–}3 \times 20) + (\text{Links } 4\text{–}10 \times 17) + (\text{Links } 11\text{–}20 \times 10)}{279 - [(20 - \min(\text{No. of links retrieved}, 20)) \times 10]}.$$

One could argue that the adjustable denominator is arbitrary, rewarding services that return between 10 and 20 links, and only gradually penalizing services that return less than 10 links. No benefit is given to a service that returned 500 rather than 500,000 links. We admit that it is arbitrary. The first 20 precision itself as a measurement is also arbitrary, as are the weights given for the ranked groups. They

[9] Since this study was conducted, HotBot, Infoseek, and Lycos have added a feature that allows the user to see pages grouped from the same site. Because of this feature, the main list can be free of clusters, while the patron can choose to access a promising cluster. This feature solves the dilemma described in our study.

are arbitrary, but they are established in order to attempt to reflect the quality of the service that users experience. It is toward that end that this investigation is undertaken at all.

Each experiment, described below, used a slightly different definition for what counts as a good or bad link. For each experiment, the numerator and denominator are calculated for each query in each search service according to the above formula. Those statistics are then analyzed to see how the services compared to each other. See Appendix C of Leighton and Srivastava (1997) for the raw numerators and denominators used to compute the metrics.

Experimental Design

Because of the possible multiple ways of defining what should count as a good or bad link, we have performed several different experiments on our data, to compare the services using a variety of definitions. The first three experiments show how the services rate if they are penalized for both duplicates and inactive links. The last two experiments show how the first two experiments would look if duplicate links were not penalized.

The first three experiments show how the services perform using different thresholds for precision. The first experiment measures a low precision threshold by assigning a 1 to all links in categories one, two, and three. It shows how well each service delivers links that minimally satisfy the search expression. The second experiment measures a moderate precision threshold by assigning a 1 to all links in categories two and three. It shows how well each service delivers links that are potentially relevant to some information need on the topic in question. The third experiment measures a high precision threshold by assigning a 1 to only category-three links, giving all others a 0. It shows how well each service delivers links that have a wide range of possible relevance in the first 20 links returned.

The fourth and fifth experiments start by eliminating duplicates from the results, and treating the remaining links as a list of less than 20 results. In other words, if there were three duplicates in the first 20 links, the duplicates would be removed, and the denominator would be calculated as if there had only been 17 links returned. Experiment four then measures a low precision threshold by giving a 1 to category one, two, and three links, whereas experiment five measures a moderate precision threshold by giving a 1 to category two and three links.

The metrics for each experiment were evaluated to see if the residuals from an ANOVA were distributed normally. Only the residuals for experiment four were normally distributed. See Table 1 for a list of the normality test results. Because the normality assumption required for the ANOVA model was violated, the Friedmann's randomized block design was used, in which the blocks were the queries and the treatment effects were the search services. The Friedmann test estimates population medians rather than means because of the skewness present. In all five experiments, the null hypothesis—that the service's medians could be equal—was rejected. See the bottom of Table 2 for *P*-values for the Friedmann's tests. Because the null hypotheses were rejected, pairwise multiple comparisons could be conducted between individual services.[10]

TABLE 1. Shapiro-Wilk test for Normality of Residuals from an ANOVA model.

Experiment	S-W *P*-value
1	0.7441
2	0.2496
3	0.0003
4	0.9510
5	0.5111

Results and Discussion

The different experiments show how stark the contrast is in scores depending on how one defines relevance. Experiment one called a link good if it at least technically satisfied the search expression. Here, the overall median was a healthy 0.81 with the best service scoring a 0.93. If ones definition of relevant is stricter, dealing with a moderate precision threshold, the overall median drops to 0.39, with the top scorer only making an estimated median of 0.51. If ones criterion is a high precision threshold, the median disappears down to 0.06, with the top scorer only rising to 0.1. (Indeed, if one performs a test for mean and standard deviation on experiment three's data, the standard deviation is higher than the mean.)

Neutralizing duplicates only raised those overall estimated medians slightly, and that rise was almost entirely accounted for in HotBot's improved score. See Table 2 for a detailed breakdown of the Friedmann's estimated medians for each service, for the Friedmann's sum of ranks for each service, for the overall medians, and for the least significant difference in sum of ranks for each experiment.

Within these ranges a definite pattern emerges. Alta Vista, Excite, and Infoseek are always the services with the three highest estimated median scores. Their scores are not always significantly higher than HotBot, but they are always significantly higher than Lycos. Only in experiment four, a version of experiment one where duplicates were neutralized rather than penalized, does HotBot have a significantly higher estimated median than Lycos.

Table 3 shows the rankings of the services for each experiment. If an underline connects two services, their estimated medians are not significantly different from one another. For instance, in experiment one, Excite and Alta Vista have significantly higher estimated medians than Infoseek, HotBot, and Lycos. So in terms of delivering nonduplicate, active links that at least technically satisfy the search expression, they are better. Infoseek is in the middle,

[10] The formula for multiple comparisons was taken from equation 15 on page 300 of Practical Nonparametric Statistics (Conover, 1980).

TABLE 2. The estimated medians and sums of ranks for the experiments.

	Experiment				
Service	1	2	3	4	5
			Medians		
Alta Vista	0.9032	0.4523	0.06022	0.9032	0.4741
Excite	0.9362	0.4717	0.07168	0.9321	0.4789
HotBot	0.7197	0.2925	0.03871	0.8246	0.3514
Infoseek	0.8659	0.5054	0.09892	0.8600	0.5347
Lycos	0.6072	0.2746	0.03154	0.6009	0.3039
Median of all	0.8065	0.3993	0.06022	0.8242	0.4286
			Sum of Ranks		
Alta Vista	59.0	53.0	46.0	56.5	52.5
Excite	66.0	55.0	50.0	63.5	55.0
HotBot	30.0	33.5	35.5	40.5	35.0
Infoseek	46.5	57.0	59.5	43.0	56.0
Lycos	23.5	26.5	34.0	21.5	26.5
LSD	10.91	17.98	9.28	12.24	14.68
Fried. *P*-value	0.000	0.000	0.006	0.000	0.001

Note. The estimated population median and sum of Friedmann's ranks is given for each service. Below the individual medians are the overall medians for the experiment. Below the individual sum of ranks are the sizes of the least significant difference between sums of ranks. If the difference between the sums of ranks for two services is greater than this number, the difference is significant. Finally, the *P*-value (adjusted for ties) is reported for each Friedmann's test.

statistically worse than the top two, but better than HotBot and Lycos. Experiment two is closest in criteria to what we ourselves would define relevance to be. In Table 3, one can see that for experiment two, Infoseek, Excite, and Alta Vista are not significantly different from one another, but are all significantly higher than HotBot and Lycos.

When one examines quantile box charts of the scores, the difference in variability is clear. See Figure 1 for quantile boxes for both experiments one and two. In experiment one, all three top services have low variability, while in experiment two, their variability increases, indicating that the availability of category two and three Web pages was less uniform across queries than the availability of category one pages. Alta Vista's change in variability is more dramatic, indicating that, while it is able to satisfy the search expression consistently, the ranking of higher category pages is uneven.

In experiment three, the score for the pages deemed most likely to be useful, Infoseek is significantly higher than all other services. Alta Vista and Excite land in the middle, being significantly better than HotBot and Lycos, which shared the bottom. It should be noted that on the Web, even the best search services deliver only an estimated median of 10% first 20 precision for a high relevance threshold. The low numbers in this experiment indicate that the free portion of the Web lacks good quality materials on a breadth of subjects. (The low numbers may also indicate that our metric formula does not behave well for poor result sets.) Figure 2 shows how, for both experiments two and three, the Friedmann's sum of ranks compares with the least significant difference (LSD) of sum of ranks. Note that the LSD for experiment three is much smaller than the LSD for experiment two.

Experiment four altered experiment one by not penalizing duplicate page references. Here, HotBot succeeded in being significantly higher in estimated median than Lycos and not significantly lower than Infoseek. But in experiment five, neutralized duplicates helped HotBot little. Alta Vista, Excite, and Infoseek again shared the upper cluster, but HotBot was not significantly higher than Lycos.

We have also conducted a correspondence analysis of the queries by the services, using the scores from experiment

TABLE 3. The rankings and significant differences among the services.

	Rankings					
Experiment	Lowest				Highest	
1	Lycos	HotBot		Infoseek	Alta Vista	Excite
2	Lycos	HotBot		Alta Vista	Excite	Infoseek
3	Lycos	HotBot	Alta Vista	Excite		Infoseek
4	Lycos		HotBot	Infoseek	Alta Vista	Excite
5	Lycos	HotBot		Alta Vista	Excite	Infoseek

Note. Each service is ranked by experiment. Underlining indicates no significant difference.

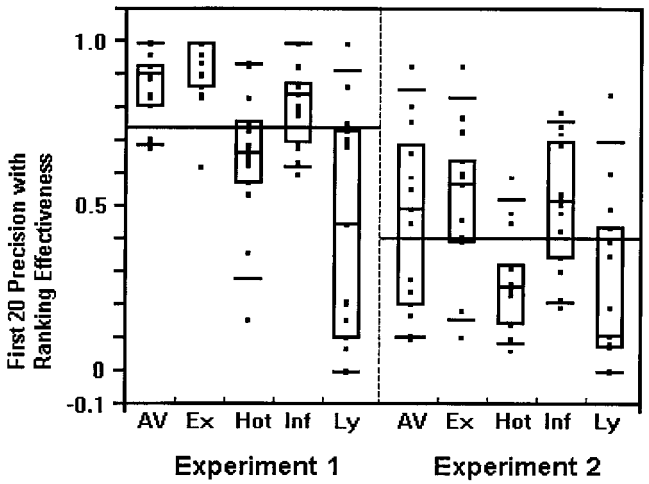

FIG. 1. Quantile box plot of median and range for each service in experiment one (left) and experiment two (right).

two as weights.[11] A correspondence analysis shows how each item (whether service or query) corresponds to the composite score of the whole. To display all of the information in the correspondence relationship, one would have to graph it in higher dimensional space. The graph in Figure 3 is a projection from many dimensions to two, yet it contains 78.7% of the information from the analysis. An item plotted toward the center of the graph has a score typical of the whole: A query near the center scored similarly in all of the services, a service near the center scored typically on most of the queries. An item plotted away from the center differs from the composite, and differs markedly from items plotted away from the center in another direction. So a service plotted close to a cluster of queries performed better on those queries than it did on queries plotted away from the center in another direction.

In our analysis, the services with better experiment two precision are closer to the center. Lycos's performance corresponds most closely with query 8, followed by queries 9 and 5, all of which are shorter and unstructured. HotBot's performance corresponds most closely with queries 12, 6 and 13. Queries 12 and 13 are structured, and 6 is a proper name. Although Lycos and HotBot have comparable scores, Lycos does best with shorter, unstructured queries, while HotBot does better with structured queries. This result is not surprising, because Lycos lacks many operators, while HotBot has many operators. Lycos's continued popularity may be in part attributed to the fact that the vast majority of Internet search expressions are short and unstructured, Alta Vista is also plotted closer to the unstructured queries,

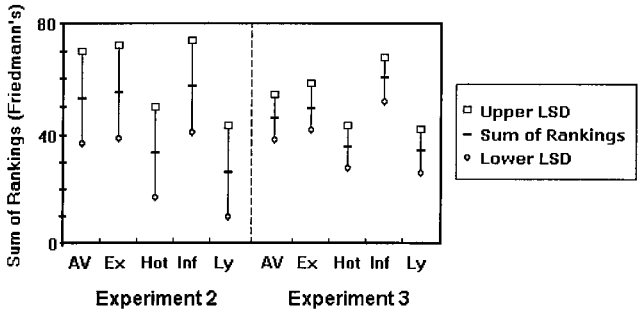

FIG. 2. The sum of ranks and the least significant difference interval for services in experiment two (left) and experiment three (right).

which is a result consistent with the hypothesis that we did not engage Alta Vista's ranking algorithm on the advanced query form.[12]

When one examines the patterns of duplicates, dead links and category zero links (see Appendix B and Appendix C), one sees differences between HotBot and Lycos. HotBot suffers from duplicates. Yet, even when they are neutralized, it is still significantly lower than the top services. Lycos has the most number of inactive links, but it does not have duplicates. Lycos has the most number of zeros, and it is evident to the evaluator that this can be in part attributed to Lycos's inability to require that a specific word be present in the results (the plus sign operator)—a feature that all of the other services had available in some form. In prelimi-

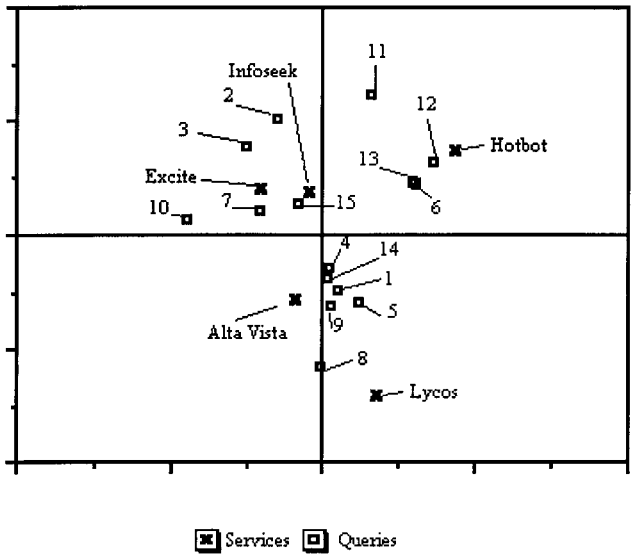

FIG. 3. A correspondence analysis of service by query using the scores in experiment two as weights.

[11] The correspondence analysis is performed in the following manner: The score (numerator over denominator) for each search engine for each query is taken and turned it into an integer between zero and 100. So a score of 0.592 becomes 59. Next, these new scores are then treated as frequencies in a contingency table, the rows being queries, the columns being search engines. We then run a correspondence analysis on that table in the statistical program JMP.

[12] Alta Vista has since then redesigned the advanced form, so that the unknowledgeable user types the search expression into the ranking text box rather than into the Boolean expression text box, solving the problem for the average user.

nary queries used to prepare for this study, we found that when we specified tight constraints in Lycos (requiring all words, for example), we often got no results. Part of Lycos's overall showing may be attributable to its policy of only indexing the 100 weightiest words in a document. It may simply not be finding documents satisfying the search expression because it did not index enough terms. Unlike Infoseek, its fortunes do not change when a higher threshold for relevance is used.

Another pattern noticed in both HotBot and Lycos was a difficulty with locating capitalized proper names (but not personal names). Repeating some of the searches later with and without capitalization, we discovered that both HotBot (unstructured) and Lycos are two-way case insensitive (capitalized match lower case), which may have effected their performance.

Summary and Future Work

It is clear looking at the data that, in general, Alta Vista, Excite, and Infoseek did a superior job delivering quality relevance. Alta Vista did its best in experiment one, then slipped relative to Excite and Infoseek in the higher categories. In retrospect, for experiments two and three, Alta Vista might have done better in the higher categories if we had explicitly added the terms to the relevancy box in advanced searches.

If one looks for why the three top services did better, one could point to common features that allow the user to control the search, such as case sensitivity for capitals and the plus operator for requiring a word. One can point to cleaner databases, with low duplicate and dead link rates. The first 20 hits for the top three contained all of the words from the search expression more often than the first 20 hits for the lower two services. To understand the reasons for these results any better, one would have to know the exact ranking algorithm used by each service, and because they are commercial companies with millions of dollars riding on their relevancy, those precise algorithms are naturally trade secrets.

Some aspects of the techniques used in our study could use more theoretical investigation. One could examine our relevance categories by comparing actual user relevancy judgments to the scores an evaluator assigns by category. One could test our metric formula against other weighing and scoring schemes, such as the traditional Coefficient of Ranking Effectiveness (Noreault, Koll, & McGill, 1977). The low median scores and high variance in experiment three indicate that our weighing formula may have boundary condition problems.

To analyze these services more thoroughly, future research should aim to conduct a study where the test suite is large enough to compare structured search expressions versus unstructured ones. Naturally, over time, the results of any precision study will become dated as the services continue to change, and a study like this will need to be repeated anyway.

Appendix A: The Queries Used in This Study

Query 1: Simple: Individuals with Disabilities Education Act
Source: Obtained locally

Query 2: Structured: Find unemployment rate statistics for Illinois
Source: Obtained locally

Query 3: Structured: The problem of domestic violence among athletes
Source: Obtained locally

Query 4: Simple: Animal husbandry
Source: Tomaiuolo and Packer study

Query 5: Simple: Cable television regulation
Source: Tomaiuolo and Packer study

Query 6: Simple: MTX International Corporation
Source: Obtained locally

Query 7: Simple: Cell cryopreservation
Source: Tomaiuolo and Packer study

Query 8: Simple: Classical architecture
Source: Tomaiuolo and Packer study

Query 9: Simple: Ecotourism
Source: Tomaiuolo and Packer study

Query 10: Simple: In Focus Systems [the company]
Source: Obtained locally

Query 11: Structured: Find information on the National Gallery of Prague
Source: Obtained locally

Query 12: Structured: The effects of caffeine on aerobic exercise
Source: Obtained locally

Query 13: Structured: Find information on the group called "Queer Nation"
Source: Obtained locally

Query 14: Simple: Find information about Prozac, but not the rock group
Source: Obtained locally

Query 15: Personal name: Find information on Maria Luisa Bombal
Source: Obtained locally

Appendix B: First Twenty: Dead Links, Duplicates, Zeros

Query\Serv.	Alta V	Excite	HotBot	Infoseek	Lycos
1 IDEA	1/0/0	0/0/0	1/11/2	2/0/0	2/0/4
2 u. rate	3/0/1	3/0/1	5/2/1	2/0/0	0/0/0 (0)
3 athletes	0/0/2	0/0/0	0/0/2	6/0/1	2/0/17
4 animal h	3/0/0	1/0/0	1/4/2	3/0/0	5/0/0
5 cable t	1/0/1	2/0/0	2/2/1	3/0/0	5/0/2
6 MTXI	0/0/9	0/0/0	1/4/1	2/0/1	6/0/11
7 cell cry	0/0/0	0/0/0	0/1/0	0/0/0	5/0/1
8 classical	0/0/0	0/0/0	2/4/2	3/0/0	0/0/0
9 ecotour	1/0/0	2/0/0	3/0/0	4/0/0	3/0/0
10 IFS	1/0/3	0/0/0	4/1/12	0/0/0	7/0/11
11 Nat Gall	3/0/2	0/1/0	1/5/1	3/3/0	7/0/13

Appendix B (Continued)

Query\Serv.	Alta V	Excite	HotBot	Infoseek	Lycos
12 caffeine (16)	2/0/1	0/0/0	1/7/0	6/2/0	4/0/9
13 QN	6/0/0	6/0/1	1/4/4	5/0/1	7/0/5
14 prozac -r	2/0/0	2/0/1	1/7/0	2/0/1	6/0/0
15 ML Bombal	0/0/1	0/0/0 (18)	4/1/1	2/0/1 (14)	5/0/12

Appendix C: Median, Mean, and Mode for Dead, Duplicate and Zero Links

Query\Serv.	Alta V	Excite	HotBot	Infoseek	Lycos
Dead median	1	0	1	3[2..5]*	4.5[3..6]*
Dead mean	1.53	1.07	1.80	2.87	4.27
Dead mode	0	0	1	2	5
Dup median	0	0	4	0	0
Dup mean	0	0.06	3.53	0.33	0
Dup mode	0	0	4	0	0
Zero median	1	0	1	0	5.5[1..8.5]*
Zero mean	1.33	0.20	1.93	0.33	5.67
Zero mode	0	0	1	0	0

* Used an estimated population median and a confidence interval for the median based on the Wilcoxon test.

Acknowledgments

This research was done for Mr. Leighton's final project for a master's degree in computer science at the University of Minnesota, Twin Cities campus. The authors thank Dr. Brant Deppa of Winona State University for suggesting the Friedmann's formula for comparing multiple comparisons, techniques for correspondence analysis and the other statistical techniques used in this study, Dr. Carol Blumberg of Winona State University for her advice on the design of the project, and Don Byrd of the University of Massachusetts at Amherst for his advice and suggestions on related literature in the field. I would like to thank the reviewers for their advice and suggestions.

References

Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: A comparative study and evaluation methodology. ASIS '96 : Proceedings of the 59th ASIS Annual Meeting (Vol. 33, pp. 127–135). Medford, NJ: Information Today, Inc. Also available at http://www.asis.org/annual-96/ElectronicProceedings/chu.html (accessed January 28, 1997).

Clarke, S.J. & Willett, P. (1997). Estimating the recall performance of Web search engines. ASLIB Proceedings, 49, 184–189.

Conover, W.J. (1980). Practical nonparametric statistics (2nd ed.). New York, NY: John Wiley & Sons.

Ding, W. & Marchionini, G. (1996). A comparative study of web search service performance. ASIS '96 : Proceedings of the 59th ASIS Annual Meeting (vol. 33, pp. 136–142). Medford, NJ: Information Today, Inc.

Gauch, S. & Wang, G. (1996). Information fusion with ProFusion. Webnet 96 Conference [online]. 19 paragraphs. Available at http://www.csbs.utsa.edu:80/info/webnet96/html/155.htm (accessed February 22, 1997).

Harman, D. (1995). Overview of the second text retrieval conference (TREC-2). Information Processing and Management, 31, 271–289.

Harman, D. (1996). The fourth text retrieval conference (TREC-4). NIST Special Publication (pp. 500–236). Gaithersburg, MD: National Institute of Standards and Technology.

Infoseek. (1997). Infoseek: Precision vs. recall [online]. Available at http://www.infoseek.com/doc?pg=prec_rec.html (accessed February 7,1997).

Lawrence, S. & Giles, C.L. (1998). Searching the World Wide Web. Science, 280, 98–100.

Leighton, H.V. (1995). Performance of four World Wide Web (WWW) index services: Infoseek, Lycos, WebCrawler, and WWWWorm [online]. Available at http://www.winona.msus.edu/library/webind.htm (accessed July 1, 1996).

Leighton, H.V. & Srivastava, J. (1997). Precision among World Wide Web search services (search engines): Alta Vista, Excite, HotBot, Infoseek, Lycos [online]. 95 Paragraphs. Available at http://www.winona.msus.edu/library/webind2/webind2.htm (accessed June 22, 1998).

Magellan Internet Guide. (1997). Real-time Magellan searches [online]. Available at http://voyeur.mckinley.com/voyeur.cgi (accessed January 24, 1997).

Maze, S., Moxley, D., & Smith, D.J. (1997). Authoritative guide to Web search engines. New York, NY: Neal–Schuman.

Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society for Information Science, 48, 810–832.

Munro, J. & Lidsky, D. (1996, Dec. 3). Web search sites. PC Magazine, 15, 232.

Noreault, T., Koll, M., & McGill, M.J. (1977). Automatic ranked output from Boolean searches in SIRE. Journal of the American Society for Information Science, 28, 333–339.

Singh, A. & Lidsky, D. (1996). All-out search. PC Magazine, 15, 213.

Scoville, R. (1996). Special report: Find it on the Net! PC World, 14, 125. Available at http://www.pcworld.com/reprints/lycos.htm (accessed February 1, 1997).

Spool, J.M., Scanlon, T., Schroeder, W., Snyder, C., & DeAngelo, T. (1997). Web site usability: A designer's guide. San Francisco, CA: Morgan Kaufmann.

Su, L.T. (1994). The relevance of recall and precision in user evaluation. Journal of the American Society for Information Science, 45, 207–217.

Sullivan, D. (1998a). Directories take center stage. Search engine report [online]. Available at http://searchenginewatch.com/sereport/9805-directory.html (accessed January 12,1999).

Sullivan, D. (1998b). Excite enhances search results. Search engine report [online]. Available: http://searchenginewatch.com/sereport/9806-excite.html [12 January 1999].

Sullivan, D. (1998c). Counting clicks and looking at links. Search engine report [online]. Available at http://searchenginewatch.com/sereport/9808-clicks.html (accessed January 12,1999).

Sullivan, D. (1998d). AltaVista debuts search features. Search engine report [online]. Available at http://searchenginewatch.com/sereport/9811-altavista.html
(accessed January 12,1999).

Sullivan, D. (1998e). Lycos buys Wired, gets facelift. Search engine report [online]. Available at http://searchenginewatch.com/sereport/9811-lycos.html
(accessed January 12,1999).

Tomaiuolo, N.G. & Packer, J.G. (1996). An analysis of Internet search engines: Assessment of over 200 search queries. Computers in Libraries, 16(6), 58–63. The list of queries that they used is in: Quantitative analysis of five WWW "search engines" [online]. Available at http://neal.ctstateu.edu:2001/htdocs/websearch.html (accessed February 1997).

Venditto, G. (1996). Search engine showdown. Internet World, 7, 78–86.