Workload Sampling for Enterprise Search Evaluation

Tom Rowlands
Dept. of Computer Science
Australian National University
and CSIRO ICT Centre
tom.rowlands@ieee.org

David Hawking
CSIRO ICT Centre
Canberra
Australia
david.hawking@acm.org

Ramesh Sankaranarayana
Dept. of Computer Science
Australian National University
ramesh@cs.anu.edu.au

ABSTRACT

In real world use of test collection methods, it is essential that the query test set be representative of the work load expected in the actual application. Using a random sample of queries from a media company's query log as a 'gold standard' test set we demonstrate that biases in sitemap-derived and top n query sets can lead to significant perturbations in engine rankings and big differences in estimated performance levels.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Performance, Measurement

1. INTRODUCTION

Test collections, e.g. those of TREC [6], are well established as the orthodox tool for evaluating, tuning and comparing retrieval systems. Test collections typically consist of documents, queries and, for each query, a set of known useful answers.

Previous research has addressed the choice of appropriate effectiveness measures for particular purposes. Buckley and Voorhees [2] have also studied the size of query sets needed to achieve stable rankings of alternative retrieval systems on those measures.

Various authors have addressed the issue of reducing the cost of obtaining judgments, e.g. Clarke and Cormack [3]. Sanderson and Zobel [5] argue that judging effort is better expended on shallower judging of a larger range of topics than on deeper judging of a small set. Hawking and Zobel [4] compare the value of different types of ranking evidence in the search of an enterprise web using a 'best page' evaluation. In this type of evaluation, retrieval systems are rated solely on their ability to retrieve the most useful document at or near the top of the ranking. Other relevant but less useful documents are ignored.

Here, we consider the choice of the queries comprising the test set. We are interested in methods for evaluating retrieval effectiveness in real search applications such as an enterprise search facility, a library search engine or a newswire retrieval system. In this context, it is important that effectiveness results obtained by testing should accurately reflect the experience of people actually using the application. Ideally, we would measure and average across every query instance submitted (the *workload*), but this is clearly impractical. We need to sample from the population.

Early Web search engine companies are believed to have focused tuning efforts on the n most frequently submitted queries.[1] This approach makes intuitive sense as the more popular the query, the more benefit gained by improving its results. However, it is likely that the n most popular queries constitute a biased sample of the workload. For example, they are typically shorter than average. More recently, some large scale studies based on Microsoft's web search engine [1] have used random sampling in order to better reflect the end-user 'experience'.

Hawking and Zobel evaluated using highly popular, medium popularity, randomly chosen and *sitemap-derived* query sets in order to confirm that conclusions about the value of topic metadata in retrieval were not dependent upon the query set. They found that while the pattern of results from sitemap-derived query sets were very similar to those from the other sets, there was some exaggeration of the performance of anchortext-based evidence.

Here, we use a methodology similar to Hawking and Zobel but look at how well popular and sitemap test sets predict the 'gold standard' measures obtained from an unbiased sample of the workload of the search facility provided on the external website of a media company. We use best page evaluation as, anecdotally, users and purchasers of enterprise search services expect simple queries to return the obvious answer at rank one. For example, a search for a product name on the manufacturer's website should return that product's home page at rank one. Further, best page evaluation allows sitemap queries to be used and permits a larger sample of queries within the bounds of judging effort.

2. EXPERIMENTS

Anonymous Media Organisation's web site (7.6×10^5 pages) and six months of query log data (2005–6) were used for the experiments.

The test sets are listed in Table 1. A random sample of three hundred queries was taken from Anonymous Media Organisation's query log. The query log contained 2.9×10^6 casefolded queries recorded between December 2005 and

[1] personal communication

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

Table 1: Test sets: n is the number of queries for which best answers were found. $\overline{terms}$ is the mean number of words in those queries. *Workload proportion* is the portion of the search engine's total workload represented by the test set.

Testset	n	$\overline{terms}$	Workload proportion (%)
sample	135	2.6	4.7
popular	103	1.6	9.4
sitemap	752	2.4	0.3

May 2006. Judgements were made by an author who was familiar with the organisation but not a domain expert. The judge found the 'best page' available to answer the query. This page was, in the judge's opinion, the page most likely to be useful to an issuer of the query. In a few cases, there was more than one 'best' page. The 'popular' test set is derived from the most popular 132 queries for the site. In both the sample and popular test sets, queries for which answers could not be found were not considered.

The 'sitemap' test uses the text of each link on the organisation's sitemaps as a query and the target as the best answer. This test has the advantage that the best answers are chosen by the organisation itself (thereby avoiding experimenter bias and/or ignorance) with no need for additional judging. The proportion of overall workload represented by the sitemap testset is low despite the large number of queries.

In all cases, redirections have been taken into account. Only the top ten results from each result set were examined. In the sitemap case, the site's main sitemap was removed before indexing, but sub-site sitemaps, from which the majority of the sitemap test's queries were derived, were not.

Each test set was run against four different retrieval engines, labelled E1–E4 and chosen to span a wide variety of different retrieval methods. Precise details of differences between engines are not important to the central issue of this paper. However, E3 relies entirely on anchortext evidence and E4 relies entirely on click data. Engines were compared using mean reciprocal rank of the first correct answer (MRR1) as MRR1 is the obvious choice for best page evaluation.

Figure 1 illustrates that the estimated performance varies substantially depending upon the test set. Compared to the unbiased sample estimate, the popular query set strongly over-estimates the effectiveness of E4 (click data based) and under-estimates the performance of E1 and E2. Similarly, the sitemap test over-estimates the performance of E3 (anchor text based) and under-estimates that of E2 and E4.

The ranking of the engines also varies considerably depending upon the test set as shown here. A '-' indicates a gap of more than 0.1 in MRR1.

```
sample:   E1 E2 E3 -- E4
popular:  E1 E2 E3 E4
sitemap:  E3 - E1 - E2 --- E4
```

3. DISCUSSION

Sitemap tests are very appealing from an experimental point of view but our study has demonstrated appreciable biases. Popular queries, on first inspection, sound like a reasonable way of tuning a search engine to the needs of many users, but like the sitemap test show significant bias. The popular tests did not, in this case, change the rankings of the search engines, but may in other cases.

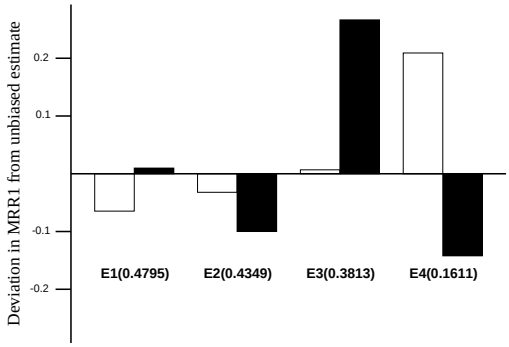

Figure 1: Deviations of performance estimates derived from popular (white) and sitemap (black) sets from the unbiased sample estimate, represented by the horizontal axis. The deviations are obtained by subtracting MRR1 scores. The unbiased sample estimate for each engine is shown in parentheses.

In this work, the *population* we have studied is actually of queries received by the search engine which have an easily identifiable best answer, rather than the total workload.

4. CONCLUSION

Traditional test collections, such as those used in TREC, facilitate the direct and reproducible comparison of search methods. However, unless the queries in a test collection form an unbiased sample of a real search workload, engine rankings and performance estimates are not likely to reflect real world performance. We suspect that there are substantial biases in the selection of the made-up queries used in many TREC and INEX evaluations, but we did not have access to a set of such queries coupled with a matching collection and query logs.

However, we have shown that, despite their attractions, neither sitemap nor top n query sets provide unbiased estimates of performance across an actual workload.

5. REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proc. ACM SIGIR '06*. ACM Press, 2006.

[2] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proc. ACM SIGIR '00*. ACM Press, 2000.

[3] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proc. ACM SIGIR '98*. ACM Press, 1998.

[4] D. Hawking and J. Zobel. Does topic metadata help with web search? *JASIST*, 58(4), 2007.

[5] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proc. ACM SIGIR '05*. ACM Press, 2005.

[6] E. Voorhees and D. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.