The Effect of Pooling and Evaluation Depth on Metric Stability

William Webber Alistair Moffat Justin Zobel

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

ABSTRACT

The profusion of information retrieval effectiveness metrics has inspired the development of meta-evaluative criteria for choosing between them. One such criterion is discriminative power; that is, the proportion of system pairs whose difference in effectiveness is found statistically significant. Studies of discriminative power frequently find normalized discounted cumulative gain (nDCG) to be the most discriminative metric, but there has been no satisfactory explanation of which feature makes it so discriminative. In this paper, we examine the discriminative power of nDCG and several other metrics under different evaluation and pooling depths, and with different forms of score normalization. We find that evaluation depth is more important to metric behaviour and discriminative power than metric type; that evaluating beyond pooling depth does not seem to lead to a misleading system reinforcement effect; and that nDCG does seem to have a genuine, albeit slight, edge in discriminative power under a range of conditions.

1. INTRODUCTION

The field of information retrieval has produced a wide variety of evaluation metrics. This fecundity raises the question of how to choose between them, inspiring the field of metric *meta-evaluation*. Several meta-evaluative criteria have been proposed, most notably that of the metric's *stability*; that is, the consistency of a metric's results from one set of topics to another. Various measures of stability have been proposed. One of these is the metric's *discriminative power*; that is, the proportion of system pairs whose difference in effectiveness is found statistically significant.

Discriminative power is desirable because it allows the experimenter to achieve reliable results with fewer topics, and so reduced effort. Statistical consistency, however, does not necessarily imply consistency in measuring true system effectiveness; it could instead arise from consistently measuring some extraneous property of the evaluation. An evaluation metric that ranked systems by name would be perfectly consistent and highly discriminative across all topics, but would hardly discern superior retrieval effectiveness. More practically, retrieval evaluation is often performed on unpooled systems, and beyond pooling depth even on pooled ones. Under these conditions, a metric's stability might be caused by system reinforcement, with similar systems drawing each others' unpooled documents into the pool.

Measures of stability repeatedly identify certain metrics as more stable than others. In particular, *normalized discounted cumulative gain* (nDCG) and *average precision* (AP) appear more stable than *rank-biased precision* (RBP), and all three of these more stable than *precision at ten*. The divergence between nDCG and RBP is of particular interest, since these are in some respects similar, *rank-weighted* metrics. They differ in that, first, nDCG is a *normalized* metric, RBP an unnormalized one; and second, RBP weights decline in consistent proportion, whereas nDCG is steep at high ranks, but flat at lower ones – exactly those ranks covered by evaluating beyond pooling depth. This suggests three hypotheses that may explain nDCG's greater stability over RBP:

1. It is normalization that makes nDCG more stable. This can be tested by normalizing RBP, and unnormalizing DCG.

2. DCG is more stable because its heavier tail picks up valid information about system performance. This can be explored by increasing RBP's p parameter, making it a deeper metric, and by evaluating DCG more shallowly.

3. The greater stability of DCG comes from evaluation beyond pooling depth causing system reinforcement, which DCG's heavy tail is consistently misled by. This can be investigated by varying evaluation and pooling depth together.

This paper probes these hypotheses, measuring the discriminative power of a variety of metrics, normalizations, and evaluation and pooling depths. The investigation illuminates the relative benefits of normalization, depth of evaluation, and depth of pooling, for metric stability.

2. PREVIOUS WORK

While there was some earlier, mostly theoretical discussion of the merits of different metrics [van Rijsbergen, 1979, Cooper, 1968], the real impetus for the development and empirical evaluation of new metrics came with the inauguration of the TREC effort in 1992 [Voorhees and Harman, 2005], which for the first time provided the large-scale runsets needed for such empirical analysis. The most widely adopted of the metrics developed at TREC is *average precision* (AP). The main justification for AP was its relationship to the traditional precision–recall curves [Buckley and Voorhees, 2005]. The combination of recall and precision in the one metric makes analysis and extension complex; in particular, extending AP to handle multi-grade relevance is possible but not straightforward. Instead, Järvelin and Kekäläinen [2002] propose the *discounted cumulative gain* (DCG) metric, where each rank has a fixed weight (or discount), which is multiplied with the multi-graded relevance (or gain) of the document at that rank. Järvelin and Kekäläinen further propose the use of *normalization*, to make each per-topic score relative to the ideal score achievable on that topic, leading to *normalized discounted cumulative gain* (nDCG). Normalization (or alternatively simple scaling based on cutoff depth) is required to

EVIA'10, June 15, 2010, Tokyo, Japan.

bring DCG scores within a defined range, because the inverse logarithm sequence of rank weights is not convergent. Moffat and Zobel [2008] propose instead to use a convergent, geometric weighting sequence, in their *rank-biased precision* (RBP) metric. Convergent weights mean that a partial evaluation bounds the scores achievable on a full one, and the uncertainty resulting from unassessed documents can be quantified as an error residual.

The TREC forum and data also provoked interest in the evaluation and comparison of metrics themselves. An early meta-evaluative study is that of Zobel [1998], who investigates the reliability of significance tests via split-topic experiments, as well as the effect of different pooling depths and of assessment bias against unpooled systems. The agreement of several different metrics is measured, though the stability of different measures is not directly determined. A more explicit measure of metric stability is introduced by Buckley and Voorhees [2000], in the form of the *swap rate*, which is the frequency of an ordering of two systems on one topic set being reversed on another. The swap rate of several metrics is compared, with recall and precision at 1,000 (R@1000, P@1000) showing the greatest stability (lowest swap rate), followed closely by R-precision and average precision, while shallower metrics such as precision at ten have the lowest stability. The effect of the number of topics is investigated, but not of different pooling depths. Alternative forms of the swap rate metric are developed by Voorhees and Buckley [2002] and Sanderson and Zobel [2005].

Another measure of metric stability, introduced by Sakai [2006], is *discriminative power*, defined as the proportion of system pairs whose difference in effectiveness is found statistically significant. Sakai employs the bootstrap test, but other tests can be used instead. Sakai compares the discriminative power of several metrics, including AP, P@1000, nDCG@1000, and his own Q-measure. He finds Q-measure, AP, and nDCG the most discriminative metrics, and P@1000 the least. The effect of pooling or evaluation depth is not directly addressed. Discriminative power is also used in studies of pooling bias in Sakai [2007] and Sakai [2008].

An alternative approach to meta-evaluation is that of *test theory* [Bodoff and Li, 2007]. Test theory was developed for analysing exam tests applied to human subjects, in which a large number of examinees are evaluated against the same set of test items. This description applies tolerably well to collaborative retrieval experiments, in which the examinees are retrieval systems, and the test items are topics. Test theory analyses are used by Carterette et al. [2008] to measure the stability of document sampling and score estimation methods, and by Kanoulas and Aslam [2009] as an objective function against which nDCG parameters can be tuned.

A final measure of stability is *predictivity*, proposed in Webber et al. [2008b]. Predictivity is the similarity of system rankings on one random set of topics compared to those on another, measured using Kendall's τ. The predictivity of a metric to itself can be measured, but also of one metric to another. Webber et al. find AP and nDCG to be the most predictive metrics, with RBP ($p = 0.95$) less predictive, and P@10 and reciprocal rank the least predictive.

3. MATERIALS AND METHODS

We begin this section by introducing the evaluation metrics examined in this paper. Next, we describe discriminative power as a measure of metric stability. Finally, we introduce the datasets this paper employs.

3.1 Evaluation metrics

An IR system is evaluated by having it run a set of queries or *topics* against a document *corpus*, returning a ranked list of documents in answer to each topic. Each document ranking is marked up for relevance, to produce a relevance vector, $\mathcal{R} = (r_1, r_2, \ldots)$. Relevance assessments are often performed in advance, and recorded as *qrels*; topics, corpus, and qrels together constitute a *test collection*. Relevance can be graded, but binary relevance is assumed here.

Exhaustive assessment of all documents for relevance to every query is not feasible. The standard solution is to *pool* and assess the top d ranked documents returned by a set of participating systems. A common *pool depth* is 100. Under pooling, unassessed documents are assumed to be irrelevant. This can lead to a bias against systems that did not contribute to the pool. Similarly, if *evaluation depth* is greater than pooled depth, there can be a bias in favour of similar systems, which drag each others' post-pool documents into the pool. The standard evaluation depth at TREC is $k = 1{,}000$.

The function of an *evaluation metric* is to convert the marked-up relevance vector $\mathcal{R}$ into a single score. Two fundamental concepts in evaluation metrics are *precision* and *recall*. Precision is the proportion of (a given set of) returned documents that are relevant; recall is the proportion of relevant documents that are returned. Calculating recall requires that the number R of relevant documents be determined. Under pooling, R is set to the number of relevant documents found in the pool; deeper pooling (or more pooled systems) will increase R, and therefore change recall values.

To evaluate precision or recall on a ranked list, the list is truncated to an *evaluation depth*, k. *Precision at depth* k is:

$$\text{P@}k(\mathcal{R}) = \sum_{i=1}^{k} \frac{r_i}{k} \ , \tag{1}$$

while *recall at depth* k is:

$$\text{R@}k(\mathcal{R}) = \sum_{i=1}^{k} \frac{r_i}{R} \ . \tag{2}$$

The metrics differ solely in the divisor. If $k > R$, then P@$k < 1.0$; conversely, if $k < R$, R@$k < 1.0$. Thus both metrics are subject to compressed score ranges for extreme disparities of k and R.

Average precision (AP) combines elements of both precision and recall. AP is calculated as the average of the precision at each rank that a relevant document is returned; unreturned (known) relevant documents receive a precision of 0. More formally:

$$\text{AP@}k(\mathcal{R}) = \frac{1}{R} \sum_{i=1}^{k} r_i \cdot \text{P@}i(\mathcal{R}) \ . \tag{3}$$

Again, if $R > k$, then AP@$k < 1.0$, meaning that AP scores can become compressed when evaluation is shallow or the number of known relevant documents is large. An alternative is to *abbreviate* the normalizing constant to the minimum of k and R:

$$\text{aAP@}k(\mathcal{R}) = \frac{1}{\min(k, R)} \sum_{i=1}^{k} r_i \cdot \text{P@}i(\mathcal{R}) \ . \tag{4}$$

This latter (non-standard) form is evaluated later.

Average precision assumes binary relevance. An alternative metric, designed in part for graded relevance (though not so used here), is *discounted cumulative gain* (DCG) [Järvelin and Kekäläinen, 2002]. In DCG, a ranking's score is the dot product of its relevance vector with an inverse logarithmic weighting vector, as follows:

$$\text{DCG@}k(\mathcal{R}) = \sum_{i=1}^{k} r_i / \log_2(\min(2, i)) \ . \tag{5}$$

(An alternative formulation divides by $\log_2(i + 1)$, but we follow the original formulation here.)

Because of its non-convergent weights, DCG is not bounded above by 1.0. Instead, the maximum achievable score for a topic varies with the number of relevant documents, impeding score comparison between topics. Järvelin and Kekäläinen [2002] deal with this by *normalizing* the metric, producing *normalized DCG* (nDCG). An ideal ranking for the topic is created, based on the known relevant documents; for binary relevance, this is a ranking with R relevant documents on top, truncated to depth k if $R > k$. The DCG score of the ideal vector is calculated, and the score of each actual document ranking for that topic is divided by the ideal score to arrive at the normalized score. Normalization can be similarly applied to most other metrics. Indeed, AP can be regarded as normalized sum of precisions, and R@k as normalized relevant-returned at k, except that, in the standard forms of AP and recall, the ideal ranking is not truncated at depth k. Untruncated ideal rankings will be referred to as *expanded*. One can consider a variant of nDCG of this form, labeled enDCG.

While DCG uses a non-convergent weighting scheme, the *rank-biased precision* metric (RBP) uses a convergent, geometric one [Moffat and Zobel, 2008]. The metric is based on a simple user model: the user, having reached any given rank in the ranking, has probability p of continuing to the next rank, where $0 \leq p < 1$ measures the user's persistence. Larger values of p lead to deeper evaluation. The full formulation is:

$$\text{RBP}(\mathcal{R}, p) = (1 - p) \sum_{i=1}^{\infty} r_i \cdot p^{(i-1)} \; . \tag{6}$$

Because of its convergent weights, RBP is not dependent on the evaluation depth k. Partial evaluation sets bounds, as a base score and a residual, on the score that a full evaluation would achieve. The residual is the sum of the weights of the unassessed documents; even to infinite depth, this sum is finite. Here, we will using RBP base values, by calculating Equation 6 to depth k. In some experiments, the p parameter will be fixed; in others, it will be adjusted so that the residual at depth k (that is, the sum of the weights from $k + 1$ to ∞) is 0.1. The formula for calculating the parameter p which will provide a residual r for evaluation to depth k is:

$$p = r^{(1/k)} \tag{7}$$

A residual of 0.1 is achieved at depth $k = 10$ evaluation by $p = 0.8$, at $k = 100$ by $p = 0.977$, and at $k = 1{,}000$ by $p = 0.9977$.

DCG and RBP differ chiefly in their rank weightings. The DCG weights decline steeply at high ranks, but are almost flat thereafter, while those for RBP decline by the same proportion at each rank, as can be seen in Figure 1. For shallow evaluation, DCG is top-weighted, whereas for deep evaluation, DCG is fat-tailed. For evaluation depth $k = 1{,}000$ and pooling depth $d = 100$, 82% of DCG's total rank weight falls between depth 101 to 1,000. In contrast, only half a percent of the weight of RBP with p set to 0.95 – the highest value considered by Moffat and Zobel [2008] – falls in this range. Raising p to 0.9977, for a residual of 0.1 beyond depth 1,000, places 77% of base RBP's weight between ranks 101 and 1,000, but even within this range, RBP is more top-weighted than that of DCG.

Normalization is one method of adjusting scores for the difficulty of a topic; an alternative is *standardization* [Webber et al., 2008a]. Under standardization, a topic's difficulty is estimated from the scores a set of reference systems on receive on that topic, and the raw metric scores achieved by systems against that topic are adjusted accordingly. If topic t produces a mean score of μ_t from the reference systems, and a score standard deviation of σ_t; and if a run s receives a raw score of X_{st} for topic t, then the normalized score X'_{st} for that run is:

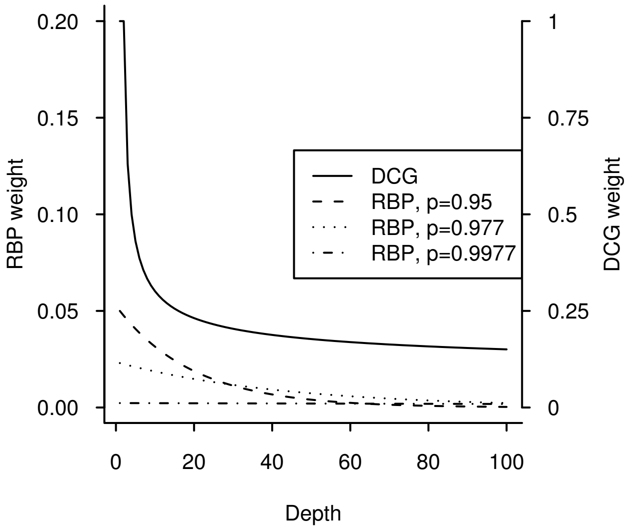


Figure 1: Rank weights of DCG and of RBP with various p values. Note the different scales for RBP (left axis) and DCG (right axis).

$$X'_{st} = \frac{X_{st} - \mu_t}{\sigma_t} \tag{8}$$

The standardized scores reported in this paper are *self-standardized*; that is, the set of systems whose scores are standardized (the original TREC runsets) also serves as the reference set used to derive the standardization factors.

3.2 Stability and associated measures

Discriminative power measures the proportion of system pairs from a set of systems found to be significantly different. The significance test employed here is a paired, two-tailed t test, at significance level $\alpha = 0.05$. The smallest topic set dealt with here is 49 topics (TREC 2004 Robust new topics), large enough to justify the use of the t test, based on the central limit theorem. Previous work has demonstrated that, in practice, the t test gives very similar results to the non-parametric bootstrap and randomized permutation tests [Smucker et al., 2007]. A metric that finds more system pairs significantly different is said to be more highly discriminative. This is a desirable statistical property, in that it means that significance can be achieved with fewer topics. Under the random sampling hypothesis, significance means that a result achieved on one topic set is likely to be true of the full population of topics, and is therefore likely to be replicated in another randomly sampled topic set; it follows that discriminative power is also a measure of consistency, predictivity, and in general stability.

3.3 Data set

The primary data set used in our experiments is the set of runs submitted to the AdHoc task of TREC 8, along with the qrel set from that year's collection. A total of 129 runs, made across the 50 topics of the test collection, were submitted by 40 different research groups. Up to 2 runs from each group were pooled, for a total of 71 pooled runs. Pooling was performed to depth 100, and the original evaluation was performed to depth 1,000. Some 13 of the runs were manual runs; the 11 best of these manual runs are the best 11 systems overall as measured by both P@10 and AP@10; one of these manual runs, however, drops several ranks for AP@1000.

Metric	T5 AH	T8 AH	T01 Web	T04 Rob	T05 TB	mean
P@10	0.628	0.645	0.594	0.516	0.555	0.588
RBP, p=0.8	0.638	0.657	0.602	0.517	0.562	0.595
RBP, p=0.95	0.661	0.691	0.627	0.598	0.658	0.647
AP@1000	0.638	**0.725**	0.627	**0.680**	0.748	0.683
nDCG@1000	**0.693**	0.718	**0.673**	0.673	**0.762**	**0.704**
mean	0.651	0.687	0.624	0.597	0.657	0.643

Table 1: Discriminative power of standard metrics on different TREC collections. The most discriminative metric for each collection is highlighted.

We use four other TREC runsets for illustrative purposes. These are the TREC 5 AdHoc task runset (61 systems); the TREC 2001 Web track runset (97 systems); the TREC 2004 Robust track runset, on the 49 new topics, Topics 651–700, only (110 systems); and the TREC 2005 Terabyte runset (58 systems). These are similar in makeup to the TREC 8 AdHoc runset, except that the TREC 2004 Robust runset lacks manual runs, and the TREC 2005 Terabyte contains only two. The Terabyte runs return up to 10,000 documents, but here evaluation is limited to depth 1,000. The "relevant" and "highly relevant" judgments of the ternary relevance schemes in the post-AdHoc collections are folded to (binary) "relevant".

In our experiments, all the runs of each runset are used. Some researchers remove the worst-performing 25% of systems, to exclude defective runs, which consistently achieve the lowest scores [Voorhees and Buckley, 2002]; however, we do not follow this practice, since we are concerned only with the relative, rather than the absolute, stability of the metrics employed. Retaining weak runs does inflate discriminative power figures, since almost all strong systems are significantly better than almost all very weak ones; under these conditions, apparently small changes in discriminative power can represent quite large changes in practical results.

4. ANALYSIS

Table 1 shows the results of a typical comparison of metric stability, using the discriminative power measure. Five metrics (two of them variants of RBP) are tested for discriminative power across five different TREC runsets. Discriminative power varies from runset to runset, and the most discriminative metric alternates between nDCG and AP. But in every case, nDCG is more (often much more) discriminative than RBP, $p = 0.95$, while the shallower metrics P@10 and RBP, $p = 0.8$ are uniformly the least discriminative metrics. The rest of this paper investigates the causes of the higher discriminative power of nDCG, in particular its relationship with depth of evaluation, depth of pooling, and normalization.

4.1 Depth and similarity

We begin by examining the effect of evaluation depth upon metric behaviour. Section 3.1 made explicit the dependence that metrics (other than RBP) have upon evaluation depth. Is this dependence a real or merely a formal one? Do metrics display noticeably different behaviours at different evaluation depths?

Shallow versus deep evaluation

The similarity between two metrics (or the one metric evaluated to two depths) can be measured by the Kendall's τ correlation between the system rankings each metric produces. Table 2 shows the correlation, on the TREC 8 AdHoc runset, between various metrics, at evaluation depths of 10 and 1,000. The depth 10 metrics are more similar to themselves as a group (mean τ of 0.905) than depth 1,000 metrics are to themselves (mean τ of 0.860), and both are much more self-similar than depth 10 metrics are to depth 1,000 metrics (mean τ of 0.767). More importantly, a given metric at depth 10 is more similar to other metrics at depth 10 than it is to the same metric at depth 1,000, and the depth 1,000 metric is more similar to other depth 1,000 metrics than it is to itself at depth 10. Indeed, depth 10 metrics are scarcely more similar to themselves at depth 1,000 than they are to other metrics at depth 1,000 (mean τ of 0.773, across the bold values, compared to a mean τ of 0.767 for the upper right quadrant as a whole). The only, partial exception to these statements is AP; AP@1000 is as close to AP@10 as it is to other @1000 metrics, possibly indicating a top-weighted metric.

Table 2 demonstrates that metric dependence on evaluation depth is more than merely formal. Depth of evaluation is an essential component of most metrics, and must be explicitly stated in reporting metric results, just as the p value of the RBP metric is. The results also reflect nDCG's peculiar weighting, top-weighted at the beginning but flat at later depths. This is reflected in Table 2 by the pronounced difference between nDCG@10 and nDCG@1000. The former is almost identical, in the ranking it produces, to RBP $p = 0.8$, which is a similarly top-weighted metric; the latter to the quite flat-weighted RBP $p = 0.9977$. If nothing else, findings about the stability of nDCG@1000 should not directly be used to support the use of nDCG@10.

Evaluating beyond pooling depth

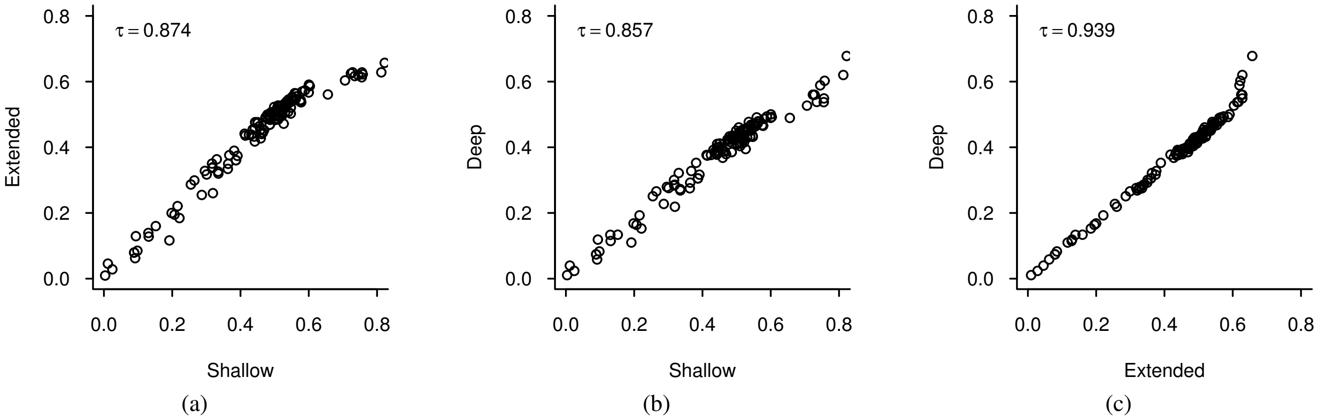
Table 2 compared evaluation and pooling to depth 10, with depth 1,000 evaluation on depth 100 pooling, the latter being the standard TREC treatment. It is interesting to tease out the relationship between evaluation and pooling depth and their effect on metrics in some more detail. Figure 2 does this for the nDCG metric, at various combinations of evaluation and pooling depth, on the TREC 8 runset. We refer to pooling and evaluating to depth 10 as *shallow*; pooling and evaluating to depth 100 as *deep*; and pooling to depth 10 but evaluating to depth 100 as *extended*. The right-hand figure shows that the relationship between deep and extended, viewed across the full set of systems, is very strong, with a Kendall's τ of 0.939. This is stronger than the relationship between shallow and extended (left, $\tau = 0.874$), and even more so than the relationship between shallow and deep (middle, $\tau = 0.857$). This indicates that, for this runset at least, extending evaluation to depth 100 on a depth 10 pool gives a reliable indication of full assessment to depth 100, thus providing evidence against the hypothesis that nDCG is misled by system reinforcement effects.

A confound to this interpretation of the reliability of extended evaluation is offered by the top-ranked runs. The second-ranked run under deep evaluation is also second under shallow evaluation; however, it drops back to being tied with three or four other systems under extended evaluation. Indeed, for pooling to depth 100 and evaluation to depth 1,000 (not shown) this system falls back into the middle of the ranking. The other top-ranked systems also show themselves to be impeded by extended evaluation. These top-ranked runs are manual runs, with distinctive results (large numbers of uniquely relevant documents), so it is not surprising that they are relatively penalised by the reinforcement effect of evaluating beyond pooling depth. Given that these runs are arguably the most important, lack of reliability on them is a significant issue.

The results from Figure 2 are, overall, reassuring for the practice of evaluating beyond pooling depth. Even for a shallow, depth 10 pool, evaluating to depth 100 seems to give quite reliable results, more reliable than restricting evaluation to depth 10. Caution is required in interpreting these results, however. First, the pool for

	R@	AP@	nDCG	RBP	P@	R@	AP@	nDCG	RBP
	10	10	@10	.8	1000	1000	1000	@1000	.9977
P@10	0.88	0.90	0.94	0.93	**0.74**	0.69	0.83	0.83	0.80
R@10		0.90	0.86	0.86	0.71	**0.68**	0.83	0.82	0.77
AP@10			0.90	0.90	0.73	0.70	**0.86**	0.85	0.79
nDCG@10				0.98	0.71	0.66	0.80	**0.81**	0.78
RBP.8					0.72	0.67	0.81	0.81	**0.79**
P@1000						0.88	0.81	0.85	0.90
R@1000							0.79	0.84	0.82
AP@1000								0.91	0.88
nDCG@1000									0.91

Table 2: Kendall's τ between system rankings on the TREC 8 AdHoc track participant systems, using different metrics. Five metrics are shown, and for each metric, evaluation to depth 10 and to depth 1,000. For RBP, p values which give a residual of 0.1 at these evaluation depths are chosen. Comparisons between a metric at evaluation depth 10 and the same metric at evaluation depth 1,000 are in bold. Pooling is performed to depth 10 for the depth 10 evaluation, and to depth 100 for the depth 1,000 evaluation.



(a)

(b)

(c)

Figure 2: Relationship of system mean nDCG scores at different pooling and evaluation depths, for the TREC 8 AdHoc runset.

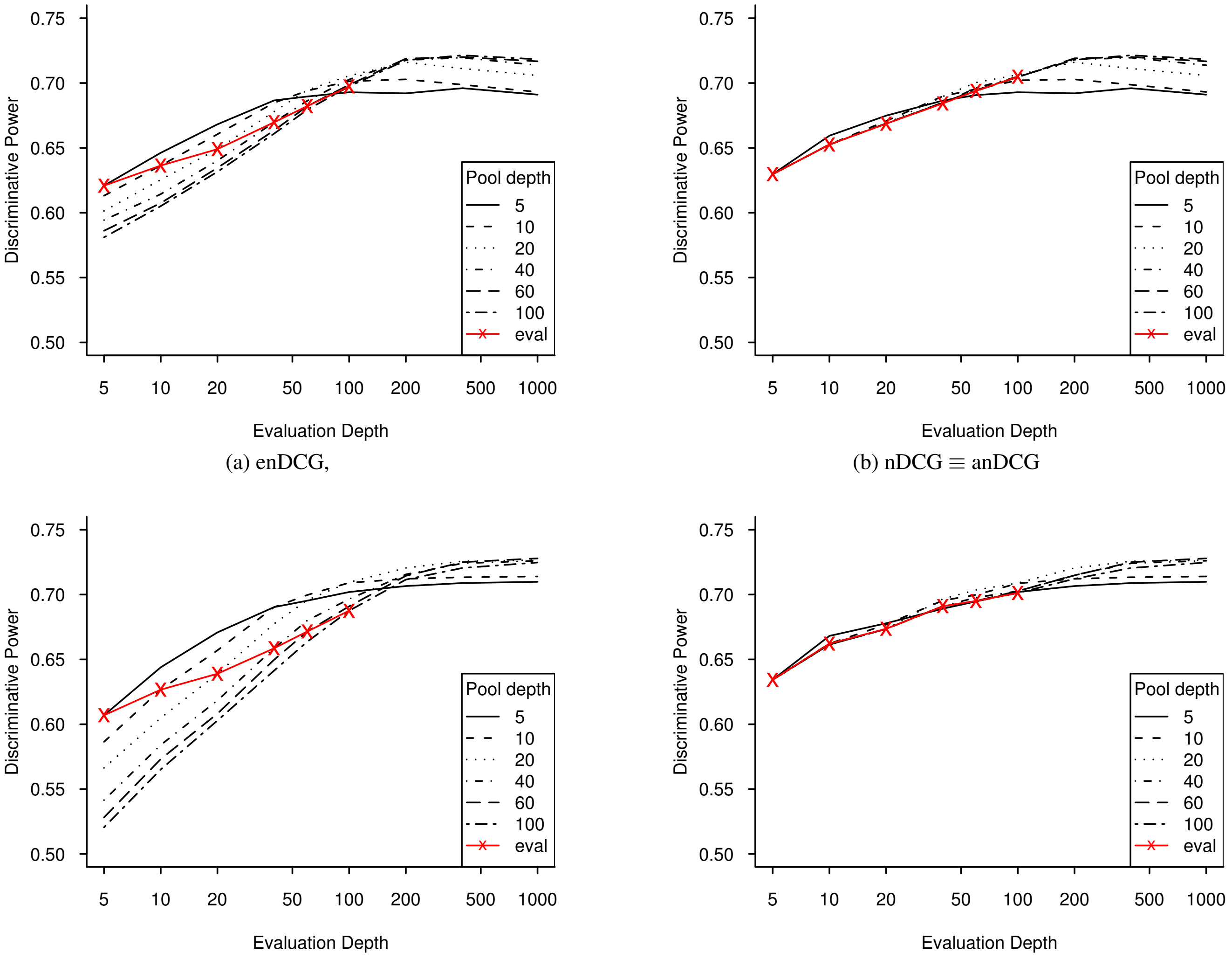
TREC 8 is a very wide one, containing many systems; the results may be less reliable for narrower pools. Second, the TREC 8 collection is, by current standards, not a large one; coverage of the set of relevant documents will be quite high compared to corpus of the size of the Terabyte collection, and so confirmation bias from extended evaluation may be weaker. And finally, the impedance experienced by the top-ranked, distinctive, manual systems under extended evaluation shows that, even with relatively comprehensive pools, distinctive systems (which are generally those that the researcher is most interested in) can suffer from confirmation bias.

4.2 Depth and stability

We now explore the effect of evaluation and pooling depth upon metric stability, as measured by discriminative power. We vary pooling depth d from 5 to 100, and the evaluation depth k from 5 to 1000. Having pooling depth greater than evaluation depth changes the normalization factors for normalized metrics, but has no effect on unnormalized metrics.

Normalization cutoff

As mentioned in Section 3.1, R normalization can be performed in one of two ways. The ideal ranking could extend to depth R, regardless of the metric cutoff depth k, referred to as *expanded normalization*; or else the ideal ranking could be cutoff at depth k, referred to as *abbreviated normalization*. Expanded normalization is standard for recall and AP precision, abbreviated normalization for nDCG. Figure 3 shows the discriminative power for each of these methods under different pooling and evaluation depths. The black lines are pooling depths; the crossed red line marks full assessment, where evaluation and pooling depth are identical. The left-hand figures show the behaviour of expanded normalization, the right-hand figures of abbreviated normalization. For expanded normalization, the pool depth lines approach the full assessment line from below, indicating that pooling beyond assessment depth hurts discriminative power. This is because deeper pooling increases the value of R, without increasing the ability of runs to retrieve more relevant documents. The range of score deltas for topics in which R exceeds k becomes increasingly squashed, inflating the variability of per-topic deltas, and making significance harder to achieve. Similarly, for expanded normalization, the pool lines rise substantially above the full-assessment curve, indicating that for a fixed evaluation depth, pooling to less than evaluation depth boosts discriminative power; perversely, less information gives better results (though note that we do not here consider reliability of discrimination). Again, this is mostly because shallower pooling allows k to exceed R for more topics. In contrast, abbreviated normalization shows no damage from pooling beyond evaluation depth, and only very slight benefit from holding pooling depth below evaluation



(a) enDCG,

(b) nDCG ≡ anDCG

(c) AP ≡ eAP

(d) aAP

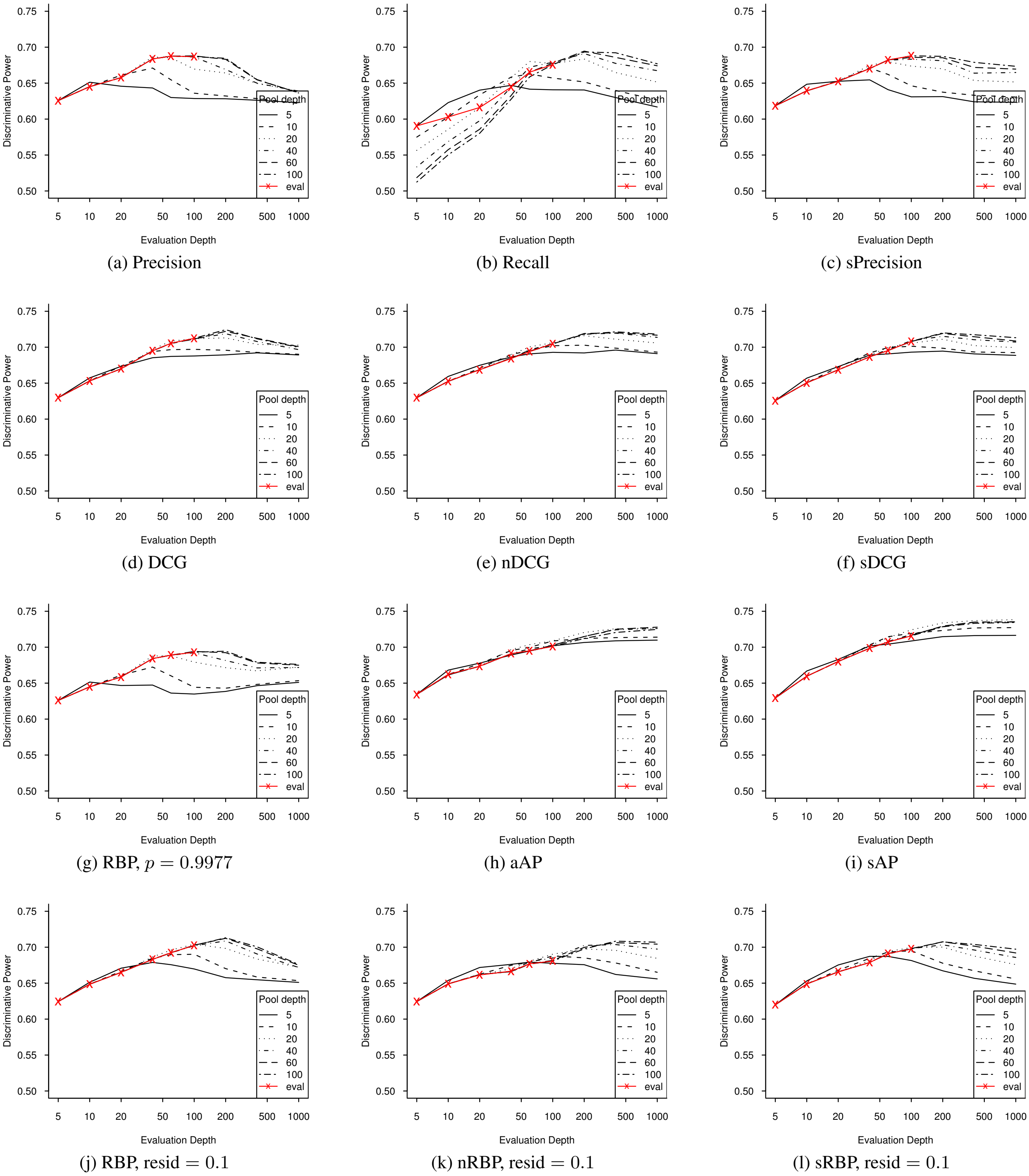
Figure 3: Effects of expanded versus abbreviated normalization on discriminative power, for the AP and nDCG metrics. The data is the TREC 8 AdHoc runset. Each curve shows discriminative power for a given pool depth, as the evaluation depth is varied. The crossed red line shows discriminative power when evaluation and pool depth are the same; this represents full assessment.

depth. Abbreviated normalization is also, depth for depth, equal to or greater than expanded normalization in discriminative power.

These results demonstrate that abbreviated normalization is to be preferred to expanded normalization, at least for metric stability. This suggests that the standard formulation for average precision, which uses expanded normalization, should be modified to use abbreviated normalization, as given in Equation 4 in Section 3.1, at least in situations where evaluation is depth is likely to be less than the number of relevant documents for a topic. Note that 20 of the 50 topics in TREC 8 have more than 100 relevant documents, making the question a pertinent one. Conversely, Figure 3 gives support to the common practice at TREC of evaluating beyond pooling depth (frequently performed as depth 1,000 evaluation on depth 100 pooling). We use abbreviated AP in the following experiments. We continue, however, to use the standard (expanded-normalization) formulation of recall. (Abbreviated normalization would likely also benefit the stability of recall, but such a metric would be quite far removed from recall as generally understood. In particular, abbreviated-normalized recall is identical to precision when $R \geq k$.)

Discriminative power

In this section, we examine the effect that varying pooling and evaluation depth has upon the discriminative power of the precision, recall, DCG, AP, and RBP metrics. We consider the raw, normalized, and standardized forms of these metrics. Figure 4 shows the results of this analysis, on the TREC 8 AdHoc runset. The first conclusion from these results is that, at least on this data set, discriminative power under full assessment (the crossed red line) is affected more by evaluation depth than it is by choice of metric. For all metrics, discriminative power rises strongly with evaluation depth, and the difference in discriminative power between any metric at depth 100 and that metric at depth 5 or 10 is greater than the difference in discriminative power between different metrics at the same depth. So, for instance, the strongest full-assessment discriminative power at depth 100 is that of standardized AP, at 0.716, while the weakest is



(a) Precision

(b) Recall

(c) sPrecision

(d) DCG

(e) nDCG

(f) sDCG

(g) RBP, $p = 0.9977$

(h) aAP

(i) sAP

(j) RBP, resid $= 0.1$

(k) nRBP, resid $= 0.1$

(l) sRBP, resid $= 0.1$

Figure 4: Discriminative power of different metrics for evaluation beyond (and before) pooling depth, at various pool depths, on the TREC 8 runset. The RBP p parameter is varied so that the residual at the specified depth is 0.1. The right column shows standardized metrics.

recall, at 0.676, a gap of 0.040. In comparison, the discriminative power of R@10 is 0.603, a fall of 0.073, while that of standardized AP@10 is 0.660, a fall of 0.056. A second point to note from Figure 4 is that only recall strongly displays the perverse behaviour observed for the expanded-normalized metrics in Figure 3; that is, for a fixed evaluation depth, of falling discriminative power with increased pooling depth.

Regarding the interaction between pooling and evaluation depth, it is notable in Figure 4 that normalized and standardized DCG and AP display behaviour of almost entirely non-decreasing discriminative power with evaluation depth. Even with depth 5 pooling, the discriminative power of these metrics is increasing, or at worst levelling off, all the way up to evaluation depth $1{,}000$. This striking result is difficult to explain away. It cannot adequately be attributed to a reinforcement effect beyond pooling depth, because a similar effect should be observed with P@1000 or with RBP, $p = 0.9977$. On the contrary, both RBP and precision show a fall in discriminative power when evaluation is pushed far beyond pooling depth. At the same time, although normalization (or standardization) does seem to help slightly (as can be seen by comparing DCG to nDCG), the result cannot be attributed solely to the effect of normalization, since normalized and standardized RBP, and standardized precision, still display a fall in discriminative power as evaluation depth is pushed to the limit. Too much weight should not, of course, be placed on the results of a single data set. Nevertheless, this is initial evidence that the particular weighting scheme explicitly adopted for DCG, and the one adopted implicitly for AP, do give these metrics a benefit in discriminative power.

Examining the question of evaluation beyond pooling depth in Figure 4, it is interesting, and perhaps a little surprising, to note that for most of the metrics examined, except for the precision/recall row, discriminative power tends to rise with increased evaluation depth even if pooling depth is not increased. Indeed, in many case it rises faster than if pooling depth is increased, too, if only by a slight amount. So, for instance, for nDCG, with the evaluation depth fixed to depth 20, pooling to depth 20 (that is, full assessment) gives a discriminative power of 0.669, whereas pooling to depth 5 gives a discriminative power of 0.675. This difference is small, and not by itself statistically significant, but it is replicated across many different metrics and evaluation depths; and in any case, even maintaining the same discriminative power with less information (decreased pooling) would require explanation. Whether the explanation lies in near-complete assessment even with shallower pools, or in a reinforcement effect, or whether it relates to normalization (since the effect is observed mostly in normalized or standardized metrics) is unclear from this high-level view, and requires further analysis.

In Figure 4, we have compared the discriminative power of different metrics at various pooling and evaluation depths, which is to say the proportion of system pairs found to be significantly different. But just because two metrics and evaluation environments find a similar proportion of system pairs to be significant, it does not mean that they find the same actual system pairs to be significant; they may be equally discriminative, but disagree in their discrimination. One way of measuring agreement on significance is to count the overlap in significant system pairs, but this is a rather insensitive measure, since it simply cuts each set of system pairs into two subsets, significant and not significant. A more sensitive measure is to consider the ordering that the p value of the significance test gives to the system pairs (roughly, the degree of significance it assigns to each system pair). These p values can be used to rank the system pairs, and then the correlation between the rankings produced by different evaluation metrics can be compared. This comparison is undertaken in Table 3, with Kendall's τ as the rank correlation metric. Note first that the ranking by p values is more similar between different metrics within the same evaluation and pooling depth (the half-blocks along the block diagonal) than it is for the one metric across different evaluation and pool depths (the values in bold). These same-depth comparisons show very high τ values, from 0.82 to 0.96 (91% to 98% of system pairs in the same order in both rankings). When the comparison is between different depths, the extended evaluation of pooling to depth 10 but evaluating to depth 100 is closer to the deep evaluation of pooling and evaluation to depth 100 than it is to the shallow evaluation of pooling and evaluating to depth 10. This once again suggests that, at least for this data set, evaluating beyond pool depth gives reliable results; or, looked at another way, that evaluation depth is more important than pooling depth, and both are more important than choice of metric.

5. CONCLUSIONS

This paper began with three, alternative (but not mutually exclusive) hypotheses for explaining the high discriminative power of nDCG: first, that it was due to greater evaluation depth; second, that it was due to normalization; and third, that it was due to a (misleading) reinforcement effect from evaluating beyond pooling depth. The experimental results reported in this paper strongly support the first hypothesis, that is, that evaluation depth is crucial to discriminative power. Provided pooling depth is raised as well (and in many cases even if it isn't), for all metrics examined here, greater evaluation depth leads to greater discriminative power. Evaluation depth is also more important, on the data set explored here, than choice of metric. Equalising the effective evaluation depth closes most (though not all) of the gap in discriminative power between precision and RBP on the one hand, and nDCG and AP on the other. Moreover, a common evaluation depth leads to greater similarity than a common base metric, both for simple system rankings, and for the ranking of system pairs by degree of significance.

Regarding the relationship between normalization and discriminative power, the results are less straightforward. Certainly, normalization without regard to evaluation depth, as is done with recall and AP, leads to the perverse behaviour that discriminative power is boosted by shallower pooling, and therefore less information; and one recommendation of this paper is that the AP formulation should be adjusted to normalize by the smaller of evaluation depth and the number of (known) relevant documents. Beyond this, normalization (and standardization) may lead to some slight improvement in consistency, but the evidence is inconclusive. (Of course, it may be that normalization gives a more reliable measure of true effectiveness, even if it does not confer greater statistical stability.)

Finally, the experiments show only slight support for the hypothesis that nDCG is, through its heavy-tailed nature, being misled by system reinforcement beyond pooling depth. At least for depth 100 evaluation, and at least on this runset, a depth 10 pool gives a reliable indication of the scores that full assessment (that is, depth 100 pooling) would give. There is evidence, though, that highly distinctive systems, such as the manual systems, can be penalized by a lack of reinforcement. Looking at the discrimination results themselves, it is distinctive that, at least for shallow depths, increasing evaluation depth while holding pooling depth fixed raises discriminative power as effectively as (and in some cases even slightly more than) it does when pooling depth is also increased. Additionally, the runset used here is one of the most fully pooled of the TREC data sets; a sparser runset might give different results.

Acknowledgment

This work was supported by the Australian Research Council.

Pool	Eval	Metric	Pool@10 Eval@10		Pool@10 Eval@100			Pool@100 Eval@100		
			nDCG	RBP	aAP	nDCG	RBP	aAP	nDCG	RBP
10	10	aAP	0.89	0.88	***0.73***	*0.72*	*0.67*	***0.74***	*0.74*	*0.73*
		nDCG		0.96	*0.73*	***0.75***	*0.70*	*0.73*	***0.76***	*0.74*
		RBP			*0.72*	*0.74*	***0.69***	*0.72*	*0.75*	***0.73***
10	100	aAP				0.88	0.84	**0.86**	0.86	0.81
		nDCG					0.88	0.79	**0.88**	0.83
		RBP						0.75	0.81	**0.85**
100	100	aAP							0.87	0.82
		nDCG								0.88

Table 3: Kendall's τ between p values assigned to TREC 8 AdHoc system pairs by paired, two-tailed t tests, for evaluation and pooling depths and metrics. Values in bold are those for which a metric is compared against itself (at different pooling or evaluation depth); values in italic are those in which the pair of metrics have different evaluation depths. Results are divided into blocks by pooling and evaluation depth.

References

D. Bodoff and P. Li. Test theory for assessing IR test collections. In C. L. A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 367–374, Amsterdam, the Netherlands, July 2007.

C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In E. Yannakoudis, N. J. Belkin, M.-K. Leong, and P. Ingwersen, editors, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, Athens, Greece, August 2000.

C. Buckley and E. Voorhees. Retrieval system evaluation. In Voorhees and Harman [2005], chapter 3.

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

W. S. Cooper. Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19(1):30–41, January 1968.

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

E. Kanoulas and J. Aslam. Empirical justification of the gain and discount function for nDCG. In D. Cheung, I.-Y. Song, W. Chu, X. Hu, J. Lin, J. Li, and Z. Peng, editors, *Proc. 18th ACM International Conference on Information and Knowledge Management*, pages 611–620, Hong Kong, China, November 2009.

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

T. Sakai. Evaluating evaluation metrics based on the bootstrap. In S. Dumais, E. Efthimiadis, D. Hawking, and K. Järvelin, editors, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 525–532, Seattle, Washington, USA, August 2006.

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

T. Sakai. Comparing metrics across TREC and NTCIR: The robustness to system bias. In J. G. Shanahan, S. Amer-Yahia, Y. Zhang, A. Kolcz, A. Chowdhury, and D. Kelly, editors, *Proc. 17th ACM International Conference on Information and Knowledge Management*, pages 581–590, Napa, USA, October 2008.

M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169, Salvador, Brazil, August 2005.

M.D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In M. J. Silvaa, A. A. F. Laender, R. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, editors, *Proc. 16th ACM International Conference on Information and Knowledge Management*, pages 623–632, Lisboa, Portugal, November 2007.

C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.

E. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In K. Järvelin, M. Beaulieu, R. Baeza-Yates, and S.-H. Myaeng, editors, *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323, Tampere, Finland, August 2002.

E. Voorhees and D. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.

W. Webber, A. Moffat, and J. Zobel. Score standardization for inter-collection comparison of retrieval systems. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, editors, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 51–58, Singapore, Singapore, July 2008a.

W. Webber, A. Moffat, J. Zobel, and T. Sakai. Precision-at-ten considered redundant. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, editors, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 695–696, Singapore, Singapore, July 2008b.

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998.