Inferred AP : Estimating Average Precision with Incomplete 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

1. INTRODUCTION

In this work, we consider the evaluation of retrieval systems using incomplete relevance information. When the document collection is dynamic, as in the case of web retrieval, new 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. Also, in the case of large collections, identifying and judging all relevant documents becomes very expensive and the relevance judgments for large collections are also usually incomplete.

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 [3]. 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.

Average precision is one of the most commonly used and cited system-oriented measures of retrieval effectiveness. It is known to be a stable [2] and highly informative measure [1]. If average precision is considered as 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 work, we describe a new evaluation measure that is both robust to incomplete relevance information and is also an estimate of average precision itself. Inferred average precision ($infAP$) 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. In this work, we show the derivation of average precision as the expectation of this random experiment, and we further show how to estimate this expectation using the random pool subsample.

Through the use of TREC data, we show that when relevance judgments are incomplete, inferred average precision provides closer estimates of average precision using the complete judgment set, both in absolute and ranking terms, than bpref.

2. INFERRED AVERAGE PRECISION

In order to derive a measure of retrieval effectiveness that is robust to incomplete judgments, 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 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

*We gratefully acknowledge the support provided by NSF grants CCF-0418390 and IIS-0534482.

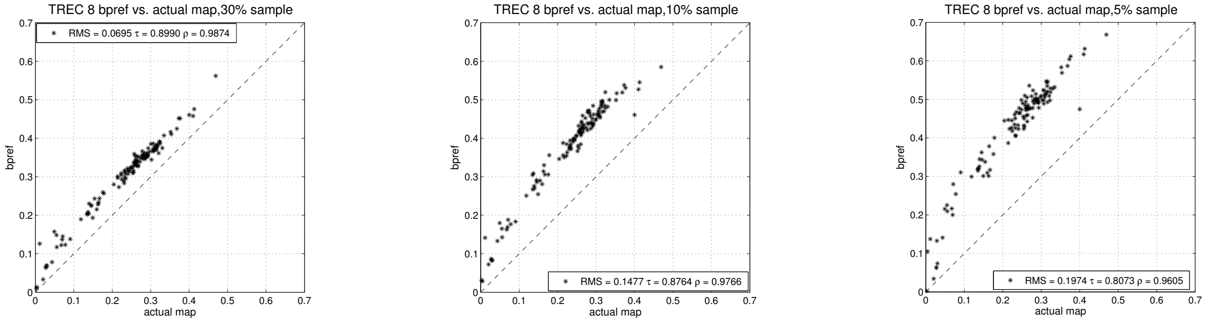

Figure 1: TREC-8 mean bpref-10 as the judgment set is reduced to (from left to right) 30, 10, and 5 percent versus the mean actual AP value (mean AP using the entire judgment set).

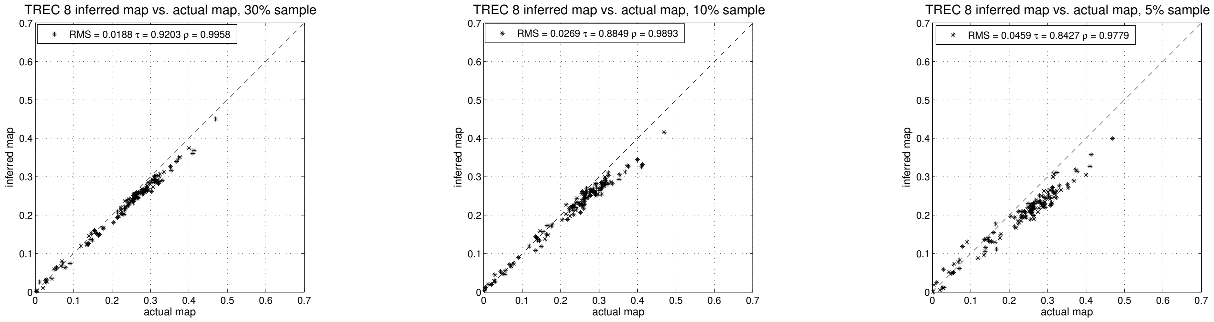

Figure 2: TREC-8 mean inferred AP as the judgment set is reduced to (from left to right) 30, 10, and 5 percent versus the mean actual AP.

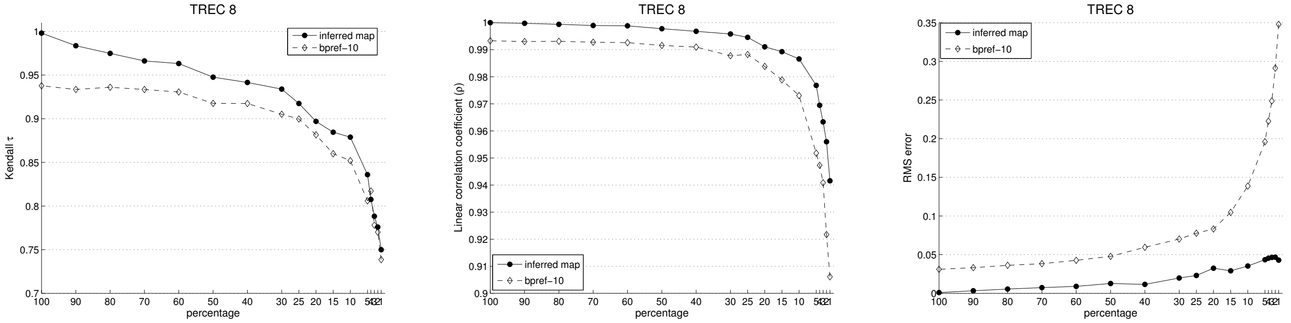

Figure 3: Change in Kendall's τ, linear correlation coefficient (ρ) and RMS errors of mean inferred AP and bpref as the judgment sets are reduced, when compared with the mean actual AP.

depth-100 pool (*non-d100*), which are assumed to be nonrelevant, 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 nonrelevant (*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*[4] 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 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.

We use TREC data to test how inferred AP performs 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 [3]. Using data from TREC-8, 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.

Since bpref [3] is a measure that is commonly used when relevance judgments are incomplete, we compare the performance of inferred AP with bpref. Figure 1 and Figure 2 show how the value of mean bpref-10 (bpref-10 averaged over all queries) and inferred map (inferred AP averaged over all queries) compare with the actual map when the relevance judgments are 30, 10, and 5% of the complete judgment set. It can be seen that 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 (actual AP). However, 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 for all percentages inferred AP is a better approximation to actual map than bpref.

In Figure 3, for different percentages of random sampling, we demonstrate the behavior of inferred AP in terms of three statistics, Kendall's tau, linear correlation coefficient and RMS error. In these experiments, we produced ten different runs (samples) for each sampling percentage and for each retrieval system, we calculated the inferred 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 reported the average of these three statistics for each percentage. It can be seen from the plot on the left that the ranking of systems obtained by inferred AP is very close to the ranking of systems using actual AP and the Kendall's τ value of inferred AP is almost better than that of bpref. The second plot shows that inferred 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 inferred map is close to the value of actual mean average precision, even when very few relevance judgments are used.

3. 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 relevance judgments are needed. Buckley and Voorhees [3] 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.

In this work, we describe a new evaluation measure named *inferred AP*. When compared to bpref, we show that this

[1] Buckley and Voorhess 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 [3] 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.

measure is more robust to incomplete relevance judgments than bpref in terms of both predicting the value of actual average precision and the rankings of systems obtained by actual average precision. Furthermore, inferred AP has the nice property that when complete judgments are available, inferred AP is exactly equivalent to actual AP.

4. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.