Is it fair to evaluate Web systems using TREC ad hoc methods?

Nick Craswell and Peter Bailey
Department of Computer Science,
The Australian National University
{nick.craswell,peter.bailey}@cs.anu.edu.au

David Hawking
Centre for Mathematical and Information Sciences, CSIRO
david.hawking@cmis.csiro.au

1 Introduction

Experiments using TREC-style topic descriptions and relevance judgments have recently been carried out for the first time over real Web data. One interesting result is that systems of TREC VLC Track participants [5] were more effective than live Web systems [6]. A number of factors could explain this, but one important possibility is that Web and TREC ad hoc systems are solving different problems. If this were the case, it would be unfair to use TREC ad hoc evaluation methods as a benchmark for the success of Web retrieval systems.

Here we informally discuss four ways in which a user's interaction with a Web information retrieval system might differ from the interaction modelled in TREC ad hoc experiments. They are: importance of hyper-links, different query topics, more variation in document quality and more documents with exact duplicates.

2 Recent results

The annual experiments in the Text Retrieval Conference (TREC) [8] ad hoc task compare the effectiveness of retrieval systems applying new queries over archived data. Each experiment is based on a test collection comprising documents, topics and relevance judgments, which serves as a laboratory model of a user performing searches. Each topic is a natural language statement of the user's information need, and is used both for generating system-specific queries and as a basis for judging the relevance of documents. Systems are compared on the basis of the quality of the ranked lists of documents which they return in response to queries derived from the topics. Effectiveness

	Web Search Engines		TREC Systems (on VLC2)	
	range	median	range	median
P@20	.231 - .377	.289	.345 - .442	.397

Table 1: This table summarises the effectiveness of Web and TREC systems, given short queries (from topic titles only).

measures are based on the positions of relevant documents in the ranked list — the more relevant documents in higher ranks the better.

The document set for the ad hoc task comprises two gigabytes of news articles and government records. The topics are research-oriented rather than, for example, questions which require a simple factual response or requests for known items. Relevance judgments are binary: every document is either relevant or irrelevant to the query.

TREC-7 VLC (Very Large Collection) experiments were very similar in design to those of the ad hoc task, using identical topics and judging criteria. However, the VLC used a 100 gigabyte snapshot of Web data and a necessarily incomplete set of relevance judgments. The lack of complete judgments meant that systems could only be compared on the basis of early precision. The measure used was *precision at 20 documents retrieved* which is defined as the proportion of the top twenty documents retrieved which were judged relevant.

At the same time as assessors were judging VLC track documents, results from five popular live Web systems [6] were presented to them for judgment. Live systems each index different and overlapping sets of Web documents, rather than the 100 gigabyte VLC. However, results were judged in the same way. A summary of a subset of the effectiveness results obtained is presented in Table 1. Please see [6] for full details.

3 Discussion of Results

The observed difference in effectiveness between TREC and live systems may mean that Web systems are truly less effective in finding relevant documents. Several of the TREC systems use the best available relevance ranking algorithms, while live systems might use inferior algorithms due to considerations of query processing cost.

The observed difference may also stem from problems with the experiment, as discussed in [6]. Each live system indexes a different and changing document set, while VLC systems indexed a fixed document set, so results are not strictly comparable. Comparisons of ranking methods in isolation will only become feasible when live systems come into the laboratory to be

tested over a standardised test collection.

The present paper considers a number of other possible differences between behaviour modelled in the TREC ad hoc task experiments and the behaviour of real users when searching on the Web. These too might explain the difference in effectiveness between TREC and live systems.

3.1 Difference 1: The importance of hyper-links

Hypothesis: Ranking methods which take link structure into account might return documents that are irrelevant but rich in links to relevant documents, and which are, in a Web context, comparable in value to a relevant document. If this value had been recognised in the VLC evaluations, then Web systems returning such documents would be rated more highly.

Inter-document references are not unique to hyper-text documents. For example, a TREC ad hoc document might say "see last week's news article". However, links are more common and much easier to follow on the Web, so we consider a hyper-text case to be different. In fact, in the Web case, it may be as easy to follow a link to a new page as to scroll down to the next section of the original document.

Because it is so easy to follow links on the Web, a large document such as a book, manual or thesis is often split into multiple Web documents, each of which is a chapter or section of the whole. It is also usual for there to be a table of contents document, which contains links to all the other documents. If several sections of a book are relevant to the user's query, then a system which returns those sections is performing well. The question is whether a system that returns the table of contents is also performing well. From the user's point of view, the table of contents document not only allows them to find the relevant documents but also provides vital information on context: who wrote the book, how else it is published and what other sections it has. However, if links are not taken into account, the table of contents is likely to be judged irrelevant and therefore considered a worthless result.

The same argument applies to site home pages. A home page may not have detailed information which would be judged relevant on its own, but may provide important context and links to relevant documents. The argument also applies for pages of links, such as bookmarks pages or Web directories like Yahoo!'s [9] Given the query "Falkland petroleum exploration", a Web search system might return a page which links to all the most authoritative sites on oil in the Falklands. Based on its text only, the link page would be judged irrelevant, but because of its links it is far from useless. Automatic methods for finding pages with useful links, pages called "hubs", have already been proposed [7] and are the subject of current research.

We have two preliminary suggestions on how to test the above hypothesis. Either could be applied using a pre-existing Web data test collection, with no special input from relevance assessors, providing that (practically) all relevant documents have been found. Each attempts to assign value to "link-relevant" pages, which do not contain relevant content (not content-relevant) but link directly to content-relevant pages.

The first suggestion is to assign all content-irrelevant pages a status of link-relevant if they contain enough good links. Perhaps two or three links to content-relevant documents would qualify a document as link-relevant. Then, both content-relevant and link-relevant pages would be considered as relevant when applying an effectiveness measure.

The other suggestion is to assign each document a non-binary relevance score ranging from zero to one. Content-relevant pages would score one and content-irrelevant pages without good links would score zero. Content-irrelevant pages with links to content-relevant pages would be given an intermediate score, perhaps 0.5. Effectiveness measures could then be applied by adding scores. For example, precision at twenty would be the sum of document relevance scores in the top twenty divided by twenty (precision would still always range from zero to one). Further work is needed to find the most appropriate method for assigning scores to link-relevant pages.

3.2 Difference 2: Different topics

Hypothesis: TREC topics are representative of the information needs of a researcher (e.g. someone writing an article for a magazine). By contrast, Web information needs can also be directory or subject catalogue lookups, known-item searches, or question-answering queries. Web search engines might not be tuned to process TREC-style queries because such queries only constitute a tiny fraction of their overall workload. Thus the fact that Web systems are less effective on TREC ad hoc topics might be unimportant.

It is important to better understand the nature of Web information needs. Participants in the TREC-8 Large Web Track will begin to do so by working with 10,000 real Web queries, instead of the 50 TREC-8 ad hoc topics. The large number of queries is to test system efficiency and to prevent tuning to particular query domains; relevance assessments will be carried out over only 50 of them.

The queries have been chosen arbitrarily from (censored) query logs of AltaVista [1] and Electric Monk [4]. Each log contained 100,000 natural language queries, as submitted to the engine. Queries with obvious adult content were removed at the request of some participants.

Natural language queries, usually comprising a sentence in the form of a question, were used instead of the very short keyword queries often sub-

mitted to Web search engines. This is because relevance judgments on very short queries might be uninteresting. If the query is "cat", all documents containing "cat" might be judged relevant. Natural language queries are representative of real Web user's information needs, our primary goal, and also to allow more strict relevance judging. Most "cat" pages do not address the information need implicit in the following query from the logs:

Where is the fun site with the cat that looks like Hitler?

3.3 Difference 3: More variation in document quality

Hypothesis: Some Web systems attempt to favour high quality documents in their ranking. However, TREC judgments are binary and give the same value to all relevant (i.e. on-topic) documents, even those of low quality. Consequently, TREC scores may under-value results produced by Web search engines.

Web documents range widely in quality. The Yahoo! directory only includes pages judged to be of high quality by Yahoo! employees. New systems such as Google [2] and CLEVER [3] attempt to automatically determine page quality, using hyper-link voting.

The concept of Web document quality is not yet well understood. It would be interesting to determine whether systems such as Google and CLEVER succeed in identifying high quality pages. It would also be interesting to build an evaluation framework that takes result quality into account. We do not know how to do this at the present time.

3.4 Difference 4: More precise duplicates

Hypothesis: The prevalence of exact duplicate pages on the Web requires Web search engines to detect and eliminate them. By contrast, TREC systems do not eliminate duplicate documents. Under ad hoc evaluation methods, every duplicate of a relevant document counts just as much as the original. Consequently, systems which return duplicates may receive artificially inflated (or deflated) scores on a TREC task, particularly if Web data is used.

Duplication of information in search results, where top ranked results all contain much the same information but not the whole picture, is a general problem we will not address here. However, on the Web several precise duplicates of a document may exist at different URLs because of host name aliasing, mirroring and HTTP server setup. This extreme form of duplication is easily detected, and is usually removed by Web search systems. Such duplication can also cause problems in evaluation. In the TREC-7 VLC track, when a few runs were unofficially judged to the 100 document mark,

one system benefitted from returning 67 copies of the same relevant document! Others were penalised for returning multiple copies of pages judged irrelevant.

Duplicates could be eliminated from TREC participants' search results any time before judging, by the participants or by TREC officials. In any case, elimination of duplicates would prevent wastage of relevance judging resources and prevent any problems in evaluation caused by the presence of duplicates. The size of the latter problems could be measured empirically, by comparing the measured effectiveness of results lists with and without duplicates. If the presence of duplicates caused unpredictable differences in systems' effectiveness, it would be important to remove them before evaluation.

4 Conclusion

We have pointed out that the user behaviour and preferences modelled in TREC ad hoc task experiments might differ from that of users searching Web documents, due to:

- The importance of hyper-links in Web documents,
- The difference between TREC and Web topics,
- The greater variation in Web document quality, and
- The presence of large numbers of precise duplicate documents on the Web.

In some cases we have also suggested ways of compensating for these differences in future evaluation experiments over Web data.

Adjusting experiments to take each of the problems into account would make some difference in experimental results, but the difference might not be large. For example, removing duplicates, so that no system can benefit or suffer from returning duplicates, might have little impact on a system's mean effectiveness over 50 topics. If duplicate elimination can be shown to have a consistently small effect, experimenters might choose to leave the duplicates in.

However, new experiments are required to determine the magnitude of such effects and to answer the question posed in the title. In addition, close attention to Web-specific issues could lead to new insights. Use of real Web queries might shed light on the needs of real Web users. Evaluation including hyper-links or page quality might reveal that link-based ranking methods of Google or CLEVER can successfully find hub pages, or that pages which relevance assessors find to be of high quality also have a high authority score or PageRank. Much more work is required to understand Web search and its evaluation.

References

[1] AltaVista. `http://www.altavista.com/`, 1999.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[3] Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[4] Electric Knowledge L. C. C. The electric monk. `http://www.electricmonk.com`, 1999.

[5] David Hawking, Nick Craswell, and Paul Thistlewaite. Overview of TREC-7 Very Large Collection Track. In Voorhees and Harman [8]. NIST special publication 500-?

[6] David Hawking, Nick Craswell, Paul Thistlewaite, and Donna Harman. Results and challenges in web search evaluation. In *Proceedings of WWW8, Toronto*, pages 243–252. Elsevier, 1999.

[7] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, San Francisco, California, 25–27 January 1998.

[8] E. M. Voorhees and D. K. Harman, editors. *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, Gaithersburg MD, November 1998. U.S. National Institute of Standards and Technology. NIST special publication 500-?

[9] Yahoo! What is the yahoo! home page?, 1999. `http://howto.yahoo.com/infodesk/yahoo_home_page.html`.