Observed Volatility in Effectiveness Metrics

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

1. School of Computer Science and Information Technology
RMIT University, Australia

2. Department of Computing and Information Systems
The University of Melbourne, Australia

ABSTRACT

Information retrieval research and commercial search system evaluation both rely heavily on the use of batch evaluation and numerical system comparisons using effectiveness metrics. Batch evaluation provides a relatively low-cost alternative to user studies, and permits repeatable and incrementally varying experimentation in research situations in which access to high-volume query/click streams is not possible. As a result, the IR community has invested considerable effort into formulating, justifying, comparing, and contrasting a large number of alternative metrics. In this paper we consider a very simple question: to what extent can the various metrics be said to give rise to stable scores; that is, evaluations in which the process of adding further relevance information creates refined score estimates rather than different score estimates. Underlying this question is a fundamental concern, namely, whether the numeric behavior of metrics provides confidence that comparative system evaluations based on the metrics are robust and defensible.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation*

Keywords

Experimentation, Measurement

1. INTRODUCTION

Information retrieval system performance is often reported in numeric terms, as evaluated relative to an effectiveness metric [11]. The goal is to construct a mapping between an ordered set of binary or graded relevance scores and the real numbers, so that system scores can be amalgamated over sets of topics, and compared using statistical techniques. Many such mappings have been devised, each with the intention of capturing some new nuance of what it means for a ranking to be "good". Section 2 briefly summarizes some of these metrics.

Much of the exploration of metrics has focused on their discrimination ability – the extent to which they give rise to system comparisons that reach some specified level of statistical confidence. Other issues that have been explored relate to the imprecision caused by the prevalence of incomplete judgments, and the extent to which any particular metric is vulnerable to uncertainty.

Our purpose in this brief paper is to examine the extent to which scores for metrics are stable, that is, are not volatile in the face of increased information. We do not (yet) have a formal definition for this concept, but one way of describing it is as the extent to which adding further judgments to a system comparison might disrupt the relativity of the initial system ordering because the metrics' scores for the individual runs are volatile. There have been previous attempts to quantify *stability* [3]. For example, if some system is measurably better than a group of other systems when evaluated using a set of topics and judgments, and then an extended set of judgments becomes available, how likely is it that the same system will still be deemed to be the best one? To help motivate this concept, consider the metric reciprocal rank (RR). As judgment resources are invested, a comparison based on RR will necessarily lead to a stable and consistent ordering of systems, since the first time a RR score for a run becomes non-zero is the last time that score changes, regardless of how many further judgments are undertaken. So, RR has a low level of score volatility, and system orderings based on RR should be stable and consistent as judgments are added to the evaluation.

In this preliminary exploration we show the differing levels of observed score volatility associated with four different TREC evaluations, using a suite of different metrics, and judgment pools of different depths. We find markedly different behavior between the TREC Newswire collections and the TREC ClueWeb collection, even when they are pooled to the same retrieval depth; and also find marked differences between utility-based effectiveness metrics such as rank-biased precision, and recall-based metrics such as average precision and normalized discounted cumulative gain.

2. BACKGROUND

Effectiveness Metrics The performance of information retrieval systems is often reported in numeric terms using one or more effectiveness metrics [11]. As a result, there has been considerable discussion of effectiveness metrics, since the choice of metric and any parameters that must be specified before it can be evaluated, can be thought of as also being a choice of user model. For example, rank-biased precision (RBP) [6] connects one set of rules describing presumed searcher behavior with a corresponding effectiveness metric; the expected reciprocal rank (ERR) metric [2] a different set of rules, and so on. Such metrics measure user perceptions of rankings based on the rate at which *utility* is gained, and assign scores based only on knowledge of the prefix of the ranking that the user inspected, relative to their presumed behavior [7].

Another family of metrics is based around the idea of measuring the ranking relative to the best that any system could have done. *Recall-based* metrics such as average precision (AP), normalized discounted cumulative gain (NDCG) [4], and the Q-Measure (QM) [10] assign scores to runs as a weighted fraction derived from both

RIGOR'15: SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results, Santiago, Chile, 13 August 2015.

Collection	Queries	d	Runs
Robust04	651–700	100	42
TREC9	451–500	100	25
CW09	1–50	12+	32
CW10	51–100	20	21

Table 1: Details of test collections used in the evaluation process. Only runs that contributed to the pool to a depth of d (or more) are included in the experimentation.

knowledge of what the user saw, and also what they did not. In these metrics, for example, if only a small number of relevant documents exist for a topic, and the user was shown those documents early in the ranking, then the metric score will be high, because the system that generated the run performed relatively well, even if in absolute terms there were only a few relevant documents provided.

Other ways of categorizing metrics have also been noted [5].

Pooling and evaluation depths We assume the relevance judgments are derived by *pooling* a set of *contributing runs* to a depth d, so that every document that appears at depth d or shallower in one or more of the runs is assigned a relevance score. Voorhees and Harman [12] describe this process.

Separate to the pooling depth, we are also interested in the *evaluation depth*, denoted by k, the depth to which each ranking is scored. Note that k will often be equal to d, but that there may also be circumstances under which $k < d$ or $k > d$ are used, with some of the judgments simply ignored in the first case, and an *extended evaluation* carried out in the second case. For example, a set of runs might be pooled to depth $d = 100$, and then evaluated using all of AP with $k = 100$ or $k = 1{,}000$; NDCG at a depth of $k = 20$; and Precision at a depth of $k = 5$.

In terms of notation used in this paper, where a single subscript is provided to a metric, for example, AP_k, it represents an evaluation depth, with an assumption that the pooling depth $d \geq k$ and hence that all required judgments are provided. Two subscripts are supplied when $k > d$ and extended evaluation is being carried out. For example, $\text{AP}_{1000,100}$ indicates that judgments to depth $d = 100$ are being used in an evaluation to depth $k = 1{,}000$. Extended evaluation raises the question of how to handle unjudged documents. Two different approaches are commonly used: either unjudged documents are deemed to be non-relevant; or unjudged documents are removed from the evaluation completely, and all deeper documents are assumed to move up the ranking by one position, to fill in the gap and form a *condensed run* [1]. Both of these approaches are used in the experiment described in the next section.

3. EXPERIMENTS

Experiment Setup We evaluate both Newswire and ClueWeb test collections. Details of these, and of the number of runs that contributed to the judgment pools, are listed in Table 1. Note that other runs were submitted in each of these rounds of experimentation, but did not take part in the pool construction, and hence are not necessarily fully judged to depth d. Only the contributing runs are included in our experimentation. The ClueWeb 2009 Ad-Hoc task adopted a sampling strategy to select documents to be judged; in this case we set d as the greatest available depth to which all runs for a topic had all of their documents judged.

Then, for each collection, for each of the contributing runs, and for a range of evaluation depths k up to and beyond the pooling depth, we evaluated five different effectiveness metrics. When $k \leq d$, all of the documents retrieved by any system have been judged; when $k > d$ only the first d documents in each run are certain to have been judged, but if other deeper documents have been judged (because, for example, they appear at a shallower depth in a different run) then those judgments are also used. Finally, we graphed metric scores as a function of evaluation depth k for the four collections and the five metrics, taking two forms of each metric: the usual one, in which unjudged documents are assumed to be non-relevant; and an alternative *condensed* version, in which unjudged documents are removed from the run, and all of the documents below them move up by one (more) position.

Average precision (AP) was evaluated using binarized relevance judgments; the other four metrics all make use of graded relevance. For RBP, we set the persistence parameter p to be 0.8, representing a relatively impatient user; and we set $\beta = 1$ for metric QM. The condensed results use the same metrics, but only the judged documents, assuming that all documents beyond the last judged one in each run are also irrelevant.

Experimental Results Figures 1, 2, 3 and 4 show the evolution of system scores for all of the contributing systems in the four rounds of TREC experimentation that are listed in Table 1. The left column of graphs in each figure shows the metric applied in the usual way, with unjudged documents taken to be non-relevant. The corresponding graphs in the right column show the corresponding condensed evaluations. The red dashed line in each plot denotes the pooling depth. Scores to the left of each red line represent fully-judged evaluations at the indicated values of the evaluation depth k. Scores to the right of the red lines are extended evaluations in which not all documents used in the evaluation have been judged.

Each graph shows one line plotted for each contributing system. Five systems are picked out with darker lines in each graph, based on the eventual system ordering arrived at when $k = 1{,}000$: the best system, the worst system, and the three systems defining the quartiles of the eventual distribution of systems. Those five systems are traced over the full range of k. Starting at the left and working to the right, crossings of the lines in a graph indicate changes of system ordering arising as more judgments are provided. Conversely, the more stable (and non-intersecting) the lines are, the more stable the system ordering induced by the metric. Note that score volatility need not correspond to system ordering volatility; and that we are interested in both.

NewsWire data Figures 1 and 2 show NewsWire data, with pooling depth of $d = 100$ in both bases. System orderings are moderately stable for the recall-based metrics in the first six panes, and very stable for the two utility-based metrics depicted in the lower four panes. In these two collections there is a clear sense that the majority of relevant documents have been identified by the deep pooling process. Even so, shallow evaluation depths of $k \leq 10$ generate confused signals, and should probably be avoided. Note the clear difference between the two types of metric – with the recall-based evaluations that are "relative to opportunity", scores can both rise and fall. The utility-based metrics give scores that can only rise.

ClueWeb data The two ClueWeb evaluations (Figures 3 and 4) use shallower judgments against a much larger collection, and it seems likely that only a minority of the relevant documents have been identified. Shallow evaluations using recall-based metrics with $k < 10$ are notably unstable, with many crossovers; the situation is only a little better at larger evaluation depths. The many intersections in the top six panes in these two figures, both left and right of the red

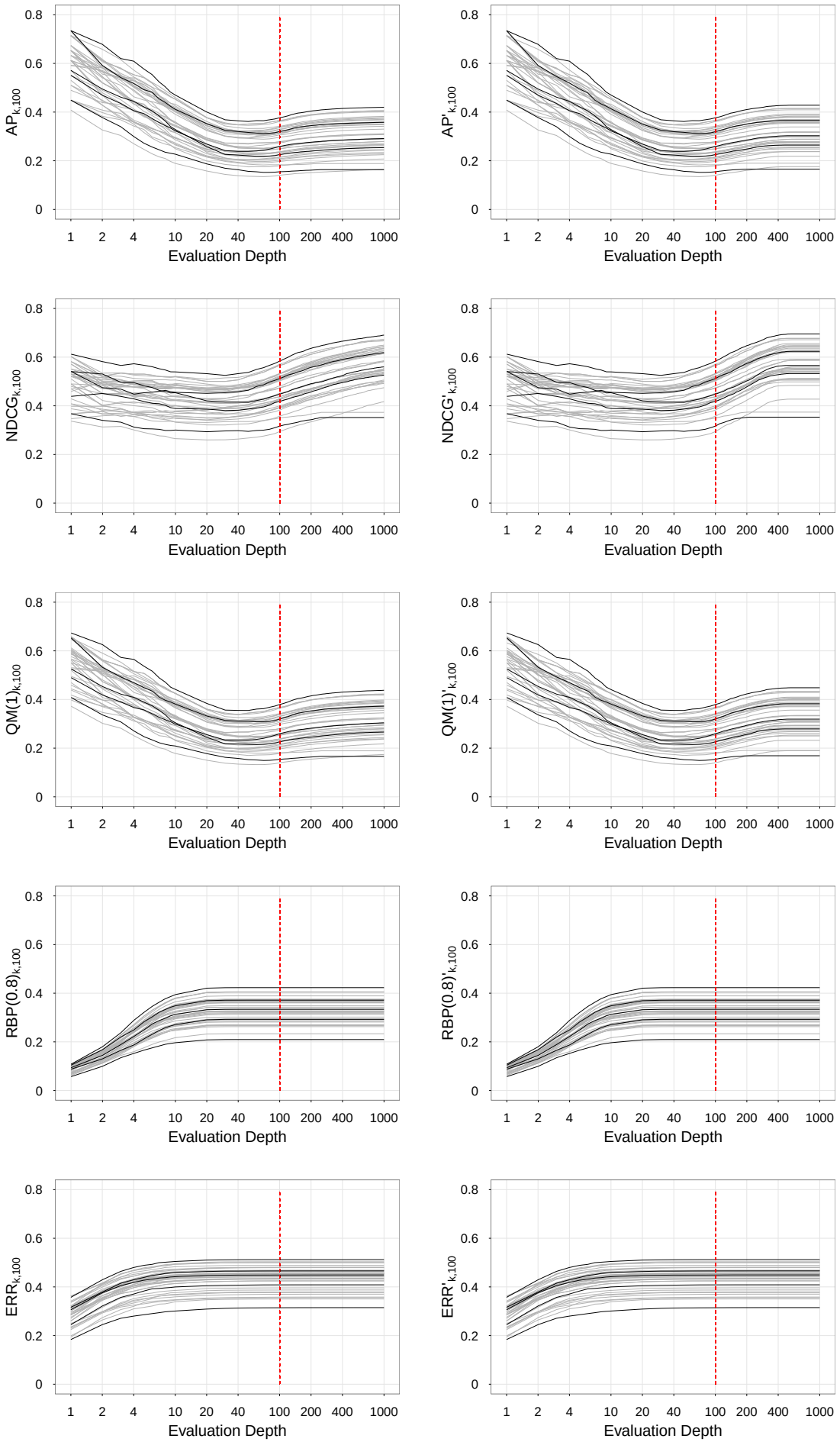

Figure 1: TREC8 Robust04 with all available judgments, using $\text{AP}_{k,100}$, $\text{NDCG}_{k,100}$, $\text{QM}(1)_{k,100}$, $\text{RBP}(0.8)_{k,100}$, and $\text{ERR}_{k,100}$. The right column shows corresponding condensed results. The dashed line denotes the pooling depth.

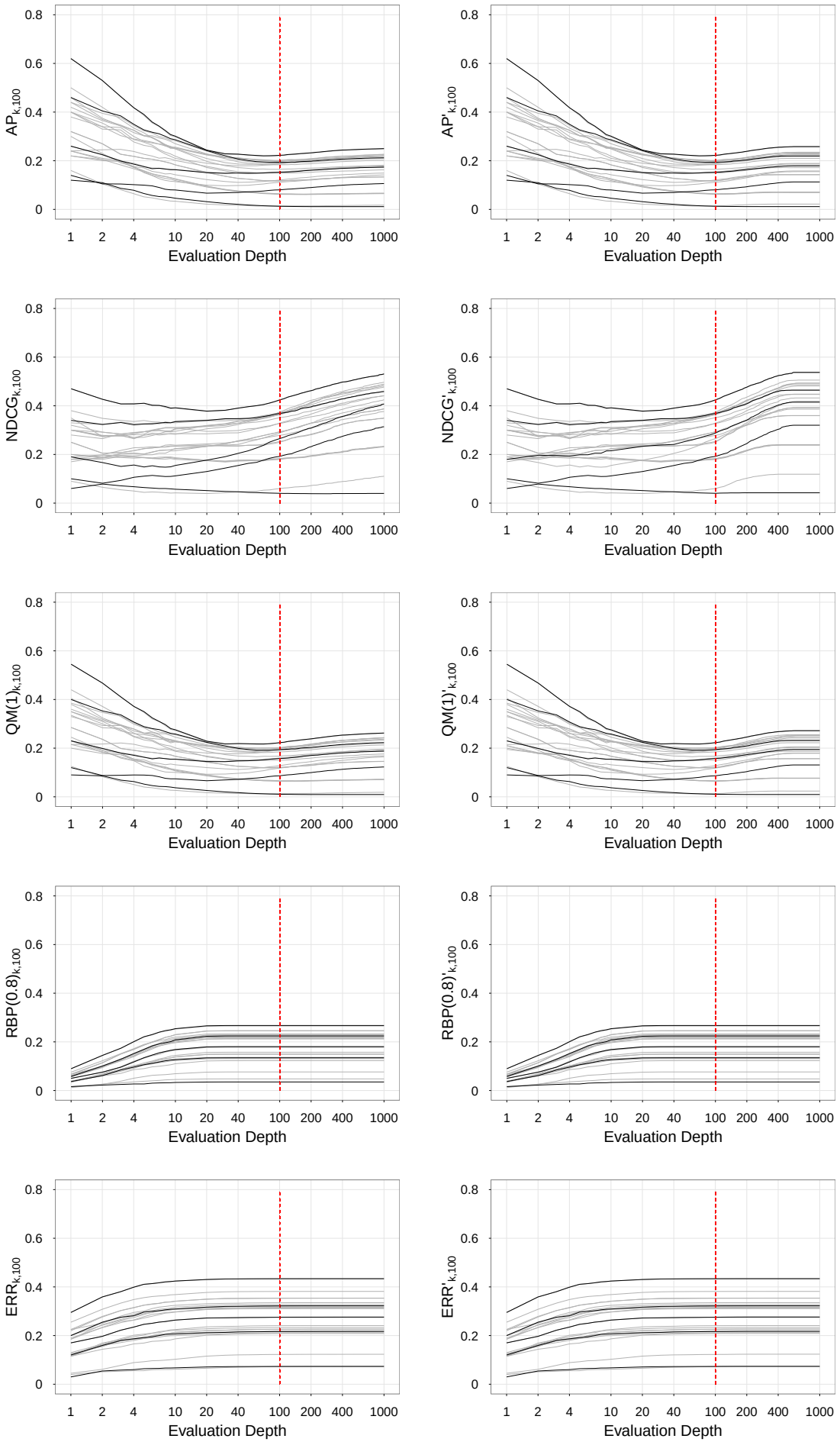

Figure 2: TREC9 systems, using $\mathrm{AP}_{k,100}$, $\mathrm{NDCG}_{k,100}$, $\mathrm{QM}(1)_{k,100}$, $\mathrm{RBP}(0.8)_{k,100}$, and $\mathrm{ERR}_{k,100}$.

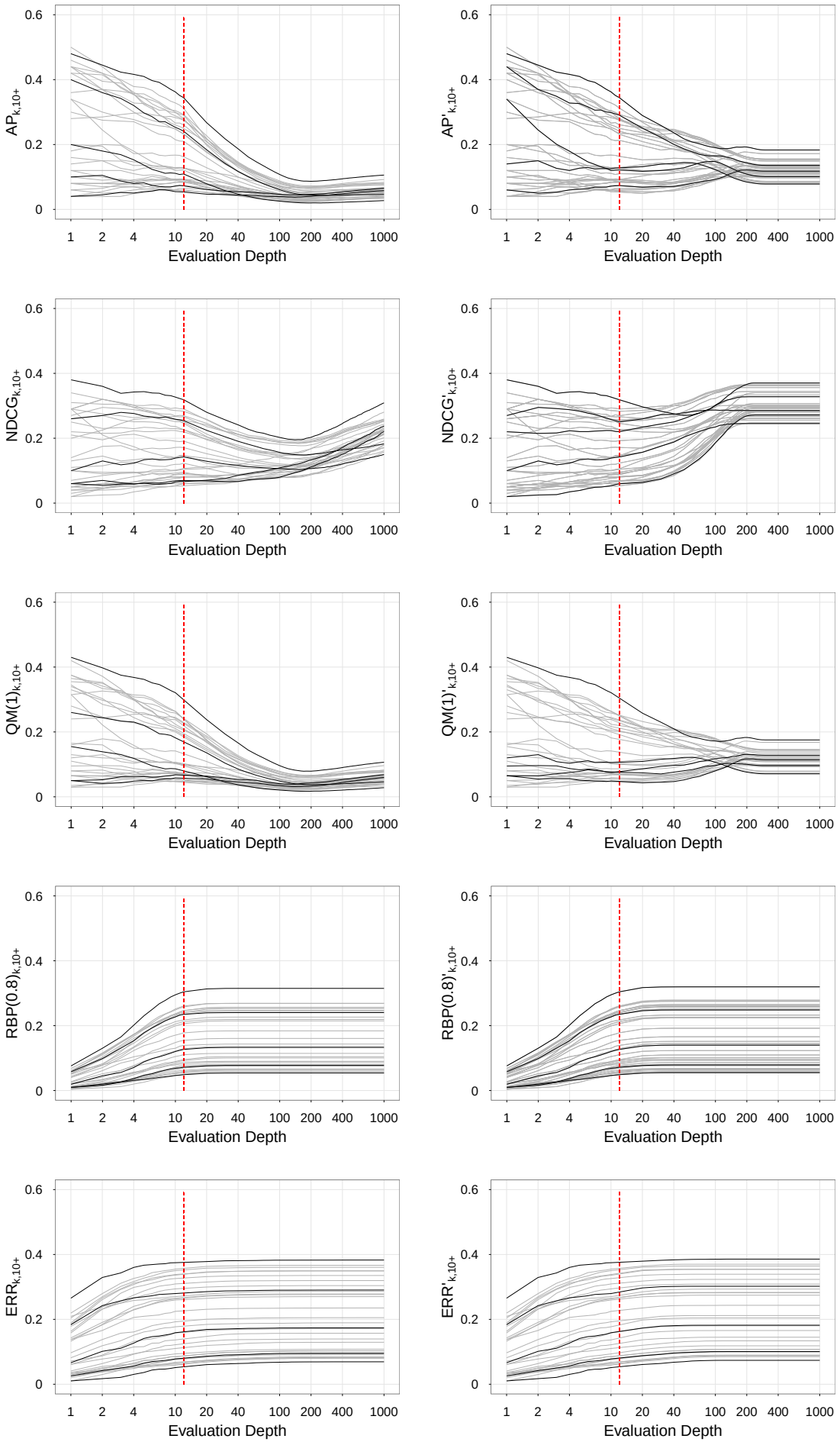

Figure 3: ClueWeb09 Ad-Hoc systems, using $\mathrm{AP}_{k,10+}$, $\mathrm{NDCG}_{k,10+}$, $\mathrm{QM}(1)_{k,10+}$, $\mathrm{RBP}(0.8)_{k,10+}$, and $\mathrm{ERR}_{k,10+}$.

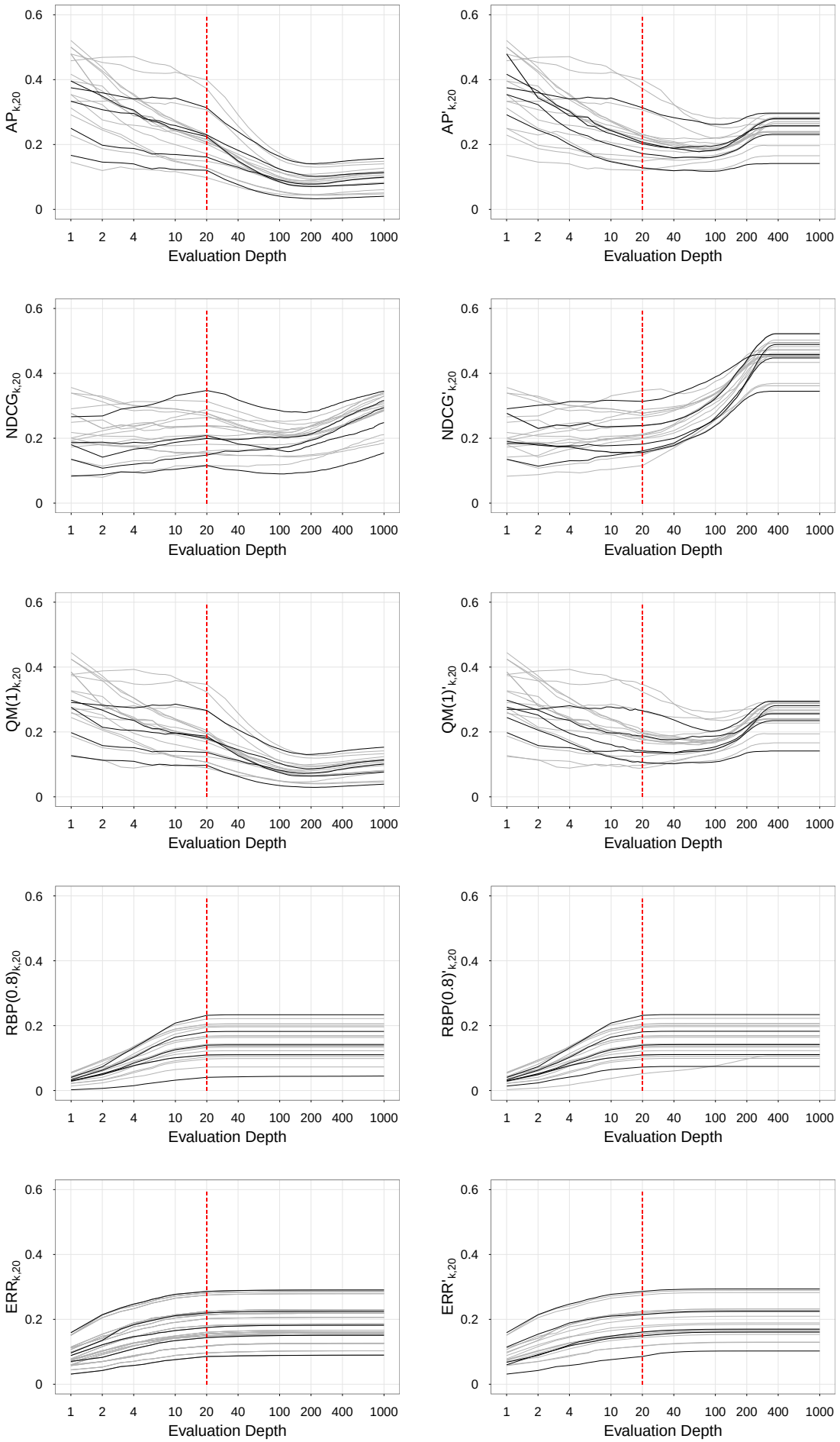

Figure 4: ClueWeb10 Ad-Hoc systems, using $\mathrm{AP}_{k,20}$, $\mathrm{NDCG}_{k,20}$, $\mathrm{QM}(1)_{k,20}$, $\mathrm{RBP}(0.8)_{k,20}$, and $\mathrm{ERR}_{k,20}$.

dashed line, mean that the evaluation depth k can play a crucial role – albeit an unpredictable and inadvertent one – in an experimental outcome.

Recall-based versus utility-based All four figures, spanning four document collections and two types of data, show the convergent behavior of the two utility-based metrics. They benefit from their relatively high top-weightedness, and so are less affected by the shallow pool depth associated with the two ClueWeb collections. Scores for these two metrics (with the chosen parameters) are largely stable by the time the pooling depth has been reached. It is also helpful that they are monotonic; and the result is that there is a much better sense of the final at "at $k = 100$" or even "at $k = 1{,}000$" scores being highly correlated with the "at $k = 10$ scores. Given that the "at $k = 10$" scores are inside the pooling depth, and hence not subject to the vagaries of unjudged documents, this is an important attribute for a metric to have.

Condensed runs The use of condensed runs as a way of dealing with unjudged documents does not alter any of these conclusions. The recall-based metrics still don't achieve stable scores until evaluation depths in the hundreds are reached, and for the two ClueWeb collections (Figures 3 and 4) the induced system orderings are markedly different from those that arise at shallower evaluation depths.

Other AP formulations Figures 1, 2, 3 and 4 use a "relative to opportunity" AP variant in which the normalizing denominator is taken as the minimum of R, the known number of relevant documents, and k, the evaluation depth:

$$\text{AP}_k = \frac{1}{\min\{k, R\}} \sum_{i=1}^{k} r_i \cdot \text{P}(i)\,, \tag{1}$$

where r_i is the relevance of the document at depth i, and $\text{P}(i)$ is the precision at depth i. In this formulation, a perfect ranker attains a score of 1.0 regardless of the relationship between R and k, mirroring the outcomes that NDCG obtains. The drawback of this form of AP is that – as is also the case with NDCG – the system scores can decrease as well as increase with k, as shown in the graphs.

In the more commonly used alternative formulation the score is monotonic with increasing evaluation depth, provided that the pooling depth is held constant:

$$\text{AP}_k{}' = \frac{1}{R} \sum_{i=1}^{k} r_i \cdot \text{P}(i)\,. \tag{2}$$

In this "standard" formulation for AP, scores of 1.0 cannot be achieved when $k < R$, even if the ranking to depth k contains nothing but relevant documents, and this inhibits the "sense" of the metric in terms of the underlying "relative to opportunity" rationale.

A third variant AP computation arises if the assumption that R is constant is challenged. For any non-trivial collection the true R cannot be determined, and all that can be established is a lower bound for it. When k and d are both large and the collection not large, a reasonable approximation might be arrived at [13]. But when d and k are small, or the collection is large, the situation is different. A third version of AP can be defined for these situations, with R_d is used to denote the value of R that emerges after pooling is carried out to depth d, so that $R_0 = 0$, $R_{d+1} \geq R_d$, and $\lim_{d\to\infty} R_d = R$:

$$\text{AP}_k{}'' = \frac{1}{R_k} \sum_{i=1}^{k} r_i \cdot \text{P}(i)\,. \tag{3}$$

Figure 5 shows the behavior of these two variants when applied to the contributed runs across the four collections. In the left column of graphs, the curves do not decrease at all, but the same density of cross-overs arises; moreover, the metrics' score are compacted in to a tighter range. Nor does the normalization regime in the right column of Figure 5 address the issues already noted in connection with AP and the other recall-based metrics. It appears that normalization itself is the issue that is creating the inconsistent outcomes, not the details of how the normalization is being carried out.

4. CONCLUSION

We report here only on preliminary measurements. But even based on these early outcomes, it seems that there are marked differences between the NewsWire and ClueWeb data sets that are more than merely a function of different pooling and judgment depths; and marked differences between the utility-based metrics such as RBP and ERR, and the recall-based ones such as NDCG and AP. The latter appear to be rather more volatile than are the utility-based methods, a difference that is not rectified by the use of condensed runs. We have now commenced a range of more detailed evaluations. These include quantifying the extent to which system orderings shift in response to the addition of judgments when measured using Kendall's τ and other correlation measures; and also quantifying discrimination ratios [9] for various combinations of d and k, including measuring the extent to which initial "findings" of statistically significant superiority get overturned as more evidence is supplied [8].

Acknowledgment This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP140101587 and DP140103256). Shane Culpepper is the recipient of an Australian Research Council DECRA Research Fellowship (DE140100275).

References

[1] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. SIGIR*, pages 25–32, 2004.

[2] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, pages 621–630, 2009.

[3] B. He, C. Macdonald, and I. Ounis. Retrieval sensitivity under training using different measures. In *Proc. SIGIR*, pages 67–74, 2008.

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

[5] A. Moffat. Seven numeric properties of effectiveness metrics. In *Proc. AIRS*, pages 1–12, 2013.

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

[7] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*, pages 659–668, 2013.

[8] S. D. Ravana and A. Moffat. Score estimation, incomplete judgments, and significance testing in IR evaluation. In *Proc. AIRS*, pages 97–109, 2010.

[9] T. Sakai. Alternatives to BPref. In *Proc. SIGIR*, pages 71–78, 2007.

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

[11] M. Sanderson. Test collection based evaluation of information retrieval systems. *Found. Trends in Inf. Ret.*, 4(4):247–375, 2010.

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

[13] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. SIGIR*, pages 307–314, 1998.

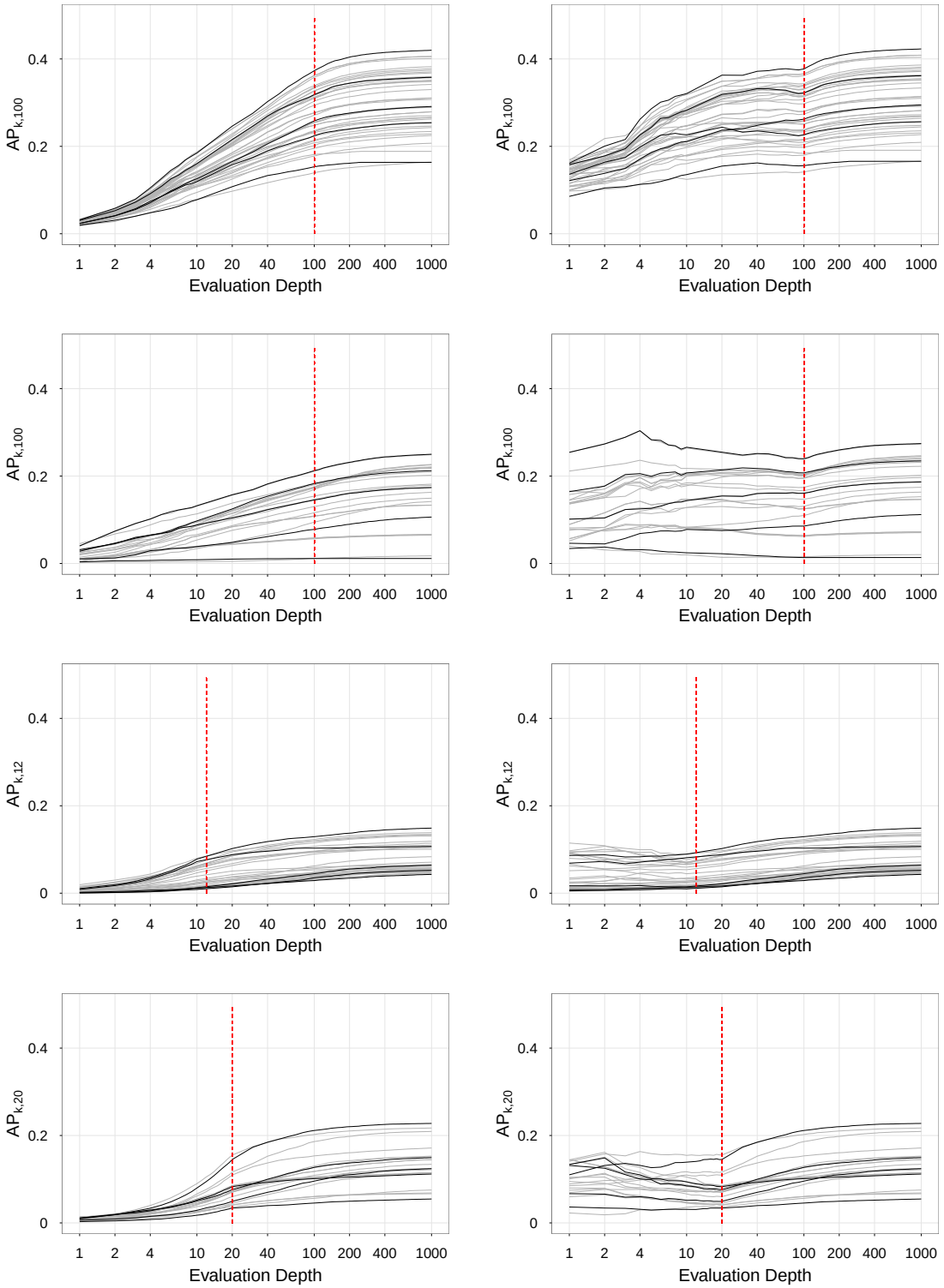

Figure 5: Scores from $AP_k{}'$ (left column) and $AP_k{}''$ (right column) for the four TREC collections listed in Table 1. The curves in each pair of graphs are identical to the right of the red line. The pooling depth for ClueWeb09 is set to $d = 12$ for all topics.