

Towards Information Retrieval Measures for Evaluation of Web Search Engines

Jacek Gwizdka and Mark Chignell
Department of Mechanical and Industrial Engineering
University of Toronto
5 Kings College Rd, Toronto, ON M5S 3G8, CANADA
jacek@ie.utoronto.ca, chignel@mie.utoronto.ca

Abstract

Information retrieval on the Web is very different from retrieval in traditional indexed databases. This difference arises from: the high degree of dynamism of the Web; its hyper-linked character; the absence of a controlled indexing vocabulary; the heterogeneity of document types and authoring styles; the easy access that different types of users may have to it. Thus, since Web retrieval is substantially different from information retrieval, new or revised evaluative measures are required to assess retrieval performance using Web search engines. This paper suggests a number of different measures to evaluate information retrieval from the Web. The motivation behind each of these measures is presented, along with their descriptions and definitions. In the second part of the paper, application of these measures is illustrated in the evaluation of three search engines. The purpose of this paper is not to give the definite prescription for evaluating information retrieval from the Web, but rather to present some examples and to initiate a wider discussion of how to enhance measures of Web search performance.

Keywords

Web information retrieval, information retrieval measures, evaluation, search engines, precision, user interaction, search length.

Introduction

Information retrieval systems have been evaluated and compared for many years. Cleverdon (1966) listed six criteria that could be used to evaluate an information retrieval system: (1) *coverage*, (2) *time lag*, (3) *recall*, (4) *precision*, (5) *presentation* and (6) *user effort*. Of these criteria, recall and precision have most frequently been applied in measuring information retrieval. Both of these measures have been criticized for a variety of reasons, and a range alternatives have been suggested (for example: Cooper, 1968; Van Rijsbergen, 1979; Meadow, 1992; Hersh and Molnar, 1995; Tague-Sutcliffe, 1996; Ellis, 1996; Hersh, Pentecost and Hickam, 1996). However, in spite of their deficiencies, recall and precision are continued to be used widely, in part because of insufficient consensus about which alternative measures might be superior.

The explosive growth of the Web has brought about an increased need to examine the fitness of these measures for evaluating information retrieval in this new environment, and to consider alternatives. The Web differs from traditional information retrieval systems in many ways. The

interconnected character of the Web and its expansive user population are two major factors affecting evaluation of information retrieval from the Web.

As pointed out by Samalpais, Tait and Bloor (1998), the "relevance"¹ in an inter-linked collection of documents is not only determined by each document considered separately, but also by the inter-linked structure of the whole system. In such an approach, even a "non-relevant" document becomes partially "relevant", when it is linked to "relevant" documents. Thus the revised measures should take into account an inter-document link structure. In addition, the typical user population has changed from trained specialists in information retrieval to an overwhelming majority of Web users who have little if any training on how to conduct information searches. Thus user interaction with the system has become crucial, because experts are generally much better at adapting to different types of interface than are novices, whose performance is greatly affected by the type of interface used (e.g., Charoenkitkarn, 1996)

Since Cleverdon listed six broad evaluative criteria in 1966, recall and precision have received the lion's share of the attention. However, presentation and user effort are two other criteria that assess information retrieval in user terms. Unlike recall and precision, these criteria have not been studied in depth, and there is no consensus agreement on how they should be measured.

Most of the discussions of measurement of information retrieval found in the literature are general and do not include aspects of particular significance to the Web. While some recent models of information seeking have explicitly considered the Web environment, they have generally not been sufficiently specific to generate evaluative measures for information retrieval from the Web. A few notable exceptions are work on relative distance retrieval measure for hypermedia libraries by Samalpais, Tait and Bloor (1998), work on time-to-view graphs applied to AltaVista by Dunlop (1997), and two comparative studies of Web search engines that incorporated some information on links between documents into their precision measures (Ding and Marchionini, 1996; and Leighton and Srivastava 1997).

Starting from the traditional information retrieval measures and building on the previous work, this paper suggests using a collection of evaluation measures to information retrieval from the Web. The measures are defined and their application to evaluation of three Web search engines is shown.

Related work

Alternative methods of information retrieval evaluation

While various measures of recall and precision have been predominant, there have been a number of attempts to develop measures of efficiency that combine recall and precision (e.g., Meadow, 1992) and to develop alternative measures. Among the less frequently used measures

¹ The evaluation measures under discussion are sensitive to the definition of relevance. Their form, however, is independent of the meaning of relevance, and, thus, argument of this paper does not require assumption of any particular relevance. The relevance was operationalized for the purpose of the reported study. In depth discussion of relevance in information retrieval can be found in for example in: Saracevic (1995) and (1996), Harter (1996), Mizzaro (1997).

is *expected search length*. Originally suggested by Cooper (1968), it was described in detail by Van Rijsbergen (1979). Assuming a linear ordering of documents retrieved by an information retrieval system, with at most one document at a given level of ordering, search length, in the simplest case, is defined as the number of non-relevant documents that a user must examine before finding n relevant documents. In the more general case, where several documents can be at the same ranking level and their ordering within that level is random, a similar and more general measure called expected search length can be defined. However, with current state-of-the-art of Web search engines, retrieval results are presented linearly. Even if an explicit relevance degree is assigned to documents and shown, the grouping into different levels of ranking is not apparent, and thus the simple search length seems to be sufficient for Web measures based on current practice.

Dunlop (1997) used the expected search length to construct graphical evaluation methods: number-to-view (NTV) and time-to-view (TTV) graphs. NTV and TTV graphs were applied by him to measuring retrieval performance from AltaVisa. These graphs were introduced as supplementary to precision-recall graphs.

Evaluation of search engines

Since the Web became a household word, many popular and trade magazines have performed comparisons of Web search engines. *Internet World* and *PC World* have performed lab tests of search engines (Overton, 1996; Scoville, 1996; Venditto, 1996), similar tests have been performed by ZDNet (e.g., Lake, 1997). Other comparisons are also published on-line by various organizations and individuals, for example Search engine watch (1998), Ralph (1997). There is no well-defined methodology for these tests, and the methods of evaluation are frequently not fully specified in the published reports.

Meghabghab and Meghabghab (1996) examined the effectiveness of five World Wide Web search engines (Yahoo, WebCrawler, InfoSeek, Excite, and Lycos) by measuring precision on five queries. They found that Yahoo obtained the best performance, followed by InfoSeek and Lycos.

Chu and Rosenthal (1996) evaluated the capabilities of Alta Vista, Excite, and Lycos in terms of retrieval performance measured by response time and precision calculated for the first ten hits. Their study also proposed a methodology for evaluating WWW search engines in terms of five aspects:

1. Composition of Web indexes (coverage) – collection update frequencies and size can have an effect on retrieval performance;
2. Search capability – they suggest that search engines should include “fundamental” search facilities such as Boolean logic and scope limiting abilities;
3. Retrieval performance (precision, recall, time lag) – such as precision, recall, and response time;
4. Output option (presentation) – this aspect can be assessed in terms of the number of output options that are available and the actual content of those options;
5. User effort (user effort) – how difficult and effortful it is to use the search engine by typical users.

The five aspects roughly correspond to the criteria listed by Cleverdon (1966), with the addition of the search capability test as a further evaluative measure.

Su, Chen, and Dong (1998) measured precision and partial precision for the first twenty hits returned by AltaVista, Infoseek, Lycos, and Open Text. They also defined an evaluative measure that compared ratings of relevance on a 5-point scale (where “1” characterized the most relevant items and “5” characterized the least relevant items). They then correlated these evaluative rankings with the machine rankings for the top 20 documents returned by each search engine. Intuitively, a higher correlation in this case would indicate that the relevance ranking by the search engine fit the human assessment of relevance better. However, one must always be careful of the metric properties of measures when carrying out this type of analysis. For instance, as discussed by Chignell, Gwizdka, and Bodner (in press), a “perfect” search engine that had almost all highly relevant scores (rated as “1” documents in the top 20 hits) would have a low correlation with the subjective ratings since the subjective ratings would mainly be “1” while the search engine rankings would go from 1 to 20. On the other hand, a search engine that had a range of high and low relevance documents in the first 20 hits might have a higher correlation, if the documents judged to be of lower relevance tended to be further down the list (in the first 20 hits). Thus a measure such as this would not be suitable for search engines that did a good job of ranking and that had mostly relevant documents at the top of their ranked output list.

Schlichting and Nilsen (1997) examined Alta Vista, Excite, Infoseek, and Lycos. They conducted a small empirical study (with five participants) and used Signal Detection Analysis to analyze the data. Signal detection analysis provided a method for distinguishing between sensitivity (d') and response bias in determining whether documents were relevant or not.

Precision measures typically employed by studies on Web search performance do not take into account links between Web documents. In contrast, Samalpais, Tait and Bloor (1998) present a systematic approach to measuring relevance in hyper-linked networks of documents. The calculation of a relative distance relevance (RDR) metric, introduced in their methodology, is based on the matrix of distances of document nodes in the network and on the binary (yes/no) relevance of these documents. This calculation of relevance takes into account both querying and browsing strategies of information seeking. For very large networks of documents, like the Web, however, calculations may be computationally too intensive to be practical. Moreover, knowledge of which of the many documents in a hyper-linked network are relevant is not feasible. Thus the proposed methodology would need to be simplified to be practically usable.

Ding and Marchionini (1996), in their study of three search engines (InfoSeek, Lycos, and OpenText), also took into account the impact of hyperlinking. They evaluated precision and result overlap among the tested search engines. Precision was calculated on the base of a scoring system (from 0 to 5) that took into account links to other relevant documents.

Leighton (1995) evaluated the performance of four index services: Infoseek, Lycos, Webcrawler, and WWWorm. Employed measures included: average top ten precision and response time. Leighton and Srivastava (1997) carried out a follow up study that compared order-weighted precision scores for Alta Vista, Excite, HotBot, Infoseek, and Lycos. In this study, the first 20 results returned for 15 queries were examined. Returned documents were

assigned relevance score in the range 0 to 3. The relevance scores in this study also took into account links to relevant pages. The final metric was composed of weighted relevance scores. The weighting took into consideration the position of a returned document in the ranked list. In this way, a measure of goodness of ranking was incorporated into the proposed metric

While a wide range of evaluation studies of Web search engines has been performed, only a few researchers used measures incorporating hyper-linked structure of the Web. Similarly, although user effort has been recognized as an important factor, only one study (Dunlop 1997) performed quantitative evaluation of user effort in the context of information retrieval from the Web.

Further studies are needed to compile a “track record” on different evaluative measures so that fair comparisons of search engine performance can be made in future. It remains to be seen whether one evaluative measure can sufficiently express the various phenomena that may occur when searching for documents or Web pages within a hyper-linked network. With the present state of knowledge, differences in observed search engine performance may reflect the properties of the evaluative measures used more than they do fundamental differences in the effectiveness of the search engines for the particular topic and information sources used.

Evaluative Measures

The six criteria given by Cleverdon provide a framework for designing information retrieval measures that is still valid over thirty years after their introduction. The discussion of measures introduced in this section is structured according to four of these criteria. The two criteria that are not used are first discussed briefly.

Time lag is obviously a very important factor for users of information retrieval from the Web. While it is important, it is also highly dependent on the level of Internet traffic at the time of measurement. This introduces a large amount of error variation into the measure, because the level of Internet traffic varies widely over both time and location. Thus, at the current stage of Internet technology, it is difficult to use time lag as a discriminative measure.

The measurement of recall on the Web is also problematic due to the extremely dynamic character of the Web, its very high changeability, and its huge size. Attempts to develop "pragmatic" methods of assessing recall on the Web, such as by using consensus or overlap in hits returned by different search engines may also produce mixed results, particularly when a search engine returns a high proportion of unique, but relevant, hits (Chignell, Gwizdka, and Bodner, in press).

Before discussing evaluative measures, the model of calculating relevance for those measures needs to be described. Documents which are directly relevant are assigned a score called the base relevance, denoted by br . The overall relevance of these documents, denoted by r , has only this one component, and thus, $r = br$. In order to incorporate information about links between documents, documents which are not directly relevant are assigned a relevance score composed of two components: the base relevance br and the distance component, denoted by dr . The overall relevance r is in this case calculated as:

$$r = \begin{cases} br - dr, br \geq dr \\ 0, br < dr \end{cases}$$

Equation 1

In this model, dr is calculated as, in the Web terminology, a number of links that need to be traversed between a “non-relevant” document, for which we calculate the relevance, and the document which has relevance br . In practical calculations $dr \leq 2$.

The model of relevance suggested by Samalpais, Tait and Bloor (1998) presents an “ideal” case for calculation of relevance in a hyper-linked document system. The rationale behind the simpler model of relevance proposed here is to obtain a method which is computationally simple and sufficient in practice. This model is based on several assumptions. It is argued that the model is sufficient because it reflects typical user behaviour during an information retrieval session on the Web. Anecdotal evidence suggests that users tend to examine the first 10 or 20 results, and they do not usually follow links very deep (1 or 2 levels). Thus, only local link information needs to be incorporated into the relevance assessment.

The relevance model is operationalized by assigning subjective relevance scores as presented in Table 1.

Relevance score	Description
3	the most relevant
2	Partly relevant or contains a link to a page with a score of 3
1	Somewhat relevant, for example, short mention of a topic within a larger page, technically correct (i.e. terms appear on a page - including META tags) or contains a link to page ranked 2
0	not relevant; no query terms found (META tags were examined as well) or a "bad" hit .

Table 1. Description of subjective relevance scores

Web search engines often return a lot of documents. As indicated earlier, users tend to view only a handful of them (10-20). It is, thus, reasonable to limit the number of hits considered in calculations. This number is denoted by n . For the purpose of this work n has been set to 20.

Precision

Precision measures the ratio of relevant documents within a given number of documents returned to the number of returned documents. By using the relevance measure described above, precision incorporates some information about the inter-document link structure. Four different precision measures are described below. The measures differ in terms of how the relevance scores are used. As noted earlier, precision is calculated for the first 20 documents, and hence called *first 20 precision*.

1. *Full* precision

This measure takes fully into account the subjective score assigned to each hit.

$$\text{precFull}(\text{minFnHits}) = \frac{\sum_{i=1}^{\text{minFnHits}} \text{score}_i}{\text{minFnHits} * \text{maxHitScore}}$$

Equation 2

where:

- score_i - score assigned to the i -th hit,
- n - number of measured hits
- $\text{minFnHits} = \min(n, \text{hitsReturned})$,
- hitsReturned - total number of hits returned,
- maxHitScore - max score that can be assigned to one hit (3).

This first measure makes an assumption that relevance scores are additive. A measure-theoretic analysis has not been yet performed to support this claim. Thus, while it is assumed that 2 documents with the highest scores (3) are equivalent to 6 documents with scores = 1, this assumption has yet to be tested or verified.

2. *Best* precision

This measure take into account only the most relevant hits. It maps relevance scores to a binary measure, and thus the additivity assumption is not necessary in this case.

$$\text{precBest}(\text{minFnHits}) = \frac{\text{count_of}(\text{score}_i = 3)}{\text{minFnHits}}$$

Equation 3

3. *Useful* precision

This measure take into account only the most relevant hits and hits containing links to the most relevant ones.

$$\text{precUse}(\text{minFnHits}) = \frac{\text{count_of}(\text{score}_i \geq 2)}{\text{minFnHits}}$$

Equation 4

4. *Objective* precision

This is an objective measure since it does not rely on human relevance judgment. It is based on computed presence or absence of required terms and on the distinction between good and bad links.

$$\text{precObj}(\text{minFnHits}) = \frac{\text{count_of}(\text{score}_i > 0)}{\text{minFnHits}}$$

Equation 5

The following variable names were used in the search engine study described below to denote the precision measures (in the same order as for the four precision measures described earlier in this section): PRECFULL, PRECBEST, PRECUSE, PRECOBJ.

Presentation - Ranking

The order of ranking search results is one of the aspects of presentation. To assess the quality of ranking, a *differential precision* measure is introduced. Variations on the measure of average precision seem particularly promising for assessment of search engines with ranked output. Differential precision measures variation in distribution of relevant hits within the 20 first returned hits. A higher concentration of relevant hits in the first 10 hits than in the second 10 hits is desired from the user point of view, since it allows users to focus more on the hits at the top of the ranking.

The differential precision is calculated as a difference in the number of documents that were relevant in the first ten (1-10) and in the second ten (11-20) ranked documents. If there were more relevant documents in the first ten hits, this would suggest that the ranking process was working.

Differential objective precision is calculated as follows:

$$\text{dpObj}(1, \text{minF20Hits}) = \text{precObj}(1, \text{minF10Hits}) - \text{precObj}(\text{minF10Hits}, \text{minF20Hits})$$

Equation 6

"Full" (dpFull) and "useful" (dpUse) differential precisions were calculated in an analogous way, using full and useful precision scores as defined above. The full differential precision was analogous to full precision (as described earlier) in summing all the scores (1, 2, and 3, as described in Table 1). The useful differential precision, calculated the differential precision based on relevance scores of two or three.

Each differential precision has the following properties:

- $\text{dpObj} > 0 \Rightarrow$ more relevant documents in the first 10 hits than in the second 10
- $\text{dpObj} = 0 \Rightarrow$ number of relevant documents in the first 10 hits and in the second 10 is the same
- $\text{dpObj} < 0 \Rightarrow$ less relevant documents in the first 10 hits than in the second 10

The following symbols were used in the search engine study to denote the differential precision measures: DPFULL, DPUSE, DPOBJ.

User Effort - Search length *i*

With a growing diversity of Web users and a growing number of casual and untrained users, the effort that the users are willing to devote to finding information on the Web decreases. Measuring this effort becomes, thus, very crucial to success of Web search services.

Search length, as described earlier in the related work section, measures user effort in terms of the number of non-relevant documents that a user must examine before finding *i* relevant documents. A modification of this metric presented below is based on the number of web pages

to be examined by the user (both relevant and non-relevant) before finding the i most relevant pages. Finding the most relevant web pages is satisfied by pages with relevance score 3 or pages with relevance score 2 which contain link to a page, or to pages, with score 3. All pages that need to be examined until i most relevant pages are found are counted as 1, with the exception of pages with links to the most relevant pages which are calculated as 2 (1 for a hit plus 1 for an additional link²).

$$fSLen_i = 1 - \frac{\max SLen_i - sLen_i}{\max SLen_i - \text{best} SLen_i}$$

Equation 7

where:

- $\max SLen_i$ - is the maximum search length for i relevant web pages within n returned search hits
- $\text{best} SLen_i$ - is the best (i.e. the shortest) possible search length for i relevant web pages
- $sLen_i$ - search length for i most relevant web pages
- The range of function $fSLen_i$ is $\langle 0;1 \rangle$. Where 0 is the best, that is the shortest search length.

Calculations of $fSLen_i$ were performed for $i=1$ and $i=3$. The following symbols were used in the search engine study to denote the above search length measures: FSLEN1, FSLEN3.

Coverage

Calculating overall coverage of search engines is in itself a complicated task. A statistical sampling method for measuring overlap among search engines and their relative coverage has been developed by Bharat and Broder (1998). The two metrics proposed here are used to measure relative ratio of coverage and overlap only for results of a given query. The metrics are: number of unique hits, which measures overlap, and the relative number of returned hits to total hits in a given domain, which measures a relative ratio of coverage³ for a given query. These measures are comparative, in that they are applicable to comparisons of Web search engines.

Hits and Hit ratio

The total number of hits returned as a result of a query is denoted *hits*. *Hit ratio* is calculated as the ratio of the total number of hits returned as a result of a query (*hits*) to the total number of hits returned by a given search engine in a given domain (*totalhits*).

Uniqueness

Uniqueness of hits is measured by analyzing the hits returned by search engines under comparison. The *unique* number of hits is calculated by counting hits returned by this search engine and not by others.

The following symbols were used in the study reported below to denote the above measures: HITS, HITRATIO, UNIQUE.

² To simplify the calculations, only one level deep was examined. As it was pointed out earlier, this simplification corresponds to the patterns of user behaviour on the Web.

³ An assumption is made here that, while search engines may index different sets of Web documents, the average coverage for sets of various topics should be similar across different search engines.

Other measures

The number of bad links, and the number of duplicate links are two additional metrics. These measures may only be applicable for the initial years of the Web, when there are "many potholes in the information highway". Given the current state of the Web, both of them seem to provide an important additional measure of Web search engine quality.

A bad link is a link of any of the following types: not found web pages, not responding servers, security protected web pages, server errors, no DNS entry errors. Duplicate links are the links with the same URLs, or with URLs that could be easily recognized⁴ by search engines. The following symbols were used in the search engine study to denote the above measures: BAD, DUPL.

Application to Evaluation of Web Search Engines

This section illustrates application of the described above measures in the comparative evaluation of three Web search engines: AltaVista, HotBot, and Infoseek. This experiment sought to examine the effect of three search engines and six Internet domains on a set of defined above performance measures. The examined Internet domains included four country specific domains: Germany (.de), Austria (.at), Poland (.pl), and the UK (.uk), the general "commercial" sites (.com), and "organizational" sites (.org).

A set of queries was posed to each search engine. One query was formulated for each of the following four topics:

Topic 1 - Find information on national museums

Topic 2 - Find currency exchange rates

Topic 3 - Find information related to the year 2000 problem, but no apocalyptic visions

Topic 4 - Find train schedules, but not training schedules

The queries were expressed in three different languages, appropriate to the country in which a given Internet domain was located (including the slight differences between American and British English). The four queries used in the experiment are shown in Appendix A.

Results

Full factorial MANOVA was carried out using search engines and domains as the independent factors and with the dependent measures described above. A significant multivariate interaction between search engines and domains was found ($F(221,425.24)=1.56, p<0.001$). Interaction between search engines and domains were found to have significant univariate effects on the following measures: number of unique hits (UNIQUE; $F(17,50)=2.11, p=0.021$), total number of hits (HITS; $F(17,50)=4.26, p<0.001$), ratio of returned hits to each search engine collection sizes (HITRATIO; $F(17,50)=1.92, p=0.038$), quality of returned hits (BAD; $F(17,50)=2.40, p=0.008$), and borderline significant effect on search length 1 (FSLEN1; $F(17,50)=1.72, p=0.069$).

⁴ Obvious case are the same URLs. Other very common cases are `<URL><path>/` and `<URL><path>/index.html` or and `<URL><path>/default.htm`

Overlap of results

There was surprisingly low overlap among the hits returned by the three search engines. The dip in the number of unique hits for Infoseek in the “de” and “at” domains reflects the low numbers of hits typically returned by Infoseek in those domains. Aside from this fact, the overlap among returned hits was small across all domains and search engines.

Number of returned hits

One of the problems in comparing search engine performance is the different coverages that search engines have. This is particularly true across the different Internet domains, as indicated in Table 2. Table 2 shows the total number of hits in each domain for each search engine. It also shows the mean number of hits in each domain returned by each search engine in the experiment.

Search Engine	Domain	Means of hits returned as a result of a query	Total number of hits for each domain
AltaVista	de	952.00	5,796,668
	at	131.00	625,174
	pl	219.75	404,604
	uk	2207.50	5,060,051
	com	16010.75	49,165,966
	org	4065.25	6,934,946
HotBot	de	775.50	4,647,297
	at	111.75	815,893
	pl	171.75	502,925
	uk	1851.00	3,471,982
	com	14948.25	33,962,466
	org	4320.50	5,538,953
infoseek	de	6.75	2,141,013
	at	1.25	227,588
	pl	31.75	88,777
	uk	35.25	2,228,112
	com	29.75	27,626,808
	org	24.25	3,651,048

Table 2. Coverage of the Search Engines across the Internet domains.

It can be seen from Table 2 that Infoseek generally returns a disproportionately small number of hits.

Both main effects (search engines and domains), and their interaction, had significant effects on the total number of hits returned as a result of each. Infoseek always returned fewer hits than

both AltaVista and HotBot (as shown in Table 2). While the absolute number of hits returned by each search engine (HITS) could vary because of the various sizes of collections indexed by each engine (TotalHits), it was reasonable to expect only relatively small differences among ratios of HITS to TotalHits (HITRATIO). However, Infoseek returned unexpectedly few hits in all tested Internet domains with the exception of "Poland" (pl), while both AltaVista and HotBot returned approximately similar percentages of the indexed collection size.

Quality of Returned hits

There was a significant interaction between domains and search engines for Bad Hits. A disproportionately large number of bad hits appeared in the results returned by HotBot from the United Kingdom (uk). That is surprising, since, according to Search Engine Watch (Search Engine Watch 1998), HotBot refreshes its database more often (about once a week) than the other two search engines (AltaVista every 1-2 weeks, Infoseek every 1-2 months), and thus it should not have such problems. A possible explanation may be that the data provided by Search Engine Watch may be applicable only to US web sites, and other domains may be re-indexed less often⁵.

Search Length

A borderline significant interaction was found on search length 1. Infoseek performed worst on this measure in the "German"⁶ domains (de, at), while HotBot had the worst performance in "Poland" (pl), as shown in Figure 1. A possible explanation may lie in the use of languages other than English. Fine tuning of indexing of web pages written in other languages, like German and Polish, may require using modified versions of algorithms. For example, word stemming is language dependent. In this study, AltaVista seemed to be generally less affected by the use of languages other than English. In general, Alta Vista at the time this study was carried out paid more attention to "foreign" languages than the other search engines, as could be seen from the availability of other language versions of the main Alta Vista search engine interface and also from the translation services that were offered. Infoseek's relatively poor performance in terms of FSLEN1 may be due to the relatively small number of pages that it indexes, particularly in countries like Poland.

Analysis of Main Effects

Full factorial multivariate analysis was carried out using search engines and domains as the independent factors, with the fourteen dependent measures described above. The multivariate main effect of search engine was significant ($F(24,78)=2.43$, $p=0.002$). The multivariate effect of domain was also significant ($F(60,186.4)=2.39$, $p<0.001$). Separate univariate analyses were then carried out to determine the source of these effects. Significant univariate effects of search engine were found on Differential objective precision (DPOBJ; $F(2,50)=5.89$, $p=0.005$), on best and full precision (PRECBEST; $F(2,50)=7.19$, $p=0.002$ and PRECFULL; $F(2,50)=5.85$, $p=0.005$, respectively), and on search length 1 ($F(2,50)=3.85$, $p=0.028$).

⁵ Effects of the less often re-indexing are highly dependent on the dynamics of web sites in a given domain. In static domains, the effects could be negligible. It is possible that the web sites located in uk and pl exhibit different kind of dynamics which cause different effects (bad hits as opposed to duplicate hits).

⁶ Note that the domains represent "virtual" countries.

Figure 1 shows the best (left panel - PRECBEST) and full (right panel - PRECFULL) precision scores (where the error bars represent 95% confidence interval) for the three search engines. It can be seen that, using a human judge, Infoseek did relatively poorly. This is in contrast to Experiment 1 in a study reported by Chignell, Gwizdka, and Bodner (in press), where Infoseek obtained relatively high precision when peer consensus review (with other search engines acting as the panel of referees) was used. Note that the relatively good performance of AltaVista in this study is consistent with the relatively good performance for Alta Vista that has been observed in previous studies.

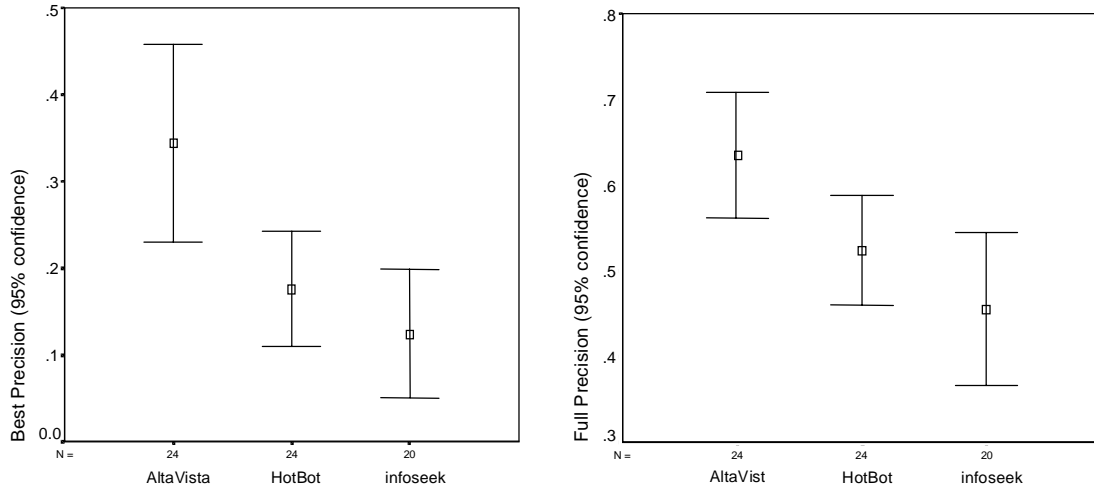


Figure 1. Best and full precisions of the three search engines.

The user effort involved in finding relevant web pages among the returned hits was indicated by the search length (FSLEN) measures. Alta Vista also did well in terms of FSLEN1 (as shown in Figure 2), with few pages needing to be read prior to finding the first relevant document.⁷

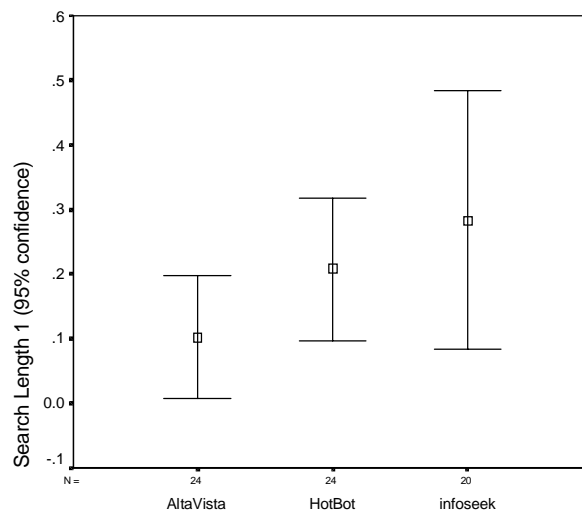


Figure 2. Search length 1 for the three search engines.

⁷ Search length was also calculated for finding three relevant documents (FSLEN3), but the effect of search engines on it were not found to be statistically significant.

Differential precision reflects how well the first 20 hits have been ranked. Figure 3 shows the differential objective precision for the three search engines. The differential objective precision (DPOBJ) was best for Infoseek, with the relevant documents tending to be strongly concentrated within the first ten returned hits ($DPOBJ > 0$).

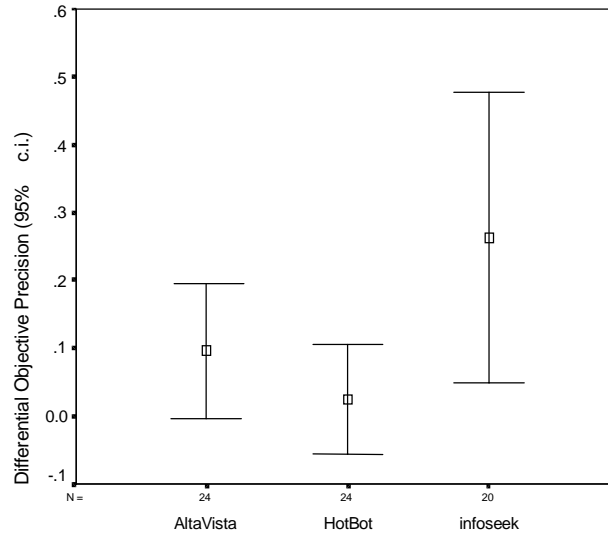


Figure 3. Differential objective precision of the search engines.

However, Infoseek often returned less than the examined twenty hits (in 13 out of 24 cases) and sometimes even less than ten hits (9 out of 24 cases). Thus, the small number of returned hits positively skewed the value of DPOBJ for Infoseek. This illustrates the type of problem that can occur when using general measures based on relevance and precision that do not take into account the specific properties of the search engines being studied.

For AltaVista and HotBot, there was little difference between the number of relevant hits on the first page of the list (items 1 through 10) versus the second page (items 11 through 20). Thus in this study there was little evidence that the ranking processes used by those search engines led to a higher density of relevant documents on the first page of hits (results). Since people typically (based on anecdotal evidence) only peruse the first few hits (or in some cases, the first few pages of hits) reported by a search engine, a search engine which returned a lot of relevant hits spread uniformly over several pages of output could subjectively appear to be "worse" than a search engine which returned fewer, less relevant hits (e.g. Infoseek), but with better relevance ranking of those hits.

Discussion of Experiment Results

AltaVista was found to have better performance in terms of precision and differential precision. This finding is consistent with the relatively good performance for Alta Vista that has been observed in previous studies (Chu and Rosenthal, 1996; Leighton and Srivastava, 1997).

The results obtained in the experiment showed surprisingly little overlap in the documents returned by different search engines, confirming observations made earlier by Ding and Marchionini (1996) and by Bharat and Broder (1998). Search engines tend to have relatively low

overlap between their result sets because they employ different means of matching queries to relevant documents, and because they have different indexing coverage.

Since most users will not exhaustively scan through hundreds or even thousands of hits, perceived precision also depends on the quality of relevance ranking of a search engine (i.e., how well it manages to put the most relevant documents for a query at the top of the list of returned hits). Since HotBot used a unique method for ranking retrieved documents (at the time of this writing) it tended to put different documents at the top of its list. For example, words found in the title were more important (were weighted more) in HotBot's relevance ranking than words found in the body of the document.

There was no significant effect of interaction between search engines and Internet domains on the precision of returned hits in the experiment. However, Alta Vista clearly had better coverage in the different domains, and Alta Vista generally seems to be more adept at handling languages other than English. The generally lower quality of hits (bad and duplicate links) in Internet domains located in Austria, Poland, and the United Kingdom may be due to a lower frequency of re-indexing web page collections located in these domains. Precision of returned hits was not found to be affected by the interaction between the three search engines and the six Internet domains.

Infoseek (at the time the study was performed, in early 1998) had relatively poor coverage outside the .com domain and should probably not have been relied upon in domains where English was not the dominant language. Infoseek had particularly bad coverage in Poland at the time of this study, indexing fewer than 90,000 pages, versus over half a million pages in the .pl domain for HotBot. HotBot had a disproportionately large number of bad hits for the .uk domain. In general, HotBot like Infoseek should have been used with caution when non-English pages were of interest. Given the present results, it seemed wise to be skeptical of mainstream search engine performance outside the .com and .org domains. The possible exception to this was Alta Vista, which appeared to be much less language and domain sensitive.

As in previous studies, the experiment found a high number of unique hits when comparing results sets across search engines. In addition there was little evidence to suggest that relevance ranking was successful in bringing a high proportion of relevant documents to the first page of output. This suggested a general problem with search engines (unreliability, and insensitivity to document relevance) which should be addressed by the search engines.

Conclusions and Recommendations

It is time to re-examine the evaluative methods that are used in information retrieval studies, particularly with respect to search performance on the Web. New measures can be defined that take into account hyperlinking and relevance ranking. The measures presented in this paper are only a few of many such measures that could be defined. Ultimately, it would be useful to derive a set of evaluative dimensions that describe search engine performance, possibly based on a principle components or factor analysis of a large set of basic evaluative measures. In the spirit of Cleverdon's (1966) paper, it seems likely that there will be a number of useful dimensions of evaluation rather than a single combined metric. In order to carry out such an exercise it is first

necessary to define a wide range of possible evaluative measures, ideally representing the views and interests of a broad range of researchers in the area of search engine and information retrieval evaluation.

Methods of evaluation will, in turn, drive search engine design. For instance, if methods are developed for evaluating the performance of relevance rankings, then these should stimulate search engine developers to tune their ranking methods to these measures. This type of tuning would work in much the same way as when hardware and software developers tune their products so as to perform well on standard or well accepted benchmarks. In addition to a marketing motivation, well defined evaluative measures may have diagnostic utility in suggesting areas for improvement in the design of web search engines and to provide guidance on how to modify searching algorithms and presentation techniques.

More extensive studies are needed to assess the properties of different search engines and evaluative measures, particularly in realistic search tasks. Studies have shown that some measures differentiate against current search engines whereas others do not. Researchers need to determine whether these lack of differences stem from the insensitivity of the measures or a genuine lack of difference between the search engines on the attribute being measured.

Further work is also needed on the development of refined measures that take link structure (Web connectendess) more fully into account. Some researchers are working on improvements to search engine algorithms that incorporate web connectedness information (e.g. Bharat and Henzinger 1998; Kleinberg 1998) and corresponding evaluative measures and studies are needed to provide foundational research that can guide design.

While many different measures can be defined, they must be used appropriately. In the experiment reported above, we provided some initial results on how well a number of measures work in practice. Further studies are needed to gather data on different evaluative measures so that fair comparisons of search engine performance can be made in future. Studies are also needed to compare and contrast different evaluative measures so that one can differentiate between those search engines that work in similar fashion, and those that represent genuinely different dimensions of evaluation.

Given the present state of knowledge, differences in observed search engine performance may reflect the properties of the evaluative measures rather than fundamental differences in search engine effectiveness. Improved evaluative tools are needed based on a solid research foundation. The development of evaluation methodologies for search engines remains a challenging but important area of research interest that is directly relevant to cutting edge design of search engines.

Acknowledgements

This research was assisted by an operating grant to the second author from the National Science and Engineering Research Council (NSERC) of Canada. Some of the experimental results were reported earlier by Chignell, Gwizdka, and Bodner (in press), in the context of meta-search.

References

- Bharat, K., and Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Proceeding from WWW7 Conference*. [On-line] Available: <http://www7.scu.edu.au/programme/fullpapers/1937/com1937.htm>
- Bharat K., and Henzinger, M.R. (1998). Improved Algorithms for Topic Distillation in Hyperlinked Environments. *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*. 104-111.
- Charoenkitkarn, N. (1996). The Effect of Markup-Querying on Search Pattern and Performance in Large-Scale Text Retrieval. Unpublished Ph.D dissertation, Department of Industrial Engineering, University of Toronto, Toronto, Canada.
- Chignell, M.H., Gwizdka, J., and Bodner, R.C. (in press). Discriminating meta-search: A framework for evaluation. *Information Processing and Management* (Special issue on Digital Libraries).
- Chu, H., Rosenthal, M. (1996) Search engines for the world wide web: a comparative study and evaluation methodology. *Proceedings of the Annual Conference for the American Society for Information Science*, 127-135.
- Cleverdon, C.W., Mills, J., and Keen, E.M. (1966). An inquiry in testing of information retrieval systems. (2 vols.). Cranfiled, U.K.: Aslib Cranfield Research Project, College of Aeronautics.
- Cooper, W.S. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19 (1), 30-41.
- Ding, W., Marchionini, G. (1996) A comparative study of web search service performance. *Proceedings of the Annual Conference for the American Society for Information Science*, 136-142.
- Dunlop, M.D. (1997). Time, Relevance and Interaction Modelling for Information Retrieval. *Proceedings of ACM/SIGIR '97*. 206-213.
- Ellis, D. (1996). The Dilemma of Measurement in Information Retrieval. *Journal of the American Society for Information Science*, 47(1), 23-36.
- Harter, S.P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1), 37-49.
- Hersh, W.R., Molnar, A. (1995). Towards New Measures of Information Retrieval Evaluation. *Proceedings of ACM/SIGIR '95*. 164-170.
- Hersh, W.R., Pentecost, J., Hickam, D. (1996). A Task-Oriented Approach to Information Retrieval Evaluation. *Journal of the American Society for Information Science*, 47(1), 50-56.
- Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*.
- Lake, M. (1997). *2nd Annual Search Engine Shoot-out: AltaVista, Excite, HotBot, and Infoseek square of f* [Online]. Available: <http://ww4.zdnet.com/pccomp/features/excl0997/sear/sear.html>
- Leighton, H.V. (1995). *Performance of four world wide web (WWW) index services: Infoseek, Lycos, WebCrawler, and WWWorm*. [On-line] Available: <http://www.winona.msus.edu/is-f/library-f/webind.htm>.
- Leighton, H. V., Srivastava, J. (1997) *Precision among world wide web search services (search engines): Alta Vista, Excite, Hotbot, infoseek, Lycos*. Unpublished master's thesis, Department of Computer Science, University of Minnesota.

- Meadow, C.T. (1992) *Text information retrieval systems*. Toronto: Academic Press.
- Meghabghab, D. B., & Meghabghab, G. V. (1996, May) Information retrieval in cyberspace. *Proceedings of American Society for Information Science Mid-Year Meeting*, 224-237.
- Mizzaro, S. (1997). Relevance: The Whole History. *Journal of the American Society for Information Science*, 48(9), 810-832.
- Overton, R. (1996). Search engines get faster and faster, but not always better [On-line]. *PC World*, 14. Available: http://www.pcworld.com/workstyles/online/articles/sep96/1409_engine.html
- Ralph, R.D. (1997). WWW Search Engines. Indexes, Directories and Libraries. [On-line] Available: <http://www.netstrider.com/search/index.html>
- Samalpais, Tait, J., Bloor, C. (1998). Evaluation of information seeking performance in hypermedia digital libraries. *Interacting with Computers*, 10. 269-284.
- Saracevic, T. (1995). Evaluation of Evaluation in Information Retrieval. *Proceedings of ACM/SIGIR '95*. 138-146.
- Saracevic, T. (1996). Relevance reconsidered. In P. Ingwersen & N. O. Pors, (Eds.), *Information science: Integration in perspectives*. Copenhagen: The Royal School of Librarianship. 201-218
- Schlichting, C. Nilsen, E. (1997). Signal detection analysis of www search engines. *Proceedings of the Second Human Factors on the Web Conference*. [On-line] Available: <http://www.microsoft.com/usability/webconf/schlichting/schlichting.htm>
- Scoville, R. (1996). Special report: Find it on the Net! [On-line]. *PC World*, 14(1). Available: <http://www.pcworld.com/re-prints/lycos.htm>
- Search Engine Watch. (1998). [On-line] Available: <http://searchenginewatch.com/>
- Su, L. T., Chen, H., and Dong, X. (1998). Evaluation of Web-based search engines from the end-user's perspective: a pilot study. *Proceedings of the Annual Conference for the American Society for Information Science*, 348-361.
- Tague-Sutcliffe, J.M. (1996). Some Perspectives on the Evaluation of Information Retrieval. *Journal of the American Society for Information Science*, 47(1), 1-3.
- Van Rijsbergen, C.J. (1979). *Information Retrieval*. London, England: Butterworths.
- Venditto, G. (1996). Search engine showdown. *Internet World*, 7(5), 79-86.

Appendix A. Queries used in search engine study

Query	Formulation in three languages
Query 1 – Find information on National Museums	English: +"national museum" German: +Nationalmuseum Polish: +"muzeum narodowe"
Query 2 – Find currency exchange rates	English: +"exchange rates" +currency German: +Wechselkurse +Währung Polish: +"kursy walut"
Query 3 – Find information related to the Year 2000 problem, but no apocalyptic visions	English: +"year 2000" +problem -apocalypse German: +"Jahr 2000" +Problem -Apokalypse Polish: +"rok 2000" +problem –apokalipsa
Query 4 – Find train schedules, but not training schedules	American English: +"train schedule" -training British English: +"train timetable" -training German: Zugfahrplan Polish: "rozklad jazdy pociagow"