Inf Retrieval (2006) 9:699–722
DOI 10.1007/s10791-006-9008-2

Evaluating the effectiveness of content-oriented XML retrieval methods

Norbert Gövert · Norbert Fuhr · Mounia Lalmas · Gabriella Kazai

Received: 15 January 2005 / Accepted: 27 February 2006 / Published online: 1 September 2006

Abstract Content-oriented XML retrieval approaches aim at a more focused retrieval strategy: Instead of retrieving whole documents, document components that are exhaustive to the information need while at the same time being as specific as possible should be retrieved. In this article, we show that the evaluation methods developed for standard retrieval must be modified in order to deal with the structure of XML documents. More precisely, the size and overlap of document components must be taken into account. For this purpose, we propose a new effectiveness metric based on the definition of a concept space defined upon the notions of exhaustiveness and specificity of a search result. We compare the results of this new metric by the results obtained with the official metric used in INEX, the evaluation initiative for content-oriented XML retrieval.

Keywords XML retrieval · Evaluation · Effectiveness · Metrics · Exhaustiveness and specificity

Computing Classification System H.3.3 Information Search and Retrieval, H.3.4 Systems and Software: Performance evaluation

N. Gövert (✉)
University of Dortmund, Germany
e-mail: norbert.goevert@uni-dortmund.de

N. Fuhr
University of Duisburg-Essen, Germany
e-mail: fuhr@uni-duisburg.de

M. Lalmas · G. Kazai
Queen Mary University of London
e-mail: mounia@dcs.qmul.ac.uk

G. Kazai
e-mail: gabs@dcs.qmul.ac.uk

1. Introduction

The *eXtensible Markup Language* (XML) is acknowledged as a standard document format for full-text documents. In contrast to HTML, which is mainly layout-oriented, XML follows the fundamental concept of separating the logical structure of a document from its layout. A major purpose of XML markup is the explicit representation of the logical structure of a document, whereas the layout of documents is described in separate style sheets.

From a content-oriented information retrieval (IR) point of view, users should benefit from the structural information inherent in XML documents. Given a typical IR style information need, where no constraints are formulated with respect to the structure of the documents and the retrieval result, XML retrieval systems aim to implement a more focused retrieval paradigm. That is, instead of retrieving whole documents, these systems aim at retrieving *document components* that fulfil the user's information need.

This raises the question of which document components, from a tree of related components, would best satisfy the user's information need. There is not yet a definitive answer to this question in the context of XML retrieval. The traditional IR view focuses on the retrieval of complete documents, and relies on the user's ability to locate the relevant content within a returned document. In our approach, and in that adopted by the INEX initiative (more about this later), we follow the view proposed in the FERMI multimedia information retrieval model: Given a user's information need, the best components to retrieve should be the deepest components in the document structure, i.e., most specific, while remaining exhaustive to the information need (Chiaramella et al., 1996). By following this approach the user is presented more specific material, and thus the effort to view it decreases.

In recent years, an increasing number of systems have been built which implement content-oriented XML retrieval in this way (Baeza-Yates et al., 2000, 2002; Fuhr et al., 2003, 2004a). The advent of such systems necessitated the development of a new infrastructure for the evaluation of content-oriented XML retrieval approaches. Traditional IR test collections, such as provided by TREC (Voorhees and Harman, 2002) and CLEF (Peters et al., 2002) are not suitable for the evaluation of content-oriented XML retrieval as they treat documents as atomic units. They do not consider the structural information in the collection, and they base their evaluation on relevance assessments provided at the document level only.

In March 2002, the *IN*itiative for the *E*valuation of *X*ML retrieval (INEX[1]) (Fuhr et al., 2003) started to address these issues. The aim of the INEX initiative is to establish an infrastructure and to provide means, in the form of a large test collection and appropriate scoring methods, for the evaluation of the effectiveness of content-oriented retrieval of XML documents. Following the "best component" view mentioned above, corresponding evaluation criteria have been defined, along with an appropriate scaling. These evaluation criteria consider retrieval at the document components level. Based on the criteria and their scaling, a metric based on traditional recall/precision metrics has been developed that facilitates statements about the effectiveness of algorithms developed for content-oriented XML retrieval.

A major limitation however arises with the metric, which has been adopted as the official metric in INEX. Returning many overlapping components (e.g., a component and its parent component) tends to lead to higher overall effectiveness performance than when adopting a more selective strategy, one which returns only the best components. In addition, XML components vary in size, which has an impact on user effort; viewing a large relevant document component is different to viewing a small one. Not considering size and overlap goes against

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

one of the main goals of XML retrieval systems, which is to provide a more focused retrieval. In this article, we develop a new metric for content-oriented XML retrieval that overcomes these shortcomings.

The article is organised as follows. In Section 2, we examine the assumptions underlying traditional IR evaluation initiatives and highlight their invalidity when evaluating content-oriented XML retrieval. Section 3 details the evaluation criteria and measures for content-oriented XML retrieval. Based on these criteria and the arguments given in Section 2, we develop a new metric for evaluating the effectiveness of content-oriented XML retrieval (Section 4). In Section 5 we give an overview on the INEX test collection. Section 6 provides the results of the new metric applied to the INEX 2002 and INEX 2003 runs and compares them to the results obtained with the official metric. We close in Section 7 with conclusions and an outlook on further issues with regard to the evaluation of content-oriented XML retrieval.

2. Information retrieval evaluation considerations

Evaluation initiatives such as TREC[2], NTCIR[3], and CLEF[4] are based on a number of restrictions and assumptions that are often implicit. However, when starting an evaluation initiative for a new type of task, these restrictions and assumptions must be reconsidered. In this section, we first pinpoint some of these restrictions, and then discuss the implicit assumptions.

Approaches for the evaluation of IR systems can be classified into system and user-centred evaluations. These have been further divided into six levels (Cleverdon et al., 1966; Saracevic, 1995): engineering level (efficiency, e.g., time lag), input level (e.g., coverage), processing level (effectiveness, e.g., precision, recall), output level (presentation), user level (e.g., user effort) and social level (impact). Most work in IR evaluation has been on system-centred evaluations and, in particular, at the processing level, where no real users are involved with the systems to be evaluated (e.g., most of the TREC tracks fall into this category—in contrast to the user-oriented evaluation of the TREC interactive track(Beaulieu and Robertson, 1996) and Web track (Craswell and Hawking, 2004)). The aim of the processing level evaluation efforts is to assess an IR system's retrieval effectiveness, i.e., its ability to retrieve relevant documents while avoiding non-relevant ones.

Following the Cranfield model (Cleverdon et al., 1966), the standard method to evaluate retrieval effectiveness is by using test collections assembled specifically for this purpose. A test collection usually consists of a document collection, a set of user requests (the so-called topics) and relevance assessments. There have been several large-scale evaluation projects, which resulted in well established IR test collections (Salton, 1971; Jones and van Rijsbergen, 1976; Voorhees and Harman, 2002; Peters et al., 2002; Kando and Adachi, 2004). These test collections focus mainly on the evaluation of traditional IR systems, which treat documents as atomic units. This traditional notion of a document leads to a set of implicit assumptions, which are rarely questioned:

1. Documents are independent units, i.e., the relevance of a document is independent of the relevance of any other document. Although this assumption has been questioned from

[2] http://trec.nist.gov/

[3] http://research.nii.ac.jp/ntcir/index-en.html

[4] http://www.clef-campaign.org/

time to time, it is a reasonable approximation. Also most retrieval models are based on this assumption.
2. A document is a well-distinguishable (separate) unit. Although there is a broad range of applications where this assumption holds (e.g., collections of newspaper articles), there is also a number of cases where this is not true, e.g., for full-text documents such as books, where one would like to consider also portions of the complete document as meaningful units, or in the Web, where often large documents are split into separate Web pages.
3. Documents are units of (approximately) equal size (or at least in the same order of magnitude). When computing precision at certain ranks, it is implicitly assumed that a user spends a constant time per document. Based on the implicit definition of effectiveness as the ratio of output quality vs. user effort, quality is measured for a fixed amount of effort in this case.[5]

In addition to these document-related assumptions, the standard evaluation measures assume a typical user behaviour:

4. Given a ranked output list, users look at one document after the other from this list, and then stop at an arbitrary point. Thus, non-linear forms of output (like e.g., in Google) are not considered.

For content-oriented XML document retrieval, most of these assumptions are not valid, and have to be revised:

1. Since we allow for document components to be retrieved, multiple components from the same document can hardly be viewed as independent units.
2. When allowing for retrieval of arbitrary document components, we must consider overlap of components; e.g., retrieving a complete section (consisting of several paragraphs) as one component and then a paragraph within the section as a second component. This means that retrieved components cannot always be regarded as separate units.
3. The size of the retrieved components should be considered, especially due to the task definition; e.g., retrieve minimum or maximum units answering the query, retrieving a component from which we can access (browse to) a maximum number of units answering the query, etc.
4. When multiple components from the same document are retrieved, a linear ordering of the result items may not be appropriate (i.e., components from the same document are interspersed with components of other documents). Single components typically are not completely independent from their context (i.e., the document they belong to). Thus, frequent context switches would confuse the user in an unnecessary way. It would therefore be more appropriate to cluster together the result components from the same document.

In this article, we are concerned with issues two and three, that is, component size and component overlap, which we view to be the most crucial for the evaluation of content-oriented XML retrieval.[6] In order to deal with component size and component overlap, we develop new evaluation criteria and a new metric (Sections 3 and 4).

[5] For example, the original TREC collection contains both newspaper articles (of the size of one or more kB) and a number of Federal Register documents (up to a few MB large) (Harman, 1993); treating both kinds of documents equally in evaluation is not appropriate from our point of view.

[6] We make no explicit assumptions about users here, due to the fact that little is known about user behaviour when searching XML documents. However, the ongoing INEX interactive track is addressing this issue (Tombros et al., 2005).

3. Relevance dimensions for content-oriented XML retrieval

In order to setup an evaluation initiative we must specify the objective of the evaluation (e.g., what to evaluate), select suitable criteria, set up measures and measuring instruments (e.g., framework and procedures) (Saracevic, 1995). In traditional IR evaluations (at the processing level) the objective is to assess the retrieval effectiveness of IR systems, the criterion is relevance, the measures are recall and precision and the measuring instruments are relevance judgements.

In XML IR evaluation, the objective remains the measurement of a system's retrieval effectiveness. However, unlike in traditional IR, the effectiveness of an XML search system will depend on both the content and structural aspects. As pointed out in Section 2, the evaluation criteria and measures rely on implicit assumptions about the documents (and users), which do not hold for content-oriented XML retrieval. It is therefore necessary to reformulate the evaluation criteria and to develop new evaluation procedures to address the additional requirements introduced by the structure of the XML documents and the implications of such a structure.

3.1. Topical exhaustiveness and component specificity

The combination of content and structural requirements within the definition of retrieval effectiveness must be reflected in the evaluation criteria to be used. The new evaluation criteria stem from the fact that XML elements[7] forming a document can be nested. Since retrieved elements can be at any level of granularity, an element and one of its child elements can both be relevant to a given query, but the child element may be more focused on the topic of the query than its parent element (which may contain additional irrelevant content). In this case, the child element is a better element to retrieve than its parent element, because not only it is relevant to the query, but it is also specific to the query.

The above relates to earlier work on hypermedia document retrieval (Chiaramella et al., 1996), which showed that the relevance of a structured document can be better described by two logical implications. The first one, $d \rightarrow q$ (the document *implies* the query), is the *exhaustiveness* of document d for the query q, and models the extent to which the document discusses all the aspects of the query. The second one, $q \rightarrow d$ (the query *implies* the document), is the *specificity* of the document d for the query q, and models to what extent all the aspects of the documents concern the query.[8] Therefore a document d can be exhaustive but not specific to a query, and vice versa. In the context of XML retrieval, some XML elements will be exhaustive but not specific to a given query; for example large document components may contain extensive relevant content and the same time may include large sections of irrelevant content. Other elements will be specific to a query, but not exhaustive; for example small components are likely to contain information that is less extensive but more focused on a single topic.

Based on the above, INEX adopted the following two criteria to express relevance:

[7] In this article, the terms elements and components are used interchangeably.

[8] Readers familiar with the classical IR literature will note that the terms 'exhaustiveness' and 'specificity' originally were introduced in the context of document indexing, where they referred to properties of the set of indexing terms assigned to a document (Lancaster, 1968); in contrast, we are regarding properties of a document (component) with respect to a query here.

Topical exhaustiveness reflects the extent to which the information contained in a document component satisfies the information need.

Component specificity reflects the extent to which a document component focuses on the information need.

Relevance is thus defined according to the two dimensions of exhaustiveness and specificity. Topical exhaustiveness here refers to the standard relevance criterion used in IR.[9] This choice is reasonable, despite the debates regarding the notion of relevance (Saracevic, 1996; Cosijn and Ingwersen, 2000)), as the stability of relevance-based measures for the comparative evaluation of retrieval performance has been verified in IR research (Voorhees, 1998; Zobel, 1998).

When considering the use of the above two criteria for the evaluation of XML retrieval systems, we must also decide about the scales of measurements to be used. For the traditional notion of relevance, binary or multiple degree scales are known. Apart from the various advantages highlighted in Kekäläinen and Järvelin (2002), we believe that the use of a nonbinary exhaustiveness scale is also better suited for content-oriented XML retrieval evaluation: It allows the explicit representation of how exhaustively a topic is discussed within a document component with respect to its sub-components. Based on this notion of exhaustiveness, a section containing two paragraphs, for example, may then be regarded more relevant than either of its paragraphs by themselves. This difference cannot be reflected when using a binary scale for exhaustiveness. In INEX, we therefore adopted the following four-point ordinal scale for exhaustiveness (Kekäläinen and Järvelin, 2002):

Not exhaustive (0): The document component does not contain any information about the topic of request.

Marginally exhaustive (1): The document component mentions the topic of request, but only in passing.

Fairly exhaustive (2): The document component discusses many aspects which are relevant with respect to the topic description, but this information is not exhaustive. In the case of multi-faceted topics, only some of the sub-themes or viewpoints are discussed.

Highly exhaustive (3): The document component discusses most or all aspects of the topic.

Our definition is different from that in Kekäläinen and Järvelin (2002) only in the sense that it refers to document components instead of whole documents.

A scale for component specificity should allow to reward XML search engines that are able to retrieve the appropriate ("exact") sized document components. For example, a retrieval system that is able to locate the only relevant section in an encyclopaedia is likely to trigger higher user satisfaction than one that returns a too large component, such as a volume of the encyclopaedia. One could think of a measure relating the sizes of the comprising components to that of the most specific one. However, we also would like to compare the specificity of components from different documents, and here size comparison would not be appropriate—e.g., due to different writing styles. Therefore, specificity has to be judged by users. As in the case of exhaustiveness, a binary scale would not be sufficient for distinguishing between the different cases mentioned above; thus, we used the following 4-category ordinal scale for component specificity:

Not specific (0): The topic or an aspect of the topic is not a theme of the document component.

[9] In this paper, we use the term 'topical exhaustiveness' instead of 'topical relevance', in order to emphasize the two dimensions of relevance regarded here.

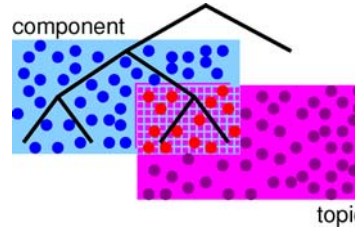


Fig. 1 Document components and topics within an ideal concept space

Marginally specific (1): The topic or an aspect of the topic is only a minor theme of the document component.

Fairly specific (2): The topic or an aspect of the topic is a major theme of the document component.

Highly specific (3): The topic is the only theme of the document component.

A consequence of the definition of topical exhaustiveness is that a container component of an exhaustive document component is also regarded as being exhaustive (since the relevant content of its child components forms part of its own content) even if it is less specific (i.e., it may also contain irrelevant child components). This clearly shows that relevance as a single criterion is not sufficient for the evaluation of content-oriented XML retrieval. For this reason, the second dimension, the component specificity criterion, is used. It measures the relation of relevant to non-relevant content within a document component.

With the combination of these two criteria it then becomes possible to differentiate between systems that return, for example, marginally or fairly specific components and systems that return the most specific relevant components, when relevant information is only contained within these sub-components.[10]

3.2. Exhaustiveness and specificity in an ideal concept space

An interpretation of topical exhaustiveness and document specificity can be done in terms of an ideal concept space as introduced by Wong and Yao (1995). Elements in the concept space are considered to be elementary concepts. Document components and topics can then be viewed as subsets of that concept space; Figure 1 uses Venn diagrams for visualisation.

If independence of the concepts in the concept space is assumed, topical exhaustiveness **exh** and component specificity **spec** can be interpreted by the following formulas:

$$\textbf{exh} = \frac{|topic \cap component|}{|topic|} \qquad \textbf{spec} = \frac{|topic \cap component|}{|component|} \tag{1}$$

Exhaustiveness thus measures the degree to which a document component covers the concepts requested by a topic. In the terminology of Wong and Yao (1995), exhaustiveness is called the *recall-oriented* measure, which reflects the exhaustiveness to which a document component discusses the topic. Values near 1 reflect highly exhaustive document components, whereas values near 0 reflect components that are not exhaustive at all with respect to the topic.

Specificity measures the degree to which a document component focuses on the topic. Wong and Yao (1995) call this the *precision-oriented* measure. Values near 1 reflect

[10] In INEX 2002, another but comparable definition of relevance was used, also based on two dimensions. The first dimension, topical relevance, corresponds to the exhaustiveness dimension defined in INEX 2003. The second dimension, coverage, is related to specificity. It has four values: no coverage, too small, too big and exact.

high specificity, while values near 0 reflect that a component is not specific at all. Values in-between reflect marginally or fairly specific components.

The interpretation of exhaustiveness and specificity in terms of an ideal concept space requires means to transform the ordinal scales (0, 1, 2 and 3) for the two relevance dimensions onto ratio scales. A quantisation function is needed for each relevance dimension. These transformations are performed by the so-called *quantisation functions*, which reflect user standpoints as to what constitutes a relevant component. For example, the *strict* quantisation functions $\mathbf{exh}_{strict}$ and $\mathbf{spec}_{strict}$ can be used to evaluate whether a given retrieval method is capable of retrieving highly exhaustive and highly specific document components:

$$\mathbf{exh}_{strict}(exh) := \begin{cases} 1 & \text{if } exh = 3, \\ 0 & \text{else.} \end{cases} \tag{2}$$

$$\mathbf{spec}_{strict}(spec) := \begin{cases} 1 & \text{if } spec = 3, \\ 0 & \text{else.} \end{cases} \tag{3}$$

In the above case, the user viewpoint is one where only highly exhaustive and specific components (i.e., both with values of 3) are of interest.

In order to credit document components according to their *degrees of* exhaustiveness and specificity (as it is done with generalised recall/precision (Kekäläinen and Järvelin, 2002)), the following *generalised* quantisation functions $\mathbf{exh}_{gen}$ and $\mathbf{spec}_{gen}$ can be used:

$$\mathbf{exh}_{gen}(exh) := \begin{cases} 1 & \text{if } exh = 3, \\ 2/3 & \text{if } exh = 2, \\ 1/3 & \text{if } exh = 1, \\ 0 & \text{else.} \end{cases} \tag{4}$$

$$\mathbf{spec}_{gen}(spec) := \begin{cases} 1 & \text{if } spec = 3, \\ 2/3 & \text{if } spec = 2, \\ 1/3 & \text{if } spec = 1, \\ 0 & \text{else.} \end{cases} \tag{5}$$

In the above case, retrieved elements that are not highly exhaustive and highly specific are rewarded, but to a lesser extent when calculating effectiveness performance. Returning such elements, which are also structurally related to a best element in a given document's XML tree, can be viewed as retrieving "near misses". The closeness of a near miss component to the best element is captured by its associated relevance values, i.e., its exhaustiveness and the specificity values.[11] Capturing near misses is very important since XML documents are accessed via both querying and browsing; thus returning elements that are near the sought-after relevant content–so, one can quickly browse to it—is better than returning elements that are far away from any relevant components.

We now look at the combinations of the different exhaustiveness and specificity values. Figure 2 shows the different possible combinations of the topical exhaustiveness degrees and component specificity values used in INEX. For example, the concept space of a highly exhaustive document component with high specificity would completely overlap the topic's concept space. It becomes clear, that not every combination makes sense. A component that is not exhaustive at all cannot be specific with respect to the topic. Vice versa, if a document component is not specific at all, then it is also not exhaustive.

[11] This comes from the fact that exhaustiveness remains or increases when going from a child element to its parent element, whereas specificity usually decreases in such a case—see Section 5.

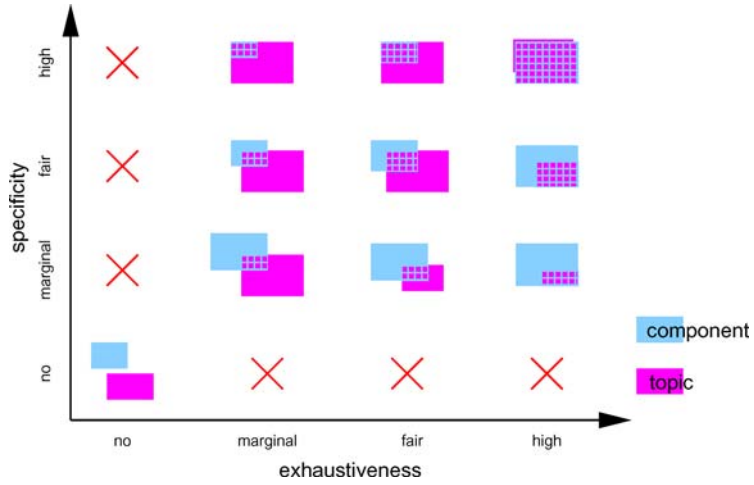


Fig. 2 Component coverage and topical relevance matrix. Components and topics are illustrated as Venn diagrams in an ideal concept space

4. A new effectiveness metric

In Section 4.1, we describe the evaluation metric developed in INEX 2002, which has been adopted as the official INEX metric. Understanding the INEX 2002 metric is important to see its shortcomings. We present our proposed new metric, the INEX 2003 metric, in Section 4.2.

4.1. INEX 2002 metric

The INEX 2002 metric applies the measure of *precall* (Raghavan et al., 1989) to document components. That is, it interprets precision as the probability $P(rel|retr)$ that a document component viewed by a user is relevant. Given that users stop viewing the ranking after having seen NR relevant document components, this probability can be computed as

$$P(rel|retr)(NR) := \frac{NR}{NR + esl_{NR}} = \frac{NR}{NR + j + s \cdot i/(r+1)}, \tag{6}$$

where esl_{NR} denotes the *expected search length*, that is the expected number of non-relevant elements seen in the rank l with the NR-th relevant document plus the number j of non-relevant documents seen in the ranks before (see Cooper (1968) for details on the derivation). Here, s is the number of relevant document components to be taken from rank l; r and i are the numbers of relevant and non-relevant elements in rank l, respectively.

Raghavan et al. (1989) give theoretical justification that intermediary real numbers can also be used (here, n is the total number of relevant document components in the collection):

$$P(rel|retr)(x) := \frac{x \cdot n}{x \cdot n + esl_{x \cdot n}} = \frac{x \cdot n}{x \cdot n + j + s \cdot i/(r+1)} \tag{7}$$

This leads to an intuitive method for employing arbitrary fractional numbers x as recall values. The metric from Raghavan has theoretical advantages over the more standard recall and precision-based metrics described in trec_eval (2002): Besides the intuitive method for

interpolation, it handles ranks containing multiple items correctly. The main advantage, however, is that it uses expectations for calculating precision, thus allowing for a straightforward implementation of the metric for the generalised quantisation function.

To apply the above metric, the two relevance dimensions are mapped to a single relevance scale by employing a quantisation function. The INEX 2002 metric employs different quantisation functions from those used for the INEX 2003 metric, whereby one quantisation function is used to map both dimensions to a single scalar value. As before, a strict and a generalised quantisation function, $\mathbf{f}_{strict}$ and $\mathbf{f}_{gen}$, respectively, are used to reflect different user viewpoints. We recall that the former, $\mathbf{f}_{strict}$, is used to evaluate retrieval methods with respect to their capability of retrieving highly exhaustive and highly specific document components.

$$\mathbf{f}_{strict}(e, s) := \begin{cases} 1 & \text{if} \quad e = 3 \quad \text{and} \quad s = 3, \\ 0 & \text{else.} \end{cases} \tag{8}$$

The generalised function, $\mathbf{f}_{gen}$, credits document components according to their *degree of* relevance, thus also allowing to reward fairly and marginally relevant elements, i.e., near misses when calculating effectiveness performance.

$$\mathbf{f}_{gen}(e, s) := \begin{cases} 1.00 & \text{if} \quad (e, s) = (3, 3), \\ 0.75 & \text{if} \quad (e, s) \in \{(2, 3), (3, 2), (3, 1)\}, \\ 0.50 & \text{if} \quad (e, s) \in \{(1, 3), (2, 2), (2, 1)\}, \\ 0.25 & \text{if} \quad (e, s) \in \{(1, 2), (1, 1)\}, \\ 0.00 & \text{if} \quad (e, s) = (0, 0) \end{cases} \tag{9}$$

For the computation of effectiveness measures, the number of relevant documents (in the retrieved set/in the whole collection) is computed as the sum of the $\mathbf{f}_{strict}$ or $\mathbf{f}_{gen}$ values of the corresponding set of components. Then the standard recall formula is applied, whereas (7) is used for computing precision.

A criticism of the INEX 2002 metric is that it does not address the problem of overlapping result elements and hence produces better effectiveness results for systems that return multiple nested components. Evidence to demonstrate this effect can be seen in Fig. 3, which shows the recall-precision graphs obtained with two simulated runs, using the generalised quantisation function. Based on the relevance assessments, a so-called "perfect" run was created containing only the elements with specificity value 3; these elements were ranked based on their exhaustiveness value. In the "ancestor" simulated run, we added to the "perfect" run all the ancestors of the elements forming it, where the "ancestor" elements are added in a single rank behind the elements of the perfect run. Hence, with the "ancestor" run, we are deliberately increasing the number of overlapping components. The graph clearly illustrates that better effectiveness is achieved by systems that return not only the most desired components (i.e., the "perfect" elements), but also their ascendant elements (i.e., the "ancestor" elements) when using the generalised quantification function.

The above problem is largely eliminated when using the strict quantisation function with the INEX 2002 metric; this is because in our simulated runs, the added ancestors will have a specificity value equal to 2 or less, and as such, they would result in a quantised score of 0. As a matter of fact, many participants prefer to use the INEX 2002 metric with the strict quantisation exactly because of this reason. However, using the strict quantisation still does not remove overlap among the highly exhaustive and specific elements, and the strict user

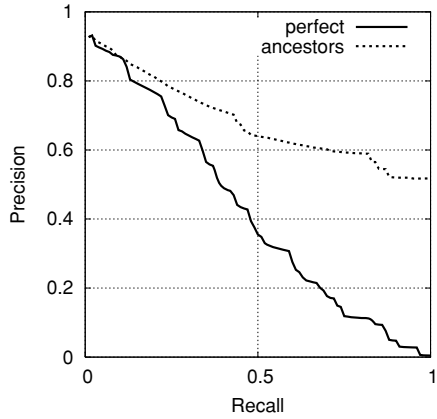


Fig. 3 Recall/precision graphs for simulated runs using the INEX 2002 metric with the generalised quantisation function. For generalised quantisation the average precision is 0.42 for the perfect run and 0.68 for the ancestors run

model also does not allow to consider near misses when evaluating content-oriented XML retrieval.

As a first solution for dealing with these issues, we developed an extended version of the 2002 metric which considered overlap and size; however, it soon became clear to us that a proper treatment of these issues is only possible when exhaustiveness and specificity are regarded separately. The INEX 2003 metric follows this idea by incorporating component size and component overlap within the definition of recall and precision.

4.2. INEX 2003 metric

Our new metric for evaluating content-oriented XML retrieval is based on the well established and understood concepts of precision and recall, but also considers component size and component overlap. We know that a direct application of recall and precision as metrics for effectiveness of XML IR systems is not suitable without additional adaptation. For this reason, we redefine the set-based measures of recall and precision in the context of XML retrieval. As pointed out in Section 2 traditional evaluation initiatives assume documents as being the atomic units to be retrieved. In the same way recall and precision have been defined as set-based measures (trec_eval, 2002):

$$\text{recall} = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents in collection}} \tag{10}$$

$$\text{precision} = \frac{\text{number of relevant documents retrieved}}{\text{total number of documents retrieved}} \tag{11}$$

These definitions do not consider the issues described in Section 2. The most crucial problems are that

- heterogeneity of component sizes are not reflected, and
- overlap of components within a ranked retrieval result is ignored.

For dealing with the amount of content of a component, the specificity dimension has been introduced into the assessments. However, this approach does not provide a solution to the

latter problem. Thus, as an alternative, we must consider component size explicitly. Instead of measuring e.g., precision or recall after a certain number of document components retrieved, we use the total size of the document components retrieved as the basic parameter. Overlap is then accounted by considering only the increment to the parts of the components already seen. In a similar way, we extrapolate the recall and precision curve for the components not retrieved, where the total size of the part of the collection not retrieved yet is then computed. We formulate the above using the concept space described in Section 3.2.

Let us assume that a system yields a ranked output list of k components $c_1, \ldots, c_k$. Let $c_i^U \subseteq U$ denote the *content* of component c_i, where U is the concept space as described in Section 3.2. In contrast, the *text* of a component c_i is denoted as c_i^T; assuming an appropriate representation like e.g., a set of pairs (term, position) (where position is the word number counted from the start of the complete document), the size of a component can be denoted as $|c_i^T|$, and the text overlap of two components c_i, c_j can be described as $c_i^T \cap c_j^T$. The complete collection consists of components $C_1, \ldots, C_N$ (where N denotes the number of all components, overlapping components not considered). Finally, $t \subseteq U$ denotes the current topic.

With these notations, we can define our variant of recall for considering document components rather than whole documents (but still ignoring overlap) in the following way: We sum up the numbers of the topic concepts in the components actually retrieved, and divide it by the sum of the numbers of topic concepts contained in all components of the collection:

$$\text{recall}_s = \frac{\sum_{i=1}^{k} \left| t \cap c_i^U \right|}{\sum_{i=1}^{N} \left| t \cap C_i^U \right|} = \frac{\sum_{i=1}^{k} \mathbf{exh}\left(c_i^U\right) \cdot |t|}{\sum_{i=1}^{N} \mathbf{exh}\left(C_i^U\right) \cdot |t|} = \frac{\sum_{i=1}^{k} \mathbf{exh}\left(c_i^U\right)}{\sum_{i=1}^{N} \mathbf{exh}\left(C_i^U\right)} \tag{12}$$

Here we use the definition of exhaustiveness ($\mathbf{exh}(c) = |t \cap c|/|t|$) from Eq. (1) in Section 3.2.

For computing precision with respect to component size, the distinction between text and content must be taken into account. Under the assumption that relevant content is distributed evenly within a given component c_i, the size of its relevant portion can be computed by $\frac{|t \cap c_i^U|}{|c_i^U|} \cdot |c_i^T|$. Using this term in the denominator and the specificity definition ($\mathbf{spec}(c) = |t \cap c|/|c|$) from Eq. (1), we obtain for precision:

$$\text{precision}_s = \frac{\sum_{i=1}^{k} \frac{|t \cap c_i^U|}{|c_i^U|} \cdot \left|c_i^T\right|}{\sum_{i=1}^{k} \left|c_i^T\right|} = \frac{\sum_{i=1}^{k} \mathbf{spec}\left(c_i^U\right) \cdot \left|c_i^T\right|}{\sum_{i=1}^{k} \left|c_i^T\right|} \tag{13}$$

The bigger the size, the higher its impact on retrieval performance; if we have two elements of equal specificity but different size, we assume that the bigger component should have a higher effect on effectiveness performance.

To take overlap into account, let us consider a component c_i (retrieved at position i in the ranking): the text not covered by other components retrieved before position i can be computed as $c_i^T - \bigcup_{j=1}^{i-1} c_j^T$. Assuming again that relevant content is distributed evenly within the component (ignoring the case where the new portion of the component does not deal with the current topic), we weigh the relevance of a component by the ratio of the component that is new.

For the denominator of the recall definition we again need to compute the maximum number of retrievable relevant concepts. In this case however, overlapping components are to be considered; relevant concepts occurring in a component are to be accounted exactly once. An upper bound can be given by the denominator in Formula 12. Instead we have to select those components of the collection, that—if being retrieved in an optimum ranking—would maximise the total number of relevant concepts $\mathbf{rel}^U$ retrieved. To do so, for a given component c we consider the number of relevant concepts and their distribution within the component as well as the number of relevant concepts in its child components:

$$\mathbf{rel}^U(c) = \begin{cases} |t \cap c^U| & \text{if } c \text{ is a leaf component} \\ \sum\limits_{c_i \in children(c)} \max\left\{\mathbf{rel}^U(c_i), |t \cap c^U| \cdot \frac{|c_i^T|}{|c^T|}\right\} & \text{else.} \end{cases} \tag{14}$$

$$= \begin{cases} |t| \cdot \mathbf{exh}(c^U) & \text{if } c \text{ is a leaf component} \\ \sum\limits_{c_i \in children(c)} \max\left\{\mathbf{rel}^U(c_i), |t| \cdot \mathbf{exh}(c^U) \cdot \frac{|c_i^T|}{|c^T|}\right\} & \text{else.} \end{cases} \tag{15}$$

$$= \begin{cases} |t| \cdot \mathbf{exh}(c^U) & \text{if } c \text{ is a leaf component} \\ |t| \cdot \sum\limits_{c_i \in children(c)} \max\left\{\frac{\mathbf{rel}^U(c_i)}{|t|}, \mathbf{exh}(c^U) \cdot \frac{|c_i^T|}{|c^T|}\right\} & \text{else.} \end{cases} \tag{16}$$

$$= |t| \cdot \begin{cases} \mathbf{exh}(c^U) & \text{if } c \text{ is a leaf component} \\ \sum\limits_{c_i \in children(c)} \max\left\{\frac{\mathbf{rel}^U(c_i)}{|t|}, \mathbf{exh}(c^U) \cdot \frac{|c_i^T|}{|c^T|}\right\} & \text{else.} \end{cases} \tag{17}$$

The maximum number of relevant components of the collection can be computed by applying $\mathbf{rel}^U$ on the collection's (virtual) root component C_{root} that connects the root components of the collection's documents to a single virtual document. Figure 4 gives an example. The topic under consideration contains four concepts. The maximum number of relevant concepts $\mathbf{rel}^U$ that can be retrieved from non-overlapping components within the illustrated collection tree is seven.

So, recall, which considers both component size and overlap, can be computed as

$$\text{recall}_o = \frac{\sum\limits_{i=1}^{k} \mathbf{exh}\left(c_i^U\right) \cdot \frac{\left|c_i^T - \bigcup_{j=1}^{i-1} c_j^T\right|}{|c_i^T|}}{\frac{\mathbf{rel}^U(C_{root})}{|t|}} \tag{18}$$

To take overlap into account in the precision measure, given a component c_i (at position i), we determine the amount of text not seen before as $c_i^T - \bigcup_{j=1}^{i-1} c_j^T$. Assuming again that relevant content is distributed evenly within the component (ignoring e.g., the case where the new portion does not deal with the current topic), we weigh the specificity of a component by the ratio of the component that is new. This way, precision accounting for component size and overlap is derived as

$$\text{precision}_o = \frac{\sum\limits_{i=1}^{k} \mathbf{spec}\left(c_i^U\right) \cdot \left|c_i^T - \bigcup_{j=1}^{i-1} c_j^T\right|}{\sum\limits_{i=1}^{k} \left|c_i^T - \bigcup_{j=1}^{i-1} c_j^T\right|} \tag{19}$$

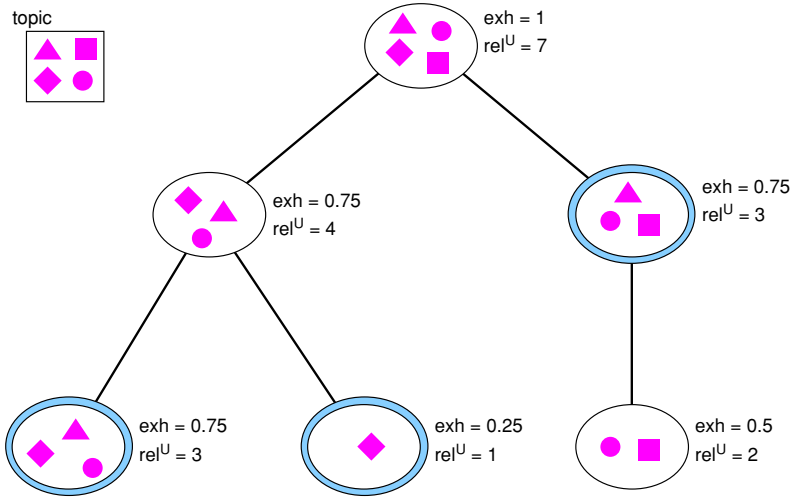


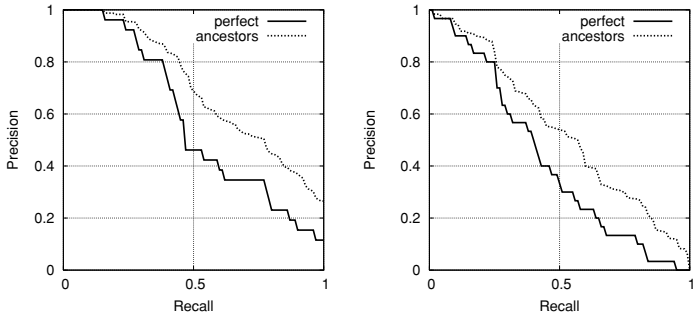
Fig. 4 $\mathbf{rel}^U$ counts the maximum number of relevant concepts, retrievable from non-overlapping components. In this example the maximum number is seven and can be achieved by retrieving the three double bordered components

These measures are generalisations of the standard recall and precision measures: In case we have non-overlapping components of equal size and no distinction between exhaustiveness and specificity, the measures are equal to the standard definitions of precision and recall.

As defined here the two INEX 2003 variants $\text{recall}_s/\text{precision}_s$ and $\text{recall}_o/\text{precision}_o$ can be applied to a single ranking. In order to yield averaged performance for a set of topics, an interpolation method is to be applied for the precision values for simple recall points. We apply the Salton method (Salton and McGill, 1983, p. 167f) here.

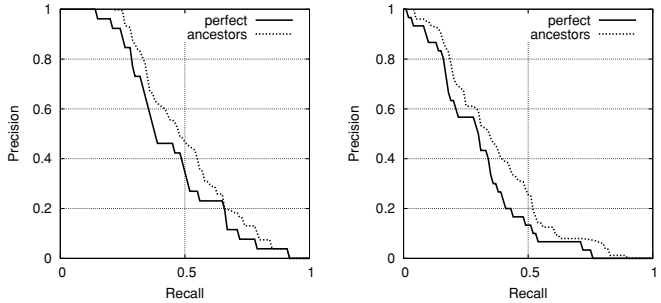
In order to show how the INEX 2003 metric behaves, the two variants of the metric were applied to the "perfect" and the "ancestors" runs described at the end of Section 4.1. The recall/precision graphs for the variant considering component size only (i.e., $\text{recall}_s/\text{precision}_s$), and for the variant considering both component size and component overlap (i.e., $\text{recall}_o/\text{precision}_o$) are given in Figs. 5 and 6, respectively. We can see that for the INEX 2003 variant that considers both component size and component overlap, the effectiveness increase is moderate compared to the increase in Figs. 3 and 5. It can be seen that considering size only (Fig. 5) is not enough; there is still a large difference between the overall effectiveness of the two simulated runs. The effect that there is at all an increase of overall effectiveness with the $\text{recall}_o/\text{precision}_o$ metric arises, because adding ancestors always means adding the siblings, cousins and so forth of the perfect elements. These components are likely to contain additional relevant material and thus on average cause a gain in effectiveness.

Applying the new metric on simulated runs shows that the proposed metric does consider overlap when calculating effectiveness performance. The next step is to compare all metrics on real runs to investigate their agreement as well as their difference in evaluating content-oriented XML retrieval. Before we do so, we describe the INEX test collection, on which we carried out this comparison.



a) strict quantisation b) generalised quantisation

Fig. 5 Recall/precision graphs for simulated runs using the INEX 2003 metric considering component size using the strict and generalised quantisation functions. For strict quantisation the average precision is 0.58 (perfect run) and 0.70 (ancestors run); for generalised quantisation the average precision is 0.42 (perfect run) and 0.54 (ancestors run)



a) strict quantisation b) generalised quantisation

Fig. 6 Recall/precision graphs for simulated runs using the INEX 2003 metric considering component size and component overlap using the strict and generalised quantisation functions. For strict quantisation the average precision is 0.45 (perfect run) and 0.51 (ancestors run); for generalised quantisation the average precision is 0.30 (perfect run) and 0.36 (ancestors run)

5. The INEX test collection

Creating a test collection requires the selection of an appropriate document collection, the creation of search topics and the generation of relevance assessments. The following sections briefly discuss these three stages of creating the INEX test collection, and provide a summary of the resulting test collection (see Fuhr and Lalmas, 2004; Fuhr et al., 2004b for full details).

5.1. XML document collection

The INEX document collection is made up of the full-texts, marked up in XML, of 12,107 articles of the IEEE Computer Society's publications from 12 magazines and 6 transactions, covering the period of 1995 to 2002, and totalling 494 megabytes in size. The collection contains scientific articles of varying length. On average an article contains 1,532 XML components, where the average depth of a component is 6.9 (more detail can be found in Fuhr et al., 2003). Overall, the collection contains over eight millions XML elements of

Table 1 Assessments at article and non-article component levels for CO topics in INEX 2003

exh	*spec*	Article	Non-article
3	3	180	1,316
3	2	112	616
3	1	150	635
2	3	24	2,105
2	2	103	1,779
2	1	222	1,358
1	3	148	5,029
1	2	50	3,872
1	1	673	8,074
0	0	10,021	70,530
	Sum	11,783	95,314

varying granularity (from table entries to paragraphs, sub-sections, sections and articles, each representing a potential answer to a query).

5.2. Search topics

In order to consider the additional functionality introduced by the use of XML query languages, which allows the specification of structural query conditions, INEX defined two types of topics:

Content-only (CO) queries are standard IR retrieval tasks similar to those used in TREC. Given such a query, the goal of an XML retrieval system is to retrieve the most specific XML element(s) answering the query in a satisfying way. Thus, a system should e.g., not return a complete article where a section or even a paragraph of the same document may also be sufficient.

Content and structure (CAS) queries contain conditions referring both to content and structure of the requested answer elements. A query condition may refer to the content of specific elements (e.g., the elements to be returned must contain a section about a particular topic). Furthermore, the query may specify the type of the requested answer elements (e.g., sections should be retrieved). The query language defined for this purpose is a variant of XPath 1.0 (Clark and DeRose, 1999).

As in TREC, an INEX topic consists of the standard *title*, *description* and *narrative* fields. From an evaluation point of view, both query types support the evaluation of retrieval effectiveness as defined for content-oriented XML retrieval, where for CAS queries the information need to be satisfied by a document component has to also consider the explicit structural constraints. The metric developed in Section 4.2 does not consider such structural constraints; thus here we restrict our study to retrieval effectiveness for CO topics. An example of a CO topic is given in Fig. 7.

The INEX topics were created by the participating institutions using their own XML retrieval systems or the system provided by the INEX organisers[12] for the collection exploration stage of the topic development process. In 2002, 30 CO were selected to be included in the INEX test collection; another 36 CO were added for the second round of INEX in 2003.

[12] In INEX 2003, the HyREX system developed in Duisburg-Essen was made available to participants for the topic creation phase, see http://www.is.informatik.uni-duisburg.de/projects/hyrex/.

```
<inex_topic topic_id="98" query_type="CO" ct_no="26">
  <title>
    Information Exchange, XML, Information Integration
  </title>
  <description>
    How to use XML to solve the information exchange
    (information integration) problem, especially in
    heterogeneous data sources?
  </description>
  <narrative>
    Relevant documents / components must talk about
    techniques of using XML to solve information exchange
    (information integration) among heterogeneous data
    sources where the structures of participating data
    sources are different although they might use the same
    ontologies about the same content.
  </narrative>
  <keywords>
    information exchange, XML, information integration,
    heterogeneous data sources
  </keywords>
</inex_topic>
```

Fig. 7 A CO topic from the INEX 2003 test collection

5.3. Assessments

Like the topics, the assessments have been derived in a collaborative effort. For each topic, the results from the participants' submissions have been collected into pools using the pooling method (Voorhees and Harman, 2002). Where possible, the author of a given topic did the assessment of the respective result pool as well. To ensure complete assessments, assessors were provided an on-line assessment system and the task of assessing every relevant document component, and their ascendant and descendant elements within the articles of the result pool (Piwowarski and Lalmas, 2004). The assessors were given detailed information about the evaluation criteria (see Section 3) and about how to perform the assessments.

Table 1 shows statistics on the assessments on article and non-article elements for CO topics in INEX 2003. The collected assessments contain a total of 163,306 assessed elements, of which 11,783 are at article level. About 96 % of the 8,802 components that were assessed as highly specific are non-article level elements. This percentage was 87% (of 3,747 components) in INEX 2002. These numbers indicate that sub-components are preferred to whole articles as retrieved units, which is not reflected when using the INEX 2002 metric for calculating retrieval effectiveness.

6. Experiments and results

We performed a number of experiments to investigate how the proposed INEX 2003 metric differs from the INEX 2002 metric.[13] We recall that the INEX 2003 metric comes in two

[13] We chose these two test sets since they differ in the nature of the assessments and the size of the runs; the INEX 2004 setting was similar to that of 2003.

variants, one which considers component size, and one which considers both component size and component overlap. We refer to these as the INEX 2003s (i.e., $\text{recall}_s/\text{precision}_s$) and INEX 2003o (i.e., $\text{recall}_o/\text{precision}_o$) metrics to follow the notation adopted in Section 4.2.

Experiments were done on three result sets, two variants of the official INEX 2003 submission runs, one with 1500 elements and the second with 100 elements, and the official INEX 2002 submission runs. For CO topics, 24 participating organisations submitted 56 runs in INEX 2003. In INEX 2002, these numbers were respectively 25 and 49. The INEX 2002 submission runs consisted of 100 elements, whereas this number was 1500 for the INEX 2003 submissions.

We first investigate the influence of the size of the result sets on all metrics in Section 6.1. We then look at the effect of the quantisation functions, i.e., strict vs. generalised, on the three result sets in Section 6.2. The two variants of the proposed new metric are compared in Section 6.3. Finally, the INEX 2002 metric and the two variants of the INEX 2003 metric are compared in Section 6.4.

We use the Pearson's correlation coefficient applied to average precision values to measure to which extent any two metrics (e.g., INEX 2002 vs. INEX2003s) or different uses of one metric (e.g., INEX 2003s applied to runs of 100 elements vs. INEX 2003s applied to run of 1500 elements) are related. A value closer to 1 shows correlation (i.e., comparable behaviour) whereas a value closer to 0 implies independence (i.e., unrelated behaviour). In some cases, we show the corresponding scatter plots and regression lines.

6.1. Number of elements in results

Here we compare whether the number of result elements has any influence on retrieval effectiveness (average precision values) as calculated by all three metrics. For this, we apply all three metrics on the two variants of the 2003 submission runs. The results are given in Table 2.

The INEX 2002 metric seems to be less sensitive than the INEX 2003 metric (both variants) to the size of result elements used to calculate retrieval effectiveness. Using the strict quantisation function rather than the generalised one also seems to be less sensitive to result size. This observation is stronger for the INEX 2003o metric. This can be further observed if we look at the scatter plot for average precision of all official INEX 2003 submissions, using the INEX 2003o metric for 100 and 1500 result elements per submission (Fig. 8).

This result is to be expected as a bigger result set is bound to have more overlapping components, which will affect retrieval effectiveness as calculated by the INEX 2003o metric. It is predominantly with the generalised quantisation that component overlap is an issue. This would suggest that a better report of the effectiveness results using the INEX 2003o metric should be done at various cut-off values (various result set sizes) so that to obtain a

Table 2 Correlation coefficients of the average precision of all official INEX 2003 submissions for 100 and 1500 result elements per submission

Metric	Quantisation	
INEX 2002	Strict	0.98257
INEX 2002	Generalised	0.96377
INEX 2003s	Strict	0.90910
INEX 2003s	Generalised	0.90207
INEX 2003o	Strict	0.93132
INEX 2003o	Generalised	0.87009

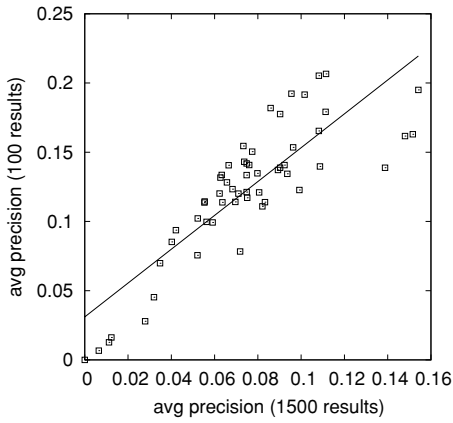


Fig. 8 Scatter plot and regression line for average precision of all official INEX 2003 submissions, using the INEX 2003o metric for 100 and 1500 result elements per submission

finer-grained evaluation, as it is well known that end-users will never look at 1500 or more hits. We will then be able to differentiate between systems that have component overlap at lower ranks, which may be considered to be better systems, and systems with overlapping components higher in the ranking.

6.2. Quantisations: Strict vs. generalised

For the INEX 2002 metric as well as for the new INEX 2003s and INEX 2003o metrics, different quantisation functions (strict and generalised) are provided. Here we examine the influence of the quantisation function on the ranking of submissions with respect to retrieval effectiveness. Results on the three submission run sets are given in Table 3.

Using different quantisation functions seems to be more of an issue with a metric that considers overlap. This can be observed for both 2003 result sets (100 and 1500 elements). For these result sets, INEX 2003s seems to be the least affected by which quantisation function is used. Now if we look at results obtained with the INEX 2002 submission runs, the two quantisation functions lead to very similar results (see also Fig. 9). This can be explained in two ways. First, the INEX 2002 submission set is smaller, and less elements usually implies less problems with overlapping components (Section 6.1). Second, the set of relevance assessments obtained in INEX 2002 is not as complete as that obtained in INEX 2003; in the latter assessors were forced to assess all ascendant and descendant elements (see Piwowarski and Lalmas, 2004), thus increasing the possible number of overlapping elements.

In INEX 2004, new quantisation metrics have been proposed to reflect other user viewpoints (see Kazai, 2004), and it would be interesting to see their effect on the various metrics. Apart from one noticeable difference (INEX 2003o on INEX 2003

Table 3 Correlation coefficients of the average precision of all the three result sets for strict and generalised quantisation

Metric	2003 run (1500)	2003 runs (100)	2002 runs (100)
INEX 2002	0.92045	0.92111	0.94799
INEX 2003s	0.97383	0.95516	0.94981
INEX 2003o	0.87410	0.89997	0.95121

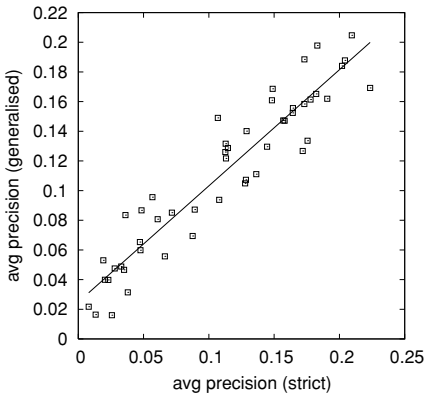


Fig. 9 Scatter plot and regression line for average precision of all official INEX 2002 submissions, using the INEX 2003o metric with strict and generalised quantisation

submission runs), the above results seem to question the need for several quantisation functions, as results tend to be relatively comparable. Further investigation is needed here.

6.3. INEX 2003 Metric: Simple vs. overlap

The INEX 2003 metric comes in two flavours: The INEX 2003s metric considers component size, but does not consider overlap, whereas INEX 2003o considers both size and overlap. We compare these variants, using both quantisation functions on the three result sets. All results are given in Table 4.

Except for the INEX 2002 runs the correlation coefficients show that considering overlap makes a real difference. This can also be seen from the scatter plots in Fig. 10. From the user's standpoint, retrieval systems should aim to retrieve relevant document components which ideally do not overlap. Given this and the relatively low correlation between the two INEX 2003 metrics, it becomes clear that it is worth using the INEX 2003o metric for evaluation of content-oriented XML retrieval.

6.4. Comparison between the INEX 2002 and INEX 2003 metrics

We compare how the results of the INEX 2002 metric deviate from the INEX 2003 metric, in its two variants. All results are given in Table 5.

It can be seen that there is a strong difference between the INEX 2002 and the INEX 2003 metric that considers overlap. The difference is stronger when the generalised quantisation function is used. The difference is still there when submission runs are composed of 100 elements. The differences are less because as we know now the size of the result sets affects the metrics.

Table 4 Correlation coefficients of the average precision for the three result sets for INEX 2003s and INEX 2003o

Quantisation	2003 runs (1500)	2003 runs (100)	2002 runs (100)
Strict	0.79631	0.82262	0.96954
Generalised	0.82443	0.80633	0.94529

Table 5 Correlation coefficients of the average precision for all three result sets for both INEX 2003 metrics compared to the INEX 2002 metric

		Result set		
Metrics	Quantisation	2003 (1500)	2003 (100)	2002 (100)
INEX 2002—INEX 2003s	Strict	0.89547	0.95233	0.94647
INEX 2002—INEX 2003s	Generalised	0.93660	0.97479	0.90292
INEX 2002—INEX 2003o	Strict	0.79004	0.80645	0.95503
INEX 2002—INEX 2003o	Generalised	0.69793	0.71330	0.93360

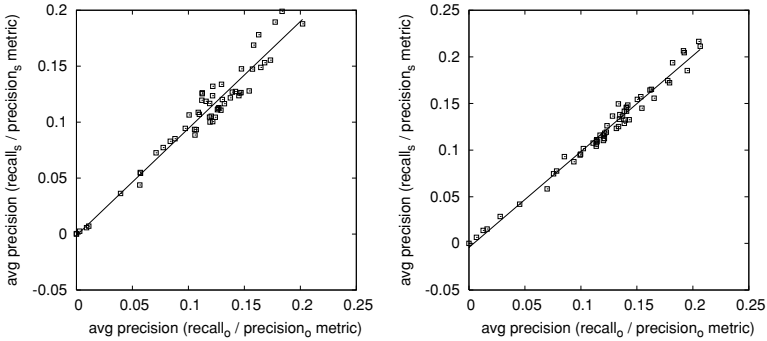


Fig. 10 Scatter plots and regression lines for average precision of all official INEX 2003 submissions (100 elements), using INEX 2003s and INEX 2003o

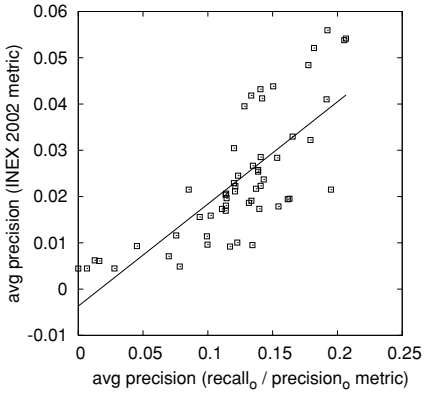


Fig. 11 Scatter plots and regression lines for average precision of all official INEX 2003 submissions (100 elements), using INEX 2002 and INEX 2003o with generalised quantisation

To further illustrate the difference between INEX 2002 and INEX 2003o, Fig. 11 shows the scatter plot for the submissions done in 2003 with average precision computed by means of generalised quantisation. We can clearly see that systems that did well according to the official INEX metric, INEX 2002, did not perform as well when overlap was considered. This indicates that we indeed need a metric that considers component size and how much overlapping components are returned by a system, in order to be able to appropriately compare XML retrieval strategies.

7. Conclusion and outlook

Evaluating the effectiveness of content-based retrieval of XML documents is a necessary requirement for the further improvement of research on XML retrieval. In this article we showed that traditional IR evaluation methods are not suitable for content-oriented XML retrieval evaluations.

We proposed new evaluation criteria, measures and metrics based on the two dimensions of content and structure to evaluate XML retrieval systems according to a re-defined concept of retrieval effectiveness. New metrics based on the well-established measures recall and precision have been developed. In order to reward systems which provide specific document components with respect to a given query, component size and possibly overlapping components in retrieval results are considered.

By applying the different metrics to the INEX 2002 and INEX 2003 submissions, we have investigated the effect of different evaluation parameters on the ranking of the submitted runs:

- The number of elements in the results (which are considered for evaluation) has an effect on the ranking, when element size or overlap are considered. Thus, for a more user-oriented evaluation, various realistic cut-off values should be considered when applying the new metrics.
- Considering overlap, in addition to component size, affects the system ranking. Also, the comparison of our new metric with the INEX 2002 metric shows significant differences in the ranking of systems, especially when overlap of components is considered. There is some preliminary evidence that users dislike overlapping results (Tombros et al., 2005); thus, this parameter should not be ignored with regard to comparing system performance.
- The type of quantisation applied has an effect on the ranking of systems when component overlap is considered. Under the presumption that component overlap is to be considered for comparing system performance, it is thus worth considering multi-valued scales for specificity and exhaustiveness as well as encoding different user standpoints by means of appropriate quantisation functions. However, multi-valued scales may reduce the reliability of assessments.

Overall, we can conclude that the new metric investigated in this article seems to be well suited for the evaluation of XML IR systems. However, like most metrics (e.g., Piwowarski and Gallinari, 2004; Kazai et al., 2004), also our approach is based on assumptions about typical user behaviour. The ongoing INEX track on interactive retrieval is collecting empirical data about user interactions with XML IR systems. The analysis of this data will provide a good foundation for the further development of appropriate metrics.

References

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

Baeza-Yates, R., Fuhr, N., Sacks-Davis, R., & Wilkinson, R. (Eds.) (2000). In *Proceedings of the SIGIR 2000 Workshop on XML and Information Retrieval*. `http://www.haifa.il.ibm.com/sigir00-xml/index.html`

Beaulieu, M., & Robertson, S. (1996). Evaluating interactive systems in TREC. *Journal of the American Society for Information Science, 47*(1), 85–94.

Chiaramella, Y., Mulhem, P., & Fourel, F. (1996). A model for multimedia information retrieval. Technical report, FERMI ESPRIT BRA 8134, University of Glasgow.

Clark, J., & DeRose, S. (1999). XML path language (XPath) version 1.0. Technical report, World Wide Web Consortium. http://www.w3.org/TR/xpath

Cleverdon, C. W., Mills, J., & Keen, E. M. (1966). Factors determining the performance of indexing systems, vol. 2: Test results. Technical report, Aslib Cranfield Research Project, Cranfield, England.

Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science, 19*, 30–41.

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing and Management, 36*(4), 533–550.

Craswell, N., & Hawking, D. (2004). Overview of the TREC 2004 Web Track. In *Proceedings of the 13th Text Retrieval Conference (TREC-2004)*. Gaithersburg, Maryland, USA, NIST. http://trec.nist.gov/pubs/trec13/papers/WEB.OVERVIEW.pdf.

Croft, W. B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R., & Zobel, J. (Eds.) (1998). In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM.

Fuhr, N., Gövert, N., Kazai, G., & Lalmas, M. (Eds.) (2003). INitiative for the Evaluation of XML Retrieval (INEX). In *Proceedings of the First INEX Workshop.* Dagstuhl, Germany, Dec. 8–11, 2002. *ERCIM Workshop Proceedings.* Sophia Antipolis, France: ERCIM. http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf

Fuhr, N., & Lalmas, M. (2004). Report on the INEX 2003 Workshop. *SIGIR Forum, 38*(1).

Fuhr, N., Lalmas, M., & Malik, S. (Eds.) (2004a). INitiative for the Evaluation of XML Retrieval (INEX). In *Proceedings of the Second INEX Workshop.* Dagstuhl, Germany, Dec. 15–17, 2003. http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf.

Fuhr, N., Malik, S., & Lalmas, M. (2004b). Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2003. In Fuhr et al. (2004a), (pp. 1–11).

Harman, D. (1993). Overview of the First Text REtrieval Conference. In D. Harman (Ed.) *The First Text REtrieval Conference (TREC-1)*, Gaithersburg, Md. 20899, National Institute of Standards and Technology Special Publication 500-207.

Jones, K. S., & van Rijsbergen, C. J. (1976). Information retrieval test collections. *Journal of Documentation, 32*(1), 59–75.

Kando, N., & Adachi, J. (2004). Report from the NTCIR workshop 3. *SIGIR Forum, 38*(1), 10–16.

Kazai, G. (2004). Report of the INEX 2003 Metrics working group. In Fuhr et al. (2004a) pp. 184–190.

Kazai, G., Lalmas, M., & de Vries, A. P. (2004). The overlap problem in content-oriented XML retrieval evaluation. In K. Järvelin, J. Allen, P. Bruza, & M. Sanderson (Eds.), *Proceedings of the 27st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 72–79) New York, ACM.

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance sssessments in IR evaluation. *Journal of the American Society for Information Science and Technology, 53*(13).

Lancaster, F. W. (1968). Evaluation of the MEDLARS demand service. Report, National Library of Medicine, Bethesda, Maryland.

Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds.) (2002). Evaluation of cross-language information retrieval systems (CLEF 2001). Vol. 2406 of *Lecture Notes in Computer Science*. Heidelberg et al., Springer.

Piwowarski, B., & Gallinari, P. (2004). Expected ratio of relevant units: A measure of structured information retrieval. In Fuhr et al. (2004a), pp. 158–166. http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf

Piwowarski, B., & Lalmas, M. (2004). Ensuring consistent and exhaustive relevance assessments for XML retrieval evaluation. In L. Gravano (Ed.), *Proceedings of the 13th International Conference on Information and Knowledge Management*. New York, ACM.

Raghavan, V. V., Bollmann, P., & Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems, 7*(3), 205–229.

Salton, G. (Ed.) (1971). *The SMART retrieval system—Experiments in automatic document processing*. Englewood, Cliffs, New Jersey: Prentice Hall.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In: E. A. Fox, P. Ingwersen, and R. Fidel (Eds.), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 138–146), New York, ACM. ISBN 0-89791-714-6.

Saracevic, T. (1996). Relevance reconsidered. In P. Ingwersen and N. O. Pors (Eds.), In *Proceedings of the 2nd International Conference on Conceptions of Library and Information Science (CoLIS 2)*, Oct. 13–16, 1996, (pp. 201–218).

Tombros, A., Larsen, B., & Malik, S. (2005). The Interactive Track at INEX 2004. In N. Fuhr, M. Lalmas, S. Malik, & Z. Szlavik (Eds.), *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004,* Dagstuhl Castle, Germany, Dec. 6–8, 2004, *Revised Selected Papers*, Vol. 3493. Springer-Verlag GmbH. `http://www.springeronline.com/3-540-26166-4`.

trec_eval (2002). Evaluation techniques and measures. In Voorhees and Harman (2002), NIST.

Voorhees, E. M. (1998). Variations in relevance judgements and the measurement of retrieval effectiveness. In Croft et al. (1998), (pp. 315–323), ACM.

Voorhees, E. M., & Harman, D. K. (Eds.) (2002). The Tenth Text REtrieval Conference (TREC 2001). Gaithersburg, MD, USA: NIST.

Wong, S. K. M., & Yao, Y. Y. (1995). On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems, 13*(1), 38–68.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments?. In Croft et al. (1998) (pp. 307–314), ACM.