Using Titles and Category Names from Editor-Driven Taxonomies for Automatic Evaluation

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman
Information Retrieval Laboratory
Illinois Institute of Technology
Chicago, IL 60616
{steve,ej,abdur,dagr}@ir.iit.edu

ABSTRACT

Evaluation of IR systems has always been difficult because of the need for manually assessed relevance judgments. The advent of large editor-driven taxonomies on the web opens the door to a new evaluation approach. We use the ODP (Open Directory Project) taxonomy 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. We compare and contrast these two methodologies by evaluating six web search engines on a sample from an America Online log of ten million web queries, using MRR measures for the first method and precision-based measures for the second. We show that this technique is stable with respect to the query set selected and correlated with a reasonably large manual evaluation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process*

General Terms

Algorithms, Measurement, Experimentation

Keywords

Automatic Evaluation, Web Search, Relevance Judgments

1. INTRODUCTION

Search engine evaluation is typically resource intensive because of the need for human-reviewed relevance assessments. Performing these assessments on very large collections like the web is impractical, since manual review can typically only be done on a very small scale. The advent of online, editor-driven taxonomies such as the ODP has enabled a new type of automated evaluation technique. The premise is to take a large sample of actual web queries and mine pseudo-relevant document sets from a taxonomy for each query. We examine two methods of doing this. The first method, called "title-match" was first developed in our prior work, and is further analyzed in this study. Title-match 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 "mortgage rates" would only have documents with exactly "Mortgage Rates" as their edited title in its pseudo-relevant set produced by title-match. In our previous efforts, title-match was shown to be unbiased in terms of the taxonomy used to mine these pseudo-relevant sets [1]. 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 the previous example, documents in categories described as "/Top/…/Mortgage_Rates" 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 mean-reciprocal rank 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. The key focus in this work is to expand on prior efforts by comparing and contrasting these two automatic evaluation methodologies, and examining their correlation with a 418-query manual "best-document" (MRR) evaluation. In addition, an expanded analysis of the title-match approach developed in [7] and shown to be unbiased in [1] is provided. Section 2 briefly reviews related work. Section 3 describes our evaluation methodologies and Section 4 gives results of evaluations performed with each. Section 5 provides an analysis of how these methodologies correlate with each other. Finally, in Section 6 contains conclusions and directions for future work.

CIKM '03, November 3-8, 2003, New Orleans, LA, USA.

2. RELATED WORK

Most of the work in evaluating search effectiveness has followed the Text REtrieval Conference (TREC) methodology which includes holding constant the test collection, using topical queries resulting from a user's information need, and using complete manual relevance judgments to compare retrieval systems based on the traditional metrics of precision and recall. Evaluating the effectiveness of web search engines provides many unique challenges that make such an evaluation problematic [2], [13]. The web is too large to feasibly perform manual relevance judgments of enough queries with sufficient depth to calculate recall. In contrast to a test collection, the web is "live" data that is continually changing, preventing experiments from being exactly

reproducible. In addition, it is believed that the set of popular web queries and the desirable results for those queries changes significantly over time and that these changes have a considerable impact on evaluation. Hawking, et al. notes "Search engine performances may vary considerably over different query sets and over time" [17]. These challenges demand a measurement that can be repeated to monitor the effect of these changing variables.

2.1 Evaluation Measures

The TREC forum is the foundation for the majority of manual evaluations as it enables researchers to pool their results for deep relevance judgments by human assessors over a common, fixed set of documents and queries. Studies of the evaluation measures used in TREC (meta-evaluations) have provided several motivating factors for this study: Although relevance is an ambiguous concept, variations in relevance judgments due to assessor disagreement have been shown not to destabilize evaluation [30]. Error rates, which measure the stability of a metric, can be calculated using multiple query sets and controlled by increasing the number of queries used in an evaluation [5]. Although assessors frequently disagree on the most relevant page for informational queries, causing instability that makes MRR unviable for informational query evaluation, Voorhees suggests, "It is likely that there would be more agreement among assessors as to the best page for navigational requests than for informational requests" [31]. Although traditional TREC methodology has provided the foundation for many interesting studies, many do not consider it relevant to the relative performance of web search engines as they are actually interacted with by searchers. Experiments in the interactive track of TREC have shown that significant differences in mean average precision in a batch evaluation did not correlate with interactive user performance for a small number of topics in the instance recall and question answering tasks [29]. In the past two years, the importance of navigational queries has led TREC to incorporate known-item evaluations as part of the web track [14][18]. These evaluations used MRR of homepages and named-pages as a metric.

There have been several studies that evaluate web search engines using TREC methodology of manual relevance judgments. Hawking and Craswell, et al. evaluated web search engines [13][15] in comparison to TREC systems involved in TREC tracks from 1998-1999 that used the 100GB VLC2 web snapshot and 50 manually-assessed informational queries each year [11][12]. They found that TREC systems generally outperformed web search engines on the informational task in 1998 and 1999; however, they acknowledged that comparing TREC systems with web engines in an ad-hoc (informational) evaluation might not be sufficient [8]. Their evaluation of the web search engines correlated with an informational task evaluation done by Gordon and Pathak in 1998 [10]. Hawking, Craswell, and Griffiths also manually evaluated web search engines on 106 transactional (online service location) queries in 2000 [17], and 95 airline homepage finding queries in 2001 [16]. Although they do not provide a direct comparison of web search services to TREC systems participating in similar transactional and navigational tasks those years, their evaluations of the two are similar and the web engines' scores are generally equivalent or slightly above those of the TREC evaluations. Leighton and Srivastava evaluated web search engine performance on an informational task using a mixture of structured and unstructured queries and found differences in the engines' effectiveness in 1997 [22]. Ding and Marchionini evaluated three web search engines on a small set of informational topics in 1996 and found no significant difference between them [9]. Other studies have used alternative methods of manually evaluating web search engines. Bruza, et al. compared the interactive effectiveness of query-based, taxonomy-based, and phrase-based query reformulation search on the web, showing that the assisted search of the latter technique could improve relevance of results, but came at the cost of higher cognitive load and user time [4]. Singhal mined homepage-finding queries from a large web query log by selecting those that contained terms such as "homepage," "webpage," and "website." He used the rank of manually judged homepages as his measure and found web engines' effectiveness to be superior to that of a TREC system in 2001 [27].

2.2 Manual Web Search Evaluation Techniques

Evaluating web search engines has traditionally been a task that requires significant resources and human intervention. Evaluations based on precision and recall of topical queries may not only be difficult on the web, but incomplete. Spink gave a basis for classifying queries [28] as informational, navigational or transactional, but we are unaware of any large-scale study that quantifies the ratio of web queries in the different categories that have been defined. Broder defines similar classifications and presents a study of Altavista users via a popup survey and self-admittedly "soft" query log analysis indicating that less than half of users' queries are informational in nature [3]. The general belief is that the majority of web searches are interested in a small number (often one) of highly relevant pages. This would be consistent with the aspects of web searching that have been measured from large query logs: the average web query is 2.21 terms in length [20], users view only the top 10 results for 85% of their queries and they do not revise their query after the first try for 75% of their queries [26]. It is also widely believed that web search services are being optimized to retrieve highly relevant documents with high precision at low levels of recall, features desirable for supporting known-item search. Singhal and Kaszkiel propose, "site-based grouping done by most commercial web search engines artificially depresses the precision value for these engines…because it groups several relevant pages under one item…" [27]. Given this, it is clear that manual evaluations and metrics other than simple precision and recall are required to effectively evaluate web search engines.

2.3 Automatic Web Search Evaluation Techniques

Although manual evaluations have provided accurate measures of web search service performance across many query tasks, they are dated very quickly as the web, search services in operation, algorithms used by those services, popular queries and desired results change rapidly. The prohibitive expense of repeating manual evaluations has led to several studies of automatic evaluation of web search systems. The least resource-intensive of the proposed methodologies is to compute a similarity measure between documents retrieved by web search services and the query to automatically estimate relevance as likeness to a known retrieval strategy. Shang and Li compared the rankings generated by using several standard IR similarity measures and one that they designed

themselves to model a ternary relevance assessment [25]. They found their evaluation correlated with a manual evaluation of a small set of queries from the academic domain [23]. Others have advocated the use of clickthrough data (which results users click on) for automatic assessment, however, there is a documented presentation bias inherent in this data: users are more likely to click on highly ranked documents regardless of their quality [2]. Joachims presents a method using a single user interface that combines rankings of results from two engines in order to remove this bias [21]. For three users of this interface to three web engines over 180 queries, he shows that the automatic evaluation correlates with a manual one. 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 [19]. 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 [24]. For 30 of these queries he found that the automatic evaluation correlated to a manual one. In 2002 we proposed a method of automatic evaluation [7] which we showed to be unbiased in [1]. What follows is an elaboration on that work, including measure stability experiments, more analysis and correlation with a new automatic technique using categories.

3. EVALUATION METHODOLOGIES

We have developed two methodologies for using web taxonomies to automatically evaluate web search engines. Each of our methodologies makes use of a reviewed collection, such as a web taxonomy, and a large sample of web queries. Title-match collects documents from the reviewed collections whose editor-supplied titles exactly match a given query. These documents are viewed as the "best" or "most relevant" documents for that query, and the mean reciprocal rank of these documents over all queries is used as the scoring metric for each engine. Category-match searches the category names in the reviewed collections, and if a category name is found that exactly matches a given query, all documents from that category are used as the relevant set. Precision measures such as P@10 are then used to rank each engine. For either methodology to yield a valid ranking of engines according to general retrieval effectiveness, the set of query-document pairs that they produce needs to be reasonable, unbiased, and large enough to satisfy both sampling and stability.

Two other factors that must be controlled in this methodology, as in any evaluation strategy, are bias in the queries sampled and the documents we select as their pseudo-relevant results. One possible approach for automatically finding best documents would be to simply select the top document retrieved by a particular engine as the pseudo-correct document for that query. However, this would bias the documents selected towards that engine's ranking scheme, resulting in inflated scores for engines using similar algorithms. Another possible solution would be to select a random document and formulate a query intended to retrieve it, as proposed by Buckley for the TREC Web Track [6]. However, the queries would then be biased and unrepresentative of real users' needs. In our methodology, unbiased queries are achieved simply through statistical sampling techniques. We ensure that the sample is large enough to be representative of the query log chosen and that the initial query log is sufficiently large, drawn from a source indicative of the domain of queries we intend to evaluate, and an accurate representation of typical queries over whatever time period in which we are interested in evaluating the engines. Although selecting documents according to the titles of random queries is not inherently biased, we have limited ourselves to editor-controlled titles of a particular collection of documents.

3.1 On-Line Taxonomies

Fundamental to our evaluation methodologies is usage of the existing manually-constructed web taxonomies. For our purposes, it is important to note that all taxonomies we've found have a common notion of categorization of entries via category names that often includes a hierarchy and inclusion of editor-entered page titles. Although the editing policies of different taxonomies vary somewhat, they all have human editors entering titles for the sites listed so that the taxonomy titles do not necessarily correspond to, and likely are more consistently accurate than, the titles of the pages themselves. In our previous efforts, we used the ODP and Looksmart taxonomies to show that title-match performs consistently no matter what taxonomy is used [1]. We found that the rankings produced by using ODP and Looksmart had a Pearson Correlation of .931.

Since we have previously shown automatic evaluation techniques like these to be unbiased in terms of taxonomy, we focused on using the ODP, the larger and more heavily-edited taxonomy, for the experiments in this paper.

In addition to eliminating taxonomy selection bias, it is crucial to the success of these automatic methodologies that they be shown to be "stable" for a reasonable sample size of queries. That is, these methods must be able to return consistent rankings for a set of engines being evaluated over any arbitrary, reasonably sized sample of queries. If the methods can be shown to be stable, they can be relied upon to produce accurate rankings over non-fixed query sets, and therefore can be used to continually evaluate web search engines even as their query traffic changes over time. To this end, we have designed a set of experiments for determining the error rate (in terms of stability) of these automatic evaluation techniques.

3.2 Engines

The web search engines that we evaluated were Google, Fast (AllTheWeb), Teoma, Inktomi (via MSN advanced search), AltaVista, and WiseNut. We assume that pages popular enough to warrant listing in the ODP are likely to be crawled by each of these engines, therefore any skewing effects due to differing index coverage are likely to be negligible. This assumption is likely reasonable, given the very large index sizes of popular search engines (Google claims over three billion pages, Alltheweb claims over two billion), and the tendency of taxonomies to list popular pages.

4. RESULTS

We began with a 10M-entry log of queries submitted to AOL Search on the first week of December, 2002. As it was from a single server of a pool that distributes queries round-robin, it is itself a sample of the total queries for that week. This 10-million

entry query log was then filtered and queries exhibiting the following characteristics were removed:

- Exact duplicates
- Queries containing structured operators, such as '+', 'AND', 'OR'
- Queries not between one and four words long
- Queries seemingly searching for pornography

The filtration process left us with a log of just over 1.5 million queries from which to draw our samples.

We then paired documents whose editor-entered title exactly matched a query (ignoring only case) with that query. To examine how heavily titles in the ODP are edited, we compared them to the titles in the web pages themselves. In the 79% of ODP query-document pairs that had URLs we were capable of crawling, only 18% of them had edited titles in the taxonomy that exactly matched (ignoring case) those of their corresponding pages. We filtered the initial set of matching query-document pairs such that we only kept 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. These constraints were intended to remove trivial matches such as the query "foo bar" matching http://www.foobar.com and limit bias that might be introduced if some engines use heuristics for matching URL text. Often, there were multiple documents in the ODP that matched a given query, creating a set of alternate query-document pairs for that query. This led to the development of four methods of scoring, all variants of Mean Reciprocal Rank computed for each engine over all queries:

- Random-match: A random candidate judgment is selected as the judgment
- Max-match: The best-scoring candidate judgment over all engines is selected as judgment
- Avg-match: The average score of all candidate-judgments is computed
- LocalMax-match (MRR1): The best-scoring candidate-judgment for an engine is selected

The numbers of initial, filtered, and average matches in the ODP per query (after filtering) are listed in Table 1.

Table 1: Number of matches on edited titles

Taxonomy	Attempted	Total Matches	After Filtering	Queries Matched	Avg. per Query
ODP	1,515,402	83,713	39,390	24,992	1.58

4.1 Manual Evaluation

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. Based on guidance from Ian Sobroff at NIST, we had 11 student evaluators manually judge the first 418 queries that matched titles in the ODP. We selected these queries from a single taxonomy with the knowledge that bias introduced through taxonomy selection was minimal [1]. We built a simple web interface which presented assessors with the next query to be judged once they had logged in. For each query, they were presented with a randomly-ordered list of all of the unique documents retrieved by each engine pooled together. Each list item consisted only of the number of that document in the list which was a link to the actual URL of the document so that users could view the live document on the web in the browser of their choice. All assessment was performed at the assessors' leisure from their personal or campus lab computers. Assessors were told to select only the best document (home page) 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 manual evaluation interface recorded 87 hours spent judging all 418 queries over a two week period. The evaluation period began the day after gathering the automatic judgments and storing the search results for each query from each engine in an attempt to minimize the effect of changes taking place in the live data.

4.2 Title Matching

Once our query-document pairs for the ODP had been constructed, and we had conducted a manual evaluation to compare to, we set about conducting automatic evaluations using the title-match method.

4.2.1 Automatic Evaluation

To get a worst-case estimate of how well the title-match 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.

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

The manual evaluation's ranking of the target engines compared to our automatic evaluation is shown in Table 2. E2 and E3 in the automatic run and E3 and E5 in the manual run are statistical ties. Even with this small number of queries the evaluations were found to have a .71 Pearson correlation, which is typically considered "moderately strong". The Spearman rank correlation (accounting for statistical ties) is .59. In a situation where a very large number of queries are available for use by the automatic evaluation system, we would expect to see these correlations increase.

4.2.2 Stability

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% relative difference in MRR for two engines to be considered to be performing differently with 95% confidence. However sampling is not the only error introduced in this methodology. The error associated with the assumption that a document whose edited title exactly matches a query is a reasonable candidate for the best document for that query is more difficult to estimate. In order to determine how many query-result pairs are necessary for a stable method we calculated error rate [5], as suggested by Buckley for this type of evaluation [6], across all query-result samples of various sizes and across five formulations of MRR according to varying uses of the sets of alternate matching documents for each query as shown in Table 3. For these error rate experiments we selected one large taxonomy (ODP) and held it constant, and produced a very large number of query-result pairs for that taxonomy. From this resulting collection of query-result pairs we constructed all possible random query samples of varying sizes, ranging from 2000-4000. Each of these sets of random query samples was then run against the 6 test search engines, and the results for each MRR measure on each sample were used in calculating the error-rate of the measure. Error rate was calculated using 0% fuzziness, meaning that any variation in the engines' rankings would count as an error, as shown in Table 3.

Table 3: Error rates across sample sizes and MRR formulas

Size / MRR	Random	Global Max	Average	Local Max (MRR1)
2000	1.11%	1.11%	0.56%	1.11%
3000	0%	0%	0%	0.83%
4000	0%	0%	0%	0%

As can be seen from the table, all of the MRR measures were very stable, leaving only near 1% probability of two engines changing places in the rankings when using different samples of the given sizes. By the time we reach sample sizes of 4000, we see no changes in the engines' ranking when using different samples. From these experiments we can conclude that these automatic evaluation approaches will be stable enough to permit the usage of changing query sets for evaluating a set of web search engines over time.

4.3 Category Matching

4.3.1 Procedure

For the "category-match" methodology, we focused on utilizing the categorical information present in the ODP for a precision-based automatic ranking method. The basic method was to exactly match queries to the most specific component of the category names and then use all documents in those matching categories as the pseudo-relevant set. For example, the query "mortgage rates" would match the categories

"/Top/Personal_Finance/Mortgage_Rates" and
"/Top/Business/Property_Assets/Mortgage_Rates".

This yields many pseudo-relevant documents for each query (see Table 4), making it suitable for precision-based measures.

4.3.2 Automatic Evaluation

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. The results of this matching can be seen in Table 4. Unlike the title-matching experiments, we did not filter the pseudo-relevant documents on the basis of their URLs being only a hostname or containing the query text.

Table 4: Number of matches on category names

Attempted	Matched	Categories per Query (avg.)	Documents per Query (avg.)
24,992	6,255	11.4	192

The target search engines were then evaluated by calculating the mean precision and reciprocal rank of the first retrieved relevant document (MRR1) over the top ten results retrieved for the entire set of queries matched. Limiting the evaluation to the top ten results from each engine (typically the first page) is consistent with the common belief that web users rarely examine more than one page of results for any given query. The intuition for using these two measures is to examine not only how many of the top ten results are relevant, but also how well those top ten are ranked (it is also believed that users often are most interested in the first relevant result). The results of this evaluation can be seen in Table 5.

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. Correlation coefficients for these are given in Table 6 and Table 7.

Table 5: Automatic category matching over 6255 Queries vs. manual over 418 queries

Automatic				Manual	
Ranking	***P@10***	*Ranking*	***MRR1***	*Ranking*	***MRR1***
E3	.0491	E3	.5017	E2	.3602
E1	.0462	E1	.4552	E1	.3184
E2	.0447	E2	.4436	E3	.2774
E5	.044	E5	.4314	E5	.2667
E6	.0401	E6	.386	E6	.2434
E4	.0347	E4	.3732	E4	.2064

5. ANALYSIS

In order to assess the extent to which the different evaluation methodologies agree and how well they correlate with actual users' judgments of relevance, we calculated correlations between them, using both on the actual evaluation measure value distributions via the Pearson correlation measure (see Table 6) and only the ranking resulting from the evaluation measure using the Spearman rank correlation measure (see Table 7). In contrast to the above results which examined a sort of worst-case performance

for the automatic methods by limiting the queries used in the automatic evaluations to the same ones evaluated manually, these correlations are 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 is a sort of best-case assessment, but it is likely the common way these techniques would be applied as it exploits one of the main benefits of automatic evaluation; namely that many queries can be used in the evaluation as the cost of producing automatic pseudo-relevance judgments is quite low (automatically string matching even the millions of queries we worked with using a naïve approach was computationally feasible). It also provides for more accurate rank correlations as the large query samples leave no statistical ties.

Table 6: Pearson correlations of measures

	Category MRR1	***Title MRR1***
Title MRR1	0.689	N/A
Manual MRR1	0.597	0.735

Table 7: 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

The only tie remaining is the one between E3 and E5 in the 418-query manual evaluation (see Table 2). This statistical tie was accounted for in our Spearman correlations.

From these experiments it can be 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. This demonstrates the main advantage of our automatic method, in that we can readily take advantage of large volumes of available queries to improve the ranking produced by our method. Additionally, both the automatic and the manual evaluations agree on which three engines are the best (E1-E3), and which three are the worst (E4-E6) out of the group as a whole.

6. CONCLUSIONS AND FUTURE WORK

We have shown a technique for automatic evaluation of search engines using manually edited taxonomies. The power of the technique is that these automatic evaluations can utilize literally thousands of queries instead of only the handful used in present TREC-style manual evaluations, and can be repeated with new queries and desired results without the cost of repeating the process of manual judgment.

We have observed that these types of automatic evaluations are consistently capable of discerning "good" engines from "bad" ones, and also that they have a very high degree of stability for query samples of size 2000 or more. As they are automatic processes, it is possible to use these techniques to judge the effectiveness of web search engines even as the content of their query traffic and web coverage changes over time. One drawback of these methods is that they are not capable of discerning whether closely performing engines are actually better or worse than each other. This limits their applicability to evaluation settings that require strict, fine-grained ranking, however, the number of advantages associated with these methods makes them, at the very least, quite suitable for deciding which engines are effective and which engines are ineffective. We have also observed that title-match has a stronger correlation with our manual evaluation than the category-match technique, however, this is likely due to the fact that both the manual evaluation and title-match used a "best-document" MRR1 ranking metric, while the category-match technique produces many pseudo-relevant documents for a query, making it fit better to a precision-based evaluation. Because of this, it is logical to expect that the correlation between category-match and our manual evaluation will be weaker.

There is a great deal of future work in this area. The most obvious extension to this work is to further the validation of these automatic methods by comparing their performance to larger manual evaluations that are more carefully controlled. We would also like to perform a traditional manual evaluation that is focused on topical relevance in order to more directly compare the performance of our precision-based category-match method to a corresponding manual evaluation. This would also allow us to examine how much the constraint of "exactly matching" document and category titles can be relaxed, as relaxing this constraint would allow is to consider an even broader domain of queries in our automatic evaluations. Most notably, we would like to pursue the development of a method that can combine varying amounts of pure manual assessment with these automatic methods. This hybrid method would then be able to take advantage of both the accuracy of manual evaluations and the ability of automatic evaluations to consider a large number of queries.

7. REFERENCES

[1] Beitzel, S., Jensen, E., Chowdhury, A., Grossman, D., and Frieder, O. Using Manually-built Web Directories for Automatic Evaluation of Known-Item Retrieval. To appear in SIGIR'03 poster session (Toronto, Canada, 2003).

[2] Boyan, J., Freitag, D., and Joachims, T. A machine learning architecture for optimizing web search engines. In AAAI'96 (August, 1996) Workshop on Internet Based Information Systems. http://www.cs.cornell.edu/People/tj/publications/boyan_etal_96a.pdf

[3] Broder, A. A Taxonomy of Web Search. SIGIR Forum 36(2) (Fall, 2002).

[4] Bruza, P., McArthur, R., and Dennis, S. Interactive Internet search: keyword, directory and query reformulation mechanisms compared. In Proceedings of SIGIR'00 (Athens, Greece, 2000), ACM Press, 280-287.

[5] Buckley, C., and Voorhees, E. Evaluating Evaluation Measure Stability. In Proceedings of SIGIR'00 (Athens, Greece, 2000), ACM Press, 33-40.

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

[7] Chowdhury, A., and Soboroff, I. Automatic Evaluation of World Wide Web Search Services. In Proceedings of

SIGIR'02 (Tampere, Finland, August, 2002), ACM Press, 421 - 422.

[8] Craswell, N., Bailey, P., and Hawking, D. Is it fair to evaluate Web systems using TREC ad hoc methods? SIGIR'99 (Berkeley, CA, 1999) Workshop on Web Evaluation. http://pigfish.vic.cmis.csiro.au/~nickc/pubs/sigir99ws.ps.gz

[9] Wei Ding and Gary Marchionini. Comparative study of web search service performance. In Proceedings of the ASIS 1996 Annual Conference (October 1996).

[10] Gordon, M., and Pathak, P. Finding information on the world wide web: The retrieval effectiveness of search engines. Information Processing and Management, 35(2) (March 1999), 141-180.

[11] Hawking, D., Craswell, N., and Thistlewaite P. Overview of TREC-7 Very Large Collection Track. In Proceedings of TReC7 (Gaithersburg, MD, 1998), NIST Special Publication 500-242, 91-104.

[12] Hawking, D., Voorhees, E., Craswell, N., and Bailey, P. Overview of the TREC-8 Web Track. In Proceedings of TReC8 (Gaithersburg, MD, 1999), NIST Special Publication 500-246, 131-149.

[13] Hawking, D., Craswell, N., Thistlewaite P., and Harman, D. Results and challenges in web search evaluation. In Proceedings of WWW8 (Toronto, Canada, May 1999), Elsevier Science, 243-252.

[14] Hawking, D., and Craswell, N. Overview of the TREC-2001 Web Track. In Proceedings of TReC10 (Gaithersburg, MD, 2001), NIST Special Publication 500-250, 61-67.

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

[16] Hawking, D., Craswell, N., and Griffiths, K. Which search engine is best at finding airline site home pages? CMIS Technical Report 01/45 (March, 2001). http://pigfish.vic.cmis.csiro.au/~nickc/pubs/TR01-45.pdf

[17] Hawking, D., Craswell, N., and Griffiths, K. Which Search Engine is Best at Finding Online Services? In Proceedings of WWW10 (Hong Kong, May 2001), Posters. Actual poster available as http://pigfish.vic.cmis.csiro.au/~nickc/pubs/www10actualposter.pdf

[18] Hawking, D., and Craswell, N. Overview of the TREC-2002 Web Track. To appear in Proceedings of TReC11 (Gaithersburg, MD, 2002).

[19] Haveliwala, T., Gionis, A., Klein, D., and Indyk, P. Evaluating Strategies for Similarity Search on the Web. In Proceedings of WWW'02 (Honolulu, HI, May, 2002), ACM Press.

[20] Jansen, B., Spink, A., and Saracevic, T. Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing and Management, 36(2) (2000), 207-227.

[21] Joachims, T. Evaluating Retrieval Performance using Clickthrough Data. SIGIR'02 (Tampere, Finland, August, 2002) Workshop on Mathematical/Formal Methods in Information Retrieval. http://www.cs.cornell.edu/People/tj/publications/joachims_02b.pdf

[22] Leighton, H., and Srivastava, J. First 20 precision among world wide web search services (search engines). Journal of the American Society for Information Science, 50(10) (1999), 882-889.

[23] Li, L., and Shang, Y. A new method for automatic performance comparison of search engines. World Wide Web, 3 (2000), Kluwer Academic, 241-247.

[24] Menczer, F. Semi-Supervised Evaluation of Search Engines via Semantic Mapping. Submitted to WWW'03 (Budapest, Hungary, 2003), ACM Press. http://dollar.biz.uiowa.edu/~fil/Papers/engines.pdf

[25] Shang, Y., and Li, L. Precision Evaluation of Search Engines. World Wide Web, 5 (2002), Kluwer Academic, 159-173.

[26] Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. Analysis of a very large web search engine query log. SIGIR Forum 33(1) (Fall, 1999), 6-12. Previously available as Technical Report TR 1998-014, Compaq Systems Research Center, Palo Alto, CA, 1998. http://www.research.compaq.com/SRC

[27] Singhal, A., and Kaszkiel, M. A case study in web search using TREC algorithms. In Proceedings of WWW10 (Hong Kong, May 2001), 708-716.

[28] Spink, A., Jansen, B.J., Wolfram, D., and Saracevic, T. From E-sex to e-commerce: Web search changes. IEEE Computer, 35(3) (2002), 107-109.

[29] Turpin, H., and Hersh, W. Why Batch and User Evaluations Do Not Give the Same Results. In Proceedings of SIGIR'01 (New Orleans, LA, 2001), ACM Press, 225-231.

[30] Voorhees, E. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In Proceedings of SIGIR'98 (Melbourne, Australia, 1998), ACM Press, 315-323.

[31] Voorhees, E. Evaluation by highly relevant documents. In Proceedings of SIGIR'01 (New Orleans, LA, 2001), ACM Press, 74-82.