Score Estimation, Incomplete Judgments, and Significance Testing in IR Evaluation

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

[1] Department of Computer Science and Software Engineering,
The University of Melbourne
[2] University of Malaya, Malaysia

Abstract. Comparative evaluations of information retrieval systems are often carried out using standard test corpora, and the sample topics and pre-computed relevance judgments that are associated with them. To keep experimental costs under control, partial relevance judgments are used rather than exhaustive ones, admitting a degree of uncertainty into the per-topic effectiveness scores being compared. Here we explore the design options that must be considered when planning such an experimental evaluation, with emphasis on how effectiveness scores are inferred from partial information.

Keywords: Retrieval evaluation, effectiveness metric, pooling.

1 Introduction

Two distinct methodologies have emerged for comparing the usefulness of information retrieval systems: subject-based approaches that observe humans as they carry out some information seeking task while aspects of their behavior are monitored; and data-based approaches, in which standard corpora of documents, and the query topics and relevance judgments that accompany them, are used and re-used. The drawback of the first approach is that human experimentation requires great care in experimental design and in data interpretation, and is both difficult to reproduce, and expensive to carry out on a large scale. The drawback of the second approach is that it requires system "usefulness" to be approximated by an effectiveness metric that may or may not adequately represent the facets of performance that can be elicited via a user study [10,19]. Data-driven experiments also require human input while the test sets are being prepared. The preparation of relevance judgments is expensive, and creation of exhaustive relevance judgments for non-trivial topics is beyond the resources of most individuals or organizations, even for relatively modest collection sizes and topic sets.

To identify a subset of the documents for judging, *pooling* is commonly used. Pooling has been the standard approach used in TREC for nearly two decades [20], and selects for judgment documents which are highly ranked by at least one of the contributing systems. The current practice is to pick the top d ranked documents for each system for each topic, identify the set of unique document-topic pairs, and judge them, with the choice of d controlling the cost of the evaluation in a non-linear manner.

This arrangement then gives rise to the question as to how unjudged documents should be handled when they are encountered during the evaluation of whatever effectiveness metric is being used. The conventional and default assumption is to presume

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 97–109, 2010.

that unjudged documents are not relevant, arguing that if they were relevant, they should have appeared within the top d for at least one of the systems that gave rise to the pool. This presumption is especially important when effectiveness metrics such as NDCG and AP are being used, since they include a normalization step by R, the total number of relevant documents for that topic. Studies have shown that when d is of the order of 100 to 200, the value of R derived from pooling is a relatively useful estimate [23]. But when a smaller number of judgments are being performed, care is required – it is not at all unusual, for example, for system runs to be generated and evaluated to depth $1{,}000$ even though $d = 100$ is used for the judgments. In this case it is entirely possible for fully 90% of the documents comprising a scored run to be unjudged.

Our work in this paper examines this tension. Specifically, we:

- explore methods for estimating effectiveness scores in the face of missing relevance judgments;
- compare the quality of the system comparisons that result from those estimation techniques; and
- explore the appropriateness of statistical tests when the number of judgments against them is small.

In particular, we demonstrate that the assumption of "not relevant" for unjudged documents, while simplistic, and giving incorrect effectiveness scores, still leads to reliable experimental outcomes.

2 Retrieval Experimentation

The design of a retrieval experiment relies on a number of critical decisions [20]. This section briefly summarizes these different facets.

Collection and topics. The starting point for data-based IR experimentation is a collection of realistic documents, where realistic refers both to content and style, and to scale. In many IR experiments documents are sourced from the public web, but commercial entities may choose to make use of private collections, such as email repositories and the like. A set of realistic topics relative to that data is also required, where a topic is expressed as a query or more detailed information need, and is accompanied by a statement as to the supposed intent of the searcher assumed to have authored that need.

The two systems being compared are then used to create a ranked list of answers, or a *system run* for each of the topics, evaluated in the context of the collection. It is those system runs – containing a thousand or more documents in order, for each of the topics – that are then evaluated in the remaining steps of the experiment.

Effectiveness metrics. A suitable effectiveness metric is then chosen to reflect the assumed model for the anticipated user behavior. For example, if users are presumed to be focussed primarily on quickly identifying a single possible resource in connection with the query, a metric such as "hit at 3" might be appropriate, which assigns "1" to any retrieval run in which one or more of the top-three ranked documents is an answer. The assessment of somewhat more patient user behavior might be modelled by the metric precision at 10 (P@10); and extended searching behaviors might be modelled by a

weighted precision metric that reaches further down the ranking, such as that offered by Discounted Cumulative Gain (DCG) [11] evaluated to depth 100 in the rankings.

Formation of relevance judgments. The next step is to form relevance judgments that allow the effectiveness metric to be evaluated. A usual method for building relevance judgments is to *pool* the runs from the systems being compared down to some depth d. This approach yields incomplete judgments, but is necessary because it is beyond most researchers' resources to carry out comprehensive judgments across a non-trivial set of topics for realistic-sized document collections.

Statistical testing. Another important facet of experimental design is the choice of statistical test, and a number of authors have commented in this regard [4,8,13,16,17,18,23]. Parametric statistics such as the t-test are used when assumptions can be made about the distribution underlying the system scores (or more correctly, underlying aggregates of scores), with predictions then possible about future behavior. For the most part, these rely on the scores being normally distributed. Parametric distributions can also be applied to score *differences* when a sufficiently large set is available for analysis, even if the two underlying score distributions are not normal. For typical purposes, most sources agree that "sufficiently large" is attained at around 30–50 independent paired observations. When there is a pre-conceived notion as to which system is being tested for superiority (because, for example, it has the higher mean score across the set of observations) a one-tailed test is appropriate.

Statistical testing also assumes that the observations are independent and drawn randomly from some universe. In the case of an IR experiment, it is hard to provide evidence that the topics are a random subset of all queries; nevertheless, it is an assumption that is made in all statistical testing on IR system scores.

3 Score Estimation

Consider the simple metric *precision at depth* k, computed as the fraction of the top k ranked documents for each system that are relevant. If d, the pool depth, is larger than k, then all of the system-topic scores are fully defined, because every required document has been judged. But when $k > d$, or when a new system that was not a contributor to the pool is being scored, there are three sets of documents, rather than two:

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

The effectiveness score for the run can then be expressed as a range $[B, T]$, where $B = r/k$ and $T = 1 - n/k$ are the bottom and top of the range, and $\Delta = T - B = 1 - (r + n)/k$ is the uncertainty, or *residual* associated with the measurement. In the presence of unjudged documents other precision-based effectiveness metrics can also be evaluated to a $[B, T]$ interval rather than a point, including *rank-biased precision* [12], and *discounted cumulative gain* [11]. Moffat and Zobel [12] make explicit reference to the benefits of tracking score uncertainty via a residual, and highlight the

lack of fidelity in effectiveness scores that arises when the pool depth d is shallow and significant numbers of unjudged documents are encountered.

In fact, what is desired in order for the system comparison to take place is a *point estimate* that reflects the interval. More precisely, if the interval is taken to be the domain of a probability density function that describes the likelihood of the final score being any value in the interval, then the required point splits the probability density into two parts each of mass 0.5.

When viewed this way, taking B as the representative point is a approach that is open to question, and there are other point estimates that could be considered. In the experiments that are described below, the following four methods are used. In all cases it is assumed that the representative point X is required to lie within $[B, T] \subseteq [0, 1]$.

Simplistic prediction. As already noted, the simplest approach is to take the minimum value of the range, $X_S = B$.

Background prediction. If a global estimate E can be computed for the background probability of a document being relevant, then the unjudged documents can be presumed to be relevant with that probability,

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

In this approach, a fixed fraction of Δ is added to B, regardless of the value of B. The question now is to determine an appropriate value of E; the value $E = 0.01$ is defended below as being a reasonable one.

Interpolated prediction. A third option is to split Δ, based on the ratio of B to $1 - T$, on the assumption that unjudged documents for a particular system are as likely to be relevant as the documents for which judgments are available. This approach yields:

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

This method is similar to the RBP projection method discussed by Moffat and Zobel [12]. An obvious drawback is that it cannot be computed when $\Delta = 1$ (that is, when $B = 0$ and $T = 1$), and in this special case $X_I = E$.

Smoothed prediction. The smaller the value of Δ, the greater the confidence in the Interpolated prediction. Conversely, the greater the value of Δ, the more attractive it may be to prefer the background model. This combination leads to point score calculated as:

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

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

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

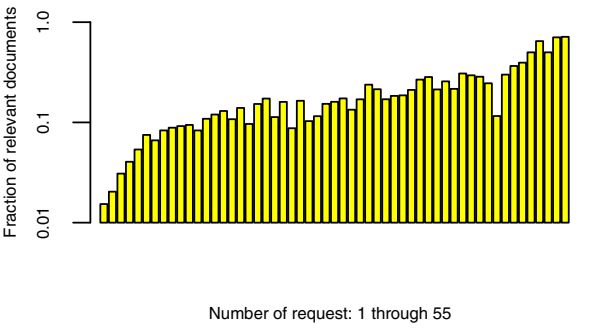


Fig. 1. Fraction of relevant documents, as a function of the number of systems that had that document in their top-d pool for $d = 100$, summed over 50 topics and the 59 systems contributing to the judgment pool in the TREC-9 Web Track

Pessimistic interval comparison. Rather than seek to represent an interval by a point value within the interval, it is also possible to compare corresponding score intervals directly. Suppose two systems are being compared, S_1 and S_2, and their score ranges are $[B_1, T_1]$ and $[B_2, T_2]$ respectively. If $T_1 < B_2$ then S_2 is clearly better (on this topic) than S_1; and vice versa if $T_2 < B_1$. On the other hand, when $B_1 < B_2 < T_1$ or $B_1 < T_2 < T_1$ (or either of two further symmetric cases) the outcome is inconclusive and there is neither evidence in support the hypothesis that S_1 is better than S_2, nor evidence to contradict it.

Estimating the background probability. It was indicated above that $E = 0.01$ would be used in our experimentation. As a justification for this, consider Figure 1, which shows the fraction of the documents in the TREC-9 Web Track that were judged relevant, categorized by the number of the pool-contributing systems that had included that document in their pool with $d = 100$. For example, seven documents were each identified as being in the top $d = 100$ by 55 different systems, of which five were judged relevant, corresponding to the rightmost bar in the graph.

The average fraction across the graph is 0.202, but is biased by the higher values at the right of the distribution. What is of more interest is the trend line, and where that trend line crosses the y-axis. That "zero requests" value can then be interpreted as being the likelihood of relevance for a document that has not been reported into the pool at $d = 100$ by any of the 59 systems. Fitting a cubic polynomial to the data gives a crossing value of 0.043. Over the whole TREC-9 judgment set (the qrels file), the probability of a document being relevant is 0.038.

Based on these values, $E = 0.01$ is not an overestimate. Note that this is not a claim that a randomly selected document in the entire collection has a 1% chance of being relevant for a randomly selected topic, that is clearly excessive. The claim is that, of the documents selected into the top $d = 100$ by a retrieval system of quality comparable to a mid-range TREC one, of the documents that have not already been judged, around 1% can be expected to be relevant.

Metrics. The estimation methods described above can be applied to all weighted-precision metrics. Other members of this family include *rank-biased precision* (RBP) [12], which in principle has no cutoff k because of the geometric weights that are used, but in practice is evaluated over a finite ranking and hence always has a residual; and *discounted cumulative gain* [11], which, like precision, must be evaluated over a finite prefix. In the results below we use both RBP and *scaled discounted cumulative gain*, SDCG, in which the DCG score is divided by the DCG score that would be achieve by an "all relevant" ranking of that depth, so as to obtain scores bounded above by 1.0.

4 Experimental Investigation

Our goal with this investigation was to determine the extent to which the quality of the outcome of an IR experiment is affected by the factors discussed in the previous sections, namely: the volume of judgments performed; the choice of score estimation technique; and the choice of metric. To measure quality, we adopt the approach that has been employed by a number of authors [13,16,23]. Using TREC data, we compare pairs of systems, and count the fraction of them that yield a significant outcome according to the 50-topic comparison. For any chosen metric, if one experimental regime results in a greater fraction of the system pairs being statistically separable in this way than does another, it is more sensitive.

The experimentation is based around the 105 system submissions over 50 topics that comprise the TREC-9 Web Track [9], and the subset of 59 systems that were used to form the pool for the relevance judgments. The `qrels` file contains ternary judgments rather than binary ones; in our experiment both the "relevant" (category 1) and "highly relevant" (category 2) document were taken to be relevant in a binary sense. The `qrels` file contains $69{,}100$ judgments, of which $2{,}614$ or 3.8%, are "relevant".

Two sets of system pairs were used in the comparisons. In the first set, each of the 59 runs that contributed to the pooling was compared to each of the other 58, as a set of $1{,}711$ system pairs. In the tables and graphs that follow, this set is called "59-con". The second set, "46-non", was generated by applying the same process to the other 46 systems, to create a set of $1{,}035$ system pairs in which neither of the two systems had contributed to the judgment pool. This latter set represents a typical "judgment reuse" situation, in which two non-contributing runs are to be compared in a post-TREC experiment. To evaluate of the effect of pool depth on metric usefulness, we sorted the `qrels` file according to the minimum depth at which each document appeared in any of the 59 contributing runs for each topic, with random ordering applied to ties. This arrangement mimics the effect of pooling and allowed, for example, the first $1{,}000$ `qrels` to be used, simulating a highly resource-limited experiment in which only shallow judgments were undertaken.

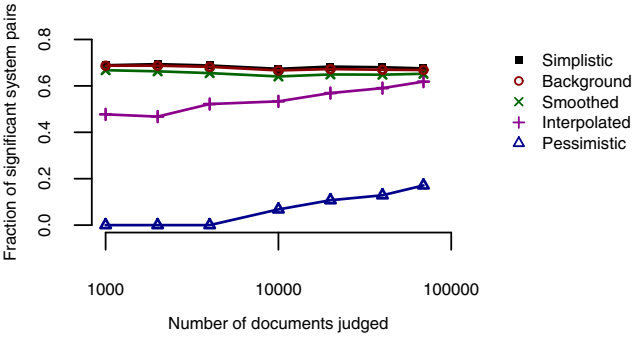
Scores and Residuals. Table 1 gives initial results for this experimental framework. Part (a) of the table shows the average base scores B computed for the two sets of systems, using three different effectiveness metrics, and evaluated using shallow, medium, and deep pooled judgments. In all cases the use of the $X_S = B$ approximation leads to non-decreasing score estimates as the number of judgments employed increases, a useful

Table 1. Base effectiveness scores, residuals, and two point estimates within the $[B, T]$ range, in all cases averaged across 50 topics and a set of system runs, for three different effectiveness metrics and three different judgment sets. In each case, two different sets of topic runs are used, the 59-con runs that led to the TREC-9 judgments; and the other 46-non system runs that did not.

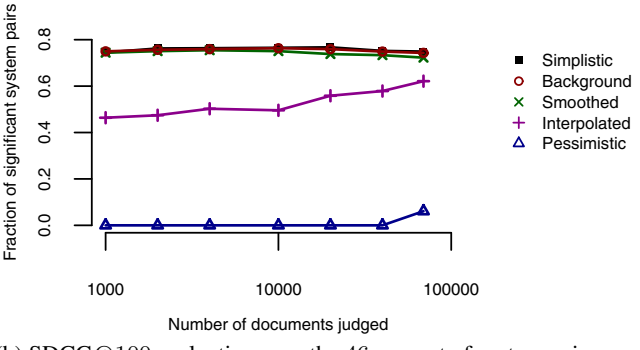
Judgments	P@10		SDCG@100		RBP, $p = 0.95$	
	59-con	46-non	59-con	46-non	59-con	46-non
1,000	0.1407	0.1184	0.0509	0.0419	0.0783	0.0647
10,000	0.2362	0.1877	0.1105	0.0900	0.1592	0.1269
69,100	0.2362	0.1923	0.1351	0.1085	0.1748	0.1398
(a) Effectiveness scores, B						
1,000	0.6181	0.6759	0.8356	0.8535	0.7638	0.7921
10,000	0.0000	0.2692	0.4760	0.5670	0.2810	0.4333
69,100	0.0000	0.1631	0.0000	0.2311	0.0019	0.1955
(b) Residuals resulting from unjudged documents, Δ						
1,000	0.3121	0.2718	0.2970	0.2691	0.3148	0.2820
10,000	0.2362	0.2201	0.1987	0.1836	0.2102	0.1980
69,100	0.2362	0.2035	0.1351	0.1257	0.1751	0.1546
(c) Interpolated scores, X_I						
1,000	0.2112	0.1786	0.0992	0.0832	0.1398	0.1169
10,000	0.2362	0.2025	0.1591	0.1318	0.1963	0.1612
69,100	0.2362	0.1981	0.1351	0.1172	0.1751	0.1470
(d) Smoothed scores, X_M						

behavior; and 10,000 judgments is mostly sufficient to get the X_S scores to within around 10% of the values attained at 69,100 judgments.

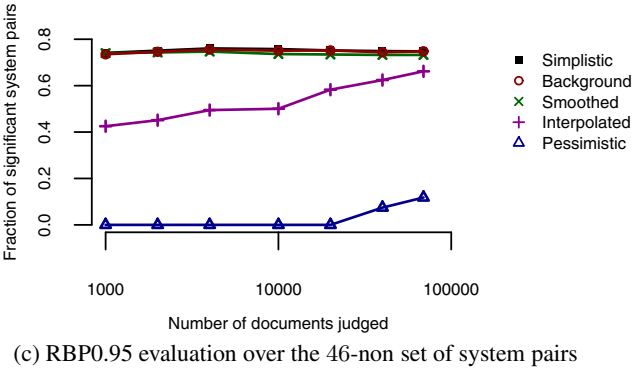
Table 1(b) lists the average residuals Δ associated with those base scores. Unsurprisingly, the 59-con set of systems has smaller average residuals than the 46-non systems, since the documents to be judged to make the partial relevance judgments were chosen from the 59-con runs. Note also that rank-biased precision has non-zero residuals even when all 69,100 judgments are used. A critical observation in Table 1(b) is that on the 46-non systems the residuals are, for the most part, comparable in magnitude to the scores B that they relate to; that is, $T \geq 2B$. It must be concluded that there is considerable uncertainty associated with the 46-non systems, and that the base scores in Table 1(a) – which are also simplistic point values X_S – may not be at all accurate. The third section of Table 1 shows the average of the X_I point estimates. Now the initial estimates based on shallow and medium judgment pools are uniformly overestimates of the final scores. Taking X_I as the point estimate – and assuming that documents at the tail of a run have the same density of relevance as documents at the start of it – is clearly too generous. Finally, Table 1(d) shows the smoothed scores X_M for the same combinations of metrics and judgments. After 1,000 judgments the point estimates are now all below the at-69,100 values; and at 10,000 judgments they are all somewhat higher than the at-69,100 values, indicating a reasonable compromise between the two options, but also perhaps indicating scope for further refinement.



(a) P@10 evaluation over the 46-non set of system pairs



(b) SDCG@100 evaluation over the 46-non set of system pairs



(c) RBP0.95 evaluation over the 46-non set of system pairs

Fig. 2. Separability rates within the 46-non set of systems for three different effectiveness metrics, 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. Use of the t-test is appropriate for this number of topics.

Significance outcomes. Figure 2 shows system separability as the number of judged documents varies, evaluated over the 46-non set of system pairs. The different curves within each graph correspond to different ways (Section 3) of rendering each of the $[B, T]$ ranges into a single score value. The vertical axis records the fraction of the

Table 2. Percentage of system pairs separable at the $p = 0.01$ level by different experimental approaches, for three sets of judgments, two sets of system pairs, and three effectiveness metrics

Judgments	P@10		SDCG@100		RBP, $p = 0.95$	
	59-con	46-non	59-con	46-non	59-con	46-non
1,000	53.0	68.9	62.0	74.5	61.0	74.1
10,000	55.8	67.3	65.5	76.5	65.2	75.7
69,100	55.8	67.5	61.8	74.9	63.8	74.8
(a) Using X_S, and the t-test						
1,000	46.8	47.7	46.3	46.4	44.0	38.6
10,000	55.8	53.3	57.5	49.6	59.6	50.0
69,100	55.8	61.8	61.8	62.1	63.8	61.9
(b) Using X_I, and the t-test						
1,000	53.4	66.8	61.4	74.4	60.8	74.0
10,000	55.8	64.1	63.6	75.1	62.1	73.6
69,100	55.8	65.2	61.8	72.3	63.8	73.2
(c) Using X_M, and the t-test						

system pairs that were significant at $p = 0.01$. The effectiveness metrics used in the three graphs are P@10, SDCG@100, and RBP (with parameter $p = 0.95$).

Except for the pessimistic interval-overlap method of handling the residuals, even as few as 1,000 judgments is sufficient to obtain relatively high rates of system separability. Perhaps surprisingly, it is the simplistic approach X_S and the background approach X_B that yield the greatest separability, followed by the smoothed approach X_M. Note that the different approaches disagree on outcomes even after the "full" set of 69,100 judgments have been used; the equivalent P@10 graph for the 59-con set of runs shows all lines converging by the time 20,000 judgments are being used, because of the top-centric nature of the P@10 metric. On the 59-con set the SDCG@100 curves also converge, but only after the full set of 69,100 judgments.

Table 2 gives more details of these separability coefficients. The Interpolated score estimation approach gives low separability, and is consistently less useful than the other point estimation methods. This inferior behavior is presumably a consequence of the fact that it badly overestimates the actual scores (Table 1). On the other hand, the simplistic X_S approach provides confident assessments even when startlingly few documents have been judged. There are only 223 relevant documents (including 76 highly relevant documents) in the first 1,000 positions of the TREC-9 `qrels` file in the ordering that is used in Figure 2 and Table 2, meaning that with 1,000 judgments, 90% of the relevant documents are *not* part of the comparison.

The success of the X_S approach to score estimation is because pooling ensures that it is tantamount to evaluating the same metric, but at a shallower depth – for example, P@10 on the shallow judgment sets is somewhat similar to evaluating P@3 (see the residuals listed in Table 1), and what is being observed in the separability graphs is that P@3 is a reasonably effective mechanism for separating systems across a set of 50 topics. Similar arguments can be made for SDCG@100 – it is sufficiently well correlated

Table 3. Percentage of significant system pairs after one set of judgments that are not identified as being significant once deeper judgments are applied

Starting with	P@10		SDCG@100		RBP, $p = 0.95$	
	10,000	69,100	10,000	69,100	10,000	69,100
1,000 judgments	7.0	8.4	4.2	7.4	4.6	6.7
10,000 judgments	–	2.2	–	3.9	–	3.1
(a) Using X_S, and the t-test						
1,000 judgments	8.0	8.4	5.0	8.5	4.3	6.6
10,000 judgments	–	1.3	–	4.3	–	3.0
(b) Using X_B, and the t-test						
1,000 judgments	9.7	6.1	17.5	11.5	18.4	8.9
10,000 judgments	–	3.3	–	4.9	–	3.3
(c) Using X_I, and the t-test						
1,000 judgments	9.0	8.1	6.0	10.5	7.7	9.1
10,000 judgments	–	0.9	–	5.8	–	2.6
(d) Using X_M, and the t-test						

with SDCG@10 (say), and the pooling approach to judgment discovery sufficiently well focussed on the top of the system runs that the latter is evaluated reasonable accurately.

Convergence. The simplistic X_S point estimate yields high separability rates even when only shallow judgments are being used, despite the actual effectiveness scores generated being under estimates. An important question then becomes the extent to which the significant pairs that are identified after $1{,}000$ judgments remain significant as more judgments are processed. Table 3 evaluates these relationships, using the 46-non system pairs, the three precision-based metrics, and the shallow, medium, and deep judgments.

To compute each value in Table 3, the set of system pairs that gave t-test p values less than 0.01 according to the test environment noted in the left-most column were then checked again in the context of the test environment recorded in the heading of the other columns. For example, with $1{,}000$ judgments performed, P@10 resulted in 713 system comparisons (of a total of $1{,}035$ system pairs) being deemed significant at the 0.01 level. Of these, 50 (or 7.0%) were *not* identified as being significant when $10{,}000$ judgments were used; and 60 (that is, 8.4%) system pairs were no longer found to be significant when all $69{,}100$ judgments were employed. What is apparent in these results is that both X_S and X_B appear to be relatively stable in their selections of significant system pairs, with less than 10% "recanting" of previous significance as further judgments are employed. The X_I approach has higher revision rates, even though it is more conservative in awarded significance when using the shallower pool depths. The latter effect is particularly marked for the two effectiveness metrics that carry out deep evaluations and are intended to reflect the behavior of very patient searchers.

Other effectiveness metrics. The use of three weighted precision metrics in the various evaluations presented in this paper is deliberate – with each of these three metrics,

discovery of additional relevant documents can only increase the effectiveness score, and so the score that is attained on partial relevance judgments is a lower bound.

But other effectiveness metrics are also amenable to this treatment, with varying degrees of credibility. Average precision (AP), defined as the average of the precision values at the ranks at which relevant documents occur, is particularly challenging. Because it is an average over relevant documents, the discovery of new relevant documents can reduce as well an increase partial scores. A similar observation holds for normalized discounted cumulative gain NDCG [11], in which the scaling factor is the best DCG score attainable given the number of relevant documents available for each topic.

5 Related Work

Other researchers have also considered the issue of partial relevance judgments, and carried out experiments in which TREC (and other) judgments are scaled back and simulated retrieval comparisons carried out. Buckley and Voorhees suggest the use of a modified effectiveness metric denoted as *bpref* in which unjudged documents are bypassed [5]; an approach commented on by Sakai [14] and by Sakai and Kando [15]. The latter work includes discussion and evaluation of a wide range of effectiveness metrics. Aslam and Yilmaz et al. [1,2,22] sample the system runs in order to derive approximate values for average precision and other metrics, and show that the variance of the estimate can be reduced as the number of samples increases. Büttcher et al. [6] consider the related problem of determining and allowing for the bias in favor of pool-contributing systems. Webber and Park [21] have also considered this issue. Bompada et al. [3] consider the similarity of system orderings when compared using incomplete relevance judgments, and demonstrate that partially-evaluated NDCG is more self-consistent than Buckley and Voorhees' bpref metric. Carterette and Smucker [7] quantify the tradeoff between pool depth and pool breadth, and conclude that shallow pooling over many topics is almost certainly more powerful than deep pooling over a restricted set of topics. Our work here, in which shallow judgment pools are demonstrated to still yield significant system comparisons, are a further validation of these various findings.

6 Conclusion

We sought to explore the extent to which the use of incomplete relevance judgments affected retrieval system comparisons. It is clear that the X_S approach of assuming unjudged documents to be irrelevant affects numeric effectiveness scores, and results in values that (for weighted-precision metrics at least) markedly underestimate the true values that would arise from a more costly evaluation. In this sense, it is appropriate to explore other estimation techniques; of the ones considered here, the smoothing approach X_M gives reasonable approximations, but still leaves room for improvement.

When the effectiveness scores are being developed purely as input to a t-test in order to carry out a paired system comparison, the X_S method shed the disadvantage of being inaccurate, and provided consistently reliable outcomes. That is, despite the fact that the scores it produces are a low-fidelity approximation of the eventual scores for that

metric, the relativities observed in the system scores can be relied on, and experimental outcomes reasonably determined.

We next plan to undertake similar experiments with AP and NDCG; with different document orderings for the purposes of creating the judgment set; and using other statistical tests, including in situations in which only small numbers of topics are in use.

Acknowledgment. This work was supported by the Australian Research Council, and by the Government of Malaysia.

References

1. Aslam, J., Yilmaz, E.: Inferring document relevance from incomplete information. In: Proc. 2007 ACM CIKM Conf. Lisbon, Portugal, pp. 603–610 (November 2007)
2. Aslam, J.A., Pavlu, V., Yilmaz, E.: A statistical method for system evaluation using incomplete judgments. In: Proc. 29th ACM SIGIR Conf. Seattle, WA, pp. 541–548 (August 2006)
3. Bompada, T., Chang, C.C., Chen, J., Kumar, R., Shenoy, R.: On the robustness of relevance measures with incomplete judgments. In: Proc. 30th ACM SIGIR Conf. Amsterdam, pp. 359–366 (July 2007)
4. Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In: Proc. 23rd ACM SIGIR Conf. Athens, Greece, pp. 33–40 (July 2000)
5. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proc. 27th ACM SIGIR Conf. Sheffield, England, pp. 25–32 (July 2004)
6. Büttcher, S., Clarke, C.L.A., Yeung, P.C.K., Soboroff, I.: Reliable information retrieval evaluation with incomplete and biased judgements. In: Proc. 30th ACM SIGIR Conf. pp. 63–70 (July 2007)
7. Carterette, B., Smucker, M.D.: Hypothesis testing with incomplete relevance judgments. In: Proc. 2007 ACM CIKM Conf, Lisbon, Portugal, pp. 643–652 (November 2007)
8. Cormack, G.V., Lynam, T.R.: Validity and power of t-test for comparing MAP and GMAP. In: Proc. 30th ACM SIGIR Conf. pp. 753–754 (July 2007)
9. Hawking, D.: Overview of the TREC-9 Web Track. In: Proc. 9th Text REtrieval Conf. (TREC-9). Gaithersburg, Maryland (November 2000)
10. Huffman, S.B., Hochster, M.: How well does result relevance predict session satisfaction? In: Proc. 30th ACM SIGIR Conf. Amsterdam, pp. 567–574 (July 2007)
11. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), 422–446 (2002)
12. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems 27(1), 1–27 (2008)
13. Sakai, T.: Evaluating evaluation metrics based on the bootstrap. In: Proc. 29th ACM SIGIR Conf. Seattle, WA, pp. 525–534 (August 2006)
14. Sakai, T.: Alternatives to Bpref. In: Proc. 30th ACM SIGIR Conf, Amsterdam, pp. 71–78 (July 2007)
15. Sakai, T., Kando, N.: On information retrieval metrics designed for evaluation with incomplete relevance assessments. Information Retrieval 11(5), 447–470 (2008)
16. Sanderson, M., Zobel, J.: Information retrieval system evaluation: Effort, sensitivity, and reliability. In: Proc. 28th ACM SIGIR Conf. Salvador, Brazil, pp. 162–169 (August 2005)
17. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval. In: Proc. 2007 ACM CIKM Conf, Lisbon, pp. 623–632 (November 2007)

18. Smucker, M.D., Allan, J., Carterette, B.: Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In: Proc. 32nd ACM SIGIR Conf. Boston, MA, pp. 630–631 (July 2009)
19. Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: Proc. 29th ACM SIGIR Conf. pp. 11–18 (August 2006)
20. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. The MIT Press, Cambridge (2005)
21. Webber, W., Park, L.A.F.: Score adjustment for correction of pooling bias. In: Proc. 32nd ACM SIGIR Conf. Boston, MA, pp. 444–451 (July 2009)
22. Yilmaz, E., Kanoulas, E., Aslam, J.A.: A simple and efficient sampling method for estimating AP and NDCG. In: Proc. 31st ACM SIGIR Conf. Singapore, pp. 603–610 (July 2008)
23. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Proc. 21st ACM SIGIR Conf. Melbourne, Australia, pp. 307–314 (August 1998)