PERGAMON

Information Processing and Management 35 (1999) 141–180

INFORMATION PROCESSING & MANAGEMENT

Finding information on the World Wide Web: the retrieval effectiveness of search engines

Michael Gordon *, Praveen Pathak

Computer and Information Systems, University of Michigan, Ann Arbor, MI 48109-1234, USA

Received 9 September 1998; accepted 16 September 1998

Abstract

Search engines are essential for finding information on the World Wide Web. We conducted a study to see how effective eight search engines are. Expert searchers sought information on the Web for users who had legitimate needs for information, and these users assessed the relevance of the information retrieved. We calculated traditional information retrieval measures of recall and precision at varying numbers of retrieved documents and used these as the bases for statistical comparisons of retrieval effectiveness among the eight search engines. We also calculated the likelihood that a document retrieved by one search engine was retrieved by other search engines as well.

1. Introduction

The Internet, and especially World Wide Web, are incredibly popular at homes and offices alike. Because of the absence of centralized control or authority, statistics about the Net lack some degree of certainty. There is no question, however, that the Net is enormous in terms of numbers of users, Web sites, and Web pages. For instance, an estimate of the minimum number of host machines on the Internet is over 16 million, nearly seven million more than a year ago, and over 11 million more than a year before that (Internet Domain Survey, 1997). Similarly, of the 220 million people in the United States and Canada over the age of 16, 23% (over 50 million) are estimated to use the Internet and 17% (over 37 million) the World Wide Web (CommerceNet/Nielsen, 1997). And there is every reason to believe these numbers will all swell significantly in the next few years, with some analysts suggesting that the Web is doubling in size every 100 to 125 days (Morgan, 1996).

* Corresponding author. E-mail: mdgordon@umich.edu

0306-4573/99/$ - see front matter
PII: S0306-4573(98)00041-7

Though many Internet-enabled applications and services are available today, the primary use of the Internet (other than e-mail) is for information retrieval. With the advance in easy to use Web page development tools, individuals have joined organizations in publishing information on almost any topic imaginable. Of course, with such a diversity of content, and with the enormous volume of information on the Internet, retrieving relevant information is far from assured.

There are four different methods for locating information on the Web[1]. First, you may go directly to a Web page simply by knowing its location. This is the reason for companies splattering their URLs over their TV, print and radio advertisements. Second, the hypertext links emanating from a Web page provide built-in associations to other pages that its author considers to provide related information. Third, 'narrowcast' services can 'push' pages at you that meet your particular user profile. Fourth, search engines allow users to state the kind of information they hope to find and then furnish information that hopefully relates to that description. This article discusses search engines to see how effective they are at furnishing relevant information. Specifically, we have studied eight major search engines to determine how effective a particular search can be in finding information on the Web. The study employed a rigorous methodology emphasizing (1) the elicitation of genuine information needs from genuine users, (2) relevance judgments made by those same individuals, (3) 'power searches' performed for those individuals by people with specialized expertise in Web search engines who sought to maximize the performance of each search engine and (4) the conduct of various statistical tests of search engine effectiveness aimed at meaningfully discriminating search engine effectiveness. In addition, this study introduces an evaluation technique that requires relevance evaluations for twenty documents per search engine, but then extrapolates those results to 200 documents per engine.

This study also investigates the degree of overlap among the pages returned by these search engines for both retrieved and relevant-retrieved documents.

2. Overview of search engines

It is fair to say that Internet-based information retrieval would collapse if search engines were not available; without search engines, searchers would be about as successful negotiating the Internet as someone trying to look up a phone number in an unsorted Manhattan phone book. While word of mouth pointers to pages from friends, acquaintances, and others are very useful, and the live hypertext links of the Web make it such a rich and convenient source of information, these means of negotiating the Internet do nothing for the user who does not even know where to begin looking: that is the job of search engines.

Search engines provide three chief facilities: (1) They gather together (conceptually, anyway) a set of Web pages that form the universe from which a searcher can retrieve information. (2) They represent the pages in this universe in a fashion that attempts to capture their content.

[1] Our chief interest in this study is the World Wide Web rather than the broader Internet. Like any information retrieval evaluation, we focus on retrieving 'documents'. For us, a document is a Web page (though, of course, a Web 'page' may contain as much text, graphics, etc. as 20 or more physical pages).

(3) They allow searchers to issue queries, and they employ information retrieval algorithms that attempt to find for them the most relevant pages from this universe. Search engines differ somewhat from each other along all these dimensions.

A search engine can gather new pages for its universe in two ways. First, individuals or companies who have created Web pages may directly contact the search engine to submit their new pages. Second, search engines employ so called Web 'spiders', 'crawlers', or 'robots' which traverse known Web pages, link to link, in search of new material. Differences among spiders determine the database of documents that a given search engine accesses as well as the timeliness of its contents. We address differences among spiders in more detail in Section 5.2.

Since computers cannot read and understand text, every page that a search engine might retrieve for a user must have its content represented in a way that a computer can process. There are two basic ways that Web search engines do this. First, the Web page, or an abbreviated version of it, may be *indexed* — or represented — by the set of words or phrases it contains (excluding overly common 'stop words' like *the*, *of*, *for*, etc.). As an example, this article, if it were a Web page, would be indexed by: *computers*, *read*, *understand*, *search*, and *search engines* (among other words and phrases), since each occurs in this paragraph. More sophisticated indexing techniques attempt to determine the *concepts* being used in a document, using statistical methods that correlate word and concept occurrences. The occurrence frequency of words (or phrases or concepts) is often maintained as well, as well as their location (in a page's title, a major heading, near the beginning of a document, etc.).

Most search engine indexes are considered to be *full text indexes* — meaning that a document is represented by and so will 'match' any query that uses any word or phrase contained within it (excluding stop words). In actuality, many search engines exclude the information in so called meta fields of Web pages (such as their author-supplied key words or descriptions, or their author fields) and in comment fields (which are used to document a page's internal structure), whereas others make use of this information. Further, the graphics on a page are not ordinarily indexed (but can be done so by special efforts by the page creator). So, a company with a home page containing its corporate logo complete with its company name may be dismayed to find that its home page is not readily retrieved by using its company name as a query. Pages containing frames and image maps or those that are password protected may also have some of their content ignored when they are indexed. And dynamically generated, 'virtual' web pages (such as a page composed on the fly containing a map, driving directions, and up-to-the-moment highway conditions) are ephemeral constructions that will not be indexed.

The second main way that Web pages are represented is by their position within a knowledge hierarchy developed and maintained by people. Yahoo!, the best known *subject directory* embracing this approach, begins with fourteen very general subject categories, then continues to subdivide them until individual Web pages are identified. For instance, the category *Arts and Humanities*, which is at the most general level, contains *Design Arts* at the next level, *Industrial Design* at the level under that, until, finally, there are lists of Web pages on very specific topics. These Web pages have been specially selected by subject specialists and assigned their position in the hierarchy by human indexers. (Along with this knowledge organization and navigation scheme, Yahoo! also allows more conventional searching for relevant Web pages. Our searchers used this facility to locate information.)

Search engine indexes and subject directory catalogs vary in the number of pages they contain, with 2 million and 100 million pages being at the small and large ends of the scale, respectively. Most major search engines contain 25–50 million pages, a figure that has stayed constant the last year even as the size of the Web has approached 150 million pages (Calafia, 1997), though other estimates put the size of the indexable Web at least double that number (Lawrence & Giles, 1998).

For many years, the field of information retrieval (IR) has devised search algorithms, which take a user's query and furnish him or her with a list of hopefully relevant documents (often ranked according to a 'relevance score' calculated by the algorithm). The eight search services in this study offer some combination of standard IR search facilities (see van Rijsbergen, 1979) for a good introduction)[2]. These include the abilities: to form Boolean queries; to specify that a term *should (not)* or *must (not)* appear in a Web page; to allow the user to use wildcards and truncation in search statements (to issue queries like *comput** to stand for *computers*, *computing*, *compute*, etc.) as well as for search algorithms to use automatic stemming (thus equating such items even if the user does not specifically ask that that be done); to search for phrases rather than individual words; to specify the importance of case sensitivity; and to do proximity searching (such as 'airplane within five words of Denver'). More advanced capabilities available in some engines include the abilities: to allow the user to write a query in the form of a complete sentence (or short paragraph) which the search engine then parses and exploits; to suggest to the user additional words or phrases to include to refine an initial query (based on his or her relevance judgments of those already presented) or to allow the user to specify certain already retrieved documents to serve as examples of the type he or she would like to see; and to show the user groupings of retrieved documents that reflect how various concepts occur among them.

The precise algorithms that search engines use for retrieval are not publicized, but one can infer their approximate workings by reading the Help, Hint or FAQ pages that accompany them as well as by being familiar with the field of IR. In most engines, a Web page typically will be highly ranked if it frequently uses many of the same words (or phrases) found in the query, especially if those are relatively rare words to begin with. The appearance of these items in a page's title, heading, or early in its text tends to raise the relevance scores of the page even further.

Web-based information retrieval also differs in some respects from more traditional retrieval. For instance, some search engines allow a user to restrict retrieval to certain domains (only .com sites) or specific domain names (such as ibm.com) or even to specify, for instance, that all the pages he or she wants to see should have a *link* to ibm.com). Further, some search engines allow the user to specify, for example, that the pages he or she is interested in should contain a plug-in, script, Java applet or embedded real audio file.

The matching algorithms that search engines employ also often embrace certain principles that apply to Web-based searching but not traditional IR. As an example, some search engines boost the relevance scores of pages that have many incoming links, the argument being that

[2] There is no end to the number of published Web pages that depict differences among search engine features. A very comprehensive guide to such information is found at: Calafia (1997). Since search engine features can change quickly, it is generally better to determine them using online resources, which can be kept more up-to-date than print resources.

these are popular pages and so are the ones people will want to retrieve. Similarly, if a search engine maintains a list of 'reviewed sites', these, too, may receive a higher relevance rating than they otherwise would, on the assumption that a reviewed site indicates that it is above average in quality and, thus, its pages are those that a searcher is most likely to want to retrieve.

Some search engines also reduce the relevance scores of certain pages for violating rules of fair play. Since most search engines favor pages with multiple occurrences of a user's query term, and since some search engines index the text contained in a page's keyword (meta) fields, some web authors load up those fields with many copies of the same, popular word(s) in hopes of getting their page retrieved. These words may be completely irrelevant to the page. More subtle gimmicks include putting blue text, say, on an identically colored blue background to make it invisible (to the eye but not a computer indexing that page). To defeat these attempts, some search engines adopt 'spamming penalties' that reduce the relevance scores for offending pages or even fail to give them a score entirely.

In short, search engines are indispensable for searching the Web, they employ a variety of relatively advanced IR techniques, and there are some peculiar aspects of search engines that make searching the Web different than more conventional information retrieval. The fundamental question motivating this research then is: How good are the different, popular search engines when used to address genuine information needs and used by highly effective searchers? By answering this question, we can learn about the optimal performance of Web search tools, as well as their comparative effectiveness. Thus, the search results reported here provide an indication of the level of performance that can be obtained today by knowledgeable, diligent searchers to meet a variety of information needs. Secondly, since the Web is so vast, and since the search engines that search it differ in both their search algorithms and the portion of the Web they index at a particular moment in time, we attempt to determine the degree to which these services overlap — in terms both of the documents they index, and those that they find relevant for a given searcher. Together, these measurements of retrieval effectiveness and retrieval overlap provide important benchmarks to understand how easily one can find information on the World Wide Web.

3. Previous web studies

Evaluations of search engines are of two types: testimonials and shootouts. Testimonials are generally conducted by the trade press or by computer industry organizations who 'test drive' and then compare search engines on the bases of speed, ease of use, interface design or other features that are readily apparent to users of the search engine. Another type of testimonial evaluation comes from looking at the more technical features of search engines and making comparisons among them on that basis. Such testimonials are based on features like the set of search capabilities different engines have (e.g., mandatory/optional, or proximity operators), the inclusiveness of their coverage, the rate at which newly developed pages are indexed and made available for searching, etc[3].

[3] There are literally dozens of testimonials on the Web and in print. For a sample of just a few see: Slot (1997), Morville, Rosenfeld, & Janes (1996), Lake (1997), Overton (1996), Courtois (1996), Calafia (1997) or Steinberg (1996).

Though testimonials can give users some useful information in making decisions about which search engine to employ, they only indirectly suggest which search engines are most effective in retrieving relevant web pages. This is where 'shoot outs' come in.

In shoot outs, different search engines are actually used to retrieve Web pages and their effectiveness in doing so is compared. Shoot outs resemble the typical information retrieval evaluations that take place in laboratory settings to compare different retrieval algorithms (see Sparck Jones (1997), e.g.), though Internet shoot outs often consider only the first ten to 20 documents retrieved, whereas traditional IR studies often consider many more.

Whereas search engines syntactically and statistically compare documents to a query, their real job, of course, is to meet a user's information need. An *information need* is an explanation of the information one would like to receive from a search. It is different from a query, which is processable by a search engine, in that it provides a full, relatively complete description of the kind of information the individual wants to obtain and then often must be translated into the appropriate syntax and even vocabulary before the search engine can process it. Each search engine places different constraints on how an information need can be expressed by a query. In addition, since different search engines employ different search algorithms, even if several engines process the identical query against the identical set of documents, the way they rank those documents may differ.

The most typical measurements of IR effectiveness are recall and precision. *Recall* measures the proportion of relevant documents in the database that is actually retrieved. *Precision* is the proportion of retrieved documents that is relevant.

To place in context some of the previously published studies on search engine shoot outs, it is useful to list seven features that make such an evaluation most accurate and informative. As in this study, we assume that the goal of such research should be to determine the power of each tool in a genuine retrieval setting.

First, the searches should be motivated by the genuine information needs of searchers. Experimenters who personally think up searches for an experiment may introduce biases into an experiment (say by composing searches which favor a particular search engine); in addition, they can never approximate the incredible diversity of information that people search for — and need — in searching the Web; nor can they reflect the nuances of these searches.

Second, if an experiment is seeking documents on a search topic someone else has identified, that person's information need should be captured as fully and with as much context as possible. A list of keywords, even with structuring grammar (like Boolean or proximity operators) can only provide an very rough approximation of the kind of information the individual requiring information really desires.

Third, a sufficiently large number of searches must be conducted to produce meaningful evaluations of search engine effectiveness.

Fourth, the shoot-out should include most major search engines.

Fifth, the effectiveness of different search engines must be analyzed by exploiting the special features of each engine. This means that the same computer-processable query (possibly with slight syntactic or stylistic variations) should *not* necessarily be used with different search engines to find Web pages for the same information need. Otherwise, the best features of a given search engine may not come into play.

Sixth, relevance judgments must be made by the individual who needs the information. Otherwise — if experimenters decide which Web pages are and are not relevant — there will

be numerous mis-evaluations due both to the experimenter's lack of familiarity with the subject being searched and to the impossibility of him or her knowing the user's needs, background, motivation, etc. with anywhere near enough detail to decide whether the user's information need is really met.

We can't emphasize enough the importance of relevance evaluations being made by those who actually require the information. In the spirit of Pierce's pragmatic theory of signs (Cherry, 1980) we note that documents are not simply relevant or not (for a given information need) — but rather they are relevant or not for a *particular person* with that need. This pragmatic level of relevance accounts for that person's particular situation, including his or her background knowledge, the general reasons he or she is requesting this information, any specific uses to which it will be put, etc. As Cherry notes, "to the pragmatic level we must relegate all questions of value or usefulness... all questions of interpretation, and all other aspects which we would regard as psychological in character" (Cherry, 1980, p. 243). In short, from a pragmatic viewpoint, the same document may mean different things to different people; this adversely colors any attempts by impartial 'relevance judges' to make decisions about whether someone else would find a particular document relevant. Relevance judges can only make semantic or even syntactic evaluations of documents and queries. But these judgments fail to involve the particular user, and so fail to identify whether the user *really* finds a particular document relevant. To modify an old phrase: relevance is in the details.

Finally, in addition to the above criteria, well-conducted experiments are necessary to obtain meaningful measures of performance. This means (1) following appropriate experimental design (for example by randomizing the order in which documents are presented to evaluators to overcome any ordering effects), (2) conforming to accepted IR measurements (like recall-precision curves) to allow results to be evaluated in a familiar context and (3) using statistical tests to measure accurately differences in performances among search engines.

There are a number of shoot-outs reported in the literature, but none satisfies the seven features we have just listed (see Table 1). These tests vary in many ways. For one, they differ in terms of the information needs they address. Westera (1996), for instance, only issued queries relating to wine, whereas Feldman (1997) tasted a set of diverse queries on topics such as cars, information retrieval, and tennis elbow. In many studies, queries were made up by experimenters, and in others, queries were drawn from training guides or reference books. Of course, in such cases, it is impossible to have a fully stated information need; instead one must simply use or adapt slightly a given query.

Many studies issued the same (or nearly the same) query to all search engines. While this might mimic the search behavior of some novice searchers, it does not test the true capabilities of search engines. In some cases, authors of studies commented that they had consulted the Help or FAQ pages in devising their queries. But in no previous study, except for Lake (1997)[4], were searchers expected to exploit the search engines under investigation to the fullest.

Also, previous studies involved users in making relevance judgments. Instead, relevance judgments were ordinarily made by the experimenters themselves. Although experimenters may try objectively to test a range of different queries that different users might pose, it is next to

[4] See also: *PC Computing* (1997) AltaVista vs. Excite vs. HotBot vs. Infoseek: Which is the one to rely on? *PC Computing* Search Engine Challenge. http://www4.zdnet.com/pccomp/srchoff/srchoff.html.

Table 1
Previous search engine evaluation studies

	Genuine search?	Information need stated?[a]	Number of searches	Number of search engines	Queries optimized per search engine?	Relevance judged by actual users?	Appropriate experimental design and evaluation?
Leighton, 1995	no	–	8	4	no	no	no
Leighton and Srivastava, 1997	yes	no	15	5	no	no	yes
Ding and Marchionini, 1996	no	–	5	3	yes	no	yes
Chu and Rosenthal, 1996	yes	no	10	3	yes	no	no
Westera, 1996	no	–	5	8	no	no[b]	no
Lebedev, 1997	no	–	8	11	no	no[b]	no
Overton, 1996	no	–	10	8	no	no	no
Schlichting and Nilsen, 1996	yes	yes	5	4	no	yes	no
Feldman, 1997	no	–	7 discussed	7	no	yes	no
Lake, 1997	no	no	40 *facts*	4	yes	evaluation = facts	no
Tomaiuolo and Packer, 1996	some	no	200	5	yes	no	no
Gordon and Pathak (current study)	yes	yes	33	8	yes	yes	yes

[a] Only applies to studies involving an information need specified by 'genuine searches' (i.e. those in which people other than the experimenters generate the information need).
[b] Study counted number of 'hits' (documents retrieved by search engine) rather than making relevance assessments.

impossible for them to determine accurately which documents would be relevant. Tomaiuolo and Packer (1996)[5] tested an impressive 200 queries; but they made relevance judgments themselves, often based just on the short summary descriptions of Web pages that search engines provide. In the Lake (1997) study, where the task was to find Web pages that provided answers to factual questions, judgments by experimenters about the accuracy of the information retrieved is acceptable[6]. In IR studies, however, the emphasis is on the ability of the retrieval system to identify different documents — not facts. So, while the Lake (1997) study is an interesting test of search engine features, it does not address the effectiveness of search engines from an IR perspective.

Finally, studies differed greatly in the amount of experimental rigor they employ. Leighton and Srivastava (1997) and Ding and Marchionini (1996) performed statistical tests to compare search engines. And Leighton and Srivastava (1997) made blind relevance assessments, meaning that they (acting as relevance judges) were unaware of the search engine that had returned any given document that they were evaluating. Other studies failed to adopt this same degree of rigor, though, in fairness, many of them were intended to be less formal studies.

[5] See also: Quantitative analysis of five WWW 'search engines'. http://neal.ctstateu.edu:2001/htdocs/web-search.html.
[6] Some sample questions from the *PC Computing* Shoot-out (Lake, 1997): How many floors are in the Sears Tower? In what year were the first faxes sent? What is the cost of a real Faberge egg? What's the most money you could make for catching an FBI fugitive?

In the next section, we describe an experiment in which we attempted a rigorous comparison of the effectiveness of eight major search engines. The experiment met all the criteria for a successful evaluation discussed in this section and listed in Table 1.

4. Experiment

We conducted an experiment both to measure the effectiveness of eight popular search engines and to determine the extent to which they retrieve the same Web pages. All searches were conducted by highly trained searchers on behalf of faculty who were looking for information from the Internet to support their research or teaching and who personally evaluated the effectiveness of the Web pages returned to them. Searchers exploited each search engine to the fullest extent possible.

4.1. Information needs

In this experiment, thirty six faculty at the University of Michigan Business School filled out a search form designed to elicit their information need. The entire faculty (approximately 125 people) was invited to participate in this experiment. They were told that well trained searchers would use powerful tools to search the Internet for items matching their information need in exchange for their evaluating the relevance of the items actually found. Thirty six faculty accepted this offer and participated in the study; three were later dropped from consideration after they failed to provide evaluations for the searches conducted for them.

The information needs of these people varied widely, their primary focus including: corporate strategy and management, communications, finance, organizational behavior, marketing, business law, accounting, communications, information systems, operations management, and international business (see Table 2). Within any one of these areas, the information needs of different individuals varied widely as well.

A search form was used to elicit faculty members' information needs. It contained five sections for describing an information need. The first of these sections allowed the faculty member to express an information need with a short paragraph. A sample information need was:

> Find information regarding entrepreneurship, especially success factors, magazines, (e-zines) and networking opportunities for entrepreneurs. The information sources should exclude franchising and business opportunities, services offering entrepreneurs an Internet presence (including web page designing, ISPs and electronic malls). It may, however, include sources for supplies and other technologies, patent and legal information, etc. The information sources should be of interest to the budding entrepreneur, helping them avoid pitfalls and provide a sense of community.

The next sections of the search form asked the faculty member to provide information that would help a searcher translate the textual description of his or her topic of interest into a search statement processable by a search engine. Specifically, faculty were asked (1) to identify any important phrases used in their search (such as *success factors* and *networking*

Table 2
Distribution of the topics areas of faculty participants

Topic area	Number of participants
Accounting	2
Communications	2
Corporate strategy and management	6
Finance	1
Information systems	5
International business	2
Law	2
Marketing	2
Operations management	2
Organizational behavior/psychology	8
Statistics	1

opportunities, in the above example); (2) to identify the most important words or phrases in their textual descriptions (*entrepreneur(ship)*; *network(ing)*; and *success factors*); (3) to identify any synonyms or related terms that they thought might help the searcher (such as *entrepreneur = small business*), as well as any terms or phrases that they thought might be confused with terms of interest (e.g., not *computer networks*); and (4) to phrase their search in the form of a Boolean query and then comment on their confidence about expressing their information need in that way.

4.2. Search engines

Seven search engines and one subject directory were used in this study. (For ease of expression, we often refer to them collectively as search engines). The search engines proper were: AltaVista, Excite, Infoseek, Open Text, HotBot, Lycos and Magellan. Yahoo! was the subject directory, though we used its search capabilities for this experiment. Although estimates put the number of search engines/services at 1800[7] (Feldman, 1997), the search engines in our study include the major search tools in use and incorporate most of the sophisticated spidering, indexing, and searching techniques being used today. Thus, they are the most sensible selection in studying search engine effectiveness and overlap.

4.3. Searchers

The searchers in this experiment were all highly qualified to search the Web on behalf of the faculty. All searchers had strong backgrounds on the Internet and in using search engines. The typical profile of a searcher in the experiment was someone either holding or earning a masters degree in library and information studies who had training in online research methods.

[7] Most of these are specialty search engines that only cover a specific subject like automobiles or sports.

Before the experiment began, all searchers received a review of effective search strategies. Each also received detailed online and written training about using all of the search capabilities of all of the search engines used in the experiment and practiced using all features of all search engines to ensure their complete familiarity with each engine. Reference materials were provided to each searcher in print and electronic form for their use throughout the experiment.

Searchers were paid on an hourly basis for the time they spent in training, for the practice searches they conducted with each search engine and for conducting the searches that constituted this experiment.

4.4. Search task

A searcher's job was to take a faculty member's information need and find the best way to express it with each search engine. Since search engines have different features, searchers almost always needed to determine a different 'best query' for each search engine. Specifically, searchers were instructed to construct a query for a given search engine that would return the most relevant documents possible among the first 200 retrieved. (In some cases, such an 'optimal' query retrieved fewer than 200 documents.)

Searchers received all the information provided on a faculty member's search form (the textual statement of an information need, phrase and synonym information, the faculty member's Boolean search statement, etc.). A searcher's task was to determine how best to use this information with a particular search engine, given its particular feature set, and keeping in mind the objective of trying to retrieve as many relevant documents as possible among the first 200 retrieved. The textual statement of information need and the other information on the form all suggested possible search approaches, any of which the searcher could follow or not. In situations where searchers had difficulty understanding or interpreting a faculty member's true information need, the faculty member and other subject experts were available to provide clarification or background information.

Searchers were instructed to search repeatedly and in an exploratory fashion with each search engine. With Lycos, for example, the searcher would try countless search variations for the same information need, each time examining the Web pages retrieved and then making search adjustments based on the effectiveness of previous search attempts as well as any new, potentially useful terminology uncovered by these searches. Typically, a searcher might compare the effectiveness (as he or she judged it) of thirty or more different Lycos searches pertinent to a given information need before determining which appeared to be the best. Each search would be an attempt to use Lycos to its fullest capability for the particular information need. For search engines that permitted such an option, searching was constrained to 'just Web pages' rather than information that might come from ftp sites, gopher sites, etc.

The seven remaining search engines were explored similarly: iteratively and with an emphasis on exploration and manually incorporating relevance feedback (or even by using a search engine's 'More pages like this' feature) — all in an attempt to find the query that would perform best for that search engine. Searchers were even instructed to allow the results of searching with one search engine to influence their use of another. For example, Web pages that appeared relevant and were retrieved by one search engine could serve as targets that another search engine could search for. Similarly, any useful terminology uncovered by one search engine could be used by another. Searchers were asked not absolutely to 'freeze' their

best search for a given search engine until all others were tried. So, even in the case where a searcher used Excite after having determined a 'best Lycos' search, he or she might learn something in performing the Excite search that would require the Lycos search to be re-done.

The searcher's task was considered complete when he or she had constructed a 'best possible' query for each search engine (eight altogether). At that point, the URLs of the top 200 Web pages were saved in rank order for each search engine. A searcher typically spent one to two 8-hour days conducting the search for a single faculty member.

4.5. User evaluations

The top 20 Web pages from each of the eight search engines were printed. (If a Web page had more than 10 physical pages of paper to be printed, only the first 10 pages were printed.) Previously, all 1600 (200 documents per search engine*8 search engines) Web pages had been given random, unique identifiers. The top 160 (20*8) printed Web pages were annotated with their identifiers and then were arranged in random order, bound in booklets, and given to the faculty member serving as the subject of the experiment[8].

The faculty member for whom the search was performed received the booklets (usually two or three telephone book-sized booklets) together with a copy of the search form he or she had submitted and a relevance evaluation form. The relevance evaluation form listed the 160 documents in the same (random) order that they occurred in the booklet. The faculty member's task was to indicate whether each Web page in the booklet was *highly relevant*, *somewhat relevant*, *somewhat irrelevant* or *highly irrelevant*. These judgments formed the basis of the results that follow.

Specifically, the faculty member was told to

> determine the degree to which each document is relevant to the topic you said you wanted documents about [on the search form]. Judge each document separately and independently of all other documents. That is, don't assume that one document can't be relevant because you just judged another one relevant, or because you've seen it before.

Further, he or she was told that judgments could be made without necessarily reading all documents in their entirety.

5. Results

The experiment conducted allowed us to test both the retrieval effectiveness of the search engines studied (Section 5.1), as well as the overlap among the documents they retrieved (Section 5.2).

---

[8] A pretest version of the experiment provided faculty with a similar, randomized list of 'live' URLs for these same 160 Web pages. Unfortunately, subjects found it too easy to visit the referenced Web page, follow its links, and then never complete the experiment.

5.1. Retrieval effectiveness

In studying the effectiveness of search engines, we are interested in two questions. (1) How effective is a given search engine in retrieving only relevant Web pages? (2) Is a given search engine good at finding most (or a high percentage) of the existing set of relevant pages?

Clearly, these questions can be asked at various times — after the first document is retrieved, after 10 are retrieved, or 20, or 200 — and the answers to them will vary as well. Accordingly, the results that follow are based on recall and precision values computed at various document cut-off values. For each of the 33 subjects' information needs, and separately for each search engine, we computed precision (proportion of retrieved documents judged relevant) and recall (proportion of relevant documents retrieved[9]) after every retrieved document. As Hull (1993) points out, one can then calculate the average precision (or recall) of each search engine, computed across queries at any fixed document cut-off value (i.e., any fixed number of retrieved documents), thereby equating evaluators' search effort across queries.

Earlier, we explained that only twenty relevance evaluations were made for each query-search engine pair, but now we have suggested that, in fact, the top 200 documents identified by any search engine were considered in our effectiveness evaluations. The explanation of this difference lies in the fact that among the 1600 potentially different Web pages retrieved by the eight search engines were some that were retrieved by more than one. A computer program was written that could read the source text of each retrieved Web page and determine which ones were retrieved by multiple search engines. This aided our experiment in two ways. First, it allowed us avoid printing and presenting the faculty member with the same Web page twice[10]. Second, and more importantly, it permitted us to tell when a Web page that a particular search engine ranked among positions 21–200 (and, thus, which would not ordinarily be printed and presented to the faculty member) was ranked one through twenty by another search engine (and, so, was printed and presented). For instance, Lycos' 87th best-ranked Web page would ordinarily not be printed and evaluated by the faculty member; but if this same Web page were the 14th best-ranked item by Excite, it would have been printed and evaluated: thus, it would be evaluated 'for free' for Lycos.

Such cross-referencing allowed us to extrapolate our findings to include all 200 Web pages identified by a given search engine, not just the first twenty. An assumption behind doing so was that the documents evaluated 'for free' were evenly distributed among the eight search engines.

Otherwise, search engines with more 'free' evaluations might enjoy inflated recall and precision scores compared to other search engines, since we assumed that Web pages with ranks 21–200 that were not evaluated for free were nonrelevant. This assumption was borne out on empirical grounds.

---

[9] Actually we computed a statistic often called *relative recall* in the IR literature. A strict calculation of recall requires knowing — for *every* page on the Web — whether or not it is relevant to a given information need. Since this is not at all possible, we calculated recall as a percentage of (i.e., *relative to*) the retrieved documents that the faculty evaluator judged to be relevant.

[10] Mirror sites (which contain identical information but with different URLs) were not detected, however.

We next consider the measures we used in our experiment. As we have mentioned, for Infoseek (as one example among eight), we were able to calculate precision for each of the 33 subjects' 'queries'[11] at any document cutoff values between 1 and 200. So, to compute the precision effectiveness of Infoseek at 'about 20' retrieved documents, we first calculated its average precision across all queries at a document cut-off value of 15, then again at 16, 17, 18, 19 and 20. Following the suggestion made by Hull (1993), we averaged these averages — thus 'smoothing' any artifactual differences that might arise by selecting one of these specific document cut-off values — to obtain what we will call Infoseek's *average precision at document cut-off value 15–20*. Other search engines were treated identically to permit comparisons among them.

We then performed an analysis of variance among search engines based on such average precision scores for this range of document cut-off values, using information needs as a blocking factor to minimize the effects of performance differences among them. With a sample size of 33 information needs, the underlying assumptions of an analysis of variance are expected to be met. Nevertheless, we conducted an assumption-free nonparametric analysis of variance for the same document cut-off value range to help corroborate our results. In the nonparametric test, average precision ranks across search engines were used as the basis of analysis — rather than average precision values. This same type of analysis was repeated for other document cut-off ranges and for average recall, instead of average precision.

The main results we present are based on a 'lenient' encoding of evaluators' relevance judgments that codes a judgment that a document is *highly relevant* or *somewhat relevant* as 'relevant', and that codes judgments of *somewhat irrelevant* or *highly irrelevant* as 'not relevant'. This decision was made after we had received the faculty judges' evaluations and was due to the fact that very few documents were judged highly relevant. For contrast, we will also briefly discuss retrieval effectiveness when the 'strict' encoding of relevance (only *highly relevant* documents) was the basis of our coding.

5.1.1. Precision results

Precision is a more important indicator of effectiveness than recall for searchers who only examine a few documents (Hull, 1993), and this is the behavior of most Web searchers (Lesk, 1997). Accordingly, we focus the discussion of precision results on low document cut-off values. For the results that we present first, a lenient coding of 'relevance' was used.

For the first document retrieved, average precision (across information needs) ranged from 66.7% (for Open Text) down to 12.1% for Yahoo. The other six search engines had average precision between 20 and 40%. For document cut-off values between 2 and 20, the precision of these six search engines generally stayed constant and remained in this 20–40% band. By 10 retrieved documents there was a spread in average precision from 40.6% for AltaVista down to 17.6% for Yahoo.

By 10 retrieved documents, a pattern was established that held through the 200th document retrieved; namely, AltaVista, Open Text, and Lycos were the top precision performers (in that

[11] Typical information retrieval studies talk about 'queries.' To emphasize that different queries were constructed based on a common information need, we prefer the term *information need* to *query*.

order), Yahoo! was the worst, and Infoseek, Excite, Magellan and HotBot occupied positions four through seven in different orders depending on the document cut-off value (Figs. 1 and 2).

Statistically, there were significant differences ($p < 0.05$) in average precision performance among search engines for all document cut-off value ranges studied 1–5, 1–10, 5–10, and 15–20, as calculated by either a standard or ranked (nonparametric) analysis of variance. Tukey's highest significant differences were calculated with $\alpha = 0.05$ to determine the subsets of three or more search engines that were statistically indistinguishable. For the range of document cut-off values 1–5, there were two different (overlapping) subsets. First, all search engines except Open Text (with its high average precision of 52.8% for this cut-off range) were statistically indistinguishable as a group; second, the top four performers for this range (Open Text, Infoseek, Lycos, and AltaVista) were indistinguishable as a group. See Table 3. The interpretation of a 'cluster' in the table is that an ANOVA would find no difference in effectiveness among just those search engines at $\alpha = 0.05$. For other document cut-off ranges, highest significant differences were also calculated (Tables 4–6). By the time 15–20 documents were retrieved, four overlapping clusters had developed. The best included AltaVista (which retrieved slightly more than 40% relevant documents through cut-off values 15–20) and Open Text; Open Text, Lycos and Excite formed the next best cluster; and the performance of Yahoo!, HotBot, Magellan, Infoseek and Excite could not be distinguished statistically at the low end of precision performance. Table 7 summarizes the 'clustering' of search engine average precision for different document cut-off values.

Pairwise for the range 1-5 documents retrieved, Tukey's multiple comparisons showed significant differences between Open Text and each of the four lowest performers and no others. For the range including document cut-off values 5–10, Yahoo!, with the lowest average precision, was statistically different than the three top performers: Open Text, AltaVista and Lycos; Open Text had a statistically significant different average precision than Excite. For document cut-off values 15–20, AltaVista was the clear winner. With precision of over 40% per query when precision was calculated and averaged at these six document cut-off values, its mean precision performance was statistically significantly better than all other search engines except Open Text. Open Text was statistically better than the bottom four performers. Yahoo! had a statistically lower mean precision for this range than Lycos as well as AltaVista and Open Text. These pairwise comparisons are summarized in Table 8. For searchers who are looking for relevant documents and will only look at the first 20 URLs a search engine furnishes (and 20 is a typical default size), these results suggest using AltaVista or Open Text.

The situation when only documents judged to be *highly relevant* were coded as 'relevant' — coding documents rated *somewhat relevant* as 'irrelevant' along with those judged *somewhat* (or *highly*) *irrelevant* — was somewhat different. Of course, the absolute level of average precision was lower compared to the more lenient coding of 'relevance', since, by definition, one will necessarily find fewer highly relevant than relevant documents (or an equal number in unusual circumstances). After just one retrieved document, average precision ranged from 24.2% (for Infoseek) down to 6% (for Yahoo! and Magellan). Interestingly, by the 20th document retrieved, the spread in average precision was just about the same: just under 18.8% (for AltaVista) down to 7.2% for Yahoo! (Figs. 3 and 4). The relatively 'flat' (and in some cases *rising*) precision curves up to document cut-off value 20 suggest that search engines are not especially good at presenting searchers with the mostly highly relevant documents first. Instead, they tend to sprinkle them rather uniformly among these first 20 positions. The lesson for

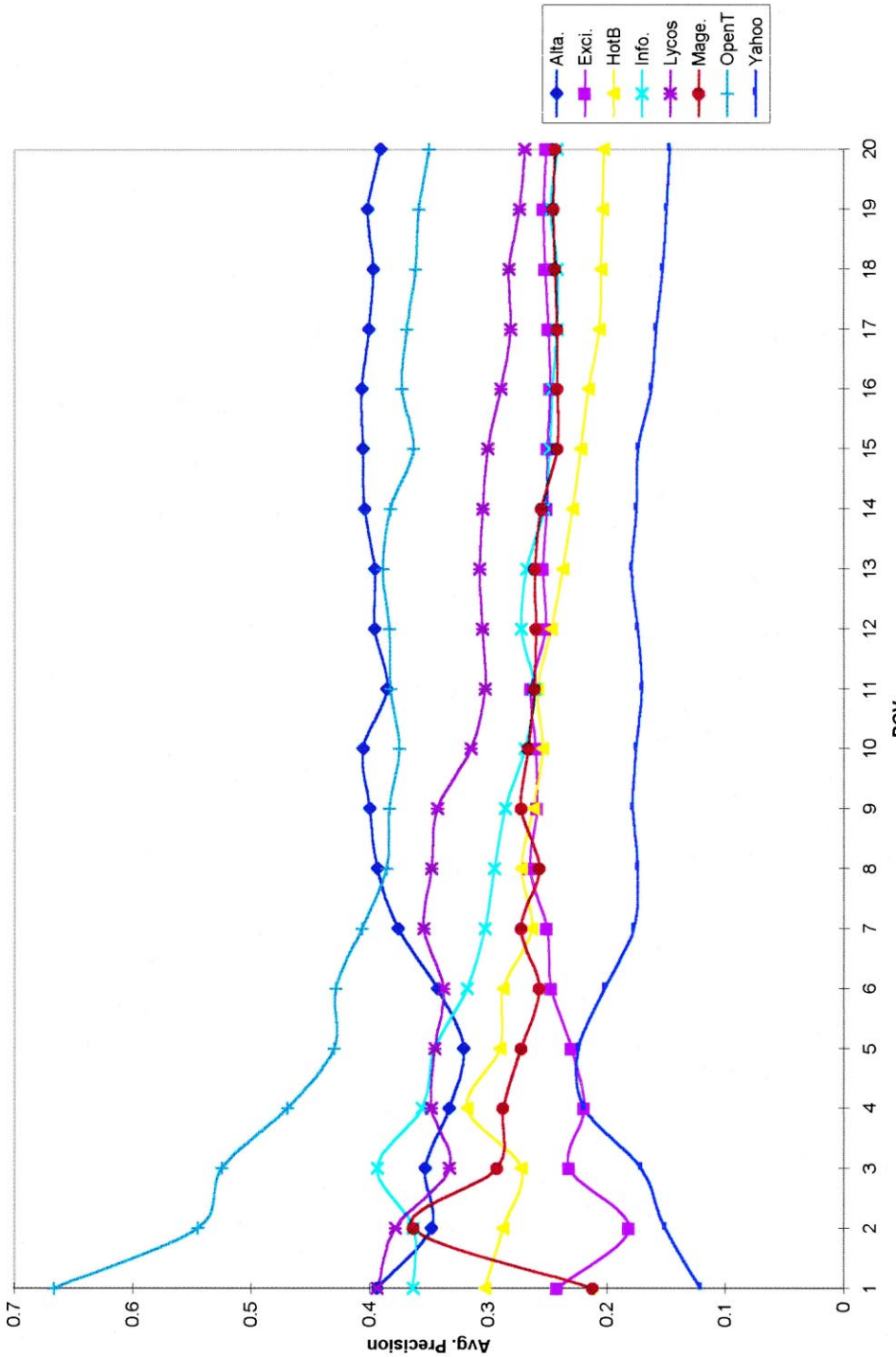

Fig. 1. Average precision vs. DCV (1–20) (lenient encoding).

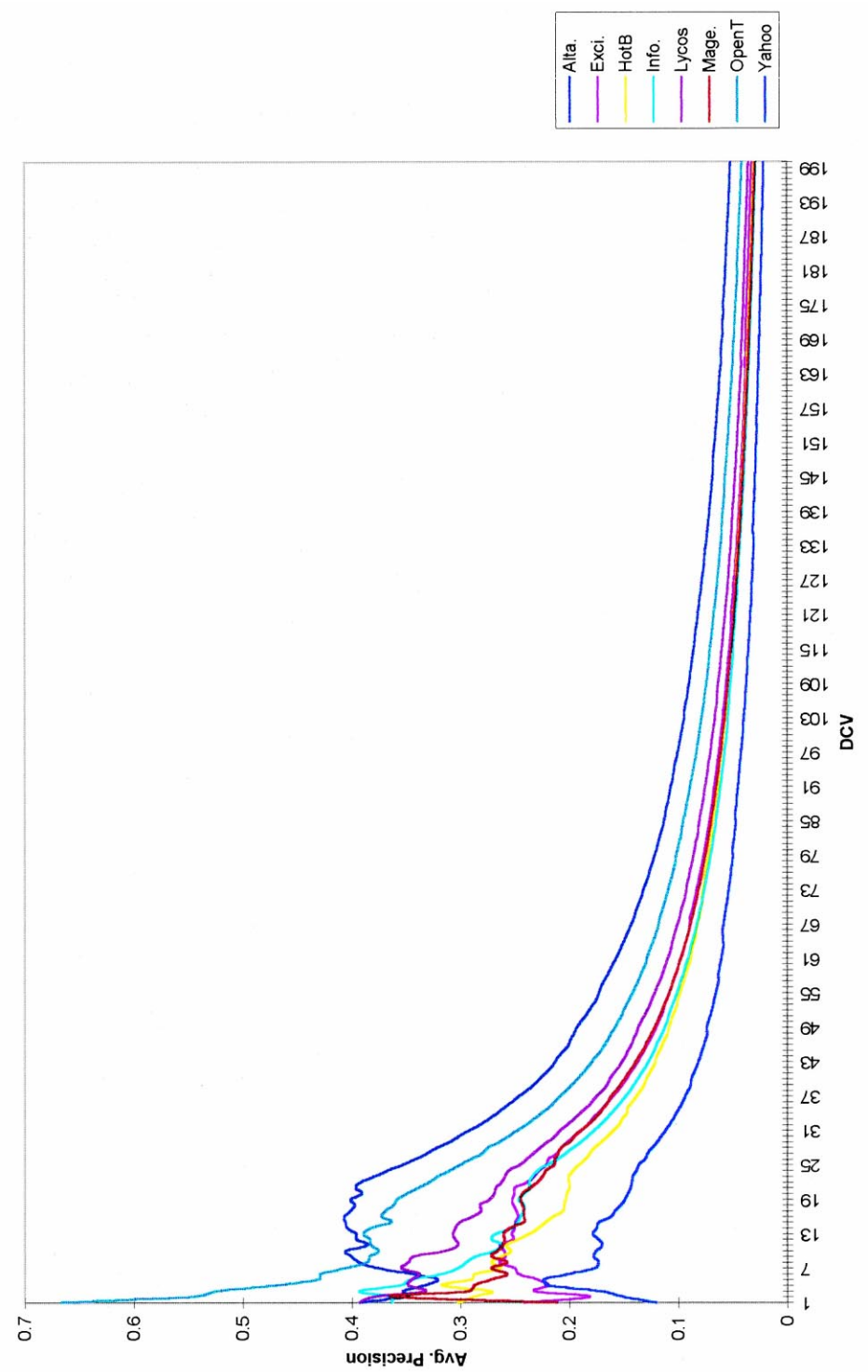

Fig. 2. Average precision vs. DCV (1–200) (lenient encoding).

Table 3
Tukey's highest significant difference for document cut-off values 1–5 based on precision. $\alpha = 0.05$.

Search engine	N	Subset	
		1	2
Yahoo	33	0.177	
Excite	33	0.2213	
Magellan	33	0.2859	
HotBot	33	0.2945	
AltaVista	33	0.3501	0.3501
Lycos	33	0.3600	0.3600
Infoseek	33	0.3645	0.3645
Open Text	33		0.5275
Sig.		0.058	0.088

Values shown are average precisions for document cut-off values 1–5.

Table 4
Tukey's highest significant difference for document cut-off values 1–10 based on precision. $\alpha = 0.05$

Search engine	N	Subset		
		1	2	3
Yahoo	33	0.1796		
Excite	33	0.2390	0.2390	
Magellan	33	0.2757	0.2757	
HotBot	33	0.2815	0.2815	
Infoseek	33	0.3295	0.3295	0.3295
Lycos	33		0.3500	0.3500
AltaVista	33		0.3671	0.3671
Open Text	33			0.4620
Sig.		0.075	0.209	0.174

Values shown are average precisions for document cut-off values 1–10.

Table 5
Tukey's highest significant difference for document cut-off values 5–10 based on precision. $\alpha = 0.05$

Search engine	N	Subset		
		1	2	3
Yahoo	33	0.1887		
Excite	33	0.2523	0.2523	
Magellan	33	0.2667	0.2667	0.2667
HotBot	33	0.2721	0.2721	0.2721
Infoseek	33	0.3030	0.3030	0.3030
Lycos	33		0.3410	0.3410
AltaVista	33		0.3737	0.3737
Open Text	33			0.4021
Sig.		0.260	0.192	0.095

Values shown are average precisions for document cut-off values 5–10.

Table 6
Tukey's highest significant difference for document cut-off values 15–20 based on precision. $\alpha = 0.05$

Search Engine	*N*	Subset			
		1	2	3	4
Yahoo	33	0.1576			
HotBot	33	0.2096	0.2096		
Magellan	33	0.2435	0.2435		
Infoseek	33	0.2449	0.2449		
Excite	33	0.2510	0.2510	0.2510	
Lycos	33		0.2832	0.2832	
Open Text	33			0.3628	0.3628
AltaVista	33				0.4007
Sig.	33	0.189	0.494	0.053	0.971

Values shown are average precisions for document cut-off values 15–20.

searchers here is simple: if you want to see the most highly relevant Web pages, don't give up after examining just the first few URLs in the ordered list a search engine presents you. Examine at least 20.

Another interesting finding is that the best precision-performer for these highly relevant documents through document cut-off value 7 is Infoseek, a distinction it did not earn when both highly and somewhat relevant documents were coded as 'relevant.' The suggestion here may be that search engines have different retrieval 'personalities,' and so possibly different preferred uses. For instance, Infoseek may be most effective at making early identifications of

Table 7
This table shows which search engines 'cluster' together according to Tukey's highest significant differences for different document cut-off value ranges (d.c.v.'s). 'Low' indicates lowest precision, and 'high' indicates highest

Yahoo	a_1		b_1			c_1			d_1			
Excite	a_1		b_1	b_2		c_1	c_2		d_1	d_2	d_3	
Magellan	a_1		b_1	b_2		c_1	c_2	c_3	d_1	d_2		
HotBot	a_1		b_1	b_2		c_1	c_2	c_3	d_1	d_2		
Infoseek	a_1	a_2	b_1	b_2	b_3	c_1	c_2	c_3	d_1	d_2		
Lycos	a_1	a_2		b_2	b_3		c_2	c_3		d_2	d_3	
AltaVista	a_1	a_2		b_2	b_3		c_2	c_3				d_4
Open Text		a_2			b_3			c_3			d_3	d_4

Legend

d.c.v. 1 -5	d.c.v. 1-10	d.c.v. 5-10	d.c.v. 15-20
a_1 = low	b_1 = low	c_1 = low	d_1 = low
a_2 = high	b_2 = medium	c_2 = medium	d_2 = low-medium
	b_3 = high	c_3 = high	d_3 = medium-high
			d_4 = high

Table 8
Pairwise average precision comparison

	Lycos	Infoseek	Excite	Magellan	HotBot	Yahoo!
AltaVista	d	d	d	d	d	b,c,d
Open Text		d	a,b,c	a,b,d	a,b,d	a,b,c,d
Lycos						b,c,d

Legend.

label	a	b	c	d
d.c.v. range	1-5	1-10	5-10	15-20

This table shows the document cut-off values for which the search engine in the column at the left of the table had a statistically significant difference (improvement) in average precision compared to the search engine named in the row across the top. Tukey's test ($\alpha = 0.05$) was used for these calculations, and there were no other statistically significant differences among means.

highly relevant documents; whereas AltaVista, for instance, may be better viewed as a search engine that will outperform the others by the 20th document retrieved, but *won't* especially shine earlier. Indeed, by the 20th retrieved document, AltaVista and Open Text had the best average precision for highly relevant documents and Yahoo! and HotBot were pulling up the rear — just as Fig. 1 showed for the more lenient encoding of relevance — while Infoseek had slid to third best, joining the same group of middle of the pack performers as in the more lenient case.

5.1.2. Recall results

Whereas average precision calculated over a range of document cut-off values favors those relevant documents that are retrieved earliest, average recall across a range of document cut-off values regards each relevant-retrieved document the same. Also, a desire for high levels of recall necessitates retrieving many documents. For these reasons it is sensible to examine average recall at both low and high document cut-off values. Accordingly, we considered various ranges of document cut-off values in our statistical analyses: 15—20, 15–25, 40–60, 90–110 and 180–200. Just as with precision, there were significant differences ($p < 0.05$) in average recall performance among search engines for all the cut-off ranges we studied, as calculated by either a standard or ranked (nonparametric) analysis of variance.

With the lenient encoding of 'relevance', AltaVista, Open Text, and Lycos were the top recall performers, in that order. The bottom two were Yahoo! and HotBot. These ranking applied to all document cut-off value ranges (Figs. 5 and 6).

Tukey's highest significant differences revealed the statistical groupings among search engines for different document cut-off value ranges (Tables 9–13). These show that AltaVista, Open Text, Lycos and Excite cluster together at the top (together with other search engines at low document cut-off values) and that Yahoo!, HotBot, Magellan, Excite and Infoseek cluster together at the bottom (together with Lycos and Open Text at document cutoff value 180–

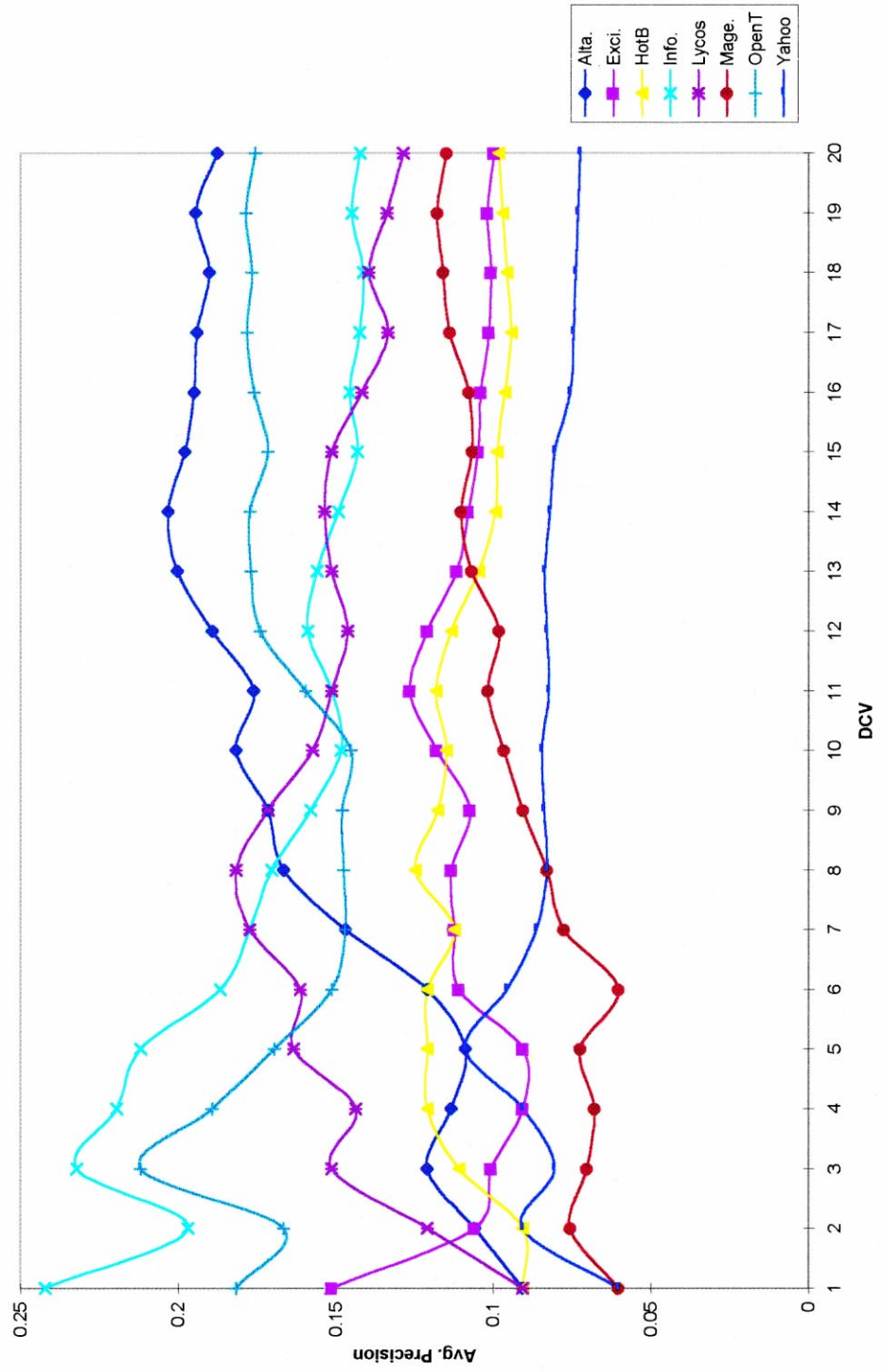

Fig. 3. Average precision vs. DCV (1–20) (strict encoding).

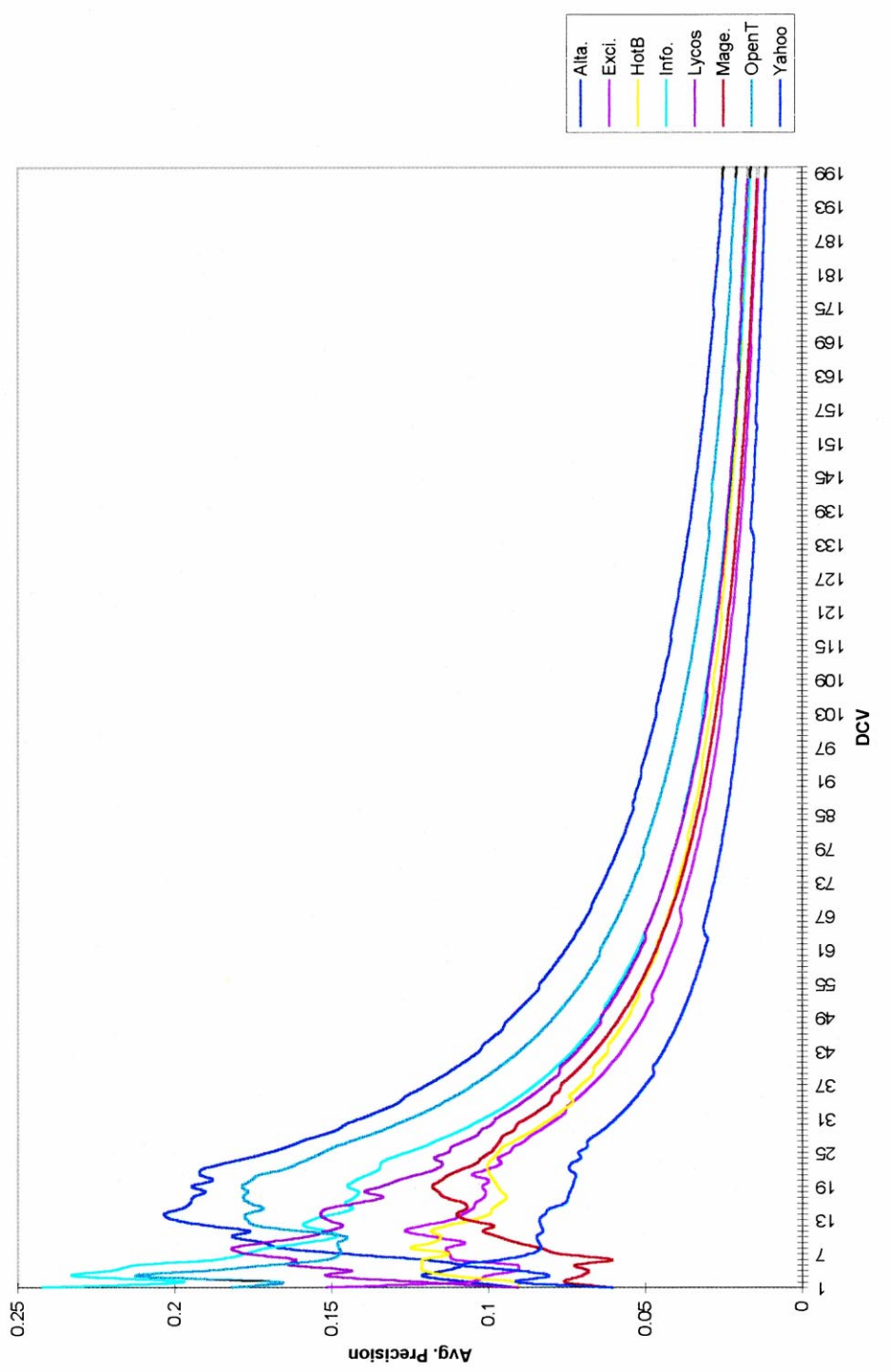

Fig. 4. Average precision vs. DCV (1–200) (strict encoding).

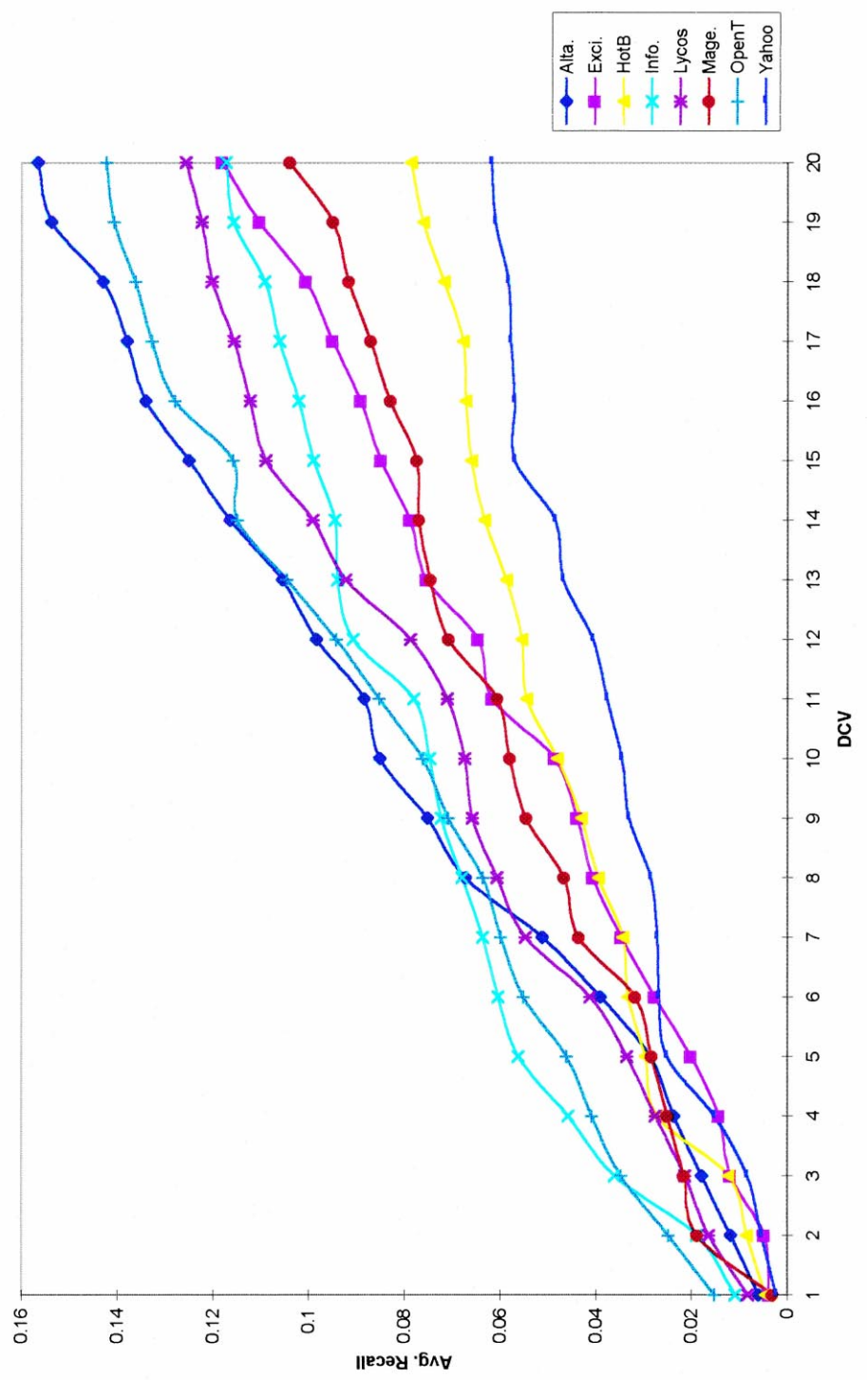

Fig. 5. Average recall vs. DCV (1–20) (lenient encoding).

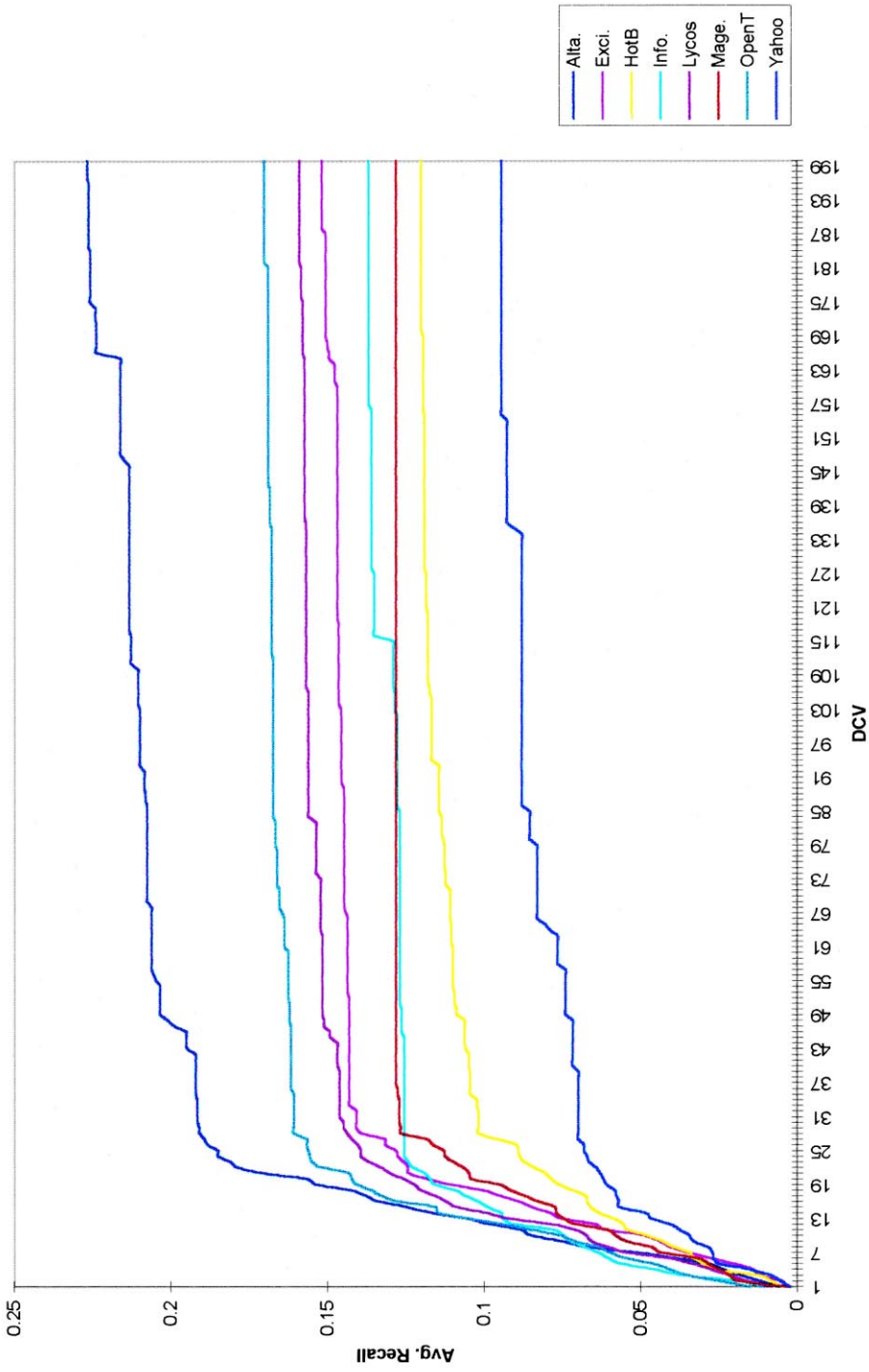

Fig. 6. Average recall vs. DCV (1–200) (lenient encoding).

Table 9
Tukey's highest significant difference for document cut-off values 15–20 based on recall. $\alpha = 0.05$

Search engine	N	Subset		
		1	2	3
Yahoo	33	0.0589		
HotBot	33	0.0713	0.0713	
Magellan	33	0.0898	0.0898	0.0898
Excite	33	0.0998	0.0998	0.0998
Infoseek	33	0.1083	0.1083	0.1083
Lycos	33		0.1176	0.1176
Open Text	33			0.1327
AltaVista	33			0.1418
Sig.		0.134	0.197	0.0940

Values shown are average recalls for document cut-off values 15–20.

200). Table 14 summarizes how these clusters vary in relation to different document cut-off value ranges.

Pairwise for 15–20 documents retrieved, there were statistically significant differences between the mean recall of each of the two top performers, AltaVista and Open Text, and either of the two bottom performers (HotBot and Yahoo!); there was a statistically significant difference between Lycos, the third best performing search engine, and Yahoo!. Pairwise for 15–25 documents retrieved, there was a statistically significant difference in average recall between AltaVista and each of the bottom three performers (Yahoo!, HotBot and Magellan, in order from worst to best); between Open Text and the bottom two; and Lycos and Yahoo!. For a range of 40–60 documents retrieved, there was a statistically significant difference between AltaVista and the four worst performers (Magellan, Infoseek, HotBot and Yahoo!). Both Open Text and Lycos had average recall scores that were statistically greater than Yahoo!'s. For the document cut–off range 90–110, there were the same statistically significant

Table 10
Tukey's highest significant difference for document cut-off values 15–25 based on recall. $\alpha = 0.05$

Search engine	N	Subset			
		1	2	3	4
Yahoo	33	0.0621			
HotBot	33	0.0780	0.0780		
Magellan	33	0.0988	0.0988	0.0988	
Excite	33	0.1117	0.1117	0.1117	0.1117
Infoseek	33	0.1153	0.1153	0.1153	0.1153
Lycos	33		0.1259	0.1259	0.1259
Open Text	33			0.1417	0.1417
AltaVista	33				0.1593
Sig.	33	0.123	0.226	0.363	0.234

Values shown are average recalls for document cut-off values 15–25.

Table 11
Tukey's highest significant difference for document cut-off values 40–60 based on recall. $\alpha = 0.05$

Search engine	N	Subset		
		1	2	3
Yahoo	33	0.736		
HotBot	33	0.1081	0.1081	
Infoseek	33	0.1265	0.1265	
Magellan	33	0.1284	0.1284	
Excite	33	0.1435	0.1435	0.1435
Lycos	33		0.1503	0.1503
Open Text	33		0.1622	0.1622
AltaVista	33			0.2006
Sig.		0.060	0.296	0.228

Values shown are average recalls for document cut-off values 40–60.

Table 12
Tukey's highest significant difference for document cut-off values 90–110 based on recall. $\alpha = 0.05$

Search engine	N	Subset		
		1	2	3
Yahoo	33	0.0880		
HotBot	33	0.1167	0.1167	
Infoseek	33	0.1283	0.1283	
Magellan	33	0.1284	0.1284	
Excite	33	0.1460	0.1460	0.1460
Lycos	33	0.1564	0.1564	0.1564
Open Text	33		0.1674	0.1674
AltaVista	33			0.2099
Sig.		0.097	0.434	0.152

Values shown are average recalls for document cut-off values 90–110.

Table 13
Tukey's highest significant difference for document cut-off values 180–200 based on recall. $\alpha = 0.5$

Search engine	N	Subset	
		1	2
Yahoo	33	0.0946	
HotBot	33	0.1202	
Magellan	33	0.1284	
Infoseek	33	0.1371	
Excite	33	0.1515	0.1515
Lycos	33	0.1591	0.1591
Open Text	33	0.1702	0.1702
AltaVista	33		0.2262
Sig.	33	0.052	0.058

Values shown are average recalls for document cut-off values 180–200.

Table 14
Significant differences between average recall

Yahoo	a_1			b_1				c_1			d_1			e_1	
HotBot	a_1	a_2		b_1	b_2			c_1	c_2		d_1	d_2		e_1	
Magellan	a_1	a_2	a_3	b_1	b_2	b_3		c_1	c_2		d_1	d_2		e_1	
Excite	a_1	a_2	a_3	b_1	b_2	b_3	b_4	c_1	c_2	c_4	d_1	d_2	d_3	e_1	e_2
Infoseek	a_1	a_2	a_3	b_1	b_2	b_3	b_4	c_1	c_2		d_1	d_2		e_1	
Lycos		a_2	a_3		b_2	b_3	b_4		c_2	c_4	d_1	d_2	d_3	e_1	e_2
Open Text			a_3			b_3	b_4		c_2	c_4		d_2	d_3	e_1	e_2
AltaVista			a_3				b_4			c_4			d_3		e_2

Legend

d.c.v. 15-20	d.c.v. 15-25	d.c.v. 40-60	d.c.v. 90-100	d.c.v 180-200
a_1 = low	b_1 = low	c_1 = low	d_1 = low	e_1
a_2 = medium	b_2 = low-medium	c_2 = medium	d_2 = medium	e_2
a_3 = high	b_3 = medium-high	c_3 = high	d_3 = high	
	b_4 = high			

This table shows which search engines 'cluster' together according to Tukey's highest significant differences for different document cut-off value ranges (d.c.v.'s). 'Low' indicates lowest recall and 'high' indicates highest.

differences among average recall levels as for the range 40–60, except Lycos was no longer statistically significantly better than Yahoo!. And pairwise for 180–200 documents retrieved, AltaVista was significantly better than Yahoo!, HotBot, Magellan and Infoseek. Pairwise differences are summarized in Table 15.

Table 15
Pairwise average recall comparison

	Infoseek	Excite	Magellan	HotBot	Yahoo!
AltaVista	c,d,e		b,c,d,e	a,b,c,d,e	a,b,c,d,e
Open Text				a,b	a,b,c,d
Lycos					a,b,c

Legend

label	a	b	c	d	e
d.c.v. range	15-20	15-25	40-60	90-110	180-200

This table shows the document cut-off values for which the search engine in the column at the left of the table had a statistically significant difference (improvement) in average recall compared to the search engine named in the row across the top. Tukey's test ($\alpha = 0.05$) was used for these calculations, and there were no other statistically significant differences among means.

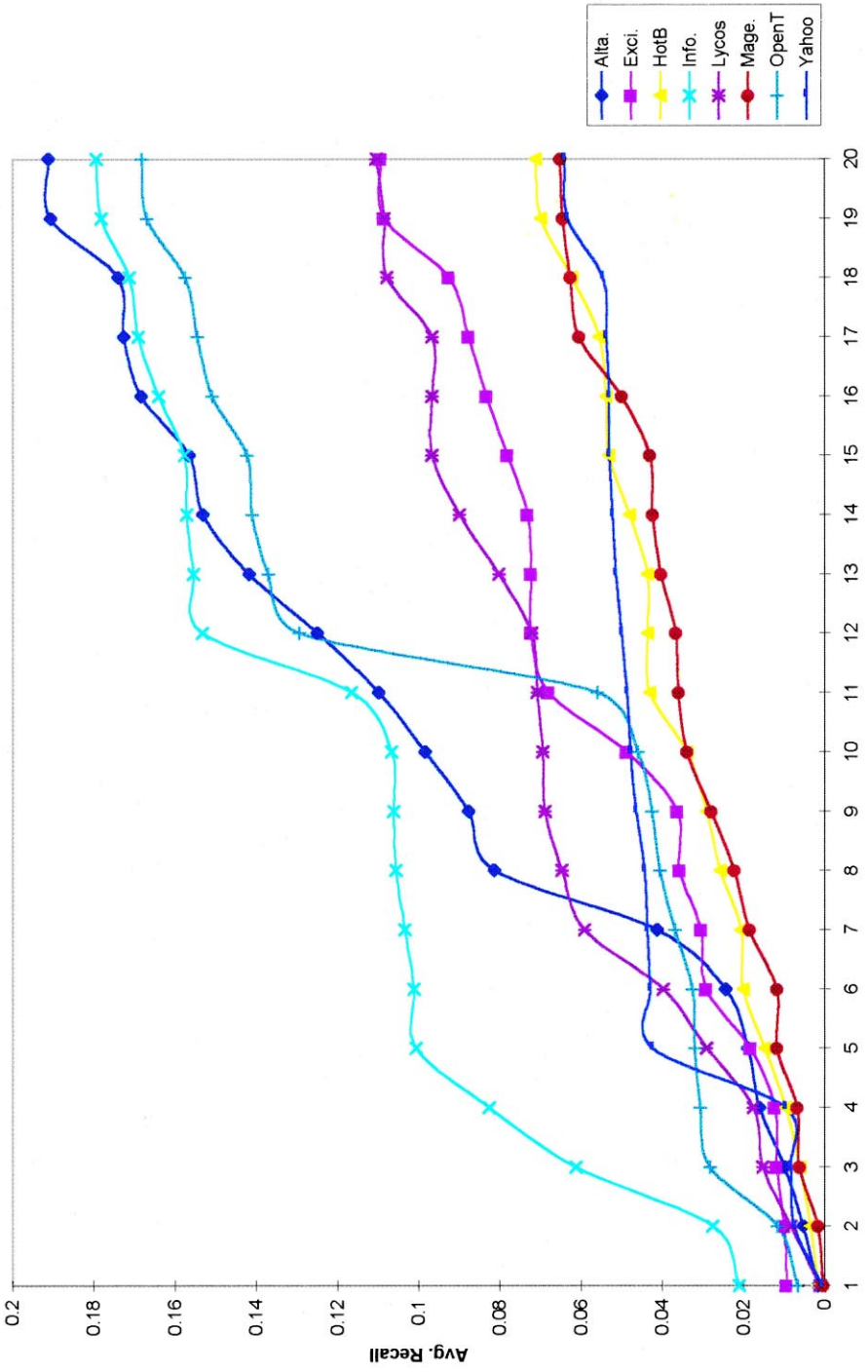

Fig. 7. Average recall vs. DCV (1–20) (strict encoding).

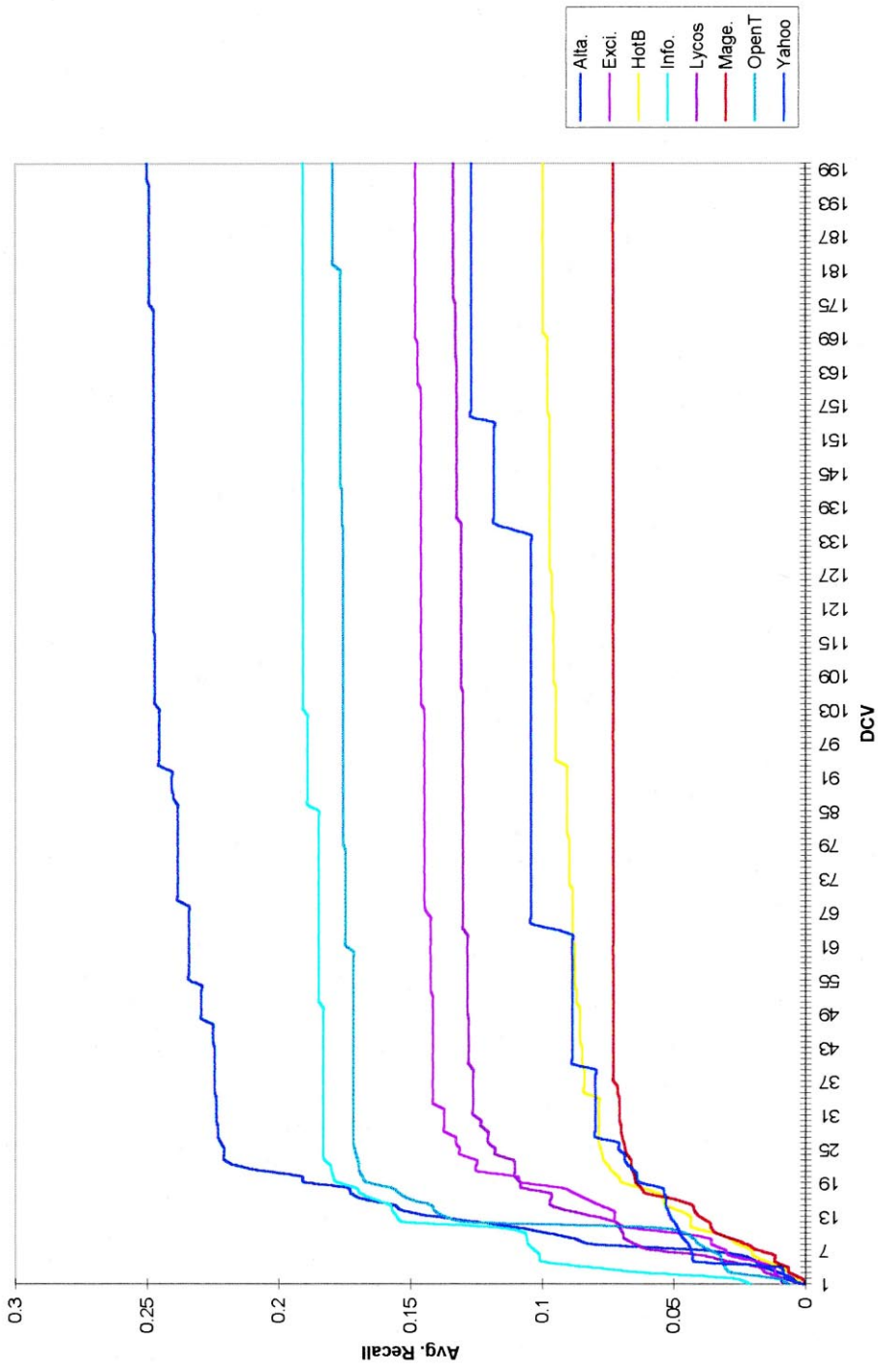

Fig. 8. Average recall vs. DCV (1–200) (strict encoding).

Overall, AltaVista and Open Text were the best performers and Yahoo! was the worst. One of the reasons for Yahoo!'s consistently poor performance may have been that the pages it retrieved were often, themselves, simply lists of URLs, which make it very difficult for faculty evaluators to determine if what they were viewing was relevant.

For recall results calculated with a *strict* encoding for relevance, two interesting observations can be made. First average recall is higher than for the more lenient encoding. This suggests that retrieval algorithms may be better suited to finding highly relevant than marginally relevant documents. In fact, Infoseek (a middle of the pack performer for a lenient encoding of relevance) is a top performer for highly relevant documents, finding many of them early and being overtaken in recall performance by AltaVista only after 18 documents are retrieved. A possible explanation of Infoseek's improved performance relative to the other search engines is that it is tuned to identify highly relevant documents, even at the expense of marginally relevant ones. Second, Yahoo!, the worst performer by all other measures, performs creditably through document cut-off value 10 (Figs. 7 and 8).

5.1.3. User differences

Although it was not a question that motivated this research, we checked to see if there were individual differences among the 'queries' (actually information needs) that faculty subjects submitted. For average relevance, there was no statistical difference among queries according to either a standard or ranked analysis of variance at any document cut-off value. For precision calculations, there were statistically significant differences among them.

By ten retrieved documents per search engine, there was a spread in the number of relevant documents a faculty member received varying from 2 up to 65 (of a possible total of 80), with an average of 23.2. By twenty retrieved documents, the range was 5 to 125 (of a possible 160), with an average of 42.2. The faculty receiving the best results had information needs that appeared quite different from each other. In some cases, they used highly specific vocabulary that the searcher exploited (such as *gage repeatability and reproducibility*; *sequential analysis of variance*; *regulatory takings*, *average reciprocity of advantage*; or *noxious use test*). But in others, the queries appeared far more general with respect to vocabulary, with the key vocabulary including phrases like *mediation or arbitration*; *existing businesses that enter similar*, *new businesses*; or *success factors for the chief information officer*.

Informal comments were sometimes supplied by users after they received their Web booklets. These ranged from "I was very disappointed by the results of the search procedures... Nothing of real value appeared" to "What a gift you have given me with this Web search" or "These three volumes [booklets of Web pages] are incredibly useful...". Other faculty remarked on issues like why, after the fact, they thought the information need they had presented was particularly difficult ("not a lot of work has been done on the topic I asked for. But that is why I'm doing research on it") or how difficult it was to distinguish 'relevance' from 'usefulness'.

5.2. Overlap study

Different search engines have different ways of 'collecting' the Web pages that comprise their databases. Most use so-called 'spiders' that traverse the Web, link to link, in search of newly

added Web pages. In addition, various ways exist for Web authors to register their pages, and then each search engine employs different search algorithms for retrieval.

As we have mentioned, popular search engine indexes differ in size, indexing as few as 5 million or more than 50 million pages. But no search engine indexes the entire Web; even large search engines exclude well over half of it.

In addition to differing in how comprehensive their indexes are (and, so, in terms of the universe of documents from which one can retrieve), search engines differ in terms of how up-to-date their indexes are. Some of these differences surround submission policies for new Web pages. Most search engines allow one to submit the home page for a Web site. A submitted page may then be added to a search engine's index within minutes or, at the other extremes, several weeks later. Pages *linked* to submitted pages are ordinarily not added right away to a search engine's index. Instead, depending on the schedule governing a search engine's spider, which will search from already indexed pages, linked pages may be detected within a few weeks or a few months (so current events are usually days to weeks or months out of date). Policies again differ from search engine to search engine concerning when such a newly detected page will be added to a search engine's index. (This can occur daily for some engines, or with a delay of up to a month for others.) Differences exist, too, concerning the 'depth' to which spiders will hunt for new pages. Some search engines attempt to find all linked pages; others select all the linked pages of all popular sites (measured by the number of incoming links), but will exclude certain pages linked to less popular sites (either by excluding pages linked too many 'jumps' from the home page, or by including linked pages on a sampling basis).

Since the Web is so dynamic in terms of the new pages that are being published (and also deleted), spiders revisit already indexed pages in an attempt to keep their indexes 'fresh'. In fact, a given search engine can be thought to have both a 'freshness spider', which visits perhaps the engine's 2 million top pages once a week, and a 'completeness spider' that visits the rest of the pages it indexes approximately once a month. The most frequently visited sites include those that have the most incoming links as well as those that change their contents and links frequently (occurrences that search engines are able to learn about).

In sum, freshness varies considerably from engine to engine. A search engine that devotes considerable resources to keeping its index fresh and 'crawls' 10 million pages a day can take almost a week to visit all its indexed sites. Even within the same search engine, freshness may vary from minutes to several months, mostly as a function of how popular a web site is considered to be.

Because search engines differ with respect to both their comprehensiveness and their freshness, it is difficult to predict the amount of overlap among the Web pages returned by different search engines processing the same information need. Thus, based on the searches conducted for the last study, we computed the degree of overlap (a) among the documents retrieved by all eight search engines; and (b) among the documents retrieved and judged relevant for all eight search engines. Measurements were taken for different document cut-off values; and the coding for 'relevant' was done twice — with both the lenient and strict codings.

We measured overlap by calculating the empirical distribution of the number of search engines that retrieved a given web page, calculated separately for each faculty participant's information need, and then averaged across them.

The average empirical distributions for document cut-off values of 20, 50, 100 and 200 retrieved documents are shown in Table 16. The data show how surprisingly little overlap there

Table 16
Overlap among retrieved documents

Frequency	d.c.v. 20	d.c.v. 50	d.c.v. 100	d.c.v. 200
1	147.65	365.85	713.75	1318.00
2	4.45	9.85	19.05	39.00
3	0.40	0.75	2.10	3.60
4	0.10	0.20	0.25	0.35
5				0.15

This table shows the average number of times a Web page was retrieved by one or more search engines.

Table 17
Overlap among relevant retrieved documents

	Document cut-off value							
	lenient				strict			
Frequency	20	50	100	200	20	50	100	200
1	35.95	40.55	36.65	38.45	17.25	19.45	18.95	18.35
2	2.30	3.70	4.50	5.85	0.95	1.60	2.00	2.80
3	0.15	0.15	0.20	0.25	0.10	0.15	0.35	0.35
4	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.15
5	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.05

This table shows the average number of times a relevant Web page was retrieved by one or more search engines. 'Lenient' means that an evaluator judged a document to be either highly relevant or somewhat relevant; 'strict' means it was judged to be the former.

is among search engines. For instance, nearly 150 of the top 160 documents[12] (8 search engines; top 20 documents per search engine) were retrieved by exactly one of the eight search engines. This percentage of approximately 93% documents being retrieved by a single search engine was practically constant across all document cut-off values.

Similar distributions were computed to determine the degree of overlap among *relevant*-retrieved documents. The results for a lenient and strict encoding of relevance are shown in Table 17. On a percentage basis, the data show that there is a bit stronger overlap among relevant retrieved documents than among all retrieved documents (whether relevant or not). Nevertheless, absolute levels are remarkably low.

The finding of lack of overlap confirms a suggestion made in the IR literature by Katzer, McGill, Tessier, Frakes, & DasGupta (1982) among others: even when retrieval performance is approximately the same, different IR systems often retrieve quite different (relevant) documents (even with a document database that is the same across IR systems). It is consistent, too, with

[12] Actually, the number was a bit less than 160, on average. Yahoo! often returned fewer than 20 documents per query. Other search engines sometimes returned fewer than 200, sometimes even fewer than 50 for certain queries.

a study reported in *Science* suggesting that individual search engines cover from 3 to 34% of the Web (Lawrence & Giles, 1998). As a practical matter, the lack of overlap among search engines suggests that searchers should use many search engines in attempting any comprehensive literature search.

6. Discussion

Changes in search engines are almost continual. These affect their interfaces, indexing policies, and their retrieval algorithms. On the other hand, a wide body of literature on information retrieval research has been available to search engine developers from the outset of developing search engines, and much has made its way into practical use. Further, this literature generally suggests that no dramatic performance improvements in retrieval will occur any time soon; and there are practical considerations of retrieval speed and disk storage that could stand in the way of implementing more effective search algorithms if they were possible (Lesk, 1997).

Thus, this paper presents results that are likely to be a fairly accurate picture of the effectiveness of Web searching for some time — even if the Web search engine industry undergoes a 'shakeout' that reduces the number of major search engine providers, as is likely. Thus, these results are best viewed as statements about currently and, likely, future retrieval effectiveness (at least for some time) — rather than as a definitive statement about the best- and worst-performing search engines.

At risk of over simplification, a few comments can summarize the retrieval effectiveness of the search engines examined. First, absolute retrieval effectiveness is fairly low. The modal number of relevant documents retrieved per search engine at document cut-off value 10 is one. Approximately half of the searches returned just one relevant document. A great majority of the searches returned five or fewer relevant documents, though two returned ten (Fig. 9). By a document cut-off value of twenty retrieved documents, most searches were returning ten or fewer relevant documents per search engine (with one or two being the most frequent), with two still being near perfect in terms of precision (Fig. 10). By 200 retrieved documents, four or fewer relevant retrieved documents per search engine was still the norm, though the most effective searches retrieved more than 20 (Fig. 11). These rather low levels of performance came about despite using highly trained searchers who used every possible retrieval strategy to coax optimal performance from each search engine.

Second there are statistical differences among search engines' retrieval- and precision-effectiveness at all document cut-off values. When extended to pairwise comparisons, AltaVista and Open Text are the best performers, with Yahoo! being the worst and the other search engines in the middle. A separate question is whether these differences are of practical import. A difference in precision of 40% vs. 20% means two to four additional relevant documents will typically be included in the list of ten to twenty top-ranked items returned by a search engine.

Third, there were no statistical differences in the retrieval effectiveness among users for recall, though there were for precision. As a practical matter, a handful of searches had precision still at or close to 100% by twenty retrieved documents (for selected search engines)— though it is hard to know whether to attribute such outstanding performance to a search algorithm, a particular search topic, or faculty evaluator's standards.

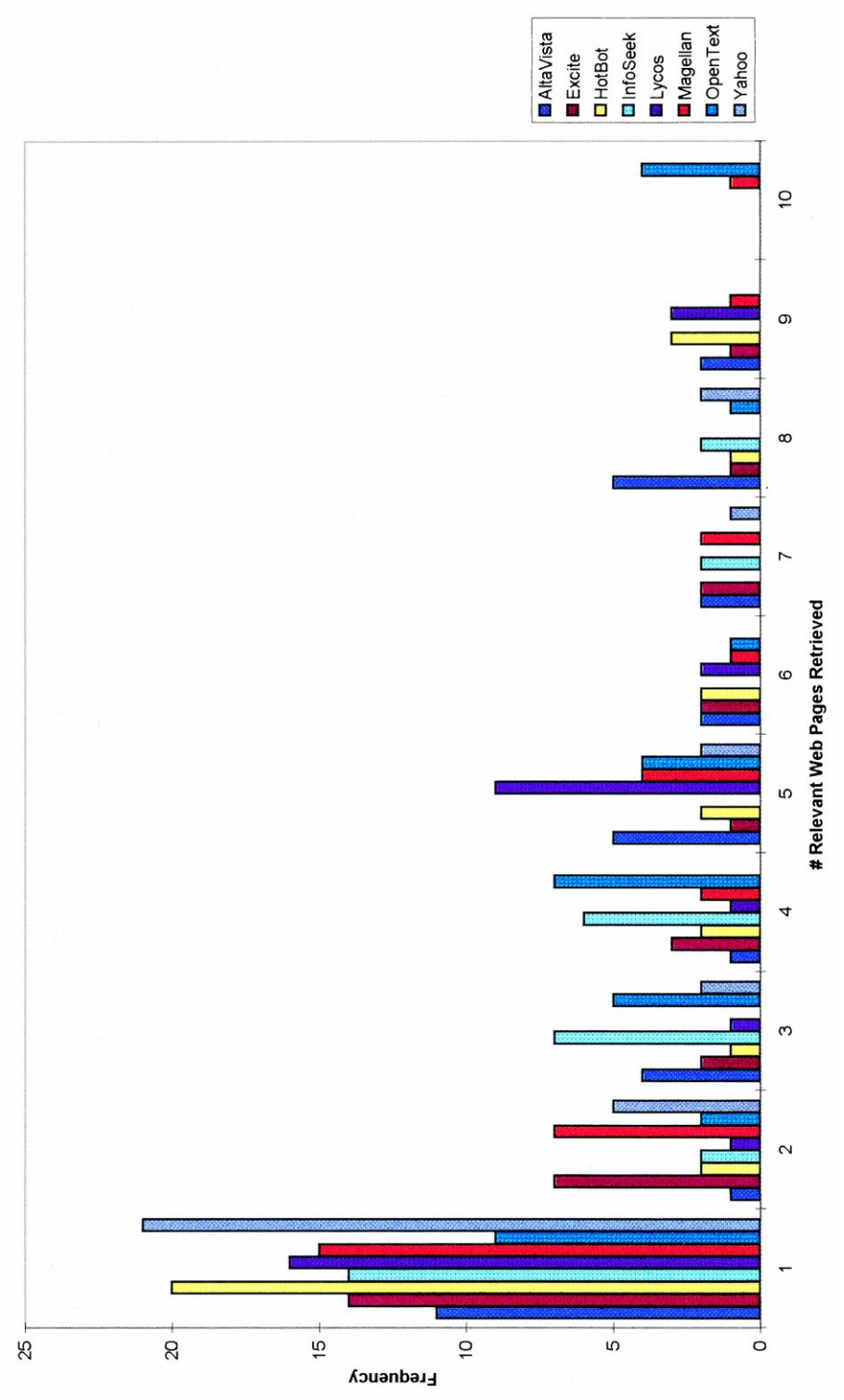

Fig. 9. Frequency vs. number of relevant Web pages retrieved DCV 10.

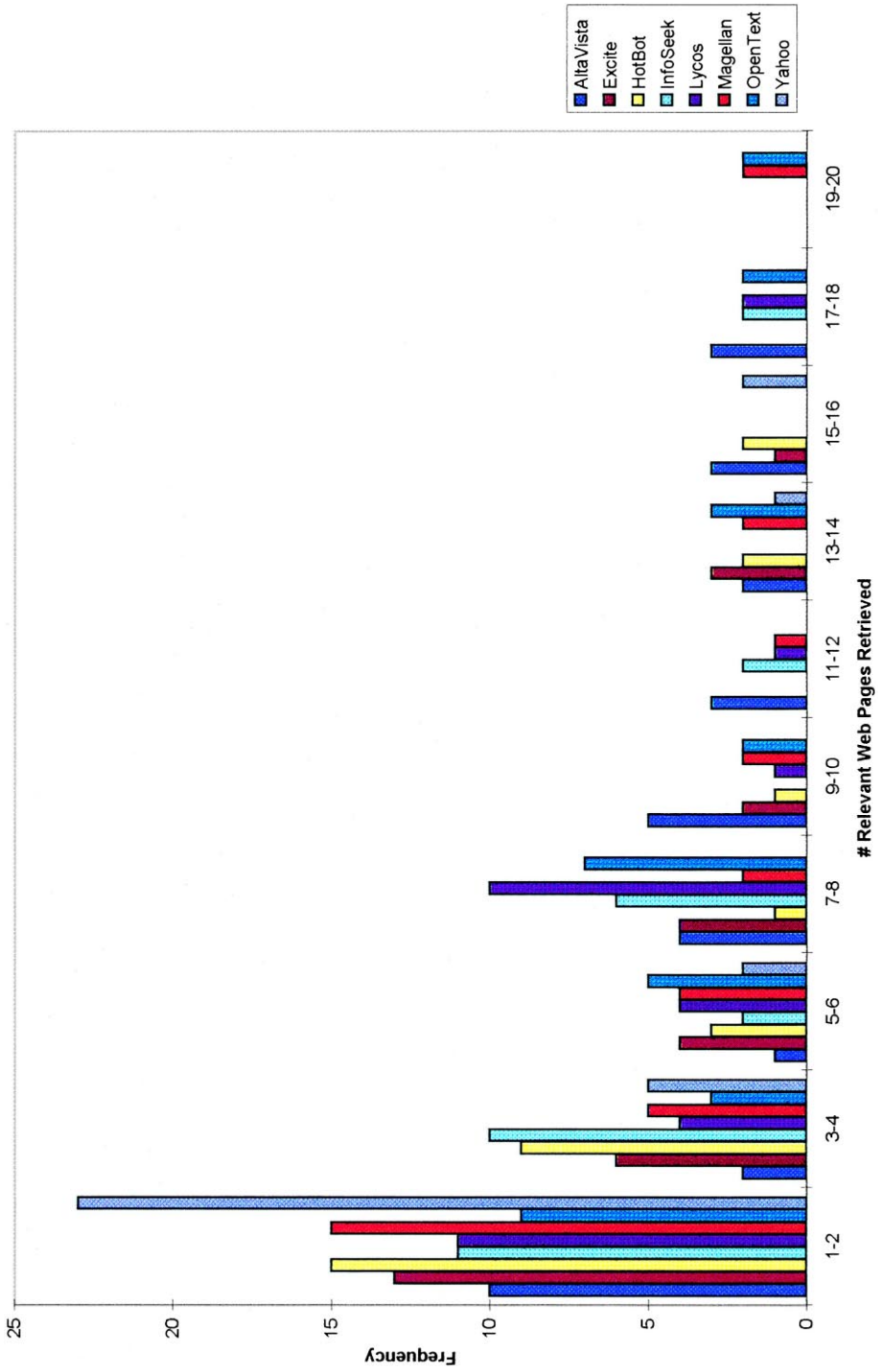

Fig. 10. Frequency vs. number of relevant Web pages retrieved DCV 20.

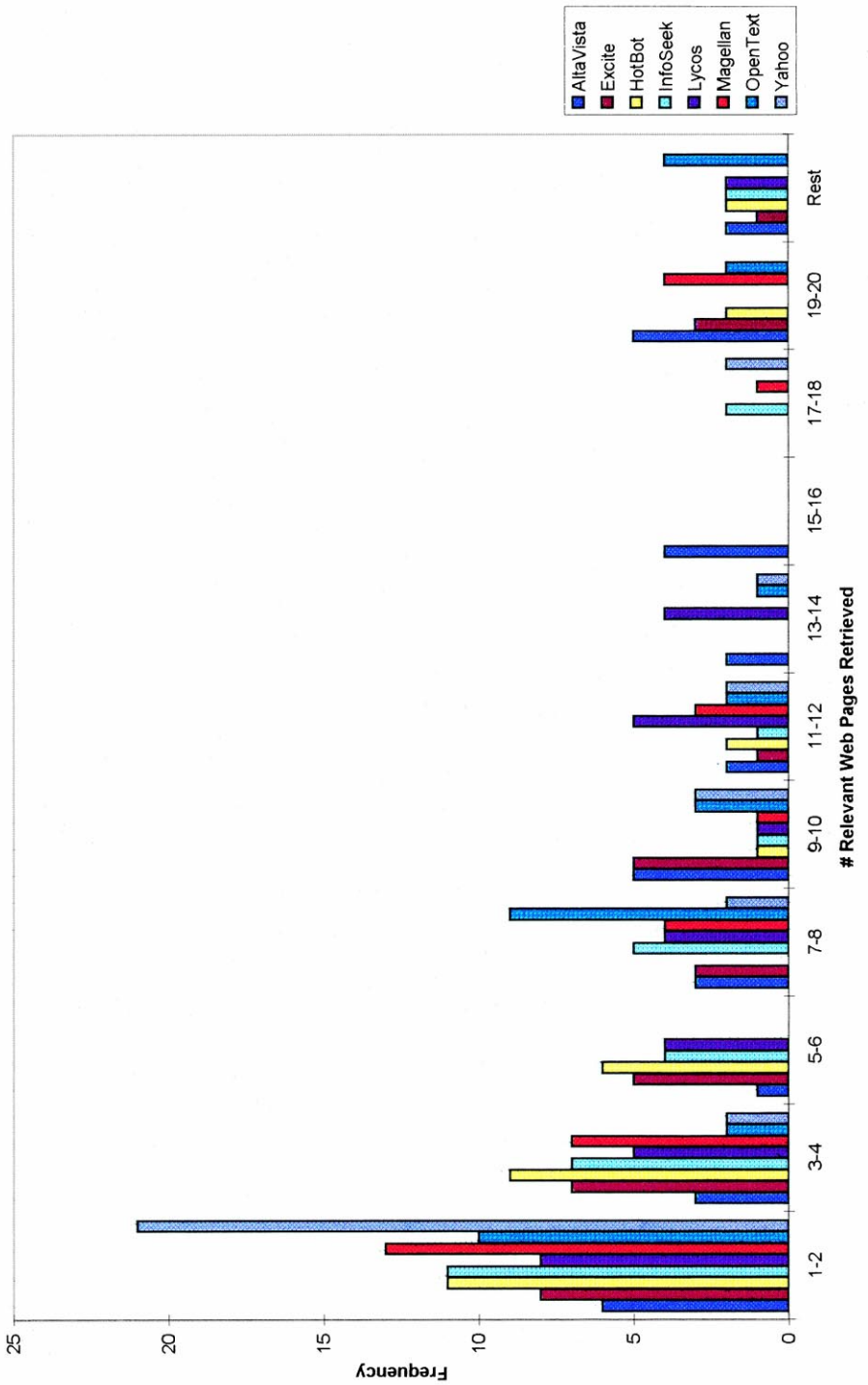

Fig. 11. Frequency vs. number of relevant Web pages retrieved DCV 200.

Information Request (Provided by Faculty Member)

The topic is formal language communication. This comes under various names, including 'agent communication languages' or 'agent communication protocols' or 'electronic data interchange' etc. I would like to see definitions of these languages. There may be other languages (or protocols) that don't come under the above names ---- I would also be interested in those. I'm particularly interested in any languages that have a linguistic foundation (i.e. philosophy of language, speech act theory, etc.).

Queries issued:

An expert search translated the above request (plus supplementary information) into eight queries — each being an attempt to achieve optimal retrieval performance for a different search engine. These queries follow:

AltaVista
("agent communication*" AND (protocol* OR language*)) OR "electronic data interchange" OR "speech act theor*"

Excite
("agent communication" AND (protocol OR language)) OR "electronic data interchange" OR "speech act theory"

HotBot
("agent communication" AND (protocol OR language)) OR "electronic data interchange" OR "speech act theory"

Options: Look for: Boolean phrase; Date: anytime; Location: anywhere; Page depth: top page

Infoseek
In the options of the Advanced Search the following were specified:
Document should contain the phrase: "agent communication"
Document should contain the words: protocol language
Document should contain the phrase: "electronic data interchange"

Lycos
("agent communication" AND (protocol OR language)) OR "electronic data interchange" OR "speech act theory"

Magellan
("agent communication" AND (protocol OR language)) OR "electronic data interchange" OR "speech act theory"

OpenText
("agent communication" AND (protocol OR language)) OR "electronic data interchange" OR "speech act theory"

Yahoo
("agent communication*" AND (protocol* OR language*)) OR "electronic data interchange" OR "speech act theor*"

Fig. 12. Optimal queries for different search engines. The strong similarity among the eight 'optimal' queries above was exhibited for most of the remaining 32 information requests.

Fourth, searching performance appears more strongly related to the matching function built into a search engine than to the type of queries it allows. For instance, Fig. 12 shows the eight queries that were optimized for a particular faculty subject — each optimized for a different search engine. For this faculty member's information need, as well as for most of the other 32, there was a strong similarity among optimal queries across search engines.

Specialized search engines may behave somewhat differently. These services provide access to special topics as diverse as health and medicine, sports, law, hobbies, politics, travel and cars[13]. Though they operate chiefly by following the indexing and retrieval principles we have outlined, their performance will likely vary from what we have described by retrieving far few totally irrelevant items, and these tools may often be used by searchers who may want to receive rather comprehensive information about a subject.

Some reservations may be raised about the generalizability of the current study. For instance, not all searches are as well defined as those in our study, and many searchers need only a few relevant Web pages. Furthermore, the unit of retrieval in our experiment was the individual Web page (up to 20 printed pages), and in some cases the page a faculty member evaluated may have contained many pointers (URLs) to highly relevant pages without the faculty member realizing it (or, at least, without being able to confidently assess the presented page as relevant). Despite these reservations, this study has answered in scientific way the question of searching effectiveness on the Web. As the Web grows, as more users begin to use it routinely, and as it begins to provide more and different services, we cannot take for granted that information retrieval from it will be successful. Instead, understanding the limits of retrieval effectiveness becomes an increasingly important issue.

The lessons from this study have additional applicability beyond using search engines to search the Web. In the first place, 'marriages' and licensing agreements often mean that the search engines described in this study are used in other Internet search contexts. For example, search.com contains a variety of search tools and subject directories used for general and special purpose searching. Searchers choose their preferred search engine from a list of nine options. Similarly, America Online uses a search engine that is actually a version of Excite with a modified interface. WebTV 'channel surfers' also are using Excite when they press the 'search' button to locate Web pages.

The overlap results we have produced support the idea of using meta search engines (see, for example, (Selberg & Etzioni, 1997)). Meta search engines allow users to issue a single query, which is then sent to various search engines before the meta search engine aggregates the URLs returned and presents the user with a unified list of them. However, meta search engines often fail to retrieve all documents formed by the union of the documents returned by the underlying individual search engines; and they often use a limited query syntax. So-called search managers operate similarly, usually with a greater emphasis on exploiting the different syntactic features of the different search engines they use or on organizing the content of the pages returned. Given the very small overlap among the pages search engines retrieve, such utilities may by quite useful for searchers needing rather comprehensive information on a

[13] A brief article in the Chronicle of Higher Education (1996) states that most scholars agree that search engines don't produce much of value for academic research — the topic of the current research. The article then gives an example of a specialty search engine supporting research on the ancient world.

topic. Of course, the challenge for meta search engines and search managers is to take advantage of the disparate results of different search engines.

Internet-based 'push' services (such as PointCast) operate by sending information to users based on standing profiles of their information needs — without the need for a query. The major browsers are also incorporating push services to complement their more customary use. The indexing and search capabilities underlying push services are being developed in alliances with search engine companies, suggesting that the results we have reported here may approximately apply when information is pushed.

Even within companies — and outside the Internet — search engines are in wide use. *Intranets* adopt Internet technologies such as TCP/IP and HTTP so that businesses and other organizations can privately publish Web pages and allow retrieval capabilities for themselves (or even customers and others). Search engines are essential to the success of these endeavors. In fact, the business models of many of the most popular search engine companies depend on licensing their search engines for private use, more than on individuals using the engine to search the Internet.

In summary, search engines are essential to the success of the Web. To some they might appear to fit the description Shakespeare applies to Gratiano:

> Gratiano speaks an infinite deal of nothing, more than any man in all Venice. His reasons are as two grains of wheat hid in two bushels of chaff: you shall seek all day ere you find them, and when you have them, they are not worth the search.

But, the truth is probably closer to John Dryden's words from the seventeenth century — especially for those who use search engines skillfully and with some persistence:

> Errors, like straws, upon the surface flow; He who would search for pearls must dive below[14].

Acknowledgements

Many people supported this research. Our thanks to Scott Moore and George Widmeyer for many relevant conversations. Thanks to Judy Roper and Terri Bell for providing administrative and personnel support. Thanks to Michelle Betts, Melissa Paul and David Cole who worked as research assistants on this projects. And thanks to the searchers (including Candace White) who tirelessly searched the Web. Finally, thanks to the University of Michigan Business School for its research support[15].

[14] Quotes from Shakespeare's *The Merchant of Venice* and John Dryden's 'All For Love. Prologue'. Quotes located on the Web using Bartlett's Familiar Quotations. http://www.columbia.edu/acis/bartleby/bartlett/.

[15] Web pages occasionally require one to infer a creation date (from the dates of other pages the page references). Also, the dates of Web pages listed above are *creation* dates, rather than last modified dates. When a reference has different dates (years) for its print and electronic versions, the date of the print version is indicated.

References

Calafia (1997). Search engine watch. http://searchenginewatch.com/.

Cherry, C. (1980). *On Human Communication: a review, a survey and a criticism* (3rd ed.). Cambridge, MA: The MIT Press.

Chronicle of higher education (1996). Oct. 18, A23.

Chu, H., & Rosenthal, M. (1996). Search engines for the world wide web: A Comparative study and evaluation methodology. *ASIS 1996 Annual Conference Proceedings*, Baltimore, MD. http://www.asis.org/annual-96/ElectronicProceedings/chu.html.

CommerceNet/Nielsen (1997). CommerceNet/Nielsen Internet demographics study. http://www.commerce.net/nielsen/index.html.

Courtois, M. (1996). Cool tools for web searching: an update. *Online, May/June, 1996,* 29–36.

Ding, W., & Marchionini, G. (1996). A comparative study of web search service performance. *ASIS 1996 Annual Conference Proceedings*, Baltimore, MD.

Feldman, S. (1997). Just the answers, please. Choosing a web search service. Information Today, Inc. http://www.infotoday.com/searcher/may/story3.htm.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. *ACM SIGIR* 1993, Pittsburgh, PA.

Internet domain survey (1997). http://nw.com/zone/WWW/top.html.

Katzer, J., McGill, M. J., Tessier, J. A., Frakes, W., & DasGupta, P. (1982). A study of the overlap among document representations. *Information Technology: Research and Development, 1,* 261–274.

Lake, M. (1997). 2nd Annual search engine shoot-out. *PC Computing* http://www4.zdnet.com/pccomp/features/exc10997/sear/sear.html.

Lawrence, S., & Giles, C. L. (1998). Searching the World Wide Web. *Science, 5360*(280), 98–100.

Lebedev, A. (1997). Best search engines for finding scientific information in the Web. http://www.chem.msu.su /eng/comparison.html.

Leighton, H. V. (1995). Performance of four World Wide Web (WWW) index services: Infoseek, Lycos, WebCrawler and WWWWorm. http://www.winona.msus.edu/is-f/library-f/webind.htm [revised 1 July 1996]; or http://www.hbz-nrw.de/schmidt/kurs/webind.htm.

Leighton, & Srivastava (1997). Precision among World Wide Web search services (search engines): AltaVista, Excite, HotBot, Infoseek, Lycos. http://www.winona.msus.edu/is-f/library-f/webind2/webind2.htm.

Lesk, M. (1997). Industrial searching panel at SIGIR 1997.

Morgan, C. (1996). The search is on — finding the right tools and using them properly can shed light on your web search efforts. *Windows Magazine*, November 01, 1996, Issue 711, Feature Section. http://www.techweb.com/se/directlink.cgi?WIN19961101S0130.

Morville, P., Rosenfeld, L., & Janes, J. (1996). *The internet searcher's handbook*. New York: Neal-Schuman Publishers, Inc.

Overton, R. (1996). Search engines get faster and faster, but not always better. (September 1996 issue of PC World). http://www.pcworld.com/workstyles/online/articles/sep96/1409_engine.html.

Salton, G. (1992). The state of retrieval system evaluation. *Information Processing and Management, 28*(4), 441–449.

Selberg, E., & Etzioni, O. (1997). The MetaCrawler architecture for resource aggregation on the Web. *IEEE Expert, 12*(1), 8–14.

Schlichting, C., & Nilsen, E. (1996). Signal detection analysis of WWW search engines. (no creation date available; document includes 1996 references). http://www.microsoft.com/usability/webconf/schlichting/schlichting.htm.

Slot, M. (1997). The matrix of internet catalogs and search engines. http://www.ambrosiasw.com/~fprefect/matrix/.

Sparck Jones, K. (1997). Summary performance comparisons: TREC-2, TREC-3, TREC-4, TREC-5. http://potomac.ncsl.nist.gov.TREC/trec5.papers/sparckjones.ps.

Steinberg, S. G. (1996). Seek and ye shall find (maybe). *Wired*, 4(05), 108 ff.

Tomaiuolo, N. G., & Packer, J. G. (1996). An analysis of Internet search engines: assessment of over 200 search queries. *Computers in Libraries, 16*(6), 58–62.

Vaughn-Nichols, S. J. (1997). Find it faster. *Internet World, 8*(6), 64–66.

Westera, G. (1996). Robot-driven search engine evaluation overview. http://www.curtin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy/index:htm.

van Rijsbergen, 1983. London: Butterworths. Information Retrieval, 2nd edition.