Author/s:
Lu, X;Moffat, A;Culpepper, JS

Title:
The effect of pooling and evaluation depth on IR metrics

Date:
2016-08-01

Citation:
Lu, X., Moffat, A. & Culpepper, J. S. (2016). The effect of pooling and evaluation depth on IR metrics. Information Retrieval Journal, 19 (4), pp.416-445. https://doi.org/10.1007/s10791-016-9282-6.

Persistent Link:
https://hdl.handle.net/11343/282638

Noname manuscript No.
(will be inserted by the editor)

The Effect of Pooling and Evaluation Depth on IR Metrics

Xiaolu Lu · Alistair Moffat · J. Shane Culpepper

Received: date / Accepted: date

Abstract Batch IR evaluations are usually performed in a framework that consists of a document collection, a set of queries, a set of relevance judgments, and one or more effectiveness metrics. A large number of evaluation metrics have been proposed, with two primary families having emerged: recall-based metrics, and utility-based metrics. In both families, the pragmatics of forming judgments mean that it is usual to evaluate the metric to some chosen depth such as $k = 20$ or $k = 100$, without necessarily fully considering the ramifications associated with that choice. Our aim is this paper is to explore the relative risks arising with fixed-depth evaluation in the two families, and document the complex interplay between metric evaluation depth and judgment pooling depth. Using a range of TREC resources including NewsWire data and the ClueWeb collection, we: (i) examine the implications of finite pooling on the subsequent usefulness of different test collections, including specifying options for truncated evaluation; and (ii) determine the extent to which various metrics correlate with themselves when computed to different evaluation depths using those judgments. We demonstrate that the judgment pools constructed for the ClueWeb collections lack resilience, and are suited primarily to the application of top-heavy utility-based metrics rather than recall-based metrics; and that on the majority of the established test collections, and across a range of evaluation depths, recall-based metrics tend to be more volatile in the system rankings they generate than are utility-based metrics. That is, experimentation using utility-based metrics is more robust to choices such as the evaluation depth employed than is experimentation using recall-based metrics. This distinction should be noted by researchers as they plan and execute system-versus-system retrieval experiments.

Xiaolu Lu, RMIT University, Melbourne, Australia, E-mail: xiaolu.lu@rmit.edu.au · Alistair Moffat, The University of Melbourne, Melbourne, Australia, E-mail: ammoffat@unimelb.edu.au · J. Shane Culpepper, RMIT University, Melbourne, Australia, E-mail: shane.culpepper@rmit.edu.au

1 Introduction

Collection-based evaluation of IR systems is an important aspect of system comparison. A set of documents and topics is required, against which matching relevance judgments are formed, often using a pooling strategy based on a set of contributing runs. Questions of the form "*Is system A better than system B?*" are answered by first applying A and B to the documents and topics to form a *run* for each system-topic pair, without requiring that A and B contributed to the judgment pool; then scoring each run relative to the available judgments using an effectiveness metric; and finally applying a statistical test to the paired scores. Voorhees and Harman [29] summarize work related to cost-effective test collection construction, and in give a range of advice in connection with such TREC-style evaluations.

Many evaluation metrics have been proposed, each with the over-arching goal of providing a better measurement of effectiveness, where "better" is subjective and dependent on a wide range of imprecise factors, not the least of which is the purpose to which the document ranking will be put, and the underlying goal of the user who formulated the information need that led to a query being issued. Those metrics fall loosely into two broad categories: those that are *recall-based*, and those that are *utility-based*. In the first group, metrics are scaled in some way that depends on the best that might be attained for the topic in question, so that the metric's score takes into account the relative difficulty (again, in an imprecise sense) of generating that ranking. In the second group, the behavior perceived by the user of the system is what is measured, and a system will be assigned a low score if it only produces a small number of relevant documents, even if there are only a small number of relevant documents across the whole of the collection, and all of them have been found. The argument in favor of recall-based metrics is that they are normalized, and hence are "fair" in terms of the scale of what is possible; the argument in favor of utility-based metrics is that they reflect what the user sees, and hence are "fair" in terms of recording the experience of the user.

A range of ways of quantifying the relative behavior of different metrics have been explored, including assessments of stability and correlation [26]. However, such evaluations are often undertaken assuming that the metric scores being combined are precise, and represent an exact numeric assessment. In practice, evaluation is usually undertaken over a relatively short prefix of a run – to an evaluation depth of perhaps $k = 20$, or to $k = 100$, or in extreme cases, to $k = 1{,}000$. An important question must then be considered: to what extent are various metrics' scores consistent, or at least correlated, as the evaluation depth k is varied? Indeed, an even simpler question is this: What is it that is being measured when a metric that motivated as measuring the quality of a whole ranking is only applied to a prefix of it?

In this paper we explore the impact of evaluation depth k on truncated evaluation metrics. We show that the evaluation depth is a critical variable in experimental IR evaluations, and plays a role that is inextricably inter-connected with the choice of metric itself. In particular, with a utility-based metric there is a clear sense that as k increases, the metric's value converges to a final value, and that a to-depth-k evaluation is thus an approximation to the final true value of the metric. But as we show with our experiments, that analogy fails for recall-based metrics, and the depth

of evaluation can play a major role in determining the outcome of an "*Is system A better than system B*" comparison. That risk is exacerbated by the difficulty of measuring an "uncertainty" or *residual*, when using recall-based metrics. Specifically, we consider these questions:

RQ1: To what extent does pooling provide an approximation of the number of relevant documents in large-scale web collections?

RQ2: How do different tie breaking strategies affect the evaluation results?

RQ3: When partial judgments based on pooling are used, to what extent do recall-based and utility-based effectiveness metrics result in reliable system comparisons?

RQ4: To what extent are system orderings generated by recall-based and utility-based metrics independent of the evaluation depth?

Zobel [36] demonstrated that for the various NewsWire collections a sufficiently large number of relevant documents are likely to have been identified by the deep pooling processes that were used, that the subsequent experimental comparisons were valid. For such collections, the greater the evaluation depth k, the more likely it is that reliable results will be obtained [30]. But as collections become larger, the challenges associated with identifying relevant documents also increase. The more recent ClueWeb collections are a case in point, with both more documents in the collection, and also a shallower pooling depth d used to construct the sets of documents to be judged. On these collections it is possible that recall-based evaluation metrics yield numeric values substantially different from the full-judgment "final" value of the metric, and that these divergences materially affect the reliability of any system comparisons built on them. A key part of our experimentation is thus to measure the extent to which system orderings at different evaluation depths are consistent, based on two quantities that we introduce, the *coverage ratio* and the *inversion ratio*, and evaluated based on both shallow pooling and deep pooling.

Our results show that extended evaluation (that is, when the evaluation depth k is greater than the pooling depth d) is not a useful strategy for the ClueWeb collections, especially when using recall-based metrics, a different outcome than when similar measurements are applied to NewsWire-type collections [30]. Further, we also show that limiting the evaluation depth to the pooling depth (that is, taking $k = d$) does not necessarily result in more consistent outcomes. Overall, recall-based metrics such as AP and NDCG are more susceptible to such issues than are utility-based ones such as RBP and ERR.

Based on our observations, we argue that, irrespective of the type of searching that is being undertaken, and hence the type of system user that is being modeled by the metric, it is useful for any results based on recall-based metrics to be augmented by parallel measurements derived from utility-based metrics. That is, we caution against the use of recall-based metrics in large-scale test environments as a sole determinant of system relativities, and urge the augmenting use of utility-based metrics that, even if not providing the "right" measurement that is desired, can at least provide a warning when imprecise outcomes might be being generated by that "right" metric.

2 Background

In this section we describe a range of existing utility-based and recall-based effectiveness metrics, and summarize the properties and issues that become evident when these are applied in practical evaluation settings.

The TREC-style Evaluation Framework Evaluations serve an important role in the field of information retrieval. A *batch evaluation* makes use of a document collection; a set of topics, or queries; and a set of relevance judgments [26]. One common way of selecting the set of documents to be judged for each query is through the use of pooling [29], in which the top d documents from each response provided by a set of *contributing systems* are judged for relevance by human assessors. The volume of work involved in this process – and hence its cost – depends on many variables, including the number of contributing systems and the extent to which they represent a diverse group of alternatives, the number of topics attached to the test collection, and, naturally, the value of d. The greater the value of d, the more likely it is that the test collection will be reusable, and hence applicable to the measurement of systems that did not contribute to the pool, but the greater too the cost of creating the resource.

For example, the leave-out-one-run experiments of Zobel [36] showed that on deeply pooled NewsWire collections, with judgments formed using $d = 100$, enough relevant documents were found to produce unbiased system comparisons, while noting also that this might not always be true in future collections. However, Buckley et al. [6] did observe a bias in the TREC AQUAINT 2005 HARD Task, and the TREC 2004 and 2005 Terabyte Track using the GOV2 test collection, which are larger than the NewsWire collections explored by Zobel [36]. The TREC Terabyte Track had a pooling depth of $d = 85$ in 2004, and $d = 100$ in 2005, while the TREC AQUAINT 2005 HARD Task had a pooling depth of $d = 55$. Buckley et al. [6] concluded that caution should be exercised when reusing these collections. On more recent large-scale document collections such as ClueWeb10, an even shallower pooling has been used ($d = 20$) [10], and the situation with regard to reusability is less clear.

It should be noted that no good rule-of-thumb currently exists to define a "good" pooling depth. Both Zobel [36] and Buckley et al. [6] explored the impact of pool depth as a function of relevant documents for each query. The total number of relevant queries each topic has, the diversity of the contributing systems, and the size of the collection can all have an impact when choosing pooling depth.

Effectiveness Metrics A wide range of effective metrics have been proposed, with several taxonomies of them also having been proposed [11, 14, 26]. One way in which metrics can be categorized is as *recall-based* and *utility-based*. In a recall-based metric, the score assigned to any particular system's run for each topic is a relative measure of how close that system came to yielding a perfect response. To obtain a perfect score, a system must rank every relevant document ahead of every non-relevant document, regardless of how many relevant documents there are in the collection, and regardless of how many of them the user may have been seeking when they issued their query. In a utility-based metric the score assigned to each run

Gain Mode	Function	Metric
Binary	$G(i) = 0$ if $g_i < \theta$, and 1 otherwise	Any: Prec, ERR, RBP, AP, NDCG, QM, etc; and required as one component in QM.
Linear	$G(i) = g_i/(\max_j g_j)$	Any, but not usually used with Prec or AP.
Exponential	$G(i) = 2^{g_i-1}/(\max_j 2^{g_j})$	Any, but common with ERR and NDCG, and not usually used with Prec or AP.

Table 1: Commonly used gain functions, and the metrics that they are typically matched with. The constant θ is a threshold for the purpose of binarizing categorical and ordinal relevance labels. The metrics themselves are defined in Table 2.

is an assessment of the utility delivered to the user (or, more commonly, the per-document rate at which utility was delivered to the user) according to some model of user behavior. Models can be defined for a wide range of different tasks and user behaviors, including (as an extreme case) the "feeling lucky" task of searching for a single relevant document, and examining only the top document in the ensuing ranking. Moffat et al. [17] and Bailey et al. [3] consider the relationship between user models and the weighting functions employed in utility-based metrics.

Where relevance assessment are categorical or ordinal labels, a *gain function* is used to generate a numeric score as part of the relevance calculation. If binary assessment judgments have been undertaken, the only labels are "not relevant" and "relevant", and the gain function maps them to zero and one respectively. If there are multiple relevance levels, for example, "0: not at all relevant", "1: somewhat relevant", "2: relevant" and "3: highly relevant", then the mapping to be used is less clear-cut; Table 1 lists three of the options that have been used to convert ordinal relevance categories to real-valued gain contributions.

Table 2 provides definitions of several standard metrics, expressed in terms of the gain values $G(i)$ associated with the ranking; the total number of relevant documents R; and the gain values $G^*(i)$ associated with the i th element in an ideal ranking in which all documents are listed in order of decreasing gain [4, 12, 15, 21]. The last three – AP, NDCG, and QM – all make use of either R, or of the ideal ordering $G^*(i)$ in which R is implicit; these are the *recall-based* metrics in Table 2. On the other hand, the first three derive their value solely from the unnormalized $G(i)$ gain values, and are *utility-based*. Note that in their original definition of NDCG, Järvelin and Kekäläinen [12] make use of a parameter b that defines the number of equal-discount items at the head of the ranking; here we use the parameter-free version in which the discounting function is taken to be $D(i) = 1/\log_2(1+i)$ and is monotonic, and for which the value of NDCG is independent of the base that is used in the logarithm. Other discounting functions can also be employed, including a Zipfian discount, $D(i) = 1/i$, which places a greater fraction of the weighting on top-ranked elements [13].

Truncated Rankings Table 2 also lists the approaches that can be used to compute "truncated at depth k" versions of these metrics. For example, average precision, AP,

M	Ideal full-depth definition for M	Truncated at depth-k options for M@k
Prec	$(1/k) \cdot \sum_{i=1}^{k} G(i)$	Unchanged.
ERR	$\sum_{i=1}^{N} \left((1/i) \cdot G(i) \cdot \prod_{j=1}^{i-1} (1 - G(j)) \right)$	Sum to k instead of N.
RBP	$(1-p) \cdot \left(\sum_{i=1}^{N} G(i) \cdot p^{i-1} \right)$	Sum to k instead of N; can compute tail residual to upper-bound the contribution of the omitted documents.
AP	$(1/R) \cdot \sum_{i=1}^{N} G(i) \cdot \text{Prec@}i$	Sum to k instead of N, plus instead of R use one of: (a) R_d; or (b) $\min\{R_d, k\}$; or (c) if $k < d$, can choose to use R_k.
NDCG	$\frac{\sum_{i=1}^{N} G(i) \cdot D(i)}{\sum_{i=1}^{N} G^*(i) \cdot D(i)}$	Sum to k instead of N, plus either: (a) use pooled-to-d judgments when forming $G^*(i)$; or, (b) if $k < d$, can choose to use pooled-to-k judgments when forming $G^*(i)$.
QM	$(1/R) \cdot \sum_{i=1}^{N} b_i \cdot \frac{\sum_{j=1}^{i} b_j + \beta \cdot \sum_{j=1}^{i} G(j)}{i + \beta \cdot \sum_{j=1}^{i} G^*(j)}$	Sum to k instead of N, plus instead of R use one of: (a) R_d, or (b) $\min\{R_d, k\}$, or (c) if $k < d$, use R_k; plus (i) use pooled-to-d judgments when forming $G^*(i)$, or (ii) if $k < d$, can choose to use pooled-to-k judgments when forming $G^*(i)$.

Table 2: Standard effectiveness metrics, in full-depth "ideal" form, and "truncated at depth k" form, where N is the number of documents in the collection; d is the pooling depth used to form relevance judgments; R is the true total number of relevant documents for the query; R_d the approximation to R that results if pooling to depth d is undertaken; and $G^*(i)$ is the gain value that would result at depth i based on a perfect ranking of all of the relevant documents. Precision makes use of a parameter k; RBP a parameter p; NDCG a discount function $D()$; and QM a parameter β. Options for the gain function $G()$ are listed in Table 1; the value b_j used in QM is a binarized value $G(j)$ for a constant threshold θ.

is defined in the ideal sense as a sum over all of the N documents in the collection (or, more precisely, a sum through until all of the R relevant documents have been included). But if a ranking of length $k < N$ is provided, and if the relevance judgments are based on pooling to depth $d < N$, that definition must be altered to accommodate the limited information that is known. One clear choice, denoted as option (a) in Table 2, is to compute

$$\text{AP}_\text{a}@k = \frac{1}{R_d} \sum_{i=1}^{k} G(i) \cdot \text{Prec}(i)\,, \tag{1}$$

where R_d is the estimated value of R derived from the pooling process to depth d. In this expression it is possible for R_d to be larger than k, and hence it may be that a score of 1.0 cannot be attained from Equation 1, even by a "perfect" ranking of length k in which every item is relevant. The second formulation for truncated AP addresses

that issue, and instead computes

$$\text{AP}_\text{b}@k = \frac{1}{\min\{R_d, k\}} \sum_{i=1}^{k} G(i) \cdot \text{Prec}(i)\,, \tag{2}$$

with a score of 1.0 always now obtainable, just as it is in the full-depth definition of AP presented in Table 2. Using Equation 2, a change from depth-k evaluation to depth-k' evaluation, with $k' > k$, might give rise to a score decrease; whereas $\text{AP}_\text{a}@k$ in Equation 1 is non-decreasing in k. That is, with $\text{AP}_\text{b}@k$, extending a ranking from a length of k to a length of $k' > k$ might result in it being numerically further from an ideal ranking of length k'; and hence, in the taxonomy given by Moffat [14], $\text{AP}_\text{a}@k$ is monotonic, and $\text{AP}_\text{b}@k$ is non-monotonic.

There is no single community consensus as to which of these two "at-k" AP variants is preferred. For example, `trec_eval` instantiates $\text{AP}_\text{a}@k$, and Sakai [24] presents $\text{AP}_\text{b}@k$. A third variant is also listed in Table 2 – in cases where $k < d$, the $\text{AP}_\text{c}@k$ version offers the option of deliberately choosing to make use only of the relevance information that could be gleaned by pooling to depth k, and hence uses $R_k \le R_d$ as the estimate for R. Like $\text{AP}_\text{b}@k$, $\text{AP}_\text{c}@k$ is non-monotonic, since an increase in k might increase R_k. In all versions of AP it is usual to make use of a binarized gain function $G()$, although it should be noted that graded AP variants have also been proposed [20].

The NDCG (normalized discounted cumulative gain) metric developed by Järvelin and Kekäläinen [12] makes use of a slightly different recall-based normalization than does AP, but like AP, is also nominally defined in the ideal sense as being a sum over all of the N documents in the collection. Adjusted to deal with truncated rankings of length k, the formulation is usually adopted is this one:

$$\text{NDCG}_\text{a} = \frac{\sum_{i=1}^{k} G(i) \cdot D(i)}{\sum_{i=1}^{k} G^*_d(i) \cdot D(i)}\,, \tag{3}$$

where $G^*_d()$ is now a "somewhat ideal" ranking derived from information accumulated from pooling to depth d. Equation 3 has more in common with Equation 2 than it does with Equation 1, since an "all maximally relevant" ranking of length k will attain a score of 1.0 regardless of R (or equivalently, regardless of $G^*_d(k')$, for $k' > k$), and if a ranking of length k is extended to $k' > k$ elements, the assigned NDCG_a score may decrease. That is, the "at-k" version of NDCG that is described by Equation 3 is, like $\text{AP}_\text{b}@k$, non-monotonic, as is the NDCG_b option that arises if only judgments to depth k are used and an "even less ideal" ranking $G^*_k()$ is formed.

The third of the recall-based measures listed in Table 2 is Sakai's Q-Measure (QM), which can be thought of as a blend of AP and NDCG, and as a consequence requires both that an estimate for R be available, and also that the cumulative gains over an ideal ranking be known [21]. Table 2 lists combinations in which the Q-Measure can be adapted to work with finite pool depths d and finite ranking lengths k. It is also non-monotonic, and can decrease as k is increased.

The general family of weighted-precision effectiveness metrics, of which rank-biased precision (RBP) [15] and expected reciprocal rank (ERR) [9] are examples,

do not attempt to provide scores that are relative to an ideal ranking, and as a consequence, are free of the requirement that R be known or estimated in some way. The scores generated by weighted-precision metrics can be interpreted as being the rate at which a user accrues gain, assuming a stochastic user model determined by the structure of the weighting function. For example, in RBP the user is assumed to transition from depth i to depth $i+1$ in the ranking with a fixed probability p; and in ERR, the user is assumed to proceed through the ranking in the same way, but instead of continuing with a fixed probability p, continuing with probability given by $1 - G(i)$. Other weighted-precision metrics have been proposed by Moffat et al. [17] and Bailey et al. [3], incorporating increasingly subtle estimates of the user's likelihood of continuing from depth i to depth $i+1$ in the ranking.

Weighted-precision metrics are monotonic, and the truncated "at-k" variants of the form described in Table 2 converge consistently to their final values. For example, if RBP@N is the true value of RBP over all N documents, then

$$\text{RBP@}N - \text{RBP@}k \geq \text{RBP@}N - \text{RBP@}(k+1) \tag{4}$$

for all values of $k < N$. Moreover, with many weighted-precision metrics it is possible to calculate an upper bound on the differences that arise in Equation 4, meaning that an upper limit on the extent of the uncertainty in the at-k metric value compared to the at-N value can be precisely calculated. For example, the imprecision in ERR@k relative to ERR@N is always bounded above by $1/(k+1)$, and for a particular ranking with gain values $G(i)$, is at most $1/(k+1) \cdot \prod_{i=1}^{k}(1 - G(i))$. Moffat and Zobel [15] derive similar limits for RBP. The ability to compute these bounds is useful if the evaluation depth k is to be chosen so as to achieve a certain level of fidelity in reported score values.

Extended Evaluation Our presumption throughout this work is that k and d are independent concepts that may or may not take on the same numeric value in any particular evaluation. To be clear on the distinction, k is the length of the system runs being evaluated (as per the formulations summarized in Table 2); whereas the pooling depth d, in conjunction with the runs generated by the contributing systems, determines the size of the pool of judged documents, and thus the number R of judged-relevant documents. In particular, note that R has a non-decreasing relationship with d.

Three evaluation scenarios are possible. First, if all of the runs being evaluated contributed to the pool, and if $k \leq d$, then relevance information is available for (at least) the first k documents in every run being scored. The at-k computations for utility-based metrics are thus completely specified, and there can be no imprecision in their calculated scores. That is, spending more judgment effort to further increase d is irrelevant, since none of the scores will change. But even in this first scenario, the scores of recall-based metrics must be regarded as being estimates, since R_d (or R_k) is being used in place of R, and spending more effort on judgments by increasing d might also increase R_d and hence *decrease* the scores generated for all of $\text{AP}_\text{a}@k$, $\text{AP}_\text{b}@k$, $\text{NDCG}_\text{a}@k$, and the various truncated QM options, even for a fixed value of k. Only $\text{AP}_\text{c}@k$ and $\text{NDCG}_\text{b}@k$ are immune to this possible source of numeric score variation.

The second evaluation scenario arises when $k > d$, but still with all runs having contributed to the pool. In this *extended evaluation* situation [30], some gain values $G(i)$ for $d < i \le k$ may be known in some of the runs, because the same document appeared within the first d of a different run and hence got judged. In other cases, $G(i)$ may be unknown for $d < i \le k$, because the documents in question were never top-d in any of the runs. But when $1 \le i \le k$, it is always the case that $G(i)$ is known.

In the third of the three cases, runs are present that did not contribute to the pool, and hence there may be ranks i at which $G(i)$ is unknown for all $1 \le i$, regardless of k and d. This is the situation that arises when a test environment – documents, topics, and judgments – is used in post-hoc evaluations of new retrieval systems. A critical step in establishing the aptness of the experimental methodology should then be to establish the extent of the mismatch between test environment and systems being tested. In particular, if a new non-contributing system is generating runs in which only a minority of the top-k documents are judged, then the mismatch between system and test-bed must be regarded as being meaningful, and any scores that are computed must be handled very carefully.

Dealing with the Unknown An important question then arises – What is the best way to handle runs which contain unjudged documents in the top k? Four techniques have emerged to deal with this issue: assumed non-relevance; predicted relevance; computed score ranges; and condensed runs. The first of these four is the easiest to implement. If the document at rank i is unjudged, then $G(i)$ is taken to be zero, on the basis that the pooling process, if implemented across a broad enough set of systems and to a sufficiently generous depth d, should have identified all – or a suitably great fraction of – the documents relevant to the query. This argument has been defended for the NewsWire collections by a range of leave-one-out experiments [36], but is also at odds with an underlying expectation of new systems, namely that one of the ways in which start-of-the-art can be improved is by retrieving previously unseen but relevant documents.

The second option is to infer an estimated gain value for unjudged documents based on a prior distribution [2, 32, 35]. For example, Yilmaz and Aslam [32] assume that relevance has a smooth distribution with regard to rank within a run, and suggest a sampling-based approach to approximating AP, denoted as infAP. Such approaches allow the judgment budget to be reduced at the risk of accuracy loss in the system ordering that is inferred. However, they do not necessarily allow test environments to be reused, since the additional runs added post-hoc might be very sparse in terms of judged documents.

The third option is to acknowledge the existence of the problems raised by unjudged documents, and accumulate an error bound, or *residual*, that quantifies the degree of uncertainty in the reported score [15]. Tail residuals for ERR and RBP were mentioned earlier, and similar computations can be applied in regard to unjudged documents appearing within the top-k for most utility-based metrics. In the simplest form, score ranges are computed by first assuming the unjudged documents have a gain of zero, and processing the whole run to get a pessimal score for the run. Then unjudged documents are scored as if they have a gain of 1.0 to get a maximally optimistic score. Finally, these two overall scores are reported as lower and

upper bounds on the true value. In cases where there are few judged documents in a run, this procedure gives rise to a wide score range, and raises a clear signal that the evaluation desired was incompatible with the underlying experimental framework.

Score ranges are significantly more complex to compute for recall-based metrics than for utility-based metrics, because of the normalization by R that is involved, which means that the maximum possible score for a run is unlikely to correspond to the situation in which all unjudged documents are deemed to be fully relevant. Moreover, the ranges that emerge are typically very broad if anything other than a very small minority of the documents in each run are unjudged, regardless of how deep in the ranking those unjudged documents appear.

The fourth and final approach for handling unjudged documents is to score *condensed runs*; that is, derivative runs from which all documents without judgments have been omitted. Any metric can be applied once the condensed runs have been formed; one that was explicitly developed for use with condensed runs is the BPref (binary preference metric) mechanism proposed by Buckley and Voorhees [5]. If a query has R_d known relevant documents, then BPref locates the first R_d judged-nonrelevant documents too, and then pairwise compares their ranks against the ranks of the R_d judged-relevant documents. Sakai [23] compares BPref with condensed recall-based metrics AP′, NDCG′, and QM′, and concludes that applying the condensing process to existing recall-based metrics gives more discriminative power (see Section 3) than BPref. Among all of the condensed metrics considered, Sakai concluded that NDCG′ and QM′ were the most suitable choice based on the collections evaluated.

3 Comparing Metrics

This section examines the question of how one metric can be compared against another, including the more subtle issue of how a metric evaluated to one depth k_1 can be compared to itself when evaluated to a different depth k_2. The underlying numerical properties [14] of the two metrics are less likely to be factors in this second type of comparison, and observed behavior on standard test collections becomes a more important determinant.

Discrimination Ratio An early empirical study by Buckley and Voorhees [4] used error rate to conclude that $\text{AP}_\text{a}@1000$ is a better metric than to Prec, where the error rate of a metric is the likely error of concluding "System A is better than system B", given a number of requests and differences. An important point made in this work is that the choice of evaluation metrics should be considered in conjunction with features of the test collection, including the number of topics. Buckley and Voorhees also examine the *the power of a measure to discriminate among systems* and compared metrics using a *fuzziness value*. The idea of formalizing the notion of *discriminative power* of metrics, defined as the fraction of the system pairs in a shared task environment for which a statistical test (often a t-test) finds that one of the two systems is superior to the other at a particular significance threshold (often $p = 0.05$), was later introduced by Sakai [22]. Sakai [22] reports further experiments using this approach.

Metrics that are only mildly top-weighted, such as AP, NDCG and RBP(0.95) tend to have higher discrimination ratios than shallow metrics such as Prec@5 and RBP(0.5) that are strongly top-weighted [25], but also have higher costs associated with the pooling activity that is needed to generate the deeper relevance judgments that are required.

If working with a particular p value as a threshold is undesirable, the *median system-pair p value* can be computed and reported. The lower that value is, the higher the discrimination ratio is likely to be. But note that the discrimination ratio and median p value are joint attributes of a metric, a test context, and a set of systems, and not a property of the metric alone. In particular, if the set of systems being compared is changed, so too might the discriminative ability of the metrics being measured. There is no sense of there being a "universe" of possible systems, or of discriminative power being an underlying quantity that can be estimated based on a sample of systems drawn from such a universe.

Metrics can also be compared based on their statistical ability to predict outcomes on held-out data. For example, Yilmaz and Robertson [33] extend the work of Aslam et al. [1] and argue that AP is more "informative" than NDCG when applied to binary judgments.

Correlation Coefficients A second criteria that might be used to quantify the relative behavior of metrics is their ability to order a set of systems relative to the ordering established by one or more reference metrics M_{ref}. There are two types of measurement that might be applied: (i) rank correlation calculations, which generate a numeric score from two system orderings; or (ii) score correlation coefficients, that take both rank and numeric scores in to account. Rank correlation coefficients such as Kendall's τ or Spearman's ρ are often calculated in IR evaluations [26]. A weakness of both is that discords that occur at the top of the ranking incur the same penalty as discords at the bottom of the ranking, whereas in IR, there is often more emphasis on having the correct system ordering at the top of the ranking than there is on having the correct ordering at the bottom.

In response to these considerations, Yilmaz et al. [34] describe an AP-based correlation coefficient they call τ_{ap}. To achieve a top-weighted emphasis, discords are discounted the deeper they arise in the ranking, in much the same way that in AP gain is discounted the deeper it occurs in the ranking. In a related thread of development, Webber et al. [31] describe a computation that they call Rank Biased Overlap (RBO), which has the additional benefit of operating over indefinite and non-conjoint rankings. As with the metric RBP, RBO is based on an underlying user model that incorporates a persistence parameter p that reflects the propensity of the user to move from depth i to depth $i + 1$ while comparing the pair of rankings.

If evaluation scores are regarded as being important, not just system orderings, and if it is assumed that two metrics that behave identically will give rise to the same numeric scores, then Pearson's coefficient and root-mean-square distance (RMSE) computations can be used. These measure are applicable if, for example, the set of scores generated by an at-k_1 evaluation are being compared with the scores generated by an at-k_2 evaluation of the same metric, with $k_1 < k_2$.

Judgment Cost A third criteria that can be applied to metrics is cost in terms of judgment effort. Shallow metrics, and steeply top-weighted metrics, are cheaper to evaluate to a given level of score fidelity than deep metrics such as Prec@100 or AP@100. Yilmaz et al. [35] proposed a sample-based approach that uses fewer judgments to estimate AP or NDCG, and later extended this idea to conduct low cost evaluations [8]. Similar ideas related to incomplete judgments have been explored by Büttcher et al. [7] and Yilmaz and Aslam [32]. The *residuals* that can be computed for weighted-precision metrics such as RBP can also be used to establish the depth of pooling required to obtain scores with some level of error [15], or to guide the choice of documents to be judged relative to a fixed judgment budget [16].

Coverage One issue with the use of discrimination ratio is that the two metrics can have the same discrimination ratio over a set of systems, but regard quite different pairs of systems as being distinct, or even regard the same pair of systems as being distinct but in two opposing directions. For example, Ravana and Moffat [19] measure the number of ordering reversals that take place as the pooling depth is increased. If the metrics are different, then such outcomes are not problematic, and can be expected to occur when metrics place their emphasis at different points in the ranking. But if the metrics are intended to be surrogates for each other, or are presumed to be strongly related, then such outcomes indicate a lack of reliability.

To capture the distinction, suppose that M_{ref} is a baseline, or reference metric; that a second metric M_{new} is being evaluated; and that a set $S = \{s_1, \ldots s_n\}$ of systems is being compared. Suppose further that M_{ref} discriminates between a set of system pairs given by

$$D_{ref} = \{(i,j) \mid \mathrm{M}_{ref}(s_i) >_p \mathrm{M}_{ref}(s_j)\}\,,$$

using the chosen statistical test at significance level p. Now consider the second metric, M_{new}, and the corresponding set of significantly different system pairs

$$D_{new} = \{(i,j) \mid \mathrm{M}_{new}(s_i) >_p \mathrm{M}_{new}(s_j)\}$$

at a same significance level p. We define the *coverage ratio* as

$$covrat_{\mathrm{M}_{ref}}(\mathrm{M}_{new}) = \frac{|D_{new} \cap D_{ref}|}{|D_{ref}|}\,. \tag{5}$$

That is, the coverage ratio is the extent to which the new metric is able to confirm the relationships that were noted by the reference metric. For example, suppose that there are 10 systems in the pool S, and hence 45 distinct system pairs; that 35 of those system pairs are identified as being different at $p = 0.05$ by M_{ref}; and that there are 28 out of the 35 that are also separated at $p = 0.05$ by M_{new}. Then the discrimination ratio of M_{ref} is $35/45 = 0.78$, and the coverage ratio of M_{new} relative to M_{ref} is $28/35 = 0.80$. The coverage ratio should be viewed as a modified discrimination ratio that is conditioned on the set of pairs that are found to be separable by the reference metric. In particular, if a metric that requires shallow judgments is intended as a surrogate for an established reference metric that requires a larger volume of judgments, then we would ideally like the coverage ratio between the two metrics to be high.

Inversions Metric M_{ref} might also deliver the opposite outcome to M_{ref}. We will say that an *observed inversion* has occurred if M_{ref} indicates that system A is significantly better than system B at some significance threshold p, but M_{new} indicates (when averaged in some manner) that system B is better than system A, with or without significance. Using similar notation as was employed to define the coverage ratio, we accumulate the pairs for which metric M_{new} assigns a higher average score to system s_j than it does to system s_i:

$$I_{new} = \{(i, j) \mid \text{M}_{new}(s_j) > \text{M}_{new}(s_i)\}$$

and then compute the observed inversion ratio as:

$$invrat_{\text{M}_{ref}}(\text{M}_{new}) = \frac{|I_{new} \cap D_{ref}|}{|D_{ref}|} \,. \tag{6}$$

Continuing the same example with $|S| = 10$ systems and $|D_{ref}| = 35$, if 6 of those 35 system pairs result in $\text{M}_{new}(s_j) > \text{M}_{new}(s_i)$, then the observed inversion ratio would be $6/35 = 0.17$.

Interpretation No particular conclusion can or should be drawn from the coverage ratio or the observed inversion ratio in cases in which the two metrics being compared are known to have different properties. The very nature of statistical testing also means that some level of inversions and some loss of coverage can be expected, even when the metrics are closely related. But when a limited-depth evaluation is being used to generate an at-k approximation to a full-depth metric, such as via the $\text{AP}_\text{a}@k$ computation shown in Equation 1 (with d regarded as being a constant), or via the $\text{RBP}(p)@k$ computation described in Table 2, low coverage ratios and/or high observed inversion ratios should be regarded as a warning that the evaluation process may not be providing the desired measurement of system performance.

Volatility Matrix A volatility matrix is a means of visualizing the strength of the relationship between two at-k versions of the some metric M. Suppose that an $\text{M}@k$ evaluation of a set of systems S results in (again, using an appropriate score aggregation regime over a set of topics, such as taking the mean) an ordering of the systems from highest overall score to lowest overall score given by the permutation $\sigma_k(S)$. Then a matrix of correlation coefficients can be constructed, where

$$T[k_1, k_2] = corr(\sigma_{k_1}(S), \sigma_{k_2}(S)) \,, \tag{7}$$

and where $corr()$ is one of the rank correlation coefficients described earlier in this section. In the experiments presented in the next section we take $corr()$ to be Kendall's τ, and plot the matrix as a heatmap; τ_{ap} and RBO can also be used, and other visualizations of the matrix. Assuming that $corr()$ is a symmetric function, then T_M is a symmetric matrix, with all diagonal elements equal to 1. Taking a Kendall's τ score of 0.9 as a threshold for equivalence of system orderings [27], we would ideally hope for $T[k_1, k_2] > 0.9$ for all $k_1, k_2 > \theta$ for some threshold depth θ that is dependent on the metric M and the set of systems S. As is demonstrated in Section 4, this requirement is met in some circumstances for some metrics, while other metrics have less

Collection	Pooling depth	Contributing runs	Topics
Robust04	100	42	651 – 700
TREC 9 Web	100	25	451 – 500
ClueWeb2009	12	32	1 – 50
ClueWeb2010	20	21	51 – 100

Table 3: Collections used in the experiments. Where tracks employed a sampling strategy to construct the pool, it is the minimum pooling depth that is listed.

consistent behavior. When the latter occurs, it is difficult to both interpret and predict the behavior of the metric as the evaluation depth changes, and it should to be taken as a clear warning that care needs to be exercised when using that metric to generate system orderings.

4 Experiments

We now describe a range of analysis that we have undertaken with TREC datasets, judgment pools, and submitted runs, as we have investigated these inter-relationships between metrics, and their various truncated evaluation methodologies.

Collections, Topics, and Systems Table 3 lists the resources used in our experiments, spanning the qrel files and submitted runs from the 2004 Robust Track; the TREC 9 Web Track; and the 2009 and 2010 ClueWeb-based Ad-Hoc Tracks. In each case only the subset of the submitted runs that was used to construct the judgment pools is employed, as shown in the third column of the table. The 2004 Robust Track and TREC 9 Web Track are "deep pooled" datasets, to depth $d = 100$; whereas the two ClueWeb datasets are "shallow pooled", to a minimum of depth 12 and 20 respectively, with some topics pooled more deeply. As has also been done by other researchers [36], we simulate shallower pooling depths $d' < d$ by restricting the qrels file to documents whose minimum depth in any contributing run is less than or equal to d'.

Relevant Documents We consider the research question RQ1 in this section. Figure 1 shows the rate at which relevant documents are identified by pooling. In the left graph, the average number of relevant documents per query is plotted as a function of pooling depth, in each case up to the minimum comprehensive depth of the original TREC pooling. The right graph shows the average relevant occurrence rate between the plotted points in the left graph, as a fraction of the new documents judged. For example, for both the 2009 and 2010 topic sets, more than a quarter of the ClueWeb documents that get inspected as the pool depth increases from $d = 4$ to $d = 10$ are found to be relevant. This behavior is in marked contrast to the TREC 8 and TREC 9 datasets, and for ClueWeb09 and ClueWeb10 (and, not shown, for 2011 and 2012 also) there can be no suggestion that the majority of the relevant documents have been identified. Rather, it seems likely that a further one-fifth of the unjudged documents that occur in the top 50 of the contributing runs could be relevant, doubling the

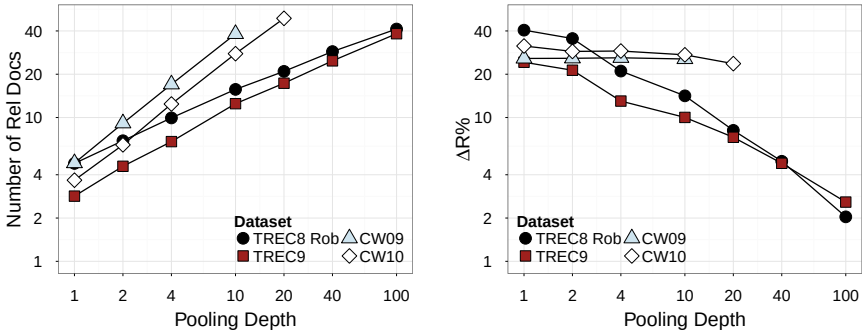


Fig. 1: Average number of relevant documents identified per topic (left), and the rate at which they are identified (right), expressed in both cases as a function of the pool depth, with the averages taken over the contributing runs listed in Table 3.

Dataset	Mode	RelInPool	RankUnj	NumRelPast
CW09	$d = 10$	2.4±3.0	12.9±2.6	13.6±12.7
CW09	$d = 12$	3.0±3.5	15.1±2.9	15.5±14.5
CW09	all	n/a	16.8±5.2	n/a
CW10	$d = 10$	3.3±3.2	12.3±2.1	17.2±13.2
CW10	$d = 20$	6.3±6.0	22.7±3.0	29.1±22.0
CW10	all	n/a	24.4±5.0	n/a

Table 4: Details of pooling outcomes, averaged across both systems and topics, with standard deviations. The column headed "RelInPool" is the average number of relevant documents per run identified within the pool construction depth indicated by the column headed "Mode"; "RankUnj" is the average rank of the first unjudged document; and "NumRelPast" is the average number of relevant documents that appear beyond the listed pool construction depth. The average number of relevant documents per topic at the given value of d is the sum of "RelInPool" and "NumRelPast".

number of known answers. Table 4 provides further details of the effect of different pool depths for the two ClueWeb collections.

The first block in Table 5 shows the results of "leave one out" experiments [36], and the second block shows the corresponding "leave one group" out results [28]. The TREC 8 collection gives robust outcomes, even with a shallow pooling depth of $d = 10$. But the two ClueWeb collections, with $d = 12$ and $d = 20$, are vulnerable to both score and system rank shifts, even when shallow metrics are used. The second block that shows results of "leave one group", which further confirms the NewsWire collections are more reusable than ClueWeb collections. Similar observations have been made in connection with the GOV2 test collection [6].

Breaking Ties In this section, we answer research question RQ2. First, we assume that the ranked results are given in a deterministic order, which may not be the case in

	TREC 8		ClueWeb 09		ClueWeb 10	
	$\Delta\%$	ΔRank	$\Delta\%$	ΔRank	$\Delta\%$	ΔRank
AP_b	–1.31	0.67	–23.30	4.13	–17.50	2.43
$NDCG_a$	–1.06	0.60	–18.54	4.03	–13.98	2.10
RBP0.8@k	–0.90	0.38	–17.12	3.66	–12.75	1.61
RBP0.95@k	–1.25	0.76	–20.51	4.56	–15.33	2.19

(a) Score and rank changes resulting from leaving out one run.

	TREC 8		ClueWeb 09		ClueWeb 10	
	$\Delta\%$	ΔRank	$\Delta\%$	ΔRank	$\Delta\%$	ΔRank
AP_b	–4.41	1.81	–48.78	9.57	–42.71	5.43
$NDCG_a$	–3.30	1.80	–41.79	8.69	–35.02	5.19
RBP0.8@k	–3.05	1.31	–40.51	7.75	–30.94	4.39
RBP0.95@k	–4.10	1.85	–44.07	9.07	–37.38	5.89

(b) Score and rank changes resulting from leaving out a related group of runs.

Table 5: Leave-one-run-out and leave-one-group-out experiments, with $k = d = 10$ for TREC 8, $k = d = 12$ for ClueWeb09 and $k = d = 20$ for ClueWeb10, where $\Delta\%$ is the percentage score change, and ΔRank is the absolute rank change among the systems. Values listed are averaged over the left out run or runs.

Task	Med. % of ties	τ	Task	Med. % of ties	τ
TREC2	12	0.742	Robust04	9	0.997
TREC3	6	0.856	Robust05	64	0.990
TREC4	14	0.838	TREC2004	44	0.999
TREC5	8	0.981	TREC2005	30	0.990
TREC6	8	0.997	TREC2006	42	0.987
TREC7	12	0.996	CW09A-2009	22	0.898
TREC8	10	0.990	CW09A-2010	35	0.959
TREC9	34	0.994	CW09A-2011	14	0.984
TREC10	42	0.995	CW09A-2012	14	1.000

Table 6: The median percentage of tied scores, and the Kendall's τ correlation between system orderings generated by TREC-style tie-breaking and system orderings generated by document-occurrence based tie-breaking, evaluated using RBP(0.8). Only the top-20 documents are considered when computing the percentage of tied scores.

practice. Approaches to tie breaking are another subtle but important point to consider when implementing evaluation metrics. Systems may submit a ranked list containing one or more tied scores, even at the top ranking positions. Table 6 shows the median percentage of ties within the top-20 documents per topic, across all submitted systems for a wide range of TREC Tracks. The median per-run percentage of ties among the top 20 documents differs widely across tasks, and the potential pitfalls of ignoring the effect are quite real.

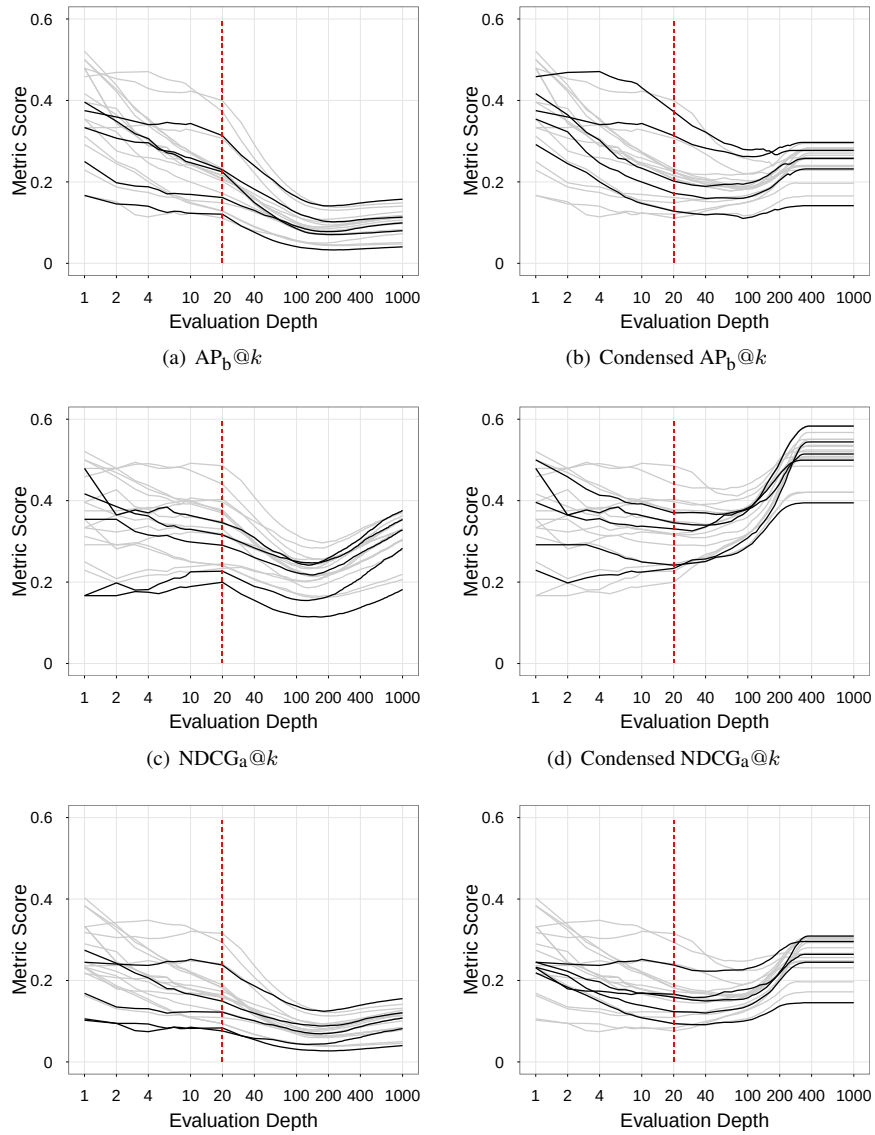
The `trec_eval` toolkit[1] breaks score ties based on reverse alphabetical ordering of document labels. This approach, while resulting in a deterministic ranking, risks promoting unjudged documents higher in the overall ranking than may be warranted by their original position in the submitted run. In particular, if a document demoted by `trec_eval` as a result of a tied score had in fact been included in the pool and judged relevant, and a corresponding promoted document had been unjudged or judged non-relevant, then the system performance will be underestimated. To minimize this risk, through most (but not all) of the TREC experimentation, the pooling process was also implemented in a way that broke ties that straddled the pooling depth so as to ensure that the documents that would be included by `trec_eval` in an at-k evaluation would also be the ones judged as part of an at-k pool. An obvious alternative is to adhere to the ordering of the lines in the run file, and not reorder score-tied documents in any way, which is the approach used in the NTCIR evaluation tools that have been used since NTCIR 7. This has the benefit of ensuring that any subsequent user-instituted pooling (as we have done in this paper, for example) implemented using the equivalent of "`head -k`" corresponds exactly to the subsequent evaluation process. Another alternative is to handle ties on document scores without regard to an external tie-breaking regime. One option that can be used with utility-based metrics is to compute score contributions across block of tied documents, summing the total weight available at that set of depths, and then sharing that total uniformly across the documents in that block. Each apportioned weight is then used to scale the corresponding per-document gain. A further variant for graded relevance applications is described by Moffat et al. [18], who suggest computing the average gain across the documents in the block of equal-score items, and then applying that gain at each rank position spanned by the block. Bevan Koopman's `inst_eval`[2] implements this suggestion.

Carrying out tie-breaking without uniformly defined rules can produce inconsistent system rankings, even when using a single evaluation metric for all comparisons. The Kendall's τ values shown in Table 6 show that such a differences can arise when performing an evaluation using RBP(0.8). Two different tie-breaking methods are used: (i) TREC-style, which breaks ties based on the reverse ordering of document IDs; and (ii) adherence to the ordering of the lines in the run file, that is, respecting the ranked lists submitted by the system authors. Although most tasks give high correlation, the potential risks cannot be ignored, given the variance in Kendall's τ values through the years. For concreteness, and to maintain consistency with the pooling processes that led to the various sets of judgments we employ, in this paper we perform all computations (except for Table 6) with tie-breaking based on the original document occurrence order.

Raw System Scores We now consider our third research question (RQ3). To illustrate score variability, Figure 2 shows computed system scores averaged over the ClueWeb10 topics, in all cases using pooling to $d = 20$, and with the evaluation depth k varying, including extended evaluation when $d < k$. Each line on each graph

[1] `http://trec.nist.gov/trec_eval/`

[2] `https://github.com/bevankoopman/inst_eval`

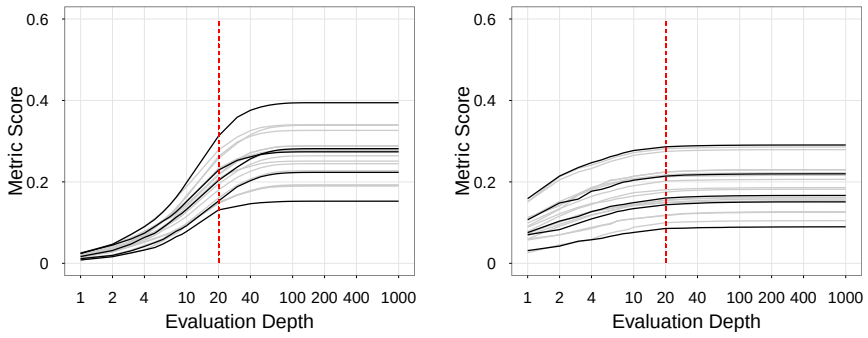


(a) $\text{AP}_\text{b}@k$ (b) Condensed $\text{AP}_\text{b}@k$

(c) $\text{NDCG}_\text{a}@k$ (d) Condensed $\text{NDCG}_\text{a}@k$

(e) $\text{QM}_{\text{a,i}}(1)@k$ (f) Condensed $\text{QM}_{\text{a,i}}(1)@k$

Fig. 2: System scores for $\text{AP}_\text{b}@k$, $\text{NDCG}_\text{a}@k$, and $\text{QM}_{\text{a,i}}(1)@k$ for the contributing runs for the ClueWeb10 evaluation, plotted as a function of k. The right column shows the corresponding condensed metrics. The pooling depth was $d = 20$ throughout.



(a) RBP$(0.95)@k$ (b) ERR@k

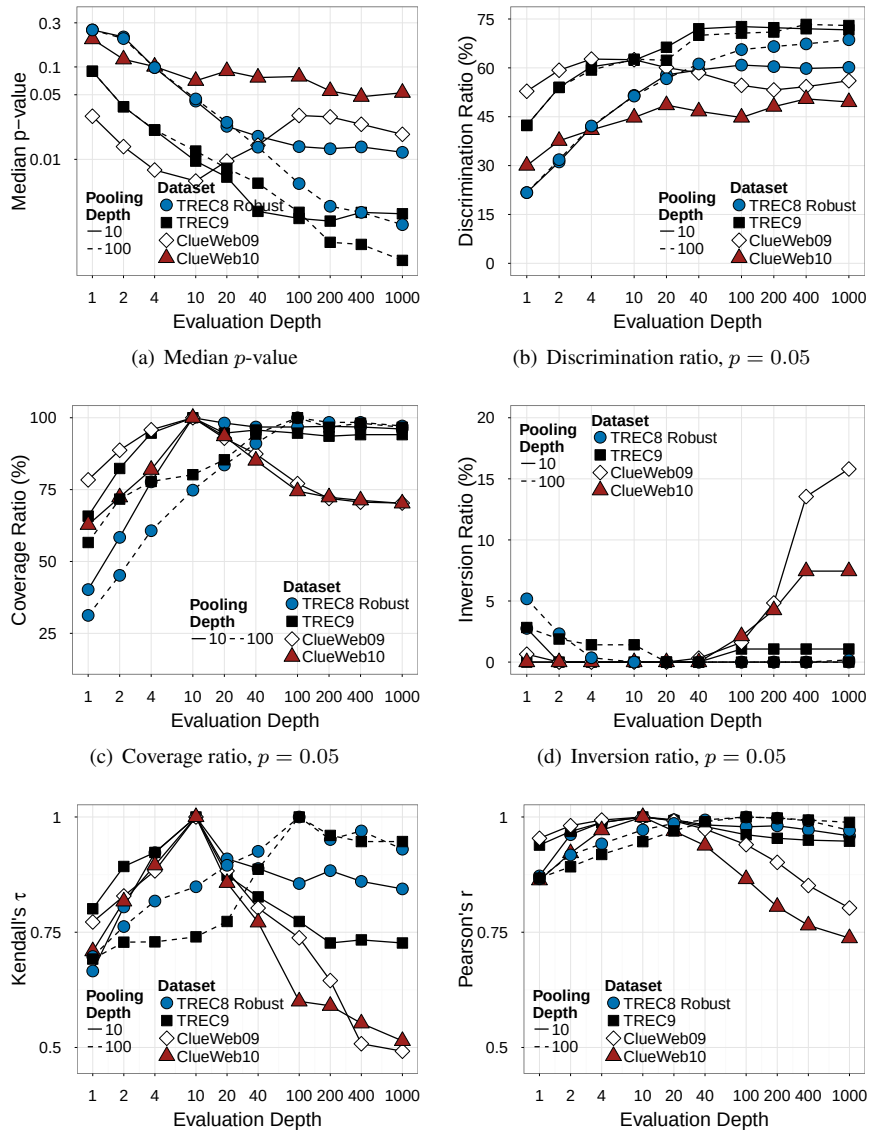
Fig. 3: System scores of RBP$(0.95)@k$ and ERR@k on ClueWeb10, using the same experimental framework as was employed in Figure 2.

represents one system, with five of the systems (the minimum, first quartile, median, third quartile, and maximum, all as of the right-hand edge of each graph) emphasized with darker lines so that the overall pattern can be discerned. In the left column of graphs, Figure 2 plots $\text{AP}_\text{b}@k$, $\text{NDCG}_\text{a}@k$, and $\text{QM}_{\text{a,i}}(1)@k$ system scores. The three graphs in the right column use the same metrics, but using condensed runs, with all unjudged documents removed. The vertical dotted line in each graph shows the pooling depth d. In very broad terms, as the evaluation depth is varied most systems exhibit the same characteristic score pattern for each of these three recall-based metrics. But there are also substantial differences, and each point at which the lines cross represents a change to the measured system orderings. These six plots also provide a tangible reminder that while these three metrics nominally converge to final values as the evaluation depth increases, that process is by no means monotonic.

Extended evaluation brings with it an increased likelihood of unjudged documents being retrieved. In the graphs in the left column of Figure 2, unjudged documents are treated as non-relevant. In the right-hand column, unjudged documents are removed, to make condensed runs; in these, when all available judgments are exhausted, the score of each system becomes constant. The scores assigned by the condensed metrics are also non-monotonic, and system cross-overs occur at around the same rate in the right-hand column as they do in the left-hand column.

Figure 3 explores score patterns for utility-based metrics, using RBP$(0.95)@k$ as a typical deep metric, and ERR@k as a typical shallow metric, using the same methodology as was applied to generate Figure 2. Compared to the recall-based metrics, more stable system orderings are observed as evaluation depth is varied, including in the case when extended evaluation is employed. The same stable pattern of system orderings for utility-based metrics is also observed on the other three test collections, in results that are not included here.

Dependence on Evaluation Depth In order to further measure how reliable conclusions can be drawn by different metrics on large-scale collections, we show the



(a) Median p-value

(b) Discrimination ratio, $p = 0.05$

(c) Coverage ratio, $p = 0.05$

(d) Inversion ratio, $p = 0.05$

(e) Kendall's τ

(f) Pearson's r

Fig. 4: Six different ways of measuring a metric's dependence on evaluation depth k, using NDCG_a throughout. Two different pooling depths are used with the two NewsWire datasets, $d \in \{10, 100\}$; the two ClueWeb collections are both presented using $d = 10$. In panes (c)–(f), the reference metric M_{ref} is $\text{NDCG}_\text{a}@d$, that is, the truncated version of NDCG pooled and evaluated to depth d. Note that the vertical scales differ in each of the panes.

experimental results of calculated correlation coefficient in this section. Figures 2 and 3 make it clear that the three recall-based metrics shown are more sensitive to the evaluation depth k than are the two utility-based ones. Figure 4 and Table 7 further quantify the extent of that dependence. In Figure 4, $\text{NDCG}_\text{a}@k$ is used in all six panes, and systems scores and system orderings are compared at different evaluation depths k using several of the approaches described in Section 3. Four different TREC experimental datasets are used (see Table 3), all adjusted so that they are pooled to a uniform depth of $d = 10$; plus the two NewsWire datasets are also used pulled to their judged depth of $d = 100$. That is, there are six experimental contexts in total, four pooled to a depth of $d = 10$; and two pooled to a depth of $d = 100$. Working through the six components in Figure 4 in order, it can be seen that:

- In pane (a), at each evaluation depth k the two shallow-pooled ClueWeb datasets have higher median p values for system-versus-system comparisons when compared to the TREC 8 Robust and TREC 9 datasets. That difference remains evident even when the latter two are shallow-pooled to $d = 10$. In general, the median p-value decreases as the evaluation depth increases towards the pooling depth d, as more information is introduced in to the computation of metric scores. But note also that in the two ClueWeb environments the median p-value then *increases* past $k = 10$, that is, once the region of extended evaluation is reached.
- In pane (b), the same relativities occur when the metric is assessed via discrimination ratio, rather than by median p value. This graph supports the earlier hypothesis that a high median p value is indicative of a low discrimination ratio. In particular, when $k > d$ and the evaluation depth is greater than the pooling depth, extended evaluation results in a smaller ratio of system-versus-system pairs being deemed significant at $p = 0.05$. That is, pane (a) and pane (b) reinforce each other in this regard.
- In pane (c), which plots coverage ratios as a function of evaluation depth k, a reference metric M_{ref} is used for each set of lines, four of them with $k = d = 10$, and two with $k = d = 100$. The same pattern of behavior is again evident: as k increases towards d, coverage increases, and the at-k evaluation converges towards the at-d measurements. But once k is greater than d, the two ClueWeb collections give rise to divergent outcomes, and the coverage ratio decreases.
- In pane (d) it can be seen that the divergence in coverage ratio beyond $k = d$ is not simply a matter of losing the ability to derive statistical significance. The relatively high inversion ratios that arise for the two ClueWeb collections is evidence that a non-trivial fraction of the significant system-versus-system outcomes identified at $k = d$ are actively contradicted at larger values of k. Note also that the two NewsWire collections, even when shallow-pooled to $d = 10$, do not behave in this manner, and do not give rise to the same high levels of inversions.
- Panes (e) and (f) show that there is also a substantial decrease in measured system order correlation, using both Kendall's τ and Pearson's r when extended evaluation is used with the two ClueWeb collections. In these two graphs the reference metric M_{ref} is again taken to be $\text{NDCG}_\text{a}@d$, where d is either 10 or 100.

Overall, all of the six depicted indicators agree that extended evaluation using the two ClueWeb collections gives rise to system comparisons that diverge from those

that arise when the evaluation depth k is restricted to the pooling depth d. Similar behavior (not shown, so as to avoid cluttering the graphs) was also found when the same experiments were carried out using the ClueWeb11 and ClueWeb12 test collections.

Table 7 shows that these effects are not limited to NDCG_a, and that other recall-based metrics give rise to the same effects. In Table 7(a), the TREC8 Robust dataset is used, with pooling to depth $d = 100$, and the reference depth for coverage ratio and inversion ratio is $d = 100$. With this collection, system relativities are stable in extended evaluation, and for $k > 100$ the coverage ratios are close to 100, and inversion ratios are very close to zero. Table 7(a) also confirms previous findings in regard to discrimination ratio: on this dataset, NDCG_a and AP_b both have discrimination ratios of around 65%, more than ten points higher than the two RBP variants.

The situation is different when the ClueWeb10 dataset is used (Table 7(b)). On this collection the recall-based metrics AP_b and $\text{QM}_{\text{a,i}}$ give rise to coverage ratios well below 100% in extended evaluation, and inversion rates that are greater than 5%. On the other hand, the two RBP variants are relatively unaffected by the exact choice of k, and give rise to system-versus-system comparisons on the ClueWeb datasets that are resilient to the evaluation depth k. Note also the discrimination ratios in Table 7(b). Now the strongly top-weighted $\text{RBP}(0.8)$ metric yields system-versus-system comparisons that match or better the corresponding rates for the three recall-based metrics.

Missing Judgments So far, we perform the comparisons by ignoring the incompleteness of judgments in large-scale collections. However, in this section, we consider the third research question by applying existing solutions to previously measured metrics. Table 7 suggests that it may be injudicious to apply extended evaluation to the ClueWeb10 dataset without seeking to address the problem of missing judgments. Table 8 gives the equivalent results for condensed runs, corresponding to the right-hand column of graphs in Figure 2, and to Table 7(b). To carry out these measurements, all unjudged documents were removed from the contributing runs, and they were then processed to the shown depths, or to a depth equal to the number of judged documents in the run if that number was less than k. For the ClueWeb10 collection, that process resulted in an average per-system per-topic run length of 244.9 documents, and a maximum run length across the systems and topics of 391 documents. In terms of discrimination ratio, coverage ratio, and inversion ratio, there is no benefit gained by condensing the runs; if anything, condensing the runs has amplified the way in which system-versus-system comparisons are affected by the choice of k.

Table 9 shows the influence of the unjudged documents in the runs. Utility-based metrics such as RBP use pre-determined weights to compute an overall score for a ranking, and it is thus straightforward to accumulate the weights associated with any unjudged documents as a *residual* and provide an exact limit on the extent of the score uncertainty attributable to incomplete judgments [15]. Table 9 records RBP residuals for two different values of the RBP parameter p. As can be seen, shallow evaluations yield both higher numeric RBP scores and lower numeric residuals, implying more accuracy in the values that are presented as being "RBP scores". Note also that as k increases, the residual cannot increase (Equation 4); this is a useful attribute of utility-based metrics, and explains the smooth score convergence already documented in

k	$AP_b@k$			$QM_{a,i}@k$			$RBP(0.95)@k$		
	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.
1	21.1	31.1	5.2	20.8	30.3	7.4	21.1	33.7	5.3
4	40.9	58.7	2.1	39.5	56.0	1.6	43.0	67.2	0.4
10	51.9	75.3	0.0	51.7	74.7	0.7	53.4	84.0	0.0
40	63.2	93.2	0.0	63.4	92.8	0.0	61.3	98.8	0.0
100	66.9	100.0	0.0	67.4	100.0	0.0	61.0	100.0	0.0
400	67.9	99.3	0.0	68.5	99.8	0.0	61.0	100.0	0.0
1000	68.4	99.7	0.0	70.0	99.8	0.0	61.0	100.0	0.0
k	$RR@k$			$ERR@k$			$RBP(0.8)@k$		
	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.
1	23.5	100.0	0.0	13.4	56.8	0.0	21.1	40.3	3.6
4	23.5	100.0	0.0	20.0	86.5	0.0	41.5	81.1	0.0
10	23.5	100.0	0.0	22.8	96.9	0.0	49.0	95.4	0.0
40	23.5	100.0	0.0	22.3	100.0	0.0	51.0	100.0	0.0
100	23.5	100.0	0.0	22.3	100.0	0.0	51.0	100.0	0.0
400	23.5	100.0	0.0	22.3	100.0	0.0	51.0	100.0	0.0
1000	23.5	100.0	0.0	22.3	100.0	0.0	51.0	100.0	0.0

(a) Evaluation using the TREC 8 Robust collection, with $d = 100$ in all cases.

k	$AP_b@k$			$QM_{a,i}@k$			$RBP(0.95)@k$		
	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.
1	30.0	53.2	8.1	36.7	54.3	6.7	30.0	55.3	1.0
4	46.2	82.9	0.0	48.6	81.9	0.0	41.4	74.8	0.0
10	52.9	100.0	0.0	50.0	100.0	0.0	49.0	100.0	0.0
40	48.6	82.9	0.0	47.1	81.9	0.0	49.0	80.6	0.0
100	49.5	79.3	0.0	48.6	74.3	1.9	47.6	79.6	1.0
400	51.0	78.4	1.8	50.5	73.3	4.8	48.6	79.6	1.0
1000	50.5	77.5	1.8	50.0	72.4	6.7	48.6	79.6	1.0
k	$RR@k$			$ERR@k$			$RBP(0.8)@k$		
	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.
1	41.4	100.0	0.0	33.8	66.7	0.0	30.0	63.4	0.0
4	41.4	100.0	0.0	48.6	87.6	0.0	40.5	82.8	0.0
10	41.4	100.0	0.0	50.0	100.0	0.0	44.3	100.0	0.0
40	41.4	100.0	0.0	48.1	96.2	0.0	45.7	97.8	0.0
100	41.4	100.0	0.0	48.1	96.2	0.0	45.7	97.8	0.0
400	41.4	100.0	0.0	48.6	96.2	0.0	45.7	97.8	0.0
1000	41.4	100.0	0.0	48.6	96.2	0.0	45.7	97.8	0.0

(b) Evaluation using the ClueWeb10 collection, with $d = 10$ in all cases.

Table 7: Discrimination, coverage, and observed inversion ratios (all expressed as percentages at $p = 0.05$) for six metrics at different evaluation depths k, and two different document collections. The reference metric M_{ref} for the coverage and inversion ratios is the corresponding metric at depth $k = d = 100$ and $k = d = 10$, respectively (see Tables 3 and 4 for pooling details).

k	$\text{AP}_\text{b}@k$			$\text{QM}_\text{a,i}(1)@k$			$\text{RBP}(0.95)@k$			$\text{RBP}(0.8)@k$		
	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.	Disc.	Covr.	Invr.
1	30.0	47.4	14.9	35.7	45.5	17.4	30.0	54.7	4.7	30.0	61.4	0.0
4	45.7	75.4	0.0	48.1	71.1	0.0	41.4	72.6	0.0	40.5	80.2	0.0
10	52.4	92.1	0.0	51.9	87.6	0.0	49.0	72.6	0.0	44.3	96.9	0.0
40	53.3	95.6	0.0	57.6	96.7	0.0	48.1	91.5	0.0	46.2	100.0	0.0
100	53.3	82.4	0.0	53.3	81.0	0.0	48.6	90.6	0.0	46.2	100.0	0.0
400	41.4	44.7	15.8	47.6	43.0	28.9	48.6	90.6	0.0	46.2	100.0	0.0

Table 8: Discrimination, coverage, and inversion ratios when two recall-based metrics and one utility-based metric are applied to condensed runs derived from the ClueWeb10 dataset. By evaluation depth $k = 400$ all applicable judgments are in use. All other settings are as for Table 7(b).

k	TREC 8 Robust04			ClueWeb10		
	$\text{RBP}(0.95)@k$	$\text{RBP}(0.8)@k$	ERR	$\text{RBP}(0.95)@k$	$\text{RBP}(0.8)@k$	ERR
1	0.031+ 0.950	0.123+0.800	0.290+0.291	0.019+ 0.950	0.074+0.800	0.088+0.385
4	0.097+ 0.815	0.315+0.410	0.404+0.060	0.063+ 0.815	0.203+0.410	0.146+0.088
10	0.183+ 0.599	0.439+0.107	0.430+0.019	0.133+ 0.599	0.301+0.107	0.170+0.044
40	0.300+ 0.129	0.470+0.000	0.437+0.003	0.249+ 0.234	0.332+0.005	0.183+0.020
100	0.317+ 0.006	0.469+0.000	0.437+0.001	0.265+ 0.190	0.332+0.005	0.185+0.020
400	0.317+ 0.002	0.470+0.000	0.439+0.001	0.266+ 0.189	0.332+0.005	0.185+0.020
1000	0.317+ 0.002	0.470+0.000	0.439+0.001	0.266+ 0.189	0.332+0.005	0.185+0.020

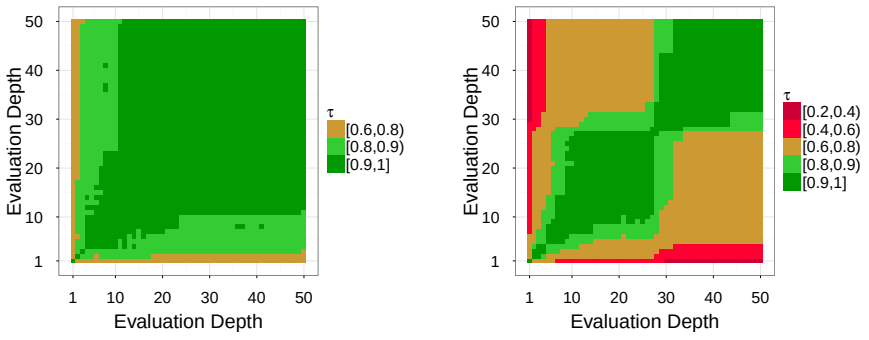
Table 9: Run scores and residuals generated by $\text{RBP}@k$, averaged across contributing systems and across topics, for two collections, two values of the RBP parameter p, and a range of evaluation depths k. In the case of the TREC 8 Robust data, pooling is to depth $d = 100$; in the case of the ClueWeb10 data, pooling is to depth $d = 20$.

Figure 3. While it is not possible to compute residuals for recall-based metrics, all of AP, NDCG, and QM are likely to be subject to score uncertainties to at least the same extent as $\text{RBP}(0.95)$.

Comparing System Orderings The previous experiments answer the third research question by assuming a reliable conclusion can be obtained by evaluating to the pooling depth, as an extension, in this section, we show the visualization of the changing system orderings at different evaluation depths.

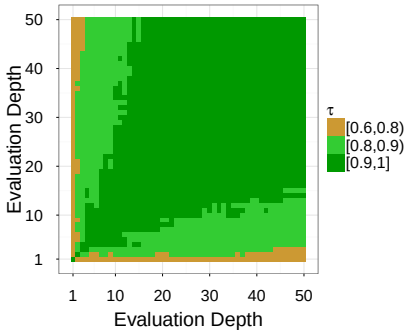
Tables 7 and 8 demonstrate that when the evaluation depth is greater than the pooling depth, system-versus-system comparisons can lead to notably different outcomes when compared to the same evaluation carried out at the pooling depth. If the evaluation depth is limited to be less than or equal to the pooling depth, system orderings still vary with recall-based metrics, as illustrated by the crossing score lines in Figure 2.

Figure 5 gives another view of the same phenomenon. To construct each of the panes in the figure, each contributing system for a dataset was scored using evaluation depths $1 \le k \le 50$ using a single metric, and with a pool depth of $d = 20$

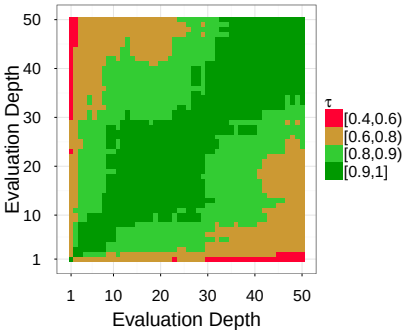


(a) $AP_b@k$ on TREC 8 Robust

(b) $AP_b@k$ on ClueWeb 10



(c) $NDCG_a@k$ on TREC 8 Robust

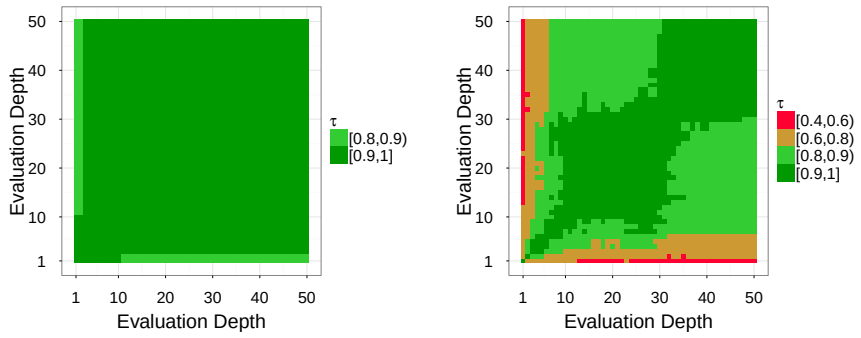


(d) $NDCG_a@k$ on ClueWeb 10

Fig. 5: Recall-based metrics applied to the TREC 8 Robust and ClueWeb10 test collections, using evaluation depths $1 \le k \le 50$, and with the resultant system orderings compared using Kendall's τ. Both collections are pooled to depth $d = 20$. The top row shows $AP_b@k$, and lower row $NDCG_a@k$, with Trec 8 Robust on the left and ClueWeb10 on the right.

throughout. Then, for each distinct value of k, a system ordering was generated using the system scores when averaged over the topic set. Kendall's τ coefficients were then computed between all pairs of system orderings, to create a volatility matrix of the form described by Equation 7 in Section 3. Finally, color coding is used to indicate different ranges for τ. If the system ordering was unaffected by evaluation depth, then all entries in the volatility matrix would be dark green. The diagonal line of each graph must be dark green; variance away from that color in other locations of the plot indicates disagreement in the induced system ordering. Correlations of 0.9 and greater represent system orderings that are very close to each other.

The four panes in Figure 5 correspond to two different metrics ($AP_b@k$ and $NDCG_a@k$) and two different datasets. On the TREC 8 Robust dataset (panes (a) and (c)) the outcome is as might be expected – the system orderings generated by the metric are largely unaffected by evaluation depth, with the only exception being that very small values of k give system orderings that compare poorly with the orderings



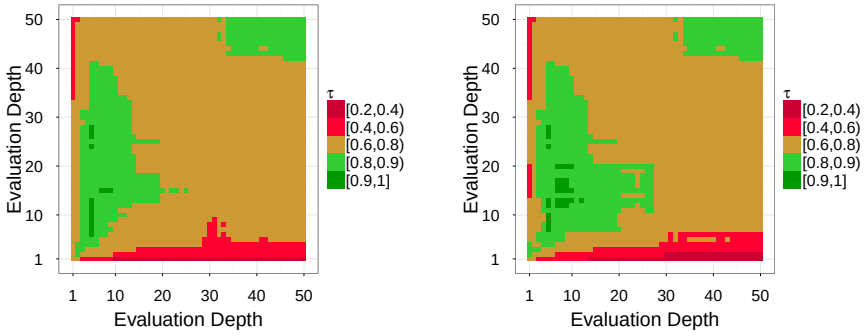
(a) ERR@k (b) RBP(0.95)@k

Fig. 6: Kendall's τ volatility matrices for two utility-based metrics on ClueWeb10, with other settings as shown in Figure 5. The left pane shows the shallow metric ERR@k; the right pane the deeper metric RBP(0.95)@k. The shallower RBP(0.8)@k has behavior that fits between the two arrangements that are shown.

generated by larger values of k. These results agree with the findings of Webber et al. [30] noted the extended evaluation give rise to increased discrimination between systems. Markedly different behavior is observed for the ClueWeb10 dataset in panes (b) and (d). Now there are many low correlation scores in evidence, and only when the evaluation depth k is approximately the same as the pooling depth d are the generated orderings in broad agreement. Similar patterns of behavior also arise when Kendall's τ is replaced by τ_{ap} or RBO.

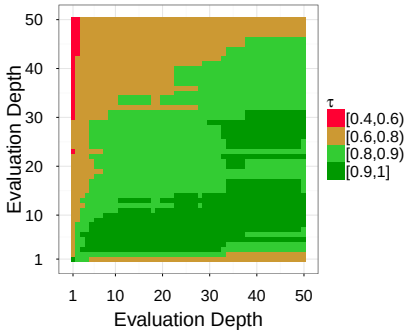
Figure 6 shows the corresponding ClueWeb10 heat maps for two utility-based metrics, and confirms their relative stability, as already shown in Figure 3. Despite the relatively low ability of ERR@k to provide discrimination in system-versus-system evaluations, Figure 6(a) demonstrates that it provides consistent system orderings across a broad range of evaluation depths k. In Figure 6(b), RBP(0.95)@k also provides consistent system orderings across a wide range of evaluation depths, but with a higher discrimination ratio (Table 7). This is a consequence of its gradated emphasis on documents beyond the pooling depth, with contributions arising from relevant documents right through until the evaluation depth k. At an evaluation depth of $k = 40$ the residuals are also relatively small (Table 9), with the exception of ClueWeb 10 and RBP(0.95). Here, the residuals are uncomfortably high when compared with the score, even at $k = 1000$.

Comparing Metrics Independent of the evaluation depth, we consider our last research question (RQ4) by first comparing between system orderings and metric choice. It is also possible to compare the relative behavior of metrics, to determine configurations in which they do and do not yield similar system rankings, by placing different metrics on the two axes. Figure 7 shows a number of such comparisons. In the first row, the normalized AP_b@k metric is compared with the two other truncated AP@k variants introduced in Table 2. The second row compares NDCG_a@k with

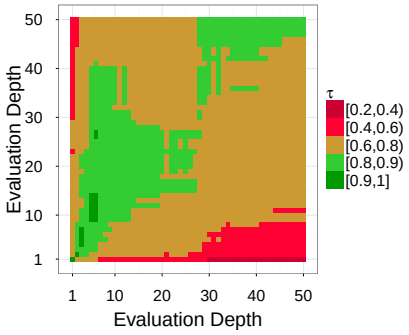


(a) $\text{AP}_\text{a}@k$ (vert.) vs. $\text{AP}_\text{b}@k$ (hori.)

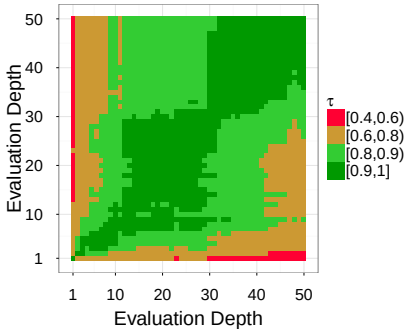
(b) $\text{AP}_\text{c}@k$ (vert.) vs. $\text{AP}_\text{b}@k$ (hori.)



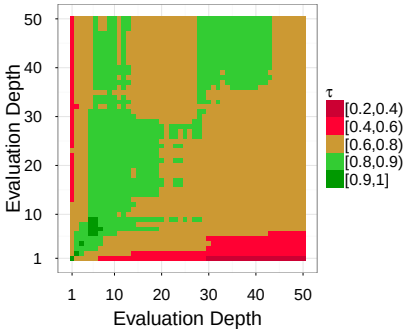
(c) $\text{NDCG}_\text{a}@k$: Log. (vert.) vs. Zipf (hori.) disc.



(d) $\text{NDCG}_\text{a}@k$ (vert.) vs. $\text{AP}_\text{b}@k$ (hori.)



(e) $\text{RBP}(0.95)@k$ (vert.) vs. $\text{NDCG}_\text{a}@k$ (hori.)



(f) $\text{RBP}(0.95)@k$ (vert.) vs. $\text{AP}_\text{b}@k$ (hori.)

Fig. 7: Kendall's τ scores on system rankings using ClueWeb10, pooled to depth $d = 20$, including using variants of AP@k and NDCG@k. The first row compares the normalized $\text{AP}_\text{b}@k$ metric used through the balance of this paper with $\text{AP}_\text{a}@k$ (the `trec_eval` variant), and with $\text{AP}_\text{c}@k$ (see Table 2). The second row compares $\text{NDCG}_\text{a}@k$ with a variant of NDCG that uses an alternative Zipfian discounting strategy, and with $\text{AP}_\text{b}@k$; and the third row compares $\text{RBP}(0.95)@k$ with $\text{NDCG}_\text{a}@k$, and with $\text{AP}_\text{b}@k$.

a variant that uses an alternative Zipfian discounting function, $D(i) = 1/i$ rather than $D(i) = 1/\log_2(1+i)$ [13]; and with $\text{AP}_\text{b}@k$. The third row completes the comparison by comparing $\text{RBP}(0.95)@k$ with $\text{NDCG}_\text{a}@k$ and $\text{AP}_\text{b}@k$. Several other comparisons such as $\text{NDCG}_\text{a}@k$ and $\text{NDCG}_\text{b}@k$ were carried out, and similar trends were observed. For example the $\text{NDCG}_\text{a}@k$ and $\text{NDCG}_\text{b}@k$ comparison graphs looks quite similar to the $\text{RBP}(0.95)@k$ and $\text{NDCG}_\text{a}@k$ comparison. In Figure 7c, evaluation of NDCG using a small value of k and the logarithmic discounting function behaves somewhat akin to a deeper evaluation using an alternative Zipfian discounting function, perhaps because the Zipfian discount means that any gains from deep-ranked items are rendered inconsequential. And in Figure 7(e), the strong symmetric pattern of correlation indicates that $\text{RBP}(0.95)$ and NDCG behave similarly across a broad range of retrieval depths, despite their different discounting and normalization regimes.

Metric Parameters Revisited The differing patterns of behavior that are evident in Figures 5 and 6 are partly a consequence of the differing nature of the NewsWire and ClueWeb datasets (see Figure 1), and partly a consequence of the fact that in $\text{AP}@k$ and $\text{NDCG}@k$ there are two different concepts combined in to a single variable. The first concept is essentially a top-weightedness parameter, corresponding to the variable p that is used in RBP. It determines the relative weighting given to each document when the relevance scores are combined to obtain an overall score for a run. For example, assuming that there are $R_d \geq k$ relevant documents and that $\text{NDCG}@k$ is being used, the first document in the ranking accounts for $1/(\sum_{i=1}^{k} 1/\log_2(1+i))$ of the total metric value. When $k = 5$, that is around 0.34; when $k = 10$, around 0.22, and when $k = 20$, around 0.14. Elements beyond depth k are always assigned a weighting of zero, in the same way that elements beyond rank k are assigned a weighting of zero by $\text{Prec}@k$. In this context, $\text{NDCG}@k_1$ and $\text{NDCG}@k_2$ must be regarded as being different metrics, in the same way that $\text{Prec}@k_1$ and $\text{Prec}@k_2$ are, or that $\text{RBP}(p_1)$ and $\text{RBP}(p_2)$ are. When k_1 is close to k_2 (or p_1 close to p_2), the numeric scores and numeric behaviors of $\text{NDCG}@k_1$ and $\text{NDCG}@k_2$ (or $\text{RBP}(p_1)$ and $\text{RBP}(p_2)$) are likely to be correlated. But when k_1 and k_2 are not close, there is no more requirement that $\text{NDCG}@k_1$ and $\text{NDCG}@k_2$ be correlated than there is that (say) ERR and AP should be correlated. A similar argument applies to $\text{AP}@k$. That is, the truncated metrics $\text{AP}@k$ and $\text{NDCG}@k$ should not be regarded as approximations of full-depth AP and full-depth NDCG; rather, they must be considered to be independent metrics in their own right.

The second concept associated with k is that it determines the degree to which what is reported as the metric's score is an approximation of the correct value. For example, in the case of $\text{RBP}(p)@k$, the parameter p indicates the strength of top-weightedness, and the value of k then controls the extent to which the sum that is computed is an approximation of its correct value. The balance is taken to be the residual. In the case of $\text{NDCG}_\text{a}@k$ and $\text{AP}_\text{b}@k$, the conflation of the two concepts – with k serving both as a top-weightedness parameter, and also a limit governing the summation – means that there is no sense of increasing k to get a "better" approximation of the metric. Increasing k changes the metric to make it less top-focused, and in doing so shifts weight further down the ranking; perversely, that shift may then

make the approximation that is computed *less* accurate, since a greater fraction of the metrics' weighting might consist of unjudged documents.

With $\text{AP}_\text{a}@k$ (Table 2), the situation is different again. Because R_d is assumed to be fixed and independent of k, $\text{AP}_\text{a}@k$ is non-decreasing in k, albeit with a possibly very large residual. With this metric, it is increasing the *pooling* depth rather than the evaluation depth that makes it less top-weighted. That is, with $\text{AP}_\text{a}@k$ the pooling depth that is used must be considered to be a parameter of the metric, rather than the evaluation depth k. A similar complex relationship exists with $\text{NDCG}_\text{a}@k$ when $k \le R_d$ – increasing the pooling depth d in order to obtain a more comprehensive evaluation makes the scores less top-weighted and hence risks increasing the amount of uncertainty that is implicit in the measured scores, and in downstream uses such as system-versus-system comparisons.

With utility-based metrics such as ERR and RBP, increasing the evaluation depth k and/or the pooling depth d serve only to reduce the uncertainty in the measured scores, and neither change can alter the degree of top-weightedness in the evaluation.

5 Conclusion

We have explored the role that the metric evaluation depth k plays in affecting metric values and system-versus-system evaluations, paying particular attention to the different characteristics between recall-based and utility-based metrics. Two collection types, TREC NewsWire data and the ClueWeb dataset have been used, together with the system runs that contributed to the relevance judgments when those collections were formed.

We have focused on the use of truncated evaluation of effectiveness metrics. Our experiments have revealed that evaluating recall-based metrics to a given depth k based on pooling to a different depth d can in some circumstances be equivalent to setting a parameter that determines the degree of top-weightedness in the evaluation, with larger values of k decreasing the weight assigned to items that appear early in the ranking. This relationship means that altering k is tantamount to making use of a different metric, and that increasing the evaluation depth is not in any way a guarantee of a more accurate evaluation. Indeed, with variants of two standard recall-based metrics, system-versus-system evaluations using the ClueWeb resources have been shown – in a range of different ways – to be highly dependent on the value of k that is used.

On the other hand, truncated evaluation of utility-based metrics, in which the top-weightedness parameter (p in the case of RBP, and the chosen gain regime in the case of ERR) are relatively stable in their behavior as k and d are varied. With these metrics, there is an explicit parameter that specifies the decay in weight as the set of documents comprising the ranking are one-by-one incorporated in to the computed metric value. This makes the scores that these metrics generate more resilient to changes to k and/or d, and hence makes comparisons based on them more robust as dataset resources are amended and augmented.

By way of conclusion, and in terms of practical advice to researchers, we offer these guidelines:

– that unless there are clear reasons to do otherwise, system comparisons should be made using standard metrics evaluated to standard depths k;
– that in the case of recall-based metrics, those depths should be fixed in advance of the experimentation being commenced, and should not be revisited as conclusions are drawn;
– that, particularly in the case of recall-based metrics, the precise form of the truncated computation be carefully described, or a standard publicly-available software tool be used (and named, together with the command-line options used to execute it);
– that if it is not possible for a standard tool to be used, attention be given to the tie-breaking regime, with a suggestion that original run orderings should be adhered to rather than altered;
– that extended evaluation, the use of evaluation depths k greater than the pooling depth d, should be avoided unless it has already been demonstrated that the dataset is amenable to it;
– that utility-based metrics should be featured alongside recall-based ones if the latter are being regarded as the primary point of comparison; and
– that if systems that did not contribute to the judgments pool are being included in any comparison, careful attention should be paid to the number and rank positions of the unjudged documents in the corresponding runs, preferably via the computation of a residual or some other indicator of score imprecision.

Many researchers have already adopted some or all of these recommendations. With this work we hope to encourage others to follow suit.

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

References

1. J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 27–34, 2005.
2. J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 541–548, 2006.
3. P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User variability and IR system evaluation. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 625–634, 2015.
4. C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 33–40, 2000.
5. C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 25–32, 2004.
6. C. Buckley, D. Dimmick, I. Soboroff, and E. M. Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval J.*, pages 491–508, 2007.
7. S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 63–70, 2007.

8. B. Carterette, E. Kanoulas, and E. Yilmaz. Low cost evaluation in information retrieval. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, page 903, 2010.
9. O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. Conf. Information and Knowledge Management*, pages 621–630. ACM, 2009.
10. C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 Web track. In *Proceedings of TREC*, volume 10, 2010.
11. G. Demartini and S. Mizzaro. A classification of IR effectiveness metrics. In *Advances in Information Retrieval*, pages 488–491. Springer, 2006.
12. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Information Systems*, 20(4):422–446, 2002.
13. E. Kanoulas and J. A. Aslam. Empirical justification of the gain and discount function for NDCG. In *Proc. Conf. Information and Knowledge Management*, pages 611–620. ACM, 2009.
14. A. Moffat. Seven numeric properties of effectiveness metrics. In *Proc. Asian Information Retrieval Societies Conf.*, pages 1–12, 2013.
15. A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Information Systems*, 27(1):2, 2008.
16. A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 375–382, 2007.
17. A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. Conf. Information and Knowledge Management*, pages 659–668, 2013.
18. A. Moffat, P. Bailey, F. Scholer, and P. Thomas. INST: An adaptive metric for information retrieval evaluation. In *Proc. Aust. Doc. Comp. Symp.*, pages 5:1–5:4, 2015.
19. S. D. Ravana and A. Moffat. Score estimation, incomplete judgments, and significance testing in IR evaluation. In *Proc. Asian Information Retrieval Societies Conf.*, pages 97–109, 2010.
20. S. E. Roberston, E. Kanoulas, and E. Yilmaz. Extending average precision to graded relevance judgments. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 603–610, 2010.
21. T. Sakai. New performance metrics based on multigrade relevance: Their application to question answering. In *Proc. NII Testbeds and Communities for Information Access and Research*, 2004.
22. T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 525–532, New York, New York, USA, 2006. ACM Press.
23. T. Sakai. Alternatives to BPref. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 71–78, 2007.
24. T. Sakai. Metrics, statistics, tests. In *Bridging Between Information Retrieval and Databases*, pages 116–163. Springer, 2014.
25. T. Sakai and N. Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval J.*, 11(5):447–470, 2008.
26. M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.
27. E. M. Voorhees. Evaluation by highly relevant documents. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 74–82. ACM, 2001.
28. E. M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, pages 355–370. Springer, 2002.
29. E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.
30. W. Webber, A. Moffat, and J. Zobel. The effect of pooling and evaluation depth on metric stability. In *Proc. Wrkshp. Evaluation Information Access*, pages 7–15, 2010.

31. W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Trans. Information Systems*, 28(4):20, 2010.

32. E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proc. Conf. Information and Knowledge Management*, pages 102–111, 2006.

33. E. Yilmaz and S. Robertson. On the choice of effectiveness measures for learning to rank. *Information Retrieval J.*, 13(3):271–290, 2010.

34. E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 587–594. ACM, 2008.

35. E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 603–610, 2008.

36. J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. ACM-SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 307–314, 1998.