Evaluating epistemic uncertainty under incomplete assessments

Mark Baillie
Dept. Computer and Information Sciences, University of Strathclyde.
Glasgow, Scotland, UK.
Mark.Baillie@cis.strath.ac.uk

Leif Azzopardi
Dept. Computing Science, University of Glasgow.
Glasgow, Scotland, UK.
leif@dcs.gla.ac.uk

Ian Ruthven
Dept. Computer and Information Sciences, University of Strathclyde.
Glasgow, Scotland, UK.
Ian.Ruthven@cis.strath.ac.uk

April 24, 2007

Abstract

The thesis of this study is to propose an extended methodology for laboratory based Information Retrieval evaluation under incomplete relevance assessments. This new methodology aims to identify potential uncertainty during system comparison that may result from incompleteness. The adoption of this methodology is advantageous, because the detection of epistemic uncertainty – the amount of knowledge (or ignorance) we have about the estimate of a system's performance – during the evaluation process can guide and direct researchers when evaluating new systems over existing and future test collections. Across a series of experiments we demonstrate how this methodology can lead towards a finer grained analysis of systems. In particular, we show through experimentation how the current practice in Information Retrieval evaluation of using a measurement depth larger than the pooling depth increases uncertainty during system comparison.

1 Introduction

In this study we revisit the implications on system comparisons that arise from incomplete relevance assessments (i.e. *incompleteness*), and in particular the assumption that unassessed documents are not relevant. Instead of assuming unassessed documents are not relevant (Salton, 1992), or more recently, ignoring documents that are not judged when estimating performance (Buckley and Voorhees, 2004), we propose an alternative method: to quantify the proportion of unassessed documents in a systems ranked list. This alternative method leads to a complementary evaluation methodology, providing a new set of measures that attempt to quantify the *epistemic* uncertainty during system comparison[1]. This method-

[1]Epistemic uncertainty refers to the amount of knowledge (or ignorance) we have about the estimate of a system's performance and is a consequence of incompleteness. In contrast, *Aleatory* uncertainty refers to

ology can provide a guide for both researchers who re-use collections, and also for designers of new test collections. Adopting such an approach is important as researchers can detect potential uncertainty during evaluation, and also identify strategies for addressing the causes of epistemic uncertainty. In this study, we illustrate the utility of this evaluation methodology, as well as highlighting the implications of using particular performance metrics, related to the depth of measurement, under incomplete assessments.

Before introducing this new methodology we first provide background context by reviewing the running debate on incompleteness, and the subsequent implications on system comparison (Section 1.1). Next we discuss the typical evaluation metrics used within the context of incompleteness (Section 2). We then introduce this new methodology which augments the current evaluation protocol (Section 3). Next, we provide an empirical analysis of this approach across a range of existing test collections (Section 4). Finally, provide a discussion of the implications of this study (Section 5) before concluding the paper (Section 6).

1.1 Background

The beginnings of laboratory based evaluation of Information Retrieval (IR) systems involved studies such as the comparison of ASITIA and Documentation Inc. indexing systems, the Cranfield I and II experiments, SMART evaluation, and the in-depth evaluation and failure analysis of the Medlars search service [see (Sparck-Jones and Willett, 1997) for further details of these studies]. The collections used in these early studies were relatively small compared to modern collections – the Cranfield collection consisted of 1 400 documents. Studies involving larger collections followed, notably the STAIRS study for IBM where over 35,000 pages of text were used as a collection (Blair and Maron, 1985). The experience gained from these studies has been incorporated into the creation of modern collections where collection size has grown considerably. The most widely used examples being the Text Retrieval Evaluation Conference (TREC) initiative (Voorhees and Harman, 2005), the Cross Language Evaluation Forum (CLEF)[2], and NTCIR[3].

Most modern test collections adhere to a well defined model based on the findings of these previous studies referred to as the *Cranfield paradigm* (Voorhees, 2002). A corpus that follows the Cranfield model will consist of a collection of documents, statements of information need (named topics), and a set of relevance judgements listing the relevant documents that should be returned for each topic. To ensure fair comparison between systems, a number of key assumptions are made, including:

- The topics are independent.
- All documents are judged for relevance (i.e. completeness).
- These judgements are representative of the target user population.
- Each document is equally important in satisfying the users information need.
- The gathering of relevance assessments is independent of any evaluation that will use these assessments.

variation in system performance across the topic set, which is often addressed through the use of statistical significance tests.

[2]http://clef.iei.pi.cnr.it

[3]http://research.nii.ac.jp/ntcir

These assumptions are made to ensure fair comparison of system performance, although to develop an "ideal" collection where these assumptions hold is unrealistic – factors such as the collection size and available (human) resources often dictate to what degree these assumptions do hold. As a consequence these assumptions are often relaxed while compensating for any potential uncertainty that could be introduced with any relaxation, such as a bias favouring one system over another. For example, under the original paradigm relevance judgements were assumed to be complete i.e. all documents were judged in the document collection per topic statement. Initially completeness was viable but as document collection size increased, assessing all documents for all topics became intractable without exhaustive resources. It has been estimated that it would take more than nine months to judge an average size TREC collection for a single topic (Voorhees, 2002). Not only is this expensive both in terms of time and resources, but over a protracted time period the criteria an assessor will use to judge a document (for relevance) could deviate significantly, resulting in inconsistencies in the relevance assessments[4].

Nonetheless, incompleteness is problematic to laboratory based system evaluations. For example, the computation of recall based performance measures require the complete set of relevant documents (in the collection) to be known. To address this limitation, techniques such as system pooling have been proposed (Spärck-Jones and Van Rijsbergen, 1975) – pooling is a process of combining a number of varying search approaches to sample the collection for relevant documents providing an unbiased estimate of recall. From this estimate, the relative system performances can be compared.

When using pooling to estimate recall it is difficult to ascertain whether the majority of relevant documents have been discovered. There have been a number of empirical studies that have attempted to quantify how many relevant documents remain undiscovered. For example, Zobel (1998) defined a method to extrapolate the potential numbers of unassessed relevant documents in TREC collections. He approximated that a large percentage of relevant documents were still to be discovered, especially across topics where a large number of relevant documents were already found through pooling, concluding that the assumption of unassessed documents as not relevant was unfounded.

For this very reason the effect that pooling, in the context of relevant document recall, has on system comparison has been investigated. Such studies have focused upon several different areas of the completeness assumption and system pooling including, the effect on system comparison and the subsequent uncertainty when using incomplete relevance judgements (Buckley and Voorhees, 2004; Voorhees and Harman, 2005; Zobel, 1998), efficient pooling strategies (Carterette et al., 2006; Cormack et al., 1998; Sanderson and Joho, 2004; Zobel, 1998), automatically generated relevance assessments (Aslam and Savell, 2003; Soboroff et al., 2001), and the importance of significance testing during system comparison (Sanderson and Zobel, 2005). A running theme throughout these studies is that it is still unclear whether the now standard assessment procedure of pooling, and the resultant evaluation measures adopted, does indeed impact upon the fair and unbiased comparison of retrieval systems, and to what extent.

A recent investigation of the TREC Robust-HARD 2005 collection identified a bias in the collection which was a result of both a shallow pool depth and similar runs forming the system pool (Buckley et al., 2006). The outcome was a bias favouring systems that retrieved

[4]Please refer to articles by (Barry, 1994; Harter, 1992; Ruthven, 2005; Saracevic, 1995; Schamber et al., 1990) for an introduction into the various criteria which influence the assessment of information.

documents containing the <title> keywords of the TREC topic. Although this does not necessarily indicate a failing of system pooling it motivates the need for a stronger evaluation framework, which considers aspects such as pooling, the status of unassessed documents, and measurement depth within the evaluation.

We now critically review this debate in further detail. We first highlight the potential implications of incompleteness within the Cranfield paradigm and any uncertainty that may be introduced as a result. We then investigate these implications within a newly proposed framework designed to quantify the uncertainty between system comparisons.

1.2 System Evaluation under Incomplete Relevance Judgements

Swanson (1988) stated as one of his postulates of impotence, a set of truisms for Information Retrieval, that it is never possible to verify if all relevant documents have been discovered for a topic, as one can never examine all documents without unlimited resources while using a strict and static set of criteria for judging relevance. This truism is related to the universal rule that empirical evidence can always be refuted but never verified. In other words, a claim of completeness can always be refuted by discovering a previously unassessed relevant document within a collection. This postulate is especially resonant given the size of modern test collections such as those created as part of the TREC initiative (Voorhees and Harman, 2005). For this very reason relative system performances are compared instead of absolute values of effectiveness, as measures of retrieval performance often require knowledge of the total number of relevant documents in a collection with respect to each topic. It is assumed that relative system performances can be compared as long as robust strategies to estimate the proportion of documents relevant to a topic are used, ensuring fair experimental conditions for the systems under comparison. One such technique for recall estimation is system pooling.

System pooling System pooling was proposed to address the intractability of the completeness assumption (Spärck-Jones and Van Rijsbergen, 1975). Pooling is a focused sampling of the document collection that attempts to discover all potentially relevant documents with respect to a search topic e.g. approximate the actual number of relevant documents for a given topic. To do so, a number of (diverse) retrieval strategies are combined to probe the document collection for relevant documents[5]. Each system will rank the collection for a given topic, then the top λ documents from the subsequent ranked lists are collated, removing duplicates, to form a pool of unique documents[6]. All documents in this pool are then judged for relevance by an assessor(s) using specified criteria[7] such as topicality or utility[8]. In the

[5]In the case of TREC, CLEF and NTCIR participating systems will submit a number of runs to be included in the pooling process. A system or group can submit numerous varying runs of the same system where a parameter or feature has changed in each run, however, only a selection of these runs will be included in the system pool while the others will be evaluated only.

[6]The cut off λ is known as the *pooling depth*. The *measurement depth* d refers to the document cut off used when estimating retrieval performance e.g. Precision at d documents.

[7]There is a running debate on the static assumption of relevance used in the system evaluation framework. This debate is outside the scope of this study. For further details on this debate please refer to the textbook by Ingwersen and Järvelin (2005), who provide an up-to-date overview.

[8]Topicality relates to whether a document is *about* a topic, while utility is a measure of the *usefulness* of the document to a given task. A decision to select which criterion the assessors will use is important. Consider an example of a document whose contents are about a topic but do not contain information which is useful for the user to complete their intended task.

case of TREC, the assessor(s) use the topic statement for guidance when determining which documents are (topically) relevant or not relevant.

Status of unassessed documents The remaining *unassessed* documents are assumed not to be relevant. This assumption follows the argument put forward by Salton, that by using a range of different IR systems the pooled method will discover "the vast majority of relevant items" (Salton, 1992). This argument is based on the assumption of diminishing returns i.e. because many runs contribute to the system pool, it is highly likely that the majority of relevant documents, or those documents representative of relevant documents, are returned through pooling. If this assumption holds then there is little need to assess those documents not included during pooling.

There has been a running debate about the validity of the assumption that unassessed documents are not relevant. Initially this assumption was made on smaller collections such as Cranfield and CACM, however, as Blair posited, the percentage of unassessed documents could be up to 99% for a given topic with respect to a modern collection, leaving a large proportion of relevant documents undiscovered (Blair, 2002b):

> .. the very hypothesis that is being tested – that the systems, collectively, order the collection with all the relevant documents ranked at the top – is used to *justify* the pooling method of recall estimation; that is, *the accuracy of relevancy ranking is assumed in order to justify a method of estimating the accuracy of relevancy ranking*. This kind of flawed reason is known, by logicians, as Petitio Principii (more commonly, "Begging the question") and most certainly cannot be taken as justification for the pooling method of recall estimation.

This argument can be explained as follows:

1. The assumption that the intellectual content of a document can be accessed by using some form of query term matching alone. Even by submitting variations of query terms, adjusted through trial and error, as in a typical search session, the likelihood of a searcher finding a substantial proportion of relevant documents has been discovered to be low across a number of various studies (Blair, 1996). An explanation for this limitation is that the intellectual content of a document is difficult to represent automatically and relates to Swanson's 5th and 6th postulates of impotence (Swanson, 1988). A document can be *about* a topic without ever mentioning key terms or phrases that a user may expect to appear. Also query terms chosen by the user may not discriminate between relevant and non-relevant documents, especially as the collection size grows (Blair, 2002a). A user searching for documents on a new subject may not select terms representative of the subject they are searching that will also discriminate such documents from the non-relevant documents which share similar vocabulary. Consequently, not all potentially relevant documents will be retrieved through keyword matching techniques alone. For example, the study by Buckley et al. (2006) illustrates how a shallow pool depth may result in a bias favouring the retrieval of relevant documents which contain title keywords in the TREC topic (i.e. short queries) over relevant documents that do not.

2. The belief that pooling is accurate because relevant documents were discovered. Locating a proportion of relevant documents is not a sole indicator of good retrieval performance, as the proportion of relevant documents missed is not known unless it is

quantified through other means. Swanson refers to this as the "fallacy of abundance" – by discovering a (substantial) number of documents about a topic creates a illusion that little remains hidden (Swanson, 1960). The searcher, or in this case the pooling process, cannot be assessed to determine how much still remains hidden. Such an assumption comes from the mistaken belief that science "seeks confirmations of a hypothesis rather than rejections". Confirmations of a hypothesis are easy to show but do not "provide deep insights unless there is some degree of risk in the predictions" (Blair, 1996). This philosophy is motivated by Karl Popper (Popper, 2000), who outlined a number of considerations concerning the verification of scientific theories including that "confirming evidence should not count except when it is the result of a genuine test of the theory; and this means that it can be presented as a serious but unsuccessful attempt to falsify the theory". In other words, system pooling alone cannot be used to confirm the accuracy of system pooling.

3. A pooling of systems may not always search in all the best places. This is a reflection of the diversity of the systems contributing to the pool. Does the selection of systems included in the pool share enough diversity as to cover all means of searching, thereby locating all potentially relevant documents that could be found through searching? If the retrieval mechanisms in the pool are (theoretically) similar then we could argue no. If they exhaustively cover many varying approaches (both manual and automatic[9]), we can argue for the converse. But paradoxically we will never be confident of which case is true until we can quantify with a high degree of confidence how accurate each retrieval system is are at discovering all relevant documents, as indicated in (2), therefore researchers should always err towards the side of caution when reporting results.

However, in a rejoinder to the comments of Blair, Voorhees and Harman (2003) highlight a key point that as pooling is a union of many different ranking approaches and because only relative system performance is measured, if the number of systems contributing to a pool is sufficiently large and these systems are diverse, bias towards one or a set of systems should be minimised, even though not all relevant documents are found. Absolute system performance may not be accurately estimated using incomplete relevance assessments, but the relative performances of systems can be fairly compared. This is related to the argument put forward by Salton (1992), where as long as the conditions remain even for all systems, then the relative differences between systems can be compared fairly and with a high degree of certainty.

However, we hypothesise that uncertainty remains when comparing the relative performance of systems due to the status of unassessed documents (i.e. those documents not in the pool of documents to be assessed for relevance). When using pooling to estimate recall it is difficult to ascertain whether the majority of relevant documents have been discovered (Zobel, 1998). It is not clear what impact the potential proportion of relevant unassessed documents may have on system comparisons. A number of studies have investigated this issue and whether system pooling is a robust solution to the completeness problem as a result of this (Buckley and Voorhees, 2004; Voorhees, 2002; Zobel, 1998; Wallis and Thom, 1996). We now summarise these studies.

[9]It should be noted that usually a typical pool is supplemented with a number of manual runs to increase the diversity of search approaches.

Potential Implications of Incompleteness on Evaluation Zobel (1998) investigated if potential bias, introduced during pooling, may effect system evaluation. Two forms of potential bias were investigated: (1) *system reinforcement* and (2) *system omission* bias. The first form of bias assumed that (theoretically) related systems involved in the pooling process, or approaches that combine a number of different retrieval strategies, could reinforce each other at the expense of more diverse systems (to these techniques), with the performance of these novel systems being underestimated as a consequence. The premise was that systems that rank similar document sets will have a higher likelihood of documents from this set being included in the judgement pool than more diverse, unrelated, techniques. These related systems as a result could share a larger representation of assessed documents per topic. Under the same assumption, techniques that combine different retrieval approaches represented in the system pool may also have performance overestimated. The assumption is that combination approaches (may) inadvertently mimic the pooling procedure thus maximising the number of judged documents in the final ranked. As a side-effect of pooling, both scenarios increase system performance artificially as the probability of an unassessed document being relevant is zero while the probability of an assessed document is significantly higher than zero.

If both scenarios do occur then this may be compounded by current evaluation practice. It is common to use metrics estimated from a measurement depth d larger than the pool depth λ for system comparison. Typically performance measurements are calculated at a particular depth of a systems document ranking (measurement depth). This is the overall rank position from which systems are normally compared. Using a fixed measurement depth is motivated by both the problem of incompleteness, and also to allow for averaging over a set of topics. The measurement depth often exceeds the pool depth. For example, the measurement depth for TREC is at $d = 1000$ documents per topic, while the pool depth is no more than $\lambda = 100$ per submitted run – although this threshold may vary depending on task, collection and available resources. The rationale is that good systems will retrieve relevant documents at ranks greater than the pool depth. These documents will, however, be included in the set of assessed documents as they will be retrieved by the remaining systems that form the system pool. Thereby measuring beyond the pool depth will provide better discrimination between systems of varying performance. Despite improved discrimination, uncertainty in the results is also increased. Such practice in turn creates potential uncertainty where similar systems will have a larger proportion of judged documents in the final ranked list to depth d.

Zobel concluded that the artificial effect of system reinforcement would be insignificant if the judging pool is sufficiently deep, although measurement depth may be prone to this effect. The current practice of using a measurement depth larger than the pool depth caused reservations. Increasing measurement depth improves discrimination between systems i.e. determining whether system A is significantly better than B. But a change in measurement depth was noted to affect system ranking in terms of performance when comparing multiple systems. Sometimes the ordering of systems in terms of performance changed by as many as 6 places in the ranking as measurement depth increased. A number of systems were also found to be returning a larger proportion of unassessed documents. Applying a measurement depth of 1000 documents could underestimate those systems that are not as well represented as others in terms of the proportion of documents in the judging pool as a consequence.

The second reservation (system omission) highlighted that the evaluation of novel systems that did not contribute to the judging pool may have their performance understated. A novel system, especially one that diverges from the pooled retrieval strategies significantly enough that it retrieves a large proportion of documents outwith the pool, could be underestimated.

When analysing the ten topics with the highest proportion of relevant documents only, average performance of a novel system was underestimated by up to 19% for the TREC-3 collection[10]. This underestimation was stated to be dependent on the pool size. The effect of system omission was thought to minimise as the document pool size increased per topic. These fluctuations were considered to be identifiable in practice through the analysis of documents in a ranked list that are not contained in the judgement pool. New novel systems in particular were warned to be aware of potential bias towards their system and under-estimation in performance.

This viewpoint was further supported with incompleteness believed to have only a negligible impact on the evaluation of new systems that were not represented in the original system pool (Voorhees, 2002). System omission was believed to be a "red herring", however the impact of system reinforcement was not formally evaluated. In conclusion, the current evaluation practice was found to be robust to the violation of the completeness assumption as the effects of system reinforcement and omission would produce only slight variations in system performance.

Incompleteness was further analysed in a following study (Buckley and Voorhees, 2004). The motivation of this investigation was to measure empirically whether incompleteness was problematic during current IR laboratory evaluation, in particular the stability of standard IR measures under incompleteness. To measure the effects of incompleteness, the ranking of systems in terms of performance across both "complete" and incomplete assessments were correlated. A number of TREC collections were used for the evaluation, with a set of incomplete relevance assessments artificially simulated through the random sampling of the original assessments, which was assumed to be "complete". However, as the TREC collections use system pooling to compile relevance judgements, these original assessments are invariably incomplete as well.

The findings of the study identified that the practice of system pooling was robust to system omission across a wide range of topics and varying levels of incompleteness. However, the standard evaluation measures were not stable under substantial levels of incompleteness, with only *bpref* found to be invariant (most of the time). Therefore, new novel systems not represented in the pool could be fairly compared with those systems represented when using *bpref*. This result was significantly important as the *bpref* measure was designed specifically to address the problem of incomplete relevance assessments by removing unassessed documents, which are normally treated as not relevant, from the ranked document list. The *bpref* metric was found to be more stable across incomplete assessments, although the measure tended to be coarse when there was a small proportion of relevant documents belonging to a topic. It was concluded overall that adopting system pooling for IR system evaluation was robust to incompleteness.

The effects on evaluation when using shallow pooling depth has also been studied. The investigation of the TREC Robust-HARD 2005 collection identified a bias in this collection which was a result of both a shallow pool depth, and (potentially) similar runs forming the system pool (Buckley et al., 2006). The outcome was a bias in the collection favouring relevant documents that contained title terms from the TREC topic compared to other relevant documents. Although this does not necessarily indicate a failing of system pooling it highlights the need for further investigation into the evaluation framework, especially aspects such as pooling, the status of unassessed documents and measurement depth.

[10]See Voorhees and Harman (2005) for further details about the various TREC collections

1.3 Focus of this Study

Based on an analysis of these studies, we posit that uncertainty remains when comparing the relative performance of systems as a result of the status of unassessed documents (being not relevant). One of the cited limitations with laboratory studies is the large amount of subjectivity or uncertainty in such evaluations. The nature of the scientific method demands as much objectivity and certainty as possible. After analysing the history of retrieval evaluation we believe that the status of unassessed documents and the resulting suitability of comparing systems with varying levels of assessed documents is still an open issue. We are especially motivated by the recommendations of Zobel (1998), who warned researchers when evaluating new systems across existing test collections for cases where performance could be underestimated. However, a standard protocol for detecting such cases has not been proposed as of yet. We therefore propose a new methodology for quantifying uncertainty during system comparisons that may exist because of incomplete relevance assessments. By doing so, we can determine when it is possible to *fairly* compare two systems using current measures, especially those systems that do not contribute to the pool. Instead of compensating for or ignoring potential uncertainty during system comparisons due to incompleteness, we believe that the proportion of unassessed documents should be captured and reported. Reporting this information can be a useful source of information to help quantify the confidence, accuracy and/or reliability of system performance. By capturing such information, we can determine whether two systems are compared under similar conditions, and flag those cases when one system may have an advantage over another due to a particular condition found in a test collection.

2 System comparisons

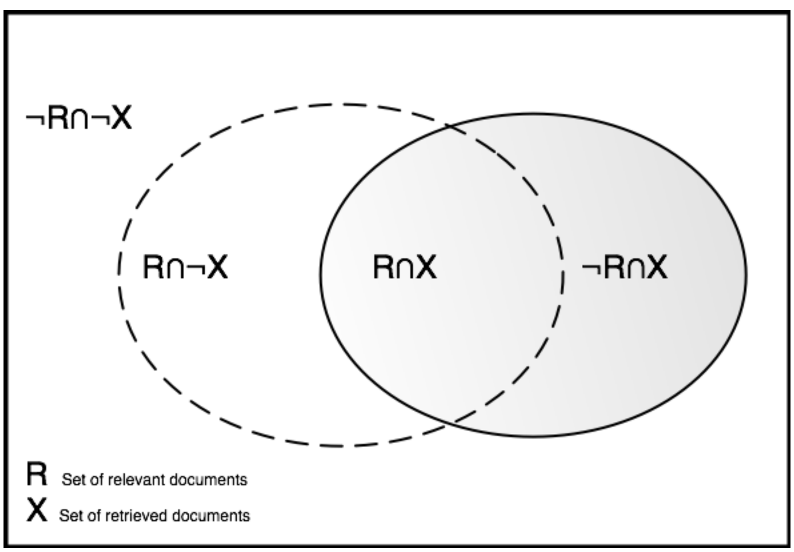

Figure 1: A Venn diagram of the sets of retrieved and set of relevant documents under the assumption unassessed documents are not relevant. Precision and Recall are derived from these two sets.

Under the Cranfield model system effectiveness is typically estimated using Precision (i.e. the proportion of retrieved documents that relevant) and Recall (i.e. the proportion of (known) relevant documents retrieved overall)[11]. We now define these metrics. Let N be the total number of documents in a collection, R be the set of relevant documents, and X be the set of retrieved documents. Under the current evaluation protocol, the status of unassessed documents is assumed to be not relevant (see Figure 1). Using the contingency Table 2, Precision can then be defined as:

	Relevant	Not Relevant	Total
Retrieved	$R \cap X$	$\neg R \cap X$	X
Not Retrieved	$R \cap \neg X$	$\neg R \cap \neg X$	$\neg X$
Total	R	$\neg R$	N

Table 1: The standard assumption of laboratory IR is to denote unassessed documents as not relevant. This allows the following four outcomes given the set of retrieved and set of relevant documents.

$$Precision = \frac{|R \cap X|}{|X|} \tag{1}$$

where $|.|$ is a counting measure (Van Rijsbergen, London, 1979).

Recall can be defined as:

$$Recall = \frac{|R \cap X|}{|R|} \tag{2}$$

Typically Precision and Recall measurements are calculated at a particular depth of a systems document ranking (measurement depth d). This is the overall rank position from which systems are normally compared. Recall is calculated at rank d rather than for all documents in the collection. Precision is taken at various ranks up until d. Using a fixed measurement depth is motivated by both the problem of incompleteness, and also to allow for averaging over a set of topics. As previously discussed the measurement depth often exceeds the pool depth to provide better discrimination between systems.

Due to incompleteness the absolute system performances are not compared over a set of topics. Instead the relative system performances are compared. If the same inconsistencies such as completeness are common for all systems, where no bias exists favouring one system over another, then the relative performances can be compared with a high degree of confidence (Salton, 1992). Due to the inconsistencies affecting all systems equally, the relative ranking of systems will remain consistent (Voorhees, 2000). Therefore it is common practice to use the relative performance for comparing systems, displaying the error bars over the set of topics, and where possible it is recommended that some form of significance test is applied i.e. Paired sample T-test, Wilcoxon Signed Ranked test or one-way ANOVA (Hull, 1993; Savoy, 1997; Zobel, 1998; Sanderson and Zobel, 2005).

[11] Recently new measures based on graded relevance judgements have been proposed which place emphasis on rewarding systems retrieving highly relevant over partially relevant documents (Järvelin and Kekäläinen, 2002). In this paper we focus on metrics that assume relevance is a binary variable (i.e. a document is relevant or not) to better understand the effect of unassessed documents on evaluation, and leave graded relevance assessments for future work.

For significance testing it is often desirable to provide a single point estimate of the performance of a system. Often precision at a predefined rank d is used ($P@d$), although the most popular metric adopted is Mean Average Precision (MAP). Average Precision (AP) is defined as:

$$AP = \frac{1}{|R|} \sum_{D_k \in (R \cap X)}^{d} \frac{|R_k \cap X_k|}{w_k} \tag{3}$$

where D_k is a sequence of documents ranked by a system, w_k is the absolute rank position of the k^{th} document, and d is the predefined measurement depth. AP measures both the Precision over the ranked list but also includes a Recall aspect in the measurement by taking Precision at the rank of each relevant document found. Therefore a system is rewarded both for finding relevant documents, and also ranking these documents towards the top of the ranked list. Over a set of topics, the mean of AP is often taken hence the name Mean Average Precision.

2.1 The binary preference measure (*bpref*)

In the previous section, we highlighted that the metrics were defined under the assumption that unassessed documents are considered not relevant. Recently, this assumption was re-addressed by (Buckley and Voorhees, 2004), who proposed a new measure called binary preference (*bpref*). Due to incompleteness, the new *bpref* measure was designed to estimate performance on the assessed documents in a system ranked list only. *bpref* is the mean number of times R assessed non-relevant documents rank above the the R relevant documents, where R is the total number of relevant documents. All unassessed documents not belonging to the judgement pool are ignored, which is intended to limit any potential uncertainty that is a result of incompleteness. Consequently, the status of unassessed documents was reconsidered (Figure 2 illustrates this change). Under this new assumption, we now have a new set of documents A that represent the assessed set of documents.

	Assessed		Unassessed		
	Relevant	Not Relevant	Relevant	Not Relevant	Total
Retrieved	$R \cap X \cap A$	$\neg R \cap X \cap A$	$R \cap X \cap \neg A$	$\neg R \cap X \cap \neg A$	X
Not Retrieved	$R \cap \neg X \cap A$	$\neg R \cap \neg X \cap A$	$R \cap \neg X \cap \neg A$	$\neg R \cap \neg X \cap \neg A$	$\neg X$
Total	$R \cap A$	$\neg R \cap A$	$R \cap \neg A$	$\neg R \cap \neg A$	N

Table 2: Under the new assumption the set of assessed A and unassessed $\neg A$ documents are also considered.

Using the updated contingency table (see Table 2.1), *bpref* can be formally defined as[12]:

$$bpref = \frac{1}{|R \cap A|} \cdot \sum_{D_k \in (R \cap X \cap A)}^{|R \cap A|} 1 - \frac{|\{D_l \in (\neg R \cap X \cap A) : w_l < w_k\}|}{\min\{|R \cap A|, |\neg R \cap A|\}} \tag{4}$$

[12]This is an updated definition of *bpref* that corrects the original definition. See Soboroff (2006) for further details.

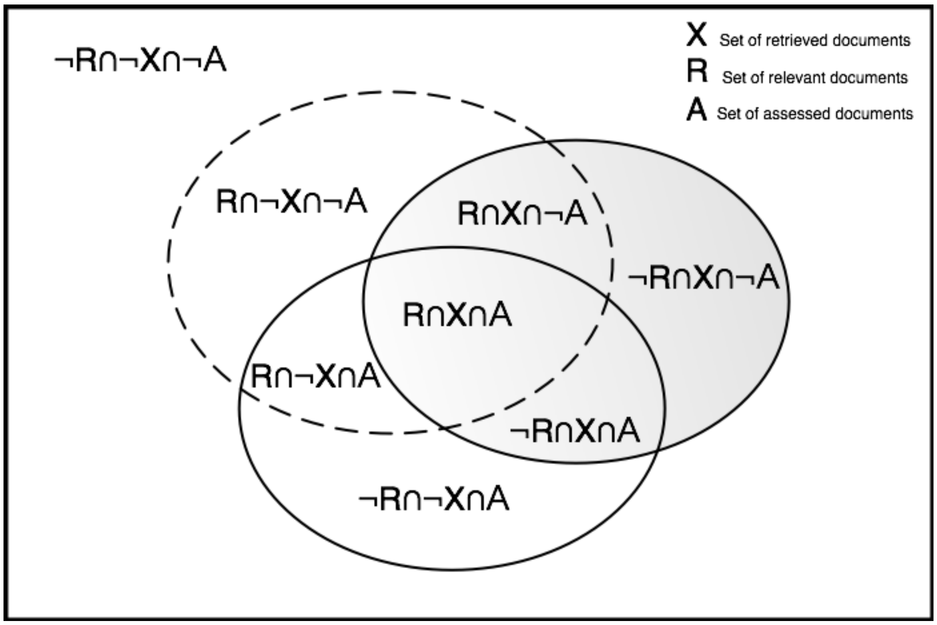

Figure 2: Illustration of the set of retrieved, set of relevant documents and set of assessed documents under the proposed assumption using a Venn diagram. Precision and Recall are based on the intersection of these three sets. With complete assessments, A will equal the total number of documents in the collection, hence reverting back to the original assumption shown in Figure 1. Also note that if $R \cap A = R$ implies $R \subset A$, then all relevant documents were discovered resulting in complete relevance assessments.

where D_k is a set of documents, w_k is the absolute rank position of a document in D_k. *bpref* examines the ranks of preference pairs of assessed relevant and non-relevant documents i.e. $w_l < w_k$, performing a summation of the difference in ranks of each pair.

We can also redefine measures such as Precision, Recall and AP under this new assumption:

$$Precision = \frac{|R \cap X \cap A|}{|X|} \tag{5}$$

$$Recall = \frac{|R \cap X \cap A|}{|R \cap A|} \tag{6}$$

$$AP = \frac{1}{|R \cap A|} \sum_{D_k \in (R \cap X \cap A)}^{d} \frac{|R_k \cap X_k \cap A|}{w_k} \tag{7}$$

2.2 Which measures to use?

The advent of *bpref* has lead to a fundamentally different assumption with respect to the status of unassessed documents. Typical Recall and Precision based metrics address the problem of incomplete assessments differently to that of *bpref*. The former suite of metrics assume unassessed documents are not relevant while *bpref* removes the unassessed documents from the performance estimation process. In other words unassessed documents are ignored.

Implicitly *bpref* assumes that unassessed documents will follow a similar ranking of non-relevant and relevant documents to what has been observed in the assessed documents.

This strategy of *bpref* – to base the performance only on what has been assessed and extrapolating that value to the unassessed results – provides a more conservative estimate of a systems performance, particularly as less information is available to estimate the value. Invariably *bpref* will lower the difference in scores between the systems. Conversely, when using Precision and Recall metrics, the problem of incompleteness is ignored. As a result, the estimate of system performance is more liberal and therefore more prone to increased levels of incompleteness and hence stability (Buckley and Voorhees, 2004).

Given this limitation we still propose to utilise existing measures instead of *bpref* for this study. This choice is influenced by a number of factors such as (i) *bpref* strongly correlates with measures such as MAP when used on test collections that are "complete" i.e. official TREC collections, (ii) *bpref* is still to become an established performance measure such as MAP or P@10, and (iii) what aspect of retrieval *bpref* is measuring is not as intuitive as Recall and Precision.

The correlation between *bpref* and MAP weakens when collections have a "substantial level" of incompleteness (Buckley and Voorhees, 2004), with standard Precision and Recall metrics becoming unstable. In this study we are interested in detecting when such situations exist particularly when using standard TREC like collections. For example, when the conditions between two systems are not even for a specific collection, thereby using a standard performance metric would not be suitable during system comparison. Instead of removing information to provide a more stable estimate, as in the case of *bpref*, we want to detect those scenarios informing the researcher when an alternative means is required such as identifying new collections, using shallower measurement depths, re-assessing unassessed documents, or utilising *bpref*.

3 Capturing the (un)certainty of System Performance

We now propose a methodology to address the epistemic uncertainty associated with system comparisons under incompleteness. Instead of compensating for or ignoring potential uncertainty during system comparisons, due to incompleteness, we believe that the proportion of unassessed documents should be captured and reported. Reporting this information can be a useful source of information to help quantify the confidence, accuracy and/or reliability of system performance. By capturing such information, we can determine whether two systems are compared under similar conditions, and flag those cases when one system may have an uneven advantage over another due to the a particular condition found in a test collection. This is especially motivated by the recommendations of (Zobel, 1998) who warned researchers when evaluating new systems over existing test collections for cases of potential *system omission*.

We hypothesise that the (un)certainty associated with estimating a measurement of a systems performance at a depth d is proportional to the number of documents that have been assessed at that depth. Conversely, the uncertainty is proportional to the number of documents that have not been assessed. The more assessed documents contained in a ranked list, the more confident we are in using the estimate of a systems performance at the corresponding depth. For example, when comparing the performance of two systems, if all documents have been assessed in the ranked list of both systems then we have the *ideal* situation of completeness i.e. the performance estimates for both systems were made with full information. If the ranked lists of both systems are incomplete, but contain similar proportions of assessed documents, then confidence in the relative comparison of these two systems would also be high. However, if we are measuring performance where one system has a substantially larger proportion of assessed documents than another, then the performance estimate of one system is based on limited information relative to the other. It is these cases that we wish to detect, where the conditions for both systems are not even, thus there is a higher degree of uncertainty in the comparison.

Case 1 P(s1)==P(s2) A(s1)==A(s2) or P(s1)==P(s2) A(s1)==A(s2) Accept null hypothesis that s1 is equal to s2, with a low degree of *uncertainty*	**Case 2** P(s1)==P(s2) A(s1)<<A(s2) or P(s1)==P(s2) A(s1)>>A(s2) Accept null hypothesis that s1 is equal to s2, with a high degree of *uncertainty*
Case 3 P(s1) >>P(s2) A(s1)<<,==A(s2) or P(s1)<<P(s2) A(s1)==,>>A(s2) Reject null hypothesis that s1 is equal to s2, with a low degree of *uncertainty*	**Case 4** P(s1)>>P(s2) A(s1)>>A(s2) or P(s1)<<P(s2) A(s1)<<A(s2) Reject null hypothesis that s1 is equal to s2, with a high degree of *uncertainty*

Figure 3: The System Evaluation Decision Matrix for system comparison.

3.1 Measure of Assessment

We propose to capture uncertainty by calculating the proportion of assessed documents in a ranked list. First, we now introduce two measures using the same notation defined in Section 2. Assessment precision A_p is defined as the proportion of assessed documents in a ranked list:

$$A_p = \frac{|X \cap A|}{|X|} \tag{8}$$

The Recall of Assessment is the proportion of assessed documents retrieved overall:

$$A_r = \frac{|X \cap A|}{|A|} \tag{9}$$

where $|X \cap A|$ is the number of documents in the set defined by the intersection of X and A, and $|X|$ is the number of documents in X. Assessment Precision relates to the confidence we place, or the certainty of a performance estimate, given a ranked result list. Note that uncertainty associated with the estimate is the complement, $1 - A_p$. We now refer to uncertainty and certainty through this measure, where a high Assessment Precision value relates to high certainty and low uncertainty.

This measure is exactly the definition for standard Precision except with respect to assessed as opposed to relevant documents[13]. Consequently, for every Precision and Recall measure there is a corresponding Assessment measure. The Average Assessment can be computed by taking the average over all ranks where an assessed document occurred. The Mean Average Assessment (MAA) then provides a summary of the assessment over all topics placing more emphasis on systems with assessed documents higher in the ranked list. This metric is analogous to Mean Average Precision (MAP), and could be used in situations where it is important to identify whether there is a difference in the proportion of assessed documents at higher ranks between systems when estimating MAP[14].

It should also be noted that the Assessment Precision metrics are functionally related to the corresponding Precision metrics. This relationship is because A is the union of the set of assessed relevant documents and assessed non relevant documents. Therefore a system retrieving more assessed documents is likely to have a higher Precision, because assessed documents are more likely to be relevant. Also, when systems have low levels of A_p there is increased uncertainty in the Precision score, and any subsequent comparison, because of the high proportion of unassessed documents. It is important to consider this context during the evaluation. Assessment Precision provides this context explicitly by capturing the proportion of assessed documents used to estimate the retrieval performance. In this paper, we concentrate on applying A_p to fairly compare systems, and leave these other issues regarding A_p for future research.

3.2 System Evaluation Decision Matrix

We now illustrate how Assessment Precision can be integrated into the current evaluation protocol. We motivate the introduction of the System Evaluation Decision matrix in the form of an example system comparison. We wish to test the performance P(), which denotes the Precision at a given measurement depth k (i.e. P@10, MAP, etc.), of two systems s_1 and s_2 over a test collection with incomplete relevance assessments.

We have the following research hypothesis:

$H_0 \,:\, P(s_1) = P(s_2)$

$H_1 \,:\, P(s_1) \neq P(s_2)$

[13]Consequently, the level of assessment can be easily computed by modifying the qrel file (i.e. the file containing all relevant documents for a topic set) file such that all documents are marked to be relevant which now denotes Assessment. This can also be achieved by using the -l option when using the trec_eval evaluation tool provided by TREC

[14]We have focused on Precision based metrics in this paper although quantifying the level of assessment can be extended to included other types of measures. For example, *bpref* itself does not quantify the proportion of assessed documents that are removed when estimating performance but a corresponding A_p measure could be derived to complement such a metric.

To determine the level of confidence we can place on this test, we test the supplementary hypothesis using a corresponding Assessment Precision metric A(), which denotes the A_p at a corresponding measurement depth k (i.e. A_p@10, MAA, etc.):

$H_0 \; : \; A(s_1) = A(s_2)$

$H_1 \; : \; A(s_1) \neq A(s_2)$

This forms a contingency table of four possible outcomes of interest displayed in Figure 3. Significance is denoted as either no difference (==) or the significant differences (<<, >>) i.e. s_1 is significantly better (>>) than s_2. We assume that statistical significance is determined using an appropriate test such as Wilcoxon sign rank test, paired T-test or ANOVA (Sanderson and Zobel, 2005).

For Case 1, the null hypothesis that $P(s_1) == P(s_2)$ and $A(s_1) == A(s_2)$ cannot be rejected. We define this a "strong" case because the level of assessment for both s_1 and s_2 are equal, that is the proportion of information used to estimate performance was comparable.

For Case 2, the null hypothesis that $P(s_1) == P(s_2)$ cannot be rejected as well, however, the proportion of information used to estimate the performance of both systems was not comparable. In other words, the result list of one system was comprised of a significantly larger proportion of assessed documents than the other, causing a degree of uncertainty in this comparison. It is unknown from this test whether, under comparable conditions, the null hypothesis $P(s_1) == P(s_2)$ would still hold or not. We therefore define this as a "weak" case.

For Case 3, also a "strong" case, the null hypothesis that $P(s_1) == P(s_2)$ is rejected. We can place a high degree of confidence in this outcome as we have either a scenario where both systems share similar proportions of assessed documents, or in special scenarios the system with significantly higher performance has significantly fewer documents assessed than the other system. In other words, even with further information about this system it could not match (or better) the opposing system.

Finally, for Case 4, another "weak" case, the null hypothesis that $P(s_1) == P(s_2)$ is rejected, although, we cannot place a high degree of confidence in this outcome, as the system with significantly higher performance also reported a significantly larger proportion of assessed documents. This does not indicate that the system with a smaller proportion of assessed documents would share similar performance under equal conditions, but instead flags a potential problem with this comparison.

Figure 4 displays an example of a pair-wise system comparison from each case using runs from an official TREC track (Robust 2005)[15]. The left-hand column is a plot of system performance (i.e. MAP) across a range of measurement depths. The right-hand column is a plot of the corresponding Assessment precision metric for each run. Of the weaker outcomes Case 2 is particularly interesting as both systems have similar performance, but this performance is based on different proportions of assessed documents. What is interesting is that the system with significantly less assessed documents could potentially be retrieving a wider diversity of documents, with respect to the pool, and some of these documents may be relevant (Zobel, 1998). A subsequent research question would be to investigate why the systems perform as well as each other. As both systems have equal system performance but unequal levels of assessment, this system may potentially improve performance when

[15] In this version of the draft the figures can be found at the end of the paper.

compared under even conditions. Further investigation may provide stronger supporting evidence.

At this stage a number of steps could be taken. If the goal of the comparison is precision orientated then system comparison could be made at a shallower measurement depth to ensure the likelihood of parity. By doing so we are assuming that at shallower depths systems will have relatively equal proportions of documents assessed. If both systems have contributed to the pooling process then this assumption would hold up until pooling depth has been reached, however, if a system has not contributed to the pool this may not be the case. The previous step may lead to the creation of test collections with an emphasis on shallow measurement depth (Sanderson and Zobel, 2005). If the goal is to compare a minimal number of systems using shallow measurements, where re-usability of the test collection is not important, such a strategy could also be adopted by research groups. For example, provided with enough resources, new relevance assessments could be generated for the collection adopting an efficient strategy such as that outlined by Carterette et al. (2006). Alternatively, comparisons could be made across different test collections where conditions remain even. This step assumes such collection(s) exist, although collections can be evaluated for such properties using the suite of Assessment measures. Finally, this reinforces the need when building test collections to include novel systems in the pooling process.

4 Experimental Analysis

To demonstrate the application of the Assessment Precision measure within the evaluation process we conducted an empirical analysis to evaluate both its utility, and to provide further justification for its introduction. Our first objective was to examine the officially submitted runs to TREC over a number of collections, spanning a range of years[16]. By using the official runs we can investigate the level of uncertainty during performance comparisons of runs included in the system pool across these collections. Our second objective was to evaluate the implications of measurement depth with respect to the level of assessment between systems at various points. Using the assessment metrics, we were investigating what effect using a measurement depth deeper than the pool depth may have on system comparisons. This is related to the argument that relative system performance can be compared if the conditions remain even for both systems. We then focus our attention on runs from particular collections, such as the Robust-HARD 2005[17], which has been identified as potentially problematic to use (Buckley et al., 2006). The aim is to better understand the problems cited with this collection. Finally, we investigate the utility of adopting the assessment metrics for explaining other phenomena in Information Retrieval such as automatic relevance assessments (Soboroff et al., 2001).

4.1 Collections

We used the official submitted runs to TREC covering a number of collections. The collections vary in size and year including the ad hoc retrieval collections TREC-2 (1993) and TREC-3 (1994) that were originally examined by Zobel (1998). We also use collections that have also been investigated in other related studies such as TREC-6, TREC-8 and Robust-HARD

[16]See http://trec.nist.gov/results.html

[17]This collection combined runs from Robust 2005 and HARD 2005 to form the pool.

2005(Voorhees, 2001; Soboroff et al., 2001; Buckley and Voorhees, 2004; Buckley et al., 2006). To complement these collections, we also examined the Robust 2003 collection that contained 250 topics, and the large Terabyte 2004 collection. For each collection we consider all official runs submitted to TREC, including the runs that formed the system pools as well as the remaining official runs submitted by each group. These remaining runs were not included in the system pools but were however evaluated. As a result, a number of these runs will be incomplete even below the pooling depth.

4.2 Decision Matrix Experiment

For each collection we first analysed each possible pair-wise system comparison of the officially submitted runs using the decision matrix (see Tables 3 and 4). To test for significance across all systems we used the ANOVA test. If significant differences in terms of performance and assessment across the systems of a collection were found, we performed a followup Bonferroni multiple comparisons to identify which systems differed significantly both in terms of performance and assessment. We repeated this experiment across numerous Performance and Assessment Precision metrics, spanning a range of measurement depths; including the Performance metrics P@d – Precision at rank d – and MAP@d – Mean Average Precision at rank d – as well as the corresponding Assessment Precision metric at the same depth A_p@d.

The reason for examining two types of performance metric is that P@d metrics are commonly used at shallow measurement depths, however, using metrics at larger depths can be misleading. Precision is essentially the proportion of relevant documents in the ranked list, it does not account for the placement of relevant documents in this list. For example, two systems could share a similar score of 0.5 for P@100. The first system ranks 50 relevant documents at positions 1-50. The other system, however, ranks the 50 relevant documents at 51-100. Hence, there is an implicit assumption that the searcher will examine all documents down to the measurement depth, although how realistic this assumption is at large measurement depths is questionable. Also, as the measurement depth becomes larger than the total number of known relevant documents for a topic, then the maximum possible precision score decreases.

Therefore, MAP is more informative at a deeper measurement depth as it accounts for the accuracy of where the corresponding relevant documents are positioned in the result list, providing more discrimination between systems, as systems are rewarded for ranking relevant documents highly. MAP also has another desirable property related to recall, where it accounts for the number of discovered relevant documents not included in a runs ranked list in the estimation. Therefore, we provide a summary of these comparisons at measurement depth $d = \{10, 30, 100, 500, 1000\}$.

For each metric, we counted the number of comparisons that fell into each outcome i.e. Cases $1 - 4$ (see Figure 3 for an outline of each case). Tables 3 and 4 present the proportion of overall system comparisons that fall into each case for Precision and MAP respectively. Rows indicate different test collections while columns represent different measures, increasing by measurement depth. Each entry represents the proportion of system comparisons that fall into that case e.g. for the TREC 3 @10 entry, 69% of pair-wise system comparisons fall in Case 1, 7% in Case 2, 17% in Case 3 and 7% in Case 4, where there were 40 runs included in as part of the pair-wise comparisons overall.

4.2.1 Results

	@10		@30		@100		@500		@1000	
TREC 3	**0.69**	0.07	**0.67**	0.14	**0.58**	0.32	**0.44**	0.49	**0.41**	0.57
40 runs	**0.17**	0.07	**0.11**	0.09	**0.05**	0.05	**0.00**	0.06	**0.00**	0.02
TREC 4	**0.72**	0.12	**0.65**	0.16	**0.43**	0.39	**0.35**	0.48	**0.33**	0.5
33 runs	**0.16**	0.00	**0.19**	0.00	**0.05**	0.13	**0.02**	0.15	**0.01**	0.16
TREC 6	**0.60**	0.16	**0.55**	0.26	**0.45**	0.42	**0.45**	0.44	**0.44**	0.45
74 runs	**0.13**	0.11	**0.10**	0.10	**0.03**	0.10	**0.01**	0.10	**0.00**	0.10
TREC 8	**0.70**	0.04	**0.69**	0.07	**0.52**	0.31	**0.51**	0.36	**0.57**	0.33
125 runs	**0.19**	0.06	**0.17**	0.07	**0.01**	0.16	**0.00**	0.13	**0.00**	0.10
ROBUST 2003	**0.83**	0.01	**0.74**	0.09	**0.55**	0.2	**0.52**	0.22	**0.57**	0.17
66 runs	**0.03**	0.13	**0.01**	0.16	**0.00**	0.25	**0.00**	0.26	**0.00**	0.27
Terabyte 2004	**0.63**	0.17	**0.54**	0.19	**0.47**	0.23	**0.35**	0.35	**0.34**	0.40
70 runs	**0.15**	0.06	**0.16**	0.11	**0.14**	0.16	**0.00**	0.30	**0.00**	0.26
ROBUST 2005	**0.77**	0.12	**0.66**	0.23	**0.53**	0.38	**0.57**	0.35	**0.6**	0.31
55 runs	**0.07**	0.03	**0.07**	0.04	**0.00**	0.10	**0.00**	0.08	**0.00**	0.08
HARD 2005	**0.71**	0.20	**0.66**	0.25	**0.57**	0.39	**0.57**	0.42	**0.57**	0.42
75 runs	**0.05**	0.05	**0.05**	0.04	**0.00**	0.04	**0.00**	0.02	**0.00**	0.01

Table 3: The proportion of cases in the System Evaluation Decision matrix for various TREC Test collections using Precision as a performance metric. Numbers in bold font indicate a "strong" comparison while numbers in a normal font indicates a "weak" comparison.

	@10		@30		@100		@500		@1000	
TREC3	**0.82**	0.11	**0.73**	0.17	**0.56**	0.30	**0.39**	0.38	**0.36**	0.40
40 runs	**0.04**	0.03	**0.05**	0.05	**0.07**	0.08	**0.05**	0.17	**0.06**	0.18
TREC4	**0.86**	0.12	**0.74**	0.17	**0.43**	0.43	**0.31**	0.44	**0.28**	0.46
33 runs	**0.02**	0.00	**0.09**	0.00	**0.05**	0.09	**0.07**	0.18	**0.07**	0.19
TREC6	**0.70**	0.25	**0.59**	0.31	**0.42**	0.42	**0.39**	0.41	**0.39**	0.40
74 runs	**0.03**	0.02	**0.05**	0.05	**0.06**	0.09	**0.07**	0.13	**0.07**	0.14
TREC8	**0.88**	0.10	**0.82**	0.13	**0.51**	0.35	**0.49**	0.31	**0.55**	0.24
125 runs	**0.01**	0.01	**0.03**	0.14	**0.03**	0.10	**0.04**	0.16	**0.04**	0.17
ROBUST 2003	**0.85**	0.09	**0.74**	0.16	**0.55**	0.31	**0.52**	0.30	**0.56**	0.24
66 runs	**0.01**	0.04	**0.01**	0.09	**0.00**	0.13	**0.00**	0.18	**0.00**	0.19
Terabyte 2004	**0.73**	0.21	**0.60**	0.24	**0.47**	0.24	**0.34**	0.25	**0.33**	0.24
70 runs	**0.05**	0.01	**0.10**	0.07	**0.14**	0.15	**0.01**	0.41	**0.01**	0.42
ROBUST 2005	**0.81**	0.15	**0.69**	0.25	**0.52**	0.38	**0.55**	0.31	**0.58**	0.27
55 runs	**0.03**	0.01	**0.04**	0.02	**0.01**	0.09	**0.02**	0.12	**0.02**	0.13
HARD 2005	**0.74**	0.23	**0.69**	0.26	**0.55**	0.37	**0.54**	0.36	**0.55**	0.35
75 runs	**0.02**	0.02	**0.02**	0.03	**0.02**	0.06	**0.03**	0.08	**0.03**	0.08

Table 4: The proportion of cases in the System Evaluation Decision matrix for various TREC Test collections using MAP as a performance metric.

The first thing we investigated was whether there was a common trend in the proportion of significant pair-wise differences, in terms of system performance, as the measurement depth increased across the various test collections. The reasoning behind the practice of increasing measurement depth is that the discrimination between systems will be greater. The intuition being that good systems will continue to retrieve relevant documents beyond the pooling depth that will have been discovered back by other runs included in the system pool.

A noticeable trend from the results of using P@d as a metric in Table 3 was the proportion of significant differences which fell into the "strong" Case 3 either decreased or remained approximately constant as the measurement depth increased. For example, across the TREC-3 collection, the proportion of significant differences dropped from 0.17 using P@10 to 0.11 when using P@30. At P@100, 0.05 comparisons feel into this case. When using MAP as a metric, Table 4, the proportion of comparisons that fell into Case 3 remained approximately constant, between $0.04 - 0.07$. This trend was consistent across other collections such as TREC-8, Robust 03, Robust 05, and HARD 05. for the Terabyte 04 track, the proportion of comparisons that fell into Case 3 increased as the measurement depth tended towards the pooling depth at approximately $\lambda = 100$, however, after the pooling depth the proportion of significant differences belonging to this case decreased.

However, from these results it would appear that discrimination between the set of systems at best remains constant and at worse lessens as the measurement depth increases. From some collections such as Trec-3, 6, 8 and the Terabyte 2004 collections this becomes more stated as measurement depth is increased beyond pool depth.

A similar trend is also followed for system comparisons where the null hypothesis that both systems have equal performance cannot be rejected. As measurement depth increases, the proportion of "strong" cases decreases, resulting in a larger proportion of cases where one system has a significantly larger number of judged documents than another.

We then examined the proportion of significant differences between systems that fall into either the "strong" or "weak" case. Overall, a common trend across collections was that, as measurement depth increased, the proportion of "strong" comparisons decreased while the proportion of "weak" cases increased. To illustrate, consider first P@10 for the TREC-3 collection in Table 3. We find a smaller proportion of comparisons falling into the 'strong" case in contrast to P@30 (0.17 to 0.11). Conversely there is an increase in "weak" cases from 0.07 to 0.087. This trend remains as we continue increasing measurement depth towards P@1000. As the measurement depth increased beyond the pooling depth for many collections, a common pattern was for a comparison to change from a "strong" to a "weak" case. *Increasing measurement depth results in a higher degree of uncertainty when comparing systems.*

4.2.2 Why the trends?

We then investigated further the effect of increasing the measurement depth had on the proportion of significant differences falling into Cases 3 and 4. To attempt to answer why this trend of decreasing significance between runs and increased uncertainty occurs, we first examined performance across increasing measurement depth across all collections. We present an summarised illustration of this experiment in Figures 5, 6, 7, 8, 9 and 10, which display system performance of all official runs submitted to the TREC-3, Robust 2003 and 2005 collections. We first ranked all runs from high to low performance with respect to MAP@10. To easily identify the change in a system performance across various metrics, this rank remains constant across the remaining plots. We also plotted system scores using the Assessment metrics A_p. For each plot, we report the mean performance and standard error across the topic set, adjusted for multiple comparisons, for measurement depth $d = \{10, 30, 100, 500, 1000\}$. Significance can be found where there are two disjoint intervals.

For the TREC 3 collection (see Figures 5 and 6), as measurement depth is increased the difference between system runs becomes less emphasised for the majority of systems. In particular those runs ranked towards the middle begin to report similar performance, converging on each other. The plot of MAP@1000 illustrates that there are approximately three groups, two small groups at the extreme and a large section of average performing systems. We tend to find discrimination between systems from both these small groups. It is noticeable that the rank order of runs from MAP@10 to MAP@1000 deviates slightly, with a small number of swaps, which is in accordance to the results found in (Voorhees, 2000).

When examining the range of A() assessment metrics, plotted on the right-hand column, we find that the coverage of assessed documents for each systems begins to vary as measurement depth increases. Initially, as measurement depth is above the pooling depth, the majority of system runs have similar levels of assessment. As the depth increases, the variance in assessment also increases, which would explain the trends we found in Tables 3 and 4. As the measurement depth goes beyond the pool depth, the number of significant differences between systems increases, resulting in higher discrimination (between runs). This would explain why there is a large increase in system comparison in the TREC-3 collection that swap from "strong" to "weak" while measurement depth increases.

In general, runs with relatively poor performance also have a lower level of assessed documents. Systems with middle ranking performance tend to have the largest coverage of assessed documents. The best two performing systems retrieving an average proportion of assessed documents. This would indicate that the best systems retrieve a larger proportion of assessed and relevant documents ($R \cap X \cap A$) compared to those with systems with equal levels of assessed documents but poorer performance, who tend to have a larger proportion of assessed non-relevant documents ($\neg R \cap X \cap A$).

For the Robust 2003 collection (see Figures 7 and 8), only a small proportion of significant differences between runs was found overall. Most of these comparisons were the result of a small group of systems with poor performance across all metrics. However, there is a pack of systems whose performance deteriorates as measurement depth increases. Examining the Assessment metrics explains why this occurs. This small group of runs only returned a partial set of documents ranging from 10 to 50 documents, instead of the 1000 documents allowed per topic in TREC. Therefore as the measurement depth increased the performance of these systems becomes frozen. As a consequence, the number of systems that significantly improved over this group of runs increased as the measurement depth increased. This analysis explains why for the Robust 2003 collection, an increase in significant differences was observed. It is important to note, however, that this increase in significant differences between the "frozen" systems and other runs fell into the category of "weak" comparison (i.e. Case 4).

We also highlight the results from the Robust 2005 collection that has reported problems with "title bias" (Figures 9 and 10). Again, as the measurement depth increases the discrimination of systems at the extremes remain but at the expense of discrimination between systems towards the middle. This is a similar trend to the other collections, where the performance between systems levels out as measurement depth increases with the majority of significant differences occurring between these systems at the extreme ends of performance. However, for this collection it is important to indicate that the system with the best performance at depths 10 and 30 may have underestimated performance at greater depths (i.e. run 1). The reasoning behind this is that the level of assessment at A_p@100 onwards for this system is one of the poorest in comparison to the other systems. Other examples include System 9, which becomes the best performing system as measurement depth increases, however, the corresponding Assessment score is also higher than average for this system – significantly more so than Systems 1 to 5.

Overall, these trends appeared to be consistent across the remaining collections.

4.2.3 What is causing the swaps in the grid?

We then examined in closer detail what conditions would result in a swap from a "strong" to "weak" comparison and vice versa when increasing measurement depth, focusing in particular on pair-wise comparisons of individual runs. As a case study we present a comparison of runs from the potentially problematic Robust 2005 collection. In Figure 11, we illustrate two examples of system comparisons where there is a swap from "strong" to "weak" comparisons as measurement depth increases. In both plots we display both the MAP@ (left) and A_p@

(right) metrics at various measurement depths for both systems. Error bars are displayed to show variation across the set of topics and significance between systems. Runs are referred to by their official TREC run-tag.

An analysis by Buckley et al. (2006) highlighted a potential bias towards documents containing the topic title keywords in this collection, although relevant documents exist that do not contain these words. For example, the sab05ror1 run (Buckley, 2005), using existing relevance assessments from another collection to generate an optimal query, reported a large change in performance when removing unique documents pooled by the system from the relevance assessments. This was a result of the sab05ror1 run discovering a large number of unique relevant documents during the pooling process.

In Figure 11 (top), we compare this system against another run, pircRB05td3 (Kwok et al., 2005), which used an external corpora of TREC newswire documents to expand an original query based on the title and description fields of the topic. Up to a measurement depth of 100, sab05ror1 has reported a higher MAP score than pircRB05td3. Beyond this depth, there was a swap in performance where pircRB05td3 reports a higher MAP. Examining the corresponding Assessment score, we find comparable levels of assessment until a depth of a 100, where the pooling depth was set at 55 for this collection. Beyond this depth, pircRB05td3 has a significantly larger proportion of documents assessed than sab05ror1. This is an example of a Case 1 comparison at depths 10 and 30, that then moves towards a Case 2 comparison after the pooling depth (see Figure 3).

This reflects the findings by Buckley et al. (2006), where the performance of s_2 is underestimated once a larger measurement depth is used. From the study of Zobel (1998), who investigated the rate of discovering new relevant documents beyond the pool depth, it is uncertain if both systems shared similar levels of assessed documents that performance would have swapped or not. However, as raised by Buckley et al. (2006), if sab05ror1 was not included as part of the pooling process, it would be highly likely that its performance would have been underestimated even more so.

In Figure 11 (bottom), we illustrate another similar example comparison from this collection. The run indri05dmmt used a combination of a term dependence model and a mixture of relevance models trained using a superset of TREC newswire articles (Metzler et al., 2005), while uic0501 expanded the original title query with using an online semantic lexicon Wordnet (Liu and Yu, 2005). The performance of uic0501 was on average higher than indri05dmmt until a depth of 100 was reached, then again there was a swap between the two runs. From the plots of assessment, this swap in performance also coincided with a significant drop in the proportion of assessed documents for the uic0501 run.

4.2.4 Automatic Relevance Assessments

We then examined another use of assessment to explain a phenomena reported by Soboroff et al. (2001). The motivation of this study was to attempt to produce automatically generated relevance assessments based on the pooling process. This would be advantageous for producing inexpensive assessments for testing systems over very large corpora such as the Web. By randomly sampling the document pool, automatic relevance assessments were generated. Documents in this sample (of the pool) were considered relevant. Using these pseudo relevance assessments it was then possible to accurately rank a large proportion of the systems. This ranking, using the automatically generated pseudo relevance judgements, correlated strongly to that of the actual judgements defined by human assessors. The only runs that were not ranked accurately tended to be found at the extremes (i.e. very good or poor systems).

In a follow up study by Aslam and Savell (2003) it was identified that the problem of ranking those systems towards the extremes could be a problem of "popularity". In other words, the pseudo relevance assessments were ranking how similar systems were to each other, where the better systems were retrieving slightly different documents to this middle section, hence the difficulty in accurately ranking them. Popularity was a measure of the set of retrieved documents from each system i.e. comparing the ranked lists of the set of systems $S = \{s_1, \ldots, s_S\}$ using the corresponding similarity of the retrieved set of documents from each system $X = \{X^{s_1}, \ldots, X^{s_S}\}$.

We believe this result can be explained further by analysing the assessment level of each run. In other words, the latent variable that is being used to rank the "popular" systems accurately is the set of assessed documents in X for a system i.e. for system s_1 the proportion of assessed documents in the ranked list $X^{s_1} \cap A$.

To test this assumption we ranked all runs for the TREC-3, TREC-6, TREC-7 and TREC-8 collections with respect to MAP performance in Figure 12, from high MAP to low. For each collection, we also plotted each systems corresponding A_p@1000 value.

From the plots, we found that it was those systems at the extremes that do not contribute to the assessment pools in comparison to the middle section of runs. Also, the better systems appear to be good at discovering relevant assessed documents ($R \cap X \cap A$) while minimising the proportion of non-relevant assessed documents in the ranked list ($\neg R \cap X \cap A$). While the poorer systems retrieved more non-relevant assessed and unassessed documents ($X \cap \neg A$).

We also calculated the Kendal tau correlation in system rankings between MAP and A@1000. This measures the concordance in the ranking of all systems in terms of Performance and Assessment metrics. A high correlation close to one would indicate high agreement in the system rankings, while a score of zero would indicate no relationship. The reported correlation for TREC-3 was 0.394, TREC-6 was 0.402, TREC-7 was 0.402, and TREC-8 was 0.462, corresponding to those of the ranked lists of pseudo relevance judgements found by Soboroff et al. (2001), and the ranking by popularity of Aslam and Savell (2003).

Therefore, those systems that were ranked accurately by the pseudo relevance assessments contribute most to the pool of judged documents. In other words they are "popular" due to both the theoretical similarity of the systems (i.e. generic systems) but also they are popular due to the contribution to the set of assessed documents A. Hence, the reasoning why those systems not accurately ranked was due the lower proportion of assessed documents A which they contribute. In effect, the pseudo judgements were generated using an IR system (the random sample of the pool A). Therefore, those systems with the highest coverage of assessed documents (the average or "popular" systems) correlated higher to this new IR system, while the systems at the extremes did not correlate as strongly, and thus, were not ranked as accurately.

5 Discussion

By proposing the System Evaluation Decision matrix we are essentially asking the question – are the conditions for both systems even? This relates to the initial defence of the current IR evaluation framework, where the relative performances of systems can be fairly compared if the experimental conditions remain even for all (Salton, 1992). Using the level of assessment metrics, A(), we can determine whether there is a significant difference in the proportion of documents assessed between two competing systems. In other words, the level of information used to estimate performance for both systems. If there is a significant difference (between systems), then this introduces some degree of uncertainty into the comparison. With such a scenario, the performance of one system may be underestimated compared to the other. At this stage, it would be beneficial to perform further evaluation of the system at different measurements depths, on different collections, or even on acquiring further relevance judgments. This is the advantage of using System Evaluation Decision matrix. Such scenarios can be detected, informing the researcher of potential uncertainty when comparing systems using incomplete judgements so that they can then act accordingly.

From the results of this study, it would appear that the use of a measurement depth larger than the pooling depth is questionable, and re-iterates the concerns raised by Zobel (1998). As the measurement depth increases beyond the pooling depth, uncertainty across many system comparisons also increases. Interestingly, the discrimination between systems becomes less. This supports the work by Sanderson and Zobel (2005), where a measure such as P@10 can be used to discriminate between systems, placing more emphasis on a larger set of test topics.

A potential explanation why there is a decrease in discrimination at lower measurement depth is because performance estimates vary more across topics, there is also less information to estimate performance (i.e. Assessment), resulting in wider confidence intervals. Also the majority of systems, what Aslam and Savell (2003) refer to as the "popular" systems, appear to discover a similar proportion of relevant documents once the measurement depth increases. The best systems, even though the relative performance is high, and they rank highly compared to other systems, performance is still underestimated because these systems return more unique documents than the "popular" set of systems. Potentially, these better systems discover more relevant unassessed documents ($R \cap X \cap \neg A$) that are not accounted for because of the necessity of fixing the pool depth. If this is the case, then it would warrant further investigation into alternative pooling strategies (Cormack et al., 1998; Zobel, 1998; Robertson and Soboroff, 2003; Sanderson and Joho, 2004).

Prior to the pooled documents being judged for relevance by assessors, the suite of assessment metrics could be deployed for assessing any irregularities such as the title bias found in the Robust-Hard 2005 collection. For example, official runs could be evaluated with respect to the level of Assessment at various ranks. Any noticeable drop off after the pooling depth could be identified and correct accordingly before documents were finally assessed, ensuring parity in conditions across all systems. It could also be envisaged that different theoretical and cognitive models, not included in the pool, be supplemented to ensure the reusability of the collection. The principle of polyrepresentation provides such a framework (Ingwersen and Järvelin, 2005), where a diverse pool of systems should reflect the different algorithmic and cognitive interpretations of the documents stored within a collection.

Finally, the relative rankings of systems will remain stable across measurement depths and varying levels of incompleteness in general. However, we believe it is the cases where systems jump in ranking because of the effects of incompleteness, in particular relating back to Case 2 in the decision matrix, where interesting cases are. It is those systems that are novel or diverse that require more consideration during Laboratory IR evaluation, because it is these systems that retrieve the larger number of unique, and potentially relevant documents. For instance, a recent study has shown the utility of defining ranking algorithms designed to provide more topical variation in a results list (Chen and Karger, 2006). This may result in the performance of such algorithms being underestimated. This is one area where graded relevance assessments and the corresponding metrics may have a greater importance in laboratory IR evaluation (Järvelin and Kekäläinen, 2002), when used alongside the System Evaluation Decision matrix and the corresponding Assessment metrics.

6 Conclusions

In this paper we argued that uncertainty when using incomplete relevance assessments should be identified during the evaluation process. Consequently, we proposed a new set of metrics based on of the level of assessment that provide an indication of uncertainty during system comparisons. If the level of assessment between systems is similar, we believe that a fair comparison can be made, otherwise uncertainty has been introduced into the evaluation process. By using the System Evaluation Decision matrix we can make stronger claims of significance (or not), and guide subsequent research to decide when further testing is required. Finally, as part of an empirical study that adopted this methodology, we provided supporting evidence which questions the practice of using a measurement depth that exceed the pooling depth. Future work will investigate this implication in greater detail, along with the relationship between Assessment and Precision.

References

Aslam, J. A., Savell, R., 2003. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In: SIGIR '03: Proceedings of the 26th ACM SIGIR. Toronto, Canada, pp. 361–362.

Barry, C. L., 1994. User-defined relevance criteria: an exploratory study. J. Am. Soc. Inf. Sci. (JASIS) 45 (3), 149–159.

Blair, D. C., 1996. Stairs redux: thoughts on the stairs evaluation, ten years after. J. Am. Soc. Inf. Sci. 47 (1), 4–22.

Blair, D. C., 2002a. The challenge of commercial document retrieval, part i: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size. Inf. Process. Manage. 38 (2), 273–291.

Blair, D. C., 2002b. Some thoughts on the reported results of trec. Inf. Process. Manage. 38 (3), 445–451.

Blair, D. C., Maron, M. E., 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. Commun. ACM 28 (3), 289–299.

Buckley, C., 2005. Looking at limits and tradeoffs: Sabir research at trec 2005. In: Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005) Notebook. NIST.

Buckley, C., Dimmick, D., Soboroff, I., Voorhees, E., 2006. Bias and the limits of pooling. In: SIGIR '06: Proceedings of the 29th ACM SIGIR. Seattle, WA, pp. 619–620.

Buckley, C., Voorhees, E. M., 2004. Retrieval evaluation with incomplete information. In: Proceedings of the 27th ACM SIGIR Conference. pp. 25–32.

Carterette, B., Allan, J., Sitaraman, R., 2006. Minimal test collections for retrieval evaluation. In: SIGIR '06: Proceedings of the 29th ACM SIGIR. Seattle, WA, pp. 268–275.

Chen, H., Karger, D. R., 2006. Less is more: probabilistic models for retrieving fewer relevant documents. In: SIGIR '06: Proceedings of the 29th ACM SIGIR. Seattle, WA, pp. 429–436.

Cormack, G. V., Palmer, C. R., Clarke, C. L. A., 1998. Efficient construction of large test collections. In: Proceedings of the 21st ACM SIGIR. pp. 282–289.

Harter, S. P., 1992. Psychological relevance and information science. J. Am. Soc. Inf. Sci. (JASIS) 43 (9), 602–615.

Hull, D., 1993. Using statistical testing in the evaluation of retrieval experiments. In: SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, New York, NY, USA, pp. 329–338.

Ingwersen, P., Järvelin, K., 2005. The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20 (4), 422–446.

Kwok, K., Grunfeld, L., Dinstl, N., Deng, P., 2005. Trec 2005 robust track experiments using pircs. In: Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005) Notebook. NIST.

Liu, S., Yu, C., 2005. Uic at trec 2005: Robust track. In: Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005) Notebook. NIST.

Metzler, D., Diaz, F., Strohman, T., Croft, W. B., 2005. Umass at robust 2005: Using mixtures of relevance models for query expansion. In: Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005) Notebook. NIST.

Popper, K., 2000. Conjectures and refutations. Readings in the Philosophy of Science, 9–13.

Robertson, S., Soboroff, I., 2003. The TREC 2002 Filtering Track Report. In: Proceedings of the Eleventh Text REtrieval Conference (TREC 2002).

Ruthven, I., 2005. Integrating approaches to relevance. In New Directions in Cognitive Information Retrieval 19, 61–80.

Salton, G., 1992. The state of retrieval system evaluation. Inf. Process. Manage. 28 (4), 441–449.

Sanderson, M., Joho, H., 2004. Forming test collections with no system pooling. In: Proceedings of the 27th ACM SIGIR conference. pp. 33–40.

Sanderson, M., Zobel, J., 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proceedings of the 28th ACM SIGIR Conference. pp. 162–169.

Saracevic, T., 1995. Evaluation of evaluation in information retrieval. In: SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, New York, NY, USA, pp. 138–146.

Savoy, J., 1997. Statistical inference in retrieval effectiveness evaluation. Inf. Process. Manage. 33 (4), 495–512.

Schamber, L., Eisenberg, M., Nilan, M. S., 1990. A re-examination of relevance: toward a dynamic, situational definition. Inf. Process. Manage. 26 (6), 755–776.

Soboroff, I., 2006. Dynamic test collections: measuring search effectiveness on the live web. In: SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, New York, NY, USA, pp. 276–283.

Soboroff, I., Nicholas, C., Cahan, P., 2001. Ranking retrieval systems without relevance judgments. In: Proceedings of the 24th ACM SIGIR Conference. pp. 66–73.

Spärck-Jones, K., Van Rijsbergen, C. J., 1975. Report on the need for and provision of an "ideal" information retrieval test collection. Tech. rep., British Library Research and Development Report 5266, University of Cambridge.

Sparck-Jones, K., Willett, P. (Eds.), 1997. Readings in information retrieval. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Swanson, D. R., 21 October 1960. Searching natural language text by computer. Science 132 (3434), 1099–1104.

Swanson, D. R., 1988. Historical note: Information retrieval and the future of an illusion. Journal of the American Society for Information Science 39 (2), 92–98.

Van Rijsbergen, C. J., London, 1979. Information Retrieval. Second Edition Butterworths.

Voorhees, E., Harman, D., 2003. Letters to the editor. Inf. Process. Manage. 39 (1), 153–156.

Voorhees, E. M., 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. Inf. Process. Manage. 36 (5), 697–716.

Voorhees, E. M., 2001. Evaluation by highly relevant documents. In: Proceedings of the 24th ACM SIGIR Conference. ACM, pp. 74–82.

Voorhees, E. M., 2002. The philosophy of information retrieval evaluation. In: CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems. Springer-Verlag, London, UK, pp. 355–370.

Voorhees, E. M., Harman, D. K. (Eds.), 2005. TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge, Massachusetts 02142.

Wallis, P., Thom, J. A., 1996. Relevance judgements for assessing recall. Information Processing and Management 32 (11), 273–286.

Zobel, J., 1998. How reliable are the results of large-scale information retrieval experiments? In: Proceedings of the 21st ACM SIGIR conference. pp. 307–314.

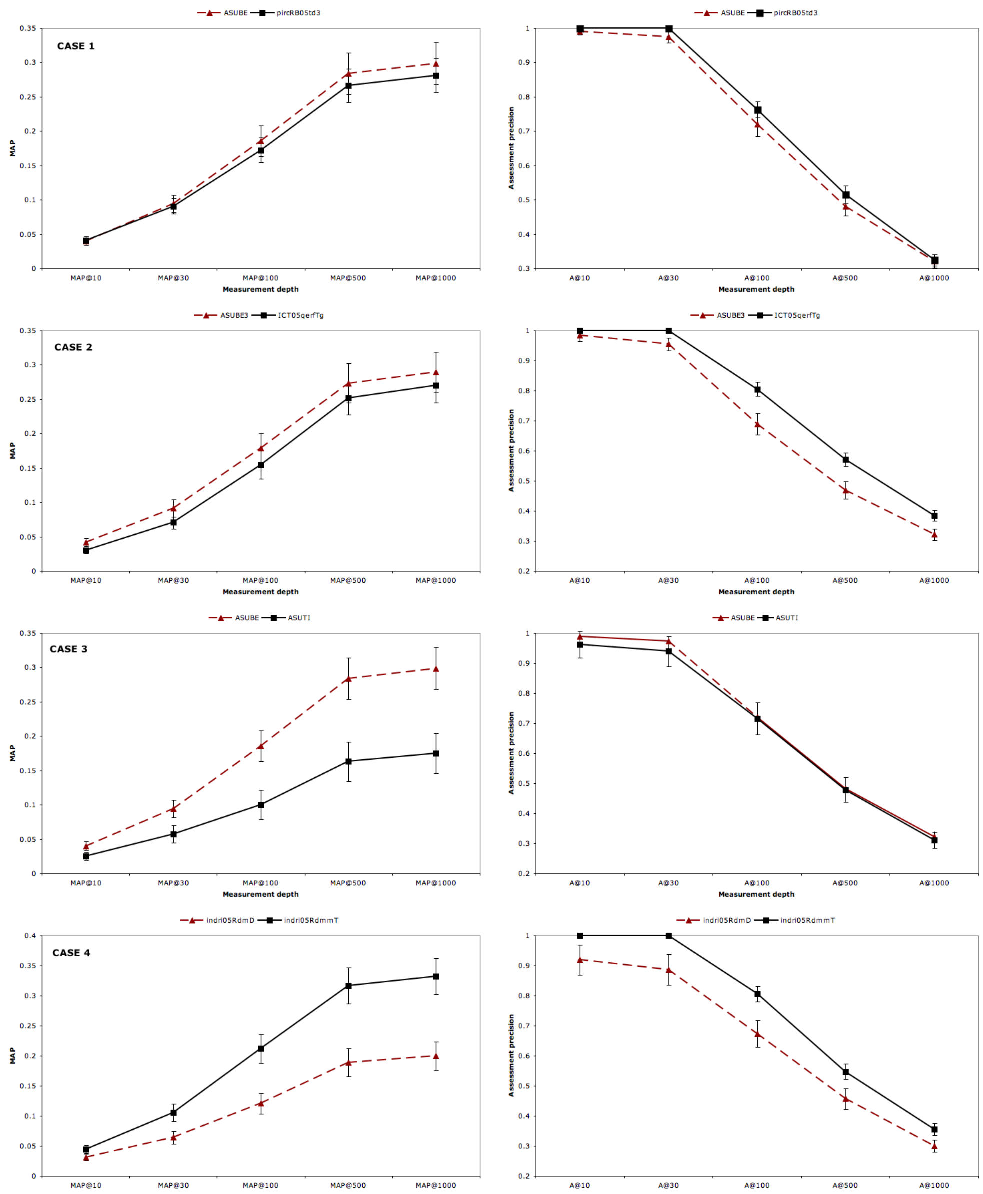

Figure 4: Example of a system comparison falling into each of the four cases. The results are taken from the Robust 2005 track.

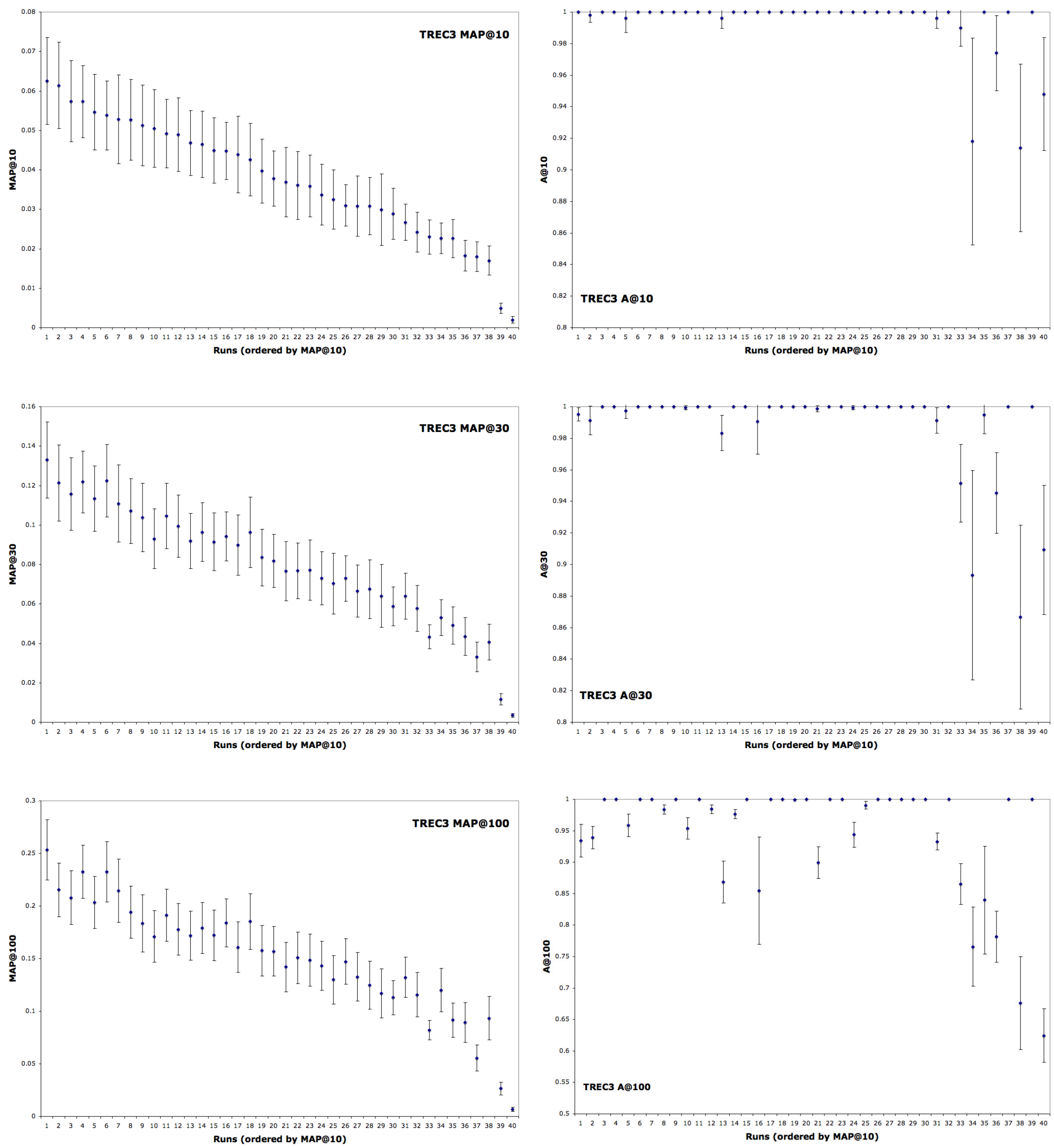

Figure 5: Trend in MAP and A_p for TREC-3, depths $d = \{10, 30, 100\}$

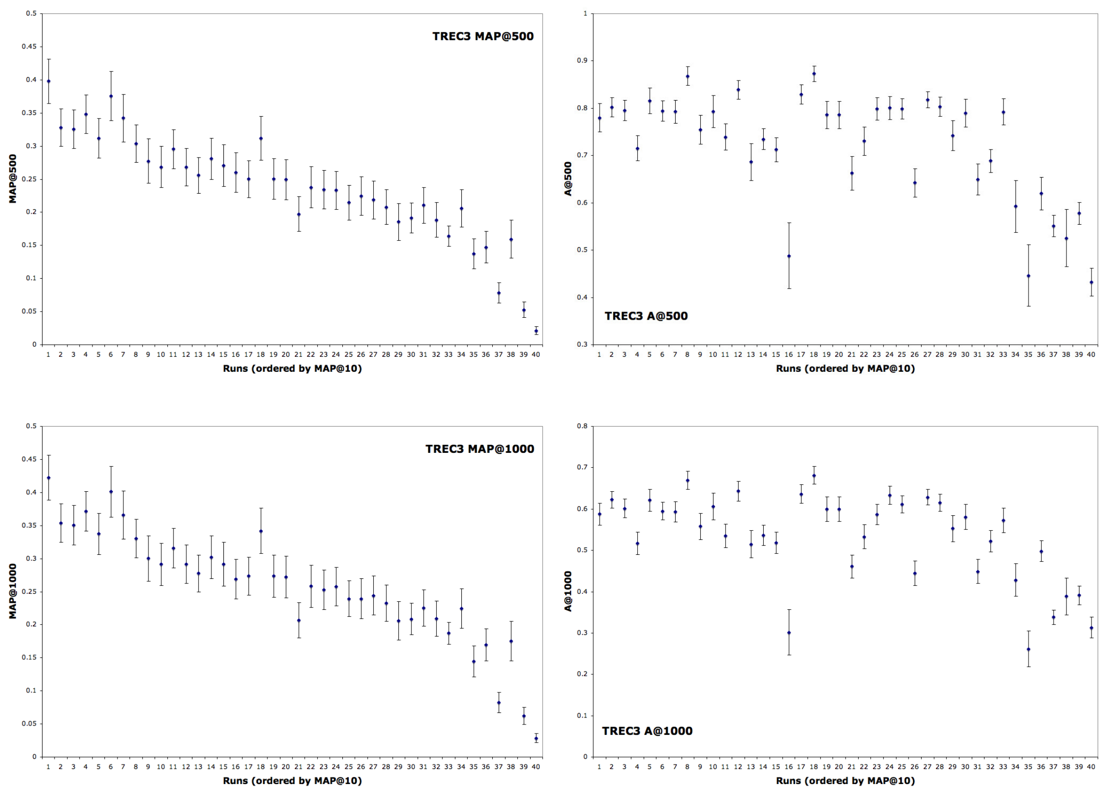

Figure 6: Trend in MAP and A_p for TREC-3, depths $d = \{500, 1000\}$

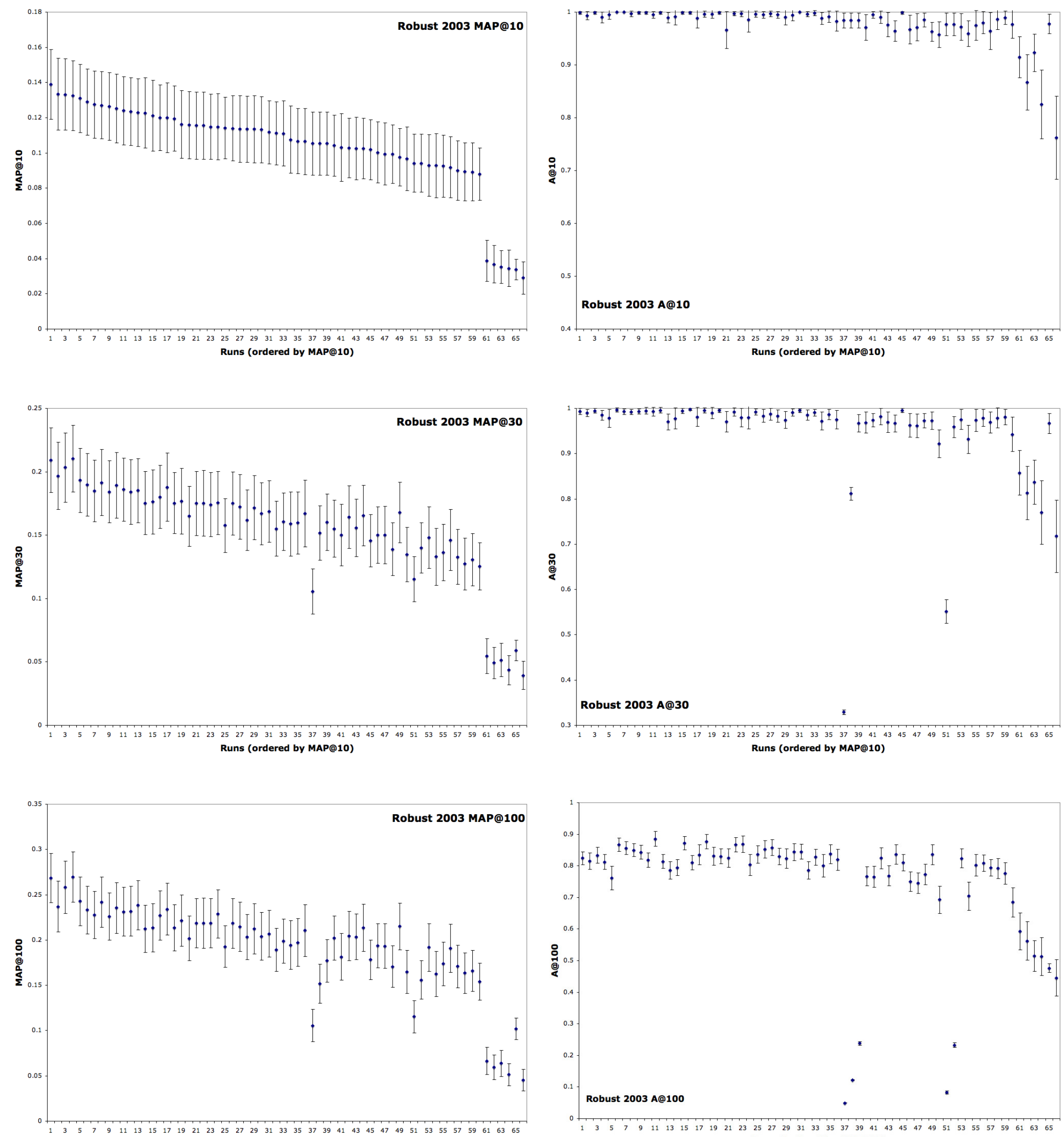

Figure 7: Trend in MAP and A_p for Robust 2003, depths d=$\{10, 30, 100\}$

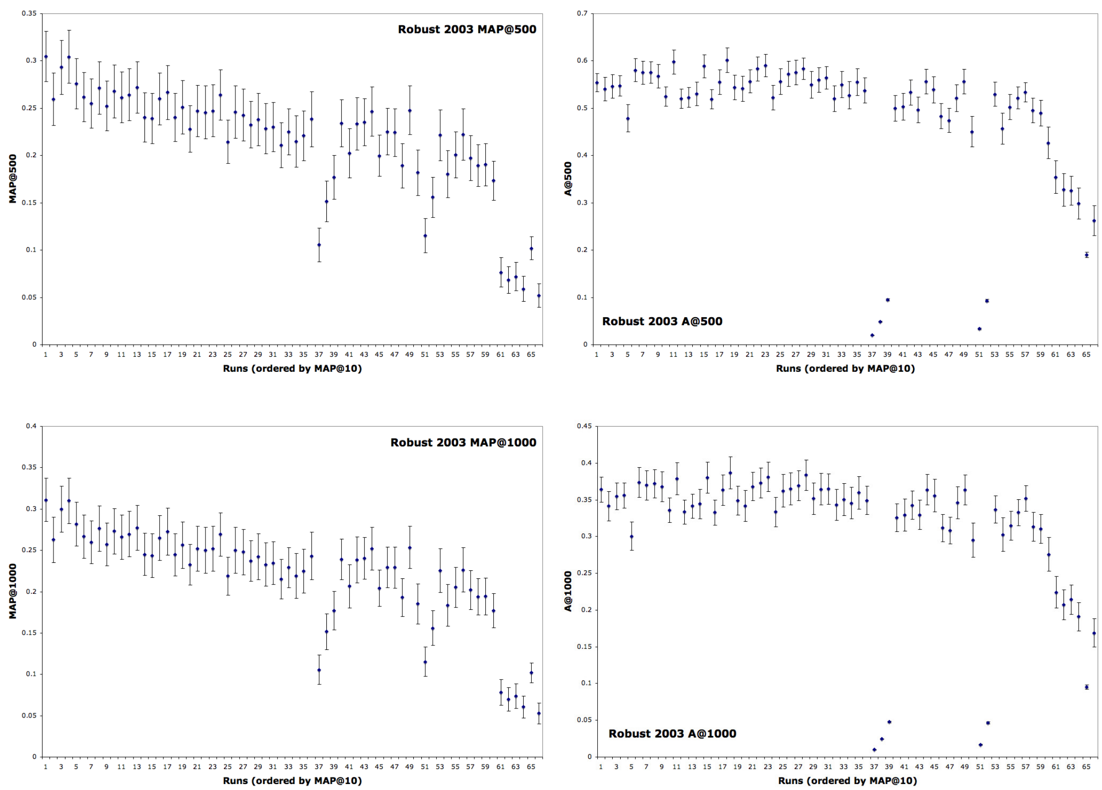

Figure 8: Trend in MAP and A_p for Robust 2003, depths d=$\{500, 1000\}$

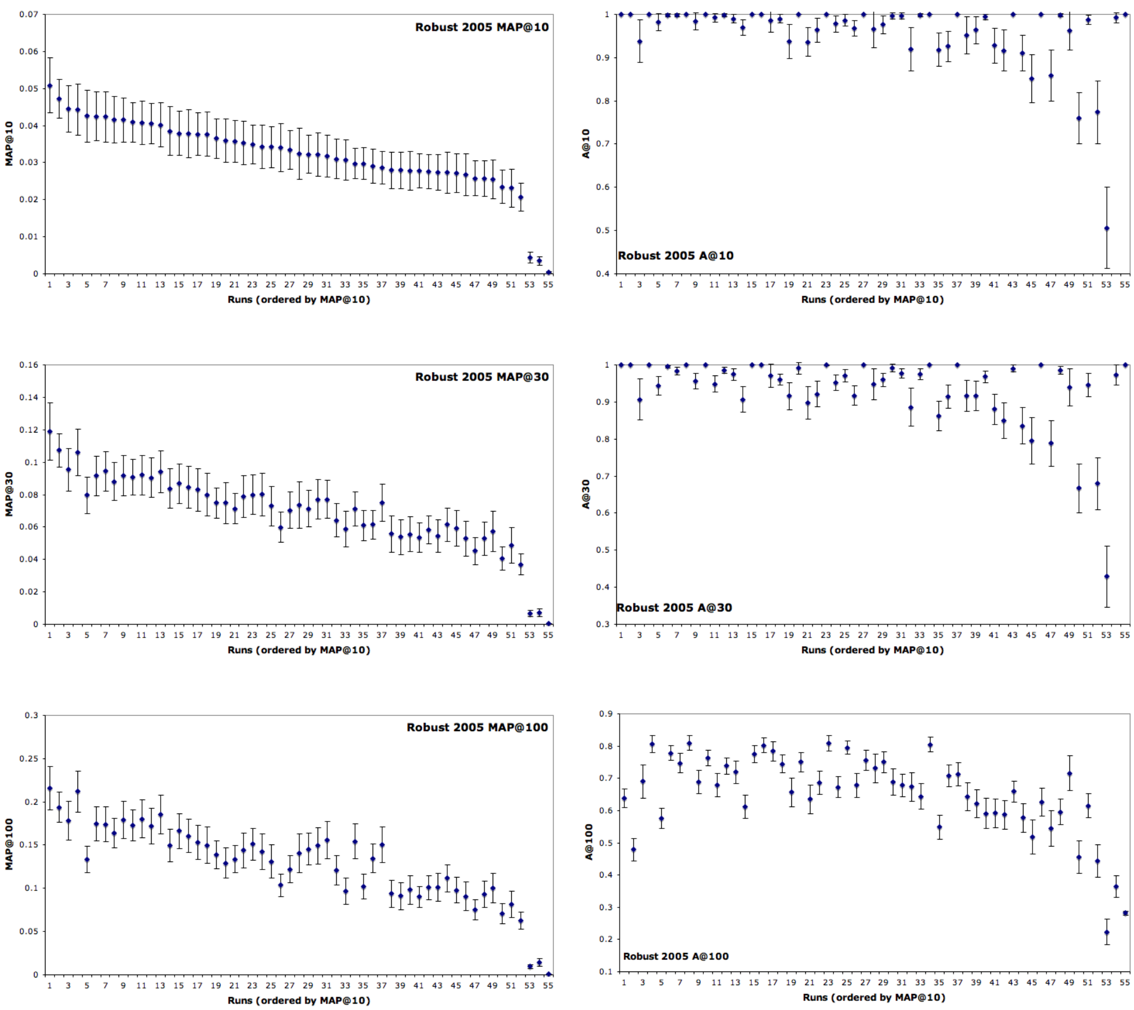

Figure 9: Trend in MAP and A_p for Robust 2005, depths d=$\{10, 30, 100\}$

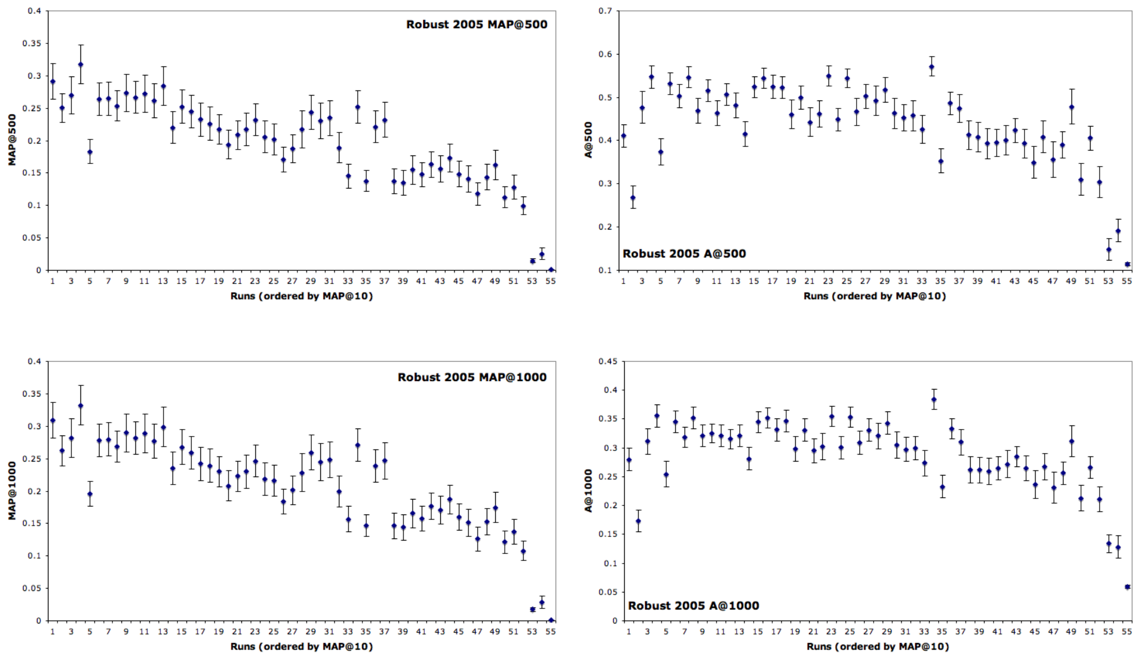

Figure 10: Trend in MAP and A_p for Robust 2005, depths d=$\{500, 1000\}$

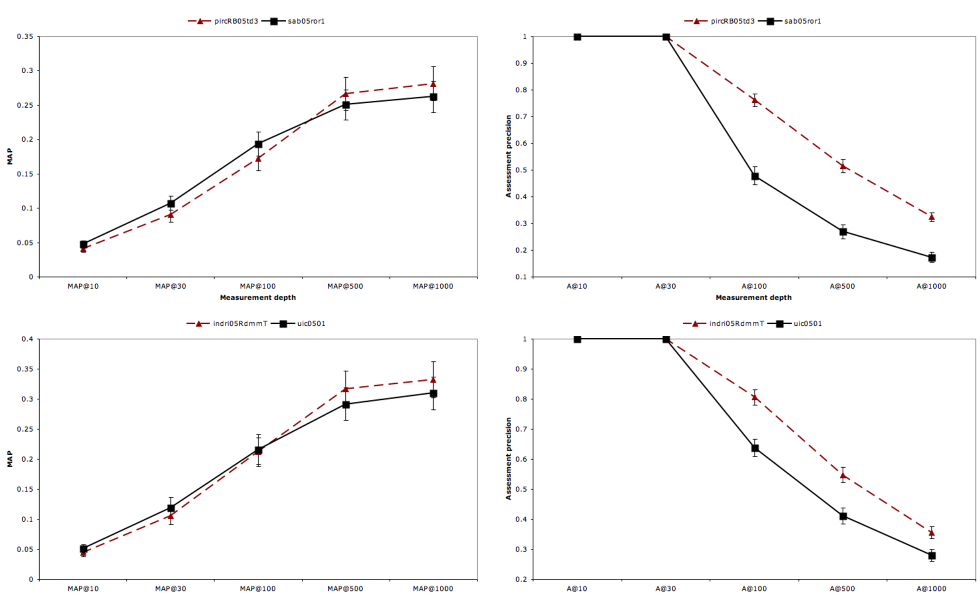

Figure 11: Example of two system comparisons from the Robust 2005 track.

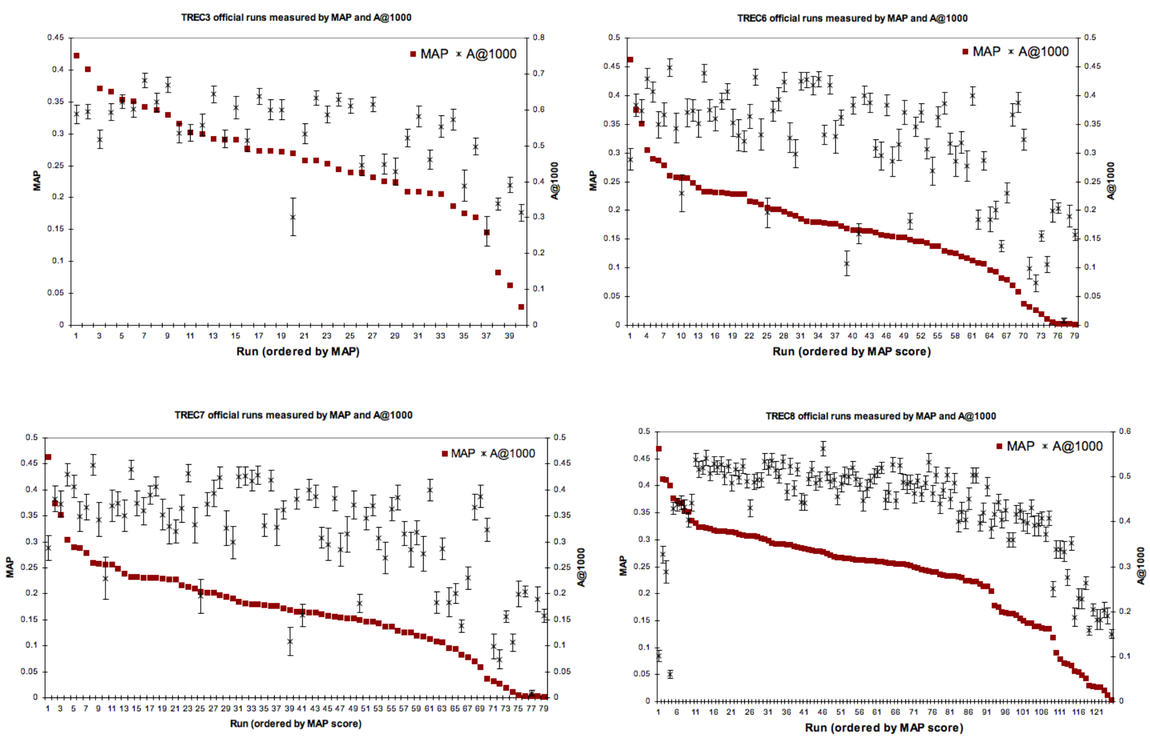

Figure 12: Systems ranked by MAP at measurement depth 1000 for the TREC-3, TREC-6, TREC-7, and TREC-8 collections. The corresponding A@1000 metric is also plotted for each system.