Which Search Engine is best at finding Online Services?

David Hawking and Nick Craswell
CSIRO Mathematical and Information Sciences
GPO Box 664
Canberra, Australia 2601
61-2-6216 7060
{David.Hawking,Nick.Craswell}@cmis.csiro.au

Kathleen Griffiths
Centre for Mental Health Research
Australian National University
Canberra, Australia 0200
61-2-6125 9723
Kathy.Griffiths@anu.edu.au

ABSTRACT

We report results for an independent, blind evaluation of the performance of 11 commercial search engines on 106 online service queries and on 54 topic relevance queries. We found a strong correlation between performance on the two types of query and significant differences between engines.

Keywords

Search Engines, Evaluation, Online Service Queries.

1. SEARCH ENGINE EVALUATION: BACKGROUND AND METHODOLOGY

Independent quality evaluation of commercial search engines facilitates informed consumer choice and may also lead to a general raising of search quality. As well as various informal and subjective comparisons in the media, a number of academic studies have attempted to compare the performance of search engines on a scientific basis. [2, 3, 4, 5, 6]

Most Information Retrieval experiments and all of the published scientific evaluations of commercial search engines have measured the ability of search systems to retrieve documents which are relevant to a topic of interest. However, queries submitted to commercial search engines reflect a range of different information needs. Recently, it has been argued that search engine evaluation methodology should be extended to reflect this broader reality [1, 4].

Accordingly, we have carried out evaluations involving query sets arising from a range of information need types. Here we report results of two evaluations conducted in October, 2000. One covers traditional topic relevance and judges a page *relevant* if it is both on topic and contributes some additional information, however small, which is not implicit in the query[1]. A newspaper report about selling flowers over the Internet would be relevant to the query "e-flowers".

The second evaluation covers search for online services[2] in which a page is judged *useful* if it allows the searcher to directly access a particular Internet service. A person wishing to send flowers via the Internet would not be satisfied with newspaper articles discussing the topic. Instead they would require a page which allowed them to initiate the desired transaction.

2. EXPERIMENTS

A set of 54 topic relevance queries and a set of 106 online service queries were submitted to commercial search engines in October 2000. The topic relevance queries were identical to those used in the TREC-8 Large Web Task [7] and in our September 1999 evaluation of commercial search engines [4]. The online service queries were identical to those used in the TREC-9 Large Web Task [7].

Examples queries (prior to stopword elimination) are:

```
Topic Relevance:
21247 where can i find information on russia?
21475 how does a digital camera work?
21826 where can i find information on the bahamas
22539 who are the current supreme court justices?
22610 thalidomide and multiple sclerosis

Online service:
20587 where can i do an iq test?
20757 where can i order flowers online?
20881 where can i find icq hacks?
20931 where can i find animal sounds?
20969 where can i download computer games
```

Queries minus stop words such as "where", "how" and "who" were submitted by automatic script with appropriate logic to separate out search results from on-site links and advertisements. Because of the risk of errors in this automated process, we subjected our results to quite thorough validity checking and in response to potential problems detected during this exercise, we eliminated a number of runs from the analysis.

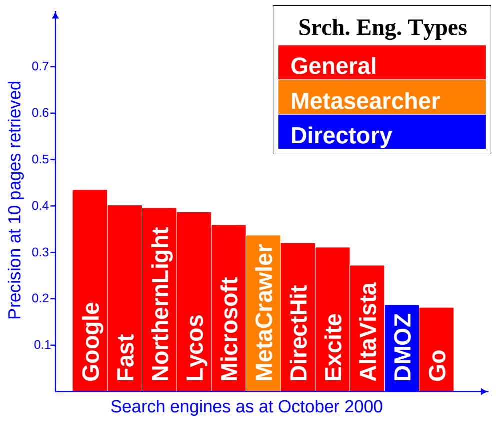

Figure 1: P@10 for 54 topic relevance queries.

[1] This fits into Andrei Broder's *informational* category [1].
[2] approximating Broder's *transactional* category

Result documents were retrieved from the appropriate server. All the documents for a query from all the search engines were pooled and presented to an independent assessor (recruited by CSIRO/ANU) via a judging interface described in [4]. Dead links were judged "not relevant" or "not useful".

Once judging was complete, the results lists for each engine were evaluated using precision at ten documents retrieved (P@10). This corresponds closely to how many good answers there are on the first page of results presented to a searcher.

3. RESULTS AND DISCUSSION

Figures 1 and 2 show the performance of the 16 commercial engines on the two different types of queries. A multiple analysis of variance (MANOVA) of the P@10 data, confirms that there is a significant difference in the performance of the search engines on both types of query. (topic relevance: $F(10, 44) = 6.28, p < 0.001$; online service: $F(10, 96) = 10.12, p < 0.001$.)

Multiple pairwise comparisons using the Least Significant Difference test were conducted. For the topic relevance queries, Google was significantly better than all the engines except Fast and NorthernLight ($p < 0.05$). For the online services queries, Google was significantly superior to all but NorthernLight ($p < 0.05$).

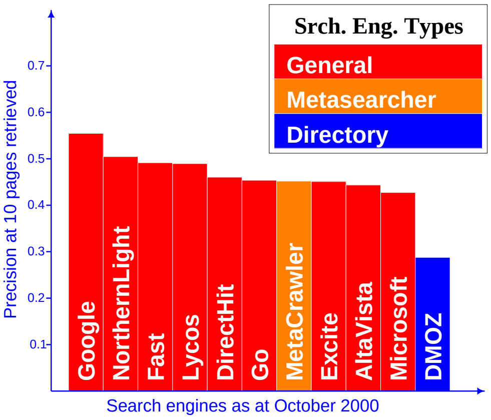

Figure 2: P@10 for 106 online service queries.

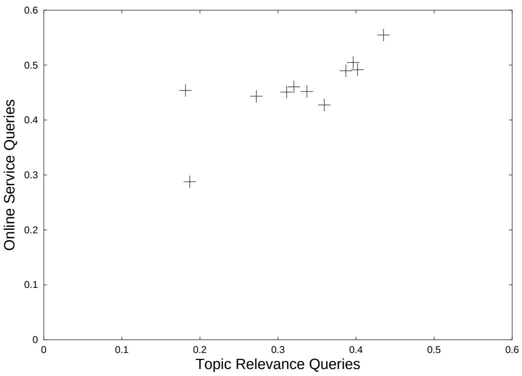

Figure 3: P@10 performance for topic relevance v. online service queries. Pearson $r = 0.76$. $p < 0.01$

As shown in Figure 3, there is a strong positive correlation between P@10 performance on the two types of query. A paired t-test revealed that performance on the online service queries was superior to that on topic relevance. ($t(10) = 7.85, p < 0.001$) This may be an indication that search engines are tuned for online service queries but it is equally possible that the topic relevance queries were harder[3].

4. CONCLUSIONS AND CAVEATS

To our knowledge this is the first published study to investigate search engine performance on online service queries. We found a strong correlation between performance on online service and topic relevance queries.

We found performance differences between engines on both types of query *for these queries, on the P@10 measure, at the time we actually submitted the queries.* However, it is important to note that performance of engines can vary considerably over time. Furthermore, factors other than ranking performance, such as response time, user interface design, and coverage and freshness of indexes may be important. We didn't evaluate *category* results returned by directory services nor did we evaluate the quality or authoritativeness of the result pages.

We are presently conducting an evaluation of how effective commercial search engines are at finding site entry pages (eg. homepages). We hope to report results in the final poster presentation.

5. REFERENCES

[1] Andrei Broder. Invited Talk at TREC-9 Conference, November 2000.

[2] Wei Ding and Gary Marchionini. Comparative study of web search service performance. In *Proceedings of the ASIS 1996 Annual Conference*, October 1996.

[3] Michael Gordon and Praveen Pathak. Finding information on the world wide web: The retrieval effectiveness of search engines. *Information Processing and Management*, 35(2):141–180, March 1999.

[4] David Hawking, Nick Craswell, Peter Bailey, and Kathy Griffiths. Measuring search engine quality. *Information Retrieval*, 2000. In Press.

[5] David Hawking, Nick Craswell, Paul Thistlewaite, and Donna Harman. Results and challenges in web search evaluation. *Proceedings of WWW8*, 31:1321–1330, 1999. `http://www8.org/w8-papers/2c-search-discover/results/results.html`.

[6] H. Vernon Leighton and Jaideep Srivastava. First 20 precision among world wide web search services (search engines). *Journal of the American Society for Information Science*, 50(10):882–889, 1999.

[7] National Institute of Standards and Technology. TREC home page. `http://trec.nist.gov/`, 1997.

[3] It is well known to Information Retrieval researchers that batches of queries vary considerably in their general degree of difficulty.