PERGAMON

Information Processing and Management 35 (1999) 337–362

INFORMATION PROCESSING & MANAGEMENT

Discriminating meta-search: a framework for evaluation

Mark H. Chignell[a,b,*], Jacek Gwizdka[a,b], Richard C. Bodner[a,b]

[a]*Interactive Media Laboratory, Department of Mechanical and Industrial Engineering, University of Toronto, King's College Road, Toronto, Canada M5S 3G8*

[b]*Knowledge Media Design Institute, University of Toronto, King's College Road, Toronto, Canada M5S 3G8*

Abstract

There was a proliferation of electronic information sources and search engines in the 1990s. Many of these information sources became available through the ubiquitous interface of the Web browser. Diverse information sources became accessible to information professionals and casual end users alike. Much of the information was also hyperlinked, so that information could be explored by browsing as well as searching. While vast amounts of information were now just a few keystrokes and mouseclicks away, as the choices multiplied, so did the complexity of choosing where and how to look for the electronic information. Much of the complexity in information exploration at the turn of the twenty-first century arose because there was no common cataloguing and control system across the various electronic information sources. In addition, the many search engines available differed widely in terms of their domain coverage, query methods and efficiency.

Meta-search engines were developed to improve search performance by querying multiple search engines at once. In principle, meta-search engines could greatly simplify the search for electronic information by selecting a subset of first-level search engines and digital libraries to submit a query to based on the characteristics of the user, the query/topic, and the search strategy. This selection would be guided by diagnostic knowledge about which of the first-level search engines works best under what circumstances. Programmatic research is required to develop this diagnostic knowledge about first-level search engine performance.

This paper introduces an evaluative framework for this type of research and illustrates its use in two experiments. The experimental results obtained are used to characterize some properties of leading search engines (as of 1998). Significant interactions were observed between search engine and two other factors (time of day and Web domain). These findings supplement those of earlier studies, providing preliminary information about the complex relationship between search engine functionality and performance in different contexts. While the specific results obtained represent a time-dependent

* Corresponding author.

PII: S0306-4573(98)00065-X

snapshot of search engine performance in 1998, the evaluative framework proposed should be generally applicable in the future.

1. Introduction

Sometime early in the twenty-first century it is likely that the one billionth document or page will be added to the World-Wide Web. It is unlikely that there will be much fanfare for that event, since keeping track of the number of documents in a dynamic and constantly changing network like the Web is a monumental task, and the exact moment at which that landmark is reached will probably never be precisely determined. Along with the billion documents, there will be hundreds, if not thousands, of search engines, and a number of digital libraries and other prominent sites that will be particularly relevant to certain types of search.

The notion that billions of documents could be directly accessible by almost everyone in the industrialized world is astounding, and particularly so when examined against the historical context. Around two thousand years ago, the greatest library in the world was in Alexandria, with thousands of books. Around A.D. 1000, the greatest library in the world was in Cordova, with a catalogued library of 600,000 books (Derry & Williams, 1960, p. 29). By the twentieth century, collections of the great research libraries routinely numbered in the millions of books.

In the twenty-first century, the vast majority of information seems destined to be stored electronically, in two fundamentally different types of repository. On the one hand there will be digital libraries, containing electronic versions of millions of books, journals and manuscripts. These digital libraries may eventually be linked into a global digital library. On the other hand there will be a vast network of billions of documents of less certain authority and with little if any indexing.

The traditional library warehouses books in one or more physical locations. There is a well defined cataloguing system, including a controlled vocabulary of subject headings, and books are checked out in an orderly fashion. Considerable research is being carried out on how to enhance future digital libraries through development of features such as interoperability, multilingual indexing, and advanced knowledge representation (Fox & Marchionini, 1998). Thus the disparity between functionality inside and outside the digital library is likely to increase over time as research results on digital libraries get incorporated into practice. In contrast to the exciting research on digital libraries, the situation for the significant portion of the Web outside digital libraries is very different. There is no definitive catalogue, and most Web pages are indexed casually (with meta-tags), if at all. Thus from the perspective of library and information science, the many millions of ad hoc Web pages are a disorganized mess and are likely to remain so for some time to come. This stands in marked contrast to the emerging digital libraries, where principles of good cataloguing and indexing are preserved for large collections of electronic documents.

Digital libraries are islands of organization and structure in a chaotic sea of unorganized Web documents (Lynch, 1997) which continues to grow exponentially, with a huge migration of critical information of all sorts on to the Web (from company reports, to government documents and college lectures). Increasingly, people depend on the Web for the information

they need to perform their tasks. However, as more information gets on to the Web, it becomes more difficult to find the precise information one wants (the problem of finding needles in ever-expanding haystacks). Since it will be a long time, if ever, before the entire Web is organized into one or more digital libraries, there is an ongoing need for search engines to provide access to general Web documents through full-text retrieval, thereby overcoming (to the extent possible) the problem of poor or nonexistent indexing.

2. A framework for discriminating meta-search engines

As of 1998, search engines were a preferred means to find specific information on the Web, and were frequently bookmarked (Abrams, Baecker & Chignell, 1998). Some of the most popular sites on the Web were the sites that provided entry points to search engines (e.g. Yahoo and Infoseek). The Web was evolving into a two-tiered system of information with digital libraries at the high (structured) end, and relatively unorganized ad hoc information at the other. Research was needed to determine how search engines could be adapted for easier and more effective access to the diverse and vast collections information in the Web.

There were many search engines available in 1998, with the precise number depending on one's definition of what a search engine was. Search engines numbered in the hundreds (and possibly in the thousands if one includes specialized search engines with narrow coverage). Information about search engines on the Web was provided at special sites such as Search Engine Watch (1998).

As search engines proliferated on the Web, meta-search engines were developed. These meta-search engines did not maintain their own index, but instead queried other search engines. The results from multiple search engines were then collated into a composite set of nonoverlapping hits that was returned back to the user. These first generation search engines each had a standard set of basic search engines that they queried, and did not use different search engines depending on the type of the query. However, previous research (as discussed below) suggests that different search engines perform better for different types of query and in different

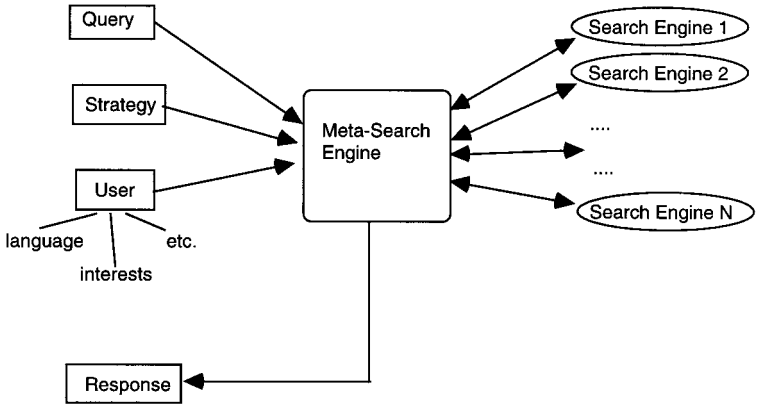

Fig. 1. Schematic overview of a discriminating meta-search engine.

situations. Thus there is an opportunity to develop more discriminating meta-search engines that search different collections of search engines according to the specifics of the query, the type of user, etc.

Fig. 1 illustrates the basic operation of a discriminating meta-search engine. Along with the query, the meta-search engine may receive information about general search strategy and user characteristics. For instance, knowing that the user is an academic, rather than business user might lead the meta-search engine to add search engines that have more emphasis on basic research in their coverage. Knowing the user's language preference might lead to hits being filtered according to language. Search strategy might include whether the user wants to get as many relevant documents as possible (i.e. a recall-oriented search) or whether he wants to get hits that are mostly relevant, even if there are fewer of them (a precision-oriented search). Topic is also a possible discriminator between search engines, and might be explicitly stated by the user or inferred in some way by the meta-search engine on the basis of the query.

On the basis of the foregoing information the meta-search engine would choose a subset of available search engines to submit the query to and then filter the resulting collection of hits into a response that is returned to the user. Note that much of this process is also carried out by first generation meta-search engines, such as meta-crawler (www.metacrawler.com). The difference is that early meta-search engines did not discriminate between the search engines they query based on the user's query and other aspects of the context (e.g. topic, search strategy, etc.) in which that query was submitted.

Discriminating meta-search engines should add value to Web-based search because they capitalize on the major differences that exist between first-level search engines. The operating characteristics of a search engine consist of a range of elements, including size of the indexed collection, coverage of each indexed site, frequency of updates of the indexed collection, indexing algorithm, search algorithm and types of supported queries. This means that different search engines will work better for different queries and query contexts, thus providing a basis for successful discrimination, providing that the knowledge about which search engines perform best under which circumstances can be acquired.

Thus the development of discriminating meta-search engines requires programmatic research on search engine performance in varying contexts. For example, which search engines work better in different Web domains? Which digital libraries are preferred sources for particular types of information? Which search engines will work better for broad versus narrow queries? Answering these questions will take concerted and extensive research effort. Our goal in this paper is to provide an evaluative framework for answering these questions, with an illustration of how this framework can be applied.

3. Related research

The fundamental organizational unit of the Web at its conception was the hypertext link (Berners-Lee, Cailliau, Luotonen, Nielsen & Secret, 1994). However, finding specific documents through a trail of hypertext links is not particularly efficient, and depends on whether or not an appropriate set of links has been included by the various Web page authors. The need to

search for specific documents, as well as browse around general topics, spawned a number of search engines and subsequent evaluation of how well these search engines worked.

Chu and Rosenthal (1996) evaluated the capabilities of Alta Vista, Excite and Lycos (Boolean logic, truncation, word and phrase search, etc.) and found that Alta Vista outperformed Excite and Lycos in both search facilities and retrieval performance. 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.

Leighton (1995) evaluated the performance of four index services: Infoseek, Lycos, Webcrawler and WWWWorm. Lycos and Infoseek had similar performance in terms of average top 10 precision scores. In terms of response time, Webcrawler had the shortest time followed by Infoseek, Lycos, while WWWWorm was a distant fourth. Leighton and Srivastava (1997) carried out a follow up study that compared precision scores for Alta Vista, Excite, Hotbot, Infoseek and Lycos. In their study, the first 25 results returned for 15 queries were examined. Leighton and Srivasta found that Alta Vista, Excite and Infoseek performed the best overall (the relative ranking of the search engines differed depending on how the authors 'interpreted the concept of 'relevant''). They also found that Lycos performed better for short, unstructured queries whereas Hotbot performed well on structured queries.

Ding and Marchionini (1996), in their study of three search engines (InfoSeek, Lycos and Open Text), found that different search engines tended to perform better for different queries, a finding that has also been noted informally in our laboratory. Ding and Marchionini also found a surprisingly low level of result overlap among the three search engines that they studied.

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 (Green & Swets, 1988) to analyze the data. They found that all of the search engines tested had poor performance. Signal detection analysis provided "an objective method of evaluating the effectiveness of extant and future technologies for resource discovery". In a recent study, Hou (1998) also found poor performance (low sensitivity) when evaluating search engine performance using signal detection analysis. He found that Alta Vista had the best sensitivity of the search engines that he tested, although its mean sensitivity (d′) was still only a comparatively low 0.42.

Other studies on search engines compared and contrasted features such as Boolean vs. natural language, number of documents indexed, etc. (see Liu, 1996; Notess, 1996; Stobart & Kerridge, 1996; Zorn, Emanoil, Marshall & Panek, 1996; Kenk, 1997). In addition to their empirical study of retrieval performance, Ding and Marchionini (1996) provided a descriptive comparison of selected search engines. Chu and Rosenthal (1996) also conducted both a descriptive comparison and empirical evaluation of retrieval performance (precision and response time). Their study also proposed a methodology for evaluating WWW search engines in terms of five aspects:

1. Composition of Web indexes — 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 — such as precision, recall and response time;
4. Output option — 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 — how difficult and effortful it is to use the search engine by typical users.

Thus, while limited in scope, previous search engine studies have found a number of differences between search engines. In addition, when assessed quantitatively (e.g. using signal detection analysis) the performance of search engines has generally been found to be poor.

Metasearch engines attempt to improve search results by utililizing multiple search engines to answer queries. A metasearch engine sends queries to remote search engines and collates the results for presentation to the user via a Web browser. As noted by Dreilinger and Howe (1997), many metasearch engines have been developed since 1994, including GlOSS (Glossary of Servers Server), the Harvest system, Discover (which queried over 500 WAIS sites) and MetaCrawler. Metacrawler took the metasearch paradigm a step further by ranking the relevance of the documents returned based on textual analysis of the HTML source of the referenced documents (Selberg & Etzioni, 1995).

Metasearch engines were developed to improve search performance. Selberg and Etzioni (1995, cited by Dreilinger & Howe, 1997) suggested that no single search engine is likely to return more than 45% of the relevant results, based on empirical findings. Thus meta-search engines were viewed as solving a problem of low recall (insufficient numbers of relevant pages or articles being retrieved). Metasearch engines can also provide a uniform interface to the user, even though the user interfaces of the first-level search engines that they query may vary widely.

While there have been few direct comparisons of metasearch and search engine performance, at least one such study has found that a metasearch engine returned the highest number of links judged relevant by the subjects (Gauch, Wang & Gomez, 1996). It seems likely that metasearch performance could be further improved by applying more knowledge to the planning process that the metasearch process uses in forming and scheduling queries to remote search engines. Dreilinger and Howe (1997, p. 203) characterized this planning process as follows:

> ...information gathering on the Web can be viewed as a simple planning problem in which the goal is to find sites satisfying specific criteria and where the actions are queries to search engines. Search plans are constrained by the resources available: how much time should be allocated to the query and how much of the Internet's resources should be consumed by it.

Dreilinger and Howe went on to describe an adaptive learning process for metasearch engines based on the development and use of a metaindex. In their approach, the search engines that could be used to answer a specific query are ranked based on the corresponding information in the metaindex and on recent data on search performance. The metaindex is developed by tracking "the effectiveness of each search engine in responding to previous queries" (Dreilinger & Howe, 1997, p. 204).

Further development of metasearch engines requires a knowledge base for how to select search engines and information sources to use in particular contexts (as shown in Fig. 1). This knowledge base may be instantiated in a number of ways, including as a metaindex (Dreilinger & Howe, 1997). Development of this knowledge base requires an evaluative framework for assessing how well different search engines and information sources do in response to different queries.

Central to the required evaluative framework is a set of measures that can characterize the effectiveness of a search engine for a given query and information source (or set of sources). Experience with research in information retrieval has shown that it can be extremely difficult to find suitable evaluative measures. For instance, while precision and recall are the most widely used measures, they tend to trade off against each other, and they are also extremely sensitive to how 'relevance' is defined and measured, which is itself a problematic issue (Harter, 1996). This has led to proposals for measures of effectiveness that combine recall and precision in various ways (e.g. Meadow, 1992). However, in spite of their deficiencies, recall and precision continued to be used widely.

The widespread use of relevance ranking (Salton, 1989) provides the opportunity for measures of relevance that incorporate assessments of the quality of ranking. In a large set of ranked documents one can assess whether or not there is a higher proportion of relevant documents in sets of documents that appear earlier (i.e. are ranked higher) in the ranked list. This has led to the development of measures of average precision for different numbers of documents in a ranked list that have been used in the TREC evaluations of information retrieval algorithms (e.g. Harman, 1995).

There are numerous measures of search effectiveness that could be defined, including time taken to find the first relevant document. For search engines with ranked output, variations on the measure of average precision seem particularly promising. Care must be taken, however, to test and calibrate the measures, so that they are proven to work as intended.

Su, Chen and Dong (1998) 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, a 'perfect' search engine that had almost all highly relevant (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.

A more suitable approach in this case would be to simply take the difference in number of documents that were relevant in different sections of the ranked list. For instance, one might compare the number of relevant documents in the first 10 versus second 10 (11–20) ranked documents. If there were more relevant documents in the first 10 hits, this would suggest that

the ranking process was working. A simple test of significance would then involve applying a binomial test where the assumed probability that any document in the first 20 returned was relevant would be the proportion of relevant documents in the first 20 hits (e.g. if there were 6 relevant documents then the estimated probability of a relevant document for each position in the ranked list of top 20 documents would be 0.30. One would then use the binomial test (e.g. Siegel & Castellan, 1988), to test the probability that the observed number of relevant documents in the first 10 positions in the ranked hit list could have occurred (assuming a coin-tossing like process where the probability of each document 'turning up' relevant was 0.30).

More generally, one can test the effectiveness of relevance ranking output using linear regression. In contrast to a binomial test which focuses on two regions of documents (e.g. 1–10 vs. 11–20) within the ranked output, a regression analysis can assess the whole of the ranked list (or at least as much of it as one can get the corresponding human judgments of relevance for). For instance, one can plot the number of relevant documents for each group of 20 hits, vs. an indexing variable that reflects the position of those hits in the ranked sequence (e.g. '1' for the first 20 hits, '2' for the second 20 hits, and so on). Successful ranking would then be indicated by a significant linear effect with a negative slope. The strength of the association in this case would be indicated by the size of the correlation.

The point of this discussion is not to make specific proposals concerning what evaluative measures should be used, but rather to show that while many different measures can be defined, one must use them appropriately. In the second experiment reported below, we provide some initial results of how well a number of measures work in practice. 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. 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.

4. Overview of experiments

Poor search performance could be due to poor query formulation (how the topic is expressed in terms of the query actually input to the search engine), poor indexing of documents (attributable either to the document author or to the search engine's indexing process), or to problems in evaluation of document relevance (e.g. a human judge may be inconsistent, or may interpret the topic differently from the topic actually implied in the text of the query as actually submitted to the search engine).

Since previous experimental studies of search engine performance have tended to use human subjects to form and evaluate queries and the resulting retrieval sets, we decided to focus on efficient evaluation methodologies that assess search engine properties with minimal use of human intervention. While we recognize that a complete understanding of search engines requires an understanding of the human role in formulating queries and evaluating results, it is also possible to gain important insights into search engine performance (relevant to the design of discriminating meta-search engines) with only limited use of human subjects in assessing relevance.

Two experiments were carried out to study the relative effectiveness of different search engines under different conditions. In the first experiment, the effect of time of day and day of week, and the effect of query strategy (general, high precision, high recall) on query processing time was examined for each of three search engines (Excite, Hotbot and Infoseek).

The second experiment used three search engines, six Internet domains, and four queries (replicates) in a fully factorial design for a total of 72 observations. Experiment 2 also used a variety of different performance measures to assess query outcomes for each of the observations collected. These will be described in detail in a later section of this paper. In contrast to experiment 1 (which used consensus peer review based on six different search engines), relevance in experiment 2 was assessed by a human judge who was fluent in English, German and Polish.

5. Experiment 1

Experiment 1 used a set of nine prespecified queries. The queries (listed below) were created based on papers written by graduate students on the topic of mobile computing. The queries were divided into categories. The three categories (types) of query used were general, higher precision, and higher recall. General query types were formulated to approximate a query that a 'typical' user (with little background knowledge on the topic) would submit to a search engine. The high recall query types were intended to represent a user with little knowledge on the topic area, who is trying to find out more information using a broad query. The last type, high precision, represents a user with a more specific query.

1. General queries
   - research 'mobile computing'
   - mobile computing

2. High precision queries
   - social implications of mobile computing on military applications
   - mobile computing quality engineering in a manufacturing facility
   - Internet-based mobile computing
   - input output methods for mobile clients computers
   - mobile computing technology office automation package tracking

3. High recall queries
   - mobile computing with PDA 'personal digital assistants' handheld palmtops
   - mobile computing software hardware technology equipment devices

These queries were then submitted to three search engines (Excite, Hotbot and Infoseek). Document relevance was assessed using a 'consensus peer review' procedure where the binary judgment of relevance (yes or no) was obtained from the results of six different (referee) search engines (Alta Vista, Lycos, Northern Light, Search.com, Web Crawler, Yahoo) to the same

query. A hit from one of the search engines in the experiment was deemed to be relevant if it was also returned by at least one of the six referee search engines in response to the query.

The nine queries were used for each combination of the three search engines, on each of four days of the week (Monday through Thursday), and at two different times of day (10 am and 9 pm), making for a total of 216 data points ($9 \times 3 \times 4 \times 2 = 216$) in a fully factorial design. In addition to query processing time and precision, the number of broken and duplicate URLs for each search was also assessed. Experiment 1 was completely automated, with computer programs being used to launch the queries and evaluate the results.

A 'run' consisted of all the search engines being queried in alphabetic sequence for all nine queries. After a run was finished, tests for duplicate and broken URLs were performed. Each run, including validation, took approximately 1.5 to 2 h to complete. After a run, the precision scores were computed.

6. Results and discussion

The data were analyzed using multivariate analysis of variance (MANOVA). There was a significant multivariate effect[1] for the two-way interaction of Query Type and Search Engines ($F(20, 604.6) = 10.6$, $p < 0.001$). This effect was due to a significant univariate interaction for precision ($F(4, 186) = 6.9$, $p < 0.001$). The univariate interactions for the other three dependent variables were not significant. Fig. 2 shows a plot of the interaction where the error bars represent 95% confidence intervals. From the figure, it can be seen that Infoseek has significantly better precision performance for general and high precision queries than both Excite and Hotbot. For high recall types of query, Infoseek and Excite had a comparable level of precision, which was higher than that obtained by Hotbot.

There was a significant main effect of search engine query time ($F(2, 186) = 65.5$, $p < 0.001$). Post hoc Tukey testing indicated a significant difference between the query processing times of Excite and Infoseek ($p < 0.001$), and Hotbot and Infoseek ($p < 0.001$), with query processing being fastest for Infoseek (with a mean time of 31.5 seconds, versus 43 and 45 s, respectively, for Hotbot and Excite).

There was a significant main effect of search engine on the number of broken URLs ($F(2, 186) = 12.3$, $p < 0.001$), with significant differences occurring between Excite and Hotbot ($p = 0.001$), and Hotbot and Infoseek ($p < 0.001$). Infoseek had an average 3% broken URLs per result set versus corresponding figures of 4 and 6% respectively, for Excite and Hotbot.

There was also a significant main effect of search engine on the number of duplicate URLs ($F(2, 186) = 65.5$, $p < 0.001$). The Tukey test revealed an ordering of the search engines as Excite, Infoseek, and Hotbot, with significant differences between Excite and Hotbot ($p < 0.001$), and between Hotbot and Infoseek ($p < 0.001$). Hotbot had an average of 8% duplicate URLs per result set, while Excite at the other extreme had only 1% duplicates per result set.

There was a significant main effect of search engine on precision ($F(2, 186) = 182.7$, $p <$

[1] All multivariate effects in this paper were assessed using Wilk's λ with a significance level of 0.05. $p < 0.1$ was considered to be borderline significant.

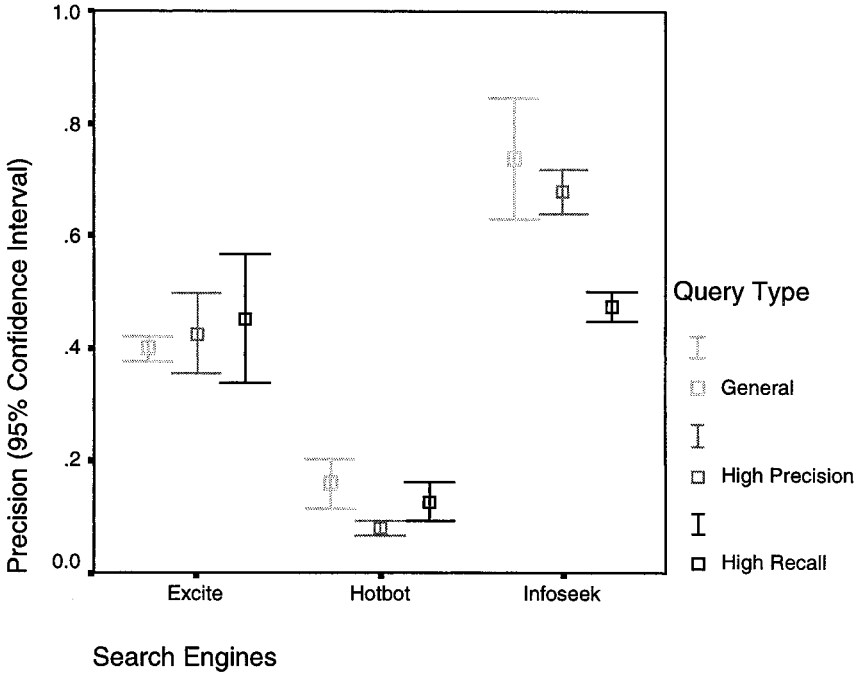

Fig. 2. The search engine by query type interaction on precision.

0.001). Tukey testing revealed an ordering of the search engines as Infoseek (mean precision score of 65%), Excite (42%) and Hotbot (11%).

Although it seems from the data that Infoseek has exceptionally good precision scores, while Hotbot has poor precision, this may be due to high overlap between Infoseek and the six referee search engines, rather than a tendency for Infoseek to return a higher proportion of relevant documents in general. It might even be argued that this result suggests that Hotbot will be more useful in meta-searches, since it tends to generate a high proportion of unique hits that are not found by other search engines.

There was only one significant main effect for time of day in experiment 1, and that was with query processing time ($F(1, 186) = 29.9$, $p < 0.001$). Evening queries were processed more quickly (mean = 33 s) and the times were less variable (Std. Dev. = 12.6 s) than daytime queries (mean = 46.7 s, Std. Dev. = 27.9 s).

There was a significant main effect of query type on the precision scores ($F(2, 186) = 3.7$, $p = 0.026$). General queries performed best over the three search engines followed by high precision and high recall. It would seem that the search engines tested here work best (in terms of precision) for simple or specific inquiries, as opposed to broader queries.

Up to this point, only the effects of the independent variables on a single dependent variable have been investigated. A *user-oriented* composite measure of performance may be defined by calculating a single measure based on the four dependent variables. Since post hoc Tukey tests were performed, the ranking of the three search engines for each dependent variable is known. Using this ranking information, a simple formula can be devised. The number of first, second

Excite: 3*(1/4) + 2*(2/4) + (1/4) = 2/3 = 66.7%

Hotbot: 3*(0/4) + 2*(1/4) + (3/4) = 1.25/3 = 41.7%

Infoseek: 3*(3/4) + 2*(1/4) + (0/4) = 2.75/3 = 91.7%

Fig. 3. Search engine composite performance measurements.

and third place rankings are added up for each search engine (e.g. for each place, the search engine will have a score out of 4). The first place results are multiplied by a coefficient of 3, the second place by a coefficient of 2, and third place is multiplied by 1. Therefore, the maximum composite performance score a search engine could receive is three (e.g. $3^*(4/4) = 3$ — first place rankings for all dependent variables). The composite performance measures for the search engines are listed in Fig. 3. The reader should note that this composite measure has two limitations. First, it does not take into consideration the case where no statistical significant difference is found between two search engines. Second, it treats all dependent variables as being equal. Some users may favor high precision scores over all other dependent variables. For such cases, different weights should be assigned to the dependent variables to reflect such preferences.

7. Experiment 2

Experiment 2 examined the effect of geographical coverage and Internet domains on search engine performance. It used AltaVista, HotBot and Infoseek as search engines. The four Web country codes used were: Germany (.de), Austria (.at), Poland (.pl) and the UK (.uk). Additionally, the general 'commercial' sites (.com), and 'organizational' sites (.org) were tested.

One query was formulated for each of four topics in experiment 2. The queries were expressed in three different languages, appropriate to the country in which a given Internet domain was located (including slight differences between American and British English). The four queries used in experiment 2 are shown in Table 1.

In addition to assessing the effect of different information sources (different Internet domains), experiment 2 also used a range of different performance measures. Most of the measures were modifications of precision. Because of the large possible number of hits returned by search engines the precision of the first 20 web pages was considered.

In contrast to experiment 1 where an automated method of relevance judgment was used based on referee search engines, human relevance judgments were used in experiment 2. The relevance judgments were based on an a priori defined set of rules as described in Table 2.

Duplicate hits, that could be easily recognized[2] by search engines, were also marked. In addition hits returned by each search engine were analyzed and compared with respect to their

[2] The same URLs present an obvious case. Other very common cases are ⟨*URL*⟩⟨*path*⟩/ and ⟨*URL*⟩⟨*path*⟩/index.html or ⟨*URL*⟩⟨*path*⟩/default.htm

Table 1
Queries used in experiment 2

Query	Formulation in three languages
Query 1 — find information on national museums	English: + 'national museum'; German: + Nationalmuseum; Polish: + 'museum 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−apocalipsa
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'

uniqueness. A unique hit (UNIQUE) was defined as one that was not reported by any of the other search engines.

A broad array of dependent variables was used to measure performance. For the full description of these measures the reader is referred to Gwizdka (1998). Measures that showed significant differences between the search engines in experiment 2 are described below.

7.1. First 20 precision

1. 'Full' precision (PRECFULL)This measure takes fully into account the subjective score assigned to each hit. Eq. (1) shows how full precision was calculated.

$$\text{precFull} = \frac{\sum_{i=1}^{\text{mF20Hitsscore}_i}}{\text{mF20Hits} * \text{maxHitScore}}, \tag{1}$$

   where: score_i is the score assigned to the ith hit, mF20Hits = min(20, hitsReturned), hitsReturned — total number of hits returned, maxHitScore — max score that can be assigned to one hit (3), PRECFULL is defined for hitsReturned > 0
2. 'Best' precision (PRECBEST)This measure takes into account only the most relevant hits (hits, that obtained score = 3).

$$\text{precBest}(1, \text{mF20Hits}) = \left. \text{count\_of}_{i=1}^{\text{mF20Hits}(\text{score}_i} = 3) \right/ \text{mF20Hits}. \tag{2}$$

Table 2
Description of subjective relevancy scores

Relevancy 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

These two precision measures were based, in part, on methodology employed by Ding and Marchionini (1996) in their study.

7.2. Difference between first 10 and second 10 precision (DPOBJ)

Differential precision measures position of relevant hits within the 20 first returned hits. Higher concentration of relevant hits in the first 10 hits than in the second 10 hits is desirable for users, since it allows them to find relevant information faster. The differential measure was based on objective calculation of precision. The measure was based on mechanically locating the presence or absence of required terms, and on a distinction between good and bad links. Differential objective precision between the first 10 and the second 10 hits was calculated as follows:

$$\text{dpObj}(1, \text{mF20Hits}) = \text{precObj}(1, \text{mF10Hits}) - \text{precObj}(\text{mF10hits}, \text{mF20Hits.}) \quad (3)$$

Differential precision has the following properties:

- dpObj > 0 = > more relevant documents in the first 10 hits than in the second 10
- dpObj = 0 = > number of relevant documents in the first 10 hits and in the second 10 is the same
- dpObj < 0 = > less relevant documents in the first 10 hits than in the second 10

7.3. Search length i (FSLENi)

Another measure used was expected search length, first suggested by Cooper, and described in detail by Van Rijsbergen (1979). It measures how many, on the average, nonrelevant documents need to be examined by users in order to find a given number of relevant ones. This study used a modification of this measure which takes into account all documents that need to be examined (relevant and nonrelevant ones) and, additionally, one level of links from returned web pages to relevant documents.

The measure is based on the number of web pages that need to be examined by a user before finding the i most relevant pages. With this measure, the most relevant web pages either have a relevance score of 3 or else are pages with a relevance score of 2 and that contain one or more links to a page, or to pages, with relevance score of 3. All pages that need to be examined until the 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 link[3]).

$$\mathrm{fSLen}_i = 1 - \frac{\mathrm{maxSLen}_i - \mathrm{sLen}_i}{\mathrm{maxSLen}_i - \mathrm{bestSLen}_i}, \tag{4}$$

where: $\mathrm{maxSLen}_i$ is the maximum search length for i relevant web pages within n returned search hits; $\mathrm{bestSLen}_i$ 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 3.

7.4. Hits and Hit ratio (HITS, HITRATIO)

The total number of hits returned as a result of a query was noted. Hit ratio was calculated as the ratio of the total number of hits returned as a result of a query to the total number of hits returned by a given search engine in a given domain. The following abbreviations were used to denote the above measures: HITS, HITRATIO.

8. 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; $F(17, 50) = 1.72$, $p = 0.069$).

8.1. Overlap of results

How much improvement can one expect to get by employing several search engines rather than just one? 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

[3] To simplify the calculations, only one level of links was examined.

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.

8.2. 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 3. Table 3 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 experiment 2.

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 3). 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 no such differences among ratios of HITS to TotalHits (HITRATIO). However, Infoseek returned unexpectedly few hits

Table 3
Coverage of the search engines across the Internet domains

Search engine	Domain	Means of hits returned as a result of a query	Total number of hits for each domain
	de	952.00	5,796,668
Alta Vista			
	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
	de	775.50	4,647,297
HotBot			
	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
	de	6.75	2,141,013
Infoseek			
	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

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.

8.3. 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 UK (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 reindexed less often[4].

8.4. Search length

A borderline significant interaction was found on search length 1. Infoseek performed worst on this measure in the 'German'[5] domains (de, at), while HotBot had the worst performance in 'Poland' (pl), as shown in Fig. 4. 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.

8.5. Analysis of main effects

Full factorial multivariate analysis was carried out using search engines and domains as the independent factors, with the 14 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 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 precisions (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* ($F(2, 50) = 3.85$, $p = 0.028$).

[4] Effects of the less often reindexing 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).

[5] Note that the domains represent 'virtual' countries.

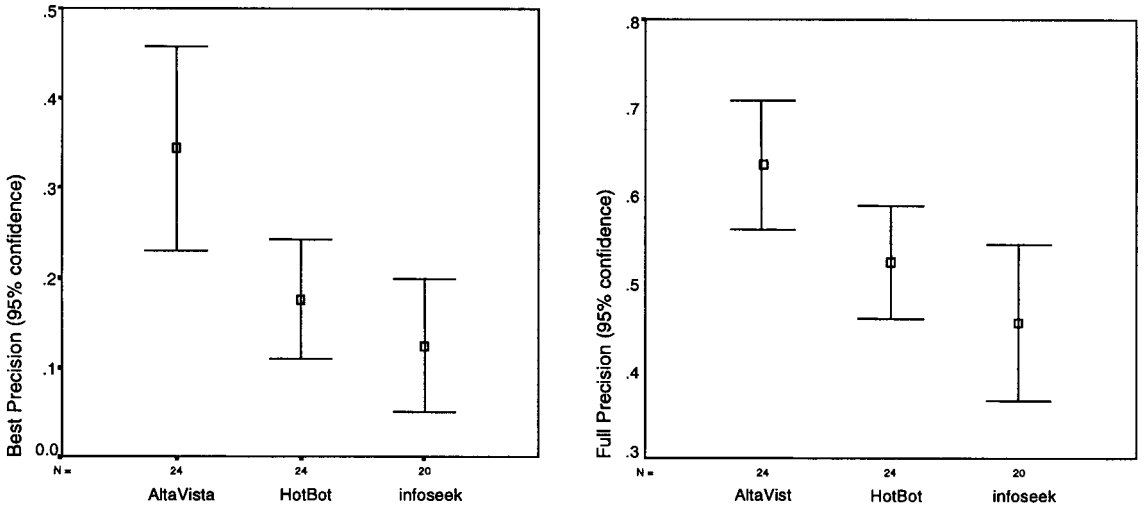

Fig. 4. *Best* and *full* precisions of the three search engines.

Fig. 4 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, 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.

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 Fig. 5), with few pages needing to be read prior to finding the first relevant document[6].

Differential precision reflects how well the first 20 hits have been ranked. Fig. 6 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 10 returned hits (DPOBJ > 0).

However, Infoseek often returned less than the examined twenty hits (in 13 out of 24 cases) and sometimes even less than 10 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.

[6] 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.

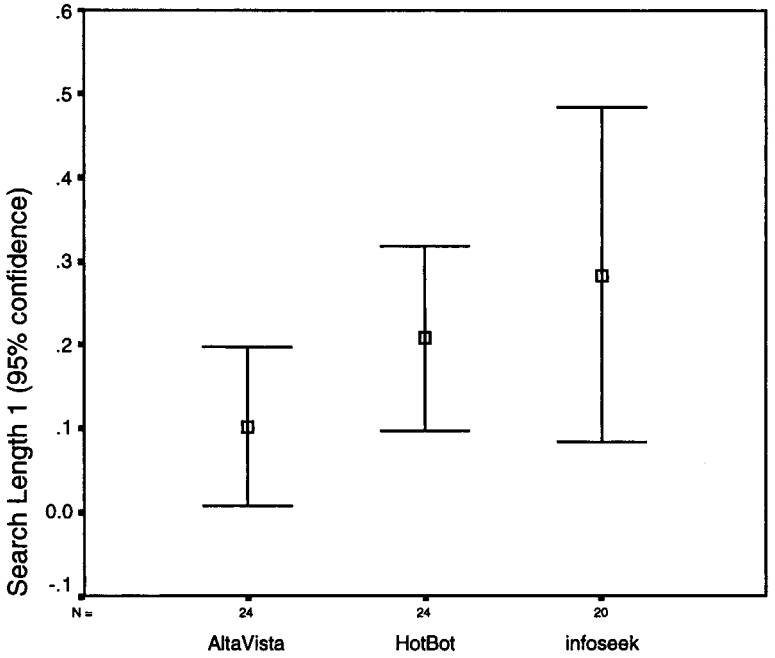

Fig. 5. Search length 1 for the three search engines.

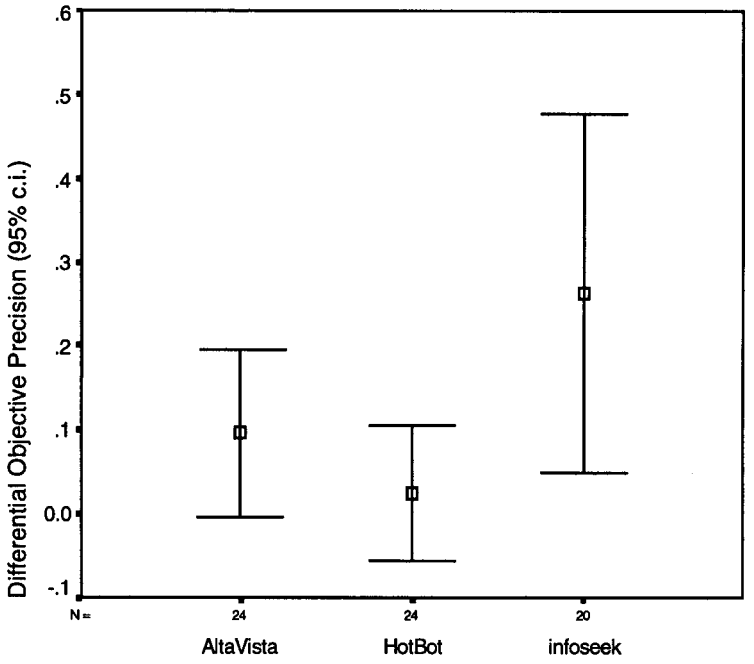

Fig. 6. Differential objective precision of the search engines.

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 with 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.

9. Discussion

Comparison of results across the two experiments clearly demonstrates how the way in which relevance is assessed, and how performance is measured, have a major impact on comparative search engine performance. In experiment 1, with relevance assessment based on peer consensus review by other search engines, and with queries on the general topic of mobile computing, Infoseek appeared to perform well with mean precision of 65%. However, when relevance ranking by a human subject, and different query topics were used in experiment 2 Infoseek performed relatively poorly on precision.

The results obtained in experiment 2 showed surprisingly little overlap in the documents returned by different search engines, confirming an observation made earlier by Ding and Marchionini (1996). This general lack of agreement between search engines calls into question the use of peer consensus review in assessing relevance. The lack of overlap also implies a useful role for discriminating meta-search engines to play in picking out the best hits from different search engines based on knowledge of how each search engine performs in different conditions.

The overlap with the referee panel of search engines in experiment 1 was smaller for Hotbot and AltaVista than for Infoseek. In the 'peer-review' method, a search engine that returns more unique documents than others receives a lower precision score. This could explain why in experiment 1 Hotbot obtained low precision scores and Infoseek obtained high precision scores, in contrast with experiment 2 (where relevance was assessed by a human judge), where Infoseek scored lowest of three search engines in terms of precision.

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. Excite uses a form of Latent Semantic Indexing (Steinberg, 1996) to index and match documents, which should lead to different results from that obtained by a search engine using Boolean matching and indexing based on raw text.

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 uses a unique method for ranking retrieved documents it will tend to put different documents at the top of its list. For example, words found in the title are more

important (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 experiment 2. However, Alta Vista and Excite 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 UK may be due to a lower frequency of reindexing 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.

9.1. Recommendations for meta-search engine design

There is considerable anecdotal evidence that people are dissatisfied with search engines (both in terms of their precision, and their ease of use). The available experimental evidence indicates that no one first-level search engine is particularly good in absolute terms, and that different search engines tend to perform better in different situations. The problems with individual search engines represent an opportunity for meta-search engines to recruit multiple search engines in carrying out queries and thereby improve performance over what any individual search engine can achieve. A good meta-search engine could also provide a consistent and easy to use interface, hiding the difficulties of dealing with idiosyncratic and complex first level search engine interfaces from the user.

The previous research literature, and the findings obtained in this study, provide a basis for a number of recommendations concerning development of knowledge bases for discriminating meta-search. In future, these knowledge bases should allow meta-search engines to choose more intelligently which first level search engines to submit queries to in particular contexts.

9.2. Recommendations from analysis of the research literature

Alta Vista has been shown to perform comparatively well in a number of studies and is an obvious choice (as of this writing) as a panel member for evaluating the performance of other search engines. Other referee search engines should be rotated according to the topic and the type of relevance being assessed, since these factors have been shown to have a major impact on the performance of different search engines. The low overlap between results sets for different search engines suggests that peer consensus review should be used conservatively, and only with referees that are known to perform reasonably well for the types of topic being used. In the language of signal detection theory, peer reviewers should have sufficient sensitivity (d') to the difference between relevant and nonrelevant documents to make their judgments meaningful.

9.3. Recommendations from experiment 1

An important factor in choosing search engines (either as panel referees or for meta-search) is whether or not the query is general, precise, or recall-oriented. For instance, while Infoseek

seemed to have an advantage over Excite (in terms of peer consensus reviewed precision) for general and high precision queries in experiment 1, it had no such advantage with high recall queries.

Another important factor for meta-search is the general quality of hits returned by search engines. In experiment 1, Hotbot had a relatively high proportion of broken links (6%) and duplicate URLs (8%). These problems could be filtered out by the meta-search engine, but they may also be indicative of problems in the coverage or indexing policy of the search engine.

Searches can be conducted more quickly (which may be important in large experimental studies) at particular times of day. While this may fluctuate over time and location, 9 pm was found to lead to faster search processing than 9 am, in the first experiment of this study. There was no evidence of an interaction between time of day or week and the relative speed of the different search engines. Infoseek returned results significantly faster than Hotbot and Excite at all times of day and week used in experiment 1. However, since search engines differ in their speed, meta-search engines should take this into account when users are in a hurry.

9.4. Recommendations from experiment 2

Infoseek has relatively poor coverage outside the .com domain and should probably not be relied upon in domains where English is 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. Meta-search engines should take account of the aberrant performance of certain search engines in certain domains. For instance, Hotbot has a disproportionately large number of bad hits for the .uk domain. In general, Hotbot like Infoseek should be used with caution when nonenglish language pages are of interest. Given the present results, it seems wise to be skeptical of mainstream search engine performance outside the .com and .org domains. The possible exception to this is Alta Vista, which appears to be much less language and domain sensitive.

As in previous studies, experiment 2 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 suggests a general problem with search engines (unreliability, and insensitivity to document relevance) which should be addressed. Discriminating meta-search can improve the overall search experience for the user, but it needs a basic level of performance from search engines in order to work effectively.

Meta-search should also enhance the presentation of search results. For instance, Ding and Marchionini (1996) pointed out the disadvantage of grouping the results from the same Web site versus scattering these results. Meta-search results from the same Web site can be scattered, even if originally grouped by the search engine they were retrieved from.

Meta-search also offers an opportunity to experiment with new types of user interface that integrate the results from various search engines. For instance visualization tools should be useful in providing overviews of large results sets. Meta-search engines could take advantage of recent developments in document clustering (e.g. Hearst & Pedersen, 1996) and in information visualization (Card, Mackinlay & Shneiderman, 1998).

Meta-search engines should also benefit from advances in the design of user interfaces and interaction dialogs. One extensive effort in this regard is the interactive track at TREC, the Text Retrieval Conference sponsored by the US National Institutes of Standards and Technology (e.g. Over, 1997). Other relevant research is work on the assessment of user experience and strategies and how they affect use of search engines (e.g. Golovchinsky, Chignell & Charoenkitkarn, 1997). Research findings from these related areas should also be incorporated into an evolving design for discriminating meta-search.

10. Conclusions

As more information becomes available in digital libraries, and the Web generally, search engines will grow in popularity. This is due to users' reliance on search engines to structure and sift through information. Since there is low overlap between the results sets returned by different search engines, there is an opportunity for discriminating meta-search engines to add significant value to the task of searching for information on the Web.

Development of discriminating metasearch engines will require qualitative and quantitative evaluative studies (cf. Ding & Marchionini, 1996) to determine what search engines should be used and when. Ideally, the best practice for this type of metasearch will be incorporated into a new generation of metasearch engines that choose which basic (first-level) search engine, or combination of search engines to use, based on the characteristics and context of the current search.

Principles for improved metasearch are likely to come from analyses of why existing search engines sometimes yield poor results. These analyses should tease apart a number of possible explanatory factors. Poor search performance could be due to poor query formulation (how the topic is expressed in terms of the query actually input to the search engine), poor indexing of documents (attributable either to the document author or to the search engine's indexing process), or problems in evaluation of document relevance. Evaluation problems, in turn, may arise from inconsistencies in human judgement, inappropriate use of automated techniques such as consensus peer review, or differences between the topic implied in the initial question versus the query actually submitted to the search engine. Further complexity arises because of the metric properties of evaluative measures, and the fact that measures that work well in one context may be inappropriate or misleading in a different context.

As demonstrated in experiment 2 of this study, poor search performance may also be due to an inappropriate matching of search engine to domain. Given this finding, it seems that digital libraries should be extremely careful in choosing search engines to provide access to their collections. Although the experiments reported in this paper were carried out on the Web in general, it seems likely that similar interactions between search engines, domains, and the impact of different evaluative measures will also apply to search within and across digital libraries.

Due to the need for more extensive studies of search engine, more automated methods for evaluation, would be helpful, but only if they can be shown to be valid. Experiment 1 illustrates one possible method for automated evaluation of a search engine's performance — a consensus peer review. The peer review did not require a human to judge the relevance of the

search engine results. While consensus peer review is very efficient, it produced questionable results in this study. Nevertheless, automatic evaluation of relevance may still be useful if more appropriate measures can be defined in future. One suggestion would be to use peer consensus review that requires more overlap between referee search engines and that uses referees that are judged to have good coverage and search performance in the topic areas being used. For instance, if 10 search engines were used, a more stringent consensus measure might require that a hit be returned by at least five of the referees before it was judged to be relevant for a query.

Peer consensus review might also be enhanced by explicitly considering the relevance rankings assigned to hits by the referee search engines. In this approach, the correlation between the rankings (presentation order) assigned to hits by the different referees would provide a more sensitive measure of relevance. Controlled studies where searches are seeded with a core set of documents that are known to be relevant may also be useful in calibrating and refining various automated techniques for relevance assessment in searches.

Based on the results of this study, and of previous studies, it is clear that there are major differences between how different search engines perform in different context, and that these should be exploited in the design of improved metasearch engines. Relevant characteristics to drive discriminating meta-search should include: the type of topic and query; the search criteria (e.g. precision-oriented vs. recall-oriented; the domains and/or countries of interest. Further studies are needed to develop a stronger foundation for advanced meta-search, both in terms of deriving the basic facts on how well different search engines work in different circumstances and in terms of developing improved and more efficient methodologies for comparing search engines and for assessing relevance in the context of Web search. These studies should be performed over time in order to establish trends for different search engines and to ensure that the rules used to map search engines to contexts are reliable.

Acknowledgements

The authors would like to thank Catherine Courage, Herman Colquhoun, Wayne Ho, Ming Hou and other members of the MIE 1402 graduate class at the University of Toronto for sharing their opinions on search engines and search engine performance. We would also like to thank the reviewers for their helpful comments. Support for this research was provided by an NSERC operating grant to the first author, and by a grant from Communications and Information Technology Ontario (CITO).

References

Abrams, D., Baecker, R. M., & Chignell, M. H. (1998). Information archiving with bookmarks: personal web space construction and organization. In: *Proceedings of the Human Factors in Computing Systems Conference* (pp. 41–48).

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H., & Secret, A. (1994). The World-Wide Web. *Communications of the ACM*, *37*(8), 76–82.

Card, S. K., Mackinlay, J., & Shneiderman, B. (1998). *Readings in information visualization*. Los Altos, CA: Morgan Kaufman.

Chu, H., & Rosenthal, M. (1996). Search engines for the world wide web: a comparative study and evaluation methodology. In: *Proceedings of the Annual Conference for the American Society for Information Science* (pp. 127–135).

Derry, T. K., & Williams, T. I. (1960). *A short history of technology from the earliest times to A.D. 1900*. New York: Dover Publications, Inc.

Ding, W., & Marchionini, G. (1996). A comparative study of web search service performance. In: *Proceedings of the Annual Conference for the American Society for Information Science* (pp. 136–142).

Dreilinger, D., & Howe, A. E. (1997). Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems*, *15*(3), 195–222.

Fox, E., Marchionini, G. (1998). Special issue on digital libraries. *Communications of the ACM*.

Gauch, S., Wang, G., & Gomez, M. (1996). Profusion: intelligent fusion from multiple different search engines. *Journal of Universal Computer Science*, *2*(9).

Golovchinsky, G., Chignell, M. H., & Charoenkitkarn, N. (1997). Formal experiments in casual attire: case studies in information exploration. *New Review of Hypermedia*, *3*, 123–157.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. New York: Wiley.

Gwizdka, J. (1998). Measuring information retrieval in hypermedia systems. Unpublished manuscript. Toronto, Ont., Canada: Department of Mechanical and Industrial Engineering, University of Toronto.

Harman, D. (1995). Overview of the third text retrieval conference (TREC-3). In D. K. Harman, *Proceedings of the Third Text Retrieval Conference (TREC-3)* (pp. 1–19). Gaithersburg, MD: National Institute of Standards and Technology.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, *47*, 37–49.

Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proceedings of ACM/SIGIR '96*.

Hou, M. (1998). Comparison for three internet search tools (Yahoo, Alta Vista, Lycos). Unpublished manuscript. Department of Mechanical and Industrial Engineering, University of Toronto.

Kenk, F. (1997). *Benchmarking www search services: a semantic portfolio analysis of popular Internet www search services* (On-line) Available: http://home.t-online.de/home/gerhard.kenk/hpgka00e.htm.

Leighton, H. V. (1995). *Performance of four world wide web (WWW) index services: Infoseek, Lycos, WebCrawler, and WWWWorm* (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.

Liu, J. (1996). *Understanding www search engines* (On-line) Available: http://www.indiana.edu/~librcsd/search.

Lynch, C. (1997). Searching the Internet. *Scientific American*, 52–56.

Meadow, C. T. (1992). *Text information retrieval systems*. Toronto: Academic Press.

Meghabghab, D. B., & Meghabghab, G. V. (1996). Information retrieval in cyberspace. In *Proceedings of ASIS Mid-Year Meeting* (pp. 224–237).

Notess, G. R. (1996). Internet 'onesearch' with the mega search engines. *Online*, *20*(6).

Over, P. (1997). TREC-5 interactive track report. In *Proceedings of the Fifth Text Retrieval Conference (TREC-5)* (pp. 29–56) NIST Special Publication 500-238.

Salton, G. (1989). *Automatic text processing*. Reading, MA: Addison-Wesley.

Schlichting, C., & Nilsen, E. (1997). Signal detection analysis of www search engines. In *Proceedings of the Second Human Factors on the Web Conference* (On-line) Available: http://www.microsoft.com/usability/webconf/schlichting/schlichting.htm.

Search Engine Watch (1998). (On-line). Available: http://searchenginewatch.com/.

Selberg, E., & Etzioni, O. (1995). Multi-service search and comparison using the MetaCrawler. In *Proceedings of the 4th International World Wide Web Conference*.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioural sciences* (2nd ed.). New York: McGraw-Hill.

Steinberg, S. G. (1996). Seek and ye shall find (maybe). *Wired*, *4*(5), 108–182.

Stobart, S., & Kerridge, S. (1996). *WWW search engine study*. Sunderland: School of Computing and Information Systems, University of Sunderland (On-line report) Available: http://osiris.sund/ac.uk/sst/se/.

Su, L. T., Chen, H., & Dong, X. (1998). Evaluation of Web-based search engines from the end-user's perspective: a pilot study. In *Proceedings of the Annual Conference for the American Society for Information Science* (pp. 348–361).

Van Rijsbergen, C. J. (1979). *Information retrieval*. London, UK: Butterworths.

Zorn, P., Emanoil, M., Marshall, L., & Panek, M. (1996). Advanced searching: tricks of the trade. *Online*, *20*(3).