Stability of INEX 2007 Evaluation Measures

Sukomal Pal   Mandar Mitra
Information Retrieval Lab, CVPR Unit
Indian Statistical Institute
203, B.T. Road, Kolkata 700108, India
{sukomal_r, mandar}@isical.ac.in

Arnab Chakraborty
arnabc@stanfordalumni.org

Abstract

This paper examines the robustness of the evaluation measures which were used at INEX 2007 to rank XML retrieval systems in the focused adhoc task. We study the behaviour of the measures when the completeness assumption of the Cranfield evaluation methodology (i.e. the assumption that all relevant information items within a test collection have been identified and included in the judgment pool) is violated. We also study how the measures behave when evaluation is based on progressively smaller sets of queries. We show that the official measure used for the Focused Task of the INEX 2007 adhoc track (Interpolated Precision at 1% recall or $iP[0.01]$*) is less stable under both types of variations, while* $MAiP$*, which is similar to the MAP measure used in traditional document retrieval, is the most stable measure among the INEX 2007 focused task evaluation measures. Our experiments are in line with their precedents in the document retrieval domain, and our findings are also in agreement with earlier findings.*

Keywords: *Evaluation, XML retrieval, INEX.*

1 Introduction

Content-oriented XML [15] retrieval is a domain of information retrieval (IR) that has been receiving increasing attention in recent times. 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, in order 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), since 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 [12].

The INitiative for the Evaluation of XML retrieval (INEX) [5], set up in 2002, has been providing an infrastructure for evaluating the effectiveness of content-oriented XML IR systems, in the form of large test collections, topic sets and relevance judgments. As the retrieval unit for XML search systems can be an element of arbitrary granularity and length, evaluation has been a challenge in 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 the complete document. Various evaluation measures have been tried over the years at INEX. The official metrics used at INEX 2002 [7] (calculated by the `inex_eval` program) were modified for INEX 2003 and 2004 [6] (cf. the `inex_eval_ng` program). Again, at INEX 2005, three new *Cumulated Gain* [8]-based metrics were taken as official metrics [11, 10]. 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 adhoc 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 bytes or characters, rather than the number of documents (see Section 3 for details). Five of these metrics, viz. $iP[0.0]$, $iP[0.01]$, $iP[0.05]$, $iP[0.10]$ and $MAiP$, were used in the official reports for the focused adhoc tasks. Among these, $iP[0.01]$ was taken as the official measure to rank the competing systems.

Since these measures are extensions of their counterparts in the standard document retrieval setting, we expect the measures to have similar properties in the domain of XML retrieval as well. To the best of our knowledge, however, no experiments have been reported that substantiate or refute this intuition. It is in this context we undertake the work reported in this paper. Our results suggest that early precision measures ($iP[0.0]$, $iP[0.01]$) are more error-prone and less stable to incomplete judgments, whereas $MAiP$ offers the least vulnerability among these metrics. Thus, our work validates our intuition that precision-recall based measures maintain similar characteristics in the domain of XML IR as far as their stability and robustness are concerned.

The paper is organized as follows. In the next section, we review past work that provides the background needed for the rest of the paper. The following section presents the test environment used in this study, definitions of the measures to be examined, and our experimental set-up. Results are reported in Section 4. In Section 5, we analyse our findings and briefly discuss issues that need further investigation. Section 6 concludes the paper.

2 Previous Work

Evaluating the evaluation measures has a well-established history in the field of document retrieval. In 2000, Buckley et al. [1] proposed a novel way to examine the accuracy of various evaluation measures and validated a number of traditional thumb-rules that address issues such as the minimum number of queries required, 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 is a more stable measure than measures based on early precision.

This work was extended in 2004 [2] with the study of evaluation measures under incomplete and imperfect relevance judgments. They 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 Mean Average Precision (MAP) is both stable and discriminatory for evaluating document-level retrieval from a static document collection.

Sanderson et al. [14] studied the error-rates of evaluation metrics, specifically *MAP* and $P@10$, in the 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 $P@10$.

However all the work discussed above was based on document-level retrieval using TREC data. Kazai et al. [10] reported similar work in the field of XML retrieval. This 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 INEX 2004 submissions (where only XML elements were permissible as units of retrieval).

Pehcevski proposed a new metric *HiXEval* for XML retrieval evaluation, based on traditional notions of precision and recall in his doctoral work [13]. This measure can accommodate both passages and elements as retrievable units. He showed that the new metric is comparable to the XCG-based official metrics at INEX 2005 with regard to fidelity tests (tests that check whether a metric indeed measures what it is intended to measure) and reliability tests (how stable a metric is at distinguishing between two different systems).

Until recently, only XML elements were considered retrievable units at INEX. How to handle overlap among the elements was an issue during this period. With the inclusion of non-overlapping passage retrieval, a new set of evaluation metrics has been introduced since INEX 2007. In this paper, we present a study of these new metrics that is essentially based on the work of Buckley et al. [1, 2], and partly supported by that of Sanderson et al. [14].

3 Test Environment

The goal of our experiments is two-fold: to observe the behaviour of the INEX 2007 metrics used for the focused adhoc task under (i) incomplete assessments, and (ii) smaller query sets. Experiments follow the Cranfield methodology [3] and use the INEX 2007 adhoc test collection.

3.1 Test Collection

The test collection consists of an XML-ified version of the English Wikipedia. The corpus contains 659,388 documents, and has a total size of 4.6 GB [4]. The original topic set for INEX 2007 contains 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.

The focused task of the adhoc track at INEX 2007 expected participating systems to return, for each topic, a ranked list of non-overlapping document parts (either passages or XML elements) that are most focused with respect to the information need expressed in the topic. Among the submitted runs, 79 were reported in the INEX 2007 website as valid runs. Each such run was supposed to retrieve 1500 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.

3.2 Evaluation Measures

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

$$\begin{aligned}\text{precision} &= \frac{\text{amount of relevant text retrieved}}{\text{total amount of } \textit{retrieved} \text{ text}}\\ &= \frac{\text{length of relevant text retrieved}}{\text{total length of } \textit{retrieved} \text{ text}}\end{aligned}$$

$$\text{recall} = \frac{\text{length of relevant text retrieved}}{\text{total length of } \textit{relevant} \text{ text}}$$

Kamps et al. [9] 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,

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

and

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

Since the notion of ranks is not clearly definable for passages, precision at recall levels, rather than at ranks is considered. Specifically, interpolated precision at various recall levels are used. Interpolated precision at recall level x is defined as follows:

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

Here $|L_q| \le 1500$.

For example, $iP[0.00]$ calculates the interpolated precision when the first 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 $(0.00, 0.01, \ldots, 1.00)$:

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

Overall performance measure: We calculate overall performance by averaging the 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

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

For the INEX 2007 focused adhoc 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.

3.3 Experimental Setup

Since relevance in the XML domain is defined at the sub-document level, a relevance judgment file contains more information than just a boolean indicator about whether a document is relevant or irrelevant. The relevance judgment file lists, for each topic, the documents that contain relevant passages, and precisely specifies the relevant elements and/or passages within each document. Elements are identified by their *xpath* [16], while passages are identified by the xpaths corresponding to their start-position and end-position.

The evaluation metrics were calculated using a pair of perl-scripts unofficially provided to INEX participants. The first script builds a huge database (approx. 14 GB) that stores normalized byte-offsets corresponding to the start and end positions of each node (identified by its *xpath*) of each document in the entire collection. The second script takes the files containing the relevance judgments along with a run file that needs to be evaluated, consults the database for each *xpath* in the relevance judgment file, and then checks the *xpath* of the retrieved units from the run file to determine the amount of overlap with any of the relevant units for that topic.

Though the first script needs to run only once, the second script has to be executed repeatedly for each submission file. This exercise takes a prohibitively long time (about 55 minutes user-time on average on an Intel Core 2 Duo (2.13GHz) workstation with 1 GB RAM). We therefore implemented a C-version of the second script that eliminated the redundant repetition of reading the relevance judgment file for each evaluation. Instead, the byte offsets corresponding to the relevant passages are extracted and saved once in a separate file, and these offsets are used when evaluating each run. Our implementation cuts down the run-time by nearly 50%, but it still takes approx. 30 minutes on an average to calculate the metrics for a single run.

In order to cut down on the total time needed for the experiments, we selected 62 runs out of the 79 valid submissions (we planned to consider all 79 runs but could manage a few less). To ensure some sort of unbiasedness in their representation, the selected set consists of runs that were ranked $1-21$, $31-50$, and $59-79$ at INEX 2007 on the basis of the official metric, $iP[0.01]$.

The actual experiments can be divided into three subtasks.

1. In the first case, we study the effect of incomplete relevance assessments on system rankings. In other words, we study how system rankings would change if some relevant passages were not actually known to be relevant. We call this task *pool sampling*.

2. In the second case, we use progressively smaller subsets of the complete topic-set, but use the complete relevance judgment information for each topic. We call this task *query sampling*.

3. The third set of experiments is an offshoot of the second task, where we look at the pairwise comparison of retrieval strategies using each of the five metrics, and study their *error-rates* as the topic-set size is reduced.

4 Results

The results of the three sets of experiments described above are similar, and are in general agreement with the earlier experiments conducted along the same lines on document retrieval. In this section, we take up each category in turn and and discuss the results in more detail.

4.1 Pool Sampling

From the INEX 2007 focused adhoc submissions, a pool was generated which was collaboratively judged by the topic creators and other assessors. This process generated a *qrel*-set which consists of 107 different files, each corresponding to a topic. Each of these files contains a list of documents retrieved by different systems contributing to the pool. If a document contains some relevant text that is highlighted by assessors, the qrel-file specifies the location of the text within the document (using *xpath* expressions). From the *qrel*-files, we extract all the relevant texts, consult the indexed database for their start and end positions, and store these positions in a file grouped by documents and then by topics in sorted order.

For each of the topics having a reasonable number of relevant units (8 topics that had less than 10 relevant units each were omitted), we chose 80% of the relevant passages at random without replacement. Though the original qrel-file contains assessed non-relevant text, it is to be noted that these entries do not figure during computation of precision-scores. Hence byte-offsets corresponding to only assessed relevant texts occur in the modified qrel-file. The selection, therefore, creates an 80% sample of the entire qrel-file. With this reduced set of assessments, we evaluated and ranked the 62 selected runs on the basis of each metric. We calculated Kendall's tau (τ) between the ranking of the runs as produced by each metric with the original (100%) pool and the same metric with this 80% pool. The process is repeated 10 times, each time choosing 80% of the qrel-entries at random for each "good" topic. For these 10 samples, we get ten τ values for each of the five metrics.

The entire exercise is repeated at 60%, 40% and 20% sampling levels. The means of the tau values along with the standard error at each sampling level are shown in Figure 1.

As the sampling level decreases, the correlation between the rankings produced by a metric with the original assessments and the reduced assessments decreases in general, so each of the curves droops. One obvious reason is that with reduced assessments, the precision-score is affected non-uniformly 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. However, 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 $MAiP$.

Error-bars for each curve tend to increase as pool-size reduces. The reason can be attributed to the fact that at smaller pool-percentage, the overlap among the samples reduces which 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. 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.

4.2 Query Sampling

The setup for this task is quite similar to what it is for *pool sampling*. First, we randomly select an

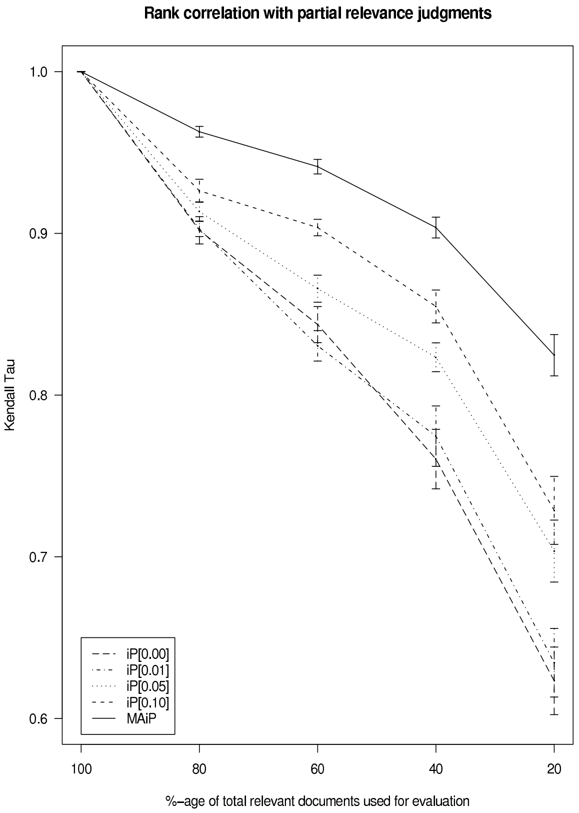

Figure 1. Kendall Tau vs. Pool-size

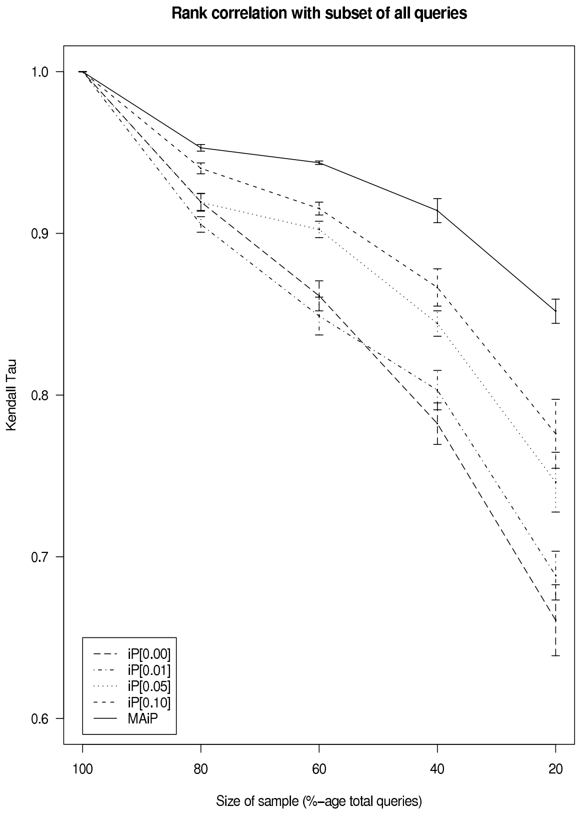

Figure 2. Kendall Tau vs. Query-size

80% sample from the total set of 107 queries. For each selected topic, all available assessment information is considered. Once again, we calculate Kendall's τ 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. We repeat the exercise with 60%, 40% and 20% of the query set. The behaviour of the metrics as query-set size varies is shown in Figure 2.

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 high precision measures ($iP[0.05]$, $iP[0.10]$) as the topic-set is gradually reduced. $MAiP$ again emerges as a clear winner both in terms of its resilience to the reduction in size of the topic-set and variation across the samples (smallest error bar). A closer look also reveals that curves are slightly more stable here in comparison to their counterparts in Figure 1. This means that the system rankings produced at reduced query-set sizes tend to agree more with the original rankings. This may be explained as follows. If a topic is included in the pool, the complete relevance judgments for the topic are considered. Thus, unlike in *pool sampling*, the query contributes to the precision score for all the systems uniformly. The reduction in τ is caused by the variation of systems across topics.

4.3 Error Rates

The experiment to examine the errors committed by the measures is designed based on the work of Buckley and Voorhees [1], but with some modifications. As we do not have the systems participating in INEX 2007 with us, it is not possible to see the retrieval results of different retrieval strategies under varying query formulations for a topic-set. We only have the submission files corresponding to different strategies, which we can evaluate using various subsets of the complete query-set. We can then determine how many times a metric correctly ranks a pair of strategies across the samples. We hence perform the test with the help of *query sampling*. The queries could be partitioned in a number of disjoint sets or can be selected with replacement containing overlap among them. The first strategy gives the upper bound in the error-rates that a metric can commit during evaluation. The second one gives a lower bound in the error-rates since the overlap in the query-set reduces the chance of error [14]. We take the second approach of query-sampling (with replacement) to find the lower-bounds of error-rates for the metrics. We take 10 random samples at each sampling level (corresponding to 20%, 40%, 60% and 80% of the complete query-set). Following the definition by Buckley et al. [1], *error rate* is defined as

$$\text{Error rate} = \frac{\sum \min(|A > B|, |A < B|)}{\sum(|A > B| + |A < B| + |A == B|)}$$

where $|A > B|$ is the number of times (out of 10) that system A does better than system B at a fixed sampling level. A and B are considered different ($A > B$ or $A < B$) only when their scores differ by more than 5%; otherwise, we say $A == B$. In our case, the denominator is 10, as we have 10 samples at each percentage point.

We find the total error rate over all the 62 runs under consideration, by considering all $\binom{62}{2} = 62.61/2 = 1891$ pairwise comparisons.

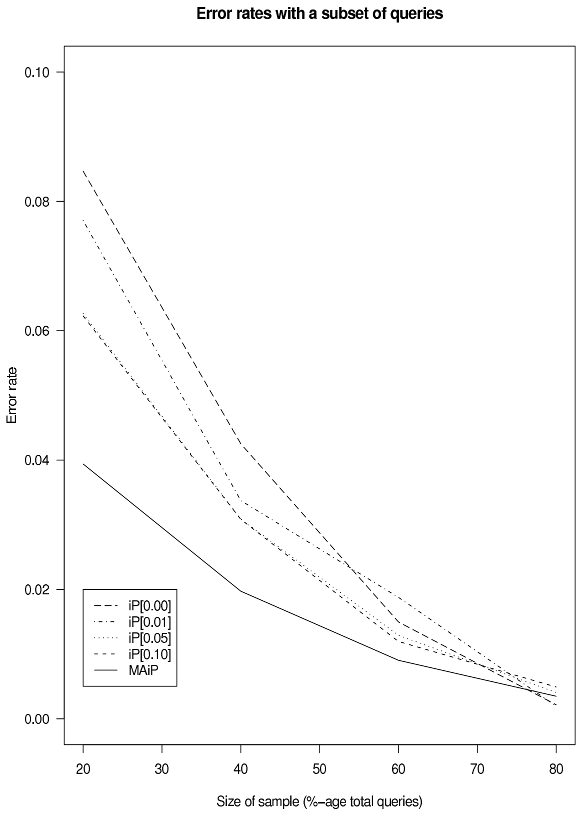

Figure 3. Error-rates vs. Query-set size

The error-rates are generally high with small query-sets and progressively decrease as overlap among the query samples increases. The graphs also suggest that 40% of the topics are sufficient to achieve less than 5% errors in ranking with the INEX 2007 test collection. In line with earlier findings, early precision measures ($iP[0.00]$ and $iP[0.01]$) are more vulnerable to errors, $iP[0.05]$ and $iP[0.10]$ have similar error-rates. However $MAiP$ appears to be the least error-prone among the metrics.

5 Limitations and Future Work

The experiments described here are basically validations of previous findings for a set of *precision-recall*-based metrics customized for XML retrieval. The fact that we achieve similar results is good evidence of the intrinsic properties of the metrics. All three experiments suggest that $MAiP$ can be used to reliably rank XML retrieval systems for the focused adhoc task. The fact that $MAiP$ averages well across the ranks or recall-levels as well as across topics gives it more shock-absorbing capabilities compared to the other metrics. For a static test environment, this metric is more reliable than other early precision measures.

One obvious limitation to the findings is that these are based on only the INEX 2007 test collection. Moreover, we could not consider all available runs of the focused task because of time constraints. However, our test-set consists of runs from three rungs (best, worst and mediocre) of the entire list of reported runs and include 62 out of 79 valid submissions.

Of course, these runs actually belong to a number of non-random categories which may affect the results (e.g., passage vs. element runs, Content-Only (CO) vs. Content-And-Structure (CAS) runs, short vs. long topics, hard vs. easy queries, query expansion vs. no expansion, same base retrieval system vs. different system etc.). In general, future work can certainly be directed in exploring the effects of these categories.

On a different note, the assessment pool (and therefore the set of relevant units) is generated from top n retrieved units from each of the submission files. We do not know the actual value of n that was used to generate the INEX 2007 pool. How the variation of n changes the pool and how this affects the behaviour of the metrics under question will be an interesting future task. Moreover, bias of qrels towards participating systems is another issue to look at. How fairly and reliably a new system, which does not contribute to the pool, is evaluated will definitely be an area of investigation.

Our study of stability (Section 4.3) is preliminary and emerged as an offshoot of *query sampling*. While at low sample rates, early precision is uniformly more error-prone than precision at higher recall levels and $MAiP$, this trend is sometimes reversed at higher sample sizes (e.g. 70% to 80%). This needs to be investigated. Further, the comments we make about the error-rates are only on the basis of the INEX 2007 test collection taking 5% difference in absolute scores. The issue needs more experimentation taking some other test collections as well.

6 Conclusion

Evaluation is a gruelling challenge for XML retrieval research. Ever since the inception of INEX, its evaluation measures have changed at regular intervals. With the inclusion of arbitrary passage as valid retrieval units besides the usual XML elements, 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 adhoc tasks. This paper studies the behaviour of five such metrics used for the focused adhoc task. Our experiments confirm that $MAiP$ is more robust to both incomplete assessments, and smaller topic sets than early precision measures ($iP[0.00]$ and $iP[0.01]$). We also observe that $MAiP$ has the lowest error-rates compared to other precision-based measures. In general, $MAiP$ appears to be the most reliable among the metrics used for the INEX 2007 focused adhoc task.

References

[1] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40, New York, NY, USA, 2000. ACM.

[2] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, New York, NY, USA, 2004. ACM.

[3] C. W. Cleaverdon, J. Mills, and E. M. Keen. *Factors determining the performance of indexing systems, Two Volumes*. Cranfield, England, 1968.

[4] L. Denoyer and P. Gallinari. The wikipedia xml corpus. *SIGIR Forum*, 40(1):64–69, 2006.

[5] N. Fuhr, M. Lalmas, A. Trotman, and J. Kamps, editors. *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, 2007. http://inex.is.informatik.uni-duisburg.de/2007.

[6] N. Gövert, N. Fuhr, M. Lalmas, and G. Kazai. Evaluating the effectiveness of content-oriented xml retrieval methods. *Inf. Retr.*, 9(6):699–722, 2006.

[7] N. Gövert and G. Kazai. Overview of the initiative for the evaluation of xml retrieval (inex) 2002. In *N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, editors, Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*, pages 1–17, Dagstuhl,Germany, 2003.

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

[9] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. INEX 2007 Evaluation Measures. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *Pre-Proceedings of INEX 2007*, pages 23–32, 2007. http://inex.is.informatik.uni-duisburg.de/2007/inex07/pdf/2007-preproceedings.pdf.

[10] G. Kazai and M. Lalmas. extended cumulated gain measures for the evaluation of content-oriented xml retrieval. *ACM Trans. Inf. Syst.*, 24(4):503–542, 2006.

[11] G. Kazai and M. Lalmas. INEX 2005 evaluation metrics. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *Advances in XML Retrieval and Evaluation: 4th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005)*, volume Lecture Notes in Computer Science vol. 3977, pages 16–29. Springer-Verlag, 2006.

[12] S. Malik, A. Trotman, M. Lalmas, and N. Fuhr. Overview of INEX 2006. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *Comparative Evaluation of XML Information Retrieval Systems*, pages 1–11, 2006.

[13] J. Pehcevski. *Evaluation of Effective XML Information Retrieval*. PhD thesis, RMIT University, 2006.

[14] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169, New York, NY, USA, 2005. ACM.

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

[16] W3C. XPath-XML Path Language(XPath) Version 1.0. http://www.w3.org/TR/xpath.