Evaluation Effort, Reliability and Reusability in XML Retrieval

Sukomal Pal and Mandar Mitra
Information Retrieval Lab, CVPR Unit, Indian Statistical Institute, 203 B T Road, Kolkata 700 108, India.
E-mail: {sukomal_r, mandar}@isical.ac.in

Jaap Kamps
Archives and Information Studies/Humanities, University of Amsterdam, Turfdraagsterpad 9, NL-1012XT, Amsterdam, The Netherlands. E-mail: kamps@uva.nl

The Initiative for the Evaluation of XML retrieval (INEX) provides a TREC-like platform for evaluating content-oriented XML retrieval systems. Since 2007, INEX has been using a set of precision-recall based metrics for its ad hoc tasks. The authors investigate the reliability and robustness of these focused retrieval measures, and of the INEX pooling method. They explore four specific questions: How reliable are the metrics when assessments are incomplete, or when query sets are small? What is the minimum pool/query-set size that can be used to reliably evaluate systems? Can the INEX collections be used to fairly evaluate "new" systems that did not participate in the pooling process? And, for a fixed amount of assessment effort, would this effort be better spent in thoroughly judging a few queries, or in judging many queries relatively superficially? The authors' findings validate properties of precision-recall-based metrics observed in document retrieval settings. Early precision measures are found to be more error-prone and less stable under incomplete judgments and small topic-set sizes. They also find that system rankings remain largely unaffected even when assessment effort is substantially (but systematically) reduced, and confirm that the INEX collections remain usable when evaluating nonparticipating systems. Finally, they observe that for a fixed amount of effort, judging shallow pools for many queries is better than judging deep pools for a smaller set of queries. However, when judging only a random sample of a pool, it is better to completely judge fewer topics than to partially judge many topics. This result confirms the effectiveness of pooling methods.

Introduction

Content-oriented XML[1] retrieval is a domain of information retrieval (IR) that has been receiving increasing attention in recent years. The widespread use of eXtensible Markup Language (XML) as a standard document format on the Web and in digital libraries has led to the continuous growth of XML information repositories. This growth has been matched by increasing efforts in the development of XML IR systems that support content-oriented XML retrieval. Besides the content, these systems also exploit structural information, both syntactic and semantic, provided by the XML markup, to return document components or XML elements instead of whole documents in response to a user query. This type of focused retrieval is particularly useful when dealing with collections of long documents or documents covering a wide variety of topics (e.g., books, user manuals, legal documents) because the effort required from users to locate relevant content can be reduced by directing them to the most relevant document components. As the number of XML retrieval systems increases, so does the need to evaluate their effectiveness (Malik, Trotman, Lalmas, & Führ, 2007).

The Initiative for the Evaluation of XML retrieval (INEX; 2009), set up in 2002, has been responsible for creating a Cranfield-style infrastructure for evaluating the effectiveness of content-oriented XML IR systems. INEX provides large test collections, topic sets, and relevance judgments. As for other document retrieval evaluation fora, the relevance assessments used for evaluation at INEX are based on a pool generated from results submitted by participants. However, as the retrieval unit for XML search systems can

---

[1]W3C, Extensible Markup Language (XML), http://www.w3.org/XML

be an element of arbitrary granularity and length, evaluation has been a challenge at INEX. Evaluation measures used in traditional IR, where a whole document is typically considered either relevant to a user query or not, are no longer tenable as the aim here is to locate the most relevant document part(s) and not complete documents. Various evaluation measures have been tried over the years at INEX. The official metrics used at INEX 2002 (Gövert & Kazai, 2003; calculated by the inex-eval program) were modified for INEX 2003 and 2004 (Gövert, Führ, Lalmas, & Kazai, 2006; cf. the inex-eval-ng program). Again, at INEX 2005, three new cumulated gain-based (Järvelin & Kekäläinen, 2002) metrics were taken as official metrics (Kazai & Lalmas, 2006a, 2006b). These metrics were also used at INEX 2006. Since 2007, however, an arbitrary passage that may span more than one XML element has also been accepted as a valid retrievable unit for the focused ad hoc task. This new definition of the task necessitated a metric that could be used to evaluate both passage-retrieval and element-retrieval systems in the same manner. This gave rise to a family of metrics that were derived from the traditional interpolated precision-recall metrics. However, these metrics are defined in terms of text length expressed in characters, rather than the number of documents (see Experimental Set-Up section for details). Five of these metrics, namely iP[0.00], iP[0.01], iP[0.05], iP[0.10], and AiP, were used in the official reports for the focused ad hoc tasks. Among these, iP[0.01] was taken as the official measure to rank the competing systems (Kamps, Pehcevski, Kazai, Lalmas, & Robertson, 2008).

Because these measures are extensions of their counterparts in the standard document retrieval setting, they may be expected to have similar properties. However, the evaluation set-up used at INEX is markedly different from that used at other fora in certain ways.

- First, at INEX, retrieval granularity is at the XML element or passage level, but pooling is done at the document level. Thus, even when a small passage is retrieved from an article, the complete article is included in the pool. This leads to considerable diversity and robustness in the pooling.
- Second, TREC (Text Retrieval Conference) generally uses a fixed pool-depth (typically the top 100 documents from each contributor are pooled) and pools vary in size across queries. In contrast, INEX pools all runs (both valid and invalid) using top-n pooling and a fixed pool size, i.e., the pool depth is dynamically chosen for each query so that a pool size of around 600 documents is reached. The dynamically chosen pool depth, which is at least 30 in terms of articles (and substantially higher in terms of elements), is assumed to be deep enough to cover a large fraction of the relevant articles for the majority of topics.
- Third, relevance in the XML domain is defined at the sub-document level. A relevance judgment file (or qrels) contains more information than just a Boolean indicator about whether a document is relevant or irrelevant for a given topic. The qrels lists, for each topic, the documents that contain relevant passages, and precisely specifies the relevant elements and/or passages within each document. Relevant elements are

identified by their xpath[2] and relevant passages either by a combination of xpath and start and end positions, or simply by their length and character-offsets from the beginning of the article. This relevance judged pool or qrels is then used for evaluation.
- Finally, there are no dedicated topic creators or assessors at INEX. Participants are responsible for creating search topics and assessing the pools generated from submissions, with the assessment for a particular topic being generally done by the participant who created the topic. Thus, each topic is judged in its entirety by a single assessor. From an operational point of view, this is possibly the most important difference between INEX and other evaluation fora.

Given these differences, it is an open question whether the INEX metrics indeed have similar properties with regard to reliability and robustness as their document retrieval counterparts. The main aim of our study is to investigate this issue. Specifically, our goal is to find answers to the following questions:

1. How reliable are the various metrics in ranking competing systems when assessments are incomplete (i.e., when some relevant documents have not been included in the judged set, and have therefore been assumed to be nonrelevant)? On a related note, what is the minimum pool size that can be used to reliably evaluate systems?
2. How reliable are the various metrics in ranking competing systems if the query set size is small? In particular, what are the error rates of the various metrics as query set size changes? (The error rate quantifies the chance of arriving at a wrong conclusion when comparing two systems using a particular set of queries.) What is the minimum number of queries that should be used to keep the error rates for the various metrics within a maximum allowable upper bound?
3. When a set of relevance assessments is used to evaluate a "new" system that did not contribute to the pool used in the relevance assessment process, are the results biased against this system? If yes, how serious is the bias?
4. For a fixed amount of assessment effort, would this effort be better spent in thoroughly judging a few queries, or in judging many queries relatively superficially?

To the best of our knowledge, no such study has been completed so far. The first two issues listed above were investigated using a selected set of runs from the INEX 2007 ad hoc focused submissions, and preliminary findings were reported earlier (Pal, Mitra, & Chakraborty, 2008). However, those experiments used partial data and had certain drawbacks (detailed in subsequent sections). It is in this context that this work was undertaken. Our aim was to do a more detailed study of these issues using the complete sets of runs from the INEX 2007 and INEX 2008 ad hoc focused submissions.

Our experiments and analyses are restricted to the focused task of INEX. This task is the closest analogue to straightforward, ad hoc document retrieval: given a query, retrieve

---

[2]W3C, Xpath – XML Path Language (XPath) Version 1.0, http://www.w3.org/TR/xpath

a ranked list of passages or elements that are most focused with respect to the information need expressed in the query. The task allows us to look at early precision (iP[0.00], iP[0.01], iP[0.05], iP[0.10]) as well as overall performance (AiP, MAiP) of a set of systems. The early precision metrics used in the task are known to be unstable and they need to be used with caution when comparing the performance of systems (Führ, Kamps, Lalmas, Malik, & Trotman, 2008; Kamps, Geva, Trotman, Woodley, & Koolen, 2009). Thus, in a sense, this task is the "weakest link" among the INEX tasks, and needs to be thoroughly investigated. Further, for many groups, the focused results form the basis for submissions to the other ad hoc tasks, suggesting that our choice of the focused task is a reasonable one.

In the next section, we review past work that provides the background for our study. In the Experimental Set-Up section, we present the test environment used in this study, definitions of the measures to be examined, and our experimental set-up. In the Pool Sampling, Query Sampling, Error Rates, and Leave-One-Out sections, we describe our experiments, results, and observations in detail. In the Discussion section, we present a comparative analysis and discuss these observations, along with some unresolved issues. In the last section, we conclude the article.

Previous Work

Evaluating evaluation metrics and methodologies has a well-established history in the field of document retrieval. Zobel (1998) examined the fairness and trustworthiness of the pooling-based evaluation methodology of IR experiments. He showed that TREC results are reliable, but at best, 50–70% of the relevant documents are identified by the pooling method used in TREC. Zobel also opined that using a large measurement depth (as compared to the pool-depth) improves measured discrimination between systems, but also introduces some uncertainty into the results. The study included experiments on pooling to study how system performances get reinforced due to pooling, and how the omission of a system's contribution from the pool affects performance measures. For recall-oriented evaluation, the author devised a way to estimate the total number of relevant documents for a query, and showed that either increasing the number of pooled systems, or using a variable pool depth will provide a better estimate of the number of relevant documents for a query.

Buckley and Voorhees (2000) proposed a novel way to examine the accuracy of various evaluation measures and validated a number of traditional rules of thumb that address issues such as the minimum number of queries required for reliable evaluation, which measures to use, and the notion of "significant" difference in the scores between two competing systems. They introduced the concept of an error rate for an evaluation measure. By repeating retrieval runs using different variations of the same query sets and comparing pairs of systems across query variations, they showed that average precision (AP) is a more stable measure than measures based on early precision. They studied the effect of topic set size on error rates more extensively in (Voorhees & Buckley, 2002). Using the TREC results and mean average precision (MAP) as the metric, error rates were directly computed for topic sets of size up to 25. The rates were then extrapolated for larger topic set sizes. The study indicated a caveat when two retrieval systems are compared to determine one's superiority over the other, and cautioned that error-rates must be taken into consideration along with the number of topics.

This work was extended in 2004 with the study of evaluation measures under incomplete and imperfect relevance judgments (Buckley & Voorhees, 2004). The authors examined the stability of system rankings produced by a metric when the size of the relevance-judged pool is gradually reduced, as well as when the topic-set size is reduced. Once again, they showed that MAP is both stable and discriminatory for evaluating document-level retrieval from a static document collection.

Sanderson and Zobel (2005) extended the study on error rates of evaluation metrics using MAP and P@10 in light of significance tests, and found the bounds (lower and upper) of these error rates. They observed that, given a set of relevance judgments, MAP is more reliable than P@10. To estimate the discriminative power of various IR metrics, Sakai (2007) used the bootstrap hypothesis test and compared it with the swap-method based on the NTCIR (NII Test Collection for IR Systems) collection.

Both approaches suggest that AP is one of the best metrics, so far as discriminative power is concerned. The power of a metric can also be seen as the statistical power of a hypothesis test as proposed by Webber, Moffat, and Zobel (2008). The authors determine the minimum number of topics necessary to detect a certain degree of superiority of one system over another by estimating between-system score differences and their standard deviations. Interestingly, they also conclude that greater statistical power is achieved for the same relevance assessment effort by judging a shallow pool for a large number of topics rather than a deep pool for a small number of topics.

In the recent past, there has been a plethora of work in the direction of low-cost evaluation (Ahlgren & Grönqvist, 2008; Baillie, Azzopardi, & Ruthven, 2008; Bompada, Chang, Chen, Kumar, & Shenoy, 2007; Carterette, 2007; Sakai & Kando, 2008; Yilmaz & Aslam, 2006). Although Yilmaz and Aslam (2006) proposed different approximations of AP (*induced AP*, *subcollection AP*, and inferred AP [*infAP*] to handle incomplete and imperfect collections, Bompada et al. (2007) compared the performance of *bpref* (a retrieval effectiveness measure), *infAP*, and *nDCG* (normalized discounted cumulated gain); Carterette (2007) devised ways to create minimal test collections and robust test collections by estimating performance differences between two systems. Ahlgren and Grönqvist (2008) proposed a measure *RankEff* and compared its performance with *bpref* and *MAP*. Sakai and Kando (2008) studied the effect of incomplete

assessments on different metrics like bpref, RBP (Rank-Biased Precision, proposed by Moffat and Zobel, 2008), Q-measure, AP, nDCG, and their condensed-list variants.

The condensed-list variant of a metric is calculated using a ranked list obtained by removing unjudged documents from the original ranked list. The condensed-list variants AP', Q', and $nDCG'$ consistently performed better than either RBP or their original counterparts under unbiased incompleteness in terms of discriminative power and ranking stability. In leave-one-group-out experiments, AP', $nDCG'$, and Q' seemed to overestimate a "new" system's performance, whereas AP, nDCG, and Q underestimate the new systems (Sakai, 2008b). However, the overestimation is higher than the underestimation. Under shallow pooling or pool-depth bias, AP, nDCG, and Q-measure are better than their condensed-list variants (Sakai, 2008a). Baillie et al. (2008) provide an overall summary of some of these approaches of low-cost evaluation and try to address the issue from a theoretical perspective.

Aslam, Pavlu, and Yilmaz (2006) proposed an elegant statistical technique to efficiently and effectively estimate standard measures of retrieval performance from random subsamples (as small as 4%) of the TREC pool, by modeling each measure using some probability distribution. They also introduced the idea of root-mean-square error for comparing evaluation metrics.

Büttcher, Clarke, Yeung, and Soboroff (2007) revisited the "leave-one-out" experiments reported by Zobel (1998), and applied two statistical learning techniques (KLD and SVM) to infer whether an unjudged document coming from a new system which did not take part in pooling is relevant or not.

Carterette, Pavlu, Kanoulas, Aslam, and Allan (2008) conducted a noteworthy experiment with topic-set size. Using the TREC Million Query Track data, the authors found that evaluation based on a large number of queries with fewer judgments is more cost effective than, but equally reliable as, using fewer queries with more judgments.

The stability of system rankings in the above studies are mainly measured by Kendall's rank correlation coefficient (τ). Recent research has shown that Kendall's τ suffers from some pitfalls. According to a widely used rule of thumb, two rankings are considered to be similar if their τ is 0.9 or higher. Sanderson and Soboroff (2007) warn that the threshold of 0.9 should be taken with caution. In more recent work, Yilmaz, Aslam, and Robertson (2008) showed that τ penalizes ranking differences at both high ranks and low ranks equally. However, the IR community is usually more concerned about differences in the top ranks than those at the bottom. Their proposed coefficient, AP correlation (τ_{AP}) gives more weight to differences at high rankings and yields smaller values than Kendall's τ for errors in the top ranks, but equals τ when errors are uniformly distributed over the entire ranked list. To the best of our knowledge, there are no reports on what should be the acceptable threshold for τ_{AP} to infer that two rankings are essentially the same. Therefore, it is prudent to use τ_{AP} along with τ, rather than using either one alone.

All the work discussed above was based on document-level retrieval using mainly TREC/NTCIR data. Kazai and Lalmas (2006a) first studied the evaluation of XML retrieval. Their work used the XCG-based metrics (e.g., MAep, nxCG, MAnxCG, etc.) and some other older metrics like Q, R, and inex-eval, with the INEX2004 submissions (where only XML elements were permissible as units of retrieval).

Trotman, Pharo, and Jenkinson (2007) experimented with the INEX pooling and assessment strategies during the INEX 2006 workshop. One of the questions addressed in this work was whether the INEX pool can be reduced in size using a shallow random pool (around 100 documents, taken in alphabetical order, per topic for 15 topics). This experiment was done on the relevant-in-context task, but was not comprehensive or conclusive. The correlation between system-rankings using the shallow pool and the original pool was highly positive (Spearman's rank correlation $= 0.97$) on one hand; but for the 10 best systems, it was near zero (Spearman's rank correlation $= -0.03$).

Piwowarski, Trotman, and Lalmas (2008) provide details about the INEX pooling and assessment exercises and their evolution during 2002–2006, along with an in-depth analysis. The authors explain why INEX shifted from an element-level pooling strategy to a document-level pooling strategy, and show that a pool containing the top-ranked 500 distinct documents per topic is large enough to ensure stable evaluation.

However both these studies used a different experimental set-up. INEX has evolved much in several respects since then. Since 2006, the test collection has changed—a collection of technical articles published by Institute of Electrical and Electronics Engineers (IEEE) has been replaced by a January 2006 dump of the English *Wikipedia*. The evaluation metrics have also changed at regular intervals. Prior to 2007, only XML elements were considered retrievable units at INEX. How to handle overlap among the elements was an issue during this period. Since 2007, the retrieval of arbitrary, nonoverlapping passages has been permitted, and a new set of evaluation metrics has been introduced. The track overview articles (Führ et al., 2008; Kamps et al., 2009) discuss evaluation results related to the Focused, Best in Context, and Relevance-in-Context tasks of the ad hoc track. The authors found that the official metric used for the focused task (iP[0.01]) was unstable; further, no significant differences were found among the 10 best systems based on the metric (one-tailed *t*-test, 95% confidence level). However, no information was provided about the other reported metrics, and to our knowledge, there are no reports of any analysis of their characteristics in the context of XML retrieval. Similarly, there is no reported study of the robustness of the INEX pool, which is unique in several respects.

In this article, we present a study of these new metrics of XML retrieval using the INEX2007 and INEX 2008 collections. Our motivation is to analyze the general characteristics of the measures to find out which of the measures is the most robust, stable, and least erroneous in reliably comparing a set of XML retrieval systems. We are also interested in examining the pool creation process, and in looking at the trade-offs between the effort involved in creating the relevance

assessments and the quality of the resultant collection. For example, we are interested in quantities like the minimum number of topics that should be used, the minimum number of documents to be judged per query, the minimum pool depth to be used, etc. We believe these observations can help in building test collections that are much larger than the ones currently used, such as the much bigger INEX 2009 corpus, for which the completeness assumption in Cranfield-based pooling is likely to be seriously challenged and possibly compromised.

Our work is in line with the earlier work of Buckley and Voorhees (2000, 2002, 2004), Zobel (1998), Büttcher, Clarke, Yeung, and Soboroff (2007), and Sanderson and Zobel (2005). Our random pool sampling and random query sampling experiments are in line with those of Buckley and Voorhees. Error rates for the measures have been obtained according to the established method of Voorhees and Buckley (2002) and Sanderson and Zobel (2005). The leave-one-out experiments are motivated by the same objective as in Zobel (1998). The methodology has been slightly modified, however, in accordance with the findings of Büttcher et al. (2007). Both these studies (Büttcher et al., 2007; Zobel, 1998) looked at the overall impact of the pooling bias, while in reality, the pooling bias varies from query to query. Beside the overall effect, we also look at the effect of pooling bias on a per-query basis (Zobel, 1998).

In most of our comparative analyses, our objective is to look at the stability of system rankings, rather than the absolute values of the various measures. Changes in ranking are quantified using the conventional Kendall's τ, as well as the more recently proposed τ_{AP} measure proposed by Yilmaz et al. (2008). Some of our results validate observations from the document retrieval domain in the context of focused retrieval, thus underlining the intrinsic properties of the metrics used, while some of the results (e.g., related to variation of pooling depth in the Discussion section) indicate some new observations that can equally hold true in the document retrieval paradigm as well.

Experiment Set-Up

Test Collection

We use the INEX 2007 and 2008 ad hoc test collections in our experiments. These test collections consist of an XML-ified version of the English *Wikipedia*. The corpus contains 659,388 documents, and has a total size of 4.6 GB (Denoyer & Gallinari, 2006). For INEX 2007, the original topic set contained 130 queries (INEX topics 414–543); however, relevance judgments were available for only 107 topics, so the remaining 23 queries were not part of our experiments. For INEX 2008, the topic set consisted of 135 queries (544–678) and relevance judgments were available for only 70 queries.

The focused task of the ad hoc track expects participating systems to return, for each topic, a ranked list of nonoverlapping document parts (either passages or XML elements) that are most focused with respect to the information need

expressed in the topic. For 2007, among the submitted runs, 79 were reported in the INEX 2007 Web site as valid runs. For 2008, this number was 61. Each such run was supposed to retrieve 1,500 passages or elements per topic, and list them in decreasing order of their relevance to the topic. The effectiveness of a strategy for a single topic is computed as a function of the ranks of retrieved and relevant texts and their relative lengths. The effectiveness of the strategy as a whole is then computed by taking into consideration its effectiveness across all the topics.

Evaluation Measures

Effectiveness is measured using metrics based on the notions of recall and precision, suitably adapted to fit the XML context:

$$precision = \frac{\text{amount of relevant text retrieved}}{\text{total amount of retrieved text}}$$

$$= \frac{\text{length of relevant text retrieved (in characters)}}{\text{total length of retrieved text (in characters)}}$$

$$recall = \frac{\text{length of relevant text retrieved (in characters)}}{\text{total length of relevant text (in characters)}}$$

Kamps et al. (2008) provide more formal definitions as follows. Let p_r be the document part at rank r in the ranked list L_q returned by a retrieval system for a topic q. Let $size(p_r)$ be the total number of characters contained by p_r and $rsize(p_r)$ be the length (in characters) of relevant text contained in p_r (as highlighted by the assessor during the relevance judgment process). If there is no highlighted text, $rsize(p_r) = 0$. Further, let $Trel (q)$ be the total amount of relevant text for topic q (this is the sum of the lengths of relevant texts across all documents). Then,

$$\text{Precision at rank } r, \ P[r] = \frac{\sum_{i=1}^{r} rsize(p_i)}{\sum_{i=1}^{r} size(p_i)} \quad (1)$$

And

$$\text{recall at rank } r, \ R[r] = \frac{\sum_{i=1}^{r} rsize(p_i)}{\sum_{i=1}^{r} Trel(q)} \quad (2)$$

Because retrieval granularity can vary, a comparison of precision values at a given rank across systems may not be meaningful. Instead, precision at various recall levels may be used. Thus, interpolated precision at various recall levels is used for comparing systems, where interpolated precision at recall level x is defined as follows:

$$iP[x] = \begin{cases} \max\limits_{\substack{1 \le r \le |L_q|, \\ R[r] \ge x}} (P[r]) & \text{if } x \le R[|L_q|] \\ 0 & \text{if } x > R[|L_q|] \end{cases}$$

$$AiP(t) = \frac{1}{101} \sum_{x \in \{0.00, 0.01, ..., 1.00\}} iP[x](t)$$

for the INEX ad hoc tasks, $|L_q| \le 1500$.

For example, iP[0.00] usually gives an estimate of the interpolated precision when the first relevant unit is retrieved, and iP[0.01] is the interpolated precision at the 1% recall level for a given topic.

Analogously, for a particular topic t, average interpolated precision, AiP, is defined as the average of interpolated precision values at 101 standard recall levels (0.00,0.01, . . . ,1.00):

$$AiP(t) = \frac{1}{101} \sum_{x \in \{0.00, 0.01, ..., 1.00\}} iP[x](t)$$

Overall performance measure. We measure overall performance of a system by averaging its scores across all the topics in the set. If there are n topics, the performance of a system at recall level x is given by:

$$iP[x]_{overall} = \frac{1}{n} \sum_{t=1}^{n} iP[x](t)$$

Similarly, mean average interpolated precision, MAiP, over n topics is expressed as

$$MAiP = \frac{1}{n} \sum_{t=1}^{n} AiP(t)$$

Since INEX 2007, for the focused ad hoc task, mean interpolated precision at four selected recall levels, $iP[x]$, $x \in \{0.00, 0.01, 0.05, 0.10\}$ and MAiP were reported, and iP[0.01] was selected as the "official" metric that was used to rank systems.

Experiments

Our experiments are organized into four groups as explained below. Each group of experiments attempts to address one of the questions raised in the Introduction.

1. In the Pool Sampling section, the effects of incomplete relevance assessments on system rankings are studied. In other words, we investigate how system rankings would change if some relevant passages were not actually known to be relevant. The motivation is to verify the robustness of the metrics when the completeness assumption of the Cranfield paradigm is grossly violated.
2. In the Query Sampling section, progressively smaller subsets of the complete topic set are randomly chosen, but for a chosen topic, all available assessments are used for evaluation. The aim here is to observe the characteristics of the metrics as the topic set size decreases, and to find the minimum number of topics required for reliable evaluation at INEX.
3. Next, in the Error Rates section we look at the above issue from a different point of view. Each of the five metrics is used in turn to compare pairs of retrieval strategies on two disjoint topic sets of the same size, and their error rates are observed as the topic set size is reduced.
4. Finally, in the Leave-One-Out section an attempt is made to simulate a scenario where a "new" system that does not

contribute to a given pool is evaluated on the basis of the resulting relevance judgments. The process involves the removal of a system's contribution from the given pool and evaluating the system with the reduced pool. This leave-one-out experiment is carried out for all the valid submissions at INEX 2007 and INEX 2008, and results are compared with those obtained when the complete pool is used for evaluation.

All experiments are driven by a common set of objectives: to find some important pooling parameters such as the minimum number of topics required, and the minimum pool depth required for a query to ensure reliable evaluation, and to determine the most robust and least error-prone evaluation metric that can be used to reliably rank a set of XML retrieval runs.

Pool Sampling

Motivation

As mentioned in the Introduction, unlike other evaluation fora, INEX does not use a fixed pool depth for all queries. Instead, an assessor judges about 600 documents per topic. Our aim is to study how results are affected if the assessment effort is reduced, i.e., when a smaller pool of documents is judged per topic. Naturally, when fewer documents are judged, the absolute values of various metrics will change. If the relative ranks of various runs remain largely unaffected, then the smaller set of assessments can still be used for evaluation. By progressively reducing the pool size, we can estimate the minimum amount of effort that yields results comparable to the current results.

Assessment effort can be reduced in two ways. First, the pool is generated as usual, but the assessors do the judgments on a best-effort basis. In this scenario, the pool for a particular topic may end up being partially judged. In the second case, the reduced pool size is fixed a priori, i.e., a smaller pool is created at the outset, and given to assessors. The following subsections describe experiments that study how evaluation results are affected in these two scenarios.

Random Sampling

The first set of experiments that we did may be taken to correspond to the following scenario. Pools are constructed in the usual way, and distributed to participants, but a participant is not able to assess all documents assigned to her. Can the partial assessments be used for evaluation? To simulate this situation, a random fraction of the qrels is discarded—these entries are regarded as unjudged and therefore assumed to be nonrelevant—and the reduced qrels are used for evaluation. Our aim here is to see how small the random sample can be so that overall evaluation reliability is not compromised for a given set of queries.

Experiments. The experiment is designed as follows. First, 80% of the relevant documents for each query are selected

TABLE 1. INEX 07: Stability of system rankings for random pool sampling (107 topics, 78 systems, 6,460 relevant documents in 100% qrels).

Pool % (# reldocs)	iP[0.00] Avg.		iP[0.01] Avg.		iP[0.05] Avg.		iP[0.10] Avg.		MAiP Avg.	
	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}
20 (1,298)	0.65810	0.64635	0.62234	0.66353	0.67128	0.68743	0.72921	0.73541	0.82164	0.80468
40 (2,589)	0.74463	0.74416	0.72792	0.76200	0.80305	0.78626	0.84945	0.84263	0.90080	0.88253
60 (3,871)	0.82469	0.81336	0.80086	0.81458	0.84878	0.83947	0.88492	0.87710	0.93062	0.91990
80 (5,164)	0.88633	0.88824	0.87556	0.88596	0.90913	0.90385	0.93584	0.92027	0.96450	0.95152

TABLE 2. INEX 08: Stability of system rankings for random pool sampling (70 topics, 61 systems, 4,887 relevant documents in 100% qrels).

Pool % (# reldocs.)	iP[0.00] Avg.		iP[0.01] Avg.		iP[0.05] Avg.		iP[0.10] Avg.		MAiP Avg.	
	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}
20 (971)	0.53104	0.53407	0.55126	0.54982	0.69268	0.67718	0.72940	0.70835	0.82131	0.79022
40 (1,942)	0.62918	0.62809	0.66383	0.65540	0.71934	0.72270	0.76536	0.76021	0.86557	0.84114
60 (2,918)	0.71716	0.72525	0.74754	0.75644	0.82448	0.81782	0.83628	0.82575	0.91268	0.88951
80 (3,889)	0.82273	0.81393	0.84885	0.84080	0.87574	0.87112	0.88601	0.87057	0.94404	0.92422

at random from the original qrels without replacement.[3] All ad hoc focused runs from both INEX 2007 and 2008 are evaluated using this reduced set of assessments, and ranked on the basis of each metric in turn. Rank correlation (both τ and τ_{AP}) values are computed between these new rankings, and the ranking produced by the corresponding metric with the original (100%) pool. The process is repeated with 10 different random samples. The entire exercise is then repeated at 60%, 40%, and 20% sampling levels.[4] From the INEX 2007 ad hoc focused task, 107 topics and 78 runs were used, while from the INEX 2008 focused task, there were 70 topics and 61 runs.

These experiments also address a flaw in our earlier experiments reported in Pal et al. (2008), where random samples were chosen directly from the entries in the qrels. Because each entry specifies relevance for a single element, it was possible for a sample to include a relevant element from a document, but exclude another relevant item from the same document (which would then be regarded as nonrelevant). This is an unrealistic situation because judgments are done one document at a time, rather than one element at a time, i.e., an assessor is given a whole document for assessment, and she or he highlights all the relevant passages/elements in it. Thus, given a particular document, all its relevant items should either figure in the pool, or be excluded from the pool.

Results. The means of the τ and τ_{AP} values across 10 random samples for each sampling level are shown in Tables 1

and 2. The same values along with the standard error at each sampling level are shown in Figure 1.

For all the graphs, as the sampling level decreases, the correlation between the original rankings produced by a metric and the rankings obtained with reduced assessments decreases in general, so each of the curves droops. One obvious reason is that with reduced assessments, the precision score is affected nonuniformly across the systems, depending upon the ranks of retrieved relevant texts that are missing in the reduced pool. This phenomenon leads to changes in comparative ranks. Further, Kendall τ drops for iP[0.00] and iP[0.01] at a much faster rate than it does for iP[0.05], iP[0.10], or MAiP. Among the metrics, MAiP clearly shows the least variation in τ values across different pool sizes and across the samples at a particular pool size.

Error bars for each curve tend to increase as pool size reduces. The reason can be attributed to the fact that at smaller pool sizes, the overlap among the samples reduces. This affects the precision scores of different systems in a very irregular fashion. This irregularity causes widely varying system rankings across the samples leading to wide variation in τ.

τ_{AP} values are on the whole, slightly lower than the corresponding τ values, indicating that the metrics cause more ranking errors among the top-performing systems than among the low-performing systems when a reduced pool is used.

On a closer look, the curves in Figure 1(a) (INEX 2007) look smoother and more regular compared to their INEX 2008 counterparts [Figure 1(c)]. There are two reasons for this: (a) the INEX 2007 dataset contains a larger number of valid runs (78 compared to 61 for INEX 2008), and (b) the INEX 2007 qrels consist of a greater number of queries than the INEX 2008 qrels (107 topics compared to 70). The changes in τ values are smoothed out through averaging over

---

[3]Though the original qrels contain assessed nonrelevant units as well, these entries do not figure during the computation of precision scores, and are therefore ignored in these experiments.

[4]For a few topics there were less than five relevant documents. For such topics, one relevant document was included in the reduced qrels at 20% sampling level.
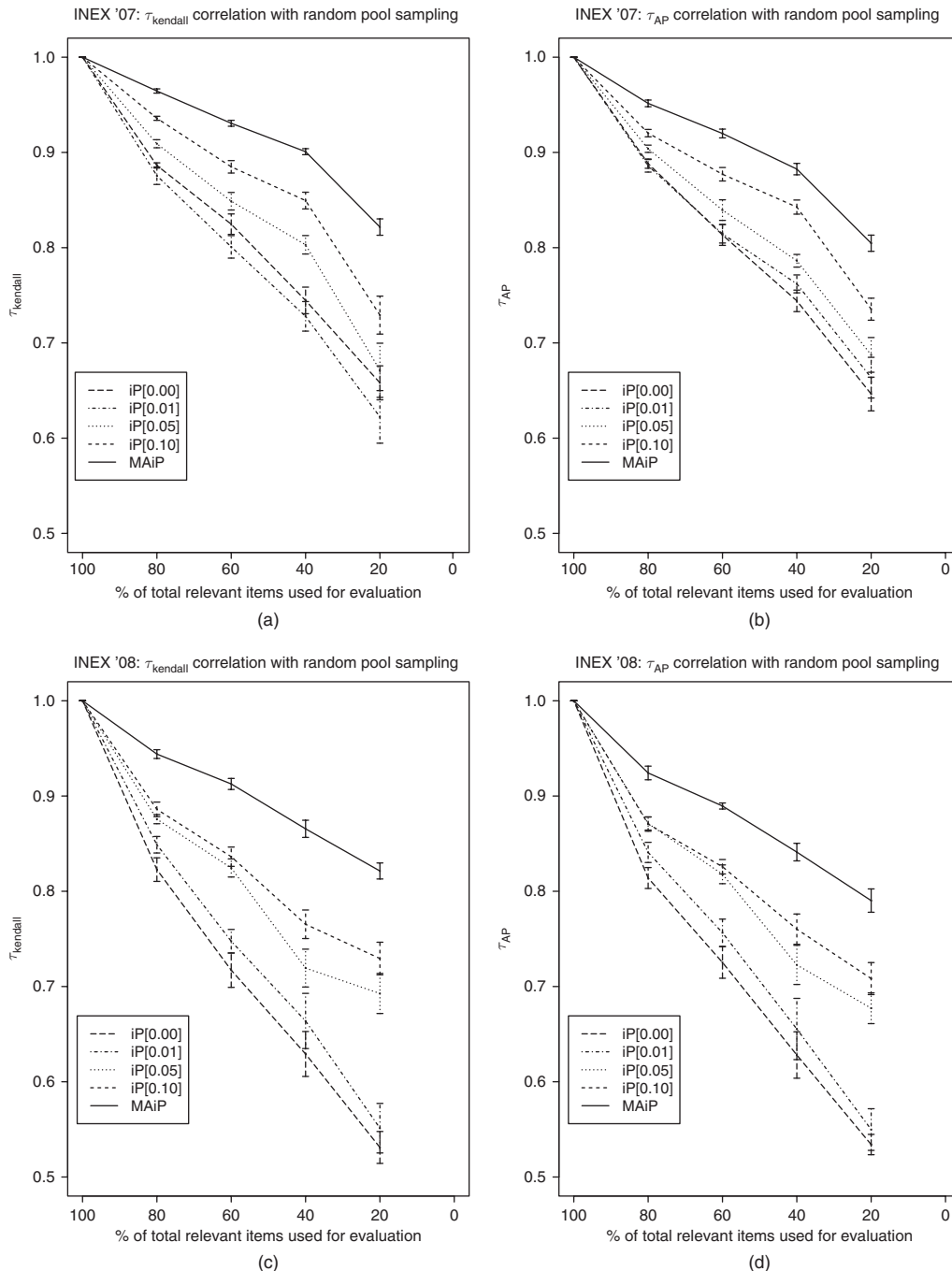
FIG. 1. Rank correlation between original system rankings and rankings obtained with randomly reduced pools.

a higher number of systems, resulting in smoother curves. Similarly, rankings disagree to a greater extent when a smaller number of queries is involved. This leads to a sharp fall in τ values, which is particularly acute for the early precision metrics (iP[0.00] and iP[0.01]) in 2008 [see Figure 1(c)].

In summary, with about 50% sample of the qrels, system rankings at INEX 2007 and INEX 2008 are not significantly affected ($\tau \geq 0.9$) if MAiP is used as the ranking metric. If iP[0.10] is used to rank systems, τ remains above 0.9 for 65% samples in case of INEX 2007, and an 83% pool for INEX 2008. For the other metrics, even a 20% random reduction

in the judged pool results in significant ranking changes (see Table 3).

Reducing Pool Depth

Our second goal is to estimate the minimum pool size that can be used to reliably evaluate a set of runs. First, pools of varying sizes are generated from a set of submissions by varying the pool depth for each query. Note that, in these experiments, a smaller pool is always a proper subset of a bigger pool. The maximum possible pool size is limited by

TABLE 3. Minimum number of reldocs required in the random pool to get $\tau \geq 0.9$ (INEX 2007: 107 topics, 6,460 relevant documents in 100% pool; INEX 2008: 70 topics, 4,887 relevant documents in 100% pool).

Metric	INEX 2007	INEX 2008
iP[0.00]	>5,164 (>80%)	>3,889 (>80%)
iP[0.01]	>5,164 (>80%)	>3,889 (>80%)
iP[0.05]	>3,871, <5164 (~75%)	>3,889 (>80%)
iP[0.10]	>3,871, <5164 (~65%)	>3,889 (>80%)
MAiP	<2,589 (<40%)	<2,918 (<60%)

the size of the original pool. Submissions are then evaluated on the basis of assessments generated from the reduced pools.

Experiments. Because some submission files were changed by participants after the pooling process was completed, we were not able to exactly replicate the original pool. We therefore take as our starting point a pool created from all valid and invalid submissions (98 for INEX 2007 and 76 for INEX 2008) in the focused category only. To create this pool, we guess the pool depth (d_Q) for each topic Q as follows: d_Q is taken to be the minimum depth at which the number of distinct documents in the generated pool is greater than (or equal to) the original pool size for Q. The restriction of the original qrels to this generated pool is taken to be the initial (or 100%) qrels.

Although this is actually a subset of the original qrels, it is a close clone (both Kendall's τ and τ_{AP} for system rankings obtained using the original qrels and our simulated 100% qrels are over 0.99 in most cases, with minimum value being 0.97).

The reduced pools are also created in a similar way. The $X\%$ pool ($X = 5, 10, 20, \ldots, 90$) is generated by first guessing an appropriate pool depth $[d_Q(X)]$ for each topic Q. At this pool depth, the pool size for Q equals (or just crosses) $X\%$ of the original pool size for Q. We refer to the corresponding qrels as the $X\%$ qrels.

All valid submissions are evaluated with the reduced qrels, and the correlation between the rankings obtained using the $X\%$ and 100% qrels—measured using Kendall's τ as well as τ_{AP}—are computed for each of the INEX metrics.

Results. Tables 4 and 5 show the variation in τ and τ_{AP} as X varies from 5 to 90. Both τ and τ_{AP} are very high (over 0.9 in almost all cases even at $X = 20\%$). This shows that system rankings are not significantly affected if a shallower pool is used for evaluation. Thus, the system rankings obtained using any of the INEX evaluation metrics is reasonably reliable even when just 20% of the original pool size is assessed. The fact that τ and τ_{AP} are in close agreement further signifies that ranking changes are more or less uniformly distributed over the entire ranked list, irrespective of the metric used for ranking.

Figure 2 displays the same information graphically. Two trends are visible from the graphs: (1) In general, correlation

values decrease from iP[0.00] to iP[0.10]; and (2) correlation decreases as shallower pools are used for evaluation.

Precision values at early recall levels are generally determined by top-ranked documents. Even when shallow pools are used, these top-ranked documents are usually included in the smaller pool. The assessments for these top-ranked documents are therefore mostly unaffected whether we use a shallow pool or the original pool. Thus, precision values at early recall levels do not change much as pool size is reduced, and the curves for iP[0.00] and iP[0.01] are relatively flat. On the other hand, a relevant document that is not highly ranked by any of the systems will be excluded from the pool at smaller pool depths. Such a relevant document will then remain unjudged, and will therefore be considered nonrelevant. If the rank of such a relevant document varies across systems (as is likely), the precision values at higher recall levels will be affected differently for different runs, and their relative ranks may change, leading to a drop in τ or τ_{AP}.

As pool size decreases, the number of such documents—which were marked relevant in the original qrels, but are regarded as nonrelevant in the reduced qrels—increases, leading to greater discrepancies in ranking. MAiP is an average of precision values at 101 recall levels, and is thus affected to an intermediate degree when pool size changes.

Random Sampling Versus Reducing Pool Depth

Of the two methods that were tried in this section, reducing pool depth is clearly the more logical, systematic, and safe way to obtain smaller pools. However, in some cases, the judged pool may be small due to forces of circumstance, rather than by design. For example, at TREC, CLEF (Cross-Language Evaluation Forum), and FIRE (Forum for Information Retrieval Evaluation), the pool of documents to be judged is given to an assessor in order of document IDs. This is done to avoid any potential bias of assessors against low-ranked documents. In such a scenario, if an assessor ends up partially judging a topic, can the judgments be of some use? In the Random Sampling section we attempt to give a quantitative answer to this question. This approach also provides a baseline that highlights the usefulness of reducing pool size by reducing pool depth.

Query Sampling

Motivation

Because INEX does not have a dedicated pool of assessors and assessment is done by participants, some topics end up not being assessed completely, or at all. In this section, we investigate the effect of eliminating queries from the topic set used for evaluation, or equivalently, of using smaller topic sets for evaluation.

Experiments

The set-up for this task is quite similar to that for random pool sampling. First, a random 80% sample of the total

TABLE 4.   INEX 07: Stability of system rankings on reducing pool depth (107 topics, 78 systems, 5,610 relevant documents in 100% qrels).

Pool % (# reldocs)	iP[0.00]		iP[0.01]		iP[0.05]		iP[0.10]		MaiP	
	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}
5 (1,123)	0.89426	0.87066	0.85995	0.87111	0.84251	0.79643	0.85910	0.81760	0.91543	0.89109
10 (1,667)	0.94394	0.90673	0.91312	0.91531	0.86520	0.81984	0.88807	0.87771	0.94363	0.91281
20 (2,497)	0.97432	0.95700	0.93247	0.93669	0.90621	0.86521	0.91579	0.89033	0.95781	0.95560
30 (3,184)	0.98432	0.98250	0.95265	0.94696	0.93224	0.91840	0.94647	0.92976	0.96848	0.96489
40 (3,757)	0.99049	0.98733	0.95697	0.95258	0.94546	0.93777	0.96164	0.95646	0.98182	0.98316
50 (4,247)	0.99232	0.98894	0.95697	0.95343	0.95662	0.94926	0.97098	0.96050	0.98766	0.98420
60 (4,697)	0.99516	0.99464	0.96898	0.96580	0.96431	0.96574	0.97765	0.96333	0.99166	0.99219
70 (5,108)	0.99750	0.99643	0.97515	0.97138	0.97565	0.97408	0.98532	0.98273	0.99366	0.99402
80 (5,375)	0.99900	0.99784	0.97849	0.97895	0.97965	0.98495	0.98716	0.98741	0.99767	0.99711
90 (5,520)	0.99933	0.99827	0.98783	0.99030	0.98815	0.99149	0.99283	0.99178	0.99900	0.99857

TABLE 5.   INEX 08: Stability of system rankings on reducing pool depth (70 topics, 61 systems, 4,667 relevant documents in 100% qrels).

Pool % (# reldocs)	iP[0.00]		iP[0.01]		iP[0.05]		iP[0.10]		MaiP	
	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}
5 (850)	0.94098	0.92886	0.86448	0.84909	0.88962	0.88182	0.83388	0.82205	0.84809	0.83739
10 (1,307)	0.96721	0.93479	0.90055	0.85759	0.90820	0.88480	0.81421	0.82937	0.88525	0.87316
20 (1,992)	0.98033	0.94805	0.93552	0.92756	0.92459	0.89686	0.92022	0.91679	0.92787	0.92271
30 (2,530)	0.98798	0.95712	0.94536	0.93295	0.94754	0.91241	0.92459	0.90303	0.94645	0.93782
40 (2,960)	0.99126	0.99187	0.95956	0.94695	0.94754	0.92723	0.93552	0.92104	0.95738	0.94647
50 (3,386)	0.99016	0.99118	0.96940	0.95657	0.95847	0.93234	0.94536	0.92924	0.96831	0.95972
60 (3,754)	0.99235	0.99254	0.97486	0.96224	0.98033	0.97733	0.96503	0.95615	0.97924	0.97223
70 (4,094)	0.99672	0.99623	0.98470	0.96899	0.97486	0.96283	0.97049	0.96344	0.99017	0.98809
80 (4,380)	0.99781	0.99845	0.99126	0.97566	0.98579	0.97958	0.97159	0.96925	0.99891	0.99928
90 (4,584)	1.00000	1.00000	0.99781	0.98258	0.99672	0.99452	0.99563	0.99580	0.99891	0.99921

set of queries in the qrels (107 for INEX 2007 and 70 for INEX 2008) is selected. For each selected topic, all available assessment information is considered. Once again, correlation is measured between the system rankings produced by each metric using the complete set of queries and the reduced query set. The process is repeated for 10 random samples. The whole exercise is repeated with 60%, 40%, and 20% of the query set.

Results

The behavior of the metrics as query-set size varies is shown in Figure 3. The curves exhibit the same drooping nature as the query-set size is progressively reduced. Early precision measures (iP[0.00] and iP[0.01]) perform poorly compared to late precision measures (iP[0.05] and iP[0.10]). MAiP emerges as a clear winner both in terms of its resilience to the reduction in size of the topic set and variation across samples (smallest error bars).

Further, τ_{AP} values are in general slightly smaller than τ for MAiP and iP[0.10] for both INEX 2007 (Figures 3(a) and 3(b)) and INEX2008 [Figures 3(c) and 3(d)], but this trend is not so prominent for early precision metrics. This again indicates that top-performing systems are more affected than low performers by the reduction in topic set size. However, such

an interpretation of the τ and τ_{AP} values should be treated with caution (Carterette, 2009). There is one straightforward explanation for this phenomenon. The top-ranked systems at both INEX 2007 and INEX 2008 are very similar (see Figure 4, for example). Thus, for iP[0.01], there is no statistically significant difference in the performance of the top 10 runs at INEX 2007 (Führ et al., 2008). Paired t-test results for the top 10 systems at INEX 2008 focused task are also similar (Kamps et al., 2009). Because these top-ranked systems are so similar to each other, even small changes in the qrels can cause swaps in the relative positions of the top-ranked systems. It is therefore expected and quite reasonable that the drop in τ_{AP} is more than the drop in Kendall's τ under reduced judgments.

Between the 2 years, the curves for INEX 2007 are flatter (i.e., correlation with the official system rankings drops more slowly) than their INEX 2008 counterparts, one of the reasons being that INEX 2007 has a larger number of participating runs (78 compared to 61). Moreover, INEX 2007 curves also have smaller error bars (range of τ values is smaller) compared to INEX 2008 curves. This is most likely due to a higher number of queries at the same percentage point (the INEX 2007 topic set has 107 queries, against 70 queries in the INEX 2008 set). One important observation is that MAiP-based rankings remain largely unchanged
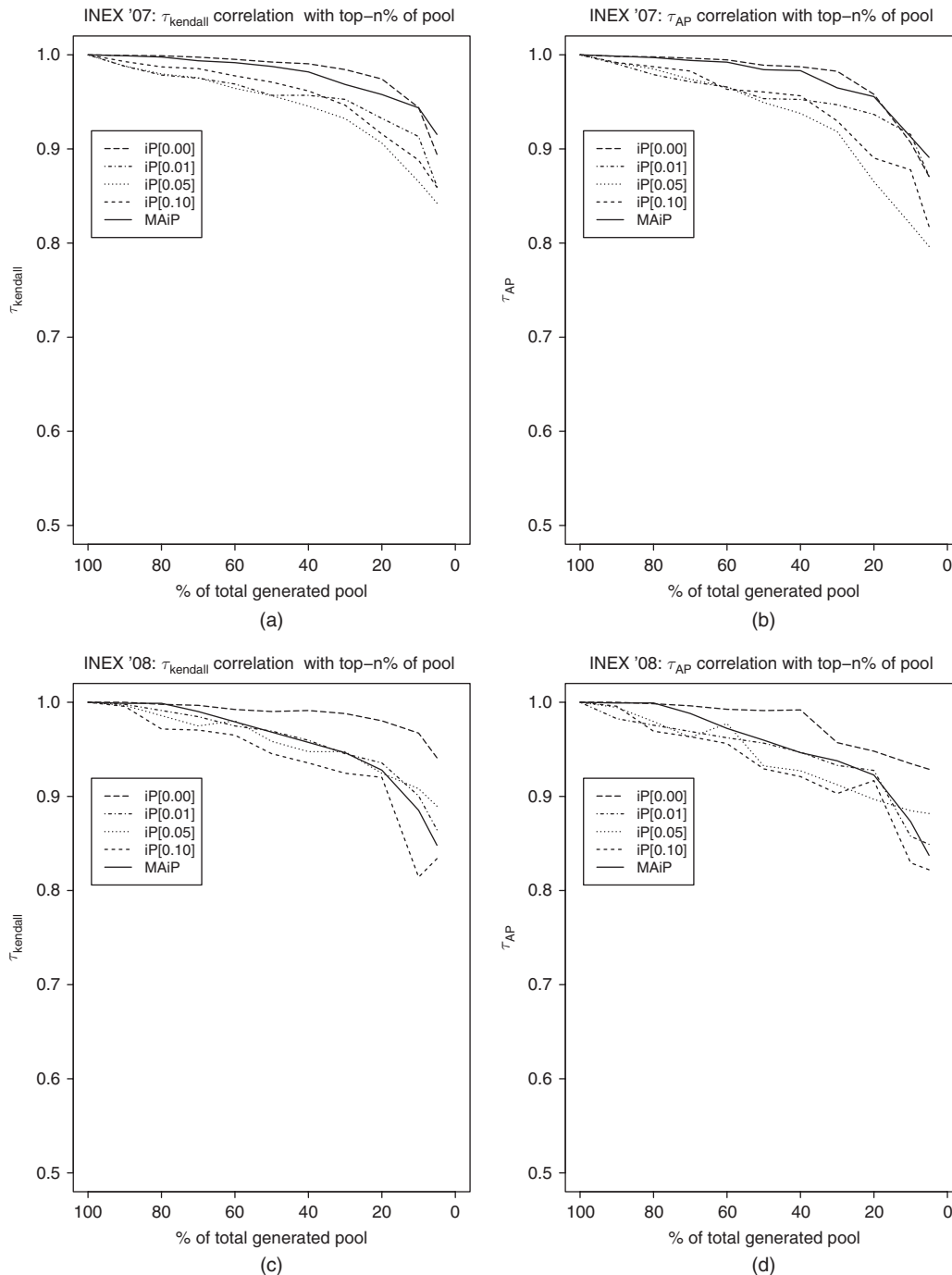
FIG. 2.   Rank correlation between original system rankings and rankings obtained with reduced pools.

($\tau \geq 0.9$) even with only 37 queries (about 35% of the original number) for INEX 2007, and 32 queries (45%) for INEX 2008.

Table 6 shows the minimum number of queries required to obtain a τ value of 0.9 or greater with the original ranking, when ranking is done on the basis of the various INEX metrics.

The behavior of the various metrics as topic set size changes is studied from a different perspective in the following section.

Error Rates

Motivation

The measured effectiveness of a system depends very much on the query used to measure effectiveness. For a particular set of queries, system A may outperform system B, whereas for a different set of queries, their relative performances can be in the opposite order. However, if two such systems are evaluated using a large number of randomly chosen sets of queries, then it is expected that the "truly
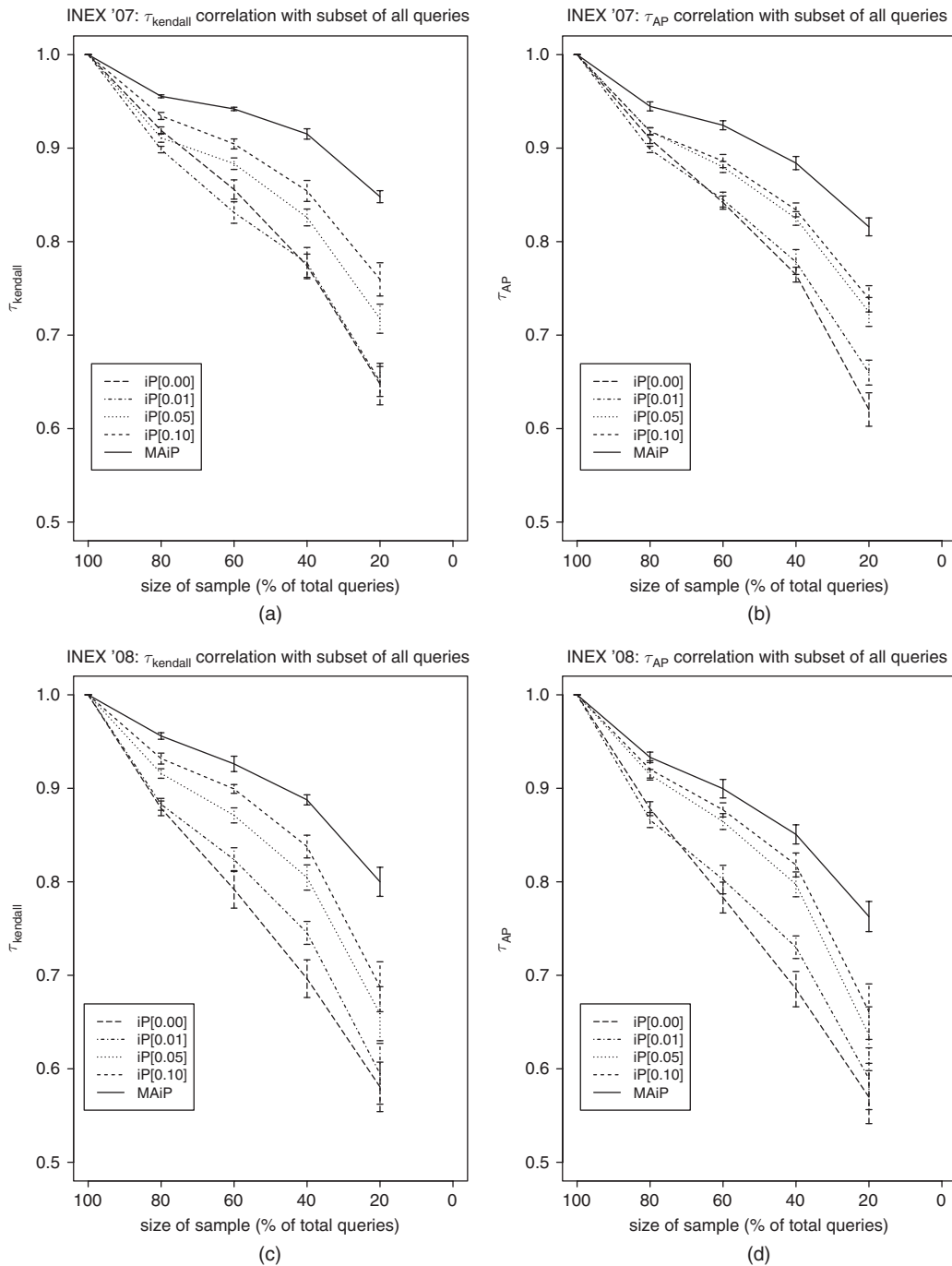
FIG. 3.   Rank correlation between original system rankings and rankings obtained with randomly reduced query sets.

better" system will outperform the other for a majority of the query sets. The remaining cases, where the better system performs worse, can be regarded as errors. How well a metric captures the intrinsic quality of systems is reflected in how often it leads to an erroneous conclusion when used to compare two systems. The fewer the errors, the better is the metric. The motivation behind the experiments reported in this section was to study the error rates of various metrics at different topic-set sizes, and then to estimate the minimum number of topics required to keep the error rate within a stipulated limit. These experiments are based on the work

of Voorhees and Buckley (2002). The results reported here present a much more extensive and comprehensive picture than the preliminary experiments reported in (Pal et al., 2008).

Computing Error Rates

The basic procedure to compute the error rate is as follows. We take two retrieval systems A and B, and two disjoint topic sets of equal size z, and compute the value of a particular evaluation metric for each system–topic-set pair. The mean scores of the two systems are compared for each topic set.
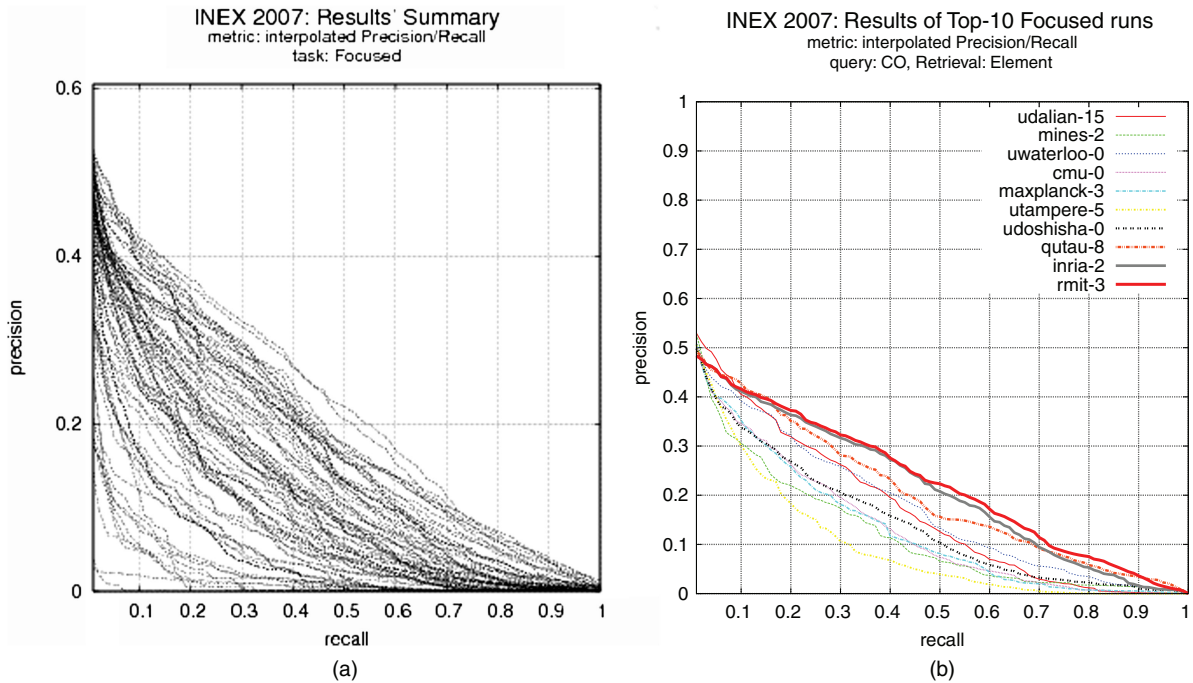
FIG. 4.   INEX 07 Precision-Recall curves for focused runs and top-10 systems.

TABLE 6.   Number of queries required to get $\tau \geq 0.9$.

Metric	INEX 2007	INEX 2008
iP[0.00]	86 (80%)	59 (84%)
iP[0.01]	86 (80%)	58 (83%)
iP[0.05]	75 (70%)	53 (75%)
iP[0.10]	59 (55%)	44 (62%)
MAiP	37 (35%)	32 (45%)

We then check whether the topic sets agree as to which of the runs is better. If they do not agree, i.e., A's score is higher than B's by at least a minimum margin p on one topic set, but B's score is higher (by at least the same minimum margin) on the other set, we mark this as a swap (or disagreement). By repeating the exercise n times with different topic sets of the same size, we calculate the proportion of swaps for a particular pair of runs. The average proportion of swaps over all possible pairs of runs is called the error rate for that particular topic-set size. The core of the algorithm for calculating error rates is similar to that in Voorhees and Buckley (2002).

The minimum topic-set size we take is five. The maximum size is roughly half of the number of topics available in the qrels (50 for INEX 2007 and 35 for INEX 2008). The number of iterations (n) is 50. We consider five different values for the tolerance p ($p = 0$, 5%, 10%, 20%, 30%). The measures M considered are the five official measures, viz. iP[0.00], iP[0.01], iP[0.05], iP[0.10], and AiP. For INEX 2007, there are $\binom{78}{2} = 78 \times 77/2 = 3{,}003$ pairwise comparisons (though there were 79 systems, two of

the systems were identical), and for INEX 2008, there are $\binom{61}{2} = 61 \times 60/2 = 1{,}830$ pair-wise comparisons for each topic-set size ranging from 5 to 50 and 5 to 35, respectively, for five different measures at five different percentage points. The whole exercise leads to a set of error curves, based on the error rates actually computed by the above algorithm.

Extrapolating to Larger Topic-Set Sizes

As explained above, error-rates can be experimentally calculated for topic-sets that contain at most half the total number of queries in the qrels. The error rate versus topic-set size graphs (see Figure 5) are initially plotted from these empirically determined error rates, and then extrapolated to estimate the error rates for larger topic sets. For each line, we fit a curve to the observed data using the FUDGIT package (Lacasse, 2001). As observed by Voorhees and Buckley (2002), the data seems to follow an "exponential decay" family of functions, given by the following equation:

$$Y = A_1 \cdot \exp(-A_2 \cdot X) \tag{3}$$

where Y is the error rate, and X is the size of the topic set ($X \in \{5, 6, \dots, 100\}$ for INEX 2007, and $X \in \{5, 6, \dots, 70\}$ for INEX 2008). A_1 and A_2 are parameters to be estimated using the observed values of Y for $X \in \{5, 6, \dots, 50\}$ (INEX 2007) or $X \in \{5, 6, \dots, 35\}$ (INEX 2008). We rewrite Equation 3 as

$$\ln Y = \ln A_1 - A_2 \cdot X \tag{4}$$

and fit a linear least squares regression model to the observed data corresponding to each line. To check goodness-of-fit,
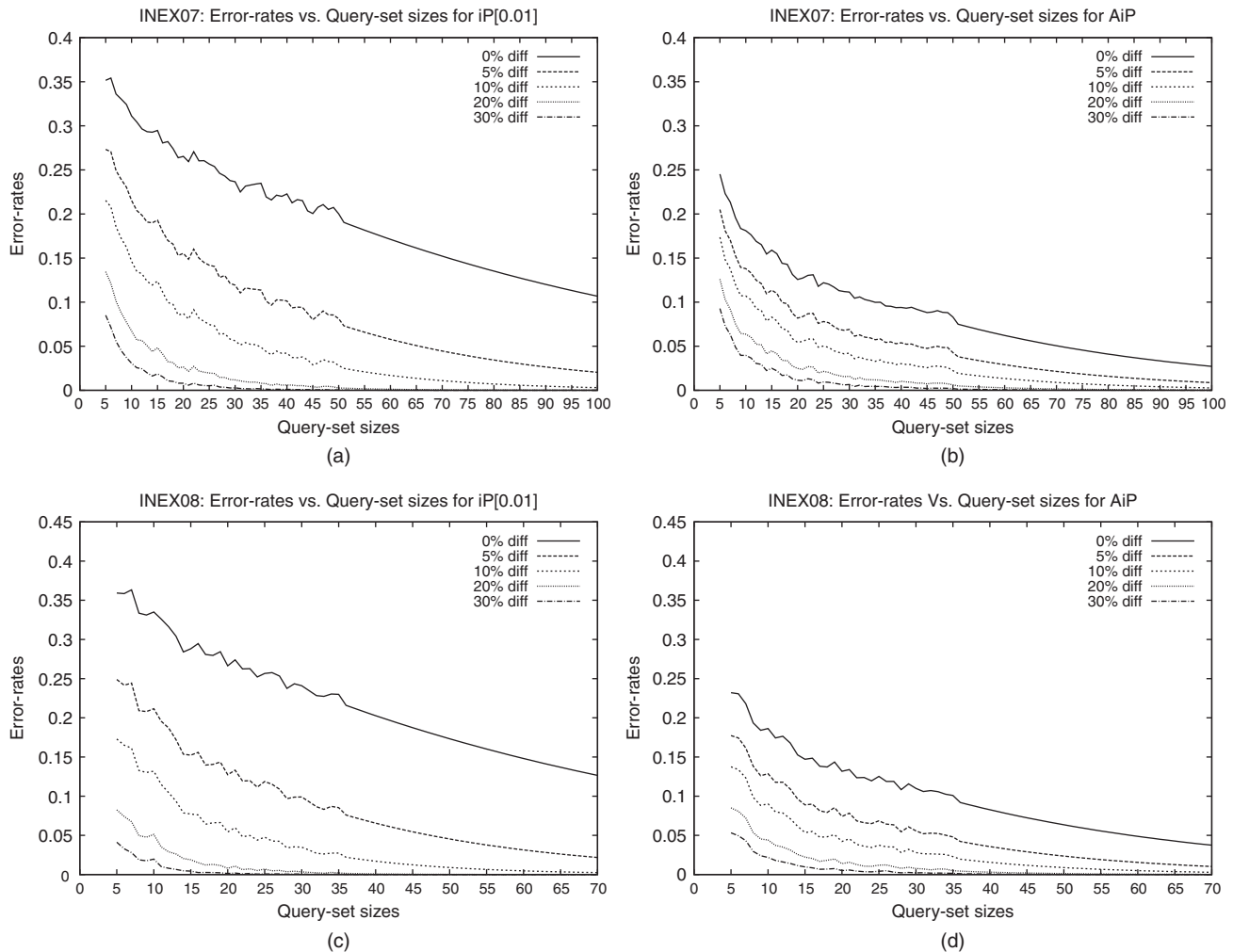
FIG. 5. Error-rates for *iP[0.01]* and *AiP* as query-set size changes (INEX 2007 and 2008).

two values for the parameters are also calculated and observed to be in the range 0 to 9.13 for both INEX 2007 and INEX 2008. The maximum allowable value for $\chi^2_{0.995,50}$ (degrees of freedom for both INEX 2007 and 2008 = number of iterations, 50) is 27.991. The model thus fits the data with at least 99.5% confidence.

Results

Figure 5 shows the results of our experiments on error rates. Because disjoint topic sets are used, the obtained error rates are the upper bounds of the error rates for the concerned metric (Sanderson & Zobel, 2005).[5] The initial part of each line is plotted based on observed error-rate values; it is then extrapolated as explained above. The error-rate plots are shown only for iP[0.01] and AiP. The error rates for iP[0.00], iP[0.05] and iP[0.10] are similar: although the error rates for iP[0.00] are slightly higher than those for iP[0.01], the

iP[0.05] and iP[0.10] curves lie in between the corresponding iP[0.01] and AiP curves. The graphs exhibit the following trends, as expected.

1. Error rates are maximum when the tolerance is 0%, and fall off as the tolerance increases.
2. Error rates are generally high with smaller query sets, and progressively decrease as query-set size increases.
3. Error rates are higher for the early precision-metrics (iP[0.00], iP[0.01]), and least for MAiP.

Though the curves for INEX 2007 and INEX 2008 follow the same pattern, the INEX 2008 curves have higher error rates for the same topic-set size because of the smaller number of total runs.

Table 7 shows the approximate minimum topic-set sizes for which error rates are 5% or less, for the various metrics. For both INEX 2007 and INEX 2008, an error rate of less than 5% with 0% tolerance is only achievable with MAiP as the metric, and indeed, for the topic-set size used at INEX 2007 and INEX 2008, the error rate for MAiP is consistently less than 5% for all tolerance values considered

---

[5]In contrast, our earlier work on error rates (Pal et al., 2008) reported minimum error rates.

TABLE 7. Minimum number of topics required to guarantee less than 5% error.

Data	% tolerance	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MaiP
INEX 2007	0	>100	>100	>100	>100	**70**
	5	60	65	65	56	**45**
	10	35	35	35	35	**25**
	20	13	15	15	17	**13**
	30	7	7	8	10	**7**
INEX 2008	0	>70	>70	>70	>70	**58**
	5	48	46	43	40	**35**
	10	22	23	22	22	**18**
	20	8	10	10	10	**9**
	30	<5	<5	<5	6	**5**

in our experiments. In comparison, the INEX results based on the official metric iP[0.01] can be taken to contain fewer than 5% errors when $p = 5\%$ or higher, i.e., if we regard a performance difference of 5% or less as insignificant.

Leave-One-Out

Motivation

One of the assumptions underlying pooling-based evaluation is completeness: the contributing systems are together successful in finding all relevant items. Thus, an unjudged document is assumed to be nonrelevant. If we use a set of assessments to evaluate a technique that did not contribute to the underlying pool, and the technique actually succeeds in finding unjudged, relevant documents that were not found by any of the systems contributing to the pool, the new technique does not get credit for this. Its effectiveness can thus be underestimated. In the current set of experiments, we try to quantify the extent of this problem within the INEX setting.

Experiments

These experiments are inspired by the work of Zobel (1998) on TREC data. Given a set of N submissions that contribute to a given pool, if each run's contribution to the pool is removed in turn, i.e., the pool is constructed on the basis of the remaining $N - 1$ runs and the qrels corresponding to this reduced pool are used to evaluate the system, then the difference between the original and new scores can give us an idea about the robustness of the pool. However, there are two practical issues here. One, it is observed that runs from a particular group are quite often similar in strategy. Second, the INEX ad hoc pool is made of runs from three separate subtasks. It is also a fact that participants tend to submit variants of runs to each subtask. Thus, leaving one focused run out of the pool, while including other related submissions from the same participant might not change the pool much, and the stability of the metric may be overrated when the difference between the original and new score is computed.

We therefore follow the suggestion of Büttcher et al. (2007), and "leave-one-group-out" at a time in our experiments. We start with the 100% pool reconstructed from the INEX focused submissions (see the Reducing Pool-Depth subsection) as our baseline. For each participating group, a pool is recreated by excluding all the focused runs from that group. Each excluded run is then evaluated using the assessments generated from this pool. We compare the new scores with those computed from baseline qrels for each query and for each run. We also check whether the overall change in score for a particular run is significant, using a paired t-test.

Results

Table 8 and Table 9 summarize the distribution of relative changes in scores. Table 8 summarizes the distribution of relative changes in overall scores for the participating systems. A negative (positive) change signifies that the new score is higher (lower) than the baseline score. As explained above, the effectiveness of a system that does not contribute to the assessed pool may, in general, be underestimated. Thus, for most systems, the score obtained with the leave-one-group-out pool is somewhat lower than that obtained with the 100% pool [see the row corresponding to the (0,10)% change in score].

On a closer examination of these results, we found that the effect of a system's contribution to the pool is query-specific, and is not uniform across queries. The changes in overall system scores shown in Table 8 hide these interesting details. In Table 9, therefore, we report the distribution of relative changes in scores across all possible query-run pairs.

As above, for a number of queries, the score of a system drops when the leave-one-group-out pool is used. In the vast majority of cases, however, the score remains unaffected, no matter what pool is used.

On the whole, the results are reassuring, as they suggest that new systems may be reliably evaluated using the INEX qrels. The actual number of queries for which there are no changes decreases from iP[0.00] to MAiP. This suggests that the top-ranked items retrieved by any system are also retrieved by at least one other system, although possibly at different ranks. As one goes down the ranked list, the number of unique contributions to the pool increases. Thus, precision scores at higher recall levels are affected for more queries.

TABLE 8. Distribution of relative change in scores for "leave-one-group-out" (overall).

Data	% change	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MaiP
INEX 2007	(−∞, −50]	0	0	0	0	0
	(−50, −10]	0	0	0	0	0
(out of 78	(−10, 0)	0	2 (2.6%)	2 (2.6%)	3 (3.8%)	4 (5.1%)
systems)	0	13 (16.7%)	9 (11.5%)	7 (9%)	8 (10.3%)	8 (10.3%)
	(0, 10]	63 (80.7%)	67 (85.9%)	69 (88.4%)	67 (85.9%)	66 (84.6%)
	(10, 20]	1 (1.3%)	0	0	0	0
	(20, 50]	1 (1.3%)	0	0	0	0
	(50, 100]	0	0	0	0	0
INEX 2008	(−∞, −50]	0	0	0	0	0
	(−50, −10]	0	0	0	0	0
(out of 61	(−10, 0)	0	2 (3.3%)	6 (9.8%)	8 (13.1%)	7 (11.5%)
systems)	0	18 (29.5%)	7 (11.5%)	4 (6.6%)	4 (6.6%)	0
	(0, 10]	43 (70.5%)	50 (81.9%)	51 (83.6%)	49 (80.3%)	54 (88.5%)
	(10, 20]	0	2 (3.3%)	0	0	0
	(20, 50]	0	0	0	0	0
	(50, 100]	0	0	0	0	0

TABLE 9. Distribution of relative change in scores for "leave-one-group-out" (per-query-level).

Data	% change	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MaiP
INEX 2007	(−∞, −50]	0	2	4	6 (0.1%)	1
	(−50, −10]	0	5	7 (0.1%)	13 (0.2%)	24 (0.3%)
(out of 78	(−10, 0)	0	6 (0.1%)	17 (0.2%)	27 (0.3%)	368 (4.4%)
systems)	0	7,890 (95%)	7,736 (93.1%)	7,612 (91.6%)	7,580 (91.2%)	6,869 (82.7%)
	(0, 10]	191 (2.3%)	326 (3.9%)	451 (5.4%)	494 (6.0%)	727 (8.8%)
	(10, 20]	61 (0.7%)	74 (0.9%)	79 (1.0%)	92 (1.1%)	135 (1.6%)
	(20, 50]	77 (0.9%)	94 (1.1%)	91 (1.1%)	60 (0.7%)	117 (1.4%)
	(50, 100]	88 (1.1%)	64 (0.8%)	46 (0.5%)	35 (0.4%)	66 (0.8%)
INEX 2008	(−∞, −50]	0	1	7 (0.2%)	7 (0.2%)	2
	(−50, −10]	0	2	7 (0.2%)	5 (0.1%)	30 (0.7%)
(out of 61	(−10, 0)	0	1	9 (0.2%)	20 (0.5%)	284 (6.7%)
systems)	0	4,135 (96.8%)	4,049 (94.8%)	3,955 (92.6%)	3,911 (91.6%)	3,395 (79.5%)
	(0, 10]	84 (1.9%)	146 (3.4%)	217 (5%)	253 (5.9%)	465 (10.9%)
	(10, 20]	15 (0.4%)	26 (0.6%)	26 (0.6%)	38 (0.9%)	47 (1.1%)
	(20, 50]	17 (0.4%)	33 (0.8%)	34 (0.8%)	17 (0.4%)	27 (0.6%)
	(50, 100]	19 (0.5%)	12 (0.3%)	15 (0.4%)	19 (0.4%)	20 (0.5%)

More interesting is the fact that the performance of some systems actually improves when their contributions to the pool are omitted. This initially appears to be counter-intuitive. However, the following (somewhat extreme) example shows how this may happen. Consider a query for which there are two relevant documents. The first document is retrieved at rank one by all systems; the second is retrieved by only one system at rank R ≫ 1. The AP for this system is $1/2(1 + 2/R) = 1/2 + 1/R$. When the system's contribution to the pool is omitted, there is only one relevant document for the given query, and the system gets a perfect score of 1.

On a closer examination of the results for individual queries, we also find that such an improvement can occur in another situation that is peculiar to the focused retrieval task (as opposed to the document retrieval task). Consider a system that retrieves only nonrelevant passages/elements from a document that contains relevant material. If this system is the only one to contribute this document to the pool,

its performance will improve when this contribution is omitted from the pool.

We also used a two-sided, paired t-test to check whether the overall change in score for each run is significant. Table 10 shows the number of runs for which the overall score is significantly affected, when the corresponding group's contributions to the pool are left out.

These results show that the early precision metrics are significantly affected for only a small number of systems. This can be accounted for in two ways. First, as stated above, most of the documents returned at low pool depth are also retrieved by other systems. By and large, this consensus decreases as one goes down the ranked list. Because the MAiP score is calculated using the entire ranked list, this metric is most significantly affected when a system's contribution is omitted from the pool.

Second, because the early precision metrics are relatively unstable, even large differences between two systems may

TABLE 10. Number of runs with significant difference in score (level of significance 0.05).

Data	Significant diff.	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MaiP
INEX 2007	YES	8 (10.2%)	14 (17.9%)	24 (30.8%)	15 (19.2%)	27 (34.6%)
	NO	70 (89.8%)	64 (82.1%)	54 (69.2%)	63 (80.8%)	51 (65.4%)
INEX 2008	YES	0	2 (3.3%)	5 (8.2%)	10 (16.4%)	15 (24.6%)
	NO	61 (100%)	59 (96.7%)	56 (91.8)	51 (83.6%)	46 (75.4%)

TABLE 11. Kendall tau values at 60% sampling for INEX 08.

Sampling	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MaiP
Random pool	0.717159	0.747541	0.824481	0.836284	0.912678
Random query	0.791913	0.823716	0.871038	0.899344	0.926120

TABLE 12. Kendall tau values at 60% sampling.

Year	Sampling	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MaiP
2007	Random query	0.855752	0.831144	0.883361	0.904501	0.941910
	Pool depth	0.995161	0.968979	0.964310	0.977652	0.991661
2008	Random query	0.791913	0.823716	0.871038	0.899344	0.926120
	Pool depth	0.992350	0.974863	0.980328	0.965027	0.979235

not be regarded as significant on the basis of a *t*-test. It would be interesting to separate the contribution of these two factors to the results in Table 10.

Discussion

Sampling experiments (both pool and query) were conducted with one common set of objectives: first, to study the relative stability of the INEX metrics, and second, to study the effect of reducing assessment effort on the overall evaluation results. Assessment effort can be reduced by reducing the number of queries judged, or by reducing the number of documents judged per query. The results given in the Pool Sampling section show the effect of reducing the number of documents judged per query, while keeping the number of queries unchanged; whereas in the Query Sampling section, we show the effect of reducing the number of queries, keeping the number of documents per query in the qrels unchanged. In this section, we compare these results to find out the safest way to reduce assessment effort without compromising the reliability of the evaluation results. This question is likely to be of importance as the corpus used at INEX grows significantly in size.

Reducing Pool Size Versus Topic-Set Size

Because the number of documents judged per query is roughly constant (just over 600 articles), a given sample ratio (say $x\%$) corresponds to similar assessment effort for both pool sampling and query sampling.

For example, at INEX 2008, a 60% sample of the pool contains 25,363 ($42,272 \times 0.6$) documents, while a pool corresponding to 60% of the query set contains roughly 25,200 articles ($70 \times 0.6 \times 600$).

Table 11 suggests that, for a given amount of assessment effort, the system rankings obtained with a smaller query set are closer to the original rankings than the rankings obtained when a subset of the documents are judged at random. This is also confirmed by a closer look at the results in the Pool Sampling and Query Sampling sections, which reveals that, in general, the curves for random query sampling are slightly more stable in comparison to their counterparts in random pool sampling (for example, see Figure 1 and Figure 3). One likely explanation for this is that, in the query sampling experiments, if a topic is used for evaluation, the complete relevance judgments for the topic are considered. Thus, unlike in random pool sampling, the query contributes to the precision scores of all systems uniformly; the reduction in τ is caused by the variation of system performance across topics.

However, much higher τ values are obtained when assessment effort is reduced by reducing pool depth, compared to when it is reduced by reducing the total number of queries (see Table 12).

In summary, if one wants to minimize the total amount of assessment effort, it is better to judge shallow pools for many queries, than to judge deep pools for fewer topics. This is in complete agreement with the observations from the document retrieval domain or the recent findings from the TREC Million Query track (Carterette et al., 2008). If, however, assessors are likely to end up partially judging their assigned queries
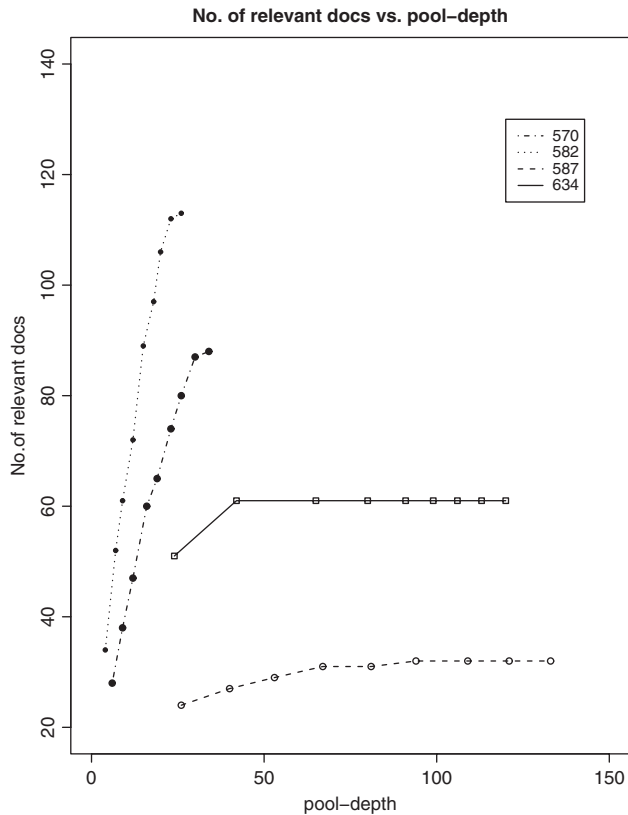
No. of relevant docs vs. pool–depth

FIG. 6. Number of relevant documents vs. pool-depth.

for query 570, pooling is stopped at a depth that is probably less than the critical pool depth. In contrast, there are queries for which the rate of finding a new relevant document is remarkably low (query 587 or 634). For these queries, critical pool depth appears to have been reached much before the pool size reaches the predetermined number of documents to be judged (about 600 in INEX). For query 634, critical pool depth seems to be achieved when the total number of documents judged is near 180, whereas for query 587, this number is 480.

It would be nice if we could balance the assessment effort across queries, by judging fewer documents for queries like 634, and using the manpower thus saved to assess more documents for queries like 582. For a given amount of assessment effort, we would then be likelier to identify more relevant documents, thus obtaining a better estimate of recall.

This is of particular importance as larger corpora begin to be used for evaluation. The following is a proposed pooling strategy that may be used to achieve this effect.

1. Start with a suitably small pool depth, say five, for each query.
2. Create the document pool, judge the pool, and note the number of relevant documents found.
3. Increase the pool depth by five.
4. Repeat steps 2–3 until the rate of finding new relevant documents drops off, or a predetermined maximum number of documents to be judged is reached.

Limitations and Future Work

Our experiments in the Reducing Pool Depth subsection and the Leave One Out section are based on a generated pool that is created from only the ad hoc-focused submissions. This pool is not identical to the actual pool used at INEX, but it is a close clone. Though our results would have been slightly different if the original INEX pool could have been regenerated, we believe that this difference would be small.

When reducing pool depth, we found that rankings do not change significantly even when the pool size is substantially reduced. However, if such small pools were actually used in practice, it would be interesting to study whether new systems could still be reliably evaluated on the basis of the resultant relevance assessments. In other words, it might be instructive to redo the Leave One Out experiments using these reduced pools as a starting point.

Our experiments on error rate follow the methodology introduced by Voorhees and Buckley (2002). One can argue that this method lacks a solid mathematical foundation, but it does provide similar results as theoretically more sound methods such as the bootstrap sensitivity method (Sakai, 2007; Sakai & Kando, 2008). Though we did not measure the discriminative power of the metrics concerned, the error rate of a metric indicates its discriminative power: a lower error rate indicates higher discriminative power. These experiments can thus be regarded as an empirical study of statistical power. A more formal study could follow the approach suggested by

(at random), it may be better to reduce the workload by giving them larger pools for fewer topics and ensuring that, if they start judging a query, they complete the assessment for that query.

Varying Pool Depth Across Queries

In our attempts to investigate how assessment effort can best be utilized, we finally look at the question of varying the amount of assessment effort across queries. We observed that the relation between pool depth d_Q and the number of relevant documents found for a query Q varies widely across queries. In general, for any query, the rate of finding new relevant documents decreases as pool depth is increased, and eventually drops to near zero after a threshold, i.e., one does not find any significant number of new relevant documents even though pool depth is increased substantially beyond this point. For a query Q, we call this threshold (which is unique for Q) the critical pool depth. The rate of finding new relevant documents and the critical pool depth vary significantly from query to query. To ensure a reasonably good estimate of recall for a given Q, d_Q should be no less than its critical pool depth.

If the rate of finding relevant documents is high for a query (e.g., queries 570, 582 in Figure 6), and pooling for that query is stopped because the target pool size has been reached, one may not reach critical pool depth. For example, for query 582, pooling stops at about $d_Q = 26$ as the pool size reaches 608 documents. Figure 6 suggests, however, that we cannot be sure to have reached critical pool depth. Similarly,

Webber et al. (2008) to investigate the number of topics (n) required to achieve a certain level of statistical power$(1 − \beta)$ at a given level of significance(say, $\alpha = 0.05$).

Conclusion

Evaluation is a grueling challenge for XML retrieval research. Ever since the inception of INEX, its evaluation measures have changed at regular intervals. With the inclusion of arbitrary passages as valid retrieval units besides the usual XML elements, and the need for a common set of effectiveness measures has gained importance. INEX 2007 therefore introduced a set of precision recall-based measures for its ad hoc tasks. The main aim of this study was to investigate the reliability and robustness of these focused retrieval measures, as well as the pooling method used at INEX. Our experiments were mainly driven by the common objective of finding a reliable, "optimum" evaluation set-up in terms of the assessment effort required, and the evaluation measures used to rank a set of XML retrieval systems at INEX. The four specific questions that we investigated (see the Introduction) and our related findings are summarized below.

1. How reliable are the various metrics in ranking competing systems when assessments are incomplete?

    The results of our experiments validate properties of precision recall-based metrics that were originally observed in a document retrieval setting. For example, our experiments reaffirm that early precision measures (iP[0.00], iP[0.01]) are more error-prone and less stable under incomplete judgments, whereas AiP is the least vulnerable among these metrics. Specifically, MAiP-based rankings remain largely unaltered ($\tau \geq 0.9$), even when evaluation effort is halved.

    On a related note, what is the minimum pool size that can be used to reliably evaluate systems?

    As evaluation is strongly dependent on a relatively small set of top-ranked results, rankings similar to the official rankings can be obtained even when pooling is limited to about 15% of the currently used pool depth.

2. How reliable are the various metrics in ranking competing systems if the query set size is small? What is the minimum number of queries that should be used to keep the error rates for the various metrics within a maximum allowable upper bound?

    As in (1) above, we find that early precision measures (iP[0.00], iP[0.01]) are more error-prone, and less stable if a small topic set is used, whereas AiP is the most stable. Our experiments also suggest that the pool size and number of queries used since INEX 2007 are large enough to reliably evaluate all submissions, i.e., the INEX results generally contain less than 5% error for all the metrics reported.

3. When a set of relevance assessments is used to evaluate a "new" system that did not contribute to the pool used in the relevance assessment process, are the results biased against this system?

    Our investigation into the effect of bias towards a system due to its contribution to the pool suggests that a new system that did not contribute to the INEX 2007 or INEX 2008 pool can be fairly evaluated on the basis of the corresponding qrels. In most cases, contributing systems get insignificant advantages due to their contribution to the qrels.

4. For a fixed amount of assessment effort, would this effort be better spent in thoroughly judging a few queries, or in judging many queries relatively superficially?

    We observe that, to reduce the amount of effort required to create usable qrels, it is better to judge shallower pools for all topics, rather than reduce the number of topics that are judged. However, it is better to completely judge a smaller number of topics, than to randomly judge many topics. The most-effective use of available manpower may be made by choosing the pool-depth/pool size on a per query basis. By reducing assessment effort for some queries, and increasing assessment effort for others, it may be possible to obtain better estimates of the 100% recall level for all queries.

These findings should be useful while formulating the evaluation strategy to be used with much larger text collections—from 2009, INEX is moving to a *Wikipedia* collection that is roughly 50 gigabyte in size, about 10 times as large as the old *Wikipedia* collection—and to address concerns regarding the INEX evaluation methodology in the coming years.

References

Ahlgren, P., & Grönqvist, L. (2008). Evaluation of retrieval effectiveness with incomplete relevance data: Theoretical and experimental comparison of three measures. Information Processing and Management, 44(1), 212–225.

Aslam, J.A., Pavlu, V., & Yilmaz, E. (2006). A statistical method for system evaluation using incomplete judgments. In E.N. Efthimiadis, S.T. Dumais, D. Hawking, & K. Järvelin (Eds.), SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 541–548). New York, NY: ACM.

Baillie, M., Azzopardi, L., & Ruthven, I. (2008). Evaluating epistemic uncertainty under incomplete assessments. Information Processing and Management, 44(2), 811–837.

Bompada, T., Chang, C.-C., Chen, J., Kumar, R., & Shenoy, R. (2007). On the robustness of relevance measures with incomplete judgments. In W. Kraaij, A.P. de Vries, C.L.A. Clarke, N. Fuhr, & N. Kando (Eds.), SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 359–366). New York, NY: ACM.

Buckley, C., & Voorhees, E.M. (2000). Evaluating evaluation measure stability. In N.J. Belkin, P. Ingwersen, M.-K. Leong (Eds.), SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 33–40). New York, NY: ACM.

Buckley, C., & Voorhees, E.M. (2004). Retrieval evaluation with incomplete information. In J. Callan, N. Fuhr, & W. Nejdl (Eds.), SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 25–32). New York, NY: ACM.

Büttcher, S., Clarke, C.L.A., Yeung, P.C.K., & Soboroff, I. (2007). Reliable information retrieval evaluation with incomplete and biased judgments. In W. Kraaij, A.P. de Vries, C.L.A. Clarke, N. Fuhr, & N. Kando (Eds.), SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 63–70). New York, NY: ACM.

Carterette, B. (2007). Robust test collections for retrieval evaluation. In W. Kraaij, A.P. de Vries, C.L.A. Clarke, N. Fuhr, & N. Kando (Eds.), SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 55–62). New York, NY: ACM.

Carterette, B. (2009). On rank correlation and the distance between rankings. In J. Allan, J.A. Aslam, M. Sanderson, C.-X. Zhai, & J. Zobel (Eds.), SIGIR '09: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 436–443). New York, NY: ACM.

Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J.A., & Allan, J. (2008). Evaluation over thousands of queries. In S.-H. Myaeng, D.W. Oard, F. Sebastiani, T.-S. Chua, & M.-K. Leong (Eds.), SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 651–658). New York, NY: ACM.

Denoyer, L., & Gallinari, P. (2006). The Wikipedia XML corpus. SIGIR Forum, 40(1), 64–69.

Führ, N., Kamps, J., Lalmas, M., Malik, S., & Trotman, A. (2008). Overview of the INEX 2007 Ad Hoc Track. In N. Fuhr, J. Kamps, M. Lalmas, & Andrew Trotman (Eds.), Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, Revised and Selected Papers (pp. 1–23). Lecture Notes in Computer Science, Vol. 4862. Berlin, Heidelberg: Springer-Verlag.

Gövert, N., Führ, N., Lalmas, M., & Kazai, G. (2006). Evaluating the effectiveness of content-oriented XML retrieval methods. Information Retrieval, 9(6), 699–722.

Gövert, N., & Kazai, G. (2003). Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2002. In N. Führ, N. Gövert, G. Kazai, & M. Lalmas (Eds.), Proceedings of the First Workshop of the Initiative for the Evaluation of XML Retrieval (INEX), Dagstuhl, Germany (pp. 1–17). Retrieved from http://www.ercim.org/publication/workshop reports.html/INEX2002.pdf

Initiative for the Evaluation of XML retrieval (INEX). (2009). Retrieved from http://www.inex.otago.ac.nz

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4), 422–446.

Kamps, J., Geva, S., Trotman, A., Woodley, A., & Koolen, M. (2009). Overview of the INEX 2008 Ad Hoc Track. In S. Geva, J. Kamps, & A. Trotman (Eds.), Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008. Revised and Selected Papers (pp. 1–28). Lecture Notes in Computer Science, Vol. 5631. Berlin, Heidelberg: Springer-Verlag.

Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., & Robertson, S. (2008). INEX 2007 Evaluation Measures. In N. Fuhr, J. Kamps, M. Lalmas, & Andrew Trotman (Eds.), Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, Revised and Selected Papers (pp. 24–33). Lecture Notes in Computer Science, Vol. 4862. Berlin, Heidelberg: Springer-Verlag.

Kazai, G., & Lalmas, M. (2006a). eXtended cumulated gain measures for the evaluation of content-oriented XML retrieval. ACM Transactions on Information Systems, 24(4), 503–542.

Kazai, G., & Lalmas, M. (2006b). INEX 2005 evaluation metrics. In N. Führ, M. Lalmas, & A. Trotman (Eds.), Advances in XML Retrieval and Evaluation: 4th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005) (pp. 16–29). Lecture Notes in Computer Science, Vol. 397. Berlin, Heidelberg: Springer-Verlag.

Malik, S., Trotman, A., Lalmas, M., & Führ, N. (2007). Overview of INEX 2006. In N. Führ, M. Lalmas, & A. Trotman (Eds.), Comparative Evaluation of XML Information Retrieval Systems, 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Revised and Selected Papers (pp. 1–11). Lecture Notes in Computer Science, Vol. 4518. Heidelberg, Germany: Springer-Verlag.

Lacasse, M.D. (2001). F U D G I T version 2.41. Retrieved from http://hpux.connect.org.uk/hppd/hpux/Maths/Misc/fudgit-2.41/

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems, 27(1), 1–27.

Pal, S., Mitra, M., & Chakraborty, A. (2008). Stability of INEX 2007 evaluation measures. In T. Sakai & M. Sanderson (Eds.), Proceedings of the Second International Workshop on Evaluating Information Access (EVIA 2008), NTCIR 7 (pp. 23–29). Retrieved from http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/EVIA2008/06EVIA2008-PalS.pdf)

Piwowarski, B., Trotman, A., & Lalmas, M. (2008). Sound and complete relevance assessment for XML retrieval. ACM Transactions on Information Systems, 27(1), 1–37.

Sakai, T. (2007). Evaluating information retrieval metrics based on bootstrap hypothesis tests. IPSJ Digital Courier, 3, 625–642.

Sakai, T. (2008a). Comparing metrics across TREC and NTCIR: The robustness to pool depth bias. In S.-H. Myaeng, D.W. Oard, F. Sebastiani, T.-S. Chua, & M.-K. Leong (Eds.), SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 691–692). New York, NY: ACM.

Sakai, T. (2008b). Comparing metrics across TREC and NTCIR: The robustness to system bias. In J.G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D.A. Evans, A. Kolcz, K.-S. Choi, et al. (Eds.), CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Management (pp. 581–590). New York, NY: ACM.

Sakai, T., & Kando, N. (2008). On information retrieval metrics designed for evaluation with incomplete relevance assessments. Information Retrieval, 11(5), 447–470.

Sanderson, M., & Soboroff, I. (2007). Problems with Kendall's tau. In W. Kraaij, A.P. de Vries, C.L.A. Clarke, N. Fuhr, & N. Kando (Eds.), SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 839–840). New York, NY: ACM.

Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. In R.A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, & J. Tait (Eds.), SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 162–169). New York, NY: ACM.

Trotman, A., Pharo, N., & Jenkinson, D. (2007). Can we at least agree on something? In A. Trotman, S. Geva, & J. Kamps (Eds.), SIGIR '07: Proceedings of the SIGIR 2007 Workshop on Focused Retrieval (pp. 49–56). Retrieved from http://www.cs.otago.ac.nz/sigirfocus/paper1.pdf

Voorhees, E.M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 316–323). New York, NY: ACM.

Webber, W., Moffat, A., & Zobel, J. (2008). Statistical power in retrieval experimentation. In CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Management (pp. 571–580). New York, NY: ACM.

Yilmaz, E., & Aslam, J.A. (2006). Estimating average precision with incomplete and imperfect judgments. In O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury, W. Teiken (Eds.), CIKM '06: Proceedings of the 15th ACM International Conference on Information and Knowledge Management (pp. 102–111). New York, NY: ACM.

Yilmaz, E., Aslam, J.A., & Robertson, S. (2008). A new rank correlation coefficient for information retrieval. In S.-H. Myaeng, D.W. Oard, F. Sebastiani, T.-S. Chua, & M.-K. Leong (Eds.), SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 587–594). New York, NY: ACM.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 307–314). New York, NY: ACM.