Modeling Relevance as a Function of Retrieval Rank

Xiaolu Lu[1], Alistair Moffat[2], and J. Shane Culpepper[1]

[1] RMIT University, Melbourne, Australia,
[2] The University of Melbourne, Melbourne, Australia

Abstract. Batched evaluations in IR experiments are commonly built using relevance judgments formed over a sampled pool of documents. However, judgment coverage tends to be incomplete relative to the metrics being used to compute effectiveness, since collection size often makes it financially impractical to judge every document. As a result, a considerable body of work has arisen exploring the question of how to fairly compare systems in the face of unjudged documents. Here we consider the same problem from another perspective, and investigate the relationship between relevance likelihood and retrieval rank, seeking to identify plausible methods for estimating document relevance and hence computing an inferred gain. A range of models are fitted against two typical TREC datasets, and evaluated both in terms of their goodness of fit relative to the full set of known relevance judgments, and also in terms of their predictive ability when shallower initial pools are presumed, and extrapolated metric scores are computed based on models developed from those shallow pools.

1 Introduction

A comprehensive set of judged documents derived from human relevance assessments is a key component in the successful evaluation of IR systems. However, growing collection sizes make it prohibitively expensive to judge all of the documents that are potentially relevant, and sampling methods such as *pooling* [15] are now commonly used to select a subset of documents to be judged. Partial judgments present an interesting challenge in carrying out reliable evaluation, and can result in subtle problems when comparing the quality of two or more systems.

The main issue arising from partial judgments is how to handle unjudged documents during evaluation. One simple rule – and the one often used in practice – is to assume that all unjudged documents are non-relevant. Although an evaluation score can be obtained using this assumption, any conclusions drawn may be a biased view of a system's relative performance. Two approaches to handling these issues have been proposed: metric-based solutions [1, 3, 5, 9, 11, 17, 18], and score adjustment [7, 10, 16]. Metric-based solutions can be further categorized as those that ignore the unjudged documents, and work only with the known documents; and those that attempt to infer the total relevance gain achieved by the system, or, at least, to quantify the extent of the uncertainty in the measured scores. Score adjustment approaches require a different type of collection pooling process, which can greatly impact the reusability of the test collection. They also seek to minimize the bias between the pooled and unpooled systems, which is different than the pooling depth bias. Pooling depth bias can occur in contributing systems as well as new systems since using a pooling depth less than the evaluation depth can result in unjudged documents occurring in any system ranking.

Here we consider traditionally pooled collections, and consider the problem from a fresh angle: *does the rank position of a previously unseen document influence the likelihood of it being relevant, and if so, can that relationship be exploited to allow more accurate system scores to be computed?* Our estimations of gain based on rank fit well with weighted-precision metrics, and allow both types of bias to be incorporated when performing evaluations. In particular, we measure the aptness of several possible models that build on existing judgments, from which we obtain an observed likelihood of relevance at different ranks. The benefit of assessing relevance as a function of rank is that the model can be applied both within the original pooling depth and also beyond it. A further advantage of the proposed approach is that in making the model topic-specific, it automatically adapts to differing numbers of relevant documents and to query difficulty, both of which can vary greatly across topics.

As a specific example of how our techniques might be employed, we consider the rank-biased precision (RBP) metric [9], which computes a *residual* as a quantification of the net metric weight associated with the unjudged documents in a ranking. Using an estimator, a value within that identified residual range can also be computed, and given as a proposed "best guess" score. To demonstrate the validity of our proposal, empirical studies are conducted on two representative TREC datasets: those associated with the 2004 Robust Track; and with the 2006 Terabyte Track. The first collection is believed to be relatively complete [13], while the second is understood to be less comprehensive [8, 12]. The proposed models are fitted using topics in the two datasets and compared using a standard goodness-of-fit criterion at different nominal pooling depths. We then explore the predictive power of those models, by comparing extrapolated system scores generated from shallow-depth pools with the corresponding scores computed using deeper pools.

2 Background

Batch IR evaluations require a set of judgments for each included topic. *Pooling* [15] is often used to generate those judgments, but has limitations, since there is no guarantee that all relevant documents for a topic are identified. The usual way of handling that problem during evaluations is to assume that unjudged documents are not relevant. Incomplete judgments have been shown to have little effect in the NewsWire collections [19], but the evaluation results in larger web collections can be biased [2]. As a result, several strategies for dealing with unknown documents have been developed [1, 3, 5, 7, 9, 10, 11, 16, 17, 18]. Broadly speaking, these strategies can be categorized into two types – metrics that deal in some way with the missing judgments, and methods for adjusting the bias. Figure 1 provides a taxonomy of approaches, which we now explore.

Metrics for Incomplete Judgments Widely used metrics such as AP and NDCG [6] were developed on the assumption that the judgments were complete. When they are used with incomplete judgments, unjudged documents are typically assumed to be non-relevant during the calculation process, an assumption that can result in underestimating the effectiveness of a system if it returns many unjudged documents, or overestimating

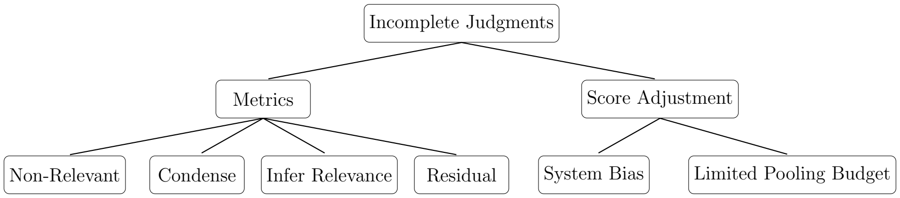

Fig. 1: A taxonomy of approaches for minimizing the effects of unjudged documents on system evaluation.

the effectiveness of all systems if there are many undetected relevant documents. Alternative approaches have been proposed that use only the documents which are judged, including condensed scoring [11, 17], and BPref [3]. Sakai [11] compared different condensed metrics with BPref and concluded that condensed Q-measure and NDCG work well in practice, and have a higher discriminative power than BPref.

In a quest to make better use of both judged and unjudged documents, metrics using inference [17, 18] have also been proposed. For example, InfAP [17] estimates the precision at ranks where relevant documents occur, and assumes that relevant documents are distributed uniformly between identified ranks. A drawback is that inferred metrics depend on pools being constructed using a predefined sampling method. A recent study by Voorhees [14] concluded that a two-strata sampling is a suitable method for constructing collections for inferred metrics.

The metric StatAP [1] embeds another approach to sampling based estimation by deploying importance sampling when judgment pools are created in order to minimize the likelihood of missing relevant documents. StatAP estimates precision based on a joint distribution derived from the relevance probability of a pair of ranks. The total number of relevant documents is estimated via a uniform sampling process over a depth 100 pool. Combining both estimates produces the final StatAP score. Both InfAP and StatAP have been shown to be highly correlated with AP when the judgments are incomplete, using a range of collections [17, 18]. However, inferred metrics and StatAP are reliant on specific sampling strategies being followed when pool construction occurs, meaning that applying these methods on unpooled systems may not be appropriate.

The final metric-based approach is to provide both the minimum and the maximum effectiveness score for a system, using the notion of a *residual* that was introduced alongside Rank-Biased Precision (RBP) [9]. Instead of generating a point effectiveness score, RBP provides a lower and an upper bound, with the gap between them representing the extent of score uncertainty associated with the unjudged documents. RBP supports traditional score-based system comparisons, and also provides quantitative evidence of the potential impact the unjudged documents may have on that comparison.

Score Adjustments Based on Estimated Relevance The alternative is to try and adjust for the bias. The first option is to compensate for system bias – the difference between pooled and unpooled systems when using a fixed pool depth – using either a metric-based approach [7, 10, 16] or a metric-independent approach [5]. Based on RBP@10, Webber and Park [16] propose adjustment methods to deal with the inference from systems and from topics. In separate work, Ravana and Moffat [10] propose estimation schemes for picking a point within the RBP residual range: a background method; an interpolation method; and a smoothing method that blends the first two. Although Ravana and Moffat primarily focus on system bias, their results also indicate that the same approaches could be applied to adjust the bias resulting from a limited pooling budget.

Recent work by Lipani et al. [7] views the problem from another perspective, proposing an "anti-precision" measure in order to determine when to correct the pooling bias. By using a Monte Carlo method to estimate the adjustment score to be added to a run, Lipani et al. empirically obtain better results than previous work. Lastly, Büttcher et al. [5] consider the problem independent of the evaluation metric. By transforming bias adjustment into a document classification problem, the relevance of a document can be predicted to minimize rank variance when a leave-one-out experiment is applied.

Most of this prior work has focused on adjusting the bias between pooled and unpooled systems. When the pooling budget is limited, condensed runs and BPref may be vulnerable to relatively high score variance. Residual-enabled metrics such as RBP at least allow this variance to be quantified, but do not necessarily provide any way of drawing useful conclusions. Sampling methods and inferred metrics may be of some benefit in this regard, but give rise to different issues when systems not contributing to the original pool are to be scored. It is this set of trade-offs that motivates us to revisit the question of system comparisons in the face of a limited pooling budget.

3 Models and Analysis

We now describe methods for modeling relevance as a function of ranking depth.

Gain Models Consider a weighted-precision metric such as RBP, which is computed as $\sum_{i=1}^{\infty} W(i) \cdot r_i$, where $W(i)$ is the ranking-independent weight attached to the item at rank i according to the metric definition, and r_i is the gain associated with that i th item in the ranking generated for the topic in question. When the judgments are incomplete, and the value r_j is not known for one or more ranks j, we propose that an estimated gain $\hat{r}_j$ be used, where $\hat{r}_j$ is computed via a model of relevance in which topic and retrieval rank j are the inputs.

Focusing on a single topic, we let $\langle r_{k,n} \rangle$ be a *gain matrix* spanning n systems that have contributed to a pooled evaluation to a maximum run length (or evaluation depth) of $k = d$, so that $r_{i,s}$ is the gain attributed to system s by the document it placed at rank i. The *empirical gain* vector $\mathbf{g} = \langle g_1, g_2, \ldots, g_k \rangle$ is then:

$$g_i = \frac{1}{n} \sum_{j=1}^{n} r_{i,j} \; . \qquad (1)$$

A *gain model* is a function $G(\mathbf{g}, k)$ that generates a value $\hat{g}_k$ as an approximation for g_k, the empirical gain at rank k. For example, one simple gain model is to assert that if a document is unjudged its predicted gain is minimal, that is, $G_0(\mathbf{g}, k) = \mathit{mingain}$, where *mingain* is the lower limit to the gain range and is usually zero. This is the pessimal approach to dealing with unjudged documents that was discussed in Section 2. Similarly, the residuals associated with RBP combine $G_0()$ at one extreme, and

Model	Description	Parameters	Assumptions
G_s	$(\mathit{maxgain} - \mathit{mingain})/2$	–	Static, constant across all ranks
G_c	$\begin{cases} \lambda_0 & 1 \le k \le m \\ 0 & k > m \end{cases}$	λ_0, m	Constant until rank m, zero thereafter
G_ℓ	$\max\{-\lambda_0 \cdot k + c, 0\}$	$\lambda_0 \ge 0,\ c$	Linear, decreasing until rank m, zero thereafter
G_z	$\lambda_0/(k^c \cdot H_{n,c})$	$\lambda_0, c \ge 0$	Zipfian, monotonic decreasing, never zero
G_w	$\lambda_0 \cdot \Big((1-\lambda_1)^{(k-1)^c} - (1-\lambda_1)^{k^c}\Big)$	$\lambda_1 \in [0,1]$, $c > 0, \lambda_0$	Weibull, might increase before decreasing, never zero

Table 1: Five possible gain models, where $k \ge 1$ is the rank, and "Parameters" lists the free parameters in the estimated model.

$G_1(\mathbf{g}, k) = \mathit{maxgain}$ at the other, where *maxgain* is the upper limit to the gain range, and is often (but not necessarily always) one.

Increasingly Flexible Models We are interested in gain models that lie between the extremes of $G_0()$ and $G_1()$, and consider five different interpolation functions in our evaluation, embodying different assumptions as to how gain varies according to rank. Table 1 lists the five options. The first model listed, $G_s()$, assumes that the gain is static and both topic and rank invariant. For early ranks this is perhaps more realistic than using G_0 or G_1, but is intuitively implausible for large ranks, since the goal of any retrieval system is to bring the relevant documents to the top of the ranking.

The second model is a truncated constant model, G_c, which is predicated on the assumption that all relevant documents appear in a random manner at the early ranks of each run, and that beyond some cutoff rank m, no further relevance gain occurs. This model is rank-sensitive in a binary sense, and because m is a parameter that is selected in the context of a particular topic, it is also topic-sensitive. That is, the constant model G_c adds a level of flexibility to the static $G_s()$, and while it may also be implausible to assert that average gain is a two-valued phenomena determined by rank for any individual topic, in aggregate over a set of topics, each with a fitted value of m, the desired overall behavior might emerge.

The third step in this evolution is the model G_ℓ. The constant model G_c allows an abrupt change in predicted gain as a function of rank, at the topic-dependent cutoff value m. If we add further flexibility and suppose that average relevance gain decreases linearly as ranks increase, rather than abruptly, we get G_ℓ. This model also has cutoff rank m beyond which the expected gain from an unjudged document is presumed to be zero, given by $m = \lceil c/\lambda_0 \rceil$. A fourth option is to allow a tapered decrease, and this is what G_z achieves, via the Zipfian distribution, in which $H_{n,c}$ is a normalizing constant determined by the controlling parameter c and the ranking length n. The expected gain rate decreases at deeper pooling depths but remains non-zero throughout, due to the long-thin tailed property of the Zipfian distribution.

Another possibility is that the gain may initially increase or be constant, and then decrease in the longer term. To achieve this option, the monotonicity expectation is relaxed, a possibility captured by the discrete Weibull distribution, model G_w. Note that this function allows the possibility of an initial increase, but does not make that mandatory. In particular, when $c = 1$, the underlying distribution becomes a simple decreasing geometric distribution. Since this model is derived from a discrete Weibull distribution, the gain rate decreases faster than G_z when the distribution of relevance by rank is similar.

Given a model G that has been determined in response to a empirical gain vector $\mathbf{g}$, we take $\hat{r}_j = G(\mathbf{g}, j)$ for unjudged documents when r_j is unavailable, and then compute a weighted-precision metric such as RBP in exactly the same manner as before. That is, the estimated gain for that topic is used whenever the actual gain is unknown.

Measuring Model Fit With a choice of ways in which relevance might be modeled, an obvious question is how to compare them and identify which ones provide the most accurate matches to actual ranking data. To measure goodness-of-fit we use root-mean-squared-error, or RMSE. That is, given a model G fitted to an empirical gain vector $\mathbf{g} = \langle g_j \rangle$ by choosing values for the controlling parameters (Table 1), we compute

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(G(\mathbf{g}, j) - g_j\right)^2}$$

as an indicator of how well that model and those parameters fit the underlying distribution. Small values of this measure – ideally, close to zero – will indicate that the corresponding model is a good estimator of the underlying observed behavior.

Measuring Model Predictive Power A second important attribute of any model is its ability to be predictive over unseen data, that is, its ability to be used as a basis for extrapolation. In particular, we wish to know if a model fitted to an empirical gain vector computed using judgments to some depth d' (training data) can then be used to predict system scores in an evaluation to some greater depth $d > d'$. Figure 2 illustrates this notion. Suppose that pooled relevance judgments to depth $d' = 10$ are available. If a weighted-precision metric such as RBP is used at an evaluation depth $k = 10$, all required judgments are available, but even so, there is a still a non-zero score range, or residual. That $d' = 10$ score range is illustrated in Figure 2 by the solid lines, plotted as a function of k, the evaluation depth. Note that as the evaluation depth k is increased beyond 10 there is still some convergence in the metric, because documents beyond depth $d' = 10$ in this system's run might have appeared in the top-10 for some other system, and thus have judgments. The endpoints of those lines, at an evaluation depth of $k = 100$, are marked LB and UB. The dotted lines in the figure show the bounds on the score range that would arise if evaluation to k was supported by pooling to $d = 100$. The final $d = 100$ LB-UB range – a subset of the wider $d' = 10$ LB-UB range – is still

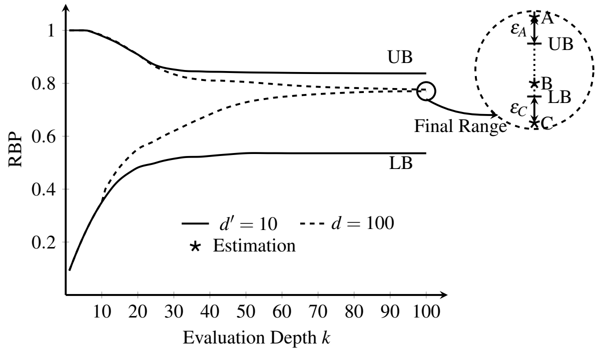

Fig. 2: An example of score bound convergence for a single system and a single topic. The two pairs of lines indicate the RBP score range at different evaluation depths k, based on two different pooling depths $d' = 10$ and $d = 100$. The "Final Range" is the metric score range at $k = 100$ using a $d = 100$ judgment pool; A, B, and C indicate three possible outcomes of a predictive model starting with the $d' = 10$ judgment pool.

non-empty, because the residual at depth k accounts for all documents beyond depth k, even if full or partial judgments beyond that depth are available.

Now consider an evaluation to depth $k = d$, but based on a model $G()$ derived from a pooling process to depth d'. If the model has strong predictive power, then the extended-evaluation using the predicted $\hat{r}_j$ values should give rise to a metric score that falls close to – or even within – the dotted-line LB-UB range that would have been computed using the deeper $d = 100$ judgment pool. That is, a metric score based on a predictive extrapolation will give rise to one of the three situations shown within the dotted circle: it will either overshoot the $d = 100$ range by an amount ϵ_A; or it will undershoot the $d = 100$ range by an amount ϵ_C; or it will fall within that range, as suggested by the point labeled B. In the latter case, we take $\epsilon_B = 0$.

The overall process followed is that for each topic we use the set of system runs for that topic, together with the depth-d' pooled judgments, and compute the parameters for an estimated gain function. We then use that gain function to extrapolate the depth-d metric scores for that topic for each system, using $\hat{r}_j$ values generated by the model in place of r_j values whenever the corresponding document does not appear within the depth-d' pool. So, for each combination of topic and system an ϵ difference is computed relative to the score range generated by a pooled-to-d evaluation.

4 Experiments

Test Collections We employ two different test collections, the 2004 Robust task (Rob04, topics 651–700) and the Terabyte06 task (TB06, topics 801–850), considering only the runs that contributed to the judgment pool. The first dataset has a pooling depth of $d = 100$ and a set of 42 contributing runs [13]; the second a pooling depth of $d = 50$, and 39 contributing runs [4]. Figure 3 provides a breakdown of document relevance in

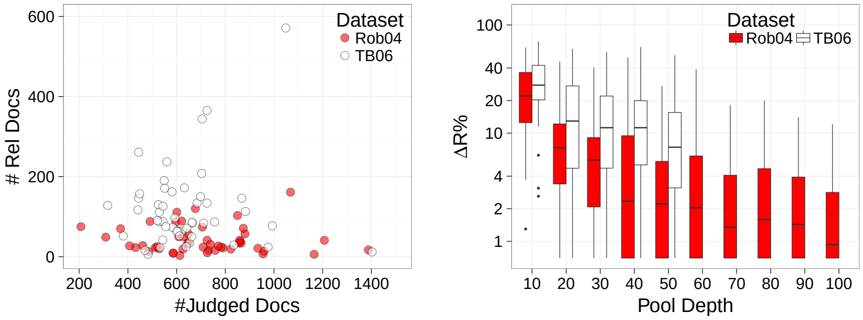

Fig. 3: Datasets used, showing the balance between judged documents and relevant documents on a per-topic basis (left); and the rate at which relevant documents are discovered by increasing pool depth bands (right, with a logarithmic vertical scale).

d	Rob04						TB06					
	G_s	G_c	G_ℓ	G_z	G_w	G_H	G_s	G_c	G_ℓ	G_z	G_w	G_H
10	0.237	0.106	0.080	0.086	0.071	0.070	0.190	0.040	0.020	0.018	$0.011^\dagger$	0.011
20	0.256	0.113	0.085	0.088	$0.072^\dagger$	0.071	0.186	0.049	0.020	0.022	$0.011^\dagger$	0.011
30	0.275	0.114	0.088	0.087	$0.071^\dagger$	0.070	0.186	0.056	0.022	0.024	$0.011^\dagger$	0.011
40	0.292	0.114	0.090	0.087	$0.069^\dagger$	0.069	0.187	0.060	0.024	0.023	0.012	0.011
50	0.307	0.112	0.090	0.087	$0.068^\dagger$	0.068	0.189	0.063	0.025	0.025	$0.012^\dagger$	0.011

Table 2: RMSE of models, evaluated to depth d, averaged across topics and systems, using parameters computed using pooling data to depth d. Model $G_H()$ is a hybrid that selects the best of the other models on a per-topic basis. Daggers indicate values not significantly worse than the hybrid model at $p = 0.05$, using a two-tail paired t-test.

the two collections. Although the TB06 dataset uses shallower pooling, on average it contains more relevant documents per topic than Rob04 (left pane); and the percentage of relevant documents decreases more slowly as a function of pool depth (right pane). For example, approximately 8% of the TB06 documents that first enter the pool as it is extended from $d = 40$ to $d = 50$ are found to be relevant.

Goodness-Of-Fit Evaluation Regression was used to compute the two or three parameters for each model (Table 1), fitting them on a per-topic-basis, and using a range of nominal pooling depths d. In the static model $G_s()$ the predicted gain was set to 0.5 at all ranks; and in the constant model $G_c()$ the cutoff parameter m was capped at the pooling depth. All of the judgments to the specified test depth d were used, in order to gauge the suitability of the various models. Note that the large volume of input data used per topic and the small number of parameters being determined means that there is only modest risk of over-fitting, even when d is small. Predictive experiments that bypass even this low risk are described shortly.

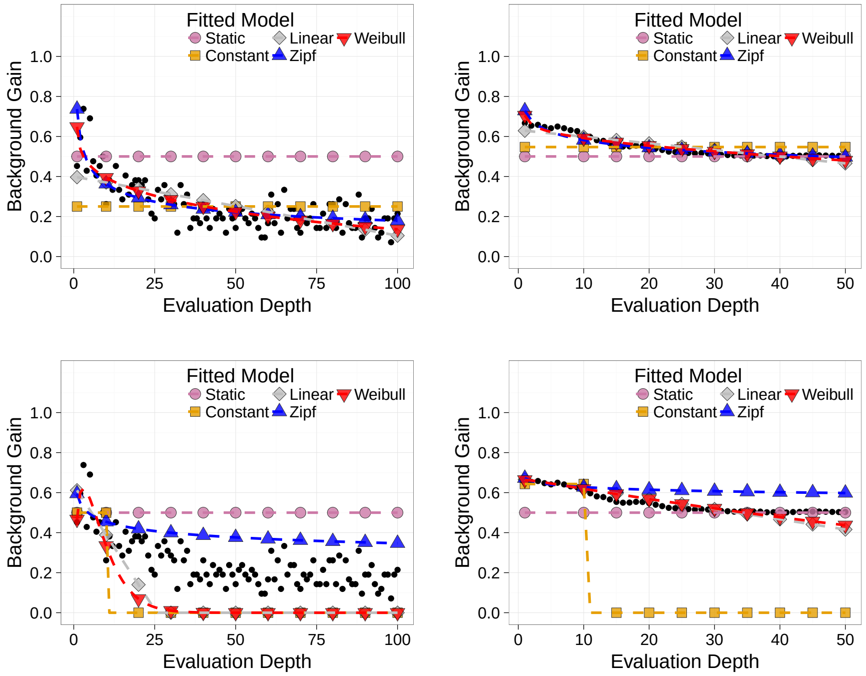

Fig. 4: Topic 683 for Rob04 (left column), and Topic 819 for TB06 (right column), with models fitted using all available judgments (top row) and using a depth $d' = 10$ pool (bottom row). The black dots show the empirical gain, and are the same in both rows.

Table 2 lists average RMSE scores, categorized by dataset, by model, and by pooling depth d. The two columns labeled $G_H()$ are discussed shortly. Two-tail paired t-tests over topics were used to compare the RMSE values associated with the five models. When all available judged documents are used, G_w has the smallest RMSE on both datasets compared to the other four models, at a significance level $p \leq 0.05$ in all cases, and is a demonstrably better fit to the observed data than are the other four approaches.

We also explored a hybrid model, denoted $G_H()$, which selects the smallest RMSE over the available fitting data for the five primary approaches on a topic-by-topic basis. Two-tail paired t-tests were also conducted between model G_H and each of the others, and in Table 2 superscript daggers indicate the RMSE measurements that were not found to be significantly inferior to the hybrid approach, again using $p \leq 0.05$. The Weibull model is a very close match to the hybrid approach, and of the per-topic selections embedded in the hybrid, the Weibull was preferred around 85% of the time.

Looking in detail at Table 2, we also conclude that the Rob04 judgments are harder to fit a curve to, with overall higher RMSE values for each corresponding depth and model compared to the TB06 judgments. It is also apparent that little separates the Zipfian $G_z()$ and linear $G_\ell()$ approaches, and that either could be used as a second-choice to the Weibull mechanism. Finally in connection with Table 2, the consistency of values down each column as data points are added confirms the earlier claim that there is only a modest risk of over-fitting affecting the results of this experiment.

Figure 4 illustrates the five fitted curves for two topics, and their approximation of the empirical gain, which is shown in the graphs as a sequence of black dots. One topic from each of the two datasets is plotted, with two different pooling depths – one graph in each vertical pair using all of the available judgments ($d = 100$ for Rob04, and $d = 50$ for TB06, in the top row), and one graph showing the models that were fitted when pooling was reduced to a nominal $d' = 10$ (bottom row). One observation is immediately apparent, and that is that empirical gain does indeed decrease with rank; moreover, in the case of TB06 Topic 819, it does so surprisingly smoothly. Also worth noting is that the empirical gain for the Rob04 topic decreases more quickly than it does for the TB06 topic as the evaluation depth k increases, which both fits with the overall data plotted in the right pane of Figure 3, and helps explain the better TB06 scores for the static model in Table 2. Comparing the top two graphs with the lower two, it is clear that the more volatile nature of the empirical gain in the Rob04 topic has meant that when only $d' = 10$ judgments are available, the models all diverge markedly from the actual g_k values when they are extrapolated beyond the fitted range. The smoother nature of the TB06 empirical gain function means that the extrapolated models based on $d' = 10$ continue to provide reasonable projections.

Predictive Strength Evaluation The most important test of the various models is whether they can be used to generate reliable estimates of metric scores when extrapolated beyond the pooling depth, the process that was illustrated in Figure 2. Table 3 lists the results of such an experiment, using RBP0.95 throughout, a relatively deep metric (at an evaluation depth of 50, the inherent RBP0.95 tail-residual is 0.07, and at an evaluation depth of 100, it is 0.006), and with $G_s()$ omitted for brevity. To generate each of the table's entries, a pool to depth d' is constructed, and the corresponding model fitted to the empirical gain values associated with that pool. Each run is then evaluated to depth $k = 100$ (Rob04) or $k = 50$ (TB06) using pooled-to-d' judgments, if they are available, or using estimated gain values $\hat{r}_j$ generated by the model for that topic. The RBP score estimate that results is then compared to the score and residual range generated using the full pool, $d = 100$ for Rob04 and $d = 50$ for TB06. If the extrapolated RBP score falls within that pooled-to-d range, an ϵ of zero is registered for that system-topic combination; if it falls outside the range, a non-zero ϵ is registered, as described in Section 3. Each value in the table is then the average over systems of the root-mean-square of that system's topic ϵ's ; with the parenthesized number beside it recording the percentage of the ϵ values that are zero, corresponding to predictions that fell within the final RBP score range. We also measured the "interpolative" method of estimating a final RBP score that was described Ravana and Moffat [10], denoted as "RM" in the table. It predicts RBP scores assuming that the residual can be assigned a gain at the same weighted rate as is indicated by the judged documents for that run.

All of the models are sensitive to the pooling depth d', and it is only when sufficient initial observations are available that it is appropriate to extrapolate. Also interesting in Table 3 is that the linear model, $G_\ell()$, provides score predictions that are as reliable as those of the Weibull model. As a broad guidance, based on Table 3, we would suggest

d'	G_c	G_ℓ	G_z	G_w	G_H	RM
Robust04						
20	0.020 (**33**)	**0.015** (**33**)	0.027 (15)	0.018 (31)	0.018 (31)	0.040 (9)
40	0.004 (60)	**0.003** (**65**)	0.005 (56)	0.004 (**65**)	0.004 (**65**)	0.010 (31)
60	**0.001** (78)	**0.001** (84)	**0.001** (**89**)	**0.001** (88)	**0.001** (88)	0.002 (71)
80	**0.000** (92)	**0.000** (95)	**0.000** (**99**)	**0.000** (97)	**0.000** (97)	**0.000** (98)
Terabyte06						
10	0.089 (24)	**0.061** (**45**)	0.082 (39)	0.065 (**45**)	0.067 (**45**)	0.065 (42)
20	0.033 (40)	**0.022** (71)	0.031 (68)	0.023 (**73**)	0.024 (**73**)	0.023 (68)
30	0.013 (58)	0.006 (88)	0.008 (87)	**0.005** (**90**)	0.006 (**90**)	0.008 (87)
40	0.004 (78)	0.001 (98)	0.001 (98)	**0.000** (**99**)	**0.000** (**99**)	0.001 (97)

Table 3: Root-mean-square of ϵ prediction errors using different pooling depths d', compared to an evaluation and pooling depth of $k = d = 100$ (Rob04) and $k = d = 50$ (TB06). The method labeled RM is the "interpolation" method of Ravana and Moffat [10]. Bold values are the best in that row, and the numbers in parentheses are the percentage of the system-topic combinations for which $\epsilon = 0$ (point B in Figure 2).

that if an evaluation is to be carried out to depth k, then pooled judgments to depth $d' \geq k/2$ are desirable, and that application of either the Weibull model $G_w()$ or the simpler linear model G_ℓ to infer any missing gain values between d' and k will lead to reliable final score outcomes. Both outperformed the previous RM approach [10].

That then leaves the choice of k, the evaluation depth to be used; as noted by Moffat and Zobel [9], k is in part determined by the properties of the user model that is embedded in the metric. In the RBP model used in Table 3, the persistence parameter $p = 0.95$ indicates a deep evaluation. When p is smaller and the user is considered to be less patient, the fact that the tail residual is given by p^k means that smaller values of k can be adopted to yield that same level of tail residual. Note that it is not possible to analyze AP in the same way, hence our reliance on RBP in these experiments.

5 Conclusions and Future Work

We have investigated a range of options for modeling the relationships between relevance and retrieval rank, calculating the probability of a document being relevant conditioned on a set of systems and the evaluation depths. Our experiments show that it is possible to use the models to estimate final scores in weighted-precision metrics with a reasonable degree of accuracy, and hence that pooling costs might be usefully reduced for this type of metric. To date the predictive score models have not been conditioned on the document itself, and the fact that it might be unjudged in multiple runs at different depths. We plan to extend this work to incorporate the latter, hoping to develop more refined estimation techniques. We also plan to explore the implications of stratified pooling, whereby only a subset of documents within the pool depth are judged.

Acknowledgment This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP140101587). Shane Culpepper is the recipient of an Australian Research Council DECRA Research Fellowship (DE140100275).

References

[1] Aslam, J.A., Pavlu, V., Yilmaz, E.: A statistical method for system evaluation using incomplete judgments. In: Proc. SIGIR. pp. 541–548 (2006)

[2] Buckley, C., Dimmick, D., Soboroff, I., Voorhees, E.M.: Bias and the limits of pooling for large collections. Inf. Retr. 10(6), 491–508 (2007)

[3] Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proc. SIGIR. pp. 25–32 (2004)

[4] Büttcher, S., Clarke, C.L.A., Soboroff, I.: The TREC 2006 Terabyte Track. In: Proc. TREC. pp. 39 – 53 (2006)

[5] Büttcher, S., Clarke, C.L.A., Yeung, P.C.K., Soboroff, I.: Reliable information retrieval evaluation with incomplete and biased judgements. In: Proc. SIGIR. pp. 63–70 (2007)

[6] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Information Systems 20(4), 422–446 (2002)

[7] Lipani, A., Lupu, M., Hanbury, A.: Splitting water: Precision and anti-precision to reduce pool bias. In: Proc. SIGIR. pp. 103–112 (2015)

[8] Lu, X., Moffat, A., Culpepper, J.S.: The effect of pooling and evaluation depth on IR metrics. Inf. Retr. 19(4), 416–445 (2016)

[9] Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Information Systems 27(1), 2 (2008)

[10] Ravana, S.D., Moffat, A.: Score estimation, incomplete judgments, and significance testing in IR evaluation. In: Proc. AIRS. pp. 97–109 (2010)

[11] Sakai, T.: Alternatives to BPref. In: Proc. SIGIR. pp. 71–78 (2007)

[12] Soboroff, I.: A comparison of pooled and sampled relevance judgments in the TREC 2006 Terabyte Track. In: Proc. EVIA (2007)

[13] Voorhees, E.M.: Overview of the TREC 2004 Robust Retrieval Track. In: Proc. TREC. pp. 69–77 (2004)

[14] Voorhees, E.M.: The effect of sampling strategy on inferred measures. In: Proc. SIGIR. pp. 1119–1122 (2014)

[15] Voorhees, E.M., Harman, D.K. (eds.): TREC: Experiment and Evaluation in Information Retrieval. The MIT Press (2005)

[16] Webber, W., Park, L.A.F.: Score adjustment for correction of pooling bias. In: Proc. SIGIR. pp. 444–451 (2009)

[17] Yilmaz, E., Aslam, J.A.: Estimating average precision when judgments are incomplete. Knowledge and Information Systems 16(2), 173–211 (2008)

[18] Yilmaz, E., Kanoulas, E., Aslam, J.A.: A simple and efficient sampling method for estimating AP and NDCG. In: Proc. SIGIR. pp. 603–610 (2008)

[19] Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Proc. SIGIR. pp. 307–314 (1998)