Baillie, M. and Azzopardi, L. and Ruthven, I. (2007) A retrieval evaluation methodology for incomplete relevance assessments. In: Advances in Information Retrieval: 29th European Conference on IR Research (ECIR 2007). Lecture Notes in Computer Science, 4425. Springer. ISBN 978-3-540-71494-1

http://eprints.cdlr.strath.ac.uk/3158/

A Retrieval Evaluation Methodology for Incomplete Relevance Assessments

Mark Baillie[1], Leif Azzopardi[2], and Ian Ruthven[1]

[1] Department of Computing and Information Sciences,
University of Strathclyde, Glasgow, UK
{mb, ir}@cis.strath.ac.uk
[2] Department of Computing Science,
University of Glasgow, Glasgow, UK
leif@dcs.gla.ac.uk

Abstract. In this paper we a propose an extended methodology for laboratory based Information Retrieval evaluation under incomplete relevance assessments. This new protocol aims to identify potential uncertainty during system comparison that may result from incompleteness. We demonstrate how this methodology can lead towards a finer grained analysis of systems. This is advantageous, because the detection of uncertainty during the evaluation process can guide and direct researchers when evaluating new systems over existing and future test collections.

1 Introduction

In this study we revisit the implications on system comparisons that arise from incomplete relevance assessments, and in particular the assumption that unassessed documents are not relevant. Instead of assuming unassessed documents are not relevant [1], or more recently, removing documents that are not judged when estimating performance [2], we propose an alternative approach: to quantify the proportion of unassessed documents in a system's ranked list. This leads to a complementary evaluation methodology, which uses a new set of measures that quantify uncertainty during system comparison. This methodology provides 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 this uncertainty. In this paper, 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 approach we first provide some context by reviewing the running debate on incompleteness, and the subsequent implications for the comparison of systems. We then introduce the proposed methodology which augments the current evaluation protocol (Section 2). Next, we provide an empirical analysis of this approach across a range of existing test collections (Section 3). Finally, we conclude with a discussion of the implications of this study (Section 4).

1.1 Background

Modern test collections adhere to a well defined model often referred to as the *Cranfield paradigm* [3]. A corpus that follows this model will consist of a collection of documents, statements of information need (named *topics*), and a set of relevance judgements listing the documents that should be returned for each topic. To ensure fair system comparisons, a number of key assumptions are made within this framework such as (i) the topics are independent, (ii) all documents are judged for relevance (completeness), (iii) these judgements are representative of the target user population, and (iv) each document is equally important in satisfying the users information need. 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 collection size and available (human) resources 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 such as system bias. For example, system pooling was proposed to address the intractability of the completeness assumption [4]. 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. Each system will rank the collection for a given topic, then the top λ documents from the subsequent ranked lists are collated[3], removing duplicates, to form a pool of unique documents. All documents in this pool are then judged for relevance by an assessor(s) using specified criteria such as topicality or utility. In the case of TREC [3], the assessor(s) use the topic statement for guidance when determining which documents are topically relevant or not. The remaining unassessed documents are assumed to be not 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" [1]. 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 pooled.

There has been much debate about the validity of this assumption. Initially this assumption was applied to smaller collections such as Cranfield and CACM. However, as Blair has posited [5], the percentage of documents not assessed for a given topic with respect to a modern collection could be up to 99%, leaving a very large proportion of unassessed relevant documents undiscovered. However, Voorhees and Harman highlight a key point [6], that as pooling is a union of many different ranking approaches and because only relative system performance

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

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 if all relevant documents were not discovered. 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 by Salton, 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.

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 this across collections such as TREC. Zobel defined a method to extrapolate the potential numbers of unassessed relevant documents, concluding that the assumption of unassessed documents being not relevant was unfounded [7]. 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. Therefore, it is not clear what impact the potential proportion of relevant unassessed documents may have on system comparisons.

For this very reason, the effect that pooling has on system comparison has been investigated in the context of relevant document recall, focusing upon several different areas of the completeness assumption and system pooling. These studies have investigated issues such as the effect on system comparison and the subsequent uncertainty when using incomplete relevance judgements [2, 3, 7], efficient pooling strategies [7–10], automatically generated relevance assessments [11, 12], and the importance of significance testing during system comparison [13]. 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, if any. For example, a recent investigation of the TREC Robust-HARD 2005 collection identified system bias, which was a result of both a shallow pool depth, and (potentially) similar runs forming the system pool [14]. The outcome was a bias in the collection favouring systems that gave more weight to query terms that appeared in the title of a document over other approaches. Although this does not necessarily indicate a failing of system pooling it motivates the need for a more robust evaluation protocol which considers aspects such as pooling, the measurement depth, and the status of unassessed documents during evaluation.

1.2 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 [7], 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 Capturing the (un)certainty of System Performance

We hypothesise that the certainty associated with estimating a measurement of a systems performance at a depth k 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 larger proportion of assessed documents contained in a ranked list, the more confident we are in the estimate of a systems performance at the corresponding measurement 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 a high degree of certainty. If the ranked lists of both systems are incomplete, but contain similar proportions of assessed documents, then confidence in the relative comparison of both systems would also be high. However, if one system has a substantially lower proportion of assessed documents than another, then the performance estimate of that 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 resulting in a higher degree of uncertainty in the comparison.

2.1 Measure of Assessment

We propose to capture uncertainty by calculating the proportion of assessed documents in a ranked list. Let A be the set of assessed documents in a collection of N documents, and X be the set of retrieved documents by a system for that topic. Then *Assessment Precision* A_p can be defined as the probability of retrieving judged documents:

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

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. 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[4].

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.

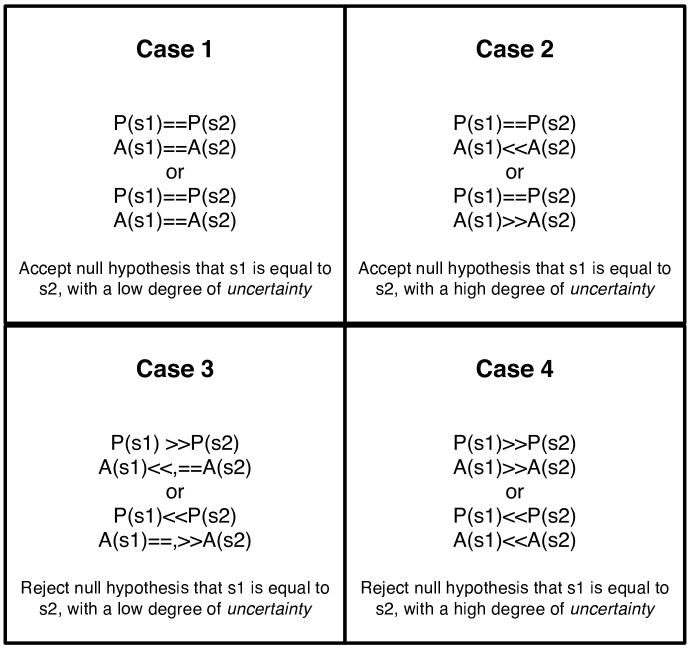
2.2 System Evaluation Decision Matrix

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:

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

[4] 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* has recently been proposed as a measure for incomplete collections [2], which removes all unassessed documents during estimation of performance. Note that, *bpref* itself does not quantify the proportion of assessed documents that are removed but a corresponding A_p measure could be derived to complement such a metric.



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*

Fig. 1. System Evaluation Decision Matrix for system comparison.

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

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@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 1. 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 [13].

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.

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 [7]. 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 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 [13]. 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, further checking of the unassessed documents can be made, adopting strategies such as that outlined by Carterette et al. [8]. 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 [14].

3 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[5]. By using the official runs we could 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 cut off values. Using the Assessment Precision metrics, we were investigating what effect using a measurement depth deeper than the pooling 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 focused our attention on runs from particular collections, such as the Robust-HARD 2005[6], which has been identified as potentially problematic to use due to title bias [14]. The aim is to better understand the problems cited with this collection, with particular focus on runs that both weight topic titles and runs that do not.

For each collection we first analysed each possible pair-wise system comparison of the officially submitted runs using the decision matrix (see Table 1). 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@10, P@30, P@100, P@500, P@1000, MAP, and the Assessment Precision metrics A_p@10, A_p@30, A_p@100, A_p@500, A_p@1000, MAA. For each metric, we counted the number of comparisons that fell into each outcome i.e. Cases 1-4 (see Figure 1 for an outline of each case). Table 1 presents the proportion of overall system comparisons that fall into each case. 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 780 pair-wise comparisons performed overall.

The first thing we were interested in was the proportion of significant pair-wise differences in terms of system performance across the various test collections, specifically to test what extent increasing measurement depth had on this proportion. The table provides the proportion of both "strong" and "weak" significant differences between systems for each metric. From the reported results, a noticeable trend was that for the majority of collections where the proportion of significant differences decreased as the measurement depth increased,

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

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

	@10	**@30**	**@100**	**@500**	**@1000**	**MAP/AA**
TREC 3	**0.69** 0.07	**0.67** 0.14	**0.58** 0.32	**0.44** 0.49	**0.41** 0.57	**0.64** 0.13
780	**0.17** 0.07	**0.11** 0.09	**0.05** 0.05	**0.00** 0.06	**0.00** 0.02	**0.04** 0.19
TREC 4	**0.72** 0.12	**0.65** 0.16	**0.43** 0.39	**0.35** 0.48	**0.33** 0.5	**0.5** 0.25
528	**0.16** 0.00	**0.19** 0.00	**0.05** 0.13	**0.02** 0.15	**0.01** 0.16	**0.02** 0.24
TREC 6	**0.60** 0.16	**0.55** 0.26	**0.45** 0.42	**0.45** 0.44	**0.44** 0.45	**0.6** 0.2
3081	**0.13** 0.11	**0.10** 0.10	**0.033** 0.10	**0.01** 0.10	**0.00** 0.10	**0.02** 0.18
TREC 8	**0.70** 0.04	**0.69** 0.07	**0.52** 0.31	**0.51** 0.36	**0.57** 0.33	**0.57** 0.22
8000	**0.19** 0.06	**0.17** 0.07	**0.01** 0.16	**0.00** 0.13	**0.00** 0.10	**0.04** 0.18
WEB 04	**0.35** 0.00	**0.03** 0.25	**0.02** 0.19	**0.01** 0.19	**0.01** 0.19	**0.02** 0.37
561	**0.65** 0.00	**0.42** 0.30	**0.51** 0.28	**0.53** 0.26	**0.53** 0.27	**0.40** 0.20
ROBUST 03	**0.83** 0.01	**0.74** 0.09	**0.55** 0.2	**0.52** 0.22	**0.57** 0.17	**0.65** 0.14
2145	**0.03** 0.13	**0.01** 0.16	**0.00** 0.25	**0.00** 0.26	**0.00** 0.27	**0.00** 0.21
ROBUST 05	**0.77** 0.12	**0.66** 0.23	**0.53** 0.38	**0.57** 0.35	**0.60** 0.31	**0.68** 0.19
1485	**0.07** 0.03	**0.07** 0.04	**0.00** 0.10	**0.00** 0.08	**0.00** 0.08	**0.00** 0.13
HARD 05	**0.71** 0.20	**0.66** 0.25	**0.57** 0.39	**0.57** 0.42	**0.57** 0.42	**0.72** 0.18
2775	**0.05** 0.05	**0.05** 0.04	**0.00** 0.04	**0.00** 0.02	**0.00** 0.01	**0.00** 0.10

Table 1. The proportion of comparisons for each case across the TREC collections.

the exception being the Robust 2003 collection. This trend is the converse of the intuition stated in [7], whereby increasing measurement depth also increased discrimination between systems. The intuition is that good systems will continue to retrieve relevant documents beyond the pooling depth, which will have been discovered by other runs. From these results it would appear that discrimination between the set of systems lessens as the measurement depth increases. For some collections such as TREC-3, 6, 8 and the Web 2004 collections this becomes more stated as measurement depth is increased beyond the pooling depth.

We then examined the proportion of significant differences between systems that fall into either the “strong” or “weak” case. 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 1. We find a smaller proportion of comparisons falling into the “strong” case in contrast to P@30 (17% to 11%). Conversely there is an increase in “weak” cases from 7% to 8.7%. This trend remains as we continue increasing measurement depth towards P@1000. Using MAP, which is calculated over all 1000 documents, 3.7 % significant differences are “strong” compared to 19% “weak” cases. This trend is common across the other collections, and appears to be an indication of the amount of information that is used to estimate system performance. *Increasing measurement depth results in a higher level of uncertainty in system comparison.*

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

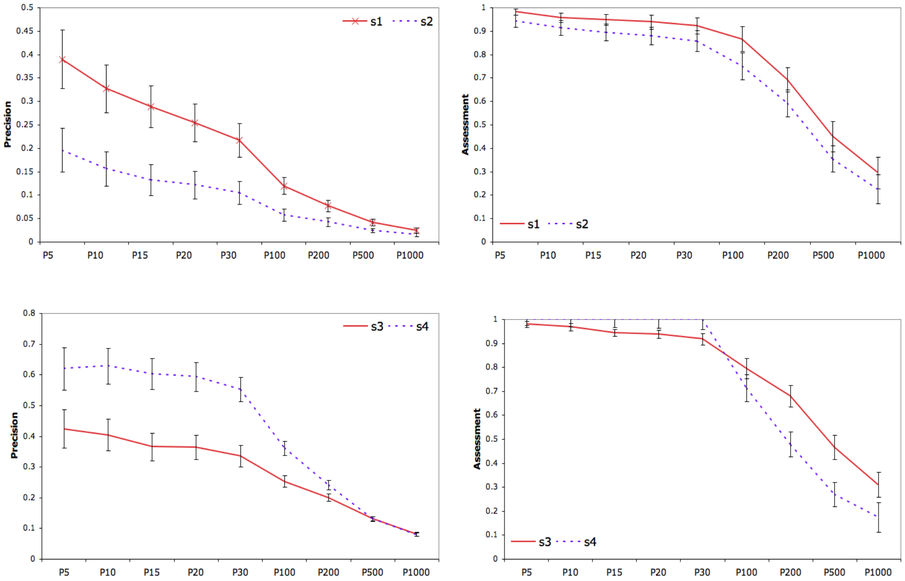


Fig. 2. Comparison of runs from Robust 2003 (top) and Robust 2005 (bottom).

in a larger proportion of cases where one system has a significantly larger number of assessed documents than another.

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. As a case study we present a comparison of two sets of systems from the Robust 2003 (Figure 2, top) and Robust 2005 tracks (Figure 2, bottom). In both figures we display both the P@ (left) and A_p@ (right) metrics at various ranks for both systems. Error bars are displayed to show variation across the set of topics and significance between systems.

The first example (Figure 2, top), illustrates a comparison of systems where conditions remain relatively even for both systems across various ranks. System s_1 has significantly higher precision than s_2 up until P@500. If we examine assessment, s_1 also has higher assessment but not significantly so with the exception of A@100. This example illustrates that even with systems that have comparable conditions in terms of assessment, the practice of using a measurement depth larger than the pooling depth can cause uncertainty in comparisons such as at A@100.

We also examined two runs from the Robust 2005 track where it has been identified that there is a bias in the collection towards documents that have query terms in the title [14]. The two runs were from the same research group, with system s_3 placing emphasis on keywords appearing in the title of documents, while s_4 uses an external collection to expand the original query. Initially system s_4 reports significantly higher precision across the 50 topics. As the measurement depth increases this improvement becomes marginal until both runs share similar performance. If we examine the levels of assessment across the same ranks, we find that assessment is equal until we reach A@100, then s_4 decreases in assessment with respect to s_3. For this collection the pool depth is 47.

This result reflects the findings in [14], where the performance of s_4 is underestimated once a larger measurement depth is used. System s_3 begins to return more assessed (and relevant) documents than s_4, which in turn is returning more unique and unassessed documents. From the study of Zobel [7], who investigated the rate of discovering new relevant documents beyond pool depth, it is uncertain if both systems shared similar levels of assessed documents that performance would converge.

4 Discussion and Conclusion

In this paper we argued that uncertainty should be identified during the evaluation process when using incomplete relevance judgements. We therefore proposed a new set of metrics based on the level of assessment which can be used to 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. 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 (e.g. shallower measurement depth, different collections, etc.). The advantage is those comparisons that may not be fair can be detected and investigated accordingly.

During the course of our empirical study, where we employed the extended evaluation methodology, we found evidence to suggest that the use of a measurement depth larger than the pooling depth weakens claims of significance in performance. This was a concern that was previously raised, but not confirmed, by Zobel [7]. Our results indicate that as the measurement depth increases beyond the pooling depth, uncertainty across many system comparisons also increases, and interestingly, the discrimination between systems weakens. Consequently, this supports the conclusions drawn by Sanderson and Zobel [13], who stated that metrics which consider early precision such as P@10 can be used to accurately discriminate between systems.

This decrease in discrimination at deeper measurement depths may result from the higher variation is performance estimates across topics stemming from the lack of assessment at these depths. Also, the majority of systems, what Aslam and Savell refer to as the “popular” systems [11], appear to discover a similar proportion of relevant documents once the measurement depth increases. However, the performance of the best systems may still be underestimated because they return documents than are unique to the documents returned by the “popular” systems. In general the relative rankings of systems tend to remain stable across measurement depths and varying levels of incompleteness. However, it is those cases where a system changes in ranking because of the effects of incompleteness that should to be detected, in particular, comparisons resulting in Case 2 in the decision matrix. This is because it is these systems, which are novel and considerably different to the "popular" systems, that require more consideration during evaluation.

In this paper we have introduced an extended retrieval evaluation methodology which uses Assessment Precision to determine whether comparisons between competing systems are made under similar conditions. The adoption of this methodology leads to a fine grained analysis during the evaluation as the necessary context is provided to draw firmer conclusions. Future work will examine the implications and usage of this methodology in greater detail as well as investigate issues relating to Assessment Precision such as the relationship between Precision and the level of assessment.

Acknowledgements

We would like to thank Fabio Crestani, Kalervo Järvelin, David Harper, and the anonymous reviewers for their valuable comments.

References

1. Salton, G.: The state of retrieval system evaluation. Information Processing Management **28** (1992) 441–449
2. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proceedings of the 27th ACM SIGIR Conference. (2004) 25–32
3. Voorhees, E.M., Harman, D.K., eds.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge, Massachusetts 02142 (2005)
4. Spärck-Jones, K., Van Rijsbergen, C.J.: Report on the need for and provision of an "ideal" information retrieval test collection. Technical report, British Library Research and Development Report 5266, University of Cambridge. (1975)
5. Blair, D.C.: Some thoughts on the reported results of TREC Information Processing Management **38** (2002) 445–451
6. Voorhees, E., Harman, D.: Letters to the editor. Information Processing Management **39** (2003) 153–156
7. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Proceedings of the 21st ACM SIGIR. (1998) 307–314
8. Carterette, B., Allan, J., Sitaraman, R.: Minimal test collections for retrieval evaluation. In: Proceedings of the 29th ACM SIGIR, Seattle, WA (2006) 268–275
9. Cormack, G.V., Palmer, C.R., Clarke, C.L.A.: Efficient construction of large test collections. In: Proceedings of the 21st ACM SIGIR. (1998) 282–289
10. Sanderson, M., Joho, H.: Forming test collections with no system pooling. In: Proceedings of the 27th ACM SIGIR. (2004) 33–40
11. Aslam, J.A., Savell, R.: On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In: Proceedings of the 26th ACM SIGIR, Toronto, Canada (2003) 361–362
12. Soboroff, I., Nicholas, C., Cahan, P.: Ranking retrieval systems without relevance judgments. In: Proceedings of the 24th ACM SIGIR. (2001) 66–73
13. Sanderson, M., Zobel, J.: Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proceedings of the 28th ACM SIGIR. (2005) 162–169
14. Buckley, C., Dimmick, D., Soboroff, I., Voorhees, E.: Bias and the limits of pooling. In: Proceedings of the 29th ACM SIGIR Conference, Seattle, WA (2006) 619–620