Information Retrieval, 4, 33–59, 2001

Measuring Search Engine Quality

DAVID HAWKING David.Hawking@cmis.csiro.au
NICK CRASWELL
CSIRO Mathematical and Information Sciences, GPO Box 664, Canberra Australia 2601

PETER BAILEY
Computer Science Department, Australian National University, Canberra, Australia 0200

KATHLEEN GRIFFITHS
Centre for Mental Health Research, Australian National University, Canberra, Australia 0200

Received July 6, 2000; Accepted November 28, 2000

Abstract. The effectiveness of twenty public search engines is evaluated using TREC-inspired methods and a set of 54 queries taken from real Web search logs. The World Wide Web is taken as the test collection and a combination of crawler and text retrieval system is evaluated. The engines are compared on a range of measures derivable from binary relevance judgments of the first seven live results returned. Statistical testing reveals a significant difference between engines and high intercorrelations between measures. Surprisingly, given the dynamic nature of the Web and the time elapsed, there is also a high correlation between results of this study and a previous study by Gordon and Pathak. For nearly all engines, there is a gradual decline in precision at increasing cutoff after some initial fluctuation. Performance of the engines as a group is found to be inferior to the group of participants in the TREC-8 Large Web task, although the best engines approach the median of those systems. Shortcomings of current Web search evaluation methodology are identified and recommendations are made for future improvements. In particular, the present study and its predecessors deal with queries which are assumed to derive from a need to find a selection of documents relevant to a topic. By contrast, real Web search reflects a range of other information need types which require different judging and different measures.

Keywords: Web search, search engines, evaluation

1. Introduction

Publication of accurate and meaningful evaluation of the quality of results returned by public Web search engines not only enables informed consumer choice but also assists and encourages search engine operators to improve their standard of service. Since many search engines claim to be using novel techniques, effectiveness comparisons between these engines and systems employing published methods are potentially of interest to the Information Retrieval (IR) research community.

It is known that some search engine operators already perform extensive internal evaluations, but it is not clear that these evaluations have the properties (blindness, independence, reproducibility) which would normally be expected of published studies in IR. On the other hand, it is possible that the model of the user retrieval task implicit in these evaluations more

accurately represents everyday Web search than do the models assumed in much published IR work.

Countless evaluations of text retrieval systems eg. (Salton and Lesk 1997, Voorhees and Harman 1998) have measured the effectiveness of relevance ranking schemes in the context of a well-defined static document collection for which sufficiently complete relevance information is available. Such evaluations are reproducible and can provide accurate measurements of recall (the proportion of all relevant documents which have been retrieved so far) as well as precision (the proportion of retrieved documents which are relevant).

Early test collections, such as those described in Cleverdon (1997) and Salton and Lesk (1997) were small enough to permit relevance judgments for every document against every request. Exhaustive judging is infeasible with collections the size of TREC (Voorhees and Harman 1998) and "almost complete" judgments are collected by assessing only the pool of documents retrieved by a large and diverse set of retrieval methods. All unjudged documents are assumed to be irrelevant.

TREC is subject to continual evolution and now models some specific features of Web search. For example, in TREC-8, the Large Web Task involved blind effectiveness evaluations of retrieval systems operating over 18.5 million pages crawled from the Web (Hawking et al. 1999). Each participant was required to process 10,000 queries taken from Web search engine logs and return the top 20 ranked documents for each. Approximately 50 queries were selected post hoc for relevance assessment using binary measures. The large number of queries and the size of the data set served to rule out the use of unrealistically inefficient algorithms.

Here, we report on the application of an extended TREC-8 Large Web task methodology to the effectiveness evaluation of 20 public search engines. When evaluating public search engines without the knowledge or cooperation of their operators, there is no fixed test collection and it is consequently necessary to evaluate crawling and document retrieval algorithms in combination.

Finally, we present an analysis of the extent to which characteristics of Web search have been captured by this experiment and argue for further methodological developments both to improve the applicability of evaluation results and to reduce the cost of obtaining them.

2. Relationship to other studies

Gordon and Pathak (1999) distinguish between two types of search engine evaluation: *testimonials*, encompassing informal and impressionistic appraisals and feature-list comparisons; and *shootouts*, which correspond more closely to traditional IR effectiveness experiments. Here, only the latter are considered.

Gordon and Pathak present a table[1] of twelve earlier shootout studies but identify only three (including their own) which make use of "appropriate experimental design and evaluation." Of these, that of Gordon and Pathak is the most comprehensive and most recent. Ding and Marchionini (1996) evaluated three engines on only five topics in 1996 and found no statistically significant difference between effectiveness means. In 1997[2] Leighton and Srivastava (1999) compared five engines using 15 topics and found that three of the engines

were superior to the other two, using a new measure based on precision at cutoff 20, but assigning different weights to the top three, next seven and next ten results.

Leighton and Srivastava constructed a set of topics from a variety of sources which were intended to model the information needs of undergraduate students. They themselves generated queries from the topics. Seven queries were simple "bag of words" queries, seven were structured in some way and one attempted to locate information about a person. Leighton and Srivastava used automatic scripts to submit queries to the engines and fetch results. After suppressing information about which engine had retrieved each page, they judged relevance themselves using four categories of relevance. They also identified dead links and duplicate pages. They analysed their data several times using different relevance thresholds and with and without penalising duplicates.

Gordon and Pathak obtained 33 real information needs from volunteers among the faculty members in a university business school. These were recorded in considerable detail and passed to skilled search intermediaries who were given the task of generating near-optimal queries for each of eight search engines by an interactive, iterative process. The top 20 live results generated by each of the engines in response to the final queries were then printed and returned to the originating faculty member for assessment on a four point relevance scale. They found that search effectiveness was generally low, that there were significant differences between engines and that the ranking of engines was to some extent dependent upon the strictness of the relevance criterion.

The characteristics of the Leighton and Srivastava and Gordon and Pathak studies are compared with the present experiment in Table 1. We compare our results with those of Gordon and Pathak in Section 4.3.

2.1. Evaluation philosophy

Gordon and Pathak[3] present a list of seven evaluation features which they claim should be present to maximise accuracy and informativeness of evaluation. Paraphrasing for brevity, these are:

1. Searches should be motivated by genuine user need.
2. If a search intermediary is employed, the primary searcher's information need should be as fully captured as possible and transmitted in full to the intermediary.
3. A large number of search topics must be used.
4. Most major search engines should be included.
5. The most effective combination of specific features of each search engine should be exploited. Ie. the queries submitted to the engines need not be the same.
6. Relevance judgments must be made by the individual who needs the information.
7. Experiments should be well designed and conducted.

Features 3 and 7 are essential features of a useful scientific study and Feature 2 is certainly desirable (when applicable). However, the value of Feature 4 is dependent upon the goals of the evaluation. For general-purpose evaluation, we propose an additional feature:

8. The search topics should represent the range of information needs both with respect to subject and to type of results wanted. (See Section 6 below.)

Table 1. Comparison of the present experiment with the two most closely related previous studies.

Detail	Gordon and Pathak 1999	Leighton and Srivastava 1997	The present study
Number of search engines	8	5	20
Number of topics	33	15	54
No. relevant found per topic	5–125, mean 42.2	Not stated	2–255, mean 88.5
Date queries submitted	1998	Early 1997	20 Sep. 1999
Information needs			
Genuine?	Yes	Yes	Presumed
Originators	Faculty members	Library clients	Anonymous searchers
Length of topic statement	Approx. 100 wds	4.9 wds	5.9 wds
Range of subjects	Business	Broad	Broad
Types of answer required	Selection of rel. pages	Selection of rel. pages	A variety
Queries			
How generated?	Human intermediary	By experimenter from verbal requests	From Web logs (verbatim)
Average number of words	Not stated	4.9	5.9[a]
Used query operators	Yes	Some	No
How many results judged	Top 20 live	Top 20	Top 20 live[b]
How were pages judged?			
By whom	Inquirers	Experimenters	Research assistants
By topic originator	Yes	No	No
Result lists merged?	Yes	Yes	Yes
Blind judging	Yes	Yes	Yes
Follow hyperlinks	No	No	No
Text/images	Both	Raw HTML	Rendered text
Presentation	Printed paper	EMACS text editor	Browser
Page truncation	After ten pages	No	No
Order of presentation	Random	Not stated	Increasing length
Auto. search aids for judging	None	Not stated	RAT[c]
Relevance categories	4	4[d]	2
Measures reported			
P@20	In graphs	Weighted	Yes[b]
P@15–20	Yes	No	No
Average P@1–5	Yes	No	Yes
P@n, $n < 20$	Yes	No	Yes[b]
Prec. based on strict rel.	Yes	Yes	No
MRRl	No	No	Yes

(*Continued on next page.*)

Table 1. (*Continued.*)

Detail	Gordon and Pathak 1999	Leighton and Srivastava 1997	The present study
Relative recall @ 20	Yes	No	No
TREC-style ave. prec.	No	No	Yes

[a]Including stopwords.
[b]Subject to the result page changeover bug for some engines (see text).
[c]The operation of the RAT (Relevance assessment tool) is described in the text.
[d]Plus duplicate and dead link categories.

The remainder of the features in the original Gordon and Pathak list merit detailed discussion.

Restricting searches to genuine user needs (Feature 1) and requiring that the person with the original need should evaluate the results (Feature 6) must inevitably limit the scope of evaluations. For example, the Gordon and Pathak study itself was limited to the information needs of faculty members in a business school.

We acknowledge that judgments of relevance vary from person to person and from time to time. However, there is empirical evidence that retrieval system rankings remain stable across different sets of relevance judgments (Voorhees 1998). We would prefer to say that searches should be *representative* of genuine user need and that relevance judging across the aggregated search results should be consistent (one judge should evaluate all responses to a topic in as short a time as possible).

Feature 5 is quite contentious. All well-known public search engines are designed to produce a list of results when a set of query words (without operators or special syntax) is typed into the search box provided. It is therefore perfectly reasonable to compare the quality of results produced by search engines given identical input queries in this form. Furthermore, such queries are far more typical of Web search than are sophisticated queries exploiting advanced query language features. Analyses of query logs (Silverstein et al. 1999, for example) show that typical search engine users very rarely use any form of query operator and, when they do, they frequently make errors.

Measures obtained by studies which adopt the approach of trying to find the best query formulation for each search engine (Feature 5) are certainly interesting, particularly if they indicated that these "near-optimal" queries performed dramatically better than simple queries. However, the Gordon and Pathak study did not compare the performance of the highly tuned queries with simpler versions such as the initial Boolean queries supplied by the search originators or simple word lists.

Unfortunately, in order to achieve the fifth desideratum within their own study, Gordon and Pathak introduced a set of confounding variables due to the introduction of skilled search intermediaries charged with converting user topic specifications into near-optimal queries for each engine. What they have evaluated is in fact a set of combinations of human intermediary and search engine, with all the human variability that that entails.

The use of human intermediaries also seems somewhat inconsistent with the reasoning behind requiring relevance assessments to be conducted by the originator of the information need (Feature 6). In order to form queries from the original user's topic statement, and to

judge which variant of a query performs best, the intermediary must make the same type of interpretations which would need to be made by relevance judges other than the topic originator.

3. The experiment

We applied TREC-style methodology to the evaluation of Web search engines, as they operated on September 20, 1999. Because we used queries taken from real query logs, our study exhibits all of the eight features listed in the preceding section, except for those numbered 2 (not applicable), 5 and 6. As discussed in the preceding section, we feel that there are strong reasons not to adopt Feature 5 and reasons why Feature 6 may not be necessary.

3.1. What was evaluated?

Most current Web search engines consist of two key components. One is a *crawler* (Koster) whose job it is to create a Web *snapshot*[4] by identifying, selecting and fetching documents and the other is a text retrieval system (TRS) operating over the snapshot collection.

The quality of search results clearly depends upon the performance of both search engine components. Relevant document documents may fail to be retrieved either because they were not fetched by the crawler or because the TRS failed to rank them appropriately. There is evidence (Hawking et at. 1999) which suggests that precision at fixed cutoff increases with collection size.

We evaluated both components in combination. This is a necessary extension of conventional evaluation methodology, in which there is a static, well-defined test collection and only the TRS is evaluated.

3.2. Which search engines?

Lawrence and Giles (1999) report measurements taken in February 1999, of the coverage and freshness of indexes maintained by Web search engines. These measurements relate principally to the performance of the crawler rather than the TRS. They found that none of the most popular search engines covered more than 16% of their estimated total of 800 million indexable web pages.[5]

To add an extra dimension to the data presented by Lawrence and Giles, we measured the result-list precision of all eleven engines considered in their study: Northern Light, Snap, AltaVista, HotBot, Microsoft, Infoseek, Google, Yahoo, Excite, Lycos and Euroseek. We added:

1. Two metasearch engines: MetaCrawler (mentioned but not studied by Lawrence and Giles and accessing About.com, AltaVista, DirectHit, Excite, GoTo.com, Infoseek, Looksmart, Lycos, Thunderstone, WebCrawler and Yahoo) and Inquirus (operated by Lawrence and Giles on a limited-access basis and accessing AltaVista, DirectHit,

Euroseek, Excite, Fast, Google, Hotbot, Infoseek, Lycos, NorthernLight, Open Directory, Snap, Thunderstone, Yahoo and Yahoo/Inktomi). Metasearchers do not index documents themselves but rather forward queries to a number of primary search engines and form a composite result list.
2. FAST (also known as "AllTheWeb", a new large search engine which aims to index the whole of the Web;
3. EuroFerret, another Europe-focused search service we wished to compare with EuroSeek;
4. DirectHit, a system which takes page popularity (in terms of access frequency) into account when ranking pages;
5. Three search engines providing search within the Australia and New Zealand area: ANZwers, ExciteAus, and Web Wombat;
6. A directory service (LookSmart Australia). This service presents Australian results first but then appends results from its Looksmart Worldwide counterpart.

Note that several of these services are not fully automatic and generate responses at least partially on the basis of classifications made and stored by human editors. Prominent examples include the Yahoo, Lycos and LookSmart Australia directory services. Given that these services accepted queries and presented results in a fashion which appeared to be ordered, we considered that it was reasonable to evaluate them alongside search services whose degree of manual intervention is less.

Our list of twenty engines included all of those in the Gordon and Pathak (1999) study except for OpenText and Magellan. The overlap between the two studies therefore comprises six engines: AltaVista, Excite, Infoseek, Hotbot, Lycos and Yahoo. Leighton and Srivastava (1999) studied the first five of these, but did not report comparable measures.

3.3. Queries and result lists

We started with two sets of 100,000 single-sentence "natural language" queries from logs supplied to us by Alta Vista and Electric Monk. We used natural language queries because we believed that it would be easier for our judges to interpret what it was that the inquirer was actually seeking. We merged the two sets and filtered out queries which were either likely to be offensive to some people or which contained fewer than two non-stopwords.[6] We then randomly selected queries from the merged set and asked an experienced judge to decide whether she and her colleagues would be able to interpret what it was the inquirer was seeking and be able to judge the relevance of results returned. She rejected 719.

A total of 59 queries were accepted but due to a pair of near-duplicate queries and problems of assessor availability, 57 were judged for the TREC Large Web track and a subset of 54 for the search engines. A sample of the judged queries is shown in figure 1. We did not correct typographical errors in the queries (such as `"slobadan"` in the figure). They were submitted to the search engines exactly as shown.

We used scripts to present each of the queries to each search engine. The first 20 "live" results from each engine for each query were retrieved and merged into a single pool of

```
who are the current supreme court justices?
historic preservation
where can i find gif files?
where can i find film reviews?
how do i create a web site?
where can i find the best jokes?
where can i find information on russia?
how does a digital camera work?
thalidomide and multiple sclerosis
slobadan milosevic
hindenburg disaster
old japanese science fictions movies
where can i find information on school violence?
what are some psychological principles and attitudes for advertising
why do feet smell?
armstrong louis
where can i find statistics for education in the united states?
methodist sermons
egyptian history
wall street journal
how to start business
find information about american anarchists
reasons for studying marketing
where can i find the saints and the catholic church?
show me a list of vegetarian restaurants in new york city.
```

Figure 1. A sample of the queries used in the experiments.

pages to be judged. To enable fair comparison with TREC systems, "dead" or inaccessible results were ignored and engines were not penalised for returning them.

3.4. Relevance assessors/Judging instructions

We employed a team of six judges, each of them an Australian with a University degree, but with no specific expertise in Computer Science or Information Retrieval, and no allegiance to any search engine or retrieval system. We asked them to judge whether documents were relevant to particular queries. The pages retrieved by all search engines were combined into a single pool for each query and presented to the judges without indication of which search engine(s) had retrieved them. The documents were rendered using a text-only browser (Lynx) and images or sounds were not presented. Hyperlinks were presented as numbered references, eg. [4], but the judges could not follow them.

We asked judges to imagine that they themselves had submitted the queries and to evaluate the answers on that basis. However, we also instructed them to judge each document independently of the others and to score as relevant any page which included material which was "on-topic" and which contributed information not contained in the query. We asked them not to make any judgment about the correctness of information given and not to require that pages returned in response to a question were in the form of an answer. In

line with past TREC methodology, judgments were binary and made on the basis of textual content alone. Either a document contained relevant content and was judged relevant or it was judged irrelevant.

To ensure consistency of results, all the documents retrieved for a query were judged by the same person. Each document retrieved in response to a query was judged by only one judge, as earlier work (Hawking and Thistlewaite 1997, Voorhees 1998) failed to demonstrate any particular benefit of multiple judgments.

3.5. Relevance assessment tool (RAT)

We used special relevance judging software known as the RAT (Relevance Assessment Tool) which was developed by Jason Haines and Paul Thistlewaite in 1996 for use in the TREC Very Large Collection and Web tracks. Since then it has been maintained by Nick Craswell. Figure 2 shows a screen snapshot of the RAT in action.

Before commencing judgments for a new query, each judge was required to enter *concepts* they would use in evaluating relevance and a *criterion of relevance*. For example, the concepts for the last query in figure 1 were "vegetarian", "restaurants" and "new york" and the criterion of relevance was the conjunction of all concepts.

We encouraged judges to accumulate lists, using RAT facilities, of words, phrases and part-words which in their view constituted *evidence* for the presence of each concept. For example, "ny", "yc"[7], "manhattan" and "new york city" as well as "New York" constituted evidence for the presence of the "New York City" concept. Judges entered evidence either by typing into an `add evidence` box or by selecting text from presented pages.

The RAT displays concept and criterion information throughout the judging process as an aid to maintaining consistency. However, the information was provided for assistance only and judges were free to make judgments regardless of whether criteria were actually satisfied or not.[8]

As shown in figure 2, occurrences of evidence were highlighted in each document by the RAT in the colour associated with the concept, enabling the judge to easily spot multi-coloured, potentially relevant sections in long documents. A hypothetical document containing the sentence, "Luigi's pasta restaurant at 2301 7th Avenue, Manhattan caters for both vegetarians and vegans." would thus show highlighting in all three colours. The judge could instantly home in on the sentence, and decide that the document is relevant. He or she might also decide to save the part-word "vegan" as additional evidence for the vegetarian concept.

Over the 54 topics, the number of concepts defined per topic by the judges ranged from 1 to 8 (mean 3.7) and the average number of pieces of evidence defined (for all concepts combined) was 30.9.

3.6. Presentation order

The order of presentation of documents for judging has been shown to affect the judgments made (Eisenberg and Barry 1998). When we first used the RAT in TREC-5 (1996), we compared three different presentation orders: ascending length, descending length and random.

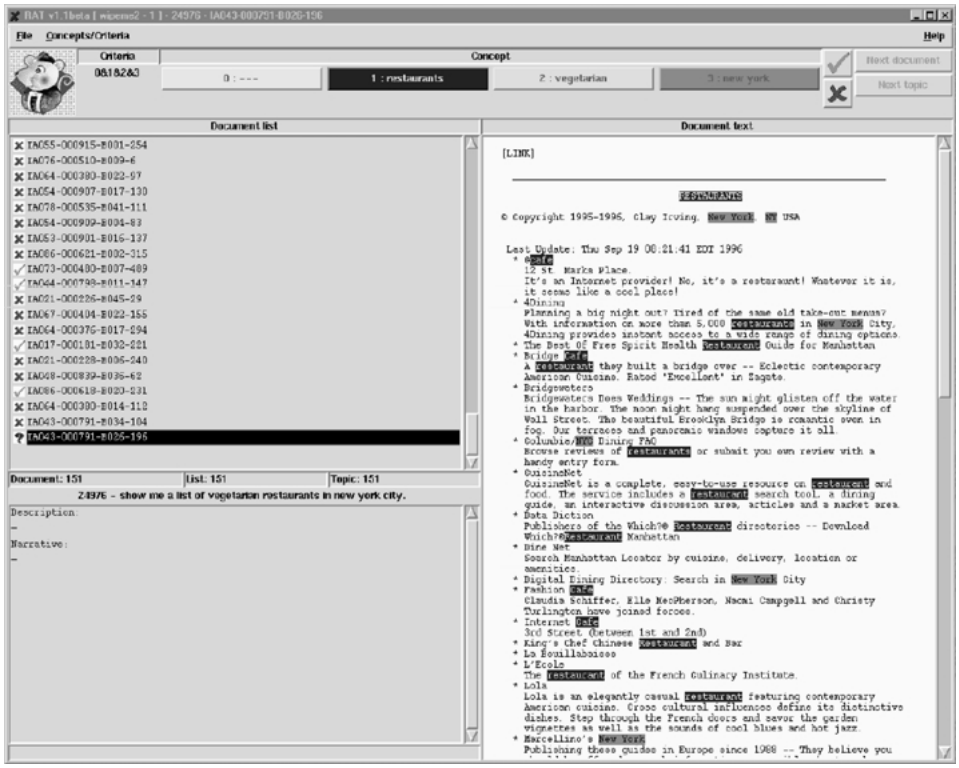

Figure 2. A screenshot of the Relevance assessment tool (RAT) used in the present experiments. Coloured rectangles representing the concepts are at the top. (In this case, Concept 0 was a dummy entry created by the assessor because she didn't like the colour assigned to it!) The list of documents seen so far is shown at top left, and the text of the current document (with highlighting of concept evidence) on the right. The bottom left panel includes the text of the current query as a heading but is otherwise unused. The description and narrative fields would be filled in if TREC ad hoc topics were being judged. A judgment for the current document is recorded by clicking on the appropriate icon at the far top right of the screen. Once the judgment has been made, the `Next document` button is enabled unless this is the last document for the topic, in which case the `Next topic` button is enabled. At any point within a topic, the judge may revisit previously seen documents (which will be displayed with all the concept evidence currently available) and, if desired, change their judgment.

We found that descending length was impractical because judges felt that they needed to read all of a long document in order to be sure that sections of it were not relevant. In the most extreme case, a judge spent more than six hours on a single document.[9]

By contrast, judges found that working from the shortest documents first enabled them to accumulate sufficient evidence to be confident that when they reached a very long document, they could use scrolling to quickly locate the paragraphs which contained potentially relevant material with a low probability of missing relevant material.

We concluded that presentation in ascending length order increased the probability that short relevant sections in long documents would be identified by maximising the high-

lightable evidence available. Accordingly, we have used this presentation order in all judging since 1996.

In the Eisenberg and Barry study, the documents had already been assessed on a multi-level relevance scale. They found that when documents were presented in relevance order, either increasing or decreasing, subjects tended to "hedge" on the first documents seen by allocating scores closer to the mean than they would have otherwise. It seems very unlikely that presentation order effects seriously affected our experiment as:

1. Hedging is not possible when there are only two possible relevance values.
2. Our judges were free to reassess previously judged documents if they felt that they had been too harsh or too lenient.
3. The association between document length and relevance is unlikely to be strong. The judge would have no reason to expect that the first document would or would not be relevant.
4. Order effects would only have confounded our comparative results if certain engines tended to disproportionately return very short or very long documents.

3.7. Measures used

Judgment of the first 20 live results allows calculation of a range of precision-oriented measures including precision at $n \leq 20$ documents retrieved (P@n), mean reciprocal rank of first relevant document (MRR1) and TREC-style average precision (TSAP).[10]

Unfortunately, data analysis long after the data was collected revealed a bug in the scripts used to send queries and fetch result lists and documents. This meant that for some engines (nine altogether) where the queries contained a question-mark (29 out of 54), the script failed to move to the second page of ten results[11] meaning that some results lists were artificially truncated. For the engines affected by this bug, P@n ($10 < n \leq 20$) would certainly be underestimated. It is also likely that comparison of engines on precision at earlier cutoff (10, 9, ...) would be unfair because dead links in the first ten results were replaced with live ones for the engines not subject to the bug.

Available evidence suggests that most Web users generally look at very few results indeed. Silverstein et al (1999) [Table 7, p. 10] report that for a sample of over half a billion queries submitted to Alta Vista, 85.2% requested only a single result page. Accordingly, and to completely avoid the effects of the bug, our main analysis is based on the first seven live results. However, we acknowledge that precision at very early cutoff is likely to result in less stable system rankings (see Buckley and Voorhees, 2000), and statistical significance is likely to be harder to achieve than for later cutoffs.

For compatibility with Gordon and Pathak's study, we performed our principal analyses on P@1–5, the average of the precisions at cutoffs 1..5. We also present results for P@1, MRR1, and TSAP and plot P@n against n separately for the groups of engines affected and not affected by the bug.

Unlike Gordon and Pathak, we did not measure recall, since the meaningfulness of the recall values depends heavily upon the accuracy of estimates of how many relevant documents there really are.

Table 2. Summary of results for the twenty search engines. These are based on the first seven live results. TSAP means TREC-style average precision, MRRl is the mean reciprocal rank of first relevant document and P@n means precision at n documents retrieved. P@1–5 is the average of precision values at each ranking cutoff from 1 to 5.

Run	TSAP	MRRl	P@1	P@5	P@1–5
ANZwers	0.1111	0.2673	0.1852	0.1667	0.1787
AltaVista	0.3714	0.582	0.5	0.4519	0.4691
DirectHit	0.244	0.4455	0.4074	0.3111	0.3474
EuroFerret	0.284	0.5553	0.4444	0.3704	0.4043
EuroSeek	0.0733	0.1185	0	0.1296	0.0858
Excite	0.2695	0.4994	0.3889	0.363	0.3593
Exciteaus	0.2288	0.4596	0.3704	0.3	0.318
Fast	0.2939	0.4843	0.4074	0.3593	0.3725
Google	0.3939	0.6133	0.463	0.4889	0.4848
HotBot	0.3064	0.5705	0.463	0.4111	0.4196
InfoSeek	0.3698	0.594	0.4444	0.4852	0.4838
Inquirus	0.404	0.5833	0.463	0.4963	0.4965
LookSmart	0.3641	0.5575	0.463	0.4333	0.4502
Lycos	0.2893	0.5244	0.3889	0.3926	0.3946
MetaCrawler	0.3075	0.5764	0.4259	0.4222	0.4329
Microsoft	0.3522	0.5974	0.5	0.437	0.4624
NorthernLight	0.3846	0.6897	0.5556	0.5037	0.5211
Snap	0.2939	0.5209	0.3519	0.3963	0.3944
WebWombat	0.1765	0.3639	0.2963	0.2333	0.2535
Yahoo	0.222	0.4336	0.3519	0.2889	0.3278

4. Results

Our main results are presented in Table 2 and in figures 3 to 10.

As shown in figure 3 the range of values recorded for P@1–5 was quite large. A multiple analysis of variance (MANOVA[12]) of the P@5 data, confirms that there is a significant difference in the performance of the search engines, $F(19, 35) = 6.865$, $p < 0.001$.

Multiple pairwise comparisons using the Least Significant Difference test were conducted. Although Northern Light was the top ranked engine on the basis of P@1–5, differences between it and the next nine engines were not statistically significant. Differences between NorthernLight and the engines below EuroFerret (Lycos, SNAP, Fast, Excite, DirectHit, etc.) were all significant, $p < 0.05$.

All 190 pairwise correlations between the engines were positive, Pearson r values ranging from 0.066 to 0.885. Of these, 155 achieved significance ($p < 0.05$).[13] Four pairs exhibited correlations in excess of 0.7. They were (Looksmart, AltaVista), (Microsoft, AltaVista), (SNAP, Hotbot) and (Microsoft, Looksmart). It is understood that at the time of the evaluation, that SNAP and Hotbot both used Inktomi search technology and that there

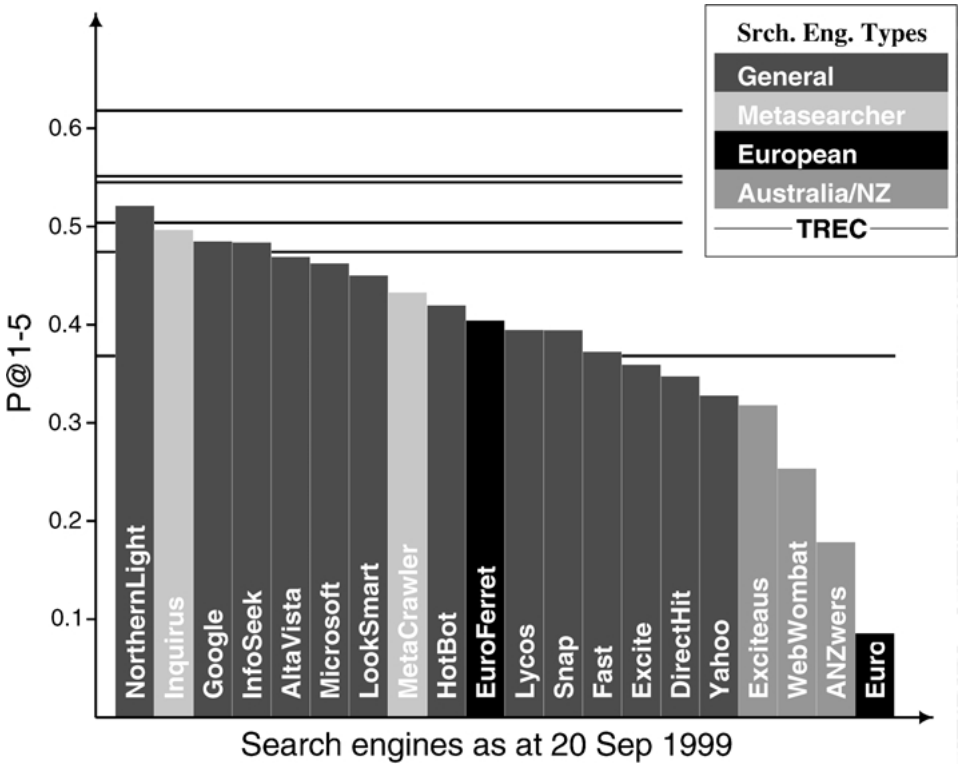

Figure 3. The 20 public search engines compared on the basis of precision averaged across cutoffs 1–5 (P@1–5). The type of search engine (general, metasearch, or regional) is color-coded according to the key at top right. Each horizontal black line gives the corresponding performance for a 1999 TREC system. See Section 4.2 and Table 4 for details.

were commercial relationships between (Microsoft, Looksmart), (Microsoft, AltaVista) and (AltaVista, Looksmart).

Table 2 shows that the search engine rated by Lawrence and Giles as having the largest coverage (NorthernLight) also scored highest on all of the measures except TSAP. However, the FAST engine which claimed a coverage nearly 50% larger than NorthernLight, performed relatively poorly. Figure 4 shows a scatter plot of coverage against P@5 for the Lawrence and Giles engines. The correlation between P@5 and coverage is not statistically significant.

A metasearcher works by broadcasting incoming queries to a large number (11 and 14 in these two examples) of primary search engines and merging their results lists. Metasearchers therefore achieve a very high effective coverage. The two metasearch engines performed creditably, but not as well as the best individual engines.

Not surprisingly, given the general nature of the test queries, the unrestricted-domain search engines generally achieved better results than those restricted to European and Australasian domains.

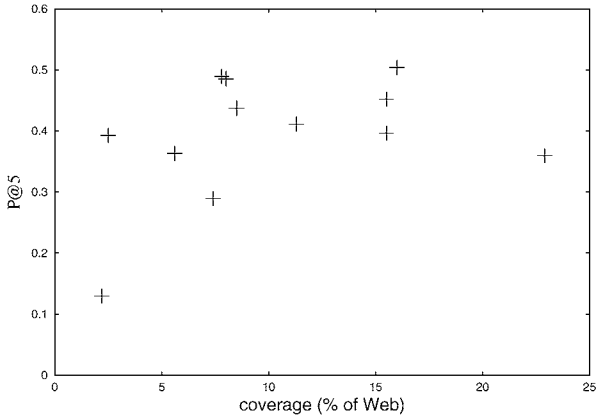

Figure 4. The relationship between coverage (percentage of estimated publicly indexable Web, as reported by Lawrence and Giles) and P@5, for the eleven Lawrence and Giles engines plus FAST, whose claimed size was known at the time of the experiments. The Pearson R coefficient of correlation was 0.370, which is not significant at the 0.05 level (two-tailed).

Figure 5 shows the mean reciprocal rank of first relevant document (counted as zero if no relevant document appeared in the first 7 results). If the first relevant document appears in the second position, the reciprocal rank is 0.5. Figure 6 compares performance on the TSAP measure (explained in Section 3.7).

Figure 7 compares performance on the basis of P@1, which is equivalent to the probability that the first result is relevant. A nonparametric test revealed a highly significant difference in the performance of the engines on this measure, Cochran $Q = 82.3$, $p < 0.001$.[14]

There are many more tied scores on P@1 than on P@20, reflecting quantisation due to the fact that there are only 54 possible scores in the former case and 1080 in the latter. Note that Euroseek did not find a single relevant document at rank 1 and scored zero on this measure.

Figure 8 documents the variation of P@n as n varies from 1 to 20 for the search engines not affected by the result changeover bug. Figure 9 shows the variation in P@n for the remaining enginges over the range of n which was not affected by the bug. Ignoring fluctuations at small n, most engines exhibited a very gradual decline in P@n across the range studied. This general pattern was also found for all of the engines shown in the corresponding Gordon and Pathak figure (figure 1, p. 156), except OpenText (not studied here.) DirectHit and Yahoo in the present study were also exceptions. For them, P@n declined more sharply.

4.1. Intercorrelation of measures

Table 3 shows the Pearson r intercorrelations between four different measures, based on the top 7 results. All are significantly intercorrelated ($p < 0.01$). Similar, and in some cases even higher, intercorrelations were observed for the combined TREC and search engine groups.

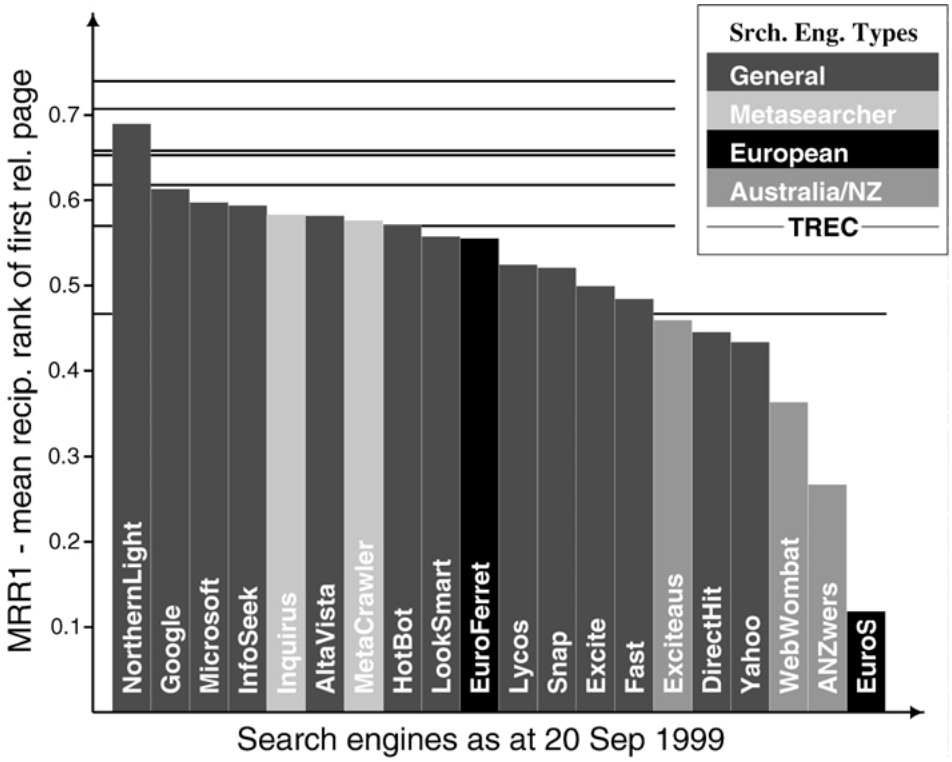

Figure 5. The 20 public search engines compared on the basis of mean reciprocal rank of the first relevant page found. Engines scored zero for a query if no relevant document was found in the first 7 results. Colors identify the type of search engine. Each horizontal black line gives the corresponding performance for a 1999 TREC system.

Considering only the eleven engines not affected by the result-page bug, there is a correlation of 0.97 ($p < 0.01$) between P@5 and P@20 and of 0.87 between P@1 and P@20 ($p < 0.01$).

4.2. How do search engines compare with 1999 TREC systems?

The performance of groups participating in the TREC-8 Large Web Task is reported in detail in Hawking et al. (1999) and summarised in Table 4. The queries judged in the Large Web Task were a slightly larger superset of the queries used in the present study. TREC results reported here have been re-analysed to take into account only the data for the 54 queries judged for both the TREC systems and the search engines.

These TREC results are also represented as horizontal black lines in figure 3, 5 and 7. Instead of crawling their own data from the current web, these runs used the first 18.5 million pages (100 gigabytes) of a crawling run carried out in early 1997. Each TREC run thus simulates the operation of a search engine with small and out-of-date coverage (comparable to the size estimated by Lawrence and Giles for Lycos).

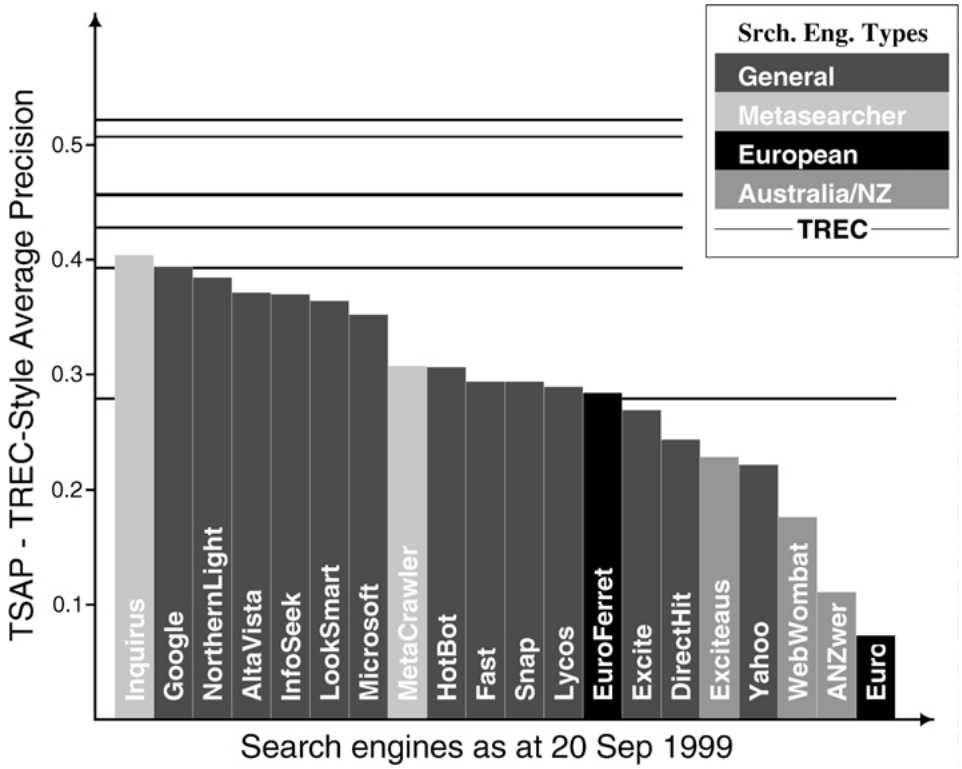

Figure 6. The 20 public search engines compared on the basis of TREC-style average precision, based on top 7 rankings. Colors identify the type of search engine. Each horizontal black line gives the corresponding performance for a 1999 TREC system.

Documents for the TREC systems were judged in a separate batch from the search engine documents but the judges assigned to each query were the same, the presentation of documents was identical and all other conditions were held constant. The TREC batch was judged first, with judging of the search engine batch following immediately afterward. Concepts and evidence accumulated with the Relevance Assessment Tool (RAT, see Section 3.5) during the TREC judging were carried over to the search engine batch.

The number of TREC documents judged was 11,654 of which 25.8% (3008) were judged relevant. The number relevant per query ranged from 1–136, with a mean of 55.7. The maximum achievable mean P@5, based on total known relevant pages, was 0.9259.

For comparison, the number of documents judged for the search engines was 14,951 of which 32.0% (4780) were judged relevant. The number relevant ranged from 2–255, with a mean of 88.5. The maximum achievable mean P@5, based on total known relevant pages, was 0.9741.

The group of TREC systems was compared with the group of search engines using separate Mann-Whitney U[15] tests for each of the following measures averaged across

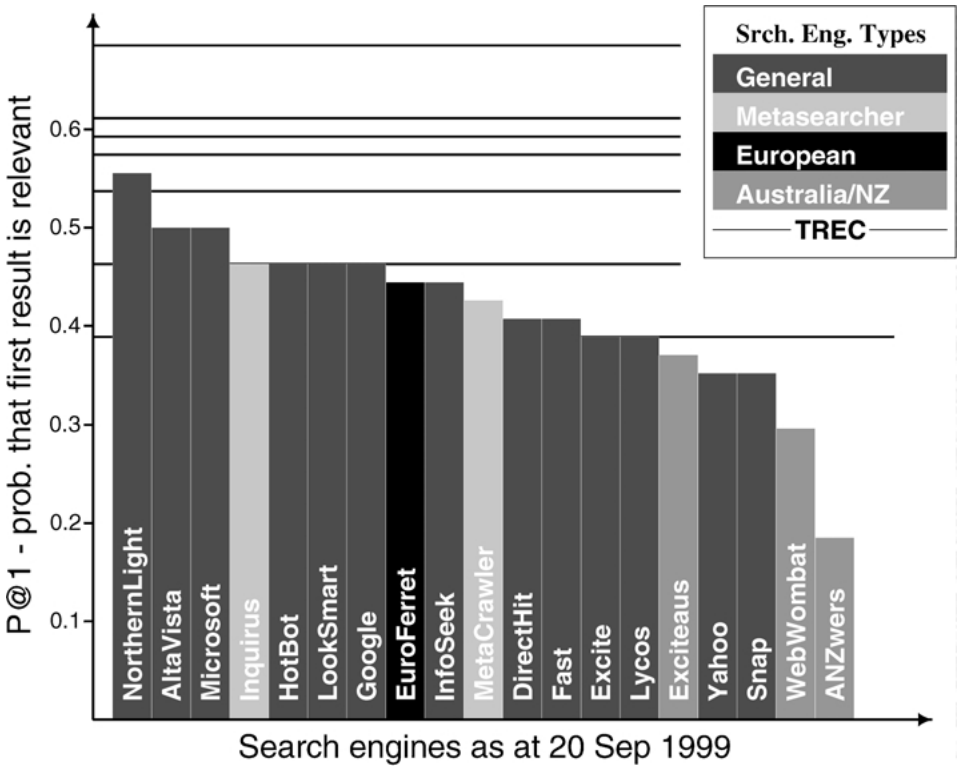

Figure 7. The 20 public search engines compared on the basis of P@1. Colors identify the type of search engine. Each horizontal black line gives the corresponding performance for a 1999 TREC system.

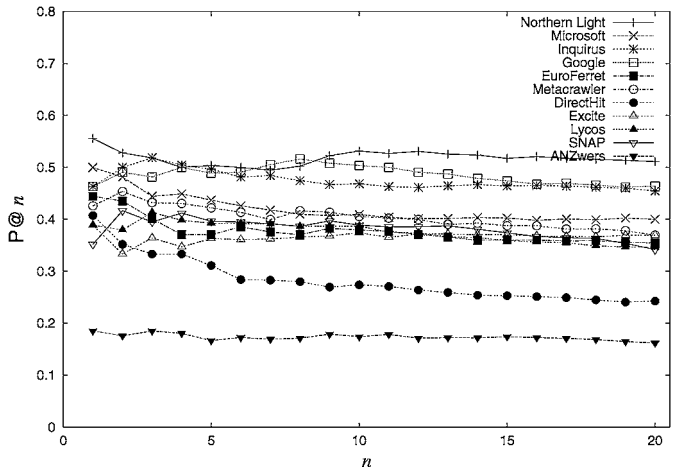

Figure 8. The variation of P@n with n for the engines not affected by the bug (ranked on P@1).

Table 3. Intercorrelation of various measures for the twenty search engines. TSAP means TREC-style average precision, MRR1 is the mean reciprocal rank of first relevant document and P@n means precision at n documents retrieved.

Measure	TSAP	MRR1	P@1	P@5
TSAP	1	0.947	0.909	0.989
MRR1	0.947	1	0.968	0.96
P@1	0.909	0.968	1	0.897
P@5	0.989	0.96	0.897	1

Table 4. Summary of results for the participating groups in the TREC-8 Large Web task. The results presented are the averages of all the runs submitted by each group. TSAP means TREC-style average precision, MRR1 is the mean reciprocal rank of first relevant document and P@n means precision at n documents retrieved.

Run	TSAP	MRRl	P@1	P@1	P@1–5
UMass	0.3927	0.5698	0.463	0.4741	0.4741
ACSys	0.2796	0.4668	0.3889	0.3519	0.3682
AT&T Research	0.507	0.7074	0.6111	0.6074	0.6178
Fujitsu Labs	0.428	0.6179	0.537	0.4926	0.5038
Microsoft Research	0.4561	0.6528	0.5741	0.5333	0.5449
MSR/City Univ.	0.457	0.6582	0.5926	0.5259	0.5512
UWaterloo	0.5218	0.7399	0.6852	0.5778	0.6183

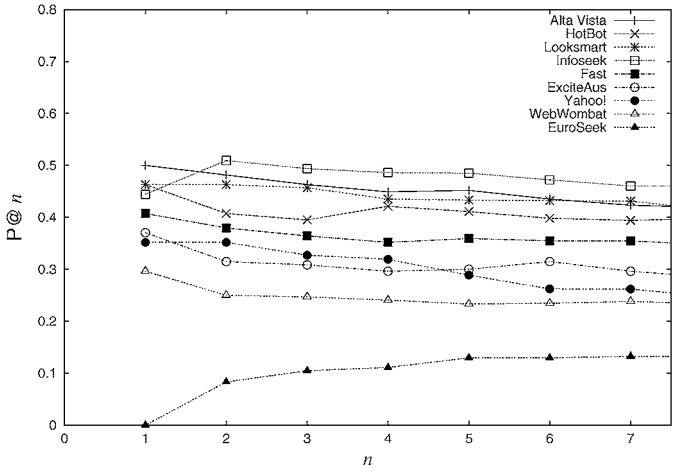

Figure 9. The variation of P@n with n for the search engines which were affected by the bug (ranked on P@1).

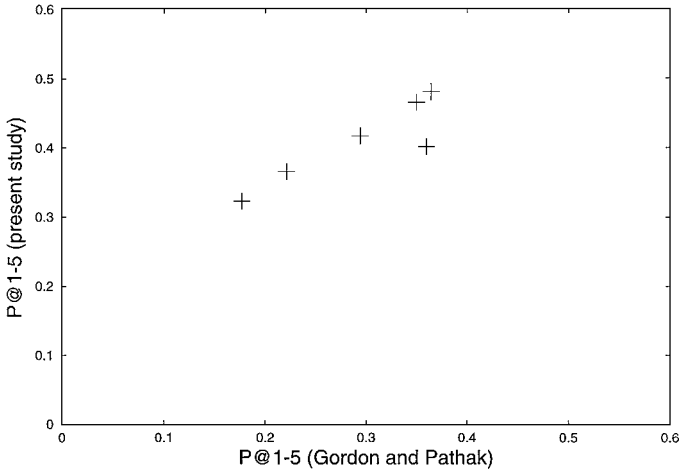

Figure 10. The 6 public search engines included in both the Gordon and Pathak study and the present experiment. The correlation coefficient of the scores is 0.89 which is significant at the 0.05 level (two-tailed).

queries: P@1, P@5, P@1–5, MRR1 and TSAP. On each measure the TREC systems are significantly better ($p << 0.05$ in all cases).

4.3. How do the present results relate to those of Gordon and Pathak?

Figure 10 compares P@1–5 scores (lenient cutoff) obtained by Gordon and Pathak with the corresponding scores for the same engines obtained in the present study. As shown, the results are highly correlated.

In figure 10 a line of best fit which passes through the origin has a slope of about 1.5. In other words, our P@1–5 scores tended to be 50% higher than those observed by Gordon and Pathak (based on their lenient encoding), despite the advantage presumably conferred by use of highly tuned queries and advanced query syntax. This may be explained either by a systematic difference in topic difficulty, or by a difference in judging standards.

5. Discussion

Our main findings are as follows:

1. There was a high correlation (0.89) between the P@1–5 scores from this study and those reported by Gordon and Pathak for their experiment conducted a year earlier than ours. This was surprising because we expected that updating of indexes and improvements to algorithms would have caused much more volatility. Furthermore, there were considerable differences in the way the two studies were conducted.
2. We found no significant correlation between index coverage and early precision (P@5) (figure 4). Provided that coverage is sufficient to include enough relevant documents for

the chosen topics, the ranking algorithm seems to be the determining factor. Coverage would become more critical if the queries had very few relevant answers or if the task were different (for example, if it involved locating obscure known items, or locating ALL documents matching a specification).

3. In the present experiment, TREC Large Web systems as a group outperformed the search engines. However, the performance of the best search engines approached (and in one case surpassed) the median performance of the participants in the TREC-8 Large Web Task (figures 3, 5, 6 and 7). This suggests that some search engines are indeed using reasonably good retrieval methods. However, it should be noted that the Large Web collection was only a small fraction of the size of the data indexed by NorthernLight and Google. Previous work (Hawking et al. 1999) shows that precision at fixed cutoff tends to increase as collection size increases. Furthermore, TREC participants almost certainly sacrificed some of the effectiveness of their methods in order to achieve speed of processing and are therefore not fully representative of TREC state-of-the-art. There is still room for improvement.

   The VLC2 content was nearly three years out of date with respect to the Web at the time of our experiments. More importantly, we believe it was out-of-date with respect to the query set as well. It is not known precisely when the queries employed were originally submitted but they were obtained from the search engine companies in late 1998 and early 1999. However, none of the queries seem to relate to subjects which would have been well covered on the Web in 1997 but not in 1999. Any slight bias due to the age of the VLC2 crawl is likely to have operated in favour of the search engines rather than against them.
4. For the great majority of engines, P@n declined very slowly with increasing n (figures 8 and 9).
5. A very high intercorrelation of measures was observed. (Section 4.1.) P@20 for the engines with complete result lists is very strongly predicted by P@5 and almost as strongly predicted by P@1.
6. Despite very large effective coverage, the metasearchers did not outperform the best individual engines. We have calculated that a hypothetical metasearcher which broadcast queries to all 20 engines in this study could achieve a P@5 score of 0.9741 on the test queries if it had used an ideal merging algorithm.[16]

   Its precision would be less than perfect because of the 3 queries for which the total number of relevant pages returned by all engines was less than 5.

   The large gap between actual and potential performance of metasearchers suggests a possibly fruitful avenue for the improvement of search engine quality. Given that some metasearchers (such as Inquirus) perform merging by downloading all the result documents and reranking on the basis of their content, it may be possible to find merging methods which are capable of substantially narrowing the gap.

5.1. Reliability of results

How subject are the above search engine evaluations to the vagaries of human judgment? It is known that agreement between human relevance judges is less than perfect. Voorhees

and Harman (1996, p. 8) reported a 71.7% rate of unanimous agreement between three assessors over 14,968 documents. However, as previously noted, Voorhees (1998) found that, while substituting relevance judgments made by different human assessors changed the magnitude of scores, it had almost no effect on the rank order of systems.

5.2. Representativeness of queries

Three potential biases affected our method for selecting queries (Section 3.3). The 200,000 queries constituting the initial pool were taken from the query logs of two different search engines. They may have been biased by user perceptions of the characteristics of the two particular engines, by the client demographics of those engines, and by the fact that the queries were in both cases submitted to a "natural language" interface. We introduced further biases by excluding adult-content queries and by allowing a single judge to select queries which she felt that she and her colleagues would be able to judge.

To what extent are these biases compatible with desirable features of search engine evaluation, particularly our own modification andextension (figure 11) of Gordon and Pathak's list? Against Feature 1, we can be reasonably confident that most of the judged queries were submitted as a result of some real information need. Apparently flippant queries such as, "who are you?", and empty or incomplete queries are occasionally found in the logs but no obvious examples were accepted for judging. Some queries may have been submitted for

1. Search topics should be representative of genuine user need.
2. If a search intermediary or third-party relevance judge is employed, the primary searcher's information need should be as fully captured as possible and transmitted in full.
3. A large number of search topics should be used.
4. If the purpose of the study is to enable informed consumer choice, the study should aim to include most major search engines.
5. If the study seeks to compare maximal as opposed to typical effectiveness of search engines, it is appropriate to exploit the full range of features of each engine and to do so may require the use of skilled search intermediaries.
6. The search topics should represent the full range of information needs over which it is desired to draw conclusions, both with respect to subject and to type of results wanted. (See type categories in Section 6.)
7. Result judging should be appropriate to the type of result sought. (See Section 6.)
8. Judging should be consistent within a topic. All judgments for a topic should be made by the same person and within a short period of time.
9. Judging should be blind and be conducted by the person with the need for information or by independent judges.
10. Documents should be presented for judging in the same way that they would have been seen by a real Web searcher, ie. rendered by a browser. Images should be viewable and it should be possible to follow links. However, careful instruction may be necessary to ensure that judges remain on task and do not waste time on fruitless link following.
11. For general comparisons of search engines, dead links should count as useless answers.
12. Experiments should be well designed and conducted.

Figure 11. Summary of desirable features of future Web search evaluations. Based on Gordon and Pathak's list, extended and modified.

testing purposes (as in our own study) but the proportion is almost certainly negligible.

Feature 4 suggests that queries should be representative of the full range of information needs over which it is desired to draw conclusions, both with respect to subject and to type of results wanted. As can be seen from figure 1, the queries we employed covered a very broad range of subjects, including politics, history, psychology, music, medicine, religion, humour, technology, geography, science and finance.

However, all our queries appear to be at the "general-knowledge" level. Our results may not, in general, accurately predict the performance of search engines when queries are posed and evaluated by subject experts, despite the fact that they are correlated with the Gordon and Pathak results for expert queries.

Despite the selection biases identified above, we are fairly confident that our results would predict search engine performance reasonably well for general, non-adult-content queries, provided the same evaluation methodology were employed.

In our view, the major limitation of our study (and most of its predecessors) lies in our failure to represent the range of result types wanted and to apply appropriate evaluation for the different types. Performance on a "find a range of relevant pages" task may not predict at all well the performance on a "find a homepage" task. This issue is discussed in detail below.

5.3. Effect of spelling errors

As noted earlier, some of the test queries contained spelling errors. These errors were not corrected for use in TREC-8 Large Web. This was something of a departure from convention for TREC participants because TREC ad hoc topics are almost free of such errors.

Judges looked for documents satisfying the apparent intent of the query, rather than literal matches. Accordingly, systems able to detect and correct spelling errors may have been at a small advantage relative to others. We have no information about which systems (if any) employed spelling correction.

6. Future Web search evaluation

Unfortunately, in the absence of a standardised Web test collection with complete relevance judgments and a willingness of search engine companies to use it, there is little prospect of reproducible Web search engine evaluation. Nonetheless, search engine effectiveness evaluation is necessary. Future Web search evaluations should adhere to sound experimental methods, such as blind judging and statistical significance testing while improving on the methodology adopted by the present study and its predecessors.

In passing, we note that, given the variation in effectiveness across Web crawlers and the potential effect of this variation on search result quality, there may be a future need for a Web test collection constructed in such a way as to enable crawling as well as retrieval to be evaluated. Such a collection would obviously comprise a set of linked pages and an initial seed list but may be difficult to construct. To have any value it would need to include examples of all the known crawler hazards such as mis-classified file types, subtle self reference, duplicate pages, page redirects, clickable image maps and framesets. Such a

collection could enable testing not only of the ability of a crawling algorithm to reach all the pages but also its ability to fetch more useful pages ahead of less useful ones (as attempted by some commercial systems).

There are a number of ways in which future Web search evaluations could more accurately model Web search.

1. Future evaluations should recognise that retrieval corresponding to different types of information need require different evaluation techniques. Searcher information needs may be broadly classified on the basis of the type of answer expected:

   A. A very short (eg. single sentence) answer to a question. For example, "The melting point of lead is 210C." Evaluation techniques for this type of information need were considered in the TREC-8 Question and Answer Track (Voorhees and Harman, 1999, various papers).
   B. A single document. For example, the home page for the Information Retrieval journal, or the current timetable for the Washington Metro Red Line, or a directory of accommodation in Athens, Greece. This type subsumes traditional *known-item* search. However, Web searchers often deduce that a particular document, such as a company homepage or a bus timetable, ought to exist without having previously seen it, a situation more accurately described as *suspected-item* search. Often, what the page *is* is more important than what it *is about* because the searcher is seeking a point from which to start browsing.
   C. A selection of documents. These documents might be sought for research purposes (for example a range of documents relating to US policy on North Korea) or in order to access on-line services (for example on-line auction sites or sites which allow download of recorded music.)
   D. Every document matching a criterion. For example, every document which discusses orbital engines. Evaluation against this type of need is obviously related to coverage, but also to the sophistication of the engine in expanding the query so as to find documents which do not literally match the query but do match one which is semantically equivalent, for example "Sarich engine", "motor developed by Ralph Sarich" etc. Another complication arises from the expression of needs in terms of explicit or implicit document metadata, such as "all documents authored by Doug Engelbart". Blair and Maron (1985) discuss retrieval effectiveness in a high-recall legal application.

   In each of the first two types, evaluation requires access to only very few documents. In Type B, relevance, even strong relevance, is usually not sufficient and evaluation must be based on asessment of whether a document is the one sought. In these cases, performance can be evaluated on the basis of the position of the desired document in the ranking and it makes little sense to talk of "relevance". To illustrate this point, the Australian National University intranet contains thousands of documents which are relevant to the topic of libraries and one of the most common queries submitted to its search engine is `library`. In the cases where the submitter of this query was looking for the library homepage or the library catalogue, the ability of the engine to retrieve pages which merely talk about libraries is unlikely to correlate well with searcher satisfaction.

In Type C, evaluation may be limited to the first 10 or 20 results, as most users would expect to find the information they require in fewer than this number of documents. However, it is important that the documents presented in a result list should not repeat the same material and evaluation methodology for this type of information need should take this into account.

Evaluation of recall-oriented[17] information needs (Type D) is difficult, given limited assessment resources. It must inevitably rely on sampling techniques.

Classification of information needs is difficult. The above four-way classification illustrates the need for a range of evaluation paradigms, but is still a gross over-simplification. Information needs are highly dynamic and are conditioned by preliminary results. For example, someone searching for a complete timetable may find that no such single document can be found and be forced to seek multiple individual timetables which together provide the needed information. A Type B search for a single homepage may or may not transform into a Type D search when it becomes clear that multiple home pages exist.

Within any of the above types there may be a need for a range of judging instructions and a need for a range of different measures.

In the case of Type C(research) needs it is probably preferable to use multi-valued relevance judgments, as Gordon and Pathak did, as it may be possible to exploit the additional relevance information. We note, however, that Gordon and Pathak do not report any results relying on the distinction between irrelevant and highly irrelevant. They also convert four-value relevance judgments into binary in order to compute their measures.

By contrast, in the case of Type C(services) needs, the judgment to be made is not whether the page is relevant or not but rather whether the page allows direct access to a service of the right type. For example, whether the page *allows* downloading of MP3 files, as opposed to *discussing the topic of* downloading MP3 files.

A defect of the present study is that we treated all queries as though they corresponded to Type C(research) information needs. The problem is not that we may have misinterpreted the intention of the searchers who submitted these particular queries. What is important is that the ability of search engines to process queries of other types was not evaluated.

In the Web context, it may be desirable to accord at least partial scores to retrieved documents which contain no useful content but which link to it. Indeed, such pages may be the best answers if the searcher is seeking a point from which to browse rather than particular content. A number of issues in indirect relevance are canvassed in the discussion of the TREC-8 Small Web Task (Hawking et al. 1999).

2. Evaluations should be carried out at frequent intervals to maintain an up-to-date picture of relative performance and to smooth out performance "glitches."
3. The cost of evaluations based on Type C queries is largely attributable to assessor salaries. Consequently, it is desirable to maximize the value of assessments made. Potentially, the high correlation between different measures, might be used to limit how many pages per topic need be judged. There may be greater value in judging fewer documents per topic for a larger number of topics, particularly as measures derived from the very top of the ranking may be the best predictors of user satisfaction for certain information need types.

4. In comparing performance from one evaluation to the next, a normalising baseline is needed. The median performance of a large number of engines could be used in this role but would be unable to detect a general increase or decline in performance. Use of a standard set of queries in each evaluation does not solve this problem because of variations in a human's judgments over time and because of the possibility that search engine companies would tune their results for these queries. The only possibility we have been able to identify involves including a known, unchanging engine, operated by the evaluators in each successive evaluation and using its (rejudged) performance as a normalising baseline.
5. To permit fair comparison with the TREC systems, the present study did not penalise search engines for returning defunct pages in the results list. Future evaluations which do not compare against TREC should treat a dead link as a worthless answer. However, it may be considered necessary to make several attempts to access the page, in case inaccessibility is only temporary.
6. Considerable care is needed in writing scripts for automatically posing queries to search engines and retrieving results. The format of result pages changes from time to time and may cause errors. Thorough sanity checks should be applied to all results obtained by such automatic means.
7. In presenting documents for assessment, it is important to do so in a way which simulates normal Web search. See point 10 in figure 11.

   Figure 11 summarises the features which we believe should be present in future Web search evaluations.

7. Conclusions and future work

We have performed a Web search engine evaluation using appropriate scientific methodology on a larger scale than previously published studies. This evaluation shows that choice of search engine makes a difference, that early precision (P@5) is not significantly correlated with index coverage, and that the best search engines are now using algorithms approaching the effectiveness (on relevance based tasks) of those used by TREC-8 Large Web task participants. We found an unexpectedly high correlation between P@1–5 results obtained by us and those obtained by a study conducted by others a year earlier. There were very high intercorrelations between four different measures derivable from the top 7 live results.

Further work is needed to refine the classification of information need types in Web search and to identify the appropriate evaluation paradigms for each type. User studies to determine the effectiveness measures which best correlate with searcher satisfaction ratings would be very valuable. User studies will also be required to determine the proportion of Web queries which are derived from each information need type, as inspection of query logs frequently fails to reveal the searchers' intentions.

In order to present a complete picture of search engine performance, future studies should include queries derived from range of commonly expressed information need types and use appropriate judgments and measures for each type. To reduce the multi-dimensional results from such a study to a single index of effectiveness will require that each dimension be weighted by frequency of occurrence of the corresponding information need type.

Notes

1. Table 1, p. 148.
2. The work was not formally published until 1999.
3. p. 146.
4. This term captures the intention but not the long-drawn-out and incomplete nature of actual crawls.
5. Any item accessible via the Web is conventionally called a *Web page*. When printed, a Web page may cover many pages or just a few lines.
6. A stopword is a functional word like "the", "and", "of", and "but".
7. We don't know why this was chosen.
8. The RAT does have a mode in which relevance of documents is automatically determined using the concepts, criteria and evidence supplied by the judge. Any discrepancies between the automatic and the manual judgments result in an alert, which requires the judge to either accept the automatic judgment or to override it and record a reason. This mode was not used in the present experiments.
9. NIST judges working with documents from the same TREC collections judge several hundred documents per day on average.
10. In TREC, average precision for a topic is based only on precision values at each point in the ranking where a relevant document is retrieved. Average precision is obtained by dividing the sum of these values by the number of relevant documents in the collection. Here, it is unknown how many relevant documents there really are and rankings are cut off at 7 rather than 1000. Accordingly, here the sum is divided in each case by 7 and the result is called TREC-style average precision at cutoff 7. TSAP is an average of precisions and unlike the true TREC measure does not include a recall component. A system for which all 7 documents retrieved for a query were judged relevant would score 1.0 on this measure; One for which only one retrieved document was relevant would score between $1/7$ and $1/49$, depending upon the position of the relevant document in the ranking. Note that although for each individual topic TSAP scores are a constant multiple of what the TREC average precision would have been at the same cutoff, the multiplier varies from topic to topic. Consequently, across-topic averages of the two measures are not simply related.
11. For some engines we requested a generous number of results in a single page. For other engines we didn't know the necessary URL syntax to do this and requested results page by page.
12. In this MANOVA, the repeated factor was Queries, the independent group factor was Search Engines and the dependent measure was P@1–5. A multiple rather than a univariate model was selected as the latter is sensitive to violations in sphericity which are common in repeated measures designs.
13. A related samples test was appropriate because each query was processed by each of the 20 engines.
14. The Cochran Q test is a non-parametric test for determining if a series of related samples differ in circumstances where data is based on a nominal or dichotomised ordinal scale of measurement. The P@1 data was of the latter type as P@1 scores for individual queries must be either 1 or 0. A related samples test was appropriate because each query was sent to each of the 20 engines.
15. The Mann-Whitney is a non-parametric test of the difference between two independent groups. A non-parametric test was selected because of the small sample size of the TREC group.
16. An ideal merging algorithm is one capable of producing a merged list in which all relevant documents in the primary results lists appear first.
17. Note that high precision is also required.

References

Alta Vista Company. Alta Vista web page. `http://www.altavista.com/`.

Blair DC and Maron ME (1985) An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28(3):289–299.

Buckley C and Voorhees E (2000) Evaluating evaluation measure stability. In: Proceedings of SIGIR'00, New York, 2000, pp. 33–40. ACM Press.

Cleverdon C (1997) The Cranfield tests on index language devices. In: Jones KS and Willett P, Eds. Readings in Information Retrieval, Morgan Kauffman, San Francisco, pp. 47–59. (Reprinted from Aslib Proceedings, 19, 173–192).

Ding W and Marchionini G (1996) Comparative study of web search service performance. In: Proceedings of the ASIS 1996 Annual Conference, Oct. 1996, pp. 136–142.

Eisenberg M and Barry C (1998) Order effects: A study of the possible influence of presentation order on user judgments of document relevance. Journal of the American Society for Information Science, 39(5):293–300.

Electric Knowledge LLC. Electric Monk home page. `http://electricmonk.com/`.

Gordon M and Pathak P (1999) Finding information on the world wide web: The retrieval effectiveness of search engines. Information Processing and Management, 35(2):141–180.

Hawking D and Thistlewaite P (1997) Overview of TREC-6 very large collection track. In: Voorhees EM and Harman DK, Eds. Proceedings of TREC-6. Gaithersburg, MD, Nov. 1997, pp. 93–106. NIST special publication 500–240, `http://trec.nist.gov`.

Hawking D, Thistlewaite P and Harman D (1999) Scaling up the TREC collection. Information Retrieval, 1(1):115–137.

Hawking D, Voorhees E, Bailey P and Craswell N (1999) Overview of TREC-8 Web Track. In: Proceedings of TREC-8. Gaithersburg, MD, Nov. 1999, pp. 131–150. NIST special publication 500–246, `http://trec.nist.gov.`

Koster M The web robots pages. `http://info.webcrawler.com/mak/projects/robots/robots.html`.

Lawrence S and Lee Giles C (1999) Accessibility of information on the web. Nature, 400:107–109.

Leighton HV and Srivastava J (1999) First 20 precision among world wide web search services (search engines). Journal of the American Society for Information Science, 50(10):882–889.

Lynx. Lynx browser home page. `http://lynx.browser.org`.

Salton G and Lesk ME (1997) Computer evaluation of indexing and text processing. In: Jones KS and Willett P, Eds. Readings in Information Retrieval, Morgan Kauffman, San Francisco, pp. 60–84. (Reprinted from Journal of the ACM, 15, 8–36).

Silverstein C, Henzinger M, Marais H and Moricz M (1999) Analysis of a very large web search engine query log. SIGIR Forum, 33(1):6–12. Previously available as Digital Systems Research Center TR 1998–014 at `http://www.research.digital.com/SRC`.

Voorhees EM and Harman DK (1998) Eds. Proceedings of TREC-7, Gaithersburg, MD, Nov. 1998. NIST special publication 500–242, `http://trec.nist.gov`.

Voorhees EM and Harman DK (1999) Eds. Proceedings of TREC-8, Gaithersburg, MD, Nov. 1999. NIST special publication 500–246, cm http://trec.nist.gov.

Voorhees EM and Harman DK (1996) Overview of the fifth Text Retrieval Conference (TREC-5). In: Voorhees EM and Harman DK, Eds. Proceedings of TREC-5, Gaithersburg, MD, Nov. 1996, pp. 1–28. NIST special publication 500–238, `http://trec.nist.gov`.

Voorhees EM (1998) Variations in relevance judgments and the measurement of retrieval effectiveness. In: Croft WB, Moffat A, van Rijsbergen CJ, Wilkinson R and Zobel J, Eds. Proceedings of SIGIR'98, Melbourne, Australia, August 1998. pp. 315–323.