Evaluating the effectiveness of content-oriented XML retrieval

Norbert Gövert*
University of Dortmund

Gabriella Kazai
Queen Mary University of London

Norbert Fuhr
University of Duisburg-Essen

Mounia Lalmas
Queen Mary University of London

The INEX initiative is a collaborative effort for building an infrastructure to evaluate the effectiveness of content-oriented XML retrieval. In this paper, we show that evaluation methods developed for standard test collection must be modified in order to deal with retrieval of structured documents. Specifically, size and overlap of document components must be taken into account. For this purpose, we use coverage in addition to relevance as evaluation criterion, in combination with multi-valued scales for both. A new quality metric based on the notion of concept spaces is developed. We compare the results of this new metric with the results obtained by the metric used in the first round of INEX, in 2002.

Category: H.3.3 Information Search and Retrieval

Keywords: Content-oriented XML Retrieval; Effectiveness; Evaluation; INEX

1 Introduction

Increasingly, the *eXtensible Markup Language (XML)* is acknowledged as the standard document format for fulltext 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 an 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, IR systems aim now to implement a more focused retrieval paradigm. That is, instead of retrieving whole documents, IR systems aim at retrieving *document components* that fulfill the user's information need. According to the FERMI Multimedia Information Retrieval model, these components should be the deepest components in the document structure, i. e. most specific, while remaining exhaustive to the information need [Chiaramella et al., 1996]. 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 implementing content-oriented XML retrieval in this way have been built [Baeza-Yates et al., 2002, 2000; Fuhr et al., 2003]. However comprehensive evaluation of the retrieval effectiveness of such systems has not yet been performed. One of the reasons for this is that traditional IR test collections, such as provided by TREC [Voorhees and Harman, 2002] and CLEF [Peters et al.,

*goevert@ls6.cs.uni-dortmund.de

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 order to set up an infrastructure for the evaluation of content-oriented XML retrieval, a number of problems have to be solved. Evaluation criteria have to be defined, which consider retrieval at the document components level. For these criteria an appropriate scaling is to be developed. Having defined criteria and their scaling, a suitable metric needs to be devised, that facilitates reliable statements about the effectiveness of algorithms developed for content-oriented XML retrieval.

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.

In this paper we describe the applied evaluation criteria, measures and metrics. First, in Section 2 we examine the assumptions underlying traditional IR evaluation initiatives and highlight their invalidity when evaluating content-oriented XML retrieval. In Section 3 we describe the INEX test collection and the methodology used for its construction. Section 4 details the selected evaluation criteria and measures. In Section 5 we develop new metrics for evaluating effectiveness of content-oriented XML retrieval. Section 6 provides the results of the new metrics applied to the INEX 2002 runs and compares them to the results obtained by the metric used in INEX 2002. We close in Section 7 with conclusions and an outlook on future INEX runs.

2 Evaluation Considerations

Evaluation initiatives such as TREC[2] and CLEF[3] 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. 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 irrelevant ones.

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 and relevance assessments. There have been several large-scale evaluation projects, which resulted in well established IR test collections [Cleverdon et al., 1966; Peters et al., 2002; Salton, 1971; Sparck Jones and van Rijsbergen, 1976; Voorhees and Harman, 2002]. 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 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 fulltext documents such as books, where one would like to consider also

[1] http://qmir.dcs.qmw.ac.uk/INEX/
[2] http://trec.nist.gov/
[3] http://www.clef-campaign.org/

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. 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.

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 (e. g. 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. 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). Since single components typically are not completely context-(document-)independent, 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 INEX, the first issue (dependence) is ignored (as e. g. in TREC, too). In order to deal with component size and overlap, we develop new evaluation criteria and metrics (Sections 4 and 5). Finally, for the document-wise grouping of components, we assume that for any cutoff point, there is an implicit grouping of the document components retrieved so far[4].

3 Construction of the INEX Test Collection

The methodology for constructing a structured document test collection, although similar to that used for building traditional IR test collections, has additional requirements [Kazai et al., 2003]. As in IR test collection building, the processes involved are the selection of an appropriate document collection, the creation of user requests and the generation of relevance assessments. The following sections discuss the first two stages of creating the INEX test collection, and provide a brief summary of the resulting test collection. The third stage, generation of relevance assessments, is described in Section 4.3.

[4]The metrics developed in Section 5 are not affected by such a grouping, as long as the order of components from the same document is not changed.

3.1 XML Document Collection

For the evaluation of XML IR systems, the document collection must consist of XML documents with reasonable structural complexity. The overall size of the collection should also be considered [Sparck Jones and van Rijsbergen, 1975].

The document collection in INEX is made up of the fulltexts, 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 1995–2002, and totalling 494 megabytes in size. Although the collection is relatively small compared with TREC, it has a suitably complex XML structure (192 different content models in DTD) and contains scientific articles of varying length. On average an article contains 1 532 XML nodes, where the average depth of a node is 6.9. A more detailed summary can be found in [Fuhr et al., 2002].

3.2 User Requests

The topic and format of the user requests should be representative of the variety of real world uses. In the case of XML retrieval systems, where the use of XML query languages is being adopted, these should include the use of queries that allow the specification of structural query conditions. According to this new criterion we can distinguish two types of queries:

Content-only (CO) queries are in a sense the standard type of query in IR. They describe a topic of interest and ignore the document structure. The need for this type of queries for the evaluation of XML retrieval stems from the fact that users often cannot or would not restrict their search to specific structural elements.

Content-and-structure (CAS) queries are topic statements that contain explicit references to the XML structure. An example of such a query is "Retrieve the title and first paragraph of sections about wine making in the Rhine region".

The format and the development procedures for the INEX queries (also called *topics*) were based on TREC guidelines, which were modified to take into account the structural requirements of the CAS topics [INEX, b]. The modified topic format of INEX allows the definition of containment conditions (concepts within specific types of components) and target elements (type of components to return to the user).

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 INEX topics were created by the participating groups using their own XML retrieval systems for the collection exploration stage of the topic development process. From the submitted 143 candidate topics, 60 (30 CO and 30 CAS) topics were selected to be included in the final set of topics (see [Fuhr et al., 2002]).

4 Evaluation Criteria for INEX

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. However, as it was pointed out in Section 2, these criteria and measures rely on implicit assumptions about the documents and users, which do not hold for XML retrieval. It is therefore necessary to reformulate the evaluation objective and to develop new evaluation procedures to address the additional requirements introduced by the structure of the documents and the implications of such a structure.

4.1 Topical Relevance and Component Coverage

For INEX, we set the objective of content-based evaluation of XML retrieval to be the assessment of a system's retrieval effectiveness, where, following the FERMI Multimedia Information Retrieval model [Chiaramella et al., 1996], effectiveness has been redefined as the system's ability to retrieve the most specific relevant document components, which are exhaustive to the topic of request. This combination of content and structural requirements within the definition of retrieval effectiveness must also be reflected in the evaluation criteria to be used. Separating the content and structural dimensions, we choose the following two criteria:

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

Component coverage reflects the extent to which a document component is focused on the information need.

Topical relevance here refers to the standard relevance criterion used in IR. This choice is reasonable, despite the debates regarding the notion of relevance [Cosijn and Ingwersen, 2000; Saracevic and Pors, 1996], 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 IR systems, we must also decide about the scales of measurements to be used. For 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 multiple degree relevance scale is also better suited for content-oriented XML retrieval evaluation as 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. Binary values of relevance cannot reflect this difference. In INEX, we therefore adopted the following four-point relevance scale [Kekäläinen and Järvelin, 2002]:

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

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

Fairly relevant (2): The document component contains information relevant w. r. t. 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 relevant (3): The document component discusses the topic of request exhaustively. In the case of multi-faceted topics, all or most sub-themes or viewpoints are discussed.

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 coverage should allow to reward XML 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 the whole encyclopaedia. Another aspect to be considered within the coverage dimension is the fact that a document component's only theme is an aspect of the topic, but on the other hand is too small to bear self-explaining information for the user and thus cannot serve as *informative unit* [Chiaramella et al., 1996]. To accommodate these requirements, we used the following 4-category nominal scale for component coverage:

No coverage (N): The topic or an aspect of the topic is not a theme of the document component.

Too large (L): The topic or an aspect of the topic is only a minor theme of the document component.

Too small (S): The topic or an aspect of the topic is the main or only theme of the document component, but the component is too small to act as a meaningful unit of information when retrieved by itself.

Exact coverage (E): The topic or an aspect of the topic is the main or only theme of the document component, and the component acts as a meaningful unit of information when retrieved by itself.

A consequence of the definition of topical relevance is that container components of relevant document components in a nested XML structure are also regarded as relevant, albeit a too large component. 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 is used, the component coverage criterion, which provides a measure with respect to the size of a component. The coverage dimension measures the ratio of relevant and irrelevant content within a document component.

With the combination of these two criteria it then becomes possible to differentiate between systems that return, for example, too large components (e. g. whole documents such as a book) and systems that return the most specific relevant components, when relevant information is only contained within these sub-components.

4.2 Relevance and Coverage in an Ideal Concept Space

An interpretation of topical relevance and document coverage 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.

If independence of the concepts in the concept space is assumed, topical relevance rel and component coverage cov can be interpreted by the following formulas:

$$rel = \frac{|topic \cap component|}{|topic|} \qquad cov = \frac{|topic \cap component|}{|component|} \tag{1}$$

Relevance thus measures the degree to which a document component covers the concepts requested by a topic. In the terminology of [Wong and Yao, 1995], relevance is called the *recall-oriented* measure which reflects the exhaustiveness to which a document component discusses the topic. Values near 1 reflect highly relevant document components, whereas values near 0 reflect irrelevant components with respect to the topic.

Coverage measures the degree to which a document component focuses on the topic. [Wong and Yao, 1995] call this the *precision-oriented* measure, i. e. it measures how *precise* a document component is about a topic. Values near 1 reflect exact coverage, while values near 0 reflect no coverage. Values inbetween reflect coverage *too large*. The value *too small* must be seen as an outlayer here. It is used to express that a component is focused on the topic of request, and therefore would get a coverage value near one, but the same time it is considered useless as a retrieval answer on its own.

Interpretation of relevance and coverage in terms of an ideal concept space means to transform the ordinal scales for the two dimensions onto ratio scales. Here, different quantisation functions $\mathbf{rel}_{quant}, \mathbf{cov}_{quant}$ are to be chosen, such that the desired user standpoint is taken into account. For example, the *strict* quantisation functions $\mathbf{rel}_{strict}$ and $\mathbf{cov}_{strict}$ are used to evaluate whether a given retrieval method is capable of retrieving highly relevant and highly focused document components:

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

In order to credit document components according to their *degree of* relevance and coverage (generalised recall / precision [Kekäläinen and Järvelin, 2002]), the *generalised* quantisation functions $\mathbf{rel}_{generalised}$ and $\mathbf{cov}_{generalised}$ are used:

$$\mathbf{rel}_{generalised}(rel) := \begin{cases} 1 & \text{if } rel = 3, \\ 2/3 & \text{if } rel = 2, \\ 1/3 & \text{if } rel = 1, \\ 0 & \text{else.} \end{cases} \qquad \mathbf{cov}_{generalised}(cov) := \begin{cases} 1 & \text{if } cov = \text{E}, \\ 1/2 & \text{if } cov = \text{L}, \\ 0 & \text{else.} \end{cases} \tag{3}$$

Given these relationships for the relevance and coverage dimensions, we now look at combinations of the different relevance and coverage values. Figure 1 shows the different possible combinations of the topical

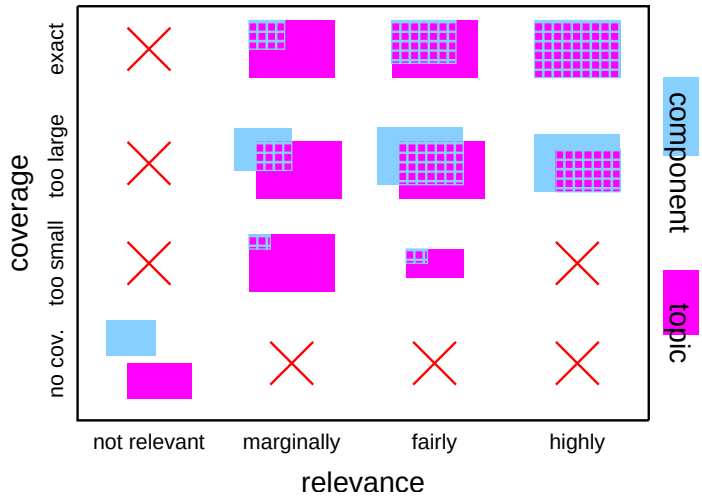


Figure 1: Document coverage and topical relevance matrix. Documents and topics are illustrated as Venn diagrams in an ideal concept space.

relevance degrees and component coverage categories used in INEX. For example, the concept space of a highly relevant document component with exact coverage would completely overlap the topic's concept space. It becomes clear, that not every combination makes sense. A component which has no relevance cannot have any coverage with the topic. Vice versa, if a document component has no coverage with a topic, it cannot be relevant to the topic at the same time. In a similar way, a document component which has a coverage too small, cannot be highly relevant, since this would assume that all or most of the concepts requested by the topic are discussed exhaustively.

4.3 Assessments

The result pools for the assessments were collected using the pooling method [Voorhees and Harman, 2002]. From the 49 participating groups, 25 submitted a total of 51 runs, each containing the top 100 retrieval results for the 60 topics. The pooled result sets contained between 1000–2000 document components from 300–900 articles, depending on the topic. Assessors had the use of 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. Due to the high resource requirement of this task, each topic was assessed only by one assessor.

The collected assessments contain a total of 67 992 assessed elements, of which 22 326 are at article level (note that this excludes statistics for six CO topics, whose assessment has not been completed at the time of writing). About 87 % of the 3747 components which were assessed with exact coverage are non-article level elements. This indicates that sub-components are preferred to whole articles as retrieved units.

5 Evaluation Metrics

As an alternative to the evaluation metric, which has been developed for INEX 2002 Gövert and Kazai [2003], we provide here metrics for content-oriented XML retrieval which deal with the issues discussed in Section 2.

Application of recall and precision as metrics for effectiveness of XML IR systems is not suitable without additional adaptation. Here we redefine the set-based measures for recall and precision for the new task. 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]:

$$\text{recall} = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents in collection}} \quad (4)$$

$$\text{precision} = \frac{\text{number of relevant documents retrieved}}{\text{total number of documents retrieved}} \quad (5)$$

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

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

For dealing with the amount of content of a component, the coverage criterion has been introduced into the assessments. However, this approach provides no 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 accounted by considering only the increment to the parts of the components already seen. In a similar way, we extrapolate the recall / precision curve for the components not retrieved: The total size of the part of the collection not retrieved yet can be computed very easily.

More formally, 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 4.2. 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 documents $d_1, \ldots, d_N$ (where N denotes the number of documents in the collection); since documents are also components, the notations d_i^U for the content and d_i^T apply here as well. Finally, $t \subseteq U$ denotes the current topic.

With these notations, recall, which considers component size (but ignores overlap), can be computed as:

$$\text{recall}_s = \frac{\sum_{i=1}^{k} \left|t \cap c_i^U\right|}{\sum_{i=1}^{N} \left|t \cap d_i^U\right|} = \frac{\sum_{i=1}^{k} rel\left(c_i^U\right) \cdot |t|}{\sum_{i=1}^{N} rel\left(d_i^U\right) \cdot |t|} = \frac{\sum_{i=1}^{k} rel\left(c_i^U\right)}{\sum_{i=1}^{N} rel\left(d_i^U\right)} \quad (6)$$

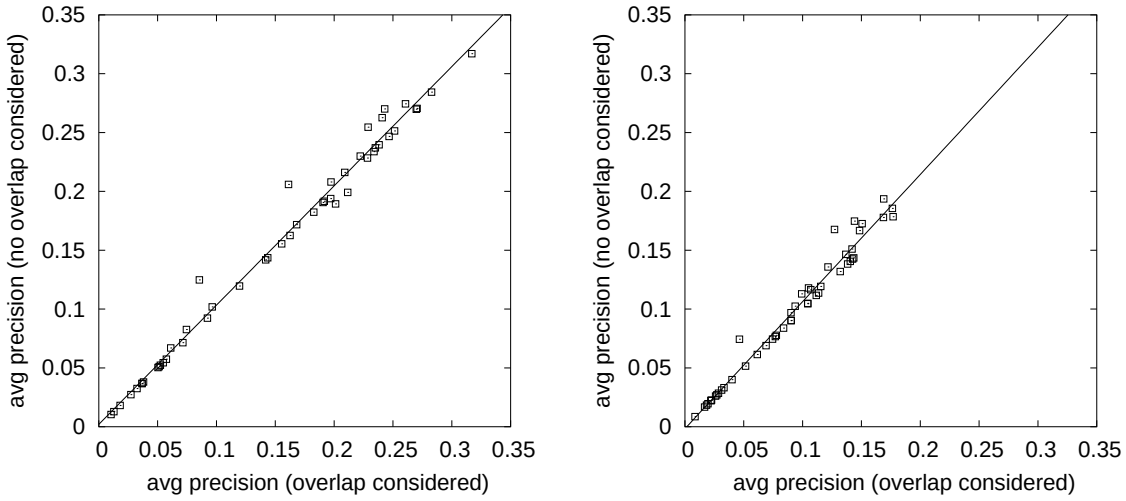
Here we use the definition of relevance from Section 4.2

For computing precision with respect to component size, the distinction between text and content must be taken into account; using the coverage definition from Section 4.2, we get

$$\text{precision}_s = \frac{\sum_{i=1}^{k} \frac{\left|t \cap c_i^U\right|}{\left|c_i^U\right|} \cdot \left|c_i^T\right|}{\sum_{i=1}^{k} \left|c_i^T\right|} = \frac{\sum_{i=1}^{k} cov\left(c_i^U\right) \cdot \left|c_i^T\right|}{\sum_{i=1}^{k} \left|c_i^T\right|} \quad (7)$$

To also take into account the overlap, regard 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 that content is distributed evenly within the component (ignoring e. g. 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. So, recall, which considers both component size and overlap, can be computed as

$$\text{recall}_o = \frac{\sum_{i=1}^{k} rel\left(c_i^U\right) \cdot \frac{\left|c_i^T - \bigcup_{j=1}^{i-1} c_j^T\right|}{\left|c_i^T\right|}}{\sum_{i=1}^{N} rel\left(d_i^U\right)} \quad (8)$$



a) strict quantisation b) generalised quantisation

Figure 2: Scatter plot and regression line for average precision of all official INEX 2002 submissions, using the recall$_o$ / precision$_o$ metric (which considers component overlap) and th erecall$_s$ / precision$_s$ metric (no overlap considered).

In a similar way, precision accounting for component size and overlap is derived as

$$\text{precision}_o \quad = \quad \frac{\sum_{i=1}^{k} cov\left(c_i^U\right) \cdot \left| c_i^T - \bigcup_{j=1}^{i-1} c_j^T \right|}{\sum_{i=1}^{k} \left| c_i^T - \bigcup_{j=1}^{i-1} c_j^T \right|} \tag{9}$$

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 relevance and coverage, the measures are equal to the standard definitions of precision and recall.

As defined here the two metrics recall$_s$ / precision$_s$ and recall$_o$ / 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] here.

6 Results

In order to discuss the evaluation results of retrieval runs that were submitted in INEX 2002, the metrics described in the previous Section 5 (in combination with the two quantisation functions proposed in Section 4.2) were applied to all official runs. We show that the different metrics (each representing a specific evaluation standpoint) affects the ranking among the runs submitted. Due to space limitations, we restrict our discussion to average precision (for 100 recall points) of the submitted runs.

For illustrating the impact of component overlap on evaluation results, Figure 2 compares the results of the recall$_o$ / precision$_o$ metric with those of the recall$_s$ / precision$_s$ metric. In this scatter plot, each point represents a run, and the coordinates are the values for the average precision with respect to the two metrics. For strict quantisation, the correlation between the two metrics is 0.993, and 0.987 for generalised quantisation; these high values show that consideration of overlap has only moderate effect on evaluation results.

The effect of the quantisation method is illustrated in Figure 3. With a value of 0.868, the correlation coefficient is lower than in the comparisons from above, indicating that there are substantial differences in the rankings. E. g., the top two runs change their position, and the run at the third rank for strict quantisation

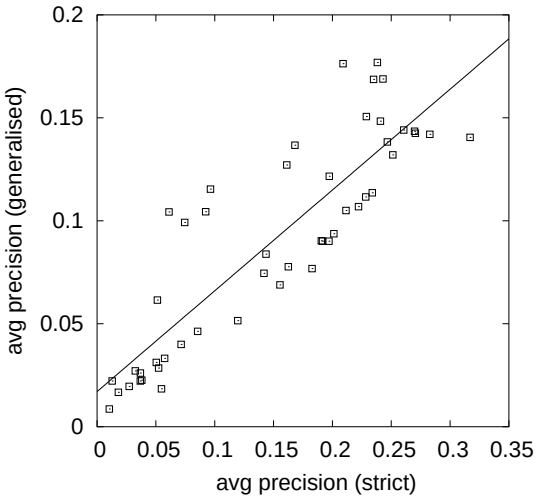


Figure 3: Scatter plot and regression line for average precision of all official INEX 2002 submissions, using the strict and generalised quantisation methods.

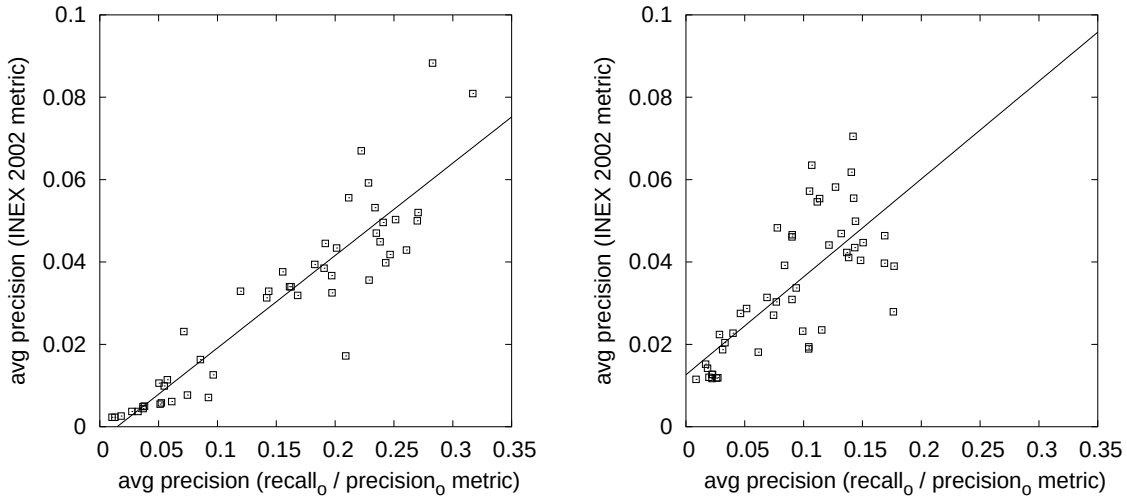
drops down to rank 9 for generalised quantisation. So quantisation is a factor that affects the relative quality of retrieval methods.

In INEX 2002 a rather heuristic approach has been taken for evaluation: The combined relevance / coverage values for a given document component have been quantised to a single relevance value. These relevance values have been used to apply standard recall / precision procedures like e. g. described in [Raghavan et al., 1989]. Again two different quantisation functions were used, comparable to the strict and generalised quantisation used for the new metric. A detailed description of the INEX 2002 metric can be found in [Gövert and Kazai, 2003], the respective official INEX 2002 results are described in [INEX, a].

We compare the results obtained from the INEX 2002 metric to those for our new recall$_o$ / precision$_o$ metric. For each submission, the average precision for 100 recall points is plotted for both metrics in Figure 4. Here the correlation coefficients are 0.917 for strict and 0.727 for generalised quantisation. For strict quantisation, the top 3 rankings are common for both metrics (although their positions are permuted), but e. g. the run at rank 4 in the INEX 2002 metric drops substantially with regard to the new metric. For generalised quantisation, the differences are even larger—except for the top position, the following 5 runs are completely disjoint for the two metrics. Given these variations in rankings, we can say that the consideration of component size and overlap in the new recall$_o$ / precision$_o$ metrics leads to more objective judgements about the quality of content-oriented XML retrieval approaches.

7 Conclusion

Evaluation of effectiveness of content-based retrieval of XML documents is a necessary requirement for the further improvement of research on XML retrieval. In this paper 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 the redefined concept of retrieval effectiveness. We adopted multiple degree scales for the measurements of both criteria of topical relevance and component coverage. A new metrics based on the well-established measures recall and precision has 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 submissions, we have shown that the choice of the metrics (and thus the underlying evaluation standpoints) has a



a) strict quantisation b) generalised quantisation

Figure 4: Scatter plots and regression lines for average precision of all official INEX 2002 submissions, using the INEX 2002 metric and the new recall$_o$ / precision$_o$ metric.

substantial effect on the relative quality of the different runs.

References

Ricardo Baeza-Yates, Norbert Fuhr, and Yoelle S. Maarek, editors. *Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval*, 2002.

Ricardo Baeza-Yates, Norbert Fuhr, Ron Sacks-Davis, and Ross Wilkinson, editors. *Proceedings of the SIGIR 2000 Workshop on XML and Information Retrieval*, 2000.

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

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

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

W. Bruce Croft, Alistair Moffat, Cornelis J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, 1998. ACM.

Norbert Fuhr, Norbert Gövert, Gabriella Kazai, and Mounia Lalmas. INEX: INitiative for the Evaluation of XML retrieval. In Baeza-Yates et al. [2002].

Norbert Fuhr, Norbert Gövert, Gabriella Kazai, and Mounia Lalmas, editors. *INitiative for the Evaluation of XML Retrieval (INEX). Proceedings of the First INEX Workshop. Dagstuhl, Germany, December 8–11, 2002*, ERCIM Workshop Proceedings, Sophia Antipolis, France, March 2003. ERCIM.

Norbert Gövert and Gabriella Kazai. Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In Fuhr et al. [2003], pages 1–17.

INEX. INEX 2002 evaluation results in detail. In Fuhr et al. [2003].

INEX. INEX guidelines for topic development. In Fuhr et al. [2003].

G. Kazai, M. Lalmas, and J. Reid. Construction of a test collection for the focussed retrieval of structured documents. In Fabrizio Sebastiani, editor, *25th European Conferenve on Information Retrieval Research (ECIR 2003)*, pages 88–103. Springer, 2003.

Jaana Kekäläinen and Kalvero Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13), September 2002.

C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors. *Evaluation of Cross-Language Information Retrieval Systems (CLEF 2001)*, volume 2406 of *Lecture Notes in Computer Science*, Berlin et al., 2002. Springer.

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

G. Salton, editor. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey, 1971.

G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

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

Tefko Saracevic and N.O. Pors. Relevance reconsidered. In *Proceedings of the 2nd International Conference on Conceptions of Library and Information Science*, pages 201–218, 1996.

K. Sparck Jones and C. J. van Rijsbergen. Report on the need for and provision of an “ideal” information retrieval test collection. Technical report, British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

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

trec_eval. Evaluation techniques and measures. In Voorhees and Harman [2002].

E. M. Voorhees and D. K. Harman, editors. *The Tenth Text REtrieval Conference (TREC 2001)*, Gaithersburg, MD, USA, 2002. NIST.

E.M. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. In Croft et al. [1998], pages 315–323.

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

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In Croft et al. [1998], pages 307–314.