Inf Retrieval (2007) 10:491–508
DOI 10.1007/s10791-007-9032-x

Bias and the limits of pooling for large collections

Chris Buckley · Darrin Dimmick · Ian Soboroff · Ellen Voorhees

Received: 12 October 2006 / Accepted: 31 July 2007 / Published online: 21 August 2007

Abstract Modern retrieval test collections are built through a process called pooling in which only a sample of the entire document set is judged for each topic. The idea behind pooling is to find enough relevant documents such that when unjudged documents are assumed to be nonrelevant the resulting judgment set is sufficiently complete and unbiased. Yet a constant-size pool represents an increasingly small percentage of the document set as document sets grow larger, and at some point the assumption of approximately complete judgments must become invalid. This paper shows that the judgment sets produced by traditional pooling when the pools are too small relative to the total document set size can be biased in that they favor relevant documents that contain topic title words. This phenomenon is wholly dependent on the collection size and does not depend on the number of relevant documents for a given topic. We show that the AQUAINT test collection constructed in the recent TREC 2005 workshop exhibits this biased relevance set; it is likely that the test collections based on the much larger GOV2 document set also exhibit the bias. The paper concludes with suggested modifications to traditional pooling and evaluation methodology that may allow very large reusable test collections to be built.

Keywords Test collections · Pooling · Sampling bias

C. Buckley
Sabir Research, Inc., Gaithersburg, MD 20878, USA
e-mail: cabuckley@sabir.com

D. Dimmick · I. Soboroff (✉) · E. Voorhees
Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899-8940, USA
e-mail: ian.soboroff@nist.gov

D. Dimmick
e-mail: darrin.dimmick@nist.gov

E. Voorhees
e-mail: ellen.voorhees@nist.gov

1 Introduction

A retrieval test collection is a laboratory tool that allows researchers to compare the effectiveness of different retrieval approaches. First used in the Cranfield experiments (Cleverdon 1967), the test collection evaluation paradigm provides an abstraction of operational retrieval tasks by substituting a static set of *relevance judgments* for the complex interactions of a live searcher. The abstraction allows researchers to quickly compare competing retrieval technologies in a controlled laboratory setting at low cost. Of course, to be an effective tool, a test collection must accurately reflect the relative quality of the different retrieval technologies.

The goal of a test collection construction method is to produce effective test collections at an affordable cost. Of the three components of a test collection—the document set, the set of information need statements called topics, and the relevance judgments that indicate which documents should be retrieved in response to a given topic—the relevance judgments are the most expensive to produce. The first test collections used complete relevance judgments, meaning that every document was judged by a human as *relevant* or *nonrelevant* for each topic (Cleverdon 1967). However, complete judgments are too expensive to obtain except for very small document sets. The majority of test collections in current use have been built through community evaluation workshops such as TREC,[1] NTCIR,[2] and CLEF,[3] which use a collection building technique known as *pooling* (Sparck Jones and van Rijsbergen 1975). In pooling, a set of documents to be judged for a topic (the "pool") is constructed by taking the union of the top λ documents retrieved for the topic by a variety of different retrieval methods. Each document in the pool for a topic is judged for relevance, and documents not in the pool are assumed to be irrelevant to that topic. The quality of the resulting collection is known to be dependent on both the diversity of the retrieval methods and the pool depth (λ) used to form the pools (Zobel 1998), but with appropriate controls enough relevant documents are found such that retrieval methods (both those that contributed to the pools and others) can be fairly compared. Pooling thus provides a way of judging only a small fraction of the document set for each topic, allowing test collections to be built that contain document sets several orders of magnitude larger than would be possible with complete judgments.

The crucial assumption of pooling is that the sample of relevant documents found by judging just the pool is unbiased with respect to different retrieval approaches. This paper will use the term "bias" when referring to a collection to mean that the relevance judgments for that collection are a biased (non-random) sample of the true relevance set, and thus the test collection may unfairly rank some class of retrieval methods. As used in current practice, pooling allows sufficiently many of the relevant documents to be found that the judgments can be considered approximately complete and hence unbiased. But collection building is done on a budget, and those budgets preclude using substantially greater pool sizes for ever larger document sets. For a constant pool size the pool represents an increasingly smaller percentage of the total document set as the document set size increases. At some point the assumption of approximately complete judgments necessarily becomes invalid. Does the assumption of an unbiased sample of judgments also become invalid?

[1] http://trec.nist.gov/

[2] http://research.nii.ac.jp/ntcir/

[3] http://www.clef-campaign.org/

This paper shows that, unfortunately, the assumption of unbiased judgments is violated when traditional pooling is used with a constant pool size and increasing document set size. In particular, we show that pools created during the TREC 2005 workshop exhibit a specific bias in favor of relevant documents that contain topic title words. These documents are retrieved by systems that are behaving reasonably, in that they rank documents containing the topic words first. As the document set size grows, these documents fill the pool, squeezing out other kinds of relevant documents. The result is a judgment set that will unfairly rank other types of queries and possibly other retrieval approaches.

The paper is organized as follows. The next section provides general background on TREC and related work on test collection building methodologies. Section 3 introduces a formal measure of the predominance of documents containing topic words in a given set of documents. Using a moderate-sized collection and related TREC submissions to illustrate the issues, Sect. 4 examines the relationship among the concentration of topic-word documents, the concentration of relevant documents, and pool depth. Section 5 then extends these findings to the significantly larger test collections built in the TREC terabyte track. The final two sections offer possible ways forward in building large, fair test collections and summarize the consequences of possible bias on retrieval experiments.

2 Related work

TREC (Text REtrieval Conference) is an annual series of workshops that since 1992 has investigated the state-of-the-art of various information retrieval tasks, and has provided resources such as test collections and evaluation methodologies for further research on those tasks (Voorhees and Harman 2005). Each year, several focus areas known as *tracks* are defined. Research groups participate by submitting one or more *TREC runs* showing the results of running their systems on these common tasks. The performance of the runs is evaluated, possibly by an evaluation methodology developed for the track, and the systems and results are presented at the annual workshop.

There are typically three main results for each TREC track:

1. A snapshot of the current state-of-the-art for the particular area addressed by the track including the retrieval approaches that currently work best.
2. An evaluation methodology with measures that can be used to compare approaches. This paper focuses on the ad hoc retrieval task in which systems produce a ranked list of documents in response to a topic, evaluated using the Cranfield paradigm. The main retrieval effectiveness measure used in the paper is Mean Average Precision (MAP) (Buckley and Voorhees 2005).
3. A test collection to enable future research in the area. Historically, test collections have driven much of the research in information retrieval, and a major goal of TREC is to establish a set of *reusable* test collections in various areas of information retrieval. A test collection is reusable if it fairly evaluates retrieval runs that did not contribute to the pools used to construct the collection.

The problem of determining an adequate set of relevance judgments for effective test collections was well-known long before TREC. Tague-Sutcliffe (1992), for example, offers several suggestions for estimating recall in large collections, some of which had been attempted (e.g., Blair and Maron 1985). However, until large test collections were constructed as part of TREC and made available to the research community, no one had

studied the problem of biased relevance judgments, or whether methods for creating these large test collections could inherently lead to bias.

Zobel (1998) looked specifically at recall underestimation and system bias in the TREC-3, 4, and 5 collections. He showed that many relevant documents are missed by pooling, especially for topics that themselves have many relevant documents. Thus, recall may be generally overestimated by pooled test collections. Furthermore, the effectiveness of a pooled system can be underestimated if that system finds disproportionately many relevant documents relative to the other pooled systems.

Additionally, if a very effective system does not contribute to the pool, it may find many relevant documents that were not pooled and thus its effectiveness will be underestimated as well. Zobel measured this effect using a test which we call the "leave out uniques" or LOU test. In this test, each run that contributed to the pools is evaluated first using the official set of judgments published for that collection, and then using the judgment set that results by removing those documents that were added to the pool by the current run only (the run's "unique relevant"). The difference in the evaluation scores in these two cases averaged over all pool runs is a measure of the how much variation in evaluation scores a subsequent user of the collection may experience. Because different runs from the same organization frequently have very high overlap in the documents retrieved, in this paper we use a more stringent variant of this test that removes all documents from the judgment set that were contributed solely by runs from the same group. Note that if all the runs use the same or very similar systems, then the overlap among them will be very high and the LOU test will not detect any bias that may be present in the judgments.

The LOU test has been used to assess the quality of a variety of the TREC test collections over the years. For example, Zobel (1998) reported average differences of 0.5% for the TREC-5 ad hoc collection and 2.2% for the TREC-3 collection.[4] The more stringent variant of the test showed average differences of 0.8% for the TREC-8 ad hoc collection (Voorhees and Harman 2000) and 1.1% for the TREC-9 web collection (Hawking 2001). Since evaluation scores can change by this amount by using different relevance assessors (Buckley and Voorhees 2005) or topic subsets (Voorhees and Buckley 2002), this level of difference has been considered to be in the noise and the collections deemed reusable. In contrast, some of the TREC cross-language collections exhibited larger average differences such as the 6.3% difference for the TREC-8 cross-language collection (Voorhees and Harman 2000) and the 8.0% average difference for the TREC 2001 cross-language collection (Voorhees and Harman 2002). These differences suggest that caution needs to be used when interpreting the evaluation results for these collections when many unjudged documents are retrieved early in the ranking.

As a result of the increased interest in creating large test collections, researchers have become very interested in mechanisms that reduce the cost and effort involved in their production. Cormack et al. (1998) proposed two techniques: iterative searching and judging (ISJ) and move-to-front pooling. With ISJ, relevance assessors perform multiple searches while judging documents for relevance, in order to try and recover as many relevant documents as possible. This approach may result in relevance judgments biased towards the search system used. Soboroff and Robertson (2003) attempted to address bias towards the system using relevance feedback and system fusion; Sanderson and Joho (2004) extended this to investigate how few systems are needed.

[4] Zobel used 11-point average precision scores in his original tests. All LOU test results in this paper are based on differences in MAP scores.

In move-to-front pooling (MTF), the pooled systems are each assigned an initial uniform priority. The next document to be judged is the highest-ranked unjudged document from the system with the highest priority. (If multiple systems all have the highest priority, one is chosen at random.) If that document is not relevant, the priority of that system is reduced, but if it is relevant, that system is assigned the maximum priority. MTF finds as many relevant documents as pooling does after judging many fewer total documents.

More recently, Aslam et al. (2006) proposed a method to accurately estimate system effectiveness using a random sample of the pool. The sample is drawn according to a distribution over pairs of pooled documents, such that the most likely documents are those which best help refine the measure being estimated. A related approach is that of Carterette et al. (2006). In this scheme, the goal is to judge the minimum number of documents necessary in order to rank the pool systems correctly. The next document selected for judging is the document whose relevance or non-relevance would best differentiate the systems at hand.

Any method of selecting documents to judge from early ranks of systems may incorporate a bias that reduces future reusability of the collection. This was Zobel's hypothesis, but all of the large collections of the time were adequately protected by their relatively large pools. None of the more recent methods proposed to scale to large collections specifically address issues of relevance judgment bias.

3 Characterizing pool documents

This examination of the problem of biased relevance judgments was motivated by the test collection constructed as a joint product of the TREC 2005 HARD (Allan 2006) and robust retrieval (Voorhees 2006) tracks known as the AQUAINT collection. This section describes the AQUAINT collection, which is based on a topic set constructed from topics used in earlier TREC collections and a new document set. One particular run that contributed to the pools during construction of the AQUAINT test collection, run `sab05ror1`, was a *routing* run that used the relevance information from the earlier collections to build queries to retrieve documents from the new set. The results of a LOU test on the AQUAINT collection showed that the `sab05ror1` run evaluated very differently depending on whether its uniquely retrieved relevant documents were included in the relevance judgment set. The section concludes with the definition of a measure to characterize pool documents by the percentage of topic words they contain.

3.1 The AQUAINT test collection

The document set of the AQUAINT test collection is the set of documents contained in *The AQUAINT Collection of English News Texts* available from the Linguistic Data Consortium.[5] The topic set is a 50-topic subset of the 249 topics that have previously been used with the documents contained on TREC disks 4 and 5. The 50 topics were selected because participating systems in past TRECs had performed poorly on them. The TREC 2005 HARD and robust retrieval tracks shared a common interest in improving retrieval effectiveness for difficult topics, so both tracks used the same 50 topics on the AQUAINT documents. The pools for the collection were created using at most two runs per HARD

[5] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T31 as of July 2007.

track participant (one baseline and one non-baseline run), two NIST-produced runs that served as baseline runs for the HARD track, and one run per robust track participant, resulting in a total of 50 runs contributing to the pools. The combined result of this effort and the earlier TREC tasks is two distinct test collections that share a common topic set. For convenience, we'll call the collection formed during the earlier TREC tasks the Disks4&5 collection.

Some of the features distinguishing the AQUAINT collection and pooling from previous collections, and in particular the Disks4&5 collection, are

1. Size: the AQUAINT collection is the largest non-web collection created within TREC. It is 1,033,461 documents and about 3.5 GBytes of text. Disks4&5 is 528,542 documents and about 2.0 GBytes of text.
2. Newswire consistency: the AQUAINT collection is entirely newswire stories while Disks4&5 also contains Federal Register and FBIS documents. The end result is that the AQUAINT collection contains about 3 times the amount of newswire stories, the source of the vast majority of relevant documents, as Disks4&5.
3. Group and run variability: The retrieval runs contributing to the AQUAINT pool were more diverse than the normal ad hoc set of runs. The HARD track non-baseline runs all involved human interaction, and several of the robust runs exploited the Disks4&5 judgments, methods that historically have produced good pools (Voorhees and Harman 2000). The runs represented many different retrieval models and many different sources of expansion terms.
4. Shallower pools: The top 55 documents per topic per pool run were added to the pools for the AQUAINT collection, as opposed to the typical 100 documents for Disks4&5. The average number of judgments per topic was roughly the same in both collections, but the large number of diverse runs for AQUAINT meant the runs could not be judged as deeply.

The AQUAINT judgment pool was expected to be a good pool based upon the experience of NIST. It was shallower than desirable, but experiments on older collections simulating a depth-50 pool showed only a small change in the reliability of the pool. In addition, the diversity of the runs was expected to ensure a reliable sample of all types of relevant documents.

3.2 The `sab05ror1` routing run

One of the robust track runs was a pseudo-routing or relevance feedback run from Sabir Research done by Chris Buckley (2006). In a routing run, in addition to the original topic a system has access to relevance information about previously seen documents. The system then tries to predict whether a new incoming document is relevant and should be routed to the user. For this run, the system constructed each query based on the original topic and the relevant documents found on the Disks4&5 collection, considering the entire Disks4&5 document collection as being previously seen. The query was then run on the AQUAINT documents which were then ranked according to how likely they were to be relevant. The entire process was automatic, though the run was classified as a manual run since it depended on the human supplied Disks4&5 judgments.

The Sabir Research system (SMART v15) constructed the queries for `sab05ror1` by taking the 250 terms that occurred most often in the relevant documents of Disks4&5, and then adjusting the weights on those terms to maximize performance on Disks4&5. The

final queries achieve a MAP score of 0.8995 when run on Disks4&5, showing that the queries do an excellent job of describing the relevant documents and distinguishing them from the non-relevant documents on the Disks4&5 collection.

The final queries when run on the AQUAINT collection achieved a MAP score of only 0.2663, slightly above the median among the pool runs and only a bit better than Sabir Research runs that did not use any relevance information. There were obvious cases of both over-fitting and under-fitting the topics, and the approach undoubtedly could be improved in the future. But what instigated the investigation here was the fact that the run uniquely contributed a very large number of relevant documents to the judgment pool.

The `sab05ror1` run was designated by Sabir Research as a pool run, and as such, 2,750 documents entered the judgment pool (50 topics to a depth of 55 documents). About 405 of these 2,750 were uniquely relevant, i.e., were relevant documents that only this run contributed to the pool. This figure was a major anomaly for several reasons. First, the ratio of number uniquely relevant to the number contributed to the pool is higher than any other run in TREC's history for the major ad hoc collections. Second, the runs in past TREC's with high numbers of unique relevant documents have been manual runs where a user has manually filtered out at least some of the non-relevant documents from the top documents from which the pools are drawn. There were no judgments done on the AQUAINT document set in creating the `sab05ror1` run. Third, the runs in past TREC's with highest numbers of unique relevant documents have tended to be among the best runs (highest MAP) in the set of runs. Runs with average effectiveness, like the `sab05ror1` run, usually do not contribute many unique relevant documents to the pools.

As would be expected, a LOU test run on `sab05ror1` shows the largest difference of any pool run, 23%. In other words, if `sab05ror1` had not contributed to the pool, its MAP score would have been 0.202 instead of 0.266 and it would not have been possible to fairly compare `sab05ror1` against other runs. The second highest difference was the University of Maryland's manual run, `MARY05C1`, a submission to the HARD track that had a 12% decrease. `MARY05C1` was a much more normal LOU test outlier, being the top run in the pool (MAP of 0.469) and having extensive user judgments filtering the top documents. All other runs had decreases of 8% or less.

This anomalous Sabir run needed to be examined in more detail to find out if it was just a fluke, or whether it indeed signaled that other runs that did not contribute to the AQUAINT pools could be unfairly evaluated with the AQUAINT relevance judgments. Looking at `sab05ror1`, we found it retrieved its unique relevant documents across most topics—it was not just a local effect confined to a couple of topics. Since the queries in a routing run are designed to describe the relevant documents in the training set, we examined how the Disks4&5 and AQUAINT relevant document sets differed and found that topic title words occur more frequently in the AQUAINT judged relevant set. Note that we are not comparing relevance judgments, but examining the relevant documents themselves.

More formally, we define the general measure *titlestat* as the percentage of a set of documents that a title word occurs in, computed as follows. For a single topic T and a set of documents C,

$$titlestat_T = \frac{1}{t_T} \sum_{t \in T} \frac{|C_t|}{\min(|C|, df_t)}$$

where t is a title word, t_T is the number of title words in that topic, and C_t is the number of documents in C that contain t. df_t is the collection frequency of t; this normalization is necessary in case t is a very rare term with a collection frequency smaller than $|C|$.

Individual per-topic titlestat values are then averaged over the set of topics. A maximum value of 1.0 is obtained when all the documents in the set contain all topic title words; a minimum value of 0.0 means that all documents in the set contain no title words at all. We now define the more specific measure *titlestat_rel* for a collection as titlestat computed over the relevant documents for each topic. The titlestat_rel value for the Disk4&5 collection is 0.588 and for the AQUAINT collection it is 0.719. More strikingly, if we compute titlestat_rel for individual topics, the per-topic value is greater for the AQUAINT collection than it is for the Disks4&5 collection for 48 of the 50 topics, an extremely highly significant difference that has essentially no probability of happening by chance ($p = 6.25 \times 10^{-10}$ according to a paired *t*-test).

Given that the change in the frequency of topic title words occurring in relevant documents did not happen by chance, what was the cause? As described earlier, there are several notable differences between the two collections: the Disks4&5 document set contains some documents that are not news stories while the AQUAINT document set consists solely of newswire articles (and from a later time period); the Disks4&5 document set is much smaller than the AQUAINT document set; and different relevance assessors judged the same topic in the two different collections. Yet both document sets are essentially news collections, and it is unlikely that any of these differences would change the frequency of words occurring in a relevant document. If anything, title words would be expected to occur more frequently in the longer Disks4&5 documents. A much more plausible explanation is that pooling failed to add to the AQUAINT pools many of the documents with fewer title words that would have been judged relevant. In other words, if we knew the true set of relevant documents for the AQUAINT collection, the difference in titlestat_rel values between the two collections would be smaller. The following section explores this contention in more detail.

4 Collection size and pooling

Our hypothesis is that the pooling used in TREC 2005 produced a sample of the true judgment set for the AQUAINT collection that is biased against documents containing few topic title words. Future systems that do not have a similar bias in their retrieved sets could be evaluated unfairly by the collection.

To make our counting argument, we define a new variant of titlestat, titlestat_rank, that uses the documents retrieved at a given rank over a set of retrieval runs as the document group. Within a single topic T and rank k, the document set C^k contains all the documents retrieved by a set of runs at rank k. Note that the same document may be retrieved by more than one run at the same rank; titlestat_rank keeps multiple instances of the same document in the group when this occurs. Following the notation from the definition of titlestat above,

$$titlestat\_rank_{T,k} = \frac{1}{t_T} \sum_{t \in T} \frac{|C_t^k|}{\min(|C_k|, df_t)}$$

where $|C_t^k|$ is the number of documents in C^k that contain term t. Titlestat_rank for a given rank is then averaged across the topics. Figure 1 shows a plot of titlestat_rank for both the AQUAINT and Disks4&5 collections. For the AQUAINT collection, all runs submitted to either the TREC 2005 HARD or robust track are used as the run set; for the Disks4&5 collection all runs submitted to the TREC 2004 robust track, restricted to the 50 topics used in the TREC 2005 track, are used as the run set.

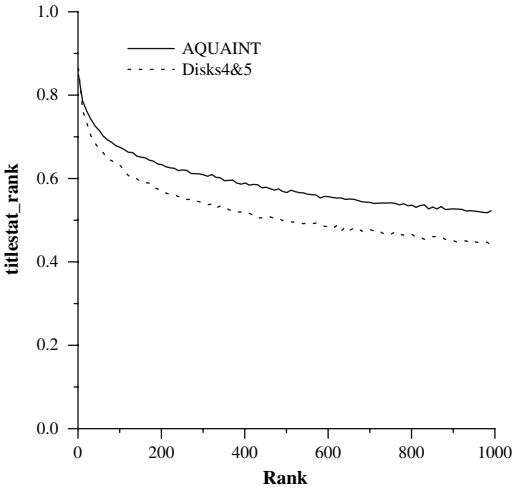

Fig. 1 titlestat_rank values at ranks 1–1,000 for the AQUAINT and Disks4&5 collections

Figure 1 shows that topic words on average occur more often in a document retrieved earlier in a ranking than a document retrieved later in the ranking. This is perfectly sensible: topic title words are intended to be highly descriptive of the subject matter being sought and many retrieval systems purposely weight these terms highly. Further, the TREC tracks that are used to build the collections frequently have conditions requiring submissions to use only topic title words as an initial query (both the TREC 2004 and 2005 robust tracks had such a condition). We should expect such an emphasis in the retrieval results, and Fig. 1 confirms that it exists. For both collections, titlestat_rank values start high, decrease relatively rapidly and then enter a much longer period of slowly declining values. The titlestat_rank values stay higher longer for the AQUAINT collection than for the Disks4&5 collection. Again, this is to be expected. Recall that the AQUAINT document set contains twice as many documents as the Disks4&5 document set. Barring artificial constructs to prevent it, in general the number of documents containing a given word will increase as the total number of documents increases.

It's fine to say that titlestat decreases with rank, but are any documents further down in the ranking, where titlestat is lower, actually relevant? Figure 2 shows the probability that

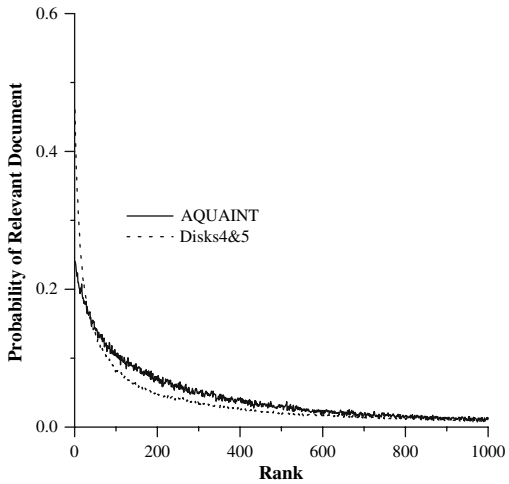

Fig. 2 Probability of retrieving a relevant document at ranks for the AQUAINT and Disks4&5 collections

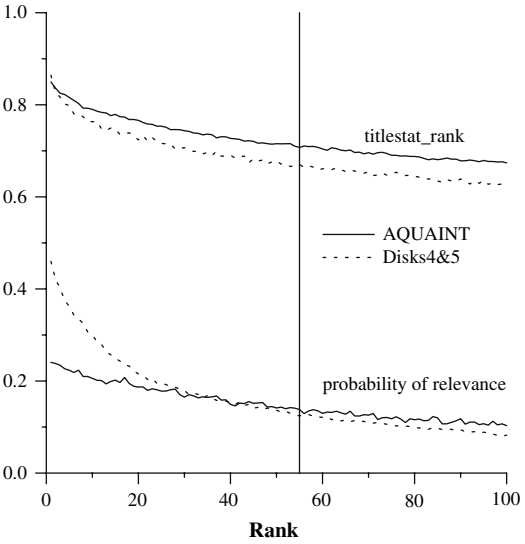

Fig. 3 titlestat_rank and probability of retrieving a relevant document by rank for the AQUAINT and Disks4&5 collections. The vertical line at 55 indicates the pool depth used for the AQUAINT collection

a document retrieved at a given rank is relevant for the same two collections and computed over the same run sets. The probability of a relevant document at a given rank is computed as the number of runs that retrieve a relevant document at that rank divided by the total number of runs, and then averaged over topics. These probabilities are necessarily computed using the known relevant set and may therefore be underestimated. The probability of retrieving a relevant document decreases as rank increases—once again demonstrating that the retrieval runs are behaving sensibly. The probability of retrieving a relevant document in the Disks4&5 collection starts high and drops quickly in the first 30 ranks; the probability of retrieving a relevant document in the AQUAINT collection is much flatter. Hawking and Robertson showed that the number of relevant documents in a collection will tend to increase as the document set size increases (Hawking and Robertson 2003, see Madigan et al. (2006) for a more in-depth discussion of this phenomenon), as the judgment sets for the AQUAINT and Disks4&5 collections also bear out, so the probability of retrieving a relevant document remains greater deeper in the ranking for the AQUAINT collection.

Figure 3 shows the titlestat_rank and probability of relevance graphs restricted to ranks 1–100 superimposed on a single graph. The vertical line at rank 55 indicates the pool depth used for the AQUAINT collection.[6] At rank 55, the retrieval runs on the AQUAINT document set are still in the high titlestat_rank section of the curve, reflecting the larger number of documents containing title words in the document set. The retrieval runs are still in the higher probability of retrieving a relevant document section of the curve as well, reflecting the larger number of relevant documents, and indicating that substantially more relevant documents are likely to be found after this cutoff where titlestat values are lower. For the Disks4&5 collection, rank 55 is after the steep descent of both curves.

[6] The Disks4&5 collection is constructed from topics that were used in different previous TRECs, and the pool for each topic was created in the first TREC in which it was used. There is no common pool depth or pool run set for the Disks4&5 collection, but all topics in the collection were pooled to at least depth 100 in its original TREC.

Zobel (1998) examined the dependence of collection quality on pool depth by down-sampling pools from early TREC collections and noted substantial degradation in collection quality for a depth of 10 but not for a depth of 50. Later TREC collections have occasionally been built using pool depths smaller than 100, but not smaller than 50 (excluding cases such as the early web tracks when there was no claim to be building a reusable test collection). The current data show that collection quality does not depend on some absolute number of judged documents. Rather, the minimum pool depth is relative to the size of the document set. A rational, high-quality retrieval system will retrieve certain types of documents—such as those containing query words—early in its ranking. As document set size increases, the number of documents of this type will also increase. For sufficiently large document sets *relative to the pool depth*, the available space in the pool is filled with this one document type, violating the assumption of an unbiased judgment set.

We can see that this phenomenon is not restricted to topics with many relevant documents. Figure 4 compares the number of relevant documents for each topic in the AQUAINT collection and the difference in titlestat_rel between the two collections. There is no relationship; topics had a larger or smaller occurrence of documents containing title words irrespective of the number of relevant documents.

The `sab05ror1` run from the TREC 2005 robust track demonstrates that this pool-crowding effect did occur during the construction of the AQUAINT collection. Recall that the run's queries used words from the relevant documents of the training collection, rather than relying on topic text. Its retrieved documents contain many fewer topic title words on average than the other runs: titlestat measured on the retrieved set of documents for `sab05ror1` is 0.388 while the average titlestat on the retrieved sets of all the other runs is 0.600. Further, 405 of the 2,750 documents the run contributed to the pools were unique relevant documents (i.e., relevant documents that only Sabir contributed to the pools). The

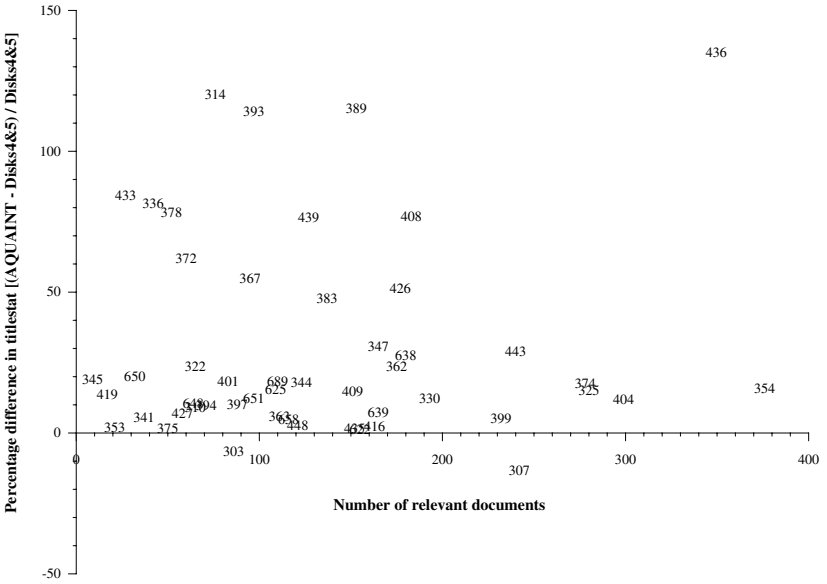

Fig. 4 Difference in titlestat_rel per topic between the AQUAINT and Disks4&5 collections plotted against the number of known relevant documents in AQUAINT. The symbol plotted is the topic number

unique relevant documents contributed by Sabir had a titlestat value of 0.530 compared to the overall titlestat_rel of 0.719 (including Sabir's unique relevant). In other words, the `sab05ror1` run produced substantial numbers of relevant documents with lower title word occurrence that the 49 other diverse runs did not contribute to the pools, demonstrating that such documents do exist, and providing strong evidence of bias in favor of documents with topic title words in the relevance judgments of the AQUAINT collection.

5 The TREC terabyte collections

The goal of the TREC terabyte track is to scale up retrieval evaluation past the gigabyte range to terabytes and hopefully beyond. The expectation when the track began was that traditional pooling would not scale to such large collections since the relevance judgments would surely be incomplete and no methodology for guaranteeing an unbiased sample was known. Organizers were perplexed, therefore, when the collections built in the first two years of the track (TREC 2004 and TREC 2005) appeared to have no obvious flaws. The LOU test did suggest that the TREC 2004 collection should be used with some caution. Groups had large numbers of unique relevant documents, but then there are a large number of relevant documents in total. The mean difference in scores is 9.6% and the maximum difference a large 45.5%, but the collection was built during the first year of a track and overall effectiveness was quite low (magnifying small absolute differences into large percentage differences). The results of the LOU test on the TREC 2005 collection were more reasonable with an average difference of 3.9% and a maximum difference of 17.7%. For both collections, MAP and bpref—a measure designed to be used in the presence of incomplete judgments and shown to be highly correlated with MAP when given relatively complete judgments (Buckley and Voorhees 2004)— ranked systems almost identically. Perhaps pooling was working just fine after all.

The argument in the previous section indicates doubts are justified regarding the viability of traditional pooling for documents sets this large. Figure 5 shows the titlestat_rank values for the set of collections described in Table 1. The collections include the AQUAINT and Disks4&5 collections, the collections built during the first two years of the terabyte track (TB04 and TB05), and two additional collections built during TREC-8 that

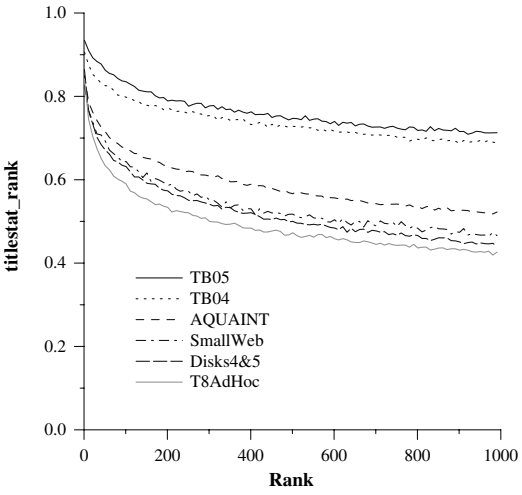

Fig. 5 titlestat_rank over top 1,000 ranks for multiple collections

Table 1 TREC test collection characteristics including number of documents in the collection, type of documents, average pool size, average percentage of the pools that were judged relevant, average percentage change in the LOU test, and titlestat_rel values

Name	Collection size	Document type	Mean pool size	Mean % of pool judged relevant	Average LOU diff in test	titlestat relevant
AQUAINT	1,033,461	Newswire	756.0	17.4	3.2	0.719
Disks4&5	528,155	Mostly news	1617.3	5.5	–	0.588
TREC-8 ad hoc	528,155	Mostly news	2508.3	5.4	0.8	0.688
TREC-8 small web	250,000	Web pages	950.1	4.8	1.8	0.850
TREC 2004 TB	25,000,000	Web pages	1189.1	18.3	9.6	0.889
TREC 2005 TB	25,000,000	Web pages	905.8	23.0	3.9	0.898

provide an interesting contrast. The TREC-8 ad hoc and small web collections are another pair of collections that share a common topic set. The ad hoc collection's document set is a news collection, while the small web collection uses a set of web pages as the document set.

The terabyte track collections were pooled using depths of 85 in 2004 and 100 in 2005. Figure 5 shows the titlestat_rank values for both terabyte collections are very high far into the ranking—well past rank 100. For both collections the titlestat_rank values are approximately 0.8 at rank 100, above 0.75 at rank 200, and approximately 0.68 at rank 1000, much larger values than for any of the other collections. Such high values are easy to understand. The majority of runs submitted to the terabyte tracks were runs using only the topic title as the initial query (such a run was a requirement of the track) and there are many thousands of documents containing an average topic title word in a collection of 25,000,000 documents. Nonetheless, such high values also mean there were exceedingly few documents with low titlestat that were ever judged since such documents never made it into the pools.

For the collections we have examined, the average percentage of the pools that were judged relevant is another indicator of bias in the judgments for the collection. For the TREC ad hoc collections built after TREC-4, approximately 6% of the documents that were judged were relevant (Voorhees and Harman 2000). Table 1 shows that the percentage for the terabyte (and AQUAINT) collections is much larger. A similar counting argument can explain this effect. Since the number of relevant documents increases as the document set size increases, more relevant documents make it into the pools. There is a total of 10,407 known relevant documents in the TREC 2005 terabyte collection, for example, which is a little more than twice as many relevant documents for the TREC-8 ad hoc collection. The problem, of course, is that there are also more relevant documents that *don't* make it into the pools, and the characteristics of the two sets are different.

Titlestat_rel values are computed over the known relevant set, so the titlestat_rel values for the terabyte collections are necessarily very high as well, 0.889 and 0.898 for TREC 2004 and 2005 respectively. This means that any single title word occurred in nearly 9 out of 10 judged relevant documents on average. Yet a close examination of the data in Table 1 shows that an alarmingly large titlestat_rel value alone does not necessarily indicate the presence of a problem with the collection. The absolute value of titlestat_rel is strongly affected by the topic set: the problem with the AQUAINT collection is indicated by the difference between AQUAINT titlestat_rel and the Disks4&5 titlestat_rel on the same topic set, not the absolute value of the AQUAINT titlestat_rel. The TREC-8 small

web collection has a titlestat_rel of 0.850 while the ad hoc collection has a titlestat_rel of 0.688 for the same topics. The titlestat_rel values are greater for the web collection than for the ad hoc collection for the vast majority of individual topics as well. But there is no evidence to suggest that either judgment set is noticeably incomplete (see below), so this implies some other factor, such as document type, also has an effect on titlestat_rel values.

Why do we believe the TREC-8 small web collection has approximately complete judgments? First, all properties of the collection other than the titlestat_rel values support the conclusion: the LOU test has a small average difference in scores, the pool size is large compared to the document set size, and a small percentage of the pool was judged relevant. Second, we went hunting for more relevant documents in the small web collection and could not find a significant number of new relevant documents. Using the TREC-8 ad hoc collection judgments, we constructed a routing run in the same manner as the `sab05ror1` run was constructed and judged the first 15 previously unjudged documents from that run for each topic. The assessor for this run differed from the original TREC-8 assessors, so we also rejudged five previously judged documents per topic as calibration. The agreement on the previously judged documents was within expected bounds, and we found less than one new relevant document per topic on average.

The terabyte collections' high titlestat_rank values explain why the LOU test shows comparatively minor variations in scores: all of the pool runs were at least implicitly targeting title-word-containing documents and so match the bias in the judgments for the collection. These collections are fine for comparing title-word-emphasizing retrieval techniques. Problems will arise for other types of runs that do not have a title word emphasis since these runs will mostly retrieve unjudged documents, and a significant fraction of those unjudged are likely to be relevant.

6 Toward large reusable test collections

Perhaps using the standard TREC pooling depth of 100 instead of 55 would have been sufficient to produce an unbiased judgment set for the AQUAINT collection, but it is clearly insufficient for the larger terabyte collections. For the terabyte document set, getting beyond the flood of title word documents by traditional pooling would require a pool depth that much larger than assessors would be willing to judge or that TREC could afford to have judged. New approaches to building very large, reusable test collections are needed. Several possible approaches, with advantages and drawbacks, are briefly discussed here.

6.1 Engineering the topics

If the topic is precise enough and has a small number of relevant documents, then judgment set bias is unlikely to be a problem. With a collection restricted to narrow topics that can be described by comparatively low frequency title words, systems should be able to go beyond those title words in finding relevant documents to add to the judgment pool. Unfortunately, the AQUAINT results show that it is not just a problem of number of relevant documents. Titlestat on the relevant documents was higher on the AQUAINT collection even on topics with low numbers of relevant documents. But the major drawback to using such engineered topics is that the applicability of any research results will then be limited to tasks where such narrow topics are required; the research results would not indicate anything about retrieval effectiveness in general.

6.2 Forming pools differently

If retrieved documents have different biases depending on the rank at which they are retrieved, then one solution toward unbiased pools is to build pools in ways designed to include documents from deeper in the rankings. Several methods have been shown to locate most relevant documents or to estimate conventional measures using a fraction of the currently judged documents; an assessment regime could apply these techniques within the current pooling "budget" and explore a much deeper pool. One such method that we have examined is move-to-front pooling (Cormack et al. 1998). If we judge the number of documents that would have been judged in a depth-50 pool, but using the move-to-front approach, we would recover 79% of the relevant documents found in the official pool while only judging 48% of the officially judged nonrelevant documents. Thus, move-to-front pooling would permit exploring down to depth 200 or so using the traditional depth-100 pool budget—a savings, but insufficient to remove bias for very large collections. An alternative method is random sampling, which can estimate MAP scores accurately with judgments from 10 to 20% of the traditional pool (Aslam et al. 2005), though how this interacts with bias is unknown. Stratified sampling would allow pushing even deeper into the system rankings. While the stratified sampling approach appears promising, there are a large number of issues that need exploring, including how the strata are determined and how topic variability is taken into account.

6.3 Encouraging different retrieval approaches

A known problem with the terabyte collections is that the runs submitted to the track were very similar to each other. Title word searches were required, and many groups focused on efficiency rather than novel retrieval strategies. It's important to get a wide variety of approaches to avoid unknown biases, and to detect when there is a problem. Participants in a collection-building exercise could be required to perform other types of runs such as manual feedback runs, routing runs, or "query track"-style runs with combinations of multiple manual queries to enrich the pools in the way that manual runs have historically done. However, run diversity is not a complete solution in itself: the AQUAINT collection was formed with many different kinds of runs which involved humans to a lesser or greater extent, and the relevance set is still biased. The run diversity did allow the bias in the AQUAINT collection to be detected. We believe bias exists in the judgments for the terabyte collections only through circumstantial evidence since none of the submitted runs directly suggests the presence of bias.

6.4 Engineering the judgment set

A more radical approach is to continue traditional pooling, but then to downsample the resulting judgments to a fair subset, discarding any judgments not in the subset. Bpref can be used to evaluate the runs with the fair partial judgment set. This approach has the very serious drawback that it is currently unknown how to construct a fair sample. The major advantage of this approach over the others is that we can experiment with it given the current collections. For example, one experiment is to start with a collection with reasonably complete judgments (e.g., TREC-8 ad hoc), construct an artificially biased subset B of its judgments, and then attempt to construct an unbiased subset U of B that gives the same system rankings as the original complete judgment set.

7 Conclusion

Obtaining human judgments is the expensive part of building retrieval test collections. Cost and assessor fatigue prevent judging ever-greater numbers of documents as document set sizes increase. Traditional pooling methodology allowed the formation of test collections that are orders of magnitude larger than collections built with complete judgments, but it, too, depends on document set size. Pooling with a constant pool size fails as collection size grows in that the resulting judgment set becomes a biased sample of the complete judgment set and thus systems might not be fairly compared.

We present evidence that one type of bias, bias towards documents containing topic title words, exists in the judgments for the TREC 2005 AQUAINT collection, and in all likelihood also exists in the TREC 2004 and 2005 terabyte collections. We suggest that given the state-of-the-art of current retrieval systems, such bias will exist in the judgments for any very large test collection built using traditional document pooling techniques.

The consequence of this bias is that some evaluations of systems using these collections may not fairly compare the systems. Comparisons between a system represented in the pool and a system not represented in the pool might be unfair to the system not in the pool. This has always been true for pooled collections to some extent, as the Leave-Out-Uniques (LOU) results have shown. However, while the typical LOU differences have been smaller than the normal experimental error associated with any test collection (see Voorhees and Buckley 2002), there is no reason to believe the unfairness found here is that small. Indeed, the Sabir run suggests it can be much larger. Comparisons between two variants of the same system, or between two systems not included in the pool, might also be affected by the unfairness, again with no known upper bound to the unfairness.

Comparisons between systems represented in the biased pools should be unaffected. In particular, there is no evidence that the track evaluations in the related years are unfair in any way. There also is no evidence that any of the earlier main TREC collections have title word biased judgments. (Some of the cross-language track collections are problematic, as evaluated by LOU numbers, but that is due to low participation in some of the languages). The main ad hoc and routing collections in the past are all on smaller collections, have lower titlestat_rel numbers, have a smaller number-relevant to number-judged ratio, and have explainable LOU numbers for all runs.

In general, the comparison problems on these unfair collections are conservative problems, in the sense that the score of a run affected by the bias in judgments will only be negatively affected (unjudged relevant documents will be counted as non-relevant). If a research publication claims evaluation results show a new method is better than published TREC results, that claim will be unaffected by the unfairness, since any unfairness would only decrease the evaluation numbers of the new method. Similarly, a publication claiming improvements of a variant of the same system when compared to its base system will almost certainly be accurate, assuming the base system level of performance is at least at the level of a typical TREC run. Given present-day systems, it's very unlikely that a decent base system would be affected by the bias while the variant would be less affected, the only situation that would yield an incorrect claim. Thus published research showing improvements on these collections remains valid.

What the existence of unfairness does imply is that some (but not all) new methods may not be able to show improvements on these unfair collections that they would be able to show on fair collections. For instance, a system based on the presence of semantic concepts instead of the presence of specific terms may retrieve a high proportion of unjudged documents which are relevant but which do not contain title words, and thus have a lower

evaluation score than warranted. Note that the mere presence of large numbers of unjudged documents in a system's run should *not* be taken as proof that the run is being evaluated unfairly. There are many reasons that a run may retrieve unjudged documents, of which only a few will cause the run to be evaluated unfairly. Researchers who suspect they are being affected by the bias will need to use other collections.

Even though these collections with biased judgments can be used with care, it is much preferable to construct unbiased judgment sets to begin with. Promising avenues to pursue to build very large reusable test collections include constructing pools using techniques designed to include documents from deeper in the systems' rankings and engineering small, but unbiased, judgment sets.

References

Allan, J. (2006). HARD track overview in TREC 2005: High Accuracy Retrieval from Documents. In *Proceedings of the Fourteenth Text Retrieval Conference* (TREC 2005).

Aslam, J. A., Pavlu, V., & Yilmaz, E. (2005). A sampling technique for efficiently estimating measures of query retrieval performance using incomplete judgments. In *Proceedings of the 22nd ICML Workshop on Learning with Partially Classified Training Data* (pp. 57–66).

Aslam, J. A., Pavlu, V., & Yilmaz, E. (2006). A statistical method for system evaluation using incomplete judgments. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)* (pp. 541–548).

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM, 28*(3), 289–299.

Buckley, C. (2006). Looking at limits and tradeoffs: Sabir Research at TREC 2005. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005).*

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on research and development in information retrieval* (pp. 25–32).

Buckley, C., & Voorhees, E. M. (2005). Retrieval system evaluation. In *TREC: Experiment and Evaluation in Information Retrieval* (pp. 53–75). MIT Press.

Carterette, B., Allan, J., & Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)* (pp. 268–275).

Cleverdon, C. W. (1967). The Cranfield tests on index language devices. In *Aslib Proceedings, 19* (pp. 173–192). Reprinted in *Readings in Information Retrieval*, K. Sparck Jones & C. van Rijsbergen (Eds.), Morgan Kaufmann, 1997.

Cormack, G. V., Palmer, C. R., & Clarke C. L. A. (1998). Efficient construction of large test collections. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on research and development in information retrieval* (pp. 282–289). Melbourne, Australia, ACM Press, New York.

Hawking, D. (2001). Overview of the TREC-9 web track. In E. M. Voorhees & D. K. Harman (Eds.), *Proceedings of the Ninth Text REtrieval Conference (TREC-9)* (pp. 87–105).

Hawking, D., & Robertson, S. (2003). On collection size and retrieval effectiveness. *Information Retrieval, 6*(1), 99–150.

Madigan, D., Vardi, Y., & Weissman, I. (2006). Extreme value theory applied to document retrieval from large collections. *Information Retrieval, 9*(3), 273–294.

Sanderson, M., & Joho, H. (2004). Forming test collections with no system pooling. In *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on research and development in information retrieval (SIGIR 2004)* (pp. 33–40).

Soboroff, I., & Robertson, S. (2003). Building a filtering test collection for TREC 2002. In *Proceedings of SIGIR 2003 the Twenty-sixth Annual Conference on research and development in information retrieval* (pp. 243–250).

Sparck Jones, K., & van Rijsbergen, C. (1975). Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge.

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management, 28*(4), 467–490.

Voorhees, E. M. (2006). Overview of the TREC 2005 robust retrieval track. In *Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005)*.

Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 316–323).

Voorhees, E. M., & Harman, D. K. (2000). Overview of the eighth Text REtrieval Conference (TREC-8). In E. M. Voorhees & D. K. Harman (Eds.), *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*(pp. 1–24). NIST Special Publication 500-246.

Voorhees, E. M., & Harman, D. K. (2002). Overview of TREC 2001. In *Proceedings of TREC 2001* (pp. 1–15). NIST Special Publication 500-250.

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on research and development in information retrieval* (pp. 307–314). Melbourne, Australia, ACM Press, New York.