Retrieval Evaluation with Incomplete Information

Chris Buckley
Sabir Research, Inc.
Gaithersburg, MD 20878
chrisb@sabir.com

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
ellen.voorhees@nist.gov

ABSTRACT

This paper examines whether the Cranfield evaluation methodology is robust to gross violations of the completeness assumption (i.e., the assumption that all relevant documents within a test collection have been identified and are present in the collection). We show that current evaluation measures are not robust to substantially incomplete relevance judgments. A new measure is introduced that is both highly correlated with existing measures when complete judgments are available and more robust to incomplete judgment sets. This finding suggests that substantially larger or dynamic test collections built using current pooling practices should be viable laboratory tools, despite the fact that the relevance information will be incomplete and imperfect.

Categories and Subject Descriptors

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

General Terms

Measurement,Experimentation

Keywords

Cranfield, incomplete judgments

1. INTRODUCTION

The predominate research methodology for information retrieval system building is the Cranfield paradigm [17]: using test collections to compare the quality of different alternative retrieval methods. A test collection consists of a set of statements of information need (called *topics* in this paper following TREC), a set of documents, and a set of relevance judgments that list which documents should be retrieved for which topics. A basic assumption of the Cranfield paradigm is that the relevance judgments are complete, i.e., that every document is judged for every topic [3]. For small test collections this assumption is literally true. For larger test collections such as the TREC collections that are built through the use of pooling [13], the assumption is not strictly true, but sufficient documents are judged such that completeness is closely approximated. The Cranfield paradigm is robust to this small violation of the completeness assumption in that retrieval systems' effectiveness can be fairly compared using pooled collections [21]. This paper examines how robust the evaluation methodology is to more gross violations of the completeness assumption.

One goal of the paper is to investigate whether unbiased incomplete relevance judgments can be used to reliably compare the relative effectiveness of different retrieval strategies. A positive finding is likely to be a prerequisite to building test collections substantially larger than the current TREC collections. The average TREC collection consists of 50 topics and approximately 800,000 documents; the largest TREC collection with pooled relevance judgments contains approximately 1.7 million web pages. These collections were built using pools between 1000 and 2000 documents per topic drawn from several dozen runs [19]. It is unlikely that assessing effort will be able to be significantly increased above this level, and yet documents sets in some operational settings are now orders of magnitude larger than these test collections. For example, at the time of this writing the Google web search engine provides access to over 3 billion pages [8].

Building test collections for dynamic environments such as the web has the additional challenge that any given set of relevance judgments will eventually include documents that are no longer contained in the collection. Such relevance judgment sets are *imperfect* [10] rather than simply incomplete. A second goal of the paper is to investigate the effect of imperfect judgment sets on the evaluation methodology.

The paper is organized as follows. The next section provides the background needed for the rest of the paper. The following section compares the behavior of the different evaluation measures when complete (pooled) relevance judgments are available. Sections 4 and 5 are the core of the paper. Section 4 analyzes the effect of incomplete judgment sets by comparing system rankings on current TREC collections when the document set is held constant but the number of judgments is progressively reduced. Section 5 analyzes the effect of imperfect judgment sets by repeating the comparisons when the judgment set is held constant but the document set is reduced. The results show that current evaluation measures are not robust to massively incomplete

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.

relevance judgments, but that a new measure is both highly correlated with existing measures when complete judgments are available and more robust to less accurate judgment sets.

2. EXPERIMENTAL CONTEXT

The Cranfield methodology was introduced as a way of performing controlled experiments on retrieval performance [3]. The large repository of retrieval results amassed from Cranfield-based evaluation efforts such as the Text REtrieval Conference (TREC), the Cross-Language Evaluation Forum (CLEF), and the NII-NACSIS Test Collection for IR Systems (NTCIR) meetings has enabled the Cranfield methodology to be refined and validated in recent years [17, 2, 18]. Since acquiring the relevance judgments needed to build a test collection is a time-consuming and expensive process, researchers have looked for ways minimize these costs. Several methods have been proposed for creating judgment pools more efficiently [4, 21], as well as methods for avoiding manual relevance judgments entirely [12, 10]. Caution must be used when adopting these methods, however, because the Cranfield methodology requires unbiased relevance judgments. The methods that avoid manual relevance judgments, for example, systematically evaluate good systems as much worse than corresponding evaluations based on pooled manual judgments.

2.1 Evaluation measures

The question of which measures to use to evaluate retrieval effectiveness has received much attention in the literature. Different evaluation measures have different properties with respect to how closely correlated they are with user satisfaction criteria, how easy they are to interpret, how meaningful average values are, and how much power they have to discriminate among retrieval results. The book by van Rijsbergen [14] contains a summary of the early work on IR evaluation measures, and Appendix A in each of the TREC proceedings describes the measures computed by `trec_eval` [1].

Many of the most frequently used retrieval evaluation measures are derived in some way from *recall* and *precision.* Precision is the proportion of retrieved documents that are relevant, and recall is the proportion of relevant documents that are retrieved. In this paper, we concentrate on precision at 10 documents retrieved (P(10)), R-precision (R-prec), and mean uninterpolated average precision (MAP). In addition, we introduce a measure based on preference relations, which we show is more robust to incomplete relevance judgments than these measures.

Precision at 10 documents retrieved counts the number of relevant documents in the top 10 documents in the ranked list returned for a topic. The measure closely correlates with user satisfaction in tasks such as web searching, and is extremely easy to interpret. However, the measure is not a powerful discriminator among retrieval methods (the only thing that matters is a relevant document entering or leaving the top 10), and averages poorly (the constant cut-off of 10 represents very different recall levels for different topics). Because of these problems, P(10) has a much larger margin of error associated with it than either R-precision or MAP [2].

R-precision is defined as the precision after R documents are retrieved where R is the number of relevant documents for the given topic. This measure addresses the main problems with using precision at a constant cutoff level by evaluating each topic at the level where precision and recall are the same. Tests of the margin of error associated with the measure show that it has a much smaller margin of error than P(10), though a larger error than MAP [2].

MAP is the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved. Geometrically, it is equivalent to the area underneath an uninterpolated recall-precision graph. MAP is based on much more information than either P(10) or R-precision, and is therefore a more powerful and more stable measure [2]. Its main drawback is that it is not easily interpreted. A MAP score of 0.4 can arise in a variety of ways, for example, whereas a P(10) score of 0.4 can only mean that four of the top 10 documents retrieved are relevant.

A preference relation is a set of statements that a user prefers one document to another in the context of a particular topic. Preferences have been suggested as the basis for retrieval system evaluation to avoid forcing assessors to pick a single category for a document, or to provide the foundation for measures that are more sensitive than recall and precision [7, 11]. Assessors can provide preferences by explicitly judging pairs of documents, by ranking a set of documents, or by judging documents on a ratio scale [6]. Frei and Schäuble based their *usefulness* measure on user preference sets [7], while Yao [20] and Mizzaro [9] each defined an evaluation measure in terms of distances between a user's ranking and the system's ranking.

Our motivation for defining a preference-based measure is to find a measure that is robust in the face of incomplete relevance information rather than to exploit a different kind of judgment. The idea is to measure the effectiveness of a system on the basis of judged documents only. Since the scores for R-precision, MAP, and P(10) are completely determined by the ranks of the relevant documents in the result set, these measures make no distinction in pooled collections between documents that are explicitly judged as nonrelevant and documents that are assumed to be nonrelevant because they are unjudged. In contrast, our proposed preference measure is a function of the number of times *judged* nonrelevant documents are retrieved before relevant documents. We call the measure "bpref" because it uses binary relevance judgments[1] to define the preference relation (any relevant document is preferred over any nonrelevant document for a given topic). Note that binary relevance judgments are an efficient way of obtaining a large unbiased set of preferences since $N + R$ judgments define $N \times R$ preferences when N is the number of nonrelevant judgments and R is the number of relevant judgments.

Naive formulations of the bpref measure such as simple counts of the number of judged nonrelevant documents retrieved before some relevant document are poor evaluation measures because the score is dependent on the absolute numbers of relevant and/or judged nonrelevant documents. Thus, the measure averages poorly across topics (confirmed by experimentation across 12 bpref variants). We can explicitly compensate for this dependency on absolute numbers by making the number of nonrelevant documents used in the computation of the score a function of the number of

[1] The WT10g collection used in the experiments described later contains three-way judgments of nonrelevant, relevant, and highly relevant. In this paper, no distinction is made between relevant and highly relevant documents.

relevant documents. For a topic with R relevant documents where r is a relevant document and n is a member of the first R judged nonrelevant documents as retrieved by the system,

$$\text{bpref} = \frac{1}{R}\sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{R}$$

.

Bpref is the theoretically attractive version of the measure that works well most of the time in practice. However, it is excessively coarse when the number of relevant documents is very small (one or two) because the evaluation is then restricted to a very few document pairs. In the experiments later in the paper, we reduce the number of relevant documents in the existing TREC collections such that there is frequently only a single relevant document. For this reason, in this paper we use a variant of the bpref measure called bpref-10 where the evaluation is guaranteed to use at least ten document pairs:

$$\text{bpref-10} = \frac{1}{R}\sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{10 + R}$$

where n is a member of the top $10 + R$ judged nonrelevant documents as retrieved by the system.

2.2 Stability of system rankings

Since the Cranfield methodology is based on comparing the relative effectiveness of different retrieval methods, we need a measure of the effect changes to the methodology have on comparative performance. We introduced an association measure based on Kendall's τ to measure the change in system rankings caused by using different relevance judges [15], and use the same procedure for this work.

Define a *system ranking* as an ordered list of runs sorted by decreasing evaluation measure score. The similarity of two rankings is defined as the Kendall's τ between them. Kendall's τ computes the minimum number of pairwise adjacent swaps to turn one ranking into the other, normalized such that two identical rankings produce a correlation of 1.0 and the expected correlation of two rankings chosen at random is 0.0. While the τ depends on the set of runs being compared, correlations computed over the same set of runs can be meaningfully compared. Previous work has considered all rankings with correlations greater than 0.9 as equivalent and rankings with correlations less than 0.8 as containing noticeable differences in general [16].

2.3 Test collections

We use three TREC test collections and the runs submitted to the corresponding TREC track as the data sets for our analysis. Details about the collections are shown in Table 1. These data sets were selected since they are the largest available test collections with pooled relevance judgments: the TREC-8 collections has the most runs associated with it as well as the most participating groups, the TREC-10 collection has the largest document set, and the TREC-12 collection has the largest topic set.

The TREC-8 and TREC-12 data sets use the same document collection, the set of documents on TREC disks 4 and 5 minus the *Congressional Record*. For the TREC-8 data set, the topic set is TREC topics 401–450 and the run set is the runs submitted to the ad hoc task during TREC-8. For the TREC-12 data set, the topic and run sets are those used in the TREC 2003 Robust track. The Robust track topic set includes 50 topics that had been used in previous TRECs (including some from 401–450) and a set of 50 new topics (numbers 601–650). The TREC-10 data set used the WT10g web page collection as the document set, the set of topics created for the ad hoc task within the TREC 2001 Web track (TREC topics 501–550), and the set of runs submitted to that task. The column label "Ave. Rels" in the table gives the average number of relevant documents per topic for the data set.

Table 1: Data sets used in the analysis.

	Documents		Topics		Runs	
TREC	#	GB	#	Ave. Rels	#	Groups
8	528k	1.9	50	94.6	124	38
10	1700k	10.0	50	67.3	77	26
12	528k	1.9	100	60.7	73	16

The last two columns in Table 1 show the combined number of runs and the number of distinct participating groups that submitted those runs. The retrieval methods used to produce the runs submitted to a TREC task can vary widely in important features such as whether queries were automatically or manually constructed, the amount of information contained within the topic statement used to construct the query, and whether relevance feedback was used. We make no distinctions among these features—all runs within a given data set are compared to all other runs in that data set. However, because the preference-based measure depends on the number of retrieved documents, we did not include runs that retrieved many fewer documents than other runs in the data set. In particular, we did not use runs that retrieved less than 95% of the maximum number of documents that could be retrieved. Since TREC runs may contain a maximum of 1000 retrieved documents per topic, this means we did not use runs that retrieved fewer than 47,500 documents for the TREC-8 and TREC-10 data sets, and runs that retrieved fewer than 95,000 documents for the TREC-12 data set. We also did not use runs that retrieved no documents for some topic. These restrictions eliminated 6 submitted runs from TREC-8, 20 submitted runs from TREC-10, and 5 submitted runs from TREC-12. The number of runs given in the table is the number of runs used in our analysis.

Different test collections have different intrinsic difficulty and different characteristics with respect to how noisy evaluation results using that collection are. One way to measure a collection's noise is to use the size of the difference in evaluation scores (δ) needed to have 95% confidence in the conclusion [18]. Table 2 gives the δ's computed using their procedure for the three test collections and the four evaluation measures used in this work. Since the δ depends on the range of absolute scores a measure obtains, its value in isolation is not meaningful. Therefore, the table also gives the best average score for a measure obtained by a run in the data set and the percentage difference of the best score that the δ represents. The TREC-12 collection requires the smallest relative differences, and thus has the least noise, which is consistent with it having twice as many topics as the other two collections. The TREC-10 collection is a much noisier collection: it requires noticeably larger relative differences than the TREC-8 collection despite the fact that each collection contains 50 topics.

Table 2: Difference in scores (δ) required to have 95% confidence in the conclusion. Also given are the best average score and the percentage difference the δ represents with respect to that maximum.

	TREC-8			TREC-10			TREC-12		
	δ	Best	% diff	δ	Best	% diff	δ	Best	% diff
MAP	0.040	0.413	9.7	0.047	0.283	16.6	0.024	0.311	7.7
P(10)	0.093	0.722	12.9	0.077	0.468	16.5	0.047	0.488	9.6
R-prec	0.039	0.436	8.9	0.055	0.302	18.2	0.027	0.322	8.4
bpref-10	0.041	0.443	9.3	0.051	0.323	15.8	0.025	0.326	7.7

Table 3: Kendall correlations using system rankings produced by MAP vs. other measures using complete judgments.

	TREC-8	TREC-10	TREC-12
MAP 1000	1.000	1.000	1.000
MAP 400	0.991	0.976	0.987
P(10)	0.813	0.729	0.721
R-prec	0.916	0.851	0.899
bpref-10 1000	0.934	0.895	0.942
bpref-10 400	0.934	0.894	0.947

3. USING COMPLETE JUDGMENTS

This section provides the baseline data for how the different measures behave with complete (pooled) relevance judgments. Table 3 gives the Kendall τ correlation between the system ranking produced when using MAP and the ranking produced by using the measure named in the row for the three different data sets and using the full set of relevance judgments. The MAP ranking was selected as the baseline ranking because MAP has been shown to be a stable measure that has good ability to discriminate among retrieval methods [2]. Correlations between system rankings do not directly address the question of measure quality—though measures that correlate poorly with established measures are probably poor measures—but rather indicate whether two measures measure the same thing when averaged over the topics. Note that the difference can come from either measures evaluating different properties of the retrieved set for a single topic, or from measures emphasizing different topics when averaged together.

The MAP and bpref-10 measures each have two rows in the table. In the first row, the measure is computed over the full set of 1000 retrieved documents per topic; in the second row, the measure is computed using only the top 400 retrieved documents. We use the reduced retrieved set size in Section 5 where we investigate the effect of imperfect judgment sets. The P(10) and R-prec measures are unaffected by using only the top 400 retrieved, so only the one figure is reported for those measures.

Table 3 shows a relatively weak correlation between MAP and P(10), and a stronger correlation between MAP and R-prec. This is consistent with previous studies [19], though the correlations between MAP and R-prec are lower here than before. The correlation between MAP and bpref-10 is at or above 0.9, the cut-off we use for essentially equivalent rankings. This shows that with complete judgments bpref-10 and MAP will in general agree as to which retrieval method has the better average effectiveness. The graphs in Figure 1 show that the two measures are also in close agreement on a per-topic basis. Each graph shows line plots of the per-topic MAP 1000 scores and bpref-10 1000 scores averaged over all runs in the data set and sorted by decreasing average MAP score. For each data set, the shapes of the two lines match extremely well indicating agreement on a per-topic basis as well as an average basis. The larger peaks in the bpref-10 lines generally correspond to topics with very few relevant documents.

4. INCOMPLETE JUDGMENT SETS

Our goal is to investigate the behavior of the evaluation measures as judgment sets become less complete. Zobel investigated the effects of varying pool sizes, and hence varying judgment set sizes, in his analysis of the reliability of pooled collections [21]. However, he was concerned with estimating the likely number of relevant documents that did not make it into the pools and the resulting bias against non-contributing systems. We assume the judgment sets are fair, though incomplete, and examine how system rankings change for the different measures.

We use a total of 17 progressively smaller judgment sets (or *qrels*) for each data set. The qrels released with the TREC collection is the largest of the qrels, which we designate as the 100% qrels. For each topic in the 100% qrels, we create a list of the relevant documents in a random order, and a separate list of the judged nonrelevant documents in a random order. We then create 16 additional qrels by taking 90, 80, 70, 60, 50, 40, 30, 25, 20, 15, 10, 5, 4, 3, 2, and 1 percent of the 100% qrels. For a target qrels that is $P\%$ as large as the 100% qrels, we select $X = P \times R$ relevant documents and $Y = P \times N$ nonrelevant documents for each topic where R is the number of relevant documents in the 100% qrels and N is the number of judged nonrelevant documents in the 100% qrels for that topic. We use 1 as the minimum number of relevant documents and 10 as the minimum number of judged nonrelevant documents per topic to include in a qrels. Thus if X or Y is less than the corresponding minimum it is set to the minimum. We add the first X relevant documents from the random list of relevant documents and the first Y judged nonrelevant documents from the random list of nonrelevant documents to the target qrels. Since we take random subsets of a qrels that is assumed to be fair, the reduced qrels are also unbiased with respect to systems. Each of the smaller qrels is a subset of a larger qrels since we always select from the top of the randomized lists.

If the same resources as are currently used to produce pooled collections are maintained to produce larger collections, then the absolute number of judged documents will

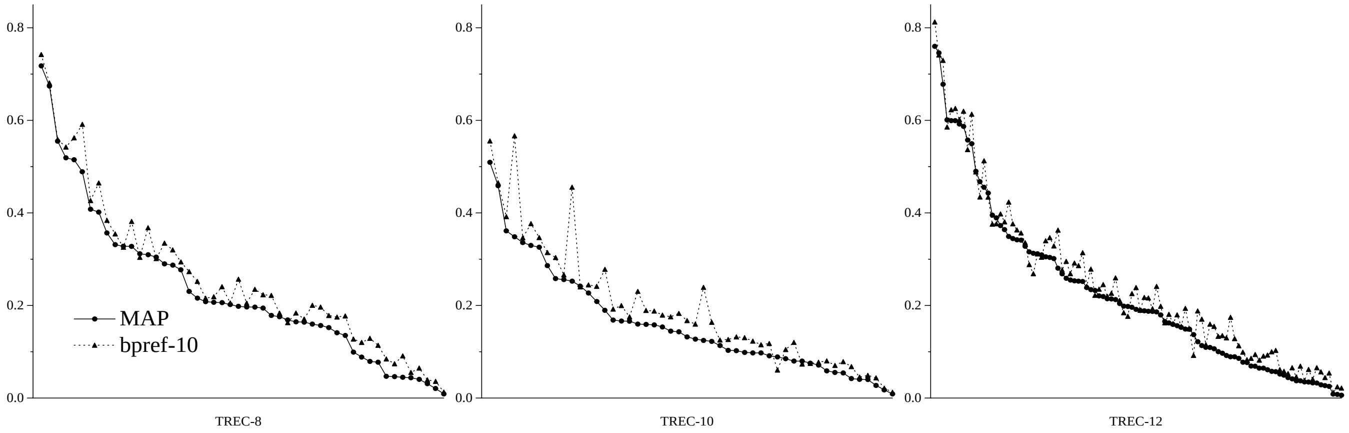

Figure 1: Agreement between MAP and bpref-10 with complete judgments on a per-topic basis. The x-axis plots topics sorted by decreasing average MAP score, and the y-axis plots the average score.

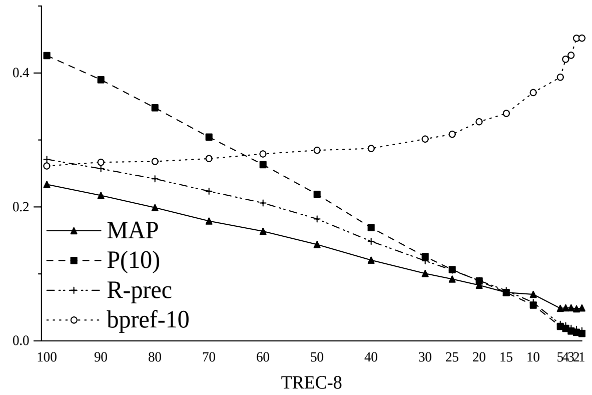

Figure 2: Change in absolute average scores as judgment sets are reduced.

stay roughly the same as current collection's judged documents, though that number will represent a smaller percentage of the total documents. Constructing the qrels as described above reflects the change in percentage of judged documents, but does not accurately reflect the absolute numbers of judged documents which are also reduced. All evaluation measures are known to be unstable for very small numbers of relevant documents [2]. The instability caused by small absolute numbers of relevant documents is likely to dominate any effects caused by incompleteness once the qrels gets small enough. We report results for all 16 reduced qrels below, but caution that the smallest qrels are most likely dominated by absolute number effects. The percentage of topics with only one relevant document in the 5% qrels is 26, 34, and 40 respectively for the TREC-8, TREC-10, and TREC-12 data sets. The corresponding percentages for the 10% qrels are 8, 16, and 18, and for the 1% qrels are 78, 92, and 90.

The graph in Figure 2 shows how the absolute scores of the different evaluation measures change as the level of incompleteness increases. In the graph a separate line is plotted for each of the four measures. The x-axis varies over the different qrels sets, from the 100% qrels to the 1% qrels. The value plotted is the average score for the measure computed over all topics and all runs. The graph shows the data for the TREC-8 collection; the graphs for the other collections are very similar. The measures that depend only on the ranks of the relevant documents (MAP, P(10), R-prec) monotonically decrease as the qrels are reduced, while the bpref-10 measure increases but at a slower rate (except for very small qrels sets). Consistent absolute scores is an important feature for practical application in an incomplete collection. In large collections built through pooling, different topics will have different levels of incompleteness. That is, unlike the qrels constructed here where a 20% qrels means all topics have 80% fewer relevant than the complete qrels, in practice the judgments for some topics will be more complete than others. The fact that bpref-10 has similar absolute scores at different levels of incompleteness means that the average score will be meaningful.

If changes in the absolute scores affect all topics and system equally, the Cranfield methodology will be robust to the increased incompleteness since the relative scores of systems will remain the same. The relative scores are affected by large amounts of incompleteness, though, as shown in Figure 3. The graphs in the figure plot the Kendall τ correlations between the system ranking produced using the 100% qrels and the system ranking produced using the same measure but a reduced qrels. The figure contains one graph for each of the three collections and each graph plots a separate line for each of the four measures. The x-axis shows the size of the reduced qrels and the y-axis the τ value.

The plot for the bpref-10 measure is flatter than the plots for the other measures, indicating that the bpref-10 measure continues to rank different systems in the same relative order as when using complete judgments for higher levels of incompleteness. The Kendall's τ scores for bpref-10 remain above 0.9 until the 50% qrels is reached in the noisiest collection (TREC-10), and remain above 0.9 until the 25% qrels is reached for the TREC-8 collection. The main reason for this stability is the fact that bpref-10 scores depend on the relative ranks of relevant and nonrelevant documents, not the absolute ranks. Adding additional unjudged documents to a collection cannot change a bpref-10 score.

The plot for the P(10) measure has an initial drop and then remains fairly flat, attaining slightly higher τ scores than bpref-10 for the 40% and lower qrels in the TREC-10 collection. Since the P(10) measure depends only on the

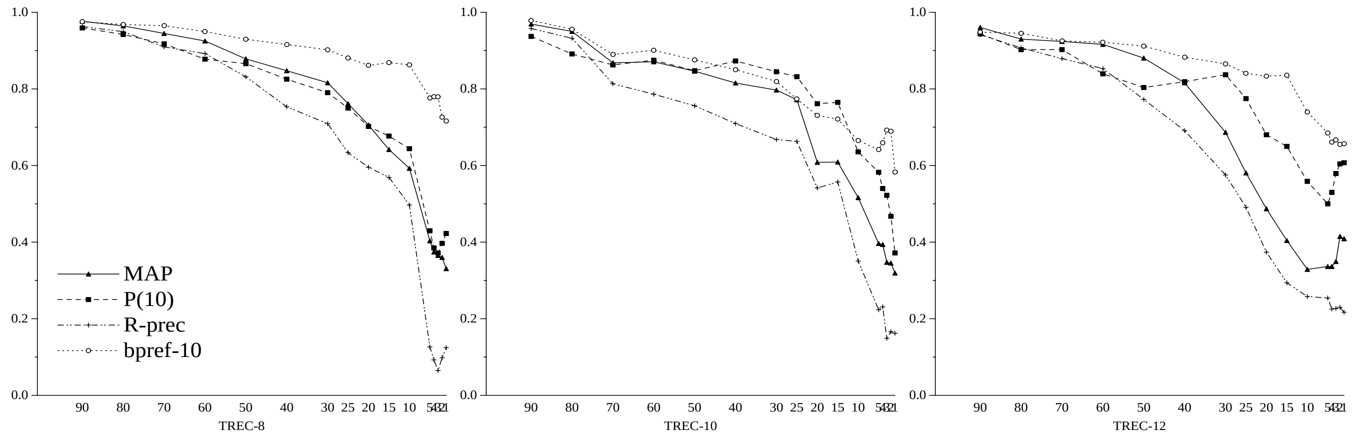

Figure 3: Change in Kendall correlation as judgment sets are reduced. The x-axis plots the percentage of the original judgment set used to evaluate systems, and the y-axis shows the Kendall τ score between the system ranking produced using the original judgment set and the system ranking prduced using the reduced judgement set.

relevant documents retrieved in the first 10 ranks, average P(10) is unstable. The value of the average score is dominated by topics that have many relevant documents (since P(10) is close to 1), but as most systems do well on those topics, the topics contribute relatively little to distinguishing among systems. The smaller correlations between system rankings as the qrels are first reduced reflects this instability. The rankings then stabilize because there is little further information to lose.

The graphs in Figure 4 show how a measure's emphasis on individual topics changes as the incompleteness levels increase. The graphs in the figure plot the Kendall's τ correlations between *topic* rankings produced using the 100% qrels and a reduced qrels. A topic ranking is a list of the topics ordered by decreasing average score where the average is computed over all the runs in the data set. Smaller correlations between topic rankings indicate that different topics receive higher scores, and thus different topics dominate the average score for a run. Once again the plot for the bpref-10 measure is the flattest among the four measures, indicating that bpref-10 emphasizes the same topics at different levels of incompleteness.

5. IMPERFECT JUDGMENT SETS

With incomplete judgments, the qrels do not contain all the relevant documents, but the documents contained within the qrels are guaranteed to be in the document collection. Qrels created for a dynamic collection are not only likely to be incomplete but also to contain judged documents that are no longer in the collection so cannot be retrieved by retrieval systems. Nuray and Can called such judgment sets *imperfect* judgment sets [10].

To investigate the effect of judged documents that are no longer contained in the collection on the Cranfield methodology, we evaluated the runs in the three data sets using the 100% qrels but a reduced document set. (Note that we do not directly address the issue of a document's content being replaced while maintaining its name; the main effect of such a replacement is the same as removing a document and adding a new one with possibly an incorrect, but unbiased, judgment.) To create the reduced document sets, we first created a randomized list of the documents in the full collection, and then selected the first $P\%$ of the documents on the list for $P \in \{90, 80, 70, 60, 50\}$. We eliminated all documents that were not contained in a given reduced document set from the retrieval results before evaluating the runs. When removing a document from the rankings, all documents retrieved at greater ranks than a removed document are in effect moved up one rank. Since a 50% random reduction of the document set means that approximately half the documents in the original rankings will be removed, the evaluation was restricted to the first 400 retrieved documents in all cases. Table 3 shows the effect of limiting the retrieval results to 400 documents retrieved with the full document set. The P(10) and R-prec measures are completely unaffected by the restriction; the MAP and bpref-10 measures are affected, but the correlations among system rankings are at or above 0.9.

The graphs in Figure 5 show that none of the measures is strongly affected by documents that appear in the qrels but are no longer contained in the document collection. The graphs plot the Kendall's τ scores between the system ranking produced using the complete document set and system rankings using the same measure but a reduced document set. τ scores generally dip below 0.9 once half the document set has been removed (though MAP and bpref-10 remain above 0.9 even then for the TREC-8 collection), and P(10) and R-prec are somewhat more affected than either MAP or bpref-10. Still, the results indicate that a judgment set needs to have almost as many missing documents as existing documents before relative scores are noticeably affected.

6. CONCLUSION AND FUTURE WORK

Building substantially larger test collections with essentially complete relevance judgments through pooling is not likely to be possible due to the amount of assessor time and the diversity of retrieval runs that would be required. This paper looked at the effect relaxing the completeness assumption has on the Cranfield evaluation methodology. It showed that the Cranfield methodology is not robust to massively incomplete relevance judgments using today's most common evaluation measures, but that a new measure based only on

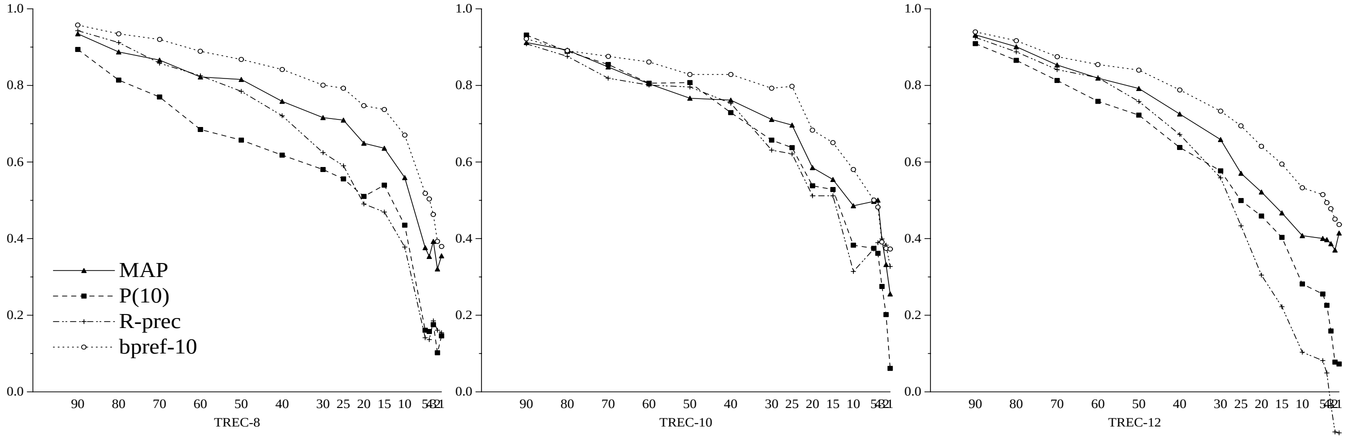

Figure 4: Kendall correlation of topic rankings. Topics are sorted by decreasing score averaged across all runs.

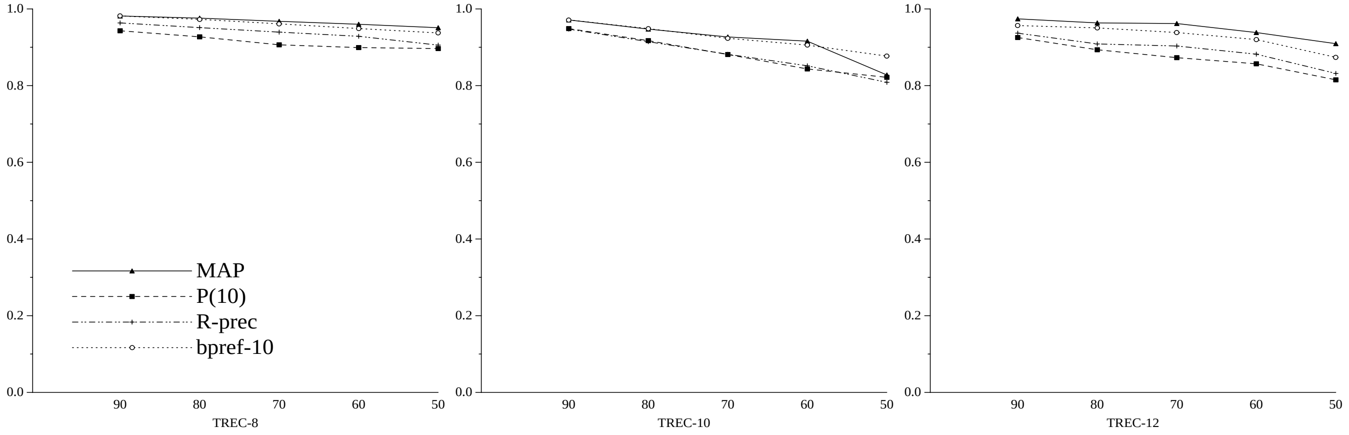

Figure 5: Change in Kendall correlation for imperfect judgments sets. The x-axis plots the percentage of the original document set used as the target document set, and the y-axis shows the Kendall τ correlation between the system ranking produced using the full document set and the system ranking produced using the reduced document set.

judged documents is highly correlated with existing measures when complete judgments are available and is more robust to incomplete judgments. The new measure, named bpref, is inversely related to the fraction of judged nonrelevant documents that are retrieved before relevant documents. A plot of the Kendall τ correlations between system rankings using the 100% qrels and the reduced qrels was much flatter for the bpref measure than for MAP, P(10) or R-precision. In the noisiest collection, bpref rankings remained essentially equivalent (i.e., τ scores remained about 0.9) until the relevance judgments were cut in half; in the less noisy collections the rankings remain equivalent even longer. The resiliency of bpref to increasing levels of incompleteness arises from the fact that the bpref measure does not depend on the specific ranks of relevant and judged nonrelevant documents, only on their relative ranks. Adding additional unjudged documents to a retrieved set can have no effect on that set's bpref score, but can have significant influence on the other measures' scores.

With complete judgments, the system rankings for bpref and MAP are essentially equivalent, indicating that on average the two measures agree on which is the better system. MAP has been shown to be a stable and sensitive measure, but has been criticized as possibly favoring the first few retrieved relevant documents. With bpref, each relevant document's score is independent of all other relevant documents' scores.

While using only judged documents with bpref to compute scores in a test collection built through pooling provides protection against incompleteness, it also puts additional requirements on the pooling process. One assumption sufficient for bpref to be a valid measure is that given a system and a retrieval rank, the chance of that document being in the judged pool is independent of whether the document is relevant or nonrelevant; i.e., that there is just as much chance of a nonrelevant document being in the pool as a relevant document. This is not an unreasonable assumption if the underlying reasons why the document being retrieved at that rank by this system match the reasons of the systems that contributed to the pool (an example reason might be that a document is "nearly relevant" by some metric that other systems have used.) However, a sufficiently novel good system could conceivably be retrieving documents for different reasons, and might therefore retrieve more nonrelevant documents not in the judged pool. Some evidence exists that this is not likely to be important: only 70 out of the 124 runs of TREC-8 contributed to the TREC-8 pool, and each run contributed only a maximum of 100 documents to the pool. Bpref agreed strongly with MAP on all runs of TREC-8, and on all topics of TREC-8, including those with over 100 relevant documents. But more investigation needs to be done before concluding that bpref can fairly evaluate novel systems that did not contribute to the judgment pool.

The properties of bpref demonstrated here make it the preferred measure to use when comparing systems over test collections with incomplete relevance judgments. Bpref is also more resilient to change than other measures when used on dynamic collections, though the limits of dynamic change that bpref can tolerate remain to be studied. A third type of test collection that bpref can be useful with is an embedded collection environment, where a test collection with known judgments is embedded in a much larger collection of similar documents with no judgments. For example, this would allow studies of operational efficiency of a large database to be studied while showing that effectiveness on an embedded subset is unchanged. Finally, bpref's theoretical properties, particularly that relevant documents' scores are independent of the rank of other relevant documents, make it more amenable to analysis than MAP. These properties may allow more direct linkage of theoretical retrieval approaches with the evaluation of those approaches.

7. REFERENCES

[1] Chris Buckley. trec_eval IR evaluation package. Available from `ftp://ftp.cs.cornell.edu/pub/smart`.

[2] Chris Buckley and Ellen 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, 2000.

[3] Cyril W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Reserach and Development in Information Retrieval*, pages 3–12, 1991.

[4] Gordon V. Cormack, Christopher R. Palmer, and Charles L.A. Clarke. Efficient construction of large test collections. In Croft et al. [5], pages 282–289.

[5] W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998. ACM Press, New York.

[6] M.B. Eisenberg. Measuring relevance judgments. *Information Processing and Management*, 24(4):373–389, 1988.

[7] H.P. Frei and P. Schäuble. Determining the effectiveness of retrieval algorithms. *Information Processing and Management*, 27(2/3):153–164, 1991.

[8] Google. Benefits of a Google search. `http://www.google.com/technology/whyuse.html`, January 2004.

[9] Stefano Mizzaro. A new measure of retrieval effectiveness (Or: What's wrong with precision and recall). In *Proceedings of the International Workshop on Information Retrieval (IR'2001)*, pages 43–52, 2001.

[10] Rabia Nuray and Fazli Can. Automatic ranking of retrieval systems in imperfect environments. In *Proceedings of the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pages 379–380, 2003.

[11] Mark E. Rorvig. The simple scalability of documents. *Journal of the American Society for Information Science*, 41(8):590–598, 1990.

[12] Ian Soboroff, Charles Nicholas, and Patrick 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, 2001.

[13] K. Sparck Jones and C. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

[14] C.J. van Rijsbergen. *Evaluation*, chapter 7. Butterworths, 2 edition, 1979.

[15] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[16] Ellen 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, 2001.

[17] Ellen M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems. Proceedings of CLEF 2001*, number 2406 in Lecture Notes in Computer Science, pages 355–370, 2002.

[18] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323, 2002.

[19] Ellen M. Voorhees and Donna Harman. Overview of the seventh Text REtrieval Conference (TREC-7). In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 1–23, 1999. NIST Special Publication 500-242.

[20] Y.Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46(2):133–145, 1995.

[21] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In Croft et al. [5], pages 307–314.