INEX 2007 Evaluation Measures

Jaap Kamps[1], Jovan Pehcevski[2], Gabriella Kazai[3], Mounia Lalmas[4], and Stephen Robertson[3]

[1] University of Amsterdam, The Netherlands
kamps@science.uva.nl
[2] INRIA Rocquencourt, France
jovan.pehcevski@inria.fr
[3] Microsoft Research Cambridge, United Kingdom
{gabkaz,ser}@microsoft.com
[4] Queen Mary, University of London, United Kingdom
mounia@dcs.qmul.ac.uk

Abstract. This paper describes the official measures of retrieval effectiveness that are employed for the Ad Hoc Track at INEX 2007. Whereas in earlier years all, but only, XML elements could be retrieved, the result format has been liberalized to arbitrary passages. In response, the INEX 2007 measures are based on the amount of highlighted text retrieved, leading to natural extensions of the well-established measures of precision and recall. The following measures are defined: The Focused Task is evaluated by interpolated precision at 1% recall (iP[0.01]) in terms of the highlighted text retrieved. The Relevant in Context Task is evaluated by mean average generalized precision ($MAgP$) where the generalized score per article is based on the retrieved highlighted text. The Best in Context Task is also evaluated by mean average generalized precision ($MAgP$) but here the generalized score per article is based on the distance to the assessor's best-entry point.

1 Introduction

Focused retrieval investigates ways to provide users with direct access to relevant information in retrieved documents, and includes tasks like question answering, passage retrieval, and XML element retrieval [18]. 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 [7]. The main change in the Ad Hoc Track at INEX 2007 was to allow the 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 can span across a range of elements). In this paper, we will use the term "document part" to refer to both XML elements and arbitrary passages. These changes address requests to liberalize the retrieval format to ranges of elements [2] and to arbitrary passages [16]. However, this

N. Fuhr et al. (Eds.): INEX 2007, LNCS 4862, pp. 24–33, 2008.

simple change had dear consequences for the measures as used up to now at INEX [6, 9, 10, 13, 14]. By allowing arbitrary passages, we loose the "natural" retrieval unit of elements that was the basis for earlier measures. At INEX 2007 we have adopted an evaluation framework that is based on the amount of highlighted text in relevant documents (similar to the HiXEval measures [15]). In this way we build directly on highlighting assessment procedure used at INEX, and define measures that are natural extensions of the well-established measures of precision and recall used in traditional information retrieval [1].

This paper is organized as follows. In Section 2, we briefly describe the ad hoc retrieval tasks at INEX 2007, and the resulting relevance assessments. Then in three separate sections, we discuss the evaluation measures used for each of the INEX 2007 tasks: the Focused Task (Section 3); the Relevant in Context Task (Section 4); and the Best in Context Task (Section 5). We finish with a some discussion and conclusions in Section 6.

2 Ad Hoc Retrieval Track

In this section, we briefly summarize the ad hoc retrieval tasks, and the resulting relevance judgments.

2.1 Ad Hoc Retrieval Tasks

The INEX 2007 Ad Hoc Track investigated the following three retrieval tasks as defined in [3]. First, there is the Focused Task.

Focused Task. This task asks systems to return a ranked list of non-overlapping, most focused document parts that represent the most appropriate units of retrieval. 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 required to choose those non-overlapping document parts that represent the most appropriate units of retrieval.

The second task corresponds to an end-user task 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 original context.

Relevant in Context. This task asks systems to return non-overlapping relevant document parts 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.

The third task is similar to Relevant in Context, but asks for only a single best point to start reading the relevant content in an article.

Best in Context. This task asks systems to return a single document part per document. The start of the single document part corresponds to the best entry point for starting to read the relevant text in the document.

Given that passages can be overlapping in sheer endless ways, there is no meaningful equivalent of the *Thorough Task* as defined in earlier years of INEX.

Note that there is no separate passage retrieval task, and for all the three tasks arbitrary passages may be returned instead of elements. For all the three tasks, systems could either use the title field of the topics (content-only topics) or the cas-title field of the topics (content-and-structure topics). Trotman and Larsen [17] provide a detailed description of the format used for the INEX 2007 topics.

2.2 Relevance Assessments

Since 2005, a highlighting assessment procedure is used at INEX to gather relevance assessments for the INEX retrieval topics [12]. 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]. After assessing an article with relevance, a separate best entry point judgment is also collected from the assessor, marking the point in the article that represents the best place to start reading.

The Focused and Relevant in Context Tasks will be evaluated against the text highlighted by the assessors, whereas the Best in Context Task will be evaluated against the best-entry-points.

3 Evaluation of the Focused Task

3.1 Assumptions

In the Focused Task, for each INEX 2007 topic, systems are asked to return a ranked list of the top 1,500 non-overlapping most focused relevant document parts. The retrieval systems are thus required not only to rank the document parts according to their estimated 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: *The amount of relevant information retrieved is measured in terms of the length of relevant text retrieved*. That is, instead of counting the number of relevant documents retrieved, in this case we measure the amount of relevant (highlighted) text retrieved.

3.2 Evaluation Measures

More formally, let p_r be the document part assigned to rank r in the ranked list of document parts L_q returned by a retrieval system for a topic q (at INEX 2007,

$|L_q| = 1,500$ elements or passages). Let $rsize(p_r)$ be the length of highlighted (relevant) text contained by p_r in characters (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 $Trel(q)$ be the total amount of (highlighted) relevant text for topic q. $Trel(q)$ is calculated as the total number of highlighted characters across all documents, i.e., the sum of the lengths of the (non-overlapping) highlighted passages from all relevant documents.

Measures at selected cutoffs. Precision at rank r is defined as the fraction of retrieved text that is relevant:

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

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.

Recall at rank r is defined as the fraction of relevant text that is retrieved:

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

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 can be biased towards systems that return several shorter document parts rather than returning one longer part that contains them all (this issue has plagued earlier passage retrieval tasks at TREC [20]). Since the notion of ranks is relatively fluid for passages, we opt to look at precision at recall levels rather than at ranks. Specifically, we use an interpolated precision measure $iP[x]$, which calculates interpolated precision scores at selected recall levels:

$$iP[x] = \begin{cases} \max\limits_{1 \le r \le |L_q|} (P[r] \wedge R[r] \ge x) & \text{if } x \le R[|L_q|] \\ \\ 0 & \text{if } x > R[|L_q|] \end{cases} \tag{3}$$

where $R[|L_q|]$ is the recall over all documents retrieved. For example, $iP[0.01]$ calculates interpolated precision at the 1% recall level for a given topic.

Over a set of topics, we can also calculate the interpolated precision measure, also denoted by $iP[x]$, 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 levels, we also calculate overall performance scores

based on the measure of average interpolated precision AiP. For an INEX topic, we calculate AiP by averaging the interpolated precision scores calculated at 101 standard recall levels (0.00, 0.01, ..., 1.00):

$$AiP = \frac{1}{101} \cdot \sum_{x=0.00,0.01,\ldots,1.00} iP[x] \tag{4}$$

Performance across a set of topics is measured by calculating the mean of the AiP values obtained by the measure for each individual topic, resulting in mean average interpolate precision ($MAiP$). Assuming there are n topics:

$$MAiP = \frac{1}{n} \cdot \sum_{t} AiP(t) \tag{5}$$

3.3 Results Reported at INEX 2007

For the Focused Task we report the following measures over all INEX 2007 topics:

- Mean interpolated precision at four selected recall levels: $iP[x]$, $x \in [0.00, 0.01, 0.05, 0.10]$; and
- Mean interpolated average precision over 101 recall levels ($MAiP$).

The official evaluation for the Focused Task is an early precision measure: interpolated precision at 1% recall ($iP[0.01]$).

4 Evaluation of the Relevant in Context Task

4.1 Assumptions

The Relevant in Context Task is a variation on document retrieval, in which systems are first required to rank documents in a decreasing order of relevance and then identify a set of non-overlapping, relevant document parts. We make the following evaluation assumption: *All documents that contain relevant text are regarded as (Boolean) relevant documents.* Hence, at the article level, we do not distinguish between relevant documents.

4.2 Evaluation Measures

The evaluation of the Relevant in Context Task is based on the measures of generalized precision and recall [11], where the per document score reflects how well the retrieved text matches the relevant text in the document. The resulting measure was introduced at INEX 2006 [8, 13].

Score per document. 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. More formally, let d be a retrieved document, and let p be a document part in d. We denote the set of all retrieved parts of document d as $\mathcal{P}_d$. Let $Trel(d)$ be the total amount of highlighted relevant text in the document d. $Trel(d)$ is calculated as the total number of highlighted characters in a document, i.e., the sum of the lengths of the (non-overlapping) highlighted passages.

We calculate the following for a retrieved document d:

- Document precision, as the fraction of retrieved text (in characters) that is highlighted (relevant):
$$P(d) = \frac{\sum_{p \in \mathcal{P}_d} rsize(p)}{\sum_{p \in \mathcal{P}_d} size(p)} \tag{6}$$
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:
$$R(d) = \frac{\sum_{p \in \mathcal{P}_d} rsize(p)}{Trel(d)} \tag{7}$$
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 [19], resulting in a score in [0,1] per document:
$$F(d) = \frac{2 \cdot P(d) \cdot R(d)}{P(d) + R(d)} \tag{8}$$

For retrieved non-relevant documents, both document precision and document recall evaluate to zero.

We may choose either precision, recall, the F-score, or even other aggregates as document score (S(d)). For the Relevant in Context Task, we use the F-score as the document score:
$$S(d) = F(d) \tag{9}$$
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).

Scores for ranked list of documents. Given that the individual document scores ($S(d)$) for each document in a ranked list $\mathcal{L}$ can take any value in $[0, 1]$, we employ the evaluation measures of generalized precision and recall [11].

More formally, let us assume that for a given topic there are in total $Nrel$ relevant documents, and let $IsRel(d_r) = 1$ if document d at document-rank r contains highlighted relevant text, and $IsRel(d_r) = 0$ otherwise. Let $Nrel$ be the total number of document with relevance for a given topics.

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

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

$$gP[r] = \frac{\sum_{i=1}^{r} S(d_i)}{r} \tag{10}$$

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

$$gR[r] = \frac{\sum_{i=1}^{r} IsRel(d_i)}{Nrel} \tag{11}$$

Based on these, the average generalized precision AgP for a topic can be calculated by averaging the generalized precision scores obtained for each natural recall points, where generalized recall increases:

$$AgP = \frac{\sum_{r=1}^{|\mathcal{L}|} IsRel(d_r) \cdot gP[r]}{Nrel} \tag{12}$$

For non-retrieved relevant documents a generalized precision score of zero is assumed.

The mean average generalized precision ($MAgP$) is simply the mean of the average generalized precision scores over all topic.

4.3 Results Reported at INEX 2007

For the Relevant in Context Task we report the following measures over all topics:

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

The official evaluation for the Relevant in Context Task is the overall mean average generalized precision ($MAgP$) measure, where the generalized score per article is based on the retrieved highlighted text.

5 Evaluation of the Best in Context Task

5.1 Assumptions

The Best in Context Task is another variation on document retrieval where, for each document, a single best entry point needs to be identified. We again assume that all documents with relevance are equally desirable.

5.2 Evaluation Measures

The evaluation of the Best in Context Task is also based on the measures of generalized precision and recall [11], where the per document score reflects how well the retrieved entry point matches the best entry point in the document. Note that at INEX 2006 a different, and more liberal, distance measure was used [13].

Score per document. The document score $S(d)$ for this task is calculated with a distance similarity measure, $s(x, b)$, which measures how close the system-proposed entry point x is to the ground-truth best entry point b given by the assessor. Closeness is assumed to be an inverse function of distance between the two points. The maximum value of 1 is achieved when the two points match, and the minimum value is zero.

We use the following formula for calculating the distance similarity measure:

$$s(x,b) = \begin{cases} \frac{n-d(x,b)}{n} & \text{if } 0 \leq d(x,b) \leq n \\ \\ 0 & \text{if } d(x,b) > n \end{cases} \tag{13}$$

where the distance $d(x, b)$ is measured in characters, and n is the number of characters representing the visible part of the document that can fit on a screen (typically, $n = 1,000$ characters).

We use the $s(x, b)$ distance similarity score as the document score for the Best in Context Task:

$$S(d) = s(x,b) \tag{14}$$

The resulting $S(d)$ score varies between 0 (non-relevant document, or the distance between the system-proposed entry point and the ground-truth best entry point is more than n characters) and 1 (the system-proposed entry point is identical to the ground-truth best entry point).

Scores for ranked list of documents Completely analogous to the Relevant in Context Task, we use generalized precision and recall to determine the score for the ranked list of documents. For details, see the above discussion of the Relevant in Context Task in Section 4.

5.3 Results Reported at INEX 2007

For the Best in Context Task we report the following measures over all topics (using $n = 1,000$)

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

The official evaluation for the Best in Context Task is the overall mean average generalized precision ($MAgP$) measure with the generalized score per article is based on the distance to the best-entry point.

6 Discussion and Conclusions

This paper described the official measures of retrieval effectiveness that are employed for the Ad Hoc Track at INEX 2007. The main innovation at INEX 2007 was a liberalization of the allowed retrieval results. Whereas in earlier years all, but only, XML elements could be retrieved, the result format was extended to ranges of elements and arbitrary passages. In order to allow for a fair comparison of the effectiveness of both element-based and passage-based runs, all INEX 2007 measures were based on the amount of highlighted text retrieved, leading to natural extensions of the well-established measures of precision and recall.

The following three measures have been defined: The Focused Task is evaluated by interpolated precision at 1% recall (iP[0.01]) in terms of the highlighted text retrieved. The Relevant in Context Task is evaluated by mean average generalized precision ($MAgP$) where the generalized score per article is based on the retrieved highlighted text. The Best in Context Task is also evaluated by mean average generalized precision ($MAgP$) but here the generalized score per article is based on the distance to the assessor's best-entry point.

Given that the Focused Task measure is defined in terms of recall rather than ranks, it is less straightforward to relate the measure to user's reading effort. As it turned out, the precision at 1% recall was indeed measuring very early precision—usually obtained after one or a few results. That is, given the total length of highlighted or relevant text per topic, and the reasonable precision of the initial results of retrieval systems, the targeted recall was reached within the first few results. Further research is needed to establish whether the chosen recall level corresponds well enough to the intuitions underlying the Focused Task.

The Best in Context Task measure used a window of 1,000 characters around the assessor's best entry point to award a generalized precision score per document, which turned out to be quite lenient. That is, given the total length of Wikipedia articles, and the large fraction of best entry points that are placed relatively early in the article, the generalized precision score is reflecting to a large degree the "article retrieval" component also already awarded in the generalized recall scores. Further research is needed to establish whether the chosen window of characters corresponds well enough to the intuitions underlying the Best in Context Task.

The results of the INEX 2007 Ad Hoc track are detailed in the track overview paper [5].

Acknowledgements

We thank Benjamin Piwowarski and James A. Thom for their valuable comments on earlier drafts of this paper.

Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.072.601, and 640.001.501), and by the E.U.'s 6th FP for RTD (project MultiMATCH contract IST-033104).

References

[1] Baeza-Yates, R., Ribeiro-Neto, B. (eds.): Modern Information Retrieval. ACM Press/Addison Wesley Longman, New York, Harlow (1999)

[2] Clarke, C.L.A.: Range results in XML retrieval. In: Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, pp. 4–5, Glasgow, UK (2005)

[3] Clarke, C.L.A., Kamps, J., Lalmas, M.: INEX 2007 retrieval task and result submission specification. In: Pre-Proceedings of INEX 2007, pp. 445–453 (2007)

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

[5] Fuhr, N., Kamps, J., Lalmas, M., Malik, S., Trotman, A.: Overview of the INEX 2007 ad hoc track. In: Fuhr, N., Lalmas, M., Trotman, A. (eds.) INEX 2006. LNCS, vol. 4518. Springer, Heidelberg (2007)

[6] Gövert, N., Kazai, G.: Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In: Proceedings of the First Workshop of the INitiative for the Evaluation of XML retrieval (INEX), pp. 1–17. ERCIM Publications (2003)

[7] INEX. INitiative for the Evaluation of XML Retrieval (2007), `http://inex.is.informatik.uni-duisburg.de/`

[8] Kamps, J., Lalmas, M., Pehcevski, J.: Evaluating Relevant in Context: Document retrieval with a twist. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 723–724. ACM Press, New York (2007)

[9] Kazai, G.: Report of the INEX 2003 metrics work group. In: INEX 2003 Workshop Proceedings, pp. 184–190 (2004)

[10] Kazai, G., Lalmas, M.: INEX 2005 evaluation measures. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 16–29. Springer, Heidelberg (2006)

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

[12] Lalmas, M., Piwowarski, B.: INEX 2007 relevance assessment guide. In: Pre-Proceedings of INEX 2007, pp. 454–463 (2007)

[13] Lalmas, M., Kazai, G., Kamps, J., Pehcevski, J., Piwowarski, B., Robertson, S.: INEX 2006 evaluation measures. In: Fuhr, N., Lalmas, M., Trotman, A. (eds.) INEX 2006. LNCS, vol. 4518, pp. 20–34. Springer, Heidelberg (2007)

[14] Malik, S., Lalmas, M., Fuhr, N.: Overview of INEX 2004. In: Fuhr, N., Lalmas, M., Malik, S., Szlávik, Z. (eds.) INEX 2004. LNCS, vol. 3493, pp. 1–15. Springer, Heidelberg (2005)

[15] Pehcevski, J., Thom, J.A.: HiXEval: Highlighting XML retrieval evaluation. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 43–57. Springer, Heidelberg (2006)

[16] Trotman, A., Geva, S.: Passage retrieval and other XML-retrieval tasks. In: Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology, Seattle, USA, pp. 43–50 (2006)

[17] Trotman, A., Larsen, B.: INEX 2007 guidelines for topic development. In: Pre-Proceedings of INEX 2007, pp. 436–444 (2007)

[18] Trotman, A., Geva, S., Kamps, J. (eds.): Proceedings of the SIGIR 2007 Workshop on Focused Retrieval, University of Otago, Dunedin New Zealand (2007)

[19] van Rijsbergen, C.J.: Information Retrieval. Butterworths, London (1979)

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