Evaluation of web-based search engines using user-effort measures

Tang, Muh-Chyun; Sun, Ying

2003

Tang, M. & Sun, Y. (2003). Evaluation of web-based search engines using user-effort measures. Library and Information Science Research E-Journal, 13(2), 1-9. https://dx.doi.org/10.32655/LIBRES.2003.2.1

https://hdl.handle.net/10356/152490

https://doi.org/10.32655/LIBRES.2003.2.1

Libres volume 13, issue 2 (2003) https://doi.org/10.32655/LIBRES.2003.2.1

Evaluation of Web-Based Search Engines Using User-Effort Measures

Muh-Chyun Tang and Ying Sun
4 Huntington St.
School of Information, Communication and Library Studies
Rutgers University,
New Brunswick, NJ 08901,
U.S.A.
muhchyun@scils.rutgers.edu Ysun@scils.rutgers.edu

Abstract

This paper presents a study of the applicability of three user-effort-sensitive evaluation measures —“first 20 full precision,” “search length,” and “rank correlation”—on four Web-based search engines (Google, AltaVista, Excite and Metacrawler). The authors argue that these measures are better alternatives than precision and recall in Web search situations because of their emphasis on the quality of ranking. Eight sets of search topics were collected from four Ph.D. students in four different disciplines (biochemistry, industrial engineering, economics, and urban planning). Each participant was asked to provide two topics along with the corresponding query terms. Their relevance and credibility judgment of the Web pages were then used to compare the performance of the search engines using these three measures. The results show consistency among these three ranking evaluation measures, more so between “first 20 full precision” and search length than between rank correlation and the other two measures. Possible reasons for rank correlation’s disagreement with the other two measures are discussed. Possible future research to improve these measures is also addressed.

Introduction

The explosive growth of information on the World Wide Web poses a challenge to traditional information retrieval (IR) research. Other than the sheer amount of information, some structural factors make searching for relevant and quality information on the Web a formidable task. The freewheeling nature of publishing on the Web is a blessing for the flow of ideas, but it has also complicated the process of retrieving relevant information. In contrast to traditional IR, there are no consistent indexing and classification principles for organizing materials on the Web. Nor are there any filtering practices at hand to ensure the quality and credibility of the documents. Furthermore, certain features of Web search situations also distinguish the Web from the traditional IR setting. It has been shown that ordinary Web searchers tend to give little input (Jasen et al. 1998) and are very sensitive to the time and effort put into the search (Silverstein et al. 1998). The issues of credibility and user efforts peculiar to the Web search environment are not addressed properly by traditional precision and recall measures. Several measures that focus on user efforts have been proposed, yet there has been little investigation of their validity.

Literature Review

Several studies have explored the applicability of traditional IR evaluation criteria, i.e., precision and recall, on search engine performance (Chu and Rosenthal 1996, Leighton and Srivastava 1997, Clarke and Willett 1997, Wishard 1998). Chu and Rosenthal studied the precision of ten queries on three search engines (AltaVista, Excite, and Lycos). Instead of a binary measure of relevance (relevant/non-relevant), they adopted a three-point scale to distinguish among relevant, partially relevant, and non-relevant documents. Clarke and Willett also used a three-point scale in assigning relevance scores, with a slight modification: pages that were considered irrelevant in themselves but that led to relevant pages were judged partially relevant. Clarke and Willett provided by far the most feasible method for measuring recall on the Web. Previously, in the absence of a predefined set of relevant documents, it had been very difficult to assess recall on the Web. Clarke and Willett constructed a relative recall measure by using the merged outputs of all three search engines tested as the pool of relevant documents.

Some characteristics of Web searching, however, require performance criteria other than the precision and recall measures developed in traditional IR. The enormous amount of information and the wide variety of sources on the Web seem to make quality of ranking a much more important dimension in assessing search engine performance since users in general spend less time and effort to sort through the retrieved pages. This is supported by studies of users' searching behaviors on the Web. Silverstein et al. found that about 85 percent of users look only at the first screen with results (Silverstein et al. 1998). Su, Chen, and Dong called for a more user-centered evaluation framework in Web searching environments (Su et al. 1998). They applied five criteria —relevance, efficiency, utility, user satisfaction, and connectivity—to evaluate the performance of four search engines (AltaVista, Infoseek, Lycos, and Open Text). Furthermore, instead of submitting simple text queries as in most search engine evaluations, they used real user search strategies and judgment in the searching and evaluating process.

In contrast to traditional IR searchers, the majority of Web users are laypersons who are more sensitive to time and effort spent on finding information. The ability to optimize search order thus becomes an even more salient dimension of search engine performance. The notion of Expected Search Length (ESL), first proposed by Cooper (1968) some 30 years ago, seems to be an ideal notion to test how well a search engine is able to deliver the most relevant documents at the top of retrieved sets (Agata et al. 1997; Su, Chen and Dong 1998; Oppenheim et al. 2000, Chignell et al. 1999). According to Cooper, the primary function of a retrieval system is to save users as much labor as possible in the search for relevant documents by perusing and discarding irrelevant ones. We tested several measures that emphasize user efforts: first 20 “full” precision, search length, and rank correlation with the view of investigating their applicability and validity. We were interested in seeing how these three measures correlated with one another. The consistency or the lack of it among these measures would be an indicator of each measure’s validity in reflecting a user’s effort and, therefore, search engines’ performance.

Research Procedures

Eight sets of topics were collected from four Ph.D. students in four different disciplines (biochemistry, industrial engineering, economics and urban planning). Each participant provided two topics along with the query terms s/he considered suitable for submission to a search engine.

Table 1 shows one of the query topics used in the study.

Subject Domain: Economics Statement 1: Find information about non-parametric estimation of factor analysis in term structure models and a test for arbitrage. Query: functional data analysis

Table 1. Topic (See Appendix 1 for entire collection of statements and queries)

The search engines selected for comparison were Google, AltaVista, Excite, and Metacrawler. The four search engines were selected mainly due to their popularity. We were also interested in seeing whether certain ranking techniques would lead to better results using our measures. Google was selected because of its growing popularity and its incorporation of citing behaviors on the Web. The only meta-search engine, Metacrawler, was selected for its claim to provide a ranking algorithm of the composite results. Its power-search feature allows the user to select up to 11 major search engines and has the option to sort results by relevance, by source search engine, or by originating site. During our run on Metacrawler, the other three search engines tested were selected as the input sources for the power-search feature, and the results were sorted according to relevance. It should be noted that the main purpose of this study is to test the applicability of user-effort measure rather than to compare search engine performance. An experiment was conducted to compare the performance of four search engines (AltaVista, Excite, Google, and Metacrawler) in academic contexts. Since different search engines have different capabilities, to ensure comparability, we decided to adopt a minimalist approach, using only simple, unstructured query terms in lower-case characters. No phrases or Boolean symbols were included. We believe that this also reflect the majority of Web search engine user's search behaviors.

The authors collected queries from the real users and submitted them on their behalf to the search engines within a two-hour time frame on December 5, 2000. All searches were conducted on computers with the same properties and located in the same LAN in order to avoid the effects of differences in computer performance and network speed. We decided to collect only the top 20 links among the thousands retrieved in light of previous studies showing that 80 percent of users view only the first two pages of results (Jansen et al. 1998). Several representation issues also need to be addressed before returned links can be present to the subjects. It was recognized that varying representation used by the search engines might factor in the subjects' judgment. Furthermore, the order in which returned links were presented might also influence subjects' judgment since they might develop different relevance criteria during the course of examining the Web pages. To avoid sequence and user preference bias, the returned hits from all four search engines for each query were mixed together and stripped of all graphic cues. We then presented the URLs in Microsoft Word files that allowed the subject to examine the real page by clicking on its URL.

Participants were then asked to judge each Web page according to its relevance and credibility on a five-point scale using "0" to indicate non-relevance or a lack of credibility and "4" to indicate high relevance or high credibility. Relevance was defined as a result that provided information that is considered useful by the participant for his or her question. The subjects were also told to judge a source's credibility by its authorship, source of its content, disclosure, and currency.

Participants were asked to mark but not to judge duplicate links (those with the same URLs). We could in this way avoid assigning different scores to the same page. Duplicate links were examined later by the authors. When duplicate links were retrieved by the same search engine, the second document was treated as non-relevant; those from different systems' result sets were assigned the same values for relevance and credibility.

Broken link ratio was an indication of how frequently and thoroughly the engine checked the links in its database for currency. In the analysis of relevance and credibility, broken links were treated as non-relevant documents with zero relevance and credibility.

Evaluation Measures

First 20 Full Precision

Precision measures the ratio of relevant documents within the total set of returned documents. The binary relevance judgment widely adopted in traditional IR evaluation, however, does not take into account the different amounts of relevant information contained in each document. In this regard, Chignell et al. proposed a "full" precision measure that sums up the total amount of relevant information contained in the first 20 documents, which seems to reflect better than binary relevance each search engine's ranking capacity (Chignell, Gwizdka, and Bodner 1999).

According to Chignell et al., the first 20 "full" precision is calculated by the following equation:

$$\mathit{First\ 20\,precision} = \frac{\sum_{i=1}^{20} userscore_i}{20 * 4}$$

Equation 1

Where:
- scorei—score assigned to the i-th hit by the judges;
- 20—number of measured hits;
- 4—maximum score that can be assigned to one hit.

User Effort Measure—Search Length i

Cooper's concept of expected search length measures user effort in terms of the number of non-relevant documents that a user must examine before finding i relevant documents. Cooper illustrated several scenarios in which different i (that is, numbers of relevant documents) may be desired based on the user's need for thoroughness. In our study we decided to set the desired number at two. The most relevant web page is defined as documents with a relevance score of three or four. Thus the search length is operationalized as
"The number of Web pages one has to examine (including relevant and non-relevant documents) before two documents with relevance score of 3 or above are found."

Rank Correlation

Su, Chen, and Dong proposed comparing the user's relevance and the system's relevance ranking in order to measure a search engine's ranking performance. The measure they proposed involved correlating the rank order assigned by the search engine and the user's preference. Our rank correlation was designed to reflect the same notion of evaluating how closely a system's ranking reflected user preference, but with a slight modification in procedure. We decided to include all 20 pages that appeared first instead of only the top five pages as was used in Su et. al. Since we do not have access to the actual ranking scores, we used a document’s position within the top 20 returned hits as the approximate ranking score. The higher a document appeared in the list, the higher the ranking score presumably assigned by the system. Thus, the top five hits were given a score of four; the next five hits were given a score of three, and so forth. The higher the correlation between the ranking scores and the user judgment scores, the more efficiently the system is able to relieve user efforts.

Results and Discussion

Overlap of Results

There was low overlap of returned hits among the three non meta-search engines. Since we customized Metacrawler on the basis of the other three systems, it is not surprising that most duplicate links occurred in Metacrawler output. For the topic Urban Planning 1, the Metacrawler returned hit sets in which 19 of 20 documents had appeared in the other three search engines' sets. For the topic of Economics 1, the comparable figure was 15 out of 20. Overall, the average overlap was 11. Notably, nearly half of the hits returned by Metacrawler were not included among the top 20 of the other three search engines, which demonstrates the re-ranking function of Metacrawler. The effectiveness of re-ranking will be discussed below in the Ranking Results section.

Currency Results

Broken-link ratio can be an indication of how frequently and thoroughly an engine checks the links in its database for currency. The numbers and percentage of broken links are shown in Table 2.

	AltaVista	Google	Excite	Metacrawler
Broken link	7 (4.38%)	5 (3.13%)	10 (6.25%)	9 (5.63%)

Table 2. Broken Links

Credibility Results

Across all four disciplines used for the test searches, the mean and standard deviation of credibility score of each search engine are listed in Table 3.

	AltaVista M(SD)	Google M(SD)	Excite M(SD)	Metacrawler M(SD)
Credibiliy	2.22(1.38)	2.68(1.28)	2.32(1.20)	2.54(1.28)

Table 3. Mean and Standard Deviation of Credibility Score

Among all pair relationships, Google was significantly better than AltaVista and Excite (Google-AltaVista: t (159) = 3.31, p < .01; Google-Excite: t (159) = 2.79, p < .01). Metacrawler provided more credible hits than AltaVista but not more than Excite (Metacrawler-AltaVista: t (159) = 2.31, p < .5; Metacrawler-Excite: t (159) = 1.89, ns). There was no significant difference between Google and Metacrawler (t (159) = 1.15, ns) nor between Excite and AltaVista (t (159) = .74, ns).

User's Effort Results: Search length 2

As discussed before, user effort was measured by a modified search length measure, that is, the number of links the user has to go through to find two relevant documents. The mean and standard deviation for this performance measure are displayed in Table 4.

	AltaVista M(SD)	Google M(SD)	Excite M(SD)	Metacrawler M(SD)
Search length 2	10.5(7.71)	4.25(1.83)	11.0(7.01)	9.00(4.78)

Table 4. Mean and Standard Deviation of Search Length

In general, at the time this study was carried out, Google significantly outperformed the other three systems on this measure as one had to go through fewer pages to find the first two satisfying and relevant documents. The performances of AltaVista, Excite, and Metacrawler were nearly equal.

Rank Correlation

	AltaVista	Google	Excite	Metacrawler
Correlations	.092	.131	.013	-.012

Table 5. Overall System Ranking vs. User Relevance Judgment

Overall, none of the correlations between system-assigned order and user's judgment score was significant although Google and AltaVista showed a slightly higher than the others. A further analysis of system performance of each individual query shows the same pattern. The only significant correlations occurred in queries submitted to AltaVista and Google. Google also had only two negative scores while the correlations of Excite and Metacrawler tended to be near zero or negative.

	AltaVista	Google	Excite	Metacrawler
Bio1	-.25	-.20	-.22	-.07
Bio2	-.11	.31	.15	.09
IE1	-.19	.33	-.24	.22

IE2	.46*	.06	-.20	.00
Econ1	.55*	.55*	-.14	-.09
Econ2	.19	-.07	-.28	-.09
Urban1	.37	.00	.30	-.14
Urban2	-.24	.11	.19	-.08
Across all queries	.09	.13	.01	-.01

Table 6. System Rank Order vs. User Relevance Judgment (* p<.05)

First 20 “full” Precision Results

	AltaVista	Google	Excite	Metacrawler
Biology	.31	.34	.21	.32
Industrial Engineering	.23	.39	.14	.21
Economics	.53	.74	.55	.60
Urban Planning	.35	.66	.46	.58
Average	.35	.53	.33	.43

Table 7. First 20 “full” Precision

Table 7 shows the full precision scores for the four search engines in the four respective domains. Figure 1 shows the profile plots of full precision across systems and domains. From the figure, one can see that Google performed best in all four domains although the difference is not necessarily significant. The varying results among the disciplines suggest that the domain will affect search engine performance. As Figure 1 makes clear, all search engines performed much better on the subset including Urban Planning and Economics than on the other.

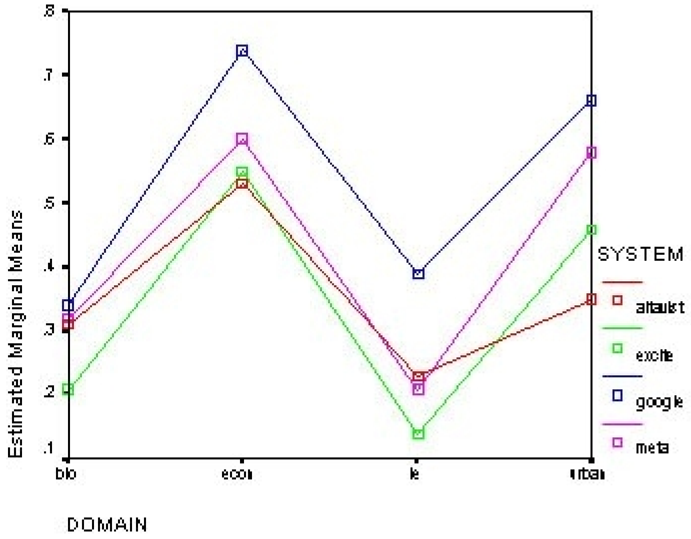

Figure 1. Profile Plots of Precision Across Systems and Domains

One should, however, be cautious about making assertions regarding any difference in search engine performance based on discipline, considering that there was only one participant in each discipline. Differing performance might be attributable to idiosyncrasies of the individual users.

Consistency

Rank	First 20 "Full" Precision	Search Length 2	Rank Correlation	Credibility
1st	Google	Google	Google	Google
2nd	Metacrawler	Metacrawler	Alta Vista	Metacrawler
3rd	Alta Vista	Alta Vista	Excite	Excite
4th	Excite	Excite	Metacrawler	Alta Vista

Table 9. Consistency — Ranking of search engine performance according to four measures.

Discussion and Conclusions

In this experiment, Google was found to outperform the other search engines on the basis of credibility, first 20 full precision, and user effort. This may be due to Google's incorporation of hyperlink information in its ranking algorithm. Metacrawler, our sample meta-search engine, failed to achieve the expected high level of performance.

The results obtained in the experiment showed little overlap in the documents returned by the different search engines except in the case of Metacrawler, confirming previous observations by Ding and Marchionini (1996). Search engines' result sets tend to have relatively low overlap because they employ different ranking techniques for indexing coverage.

There was no significant effect of interaction between search engines and subject domains on the first 20 full precision of returned hits in the experiment. Beyond its overall superior performance, however, Google seems to be particularly adept at handling natural science and engineering topics. Further analysis of the returned hits showed that Google returned more academic articles given high scores for relevance in biochemistry and industrial engineering while AltaVista and Excite returned more pages with faculty or departmental information.

The concern of the study, however, was not so much to determine the best search engine as to test various performance measures and investigate their theoretical implications. As pointed out earlier, we felt that the most widely used measures in traditional IR evaluation, namely, precision and recall, do not seem to be well suited to the Web search environment. During our literature review, we found that Cooper's notion of search length provided an appropriate evaluative framework for the Web environment, in which users' search behaviors suggest that ranking capacity is crucial. Our three ranking measures can all be properly explained by Cooper's principle of optimizing search order, that is, an ideal search engine should be able to deliver most relevant documents higher in the rank list. We were able to demonstrate in our findings a certain degree of consistency in these ranking measures and in credibility and currency measures (see Table 9). More important, the consistency also presents among three ranking evaluation measures, more so between first 20 full precision and search length 2 and to a less degree between rank correlation and the other two measures. We suspect the lower level of consistency between rank correlation and the other two measures is the result of the imprecise representation of system-assigned scores. Our transformation of system assigned scores (1-20) into the same scale as user judgment score (0-4) might not correctly capture the real ranking scores assigned by the systems. An accurate assessment of the applicability of the rank correlation measure is more likely when using actual ranking scores assigned by the search engines.

Whereas Cooper discussed different scenarios of search thoroughness, in our study the measure of search length 2 only took into account one specific situation in which the thoroughness of the search was limited. We feel this is proper considering that the evidence shows that average Web users are sensitive to the effort they have to expend when using search engines. This may not be the case when thoroughness of search is desired. Further studies are needed to explore the tradeoff between user effort and information acquired when different degrees of thoroughness of search are desired.

References

Agata, T., Nozue, M., Hattori, N., & Ueda, S. (1997). A measure for evaluating search engines on the World Wide Web: Retrieval test with ESL. *Library and Information*, 37, 1-11.

Chignell, M. H., Gwizdka, J., & Bodner, R. C. (1999). Discriminating meta-search: A framework for evaluation. *Information processing and management*, 35(3), 337-362.

Chu, H., & Rosenthal, A.U. (1996). Search engines for the world wide web: A comparative study and evaluation methodology. *Proceedings of the 59th Annual Meeting of the American Society for Information Science, Baltimore, M.D.*, 127-135.

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

Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *Journal of American Society of Information Science*, 19(1), 30-41.

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

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

Leighton, H. V., & Srivastava, J. (1997). Precision among World Wide Web search services (search engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos. [online] available at www.winona.edu/library/webind2/webind2.html

Oppenheim, C., Morris, A., & McKnight, C. (2000). The evaluation of WWW search engines. *Journal of Documentation*, 56 (1), 71-90.

Silverstein, C., Henzinger, M., Marais, J. & Moricz, M. (1998). Analysis of a very large Alta Vista query log. *Technical Report 1998-014, COMPAQ Systems Research Center, Palo Alto, Ca, USA*, 1998.

Su. L. T., Chen, H. L., & Dong, X. Y. (1998). Evaluation of Web-based search engines from an end-user's perspective: A pilot study. *Proceedings of the 61st Annual Meeting of the American Society for Information Science, Pittsburgh, PA*., 348-361.

Wishard, L. (1998). Precision among Internet search engines: An earth sciences case study. *Issues in Science and Technology Librarianship 1998*.
[online] available at http://www.library.ucsb.edu/istl/98-spring /articles5.html