A Comparison of Pooled and Sampled Relevance Judgments in the TREC 2006 Terabyte Track

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, Maryland, USA

Abstract

Pooling is the most common technique used to build modern test collections. Evidence is mounting that pooling may not yield reusable test collections for very large document sets. This paper describes the approach taken in the TREC 2006 Terabyte Track: an initial shallow pool was judged to gather relevance information, which was then used to draw a random sample of further documents to judge. The sample judgments rank systems somewhat differently than the pool. Some analysis and plans for further research are discussed.

Keywords: *random sampling, pooling, bias.*

1 Introduction

Test collections are tools for comparing the effectiveness of two or more retrieval systems. A test collection consists of a set of documents, a set of search needs called *topics* and a mapping from topics to the documents that are relevant to them called *relevance judgments* or *qrels*. A retrieval system transforms each topic into a query which is then used to search the collection. The system returns a ranked list of documents predicted to be relevant to each topic. This set of ranked lists is called a *run*. Using the relevance judgments, we can measure the quality of the runs, and quantitatively compare the outputs of multiple retrieval systems. This experimental approach to measuring search effectiveness is called the Cranfield paradigm, after a pioneering study by Cleverdon at the Cranfield Aeronautical College [10].

Historically, test collections were completely judged, meaning that every document was assessed for relevance to every search topic. While allowing for exact measurement of search effectiveness, complete judgments limits the scalability of test collections to perhaps on the order of ten thousand documents. In 1975, Sparck Jones and Van Rijsbergen proposed *pooling* as a method of scaling retrieval collections [12]. Rather than judging the entire collection, we search for relevant documents using a number of retrieval systems, combine the top-ranked documents from each of them, and judge only those documents. The top λ documents from each system for a single topic are called a *pool*. If the pools are large and diverse enough, the judged documents that come from them represent an unbiased sample of the relevant and nonrelevant documents that exist for those topics. If the sample is unbiased, we can use it to compare two or more systems fairly.

Pooling has been used for many years to build test collections within the Text REtrieval Conference (TREC), the NTCIR conferences, and other venues [15]. These collections range in size from about half a million documents up to tens of millions. Even a collection of 500,000 documents can never have complete judgments, since an assessor would need to work non-stop for months on end in order to read every document in the collection for even a single topic. Thus, pooling has become the standard method for building large test collections.

Test collections are most useful when they are *reusable*, that is, when they can be reliably used to rank systems that did not contribute to the pools. The systems that contribute to the pools can be measured exactly up to the pooling depth; when $\lambda = 100$, for example, precision at ranks up to 100 is exactly known, and other measures including mean average precision (MAP) can be estimated very accurately. However, systems that did not contribute to the pools may use different retrieval algorithms and techniques, and consequently may retrieve unjudged documents. These unjudged documents are normally assumed to be nonrelevant. If the pooled relevance judgments are indeed sufficiently complete and unbiased, this is a safe assumption, but if not, the effectiveness of systems that did not contribute to the pool will be underestimated.

Justin Zobel investigated the reliability of the pooling assumption and the reusability of the earlier TREC collections by removing runs from the pool. For each pool run, he removed the relevant documents retrieved by only that run, and measured that run using the remaining relevance judgments. This procedure simu-

lates what would happen if that run had not contributed to the pools. He found that MAP scores changed only slightly, and based on this as well as estimates of full recall that the TREC collections of that time were sufficiently complete to be reusable [17].

Zobel's result is important because the success of the Cranfield paradigm is entirely due to the reusability of test collections. The ranking of systems based on a single test collection only holds for that particular collection; results must be reproduced in multiple test collections before drawing conclusions. If a collection is only fair to the systems that participated in its creation, then every experiment must create its own test collection, making it very difficult if not impossible to achieve reliable and comparable experimental results.

2 The scalability of pooling

Does pooling scale as collections grow beyond the gigabyte range? A workshop held at SIGIR in 2003 hypothesized that very large collections would have very many relevant documents, and that pooling would discover only a very small fraction of them. The resulting relevance judgments would then be biased towards participating retrieval systems and might unfairly rank new or alternative approaches [11]. The goal of the TREC terabyte track, which grew out of that workshop, was to build a very large test collection and attempt to determine the effectiveness of pooling in that collection [8].

Pooled relevance judgments for very large collections may not be reusable for two reasons. The first is that the judgments will be very sparse and thus not sufficiently complete to accurately measure new runs. A solution to this problem is to design measures to be robust when judgments are not complete. One such measure is *bpref* [7]. Whereas MAP assumes that unjudged documents are not relevant, bpref only considers the ranks of retrieved documents that have been judged either relevant or nonrelevant. bpref is defined as:

$$bpref = \frac{1}{|R|} \sum_{r} (1 - \frac{|n \ ranked \ higher \ than \ r|}{\min(|R|, |N|)})$$

where R is the total number of judged relevant documents, N is the number of judged nonrelevant documents, r is a relevant retrieved document, and n is a member of the first R nonrelevant retrieved documents.[1] Bpref can be thought of as the inverse of the fraction of judged nonrelevant documents that are retrieved before relevant ones. In experiments where older TREC collections were subsampled to simulate sparse relevance judgments, bpref strongly correlates with the full-collection MAP score. In the TREC terabyte track, bpref has been reported alongside MAP as an official effectiveness measure.

The second concern with pooling in very large collections is that the judgments will be biased. "Biased" here does not mean statistical bias but rather that a test collection will unfairly rank some class of systems. This can happen because only a certain kind of retrieval algorithm contributes to the pool, or because some kinds of documents are pooled and not others, for example. Zobel's leave-one-out experiment examines bias against a run which was not pooled (but which might have been pooled at that point in time). Buckley et al. considered bias towards a particular kind of query formulation strategy, namely queries that focus strongly on the title words of topics [6]. They proposed a measure of this bias, *titlestat*, and found that larger collections seem to be more likely to exhibit this bias than smaller ones.

Given a set of topics and a set of documents, titlestat represents the occurrence of an average topic title word in that set of documents. Formally, 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. Individual per-topic titlestat values are then averaged over the set of topics. A titlestat of 1.0 indicates that every document in the set contains a title word. Special types of titlestats discussed by Buckley et al. include *titlestat_rel*, the titlestat of the relevant documents in a collection, and *titlestat_rank*, the titlestat of relevant documents retrieved at a given rank by a set of runs.

Buckley et al. used the titlestat measure to analyze the results of the 2004 HARD and Robust tracks in TREC. These two tracks used a set of old topics originally developed using TREC disks 4 and 5, and asked participants to search for those topics in the newer AQUAINT collection, which has about twice as many documents. The participating runs found many more relevant documents than originally existed for those topics, and generally performed much better than systems have done on those topics in the older collection. Moreover, one particular run found a very high number of unique relevant documents by training a highly optimized routing query for each topic using the old relevance judgments. The overall titlestat_rel value for the topics in the new collection was much higher than for the same topics in the older collection, indicating that most relevant documents could be located with a simple query based on the topic title. The run with many unique relevant documents had a title-

[1] This formula differs from that given in [7] and solves a bug when there are very few judged nonrelevant documents. The formula given here matches the current implementation in `trec_eval`. See the file `bpref_bug` in the `trec_eval` documentation.

stat_rel value closer to that of the old collection. This run would have been measured unfairly in the new collection if it had not contributed to the pools because the high-titlestat documents found by the other runs did not reflect the full landscape of relevant information in the collection.

In the TREC 2006 terabyte track, we selected documents to judge in a novel hybrid way, in order to measure the effects of sparseness and bias. This approach included a pooled component along with a deeper random sample. The following sections introduce the terabyte track, focusing on issues of reusability and fairness, describe the 2006 sampling approach in detail, quantify how sampling and pooling differ in their sets of relevant documents and in how they rank the systems, look explicitly at reusability, and close with a number of questions that remain to be answered.

3 The TREC terabyte track

The terabyte track began as part of TREC 2004 and has run for three years. It uses the GOV2 document collection, a fairly complete crawl of websites in the `.gov` domain, which includes sites from many U.S. federal, state, and local govenrnment agencies. Crawled in early 2004, GOV2 contains over 25 million documents and 426GB of text.[2]

The main task for the terabyte track is a traditional adhoc search task. Adhoc search was chosen as a task rather than something more web-centric because it is well-understood and because earlier collections for this task are considered reliable and reusable. Indeed, all the studies of modern collection reliability which had been done at that time had examined adhoc test collections. Other tasks in the track include named-page finding and tests to benchmark the efficiency of systems.

The terabyte track has created a total of 149 adhoc search topics over the course of TRECs 2004–2006. These topics are numbered 701–850; topic 703 has no relevant documents and should not be used.[3] Topics 701-800, from 2004 and 2005, have relevance judgments collected using the standard pooling approach described above. In 2004, pools were formed from the top 85 documents from two runs per group. In 2005, the pools went down to rank 100 for two runs per group. Additional details on the 2004 and 2005 terabyte tracks can be found in the respective track overviews [8, 9].

Initial investigation seemed to indicate that the 2005 terabyte collection is probably reusable, and the 2004 collection less so. We applied Zobel's leave-one-out procedure with the modification that we hold out the unique relevant documents retrieved by all of the runs from a single group. The reason for this is that runs from a single group tend to be very similar, and if just a single run is held back, the other runs from that group tend to contribute enough other documents to cover the held-out run even without that run's unique relevant documents [14].

Year	Measure	maximum abs. diff.	maximum rank change
2004	MAP	0.03	$-10/+3$
	bpref	0.08	$-1/+16$
2005	MAP	0.04	$-4/+0$
	bpref	0.02	$-1/+8$

Table 1. Maximum absolute differences in score and movement in ranking when a group is held out of the terabyte track pools.

Collection	Judged	Relevant	Unique rel per group
TREC-8 (v4&5)	82102	4728	29.1
TREC-9 (wt10g)	70070	2617	38.4
Terabyte 2004	58077	10617	309.4
Terabyte 2005	45291	10407	201.6

Table 2. Counts of judged, relevant, and the average number of unique relevant documents per group in four TREC collections.

Table 1 shows the results of the leave-a-group-out (GLOO) experiment for the MAP and bpref measures. Along with the observed maximum absolute difference in score, the table also shows the maximum number of places in the official ranking that a run moves when ranked using its leave-one-out score rather than its official score. The absolute differences are small, except for bpref in 2004. The large movements in the ranking for some runs deserve closer attention. For example, the run that moved downwards by 10 ranks by MAP score in 2004 was also the run with the maximum MAP score difference, decreasing from 0.2311 to 0.2037.

We also observed that the terabyte collections have many more relevant documents and many more unique relevant documents per group than older collections. Table 2 compares the 2004 and 2005 terabyte collections to the TREC-8 adhoc collection and the TREC-9 web collection, smaller collections that nevertheless have more judged documents. We can see that in the terabyte collections, nearly 20-25% of the pool is relevant, and the groups find many more unique relevant documents. Large numbers of relevant documents can indicate a less reusable collection, because it implies

[2] `http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm`

[3] See `http://trec.nist.gov/data/terabyte.html` to obtain these topics and their relevance judgments.

that even more relevant documents may exist that are not currently judged.

Additionally, we looked for title-word bias which might indicate that the collection would favor simple retrieval strategies. The *titlestat_rel* for the 2004 collection is 0.889 and for the 2005 collection is 0.898. As absolute figures, these indicate that nearly every document judged relevant contains the topic title words, and consequently we are concerned that there exist unjudged but relevant documents that do not contain the title words. Because TREC runs quite reasonably use the title words as strong indicators of relevance, the top ranked documents tend to contain title words. However, since these words are much more frequent in the terabyte collections than in smaller collections, documents with title words fill the pool. Put another way, our pool depth is too shallow with respect to the collection size to capture the full range of relevant documents that might exist.

Buckley et al. had a "smoking gun" which indicated title-word bias in the smaller AQUAINT collection, in the form of a run with many unique relevant documents and a much lower *titlestat_rel* than any other run contributing to that pool [6]. It was clear that the collection would have been biased against that run had it not been pooled. However, none of the terabyte track runs has this property. Thus while we had strong circumstantial evidence of bias in the terabyte collections, there was no proof.

4 Relevance-based sampling

For the 2006 terabyte track collection, we tried a different approach to see if we could reduce bias and increase reusability. The relevance assessment process was divided into two phases. In the first phase, a depth-50 pool was created from up to three runs per group (one manual and one automatic run, and one run from the track's efficiency task). Judgments from this pool would certainly be adequate to compute MAP to a reasonable precision for the participating runs. In the second phase, a random sample starting from rank 1 and including 200 not-previously-judged documents was drawn from the pool runs. This sample reached to a depth that varied per topic depending on the number of relevant documents in the depth-50 pool. Ideally, this sample should accurately estimate MAP for the pool runs, and provide a more reusable collection for future runs.

The sampling strategy, which we call *relevance-based sampling*, has two critical parameters: the depth or maximum rank to sample to, and the sampling rate. These parameters are estimated with the goal of finding an additional 20 relevant documents in the 200 new documents we plan to judge (*rpt*, for "relevant-percent-target"):

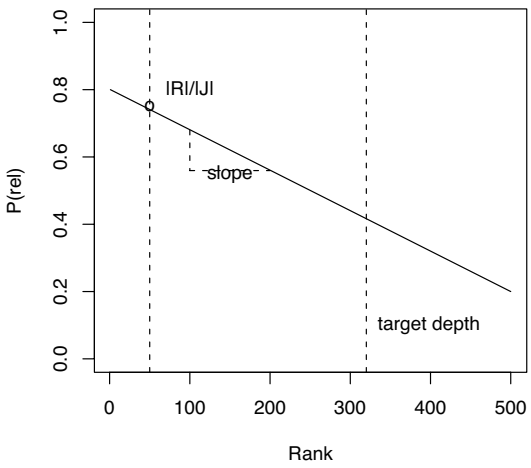

Figure 1. Schematic view for computing the target sampling depth.

$$rpt = 20/200$$

The depth for a single topic is computed as follows. From the relevance judgments from the depth-50 pool, we compute the probability of relevance in the pool at depth 50:

$$P(rel) = |R|/|J|$$

where $|R|$ is the number of relevant documents and $|J|$ the number of judged documents.

As we go deeper into the pool, this probability of relevance drops off exponentially, but we approximate this trend with a linear fit. This fit is simpler, but for the 2004 and 2005 terabyte track collections $P(rel)$ is actually fairly flat up to rank 1000. Our fit has a fixed slope for $\Delta P(rel)$ per 50 ranks, and compute a rank x' where the relevant fraction $y = rpt$:

$$x' = (rpt - slope - P(rel)) \cdot |J|/(-slope)$$

Following this fit, we estimate the size of a pool that contains 20 additional relevant documents per 200 judged:

$$poolsize = |J| + (x' - |J|) \cdot 2$$

For topics with very few relevant documents in the depth-50 pool, this size can by small or even actually be negative, and so we additionally require that the estimated pool size have 200 more documents to judge. We then estimate the depth for a pool of that size:

$$depth' = poolsize \cdot \lambda/|J|$$

	judged	relevant
depth-50 pool	639.7	117.9
sample	491.8	35.1
(not in pool)	210.2	14.0

Table 3. Average number of judged and relevant documents per topic in the pool and in the sample. The third line counts the sampled documents not present in the depth-50 pool.

where λ is 50. Once this depth is determined, we pool the runs to that depth, and compute the sampling rate as 200 divided by the number of unjudged documents in that pool.

The maximum depth of the sample for the 2006 terabyte topics varies from 57 to 1252, with an average of 314. Those topics with a shallow sample depth had very few relevant documents in the depth-50 pool, and consequently we judged additional documents from the next few ranks. For those topics, the sampling rate is very close to 100%, and the sampled judgments cover the pooled judgments nearly completely. Topics with a deep sample depth have correspondingly smaller sampling rates and diverge more from the pooled judgments.

Table 3 shows the number of judged and relevant documents per topic in the depth-50 pool and in the sample. The average number of new judged documents in the sample is not exactly 200 because the depth of the sample is based on an estimated pool size, and in any event we draw documents randomly according to the computed sampling rate. Note also that while we did not obtain 20 new relevant documents on average, we did find at least one new relevant document for 46 out of 50 topics, and 20 or more new relevant documents for 10 topics. Thus, most of these topics do continue to have relevant documents below our initial pool depth.

Zobel actually proposed an approach very similar to this one in his 1998 paper [17]. He fit an exponential curve to various estimates of $P(rel)$ and found very good fits up to the pool depths. Our linear estimator is sufficient for the terabyte collections because the sheer number of relevant documents makes the probability of relevance nearly linear at the ranks where we are looking.

Aslam et al. proposed an intricate sampling scheme intended to draw the best sample for estimating mean average precision in the runs being pooled [4]. The prior document probabilities favor documents at higher ranks, again so as to accurately estimate MAP [3]. Because of this early-rank prior, this method is not useful for probing deeply into the runs; because it is always more likely to draw a relevant document from earlier in a ranking, their sampling method will draw an earlier-retrieved document before one retrieved later.

5 Ranking Differences

Given these two sets of relevance judgments, one from a depth-50 pool and one a relevance-based random sample, we first examine if they rank the systems in the 2006 terabyte track differently. The systems were scored using mean average precision (MAP) with the pooled judgments, and inferred average precision with the sampled judgments.

Since the sampled judgments are by definition incomplete, and in practice the documents in the top ranks of the pool runs are fully judged for only a fraction of the topics, traditional measures such as mean average precision and precision at fixed rank cutoffs can't be computed exactly. While bpref is certainly usable in this situation, we have begun to experiment with a new measure, inferred average precision (infAP) [16]. infAP differs from bpref in that it is an estimate of average precision. When judgments are complete, infAP and MAP are equal. In the presence of unjudged documents, instead of assuming them to be nonrelevant, infAP estimates the set precision at those ranks using the precision at earlier ranks. This "interpolation" is only done across documents which could conceivably have been judged because they were contained within the sample range of the pool runs. Formally, infAP is defined as follows. Assume that the relevance judgments represent a uniform sample of a pool drawn up to some depth The expectation of precision of the set of documents up to rank k where k is the rank of a retrieved relevant document is

$$E[precision\ at\ rank\ k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} \left(\frac{|J_{k-1}|}{k-1} \cdot \frac{|R_{k-1}| + \epsilon}{|R_{k-1}| + |N_{k-1}| + 2\epsilon} \right)$$

where J_{k-1} is the set of retrieved documents present in the pool above rank k, R_{k-1} is the set of judged relevant documents above rank k, and N_{k-1} is the set of judged nonrelevant documents above rank k. Just as with mean average precision, this expectation is computed at each relevant document, and averaged over the known relevant documents to yield infAP.

Because infAP is an estimate of average precision, it provides a strong basis for comparing the ranking from the sampled set of relevance judgments to the MAP ranking based on pooled judgments. This is indeed the case in Yilmaz and Aslam's experiments [16]. The sampling scheme described above makes the situation somewhat different. Rather than being an estimate of the depth-50 MAP, our infAP scores will be estimates of the MAP from a pool of topic-specific depth.

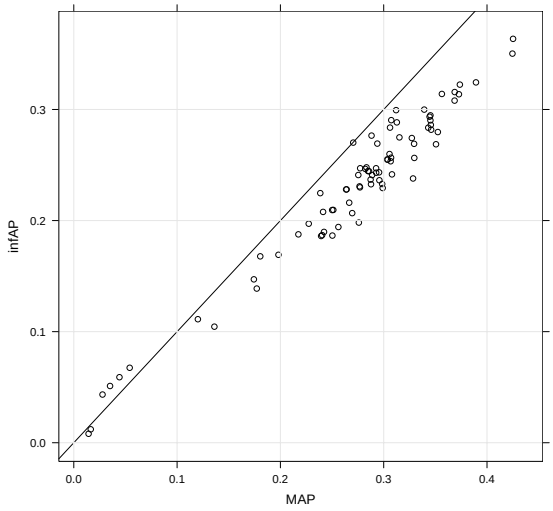

Figure 2. MAP and infAP scores for the terabyte 2006 runs.

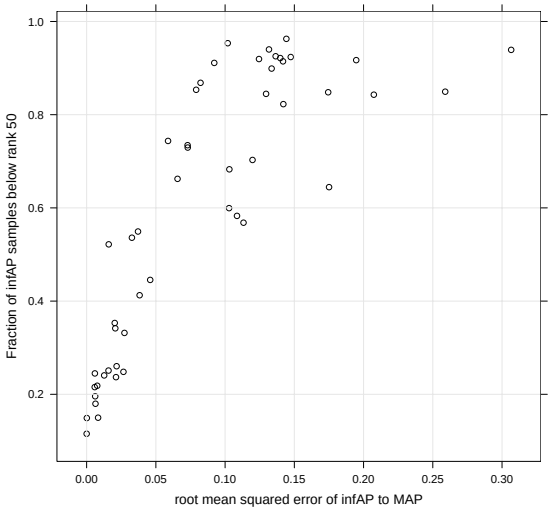

Figure 3. Per-topic RMS difference between MAP and infAP versus the fraction of the sampled judgments coming from below rank 50.

Figure 2 plots each run's MAP score based on the depth-50 pool against its infAP score based on the sampled judgments. The diagonal indicates where MAP equals infAP; for all but four runs, the infAP score is lower than the MAP score. The mean squared difference is 0.047. This difference is higher than Yilmaz and Aslam found in their experiments. We might attribute this to the fact that we are comparing to the depth-50 MAP rather than the "true" variable-depth MAP, but in truth MAP at depth 50 should be a good estimate of the variable-depth MAP for the pooled runs.

A common metric for comparing retrieval rankings is Kendall's tau (τ) rank correlation. Tau is equivalent to the number of pairwise swaps needed to convert one ranking into another. Voorhees established as a rule of thumb that a tau of 0.9 represents essentially identical rankings, with any differences that exist being in the noise of differences between assessor opinions [13]. The tau correlation between the two rankings we have here is 0.8, implying that the two rankings have notable differences despite being very highly correlated. Where does the difference in these rankings come from? Since those topics with few relevant documents in the depth-50 pool are represented by a nearly 100% sample in the sampled judgments, any difference must come from those topics where we sampled deeply.

Figure 3 illustrates this trade-off. The root mean squared difference between MAP and infAP scores for each topic are plotted against the percentage of the sampled documents that were first retrieved for that topic below rank 50, and thus could not have been in the depth-50 pool. It is clear that the infAP scores diverge the most from MAP when the sample mostly contains deep documents.

But should this necessarily be the case? On the one hand, a deep sample in our scheme requires a low sampling rate, and thus the documents in the top 50 ranks will not be well represented in the sample. But recall the hypothesis from the SIGIR 2003 workshop: the judgments will be incomplete simply because the pools are too shallow to contain all the relevant documents. But if the documents are essentially like those retrieved in the higher ranks, and were only left out of the pool because there were so many of them, then we would expect that the runs would rank the deep documents about the same as they rank the earlier ones. The overall tau is too low for this to be the case.

The most likely reason that infAP scores diverge from MAP is if very highly-ranked documents are not in the sample. If a relevant document is retrieved at rank 1, this has a very large effect on MAP, but if that document is not sampled, then infAP will necessarily be lower than MAP. This problem is particularly acute when the sampling rate is low, as it is in the deeply sampled documents. Simulation experiments with TREC-8 judgments show that forcing the sample to include everything retrieved at rank 1 has a big impact on the difference between infAP and MAP. Based on this, we are currently investigating stratified sampling approaches that combine good coverage at the very top of the ranking with deep samples where needed.

6 Reduced bias

The main goal of sampling deeply into the runs was to try and get different relevant documents than those higher in the rankings, specifically those with lower titlestat. Buckley et al. conjectured that the pools from smaller collections are deep enough (with respect to the collection as a whole) to contain a sufficient variety of documents such that a bias towards title-only queries is avoided [6]. If the terabyte collections only have high-titlestat judgments, then they are less reusable than if we had a variety of documents judged.

The *titlestat_rel* for the sampled judgments is 0.899, compared to 0.93 for the depth-50 pool. When we consider only the sampled documents below depth 50, the *titlestat_rel* is 0.851. Did this lower titlestat come from sparse sampling, or from going more deeply into the runs?

To try to answer this question, we divided the sampled documents into ten chunks based on the rank that the document was first retrieved by a pool run. We then computed the *titlestat_rel* for each chunk, and plotted that value against the median rank of the documents in that chunk. The plots for each topic are shown in Figure 4.

The graphs show that the sampling scheme indeed found lower titlestat documents when we sampled deeply, but not always, and sometimes they were found without needing to search so deeply. For example, topic 822 has low titlestat documents at rather shallow ranks. In contrast, topic 834 only achieves similarly low titlestat at a much greater depth. For topic 832, we sampled quite deeply without ever finding very low titlestat. Rarely do we see a topic with what we might think of as the expected pattern, a steady drop-off in titlestat values as documents come from deeper ranks.

All of this illustrates that title-word bias has a strong topic effect. For some topics, the title words are really the best indicators of document relevance. For others, there are other useful words not in the topic title. The number of relevant documents also plays a role.

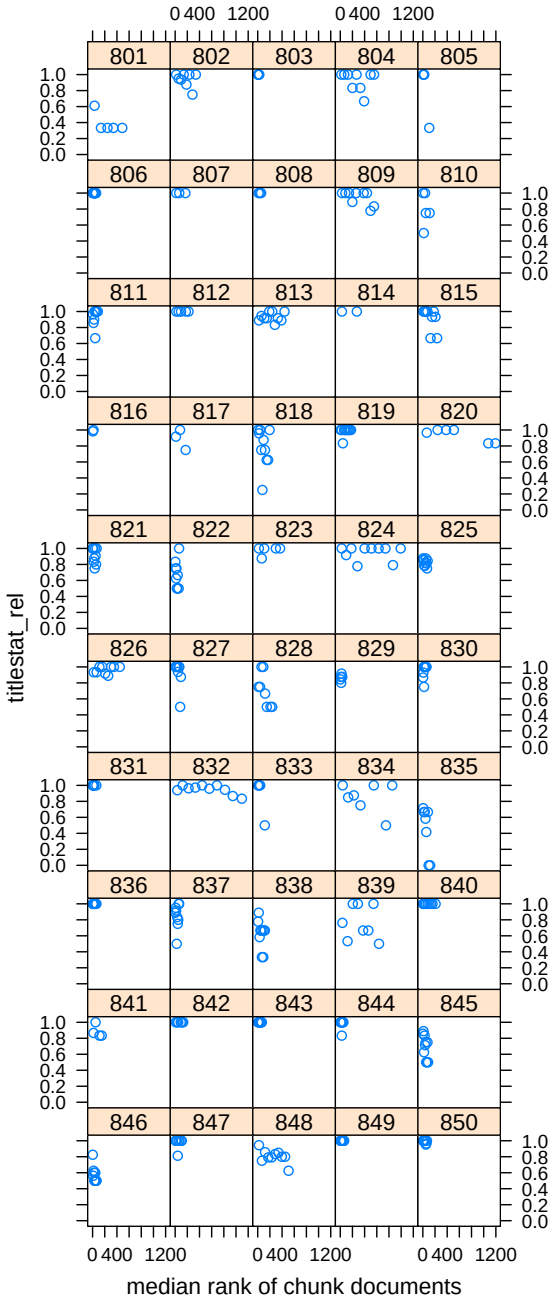

Figure 4. *Titlestat_rel* **at different ranks in the sampled judgments for each topic.**

7 Reusability

We next looked to see if the sampled judgments are any more or less reusable than the depth-50 pooled judgments. Table 4 shows the results of the leave-a-group-out test. The systems' infAP scores change somewhat less than do their MAP scores, and there is much less movement in the ranking with infAP. Bpref again is a much less reusable measure.

Note that this is a somewhat artificial experiment, because even though a group is held out of the infAP sample, it still contributed to the depth-50 pool which was used to decide the infAP sampling depth and rate.

Qrels	Measure	maximum abs. diff.	maximum rank change
Depth-50	MAP	0.03	$-15/+2$
	bpref	0.13	$-0/+29$
Sample	infAP	0.02	$-8/+2$

Table 4. Maximum absolute differences and ranking movement when leaving a group out of the 2006 judgment pools and samples.

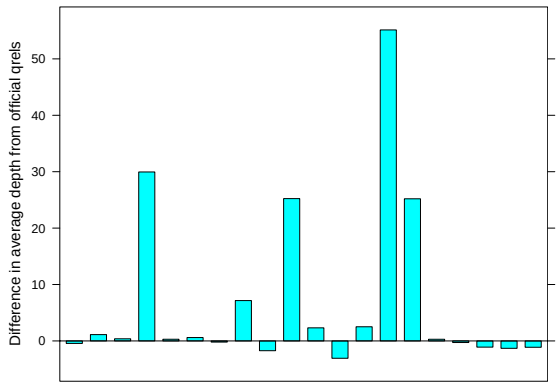

Figure 5. Difference in average per-topic sampling depths for each participating group when it is held out of the depth-50 pool. Each bar represents a group.

If in fact a group did not participate in the pooling process, then our sampling depths might have been somewhat different. It's not possible to draw a new sample after holding each group's runs out, so instead in order to measure this effect we recomputed the sampling depths for the residual judgments when each group was held out.

Figure 5 shows the difference for each group in the average per-topic sampling depth from that of the official qrels.[4] For most groups, the difference is minor (three ranks or less), but for some groups, we would have sampled much more deeply for some topics had they not participated. The runs from those groups contributed many unique documents (thus increasing the total number of judged documents), but relatively few of them were actually relevant.

The complete effect on these groups of being held out of the entire process is hard to gauge. The scores for these groups' runs do not change very much, and they cause minor ranking changes if any, when their unique relevant documents are held out. If they had not participated, and consequently we had sampled more deeply, we might have discovered their unique documents. However, deeper sample depths mean a lower sampling rate overall.

The sampling mechanism itself presents another risk, namely that the sample that happens to be drawn at evaluation time will be infelicitous and not score a run or a group fairly. This is a risk to the evaluation itself, and also to future reuse.

To measure any effect that random sampling might have on infAP scores, we drew 100 random subsamples of the depth-50 pool, and used these qrels subsets to score the runs using infAP. The samples were drawn at the same rates as were used to select the official sampled judgments; the difference here is that we only drew the sample within the pool, so that all sampled documents would have a relevance judgment.

From this data, we conducted two analyses of variance. The first ANOVA looked at the variance of all the infAP scores as a function of topic and sample. The sample was never a significant effect (even at $\alpha = 0.05$) but topic was significant for 45 out of 50 topics. Incorporating the runs into the model is complicated simply because runs normally vary in effectiveness in a topic-dependent fashion. In the end we ran a second ANOVA of infAP score by topic and sample within each run. For two runs, sample was moderately significant ($p = 0.016$ and 0.014); these were two runs from the same group, and their maximum average infAP scores in any sample were 0.1002 and 0.0866. From this, we conclude that the variance across samples such as those we are drawing should not be a worry.

8 Future work

There are a number of unanswered questions in this work. Chief among them is whether or not the sampled judgments are "more fair" to future runs than the pooled judgments. This seems intuitively like it should be so, but it's actually quite a difficult question to answer. In contrast to smaller TREC collections, where we are more sure of having a sufficiently complete set of judgments that can be thought of as "truth", we have no sufficiently complete qrels for the terabyte collections. We have done a number of simulation experiments in the older collections, but the relatively low occurrence of relevant documents in those collections is itself challenging. For example, MAP computed from a depth-20 pool in TREC-8 is actually very close to MAP in the official depth-100 pool. So creating equivalent conditions for simulation is difficult.

Furthermore, it's hard to know for sure if the sampled judgments are "more true" than the pool because we observe a large topic effect in titlestat. Whether we can sample deeply to overcome bias depends on if the bias exists for the topic, which itself depends somewhat on how the title section is stated. Sometimes title-word bias is exactly what systems should always do, because that is simply the best articulation of the search need given the document collection, and there really are no other kinds of relevant documents than those which contain the topic title words.

Another question is that of an optimal sampling strategy that balances meaningful measures of effectiveness with reusability and low bias. Uniform random sampling is not usable because it will not recover

[4] The "official qrels" sampling depth is slightly different than that derived from the depth-50 pool. This is because after the both the pools and samples were judged, a second judging pass was made over both qrels sets to rectify conflicting judgments between near-duplicate documents, as recommended by [5]. In the graph here, we are comparing to the depth that would have been used based on the duplicate-corrected qrels. The difference involved is very small.

enough relevant documents at any reasonable sampling rate. Sampling strategies such as that of Aslam et al. focus too strongly on the early ranks, and as such fall prey to title-word bias [4]. However, early ranks need to be covered if MAP is to be accurately estimated; if the sample rate is too sparse, then it is likely we will miss judging documents from the first two or three ranks, which are critical to MAP. The sampling strategy described here can be too sparse to cover those early ranks. At the "deep end" of the sample, because we have no complete judgments to compare to, we can't be certain that we're locating enough low-titlestat documents to make the collection sufficiently more fair. Currently, we are investigating whether stratified sampling strategies (and measures that can cope with them) can solve this problem.

Related to this is the amount of time we spent judging pool. We almost certainly did not need to judge to rank 50 to train our sampler adequately. The depth-50 pool is much larger than the sample. We would prefer to spend more time on deep samples and less time in the pool.

Lastly, our present understanding of test collection reusability is very limited. The leave-a-group-out test only measures reuse for systems that could have contributed to the pool. However, retrieval algorithms do improve with time (we hope!). A leave-a-group-out study of the TREC-8 collection shows it to be reusable, but that collection is now nearly eight years old, and in the meantime completely different retrieval models have been developed, largely using that collection and others as measuring devices.

9 Conclusion

The TREC 2006 terabyte track created two sets of relevance judgments, one a depth-50 pool, the other a random sample of a "virtual" pool where the depth and sampling rate varied per topic. The sampling model was trained from the judgments on the depth-50 pool, an approach we call *relevance-based sampling*. Systems are ranked differently between the two sets of judgments. At present, we can't determine if one is more "true" or "correct" than the other, because there is no sufficiently complete set of judgments in the terabyte collection for comparison, and also because of topic effects. Intuitively, it seems that the sampled judgments should offer a more complete measure of effectiveness, but several issues in the sampling strategy need to be more closely examined. Most of the discrepancy is likely due to sparse samples missing documents retrieved at rank 1.

A short series of experiments, including holding a group out of the pool and looking at sample variance, seems to indicate that the sampled judgments are more reusable than the pooled ones. This is only an argument for using the sampled judgments if you believe the sampled judgment ranking, but it is certainly good to know that sampling doesn't make a collection *less* reusable.

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*, Seattle, WA, August 2006.

[3] J. A. Aslam, V. Pavlu, and E. Yilmaz. Measure-based metasearch. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, Salvador, Brazil, August 2005. ACM Press.

[4] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* [2], pages 541–548.

[5] Y. Bernstein and J. Zobel. Redundant documents and search effectiveness. In *Proceedings of the 14th International Conference on Information and Knowledge Management (CIKM 2005)*, pages 736–743, Bremen, Germany, November 2005. ACM Press.

[6] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* [2], pages 619–620.

[7] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 25–32, Sheffield, UK, July 2004. ACM Press.

[8] C. L. A. Clarke, I. Soboroff, and N. Craswell. Overview of the TREC 2004 terabyte track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, Gaithersburg, MD, November 2004.

[9] C. L. A. Clarke, I. Soboroff, and N. Craswell. The TREC 2005 terabyte track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, November 2005.

[10] C. W. Cleverdon. The cranfield tests on index langauge devices. In *Aslib Proceedings*, volume 19, pages 173–192, 1967.

[11] I. Soboroff, E. Voorhees, and N. Craswell. Summary of the SIGIR 2003 workshop on defining evaluation methodologies for terabyte-scale test collections. *SIGIR Forum*, 37(2), Fall 2003.

[12] 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.

[13] E. 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].

[14] E. 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.

[15] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiments in Information Retrieval Evaluation*. MIT Press, 2005.

[16] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 102–111, Arlington, Virginia, November 2006.

[17] J. 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.