

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

Motivation

- Manual search effectiveness evaluation is very labor-intensive.
- The dynamic nature of the World Wide Web demands conclusions that are repeatable across query sets.
- Typical web search engine traffic consists of many hundreds of millions of queries per day.
- Web queries are diverse and heterogeneous: 55% of all queries are repeated less than 5 times in a week.
- The set of popular web queries changes significantly over time: only 25% of the 30,000 most popular queries remain that popular one year later.
- The web collection is constantly changing: 8% each week, 55% each year (Ntoulas, et. al, WWW-2004).
- The web is too large to reliably calculate recall.
- Web search results change dramatically over time: on average for our 10 engines, only 61% of the results in the top 10 were still there three months later.

Precision-Oriented Web Test Collection

- Sampled 896 queries from an AOL web search log, preserving natural frequency (popularity) and query length distributions.
- Engines: Google™, Yahoo™, AltaVista™, AllTheWeb™, MSN™, Lycos™, Teoma™, MSN Tech Preview™, Wisenut™, and Gigablast™ (anonymized in no particular order as E1, E2, ... E10).
- Pooled all of the top 10 results from each engine, averaging 43 distinct URLs per pool.
- AOL assessors and students judged binary relevance for every result and the best page in the pool using a uniform interface that displayed title and snippet and allowed clickthroughs, but hid the originating engine.
- Total assessment time for this evaluation was 225 man-hours or ~15min per query.
- Across all engine pairs and queries, only 3.7 of the top 10 results are returned by both engines on average.
- Collection available at <http://ir.iit.edu/collections>

Methodology

- Randomly sample a distinct set of queries Q with size n from a query log.
- For each query in Q , manually evaluate the union of the top X retrieved results from each of the engines.
- 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.
- For B iterations:
 - Randomly sample, with repetition, a set of queries Q^* with size m from the original set Q .
 - For each pair of engines E_A, E_B
 - If one-sided test with $H_A : E_A > E_B$ over Q^* yields $p < \alpha$, increment $C_{E_A > E_B}$
 - If one-sided test with $H_A : E_B > E_A$ over Q^* yields $p < \alpha$, increment $C_{E_B > E_A}$
- $P_{E_A > E_B}(p < \alpha | Q, m) = \frac{C_{E_A > E_B}}{B}$ is the confidence that E_A will outperform E_B with significance α over any randomly chosen set of queries with size m .

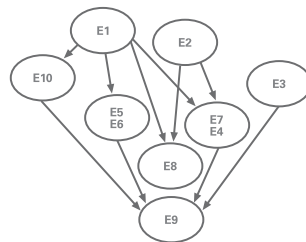
Conclusions

- At least 650 queries are needed to reliably estimate confidence in a collection such as this using bootstrapping.
- Single hypothesis tests are not reliably repeatable across query sets as large as 850 queries: up to 12% of results significant with 95% confidence on sets of size 850 are not significant over all 896 queries.

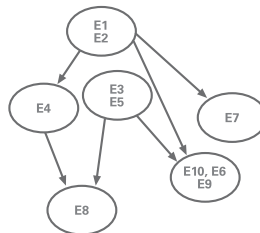
Mean Scores over 896 Manually Evaluated Web Queries

	AvgP	P@10	MRR
E1	0.632	0.690	0.359
E2	0.620	0.681	0.338
E3	0.611	0.676	0.312
E5	0.607	0.672	0.311
E4	0.600	0.667	0.283
E7	0.585	0.657	0.300
E10	0.573	0.634	0.313
E6	0.572	0.625	0.291
E9	0.568	0.635	0.241
E8	0.562	0.634	0.282

Repeatable Differences from all 896 MRR $P(p < 0.10) > 0.99 @ 850$

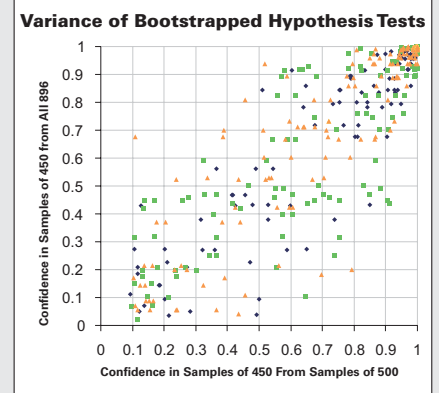
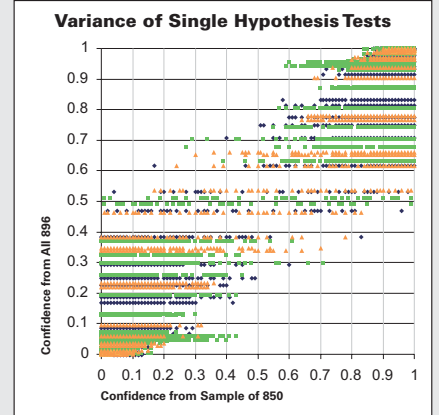
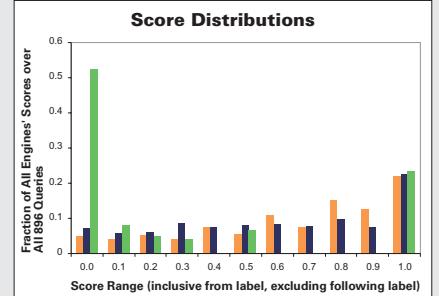


Repeatable Differences from all 896 Avgp $P(p < 0.10) > 0.99 @ 850$



Legend

Legend: Precision at 10 (orange square), Average Precision (blue square), Reciprocal Rank (green square)



Only p -values found at least 25 times out of 2,401 samples using single hypothesis tests are shown, as less common errors fill the entire graph. All values of bootstrapped tests found are shown.

Growth of confidence for example pairs using AvgP

