INEX 2007 Evaluation Measures

Jovan Pehcevski, Jaap Kamps, Gabriella Kazai, Mounia Lalmas, Paul Ogilvie, Benjamin Piwowarski, Stephen Robertson

INEX 2007 Evaluation Measures (Draft)

Jovan Pehcevski[1], Jaap Kamps[2], Gabriella Kazai[3], Mounia Lalmas[4], Paul Ogilvie[5], Benjamin Piwowarski[6], and Stephen Robertson[3]

[1] INRIA Rocquencourt, France
jovan.pehcevski@inria.fr
[2] University of Amsterdam, The Netherlands
kamps@science.uva.nl
[3] Microsoft Research Cambridge, United Kingdom
{gabkaz,ser}@microsoft.com
[4] Queen Mary, University of London, United Kingdom
mounia@dcs.qmul.ac.uk
[5] Carnegie Mellon University, USA
pto@lti.cs.cmu.edu
[6] Yahoo! Research Latin America, Chile
bpiwowar@yahoo-inc.com

Abstract. This paper describes the official measures of retrieval effectiveness that are planned to be employed for the ad hoc track of INEX 2007.

1 Introduction

Focused retrieval, including question answering [19], passage retrieval [1, 2, 5, 20], and XML element retrieval [8, 11, 12, 16], investigates ways to provide users with direct access to relevant information in retrieved documents. Since its launch in 2002, INEX has studied different aspects of focused retrieval by mainly considering XML element retrieval techniques that can effectively retrieve information from structured document collections [11]. The main change at INEX 2007 is allowing retrieval of arbitrary document parts, which can represent XML elements or passages [3]. That is, a retrieval result can be either an XML element (a sequence of textual content contained within start/end tags), or an arbitrary passage (a sequence of textual content that can be either contained within an element, or it can span across a range of elements).

How to properly evaluate XML retrieval effectiveness is still an ongoing problem within INEX. To alleviate this problem, in INEX 2007 we will adopt an evaluation framework where different aspects of focused retrieval can be consistently evaluated and compared. To measure the extent to which an XML or passage retrieval system returns relevant information, we will employ evaluation measures that only consider the amount of highlighted text in relevant documents [13, 14]. This is motivated by the need to directly exploit the highlighting assessment procedure used at INEX 2007, which as we demonstrate in this paper leads to measures that are natural extensions of the well-established measures used in traditional information retrieval [6, 17].

This paper is organised as follows. In Section 2, we briefly describe the ad hoc retrieval tasks of INEX 2007 and their motivations. In Section 3, we describe how relevance is defined in INEX 2007. The evaluation measures used for each of the INEX 2007 tasks are described in the last two sections (Sections 4 and 5).

2 Ad hoc retrieval tasks

INEX 2007 will investigate the following three ad hoc retrieval tasks, which are defined as follows [3]:

- *Focused* task: This task asks systems to return a ranked list of the most focused document parts (XML elements or passages), where the resulting document parts should not overlap. For example, in the case of returning XML elements, a paragraph and its container section should not both be returned. For this task, from all the estimated relevant (and possibly overlapping) document parts, systems are forced to choose those non-overlapping document parts that represent the most appropriate units of retrieval.
- *In context* tasks: These tasks correspond to end-user tasks where focused retrieval answers are grouped per document, in their original document order, providing access through further navigational means. This assumes that users consider documents as the most natural units of retrieval, and prefer an overview of relevance in their context. Two *in context* tasks are distinguished at INEX 2007, depending on whether a set of document parts or a single answer part are returned per document.
  - *Relevant in context*: This task asks systems to return non-overlapping relevant document parts (XML elements or passages) clustered by the unit of the document that they are contained within. An alternative way to phrase the task is to return documents with the most focused, relevant parts highlighted within.
  - *Best in context*: This task asks systems to return a single document part (XML element or passage) per document. The single document part corresponds to the best entry point for starting to read the relevant text in the document.

At INEX 2007 there is no separate passage retrieval task, and for all the three tasks arbitrary passages may be returned instead of elements. Note that a run submitted by an INEX 2007 participating group can contain either elements or passages, but not a mixture of both. For all the three tasks, systems could either use the title field of the topics (content-only topics) or the castitle field of the topics (content-and-structure topics). Trotman et al. [18] provide a detailed description of the format used for the INEX 2007 topics.

3 Relevance Assessments

Since 2005, a highlighting assessment procedure is used at INEX to gather relevance assessments for the INEX retrieval topics [10]. In this procedure, assessors

from the participating groups are asked to highlight sentences representing the relevant information in a pooled set of documents of the Wikipedia XML document collection [4]. An assessment program then computes the relevance of the judged document parts (including whole documents) as the ratio of highlighted to fully contained text; thus, the relevance values for document parts are drawn from a continuous scale in the range 0 to 1, where 0 corresponds to a document part that does not contain any highlighted information, while 1 corresponds to a fully highlighted document part.

The highlighting assessment procedure will also be used in 2007. For each relevant document part (XML element or passage), the INEX 2007 relevance assessments will record the size of the highlighted text contained by the document part (in number of characters) as well as the total text size of the document part (again, in number of characters). These two statistics form the basis for calculating the relevance score of the document part.

4 Evaluation of the *focused* task

4.1 Assumptions

In the *focused* task, for each INEX 2007 topic, systems are asked to return a ranked list of the top 1500 non-overlapping most focused relevant document parts. The retrieval systems are required not only to rank the document parts according to their estimated likelihood of relevance, but to also decide which document parts are the most focused non-overlapping units of retrieval.

We make the following evaluation assumption about the *focused* task: *Users want to see as much relevant text as possible with as little irrelevant text as possible.* Such an assumption is the basis of methods for evaluating the effectiveness of information retrieval systems based on recall and precision. However, instead of counting the number of relevant documents retrieved, in this case we measure the amount of relevant (highlighted) text retrieved [13, 14]. This assumption implies that, if two systems return document parts containing the same proportion of relevant text, the system that returns larger amount of relevant text will be preferred over the system that returns smaller amount of relevant text.

4.2 Evaluation measures

More formally, let p_r be a document part (XML element or passage) assigned to a rank r in a ranked list of document parts $\mathcal{R}$ returned by a retrieval system.[7] Let $rsize(p_r)$ be the amount of highlighted (relevant) text contained by p_r (if there is no highlighted text, $rsize(p_r) = 0$). Let $size(p_r)$ be the total number of characters contained by p_r, and let $Trel$ be the total amount of (highlighted) relevant text for a given INEX 2007 topic. $Trel$ is calculated as the total number of highlighted characters across all documents, which means that the total amount of highlighted relevant text for the topic represents the sum of the sizes of the (non-overlapping) highlighted passages contained by all the relevant documents.

[7] At INEX 2007, $|\mathcal{R}| = 1\,500$ elements or passages.

Measures at selected cutoffs We measure the fraction of retrieved relevant text at rank r as:

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

To achieve a high precision score at rank r, the document parts retrieved up to and including that rank need to contain as little non-relevant text as possible.

We measure the fraction of relevant text retrieved at rank r as:

$$R[r] = \frac{1}{Trel} \cdot \sum_{i=1}^{r} rsize(p_i) \quad (2)$$

To achieve a high recall score at rank r, the document parts retrieved up to and including that rank need to contain as much relevant text as possible.

An issue with the precision measure $P[r]$ given in Equation 1 is that it could be biased towards systems that return shorter document parts [20]. We therefore use an interpolated precision measure $iP[jR]$, which calculates interpolated precision scores at selected recall levels (such as $iP[0.1R]$, which calculates interpolated precision at 10% recall level). By calculating interpolated precision scores at selected recall levels, it would be possible to measure which system is more capable of retrieving as much relevant text as possible (the selected recall level), without also retrieving a substantial amount of non-relevant text.

With the interpolated precision measure at selected cutoffs, the performance across a set of topics is measured by calculating the mean of the scores obtained by the measure for each individual topic.

Overall performance measure In addition to using the interpolated precision measure at selected recall level cutoffs, for an INEX 2007 topic we also calculate scores with an overall performance measure: average precision.

Average precision (AP) is a measure that combines precision and recall to produce a single value for the overall performance of a retrieval system. We calculate AP as follows: first, the precision is calculated at each natural recall level (after a relevant document part is retrieved). If a relevant document part is not retrieved, the precision is taken to be zero. The precision values are then averaged such that a single value for the overall retrieval performance is produced for a topic. Let $rel(p_r)$ indicate the relevance of a document part p assigned to the rank r, such that $rel(p_r) = 0$ if the document part does not contain any highlighted information, and $rel(p_r) = 1$ if there is highlighted information contained by the document part.

We formally define AP as follows:

$$AP = \frac{\sum_{r=1}^{|\mathcal{R}|} rel(p_r) \cdot P[r]}{\sum_{r=1}^{|\mathcal{R}|} rel(p_r)} \cdot \frac{\sum_{r=1}^{|\mathcal{R}|} rsize(p_r)}{Trel} = \frac{\sum_{r=1}^{|\mathcal{R}|} rel(p_r) \cdot P[r]}{\sum_{r=1}^{|\mathcal{R}|} rel(p_r)} \cdot R[|\mathcal{R}|] \quad (3)$$

We use the corresponding overall performance measure iAP, which represents interpolated average precision calculated at 101 recall levels.

With the overall performance iAP measure, the performance across a set of topics is measured by calculating the mean of the values obtained by the measure for each individual topic ($iMAP$).

4.3 Results reported at INEX 2007

For the *focused* task we report the following measures over all INEX 2007 topics:

- Interpolated precision at selected recall level cutoffs:
  $iP[jR]$, $j \in [0.0, 0.01, 0.05, 0.1]$; and
- Interpolated mean average precision ($iMAP$).

The official evaluation for the *focused* task will be based on the overall $iMAP$ measure.

5 Evaluation of the *in context* tasks

5.1 Assumptions

The two *in context* tasks are document retrieval tasks, where not only the relevant documents should be retrieved, but also either a set of relevant answer parts (*relevant in context*) or a single answer part (*best in context*) should be correctly identified. In both tasks, the documents should be ranked in a decreasing order of their estimated likelihood of relevance. In the *relevant in context* task, for each relevant document, systems are expected to return a set of non-overlapping document parts representing the relevant text within the document. In the *best in context* task, for each relevant document, systems are expected to return a single document part representing the best entry point (BEP) for starting to read the relevant text in the document.

We make the following evaluation assumption about the two *in context* tasks: *Users consider all relevant documents to be equally useful answers.* This assumption models users that place equal value on each relevant document that has been retrieved as an answer.

5.2 Evaluation measures

The evaluation of the *in context* tasks calculates scores for ranked lists of documents, where per document we obtain a score reflecting how well the retrieved text corresponds to the relevant text in the document.

Score per document Two different scores per document are calculated, depending on whether a set of answer parts (*relevant in context*) or a single answer part (*best in context*) are retrieved from the document.

Relevant in context

For a retrieved document, the text identified by the selected set of non-overlapping retrieved parts is compared to the text highlighted by the assessor [7, 15]. More formally, let d be the retrieved document, and let p be a part (element or passage) that belongs to $\mathcal{P}_d$, the set of retrieved parts from document d. Let $Trel(d)$ be the total amount of highlighted relevant text for the document d.

We calculate the following:

- Document precision, as the fraction of retrieved text (in characters) that is highlighted:

$$P(d) = \frac{\sum_{p \in \mathcal{P}_d} rsize(p)}{\sum_{p \in \mathcal{P}_d} size(p)} \tag{4}$$

  The $P(d)$ measure ensures that, to achieve a high precision value for the document d, the set of retrieved parts for that document needs to contain as little non-relevant text as possible.
- Document recall, as the fraction of highlighted text (in characters) that is retrieved:

$$R(d) = \frac{\sum_{p \in \mathcal{P}_d} rsize(p)}{Trel(d)} \tag{5}$$

  The $R(d)$ measure ensures that, to achieve a high recall value for the document d, the set of retrieved parts for that document needs to contain as much relevant text as possible.
- Document F-Score, as the combination of the document precision and recall scores using their harmonic mean, resulting in a score in [0,1] per document:

$$F(d) = \frac{2 \cdot P(d) \cdot R(d)}{P(d) + R(d)} \tag{6}$$

For retrieved non-relevant documents, all the above scores evaluate to zero: $P(d) = R(d) = F(d) = 0$.

We use the F-score as an appropriate document score for the *relevant in context* task:

$$S(d) = F(d) \tag{7}$$

The resulting $S(d)$ score varies between 0 (document without relevant text, or none of the relevant text is retrieved) and 1 (all relevant text is retrieved without retrieving any non-relevant text).

Best in context

The document score $S(d)$ for this task is calculated with a distance similarity measure, $s(x, b)$, which is constructed as follows. For each document in a ranked list, $s(x, b)$ measures how close the system-proposed entry point x is to the BEP b. Closeness is assumed to be an inverse function of distance, with a maximum value of 1 if and only if the system hits the BEP and a minimum value of zero. We first measure the distance $d(x, b)$ in arbitrary units (characters). Next, we remove the arbitrariness by normalising the distance $d(x, b)$ by the actual document length L in characters: $d'(x, b) = d(x, b)/L$. Finally, we make an inverse transformation to a $[0, 1]$ scale: $f(d'(x, b)) = A/(A + d'(x, b))$. The controlling parameter $A > 0$ can be turned up to allow longer distances without much penalty, or down to reward systems which get very close to the BEP. The resulting formula is:

$$s(x, b) = \frac{A \cdot L}{A \cdot L + d(x, b)} \tag{8}$$

A value of $A = 10$ will result in a score close to 1 for any answer in a relevant document, while a value such as $A = 0.1$ will favour systems that return answer parts very close to the BEPs. The official distance similarity score will very likely be based on the value $A = 0.1$.

An alternative formula for calculating the distance similarity measure $s(x, b)$ could be the following:

$$s(x, b) = \begin{cases} \frac{n - d(x,b)}{n} & \text{if } 0 \leq d(x, b) \leq n \\ \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where n is the number of characters representing the visible part of the document that can fit on a screen (typically, $n = 1000$ characters).

We use the $s(x, b)$ distance similarity score as an appropriate document score for the *best in context* task:

$$S(d) = s(x, b) \tag{10}$$

The resulting $S(d)$ score varies between 0 (document without relevant text, or in the case of the alternative formula, the distance of the starting point of the answer part is more than n characters from the BEP) and 1 (the starting point of the answer part is identical to that of the BEP).

Scores for ranked list of documents We have a ranked list of documents $\mathcal{D}$, and for each document we have a document score $S(d_r) \in [0, 1]$, where d_r is the document retrieved at rank r ($1 \leq r \leq |\mathcal{D}|$). Hence, we need generalized evaluation measures, and we utilise the most straightforward generalization of precision and recall [9]. More formally, let us assume that for an INEX 2007 topic there are in total $Nrel$ relevant documents, and let $rel(d_r) = 1$ if document d_r contains highlighted relevant text, and $rel(d_r) = 0$ otherwise.

Over the ranked list of documents, we calculate the following:

– generalized precision ($gP[r]$), as the sum of document scores up to a document-rank r, divided by the rank r:

$$gP[r] = \frac{\sum_{i=1}^{r} S(d_i)}{r} \tag{11}$$

– generalized Recall ($gR[r]$), as the number of relevant documents retrieved up to a document-rank r, divided by the total number of relevant documents:

$$gR[r] = \frac{\sum_{i=1}^{r} rel(d_i)}{Nrel} \tag{12}$$

These generalized measures are compatible with the standard precision/recall measures used in traditional information retrieval. Specifically, the average generalized precision for an INEX 2007 topic can be calculated by averaging the generalized precisions at natural recall points where generalized recall increases (the generalized precision of non-retrieved relevant documents is zero):

$$AgP = \frac{\sum_{r=1}^{|\mathcal{D}|} rel(d_r) \cdot gP[r]}{\sum_{r=1}^{|\mathcal{D}|} rel(d_r)} \cdot gR[|\mathcal{D}|] \tag{13}$$

When looking at a set of topics, the mean average generalized precision ($MAgP$) is simply the mean of the average generalized precision scores per topic.

5.3 Results reported at INEX 2007

For the *in context* tasks we report the following measures over all topics:

– Non-interpolated generalized precision at early ranks:
$gP[r]$, $r \in [5, 10, 25, 50]$; and
– Non-interpolated mean average generalized precision ($MAgP$).

The official evaluation for the *in context* tasks will be based on the overall $MAgP$ measure.

Acknowledgements

We thank James A. Thom for his valuable comments on a draft of this paper.

Bibliography

[1] J. Allan. HARD track overview in TREC 2003 high accuracy retrieval from documents. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 24–37, 2004.

[2] J. Allan. HARD track overview in TREC 2004 high accuracy retrieval from documents. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2004.

[3] C. Clarke, J. Kamps, and M. Lalmas. INEX 2007 retrieval task and result submission specification. In *INEX 2007 Workshop Pre-Proceedings*, 2007 (to appear). http://inex.is.inf.uni-due.de/2007/adhoc-protected/submissions.html.

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

[5] W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli. TREC 2006 genomics track overview. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, 2006.

[6] D. Hiemstra and V. Mihajlovic. The simplest evaluation measures for XML information retrieval that could possibly work. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 6–13, Glasgow, UK, 2005.

[7] J. Kamps, M. Lalmas, and J. Pehcevski. Evaluating Relevant in context: Document retrieval with a twist. In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, 2007 (to appear).

[8] G. Kazai and M. Lalmas. eXtended Cumulated Gain measures for the evaluation of content-oriented XML retrieval. *ACM Transactions on Information Systems*, 24(4):503–542, 2006.

[9] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.

[10] M. Lalmas and B. Piwowarski. INEX 2006 relevance assessment guide. In *INEX 2006 Workshop Pre-Proceedings*, pages 389–395, 2006.

[11] S. Malik, G. Kazai, M. Lalmas, and N. Fuhr. Overview of INEX 2005. In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2005*, volume 3977 of *Lecture Notes in Computer Science*, pages 1–15, 2006.

[12] P. Ogilvie and M. Lalmas. Investigating the Exhaustivity dimension in content-oriented XML element retrieval evaluation. In *Proceedings of the Fifteenth ACM Conference on Information and Knowledge Management (CIKM '06)*, pages 84–93, Arlington, USA, 2006.

[13] J. Pehcevski. *Evaluation of Effective XML Information Retrieval*. PhD thesis, RMIT University, Melbourne, Australia, 2006. http://www.cs.rmit.edu.au/~jovanp/phd.pdf.

[14] J. Pehcevski and J. A. Thom. HiXEval: Highlighting XML retrieval evaluation. In *Advances in XML Information Retrieval and Evaluation: Fourth*

Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2005, volume 3977 of *Lecture Notes in Computer Science*, pages 43–57, 2006.

[15] J. Pehcevski and J. A. Thom. Evaluating focused retrieval tasks. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, Amsterdam, The Netherlands, 2007 (to appear).

[16] B. Piwowarski and G. Dupret. Evaluation in (XML) information retrieval: Expected precision-recall with user modelling (EPRUM). In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 260–267, Seattle, USA, 2006.

[17] S. Robertson. Evaluation in information retrieval. In *European Summer School on Information Retrieval (ESSIR)*, volume 1980 of *Lecture Notes in Computer Science*, pages 81–92, 2001.

[18] A. Trotman and B. Larsen (et al.). INEX 2007 guidelines for topic development. In *INEX 2007 Workshop Pre-Proceedings*, 2007 (to appear). http://inex.is.inf.uni-due.de/2007/adhoc-protected/topics.html.

[19] E. Voorhees. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68, 2004.

[20] C. Wade and J. Allan. Passage retrieval and evaluation. Technical report, CIIR, University of Massachusetts, Amherst, 2005.