A Framework for Determining Necessary Query Set Sizes to Evaluate Web Search Effectiveness

Eric C. Jensen, Steven M. Beitzel, Ophir Frieder
Information Retrieval Laboratory
Illinois Institute of Technology
Chicago, IL 60616
{ej,steve,ophir}@ir.iit.edu

Abdur Chowdhury
Search & Navigation Group
America Online, Inc.
Dulles, VA 20166
cabdur@aol.com

ABSTRACT

We describe a framework of bootstrapped hypothesis testing for estimating the confidence in one web search engine outperforming another over any randomly sampled query set of a given size. To validate this framework, we have constructed and made available a precision-oriented test collection consisting of manual binary relevance judgments for each of the top ten results of ten web search engines across 896 queries and the single best result for each of those queries. Results from this bootstrapping approach over typical query set sizes indicate that examining repeated statistical tests is imperative, as a single test is quite likely to find significant differences that do not necessarily generalize. We also find that the number of queries needed for a repeatable evaluation in a dynamic environment such as the web is much higher than previously studied.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Traditional evaluation methodologies based on static test collections are difficult to employ in dynamic environments such as the web because the document collection, popular queries, and search systems themselves are constantly changing. The large size of the collection also makes it difficult to calculate recall, because the amount of effort required per query to evaluate large result pools is incompatible with the large number of queries that must be examined to represent the highly diverse query population (in which, for example, ~55% of all queries are repeated five or less times over a week) [1].

This confluence of factors, coupled with the associated need for repeating evaluations as conditions change, motivates the use of precision-oriented evaluation, where only a small number of the highest ranked results (top 10, etc.) are evaluated. In contrast to pooling results from all engines and only evaluating the best of the pool, these approaches evaluate every retrieved document at a shallow depth, enabling independent evaluation of any set of engines with reduced effort when focusing on specific questions such as "Does engine A significantly outperform engine B?" or "What is the best engine from this set?"

WWW 2005, May 10--14, 2005, Chiba, Japan.

2. PRIOR WORK

The Text Retrieval Conference (TREC) is the benchmark for developing static test collections. Although it focuses on deep pooling for recall-oriented evaluation, recent meta-evaluation has shown that precision-oriented metrics require more queries to provide a stable evaluation [2]. Attempts to employ the TREC methodology in evaluating web search engines have found that "search engine performances may vary considerably over different query sets and over time" [3]. These studies have not addressed the problem of determining the confidence that precision-oriented evaluations will generalize across query sets and have produced test collections with only 50-300 queries. Although classical statistical methods such as hypothesis testing (t-tests, etc) and formulas for calculating sampling error answer the questions of whether two engines are significantly different over a particular query set and if the sample size is sufficient, they do not answer the combined question of whether we should be confident that one engine will significantly outperform another across any randomly-selected query set of a given size.

3. METHODOLOGY

We propose a framework based on bootstrapping statistical hypothesis tests to estimate the likelihood that their significance values will be repeatable across other query sets [4]:

```
1. Randomly sample a distinct set of queries $Q$ with size $n$ from a query log.
2. For each query in $Q$, manually evaluate the union of the top $X$ retrieved results from each of the engines.
3. Calculate each engine's score for each query using the metric of interest, e.g. average precision (AvgP), reciprocal rank of the best page (MRR), etc.
4. For $B$ iterations:
    a. Randomly sample, with repetition, a set of queries $Q^*$ with size $m$ from the original set $Q$.
    b. For each pair of engines $E_A, E_B$
        i. If one-sided test with $H_A : E_A > E_B$ over $Q^*$ yields $p < \alpha$, increment $C_{E_A > E_B}$
        ii. If one-sided test with $H_A : E_B > E_A$ over $Q^*$ yields $p < \alpha$, increment $C_{E_B > E_A}$
```

5. $\mathrm{P}_{\mathrm{E_A > E_B}}(\mathrm{p} < \alpha \mid Q, m) = \dfrac{\mathrm{C}_{\mathrm{E_A > E_B}}}{B}$

This resultant probability estimate is the confidence that $\mathrm{E_A}$ will outperform $\mathrm{E_B}$ with significance α over any randomly chosen set of queries with size m.

4. EXPERIMENTATION

In order to validate our framework, we evaluated ten web search engines: Google, Yahoo, Wisenut, Teoma, Altavista, AllTheWeb, Lycos, Gigablast, MSN, and the MSN TechPreview. We randomly sampled a set of 896 queries from a log of all the queries submitted to AOL search for two days. We then submitted these queries to each of the engines, pooled the top 10 results from each engine together in a uniform interface (averaging 43 results per pool), and had assessors manually assign each result as relevant, non-relevant, or the single best result for that query. For each engine, we calculated the per-query scores for average precision cut off at 10 results and reciprocal rank of the best page. We found that average precision over the top 10 results produces more stable hypothesis test statistics than P@10, likely because it also measures the quality of the ranking, producing less discretized scores and therefore fewer ties.

We begin our analysis by examining the repeatability of hypothesis tests over varying samples using step 4 of our methodology. As the distributions of our metrics do not appear to fit any apparent distribution, we chose the non-parametric Wilcoxon paired signed rank test. We set the number of iterations $B = 2{,}401$ (which would yield a sampling error of 0.02 with 95% confidence if our p-values were normally distributed). We set $m = 850$, approximately 50 less than the entire set of 896 used as Q. In performing all the hypothesis tests between each pair of engines on each sub-sample, we found that of the tests that yielded significance of $p < 0.05$, 12% for MRR and 7% for AvgP were from engine pairs for which no such significant difference exists in a hypothesis test over all 896 queries. When using only a single hypothesis test to decide that one engine outperforms another, as is traditionally done, even sets as large as 850 queries have a disturbing likelihood of finding a significant difference that does not necessarily exist on other query sets of that size.

The next phase of our analysis focuses on examining $\mathrm{P}_{\mathrm{E_A > E_B}}(\mathrm{p} < 0.10 \mid Q, m)$ using step 5 of our methodology. Figure 1 shows the growth of this confidence with increasing sub-sample size m for two example engine pairs (anonymized). In the number of queries we have evaluated, we are able to conclude that E2 repeatably outperforms E3 with significance of $p < 0.10$ (their mean AvgP are .620 and .611, respectively). When examining E5 and E3, however, it is clear that it would take an unreasonable number of queries to find a repeatable difference between the two engines, and we would therefore conclude that those engines are tied. This sort of analysis permits evaluators to focus their efforts on those engines for which repeatable differences are likely to appear, quickly ruling out those engines that are not likely to be the best.

To invoke the framework for that sort of prediction, these confidence estimates must be reliable at the number of queries that have been evaluated. To determine the number of queries necessary to provide reliable estimates of confidence, we include the range of confidences (shown as error bars) estimated by employing our bootstrapping methodology on several distinct random samples used as the initial query set Q. As we did when examining the reliability of single hypothesis tests, we set the size of these initial samples $Q_1, Q_2, \ldots$ to be 50 more queries than the sub-samples we estimate on ($n = m + 50$). Estimates are only provided for sizes less than 800 as larger sets' error ranges may be artificially low due to their nearing the size of the available set. In contrast to single hypothesis tests, which were unreliable even with sets of 850 queries, bootstrapped confidence estimates for any pair of engines with $\mathrm{P}_{\mathrm{E_A > E_B}}(\mathrm{p} < 0.10 \mid Q_x, m = 650) > 0.984$ from any seed set Q_x of size $n = 700$ guarantee the confidence from using the entire set of 896 as Q is greater than or equal to 90% for that pair (using AvgP; this minimum threshold is 0.978 for MRR). We could not find reliable thresholds for higher levels of confidence with the available number of queries.

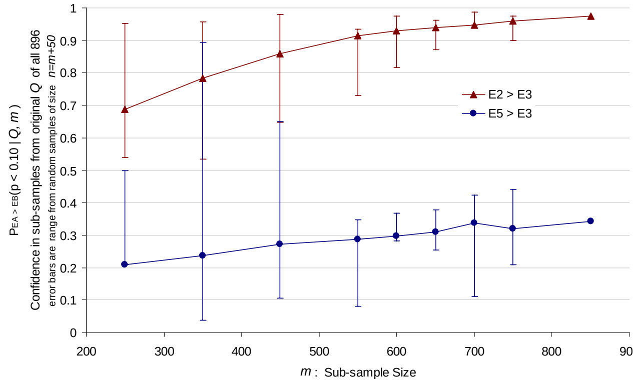

Figure 1: Growth of confidence for example pairs using AvgP

5. CONCLUSION

We have developed a framework for estimating the confidence that significant differences are repeatable across query sets. By building and making available (at http://ir.iit.edu/collections) a precision-oriented web search test collection of 896 queries, we have demonstrated the utility of our framework in reducing evaluation effort and also validated it; finding that at least 650 queries must be evaluated to reliably estimate significance.

6. REFERENCES

[1] Beitzel, S., et al. *Hourly Analysis of a Very Large Topically Categorized Web Query Log*. in *SIGIR*. 2004.

[2] Buckley, C. and E. Voorhees. *Evaluating Evaluation Measure Stability*. in *SIGIR*. 2000.

[3] Hawking, D. and N. Craswell, *Measuring Search Engine Quality*. Information Retrieval, 2001. **4**(1).

[4] Efron, B. and R.J. Tibshirani, *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. 1993.