Sound and Complete Relevance Assessment for XML Retrieval

BENJAMIN PIWOWARSKI
Yahoo! Research Latin America
ANDREW TROTMAN
University of Otago
and
MOUNIA LALMAS
Queen Mary, University of London

In information retrieval research, comparing retrieval approaches requires test collections consisting of documents, user requests and relevance assessments. Obtaining relevance assessments that are as sound and complete as possible is crucial for the comparison of retrieval approaches. In XML retrieval, the problem of obtaining sound and complete relevance assessments is further complicated by the structural relationships between retrieval results.

A major difference between XML retrieval and flat document retrieval is that the relevance of elements (the retrievable units) is not independent of that of related elements. This has major consequences for the gathering of relevance assessments. This article describes investigations into the creation of sound and complete relevance assessments for the evaluation of content-oriented XML retrieval as carried out at INEX, the evaluation campaign for XML retrieval. The campaign, now in its seventh year, has had three substantially different approaches to gather assessments and has finally settled on a highlighting method for marking relevant passages within documents—even though the objective is to collect assessments at element level. The different methods of gathering assessments at INEX are discussed and contrasted. The highlighting method is shown to be the most reliable of the methods.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Measurement, Standardization, Performance

Additional Key Words and Phrases: XML, XML retrieval, passage retrieval, evaluation, relevance assessment, INEX

The INEX initiative is funded by DELOS, a network of excellence for digital libraries.

Authors' current addresses: B. Piwowarski, Department of Computing Science, Sir Alwyn Williams Building, 18 Lilybank Gardens, University of Glasgow, Glasgow G12 8QQ, U.K.; email: benjamin@bpiwowar.net; A. Trotman, Department of Computer Science, University of Otago, P.O. Box 56, Dunedin, New Zealand; email: andrew@cs.otago.ac.az; M. Lalmas, Department of Computer Science, Queen Mary, University of London, Miles End Road, E1 4NS, U.K.; email: mounia@dcs.qmul.ac.uk.

ACM Reference Format:
Piwowarski, B., Trotman, A., and Lalmas, M. 2008. Sound and complete relevance assessment for XML retrieval. ACM Trans. Inform. Syst. 27, 1, Article 1 (December 2008), 37 pages. DOI = 10.1145/1416950.1416951 http://doi.acm.org/10.1145/1416950.1416951

1. INTRODUCTION

The aim of content-oriented XML (eXtensible Markup Language) retrieval is to exploit the explicit logical structure of documents to retrieve XML elements (instead of whole documents) in response to a user query [Baeza-Yates et al. 2002; Blanken et al. 2003; Carmel et al. 2000; Luk et al. 2002] . This means that XML retrieval systems must not only find relevant XML documents but must also determine where, within the document, this information is found, as well as the granularity (or size of the fragment) of relevant information. A consequence of this is that the relevance of a result, in this case an XML element, depends both on the content of the result and the granularity of the result. Indeed, the information the user seeks might be contained in a single paragraph within a document, thus when evaluating retrieval effectiveness, finding this paragraph should be rewarded over finding a less appropriately sized element such as the embedding section. Consequently, the relevance assessments should distinguish between the relevance of different information units within a single document.

Evaluating the effectiveness of XML retrieval systems requires a test collection where the relevance assessments are provided according to a relevance criterion that takes into account two aspects: relevance and size. Such a test collection has been developed at INEX,[1] the INitiative for the Evaluation of XML Retrieval. The initiative aims to establish an infrastructure and means, in the form of large XML test collections and appropriate effectiveness measures, for the evaluation of content-oriented retrieval of XML documents.

Following a system-oriented evaluation viewpoint, effectiveness is a measure of a system's ability to retrieve as many relevant and as few non-relevant results as possible. Evaluating effectiveness relies on appropriate measures of relevance. In traditional document-oriented information retrieval, which mainly deals with flat text documents, this can and is often done assuming document independence. An important difference between flat text retrieval and XML retrieval is that, in the latter, the relevance of a result (an XML element) is not independent of other possible results. This is because XML elements can be nested within each other exhibiting a parent-child relationship. The fact that an element has been deemed relevant implies that its parent element must also be relevant, although perhaps to a different extent. When constructing a test collection for evaluating XML retrieval effectiveness it is essential to consider this dependency to obtain appropriate relevance assessments.

The standard document-centric evaluation of retrieval effectiveness is based on the Cranfield methodology [Cleverdon et al. 1966]. This methodology relies on a sound and complete test collection, which comprises a set of documents, a

[1] http://inex.is.informatik.uni-duisburg.de/.

set of information needs (stated in topics), and a set of relevance assessment describing which documents in the collection are relevant to which topics. In this article we describe the methodologies used at INEX (from 2002 through to 2006) for gathering sound and complete relevance assessments. The process is more complex than for document retrieval because of the relationship between XML elements, which must be taken into account.

For large document collections, such as those used since TREC began [Harman 1992], it is not feasible, even with a large number of human assessors, to assess the relevance of all the documents in the collection for each of the topics. Instead, a process called pooling is used to select which documents in the collection should be assessed. In TREC, these documents correspond to those that were highly ranked by participating retrieval systems. INEX, for the same reason, also uses a pooling process, which is adapted to elements and considers element dependency. The pooling process has an impact on the completeness of the assessments, which is crucial for the reusability of the test collection. By completeness, we mean whether a high enough proportion of relevant elements have been identified to perform meaningful comparisons of systems. Completeness is related to the number of elements judged for each topic, and the difficulty is balancing between this number (the higher the better) and resources (often limited, e.g., a small number of human assessors).

The assessments of the relevance of the documents for each given topic of the test collection is often done through the use of an online assessment tool that interfaces between the documents to be assessed and the human assessors. INEX also uses a tool, where elements to assess and the documents containing these elements are displayed. The tool also takes into account the dependency of the elements when their relevance is assessed. The interface between the elements to assess and the human assessors has an impact on the soundness of the assessments, which is also crucial for the reusability of the test collection. By soundness, we mean whether the assessments for a given topic are consistent across assessors. Soundness is related to the amount of freedom given to the assessors to carry out the judging task in the least obstructive way, and the difficulty is balancing between this freedom and the reliability of the assessments.

Determining a methodology for gathering sound and complete relevance assessments has been ongoing at INEX. It should be pointed out that, at INEX, the assessors are not paid to perform the assessment task. The assessors are the INEX participants themselves who are otherwise expected to carry out their ordinary duties. Obtaining sound and complete assessments with minimum impact on the assessors is thus important. Getting reliable assessments in a short period of time is a necessary goal for INEX. Herein (Section 1.1) we give some historical perspective on methods used at INEX. In this section we also list the limitations of the process used to gather the assessments up to 2005, when an improved process was put in place. The remainder of the article is then organized as follow. In Section 2, the two document collections used at INEX are described. The definition of relevance in the various rounds of INEX is described in Section 3. We then describe in Section 4 the current methodology adopted by INEX to ensure complete and sound relevance assessments, and the reasons for it. We analyze in Section 5 the soundness and completeness of the assessments. Finally, in Section 6, we conclude this work, where we include the lessons learned.

1.1 Historical Perspective and Limitations

Assessments have been incomplete since the early days of TREC. There, the top (typically) 100 results from each submitted run are pooled and judged for relevance. All nonjudged documents are considered irrelevant. The validity of pooling has been questioned many times. Indeed, it is possible though unlikely for a new retrieval system to appear that identifies 100 results to a topic, none of which has been previously judged but all of which are relevant. This retrieval system would be considered to have found no relevant results. Nonetheless, thorough investigation into pooling has shown that it is a sound process even though the judgments are incomplete [Zobel 1998].

At INEX, a variant of the TREC pooling method is used—necessitated by the element-centric nature of XML retrieval. Participants are asked, for each topic, to submit a ranked list of up to 1500 elements considered relevant by their retrieval system. At INEX 2002, the top 100 elements were taken from each run to form the pool, exactly as is done at TREC. The assessor was then asked to assess the pool on a document-by-document basis, and within each document, on an element-by-element basis. A criticism of this method was that it is possible (but unlikely) that the pool for a given topic is covered by a single document. As assessors assess on a document-by-document basis, perhaps the pools should contain the same number of documents and not elements. An obvious solution of this problem is to build the pools at document level. We describe this procedure and discuss some alternatives in Section 4.1.

At INEX 2002, the assessment process was long and laborious. An online assessment tool was provided but the process of loading documents and assessing elements was time consuming. In INEX 2003, a new interface designed for the purpose of assessing XML documents was introduced [Piwowarski and Lalmas 2004]. Judges assessed XML elements with respect to two dimensions (see Section 3), each defined on a four-graded scale. So-called enforcement rules (see Section 4.2.2) were used to contribute to the completeness and the soundness of assessments. While the assessment process was better than in 2002, substantial limitations remained:

(1) It was not possible to assess an element as nonrelevant if it was the child of a fully relevant element. For example, a numeric citation in a paragraph of relevant text is itself only a number. The paragraph is relevant but contains a nonrelevant piece of text

(2) Using a four-graded scale for each dimension was complex, which resulted in low agreement levels between assessors.

(3) The two dimensions were interdependent and the assessors had to handle both at the same time. It became clear that the assessors found it difficult to distinguish between the dimensions [Pehcevski et al. 2005].

(4) An XML element was not always the best unit of retrieval. In some cases a sequence of paragraphs (not collectively a single element) fulfilled the user's need.

The enforcement rules also contributed to limitations:

(5) It was often the case that only part of a document was assessed. As only elements in the pool were required to be assessed, those parts of a document not in the pool were not assessed.
(6) The enforcement rules were not well understood by assessors, who questioned whether the rules increased or decreased the reliability of the assessments—certainly they interfered with assessor behavior.

None of these limitations were addressed until 2005 as the assessment procedure remained unchanged from 2003 to 2004.

Throughout the remainder of this article, we will refer to these limitations, discuss the changes made between 2002 and 2006, and show that each has been addressed with the current assessment procedure established for 2006. We believe the procedure is now stable and unlikely to change to any great extent.

The INEX relevance scale has been modified many times and is discussed in Section 3. Each time the scale was changed in an effort to overcome limitations (1) and (2). The simplifications have continued and at INEX 2006, although the two dimensions remained, one dimension became binary.

In Section 4, we detail the pooling process and the interface used to judge XML documents and elements. Ensuring that the pool is as complete as possible (i.e., maximizing the number of elements to be assessed) is detailed in Section 4.1. The interface used for assessment is described in Section 4.2.1 along with changes to alleviate limitation (3). The enforcement rules that contributed to the gathering of sound and complete assessments were adapted to the interface, and are described in Section 4.2.2.

Substantial changes in the judgments method include the change from manually assigning relevance values to elements to one of highlighting passages of relevant text within a document. A direct consequence of this is that all relevant (and irrelevant) elements in a document are now identified (which addresses limitation (5)), and one simple enforcement rule now suffices (limitation (6)). Overall, the changes resulted in a much more natural and non-intrusive way of assessing (addressing limitations (3) and (4)).

The changes also have consequences on two important factors of assessments: the completeness (i.e., the proportion of relevant elements that have been assessed) and the soundness (i.e., the reliability of the assessments). In Section 5, we present an analysis of the soundness and completeness of the collected assessments.

2. THE INEX DOCUMENT COLLECTIONS

Between 2002 and 2004 the INEX test collection consisted of the full text of 12,107 articles, marked up in XML, from 12 magazines and six transactions of the IEEE Computer Society's publications, covering the period of 1995–2002, and totaling 494 MB, and 8 million elements. The collection contained scientific articles of varying length. On average, an article contained 1532 XML nodes, where the average depth of the element was 6.9.

The overall structure of a typical article consists of front matter, a body, and back matter. The front matter contains the article's metadata, such as the title, author, publication information, and abstract. Following it is the article's body, which contains the actual content of the article. The body is structured into sections, subsections, and subsubsections. These logical units start with a title and are followed by a number of paragraphs. The content has additional markup for references (citations, tables, and figures), item lists, and layout (emphasized, bold, and italic text). The back matter contains a bibliography and further information about the article's authors.

For 2005 the collection was extended with a total of 4712 new articles from the period 2002–2004, giving a total of 16,819 articles, leading to a total of 764 MB, and 11 million elements.

In 2006 the collection changed to the Wikipedia collection of 659,388 articles taken from the English Wikipedia totaling about 4.6 GB of data. There are about 52 million elements and about 1200 unique tags (compared to 176 in the IEEE collection) [Denoyer and Gallinari 2006]. On average, an article contains 161.35 XML nodes, where the average depth of an element is 6.72.

The structure of a Wikipedia article contains a body, which contains the article content. The body is structured into sections with a title and followed by a number of paragraphs. Additional markup for cross references, tables, and emphasis is also present. The XML elements are of varying size and nested.

3. RELEVANCE IN INEX

In many information retrieval evaluation frameworks, the relevance value of a document is restricted to 0 (not relevant) or 1 (relevant), where the basic threshold for relevance is defined as a mention of the topic at hand [Harman 1995]. In XML, elements are of varying size and nested. As relevant elements can be at any level of granularity, an element and one of its children can both be relevant to a given query, but to a different extent. Indeed, if there is a relevant section in a document then the document must also be relevant. In addition, if the section is a more focused answer to the query then it is a better answer for that query.

Using binary relevance for XML retrieval, an article would be just as relevant as a section and a paragraph. It is, thus, not possible to state that the article is relevant, but to a lesser or greater extent than the section. INEX consequently opted for graded relevance assessments. A summary of the relevance scales at INEX along with the document collection it was used for is given in Table I.

Relevance at INEX 2002 was defined on a two-dimensional scale of topical relevance and component coverage. The former was marked on a four-point scale of irrelevant (0), marginally relevant (1), fairly relevant (2), or highly relevant (3). The latter was marked on a separate four-point scale of no coverage (N), too large (L), too small (S), or exact coverage (E). The component coverage dimension was included as a method of rewarding systems capable of

Table I. Relevance Dimensions, Scales, and Collections Used at INEX 2002–2006

INEX	Dimensions	Interface	Collection
2002	Relevance $\in \{0, 1, 2, 3\}$ Coverage $\in \{N, L, S, E\}$	XML view Manual annotation	IEEE
2003	Exhaustivity $\in \{0, 1, 2, 3\}$ Specificity $\in \{0, 1, 2, 3\}$	Document view Assessment selection	IEEE
2004	Exhaustivity $\in \{0, 1, 2, 3\}$ Specificity $\in \{0, 1, 2, 3\}$	Document view Assessment selection	IEEE
2005	Exhaustivity $\in \{0, 1, 2, ?\}$ Specificity $\in [0, 1]$	Document view Highlight then assess	Extended IEEE
2006	Exhaustivity $\in \{0, 1\}$ Specificity $\in [0, 1]$	Document view Highlight then add BEP[a]	Wikipedia

[a] BEP is discussed in Section 5.2.5.

retrieving the appropriate (or "exact") sized elements for a given query. For example, a retrieval system that is able to locate the only relevant section within a book is more effective than one that returns a chapter (or the whole book). For the purpose of evaluation these relevance scores were quantized into a single scale—several different quantizations have been proposed, but a discussion of quantization functions and effectiveness measures is beyond the scope of this article (see Lalmas and Tombros [2007] for more details).

In a study of the assessments from INEX 2002, the component coverage dimension was shown to have been misunderstood by the assessors [Kazai et al. 2004]. Whereas the scale was designed to describe the relationship between the relevant and irrelevant content of an element, it was incorrectly used to describe the relationship between the result element and the preferred result element. In other words, even if an answer fully satisfied the user's information need, it was judged too small if it was in a subsection but the user thought the answer would be in a section. More technically, the scale was used to describe the relation between the returned element and the target element, not the information content of the element [Kazai et al. 2004].

Consequently, at INEX 2003, the two dimensions were changed to exhaustivity and specificity. Exhaustivity measured the extent to which the given element covered or discussed the topic of request. Specificity measured the extent to which the given element was focused on the topic of request. Since exhaustivity is analogous to topical relevance, the scale was redefined by simply replacing the name of one with the name of the other. As specificity is not analogous to coverage, the scale changed so that specificity was also defined on an ordinal scale (see Table I).

Some XML elements are exhaustive but not specific to a given query; they might be too large or additionally contain information not relevant to the query. Other elements will be perfectly specific to a query, but not exhaustive, as they satisfy only a part of the information need. By combining the two criteria it becomes possible to identify those relevant elements that are both exhaustive and specific to the topic of request, and hence represent the most appropriate unit to return to the user. When evaluating XML retrieval effectiveness with the two dimensions (and their scale), it is possible to reward systems that are able to retrieve these most appropriate elements.

The relevance scale itself was based on the work of Kekäläinen and Järvelin [2002], and four points (not (0), marginally (1), fairly (2), highly (3)) were used for each dimension. A nomenclature for discussing the relevance of an element to a query was also adopted. An assessment is denoted *EeSs* where e and s are integers between 0 and 3 for each of exhaustivity (E) and specificity (S), respectively. For example, an *E2S3* element is "fairly exhaustive and highly specific" to the topic.

Exhaustivity and specificity values are not independent of each other. A nonexhaustive element (*E0Ss*) is also, by definition, not specific (must be *EeS0*), and vice versa. There are therefore only 10 valid relevance points. An element is considered relevant if $e > 0$ and $s > 0$. An element is not relevant if its assessment is *E0S0*.

In 2004, 12 topics were each assessed by two judges, each without knowledge of the other. Trotman [2005] converted the 10-point relevance scale into binary relevance judgments (relevant or not) and computed the whole-document-only agreement level. About 27% of judged elements had the same relevance assessment. Although low, this is in line with those of TREC 6 (0.33) and TREC 4 (0.42–0.49) [Cormack et al. 1997; Voorhees 2000]. The exact *EeSs* agreement of elements was found to be 0.16. Compared to the 0.27 binary whole-document agreements, this is extremely low. It is reasonable to conclude that the judges do somewhat agree on which documents are relevant, but not on why or how relevant those documents are.

As part of the interactive track at INEX 2004 [Tombros et al. 2004], participants were asked to judge elements for a small number of topics. The obtained relevance assessments were then compared to those of the INEX assessors [Pehcevski et al. 2005]. It was shown that the participant and INEX assessor's agreement levels are high only at the extreme ends of the relevance scale (*E0S0* and *E3S3*).

Furthermore, an observation made by Clarke [2005] was that the assessment process could be simplified if first relevant passages of text were identified by highlighting, then the elements within these passages were assessed. However, this leads to a paradox: an element can be a part of a relevant passage, while at the same time not relevant on its own (it might, for example, be a single italicized word). It is *both relevant and not relevant* at the same time.

As a consequence of the studies into assessment agreement, the assessment method was changed for INEX 2005. The definition of the relevance dimensions remained unchanged, but the scale was revised. In addition, assessments were gathered using the highlighting method (see Section 4.2.1) and a too small assessment (denoted "?") for paradoxical elements was adopted in the exhaustivity dimension (maintaining the four-point scale). Specificity became implicit because it could be measured directly from the amount of highlighted content. These changes directly addressed limitation (2) and limitation (3) from Section 1.1—the relevance scale was simplified and a clear distinction was made between the two relevance dimensions.

Summarizing, the INEX 2005 exhaustivity dimension was defined according to the following scale:

- Highly exhaustive (2): the XML element discussed most or all aspects of the topic of request.
- Partly exhaustive (1): the XML element discussed only few aspects of the topic of request.
- Not exhaustive (0): the XML element did not discuss the topic of request.
- Too-small (?): the XML element contains relevant material but is too small to be relevant on its own.

It is important to distinguish the exhaustivity value of too small (at INEX 2005) from the coverage value of too small (at INEX 2002). The former was introduced to allow assessors to label elements, which, although they contained relevant information, were too small to sensibly reason about their level of exhaustivity (such as reference numbers, or single highlighted words). In 2002 the too small category was with respect to the coverage—in other words, regardless of the size of the element, the content was insufficient to be useful to the user.

The specificity of an element was measured on a continuous scale. The assessor highlighted relevant passages and the specificity value of an element was computed directly from this. An element that was completely highlighted had a specificity value of 1. An element not highlighted at all had a specificity value of 0. For all other cases, the specificity value was defined as the ratio (in characters) of the highlighted text (i.e., relevant information) to the element size.

The relevance of an element continued to be expressed *EeSs*, but, $e \in \{2, 1, 0, ?\}$, and $s \in [0, 1]$. An element was relevant if it intersected a relevant passage, that is, if both specificity and exhaustivity were nonzero. By extension, a document was relevant if it contained relevant content.

In INEX 2006 the exhaustivity dimension was transformed to a binary scale because an investigation by Ogilvie and Lalmas [2006] showed that a multipoint scale was not necessary to soundly rank retrieval systems relative to each other. The scale for specificity remained unchanged.

4. OBTAINING COMPLETE AND SOUND RELEVANCE ASSESSMENTS

Comparing the effectiveness of retrieval approaches requires test collections where the relevance assessments are as sound (i.e., accurate) and as complete as possible as it is against these judgments that retrieval approaches are evaluated. In the context of XML retrieval, producing complete assessments for collections the size of the IEEE or Wikipedia collections is prohibitively time consuming as every element in every document needs to be assessed for a statistically large enough number of topics. It would also be a complex and tedious task because the relevance of an XML element cannot be assessed in isolation from other elements. This section discusses how INEX elicits the elements to assess using a pooling method similar to that of TREC, then how INEX contributes to the gathering of sound assessments. This is based on an especially designed online interface for assessing elements and the use of enforcement rules.

4.1 Completeness

A criticism of the INEX 2002 (element-based) pooling method (see Section 1.1) was that all relevant elements might be found in a single document; therefore, in 2003 a change was made and document-based pooling has been used since. From each run, the document in which the top-ranked element is found is added to the pool. Then the document for the second top element is added, and so on until the pool contains 500 unique documents. This method is known as the *top-n method* as the pool contains the top ($n = 500$) most highly ranked documents identified in the participants' submissions. Of course, if n unique documents are not identified by the runs, the pool can be short—however, this will not adversely affect the evaluation as all documents in all runs will be in the pool. Piwowarski and Lalmas [2004] showed that assessing elements in the same document is less time-consuming than assessing elements in different documents. They also showed that this approach increases the number of assessed elements without adversely impacting the total assessment load.

Alternative pooling methods have been suggested. Woodley and Geva [2005] suggested using a metasearch engine to choose part of the pool and top-n to choose the remainder. They experimented with the Borda Count metasearch approach (a sum of inverse ranks) and showed that the approach was at least as good as the top-n approach. By setting n to 50, they measured the precision of the two methods against the official assessments and showed that their method had a higher precision—that is, there were more relevant results in their top 50 than the top-n top 50. Advantageously it also adds some elements that were chosen by many search engines, but were not in the top 100 of any one search engine. Similar results were found independently by Vu and Gallinari [2005] using the RankBoost algorithm to merge the (whole document) results. Nonetheless, top-n ($n = 500$) was used at INEX between 2003 and 2006, perhaps in an effort to avoid changing too much in any one year.

In 2002 and 2003, due to the burden of assessing elements one by one, the number of assessed elements within a document was limited. In 2002, for each element, the assessor had to manually write a two-letter code (one letter for each relevance dimension). In 2003, thanks to a set of new rules, some parts of the documents could be automatically assessed. In 2003, the element pool was also dynamic. Some elements were added while the judge was assessing whereas others were removed as a consequence of assessing. We give two examples. For example, judging an element as relevant but not highly specific would add its children in order to find more specific elements. The rationale was that, as highly specific elements are more likely to be of interest for a real user, it is important to identify such elements. As another example, if the assessor judged the whole article to be not relevant then they were not required to assess each and every element within that document.

Using an online assessment tool (see Section 4.2.1), the assessor loaded a document from the pool and judged the elements from the submitted runs. The process involved reading an element and manually assigning a relevance score (*EsSs* value) to it. The process was laborious. If, for example, a paragraph was identified by a retrieval system, then it needed to be assessed, as did any subsection containing it, that is, the subsection, the section, the body, and then the document. This propagation of relevance up the XML tree caused the element pool to constantly change throughout the assessment process.

Clarke [2005] suggested that using a highlighting method might decrease the assessment load. It became obvious that this had an additional advantage over previous methods. Not only were all run elements judged but, by highlighting all relevant passages in a document, all elements within a document were judged, thus leading to more complete assessments (see Section 5). The irrelevance of all nonhighlighted elements was implicit in the fact that they had not been highlighted. Furthermore, there was no longer any dynamic propagation of relevance to resolve during assessment.

In summary, by INEX 2005, the document-based top-*n* pooling method was in use and assessors were assessing full documents with the highlighting method. Assessors were then manually assigning exhaustivity values to the highlighted elements. Just like in 2003 and 2004, elements identified by a participant's retrieval system (within a pooled document) were shown to assessors while assessing. Unlike in 2003 and 2004, assessors were asked to identify relevant passages rather than to make decisions as to the relevance of a previously identified element. Finally, for 2006, as the result of the investigation carried out by Ogilvie and Lalmas [2006], the exhaustivity dimension was reduced to binary and the second pass of manually assigning exhaustivity values was dropped.

Comparing the pooling and assessment method of 2006 with that of previous years, more documents are now being assessed, and all elements within a document are now assessed. In addition, as shown in Section 5.1, it takes less time to assess. It is reasonable to conclude that the assessments are more complete than they have been in previous years.

As yet, no experiments have been conducted in INEX to assess previously un-assessed elements to see exactly how many relevant elements remain and the effect on relative system performance of having these assessments. Such a comparison would involve assessing further documents as the top-*n* pooling method results in whole document assessment. Section 5.2.1 provides, nonetheless, some insights into this issue. To conclude, the 2005 and 2006 assessment methods ensured that all elements within a document were assessed, even if not identified as relevant (i.e., retrieved) by any retrieval system.

4.2 Soundness

The judgments are a snapshot of the mind of the assessor at the moment the assessment decision is made. If presented with the same element and query at a later date, the assessor might make a different decision. This is a well-known problem with the Cranfield methodology and has been thoroughly explored by others (e.g., Saracevic [1991]). Nonetheless, the principle of comparing the results of a retrieval system to those of a human is believed to be sound because any errors are likely to be random and thus equally fair (or unfair) to all systems. However, the judgments themselves must be sound—something that can be confirmed by comparing the assessment sets of multiple judges assessing the same topics. Of course enough topics have to be multiply assessed to draw any firm conclusions. Further, if only a subset of topics is chosen for multiple assessments then the subset must be representative of the whole, and it must be judged by a large enough and representative sample of users. Further investigation is necessary to determine if this is the case. This section is concerned with the soundness of the assessments under the assumption that all these requirements are met.

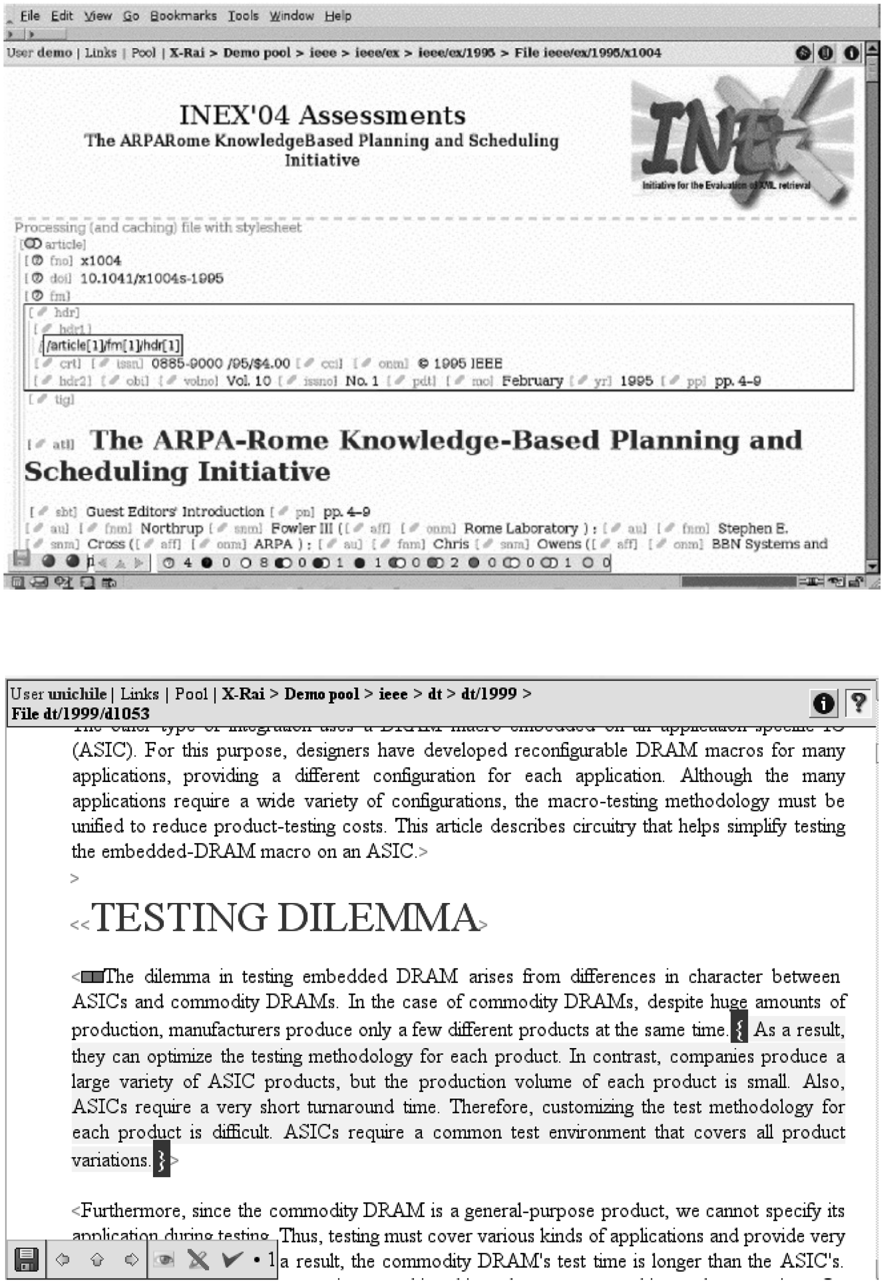
At INEX, the agreement levels between judges have been low when compared with those at TREC (see Section 5.2.6). If the INEX collection is to be reliable then cross-judge agreement should be comparable to that seen in known reliable test collections—when measured on a like-to-like basis.

At INEX there are only two ways the agreement level can be influenced. First, the topic contains a <narrative> field, which is the definition (by the topic author) as to what makes a piece of information relevant. As INEX has progressed, more care has been taken to ensure that both relevance and irrelevance are discussed in the <narrative>. A description of the work task, that is, why the information is needed is also expected in the <narrative>. But, being natural language, the <narrative> can remain ambiguous. Second, the agreement level can also be influenced by the assessment interface (Section 4.2.1), and the rules enforcing internal consistency of assessments by a single assessor (Section 4.2.2). Although these rules remove the ability of the assessor to create contradictory judgments, they can lead to inconsistent assessments across assessors.

4.2.1 *The Assessment Interface.* In this section we describe the interface developed for INEX in 2003 for collecting the relevance assessments. The purpose of the interface was to not only ensure sound and complete relevance assessments, but also to ease the assessment process. This is crucial because the assessors are the INEX participants, who are not paid to perform the task but do it in addition to their ordinary duties.

Although an assessment interface was used in 2002, it was laborious to use. In 2003, a project to log and study judge behavior was initiated and, as part of that project, completely new purpose-designed tools were put in place. For 2003 and 2004, the interface displayed the document with each pool element from the runs identified in a box. Assessors were asked to click on each box and assign a relevance score (*EeSs* value) to the box (see Figure 1). This method of assessing was replaced with highlighting in 2005.

The 2005 assessment procedure had two phases. In the first phase (as shown in Figure 1), assessors highlighted text fragments (or passages) that contained only relevant information. Highlighting was based solely on relevance and was irrespective of the two dimensions and their scales. Assessors were asked not to highlight larger fragments if they contained irrelevant fragments; only purely relevant information fragments were to be highlighted. To decide which text to highlight, assessors were asked to skim-read the whole document and to identify all relevant information. The interface assisted the assessor by highlighting assessor-chosen keywords within the article and by showing all elements identified by the retrieval systems within the document. In the second phase,


File Edit View Go Bookmarks Tools Window Help
Initiative for the Evaluation of XML retrieval
◷ 4 ● 0 ○ 8 0 1 ● 1 0 2 ● 0 0 1 ○ 0
• 1

User demo | Links | Pool | X-Rai > Demo pool > ieee > ieee/ex > ieee/ex/1995 > File ieee/ex/1995/x1004

INEX'04 Assessments
The ARPARome KnowledgeBased Planning and Scheduling Initiative

Processing (and caching) file with stylesheet
[article]
[fno] x1004
[doi] 10.1041/x1004s-1995
[fm]
[hdr]
[hdr1]
/article[1]/fm[1]/hdr[1]
[crt] [issn] 0885-9000 /95/$4.00 [cci] [onm] © 1995 IEEE
[hdr2] [obi] [volno] Vol. 10 [issno] No. 1 [pdt] [mo] February [yr] 1995 [pp] pp. 4–9
[tig]

[atl] The ARPA-Rome Knowledge-Based Planning and Scheduling Initiative

[sbt] Guest Editors' Introduction [pn] pp. 4–9
[au] [fnm] Northrup [snm] Fowler III ([aff] [onm] Rome Laboratory) : [au] [fnm] Stephen E. [snm] Cross ([aff] [onm] ARPA) : [au] [fnm] Chris [snm] Owens ([aff] [onm] BBN Systems and

User unichile | Links | Pool | X-Rai > Demo pool > ieee > dt > dt/1999 >
File dt/1999/d1053

The other type of integration uses a DRAM macro embedded on an application specific IC (ASIC). For this purpose, designers have developed reconfigurable DRAM macros for many applications, providing a different configuration for each application. Although the many applications require a wide variety of configurations, the macro-testing methodology must be unified to reduce product-testing costs. This article describes circuitry that helps simplify testing the embedded-DRAM macro on an ASIC.>
>

<<TESTING DILEMMA>

<The dilemma in testing embedded DRAM arises from differences in character between ASICs and commodity DRAMs. In the case of commodity DRAMs, despite huge amounts of production, manufacturers produce only a few different products at the same time.{ As a result, they can optimize the testing methodology for each product. In contrast, companies produce a large variety of ASIC products, but the production volume of each product is small. Also, ASICs require a very short turnaround time. Therefore, customizing the test methodology for each product is difficult. ASICs require a common test environment that covers all product variations.}>

<Furthermore, since the commodity DRAM is a general-purpose product, we cannot specify its application during testing. Thus, testing must cover various kinds of applications and provide very a result, the commodity DRAM's test time is longer than the ASIC's.

Fig. 1. INEX 2004 (top) and 2005 (bottom) assessment interfaces.

assessors were asked to assess the exhaustivity of relevant elements (those elements that intersected with any highlighted passage).

By 2006, this second pass was eliminated as exhaustivity scores were implicit in the highlighting. This was done by assigning every element with highlighted

Table II. Number of Soundness Enforcement Rules at Each Campaign

INEX	2002	2003	2004	2005	2006
Rules	1	6	8	4	1

content an exhaustivity value of 2 and those without highlighted content a value of 0. The assessment process involved just the highlighting phase.

4.2.2 *Enforcement Rules.* Enforcement rules were first introduced in 2002. That year only one rule was used, which ensured that a relevant element could not be more exhaustive than its parent:

2002-Rule [Exhaustivity Growth]. An XML element cannot be more exhaustive than its parent element.

This rule expanded to a set of rules for subsequent rounds of INEX, including rules for when an element was to be assessed, and what possible relevance values it could take. Details were reported by Piwowarski and Lalmas [2004]. By INEX 2004, the enforcement rules had became too numerous (see Table II) and too demanding for the assessors, who were obliged to explicitly assess most of the elements (each individually) of a document if it contained any relevant elements. Worse, as the rules were enforced while the assessors were assessing, the assessors were often assessing elements out of sequential reading order. If, for example, the second paragraph of a section was assessed relevant, then, because the section had to be relevant, a separate manual assessment of the section was needed that necessitated that each child of the section required assessment—except the second paragraph, which had already been assessed.

By moving to a highlighting method, most of the 2004 rules became void. However a new set of rules was needed to maintain the relationship between highlighted text and the document elements. These rules fell into three categories: those enforcing highlighting (H-Rules), those enforcing completeness (C-Rules), and those enforcing soundness (S-Rules). For 2005, the rules were the following.

H-Rule 1 [ATOM]. Two relevant passages cannot be immediately adjacent.

This rule was enforced by merging adjacent highlighted passages. Should the assessor highlight two disjoint passages and the adjoining text, then the three passages were merged into a single passage encompassing the entire highlighted text. We expected this situation to occur if the assessor was highlighting a very long passage of text and was forced to scroll while doing so. Furthermore, highlighting was related to specificity only and, in order to alleviate as much as possible the assessment burden, assessors were not asked to identify passages but rather to highlight relevant material; isolating adjacent passages was thus neither necessary nor desirable.

C-Rule 1 [ALL]. All highlighted or partly highlighted elements must be assessed.

C-Rule 2 [ONLY]. Only highlighted or partly highlighted elements may be assessed.

These two rules ensured that only relevant elements and all relevant elements within a document were assessed (i.e., they are given an exhaustivity value).

S-Rule 1 [GAIN]. An element cannot be more exhaustive than its parent.

This rule ensured that exhaustivity did not decrease as more information (highlighted content) was added to an already assessed element.

In 2005, there was no checking on the allowed exhaustivity values during assessment. If there was a conflict after the judge assessed an element, then the conflicting assessments were reset and identified for reassessment. If, for example, a section containing one paragraph was assessed *E1* and the paragraph was then assessed *E2*, then the section assessment was reset. The reasoning was that this would decrease the obtrusiveness of the interface and thus address limitation (6) listed in Section 1.1.

In INEX 2006, exhaustivity was binary so only the first rule was needed, which is nonintrusive. The new definition of exhaustivity completely removed the burden of the rules as the only effect on the assessor was that adjacent highlighted passages were automatically merged. Limitation (6) was then fully addressed.

Since 2002 the rules evolved reflecting the different observations made at the INEX workshops. In 2006, they reached a state from which it appears that they cannot be simplified further. Completeness (within a document) has been addressed in full as judges now assess whole documents. This, in turn, was made possible by the successive simplification of the assessment process, and especially the use of highlighting. Highlighting hides the specificity dimension from the judge and is a much more intuitive and accurate process than before.

It is difficult to be certain that soundness has been enforced, but the more natural way of assessing has decreased the judge's cognitive load. Estimating the specificity, jumping within a document to assess distant elements, and so on no longer occurs. The increase in both soundness and completeness are a direct effect of the pooling process and assessment interface.

5. ANALYSIS

This section presents an analysis of the effect of the assessment procedure used at INEX 2005 and 2006 on the soundness and completeness of the relevance assessments. In Section 5.1, an analysis of the logs gathered during the assessment process is presented showing that the time to assess a document has decreased with each change made to the assessment process. In Section 5.2, an analysis of the completeness of the assessments is given along with details of the agreement levels. It is also shown that the agreement level, and therefore soundness, have increased with each change made to the assessment process.

5.1 Assessing

To analyze assessor behavior, we used logs generated by the assessors using the interface to perform the actual judgments during the multiple INEX campaigns.

Table III. Average Time Taken to Assess a Topic at Each Round of INEX (Times Given Are in Hours, Minutes, and Seconds; No Logging Mechanism Was in Place in 2002.)

INEX	Relevant Document	Irrelevant Document	Whole Topic
2002	—	—	—
2003	8 min	1 min	21 h
2004[a]	2 min	2 min	13½ h
2005	5 min	50 s	11 h
2006	1 min	45 s	7 h

[a] The full logs for 2004 are not available for analysis; these are estimates based in partial logs (no seconds were available and the exact time the document was loaded is not known). The results for 2004 should be considered nothing more than estimates included for completeness.

In 2003 and 2004, the log file contained the details of which elements in which documents were assessed with which relevance scores, and at what time. For 2005 each line of the log corresponded to an action of loading, highlighting, unhighlighting, or assessing exhaustivity, along with the time of the action. In 2006, exhaustivity was implicit so the log files do not contain this last action.

5.1.1 *Time to Assess.* The time to assess is presented in Table III. In INEX 2003, assessors spent an average of 8 min for a relevant document and 1 min for an irrelevant one [Piwowarski and Lalmas 2004]. In 2005, the average assessment time was about 5 min for a document containing a highlighted passage (i.e., a relevant document). This drops to about 50 s for a nonrelevant document (i.e., a document that was part of the pool but contained no highlighted passages). The total overall assessment time for a single topic was about 11 h. In 2006, the average assessment time per topic was about 7 h, with an average of about 50 s per document. The average assessment time for a nonrelevant document was 44 s while it was about 1 min per relevant document.

Table III shows a downward trend in the mean time taken to assess a relevant document. The time to assess an irrelevant document has remained essentially constant regardless of the fact that the document collection changed between 2005 and 2006. It is reasonable to believe that the decrease for relevant documents is directly attributable to the changes made in the assessment procedures.

We can measure the time saved as follows. At INEX 2005, there was an average of 499 documents in each topic assessment pool. Of these, 60.7 were relevant. An estimate of how long it would have taken in 2005 if changes had not occurred can be made using the times from 2003. In 2003, it took, on average, 8 min to assess a relevant document and 1 min to assess a nonrelevant document. For relevant documents, it would have taken 60.7×8 min, and for the remainder (438.3 irrelevant documents per topic), 438.3×1 min. This sums to 15 h. Compared to the 2005 time of 11 h, this is a saving of almost 4 h/topic (36%). A similar comparison to the 2006 data is unreasonable because the document collection changed; in particular there was a substantial decrease in average document size in 2006.

It is reasonable to conclude that the adoption of highlighting is responsible for the decreased assessment load. This is especially true since, as a consequence

Table IV. The Average Number of Relevant Elements per Topic (For 2005, This Number Is Exclusive of Too Small Elements; for 2005 and 2006, the Number of Relevant Passages is Shown in Brackets.)

INEX	2002	2003	2004	2005	2006
Relevant elements	534	1190	739	522 (156)	1908 (80)

of highlighting, all pool documents are now fully assessed (whereas in previous years they were not necessarily so). Prior to the highlighting method, assessors were required to assess many XML elements for every relevant element they judged (such as child elements in the document tree). The complex enforcement rules that propagated the relevance (*EeSs*) scores often made this process time consuming, whereas with highlighting the relevance scores were often implicit in the highlighted passage.

Additionally, with the interface used prior to the highlighting method, assessors were not asked to assess every element of a document (although enforcement rules could make this the case) while with highlighting they were asked to highlight every relevant passage in a document. This resulted in *all* elements within a document being assessed (assuming the assessors was, indeed, assessing correctly) because those that were not highlighted were implicitly assessed as nonrelevant.

Table IV presents the average number of relevant elements seen per topic. At INEX 2002, there was an average of 534 assessed relevant elements per topic, while in INEX 2005, this number was 522. Since 2003, enforcement rules have been used to automatically infer some of these assessments. In INEX 2005, specificity was computed automatically and S-Rule 1 (GAIN) was used to infer exhaustivity in two cases: first, if an element was assessed as too small, then by definition all its descendants were too small; and second, if an element was assessed as highly exhaustive, then all its ancestors were also assessed as highly exhaustive. In 2006, all assessments were inferred because both specificity and exhaustivity were inferred from the highlighted text.

5.1.2 *Assessor Behavior.* In 2003 and 2004, assessors were forced to assess a single document, and within each document, they were forced to assess elements in whatever order the enforcement rules imposed. The highlighting method used in 2005 required a two-phase process as outlined in the detailed guidelines given to the assessors [Malik et al. 2005]: first highlight, then assign exhaustivity values to the highlighted elements. An analysis of the logs shows that, in general, the two-phase process was followed.

The assessors were asked to read the document and to highlight passages while doing so. The logs show that, 85% of the time, an assessor highlighted a passage located somewhere in the document after the previously highlighted passage—they were, indeed, highlighting while reading. They then switched to the exhaustivity assessment mode and gave values to relevant elements; they did not generally return to highlighting. On average, there were 1.13 highlight-then-assess cycles per assessed document.

When assessing exhaustivity, the judges also followed the natural document order. 95% of the time the assessor went further down the document when assessing exhaustivity. Among these 95%, in 35% of cases, the assessor judged a contained element (i.e., assessed first a section and then one of its paragraphs).

As a conclusion, the assessor behavior was close to what was expected when the interface was designed, which in turn was inspired by the way one reads and highlights a book. It appears as though assessors discriminated well the two phases (highlight then assess), which in turn indicates a good understanding of the two relevance dimensions. This has directly addressed limitation (2) in Section 1.1.

In 2006, the assessment task did not require the two-phase process as there was no manual assignment of scores after the highlighting process.

5.2 Assessment Characteristics

In this section, an analysis of the assessments is given along with a comparison between years. In particular the focus is on the effect of using highlighting against the prior method of individually assessing each element. In Section 5.2.1, completeness is discussed. An analysis of the different specificity scales is provided in Section 5.2.2. How the too small exhaustivity value was used is described in Section 5.2.3. The relationship between highlighted passages and XML elements is described in Section 5.2.4. Finally, an analysis of the consistency of the assessments is given by comparing the cross-judge agreement level (in Section 5.2.6) and passage agreement level (in Section 5.2.7).

5.2.1 *Completeness.* Pooling implies that only a subset of results returned by the participating retrieval systems will be assessed. It is thus crucial to ensure that the pool is large enough to capture the vast majority of relevant content. This is controlled by choosing a suitable value n, the controlling factor of the top-n pooling process (described in Section 4.1). Since INEX 2003, n has been the number of documents (not elements) in the pool; the process stops whenever the number of documents is above $n = 500$ at the end of a complete round of the pooling process. It is important to check the validity of the choice of 500 for n. As neither the document collection nor n changed between 2003 and 2005, any year could be selected for analysis; we chose 2005. As the document collection changed in 2006, the analysis is also included for that year.

Figure 2 is a plot of the cumulative number of highlighted passages against the number of documents in the pool seen at INEX 2005 and 2006. For each round of the top-n algorithm (until the pool is full), the number of documents and the number of highlighted passages within those documents were computed. We observe that the number of highlighted passages first increases rapidly and then increases linearly. For 2005 and 2006, the first 20 documents contained 48 and 23 passages, respectively, whereas the last 20 documents contained three new passages both in 2005 and 2006.

It is not straightforward to find a threshold of documents for pooling, since the curve continues to grow until 500 documents are reached (n). The increase is more or less constant after rank 250. Before that point, there are an average of 0.63 and 0.48 passages per document, respectively, in 2005 and 2006, but after that point the average is 0.14 and 0.20 passages, again in 2005 and 2006, respectively.

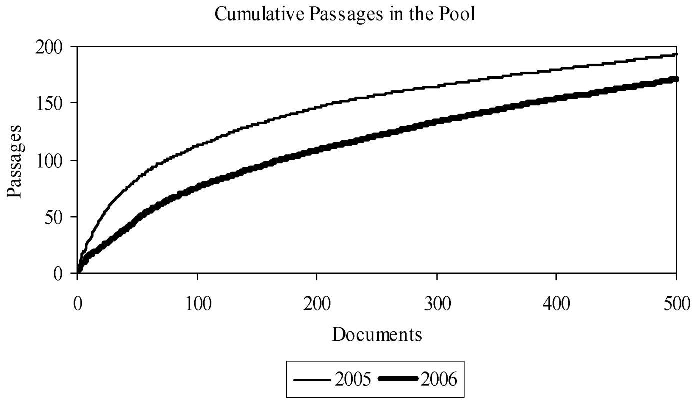


Fig. 2. Cumulative number of highlighted passages in function of the number of pool documents.

If the pools of INEX 2005 were computed with $n = 250$ documents (the pool is half the size), then 81% of the relevant elements would have been found (71% in 2006). However, as the time needed to assess a nonrelevant document is substantially less than the time to assess a relevant document (seconds vs. minutes), this is not a sufficient argument to decrease the threshold. It is, perhaps, an argument to increase the pool size to see how large the pool must be before the increase (in the number of highlighted passages) becomes negligible.

Zobel [1998] measured the effects of the pool size on relative system performance for document-centric retrieval at TREC. Using runs submitted to TREC, he first generated an assessment pool (the set of documents to assess). He then removed a run used for the pooling process and generated a new pool. A comparison of the excluded-run performance using both pools gives an estimate of the stability of the collection under the conditions of seeing a hitherto unseen run. A comparison of the two pools gives an indication of the expected number of nonassessed documents that will be seen with a new run.

To investigate the same effect at INEX, we first carried out an experiment using the INEX 2004 test collection. This test collection was selected because it contains both official runs, which were used in the pooling process (122 runs in total), and additional runs submitted latter by participants, which were not used in the pooling process (561 runs in total). We can then look at the percentage of elements from these additional runs that are in the pool and thus assessed, using the element pooling and document pooling strategies. To this end, in Figure 3, we plot the percentage of these elements that are in the pool (thus assessed) as a function of the pool depth (pooling documents (left) or elements (right)).

From the figure (left), it can be seen that for a pool depth of 500 documents, about 90% of the elements at rank 50 or above (in the additional runs) are in the pool and thus were judged. When pooling elements (as was done at INEX 2002), the result was catastrophic, as shown in the figure (right). A large proportion of the elements in the additional runs were not part of the pool using the element pooling strategy. This clearly shows that, although it may not be obvious to select an optimal pool depth threshold, pooling must be with respect

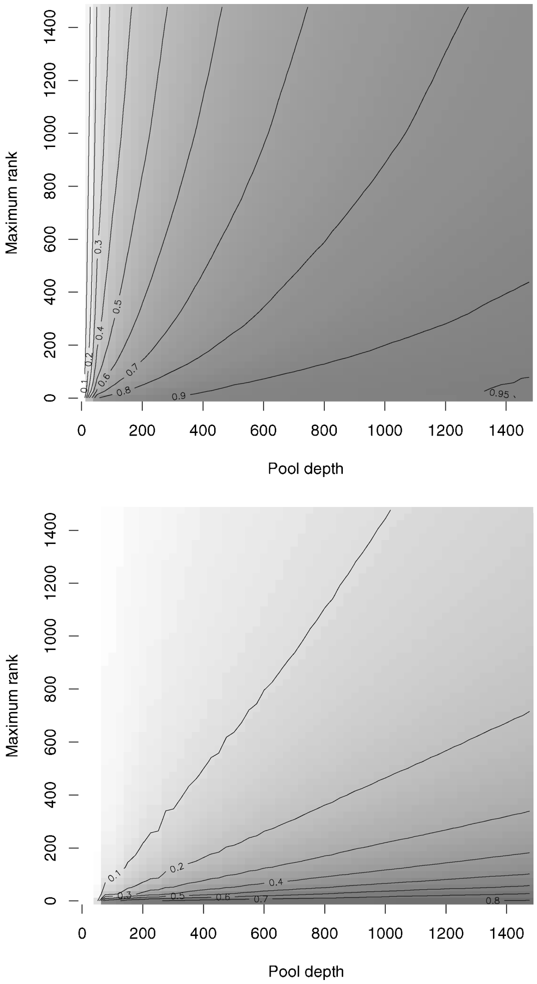


Fig. 3. For INEX 2004, the proportion of elements in the pool at a given pool depth (left: documents; right: elements) as a function of maximum run rank (for runs not used in the pooling process). For example, about 40% of the elements at rank 200 or below in a new run would be judged if the pool depth had been 450.

to documents and not elements. In addition, although enforcement rules dynamically add elements to the pool (as described in Section 4.2.2), using very deep pools still resulted in a low chance of these elements being part of the pool (and thus assessed), further reinforcing the above conclusion.

We also applied Zobel's [1998] technique to the INEX runs from 2003 to 2006. For each run used in the pooling process (from now on we only consider the document pooling strategy), we measured the performance against two different sets of assessments. The first set, referred to as set A, corresponds to the official assessments used for that year at INEX. To construct the second set, referred to as set B, all runs from a given participating organization were excluded from the pooling process and a new pool generated. It is important to remove all the

Table V. The Percent of Runs Showing a Significant Performance Difference as the Pool Size Increases (Rows Labeled *Same* Represent the Run Measured Against Itself, Whereas the Rows Labeled *Other* Report Percentage of Pairs of Runs Showing a Significant Order Change.)

		50	100	200	300	400	500
2002	Same	—	—	—	—	—	—
	Other	—	—	—	—	—	—
2003	Same	17.3%	38.5%	55.7%	33.7%	24.0%	12.5%
(120)	Other	9.0%	5.8%	2.9%	1.8%	0.5%	0.9%
2004	Same	32.0%	32.8%	32.8%	18.9%	17.2%	15.6%
(122)	Other	3.9%	2.7%	1.4%	0.6%	0.3%	0.3%
2005	Same	40.6%	37.7%	33.4%	32.8%	18.5%	2.3%
(325)	Other	7.2%	4.8%	2.0%	0.9%	0.3%	0.2%
2006	Same	39.3%	39.3%	38.7%	29.7%	18.5%	33.6%
(333)	Other	8.1%	6.1%	3.9%	2.8%	1.1%	0.6%

runs from a participant and not just a single run because participants typically use the same retrieval system to generate multiple runs and so their runs are often simply different permutations of the same elements. By excluding a single run in set B, the pools remain almost identical; by removing a participant we are simulating the arrival of a new participant and the consequences of the development of a new XML retrieval system.[2] Pool depths varying from 50 to 500 documents were used.

Retrieval performance was measured using a metric based on generalized precision recall [Kazai and Lalmas 2005], which was used at INEX to evaluate retrieval effectiveness. This metric rewards systems according to their ability to return specific and exhaustive elements. We computed the following:

- The percentage of runs that performed significantly differently between assessment sets A and B (using a paired t-test with confidence of 0.95). A sufficient pool depth should not lead to a significant change in the performance of a run.
- The percentage of run pairs whose order changed significantly when evaluated with sets A and B. Two paired t-tests were used to determine if a run performed significantly better, worse, or similar to another, one with set A and the other with set B. We then computed the percentage of change of the order between two runs. A sufficient pool depth should maintain a low probability of a switch in the order of the runs.

Table V presents the results for the different pool depths (50 to 500 documents). 2002 is not included as element pooling was used in that year.

Several conclusions can be drawn from the table. First, a pool depth of 500 is sufficient to ensure stable evaluation. When a pool depth above 300 is used, the probability that two runs are ordered (significantly) differently is small (generally below 1%, but up to 2.1%). As expected, this percentage drops by

[2] Note that we did consider run performances to be significantly different based on the result of the paired t-tests only, and did not consider very close scores as similar. Such distinction was made by Vorhees [2000]. This means that reported percentage of change are overestimations of the percentage one would get following the latter method. However, we believe that the same conclusions would be reached.

Table VI. Average Run Depth for Different Pool Depths for Each INEX Year

	Submission				Pool					
	Topics	Runs	Answers	Runs per Topic	50	100	200	300	400	500
2002	—	—	—	—	—	—	—	—	—	—
2003	66	120	3679	56	3.4	6.3	13.9	22.1	30.1	38.0
2004	75	122	4423	59	2.4	4.6	10.0	16.1	22.6	29.6
2005	87	325	13164	151	4.3	13.3	74.3	134.0	136.0	151.0
2006	125	333	41203	330	1.2	1.9	3.9	6.2	8.4	10.7

increasing the pool depth. A pool of 500 (as used at INEX) results in a chance below 1% of a significant change of the order of the runs.

With respect to single runs (the *same* rows in Table V), as the pool size increases (from 50 to 500), the proportion of runs showing a significantly different performance decreases. With a pool size of 500, the scores are generally low, except for 2006. 2006 was the year that the document collection changed from the IEEE to Wikipedia. The two collections are different; in particular, the Wikipedia collection contains many more documents with many small parts in which relevant material might be found. Moreover, systems were trained on the prior (IEEE) collection but evaluated on the new (Wikipedia) collection, which might have resulted in the general instability of the retrieval systems for that year.

Table VI supports this latter view. From 2003 through to 2006, the number of topics has increased from 66 to 125, while at the same time the number of runs submitted has increased from 120 to 333. However, not all runs provide answers to all topics. To show this, we show in column 4 (*answers* row) the total number of topics with ranked lists containing answers (relevant elements). For example, in 2004, we have 75 topics and 122 submitted runs, giving a total of 9150 ranked lists for all the topics, of which 4423 contained answers. We also show in column 5 (*runs per topic*) the mean number of runs that answered a topic. The change of collection in 2006 led to a jump in the proportion of all topics for which answers were provided by the runs (from 47% in 2005 to 99% in 2006). On average, over double the number of runs were used in pooling in 2006 than in 2005.

In Table VI, we also show the average rank (over participant runs) of the rank of the last document included in the pool (for different pool depths). Values can be found under the label *Pool*. For instance, in 2004, for a pool depth of $n = 300$, on average the top 16.1 documents of each participant run is included in pool. From the table it can be seen that in 2005 the top 151 results were taken from each run to form the 500-document pool; however, in 2006 only the top 10.7 were taken. That is, only the top 10 results from each run were added to the pool and assessed. This small number comes from the fact that we have 52 million elements in the Wikipedia collection, compared to 11 million in the IEEE (2005) collection, and is likely to have contributed to instability in the performance, which is reflected in Table V. Analysis on the 2007 test collection will be carried out to investigate this further.

Finally, we further analyze the impact of increasing the pool depth. Extrapolating from Figure 2, for the Wikipedia collection an average of five documents

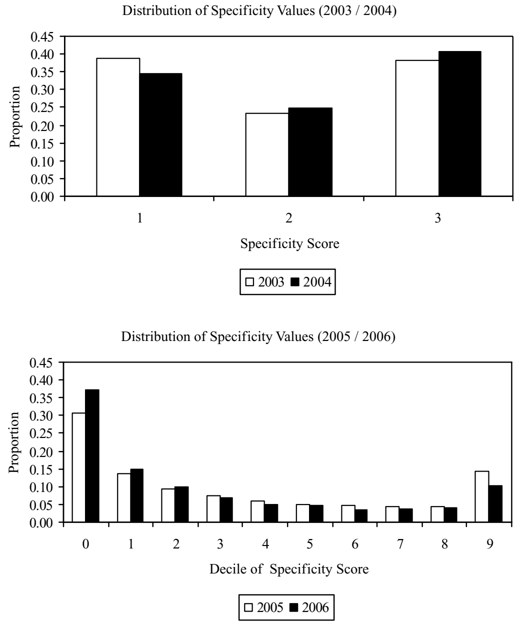


Fig. 4. Specificity distributions. Specificity was not a relevance dimension in 2002. In 2003 and 2004 it was quantized on a four-point scale (0 = not relevant). In 2005 and 2006, it was a continuous scale and is shown here in 10 equal sized ranges (specificity of 1 is removed).

would need to be assessed to find one new relevant passage. Increasing the pool from 500 to 600 documents, 80 irrelevant documents are expected (taking 48 s each to assess) and 20 new relevant documents are expected (taking 1 min each to assess). This is an additional 1$\frac{1}{2}$ h/topic (21%), and would increase the total average assessment time per topic to over 8 h.

In conclusion, it is reasonable to believe that 500 documents are sufficient to catch enough relevant passages (in both document collections) to allow meaningful comparisons of XML retrieval systems. However, our investigation for the 2006 pools supports prior suggestions of the introduction of new pooling techniques (see, e.g., Vu and Gallinari [2005] or Woodley and Geva [2005]). With a large number of participants and an increased potential for noise (highly ranked nonrelevant documents), it is important to find better ways of identifying potentially relevant documents for the pool.

5.2.2 *Specificity Distribution.* At INEX 2002, specificity was not a relevance dimension. At INEX 2003 and 2004, specificity was defined on a four-point scale. From INEX 2005, it has been defined as a continuous scale derived from the highlighted text. A change in the distribution of the specificity scores was expected with the change from manual (explicit) assessment to highlighting (implicit assessment). Plotted in Figure 4 is the distribution of specificity values obtained for the ordinal point scale (INEX 2003 and 2004) and the continuous scale (INEX 2005 and 2006). The distribution for the latter is shown distributed into 10 equal sized buckets ((0.0–0.1), etc.).

Table VII. Average Specificity Values for Four Major Categories of (Relevant) Elements (Between Parentheses, the Value is Divided by the Average Specificity for the Paragraphs (in Order to Give Comparable Values). Columns Are Categories as Multiple Tags Are Used to Represent the Same Concept in the Data (Section Represents Sections and Subsections, for Example). Note That Specificity Was Not a Relevance Dimension at INEX 2002.)

INEX	Article	Body	Section	Paragraph
2002	—	—	—	—
2003	0.56 (83%)	0.58 (86%)	0.66 (97%)	0.68 (100%)
2004	0.45 (63%)	0.44 (62%)	0.63 (88%)	0.71 (100%)
2005	0.12 (13%)	0.15 (16%)	0.51 (55%)	0.94 (100%)
2006	0.32 (33%)	0.32 (34%)	0.67 (71%)	0.94 (100%)

Most relevant elements have a specificity value of 1 in 2005 and 2006 (92% and 91%, respectively) and are not shown. This is because there are a large number of very small elements (footnotes, typographic elements, hypertext-links, etc.) that contain relevant information. These same elements were generally given an exhaustivity value of too small in 2005 and obscure the picture. By excluding these elements from the graph, it becomes clear that assessors used the full specificity scale. The change from a four-point scale to a continuous scale resulted in a clearer differentiation between levels of specificity—thus supporting the decision to use the continuous scale.

With the four-graded scale, all marginally specific elements were rewarded equally, regardless of how marginal they were. Although the grades allowed the assessor to differentiate a document from a paragraph, it did not allow assessors to differentiate between a section and a subsection. With a continuous scale, it is possible to distinguish many different levels of marginal specificity: it is now possible to distinguish between not only the document and the paragraph but also the section from the subsection. Most importantly, for XML retrieval evaluation, this enables the distinction between two systems returning, for example, the section versus the subsection, when the paragraph was the preferred answer. With the implicit (continuous) scale, it also becomes possible to check the soundness of the manual (four-point) judgments by comparing the scores for certain given elements.

Table VII presents the average specificity values for four[3] categories of elements. The discrete specificity values have been scaled to a continuous scale by dividing each value by 3. A drop in specificity for large elements (article and body) and to a lesser extent for medium sized elements (section) is observed. If we accept that specificity was more naturally computed using the continuous scales of INEX 2005 and 2006, this drop suggests that the specificity values for larger elements have been overevaluated in previous years (before highlighting was introduced) and also have been underevaluated for smaller elements such as paragraphs.

[3] INEX holds a notion of tag equivalence in which certain tags have the same meaning and are consequently considered equal. In the discussion and the table, article ∈ {article}, body ∈ {body, bdy}, section ∈ {section, sec, ss1, ss2}, paragraph ∈ {p, ip1, ip2, list}.

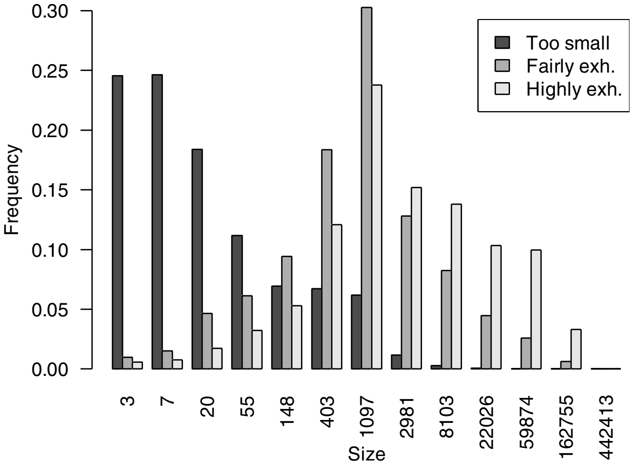


Fig. 5. Size (in characters) distribution with respect to the exhaustivity values (too small, partly, and highly). Average paragraph size is about 308 characters; average section size is about 5283. Results are only for 2005 as no other year included the too small assessment value.

Since the introduction of the continuous scale, paragraphs have had a near-1 specificity; they are either not relevant or totally specific (i.e., they either contain no highlighted text or are completely highlighted). They are also the largest elements for which the specificity is near 1. This suggests that the natural atomic level of granularity of an assessor is the paragraph. The consequence of this is not clear, but it does suggest that paragraphs are the smallest meaningful unit of retrieval in XML. Higher specificity values are seen for larger elements in 2006 because Wikipedia documents are generally smaller than IEEE articles.

5.2.3 *Exhaustivity and Size.* It is a reasonable criticism that assessors might mark elements they did not want to assess as too small—as a means to a fast end (especially due to relevance propagation from the enforcement rules). This problem could only happen at INEX 2005 as other INEX campaigns did not use the too small assessment value. This concern was checked by computing the average size of a too small element. About 79% of too small elements are shorter than 55 characters, suggesting the value too small was used mainly as intended. For comparison, the average size of a paragraph is 308 characters.

The distribution of element sizes with respect to the different exhaustivity values is presented in Figure 5. The exhaustivity level is correlated to the element size (assuming too small as the lowest). This is as expected, as having more text also potentially means having more space to discuss a topic, and therefore more space in which to thoroughly (exhaustively) cover a topic.

Too small was not an exhaustivity score in 2006. However, there remain elements in the document collection that are too small to be meaningful. These included varieties of hypertext links within Wikipedia. To overcome this shortfall, these hyperlinks were declared too small and not used for the formal ranking of information retrieval systems. Further investigation is needed to determine if it is possible to automatically infer that an element is too small. Such an investigation might take the mean length of a given element into consideration and (based on Figure 5) exclude those elements shorter than some given length (such as 55 characters). Such a method would catch most too small elements, but would also catch some elements that are not too small. An analysis on the change of relative performance of the retrieval systems would be necessary to validate this technique.

5.2.4 *Passages and Elements.* In 2005, the assessment method involved highlighting relevant passages then assigning an exhaustivity value to elements in the passage. The assessors might, however, simply highlight whole elements (as relevant) and then assign exhaustivity values to them. If this were the case then highlighting would not be a sound method of determining the specificity value of an element. We can determine whether assessors were highlighting relevant passages only by examining whether the passages are typically whole elements or not. We recall that the retrieval system identified elements and the assessors were aware of which elements they were.

We define an elemental passage as a passage that has exactly the same span as an XML element. That is, the passage starts on an element boundary b_s of element e_s and finishes on an element boundary b_f of element e_f where $e_s = e_f$.

In practice, computing whether a passage is elemental or not is not entirely trivial. A passage might, and many do, cross tag boundaries. Passages are not required to maintain the hierarchy of the document and can overlap tag boundaries. For example, given the XML fragment "<a><b>text</b></a>", the passage might be the segment "<b>text</b></a>". This is intuitively an elemental passage as the highlighted text is elemental even though the start tag "<a>" is not in the passage. Equally, for the XML "<a><b>text</b><b>text</b></a>", the passage "text</b><b>text" is elemental even though it starts in one element and ends in another.

The method used to determine if, or not, a passage is elemental was as follows:

—Compute the lowest common ancestor of the starting and finishing points of the passage.

—Compare the quantity of relevant text in the passage to that of the ancestor.

—If they are equal, the passage is elemental; otherwise the passage is not elemental.

If the assessors highlighted elements and marked those as relevant, it is reasonable to expect most passages to be elemental. If the assessor highlighted relevant passages then marked the relevance of the elements within those passages, it is reasonable to expect most passages will be nonelemental. The latter is expected, as it is reasonable to believe that relevance in a document is a function of the text and should not (necessarily) be bound by the structure of the document.

In Table VIII, the total number of passages along with the number that are elemental and nonelemental is presented for the 2005 and 2006 assessments (highlighting was not used in prior years). The proportion of elemental passages is 36% for 2005 and 45% for 2006. From this, it is reasonable to conclude

Table VIII. Breakdown of Elemental and Nonelemental Passages (Passage highlighting was not used prior to 2005.)

	2002	2003	2004	2005	2006
Elemental	—	—	—	2183	4093
Nonelemental	—	—	—	3918	4993
Total	—	—	—	6101	9086
% Elemental	—	—	—	36%	45%

that assessors are, indeed, highlighting relevant passages using the information content of the document. They are not just highlighting elements. The realization that assessors assessed passages led Trotman and Geva [2006] to suggest a passage retrieval task for INEX 2007 (this is further discussed later in this section).

As a consequence of assessing through highlighting, it is now possible to know how many relevant passages to expect in a document. As the assessment interface conglomerates overlapping and adjacent passages of text into a single passage (H-Rule 1 (ATOM)), such an analysis is not biased toward how the assessor highlighted the text. This information might be used for two purposes. First, if relevant information is only ever seen in a single passage within a document, then the assessment process can be simplified further, to marking the beginning and end of *the* relevant passage. Second, for relevance ranking purposes it is useful to know if, once a relevant part of a document has been identified, the remainder can be ignored.

Figure 6 shows the proportion of relevant documents that contain a given number of relevant passages for the years in which highlighting was used (2005 and 2006). In 2005, just over 50% of all relevant documents contained only one relevant passage. Nearly 90% of relevant documents contained five or fewer relevant passages. Outliers are seen, with one document containing 49 passages. In 2006, over 70% contained only one relevant passage with nearly 98% of documents containing five or fewer passages. The outlier is one document containing 74 passages. Wikipedia (2006) articles are typically smaller than the IEEE (2002–2005) articles, and so fewer relevant passages per document would be expected. The shape of the distributions, however, is similar.

5.2.4.1. *Further XML Retrieval Tasks.* In 2005 element boundaries were shown in the interface. This was removed in 2006, making it possible for the first time for assessors to assess without subtle hints as to where they should assess. This, in turn, made it possible to examine the relationship between relevant elements, passages, and documents. Figure 6 shows that there are in general only a small number of relevant passages in a relevant document. Most relevant documents contain more than one relevant passage, and those relevant passages are not, in general, elements. From this conclusion, it is appropriate to suggest a new retrieval task at INEX, one that matches the behavior we observe in the assessors. In this new INEX task, an XML retrieval system would identify a passage of relevant text (perhaps delineated by element boundaries) as the answer to a topic. Such a task was initially suggested by Clarke [2005], and has been investigated at INEX 2007 [Fuhr et al. 2007].

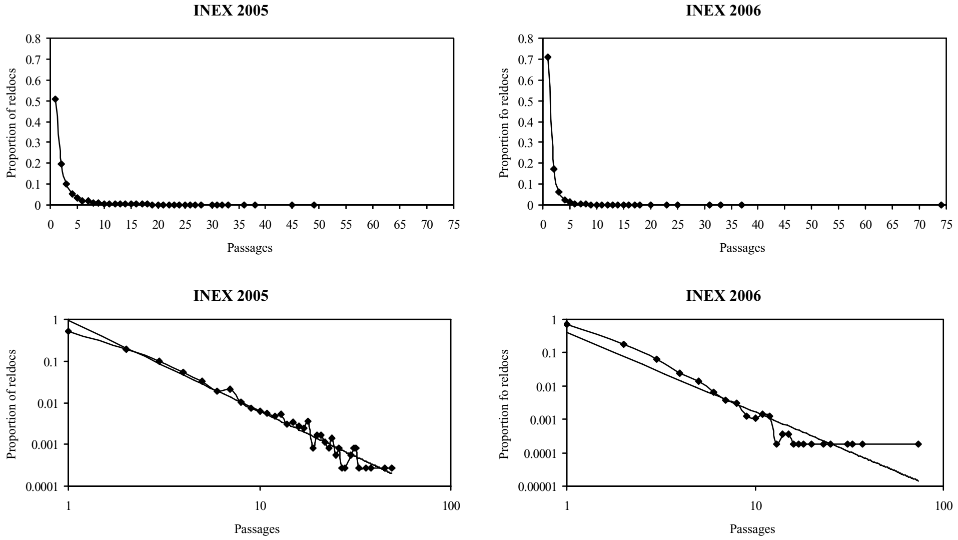


Fig. 6. Proportion of relevant documents containing the given number of relevant passages. Linear scale (top) and log-log scale (bottom). The distribution is approximately power law (shown). Highlighting was not used prior to 2005.

5.2.5 *Passages and Best Entry Points.* INEX 2006 introduced a task known as the *Best in Context task*, where the XML search engine must not only identify and rank relevant documents, but also identify the Best Entry Point (BEP) in the document from which the user should start reading.

In 2006, judges were asked to locate, within each relevant document, the location of one BEP. Section 5.2.4 examined the relationship between element boundaries and passages, and discussed the introduction of passage retrieval as an XML retrieval task. Studying the relationship between the BEP and passage positions is equally informative, since there may be a strong relationship between the two. If an unambiguous and deterministic method for finding the BEP, given the position of highlighted passages, can be found, then it would imply that the Best in Context task of INEX is not needed because BEPs can be found as a simple consequence of locating relevant passages. It also implies that users do not need BEPs as highlighting relevant content would suffice.

Kamps et al. [2007] analyzed the relationship between BEPs and relevant content. Their main findings are that BEPs are generally not located *exactly* at the beginning of a relevant passage (justifying the Best in Contest task) but do generally *coincide* with the first character of the first relevant passage in a document. In Figure 7 the distance from the best entry point to the first, biggest, and nearest passage is plotted. The results have been placed into buckets of 100 characters in size and the vertical axis is logarithmic. Our result confirms the results of Kamps et al. [2007].

This result suggests that the best simple deterministic strategy for BEP location in an element retrieval system is to select the start of the first relevant element. This strategy could, however, be enhanced by using the start of the first relevant passage (again suggesting that passage retrieval is an important

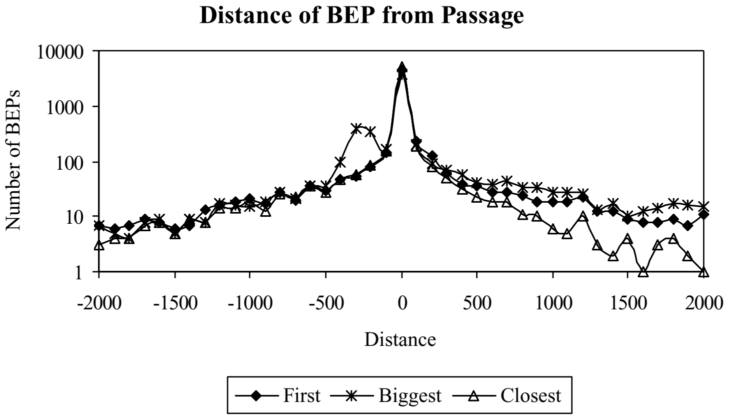


Fig. 7. Distance of best entry point to the first, biggest, and nearest relevant passage in the document.

task). This finding suggests that the Best in Context task may not be a separate task from passage retrieval, but a more thorough analysis of both passage retrieval and best entry points is needed before any firm conclusion can be drawn.

5.2.6 *Element Agreement Levels.* We expect the agreement levels between judges using highlighting to be much higher than those of judges individually assessing elements within a document. This was investigated by Pehcevski and Thom [2005] immediately after INEX 2005, and indeed this appears to be the case (only five topics were double judged so no statistically sound conclusion should be drawn).

Following the convention used at TREC, agreement levels at INEX are reported as the intersection divided by the union. For two judges, if the first judge identified two relevant elements and the second judge identified two relevant elements, but between them they only agreed on one element, then the agreement level is 0.33.

To compare the agreement levels at INEX to those of TREC, they must first be converted from multilevel element-centric judgments to binary document-centric judgments. Because relevance propagates up the XML document tree, if a document contains any relevant elements, the document is by definition also relevant. The conversion to document-centric judgments can be done by discarding all elements that are not the root element of a document.

The judgments must also be converted from the two-dimensional relevance scale of exhaustivity and specificity into a binary decision of relevant or not. This is done by considering all *E0S0* documents to be nonrelevant and all others to be relevant. Pehcevski and Thom [2005] found that, by using highlighting, the nonzero agreement levels for whole documents was 0.39 (compared to 0.27 from the previous year). This is in line with TREC, as shown in column 3 of Table IX, especially considering that the experiments reported by Voorhees [2000] computed the agreement levels using only a subset of the 200 documents judged by the first assessor as nonrelevant.

Table IX. Cross-Judge Agreement Levels at INEX and TREC (In 2006, Passage Highlighting Was Used So Element Agreement Levels Are not applicable.)

Evaluation Forum	Topics Double-Judged	Document Agreement Level	Exact Element Agreement Level
INEX 2002	2	—[a]	0.22
INEX 2003	0	—	—
INEX 2004	12	0.27	0.16
INEX 2005	5	0.39	0.24
INEX 2006	15	0.57	N/A
TREC 4		0.42—0.49	—
TREC 6		0.33	—

[a] The alternative assessments for 2002 are not available for analysis so the document agreement level cannot be computed. The value we report for exact element agreement level is from Piwowarski and Lalmas [2004].

Table X. Element Agreement Levels at INEX 2005 (Prior to 2005 Specificity Was Quantified on a Four-Point Scale and Post-2005 Exhaustivity Became Binary.)

Topic	∩/∪	Exh.	s < 0.1	s < 0.5	s < 0.9
209	0.12	0.08	0.09	0.11	0.12
217	0.18	0.16	0.17	0.18	0.18
234	0.49	0.15	0.46	0.48	0.49
237	0.13	0.09	0.10	0.13	0.13
261	0.3	0.29	0.30	0.3	0.3
Mean	0.24	0.15	0.22	0.24	0.24

To compare the exact agreement level to that of the previous year (0.16; see column 4 of Table IX), Pehcevski and Thom [2005] divided the continuous judgment scale of specificity into a three-point scale. Those elements that had specificity in (0.00,0.33] were assigned a specificity of 1, those in (0.33, 0.67] to 2, and those in (0.67, 1.00] to 3. As exhaustivity was already an integer, it was not necessary to further process the judgments. An agreement of 0.24 was reported. In 2006, only passages were assessed, so it is not appropriate to examine agreement levels for that year.

By introducing highlighting as a method of identifying relevant elements, the agreement levels between judges has increased considerably. This is for several reasons. First, the judges can more easily identify relevant passages than relevant elements. Second, assigning only a single relevance value (exhaustivity) is an easier task than assigning both a specificity value and an exhaustivity value. Third, the meaning of relevance is clearer on a one-dimensional scale than on a two-dimensional scale. In conclusion, the two-phase assessment method has overcome limitations (2) and (3) listed in Section 1.1.

Table X presents the element agreement levels for INEX 2005. Column 3 reports the ratio of relevant elements with the same exhaustivity level, while columns 4, 5, and 6 report the ratio of relevant elements with a difference in specificity (s) of less than m (i.e., $m < 0.1, m < 0.5$, and $m < 0.9$)

Within the double-assessed topics, the (implicit) agreement of specificity is always greater than the (explicit) agreement of exhaustivity. Assessors agree more on the highlighted passages than they do on the exhaustivity values.

Table XI. Passage Agreement Levels (Highlighting Was Not Used Prior to 2005.)

Topic	J1	J2	∩	∩/J1	∩/J2	∩/∪
INEX 2005						
209	540218	998147	213864	0.40	0.21	0.16
217	473400	614479	60387	0.13	0.10	0.06
234	163489	214266	127231	0.78	0.59	0.51
237	21412	108283	9366	0.44	0.09	0.08
261	219353	286842	125163	0.57	0.43	0.33
Mean						0.23
INEX 2006						
304	57908	55789	20377	0.35	0.37	0.22
310	133293	82421	77870	0.58	0.94	0.56
314	143759	30181	18397	0.13	0.61	0.12
319	89964	9045	5263	0.06	0.58	0.06
321	6693	14184	6216	0.93	0.44	0.42
327	18279	5736	3578	0.20	0.62	0.18
329	65687	53031	33480	0.51	0.63	0.39
355	27814	73721	20505	0.74	0.28	0.25
364	66150	43342	23144	0.35	0.53	0.27
385	32971	24044	10973	0.33	0.46	0.24
403	53557	40332	22362	0.42	0.55	0.31
404	154338	132758	74762	0.48	0.56	0.35
405	42610	17443	17256	0.40	0.99	0.40
406	47109	65216	13701	0.29	0.21	0.14
407	17970	15006	14587	0.81	0.97	0.79
Mean						0.31

This raises the question of the validity of the exhaustivity dimension. Ogilvie and Lalmas [2006] examined a binary scale (relevant or not) in place of the multipoint scale and determined that it was sound to do so. Exhaustivity became binary and implicit for INEX 2006.

Element agreement levels and passage agreement levels are inextricably related in INEX 2006 because the element relevance scores are derived from the highlighted passages; consequently they are not presented in this section but are described in the next section.

5.2.7 *Passage Agreement Levels.* Table XI shows the quantity of text (in characters) identified as relevant for each of the assessors (J1 and J2) and the quantity of text they agree is relevant. The final column presents the agreement level (intersection divided by union) for each topic. The two preceding columns are the ratio of the quantity of highlighted text of each judge to the common text they agree on.

Low levels of agreement can be observed in some topics for two reasons. First, and by example, in topic 217 the two judges simply did not agree on the location of the relevant text. Second, again by example, in topic 237 one of the judges identified a subset of the other's in as many as half the passages.

Although only five topics were double judged at INEX 2005, it is reassuring that the overall mean agreement level was reasonably high. The double judging was again performed at INEX 2006, but on a larger number of topics. In this case for the 15 topics that were double judged, the agreement level was 0.31, higher than the 0.23 of the previous year. Although the samples were small and caution must be taken when drawing conclusions from them, there are several possible reasons for the improvement. First, the assessment procedure was one-dimensional and so had a lower impact on the assessment task. Second, the document collection changed from the (technical) IEEE collection to the (general) Wikipedia collection, which might have made decision making easier for the assessor.

5.3 Conclusions

The assessments were examined in this section and it was shown that, although they are not complete, a pool of 500 documents contains enough of the relevant documents, and more importantly, the size of the pool is sufficient to provide meaningful comparisons of XML retrieval systems. The change to assessing with highlighting ensured that each document that was assessed was assessed in full. The shift to highlighting also resulted in finer-grained specificity scores, allowing finer-grained measurement of a retrieval run than before. Progressive changes over several years have led to increased assessor agreement levels, and consequently increased soundness of the assessments.

The assessors themselves were shown to be assessing passages (and not just elements) in the natural reading order of a document, and to assign too small values only to elements that were, indeed, too small to be meaningful on their own. This has, no doubt, decreased the cognitive load on the assessor, which in turn has resulted in a decreased time to assess and increased the soundness of the assessments.

6. DISCUSSION AND FUTURE WORK

Evaluating information retrieval effectiveness is mostly done through the Cranfield methodology [Cleverdon et al. 1966]. One component of the methodology is the gathering of appropriate (sound and complete) relevance assessments, which state which information objects (documents, elements, or passages) are relevant to which topics. The topic, in turn, is the definitive statement of the information need. For content-oriented XML retrieval, the process of gathering assessments is more complex than it is for document retrieval because the relationship between information objects (XML elements) must be taken into account.

In this article, we described the methodologies used at INEX (from 2002 through 2006) for gathering sound and complete relevance assessments. By *sound* we mean reliable. By *complete* we mean that a high enough proportion of relevant elements have been identified to perform meaningful comparison of systems. In XML retrieval, gathering sound and complete assessments is especially difficult because the relevance assessments of two elements within the same document are often inextricably related. If a section in a document is relevant then the document is also relevant, but just because the document is relevant does not mean that all sections in that document are.

Although INEX topics are (as often as possible) assessed by their authors, the assessors are not paid for the task. The assessors are the INEX participants themselves, who are otherwise expected to carry out their ordinary duties. Obtaining sound and complete assessments with minimum impact on the assessors is thus important. Getting reliable assessments in a short period of time is a necessary goal for INEX.

We described the different relevance scales used by INEX throughout the years. One reason for the changes was the goal of creating an easier assessment process—such as the introduction of inferred assessment and the too small elements. Another was the need to increase the soundness of assessments, that is, to increase assessor agreement levels to those similar to known robust evaluation forums such as TREC. We found that the introduction of binary exhaustivity had a substantial effect on soundness. Experiments reported by Ogilvie and Lalmas [2006] already showed that the impact of binary exhaustivity on the evaluation of systems is negligible, since it did not substantially change the relative order of systems submitted to INEX. The impact on the agreement levels between assessors did change as expected, primarily because assessors can more easily agree on binary relevance than on multidimensional graded relevance. While it might seem important to have highly expressive multilevel judgments, the disagreement between different assessors led to questions of the validity of the assessments and hence the measure of a system performance.

The aim of easing the assessment procedure also led to a more appropriate specificity scale. From four levels in 2002, the scale became continuous (between 0 and 1) in 2006. Even though there are more possible values in the continuous scale, this does not adversely impact agreement levels because the relevance score is computed automatically from highlighted relevant passages. Because it is computed from passages, it can be viewed as more objective than arbitrary assessment values chosen by the assessor. We also show that the specificity of different categories of elements appears to be more realistic since the introduction of highlighting. We cannot generalize this to the exhaustivity dimension, as there is yet no obvious way to measure this effect.

Although not an original goal, we have shown that a good way to increase the soundness of the assessments is to design an assessment process that results in a low cognitive effort for the user. The harder the assessor is forced to think—about the relevance of an element—the worse the agreement levels. Even the enforcement rules aimed specifically at increasing soundness in 2003 and 2004 were not able to raise the agreement to those seen in 2005 and 2006 where the less cognitively intensive method of highlighting was used to assess relevance.

While the soundness of the assessments increased from year to year, the time to assess a document decreased. Assessments in 2002 took a considerable period of time and the cognitive load was very high, both presumably because assessors were forced to manage two relevance dimensions and two views of the same document, and to perform manual assessments. In 2003 and 2004, the two views of the XML document were merged and the assessment process involved clicking on an icon—much less work. Although we have no firm data on the improvement from 2002 to 2003, anecdotes from participants suggest the improvement was considerable. Finally, in 2005 and 2006 the highlighted procedure measurably decreased the assessment time. We believe that this decrease also had an impact on the agreement levels (soundness), since the judges could spend more time focusing on the document and less on the assessment procedure. An important conclusion from this is that the assessment time should be kept as close as possible to the time a user might spend in making the same decision.

Before 2005 and 2006, the time needed for assessment was prohibitive—and there was no possibility of increasing the pool size. In fact, others examined the possibility of reducing the pool size in an effort to reduce the assessment time. Since 2006, the time needed to assess a single topic has dropped to a sufficiently low level that it could be viable and interesting to examine the effect of increasing the pool size. Of particular interest is the question of the number of relevant passages—whether this will stabilize at some sufficiently small point or if it will continue to increase even after a large number of documents have been assessed. Of course the effect of increasing the pool size on relative system performance is important too.

In order to ensure completeness, we found that the granularity of the units in the pool should be well suited to the task at hand. At INEX, an assessor can often judge more than one element at a time if the element belongs to the same context (for example, the document). We found that a good granularity for the assessment of relevance in XML retrieval is the document, and not the element. In the context of Web search, this result suggests that it might be easier for an assessor to judge a set of related pages (perhaps a single site) than a single Web page. Pool depth, however, remains important. We used Zobel's [1998] methodology to validate the pool depth of 500 used at INEX. Our analysis shows that the variability of runs must be considered before the pool algorithm and depth are chosen, especially when runs are very different (e.g., when new collections or tasks are introduced). This reraises the possibility of using alternative pooling strategies to the top-*n* used at INEX since 2003.

Finally, we examined the assessments in the light of two new INEX tasks: the passage retrieval task, and the Best in Context task. We first provided some insight about the benefit of using highlighting by showing that highlighted passages do not generally correspond to whole XML elements. This led to a new XML retrieval task of identifying relevant passages from the content and structure within a document. We believe this task is well suited to how a user might use an XML retrieval system, for example, reading the abstract and conclusions of an article before making a decision about the benefit of reading the whole article. We analyzed how Best Entry Points are related to highlighted passages and raised the question of the validity of the Best in Contest task as a separate task from passage retrieval because of the enormous quantity of Best Entry Points being so close to the beginning of the first relevant passage. This is an important question for assessment as we have shown that any additional cognitive load on the assessor is likely to have a negative effect on the soundness of the assessments.

INEX started in 2002 and, at the time of writing, is now in its seventh year. There have been substantial changes to the methodology from year to year, but we believe it is not likely to change much in future years. We have shown that the assessments are sound and are nearly complete. In future work, we will increase the pool size in an effort to measure exactly how complete the assessments are and whether more complete assessments will affect relative system performance. Previous pooling experiments at INEX have concentrated on shallow pooling due to the (until recently) effort needed to assess. Now it is possible to increase pool sizes and to see the effect.

To recap, the lessons learned are as follows:

—*With respect to soundness*, we have shown that the cognitive processes of judging relevance and of using the assessment interface should interfere as little as possible with the task at hand. This implies that an online assessment interface should not be invasive. In addition, the number of possible judgments for each element should be as few as possible (ideally, binary) or should be implicitly collected. Multilevel judgments may be beneficial in information retrieval evaluation, but eliciting them may—and often will—impact on the reliability of the assessments.

—*With respect to completeness*, we have shown that, although we may be measuring how effective XML retrieval systems are at returning relevant elements, we do not necessarily need to assess at the element level. This means that the granularity of the judged unit should match the assessment task and not necessarily the evaluation task. We also have shown that assessors can often judge more than one element concurrently. For example, if a document is not relevant then the elements within the document cannot be relevant. In addition, we found that highlighting text within a whole document and determining the relevance of an element from this was better than individually assessing all elements. Finally, the pooling process should be chosen with respect to the variability of runs, especially when the number of runs used for pooling is high. High run variability in conjunction with a large number of runs can result in pools drawn from only very highly ranked documents. We suggest that alternative pooling strategies should be used in these cases.

To conclude, we are confident that after 5 years of experimentation we have found the appropriate way of gathering assessments that produces both sound and complete enough assessments for XML retrieval. As INEX examines a greater range of focused retrieval questions (including passage retrieval, question answering, and element retrieval), we will be interested in how this methodology will have to change. We expect that a single common method of gathering assessments for all Cranfield-based experiments can be found, thus unifying the diverse set of methods currently seen.

ACKNOWLEDGMENTS

The INEX organizers would like to thank the participants in INEX for their assessments, feedback, and discussions. We would also like to thank the anonymous reviewers for their excellent comments.

REFERENCES

Baeza-Yates, R., Fuhr, N., and Maarek, Y. S. (Eds.). 2002. *Proceedings of the ACM SIGIR Workshop on XML*.

Blanken, H. M., Grabs, T., Schek, H.-J., Schenkel, R., and Weikum, G. (Eds.). 2003. *Intelligent Search on XML Data, Applications, Languages, Models, Implementations, and Benchmarks*. Lecture Notes in Computer Science, vol. 2818. Springer, Berlin, Germany.

Carmel, D., Maarek, Y., and Soffer, A. (Eds.). 2000. *Proceedings of the ACM SIGIR Workshop on XML*.

Clarke, C. 2005. Range results in XML retrieval. In *Proceedings of the INEX Workshop on Element Retrieval Methodology*, 2nd ed. 4–5.

Cleverdon, C. W., Mills, J., and Keen, M. 1966. *Factors determining the performance of indexing systems*: Cranfield University, Cranfield, Bedforshine, U.K.

Cormack, G. V., Palmer, C. R., To, S. S. L., and Clarke, C. L. A. 1997. Passage-based refinement (multitext experiements for TREC-6). In *Proceedings of the 6th Text REtrieval Conference* (TREC-6), 171–186.

Denoyer, L. and Gallinari, P. 2006. The Wikipedia XML corpus. In *Proceedings of the INEX Workshop*.

Fuhr, N., Kamps, J., Lalmas, M., Malik, S., and Trotman, A. 2007. Overview of the INEX 2007 ad hoc track. In *Proceedings of INEX*. 1–22.

Harman, D. 1992. Overview of the first text retrieval conference (TREC-1). In *Proceedings of the 1st Text REtrieval Conference* (TREC-1).

Harman, D. 1995. The TREC conferences. In *Proceedings of the HIM International Conference*.

Kamps, J., Koolen, M., and Lalmas, M. 2007. Where to start reading a textual XML document? In *Proceedings of the 30th ACM SIGIR Conference on Information Retrieval*.

Kazai, G. and Lalmas, M. 2005. INEX 2005 evaluation metrics. In *Proceedings of INEX*.

Kazai, G., Masood, S., and Lalmas, M. 2004. A study of the assessment of relevance for the INEX'02 test collection. In *Proceedings of the 26th European Colloquium on Information Retrieval Research* (ECIR 2004).

Kekäläinen, J., and Järvelin, K. 2002. Using graded relevance assessments in IR evaluation. *J. Amer. Soc. Inform. Sci. Tech. 53*, 13, 1120–1129.

Lalmas, M. and Tombros, A. 2007. Evaluating XML retrieval effectiveness at INEX. *SIGIR For. 41*, 1, 40–57.

Luk, R. W. P., Leong, H. V., Dillon, T. S., Chan, A. T. S., Croft, W. B., and Allan, J. 2002. A survey in indexing and searching XML documents. *J. Amer. Soc. Inform. Sci. Tech. 53*, 6, 415–437.

Malik, S., Kazai, G., Lalmas, M., and Fuhr, N. 2005. Overview of INEX 2005. In *Proceedings of the INEX Workshop*. 1–15.

Ogilvie, P. and Lalmas, M. 2006. Investigating the exhaustivity dimension in content oriented XML element retrieval evaluation. In *Proceedings of the 15th ACM Conference on Information and Knowledge Management* (CIKM 2006).

Pehcevski, J. and Thom, J. A. 2005. HiXEval: Highlighting XML retrieval evaluation. In *Proceedings of the INEX Workshop*.

Pehcevski, J., Thom, J. A. and Vercoustre, A.-M. 2005. Users and assessors in the context of INEX: Are relevance dimensions relevant? In *Proceedings of the INEX Workshop on Element Retrieval Methodology*, 2nd ed. 47–62.

Piwowarski, B. and Lalmas, M. 2004. Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. In *Proceedings of the 13th ACM Conference on Information and Knowledge Management*. 361–370.

Saracevic, T. 1991. Individual differences in organizing, searching and retrieving information. *Proc. Amer. Soc. Inform. Sci. 28*, 82–86.

Tombros, A., Larsen, B., and Malik, S. 2004. The interactive track at INEX 2004. In *Proceedings of the INEX Workshop*. 410–423.

Trotman, A. 2005. Wanted: Element retrieval users. In *Proceedings of the INEX Workshop on Element Retrieval Methodology*, 2nd ed. 63–69.

Trotman, A. and Geva, S. 2006. Passage retrieval and other XML-retrieval tasks. In *Proceedings of the SIGIR Workshop on XML Element Retrieval Methodology*. 43–50.

VOORHEES, E. M. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inform. Process. Manage. 36*, 5, 697–716.

VU, H.-T. AND GALLINARI, P. 2005. Using rankboost to compare retrieval systems. In *Proceedings of the 14th Conference on Information and Knowledge Management* (CIKM'05).

WOODLEY, A. AND GEVA, S. 2005. Fine tuning INEX. In *Proceedings of the INEX Workshop on Element Retrieval Methodology*, 2nd ed. 70–79.

ZOBEL, J. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st ACM SIGIR Conference on Information Retrieval*. 307–314.

Received June 2007; revised February 2008; accepted March 2008