Semi-Automatic Evaluation via Editor-driven Taxonomies

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman
Illinois Institute of Technology
{steve,ej,abdur,dagr}@ir.iit.edu

ABSTRACT

We propose an evaluation methodology that augments traditional manual relevance judgments with automatic judgments in order to scale to evaluations over terabyte collections. Participants would submit results for a large (minimum of 2000) set of queries from a real web query log. These queries would first be paired with pseudo-relevant results using an automatic evaluation method such as matching queries with elements of a taxonomy. We have developed automatic judgment techniques for both best-document and precision-based evaluations. Assessors would inspect the automatic evaluations and select a number of the worst of them for a traditional manual evaluation. This methodology allows for not only an evaluation of the engines themselves, but also a study of real web queries and evaluation of automatic evaluation techniques. Although it does not solve the pooling problem associated with very large collections, it proceeds under the hypothesis that evaluating over a very large query set can provide the stability to compensate.

1. SEMI-AUTOMATIC EVALUATION

Our proposed methodology centers on examining search engine performance on a set of real web user's queries. Our hypothesis is that the large size of this query set will allow us to perform a stable evaluation on a very large collection. Our past experimentation has shown that query sets of size 2000 provide error rates near 1% for a fully automatic best-document MRR evaluation. In order to facilitate relevance assessments over this large query set, an automatic assessment component must be used. In order to ensure a reasonable set of relevance judgments, assessors will review the judgments made automatically and select the top X worst automatically judged queries for manual assessment. Obviously, this process would be much simpler in a best-document style evaluation as there would then be a small number of documents to examine for each query. This process of examining queries and automatic judgments yields several possible directions. Assessors could decide on a query-by-query basis which queries should be judged with a precision-based (large relevant set) metric and which with a best-document (MRR) metric. This would implicitly entail some study of the query log itself, giving an idea of what fraction of queries are topical in nature. Also implicit in examining the quality of the automatic judgments is the ability to iteratively improve the automatic judgment methodology both during the examination process and after the manual judgments are performed in anticipation of next year. With a large enough set of manual judgments, correlations between manual and automatic evaluation methods can be used as a metric for improving future automatic judgment.

2. AUTOMATIC JUDGMENT METHODS

The advent of online, editor-driven taxonomies such as the ODP and Looksmart has enabled methods for automatic judgment. In past work, we used such taxonomies to find sets of pseudo-relevant documents via one of two assumptions: 1) taxonomy entries are relevant to a given query if their editor-entered titles exactly match the query, or 2) all entries in a leaf-level taxonomy category are relevant to a given query if the category title exactly matches the query. The first method, referred to as "title-match," was first developed by Chowdhury and Soboroff [5]. Basically, it finds queries that exactly match the editor-entered title of taxonomy entries and uses these entries as a "best document" assessment. For example, the query "information retrieval" would only have documents with exactly "Information Retrieval" as their edited title in its pseudo-relevant set. The second method, called "category-match," finds leaf-level taxonomy categories with names that exactly match the query and treats all documents in that category as relevant, allowing for a precision-based assessment. Referring back to our previous example, documents in categories described as "/Top/…/Information_Retrieval" would be used as the pseudo-relevant set for category-match. Because of the relatively few matches found with title-match (less than two on average in our experiments) it lends itself to a best-document MRR evaluation scheme. By contrast, category-match yields large pseudo-relevant sets (of size 192 on average in our experiments), making it more suitable for a precision-based evaluation. We showed that purely automatic instantiations of these methodologies correlate moderately strongly with a manual evaluation by evaluating six web search engines on a sample from an America Online log of ten million web queries [1]. We have also shown that such an evaluation is unbiased in terms of the chosen taxonomy and stable with respect to the query set selected when that set is of sufficient size [2].

3. RELATED WORK

Prior studies have shown that variations in relevance judgments due do not de-stabilize evaluation and error rates, measuring the stability of a metric, can be calculated using multiple query sets and controlled by increasing the number of queries used in evaluation [3][8]. One possible semi-automatic evaluation approach is to select a random document and formulate a query intended to retrieve it, as proposed by Buckley [4]. However, the queries would then be unrepresentative of real users' needs. Others have made use of web taxonomies to fuel automatic evaluation. Haveliwala, et al. used the categories in the ODP to evaluate several strategies for the related page (query-by-example) task in their own engine by selecting pages listed in the ODP and using distance in the hierarchy as a measure of how related other pages are [6]. Menczer used distance in the ODP hierarchy as a part of an estimate of precision and recall for web search engines using TReC homepage-finding qrels to bootstrap his evaluation [7]. For 30 of these queries he found that the automatic evaluation correlated to a manual one.

4. PAST EXPERIMENTATION

We began with a 10M-entry log of queries submitted to AOL Search. 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. In order to assess how well our automatic evaluation measures estimate the evaluations of real users, we created a set of manual best-document relevance judgments for our evaluation of six web search engines over the queries from an AOL log. We had 11 student evaluators manually judge the first 418 queries that matched titles in the ODP. For each query, they were presented with a randomly-ordered list of all of the unique documents retrieved by each engine pooled together. Assessors were told to select only the best document and any duplications or equivalently probable interpretations (i.e. an acronym that could be expanded to multiple equally-likely phrases). On average, they selected 3.9 best documents per query.

Our first method paired documents whose editor-entered title exactly matched a query (ignoring only case) with that query. 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 use the reciprocal rank of the highest ranked matching document, referred to as MRR1 in prior work. To get a worst-case estimate of how well our title-matching automatic evaluation tracked with the manual one, we performed the automatic evaluation on only those queries which we had manually judged. With only 418 queries, a difference of 4.8% is necessary for two engines to be considered to be performing differently with 95% confidence. Even with this small number of queries the evaluations were found to have a .71 Pearson correlation.

For our second method, we focused on utilizing the categorical information present in taxonomies for a precision-based automatic evaluation method. For the sake of comparison, we began with the set of 24,992 distinct queries that matched titles of documents in the ODP. We then attempted to match each of those with category names as stated. Again, for a worst-case estimate of how this automatic strategy tracks a manual one, we initially limited the automatic and manual evaluations to only those queries they had in common. However, since not all manually judged queries also matched category names, this only left 94 queries, demanding a 10.1% difference between two engines' scores for them to be considered to have performed statistically different with 95% confidence. Examining those results, there were too many ties for correlations to be meaningful. Therefore, we present instead the entire set of automatic category matches in comparison with the entire set of manual judgments.

In addition to the above title-match results which examined a sort of worst-case performance, we calculated Pearson (see Table 1) and Spearman rank (see Table 2) correlations between evaluations performed on all of the queries we were able to (automatically or manually) judge: 24,992 matching the ODP for title-matching, the 6,255 in the subset of those that matched categories, and all 418 manual judgments we performed. This provides for more accurate rank correlations as the large query samples leave no statistical ties. The only tie remaining is one in the 418-query manual evaluation.

Table 1: Pearson correlations of MRR1 measures

	Category	***Title***
Title	0.689	N/A
Manual	0.597	0.735

Table 2: Spearman correlations of rankings

	Category MRR1	***Category P@10***	***Title MRR1***
Category P@10	1.0	N/A	N/A
Title MRR1	.6571	.6571	N/A
Manual MRR1	.7000	.7000	.7714

From these experiments, we can see that, as expected, the correlations between the title-match automatic evaluation and the manual evaluation increased when a larger number of queries were used. All evaluations agree on which three engines are the best, and which three are the worst. It can also be seen that title-match has a stronger correlation with our manual evaluation than category-match. This is likely due to the fact that both the manual evaluation and the title-match technique used a best-document style method yielding few documents in the relevant set, while the category-match technique produces many pseudo-relevant documents for a query.

5. REFERENCES

[1] Beitzel, S. et al. Using Manually-built Web Directories for Automatic Evaluation of Known-Item Retrieval. To appear in SIGIR'03.

[2] Beitzel, S. et al. Using Titles and Category Names from Editor-driven Taxonomies for Automatic Evaluation. Submitted to CIKM'03.

[3] Buckley, C., and Voorhees, E. Evaluating Evaluation Measure Stability. In Proceedings of SIGIR'00, 33-40.

[4] Buckley, C. Proposal to TReC Web Track mailing list (November, 2001). http://groups.yahoo.com/group/webir/message/760

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

[6] Haveliwala, T. et al. Evaluating Strategies for Similarity Search on the Web. In Proceedings of WWW'02.

[7] Menczer, F. Semi-Supervised Evaluation of Search Engines via Semantic Mapping. Submitted to WWW'03. http://dollar.biz.uiowa.edu/~fil/Papers/engines.pdf

[8] Voorhees, E. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In Proceedings of SIGIR'98, 315-323.