Test Collection Diagnosis and Treatment

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

ABSTRACT

Test collections are a mainstay of information retrieval research. Since the 1990s, large reusable test collections have been developed in the context of community evaluations such as TREC, NTCIR, CLEF, and INEX. Recently, advances in pooling practice as well as crowdsourcing technologies have placed test collection building back into the hands of the small research group or company. In all of these cases, practitioners should be aware of, and concerned about the quality of test collections. This paper surveys work in test collection quality measures, references case studies to illustrate their use, and provides guidelines on assessing the quality of test collections in practice.

1. INTRODUCTION

Since Cleverdon's Cranfield experiments [8], test collections have become a mainstay of information retrieval research. Early test collections were fully judged, avoiding concerns about unjudged relevant documents, but nevertheless had serious issues originating from topic selection bias, assessor background and training, and domain limits. These problems were discovered through deep analysis of experiments with those collections, and those experiences informed the techniques used today to build large test collections.

Today, information retrieval researchers have any number of sources of test collections. The most commonly-used are those created as part of the major community evaluation forums: the Text REtrieval Conference (TREC), the Cross-Language Evaluation Forum (CLEF), the NII NACSIS Test Collection for IR systems (NTCIR), and the Initiative for the Evaluation of XML Retrieval (INEX). These collections have become the de facto standards for measuring retrieval systems.

Recently there have been several important areas of research and consequent advances in building test collections. First, Cormack et al. [10] and Soboroff et al. [20] spurred a line of research into efficient pooling practices (e.g. [7]) now being used in some evaluations (e.g., [3]). Second, several researchers have been probing the reliability of test collections and effectiveness measures (e.g., [24, 29, 6, 27] and many more). Third, evaluation forums have been moving away from collections of news articles and the adhoc search task and are attempting to build test collections for novel data and novel tasks. Many of these efforts are exploring the use of crowdsourcing and other inexpensive methods of collecting relevance judgements.

As information retrieval researchers propel the Cranfield paradigm headlong into uncharted territory, we are beginning to realize the importance of measuring and understanding our tools, both test collections and evaluation measures. When a test collection is built for a new search task, or a new medium, or using a new methodology, we need to have ways to measure the quality of that collection, so we can understand the reliability of measures computed from it.

This paper seeks to fill this need by drawing together both published research on test collection quality and tacit experience from the author in building test collections. The research is scattered in the literature and no single resource exists that gathers it into one place. Notes on test collection building practices are likewise scattered in prose and footnotes in various places, and badly need compiling, as the research community creates new tools that allow anyone to build a test collection for their particular need. This paper is not meant to be a comprehensive survey but to indicate major results and best practices.

This paper is written mostly within the context of TREC, since that is the family of test collections most familiar to the author. This should not detract from the generality of what is presented.

Within the scope of this paper, a test collection should ideally have the following properties.[1] It should be **reusable**; the test collection should provide reliable measures both for systems involved in its creation and those not so involved. It should be **diagnostic**; differences between systems should be revealed when they exist, and the collection should support analysis of these differences. It should be **unbiased** toward specific retrieval algorithms or strategies.

This survey begins with a discussion of pooling practice and the various parameters set during pooling, then discusses issues surrounding topic creation and balance within a test collection. The paper then presents several measures of test collection quality, along with small case studies to illustrate their use.

2. POOLING

In the classic Cranfield approach, every document's relevance must be assessed against every topic. Cleverdon maintained that the completeness of relevance judgments was more important than the number of topics, or search

[1]Other properties of test collections certainly exist and are desirable, such as task, media type, availability and realism; these properties are not considered in this paper.

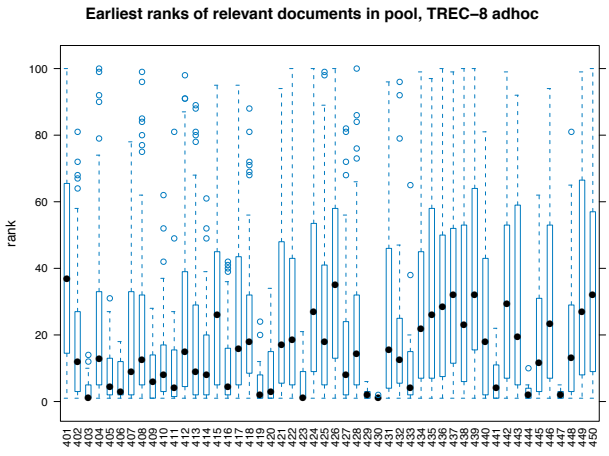

Figure 1: Distribution of the ranks at which relevant documents were added to the TREC-8 adhoc pools.

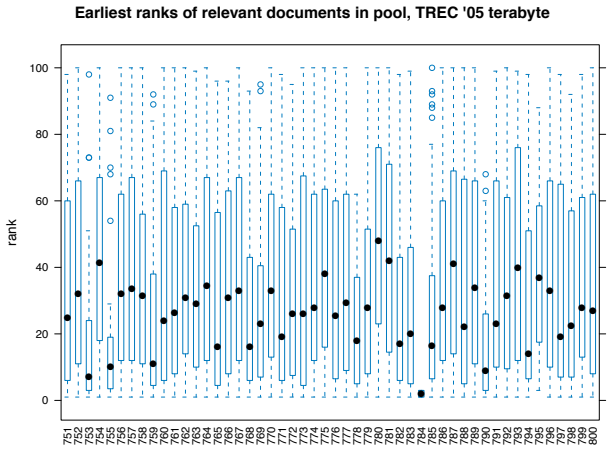

Figure 2: Distribution of the ranks at which relevant documents were added to the TREC 2005 terabyte pools.

needs, present in the collection [9]. Obtaining complete relevance judgments for any collection of more than ten or twenty thousand documents is infeasible and in the information retrieval community this has not been done in over ten years.[2]

Instead, nearly all test collections in current use were built using pooling [23, 22, 25, 28][3]. In pooling, rather than judge every document in the collection, only a small sample is judged. This sample is obtained from a set of rankings of the top documents for each topic. In TREC, these rankings are called *runs*. The runs may be created by manual searching but are more often outputs from automatic retrieval systems. In all cases the number of documents returned by a run is fixed at some maximum N, typically 1000 documents in a TREC adhoc search task. The top-ranked $n < N$ documents retrieved for a topic by each run are combined and duplicates removed to form the *pool*. The documents in the pool are judged, and all unpooled documents are assumed to not be relevant.

This is the *pooling assumption* — all unjudged documents are not relevant. Clearly it is not true, and historically this has been a criticism of pooling; the debate over completeness in modern test collections is summarized well by Baillie et al. [4]. Experience seems to indicate that if the pool is deep enough and the runs are good enough and diverse enough, then the pool will be unbiased and provide reasonably complete relevance judgments.

The quality of the runs is of course hard to judge without an existing test collection. New test collections developed for new tasks or new media must be used carefully until the behavior of systems in the new setting is understood. For example, the TREC-1 topics and relevance judgments should probably not be used in any experiment today — the document collection, the topic creation process, and the relevance assessment procedures were all new, even though Cranfield-style evaluations on the identical task had been conducted since the 1960s. Moreover, the document collection in TREC-1 was much larger than existing systems were designed for. As a result, TREC-1 participants were doing significant systems engineering in order to retrieve documents whose relevance they had only limited basis to estimate [28, chap. 2].

One hedge against run quality is run diversity. Gathering runs from different systems, different preprocessing components, and using different query formulations all adds to the diversity of the pool. If a variety of retrieval approaches is not represented in the pool, then it is more likely that the resulting test collection will not be able to measure new runs. A pool with few runs, or runs from only a few participating groups, is likely to be biased towards those systems. Runs from a single participating group are often parameter variations of the same system or otherwise very similar to each other; rather than pooling more runs from each group, pooling deeper from fewer runs is usually a more effective practice.

Manual searches can be extremely valuable additions to the pool. In NTCIR-1, manual interactive searches were performed and located 17% of the unique relevant documents in the collection [15]. We will have more to say on the role of unique relevant documents later in the paper.

Pooling to depth n allows exact computation of precision at rank n for the runs that were pooled, and related measures such as reciprocal rank (RR). However, we often want to compute measures that include a recall component, such as average precision (AP), and we want to be able to measure runs that were not pooled. For these reasons, n needs to be deep enough to give a good estimate of the number of relevant documents.

Figures 1 and 2 show the effect of the choice of pool depth given the number of relevant documents. Each box-and-whisker illustrates for a single topic the distribution of the ranks at which the relevant documents for that topic were added to the pool. (A relevant document may be found by more than one run; this paper considers the shallowest ranked occurrence, the rank at which it was first pooled.) Figure 1 shows the topics for the TREC-8 adhoc collection, and Figure 2 shows the topics for the TREC 2005 terabyte collection. In both cases, the pools went to depth 100, but we can see that there are relevant documents at much deeper

[2]To the best of this author's knowledge, the last test collection for a retrieval task built with complete relevance judgments was TDT-3, with 60 topics and 43,000 documents [13]. The TDT-3 corpus was built in late 1998.

[3]Sparck Jones and Van Rijsbergen's initial proposal is in [23]. A detailed analysis of their design appears in [22].

ranks in the terabyte collection, and as such it is more likely that more relevant documents exist below the pool line, where they were not judged.

Usually, TREC collections are pooled to depth $n = 100$, but in practice, choosing a pool depth is a trade-off. Deep pools improve reusability, diagnostic power, and decrease bias. However, limited evaluation resources will allow for some number of documents to be judged, and pool sizes are tweaked either by tuning the number of runs pooled or the depth of the pool. If large numbers of relevant documents are expected, one should choose a deeper n, but when a task is novel, one can choose to pool more runs, perhaps to a shallower depth, to enable measuring all participating runs at least to a minimal degree. Measuring the document overlap among runs at different pool depths can help support this decision.

The TREC 2002 filtering collection is an example of a novel pooling scenario that was confronted with all these challenges [21]. In the TREC adaptive filtering task, the system is provided a topic statement and a handful of known relevant documents. The system is then shown each document in the collection in chronological order, and must make a binary decision at each document whether to show it to the user or not. If the system decides to show the document, it may access the relevance judgment for it, if it exists. Thus, systems must simultaneously learn the topic model as well as the retrieval threshold.

Building a test collection for the adaptive filtering task is tricky because systems need the relevance judgments while they run, so pooling is not an option. In most years, older collections with existing relevance judgments were used, but in 2002 it was decided to create a new test collection from scratch on top of the Reuters Corpus (RCV1, [16]). The solution was to use relevance feedback from the human assessors with seven different retrieval and classification systems, over the course of a week, to gather relevance judgments during topic development. This procedure was somewhat similar to "iterative searching and judging" as proposed by Cormack et al. [10]. The final runs were also sampled for further judgment (pooling wasn't an option since the runs don't provide a ranking of documents) and relatively few new relevant documents were found (see Figure 3, from [21]), mostly in topics which already contained many relevant documents. Unfortunately given the dynamic nature of the systems it's difficult to say retrospectively how those systems would have fared given the additional relevance judgments. Nevertheless, by using a variety of search systems combined with relevance feedback to judge thousands of documents, the authors were able to simulate the richness of a pooling scenario and build a reusable test collection.

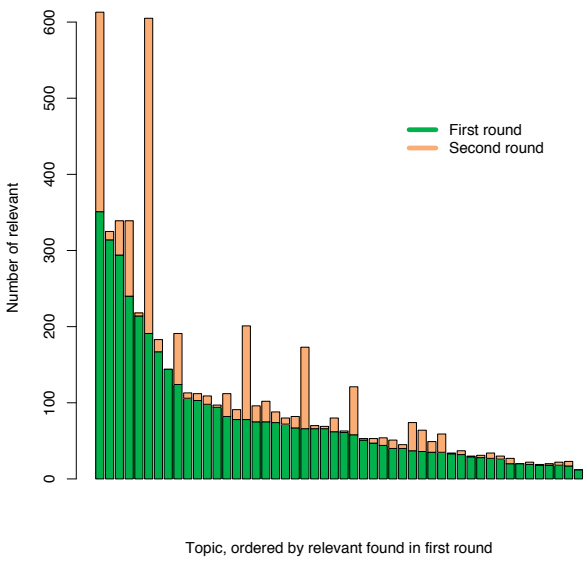

Figure 3: Number of relevant documents found in feedback development (green) and new relevant documents found in sampled final runs (yellow)[21].

3. TOPIC DEVELOPMENT

Few recent works on test collections, with the notable exception of Harman [28, chap. 2], discuss the role of topic development in building a test collection. Notionally, the topics should represent a sample of the universe of information needs expected in the context of the user task that the collection models. Clearly, most test collections contain only a minuscule sample of this very large space. Nevertheless, the topic set dictates what the collection can measure, and if the topic set is biased by design, then this can invalidate or at the very least circumscribe results from that collection. As Harman notes, in several of the early test collections, the topics were developed in ways which, in retrospect, predicted certain evaluation outcomes, but these aspects of the collections were not well-known in the research community. In some sense, detailed information about the topic set is the "missing manual" of many test collections. The recommendations below are gleaned from experience and to the best of the authors knowledge have not been compiled in any form.

In TREC, the topic development process has necessarily evolved as tasks and data changed, but certain core desiderata remain. The central principle is balance. Since most effectiveness measures in the Cranfield paradigm are averaged across the topic set, unbalanced topic sets will be reflected in the measure. Balance of difficulty was explored in the TREC robust track; topics that are "easy" tend to dominate the average, and techniques that improve a topic where effectiveness is very low aren't detected. The GMAP (geometric mean average precision) measure attempts to solve this by averaging differently.

In practice, balancing for difficulty is hard to do because difficulty cannot be known precisely without retrieval results in hand. TREC applies heuristics during topic development to try and control for topics that are too "hard" or too "easy". One heuristic is to control the number of expected relevant documents. The topic creator searches the collection during topic development and keeps track of how many unique relevant documents are found, and if this exceeds a threshold the topic is not used. TREC also tries to avoid topics which may either be trivial or intractable given the current state of the art in retrieval systems. Trivial topics inflate scores and thus the average, while intractable topics which require human understanding and intuition can be a diversion during failure analysis.

Topics should be balanced in subject matter. If there are too many topics within a similar subject, then the collection may be biased towards some subset of the collection.

This can be hard to achieve for example if a collection of news articles is dominated by a major story. Typical TREC adhoc collections have very little overlap in relevant documents across topics. Oard et al. [18], in investigating the possibilities of collaborative filtering as a TREC routing algorithm, found that in the TREC-8 filtering track the topics had only minimal overlap in relevant documents. While in that case the lack of overlap among topics was a detriment to the success of their approach, in general, avoiding topics that are too close to each other is good practice.

A related issue of balance is across rough classes of topics. The definition of class in this context is highly task-specific. As an example, the TREC 2002-4 web tracks had a task called "topic distillation", where the goal was to identify the key sites for a topic rather than all the relevant web pages. Early in the topic development process, several topics about various medical conditions, surgeries, etc. were independently composed, whereupon it was found that all medical subjects had the same handful of key sites: Medline, an NIH laboratory, and so forth. The topic sets were thus limited to one topic of this class at most.

The TREC topics are usually developed by a group of six to eight contractors (called "assessors"), and in most cases the number of topics is balanced across the assessors. This is primarily done to avoid an unbalanced workload during relevance assessment, but it also has the effect of balancing the topics among topic authors. If instead the topics came primarily from a single author, the topic sets would be skewed towards the interests of that assessor. If that assessor is inclined to design topics in certain subjects of interest or topics that have a characteristic difficulty, this can compound the problem.

4. THE UNIQUES TEST

An important group of diagnostic criteria for test collections are the number of unique relevant documents, how unique documents are spread across the pooled runs, and the effect of ignoring those documents. These criteria form what is sometimes called the "uniques test". A "unique relevant document" is a document which only one group retrieves; if that group's runs had not been pooled, that document would be regarded as not relevant.

The problem of unique relevant documents is central to the pooling approach. The goal of pooling is to create a reusable test collection while minimizing the number of irrelevant documents judged. If unjudged documents are to be considered as not relevant, then the pools need to reasonably span the range of retrieval outputs that are measurable by that collection. Systems that retrieve relevant documents which no pooled system finds are theoretically unmeasurable. In practice the impact depends on how many unique relevant documents are found.

Harman [14] describes a search for unpooled relevant documents in the TREC-2 and 3 pools. The original pools contained the top 100 documents; in the re-investigation, the second 100 documents were pooled and judged by the same assessors. On average, one new relevant document per run was found, with a high variation across topics. Voorhees [25] mentioned this study and showed the statistics of unique relevant documents retrieved in the TREC-5, 6, 7, and 8 adhoc tasks.

Zobel [29] examined the impact of unique relevant documents directly, by taking each run in turn and removing its unique relevant documents from the relevance judgments, and measuring the percentage loss in 11-point average precision (11AP). He found that in TREC-3, the average run would have lost 2.2% 11AP, and 0.5% 11AP in TREC-5. The effect was much larger for the topics with the most known relevant documents. Zobel's conclusion was that the test collections he examined were reliable under this test.

In practice, because runs from the same participating group often have highly-overlapping sets of retrieved documents, one should conduct this test by holding all runs from a group out at a time. Voorhees [25] reports that conducting the test in this manner for TREC-8 adhoc collection yields an average loss in mean average precision (MAP) of 0.78%, and a maximum loss of 9.9%. The maximum here is from a manual run; the maximum loss for an automatic run was reported to be 3.85% MAP. This shows that manual runs contribute disproportionate numbers of unique relevant documents to the pool, thus greatly enriching the test collection for future users.

An additional consideration in performing the uniques test is that the effect on very poorly-performing runs should be ignored, because any change in a very low MAP score will be a large percentage change. For example, a reduction from 0.002 to 0.001 is 100%, but not very meaningful as a predictor of the effect of that run's unique documents on the measurement of future systems.

In 2001 and 2002, the TREC cross-language track created the first very large Arabic test collections, with queries translated into English and (in 2001) French for cross-language retrieval experiments [12, 17]. Because no such collection was widely available prior to the track, system performance was suspected to be lower than might be achieved with training resources. The coordinators found that 9 out of 28 participating runs experienced a reduction in MAP of greater than 10% under the uniques test, much higher than Zobel had observed. For 7 out of the total 25 topics, more than half the relevant documents were unique and additionally all found by one group, and for another 6 topics, 40-50% of the relevant documents were uniques. The coordinators supposed that this could be due to four possible factors: using a pool depth of 70 rather than 100; unusually large topics with many relevant documents; a small number of participating groups (10); and a diversity of techniques represented among those groups' runs. They hypothesized that the last two factors were the most important, because test collections had been built in other cross-language settings with somewhat shallow pools and large topics, without a high uniques test effect [12].

In 2002 collection, fifty topics were developed and 41 runs were submitted by nine groups. The coordinators found that no group found more than 6% of the unique documents in the collection, and that only a single run had a greater than 5% reduction in MAP under the uniques test. These statistics are much closer to those of a standard TREC collection as described by Voorhees [25] and Zobel [29]. The coordinators did not determine what precisely affected this outcome, but systems were certainly more mature in 2002 than they were in 2001. Additionally, the track made a number of common resources available, including an Arabic light stemmer, translation dictionaries, and a web-accessible machine translation system. Groups were encouraged to submit runs using the common resources. These resources may have encouraged some convergence in results among the participating

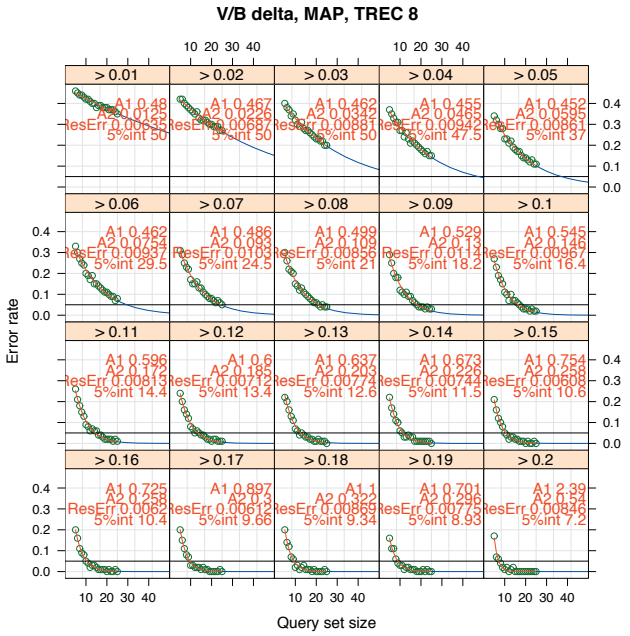

Figure 4: Data from the Voorhees and Buckley minimum-delta test for the MAP measure in the TREC-8 adhoc collection. Computed topic set sizes are circles, the line fitting them shows the extrapolation, and the small text in each subplot shows the parameters, residual error, and confidence interval of the fit.

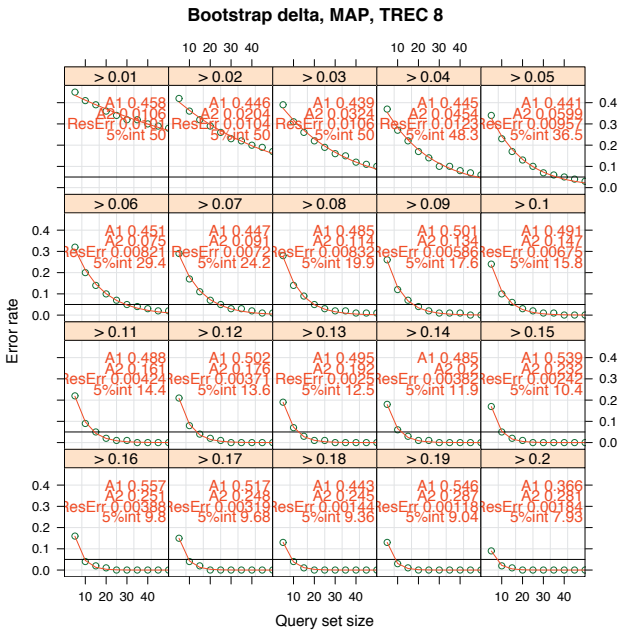

Figure 5: Data from the bootstrap version of the minimum-delta test. It is equivalent to the plots in Figure 4, but with no need for extrapolation. The fit parameters and statistics are provided for comparison.

groups.

5. THE MINIMUM-DELTA TEST

Voorhees and Buckley [27] investigated the minimum number of topics that would lead to a stable measure of effectiveness. The definition of stability in this context is the probability of two systems swapping their places in an ordering of the systems by the measure, if a different set of topics of the same size were used. The method they proposed also computes a minimum difference in the measure, greater than which one should not expect systems to swap with high probability.

Their method is as follows. Given a test collection with N topics and the top 75% of systems pooled to create it, draw at random two disjoint subsets of the topics, of equal size up to $N/2$ topics each. Evaluate the systems using each set of topics separately, and ranking the runs according to their effectiveness on each topic subset, count the number of pairwise swaps between the two rankings. The swap probability is expressed as the observed frequency of swaps out of all possible pairwise swaps. The topic subsets are drawn randomly multiple times, and the swap probability averaged across trials. Swap probabilities are extrapolated from a fitted exponential model to the full set of N topics.

Voorhees and Buckley did not name their test in the paper, and it is variously called the "swaps test" or "the Voorhees-Buckley swaps test" in other work. We propose the name "minimum-delta test" as one that is more clearly descriptive and indicates its diagnostic use.

Figure 4 illustrates the output of the procedure for the TREC-8 adhoc collection. Each curve plots the probability of a swap between two systems against the number of topics. As the number of topics grows in each case, the swap probability goes down; more topics make the measure (in this case, MAP) more stable. The swaps are binned by the difference in MAP between the systems that swapped, and each subgraph shows the swaps in that bin. As a desired "minimum effective difference" between systems gets smaller, more topics are needed to get the swap rate to an acceptable level. So, the subgraph labeled "> 0.04" shows that at 50 topics (the right edge of the x-axis), the swap rate is predicted to be 5% for differences in MAP of 0.04 or greater. In other words, differences of less than 0.04 in MAP (not percentage, but in score) have a 5% chance of occurring the other way around in an "equivalent" set of 50 topics on this collection, and should be regarded as not meaningful.

Sakai [19] showed that this method can be computed using the bootstrap [11]. Bootstrapping is a statistical resampling procedure where (in this case) topics are chosen with replacement from the full topic set. This allows direct sampling up to the full topic set size, eliminating the need for extrapolation, and also offers a more strongly statistical perspective when regarding the minimum-delta test. Figure 5 shows the bootstrapped values in the TREC-8 adhoc collection; they are equal to the outputs of the classic Voorhees and Buckley test.

The minimum-delta test is a useful diagnostic for test collections because it indicates the resolution power of the collection given the systems pooled to create it. A high minimum difference to obtain a 5% swap rate on the full topic set can indicate a difficult task, or very highly variable performance across topics among the pooled systems.

Voorhees and Buckley do not provide guidance on using the minimum-delta test on runs that were not pooled. A straightforward approach might be to add the new runs to the top 75% of pooled runs (provided that the new runs fall within or above the range of effectiveness of that set), re-run the minimum-delta test, and examine the swap rates of the new systems. This is tricky because a small handful of systems of interest would have comparatively few swaps overall, so it's not clear how to use this approach as a "significance test" per se.

6. THE TITLESTAT TEST

In the TREC 2004 robust and HARD tracks, systems submitted runs over a set of topics developed for the TREC newswire collections from TREC CDs 4 and 5 (without the Congressional Record subcollection). In 2005, the same topics were used but with a new collection of newswire, the AQUAINT collection [26]. Chris Buckley created a run for the 2005 track that used highly-optimized relevance feedback queries trained from all the 2004 collection relevance judgments. This run performed around the median of systems, but yielded hundreds of unique relevant documents, leading to a 23% reduction in MAP under the uniques test.

This uniques test result would lead to questions about the reusability of the AQUAINT robust track collection, but Buckley and others felt that something more was happening. By largely disregarding the topic text, and training feedback queries based on massive data from a different collection, the SABIR run was effectively searching the collection in a very different manner than any other run. They devised a measure called "titlestat" which illustrated that the difference in the SABIR run could be explained by a lack of retrieved documents containing words from the "title" field of the search topics [5].

Titlestat is computed as follows. Given a set of documents and a set of topics, compute the fraction of the set of documents that contain a word in the "title" field of the topics. This fraction is averaged across terms in a topic's title field, then across topics, to compute the occurrence of the "average" title field word. Titlestat can be computed for any set of documents, such as the set of relevant documents in a collection, or the documents retrieved by a run, or all documents retrieved by all runs at or above some given rank.

Buckley et al. found that the titlestat of the relevant documents in the 2004 collection (using TREC CDs 4 and 5) was 0.588, whereas the titlestat of the relevant documents in the 2005 collection (using AQUAINT) was 0.719. The titlestat of the unique relevant documents retrieved by the SABIR run is 0.53, much lower than the greater set of all relevant documents in the AQUAINT robust collection.

The authors of that study found a convergence of factors explained the data. First of all, the AQUAINT collection is twice as large as the older collection from TREC CDs 4 and 5. Secondly, the 2005 collection (using AQUAINT) was only pooled to a depth of 55. Thirdly, when the topics were originally developed on the older collection, there was a reasonably small number of relevant documents in the collection, but no such guarantee existed in the AQUAINT collection. The combination of the shallow pool, the larger collection, and no development-time control for topic size meant that very large topics with relevant documents trivially matching the title query field could crowd out the pool, disadvantaging future runs with more adventurous search approaches.

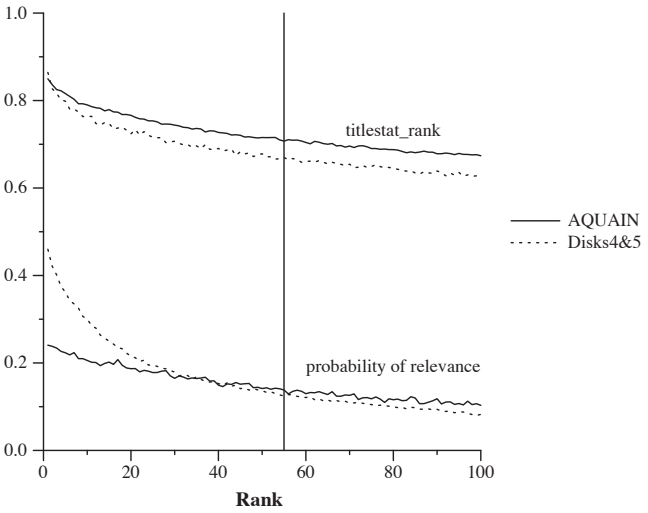

Figure 6: Titlestat of pooled documents by rank cutoff, and also probability of relevance of documents at that rank, for the 2004 and 2005 robust collections (from [5]).

Figure 6 (from [5]) illustrates this phenomenon. As documents enter the pool rank-by-rank, the titlestat of those documents decreases, at about the same rate in the two collections. However, the probability of those documents being relevant decreases much faster in the older collection from CDs 4 and 5 than it does in the AQUAINT collection. As collections grow in size, this phenomenon becomes more pronounced — in the TREC terabyte collections, the curve for probability of relevance is almost flat, indicating very large topics that are highly titlestat-biased.

At present there is no good framework for understanding the impact of the difference between two titlestat values, but they are valuable exploratory measures that indicate how "future-proof" a collection might be. Collections with high titlestat and topics with many relevant documents may be biased against systems that search off the beaten path, as it were. Suspicious uniques test values can be combined with a titlestat analysis to demonstrate such a bias more convincingly, but in general, compelling runs such as the SABIR run described here are hard to create and don't happen very often.

7. CONCLUSIONS

Creating test collections is more of an art than a science at present. Over many years and through the creation of many test collections, a small body of techniques and analysis tools have been developed to help diagnose when a test collection may have problems. These techniques are compiled in various articles and in the tacit knowledge of test collection builders. The goal of this paper is to bring together this information and create a guide to the prominent results in this area. We plan to make portable code available for computing the tests described here, using an open-source code repository, further aiding practitioners in the field.

This paper is far from the final word on test collection diagnosis. If one thing is clear, it is how little we still know. Future-proofing test collections is a hard problem, made all the harder by new large-scale collections that dis-

courage computationally intensive but possibly revolutionary results. New collections for novel tasks and media domains are also at risk of poor reliability due to immature participating systems. Crowdsourcing and other methods for compiling inexpensive relevance judgments force us to consider the quality of that data. We need to develop a strong suite of tools, of which this paper describes a small part, to measure the quality of test collections and improve the reliability of modern, Cranfield-style experiments.

Much of what is known about building test collections comes from evaluation forums where the system outputs from multiple research groups and many systems are combined using pooling or some close equivalent. However, small research groups as well as companies need purpose-built test collections to measure search quality, and in those contexts the diversity and richness of an evaluation forum is impossible to achieve. We need to study this area and understand how our practices can be translated to those communities.

8. ACKNOWLEDGEMENTS

I gratefully acknowledge the kind invitation to speak at the 2010 Indian Forum for Information Retrieval Evaluation (FIRE), an opportunity which allowed the compilation of much of the information in this paper. I am also grateful to my colleagues Donna Harman and Ellen Voorhees, to Chris Buckley, and to many others who have helped me to build test collections and to try to understand them.

9. REFERENCES

[1] *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, Melbourne, Australia, August 1998. ACM Press.

[2] *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, Seattle, WA, August 2006.

[3] James Allan, Ben Carterette, Javed A. Aslam, Virgil Pavlu, Blagovest Dachev, and Evangelos Kanoulas. Million query track 2007 overview. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, Gaithersburg, MD, November 2007.

[4] Mark Baillie, Leif Azzopardi, and Ian Ruthven. Evaluating epistemic uncertainty under incomplete assessments. *Information Processing and Management*, 44(2):811–837, 2008.

[5] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10:491–508, 2007.

[6] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pages 33–40, Athena, Greece, July 2000. ACM Press.

[7] Ben Carterette, James Allan, and Ramesh Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)* [2], pages 268–275.

[8] Cyril W. Cleverdon. The Cranfield tests on index langauge devices. In *Aslib Proceedings*, volume 19, pages 173–192, 1967. Reprinted in K. Sparck Jones and P. Willet, eds. *Readings in Information Retrieval*, Morgan Kaufmann, 1997.

[9] Cyril W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Chicago, October 1991.

[10] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)* [1], pages 282–289.

[11] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, NY, 1993.

[12] Fredric C. Gey and Douglas W. Oard. The TREC-2001 cross-language information retrieval track: Searching Arabic using English, French or Arabic queries. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, NIST Special Publication 500-250, Gaithersburg, MD, November 2001.

[13] David Graff, Chris Cieri, Stephanie Strassel, and Nii Martey. The TDT-3 text and speech corpus. In *Proceedings of DARPA Broadcast News Workshop*, pages 57–60. Morgan Kaufmann, 1999.

[14] Donna Harman. Overview of the fourth Text REtrieval Conference (TREC-4). In Donna K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, NIST Special Publication 500-236, pages 1–23, Gaithersburg, MD, November 1995.

[15] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Soichiro Hidaka. Overview of IR tasks. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, Tokyo, Japan, August 1999.

[16] David D. Lewis, Yiming Yang, Tony Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[17] Douglas W. Oard and Fredric C. Gey. The TREC 2002 Arabic/English CLIR track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, NIST Special Publication 500-251, Gaithersburg, MD, November 2002.

[18] Douglas W. Oard, Jianqiang Wang, Dekang Lin, and Ian Soboroff. TREC-8 experiments at Maryland: CLIR, QA, and routing. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, NIST Special Publication 500-246, Gaithersburg, MD, November 1999. National Institute of Standards and Technology.

[19] Tetsuya Sakai. Evaluating evaluation metrics based on

the bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)* [2], pages 525–532.

[20] Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 66–73, New Orleans, LA, September 2001. ACM Press.

[21] Ian Soboroff and Stephen Robertson. Building a filtering test collection for TREC 2002. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pages 243–250, Toronto, Canada, July 2003. ACM Press.

[22] K. Sparck Jones and R. G. Bates. Report on a design study for the "ideal" information retrieval test collection. British Library Research and Development Report 5428, Computer Library, University of Cambridge, 1977.

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

[24] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)* [1].

[25] Ellen M. Voorhees. The philosophy of information retrieval evaluation. In *Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum (CLEF 2001)*, Darmstadt, Germany, 2002.

[26] Ellen M. Voorhees. Overview of the TREC 2005 robust track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, November 2005.

[27] Ellen M. Voorhees and Chris Buckley. 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 (SIGIR 2002)*, pages 316–323, Tampere, Finland, August 2002. ACM Press.

[28] Ellen M. Voorhees and Donna K. Harman, editors. *TREC: Experiments in Information Retrieval Evaluation*. MIT Press, 2005.

[29] Justin Zobel. How reliable are the results of large-scale retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)* [1], pages 307–314.