Automatic Evaluation of World Wide Web Search Services

Abdur Chowdhury
America Online Inc.
cabdur@aol.com

Ian Soboroff
National Institute of Standards and Technology
ian.soboroff@nist.gov

ABSTRACT

Users of the World-Wide Web are not only confronted by an immense overabundance of information, but also by a plethora of tools for searching for the web pages that suit their information needs. Web search engines differ widely in interface, features, coverage of the web, ranking methods, delivery of advertising, and more. In this paper, we present a method for comparing search engines automatically based on how they rank known item search results. Because the engines perform their search on overlapping (but different) subsets of the web collected at different points in time, evaluation of search engines poses significant challenges to the traditional information retrieval methodology. Our method uses known item searching; comparing the relative ranks of the items in the search engines' rankings. Our approach automatically constructs known item queries using query log analysis and automatically constructs the result via analysis of editor comments from the ODP (Open Directory Project). Additionally, we present our comparison on five (Lycos, Netscape, Fast, Google, HotBot) well-known search services and find that some services perform known item searches better than others, but the majority are statistically equivalent.

Categories & Subject Descriptors: H.3.4 [Information Storage and Retrieval]: Systems and Software -- Performance evaluation; H.3.5 [Information Storage and Retrieval]: Online Information Services -- Web-based services

General Terms: Experimentation, Performance.

Keywords: IR Evaluation, Automatic Mean Reciprocal Ranking.

1. INTRODUCTION

Web search evaluation poses a considerable number of challenges to traditional IR evaluation methods. First, the collection is constantly changing, i.e. any evaluation is not reproducible in the future. Since the collection is so large, it is not possible to manually judge enough queries to a sufficient result depth to be able to measure recall in any reasonable way. Other researchers have also enumerated these fundamental challenges of web evaluations, however they have focused on the ad-hoc search task [1]. While other researchers have examined the effectiveness of various search services, their primary evaluation technique has been some precision variant evaluation of informational queries [3, 4]. Singhal et al. evaluated the search tasks of web users and proposed that navigational queries were more significant to web search than traditional TREC ad-hoc information gathering [2]. In addition, Singhal compared the effectiveness of TREC systems for navigation queries in comparison to web search engines and concluded that the search services were performing known item search better than traditional ad-hoc approaches. Navigational queries were examined in the TREC 2001 web track so a standard corpus could be created with relevance judgments and reproducible results [5]. Most recently, Spink classified the types of data that people are looking for but did not examine the effectiveness of their searching [8].

Given the dynamic nature of the web, the difficulty of creating a large representative test collection and the resources needed for traditional evaluation methodologies, we present a technique that is able to automatically compare search services at regular, short intervals. Since our technique uses a large number of search queries that are automatically assessed, the problems of having to devote large numbers of assessors to determine the effectiveness of the various systems is removed. In addition, since this is an automatic task it can be repeated giving general rankings of effectiveness. In the next section, we present our evaluation method and results from the known item search task.

2. EVALUATION METHOD

Our method for evaluating search engine rankings is as follows and is based on a proposal by Chris Buckley made on the TREC Web Track mailing list [6]. We construct a large number of query-document pairs. Queries are mined from search service query logs and documents are mined from the Open Directory Project (ODP). We then issue the queries to the search engines, collect the results, and find the rank at which the engine returns the document we have paired with that query. We score each ranked list using the reciprocal rank of the target document. The overall score for a search engine is the mean reciprocal rank over all query-document pairs.

For this method to be useful, the query-document pairs need to be both reasonable and unbiased. Since we do not require that the document be the most relevant for a query (this might be difficult to determine, although it is commonly the goal in known item searching), it could be the case that the best search engine for that query is the one that ranks the document *lower* than the others. However, if the documents are reasonably good matches to the query, then in the aggregate, the better engines will be those that rank the documents higher.

If the documents are biased such that they favor some particular search engine, then the results will not be reliable. For example, if we chose as a query's target document the web page retrieved at rank 1 by Google, then it would not be fair to include Google, or metasearch engines which use Google, in the comparison. One method to avoid engine bias might be to manually construct a query according to a random web page. The problem with that is that the queries are then biased and may not be representative of user needs.

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

2.1 Query-document Pairs

To construct the query-document pairs, we began with a 12M-entry log of queries submitted to AOL Search. We eliminated the very small fraction of queries which used special query operators such as '+' and quoted phrases. The log was from taken from a single server over several days. Each server is presented user queries in a round robin fashion, thus each server log is a sub-set sample of the larger set of queries.

We drew our target documents from the Open Directory Project, also known as the ODP. Each page in the ODP is compiled by a human editor, and consists of a list of titles for web pages linked from that directory page, each accompanied by a short description. The browser formatting is similar to Yahoo!, but the data itself (which can be freely downloaded) is stored in RDF format [7]. A directory entry title does not necessarily correspond to the title of the web page pointed to, since the directory page is composed by a human editor. Since it is freely available, the Open Directory is used by several search engines in different ways, but rarely as the prime search collection.

We indexed the entry titles and links in the ODP, excluding the Adult, World, Netscape, and Kids & Teens sub-trees. We then matched each query in the query log to an exact directory entry title in the ODP. For example, we matched the query "alpha technologies" to the ODP entry titled "Alpha Technologies", which points to the web site http://www.alphafittings.com/. Most queries in the log did not match any title in the ODP. This matching process resulted in nearly 41,000 query-document pairs.

We then constructed three random samples of 500, 1000, and 2000 pairs respectively. The pairs in the samples were constrained such that:

1. The query is between one and four words long
2. The document URL is longer than just a hostname (i.e., there is at least one path component)
3. The query does not appear verbatim in the URL

The last two constraints were intended to avoid a query like "foo bar" matched trivially to http://www.foobar.com/. The process of constructing the query-document pairs described above is completely automated from a query log and the Open Directory resource.

2.2 Search Engine Results

Each engine was queried for the respective query sets, 500, 1000, and 2000. The MRR was calculated for each engine for each set and Table 1 gives the details of those query sets. Using our original query log of 12 million as a population size, and limiting sampling error to 3%, a sample size of 756 pairs would be needed for a 90% confidence interval. For a 95% confidence interval a sample size of 1067 and for 99% a query sample size of 1843 would be needed.

Since our sample size of 2000 exceeds our sample size needed for a 99% confidence interval we examine those results in more detail. With 2000 samples, our sampling error is 2.2% thus showing that E1 and E3 are equivalent with about a 1% difference in MRR. E3 and E4 are only about 2.5% different. Thus, three of the five engines are within 3.8% difference or equivalent given the sampling error. Further E4 and E5 are 14% different putting E5 as clearly the lowest performing service for known item search. E2 does 55% better than the next best search service and 91% better than the worst showing that for known item searching it is clearly out-performing the other services.

Table 1 Mean Reciprocal Ranking of Search Services

	500		1000		2000	
	MRR	Found	MRR	Found	MRR	Found
E1	26.69%	200	24.94%	356	22.70%	673
E2	38.39%	254	36.53%	484	35.68%	972
E3	26.23%	201	24.63%	347	22.96%	676
E4	22.11%	171	24.01%	335	21.66%	627
E5	18.82%	128	18.65%	250	18.63%	497

3. CONCLUSION

We presented an automatic navigational task evaluation approach that finds the effectiveness of various web search services automatically. Our approach builds upon the idea that web searches have a navigational intent and providing an automatic way to evaluate effectiveness can guide both research and commercial solutions to better effectiveness for end users. Additionally we show that while most engines are roughly the same in terms of effectiveness, there is a considerable gap between the best and worse in terms of MRR.

4. REFERENCES

[1] D. Hawking, N. Craswell, P. Thistlewaite, and D. Harman. "Results and challenges in web search evaluation. ", in Proc. Eighth Int'l World Wide Web Conf., pages 243-252, May 1999.

[2] Amit Singhal and Marcin Kaszkiel, "A case study in web search using TREC algorithms", in Proc Tenth Int'l World Wide Web Conf., pages 708-716, 2001.

[3] M. Gordon and P. Pathak, "Finding information on the world wide web: the retrieval effectiveness of search engines", IP&M, 25(2):141-180, 1999.

[4] Leighton, H. Vernon and Jaideep Srivastava. "First 20 Precision among World Wide Web Search Services (Search Engines)," JASIS 50 (July 19, 1999): 870-881.

[5] D. Hawking, TREC Web Track Guidelines, 2001, http://www.ted.cmis.csiro.au/TRECWeb/guidelines_2001.html

[6] Chris Buckley, TREC Web Track mailing list, 2001.

[7] Open Directory Project, http://www.dmoz.org

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