IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4500

The Noise Factor: Irrelevant Search Results on the World Wide Web

Wendy Lucas, Assistant Professor, Department of Computer Information Systems
Bentley College, 175 Forest Street, Waltham, MA 02452-4705, Ph: 781-891-2554, Fax: 781-891-2949, wlucas@bentley.edu

ABSTRACT

Finding information on the World Wide Web is easy; finding relevant information is not. While search engines provide a more directed approach to resource discovery than browsing, the pages they identify as "matching" a query often have little relevance to the information being sought. To examine the relationship between query terms and the pages they match, the queries that were used to locate five Web pages were collected over a five-month period. The relevancy of page content to the query terms was then analyzed.

Page content was judged as being irrelevant to more than one-third of the queries. Search engines disagreed; a test page link appeared within the first one percent of the total number of links retrieved in response to forty percent of those queries. This supported the supposition that conducting searches with popular search engines often results in too many links with little relevance to the query terms. An alternative approach is therefore proposed in which metadata, hyperlinks, and other subject-related HyperText Markup Language (HTML) tags are used to improve the effectiveness of Web queries. By relying on the structural components of HTML documents, it should be possible to conduct intelligent searches that yield more relevant results.

INTRODUCTION

The ability to find information on the World Wide Web was greatly enhanced by the introduction of search engines in the mid-1990s. Anyone using a search engine, however, has submitted a query that returned thousands of documents of questionable relevancy. As the number of Web pages continues to grow, it becomes both more important and more difficult to ensure the relevance of the results returned. In [1], it was estimated that there were at least 275 million distinct, static pages on the Web, and that the size of the Web was growing by about 20 million pages per month.

The major search engines all claim to maintain full-text indices of Web pages, up to some maximum limit on the file size. Maintaining these indices for an ever-expanding Web is a daunting task. At the present time, each engine indexes small, varying portions of the Web [9]. Without a complete index, it is impossible for any one search engine to find all of the relevant pages that exist in response to a query. Given that the indexing of new or modified pages can take several months or more [9, 11], those pages that are found may no longer meet the search criteria if their contents have changed but have not been updated in the index.

When viewing search results, links to matching pages are listed in decreasing order of relevance, with relevancy ratings determined by proprietary algorithms [7]. Those pages whose links appear on the first page of results are the most likely to be viewed [15, 16]. If the rating algorithm used is not reliable, then a link to a truly relevant page may never be viewed because of all the irrelevant links that precede it.

The research presented here begins with the hypothesis that the ranking methods used by the most popular search engines are inadequate for finding relevant pages. Five Web pages were created to test this hypothesis. The relevancy ratings assigned by the search engines to these pages in response to queries entered by their users are compared to judged relevancy ratings. Then, a methodology that relies on the structure of HTML documents for improving the relevance of search results is developed and tested. Its premise is that the content of META and other HTML tags plus the network of hyperlinks joining Web pages can be used to eliminate many of the irrelevant links that appear in response to queries.

RELATED WORK

There are many studies that analyze the precision of search engine results, where precision is defined as the ratio of the number of relevant documents retrieved to the total number of documents retrieved. To make these studies feasible, analysis is limited to the top ten or twenty results returned in response to a query. Two examples of such studies are [10] and [5]. The former found AltaVista, Excite, and Infoseek to be the top three performers, while the latter found that AltaVista outperformed Excite and Lycos. The focus of the research presented here is on the queries used to find Web pages, regardless of where links to those pages fell in the search results. The total number of links returned is also considered when judging the accuracy of the search engine-assigned ratings.

A variety of methodologies for improving query results are currently under development. LASER [2] uses HTML formatting and tags plus hypertext links to improve search engine performance. Google [4] takes advantage of link structure and link text to increase the relevancy of search results. Other features include its use of term proximity and the font size of terms in computing rank. In the methodology presented later in this paper, the content of HTML tags and link structure are the sole criteria for determining page relevance.

RESEARCH METHODOLOGY

In order to evaluate the relevance of Web page content to the queries that led searchers to them, five Web pages were registered with the five most-visited search engines [13]: AltaVista,

This paper appears in *Challenges of Information Technology Management in the 21st Century*, the proceedings of the Information Resources Management Association International Conference.

Excite, HotBot, Infoseek (now part of the GO Network), and Lycos. Estonian recipes were chosen as the topic for these pages because of the many unique keywords that could be used to search for them. At the same time, the topic is not so popular that thousands of Web pages would include those keywords, as that would decrease the possibility of anyone finding and clicking on a link to one of the test pages.

The queries submitted to search engines that led visitors to these pages were collected for analysis. Unfortunately, the five search engines to which the page registration forms were directly submitted were often unable or unwilling to correctly index the five Web pages [11]. Having only some of the pages indexed by some of the leading search engines at any point in time limited the number of queries available for analysis.

ANALYSIS OF RESULTS

The analysis presented here examines the relevance of page content to query terms. First, the relevancy of the test pages' contents to the query terms that resulted in their retrieval is judged. Then, the accuracy of the relevancy rankings assigned by the search engines to the test pages in response to those queries is compared to the judged ratings.

The Queries

A total of one hundred and eighteen queries submitted to fifteen search engines and metasearchers, which submit queries to multiple search engines, led people to the test pages, as shown in Figure 1. Of those queries, 38% were unique (data available upon request).

Search engines support a variety of query styles. The majority (77%) of the unique queries used to locate the test pages were term queries, in which a disjunctive comparison is performed to determine if at least one of the terms matches the page's content. The remaining queries were phrase, Boolean, and Boolean-like (i.e., used the + and - symbols).

Figure 1: Summary of Queries by Search Engine

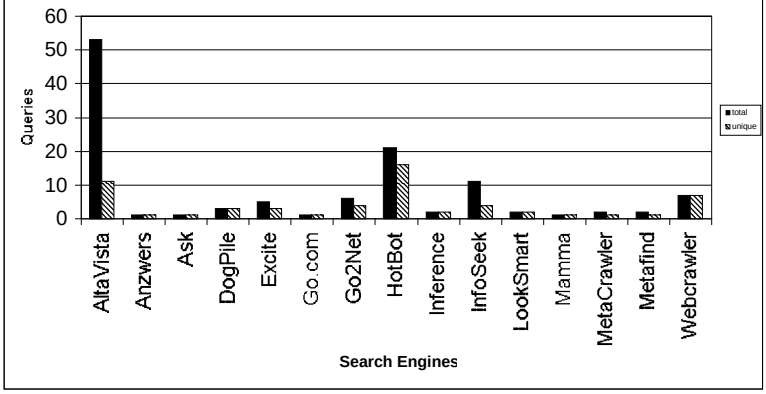
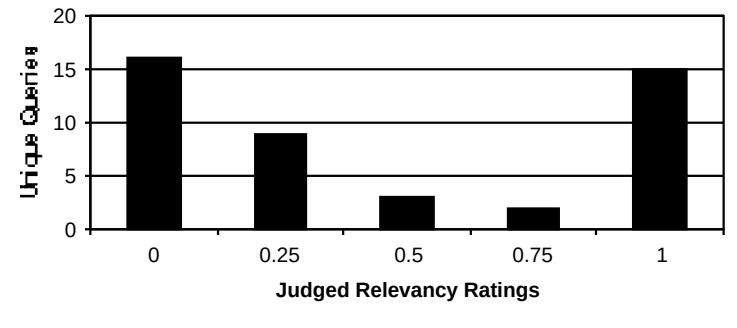


Figure 2: Summary of Judged Relevancies



Judged Relevancy Ratings

A three-level scoring method was used for judging the relevancy of the test pages' contents to each unique query. The levels assigned were 1.0 for *relevant*, 0.5 for *somewhat relevant*, and 0 for *irrelevant*. As in [5], two independent evaluators made these judgments in order to reduce bias. Their level of agreement was 88%. Averages of the two independently arrived-at scores were then calculated for each query. Figure 2 shows the distribution of these scores.

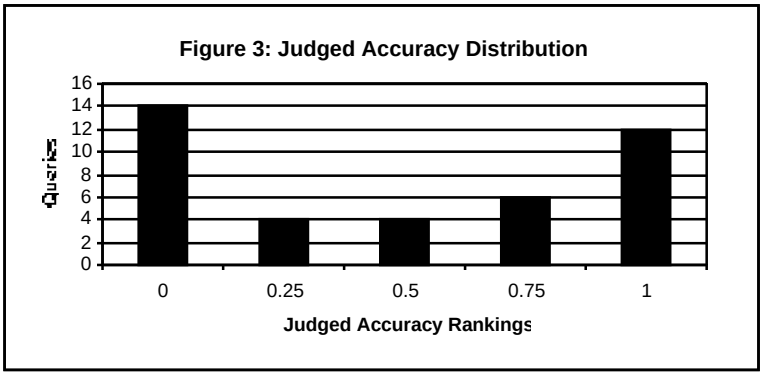
The average judged relevancy rating for how well the test pages' contents match the query terms is 0.46 ± 0.054 (using a 95% confidence interval). On average, the test pages are therefore somewhat relevant to the queries that retrieved them. Test page content was judged as being irrelevant to 35% of the unique queries. Links to those pages should therefore not have been listed in response to those queries.

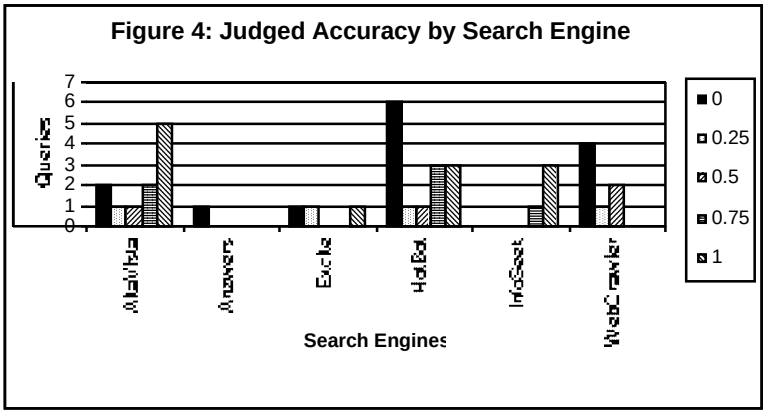
Test Page Relevancy Rankings

The next step is to examine the relevancy rankings assigned by the search engines to the test pages. The "judged accuracy" of these rankings refers to the level of agreement with them. Eighty-nine percent of the unique queries had ranking data available from one or more search engines. This data was either present in the information stored with each submitted query, or was obtained by resubmitting the original query to the search engine. In the latter case, the position of a test page link in response to a query could have changed from when that query was originally submitted. The range into which a test link fell, rather than an absolute value, was therefore used in judging the accuracy of a relevancy ranking. Percentage rating information, available from one or more search engines for 44% of the unique queries, was also considered in judging accuracy.

A three-level scoring method was used once again in making these judgments. The levels assigned were 1.0 for *agree*, 0.5 for *somewhat agree*, and 0 for *disagree* with the rankings assigned by the search engines. In each case, scores were arrived at independently and then averaged. The level of agreement on the independent scores was 80%.

Figure 3 shows the distribution of the judged accuracy rankings by scoring level, and Figure 4 shows the distribution by search engine. The overall average for agreement with the search engine-assigned relevancy rankings is 0.49 ± 0.056 (using a 95% confidence interval), which corresponds to "somewhat agreeing" with the rankings provided by the search engines in response to the queries. Agreement is highest with the relevancy rankings of AltaVista and Infoseek, but the sample sizes are too small to form any significant conclusions. Most of the disagreement with assigned rankings comes from the fact that test page links appeared in the first 10% of the total number of links returned in response to all queries whose content was judged as being irrelevant to the page content. Forty percent of those links were within the top 1% of the total number of links retrieved.





Inflated ratings could be somewhat reduced by better query formation on the part of submitters. For example, a test page was listed in response to a query for Bath Oil Making. Both "oil" and "making" appear in a pancake recipe on that page. "Bath" never appears at all. If the terms "Bath" and "Oil" had been entered as "Bath Oil", then the test page would probably have appeared later in the listing than within the range of 26-50 out of 126,788 matches. Unfortunately, it still would have appeared because of the match on the remaining query term.

Entering "Bath Oil Making" would have excluded the test page, but would also have excluded documents on "making bath oil". The best way to exclude irrelevant pages would be to enter the query as +"Bath Oil" +Making. This requires some degree of sophistication on the part of the query submitter, plus an accurate and up-to-date index in order for the results to be correct.

The relevancy of a test page in response to a query was judged as being underrated by a search engine in only one case. This was for a query on estonian food submitted to AltaVista. The test page did not appear until the fifth page of matches, and was far more relevant than many of the links that appeared before it, such as the one leading to: Estonian choir finds success in a post-Soviet world.

RETRIEVAL HEURISTICS

Even if queries to search engines were optimally formed and correctly executed, it is likely that too many irrelevant pages would still be listed. Indexing all the terms in a Web page and retrieving any page that contains at least one search term has a positive effect on recall, which is defined as the ratio of the number of relevant documents retrieved to the total number of relevant documents in the collection. Precision, however, suffers.

To maximize precision, terms with higher correspondence to page content must be identified and used as the basis for determining how well a query matches a document. While this method can have a negative impact on recall, it will result in the listing of fewer irrelevant links. This tradeoff of recall in favor of precision becomes increasingly attractive as the Web continues to grow. The goal of the methodology described in this section is therefore to improve the relevancy of search engine results by listing fewer links of questionable value.

Evaluation Metric

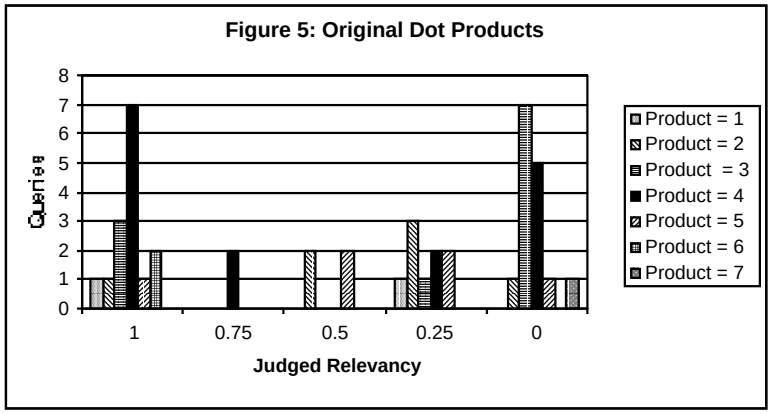
An evaluation metric is needed to analyze how changes to indexing and ranking practices affect the number of matches found and their relevancy rankings. A simple method for determining if a page and a query match is to represent each by a vector of terms. If a term is present in the page or query, then its vector value is 1. If it is not present, its value is 0. A dot product of the query vector and each page vector is then performed, and the results are used to rank the pages [8]. Weighted matches can also be computed by increasing the weights of certain terms found in the document on the basis of frequency, position, or some other criteria.

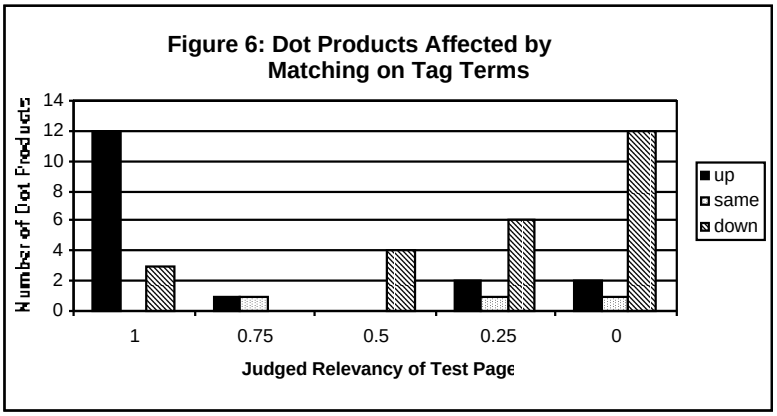
In composing vectors for the test pages, each occurrence of a query term in the page was assigned a value of 1, while each phrase was assigned a weighted value of 2. If a term or phrase occurred more than once in the document, its value was doubled. Search engines will often exclude Web pages that repeat the same term many times from their index. Needless repetition is viewed as "spamming", or an attempt to manipulate a page's rank. Having a term or phrase appear two or more times may be beneficial, but beyond some limit it has a negative effect. Given the short length of the test pages, a limit of two occurrences was chosen.

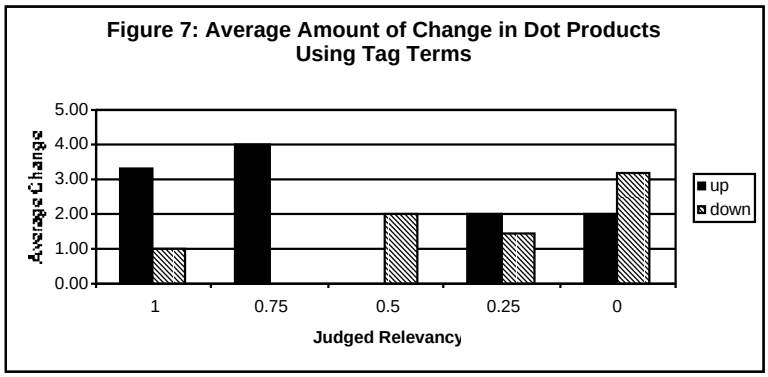
Terms that were identified as "stop words" were excluded from document and query vectors. These are terms that search engines will not "stop for". They are either not included in indices or are indexed but ignored during searches because they are too common and have little subject-related significance to add meaning to the search.

Figure 5 shows the dot products calculated for the forty-five unique queries and the test pages they retrieved. For Boolean queries containing an AND condition, the dot product was 0 if both query terms did not appear in the test page. For an OR condition, if both query terms appeared in the page, the higher score of the two was used. These dot products provide a baseline for comparing the effects of the methodologies described next.

HTML Tags







HTML is a universal language that enables the publishing and retrieval of documents on the Web. Members of the World Wide Web Consortium recommend its specification. The document type definition (DTD) declares the element types that are may appear as tags embedded within HTML documents [14]. Some of these tags contain information directly related to the subject matter of a Web page. The methodology for page ranking described here relies on the contents of those tags as the sole determinants of page rank.

The most popular search engines already give increased relevancy weightings to terms found in title tags. Headings may also convey information about page content. The DTD specifies six levels of headings, ranging from most to least important. Metadata is another source of information about a page's content. It is specified in HTML by the META tag.

To see the effects of using only terms found in HTML tags, rather than all terms from the page, on the ranking process, a new dot product calculation was performed. The weighting of terms found in title and META tags was increased to 2, while phrases were increased to 4. For heading tags, term weightings were increased to 1.5, while phrase weightings were increased to 3. Once again, up to two occurrences per query term or phrase were counted. Figure 6 shows how the dot products change as a result of these calculations.

The dot products that are most affected are those for which the test pages were either judged to be the most relevant or the least relevant in response to the queries. The match between query content and page content typically rose for the former and fell for the later. The desired effect of improving the rank of the test page when it is relevant to a query and decreasing its rank when it is irrelevant is therefore achieved.

Figure 7 shows the average amount of change to the dot products by relevancy rating. Overall, the average change was -0.122 ± 0.841 (using a 95% confidence interval) when only terms from HTML tags were used in the matching process.

Table 1 shows the fourteen queries (as entered by search engine users) to which test page content is no longer relevant when the dot product is calculated using terms from the title, heading, and META tags. Page content has judged relevancy ratings of 0 for nine of these queries, and from 0.25 to 0.50 for the remaining five. For 56% of the cases where page content is irrelevant to the query and for 42% of the cases where it is slightly to somewhat relevant, the relevancy rating falls to 0.

Hyperlinks

The hypertext structure of the Web provides valuable information for indexing and ranking pages. Graphically, the Web can be represented as nodes joined together by directed links. Anchor tags within HTML documents define the links between Web pages and include the address of the referenced page as well as a label describing it. The ancestors of a page, its descendents, and the contents of its anchor tags all provide information about a page's subject matter.

Many leading search engines, including Excite, HotBot, Infoseek, and Lycos, include link popularity as a determinant of relevance to a query. Higher ratings are given to pages with several links leading to them, particularly if those links originate at popular pages themselves. While a valuable source of information, link popularity will favor more established pages over newer, less established ones.

The links leading from a page should also be incorporated into the ranking algorithm, as pages that serve as portals to other relevant pages provide a valuable service. The addition of an internal link from one of the test pages to another Web page about Estonian recipes, for example, would increase the usefulness of that test page for any searcher looking for information on that topic. In [4], it is demonstrated that the use of link structure and link text in indexing pages improves the quality of search results.

Table 1: Queries with Dot Products of 0

Query	Judged Relevancy
cleargelatin	0.50
sauerkraut+cooking	0.50
"pork"	0.25
potatoes	0.25
sauerkraut	0.25
Bath Oil Making	0.00
hard boil eggs	0.00
making butter	0.00
salt pickles	0.00
crock pot	0.00
neck bones	0.00
reefer pot marijuana chronic not trucks and not houseware and not pain and not pottery and not sex	0.00
ruff salad	0.00
small intestines	0.00

Results and Implications

Using only the text from a document's HTML tags in determining relevance to a query can improve the precision of the matching process by decreasing the number of irrelevant links. The match between query terms and page content, as calculated by the dot product, increased for the majority of cases where the page is relevant to the query, and decreased for the majority of cases where it is not.

Objections to the use of META tags in indexing documents include the additional burden they place on authors of Web pages, the fact that they are not compulsory, and the need for better classifications of Web objects [12]. Many of these objections can be overcome by providing a metadata form to be completed when submitting a page to a search engine or resubmitting an updated one. Some search engines, such as Excite and WebCrawler, already ask for keywords during the submission process.

Standards for metadata classification are also needed and are currently being developed. The W3C Resource Description Language (RDF) is working on a common framework for metadata [3]. One of its intended uses is to improve search engine capabilities by providing machine-understandable information about Web resources. Many other metadata classification schemes, such as the Dublin Core metadata element set [6], are also being developed for this purpose.

CONCLUSIONS AND FUTURE WORK

Maintaining full-text indices is becoming less and less viable as the Web continues to grow. The research presented here supports the hypotheses that current ranking methods used by today's most popular search engines are inadequate for finding relevant results. An alternative approach that relies on indexing only the content of subject-related HTML tags was explored. Matching query terms to tag content was shown to reduce the number of links to pages with little or no relevancy. This finding supports the results sited in [2], in which indexing the structural components of HTML documents was shown to improve the performance of the ranking function.

Directions for future work include an analysis of how the query formation process affects the relevancy of search results, and how that process can be enhanced. Alternative presentations of search results in order to increase their usefulness will also be studied.

ACKNOWLEDGEMENTS

My appreciation goes to Madis Rehepapp, who authored and maintained the five test pages, collected the query data, and assisted in its analysis. I am also grateful for the many helpful discussions, comments and suggestions from Jay Cooprider and William Schiano.

REFERENCES

[1] K. Bharat and A. Broder. Measuring the Web, June 1999; http://www.research.digital.com/-SRC/whatsnew/sem.html.

[2] J. Boyan, D. Freitag, and T. Joachims. A Machine Learning Architecture for Optimizing Web Search Engines. In *AAAI Workshop on Internet-Based Information Systems*, 1996.

[3] D. Brickley and R.V. Guha, Eds. W3C Resource Description Framework (RDF) Schema Specification, Proposed Recommendation 03 March 1999; http://www.w3.org/-TR/PR-rdf-schema/

[4] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *7th International World Wide Web Conference*, 1998.

[5] H. Chu and M. Rosenthal. Search engines for the World Wide Web: A comparative study and evaluation methodology. In *ASIS 1996 Annual Conference Proceedings*, pp. 127-135, October 19-24, 1996.

[6] Dublin Core Metadata Initiative. The Dublin Core: A Simple Content Description Model for Electronic Resources, May 1999; http://purl.-oclc.org/dc/index.htm

[7] V. N. Gudivada, V. V. Raghavan, W. I. Grosky, and R. Kasanagottu. Information Retrieval on the World Wide Web. In *IEEE Internet Computing*, pp. 58-58, September-October 1997.

[8] D. Harman. Ranking Algorithms. In W. B. Frakes and R. Baeza-Yates, Eds., *Information Retrieval: Data Structures and Algorithms*, pp. 363-392. Prentice Hall, Inc., 1992.

[9] S. Lawrence and C. L. Giles. Searching the World Wide Web. In *Science Magazine*, Volume 280, Number 5360, April 3, 1998.

[10] H. V. Leighton, H. V. and J. Srivastava. Precision among World Wide Web Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos, June 1997; http://www.winona.msus.-edu/library/webind2/webind2.htm

[11] W. T. Lucas. (Mis)Management of Information on the World Wide Web. In *Proceedings of the ISAS/SCI'99*, Vol. 4, pp. 511-518, July 1999.

[12] M. Marchiori. The Limits of Web Metadata, and Beyond. In *7th International World Wide Web Conference*, 1998.

[13] Media Metrix. Sept. 1999; http://www.mediametrix.com/PressRoom/Press_Releases/-07_20_99.html

[14] D. Raggett, A. Le Hors, I. Jacobs, Eds. HTML 4.0 Specification, W3C Recommendation, revised on 24-Apr-1998; http://www.w3.org/-TR/REC-html40

[15] C. Silverstein, M. Henzinger, J. Marais, and M. Moricz. Analysis of a Very Large AltaVista Query Log. Technical Report 1998-014, Compaq Systems Research Center, Palo Alto, California, 1998.

[16] A. Spink, C. Chang, and A. Goz. Users' Interactions with the Excite Web Search Engine: A Query Reformulation and Relevance Feedback Analysis. In *Proceedings of WebNet 99 Conference* (to appear).