Can Deep Effectiveness Metrics Be Evaluated Using Shallow Judgment Pools?

Xiaolu Lu
RMIT University
Melbourne, Australia

Alistair Moffat
The University of Melbourne
Melbourne, Australia

J. Shane Culpepper
RMIT University
Melbourne, Australia

ABSTRACT

Increasing test collection sizes and limited judgment budgets create measurement challenges for IR batch evaluations, challenges that are greater when using deep effectiveness metrics than when using shallow metrics, because of the increased likelihood that unjudged documents will be encountered. Here we study the problem of metric score adjustment, with the goal of accurately estimating system performance when using deep metrics and limited judgment sets, assuming that dynamic score adjustment is required per topic due to the variability in the number of relevant documents. We seek to induce system orderings that are as close as is possible to the orderings that would arise if full judgments were available.

Starting with depth-based pooling, and no prior knowledge of sampling probabilities, the first phase of our two-stage process computes a background gain for each document based on rank-level statistics. The second stage then accounts for the distributional variance of relevant documents. We also exploit the frequency statistics of pooled relevant documents in order to determine a threshold for dynamically determining the set of topics to be adjusted. Taken together, our results show that: (i) better score estimates can be achieved when compared to previous work; (ii) by setting a global threshold, we are able to adapt our methods to different collections; and (iii) the proposed estimation methods reliably approximate the system orderings achieved when many more relevance judgments are available. We also consider pools generated by a two-strata sampling approach.

KEYWORDS

Test collection; relevance assessment; pooling; shallow judgments.

1 INTRODUCTION

Batch evaluations are performed by calculating a metric score based on a set of judged documents. Despite five decades of success, this "Cranfield/TREC" paradigm also faces challenges. One of the key issues is that realistic collection sizes now greatly exceed the budget available to perform human judgments. "Pooling-to-depth-d" is one widely-used approach [25], in which documents in the union of the top-d lists returned from a set of contributing systems are judged, but other documents are not. The pooling depth d is ideally determined by the needs of the effectiveness metric to be used, but in reality is also constrained by the experimental budget. Although pooling has identified the majority of relevant documents in earlier collections [30], there is growing evidence that this is not true for the web collections that are now the norm [2, 11].

The uncertainty in effectiveness measurement in large collections is the key emphasis of our work here, focusing on how to estimate evaluation scores when reduced judgment sets are used. This is not a new problem, and a range of prediction mechanisms have been proposed [1, 22, 23, 27, 28], mainly focusing on predicting system orderings. We focus on prevailing pool-based test collection construction methods, as these best match our methodology, and on deep evaluation metrics, noting that pool depth has a lesser impact on shallow evaluation metrics such as ERR [6]. Alternative approaches using direct sampling exploit prior knowledge of the probability of each document being judged, and are applied during pool construction, on the assumption that all systems requiring measurement have been identified. But that process makes it difficult to infer scores for any new systems that get added later. On the other hand, pooling selects documents based on the assumption that top-ranked documents are both more likely to be relevant, and hence more influential in computing effectiveness scores. In this more general setting there is no *a priori* knowledge of the system scores, and while that means that regression cannot be applied, new systems can be considered. We also argue that the decision to apply score adjustment should be done on a per topic basis. Robertson [17] notes that topics vary in terms of the number of potential relevant documents, and that this can have a significant impact on evaluation scores. Dynamically identifying when to perform score adjustment is thus a second challenge that must be considered.

The end objective of an evaluation goes beyond the metric scores, of course; in the end we wish to be able to compare and choose between systems, meaning that it is also important that the score estimations are concordant with the system orderings that would arise if full knowledge were available. Since the latter is measured according to a reference point which may not be known, there is no clear optimization goal, another complication that we address. These various considerations lead to two questions:

Research Question 1: *For each topic, how can we estimate the evaluation score of a system using a shallow pooling depth?*

Research Question 2: *Can stable system rankings be achieved using the adjusted scores?*

In considering these two questions, we perform experiments using several different ad-hoc test collections and a range of modeled pool depths. Our results show that: (i) a two-stage optimization framework generates more accurate score estimations than previous approaches; (ii) topic-based adjustment thresholds identified using early TREC collections allow additional improvements in

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080793

estimation accuracy; and (iii) the adjusted evaluation scores yield better approximations of the "true" system rankings than do the unadjusted scores. In addition to standard pooling methods, we also consider two-strata sampling [24].

2 RELATED WORK

Incomplete Judgments and Evaluation Bias. Two types of bias arise in batch evaluations: pooling depth bias [19] and system bias [20]. The first is caused by the use of shallow pools, and the second by performance underestimation for systems that did not contribute to the pool. Both are a result of documents appearing in the ranking for which judgments are not available. The simplest response to unjudged documents is to stipulate that anything not examined in the pooling process is not relevant. Zobel [30] challenged this notion using a series of leave-one-out experiments, and showed for several early TREC collections that while it was likely there were indeed further relevant documents that had not been identified, system bias was nevertheless within acceptable levels. However, on more recent web collections, there is growing evidence that this situation may not be assumed [2, 11].

Other responses to the issue of unjudged documents have been proposed. Buckley and Voorhees [3] describe BPref, which balances the rank positions of documents judged as non-relevant and relevant, and ignores unjudged documents. In a related approach, Sakai [18] considers *condensed lists*, which compute scores using a filtered ranking containing only judged documents, and finds that standard metrics give higher discriminative ratios than achieved by BPref. However, the condensed list methodology has not been shown to be stable when comparing relative system orderings using Kendall's τ or discrimination ratios [19, 20].

Score Estimation / Collection Construction. Documents without judgments are not distributed randomly in ranked result lists. Therefore, sample-based collection construction approaches have been suggested to support statistical inference [1, 22, 23, 27, 28]. Yilmaz and Aslam [27] present an inferred Average Precision (AP) metric that uses an expectation model, and can be coupled with a sampling process to select documents to be judged. Their InfAP metric uses uniform random sampling during collection construction. When compared with standard TREC-style pooling, the results produced by InfAP were strongly correlated with AP. However, this sampling process is random, and retrieval systems return documents in rank order, meaning that relevant documents are more likely to be returned at the top of the list if the system is effective.

The use of non-random sampling has also been explored. Yilmaz et al. [28] extended their previous work, proposing metrics XInfAP and XInfNDCG, based on a stratified sampling process. In contrast, Aslam et al. [1] consider the use of importance sampling for the same task, proposing statAP, which estimates the expectation of AP. The key difference between InfAP and statAP is that statAP is designed to generate the optimal distribution estimates using all of the contributing systems. Voorhees [24] further examines the effect of sampling methods on inferred metrics.

A recent study by Schnabel et al. [23] also used importance sampling, this time in conjunction with Discounted Cumulative Gain (DCG). The key idea in their approach was to use the probability of relevance with respect to rank information when determining the sample distribution. They provide an analysis on how to derive the optimal sampling distributions under different system comparison settings [22]. Using the proposed framework, any metric can be reformulated in the form of expectations and be estimated directly from the sampling process. Moffat et al. [14] had earlier examined targeted pooling and document judgment order in conjunction with the Rank-Biased Precision (RBP) metric.

Score Estimation Based on Pooling Methods. Estimation in traditional pooling techniques has also received considerable attention [4, 7, 9, 10, 16, 26]. Most existing techniques focus on adjusting the bias which exists between pooled and unpooled systems. Webber and Park [26] proposed two methods to perform score adjustment. The first uses an adjustment factor, which is computed from the contributing systems. Each contributing system has an error value assigned when it is left out of the training process, and the mean of those values is applied to any new system to be measured. The second approach requires a set of common topics with "complete judgments". A similar calculation is performed in order to obtain the adjustment factor, but restricted to the subset of common topics. To obtain additional adjustment accuracy, Webber and Park introduced randomization to build an unbiased estimator.

Recent work by Lipani et al. [9] using a precision metric outperformed the first method of Webber and Park. Their "anti-Precision" measurement is similar in spirit to the residual computed by RBP [15]. Lipani et al. [9] compute adjustment factors using the leave-one-run out methodology, and then improve their previous approach by computing an average distribution [10].

The closest work to our current approach is that of Ravana and Moffat [16]. They focus on pooling depth bias, proposing three methods to estimate the effect of unjudged documents, using the residual that can be computed for weighted-precision metrics [15]. Their first method uses a background estimation based on a static scaling factor; the second assumes that the percentage of relevant but unjudged documents can be derived directly from the known score component; and the third uses a parametric combination of the first two. Lu et al. [12] subsequently define the same problem in terms of the anticipated effectiveness gain as a function of ranking depth. Based on different assumptions derived from the underlying gain distributions, they propose several alternatives, and compare the estimates achieved. They empirically show that relatively simple models can be used to estimate gain values for unjudged documents.

An approach due to Büttcher et al. [4] directly predicts the relevance of unjudged documents, using two types of classifiers trained with the existing pool to predict the relevance of unjudged document in a new system. Although the effectiveness of the classifier is low, their results show that classification does help maintain similar system orderings when measured via Kendall's τ. Jayasinghe et al. [7] take a similar approach, and show that reliably predicting document relevance is often difficult.

3 PRELIMINARIES AND BASELINES

Pools. Figure 1 shows the construction of a pool for one topic, with $s_{j,i}$ (on the left) corresponding to the jth document in the run for system S_i, and with the corresponding documents (on the

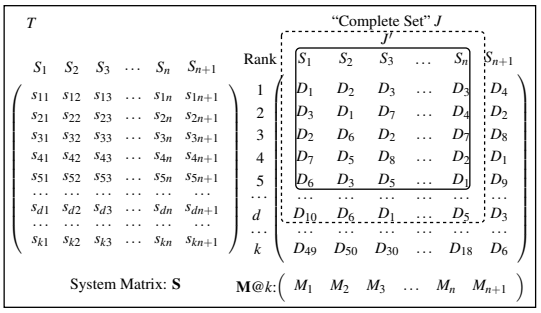

Figure 1: Pooling process for a topic T. The left matrix is a rank-based representation; the right one shows the equivalent document identifiers. The two boxes indicate two possible sets of pooled documents, the larger to depth d, and the smaller to some depth $d' < d$. The metric M is evaluated at some depth k, where k may or may not be less than or equal to d or d'.

right), each potentially retrieved by multiple systems at different rank positions. Hence, a document D can also be represented by its rank-position information, $\langle D, (\mathsf{p}_{D,1}, \mathsf{p}_{D,2}, \ldots, \mathsf{p}_{D,n})\rangle$, in which $\mathsf{p}_{D,i}$ is the rank returned for D by contributing system S_i. Metric evaluation to depth d for systems S_1 to S_n requires that the documents in the set $J = \{D \mid \min_{i=1}^{n} \mathsf{p}_{D,i} \le d\}$ be judged. That is, both matrices can be further mapped to a matrix of relevance $\mathbf{R}_{d\times n}$ in which $r_{j,i}$ is a relevance, or *gain*, value.

If there is insufficient judgment volume available, a shallower pool J' might be formed, with documents D for which $\min_{i=1}^{n} \mathsf{p}_{D,i} > d'$ not judged, and elements in $\mathbf{R}_{d\times n}$ left without values. Unknown relevance labels may also arise for a new system S_{n+1}, regardless of the pooling and evaluation depths. In this framework, the first of the two research questions proposed in Section 1 can be split into two aspects: (1a) for each topic, how do we estimate the scores of a system using a set of shallow pooled judgments; and (1b) which topics may assume that unjudged documents are not relevant and which ones should not.

One method for dealing with missing data is to compute expected gains as a function of retrieval rank [12]. However, modeling relevance as a function of rank only considers the LHS representation in Figure 1, and ignores that documents can have multiple ranks. Addressing that limitation is a key part of our work here.

Metric Residuals. Suppose that for some topic T, a set of documents J results from pooling to depth d (Figure 1). Consider the ranked list returned by some system $S_i = (s_{1,i}, s_{2,i}, \ldots, s_{k,i})$ and let $r_{j,i}$ represent the gain of the document at rank j, normally (but not necessarily) a value in $[0, 1]$. The effectiveness M_i of S_i when computed to depth k by a weighted-precision metric M is:

$$\mathrm{M}_i = \mathrm{M}_{@k}(S_i, J) = \sum_{\substack{j=1 \\ s_{j,i}\in J}}^{k} r_{j,i} \cdot W_{\mathrm{M}}(j)\,, \tag{1}$$

where $W_{\mathrm{M}}(j)$ is the weight assigned by the metric at depth j, with $\sum_{j=1}^{\infty} W_{\mathrm{M}}(j) = 1$ [13, 15]; and where the restriction $s_{j,i} \in J$ is required to ensure that only defined values of $r_{j,i}$ are included. A corresponding residual Δ_i can then be computed, quantifying the metric weighting associated with the unjudged documents [15]:

$$\Delta_i = \sum_{\substack{j=1 \\ s_{j,i}\notin J}}^{k} r_{max} \cdot W_{\mathrm{M}}(j) + \sum_{j=k+1}^{\infty} r_{max} \cdot W_{\mathrm{M}}(j)\,, \tag{2}$$

where r_{max} is the maximum possible gain. Either term might be zero, depending on whether S_i contributed to the pool, on the relationship between the evaluation depth k and the pooling depth d, and on whether $W_{\mathrm{M}}(j) = 0$ when $j > k$, as occurs with truncated metrics.

There is a three-way tension between metric depth (quantified as the expected point reached in the ranking in the corresponding user model [13]); accuracy of measurement, captured by the residual; and the cost $|J|$ of performing the judgments. For example, in RBP the tail residual (the second component in Equation 2) is given by p^k, and if $p = 0.5$, $k \approx 10$ is sufficient. Similar calculations apply for ERR [6]. But in either case, the first term of Equation 2 might be non-zero for new runs. Furthermore, even the tail residuals might become large for deeper metrics, for example, RBP with $p = 0.95$. Truncated (that is, non-infinite) metrics such as Scaled DCG at depth 100, $\mathrm{SDCG}_{@100}$, also require deep pools if the residual is to be moderately bounded. The same requirement must apply by implication to other deep metrics such as Average Precision.

Problem Definition. Consider a set of n contributing systems $\{S_1, S_2, \ldots, S_n\}$. For one topic T, let d be a pooling depth at which it is believed that a majority of the relevant documents occurring in the runs of those systems have been identified. We refer to this set of judgments J as the "complete set". Let $d' < d$ be a shallower pooling depth, with judgments forming an incomplete set $J' \subseteq J$. Given a weighted precision metric M, the effectiveness score of S_i evaluated using M and J to depth d is denoted as $\mathrm{M}_i = \mathrm{M}_{@d}(S_i, J)$, with a residual of Δ_i. Similarly, an estimated metric score based on judgments to depth $d' < d$, is denoted as $\hat{\mathrm{M}}_i = E_d(\mathrm{M}_{@d'}(S_i, J'))$ where $E_d(\cdot)$ is an estimation function for the same metric at depth d. Following Lu et al. [12], the estimation error ϵ_i is then defined as:

$$\epsilon_i = \begin{cases} \mathrm{M}_i - \hat{\mathrm{M}}_i & \text{if } \hat{\mathrm{M}}_i < \mathrm{M}_i\,, \\ 0 & \text{if } \mathrm{M}_i \le \hat{\mathrm{M}}_i \le \mathrm{M}_i + \Delta_i\,, \\ \hat{\mathrm{M}}_i - (\mathrm{M}_i + \Delta_i) & \text{if } \hat{\mathrm{M}}_i > \mathrm{M}_i + \Delta_i\,. \end{cases} \tag{3}$$

This definition respects the residual range, and only gives non-zero values if the estimated effectiveness falls outside the score range arising from the use of J at depth d. The challenge is to develop a method $E_d(\cdot)$ that estimates the depth-d effectiveness score of a contributing system based on a subset J' of the judgments, and minimizes the average value of ϵ_i.

In the experiments in Section 6 we report the RMS aggregate of the ϵ_i values computed, across systems and topics; and, as a "percentage accurate", the fraction of those values that are zero.

Lower-Bound Estimation. A simple approach is to take $E_d(x) = x$, that is $\hat{\mathrm{M}}_i = \mathrm{M}'_i$, where M'_i is the score for system S_i when evaluated using J', and assert that documents outside J' do not alter the score. Taking unjudged documents to be not relevant is the normal default in batch evaluation, and is a valid estimator. But the estimation quality depends on the breadth of the pool, and whether a majority of relevant documents have been identified. When there

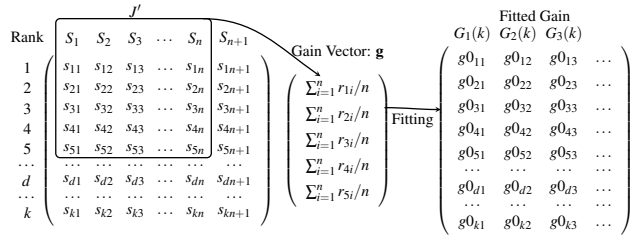

Figure 2: Overview of rank-based estimation for a single topic. The judgments J' are used to infer an observed gain vector $\mathbf{g}$; each of a set of m functions $G_\ell(k)$ is then fitted to $\mathbf{g}$.

are still many unjudged relevant documents, this estimator results in underestimation of system performance. Reasonably good rank correlation between the estimates and the true score over a set of systems can be obtained, but there is no guarantee that the performance of each of the systems has been accurately measured.

Interpolative Estimation. A second baseline is provided by the RM interpolative estimator proposed by Ravana and Moffat [16], who scale the metric score across the residual, assuming that unjudged documents are relevant at the same rate as judged ones:

$$\hat{\mathrm{M}}_i = \mathrm{M}'_i/(1-\Delta'_i)\,. \tag{4}$$

A collection-based background probability is used when $\Delta'_i = 1$. This estimator assumes that gain is accrued at the same rate across all of the documents retrieved by the system, both judged and unjudged. Although more robust than the LB estimator, it does not allow the likelihood of relevance to decrease as the pool is extended from d' to d.

Rank-Based Estimators. Lu et al. [12] introduce rank-based estimation, illustrated in Figure 2. The judgments J' are used to estimate expected gain as a function of rank on a per-topic basis. Those rank-based fractional gain predictions are then used for unjudged documents – interpolated at depths up to d', and extrapolated from d' to the metric evaluation depth k. Lu et al. explore alternative estimation functions, measuring the prediction error using the mechanism described in Equation 3, and find that while improvements are possible, no single estimator works consistently well across all collections and topics. Rank-based estimation also has the drawback of ignoring the fact that a document can appear at different ranks for different systems; and hence potentially assigns different gain estimates to the same document in different runs, a representational issue that usually leads to a biased estimation [29]. As a further drawback, an entire row in the system matrix S (see Figure 1) must be judged in order to compute the expected gain, limiting construction methods to pooling or sampling by rank, and possibly excluding stratified sampling processes.

Sampling-Based Estimation. Other sampling approaches can also be used when forming the judgment set. Voorhees [24] describes a two-strata sampling method, which consists of shallow pooled judgments J' to some depth d', and then 10% random sampling to depth d in a second set J_s. These judgments allow computation of inferred recall-based metrics, and also inferred versions of weighted-precision metrics, with $\hat{\mathrm{M}}_i$ for system i calculated as:

$$\hat{M}_i = \sum_{\substack{j=1 \\ s_{j,i}\in J'}}^{k} r_{j,i}\cdot W_{\mathrm{M}}(j) + \lambda\cdot \sum_{\substack{j=1 \\ s_{j,i}\in J_s}}^{k} r_{j,i}\cdot W_{\mathrm{M}}(j)\,, \tag{5}$$

where

$$\lambda = \left(\sum_{\substack{j=1 \\ s_{j,i}\notin J'}}^{k} W_{\mathrm{M}}(j)\right)\cdot\left(\sum_{\substack{j=1 \\ s_{j,i}\in J_s}}^{k} W_{\mathrm{M}}(j)\right)^{-1}$$

and where the second term in Equation 5 estimates the total gain associated with documents contained in the second stratum. Here, λ is the interpolation estimator. Note that Equation 5 only adapts the RM method for sample based judgments.

4 TWO-STAGE ESTIMATION

Overview of the Framework. To compute score estimates, we propose a two-stage framework, guided by a unified optimization goal, and built on a set of $m \geq 1$ per-topic rank-level estimators. The overall structure of this mechanism is described in Algorithm 1. We omit the process of obtaining rank-level estimations, discussed briefly in the previous section, and in detail by Lu et al. [12]. That is, we assume as our starting point here that m different rank-based estimators have been generated, each derived from the judged documents $D \in J'$, and that values for a set of gain functions have been computed, with $g0_{j,\ell}$ the gain associated with an unjudged document that appears in the j th position of any of the n system

Algorithm 1 Estimation Framework

Input: System matrix $\mathrm{S}_{k\times n}$; partial relevance judgments J' with $g2[D]$ the gain associated with document D for $D \in J'$ and undefined otherwise; and a set of m rank-level background gain estimates, $g0_{j,\ell}$ for $1 \leq j \leq k$ and $1 \leq \ell \leq m$, with $g0_{*,\ell} \equiv \langle g0_{j,\ell} \mid 1 \leq j \leq k\rangle$ and $g1_*[D] \equiv \langle g1_\ell[D] \mid 1 \leq \ell \leq m\rangle$.

Output: Values $g2[D]$, gain estimates for the documents $D \in J \setminus J'$

```
for D ∈ J \ J′ do g2[D] ← 0
γ ← ComputeCV(J′, S)        // compute coefficient of variance
if γ > θ then               // adjust only if γ exceeds threshold
    for ℓ ← 1 to m do
        for D ∈ J′ do g1_ℓ[D] ← 0
        w1_opt ← argmin_{w1∈[0,1]^n} L(h1(g0_{*,ℓ}, w1) | D ∈ J′)
        for D ∈ J do
            g1_ℓ[D] ← h1(g0_{*,ℓ}, w1_opt)
        end for
    end for
    w2_opt ← argmin_{w2∈[0,1]^m} L(h2(g1_*[D], w2) | D ∈ J′)
    // get final per-document estimation
    for D ∈ J \ J′ do
        g2[D] ← h2(g1_*[D], w2_opt)
    end for
end if
return g2
```

rankings, as predicted by the ℓ th of the m different estimators. Prior to forming the new combined estimates, we first compute the coefficient of covariance γ from the judgment set [5], in order to determine whether to use a background "unjudged are not relevant" predictor. Estimation is computed by steps 4 to 15, with $h_1(\cdot)$ and $h_2(\cdot)$ two parametric combining functions, in which the parameters are obtained by minimizing a loss function $L(\cdot)$. We discuss the details of Algorithm 1, including the rationale behind the use of γ, in the next few paragraphs.

First Stage. As noted already, one problem with rank-based estimators is the potential inconsistency across runs of the gain attached to any particular document. As always, we assume that one topic is being addressed; the goal in the first stage is to aggregate the $m \times n$ per-document estimates across the m estimators and n systems into a smaller set of m estimates per document. That is, the m rank-level estimators are treated separately at first, in the loop at step 4, to obtain a consistent background gain for each document D for each model, denoted $g1_\ell[D]$. This is done via a combining function $h_1(\cdot)$ that maps a vector to a single value. Several options for $h_1(\cdot)$ are available, with the choice between them depending on assumptions about system quality and the degree to which the systems are correlated. For simplicity, we assume that the systems are independent and that they vary in quality. Therefore, for each document D, a natural combining function is to compute a weighted average, with h_1 (step 6) parameterized by an n-element weighting vector $\mathbf{w1}$ that is specific to the ℓ th estimator:

$$\begin{gathered} \forall D \in J',\ h_1(g0_{*,\ell}, \mathbf{w1} \mid D) = \sum_{i=1}^{n} g0_{\mathsf{p}_{D,i},\ell} \cdot \mathbf{w1}_i \\ \text{with } \sum_{i=1}^{n} \mathbf{w1}_i = 1 \text{ and } \mathbf{w1}_i \in [0,1]\,, \end{gathered} \tag{6}$$

and where $g0_{\mathsf{p}_{D,i},\ell}$ applies the ℓ th estimator to the rank at which document D appears in the i th of the n runs. One practical issue is that a document may not be retrieved by all systems in their top-k ranked lists, where k is the maximum depth of lists returned. In such cases the rank-based background gain of that document for that system is set to the modeled gain at depth k.

To compute a value for $\mathbf{w1}$, we consider the aggregation process as an optimization problem, where the goal is to minimize the estimation error. The estimation error has two granularities: (i) the total error of system effectiveness score calculated using J'; and (ii) the total error of estimating the background gain of the labeled documents. From either perspective, we can formalize an objective function L and use it at step 6 of Algorithm 1. Consider the first case, with the system matrix as shown in Figure 1. We define L as:

$$L_a(\cdot) = \sqrt{\sum_{i=1}^{n} \left(\sum_{\substack{j=1 \\ s_{j,i} \in J'}}^{k} \Big(W_{\mathrm{M}}(j) \cdot (h_1(\cdot, \mathbf{w1} \mid s_{j,i}) - r_{j,i}) \Big) \right)^2}\,, \tag{7}$$

where $W_{\mathrm{M}}(j) \cdot h_1(\cdot, \mathbf{w1} \mid s_{j,i})$ is the estimated background gain for document $s_{j,i} \in J'$, and $r_{j,i}$ is the known relevance value of that same document. As noted, L_a minimizes the overall estimation error of the evaluation scores for the set of systems.

The second alternative uses the document-position representation $(\mathsf{p}_{D,1}, \mathsf{p}_{D,2}, \ldots, \mathsf{p}_{D,n})$:

$$L_b(\cdot) = \sum_{D \in J'} \sqrt{\sum_{i=1}^{n} \Big(W_{\mathrm{M}}(\mathsf{p}_{D,i}) \cdot (h_1(\cdot, \mathbf{w1} \mid D) - r_D) \Big)^2}\,, \tag{8}$$

in which r_D is the relevance value of document D and is included only once per document, rather than once per document-rank. When compared to Equation 7, which considers estimation errors at the system level, this loss function is focused at the per-document level, seeking to minimize the overall estimation error for the weighted gain of each document. Either Equation 7 or Equation 8 can be used at step 6 of Algorithm 1, with the combination function $h_1(\cdot)$ and constraints defined in Equation 6. The result is the computation of a sequence of $\mathbf{w1}_{\text{opt}}$ vectors, one for each of the m different rank-level estimators.

Second Stage. Multiple fitting models have been proposed because different assumptions about the underlying relevance distributions across all systems are plausible, with a risk that no single model covers the true hypothesis space. Indeed, the limited non-random training data means that we may suffer from a high variance if only one model is considered. Therefore, a "meta" optimizer is also used, combining results from the first stage, as described by steps 11 to 15. A weighted average is used in this role too, considering each document D, together with the estimated background gains generated by the m previous computations, $g1_*[D]$. That combiner, $h_2(\cdot)$ (step 11), is defined via the m-vector $\mathbf{w2}$ as:

$$\begin{gathered} \forall D \in J',\ h_2\left(g1_*[D], \mathbf{w2}\right) = \sum_{\ell=1}^{m} g1_\ell[D] \cdot \mathbf{w2}_\ell\,, \\ \text{with } \sum_{\ell=1}^{m} \mathbf{w2}_\ell = 1 \text{ and } \mathbf{w2}_\ell \in [0,1]\,. \end{gathered} \tag{9}$$

Both L_a and L_b can be used in step 11, but may not necessarily be the same. Note that the m-vector $\mathbf{w2}_{\text{opt}}$, computed at step 11 as the minimizing value for Equation 9, provides an indication of the importance of individual optimizers from the previous stage. Previous work has shown that the expected error of combining loss functions is smaller than the average error on results output by each optimizer in isolation from the first stage [29].

Computing the Coefficient of Variance. The score adjustment and estimation process has been presented on a per-topic basis, with an underlying assumption that a shallow judgment pool cannot identify a majority of the relevant documents. However, some topics may have only a small number of relevant documents, and a shallow depth may be sufficient to identify most of them, with adjustment unnecessary. Only if deeper pooling would identify further relevant documents can score adjustment have an effect on system effectiveness scores. Hence a coefficient of variance [5] is computed for the relevant documents in the shallow pool and used as an indicator, as described in step 2.

Pooling is treated as a sampling with replacement process, with an unknown probability of a relevant document being sampled. Although the final judgment process considers only the documents in the pool, a document returned by multiple systems has a selection

frequency. The intuition behind γ is to make use of that frequency information to describe the sample coverage of relevant documents.

Consider the system matrix S in Figure 1 and a pooling depth d'. Each document $s_{j,i}$ ($1 \le j \le d'$, $1 \le i \le n$) has a multiplicity in $\mathbf{S}_{d' \times n}$; we then group them by that frequency count. Let f_i be the number of relevant documents appearing i times in $\mathbf{S}_{d' \times n}$, $R' = \sum_i f_i$ the number of relevant documents, and $C = \sum_i i \cdot f_i$ be the total occurrence count of relevant documents. For example, if only D_8 and D_1 in Figure 1 are identified as relevant documents, then we have $f_1 = 1$, $f_3 = 1$, and $R' = 2$ and $C = 4$. Based on these elements, the coefficient of variance, γ, is estimated via [5]:

$$\gamma^2 = \max\left\{ \frac{\frac{|R'|}{1-f_1/C} \sum_i i \cdot (i-1) \cdot f_i}{C \cdot (C-1)} - 1, 0 \right\}. \tag{10}$$

When $\gamma = 0$, the probability of sampling a relevant document follows a uniform distribution; and when γ is high, the distribution is skewed, and it is likely that more relevant documents exist due to the low sampling coverage. Based on this, we have two hypotheses:

Hypothesis 1: *γ tends to decrease as pooling depth increases.*

Hypothesis 2: *There is a threshold θ, where if $\gamma < \theta$, then the existence of unjudged documents will only negligibly affect the estimate of the system performance, and they can be ignored.*

The first hypothesis is easy to understand, because increasing the pooling depth increases the sample size, and increases the sampling coverage. The second hypothesis assumes that the score can be dynamically adjusted based on a threshold. If this is correct, then a point at which the total estimation error is minimal can be observed. Otherwise, we must conclude that a shallow pool is not sufficient for finding relevant documents, and adjustment must be applied to all topics in all evaluations.

Discussion. We have described two possible realizations of loss functions, and one option for the combining functions $h_1(\cdot)$ and $h_2(\cdot)$. More sophisticated mechanisms are also possible. For example, the relationship between systems might be leveraged to derive a better $h_1(\cdot)$ and its constraint.

Note also that although our process targets the problem of estimating the effectiveness of runs that contribute to the pool, it is possible to apply the same process to estimate the score of a new system, and is demonstrated empirically in Section 6. Section 6 also shows that the framework can be applied to the judgments constructed using two-strata sampling [24], incorporating the additional information provided in the second stratum.

5 COMPARING SYSTEM RANKINGS

Section 3 already defined ϵ_i, a score-based evaluation criterion. But we are also interested in comparing system orderings as a measure of usefulness of an estimation regime.

Kendall's Distance. This distance metric is widely used to measure the similarity between ranked lists, and counts the number of inverted pairs between two n-item orderings. Let $\sigma_{i,j}$ represent the pairwise relationship between the effectiveness metric means $\bar{S}_i$ and $\bar{S}_j$ of systems S_i and S_j over a set of topics according to one measurement regime, with $\sigma_{i,j} \in \{-1, 0, 1\}$ indicating that $\bar{S}_i < \bar{S}_j$, that $\bar{S}_i = \bar{S}_j$, and that $\bar{S}_i > \bar{S}_j$, respectively; and let $\sigma'_{i,j}$ be the corresponding values for a second measurement regime and the system means that it induces, for example, using pooling to a different depth. Then Kendall's normalized τ distance is the number of pairs $1 \le i < j \le n$ in which $\sigma_{i,j} \cdot \sigma'_{i,j} < 0$, divided by $n(n-1)/2$ to bring it into the range $0 \le \tau \le 1$, with 0 meaning "identical".

Statistical Weighting. Paired t-tests are often used to quantify the strength of the relationship between two systems, and the values $\sigma_{i,j}$ and $\sigma'_{i,j}$ might be thought of as being continuous rather than ternary. Kumar and Vassilvitskii [8] describe a weighted τ distance that counts the strength of each discordant pair, focusing solely on cases where $\sigma_{i,j} \cdot \sigma'_{i,j} < 0$. In practice we are not only interested in the discordant pairs, but also in pairs that are deemed to be significantly different according to one of the measurement regimes but not the other, even if their overall relationship is concordant.

Suppose that $\bar{S}_i > \bar{S}_j$ according to the first measurement, and that a paired one-tail statistical test across topics yields $p_{i,j}$. Values of $p_{i,j}$ near zero indicate a significant superiority of S_i over S_j; values close to 0.5 indicate that it is by chance. If we define

$$\sigma_{i,j} = \begin{cases} 0.5 - p_{i,j} & \text{if } \bar{S}_i > \bar{S}_j \\ 0.0 & \text{if } \bar{S}_i = \bar{S}_j \\ p_{j,i} - 0.5 & \text{if } \bar{S}_i < \bar{S}_j \end{cases},$$

then $-0.5 \le \sigma_{i,j} \le 0.5$ is a real-valued quantity that captures both the direction and strength of the relationship between the two systems according to the first measurement regime. We compute $\sigma'_{i,j}$ similarly using a second measurement approach, and then, to compare the alternative rankings of n systems induced by the two measurement techniques, calculate

$$dist = \sum_{1 \le i < j \le n} \alpha \cdot |\sigma'_{i,j} - \sigma_{i,j}|, \tag{11}$$

where $\alpha \ge 0$ is an additional scaling factor. For example, if $\alpha = |\sigma_{i,j}|$ then the strength of the relationship between S_i and S_j according to the first measurement regime also influences the measured distance. Overall, if $dist \approx 0$, then the two measurement regimes agree in terms of both the direction of each pairwise relationship S_i versus S_j, and also its strength. If *dist* is substantially greater then zero, then the two measurement regimes give rise to many system pairs for which there are non-trivial disagreements (including in both discords and in concords) over the strength of the measured relationships. Compared with Kendall's τ distance, Equation 11 operates over continuous values, which makes it both resistant to inconclusive changes in rank position, and also sensitive to differences in which the direction of the relationship between S_i and S_j stays the same, but the statistical strength varies markedly.

6 EXPERIMENTS

The experiments described in this section include: (i) a post-hoc analysis for testing two hypotheses proposed in Section 4, and setting the threshold θ; (ii) evaluating prediction accuracy using RMSE and Acc% as defined by Lu et al. [12]; (iii) system ordering stability evaluation using the distance metric defined in Equation 11 with $\alpha = 1$, and using normalized τ distance; and (iv) a case study covering the ClueWeb 2010 (CW10) task.

Dataset	d	$\|S\|$	Judgments per topic				2-strata
			$d' = 10$	$d' = 20$	$d' = 30$	$d' = d$	
TREC5	100	76	272 (13)	512 (10)	747 (8)	2298 (4)	–
TREC9	100	59	174 (11)	322 (8)	462 (7)	1382 (4)	294 (7)
TREC10	100	54	182 (13)	335 (10)	480 (9)	1402 (5)	303 (9)
Rob04	100	42	75 (25)	139 (18)	206 (15)	710 (7)	134 (15)
TB04	80	33	164 (31)	313 (27)	453 (25)	1121 (19)	270 (25)
TB05	100	34	111 (41)	202 (36)	291 (33)	878 (25)	187 (33)
TB06	50	39	141 (31)	270 (26)	394 (23)	633 (19)	–
CW10	20	21	98 (30)	–	–	187 (28)	–

Table 1: Datasets used: d is the original pooling depth and provides the reference point for metric scores; d' is a notional pooling depth used our experimentation; and $|S|$ is the number of contributing runs. Only Adhoc Task runs are used. The middle four column pairs show the number of judgments averaged across topics at each pooling depth d', and the percentage of relevant documents. The last column shows the statistics when using two-strata sampling [24], averaged over topics and over ten random iterations.

Experimental Setup. The collections and configuration parameters used in our experiments are shown in Table 1. We also measured a range of behavior using the TREC7 and TREC8 collections, but do not include them here because those two collections were used as part of the post-hoc analysis and parameter setting. Scripts are available to reproduce all of the various results given here[1].

Pooling to different depths is simulated using the identified contributing systems, and the average number of judgments required per topic at different pool depths is also shown in Table 1, together with the corresponding percentage of documents identified as being relevant. In the experiments measuring rank stability, we also examine the two-strata sampling method described by Voorhees, and averages over ten runs for this randomized approach are included in the table. For the Robust04 task the last 49 topics are used, and judged to a depth of 100; for other tasks, we use all of the original topic set and judgments. Our goal in collection selection was to capture as much variety as possible. TREC5 and Rob04 use the NewsWire document collection, TREC9 and TREC10 use WT10G, a small web collection, and TB04/05/06 use the GOV2 web collection. The ClueWeb 2010 task (CW10) uses the largest web collection but also has fewer contributing systems and a shallow pool depth. It is representative of newer collections, which are large, and have more uncertainty associated with the judgment coverage – the core issue which motivated our investigation. We show results for this dataset as a practical application of our work, noting that a pooling depth of $d = 20$ cannot provide a ground truth for a deep metric [11].

We use RBP with $p = 0.95$ for training and for all testing, as a representative weighted-precision metric. RBP supports graded relevance (needed to make use of the estimated background gains we generate); allows residuals to be computed; and with $p = 0.95$ gives similar system orderings to AP and NDCG [15]. The estimated background gain of each document generated via training using RBP0.95 can also be used to compute other weighted-precision measures, such as the truncated metric $\text{SDCG}_{@k}$ when $k > d$.

[1] https://github.com/xiaolul/opt_est

We consider five methods for predicting effectiveness scores, three of which are baselines. The first baseline is the lower bound, LB, which assumes unjudged documents are not relevant; the second is the interpolative estimator of Ravana and Moffat [16] (Equation 4), denoted RM; and the third is the linear model Lin. that is the best of the rank-based approaches described by Lu et al. [12]. They are compared to the loss functions defined in Equations 7 and 8, denoted L_a and L_b respectively, with the same loss function used in both stages, and aggregation via Equations 6 and 9.

We use Linear, Zipf and Discrete Weibull models as initial rank-based estimators [12], and hence have $m = 3$. Two experiments explore rank stability, categorized by how the judgment set is constructed: (i) pooling based judgments; and (ii) two-strata sampling based judgments. Rank stability is measured using the approaches discussed in Section 5. The same baselines are used in the first rank stability evaluation. However, for the sample-based judgments, we consider the metrics InfRBP ($p = 0.95$) defined by Equation 5, and Yilmaz and Aslam's InfAP [27] as baselines. Throughout the experiments, the system scores (plus residuals) and system orderings computed using the same metric, but evaluated at the full pool depth (that is, at $k = d$), are taken as the "gold standard". The truncated metric $\text{AP}_{@d}$ is computed as described by Sakai [21].

Setting θ. We first test the two hypotheses in Section 4, with γ in Equation 10 normalized by the number of systems. Average (over topics) γ values are plotted against pool depth in the left-hand plot in Figure 3, showing that γ decreases as the pooling depth increases. This is as expected, since the increasing pooling depth results in a more complete judgment set. Among the plotted datasets the TB06 collection has the largest average γ, and corresponds to a high relevance rate (Table 1). TREC5 is a relatively complete test collection, and hence has the lowest γ among the datasets plotted.

The center pane in Figure 3 shows the distribution of γ across topics for $d' = 10$. Although γ is usually low for TREC5, there are still some topics that have high values. The same pattern is also observable for Rob04 and TREC10. Based on our hypothesis, this observation indicates that, on earlier TREC collections, not all topics necessarily require score adjustment even at $d' = 10$.

To set θ we use the earlier datasets TREC5, TREC7 and TREC8 and perform a post-hoc analysis, noting that the majority of relevant documents have been identified in these collections, and hence that the computed RMSE should be close to the true error. The right-hand pane in Figure 3 shows TREC5 outcomes, with three rank-based models plotted. Weibull (Wei.) may be an overestimate due to the shaping parameter, and Linear (Lin.) tends to provide low estimates due to the monotonically decreasing nature of the model [12]. At first, neither of the score estimation methods works better than the lower bound LB, but as θ increases, fewer topics need to be estimated, and when $\theta = 0.018$, both estimation methods outperform LB. Similar cross-overs occur for TREC7 and TREC8.

Prediction Accuracy. We then employed $\theta = 0.018$ for the other datasets, obtaining the results shown in Table 2. When $\theta = 0$ the L_b method outperforms all three baselines (LB, RM and Lin.) in terms of RMSE and Acc%, while the L_a approach has a higher RMSE than L_b and on earlier datasets (TREC9, TREC10) is slightly worse than the LB and Lin. baselines. That is, the loss function L_a provides poorer coverage of the true hypothesis space than does L_b. The RM

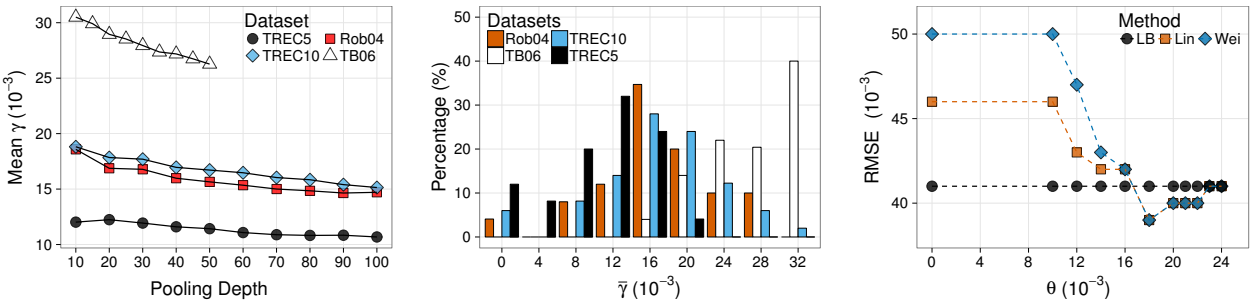

Figure 3: Left: γ relative to pooling depth d'. Middle: distribution of γ per topic when $d' = 10$. Right: impact of the threshold on the training set TREC5, with $d' = 10$.

Dataset	d'	LB	RM	Lin.		L_a		L_b	
				$\theta = 0$	$\theta = 0.018$	$\theta = 0$	$\theta = 0.018$	$\theta = 0$	$\theta = 0.018$
	10	0.031 (45)	0.056 (22)	0.037 (31)	0.031 (44)	0.038 (26)	**0.030** (44)	0.031 (41)	0.031 (**46**)
TREC9	20	0.012 (59)	0.025 (32)	0.013 (51)	0.012 (60)	0.013 (42)	0.012 (58)	**0.010** (**62**)	0.012 (61)
	30	0.006 (67)	0.012 (45)	0.006 (66)	0.006 (68)	**0.005** (63)	0.006 (68)	**0.005** (**71**)	0.006 (68)
	10	0.038 (**39**)	0.064 (16)	0.034 (25)	0.033 (31)	0.036 (13)	0.031 (27)	**0.027** (34)	0.028 (38)
TREC10	20	0.016 (53)	0.030 (25)	0.015 (45)	0.014 (51)	0.016 (36)	0.014 (50)	**0.012** (53)	**0.012** (**55**)
	30	0.007 (64)	0.015 (37)	0.007 (61)	0.007 (63)	0.007 (55)	0.007 (62)	**0.006** (**66**)	**0.006** (**66**)
	10	0.046 (**21**)	0.088 (5)	0.043 (**21**)	0.039 (20)	0.045 (9)	0.039 (17)	0.039 (20)	**0.035** (20)
Rob04	20	0.020 (34)	0.040 (9)	0.015 (33)	0.016 (34)	0.016 (26)	0.015 (32)	**0.013** (**38**)	0.016 (35)
	30	0.008 (49)	0.020 (14)	0.007 (49)	0.007 (52)	0.007 (47)	0.006 (53)	**0.005** (**59**)	0.006 (55)
	10	0.117 (14)	0.082 (14)	0.082 (15)	0.087 (14)	**0.072** (15)	0.077 (15)	0.073 (**16**)	0.077 (**16**)
TB04	20	0.053 (21)	0.039 (23)	0.039 (25)	0.041 (25)	0.035 (28)	0.037 (28)	**0.033** (**32**)	0.036 (31)
	30	0.026 (26)	0.020 (39)	0.018 (39)	0.019 (38)	**0.015** (**45**)	0.016 (44)	**0.015** (44)	0.016 (43)
	10	0.125 (6)	0.080 (5)	0.085 (7)	0.085 (7)	0.070 (6)	0.070 (8)	**0.067** (7)	**0.067** (**9**)
TB05	20	0.056 (10)	0.041 (10)	0.039 (13)	0.039 (13)	0.034 (14)	0.034 (14)	**0.033** (18)	**0.033** (**19**)
	30	0.028 (16)	0.022 (18)	0.021 (24)	0.021 (24)	0.018 (24)	0.018 (24)	**0.017** (**29**)	**0.017** (**29**)
	10	0.089 (24)	0.065 (43)	0.059 (43)	0.059 (43)	**0.047** (**55**)	**0.047** (**55**)	0.053 (51)	0.053 (50)
TB06	20	0.033 (40)	0.023 (68)	0.021 (66)	0.021 (66)	**0.013** (**81**)	**0.013** (**81**)	0.017 (73)	0.017 (73)
	30	0.013 (58)	0.007 (87)	0.006 (85)	0.006 (85)	**0.003** (**94**)	**0.003** (**94**)	0.005 (89)	0.005 (89)

Table 2: RMSE and Acc% scores for RBP0.95 for all estimation methods, with d' the depth of the reduced pool, and the reference depth d of each dataset as listed in Table 1. Bold numbers are the lowest RMSE and highest Acc% for that collection at that depth.

approach performs poorly on all of the earlier datasets, for which the assumption that unjudged documents are equivalent to judged ones is inappropriate. On the larger collections such as TB04/05/06, the gain decreases at a slower rate, making the assumptions in RM more appropriate. The LB approach has similar issues, seen in the TB04/05/06 collections. However, for TB06, smaller RMSE (and larger Acc%) values are achieved when compared to the other collections. This is because the reference depth $d = 50$ is smaller, resulting in larger residuals. As shown in Figure 3, some of the topics may not necessarily require a score adjustment process, especially in the earlier test collections. This explains why the LB estimator works well on those collections. As expected, applying a threshold θ improves the estimation for both L_a and for the Lin. model, on TREC9, TREC10 and Rob04 test collections. Unsurprisingly, on TB04/05/06, only minor score changes are observed when $\theta = 0.018$ is used, because the computed γ values are larger than the threshold, indicating low coverage of the relevant documents identified. The only unexpected observation occurs on the TB04 test collection, where the threshold falsely identifies Topic 734 as having a "sufficient" sampling of relevant documents, but around 48% in the final judged set are relevant, which increases the RMSE value. Table 3 shows the results for a leave-one-group-out experiment at $d' = 10$ (with $\theta = 0$), demonstrating the applicability of the framework in adjusting for both system and pooling depth bias.

System Ordering Stability on Pooling-Based Judgments. The system orderings derived from the score estimates when compared against the orderings at the reference depth of $k = d$ are shown in Figure 4. Kendall's τ correlation was also computed, but the closely-related τ distance is used here since it has a strictly positive value. In the first row, when normalized τ distance is measured, the estimation framework gives orderings close to the reference

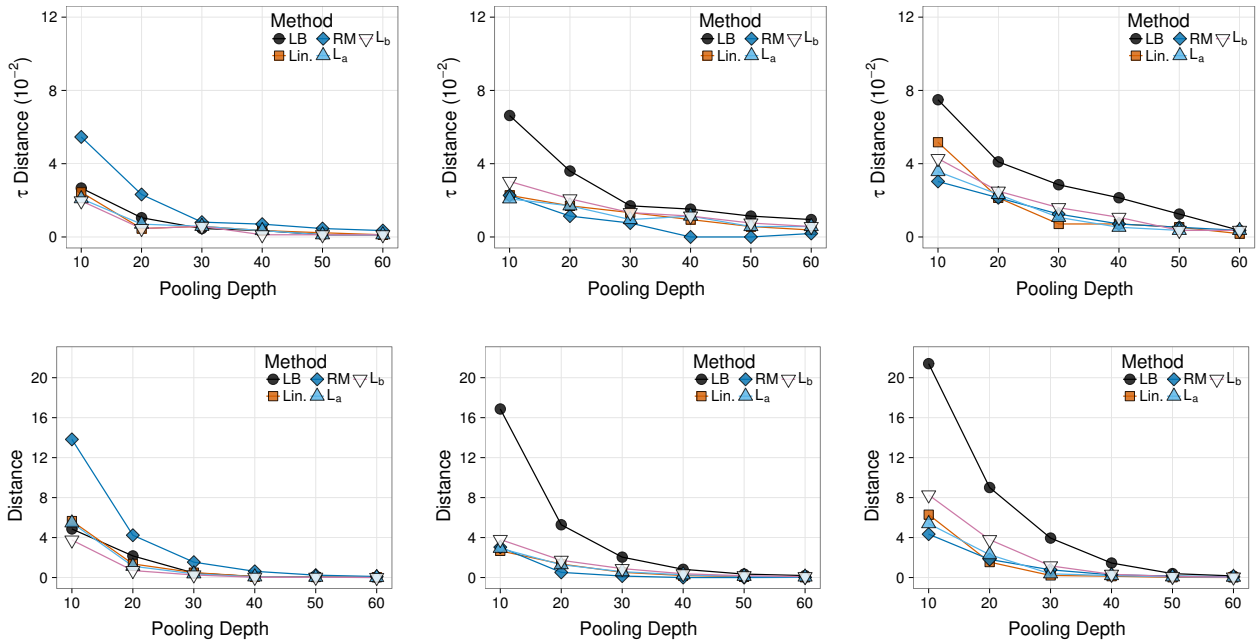

Figure 4: System ordering comparisons (RBP0.95) for five estimators. The first row uses normalized τ distance; the second row uses *dist* (Equation 11). The columns (from left) show Rob04, TB04, and TB05, with reference lists using LB at $d = 100$, $d = 80$ and $d = 100$, respectively.

Dataset	LB	RM	Lin.	L_a	L_b
Rob04	0.060 (**19**)	0.126 (4)	0.068 (11)	0.060 (9)	**0.050** (**19**)
TB04	0.181 (**11**)	0.202 (**11**)	0.131 (9)	**0.117** (**11**)	0.119 (**11**)
TB05	0.170 (**6**)	0.141 (5)	0.125 (4)	**0.110** (4)	**0.110** (5)
TB06	0.125 (22)	0.185 (32)	0.112 (35)	0.090 (**48**)	**0.086** (46)

Table 3: RMSE and Acc% for leave-out-one-group experiments with $d' = 10$ throughout, averages across groups assuming that each group in turn is omitted from pool construction (RBP0.95).

ordering across a range of nominal pool depths d'. The RM approach performs well on TB04/05, agreeing with the results in Table 2. However, as noted above, τ is sensitive to swaps that might be inconclusive. The bottom row of Figure 4 shows the *dist* measure of Equation 11. Overall, there are situations in which LB performs poorly, and situations in which RM performs poorly. The Lin., L_a, and L_b methods consistently provide the highest agreements.

We also carried out paired t-tests and calculated the discrimination ratio for a significance level $p = 0.05$, and compared against the original discrimination ratios. The Lin., L_a, and L_b estimation methods used with J' all have only a small effect on discrimination ratio when compared to the use of LB and J.

System Ordering Stability on Sample-Based Judgments. We also show the applicability of our methods on the judgment set constructed using a two-strata sampling method [24], which has been empirically shown to assist when computing inferred metrics. On this set of judgments we compute InfAP using trec_eval, and InfRBP as defined in Equation 5. Figure 6 shows that L_a, L_b and InfRBP give rise to stable system orderings, with normalized τ

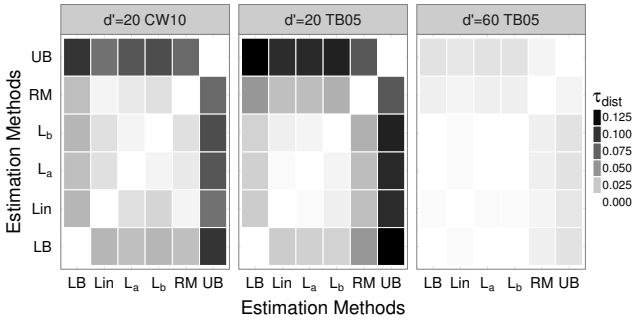

Figure 5: Normalized τ distance between system orderings generated by different estimation methods based on a pool of depth $d' = 20$, and on TB05 based on pool depths of $d' = 20$ and $d' = 60$.

distance scores below 0.05 across all collections. When *dist* is measured, L_a outperforms InfRBP on all collections but TREC9, while L_b outperforms InfRBP except on TB05. The slightly worse outcome for L_a on TREC9 is a consequence of the increase in the number of significantly different system pairs. Note the more variable outcomes generated when InfAP is used as the metric driving the system orderings.

Predictions in ClueWeb. As a final test of our approach, we examine the CW10 collection. It has a shallow pool depth ($d = 20$), meaning that validation is not possible, as there is no deep-pool reference ordering. Instead, we compute the normalized τ distance between each pair of estimation methods, and simply record how much the rankings differ, as shown in Figure 5. The UB estimator assumes that all unjudged documents are relevant. As a reference

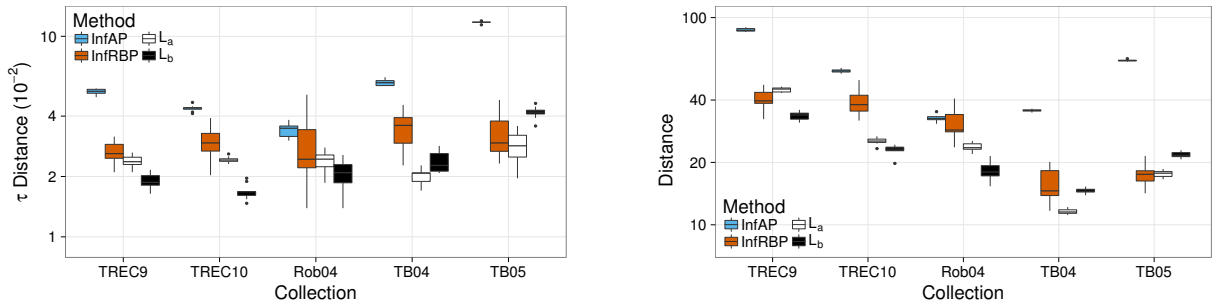

Figure 6: System ordering comparisons on a two-strata sampled judgment set, repeated ten times. Judgments are to depth $d' = 10$, plus a 10% random sample of remaining documents to depth 100 to form the second stratum. Note the logarithmic vertical scales.

point, we also compute the same values for TB05, at two depths, $d' = 20$ and $d' = 60$. At the latter depth all estimation approaches tend to agree with each other. On TB05, all of the estimation results, including UB, tend to agree on the system ordering. However, on CW10, there is clear uncertainty, confirming that $d = 60$ is a more robust pool depth for TB05 than is $d = 20$ on either TB05 or CW10 when seeking to apply RBP0.95 as an evaluation metric. Great caution should be exercised when the $d = 20$ CW10 judgments are used for anything other than shallow metrics.

7 CONCLUSIONS

We have presented new methods to improve system comparisons in batch IR evaluation, with the key idea being to predict a gain value for each unjudged document. We show that estimation is a viable technique to predict scores for deep evaluation metrics when limited judgments are available, including the case when the judgments are obtained using stratified sampling rather than pooling. One important aspect of our approach is to make decisions on *when* to adjust topics, instead of treating all topics equally.

A secondary contribution is the development of a new technique to more precisely compare system orderings. By focusing on swaps that are conclusive, our weighted rank correlation coefficient *dist* can be used to measure the stability of a variety of estimation techniques. Using *dist*, we show that estimation improves our ability to score and compare systems using limited judgments.

It must be noted, however, that the estimation is built on the m rank-based fitted models, each of which requires that when constructing the judgment set, documents up to some rank d' be fully judged. This means that for some sampling-based judgment approaches, the proposed method is not applicable. Second, while we show that our estimation methods can also account for system bias to some extent, outcomes might be further improved by introducing more randomization into the optimization framework. Hence, in answer to the question posed in the title, our answer remains a somewhat cautious "better than before", rather than a "yes".

Funding. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP140103256 and DP170102231).

REFERENCES

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. 2006. A statistical method for system evaluation using incomplete judgments. In *Proc. SIGIR*. 541–548.
[2] C. Buckley, D. Dimmick, I. Soboroff, and E. M. Voorhees. 2007. Bias and the limits of pooling for large collections. *Inf. Retr.* 10, 6 (2007), 491–508.
[3] C. Buckley and E. M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proc. SIGIR*. 25–32.
[4] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. 2007. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proc. SIGIR*. 63–70.
[5] A. Chao and S. Lee. 1992. Estimating the number of classes via sample coverage. *J. American Statistical Association* 87, 417 (1992), 210–217.
[6] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proc. CIKM*. 621–630.
[7] J. K. Jayasinghe, W. Webber, M. Sanderson, and J. S. Culpepper. 2014. Improving test collection pools with machine learning. In *Proc. Aust. Doc. Comp. Symp.* 2–9.
[8] R. Kumar and S. Vassilvitskii. 2010. Generalized distances between rankings. In *Proc. WWW*. 571–580.
[9] A. Lipani, M. Lupu, and A. Hanbury. 2015. Splitting water: Precision and anti-precision to reduce pool bias. In *Proc. SIGIR*. 103–112.
[10] A. Lipani, M. Lupu, E. Kanoulas, and A. Hanbury. 2016. The solitude of relevant documents in the pool. In *Proc. CIKM*. 1989–1992.
[11] X. Lu, A. Moffat, and J. S. Culpepper. 2016. The effect of pooling and evaluation depth on IR metrics. *Inf. Retr.* 19, 4 (2016), 416–445.
[12] X. Lu, A. Moffat, and J. S. Culpepper. 2016. Modeling relevance as a function of retrieval rank. In *Proc. AIRS*. 3–15.
[13] A. Moffat, P. Thomas, and F. Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*. 659–668.
[14] A. Moffat, W. Webber, and J. Zobel. 2007. Strategic system comparisons via targeted relevance judgments. In *Proc. SIGIR*. 375–382.
[15] A. Moffat and J. Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Information Systems* 27, 1 (2008), 2:1–2:27.
[16] S. D. Ravana and A. Moffat. 2010. Score estimation, incomplete judgments, and significance testing in IR evaluation. In *Proc. AIRS*. 97–109.
[17] S. E. Robertson. 2007. On document populations and measures of IR effectiveness. In *Proc. ICTIR*. 9–22.
[18] T. Sakai. 2007. Alternatives to BPref. In *Proc. SIGIR*. 71–78.
[19] T. Sakai. 2008. Comparing metrics across TREC and NTCIR: The robustness to pool depth bias. In *Proc. SIGIR*. 691–692.
[20] T. Sakai. 2008. Comparing metrics across TREC and NTCIR: The robustness to system bias. In *Proc. CIKM*. 581–590.
[21] T. Sakai. 2014. Metrics, statistics, tests. In *Bridging Between Information Retrieval and Databases*, N. Ferro (Ed.). Springer, 116–163.
[22] T. Schnabel, A. Swaminathan, P. I. Frazier, and T. Joachims. 2016. Unbiased comparative evaluation of ranking functions. In *Proc. ICTIR*. 109–118.
[23] T. Schnabel, A. Swaminathan, and T. Joachims. 2015. Unbiased ranking evaluation on a budget. In *Proc. WWW*. 935–937.
[24] E. M. Voorhees. 2014. The effect of sampling strategy on inferred measures. In *Proc. SIGIR*. 1119–1122.
[25] E. M. Voorhees and D. K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval.* The MIT Press.
[26] W. Webber and L. A. F. Park. 2009. Score adjustment for correction of pooling bias. In *Proc. SIGIR*. 444–451.
[27] E. Yilmaz and J. A. Aslam. 2008. Estimating average precision when judgments are incomplete. *Knowledge and Information Systems* 16, 2 (2008), 173–211.
[28] E. Yilmaz, E. Kanoulas, and J. A. Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In *Proc. SIGIR*. 603–610.
[29] Z. Zhou. 2012. *Ensemble Methods: Foundations and Algorithms.* CRC press.
[30] J. Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proc. SIGIR*. 307–314.