Estimating System Effectiveness Scores With Incomplete Evidence

Sri Devi Ravana[1,2] *Alistair Moffat*[1]

1. Department of Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia

2. Department of Information Science
University of Malaya
Kuala Lumpur 50603, Malaysia

{*sravana,alistair*}*@csse.unimelb.edu.au*

Abstract *It is common for only partial relevance judgments to be used when comparing retrieval system effectiveness, in order to control experimental cost. Using TREC data, we consider the uncertainty introduced into per-topic effectiveness scores by pooled judgments, and measure the effect that incomplete evidence has on both the systems scores that are generated, and also on the quality of paired system comparisons. We measure system behavior from three different points of view: the trend in effectiveness scores; the separability of system pairs; and the number of reversals in significance outcomes as the depth of judgments increases. Our results show that when shallow pooled judgments are used system separability remains relatively high, but that there is also a high rate of significance reversal. We then show that explicitly adjusting effectiveness scores to allow for the known amount of uncertainty gives a reduced number of reversals, and hence more consistent experimental outcomes.*

Keywords Retrieval evaluation, effectiveness metric, pooling

1 Introduction

It is now nearly twenty years since TREC-style large-scale experimentation comparing retrieval techniques was commenced. One facet of such experiments that has remained constant over these two decades is the tension between the cost of undertaking relevance judgments, and the desire for accuracy of measurement. An experiment can be relatively low-cost if only shallow judgments are undertaken, but that then means that "deep" effectiveness metrics (such as average precision, AP, and normalized discounted cumulative gain, NDCG) cannot be properly computed. As a result, a range of work has been undertaken to quantify the extent to which the values of deep metrics are correlated to shallow metrics, see, for example, Webber et al. [22].

In this paper we take a different approach, and seek to quantify the extent to which system pair comparisons are inaccurate when only shallow judgments are performed. In particular, we make use of TREC experimental runs and TREC relevance judgments to investigate whether pairwise relativities that are deemed significant when only shallow judgments are available remain significant when deeper judgments are provided, for a range of effectiveness metrics. We call this the *reversal rate* of an experiment – the extent to which the use of shallow judgments leads to conclusions of statistical significance that are not in fact supported when a fuller set of relevance judgments is used in the calculation of the effectiveness metric.

The results presented below show that the usual simplistic pooling assumption – that documents that are unjudged are irrelevant – leads to a higher reversal rate than methods that attempt to infer effectiveness scores based on other assumptions about unjudged documents. This outcome suggests that if shallow pooling is being used during an experiment, an appropriate mechanism for estimating effectiveness scores should also be employed.

2 Retrieval experimentation

Retrieval systems are often evaluated using prescribed test collections, fixed topic sets, and matching relevance judgments. This is particularly true in non-commercial research environments, in which access to query and click logs, and to other large-scale user interaction data, is limited by competition or privacy concerns. But formation of relevance judgments is costly, and so it is also usual for the relevance judgment sets to be generated once in a shared effort, and then reused as ground-truth by subsequent experimentation. This section briefly summarizes this type of collection-based retrieval experimentation, and outlines the various facets of the process that have been subject to scrutiny.

Proceedings of the 15th Australasian Document Computing Symposium, Melbourne, Australia, 10 December 2010.

Collections and topics The first step is to compile a suitable document collection. Among others, the annual TREC rounds have used newswire collections, government web pages, and patent repositories. Topics have been based on a range of statements of information need, matched to the collection. For details of these aspects of experimental design, see Voorhees and Harman [20].

Pooling Collection-based testing had its origins in the Cranfield collection about 40 years ago [2, 20, 5] which consists of about 1,400 abstracts and 225 requests. At that scale it was possible to be confident that all of the relevant documents were known, and that the relevance judgments were *complete*. With the use of larger (and hence more realistic) collections, it is impractical to generate complete judgments [17]. Instead, the documents are triaged into three sets in regard to each topic: those that have been judged and are relevant; those that have been judged and are irrelevant; and those that have not been judged.

To select the subset of the documents that will be judged for each topic, *pooling* is used [15]. To form a pool for each topic, each system in the set of s participating systems ranks the documents in relation to that topic. These s ranked lists are then truncated to some fixed *pool depth* d, and the list prefixes are combined and de-duplicated. This process focusses the judgment effort on at most sd documents, and means that, as a minimum, in each of the s system runs, the highest ranked d documents have all been judged. Presuming that each of the s systems prioritizes its ranking on the documents perceived as being most likely to be relevant, the overall set of judgments is similarly focussed on the documents that are most likely to be relevant. Test collections developed using this technique have been investigated in a number of ways and found to be relatively reliable in terms of their ability to predict system behavior on unknown topics [19, 26].

In contrast to these earlier evaluations, recent work has suggested that even pooling cannot completely eliminate the need to assess significant numbers of documents when large collections are being used, because the fraction of documents pooled compared to the collection size is small. In these cases the results generated using these relevance judgments may not be reliable [4].

Related work Many new methods and techniques have been introduced in recent years that seek to overcome the problems generated by incomplete judgments, including: alternative strategies that seek to increase the number of relevant documents identified [26], or otherwise adjust the order in which documents are added into a queue for judgment [9]; the use of multiple assessors per topic [16]; incorporating prior system scores into extended experiments [11]; reducing judging effort while maintaining a large number of topics [5]; identifying topic difficulty to provide reliable results [25]; evaluation without relevance judgments [1, 24]; and score adjustment for pooling bias [21]. Ali et al. [1], Sanderson and Zobel [14], Trotman and Jenkinson [16], and Voorhees [18] examine other aspects of test collection construction, and of pooling as a technique for identifying documents to be judged.

Effectiveness metrics Closely coupled with the issue of pooling is the question of which effectiveness metric should be used. Shallow metrics, such as precision at depth k (P@k, with k often chosen to be a small number such as 10) are completely determined provided d, the pool depth, is chosen such that $d \geq k$. Finite-depth metrics of this type do not include any normalization factor that scales them against the best that any system might do on this topic; this absence means that the scores generated are absolute values. When $k > d$, the extent of the uncertainty in any P@k score can also be exactly known, since the possible contribution of each unjudged document is exactly $1/k$. This uncertainty is denoted as a *residual*, and represents the magnitude of the range in which the P@k score might ultimately sit. This notion of residuals is taken up in more detail in the next section.

On the other hand, deeper "system" metrics such as average precision (AP) [2] and normalized discounted cumulative gain (NDCG) [7] give rise to normalized scores that are scaled against the best that any system might achieve, with a "perfect" ranking always attaining a score of 1.0, regardless of how many relevant documents there are for the query. For these metrics to be correctly computed, the number R of relevant documents for each topic must be known, with the implication that any pooling-based approach to depth d will identify an approximation $R_d \leq R$ of that number.

As a compromise between these classes of metric, rank-biased precision (RBP) [8] computes an absolute score rather than a relative score, over any finite prefix of a presumed-infinite ranked list. Because only a finite prefix is ever scored, and because there is no normalization by R, it is again possible to compute a residual that indicates the extent of the uncertainty generated by the truncated tail of the ranking, or by any other unjudged documents within the supplied prefix.

Other effectiveness metrics have also been proposed, including ones that expressly seek to ameliorate the problems cased by incomplete judgments [3, 12, 13]. We do not consider these approaches further in this work; rather, we seek to apply score estimation techniques to the more traditional effectiveness metrics. In particular, we consider the problem of uncertainty introduced into per-system per-topic effectiveness scores when incomplete judgments are used in experiments. In our work we present results of investigating pairwise comparison of systems when estimating system scores in face of incomplete evidence.

3 Score Estimation

As has already been introduced, suppose that d is the pool depth used in the development of a set of relevance judgments, and that k is the run depth for some system that contributed to the pool. Then, as an example, P@k assesses that fraction of the top k ranked documents for each system that are relevant for each topic. When $k > d$ there are three sets of documents identified by this process:

- those judged relevant, r in total;
- those judged irrelevant, n in total; and
- those not judged, $k-(r+n)$ in total.

The P@k score for this system on this topic can then be expressed as the interval $[B,T]$, where $B = r/k$ and $T = 1-n/k$ are the lower and upper bound on the P@k score. The P@k residual is then defined as $\Delta = T-B = 1-(r+n)/k$. Computation of a residual for RBP and other weighted-precision metrics (including DCG, the unnormalized version of NDCG) is only a little more complex.

On the other hand, when metrics such as NDCG and AP are being used, the unjudged tail of any ranking can (at least potentially) dominate the score established by any finite prefix and it is not possible to establish a range for the score. Indeed, with these metrics, all that can be said is that in a pathological situation, the eventual score calculated for any document ranking lies between $B = 0$ and $T = 1$. More specifically, if the judgment depth d is extended to a new value $d' > d$, then computed AP and NDCG scores might increase or might decrease, whereas P@k, RBP, and DCG scores can only increase.

In practice, use of score intervals is unwieldy, and scores ranges are represented by a *point estimate* that is computed as some function of the available information, with the estimate X associated with a range $[B,T]$ required to satisfy $B \le X \le T$. Taking as a starting point the work of Ravana and Moffat [10], we explore four different estimation techniques in the experiments discussed below.

Simplistic prediction The simplest method is to take the lower bound of the interval, as the score estimate X,

$$X_S = B\,.$$

This is the "conventional" way of dealing with judgment uncertainties, and is best summarized as "if it ain't judged, it ain't relevant".

Background prediction A second option is to make use of a global estimate E that represents the background probability of a document being relevant given that it has been retrieved. The score associated with a $[B,T]$ interval can then be estimated as

$$X_B = B + \Delta E\,.$$

In this method, a fixed fraction of Δ is uniformly added to B. The value E is a constant and it can take any value from 0 to 1 although through experiments we observed that the $0.01 \le E \le 0.05$ is a reasonable range [10].

Interpolated prediction Assuming that the unjudged documents for a system are – to within some constant factor C – as likely to be relevant as the documents for which judgments are available leads to interpolated scores to be computed as:

$$X_I = B + C\Delta\frac{B}{1-\Delta}\,.$$

Constant C is a value between 0 to 1, and suitable values are discussed shortly. A value for X_I cannot be computed when $\Delta = 1$ (that is, when $B = 0$ and $T = 1$), and in this special case $X_I = E$ is assumed.

Smoothed prediction Assuming that the lower the uncertainty Δ, the greater the confidence is in the Interpolated prediction, and in contrast the higher the uncertainty, the more the background model should be preferred, leads to a smoothed approach Ravana and Moffat [10]:

$$\alpha X_I + (1-\alpha)X_B\,,$$

where α is a parameter that reflects the level of confidence in the Interpolated prediction. If α is chosen to be $\alpha = 1-\Delta$, this simplifies to

$$X_M = B + C\Delta B + \Delta^2 E\,,$$

where, as before, C is a constant between zero and one.

Computing score ranges We experimented with a total of five different effectiveness metrics: P@10, a typical shallow metric; AP, the average precision when evaluated relative to R_d, where R_d is the number of relevant documents encountered in the first d items of any of the pooled system runs; NDCG, normalized cumulative discounted gain, again using R_d; SDCG at depth $k = 100$, the discounted cumulative gain (see Järvelin and Kekäläinen [7]) scaled by the maximum possible score possible at depth 100; and RBP, rank-biased precision, with parameter $p = 0.95$.

With all of P@10, SDCG, and RBP, the base value B and top value T that bookend each score interval are relatively straightforward to compute. With AP and NDCG neither B nor T is easy to compute, and instead we use an approximation to gauge the breadth of the $[B,T]$ interval. To establish a lower estimate B, the usual approach to computing the scores was followed, making the assumption that none of the unjudged documents that appeared in the ranking were relevant. As already noted, this is not a strict lower bound on the eventual score.

To calculate an upper estimate T of the score range, the number of judged relevant documents in the run was subtracted from R, the total number of relevant documents for the topic. The remaining relevant documents not already accounted for in the ranking were

then assumed to be inserted into the ranking at the earliest possible locations at which unjudged documents appeared. For example, with $R = 5$ and $k = 10$, the ranking

```
1  0  ?  0  1  1  0  ?  0  ?
```

(in which "0" is an irrelevant document, "1" is a relevant document, and "?" is an unjudged document) gives rise to an AP-based $[B,T]$ interval computed as

$$B = \frac{1}{5}\left(\frac{1}{1}+\frac{2}{5}+\frac{3}{6}\right) = 0.38\,,$$

which is the usual computation with two terms completely absent; and

$$T = \frac{1}{5}\left(\frac{1}{1}+\frac{2}{3}+\frac{3}{5}+\frac{4}{6}+\frac{5}{8}\right) = 0.71\,,$$

now with those two terms inserted at the first available locations.

While this arrangement is in fact not feasible (since, in the example, it is known that neither of the two relevant documents that are absent from the ranking appear in positions 3 and 8), in conjunction with the B value, this method of estimating an approximation of T does give reasonable guidance as to the level of imprecision in the computed AP score. In particular, when $\Delta = T - B$ is large, then the B score may not be a good estimate of the final AP score. A similar approach allows estimates of B and T to be made for NDCG.

4 Experimental Investigation

Test Data We make extensive use of the relevance judgments and submitted runs that were generated during the TREC9 Web Track undertaken in 2000 [6]. This track had 105 runs submitted in response to a set of 50 topics. Of these, 59 runs were used in the pooling stage during which the set of judgments was generated, and 46 of the runs were not. From the 59 contributing runs (the set denoted as "59-con"), for each topic, the top 100 document identifiers from each run were pooled and judged, following the standard TREC methodology of generating relatively deep judgment sets.

The complete set of judgments for the TREC9 Web track contains 69,100 recorded outcomes, which means that on average each judged document got nominated by 4.3 of the 59 contributing systems. To generate simulations of shallower pools for experimental purposes, each judgment in the full set was tagged with the minimum depth at which that document was located in any of the 59-con runs, and then the judgments sorted into increasing order of minimum encountered depth. Prefixes of length 1,000 and 10,000 judgments were then taken, as a simulation of the outcomes that would arise if shallow and medium judgments were used.

The division into contributing (set 59-con) and non-contributing (set 46-non) runs is a useful one, and we report results separately for the two sets of systems. In particular, use of the 46-non set of system runs allows exploration of issues that arise when systems are being compared without any of them having contributed to the judgment set.

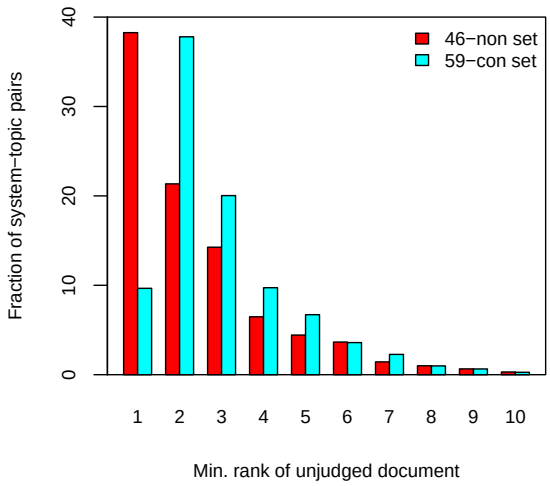

Figure 1: Distribution of ranks of first unjudged document in each run, categorized by whether or not the run contributed to the pool. A total of 1,000 judgments is assumed.

Figure 1 highlights this distinction. It shows, averaged across the systems and topics, the depth of the first unjudged document in each run when only 1,000 judgments are used. With this number of judgments to be distributed across 59 systems and 50 topics, the majority of runs in the 59-con set have their top-ranked document judged, but not always the second, so the effective pool depth is around $d = 1$. On the other hand, in the 46-non set, more than a third of the runs do not even get their top-ranked document judged.

Shallow pooling and uncertainty The first phase in our evaluation was to simply score the sets of runs using the three judgment sets, but measuring the extent of the uncertainty generated by the incomplete judgments. Table 1(a) shows the average base scores B computed for the 46-non systems, using five different effectiveness metrics, and evaluated using shallow, medium, and deep pooled judgments. In the case of the three weighted precision (and hence score accretive) metrics P@10, SDCG and RBP, the use of the $X_S = B$ approximation leads to non-decreasing score estimates as the number of judgments increases. On the other hand, the base AP score estimates decrease as the pool depth increases. This behavior is a consequence of R increasing, but those additional relevant documents not appearing in the majority (or even any) of the runs actually being scored. In between is NDCG, where it appears that the base score estimate B is relatively stable even from very shallow pool depths.

Table 1(b) lists the average residuals Δ associated with those base scores. The best that can be said about these values is that for the three weighted-precision metrics they decrease as the judgment pool increases in size. But as a general indication of scoring certainty, they provide very weak evidence. In particular, even

Judgments	P@10	SDCG	RBP	AP	NDCG
1,000	0.1184	0.0419	0.0647	0.1822	0.3282
10,000	0.1877	0.0900	0.1269	0.1541	0.3494
69,100	0.1923	0.1085	0.1398	0.1258	0.3237
(a) Base effectiveness scores, $X_S = B$					
1,000	0.6759	0.8535	0.7921	0.2744	0.2903
10,000	0.2692	0.5670	0.4333	0.2789	0.3240
69,100	0.1631	0.2311	0.1955	0.2309	0.3229
(b) Residuals resulting from unjudged documents, Δ					
1,000	0.1840	0.1373	0.1556	n/a	n/a
10,000	0.2019	0.1293	0.1566	n/a	n/a
69,100	0.1974	0.1157	0.1460	n/a	n/a
(c) Interpolated scores, X_I with $C = 0.42$ and $E = 0.01$					
1,000	0.1823	0.0995	0.1255	n/a	n/a
10,000	0.2064	0.1296	0.1589	n/a	n/a
69,100	0.2020	0.1216	0.1509	n/a	n/a
(d) Smoothed scores, X_M with $C = 0.91$ and $E = 0.05$					

Table 1: Base effectiveness scores B; residuals Δ; and two point estimates within the $[B,T]$ range, in all cases averaged across 50 topics and the 46-non set of system runs.

when the full set of 69,100 judgments is applied to the most focussed of the five metrics, P@10, not even one decimal digit of accuracy can be relied on. This clearly suggests that the simplistic point values X_S may not be accurate. (When the 59-con set of systems is used with $d = 100$ relevance judgments, the average residual for P@10 and SDCG@100 is zero, see Ravana and Moffat [10] for these and related results.)

The approximated residuals for the two normalized metrics, AP and NDCG , are both very large; nor do they decrease as the judgment pool increases in size. This suggests that either AP and NDCG scores are intrinsicly imprecise, or that the estimation methodology is inaccurate. Further work is required to determine which explanation is the correct one.

RMS error To quantify the difference between true and estimated values, we computed root-mean-square (RMS) errors. If Y is a set of n "true" values, $Y = [y_1, y_2, \ldots, y_n]$, and X is a corresponding set of estimated values, $X = [x_1, x_2, \ldots, x_n]$, then the root-mean-square difference between Y and X is computed as:

$$RMSE(X,Y) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}}.$$

The smaller the RMSE value the better the predictive quality of the estimation method.

To determine constants C and E to be used in the Interpolative and Smoothed predictions methods X_I and X_M respectively, we took the set Y to be the 50×59 at-69,100 system-topic scores achieved by the 59-con systems. The set of estimates X was then computed for each system and each topic, based on six different pool depths of, variously, 1,000, 2,000, 4,000, 10,000, 20,000, and 40,000 judgments.

Figures 2 and 3 show how *RMSE* varies as C and E are altered, using the Interpolated method to predict

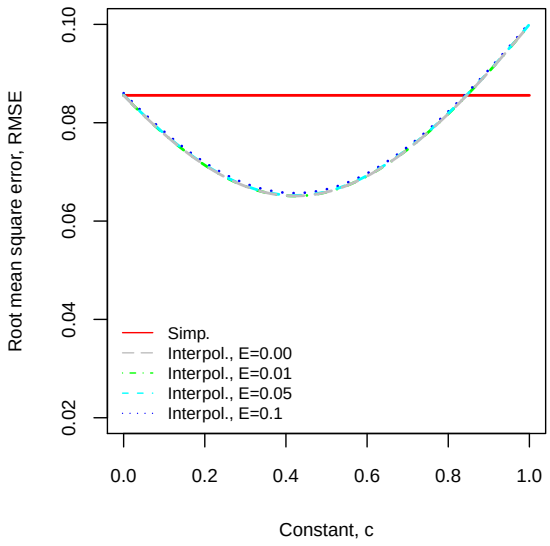

Figure 2: Prediction quality of P@10 scores estimated using the X_I Interpolated approach, plotted as RMSE values. The 59-con set was used with C and E varying, with results aggregated over six different pooling depths. The minimum point arises with $C = 0.42$ and $E = 0.01$.

scores from $[B,T]$ ranges. The minimal X_I-RMSE value is 0.065, while the minimum X_M-RMSE is 0.086, achieved with different C and E values for the two different methods. Both of these two figures were generated using the metric P@10; broadly similar curves resulted for the SDCG and RBP metrics.

For the purpose of the experimentation, $C = 0.42$ and $E = 0.01$ are used in the Interpolated method X_I; and $C = 0.91$ and $E = 0.05$ are used in the Smoothed method X_M. Both combinations give smaller RMSE values than the simplistic predictor X_S, for all of P@10, SDCG, and RBP.

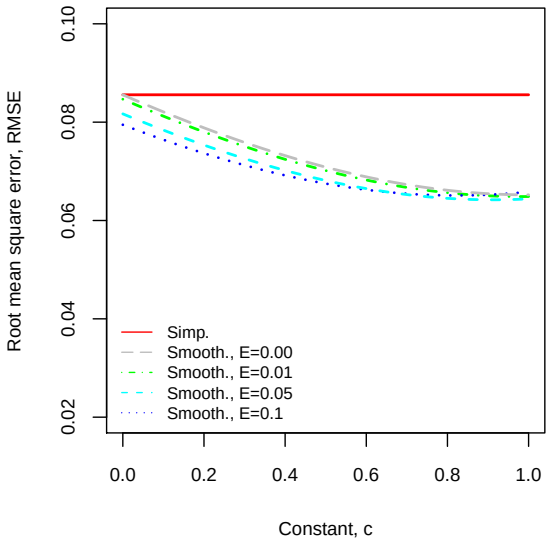

Figure 3: Prediction quality of P@10 scores estimated using the X_M Smoothed approach, plotted as RMSE values. The 59-con set was used with C and E varying, with results aggregated over six different pooling depths. The minimum point arises with $C = 0.91$ and $E = 0.05$.

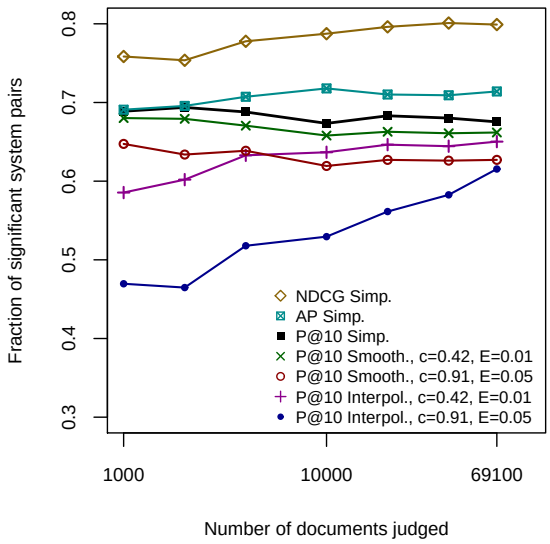

Figure 4: Separability rates within the 46-non set of systems for P@10 as a function of the total number of relevance judgments performed, with pooling across 50 topics and 59 systems, and with the comparison based on use of the t-test at the 0.01 confidence level.

Trends in effectiveness scores Table 1(c) applies these learnt constants to the 46-non set of systems, showing the average of the X_I point estimates for three metrics with $C = 0.42$ and $E = 0.01$. The final at-69,100 scores are now being overestimated when the judgment pool is shallow, but by less than the previous underestimates. Similarly, Table 1(d) shows the smoothed scores X_M with $C = 0.91$ and $E = 0.05$, for the same combinations of metrics and judgments. The smoothed estimates seem to have a more consistent trend of scores, especially from the 10,000 judgment starting point. The Interpolated and Smoothed predictors were not applied to the AP and NDCG metrics. Indeed, NDCG is relatively consistent in its value as the pool depth increases.

Separability Figure 4 shows system separability using P@10 as the number of judgments employed increases, where separability (sometimes also called discrimination) is the fraction of the possible system pairs that are identified as being statistically separable at the $p = 0.01$ confidence level. In this case, the fraction shown is relative to the $46 \times 45/2 = 1{,}035$ possible system pairs among the 46-non data set. The different curves within correspond to different score estimation method, C and E values used.

Surprisingly, it is the simplistic predictor X_S that generates the highest fraction of significant pairs at all depths of judgments compared to the other estimation methods when the underlying metric is P@10. Indeed, the high level of separability is attained despite the non-trivial residuals documented in Table 1(b) – the high separability is not just a matter of P@10 being a shallow metric and hence capable of being fully evaluated from a shallow pool. The same also holds true of the SDCG and RBP metrics – the highest separability arises with the simplistic predictor, and the Interpolative and Smoothed predictors give lower levels of separation between system pairs.

In behavior that is in agreement with the experimentation of other researchers (see, for example, Webber et al. [23]), AP tends to generate a greater fraction of separable system pairs than P@10, and NDCG is better again than AP. Figure 4 shows that this consistency happens irrespective of pool depth, and may be a consequence of the relative stability of the numeric values for the NDCG, as noted in Table 1(a). Similar outcomes are also noted by Webber et al. [23].

Reversals High separability rates are desirable, but only if the outcomes that are found to be significant are genuine ones. In particular, it is of concern if a metric asserts that some system significantly outperforms another when evaluated using a shallow pool, but the same conclusion cannot be reached when a more extensive set of relevance judgments is used. We call this situation a *reversal* – a system pair that is separable based on a prefix of the judgment set, but cannot be separated using the full set of 69,100 judgments (which is, of course, a prefix of the full all-documents all-topics judgment set). A system with a high separability rate might also suffer from a high reversal rate – in which case it is identifying spurious relativities between systems. Of course, the nature of significance testing itself allows some leeway in this regard – if 1,000 system pairs are evaluated, and 800 of them yield significance at the $p = 0.01$ level, then it would be unsurprising if a dozen of the system pairs did not yield significance on fresh topics and judgments.

Figure 5 shows the set of p values of the $46 \times 45/2 = 1{,}035$ system pairs making up the 46-non set, with the comparison based on use of the

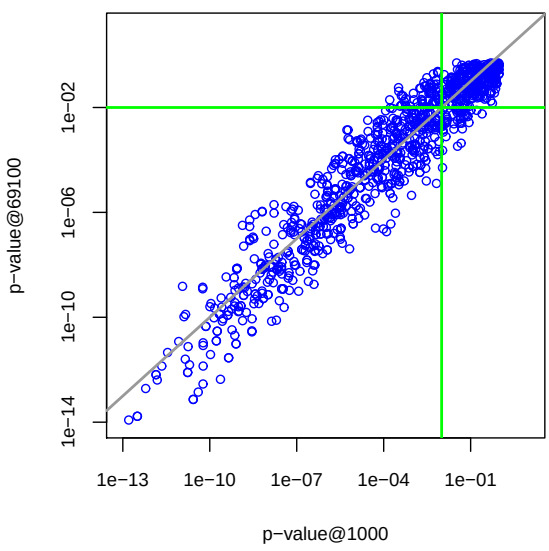

Figure 5: Using P@10 and X_S, with each point representing one system pair within the 46-non set, plotted according to the *p*-value computed over 50 topics, using a judgment pool of 1,000 outcomes (horizontal axis) and a judgment pool of all 69,100 outcomes (vertical axis).

t-test at the 0.01 confidence level. Each point plotted represents one system pair, with the horizontal location of the point determined by the p value computed for that pair when 1,000 judgments are being used (as previously, derived from the 59-con set over 50 topics), and the vertical location being determined by the p value that arises when the full 69,100 judgments are used.

Points in the lower-left quadrant of the graphs represent system pairs that are separable at the $p = 0.01$ level using both 1,000 and 69,100 judgments – that is, evaluations that are stable with respect to pool depth. The lower-right quadrant is also of interest – it indicates situations in which supplying more judgments improves separability, and points plotted in this zone can be regarded as being the payoff for performing deep judgments.

The quadrant of concern in Figure 5 is the upper-left one, which shows system pairs that were identified as being significantly different using shallow judgments, but for which that assessment was retracted once the full set of judgments was made available.

Table 2 draws all these ideas together, and lists separability percentages and reversal percentages (both as fractions of the $46 \times 45/2 = 1{,}035$ system pairs in the 46-non set of systems) for a wide range of metrics and point estimation mechanisms. The Simplistic prediction mechanism gives the greatest separability in each of the metrics, but also has a high rate of reversals. It thus appears that at least some of the separability advantage is illusory. On the other hand, the Interpolated predictor is less likely to yield a significant outcome at shallow pool depths, but compensates with a lower rate of reversed assessments. The two "deep" evaluation metrics, AP and NDCG, have the highest separability rates; but also have a relatively high rate of reversals.

Similar results were obtained for the metric P@10 when the same experiment was carried out using the TREC-8 Ad-Hoc Track data, consisting of 50 topics, 86,830 documents in the pool, and 71 runs contributed to the pooling (of a total 129 runs submitted).

5 Conclusion

All measurement involves uncertainty. When the measurement is of opinion-based outcomes, the uncertainties must be incorporated and managed; and when the cost of undertaking the measurement can be traded against repeatability and fidelity, the set of issues to be balanced becomes very large indeed. In this paper we have explored some of the consequences of shallow pooling in information retrieval experiments, and demonstrated that while simplistic predictions allow relatively high separability coefficients, there is also a higher rate of retraction of significance relationships as more judgments are performed. The Interpolative approach to predicting system scores is more robust in terms of reversals, but also less likely to find significance when the pool is shallow. It may be that these two facets of behavior are inevitable consequences of each other.

References

[1] K. Ali, C.-C. Chang, and Y. Juan. Exploring cost-effective approaches to human evaluation of search engine relevance. In *Proc. 32nd ACM SIGIR Conf.*, pages 802–803, Boston, USA, July 2009.

[2] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proc. 23rd ACM SIGIR Conf.*, pages 33–40, Athens, Greece, July 2000.

[3] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. 27th ACM SIGIR Conf.*, pages 25–32, Sheffield, England, July 2004.

[4] C. Buckley, D. Dimmick, I. Soboroff, and E. M. Voorhees. Bias and the limits of pooling. In *Proc. 29th ACM SIGIR Conf.*, pages 619–620, Seattle, WA, August 2006.

[5] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over thousands of queries. In *Proc. 31st ACM SIGIR Conf.*, pages 651–658, Singapore, July 2008.

[6] D. Hawking. Overview of the TREC-9 Web Track. In *Proc. 9th Text REtrieval Conf. (TREC-9)*, Gaithersburg, Maryland, November 2000.

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

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

Metric	Predictor	Collection	Separability		Reversals
			1,000	69,100	
P@10	Simplistic	TREC-9	68.9%	67.5%	5.8%
P@10	Interpolated	TREC-9	58.6%	65.0%	2.2%
P@10	Smoothed	TREC-9	64.7%	62.7%	6.1%
SDCG	Simplistic	TREC-9	74.5%	74.9%	5.5%
SDCG	Interpolated	TREC-9	56.2%	71.6%	4.1%
SDCG	Smoothed	TREC-9	74.1%	67.7%	11.3%
RBP	Simplistic	TREC-9	74.1%	74.8%	4.9%
RBP	Interpolated	TREC-9	56.9%	72.9%	3.1%
RBP	Smoothed	TREC-9	72.1%	70.0%	7.8%
AP	Simplistic	TREC-9	69.1%	71.4%	5.4%
NDCG	Simplistic	TREC-9	75.8%	79.9%	4.3%

Table 2: Separability of metrics and predictors, and the fraction of reversals that arise when statistical testing based on a pool of 1,000 judgments is then extended to use all 69,100 judgements. In each case the evaluation is over all system pairs in the 46-non set of systems, with the percentages expressed as fractions of 1,035.

[9] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *Proc. 30th ACM SIGIR Conf.*, pages 375–382, Amsterdam, July 2007.

[10] S. D. Ravana and A. Moffat. Score estimation, incomplete judgments, and significance testing in IR evaluation. In *Proc. AIRS Asia Information Retrieval Societies Conf.*, Taipei, Taiwan, December 2010. To appear.

[11] S. D. Ravana, L. A. F. Park, and A. Moffat. System scoring using partial prior information. In *Proc. 32nd ACM SIGIR Conf.*, pages 788–789, Boston, USA, July 2009.

[12] T. Sakai. Alternatives to Bpref. In *Proc. 30th ACM SIGIR Conf.*, pages 71–78, Amsterdam, July 2007.

[13] T. Sakai and N. Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11(5):447–470, 2008.

[14] M. Sanderson and J. Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proc. 28th ACM SIGIR Conf.*, pages 162–169, Salvador, Brazil, August 2005.

[15] K. Sparck Jones and C. J. Van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32 (1):59–75, 1976.

[16] A. Trotman and D. Jenkinson. IR evaluation using multiple assessors per topic. In *Proc. 12th Australasian Document Computing Symp.*, pages 9–16, Melbourne, Australia, December 2007.

[17] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Proc. 2nd Workshop of the Cross-Language Evaluation Forum (CLEF)*, pages 355–370, Darmstadt, Germany, September 2001.

[18] E. M. Voorhees. Topic set size redux. In *Proc. 32nd ACM SIGIR Conf.*, pages 806–807, Boston, USA, July 2009.

[19] E. M. Voorhees. Variations in relevance judgements and the measurements of retrieval effectiveness. In *Proc. 21st ACM SIGIR Conf.*, pages 315–323, Melbourne, Australia, August 1998.

[20] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, Cambridge, Mass., 2005.

[21] W. Webber and L. A. F. Park. Score adjustment for correction of pooling bias. In *Proc. 32nd ACM SIGIR Conf.*, pages 444–451, Boston, USA, July 2009.

[22] W. Webber, A. Moffat, J. Zobel, and T. Sakai. Precision-at-ten considered redundant. In *Proc. 31st ACM SIGIR Conf.*, pages 695–696, Singapore, July 2008.

[23] W. Webber, A. Moffat, and J. Zobel. The effect of pooling and evaluation depth on metric stability. In *Proc. 3rd EVIA Int. Work. Evaluating Information Access*, pages 7–15, Tokyo, Japan, June 2010.

[24] S. Wu and F. Crestani. Methods for ranking information retrieval systems without relevance judgements. In *Proc. ACM SAC Symp. on Applied Computing*, pages 811–816, Florida, USA, March 2003.

[25] J. Zhun, J. Wang, I. Cox, and V. Vinay. Topic (query) selection for IR evaluation. In *Proc. 32nd ACM SIGIR Conf.*, pages 802–803, Boston, USA, July 2009.

[26] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. 21st ACM SIGIR Conf.*, pages 307–314, Melbourne, Australia, August 1998.