Test collections used to estimate real world retrieval performance must use queries that are representative of the real world workload.

We used several samples of queries from a media company’s query log to demonstrate biases in performance estimates.

Test collections, like in TREC, are a well established tool. They comprise documents, queries and known good answers. Previous research has addressed the choice of appropriate effectiveness measures, number of required queries and reducing judgement cost.

Here, we consider choice of query set.

In real world search evaluation, test results should accurately reflect users’ experiences. Ideally, we would measure every query instance submitted (the workload), but this is impractical. Instead, as in other sciences, we sample from the population.

Experiments

We used an anonymous media organisation’s web site (760 000 pages) and a sample from six months of query logs. Best page judgements were made by an author.

The table shows the test sets. The ‘popular’ test set is derived from the most popular 132 queries. The ‘sitemap’ test uses the anchortext of each link on the sitemap as a query and the target as the best answer. In both the sample and popular test sets, queries for which answers could not be found were not considered.

Sitemap tests have the advantage that the best answers are chosen by the organisation. Only the top ten results from each result set were examined.

Each test set was run through four different engines, labelled E1–E4. Precise details are not important, but E3 relies on anchortext and E4 relies on click data. Engines were compared using mean reciprocal rank of the first correct answer (MRR1).

Results

The graph shows deviation from the sample test set. The red bars are the deviation of popular queries and the blue bars the deviation of the sitemap queries.

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 (anchortext based) and under-estimates that of E2 and E4. The ranking of the engines also varies considerably, as shown in the table.

Discussion

Sitemap tests are appealing but our study has demonstrated appreciable biases. Popular queries sound like a reasonable way of measuring an engine’s performance against the needs of many users, but like the sitemap test, show significant bias. In this work, the population we have studied is of queries received by the search engine which have an easily identifiable best answer, rather than the total workload.

Traditional test collections facilitate the reproducible comparison of engines. However, unless the queries in a collection form an unbiased sample of a real search workload, engine rankings and performance estimates are unlikely to reflect real world performance. We have shown that, despite their attractions, neither sitemap nor popular query sets provide unbiased estimates of performance across a real workload.