Estimating Average Precision with Incomplete and Imperfect Judgments

Emine Yilmaz, Javed A. Aslam*
College of Computer and Information Science
Northeastern University
360 Huntington Ave, #202 WVH
Boston, MA 02115
{emine, jaa}@ccs.neu.edu

ABSTRACT

We consider the problem of evaluating retrieval systems using incomplete judgment information. Buckley and Voorhees recently demonstrated that retrieval systems can be efficiently and effectively evaluated using incomplete judgments via the bpref measure [6]. When relevance judgments are complete, the value of bpref is an approximation to the value of average precision using complete judgments. However, when relevance judgments are incomplete, the value of bpref deviates from this value, though it continues to *rank* systems in a manner similar to average precision evaluated with a complete judgment set. In this work, we propose three evaluation measures that (1) are approximations to average precision even when the relevance judgments are incomplete and (2) are more robust to incomplete or imperfect relevance judgments than bpref. The proposed estimates of average precision are simple and accurate, and we demonstrate the utility of these estimates using TREC data.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Performance evaluation (efficiency and effectiveness)*

General Terms

Theory, Measurement, Experimentation

Keywords

Evaluation, Sampling, Incomplete Judgments, Average Precision, Bpref

1. INTRODUCTION

We consider the problem of evaluating retrieval systems. The test collection methodology adopted by TREC (the Cranfield paradigm [9]) is one of the most commonly used methodologies in evaluating retrieval systems. In this test collection methodology, the evaluation of retrieval systems typically involves (1) assembling a document collection, (2) creating a set of information needs (topics), and (3) identifying the set of documents relevant to the topics.

There are several simplifying assumptions made by the Cranfield paradigm. One of the main assumptions is that the relevance judgments are complete, i.e., for each topic, all relevant documents in the document collection are identified. When the document collection is large, obtaining complete relevance judgments is problematic due to the need for significant human effort.

In order to avoid judging the entire document collection, TREC uses a technique called *pooling*, where in the case of depth-100 pooling, for example, only the top 100 documents retrieved by the systems are judged and the rest of the documents in the collection are assumed to be nonrelevant. In standard TREC settings, depth-100 pools tend to include many or most of the relevant documents, and depth-100 pooling has been shown to be an effective way to evaluate the relative performance of retrieval systems while avoiding judging the entire collection [12, 17].

Much recent research has attempted to address the assessment effort required for large-scale retrieval evaluation. Soboroff et al. [15] describe a technique for ranking systems without relevance judgments, though it has been argued that without any relevance assessments, such evaluations tend to rank systems by "popularity" rather than "performance" [3]. Cormack et al. [10] and Aslam et al. [2] describe greedy algorithms for creating pools likely to contain large numbers of relevant documents, but evaluations using such pools tend to produce biased or unprovable estimates of standard retrieval measures, especially when the total number of judgments is limited.

In this paper, we consider the evaluation of retrieval systems using incomplete relevance information. When the document collection is dynamic, as in the case of web retrieval, documents are added to the collection over time. Hence, the relevance judgments become *incomplete*, and the judged relevant documents become a smaller random subset of the entire relevant document set. Another problem associated with dynamic collections is that as time passes, some documents may be deleted from the collection; one example of this is broken links in the case of the web. Even

though these documents are no longer in the collection, the relevance judgments will still include these documents. Such relevance judgment sets are said to be *imperfect* [14, 6].

When the document collection is large, there may be many relevant documents in the collection, and the depth-100 pooling method may not be able to identify all relevant documents. Also, in the case of large collections, identifying and judging all relevant documents becomes very expensive. Hence, the relevance judgments for large collections are also usually incomplete. Therefore, in the case of dynamic or large document collections, if current pooling methodologies are used to identify the relevant documents in a collection, an evaluation measure that is both robust with respect to *incomplete* and *imperfect* relevance information and that is also highly correlated with standard measures of retrieval performance is desirable. Such a measure would enable the safe usage of dynamic or larger document collections using the current evaluation methodology.

Recently, Buckley and Voorhees showed that average precision and other current evaluation measures are not robust to incomplete relevance judgments, and they proposed a new measure for efficiently and effectively evaluating retrieval systems [6]. When complete relevance judgments are available, this new measure, *bpref*, is shown to rank systems in a manner similar to average precision. Furthermore, bpref is shown to be relatively stable even when the relevance judgments are highly incomplete or imperfect. Thus, bpref holds promise for the efficient and effective evaluation (ranking) of retrieval systems using large or dynamic document collections.

After bpref was proposed, it became a standard evaluation measure in `trec_eval` (the evaluation program used in TREC), and it has been commonly used in places such as the Terabyte track [8] and the HARD track [1]. When complete relevance judgments are present, the value of bpref is close to average precision. However, when the judgment set is incomplete, the value of bpref deviates from the value of average precision, both in absolute terms and in its ability to rank systems via the evaluation returned.

Average precision is one of the most commonly used and cited system-oriented measures of retrieval effectiveness. It is known to be a stable [5] and highly informative measure [4]. If average precision is considered the "gold standard" for evaluating retrieval effectiveness, an evaluation measure that is both highly correlated with average precision and also robust to incomplete and imperfect relevance judgments is desired.

In this paper, we propose three new evaluation measures that are both robust to incomplete and imperfect relevance information and are also estimates of average precision itself. In order of complexity and accuracy, these measures are (1) *Induced Average Precision*, (2) *Subcollection Average Precision*, and (3) *Inferred Average Precision*. Induced average precision only considers the documents present in the judged subsample, and computes the average precision of the induced list once the unjudged documents are removed. Subcollection average precision also considers those documents present in the judged subsample; additionally, it considers a random subsample of the *non-pool* documents, sampled at the same rate as the pool documents. Thus, a random subsample of the entire collection is effectively considered. Inferred average precision estimates the full collection average precision from the pool subsample directly.

Inferred average precision has the nice property that it is based on defining average precision as the outcome of a random experiment. A derivation of average precision as the expectation of this random experiment is shown, and it is further shown how to estimate this expectation using the random pool subsample.

We show through the use of TREC data that when relevance judgments are incomplete or imperfect, these measures provide closer estimates of average precision using the complete judgment set, both in absolute and ranking terms, than bpref.

In the sections that follow, we describe bpref and induced, subcollection, and inferred average precision in detail, and we describe and discuss our extensive experiments with the TREC collections.

2. EVALUATION WITH INCOMPLETE JUDGMENTS

In this section, we describe bpref and our three proposed measures, and we discuss the results of experiments with these measures using the TREC data collections. We begin by discussing our experimental design using TREC data, and we then discuss each measure in turn together with experimental results.

2.1 Experimental Setup

We use TREC data to test how the proposed measures perform when the relevance judgments are incomplete; for example, if documents are added to a collection over time, the initial (effectively complete) judged set may be modeled as a random subset of the "new" collection. To imitate this effect of incomplete relevance judgments, we use a sampling strategy effectively identical[1] to one proposed by Buckley and Voorhees [6]. For each TREC, we form incomplete judgment sets by randomly sampling from the entire depth-100 pool over all submitted runs.[2] This is done by selecting $p\%$ of the complete judgment set uniformly at random for each topic, where $p \in \{1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100\}$. Note that especially for small sampling percentages, the random sample may not contain any relevant documents. In this case, we remove the entire random sample and pick another $p\%$ random sample until a random sample with at least one relevant document is obtained.

In order to evaluate the proposed measures obtained using incomplete relevance judgments and to compare them with average precision using the entire judgment set (which we refer to as *actual AP*), we use three different statistics: Kendall's τ, linear correlation coefficient ρ, and root mean squared (RMS) error.

Kendall's τ evaluates how the ranking of systems using the estimated values compare to the ranking obtained from the actual values. It is related to the minimum number of pairwise adjacent interchanges needed to convert one ranking into another. Kendall's τ ranges from -1 (perfectly anti-correlated) to $+1$ (perfectly correlated); a τ value of

---

[1]Buckley and Voorhees employ *stratified random sampling* while we employ standard random sampling; these are identical in expectation. Other minor differences exist as well; see Buckley and Voorhees [6] for an exact description of their sampling method.

[2]Note that we consider all *submitted* runs rather than all *pooled* runs, these two sets may be different for some TRECs.

zero corresponds to no correlation. Pairs of rankings whose Kendall's τ values are at or above 0.9 are often considered effectively equivalent [16]. The *linear correlation coefficient* ρ evaluates how well the actual and estimated values fit to a straight line. As with Kendall's τ, the linearly correlation coefficient ρ ranges from -1 to $+1$ with similar interpretations. *RMS error* is related to the standard deviation of the estimation error. Let $(a_1, a_2, ...a_N)$ be actual values and $(e_1, e_2, ..., e_N)$ be estimates of these values. Then the RMS error of the estimation is computed as

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (e_i - a_i)^2}.$$

RMS error is measured in the same units as the underlying data (in our case, units of "average precision").

In our experiments, we use data from TRECs 7, 8 and 10. We report detailed results for TREC8 and overall results from TRECs 7 and 10, due to space constraints.

2.2 Bpref

Buckley and Voorhees [6] show that commonly used evaluation measures such as average precision, R-precision and precision-at-cutoff 10 are not robust to incomplete relevance judgments. They propose another measure named *bpref* which is more robust to incomplete relevance judgments.

Given a query with R relevant documents, the bpref value of a ranked list is calculated as follows

$$\text{bpref} = \frac{1}{R} \sum_r \left(1 - \frac{\text{number of } n \text{ above } r}{R} \right) \qquad (1)$$

where r is a relevant document and n is a nonrelevant document within the first R judged nonrelevant documents in the output of the retrieval system.

However, when the number of relevance judgments is small, a variation of the measure named *bpref-10* is preferred. The bpref-10 measure is calculated as follows

$$\text{bpref-10} = \frac{1}{R} \sum_r \left(1 - \frac{\text{number of } n \text{ above } r}{10 + R} \right)$$

where n is a nonrelevant document within the top $10 + R$ judged nonrelevant documents in the output of the system.

When the relevance judgments are incomplete, Buckley and Voorhees [6] show that the rankings of systems obtained using bpref-10 are more robust and correlated with the rankings of the systems obtained using average precision and the complete judgment set than the rankings of systems obtained using R-precision and precision-at-cutoff 10. However, one potential drawback of bpref (or bpref-10) is that the value of bpref does not have a theoretical basis, as have average precision (an approximation to the area under the precision-recall curve) and R-precision (the break-even point in the precision-recall curve). Buckley and Voorhees [6] show that when the entire judgment set is used, the value of bpref is closely related to the value of average precision. However, as the relevance judgment sets become more and more incomplete, the value of bpref deviates from the value of average precision computed using the entire judgment set. This behavior can be seen in Figure 1. The figure shows the value of mean bpref obtained using 30, 10, and 5% of the entire judgment set versus the value of mean average precision using the entire judgment set. Each plot in the figure

reports the root mean squared (RMS) error, the Kendall's τ correlation, and the linear correlation coefficient ρ. The plots also include the line $y = x$ for purposes of comparison.

Ideally, one might well prefer a measure that is both robust to incomplete judgments and that has longstanding usage, a theoretical basis, and/or exhibits a high correlation to standard measures of retrieval effectiveness. In this paper, we propose three measures based on average precision that are approximations to average precision itself and that are robust to incomplete relevance judgments.

2.3 Induced AP (indAP)

The average precision of the output of a retrieval system is the average of the precisions at each relevant document; the precisions at unretrieved relevant documents are assumed to be zero, by convention. It is known that average precision is an approximation to the area under the precision-recall curve.

Buckley and Voorhees [6] show that as the number of judgments is reduced, the average precision value decreases and the rankings of the systems based on average precision also change. This can be explained by the fact that all unjudged documents are assumed to be nonrelevant by average precision in a typical evaluation setting such as TREC. Therefore, as the number of judgments is reduced, the number of relevant documents retrieved before a relevant document, hence the precision at a relevant document, is also reduced. Therefore, average precision, the average of the precisions at relevant documents, is reduced.

However, one can obtain a different version of average precision, which we call *induced AP (indAP)*, that does not make any assumption about the unjudged documents. For a query with R relevant documents, induced AP can be calculated in exactly the same way as average precision with a slight difference: in induced AP, the documents that are unjudged are removed from the list and are not considered in evaluation.

Once the unjudged documents are removed from the retrieval system's output, induced AP can be calculated in exactly the same way as traditional average precision. Induced AP has the nice property that it is an approximation to the area under the precision-recall curve of the output of a retrieval system when only the judged documents are considered.

Given a query with R judged relevant documents, induced AP can be calculated as

$$\text{indAP} = \frac{1}{R} \sum_r \frac{\text{number of relevants up to } rank(r)}{rank(r)}$$

where r is a relevant document and $rank(r)$ is the rank of a document when only judged documents are considered. Note that the above formula can be written as a preference based measure:

$$\text{indAP} = \frac{1}{R} \sum_r \left(1 - \frac{\text{number of } n \text{ above } r}{rank(r)} \right)$$

where n is a nonrelevant document retrieved above a relevant document when only judged nonrelevant documents are considered.

Note the similarity between induced AP and bpref (Equation 1). In bpref, at each relevant document, the number of nonrelevants above a relevant document is scaled by a factor of $1/R$ (or with $1/(R + 10)$ in the case of bpref-10),
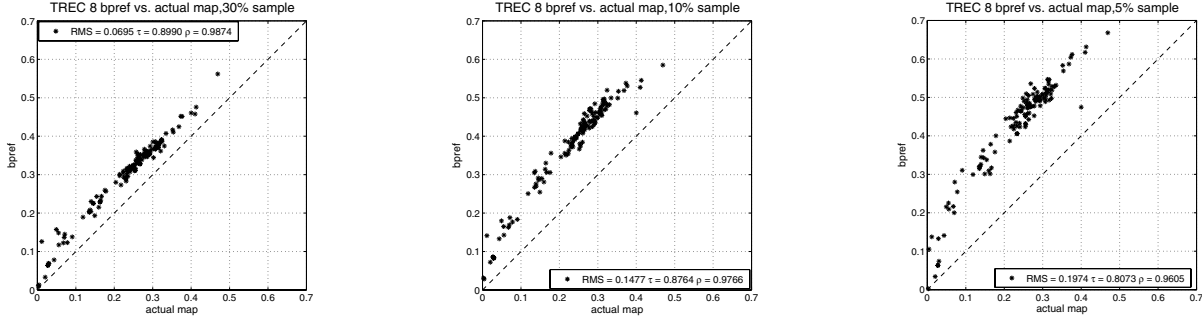
Figure 1: TREC-8 mean bpref-10 as the judgment set is reduced to (from left to right) 30, 10, and 5 percent versus the mean actual AP value (mean AP using the entire judgment set).
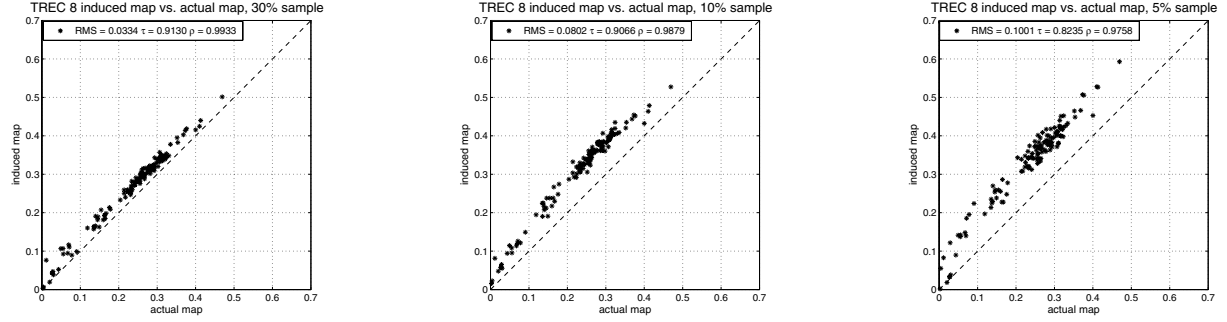

Figure 2: TREC-8 mean induced AP as the judgment set is reduced to (from left to right) 30, 10, and 5 percent versus the mean actual AP.
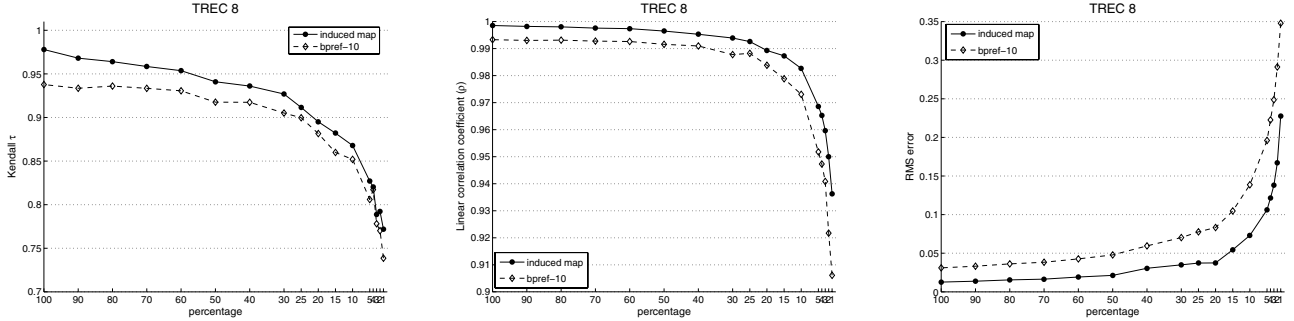

Figure 3: Change in Kendall's τ, linear correlation coefficient (ρ) and RMS errors of mean induced AP and bpref as the judgment sets are reduced, when compared to the mean actual AP.

where R is the total number of judged relevant documents, whereas in the case of induced AP, it is scaled by a factor of $1/rank(r)$, where $rank(r)$ is the rank of the relevant document when only the judged documents are considered. Also, while calculating the number of judged nonrelevant documents above a relevant document, bpref only considers the top R (or $R+10$) judged nonrelevant documents, whereas induced AP considers all such documents. Due to this similarity, bpref can be considered as an approximation to induced AP.

We note that induced AP is an available (though seldom, if ever, used) evaluation measure in TREC's `trec_eval` pro-

gram. However, the robustness of induced AP with respect to incomplete relevance judgments has never been analyzed. In the following experiments, we describe the behavior of induced AP when the relevance judgments are incomplete.

Figure 2 shows how induced map (induced AP averaged over all topics) computed using 30, 10, and 5 percent of the judgments compares with actual map calculated using the complete relevance judgment set. In order to be fair in comparison, we create 10 different runs (samples) for each percentage and pick a sample on which the correlation of induced AP and average precision is close to the average over all 10 samples in terms of RMS error. It can be seen

from this figure that induced AP is an over-approximation to actual AP since when removing the unjudged documents from the depth-100 pool, we also remove the documents that are not in the depth-100 pool, which are mostly nonrelevant. However, when compared to bpref using Figure 1, it can be seen that induced AP is always a better approximation to actual AP as seen by the RMS errors in the plots. Also, induced AP *ranks* systems closer to the ranking obtained from actual AP than bpref does.

In Figure 3, for different percentages of random sampling, we demonstrate the behavior of induced AP in terms of all three statistics. In these experiments, we produced ten different runs (samples) for each sampling percentage and for each retrieval system, we calculated the induced map averaged over all queries. We also calculated the bpref-10 measure in the same way and using the same sample for comparison purposes. Then, we calculated all three statistics for each run and report the average of these three statistics for each percentage. It can be seen from the plot on the left that the ranking obtained by induced AP is very close to the ranking of systems using actual AP and the Kendall's τ value of induced AP is almost always better than that of bpref. The second plot shows that induced AP is highly correlated with actual AP in terms of linear correlation coefficient. The rightmost plot shows via the RMS error that the value of induced map is close to the value of actual mean average precision, even when very few relevance judgments are used.

Note that induced AP is an approximation to average precision, and it is highly robust with respect to incomplete judgment sets. However, one can obtain better estimates of average precision that are still robust with respect to incomplete judgments, as described below.

Note that both induced AP and bpref make use of only the judged documents. However, the unjudged documents also provide some information and one can obtain better estimates of average precision using this additional information. The next two measures, subcollection AP and inferred AP makes use of this extra information provided by the unjudged documents.

2.4 Subcollection AP (subAP)

Hawking and Robertson [13] show that average precision is highly stable with respect to random sampling from the entire document collection. One can use this fact to derive a measure that is also robust with respect to incomplete relevance judgments. In our current setup, we randomly sample from the complete judgment set since this is similar to building larger test collections while the relevance judgments are kept constant. The effect of sampling from the complete judgment set (depth-100 pool) and considering only the judged documents is not exactly the same as sampling documents from the entire document collection. While sampling documents from the entire document collection, the documents that are unjudged, i.e., the documents that are not in the depth-100 pool, are also sampled. It is known that even if the complete relevance judgment set was available, these documents would be considered nonrelevant, since they are not in the depth-100 pool.

One can use this fact and the fact that average precision is stable with respect to subsampling from the entire document collection to derive a new evaluation measure, subcollection AP (subAP), which is a better approximation to

average precision. Note that in computing induced AP, we only considered the judged documents. However, while calculating the average precision estimate for a $p\%$ judgment set, instead of removing all documents that are not in the depth-100 pool together with all documents that were not sampled from the depth-100 pool, one could keep the documents that are not in the depth-100 pool with probability p, randomly and independently for each such document. Note that this has the effect of sampling from the depth-100 pool while also sampling from the non- depth-100 pool, which is equivalent to forming a $p\%$ subcollection from the entire document collection. Since average precision is highly stable with respect to subcollections [13], subcollection AP is also expected to be robust with respect to incomplete relevance judgments.

Note that subcollection AP has the appealing property that when the complete relevance judgment set is available, it is exactly equivalent to average precision. We further note that to employ subcollection AP in practice, one needs knowledge of p; in the case of a dynamic (growing) collection, this is the relative size of the original vs. current collection.

Using the same setup for induced AP and the same randomly generated judgment sets, Figure 4 and Figure 5 show how correlations of subcollection AP with respect to actual AP change as the judgment sets become more incomplete. Figure 4 illustrates that subcollection AP is a better approximation to actual AP than is induced AP. The plots show that subcollection AP is much less biased when compared to bpref and that it is less biased compared to induced AP. One can also see this by comparing the RMS error values in the rightmost plots in Figure 5 and Figure 3 since subcollection AP has lower RMS value. In terms of Kendall's τ and linear correlation coefficient ρ, subcollection AP is also better than induced AP and bpref.

2.5 Inferred AP (infAP)

In order to derive our final robust measure of retrieval effectiveness, we consider the following random experiment whose expectation is average precision. Given a ranked list returned with respect to a given topic:

1. Select a relevant document at random from the collection, and let the rank of this relevant document in the list be i (or ∞ if this relevant document is unretrieved).

2. Select a rank at random from among the set $\{1, \ldots, i\}$.

3. Output the binary relevance of the document at rank i.

In expectation, steps (2) and (3) effectively compute the *precision* at a relevant document, and in combination step (1) effectively computes the *average* of these precisions. One can view average precision as the expectation of this random experiment, and in order to *estimate* average precision, one can instead estimate this expectation using the given sampled relevance judgments.

Consider the first part of this random experiment, picking a relevant document at random from the collection. Since we uniformly sample from the depth-100 pool (which contains all documents assumed to be relevant), the induced distribution over relevant documents is also uniform, as desired. Now consider the expected precision at a relevant document retrieved at rank k. When computing the precision at rank k by picking a document at random at or above k, two cases can happen. With probability $1/k$, we
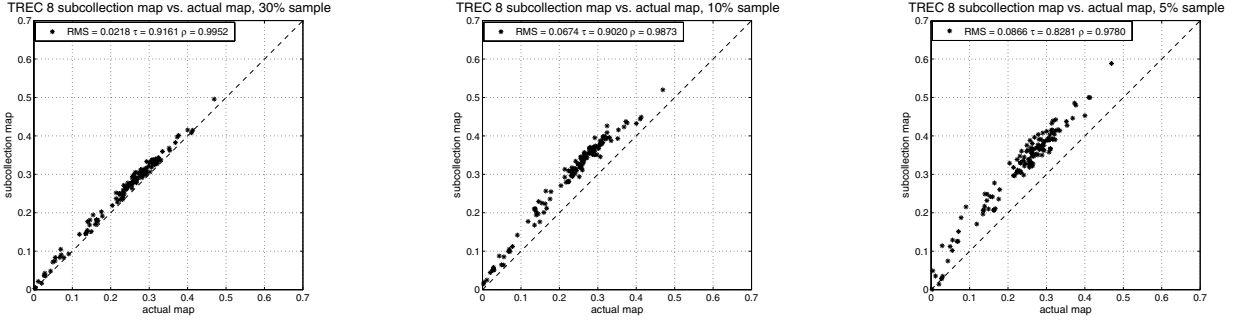
Figure 4: TREC-8 mean subcollection AP as the judgment set is reduced to (from left to right) 30, 10 and 5 percent versus the mean actual AP.
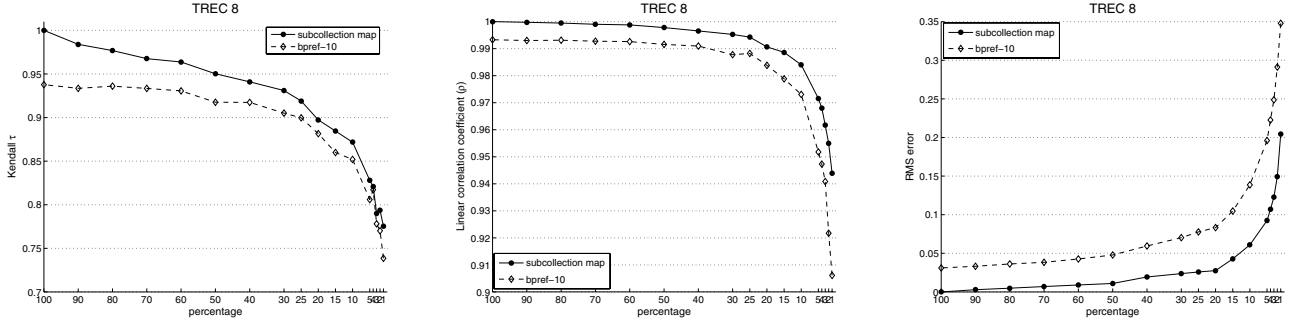

Figure 5: Change in Kendall's τ, linear correlation coefficient (ρ) and RMS errors of mean subcollection AP as the judgment sets are reduced, when compared to the mean actual AP.

may pick the current document, and since this document is known to be relevant, the outcome is 1, by definition. Or we may pick a document above the current document with probability $(k-1)/k$, and we calculate the expected precision (or probability of relevance) within these documents. Thus, for a relevant document at rank k, the expected value of precision at rank k can be calculated as:

$$E[\text{precision at rank } k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} E[\text{precision above } k]$$

Now we need to calculate the expected precision above k. Within the $k-1$ documents above rank k, there are two main types of documents: documents that are not in the depth-100 pool (*non-d100*), which are assumed to be non-relevant, and documents that are within the depth-100 pool (*d100*). For the documents that are within the depth-100 pool, there are documents that are unsampled (unjudged) (*non-sampled*), documents that are sampled (judged) and relevant (*rel*), and documents that are sampled and non-relevant (*nonrel*). While computing the expected precision within these $k-1$ documents, we pick a document at random from these $k-1$ documents and report the relevance of this document. With probability $|non\text{-}d100|/(k-1)$, we pick a document that is not in the depth-100 pool and the expected precision within these documents is 0. With probability $|d100|/(k-1)$, we pick a document that is in the depth-100 pool. Within the documents in the depth-100 pool, we estimate the precision using the sample given. Thus, the expected precision within the documents in the depth-100 pool

is $|rel|/(|rel| + |nonrel|)$. Therefore, the expected precision above rank k can be calculated as:

$$E[\text{precision above } k] =$$
$$\frac{|non\text{-}d100|}{(k-1)} \cdot 0 + \frac{|d100|}{k-1} \cdot \frac{|rel|}{(|rel| + |nonrel|)}$$

Thus, if we combine these two formulae, the expected precision at a relevant document that is retrieved at rank k can be computed as:

$$E[\text{precision at rank } k] =$$
$$\frac{1}{k} \cdot 1 + \frac{(k-1)}{k} \left(\frac{|d100|}{k-1} \cdot \frac{|rel|}{(|rel| + |nonrel|)} \right)$$

Note that it is possible to have no documents sampled above rank k ($|rel| + |nonrel| = 0$). To avoid this $0/0$ condition, we employ *Lidstone smoothing*[7] where a small value ϵ is added to both the number of relevant and number of nonrelevant documents sampled. Then, the above formula becomes:

$$E[\text{precision at rank } k] =$$
$$\frac{1}{k} \cdot 1 + \frac{(k-1)}{k} \left(\frac{|d100|}{k-1} \cdot \frac{|rel| + \epsilon}{(|rel| + |nonrel| + 2\epsilon)} \right)$$

Since average precision is the average of the precisions at each relevant document, we compute the expected precision at each relevant document rank using the above formula and calculate the average of them, where the relevant documents that are not retrieved by the system are assumed to have a
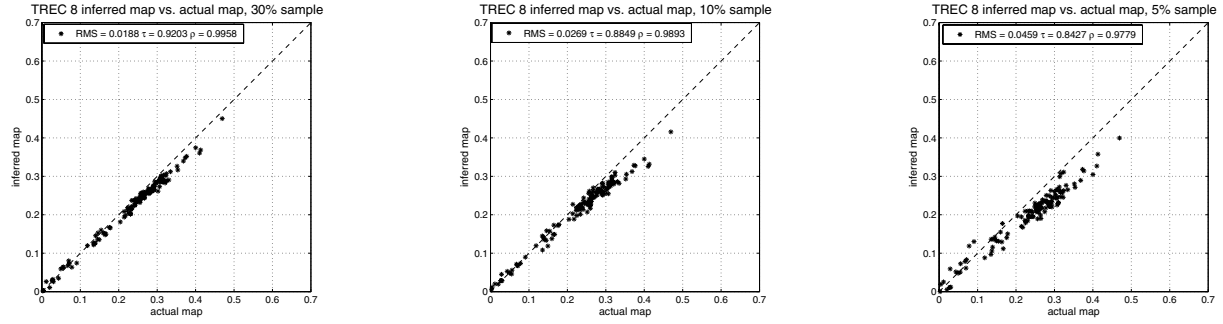
Figure 6: TREC-8 mean inferred AP as the judgment set is reduced to (from left to right) 30, 10, and 5 percent versus the mean actual AP.
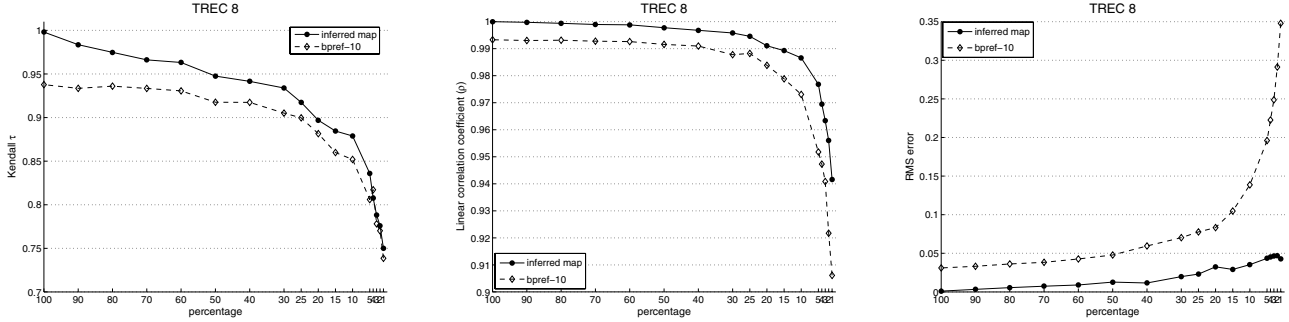

Figure 7: Change in Kendall's τ, linear correlation coefficient (ρ) and RMS errors of mean inferred AP and bpref as the judgment sets are reduced, when compared with the mean actual AP.

precision of zero. We call this new measure that estimates the expected average precision *inferred AP* (infAP).

Note that in order to compute the above formula, we need to know which documents are in the depth 100-pool and which are not. However, the above formula has the advantage that it is a direct estimate of average precision.

Figure 6 shows how the inferred map compares with the actual map. It can be seen that with as few as 5% of the complete relevance judgments, inferred map is a reasonable approximation to actual map. When 30% of the relevance judgments are available, inferred map is a highly accurate approximation to actual map as seen by the RMS error in the plot. Also, based on the RMS error, one can see that inferred map is a better approximation to actual map than all previous measures.

Similarly, Figure 7 shows that the ranking of the systems using inferred map is very close to the ranking obtained using actual map (left plot), that inferred map is highly correlated with actual map (middle plot) and that inferred map is a very close approximation to actual map (rightmost plot). Note that even at very small percentages when the judgment sets are very incomplete inferred map is a very close approximation to actual map.

3. EVALUATION WITH IMPERFECT JUDGMENTS

In the case of dynamic collections such as web, as time passes, some documents may be removed from the collec-

tion (e.g., broken links in the case of web). However, the relevance judgments for these documents are still available, although systems can no longer retrieve these documents. Such relevance judgment sets are referred to as *imperfect* [14]. Buckley and Voorhees [6] show that standard evaluation measures such as precision-at-cutoff, R-precision, and average precision as well as bpref are robust to imperfect relevance judgments.

In this section, we test the behavior of the three proposed measures with respect to imperfect relevance judgments. To imitate the effect of imperfect judgments, we employ an experimental setup similar to that proposed by Buckley and Voorhees [6]. First, we randomly form a $p\%$ sample of the complete document collection, where $p \in \{50, 60, 70, 80, 90, 100\}$. Then, we remove the unsampled documents from the output of retrieval systems and evaluate the three proposed measures using the judgment set for the complete document collection. Finally, we compare the mean values of these measures with the mean actual AP in terms of their ability to properly rank the systems in question. Figure 8 shows the results of this experiment. It can be seen that for all level of imperfectness, all three measures as well as bpref have a Kendall's τ correlation greater than 0.9 when compared to actual mean average precision. However, for all percentages, the proposed measures are more correlated with actual mean average precision than bpref. Hence, the proposed measures are more robust to imperfect relevance judgments in terms of predicting the rankings of systems by actual average precision than bpref, with subcol-

lection and inferred AP being more robust than the induced AP.

4. DISCUSSION AND SUMMARY

We propose three different evaluation measures induced AP, subcollection AP and inferred AP for evaluation using incomplete and imperfect relevance judgments. The proposed measures are estimates of average precision and are robust to both incomplete and imperfect relevance judgments. In this section, we compare the proposed measures with respect to incomplete relevance judgments and summarize.

Figure 9 shows the comparison of the proposed three evaluation measures and bpref in terms of Kendall's τ (first row), linear correlation coefficient (second row), RMS error (third row) and the change in average value over all topics (over all runs and all topics, i.e., a single value per TREC) for TRECs 7, 8 and 10 when the relevance judgments are incomplete. Note that it is important to have stability for the per topic averages since different topics may have different level of incompleteness. In order for the mean values of the measures to have valid meanings, the measures should have similar values even when there are different levels of incompleteness. The plots in the last row of the figure are included to show how the proposed measures perform with respect to different levels of incompleteness in terms of average value. Note that the RMS error and per topic averages have a similar trend as expected.

Among the three measures proposed, induced AP is the simplest one. It uses the same underlying data and information as bpref, and yet it is more robust to incomplete relevance judgments. It can be seen in the plots in the first and second row of Figure 9 that induced AP has a better Kendall's τ and linear correlation with actual AP than has bpref. Induced AP is also a better approximation to actual AP than is bpref. This can be seen in the plots in the third row of the same figure. For all percentages, induced AP has a lower RMS error than bpref. However, when the complete judgment set is available, the value of induced AP is slightly different than actual AP since induced AP does not make any assumptions about the unjudged documents. The plots in the last row show that the value of induced AP increases when the collection becomes highly incomplete and has a similar trend to bpref.

Subcollection AP is slightly more complicated than induced AP. Compared to bpref and induced AP, subcollection AP requires the additional knowledge of the documents that are in the depth-100 pool (even if they are not judged) and the percentage p of the depth-100 pool that is judged. However, subcollection AP is more robust to incomplete relevance judgments than both induced AP and bpref (first and second rows of plots in Figure 9), and it is a better approximation to actual AP (third row plots in Figure 9). Subcollection AP is exactly equivalent to actual AP when complete relevance judgments are present. The plots in the last row show that the value of subcollection AP increases when the collection becomes highly incomplete and has a similar trend to bpref.

Out of the three measures proposed, inferred AP is the most complex, yet it is the closest approximation to actual average precision. In order to compute inferred AP, one needs the additional knowledge of the documents in the depth-100 pool. However, it does not require knowledge of the percentage p of the depth-100 pool that is judged as

subcollection AP does. In terms of Kendall's τ and linear correlation coefficient ρ, inferred AP has a similar performance to subcollection AP and is better than induced AP and bpref (plots in the first and second rows of Figure 9). Inferred AP consistently has much less RMS error compared to all the rest of the measures for all levels of incompleteness. Even when the number of judgments is very small (even as small as 1%), inferred AP has an RMS error less than or equal to 0.05, which means that inferred AP is a very close approximation to actual average precision. As was the case with subcollection AP, inferred AP is exactly equivalent to actual AP when complete relevance judgments are present, except for a slight difference due to the effect of smoothing. The value of inferred AP is also highly stable with respect to different levels of incompleteness and does not much vary even if the collection becomes highly incomplete.

Considering these differences among the proposed three measures, one might prefer to use induced AP if the aim is to have a simple measure that is both an approximation to actual AP and is also robust with respect to incomplete relevance judgments. If one is looking for a simple but better approximations to actual AP, and one knows (or can estimate) the proportion of the size of incomplete judgment set as compared to the complete judgment set, then subcollection AP might be preferred. If the aim is a measure that is both robust and is a very close approximation to average precision, then inferred AP might be preferred.

5. CONCLUSIONS

When document collections are large or dynamic, it is more difficult to evaluate the retrieval systems since obtaining complete relevance judgments becomes more and more difficult. Therefore, evaluation measures that are robust to incomplete and imperfect relevance judgments are needed. Buckley and Voorhees [6] show that most commonly used evaluation measures such as average precision, R-precision and precision-at-cutoff k are not robust to incomplete relevance judgments, and they propose another measure, *bpref*, which is more robust to incomplete relevance judgments. After bpref was proposed, it became a commonly used evaluation measure in TREC, such as in the Terabyte and Hard tracks.

In this paper, we instead propose three different evaluation measures, namely *induced AP*, *subcollection AP*, and *inferred AP*. When compared to bpref, we show that all of these measures are more robust to both incomplete and imperfect relevance judgments than bpref in terms of both predicting the value of average precision and the rankings of systems obtained by average precision. Apart from this, these measures have the nice property that they are different approximations to average precision itself when computed using the entire judgment set (actual AP). Furthermore, when complete judgments are available, the measures subcollection AP and inferred AP are exactly equivalent to actual AP. Finally, we describe the cases where each of one of these evaluation measures might well be preferred over one another.

6. REFERENCES

[1] J. Allan. HARD track overview in TREC 2004: High accuracy retrieval from documents. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2004.
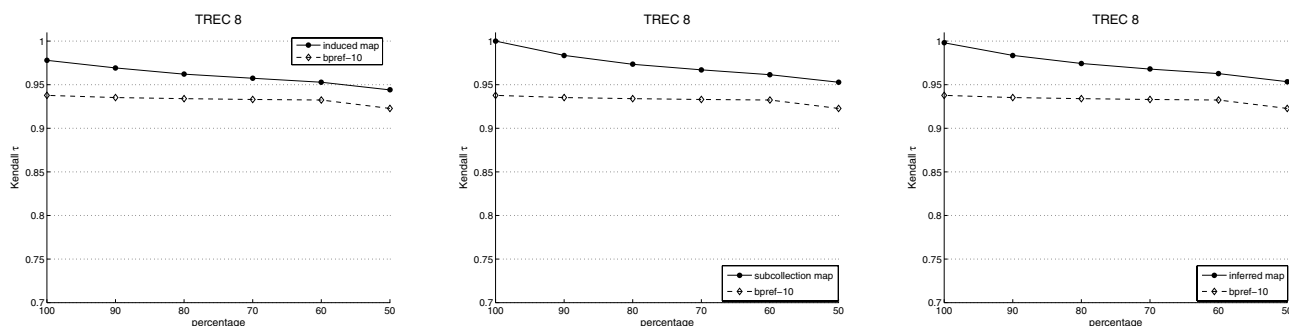
Figure 8: Change in Kendall's τ correlation of mean induced AP, mean subcollection AP and mean inferred AP with mean actual AP as the document collection is reduced (judgments become more imperfect).

[2] J. A. Aslam, V. Pavlu, and R. Savell. A unified model for metasearch, pooling, and system evaluation. In O. Frieder, J. Hammer, S. Quershi, and L. Seligman, editors, *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 484–491. ACM Press, November 2003.

[3] J. A. Aslam and R. Savell. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In J. Callan, G. Cormack, C. Clarke, D. Hawking, and A. Smeaton, editors, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 361–362. ACM Press, July 2003.

[4] J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, August 2005.

[5] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40. ACM Press, 2000.

[6] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, New York, NY, USA, 2004. ACM Press.

[7] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, 1996. Morgan Kaufmann Publishers.

[8] C. L. A. Clarke, F. Scholer, and I. Soboroff. The TREC 2005 terabyte track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[9] C. Cleverdon. The cranfield tests on index language devices. In *Readings in Information Retrieval*, pages 47–59. Morgan Kaufmann, 1997.

[10] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In Croft et al. [11], pages 282–289.

[11] W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors. *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, Aug. 1998. ACM Press, New York.

[12] D. Harman. Overview of the third text REtreival conference (TREC-3). In D. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 1–19, Gaithersburg, MD, USA, Apr. 1995. U.S. Government Printing Office, Washington D.C.

[13] D. Hawking and S. Robertson. On collection size and retrieval effectiveness. *Information Retrieval*, 6(1):99–105, 2003.

[14] R. Nuray and F. Can. Automatic ranking of retrieval systems in imperfect environments. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 379–380. ACM Press, 2003.

[15] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 66–73, New Orleans, Louisiana, USA, Sept. 2001. ACM Press, New York.

[16] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82. ACM Press, 2001.

[17] J. Zobel. How reliable are the results of large-scale retrieval experiments? In Croft et al. [11], pages 307–314.
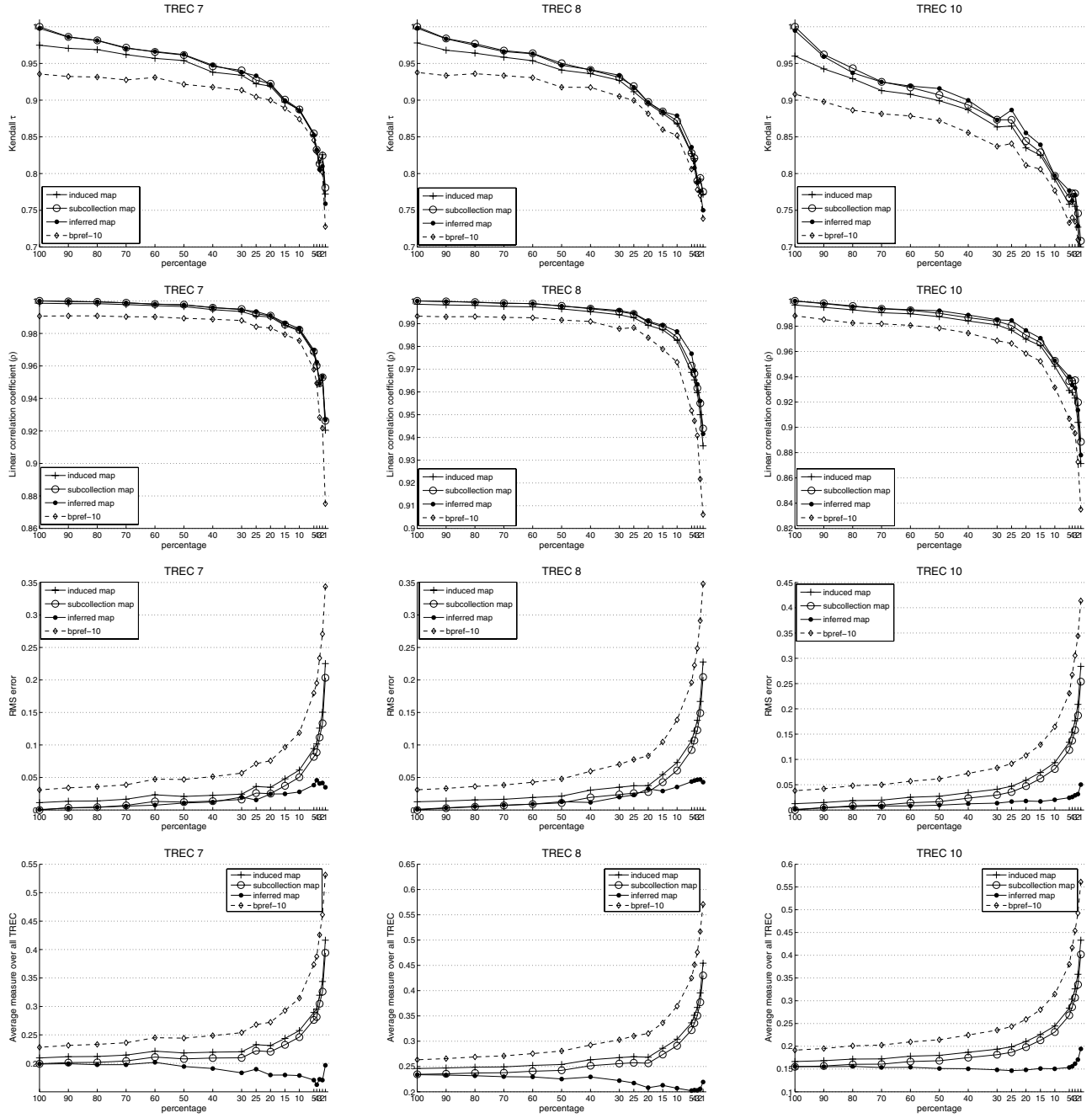
Figure 9: Comparison of induced map, subcollection map, inferred map, and bpref using Kendall's τ (first row), linear correlation coefficient (second row), RMS error (third row) and per topic averages (fourth row) for TREC 7, 8 and 10.