Score Adjustment for Correction of Pooling Bias

William Webber and Laurence A. F. Park

Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia
{wew,lapark}@csse.unimelb.edu.au

ABSTRACT

Information retrieval systems are evaluated against test collections of topics, documents, and assessments of which documents are relevant to which topics. Documents are chosen for relevance assessment by pooling runs from a set of existing systems. New systems can return unassessed documents, leading to an evaluation bias against them. In this paper, we propose to estimate the degree of bias against an unpooled system, and to adjust the system's score accordingly. Bias estimation can be done via leave-one-out experiments on the existing, pooled systems, but this requires the problematic assumption that the new system is similar to the existing ones. Instead, we propose that all systems, new and pooled, be fully assessed against a common set of topics, and the bias observed against the new system on the common topics be used to adjust scores on the existing topics. We demonstrate using resampling experiments on TREC test sets that our method leads to a marked reduction in error, even with only a relatively small number of common topics, and that the error decreases as the number of topics increases.

Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and software – *performance evaluation*.

Keywords

Retrieval experiment, evaluation, system measurement

General Terms

Measurement, performance, experimentation

1. INTRODUCTION

Information Retrieval (IR) systems are evaluated using test collections, which contain a document corpus, a set of topics to run against that corpus, and judgments (called qrels) as to which documents are relevant to which topics. A system returns a ranked list of documents or run for each query. The documents are marked for relevance using the qrels, and an evaluation metric is applied to the resulting vector of relevancies to calculate an effectiveness score for the run. The mean of the per-topic scores becomes the effectiveness score for the system against the test collection.

Relevance judgments are performed by human assessors, and are expensive to collect. It is not in general practical to assess every document in the corpus for relevance to every topic. Instead, the top documents from the runs of the systems participating in the test collection's original experiment are *pooled*, and only these documents are assessed. The assumption behind pooling is that, if a diverse enough range of good systems contribute to the pool, and if the systems are pooled to a sufficient depth, then the pool should contain "almost all" the documents in the corpus relevant to each topic. If a new, unpooled system is run against the test collection, and it returns unpooled, unassessed documents, these documents can reasonably be *assumed irrelevant*. As a result, despite its incompleteness, the test collection is acceptably robust to reuse.

As document corpuses have grown in size, however, the assumption that pooling will retrieve nearly all relevant documents has become increasingly suspect. Unassessed documents returned by new systems may therefore in fact be relevant, and assuming them to be irrelevant will lead to an evaluation bias against new systems. In addition, the bias can be systematic against systems that are different in nature from those which contributed to the pool. For instance, Buckley et al. [2007] suggest that recent large TREC collections have their pools flooded with documents rich in query keywords, and are biased against retrieval methods that attempt to go beyond keyword matching. Such systematic bias is not merely unfair to certain systems, but is an obstacle to an entire direction of potential retrieval improvement.

The pooling approach requires deep assessments of the pooled runs, in order to provide good coverage of the set of relevant documents and make the collection reusable. Users, though, rarely look beyond the first page of results [Joachims et al., 2005], so deeper assessment is not necessary to capture the typical user experience. Assessing a large number of queries to a shallow depth gives greater experimental power than assessing a small number of queries deeply [Webber et al., 2008]. One would rather assess $1{,}000$ queries to depth 10 than 100 queries to depth 100. Anecdotal evidence suggests that shallow, broad assessment is indeed the approach taken by large search engines; see, for instance, the data set described in Najork and Craswell [2008]. Methods that allow for the reliable reuse of such shallowly-assessed topics are attractive.

Rather than assuming unassessed documents to be irrelevant, which is biased against new systems, unassessed documents can instead be ignored. This can be done either by using a special-purpose metric such as BPref [Buckley and Voorhees, 2004], or else by excising the unassessed documents from the run, shifting

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

the remaining (assessed) documents up, and then evaluating the resulting *condensed list* with a standard evaluation metric [Sakai, 2007]. However, the fact that a document has not been returned by other systems is evidence for its not being relevant. Therefore, excising it from the run, and promoting in its place a document that has been returned by other runs and so is more likely to be relevant, leads to a bias in favor of unpooled systems [Sakai, 2008].

Assuming unassessed documents to be irrelevant, then, is biased against new systems, while condensing runs by excising unassessed documents is biased in favor of new systems. Nor is the degree of bias fixed. Rather, it depends on the comprehensiveness of the pool, and therefore on the number, quality, and variety of the pooled systems. Where a small number of similar systems are shallowly pooled, the bias of assumed irrelevance is strong, while that of condensed lists is relatively weak. As the comprehensiveness of the pool increases, the bias of assumed irrelevance decreases, and the bias of condensed lists increases. Therefore, neither assumed irrelevance nor condensed lists are appropriate and bias-free in all circumstances, nor will a static adjustment method work.

1.1 Our contribution

In this paper, we propose to address the problem of bias against unpooled systems by estimating that bias, and adjusting the unpooled score accordingly. The estimation is made empirically, from the systems under evaluation, and rests only on sampling theory; it requires no prior information or model fitting. The bias estimate becomes an adjustment factor, and this factor is added to the score of the unpooled system, to derive the adjusted score.

The first method of adjustment we examine does not require the new system to be fully assessed for any topics. Instead, the bias estimate is derived from the existing, fully-pooled systems, using a leave-one-out experiment. This method is straightforward and can be applied with entirely static collections such as those produced by the TREC effort. However, the resampling technique applied assumes that the existing, fully-pooled systems and the new, unpooled system have been randomly sampled from a common system population. If the new system is significantly better, or just significantly different, then the random sampling assumption is invalid, and the adjustment method is unreliable and likely to understate the system's effectiveness.

A more robust method of bias estimation and adjustment can be applied if full assessment of the new system, along with the existing ones, is available on a set of common topics. The error on the unpooled score of the new system can be directly observed on the common topics, and used as an estimate of unpooled bias for that system. The estimate is applied as an adjustment to the unpooled scores achieved by the new system on existing topics. The resulting adjusted scores are unbiased, and have markedly less mean error than the unadjusted ones. Lower mean error is achieved with only a few common topics, and the error decreases as the number of common topics increases. Estimating bias based on a set of fully-assessed topics is preferable in its assumptions to the leave-one-out estimation upon the fully-pooled systems because the inference is being made not from systems to systems, but from topics to topics. Therefore, the underlying assumption is not the dubious one that the systems have been randomly sampled, but the more reasonable one that the topics have. Using our method, the system evaluator can leverage a small number of common topics to reuse the assessment effort already spent on a large number of existing topics.

1.2 Related work

Buckley and Voorhees [2004] propose BPref as a special-purpose metric for handling incomplete relevance information. Yilmaz and Aslam [2006] calculate AP directly on lists from which unjudged documents have been excised, a method which they refer to as Induced AP. Sakai [2007] suggests applying general-purpose metrics to runs with unjudged documents excised, and introduces the expression *condensed lists* to describe such runs. Sakai also demonstrates that BPref is in fact a restricted form of AP on condensed lists, where evaluation is cut off once a certain number of nonrelevant documents are seen in a run. All of these three papers perform experiments in which incomplete relevance information is formed by randomly sampling documents from the full qrels set. This construction method is inherently unbiased and therefore highly artificial; it does not simulate the effect of shallow pooling or of comparing unpooled against pooled systems. Sakai [2008] instead creates incomplete relevance information by partial pooling, and demonstrates that in this circumstance, condensed lists lead to bias in favor of new systems.

Unassessed documents pose such a thorny problem in part because most existing evaluation metrics do not directly express the degree of uncertainty that arises from the presence of unassessed documents in a run. For many metrics, indeed, the uncertainty is difficult to quantify, particularly where the metric in normalized by the number of relevant documents. Moffat and Zobel [2008] propose a new metric, Rank-Biased Precision (RBP), which is unnormalized but naturally convergent, with the contribution of each rank having a fixed weight. A run's RBP score is expressed not as a single value, but as a base value and residual. The residual exactly quantifies the uncertainty that results from incomplete assessment. If two systems have overlapping residuals, then it cannot be concluded for certain that one is superior to the other.

Yilmaz and Aslam [2006] propose that documents for assessment be chosen by uniform random sampling from the pool. The sampled documents are then used to estimate the true score. The estimator is unbiased, but has relatively high variance. They apply this sampling method to AP, referring to the resulting metric as Inferred AP. Aslam et al. [2006] instead use a lower variance unequal sampling scheme, in which a document is sampled with probability proportional to its weight under the evaluation metric employed. These sampling methods cannot be applied in environments where incomplete assessments have been chosen by non-random means, such as pooling of a subset of systems. Aslam and Pavlu [2008] combine the pooling and random sampling using stratified sampling. Stratified sampling is applied by Yilmaz et al. [2008] to an environment which mixes pooling and random sampling. Their finding that the method is not subject to pooling bias is not confirmed in application [Carterette et al., 2008], possibly because aggregating probability of inclusion across multiple runs by taking the mean of the per-run probabilities may not properly account for reinforcement by like systems.

Instead of pooling on the one hand, or random sampling on the other, a number of authors have proposed that documents for assessment should be chosen in an attempt to maximize some evaluation goal; for instance, to boost the proportion of relevant documents [Cormack et al., 1998], or to focus on the score accuracy of the best-performing systems [Moffat et al., 2007]. Carterette et al. [2006] select documents so as to maximize confidence that one system does or does not have a positive score delta with another, using a simple fixed probability of relevance model. A more complex model is developed in Carterette [2007]. Each system is treated as an expert, and in returning or failing to return a document, a system is asserting its judgment as to the probability that a document is relevant. The higher the document is returned in a ranking, the stronger the assertion of its probability of relevance. Then, as documents are incrementally assessed, the reliability of each system can be progressively calibrated. Multiple logistic regressions are used to aggregate the evidence and formulate a probability of relevance. These probabilities of relevance can be used to directly estimate a score for a system. When employed in practice, estimated scores were consistently a third of actual scores [Carterette et al., 2008]; this suggests that a strong bias would occur if the method were used to compare pooled and unpooled systems, particularly if the number of pooled systems was small.

2. UNPOOLED BIAS

Our approach is to estimate the degree of bias that a system suffers from not being pooled based on a leave-one-out experiment. Estimation can be undertaken solely on the existing, pooled systems, and then the result applied to the new, unpooled system; however, inference from systems to systems is problematic, as we demonstrate. Preferably, if a common set of topics for which the new system is also fully assessed is available or can be created, the bias against the new system can be directly measured on that subset. The resulting bias estimate is then applied as a score adjustment to the unpooled score. The unpooled score can be calculated either by assuming that all unassessed documents are irrelevant, or else by excising unpooled documents and evaluating the condensed lists.

This section begins by introducing the materials and methods employed. The degree of bias that unpooled systems suffer under assumed irrelevance, and enjoy under condensed lists, is then illustrated on test set data.

2.1 Materials

Two test sets from the TREC effort are used in this paper. The first is from the 2004 Robust Track [Voorhees, 2004]. It consists of 110 systems submitted by 14 groups, run against 249 topics. The large number of topics makes this data set particularly attractive for use in meta-evaluation studies such as this. Of the topics, 200 were drawn from earlier tracks of TREC, and the relevance assessments from these tracks reused, without new assessments being performed, meaning that not all documents returned by 2004 Robust Track systems have assessments. Additionally, only a subset (albeit a plurality) of systems were pooled for the 49 new topics. To avoid confusion between documents unassessed in the original collection, and documents unassessed because of experimental withdrawal from the pool, the former are eliminated by expanding the original qrel set with non-relevant judgments for all unassessed documents. This affects only 3% of total returned documents for the old topics, and 1.5% for the new.

The second data set used is from the AdHoc track of TREC-8 [Voorhees and Harman, 1999], consisting of 129 systems submitted by 40 groups, run against 50 topics. The additional value contributed by this data set is the manual runs it contains. Manual runs allow for human involvement in query formulation and reformulation. They typically outperform automatic runs, and in particular find a much higher proportion of unique relevant documents. In TREC-8, the 13 manual runs find 24% of the relevant documents, while the 116 automatic runs between them only find 17% (the remainder are returned by both categories of runs). Similarly, the best 11 manual systems are also the best 11 systems over all, at least under some metrics. The Robust test set contains no manual runs. It will be observed later in this paper that methods of score estimation that appear to perform well on homogeneous system sets often perform poorly on the more interesting case of heterogeneous sets. We test this by attempting to use information from automatic systems to estimate scores on manual ones.

An evaluation metric is a function that takes a vector of relevancies and produces a real-valued score that summarizes the vector, rewarding the return of relevant documents, and generally giving higher weight to higher rankings. Topics have differing degrees of difficulty. Many metrics attempt to compensate for this by normalizing scores based on the number of relevant documents for a topic. Also, metrics can be evaluated to a greater depth than runs are pooled, with assessments for documents beyond pool depth in one ranking being available if those documents were returned before pool depth in another ranking. Both normalization and evaluation beyond pooling depth add to the complexity of score estimation, since finding relevant documents in one run affects the scores of other runs.

The metric employed in this paper is *rank-biased precision at ten* (RBP@10). Rank-biased precision assigns geometrically decaying weights to each rank, with the score being the inner product of the relevance and weight vectors. Since the geometric sequence is convergent, RBP scores fit within the range of $[0, 1)$. Also, because each rank has a fixed weight (due in part to the exclusion of a normalization step), the degree of uncertainty arising from partial unassessment (either because evaluation stops at a certain rank, or because the relevance of some documents up to that rank is unknown) can be precisely stated as a residual value. For this paper, we take the base RBP value as our point score. Additionally, we cut off evaluation at depth 10, and adjust the rank weights accordingly, assigning no weights (or residual) to ranks 11 and beyond. The precision-at-ten metric was also employed in experiments, but since the outcomes obtained are very similar to RBP@10, the results are not separately reported here.

As mentioned in the introduction, unassessed documents can be handled either by assuming them irrelevant or by condensing the lists. We implement condensing of lists by shuffling in documents from beyond evaluation (hence pooling) depth. If there are insufficient assessed documents in the ranking, the trailing positions are filled with placeholder irrelevant documents.

2.2 Bias of exclusion from the pool

In this section, we observe the empirical bias of excluding systems from the pool. This can be derived from a leave-one-out experiment. We sample a set of fully-pooled systems from the test set, and randomly select one system s from the sampled set. The score of s under full assessment is calculated. Then, we form a pool consisting of the sampled systems, but excluding s, which is equivalent to marking all documents uniquely returned by s to pool depth as unassessed. The partial score of s is calculated, using either presumed irrelevance or list condensing. The bias is then the difference between the partial and true score. This is calculated for system s on every topic. The random sampling of system sets and unpooled system is repeated multiple times, to give a distribution of biases. The whole experiment is performed for different sizes of the fully-pooled set, allowing us to empirically relate bias to pooled system set size. Note that this is an observational study based on sub-sampling. The full set of systems being sub-sampled from is not itself randomly sampled from the universal population of systems, so we cannot infer that the observed mean biases and bias distributions hold for all systems. This investigation supplements that undertaken in Sakai [2008]; however, our focus is on a much smaller number of pooled systems than are investigated there.

Figure 1 graphs the mean and quartile biases for rank-biased precision; the figure for precision at ten (not shown) is very similar. Using condensed lists is biased in favor of the unpooled system, while assuming unassessed documents to be irrelevant is biased against it. For assumed irrelevance, the bias steadily decreases as the number of pooled systems or *pool width* increases, roughly halving when the pool width is doubled, as the number of unassessed

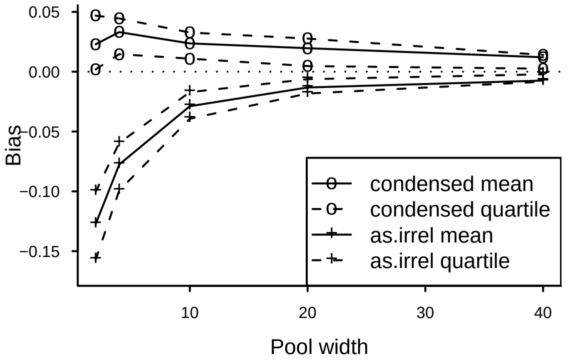

Figure 1: Empirical RBP ($p = 0.8$) bias for unpooled system for different numbers of pooled systems. Pooling is to depth 10. The TREC 2004 Robust Track data set is used. Graphed is the mean and quartiles of the difference in mean system RBP score between the true score for the unpooled system and the score using either condensed lists or assumed irrelevance. Each data point represents 100 system set subsamplings.

documents (not shown) steadily drops. The behavior of condensed lists is more complex. Fewer unassessed documents means that the condensed and true relevance vectors are more similar. At the same time, more pooled systems omitting a document strengthens the odds that the document is in fact irrelevant, and therefore strengthens the bias resulting from excising it and replacing it with a pooled document, partially counteracting the effect of better accuracy. As a result, condensed lists have less bias than assumed irrelevance for small numbers of pooled systems, but assumed irrelevance has less bias for larger ones. We also observe that for wide pools, the distribution is quite skewed (mean is close to third quartile), as the frequency with which the unpooled system is largely "covered" by an almost-identical runset from the same family increases.

3. ADJUSTING SCORES FOR BIAS

Two methods of estimating unpooled bias are discussed in this section. The first, *bias inference from systems*, requires no fully-assessed common topics, and instead takes its estimate from a leave-one-out experiment on the fully-pooled systems. The method is simple and requires no additional assessment effort. However, in inferring from systems to systems, it dubiously assumes that the systems have been randomly sampled from the same population. The second, *bias inference from topics*, requires that a subset of topics be fully assessed for all systems, and directly observes the bias against the otherwise unpooled system on those topics. While this requires extra assessment effort, the method makes the more reasonable assumption that the topics have been randomly sampled from the same population. A formal analysis of the common-topics method is provided, and experimental results are given.

3.1 Bias inference from systems

Section 2.2 examined the empirical bias that results from excluding a system from the pool, both under assumed irrelevance and using condensed lists, for a particular test set. This observed bias, or perhaps an average across a few different test sets, could be directly used as a global score adjustment factor for unpooled systems. The problem is that we could not be certain that the results for one test set would be transferable to another. An alternative is to derive the estimate directly from the experimental set at hand.

Algorithm 1 Adjust scores based on inference from systems

```
$T \leftarrow$ set of topics
$S \leftarrow$ set of (pooled) systems
$Q \leftarrow$ set of qrels on $T$ derived from pool of $S$
$r \leftarrow$ (unpooled) system
for $s \in S$ do
    $Q' \leftarrow Q \backslash \{$documents uniquely pooled from $s\}$
    $Q' \leftarrow Q' \cup \{$documents returned by $r\}$
    $p_s \leftarrow$ mean (pooled) score of $s$ evaluated against $Q$
    $u_s \leftarrow$ mean (unpooled) score of $s$ evaluated against $Q'$
    $\beta_s \leftarrow p_s - u_s$                    $\triangleright$ unpooled bias against $s$
end for
$a \leftarrow \sum_{s \in S} \beta_s / |S|$                    $\triangleright$ adjustment factor
$u_r \leftarrow$ mean (unpooled) score of $r$ evaluated against $Q$
return $u_r + a$
```

Pool Width	Robust		Manual	
	Raw	Adjusted	Raw	Adjusted
2	0.127	0.041	0.451	0.302
4	0.078	0.028	0.384	0.294
10	0.029	0.015	0.283	0.237
20	0.013	0.008	0.231	0.203
40	0.007	0.006	0.177	0.159

Table 1: Bias inference from systems. Mean absolute error (MAE) of leave-one-out score adjustment and unadjusted scores for RBP ($p = 0.8$) under presumed irrelevance, for different numbers of pooled systems. The left columns are for all systems from the TREC 2004 Robust Track. The right columns show estimation of unpooled manual system scores from pooled automatic systems on the TREC-8 AdHoc Track data set.

Adjustment factors for an unpooled system being compared to a set of pooled systems can be derived by a leave-one-out experiment. Say that an unpooled system r is being compared to S, a set of pooled systems. We remove each of the systems in S from the pool in turn, and calculate its unpooled score, either by assuming unassessed documents to be irrelevant or by condensing the run. The difference between mean unpooled and pooled scores of each system in S is the observed pooling error for that system. The average across the observed errors of S provides an estimate of the pooling bias for the test set, and therefore of the adjustment factor a that should be added to the mean score of unpooled systems to correct for this bias. An important refinement to this method is that when each system s is withdrawn from the pool S for its unpooled score to be calculated, the new system r is added to the pool to replace it. This has the effect of retaining any documents common to s and r but not found in $S \backslash s$. Otherwise, we would be estimating the penalty against an unpooled system in an $n - 1$ pool, and our adjustment would be biased; specifically, it would tend to overestimate the adjustment a. This leave-one-out adjustment method is described in Algorithm 1.

The effectiveness of leave-one-out score adjustment can be experimentally assessed by resampling on an existing test set. For this purpose, we take the 2004 TREC Robust Track data set. A set of $n + 1$ systems are randomly sampled from the full system set, with uniform probability. One of these systems is selected to act as the unpooled system r, and the remaining n to form S, the set of pooled systems. Judgments from the original qrel set are reused. The mean score of r across all 249 topics is adjusted based on a leave-one-out experiment on S. The system sampling is repeated 100 times for each pool size. For this experiment, unassessed documents are assumed irrelevant. The resulting mean absolute error (MAE) between the true score on the one hand, and the unpooled score (raw or adjusted) on the other, is then calculated. Let t_i be the true score for topic i, let s_i be the unpooled score (raw or adjusted), and let N be the number of topics; then:

$$MAE = \frac{1}{N} \sum_{i}^{N} |t_i - s_i| \, .$$

The MAE figures for the experiment are reported in the left-hand columns of Table 1. Adjustment leads to much greater accuracy of scores, particularly with smaller pools. In addition, it is unbiased (as likely to over as to underestimate), whereas the unadjusted scores are all underestimates, meaning that bias or mean error (not separately reported in the table) is identical to mean absolute error.

The apparently good results obtained on the 2004 TREC Robust Track data set are, however, misleading. The uniform random sampling employed is artificially beneficial to the adjustment method being examined. For instance, it is not surprising that the adjusted scores are unbiased, because for every randomly-selected set of systems that leads to a high estimate, there will be another randomly-selected set that leads to a compensating low one. This would not be a problem if the real-world evaluation environment in which this technique was used were indeed one in which systems were being randomly sampled for evaluation, but in general this will not be the case. Rather, the new system under the evaluation will be one that the developer has consciously tried to make better than the existing ones. What can happen when the leave-one-out score adjustment method is employed in such a situation is illustrated by the right-hand columns of Table 1. Here, the TREC8 AdHoc Track data set is employed. The new, unpooled system whose score is to be adjusted is randomly selected from the 11 best manual systems, while the pooled systems are sampled from the remainder of the system set. As described previously, the manual systems are significantly different from and better than the automatic ones, as shown by the large number of unique relevant documents they return. Exclusion from the pool and the use of presumed irrelevance greatly underestimates the performance of these manual runs, and while leave-one-out adjustment helps, there is still a strong error, even with large pool sizes.

3.2 Bias inference from topics

Section 3.1 has examined the derivation of adjustment factors from a leave-one-out experiment on the fully-pooled systems. The basic principle was inference from one set of systems to another system. As was pointed out, the more the inferred-to system differs from the inferred-from systems, the more tenuous this inference becomes. And when performing evaluation on a new system, that system is generally only interesting to the degree that it is different from, and better than, the existing ones.

A more robust inference can be performed if there exists, or can be created, a set of *common topics* for which both the existing systems S and the new system r are fully assessed. In this case, the bias against r of being omitted from the pool can be directly observed on the common topics, since both true and unpooled scores are known. Then, this observed bias can be generalized as an adjustment factor for that system's scores on the topics for which it is genuinely unpooled. The process is described in Algorithm 2. Inference from common topics to unpooled topics is more robust than from pooled systems to unpooled systems because we are inferring from one set of topics to another, rather than from systems to systems, and it is more reasonable to assume that the topics are randomly sampled from the same population than that systems are; indeed, in some experimental settings, random sampling of topics can be directly enforced.

Algorithm 2 Adjust scores based on inference from topics

```
$T \leftarrow$ set of topics
$S \leftarrow$ set of (pooled) systems
$Q_T \leftarrow$ qrels on $T$ derived from pool of $S$
$r \leftarrow$ (unpooled) system
$C \leftarrow$ set of common topics
$Q_C \leftarrow$ pool $S \cup r$ on $C$ and assess for relevance
for $c \in C$ do
    $Q'_C \leftarrow Q_C \backslash \{$documents uniquely pooled from $r\}$
    $p_{r,c} \leftarrow$ (pooled) score of $r$ on $c$ evaluated against $Q_C$
    $u_{r,c} \leftarrow$ (unpooled) score of $r$ on $c$ evaluated against $Q'_C$
    $\beta_{r,c} \leftarrow p_{r,c} - u_{r,c}$        ▷ unpooled bias against $r$ on $c$
end for
$a \leftarrow \sum_{c \in C} \beta_{r,c}/|C|$        ▷ adjustment factor
$u_r \leftarrow$ mean (unpooled) score of $r$ evaluated against $Q_T$
return $u_r + a$
```

We begin by offering a formal analysis of the common-topics adjustment method as a form of sample-based ratio estimation. Then an experimental assessment is performed, which validates the formal analysis and demonstrates that common-topics adjustment leads to greatly improved accuracy over unadjusted scores, even if the new system is quite different from the existing ones.

Analysis

The proposed method is a form of *ratio estimator* [Thompson, 2002, Chapter 7]. Ratio estimators are of use where a cheap but inaccurate measure x is available for every element of a population, while the more costly true value y is only known for a sample. The mean ratio r between x and y is estimated from the sample, and applied to the x values across the population to estimate the true mean value of y; that is, $\hat{\mu}_y = r\mu_x$. For us, the desired value is the mean of the true scores μ_t (or simply μ), and the cheap approximations are the unpooled scores u. We use arithmetic difference rather than ratio for the adjustment, since it can readily happen that the unpooled score u_i for a topic i is 0 when the true score t_i is greater than 0, in which case the ratio t_i/u_i is undefined. Let N be the total number of topics, unpooled and common, and n be the number of common topics, that is, the topics for which all systems are fully assessed. Analytically, we will be treating the n common topics as randomly sampled from the full set of N topics. So the estimated adjustment a, derived from the n common topics, is:

$$a = \frac{1}{n} \sum_{i=1}^{n} (t_i - u_i) \, . \tag{1}$$

The estimate of the true mean score μ, using the *adjusted estimator*, for all N topics based on the unpooled scores u_i is:

$$\hat{\mu}_a = \frac{1}{N} \sum_{i=1}^{N} (u_i + a) \, , \tag{2}$$

where the quantity defined on the left of the equation should be understood as "the a-based estimator of true mean μ", not "the estimator of the mean of a". Of course, for n of these N topics we know the true score; however, since by derivation the adjustment a will be exactly correct for the mean of these n unpooled scores, we do not need to separately account for these n topics in the estimator. The adjustment a is an estimate of the true adjustment A that should be applied to the unpooled scores of all N topics in order to get the true mean score. As a is derived by sampling n deltas from the full N deltas whose mean is A, it follows that a is an unbiased estimator of A, and therefore that $\hat{\mu}_a$ is an unbiased estimator of μ, the true mean score. The variance of this estimator is:

$$\mathrm{var}(\hat{\mu}_a) = \frac{N-n}{N} \cdot \frac{\sigma_a^2}{n} \,. \tag{3}$$

The left-hand fraction here and subsequently adjusts for the small population; that is, for the fact that the exact values are known for n of the N elements in the population, and estimation is only be applied for the remaining $N - n$. The numerator of the right-hand fraction is:

$$\sigma_a^2 = \frac{1}{N-1} \sum_{i=1}^{N} (t_i - (u_i + a))^2 \tag{4}$$

namely, the mean squared error of the per-topic adjusted scores against the true scores across all N topics (loosely speaking, the variance of the adjusted scores).

Instead of the adjusted estimator $\hat{\mu}_a$, we could take the mean true score $\bar{t}$ from the n topics for which full assessment has been performed. This is also an unbiased estimator of μ, the true mean score. The variance of this *sampled estimator* $\hat{\mu}_n$ is given by elementary sampling theory, and is:

$$\mathrm{var}(\hat{\mu}_n) = \frac{N-n}{N} \cdot \frac{\sigma_t^2}{n} \,, \tag{5}$$

where σ_t^2 is the variance of the true scores across all N topics. Comparing Equations 3 and 5 shows that adjusted estimator is more accurate than the sampled estimator when σ_a^2, the MSE of the adjusted scores, is less than σ_t^2, the variance of the true scores. And this is something which, due to the high inter-query variance of most metrics, is quite generally the case. That is, adjusted scores are in general closer to true scores than true scores are to their mean. Indeed, of the 500 different randomly-sampled system sets used in the experimental section that follows, in not one is the variance of true scores less than the MSE of adjusted scores, either with unassessed documents assumed irrelevant or condensed lists employed.

The unadjusted scores could be used instead of the adjusted ones as a (generally biased) *unadjusted estimator*. The error on the unadjusted scores is A, of which a is an estimator. The error on the adjusted scores is $A - a$; that is, it is dependent on the degree to which a's estimation of A is incorrect. Therefore, the adjusted scores will be more accurate than the unadjusted scores if $0 < (a/A) < 2$; that is, if the following two conditions are met:

1. the estimated adjustment a is the same sign as the true adjustment A
2. the estimated adjustment a is no more than twice the true adjustment A

Where the unadjusted scores always misestimate the true scores in the same direction (always underestimate them, or always overestimate them), as occurs when unassessed documents are assumed irrelevant, Condition 1 is met. And since the expected value of a is A, Condition 2 will be met the vast majority of the time (exactly how often depends on the distribution of a).

However, where the unadjusted scores can underestimate the true scores for some topics, and overestimate them for others, Condition 1 is not guaranteed. Additionally, although $a = A$ in expectation, its distribution may spread beyond 0 at one end, potentially violating Condition 1, and above $2A$ at the other, potentially violating Condition 2. The cumulative density beyond these limits gives the probability that the adjusted scores are less accurate than the unadjusted ones, and this probability depends on the distribution of a. If we assume a to be normally distributed under the central limit theorem (CLT), then to have 68% confidence that the adjusted scores are more accurate than the unadjusted ones requires that the standard deviation of the adjustment estimator be less than the true adjustment. That is,

$$\sigma_{\hat{\mu}_a} = \sigma_a \cdot \sqrt{\frac{(N-n)}{N \cdot n}} < |A| \,, \tag{6}$$

which is derived by taking the square root of Equation 3. For different degrees of confidence, different percentiles of the normal cumulative distribution function should be checked. For small n, where the assumption of normality is dubious, a bootstrap can be used instead.

Algorithm 3 Sample systems, topics to assess adjustment accuracy

```
$T \leftarrow$ 249 TREC Robust 2004 topics
$X \leftarrow$ 110 TREC Robust 2004 systems
$I \leftarrow 100$                        ▷ number of system sampling repeats
$J \leftarrow 200$                        ▷ number of topic sampling repeats
for $w \in \{2, 4, 10, 20, 40\}$ do                           ▷ pool widths
    for $i \in 1 \rightarrow I$ do
        $S \leftarrow \mathrm{sample}(X, w)$
        $r \leftarrow \mathrm{sample}(X \backslash S, 1)$
        $Q \leftarrow$ pool $S \cup r$ on $T$
        $Q' \leftarrow$ pool $S$ on $T$
        $t_r \leftarrow$ mean (true) score of $r$ evaluated against $Q$
        $u_r \leftarrow$ mean (unpooled) score of $r$ evaluated against $Q'$
        for $n \in \{10, 20, 40, 100\}$ do
            for $j \in 1 \rightarrow J$ do
                $C \leftarrow \mathrm{sample}(T, n)$               ▷ common topics
                $a \leftarrow$ estimate adjust. on $C$ as in Algorithm 2
                $e_r \leftarrow t_r - (u_r + a)$             ▷ adjustment error
                $E_{w,n} \leftarrow E_{w,n} + |e_r|$
            end for
        end for
    end for
end for
$E \leftarrow E/(I * J)$             ▷ Take mean error over $I * J$ repeats
return $E$
```

Of course, in most evaluation settings, the true values of A and σ_a^2 are unknown, so Equation 6 cannot be directly calculated. The value of σ_a^2 can be estimated as:

$$s_a^2 = \frac{1}{n-1} \sum_{i=1}^{n} (t_i - (u_i + a))^2 \tag{7}$$

that is, the observed MSE on the sampled topics. This would enable us to assess whether using the adjusted scores across all N topics rather than just the true scores on the n topics was likely to improve accuracy — which, as noted before, will usually be the case. As for the adjustment A, the estimator for it is a itself.

Experiments

The purpose of this section is to empirically assess the improvement in accuracy that score adjustment, based on bias inference from topics, provides over using the unadjusted, unpooled scores. Necessarily, the precise results achieved apply only to the particular test sets examined, but they are in accordance with the preceding formal analysis, and are indicative of what might be expected in general.

Pool Width	Unadj	Mixed True and Unadjusted				Adjusted			
		Common Topics				Common Topics			
		10	20	40	100	10	20	40	100
2	0.127	0.122	0.117	0.107	0.076	0.044	0.032	0.024	0.019
4	0.078	0.074	0.071	0.065	0.046	0.033	0.024	0.019	0.014
10	0.029	0.028	0.027	0.024	0.017	0.018	0.013	0.010	0.008
20	0.013	0.013	0.012	0.011	0.008	0.010	0.008	0.006	0.005
40	0.007	0.007	0.007	0.006	0.004	0.007	0.005	0.004	0.003

Table 2: Bias estimates from topics. MAE of unadjusted, mixed ($N - n$ unadjusted, n true), and adjusted scores. Metric is RBP@10 ($p = 0.8$). Columns denote number of common topics used to derive adjustment factors. Rows denote number of systems in the pool. Values are the MAE of the error between the mean estimated score and the mean true scores, across the non-common topics.

Pool Width	Unadj	Mixed	Adjusted			
		10	10	20	40	100
2	0.034	0.033	0.041	0.028	0.019	0.010
4	0.035	0.034	0.031	0.022	0.015	0.008
10	0.024	0.023	0.018	0.013	0.009	0.005
20	0.020	0.019	0.013	0.009	0.007	0.003
40	0.012	0.012	0.009	0.007	0.005	0.003

Table 3: MAE of unadjusted, mixed, and adjusted scores, based on condensed lists. Metric is RBP ($p = 0.8$). Other details are as for Table 2.

Pool Width	Unadj	Mixed	Adjusted		
		10	10	15	20
2	0.451	0.361	0.060	0.046	0.037
4	0.384	0.308	0.059	0.045	0.037
10	0.283	0.226	0.054	0.041	0.033
20	0.231	0.185	0.050	0.038	0.031
40	0.177	0.141	0.045	0.035	0.028

Table 4: MAE of unadjusted and adjusted scores, on the TREC 8 manual dataset. Metric is RBP ($p = 0.8$). Other details are as for Table 2.

We begin by exploring the relative accuracy of unadjusted and adjusted scores, with unassessed documents assumed to be irrelevant. The error on adjusted scores is observed using the experiment described in Algorithm 3. The mean absolute error (MAE) from the true score is taken as the measure of accuracy of the adjusted scores, with each MAE entry averaged from $20{,}000$ subsamples. We also record the error on unadjusted scores for each system sample. In addition, the true scores of the common topics are mixed with the unadjusted scores of the remaining topics and their error is calculated. This is necessary for a fair comparison, since in the adjusted scores the adjustment is precisely correct for the mean of the common topics.

The results of the experiment described in Algorithm 3 are given in Table 2. We first observe by following along the rows of the adjusted score results that the error of adjusted scores is proportional to $\sqrt{(N-n)/(n \cdot N)}$, as the analysis predicts. (To be exact, the analysis makes this prediction of RMSE, but it holds true of MAE as well.) In contrast, the error of the unadjusted scores, with true scores mixed in, is proportional to $(N-n)/n$, and hence declines at a slower rate. For narrow pools, score adjustment offers a marked improvement over unadjusted scores, even for a handful of topics in the common topic set. If the number of common topics is increased, quite high fidelity can be achieved even with only a couple of fully-pooled systems. As the pool width increases, the error of the unadjusted scores decreases, and while adjustment still improves accuracy, the relative benefit for the same number of common topics is decreased. On the other hand, as the number of unassessed documents decreases, the effort involved in taking a given number of topics for which r was unpooled and filling in the unassessed documents decreases, allowing a larger common topic set to be created for the same amount of overall effort.

Table 3 reports the same experiment as Table 2, but this time using condensed lists to handle unassessed documents in the unpooled system, both for the unadjusted scores and as the base for the adjusted scores. Only the mixed scores with 10 true scores mixed in have been reported; the remainder decline at the same rate as for Table 2. Condensed lists tend to be biased in favor of the unpooled system (see Figure 1), but the error is not uniform. Particularly in the case of very narrow pools, condensed scores on some topics are less than true scores. This means that variability is high relative to bias, which in turn leads to high variance in the adjustment estimator (see Equation 4). Equation 6 predicts that this will diminish the accuracy of the adjusted scores relative to the unadjusted, condensed ones, and this is indeed what we observe in Table 3. For a pool width of 2, the unadjusted scores are more accurate than the adjusted scores with 10 common topics, and are roughly as accurate for a pool width of 4. As the pool width increases, though, the error becomes more consistently one-sided (that is, condensed scores are higher than true ones), and the adjusted scores become more reliable. In any case, increasing the number of common topics improves the quality of the adjusted scores more rapidly than of the unadjusted ones.

Section 3.1 observed the problems caused for inferring the adjustment from the pooled systems to the unpooled system when the latter is significantly different from the former. The point was demonstrated using the manual runs of the TREC 8 AdHoc track as the unpooled systems. In Table 4, the unpooled scores of these same manual runs are adjusted using the common-topic method instead. Since there are only 50 topics altogether in this test set, the number of topics that can experimentally be held as common is limited. Nevertheless, the utility of the common-topics adjustment method is clear. The error of the estimate with only 10 common topics and 2 pooled systems is well under half that of the unadjusted score with 40 pooled systems. Even if the unpooled system is distinctly different from the pooled ones, score adjustment from common topics provides serviceable accuracy.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a simple and robust sampling method for correcting the score bias suffered (or enjoyed) by unpooled systems. The method proposed does require that a certain number of topics be fully assessed for the otherwise unpooled system, as well as for the systems in the pool. Additionally, the fully-assessed topics must be randomly sampled from the same population as the existing, partially-unpooled ones. However, we suggest that both of these conditions will often exist already in private lab work, and can in most cases be cheaply and reliably attained if they do not. In return, our method offers an adjustment method that is unbiased and much lower variance than either using the unadjusted scores, or making do with only the fully-assessed topics. In addition, the degree of variance itself can be estimated from the topics used, and additional common topics added to reduce it if desired.

The unpooled score approximations dealt with in this paper have been the simple ones of, on the one hand, assuming unassessed documents to be irrelevant, or on the other hand, excising them and condensing the rankings. However, the method can be applied to any approach to approximating unpooled scores — or indeed any situation in which an expensive, exhaustive assessment on one small set of topics might be supplemented by a cheaper, approximate assessment on a second, larger set. This might indeed include methods where the approximated evaluations are made with little or no human assessment at all.

Score adjustment could also be incorporated with more complex sampling and inferential schemes, as one form of evidence amongst many. Score adjustment has the particular advantage that, whereas pooling or other sampling bias is a problem for many schemes, adjustment directly addresses and substantially solves the issue. Perhaps its best use in this field might be as a sort of sanity check for more complex methods.

However, the main attractions of the sampling method proposed here are the minimal inferential assumptions it makes, and its robustness to pooling bias. This recommends it to evaluators working for the most part in a traditional pooled setup, and who are wary of more complex, potentially more fragile inferential methods.

Acknowledgments

This work was supported by the Australian Research Council.

References

J. Aslam and V. Pavlu. A practical sampling strategy for efficient retrieval evaluation. Technical report, Northeastern University, 2008.

J. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In S. Dumais, E. Efthimiadis, D. Hawking, and K. Järvelin, editors, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 541–548, Seattle, USA, August 2006.

C. Buckley and E. Voorhees. Retrieval evaluation with incomplete information. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, Sheffield, United Kingdom, August 2004.

C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10(6):491–508, December 2007.

B. Carterette. Robust test collections for retrieval evaluation. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–62, Amsterdam, the Netherlands, July 2007.

B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In S. Dumais, E. Efthimiadis, D. Hawking, and K. Järvelin, editors, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 268–275, Seattle, USA, August 2006.

B. Carterette, V. Pavlu, E. Kanoulas, J. Aslam, and J. Allan. Evaluation over thousands of queries. In S. Myaeng, D. Oard, F. Sebastiani, T. Chua, and M. Leong, editors, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 651–658, Singapore, Singapore, July 2008.

G. Cormack, C. Palmer, and C. Clarke. Efficient construction of large test collections. In W. Croft, A. Moffat, C. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 282–289, Melbourne, Australia, August 1998.

T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161, Salvador, Brazil, August 2005.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):1–27, 2008.

A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 375–382, Amsterdam, the Netherlands, July 2007.

M. Najork and N. Craswell. Efficient and effective link analysis with precomputed SALSA maps. In *Proc. 17th ACM International Conference on Information and Knowledge management*, pages 53–61, Napa, USA, October 2008.

T. Sakai. Alternatives to Bpref. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 71–78, Amsterdam, the Netherlands, July 2007.

T. Sakai. Comparing metrics across TREC and NTCIR: The robustness to system bias. In *Proc. 17th ACM International Conference on Information and Knowledge management*, pages 581–590, Napa, USA, October 2008.

S. Thompson. *Sampling*. John Wiley & Sons, New York, 2nd edition, 2002.

E. M. Voorhees. Overview of the TREC 2004 robust retrieval track. In E. M. Voorhees and L. P. Buckland, editors, *Proc. TREC-13*, November 2004. NIST Special Publication 500-261.

E. M. Voorhees and D. K. Harman. Overview of the eighth text retrieval conference. In E. M. Voorhees and D. K. Harman, editors, *Proc. TREC-8*, November 1999. NIST Special Publication 500-246.

W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *Proc. 17th ACM International Conference on Information and Knowledge management*, pages 571–580, Napa, USA, October 2008.

E. Yilmaz and J. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proc. 15th ACM International Conference on Information and Knowledge Management*, pages 102–111, Arlington, Virginia, USA, November 2006.

E. Yilmaz, E. Kanoulas, and J. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In S. Myaeng, D. Oard, F. Sebastiani, T. Chua, and M. Leong, editors, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 603–610, Singapore, Singapore, July 2008.