On Evaluating Web Search With Very Few Relevant Documents

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

ABSTRACT

Many common web searches by their nature have a very small number of relevant documents. Homepage and "named page" searching are known-item searches where there is only a single relevant document. Topic distillation is a special kind of topical relevance search where the user wishes to find a few key web sites rather than every relevant web page. Because these types of searches are so common, web search evaluations have come to focus on tasks where there are very few relevant documents. Evaluations with few relevant documents pose special challenges for current metrics. In particular, the TREC 2003 topic distillation evaluation is unable to distinguish most submitted runs from each other.

Categories & Subject Descriptors:
H.3.4 [Information Storage and Retrieval]: Information Search and Retrieval – *performance evaluation*

General Terms: Experimentation, Measurement

Keywords: measurement error, web search

1. INTRODUCTION

Web search evaluations in the Cranfield tradition began in TREC-7 with the Very Large Corpus track and have continued in the Web Track. The early search tasks were either classical ad hoc search or high-precision search, but following trends on the web, recent TREC Web evaluations have focused on known-item search and topic distillation. These latter search tasks both presume a very small set of relevant documents. For topic distillation, typically less than 10 documents are relevant to a topic. For known-item search, only a single item is desired [1].

Recently, Voorhees and Buckley advocated measuring the sensitivity of test collections in order to empirically determine how well they distinguish the evaluated systems [2]. Their procedure determines the minimum number of search topics needed to differentiate two systems by a minimum amount at a given error rate. The procedure is as follows: select two disjoint subsets of the topics at random and measure the performance of the systems in each subset, counting the number of times two systems swap their relative order in the ranking of systems. The "swap rate" indicates the lack of confidence in a system ranking derived from topic subsets of that size. Topic subsets are sampled up to half of the full topic set, since each pair must be disjoint. By extrapolating to the full topic set, one can determine the minimum absolute difference in score which can be determined within the confidence interval. In this paper, we conduct this analysis for several recent TREC Web test collections, and show that while known-item search is quite stable, topic distillation with few relevant documents presents serious problems for current metrics.

2. KNOWN-ITEM SEARCH

A known-item search is where the user is looking for one particular page. In TREC there have been two kinds of known-item search evaluations. In homepage search (HP), the target is the homepage of a person, a project, an organization, etc. In named page search (NP), the user is looking for one particular page that they recall seeing before, a sort of "mental bookmark." Named pages can include homepages, and did in the TREC 2002 named page evaluation. In TREC 2003, known-item search topics were broken into equal numbers of homepage and (non-homepage) named page topics.

In TREC, known-item searching performance is measured using the reciprocal rank (RR) of the target document in the system's ranked list. Thus, if the system retrieves the document at rank 4, the RR is 0.25 for that topic. If the target is not retrieved then the system receives 0 for that topic. The overall measure is the mean reciprocal rank (MRR) computed over all topics. In practice, because of URL aliasing and page duplication, there can be more than one target document, in which case the RR is that of the highest ranked target retrieved.

Table 1 shows the minimum absolute differences in score required to distinguish two TREC known-item search systems with a 5% chance of error. Smaller differences are not stable because there is a good chance that the two systems would compare differently using two disjoint sets of topics. Since experimental results are often stated in terms of relative difference in score, rather than absolute, Table 1 also gives the top MRR score in each evaluation and the percent

Collection		#Topics	Abs.diff	Top MRR, % diff
2001	HP	145	0.045	0.774, 5.8%
2002	NP	150	0.045	0.719, 6.3%
2003	NP/HP	300	0.04	0.727, 5.5%
	NP only	150	0.065	0.688, 9.4%
	HP only	150	0.06	0.815, 7.4%

Table 1: Minimum absolute differences in MRR in the TREC known-item web evaluations.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK
ACM 1-58113-881-4/04/0007.

improvement that the absolute difference would represent.

The behavior of the MRR measure on these three collections are remarkably similar compared to the TREC ad hoc collections, where the minimum MAP difference varies more widely. This implies that the known-item topics are fairly even in terms of their difficulty.

In addition to the full known-item test collections, we looked at the named-page and homepage subsets of the 2003 collection. Table 1 indicates the minimum difference needed for an evaluation using those topics alone. The subsets have a higher required difference than the full set because they only have 150 topics. In contrast, if a random sample of 150 topics from the full set (75 of each type) is used, an absolute difference of 0.07 is required for 95% confidence. The systems in TREC 2003 appear to perform differently on the two types of topics, and it is probably a good idea to evaluate named-page and homepage search separately if as few as 150 topics are used.

3. TOPIC DISTILLATION

Topic distillation is a variation on topical relevance search. Classically, a relevant page is one which contains even a very small amount of topical information. A user in a topic distillation scenario is searching within a broad subject, and seeks a small set of key web sites on the topic which cover it broadly; the homepages of these sites are the "relevant documents" for topic distillation. The first topic distillation evaluation in TREC 2002 had between 1 and 188 relevant documents per topic (mean=32), which was larger than desired although still much less than a typical ad hoc collection. In TREC 2003, topics had between 1 and 84 relevant documents (mean=10).

Topic distillation performance is measured with precision at the top 10 or 20 documents retrieved (P@10, P@20), corresponding to a user model of success within the first page of hits from a web search engine. However, these measures have problems when there are very few relevant documents. If there are fewer than 10 relevant documents for a topic, then a system can never achieve a P@10 of 1.0. When a measure has a different range for each topic, it averages poorly across them. Also, precision at a fixed cutoff yields scores (and averages) that are quantized, that is, they can only take a discrete set of values in their full range. This problem is not specific to topic distillation, but applies whenever these measures are used with very few relevant documents per topic.

An alternative used in TREC 2003 is R-precision, which measures the precision in the top r documents where r is the number of relevant documents for that topic. A topic with 8 relevant documents would be measured using P@8, while a topic with 15 would use P@15. This solves the averaging problem, since per-topic score always lie within [0,1]. However, when there are so few relevant documents R-precision exhibits even worse quantization than P@10 within a topic. Moreover, R-precision is a harder measure to do well on: for the topic with 8 relevant documents, a system could achieve 0.5 P@10 with four correct documents in any of the top 10 ranks; R-precision holds the system to the top 8.

Table 2 shows the sensitivity of P@10, R-precision (R-Prec), and average precision in the 2002 and 2003 topic distillation collections. The difference between the two collections is striking. The measures in the 2002 task behave as they do in ad hoc collections: average precision and R-precision are less sensitive than P@10, and the relative difference over the top score is very similar. However, in the 2003 topic distillation collection average precision and R-precision are much more sensitive than P@10. Moreover, the minimum absolute difference in P@10 covers nearly a third of the entire score range, meaning that most TREC 2003 topic distillation systems are indistinguishable by score.

	2002	Top score, % diff	**2003**	Top score, % diff
P@10	0.05	0.251, 19.9%	0.035	0.128, 27.3%
R-Prec	0.04	0.215, 18.6%	0.09	0.164, 55%
Ave. Prec	0.035	0.190, 18.5%	0.075	0.154, 48.6%

Table 2: Minimum absolute differences (along with top scores and relative differences from the top score) in topic distillation for various proposed measures.

This difference is entirely due to having fewer relevant documents in TREC 2003. This makes the task much harder (as evidenced by the decrease in top scores), which in turn affects the sensitivity because most systems are performing poorly. The quantization of R-precision further causes that measure to be much less stable than P@10.

For TREC 2003, we also computed the stability of "interpolated precision at 0% recall" (P@0) which is the highest precision achieved at any point in the ranking. This measure is the same as MRR when there is only one relevant document, and is achieved early in the ranking otherwise. The minimum absolute difference for P@0 is 0.11, much too unstable to consider using in this evaluation. However, if we compare it to MRR, it turns out that P@0 in topic distillation is about as sensitive as MRR if only 50 topics are used. The sensitivity of MRR for 50 known-item topics reaches the 5% error rate at a minimum absolute difference of about 0.145, and P@0 for topic distillation at about 0.115. This implies that the current measures for topic distillation could be stable if there were more topics. Since the known-item search tasks seem to be stable with as few as 100 topics, we recommend this as a minimum for topic distillation.

4. CONCLUSION

Web search evaluations have recently focused on search tasks where only a few documents are relevant to a topic. Most common evaluation measures, such as precision at high rank, are not well-behaved in this situation. In known-item searching, where the user is looking for one particular document, evaluations achieve stability by using many topics. For topic distillation, fifty topics is not enough to effectively distinguish systems using these measures. In the future, such evaluations should employ more topics. Alternatively, new effectiveness measures which are less sensitive to the number of relevant documents should be explored.

5. REFERENCES

[1] Nick Craswell, David Hawking, Ross Wilkinson, and Mingfang Wu. Overview of the TREC 2003 Web track. In *Proceedings of TREC 2003*, Gaithersburg, MD, November 2003.

[2] Ellen M. Voorhees and Chris Buckley. The effect of topic size on retrieval experiment error. In *Proceedings of ACM SIGIR 2002*, pp. 316–323, Tampere, Finland, August 2002.