Using Manually-Built Web Directories for Automatic Evaluation of Known-Item Retrieval

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, Ophir Frieder
Illinois Institute of Technology
10 W 31 St.
Chicago, IL 60616
(01) 312-567-5150
{steve,ej,abdur,grossman,frieder}@ir.iit.edu

ABSTRACT

Information retrieval system evaluation is complicated by the need for manually assessed relevance judgments. Large manually-built directories on the web open the door to new evaluation procedures. By assuming that web pages are the known relevant items for queries that exactly match their title, we use the ODP (Open Directory Project) and Looksmart directories for system evaluation. We test our approach with a sample from a log of ten million web queries and show that such an evaluation is unbiased in terms of the directory used, stable with respect to the query set selected, and correlated with a reasonably large manual evaluation.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*

General Terms: Experimentation, Measurement

Keywords: IR Evaluation, Automatic Ranking

1. INTRODUCTION AND PRIOR WORK

Most of the work in evaluating search effectiveness has followed the Text REtrieval Conference (TREC) methodology of using a static test collection and manual relevance judgments to evaluate systems. Unfortunately, evaluating the effectiveness of web search engines creates many unique challenges that make a TREC-style evaluation problematic: the web is too large to perform deep manual relevance judgments of enough queries. In contrast to a test collection, the web is "live" data that is continually changing, and studies [1] have found that less than half of queries on the web are informational in nature. In the past two years, the importance of navigational queries has led TREC to incorporate manual known-item evaluations as part of the web track. However, these operate on a static test collection. These issues demand a new evaluation methodology that can be practically, repeatedly applied to evaluating search services on the live web.

The emergence of web directories such as the ODP and Looksmart enable a new type of automated assessment that allows relevant documents to be found on the "live" web. These directories are human-edited, category-driven collections of links. The basis for our approach is the assumption that human editors ensure high-quality, relevant content in a web directory. We lookup pages in directories whose hand-edited directory titles exactly match queries, and use the corresponding pages as a set of relevance judgments. Chowdhury and Soboroff have shown that this basic approach is viable using a single directory [4]. We show that the directory used does not significantly bias automatic evaluations. Additionally, we investigate the stability of the measure to ensure that it does not vary significantly when different query sets are used. One of the key advantages of our automated approach is that we are able to run thousands of queries where a manual approach is generally limited to a handful of queries. It has been shown that the ability to execute this volume of queries allows the error rates of evaluation measures to be examined [2]. Finally, we build a large set of manual relevance judgments to compare with our automatic evaluation method and find a moderately strong (.71 Pearson) positive correlation.

2. EVALUATION METHODOLOGY

We sampled a query log and pair queries with documents from an annotated collection, such as a web directory, whose edited titles exactly match the query. Queries that were successfully paired are issued to the search engines and the reciprocal rank of the corresponding document is stored. The mean reciprocal rank for each engine is used as a metric. For this methodology to yield a valid ranking of engines according to general known-item effectiveness, the set of query-document pairs needs to be reasonable, unbiased, and large enough to satisfy both sampling and stability. It has been shown that even manual assessors rarely agree on which document is the best for a query [6]. The heuristic of selecting documents as "best" by exact title matches is a source of error in our method, but we hypothesize that we can control this error if our selected documents are reasonably good, and we use enough of them. Two other factors that must be controlled in this methodology are bias in the queries sampled and the documents we select as their pseudo-correct results.

We performed several evaluations of web search services to examine stability and bias of our method as we varied the number of query-document pairs and the directory from which pseudo-best documents were selected. We began with a 10M-entry log of queries submitted to a major web search engine on the first week of December 2002. We then filtered queries that were exact duplicates, contained structured operators, were not between one and four words long, or contained adult content. This left us with 1.5 million remaining queries. We then paired queries that exactly matched (ignoring only case) documents' directory-edited title with those documents. We did this for both the ODP and Looksmart directories. We excluded the "Adult", "World", "Netscape", and "Kids & Teens" sub-trees of the ODP and took entries from the "Reviewed Web Sites" section of the queries' results pages of Looksmart.

Copyright is held by the author/owner(s)
SIGIR '03, July 28-August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007

In 2001, ODP was estimated to have 2.6M links and has been built primarily by 36,000 volunteer editors, whereas 200 paid editors annotated many of the 2.5M links in Looksmart. Although the editing policies of the directories vary somewhat, each has human editors entering titles for the sites listed so that the directory titles do not necessarily correspond to, and likely are more accurate than, the titles of the pages themselves. In the 79% of the ODP query-document pairs that had URLs we were capable of crawling, 18% of them had edited titles in the directory that exactly matched (ignoring case) those of their corresponding pages. There were 83,713 matching query-document pairs for ODP and 33,149 for Looksmart. We filtered these, only keeping pairs whose result URLs have at least one path component (not just a hostname) and for which the query does not appear verbatim in the URL. This left 39,390 pairs over 24,992 queries for ODP and 10,902 pairs over 10,159 queries for Looksmart. As can be seen, often there were multiple documents in a directory that matched a given query, creating a set of alternate query-document pairs for that query. We therefore used the reciprocal rank of the highest ranked matching document, referred to as MRR1 in prior work [5].

The web search engines that we evaluated were Google, Fast (AllTheWeb), Teoma, Inktomi (via MSN advanced search), AltaVista, and WiseNut. Although we hypothesize that pages popular enough to be listed in directories would likely be crawled by each of these engines, index coverage affects their scores. Using our original query log of 10 million as a population size, and limiting sampling error to 3%, a sample size of 1067 pairs is needed for 95% confidence in our representation of the population. Using a sample of 2000, our sampling error is 2.2%, demanding at least a 2.2% difference in MRR1 for two engines to be considered to be performing differently. However, as stated above, sampling is not the only error introduced in this methodology. To determine how many query-document pairs are necessary for a stable evaluation we calculated the error rate [2], as suggested by Buckley for this type of evaluation [3], across all non-overlapping (no queries in one sample are re-used in another) query-document samples of various sizes from the entire set of 39,390 query-document pairs matched in the ODP. The error rate estimates the probability that varying query sets will cause a swap in the engines' rankings by dividing the number of swaps by the total number of pair-wise comparisons across all samples. Error rate was calculated using zero fuzziness, meaning that any MRR1 score difference causing a variation in the engines' rankings would count as an error (no scores were counted as ties). At a sample size of 2000, the error rate was 1.11%, at 3000, 0.83%, and at 4000 there were zero differences in rankings (errors) across all samples.

We also designed a series of experiments to estimate any possible bias introduced by selection from particular directories. Since different queries matched on each directory and the number of overlapping pairs was only 734 (1.5% set overlap), we used the first 2000 matching queries from each directory. These had 68 pairs in common. As shown in Table 1, the ranking of the engines is nearly identical for each directory, having a .93 Pearson correlation.

As a final method of evaluating our methodology, we turned to manual evaluations. Based on guidance from Ian Soboroff at NIST, we had 11 student evaluators manually judge 418 queries from the first 2000 queries matched in the ODP. We selected these queries from a single directory with the knowledge that bias introduced through directory selection was minimal. Assessors were told to select only the best document (home page) and any duplications or equivalently probable interpretations (i.e. an acronym expandable to multiple equally-likely phrases). As per Table 2, our automatic evaluation MRR1 scores have a moderately strong positive Pearson correlation of .71 to our manual evaluation.

Table 1: First 2000 query-document pairs from each directory

ODP			Looksmart		
Ranking	*MRR1*	*Found in top 10*	*Ranking*	*MRR1*	*Found in top 10*
E1	.3282	1095	E1	.3078	982
E2	.2720	939	E2	.2866	946
E3	.2647	796	E3	.2327	712
E4	.1784	720	E5	.2081	776
E5	.1610	632	E4	.2061	720
E6	.1391	517	E6	.1958	661

Table 2: Automatic vs. Manual for 418 queries

Automatic			Manual		
Ranking	*MRR1*	*Found in top 10*	*Ranking*	*MRR1*	*Found in top 10*
E1	.3254	220	E2	.3602	307
E2	.2475	191	E1	.3184	275
E3	.2429	151	E3	.2774	237
E4	.1608	144	E5	.2667	235
E5	.1472	118	E6	.2434	224
E6	.1216	100	E4	.2064	196

3. CONCLUSION

In contrast to manual judgments, evaluations using our automatic methodology can contain literally thousands of queries and can be repeated frequently. We demonstrated that our automatic evaluation methodology is stable and unbiased with regard to directory used. Our automatic evaluation has a moderately strong positive correlation to a reasonably large manual evaluation.

4. REFERENCES

[1] Broder, A. A Directory of Web Search. SIGIR Forum 36(2) (2002).

[2] Buckley, C., and Voorhees, E. Evaluating Evaluation Measure Stability. SIGIR'00, 33-40.

[3] Buckley, C. TREC Web Track mailing list, 2001.

[4] Chowdhury, A., and Soboroff, I. Automatic Evaluation of World Wide Web Search Services. SIGIR'02, 421-422.

[5] Hawking, D., and Craswell, N. Measuring Search Engine Quality. Information Retrieval, 4(1) (2001), 33-59.

[6] Voorhees, E. Evaluation by highly relevant documents. SIGIR'01, 74-8.