INEX 2006 Evaluation Measures

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

INEX 2006 Evaluation Measures

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

[1] Queen Mary, University of London, United Kingdom
`mounia@dcs.qmul.ac.uk`
[2] Microsoft Research Cambridge, United Kingdom
`{gabkaz,ser}@microsoft.com`
[3] University of Amsterdam, The Netherlands
`kamps@science.uva.nl`
[4] AxIS project group team, INRIA Rocquencourt, France
`Jovan.Pehcevski@inria.fr`
[5] Yahoo! Research Latin America, Chile
`bpiwowar@yahoo-inc.com`

Abstract. This paper describes the official measures of retrieval effectiveness employed at the ad hoc track of INEX 2006.

1 Introduction

Since its launch in 2002, INEX has been challenged by the issue of how to measure an XML retrieval system's effectiveness. The main complication comes from how to consider the dependency between elements when evaluating effectiveness.

As discussed in Section 2, the ad hoc track at INEX 2006 has four retrieval tasks, namely focused task, thorough task, relevant in context task, and best in context task. INEX 2006 uses various sets of measures to evaluate these tasks:

- The XCG measures introduced at INEX 2005 [4] were used to evaluate the thorough and the focused retrieval tasks (Sections 4 and 5, respectively).
- A new generalized precision measure was introduced to evaluate the relevant in context retrieval task (Section 6). This measure is based directly on the text highlighted by the assessors, just as the HiXEval measures [9].
- A distance measure, BEPD, was defined to evaluate the best in context retrieval (Section 7).
- The EPRUM measures originally defined in [10] were adapted to also evaluate the best in context retrieval task (Section 7).

This paper is organized as follows. In Section 2, we describe the INEX 2006 ad hoc retrieval tasks, including their motivations. In Section 3, we describe how relevance is defined in INEX 2006. The evaluations of each task are described in the next four sections (Sections 4 to 7). We finish the paper with some discussions.

2 Ad hoc retrieval tasks

The main INEX activity is the ad hoc retrieval task, where the collection consists of XML documents, composed of different granularity of nested XML elements, each of which represents a possible unit of retrieval. A major departure from traditional information retrieval is that XML retrieval systems need not only score elements with respect to their relevance to a query, but also determine the appropriate level of element granularity to return to users. The user's query may also contain structural constraints or hints in addition to the content conditions. In addition, the output of an XML retrieval system may follow the traditional ranked list presentation, or may extend to non-linear forms, such as grouping of elements per document.

Up to 2004, ad hoc retrieval was defined as the general task of returning, instead of whole documents, those XML elements that are most relevant to the user's query. In other words, systems should return elements that contain as much relevant information and as little irrelevant information as possible. Within this general task, several sub-tasks were defined, where the main difference was the treatment of the structural constraints.

However, within this general task, the actual relationship between retrieved elements was not considered, and many systems returned overlapping elements (e.g. nested elements). What most systems did was to estimate the relevance of XML elements, which is different to identifying the most relevant elements. This had very strong implications with respect to measuring effectiveness, where approaches that attempted to identify the most relevant elements, and to return only those, performed poorly. As a result, the focused task was defined in 2005, intended for approaches aiming at targeting the appropriate level of granularity of relevant content that should be returned to the user for a given topic. The aim was for systems to find the most relevant element on a path within a given document containing relevant information and return to the user only this most appropriate unit of retrieval. Returning overlapping elements was not permitted. The INEX ad hoc general task, as carried out by most systems up to 2004, was renamed in 2005 as the thorough task.

Within the focused and thorough tasks, the output of XML retrieval systems was assumed to be a ranked list of XML elements, ordered by their presumed relevance to the query. User studies [11] suggested that users were expecting to see returned elements grouped per document, and to have access to the overall context of an element. The fetch & browse task was introduced in 2005 for this reason. The aim was to first identify relevant documents (the fetching phase), and then to identify the most relevant elements within the fetched documents (the browsing phase). In 2005, no explicit constraints were given regarding whether returning overlapping elements within a document was allowed. The rationale was that there should be a combination of how many documents to return, and within each document, how many relevant elements to return.

In 2006, the same task, renamed the relevant in context task, required systems to return for each document an unranked set of non-overlapping elements,

covering the relevant material in the document. These elements could be shown to the users, for example, as highlighted text, or through the use of a heat-map.

In addition, a new task was introduced in 2006, the best in context task, where the aim was to find the best entry point, here a single element, for starting to read documents with relevant information. This new task can be viewed as an extreme case of the fetch & browse approach, where only one element is returned per document.

To summarize, INEX 2006 investigated the following four ad hoc retrieval tasks, defined as follows [1]:

- Thorough: This task asks systems to estimate the relevance of all XML elements in the searched collection and return a ranked list of the top 1500 elements.
- Focused: This task asks systems to return a ranked list of the most focused XML elements, where result elements should not overlap (e.g. a paragraph and its container section should not both be returned). Here systems are forced to choose from overlapping relevant elements those that represent the most appropriate units of retrieval.
- Relevant in context: This task asks systems to return to the user the most focused, relevant XML elements 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 elements indicated (e.g. highlighted) within.
- Best in context: This task asks systems to return a single best entry point to the user per relevant document.

For all tasks, systems could use the title field of the topics (content-only topics) or the castitle of the topics (content-and-structure topics) - see [8] for description of the INEX 2006 topics.

3   Relevance Assessments

In INEX 2006, relevance assessments were obtained by assessors highlighting relevant text fragments in the documents, which correspond to wikipedia articles (see [8] for description of the document collection). XML elements that contained some highlighted text were then considered as relevant (to varying degree). A default assumption here is that if an XML element is relevant (to some degree), then its ascendant elements will all be relevant (to varying degrees) due to the subsumption of the descendant elements' content. For each relevant XML element, the size of the contained highlighted text fragment (in number of characters) is recorded as well as the total size of the element (again, in number of characters). These two statistics form the basis of calculating an XML element's relevance score, which in 2006 corresponds to its specificity score [7].

The specificity score, $spec(e_i) \in [0, 1]$, of an element e_i is calculated as the ratio of the number of highlighted characters contained within the XML element,

$rsize(e_i)$, to the total number of characters contained by the element, $size(e_i)$:

$$spec(e_i) = \frac{rsize(e_i)}{size(e_i)} \tag{1}$$

4 Evaluation of the thorough task

4.1 Assumptions

This task is based on the assumption that all XML elements of a searched collection can be ranked by their relevance to a given topic. The task of a system here is then to return a ranked list of the top 1500 relevant XML elements, in descreasing order of relevance. No assumptions are made regarding the presentation of the results to the user: the output of a system here can simply be considered as an intermediate stage, which may then be processed for displaying to the user (e.g. filtered, clustered, etc.). The goal of this task is to test a system's ability to produce the correct ranking. Issues, like overlap (e.g. when a paragraph and its container section are both returned) are ignored during the evaluation of this task.

4.2 Evaluation measures

Two indicators of system performance were employed in the evaluation of the thorough task: effort-precision/gain-recall (ep/gr) graph and mean average effort-precision ($MAep$). These are both members of the eXtended Cumulated Gain (XCG) measures [4], which are extensions of the Cumulated Gain based measures [3]. These were developed specifically for graded (non-binary) relevance values and with the aim to allow information retrieval systems to be credited according to the retrieved documents' degree of relevance.

From the family of XCG measures, ep/gr and $MAep$ were selected as they provide an overall picture of retrieval effectiveness across the complete range of recall. The motivation for this choice is the recall-oriented nature of the task, e.g. rank all elements of the collection and return the top 1500 results. $MAep$ summarizes retrieval effectiveness into a single number, while an ep/gr graph allows for a more detailed view, ploting ep at 100 recall points $\{0.01, 0.02, ..., 1\}$.

Gain value. The definition of all XCG measures is based on the underlying concept of the value of gain, $xG[i]$, that a user obtains when examining the i-th result in the ranked output of an XML retrieval system. Given a ranked list of elements, where the element IDs are replaced with their relevance scores, the cumulated gain at rank i, denoted as $xCG[i]$, is computed as the sum of the relevance scores up to that rank:

$$xCG[i] = \sum_{j=1}^{i} xG[j] \tag{2}$$

Assuming that users prefer to be returned more relevant elements first, an ideal gain vector, xI, can be derived for each topic by filling the rank positions with the relevance scores of the relevant elements in decreasing order of their relevance scores. The corresponding cumulated ideal gain vector is denoted as xCI and is calculated analogue to $xCG[i]$. Both $xG[i]$ and $xI[j]$ are calculated using the element's specificity value:

$$xG[i] = spec(e_i) \tag{3}$$

$$xI[j] = spec(e_j) \tag{4}$$

where e_i is the i-th element in the system ranking, e_j is the j-th element in the ideal ranking, and the specificity score is given in Equation 1.

Effort-precision/gain-recall. Effort-precision at a given cumulated gain value, r, measures the amount of relative effort (where effort is measured in terms of number of visited ranks) that the user is required to spend when scanning a system's result ranking compared to the effort an ideal ranking would take in order to reach the given level of gain (illustrated by the horizontal line drawn at the cumulated gain value of r in Figure 1):

$$ep[r] = \frac{i_{ideal}}{i_{run}} \tag{5}$$

i_{ideal} is the rank position at which the cumulated gain of r is reached by the ideal curve and i_{run} is the rank position at which the cumulated gain of r is reached by the system run.
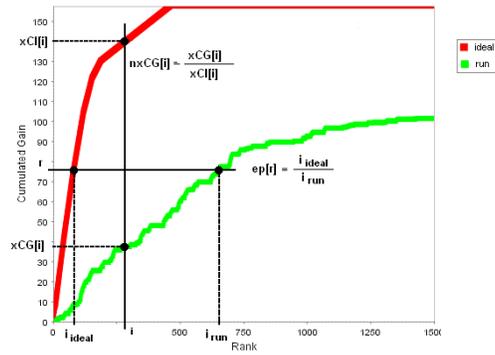

Fig. 1. Calculation of $nxCG$ and effort-precision ep

By scaling the recall axis to $[0, 1]$ (i.e. dividing by the total gain), effort-precision can be measured at arbitrary recall points, $gr[i]$ [5]:

$$gr[i] = \frac{xCG[i]}{xCI[n]} = \frac{\sum_{j=1}^{i} xG[j]}{\sum_{j=1}^{n} xI[j]} \qquad (6)$$

where n is the total number of relevant elements in the full recall-base of the given topic. The range for i is $[0, 1500]$, where 1500 is the maximum length of a result list that participants could submit.

As with standard precision/recall, for averaging across topics, interpolation techniques are necessary to estimate effort-precision values at non-natural gain-recall points, e.g. at standard recall points $\{0.1, .., 1\}$.

The non-interpolated mean average effort-precision, denoted as $MAep$, is calculated by averaging the effort-precision values obtained for each rank where a relevant document is returned. For not retrieved relevant elements, a precision score of 0 is used.

4.3 Results reported at INEX 2006

For the thorough task we report the following measures over all topics:

- non-interpolated mean average effort-precision ($MAep$)
- effort-precision/gain-recall up to rank 1500 (ep/gr)

5 Evaluation of the focused task

5.1 Assumptions

In this task, systems are asked to return the ranked list of the top 1500 most focused, relevant XML elements for each given topic, without returning over-lapping elements. The task is similar to the thorough task in that it requires a ranking of XML elements, but here systems are required not only to estimate the relevance of elements, but also to decide which element(s), from a tree of relevant elements, are the most focused non-overlapping one(s).

5.2 Evaluation measures

The normalized cumulated gain $nxCG[RCV]$ measure, from the XCG family of measures, was used in the evaluation of the focused task. System performance was reported at several rank cutoff values (RCV).

Normalized cumulated gain. For a given topic, the normalized cumulated gain measure is obtained by dividing a retrieval run's xCG vector by the corresponding ideal xCI vector (see Section 4 for the definition of these two vectors):

$$nxCG[i] := \frac{xCG[i]}{xCI[i]} \tag{7}$$

$xCG[i]$ takes its values from the full recall-base of the given topic and $i \in [0, 1500]$ where 1500 is the maximum length of a result list that participants could submit. $xCI[i]$ takes its values from the ideal recall-base (described below) and i ranges from 0 and the number of relevant elements for the given topic in the ideal recall-base. The gain values $xI[j]$ used in $xCI[i]$ are given by Equation 4. The gain values used in $xCG[i]$ are normalized as follows. For the j-th retrieved element, where j ranges from 1 to i:

$$xG_{norm}[j] = min(xG[j], xG[j_{ideal}] - \sum_S xG[k]) \tag{8}$$

where $xG[\cdot]$ is given by Equation 3, j_{ideal} is the rank of the ideal element that is on the same relevant path as the j-th relevant element, and S is the set of elements that overlap with that ideal element and that have been retrieved before rank j. The normalization ensures that a system retrieving all descendant relevant elements of an ideal element cannot achieve a better overall score then if it retrieved the ideal element.

For a given rank i, $nxCG[i]$ reflects the relative gain the user accumulated up to that rank, compared to the gain he/she could have attained if the system would have produced the optimum ranking. As illustrated in Figure 1, $nxCG$ is calculated by taking measurements on both the system and the ideal rankings' cumulated gain curves along the vertical line drawn at rank i. Here, rank position is used as the control variable and cumulated gain as the dependent variable.

Recall-bases. The evaluation of the focused retrieval task requires two recall-bases. The full recall-base is the list of all elements that contains any relevant information (which therefore includes all parents of any such element), already used in the thorough task. The ideal recall-base is a subset of the full recall-base, where overlap between relevant reference elements is removed so that the identified subset represents the set of ideal answers, i.e. the most focused elements that should be returned to the user.

The selection of ideal elements into the ideal recall-base is done by traversing an article's XML tree and selecting from the set of overlapping relevant elements, those with the highest gain value. The methodology to traverse an XML tree and select the ideal elements is as follows [10]: Given any two elements on a relevant path,[6] the element with the higher score is selected. In case two elements' scores are equal, the one higher in the tree is chosen (i.e. parent/ascendant).

---

[6] A relevant path is a path in an article file's XML tree, whose root element is the article element and whose leaf element is a relevant element.

The procedure is applied recursively to all overlapping pairs of elements along a relevant path until one element remains. After all relevant paths in a document's tree have been processed, a final filtering is applied to eliminate any possible overlap among ideal elements, keeping from two overlapping ideal paths the shortest one.

5.3 Results reported at INEX 2006

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

– normalized cumulative gains at low RCV (i.e. early ranks) ($nxCG$ [5, 10, 25, 50])

6 Evaluation of the relevant in context task

6.1 Assumptions

The relevant in context task is document (here Wikipedia article) retrieval, where not only the relevant articles should be retrieved but also a set of XML elements representing the relevant information within each article. In this task, there is a fixed result presentation format defined. Systems are expected to return, for each relevant XML Wikipedia article, a set of elements that focused on the relevant information within the article. The Wikipedia articles should be ranked in decreasing order of relevance, but there should not be a ranking of the contained XML elements. The set of result elements should not contain overlapping elements.

6.2 Evaluation measures

The evaluation of this task is based on a ranked list of articles, where per article we obtain a score reflecting how well the retrieved set of elements corresponds to the relevant information in the article.

Score per article. For a retrieved article, the text retrieved by the selected set of elements is compared to the text highlighted by the assessor [9]. We calculate the following:

– Precision, as the fraction of retrieved text (in bytes) that is highlighted;
– Recall, as the fraction of highlighted text (in bytes) that is retrieved; and
– F-Score, as the combination of precision and recall using their harmonic mean, resulting in a score in [0,1] per article.

More formally, let a be a retrieved article, and let e be an element that belongs to $\mathcal{E}_a$, the set of retrieved elements from article a. Let $rsize(e)$ be the amount of highlighted (relevant) text contained by e (if there is no highlighted text in the element, $rsize(e) = 0$). Let $size(e)$ be the total number of characters

(bytes) contained by e, and let $Trel(a)$ be the total amount of (highlighted) relevant text for the article a.

We measure the fraction of retrieved text that is highlighted for article a as:

$$P(a) = \frac{\sum\limits_{e \in \mathcal{E}_a} rsize(e)}{\sum\limits_{e \in \mathcal{E}_a} size(e)} \qquad (9)$$

The $P(a)$ measure ensures that, to achieve a high precision value for the article a, the set of retrieved elements for that article needs to contain as little non-relevant information as possible.

We measure the fraction of highlighted text that is retrieved for article a as:

$$R(a) = \frac{\sum\limits_{e \in \mathcal{E}_a} rsize(e)}{Trel(a)} \qquad (10)$$

The $R(a)$ measure ensures that, to achieve a high recall value for the article a, the set of retrieved elements for that article needs to contain as much relevant information as possible.

The final score per article is calculated by combining the two precision and recall scores in the standard F-score (the harmonic mean) as follows:

$$F(a) = \frac{2 \cdot P(a) \cdot R(a)}{P(a) + R(a)} \qquad (11)$$

The resulting F-score varies between 0 (article without relevance, or none of the relevance is retrieved) and 1 (all relevant text is retrieved and nothing more).[7] For retrieved non-relevant articles, $P(a) = R(a) = F(a) = 0$.

Scores for ranked list of articles. We have a ranked list of articles, and for each article we have an F-score $F(a_r) \in [0, 1]$, where a_r is the article retrieved at rank r. Hence, we need a generalized measure, and we utilise the most straightforward generalization of precision and recall as defined in [5].

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

- generalized precision ($gP[r]$), as the sum of F-scores up to an article-rank r, divided by the rank r; and
- generalized recall ($gR[r]$), as the number of articles with relevance retrieved up to an article-rank r, divided by the total number of articles with relevance.

---

[7] This task is very similar to the INEX assessors' task, who are highlighting relevant information in a pooled set of articles. Note that the assessors can highlight sentences, whereas systems can only return XML elements. This makes it impossible for a system to obtain a perfect score of 1 (although the theoretical maximum will be close to 1).

More formally, let us assume that for an INEX 2006 topic there are in total *Numrel* articles with relevance, and let us also assume that the function $rel(a_r) = 1$ if article a_r contains relevant information, and $rel(a_r) = 0$ otherwise. At each rank r of the list of ranked articles, generalized precision is defined as:

$$gP[r] = \frac{\sum\limits_{i=1}^{r} F(a_i)}{r} \qquad (12)$$

At each rank r of the list of ranked articles, generalized recall is defined as:

$$gR[r] = \frac{\sum\limits_{i=1}^{r} rel(a_i)}{Numrel} \qquad (13)$$

These generalized measures are compatible with the standard precision/recall measures used in traditional information retrieval. Specifically, the average generalized precision (AgP) for an INEX 2006 topic can be calculated by averaging the generalized precision at natural recall points where generalized recall increases. That is, averaging the generalized precision at ranks where an article with relevance is retrieved (the generalized precision of non-retrieved articles with relevance is 0). 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.

6.3 Results reported at INEX 2006

For the relevant in context task we report the following measures over all topics:

- mean average generalized precision ($MAgP$)
- generalized precision at early ranks ($gP[5, 10, 25, 50]$)

The official evaluation is based on the overall $MAgP$ measure.

7 Evaluation of the best in context task

7.1 Assumptions

In this task, systems are required to return a ranked list of best entry points (BEP), one per article, to the user, representing the point in the article where they should start reading the relevant information in the article. The aim of the task is to first identify relevant articles, and then to identify the elements corresponding to the best entry points for the returned articles. Articles should be ranked according to their relevance.

7.2 Evaluation measures

The evaluation of this task is performed with two measures, a distance measure, BEPD (for BEP-distance), and an extension of precision/recall (EPRUM) [10]. Both measures give a score of 0 for a ranked article that is not relevant, i.e. does not contain a BEP for the current topic.

BEPD. This measure is constructed as follows. For each document in a ranked list, $s(x, b)$ will measure how close the system-proposed entry point x is to the BEP b (as above, s is 0 if the article is not relevant). 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 normalizing d by the average article length L in characters ($d' = d/L$). Finally we make an inverse transformation to a $[0, 1]$ scale ($f(d') = A/(A + d')$), with a controlling parameter $A > 0$, which 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:

$$s(x, b) = A \times \frac{L}{A \times L + d(x, b)} \tag{14}$$

A value of $A = 10$ will give a score close to 1 for any answer in a relevant article; a value such as $A = 0.1$ will favour systems that return elements very close to the BEP.

BEPD for a single topic/ranked list is the sum of s values for the articles in the list divided by the total number of BEPs for this topic. Thus a system is penalized both for not retrieving the right articles and (to some extent controlled by A) for not pointing to the right places in the articles it does retrieve. This measure is averaged in the usual way over topics.

EPRUM-BEP-Exh-BEPDistance. The EPRUM measure is an extension of precision/recall developed for structured document collections and fine-grained user models [10]. While standard precision-recall assumes a simple user model, where the user consults retrieved elements (elements returned by the retrieval system) independently, EPRUM can capture the scenario where the user consults the context of retrieved elements. Most measures assume that a user sees the elements in their order of appearance in the result list. EPRUM on the other hand considers these elements as entry points to the collection from where the user can navigate to find relevant elements.

As in the classical precision at a given recall definition, the recall value R is the number of relevant elements the user wants to see. The recall level ℓ ($0 < \ell \leq 1$) is defined as the ratio of a recall R to the total number T of relevant units. The generalisation lies in the definition of the minimum number of ranks m the user needs to consult in the list to reach a recall level ℓ, or said otherwise a recall value of ℓT.

The user starts considering the first rank of the list. If (s)he finds more than ℓT relevant elements at this rank, then the information need is satisfied and (s)he stops. In this case, the user effort has been restricted to the consultation of the first rank of the list (m is 1). If not, (s)he proceeds to the second rank, etc. The definition of precision is based on the comparison of two minimum values: the minimum rank that achieves the specified recall over all the possible lists, and over the evaluated list. For a given recall level ℓ, precision is defined as as:

$$\text{Precision@}\ell = \mathbb{E}\left[\begin{array}{c}\text{achievement}\\\text{indicator}\\\text{for a recall }\ell\end{array} \times \frac{\begin{array}{c}\text{minimum number of}\\\text{consulted list items for}\\\text{achieving a recall }\ell\text{ over all lists}\end{array}}{\begin{array}{c}\text{minimum number of}\\\text{consulted list items for achieving}\\\text{a recall }\ell\text{ over the evaluated list}\end{array}}\right] \quad (15)$$

where the achievement indicator is used to set the precision to 0 if the recall level cannot be reached for the evaluated list. This is compatible with the classical definition of precision at a given recall where the precision is set to 0 if the list does not contain enough relevant elements.

Similarly, we can extend the definition of precision at a given rank r with this definition:

$$\text{Precision@}r \quad = \quad \frac{1}{r} \quad \times \quad \mathbb{E}\left[\begin{array}{c}\text{minimum number of consulted}\\\text{list items (over all lists)}\\\text{for achieving the same level of}\\\text{recall as the evaluated run}\end{array}\right] \quad (16)$$

EPRUM is defined by three parameters stating: (1) the rewarded elements, i.e. here the BEPs. (2) the relevance value of an element, which is set to 1 since there was only one relevance level (i.e. exhaustivity value [7]) in INEX 2006; and (3) the probability that the user goes from one element in the list to a target (BEP). For the best in context retrieval task, this probability is defined as $s(x, b)$, as defined in Equation 14, for any BEP b. This behaviour is defined stochastically, i.e. we only know that a random user sees the BEP with probability $s(x, b)$ if presented the element x in the list. With these settings, a ranking only made of BEPs will obtain a constant precision of 1 for all recall levels. The performance slowly decreases when returned elements are further away from the BEPs, and reach 0 when returned elements are not in relevant articles.

7.3 Results reported at INEX 2006

For the best in context task we report the following measures over all topics.

- BEPD
- EPRUM-BEP-Exh-BEPDistance precision recall graph
- EPRUM-BEP-Exh-BEPDistance precision averaged over all recall values (mean average precision)

Although we reported results for the values 0.01, 0.1, 1, 10, 100 for the parameter A, the official evaluation is based on the value $A = 0.1$.

8 Discussions

8.1 Too small elements

Because of how relevance was assessed in INEX 2006, a high number of fully highlighted elements – the figure reported at the INEX workshop was 18% – (which will then obtain a specificity score of 1) were of link type (i.e. collection-link, wikipedialink, redirectlink, unknownlink, outsidelink, weblink, etc.). This led to concerns regarding the use of such a set of relevance assessments to evaluate retrieval performance using the XCG measures.

Using the INEX 2005 assessment process would have avoided this problem because any element with some highlighted (relevant) content would have to be further assessed according to how exhaustive it was. The exhaustivity value of "?" was used to express that an element was too small to provide any meaningful information.

We therefore created a second set of assessments, where all link element types were ignored. For both the focused and the thorough tasks, the XCG measures were applied using this filtered assessment set. We then examined correlation when using the two sets of relevance assessments (the full set and the filtered set) by calculating the Kendall τ correlation [2] between their resulting respective system rankings. Previous work has considered all rankings with correlations greater than 0.9 as equivalent and rankings with correlation less than 0.8 as containing noticeable differences [12].

Table 1. Correlation of the XCG measures using the full and the filtered assessments

focused task	
$nxCG[5]$	0.9292135
$nxCG[10]$	0.933427
$nxCG[25]$	0.8989332
$nxCG[50]$	0.8748597
thorough task	
$MAep$	0.9484281

Table 1 shows that for the focused retrieval task, as evaluated by the XCG measures, how to consider the so-called too small elements seems important, as they can affect, to some extent, effectiveness measures. The change between the full to filtered set of assessments does not affect the relevant in context, and best in context tasks, because the metrics used to evaluate these tasks were not affected by the problem of too small elements. The official results of INEX 2006 for the focused and thorough tasks are based on the filtered assessments.

8.2 Ideal recall-base

For the focused retrieval task, as described in Section 5, the cumulated gain $xCG[i]$ at rank i take its values from the full recall-base, whereas the cumulated

gain $xCI[i]$ (for the ideal vector) takes its values from the ideal recall-base. One possible approach is for $xCI[i]$ to also take its value in the full recall-base. The resulting system ranking was compared to the initial one, also using Kendall τ correlation measure.

Table 2. Correlation of the $nxCG[DCV]$ results for the focused task, when the ideal recall-base is build as defined in Section 5 and when it corresponds to the full recall-base

$nxCG[10]$	0.8025281
$nxCG[25]$	0.844944
$nxCG[5]$	0.8185393
$nxCG[50]$	0.8420758

We can see that there is a small difference in the ranking of retrieval approaches. Given that the ideal gain vector is then built from the full set of relevant elements, i.e. the full recall-base, which contains overlapping XML elements, systems can never achieve 100% recall (as the task did not allow runs to return overlapping results). This, however, presents less of an issue when performance is measured at low rank cutoffs. A more serious problem is that what is measured here does not correspond to the task. Since all relevant elements in the full recall-base are considered ideal, systems are in effect rewarded for the retrieval of any relevant element, not just the most focused elements. Therefore, any improvement in performance cannot be attributed to a system's ability in locating the most focused element. We can only say that well performing systems were able to return relevant elements within the top RCV ranks. Furthermore, given that measuring performance at low rank cutoffs is highly sensitive to the individual measuring points, especially those very early in the ranking, the retrieval of relevant, but not ideal elements, can impact on the score quite significantly.

References

1. C. Clarke, J. Kamps, and M. Lalmas. INEX 2006 retrieval task and result submission specification. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *INEX 2006 Workshop Pre-Proceedings*, pages 381–388, 2006.
2. W. Conover. *Practical Non-Parametric Statistics, 2nd edn.* John Wiley & Sons, Inc., New York, NY, USA, 1980.
3. K. Järvelin and J. Kekäläinen. Cumulated Gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (ACM TOIS)*, 20(4):422–446, 2002.
4. G. Kazai and M. Lalmas. eXtended Cumulated Gain Measures for the Evaluation of Content-oriented XML Retrieval. *ACM Transactions on Information Systems (ACM TOIS)*, 24(4):503 − 542, October 2006.
5. 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.

6. M. Lalmas. INEX 2005 retrieval task and result submission specification. In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors, *INEX 2005 Workshop Pre-Proceedings*, pages 385–390, 2005.

7. M. Lalmas and A. Tombros. INEX 2002 - 2006: Understanding XML Retrieval Evaluation. In *DELOS Conference on Digital Libraries*, Tirrenia, Pisa, Italy, 2007.

8. S. Malik, A. Trotman, M. Lalmas, and N. Fuhr. Overview of INEX 2006. In *Comparative Evaluation of XML Information Retrieval Systems, 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*, 2007.

9. 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*, pages 43–57, 2006.

10. B. Piwowarski and G. Dupret. Evaluation in (XML) information retrieval: Expected precision-recall with user modelling (EPRUM). In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

11. T. Tombros, B. Larsen, and S. Malik. The interactive track at INEX 2004. In *Proceedings of the 3rd Workshop of the INitiative for the Evaluation of XML retrieval (INEX), Dagstuhl, Germany, December 2004*, 2005.

12. E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, 2001.