A study of the assessment of relevance for the INEX'02 test collection

Gabriella Kazai, Sherezad Masood and Mounia Lalmas

Department of Computer Science, Queen Mary University of London
{gabs,shm1,mounia}@dcs.qmul.ac.uk

Abstract. We investigate possible assessment trends and inconsistencies within the collected relevance assessments of the INEX'02 test collection in order to provide a critical analysis of the employed relevance criterion and assessment procedure for the evaluation of content-oriented XML retrieval approaches.

1. Introduction

In information retrieval (IR) research, the evaluation of a retrieval system's performance typically focuses on assessing its retrieval effectiveness. Based on a traditional IR task, such as ad-hoc retrieval, and on a system-centred evaluation viewpoint, effectiveness provides a measure of a system's ability to retrieve as many relevant and as few non-relevant documents to the user's query as possible. Such an evaluation criterion relies on appropriate measures of relevance. In typical IR evaluation experiments, where the predominant approach to evaluate a system's retrieval effectiveness is with the use of test collections, the measure of relevance is provided by human judges. For example, the Text REtrieval Conference (TREC), which is one of the largest evaluation initiatives, employs and trains assessors who then follow guidelines that define relevance and detail the assessment procedure [1].

Traditional IR, however, mainly deals with flat text files. Due to the widespread use of the eXtensible Markup Language (XML), especially the increasing use of XML in scientific data repositories, Digital Libraries and on the Web, brought about an explosion in the development of XML tools, including systems to store and access XML content. The aim of such retrieval systems is to exploit the explicitly represented logical structure of documents, and retrieve document components, instead of whole documents, in response to a user query. Implementing this, more focused, retrieval paradigm means that an XML retrieval system needs not only to find relevant information in the XML documents, but also to determine the appropriate level of granularity to return to the user [14].

A fundamental consequence of the XML retrieval paradigm is that the relevance of a retrieved component is dependent on meeting both content and structural conditions. Evaluating the effectiveness of XML retrieval systems, hence, requires a test collection where the relevance assessments are provided according to a relevance criterion that takes into account the imposed structural aspects. A test collection as such has been built as a result of the first round of the INitiative for the Evaluation of

XML Retrieval (INEX)[1]. The initiative was set up at the beginning of 2002 with the aim to establish an infrastructure and provide means, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented retrieval of XML documents. In this paper, we make use of the constructed test collection, and in particular the collected relevance assessments. Our aim is to investigate the assessment of relevance for XML documents, explore possible assessment trends both with respect to individual and related components and examine the consistency and exhaustiveness of the assessments. Our study provides a critical analysis of the relevance criterion and assessment procedure employed in INEX'02, which serves as input to future INEX runs.

The paper is organized as follows. In section 2, we discuss the concept of relevance both in general and as defined in INEX'02. In Section 3, we describe the INEX test collection, and the methodology used to obtain the relevance assessments. Our investigation of the collected assessments consists of three parts. First, in Section 4, we examine the distribution of the relevance assessments. Second, in Section 5, we look at the assessments of related elements. Finally, in Section 6, we investigate the consistency and exhaustiveness of the assessments. The paper concludes in Section 7 with guidelines for future runs of INEX.

2. The concept of relevance in information and XML retrieval

Dictionaries define relevance as "pertinence to the matter at hand". In terms of IR, it is usually understood as the connection between a retrieved document and the user's query. With respect to the evaluation of IR systems, relevance plays a fundamental role as "relevance judgments form the bedrock on which the traditional experimental evaluation model is constructed" [2]. Due to its importance in IR theory, the concept of relevance has been the subject of numerous studies over the years. Despite it being a "primitive concept" that people understand intuitively [3], several interpretations of relevance, such as "aboutness" or "utility", have been explored in the past. Of these, a recent study found topicality as the most important criterion for relevance [4]. However, many researchers agree that relevance is a multidimensional cognitive concept whose meaning is dependent on the users' perceptions of information [5,6]. Several studies examine these dimensions [7], while others concentrate on defining various manifestations of relevance, e.g. algorithmic, situational, or motivational [8]. Relevance is also considered as a dynamic notion reflecting the finding that a user's perception of relevance may change over time. A relevance judgement, hence, is described as an assignment of a value of relevance by a judge or assessor at a certain point in time [9]. Furthermore, relevance is described as a multilevel phenomenon, according to which some documents may be more relevant than others [10,11,12].

Despite its many characteristics, relevance is considered a systematic and measurable concept when approached conceptually and operationally from the user's perspective [6]. It has also been successfully employed in IR evaluations as the standard criteria of evaluation, where research showed that, despite its dynamic nature, which can lead to large differences between relevance judges, the comparative

[1] http://qmir.dcs.qmul.ac.uk/inex/

evaluation of retrieval systems based on test collections provides reliable results when based on large number of topics [13].

In XML retrieval, the relationship between a retrieved item and the user's query is further complicated by the need to consider an additional dimension brought upon by the structural knowledge inherent in the documents and the possible structural conditions specified within the user's query. Given the need to accommodate both content and structural aspects, INEX defined a relevance criterion with the two dimensions of topical relevance and component coverage (both based on the topicality aspect of relevance) [14, pp184]. Topical relevance was defined as a measure of how exhaustively a document component discusses the topic of the user's request, while component coverage was defined as the extent to which a document component focuses on the topic of request (and not on other, irrelevant topics). Topical relevance adopted a four-point ordinal relevance scale based on the one proposed in [11]:

- Irrelevant (0): the document component does not contain any information about the topic of request.
- Marginally relevant (1): the document component refers to the topic of request but only in passing.
- Fairly relevant (2): the document component contains more information than the topic description, but this information is not exhaustive.
- Highly relevant (3): the document component discusses the topic of request exhaustively.

For component coverage, a nominal scale was defined:

- 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.

The combination of these two relevance dimensions was used to identify those relevant document components, which were both exhaustive and specific to the topic of request and hence represent the most appropriate unit to return to the user.

3. The INEX'02 test collection

The document collection of the INEX'02 test collection consists of the full texts of 12107 articles of the IEEE Computer Society's publications, from 1995 to 2002, totalling 494 megabytes [14, pp1]. On average, an article contains 1532 XML nodes (including attribute nodes), where the average depth of a node is 6.9. The overall structure of a typical article consists of a frontmatter (fm), a body (bdy) and a backmatter (bm). The frontmatter contains the article's metadata, such as title, author, publication information, and abstract. The body is structured into sections (sec) and sub-sections (ss1). These logical units start with a title, and contain a number of paragraphs (para[2]), tables, figures, lists, citations, etc. The backmatter includes bibliography and author information. The “collection” column of Table 1 shows the occurrence frequency of different element types in the collection.

In INEX’02, two types of topics were used: 1) content-only (CO), which are typical IR queries where no constraints are formulated with respect to the structure of the retrieval results, and 2) content-and-structure (CAS), which are XML queries that contain explicit references to the XML structure, i.e. specifications of target elements (e.g. what should be returned to the user) and containment conditions (e.g. element types that should be about a given concept).

During the retrieval sessions participating groups produced a ranked list of XML elements in answer to a topic. The top 100 result elements from all 60 sets of ranked lists (one per topic) formed the results of one retrieval run. A total of 51 runs from 25 groups were submitted. For each of the 60 topics, the results from the submissions were merged to form the pool for assessment (see Table 1, “Result pool” columns).

The result pools were then assigned for assessment either to the original topic authors or, when this was not possible, on a voluntary basis, to groups with expertise in the topic’s subject area. The assessments were done along the two dimensions of topical relevance and component coverage. Assessors were asked to judge each and every relevant document component by following a two-step process [14, pp184]. During the first step, assessors were required to skim-read the whole article that contained a result element and identify any relevant information within. In the second step, assessors had to judge the relevance of the found relevant components and of their ascendant and descendant elements. Assessors were allowed to stop assessing ascendant elements once a container component was judged as too large. Similarly, descendant elements only needed to be judged until an irrelevant component or a component with too small coverage was reached. To lessen the workload, the system implicitly regarded any non-assessed elements as irrelevant. For CAS topics with target elements, the procedure was modified stating that any elements other than the target elements had to be considered irrelevant.

Assessments were recorded using an on-line assessment system, which allowed judges to view the result pool of a given topic (listing result elements in alphabetical order), browse the document collection and view articles and result elements both in XML (i.e. showing the tags) and document view (i.e. formatted for ease of reading). Other features included keyword highlighting and consistency checking.

Assessments were collected for 55 (30 CAS and 25 CO) of the 60 topics, for a total of 48849 files containing 71086 elements, of which 22719 are at article level. The last three columns of Table 1 show a breakdown of the collected assessments by element type for both topic types. Note that these statistics only include explicit assessments, i.e. implicitly irrelevant elements are not considered. In addition, the assessments of 8246 article files for CAS and 10717 for CO are also excluded. The reason for this is that these files were assessed using a “quick assess” option of the

[2] Paragraphs are elements of the “para” entity as defined in the document collection’s DTD: <!ENTITY % para "ilrj|ip1|ip2|ip3|ip4|ip5|item-none|p|p1|p2|p3">.

on-line assessment system, which allowed judges to skip the explicit assessment of result elements in an article and only mark the article file as irrelevant.

Table 1. Number of element types in collection, result pool and assessments

Component	**Collection**	**Result pool**			**Assessed**		
		CAS	CO	Total	CAS	CO	Total
Article files	12107	23375	30275	53650	24237	24612	48849
All elements	**8239997**	**47419**	**60066**	**107485**	**34130**	**36956**	**71086**
article	12107	12418	22630	35048	10379	12340	22719
bdy	12107	1215	4133	5348	526	2231	2757
sec	69735	4182	7329	11511	2004	3999	6003
ss1	61492	726	1313	2039	492	1425	1917
para	983737	5349	10133	15482	3645	7505	11150
d.o.[3] para	2835975	1723	1828	3551	2336	1227	3563
fm	12107	4439	2639	7078	1325	862	2187
d.o. fm	383575	7188	1472	8660	4270	1041	5311
bm	10065	268	463	731	361	501	862
d.o. bm	2483446	8175	6066	14241	6460	3632	10092

From Table 1, we can obtain that for CAS 0.58% and for CO 0.73% of all XML elements in the collection were included in the result pools. The difference between the sizes of the CAS and CO result pools is due to the fact that more runs were submitted for CO topics, where the runs also contained on average more results. On average, 2.0 elements were retrieved from an article file, 68% of which were related by way of ascendant, descendant or sibling relations. Looking at the distribution of the elements in the result pools according to their element types, we can see that article (32.61%), section (10.71%), paragraph (12.07%) and bibliography sub-components (13.25%) were the most frequently returned elements. Another trend that can be observed is that the result pools of CAS topics consisted of approximately equal numbers of "small" (i.e. para, sub-components of para, fm and bm) and "large" (i.e. article, sec, ss1, fm and bm) components, while for CO topics this ratio is around 35% and 65%, respectively. From the result pools, 72% of the results were explicitly assessed for CAS and 61.5% for CO. Note that this only indicates that assessors more often used the "quick assess" option with CO topics (all results were assessed and even some additional article files that were not included in the result pool).

4. Investigation of the INEX'02 relevance assessments

In this section we investigate the distribution of the collected assessments at different levels of granularity in order to derive conclusions of possible assessment trends.

4.1 Distribution of topical relevance

We first look at the distribution of assessments for the topical relevance dimension (Table 2). In general, for both topic types a large proportion of the results were judged

[3] d.o. = descendant elements of …

irrelevant (71% for CAS and 48% for CO) and only a small portion were perceived as highly relevant (9% and 8%). This is not surprising and it correlates with findings of previous studies [15]. What is interesting, however, is the relatively high percentage of irrelevant CAS assessments compared with CO. The reason for this lies in the definition of the CAS relevance criterion, which explicitly states that any components other than target elements must be treated as irrelevant. Since a number of CAS submissions were produced by systems that did not support XML style queries, a high percentage of the CAS result elements were hence irrelevant, non-target elements.

With respect to the highly relevant elements, although their numbers are the same for CAS and CO, they represent a very different distribution, making up for CAS 32% and for CO 16% of the total number of their respective relevant elements. These ratios are also magnitudes higher in INEX than in flat text test collections [15]. This is mainly due to the structure of the collection and the definition of the relevance criterion. As mentioned earlier, assessors in INEX had to assess both ascendant and descendant elements of a relevant component. Furthermore, due to the definition of topical relevance, we know that the relevance degree of an ascendant node is always equal to or greater than the relevance degree of its sub-nodes. This means that any elements assessed as highly relevant will have highly relevant ancestor nodes. This propagation effect of topical relevance, combined with its cumulative nature, provides an answer to the increased level of highly relevant elements for CO topics.

For CAS topics, however, no such relation between the relevance degrees of related components exists (as only target elements are relevant). Furthermore, looking at the ratio of marginally and fairly relevant CAS assessments, we can see that, while the proportion of marginally relevant components is the same as for CO topics (41%), the relatively high ratio of highly relevant elements is complemented with a low percentage of fairly relevant elements (27%). A plausible reason, given the question-answering nature of most CAS topics and that 47% of CAS result elements were small components, is that the degree of fairly relevant was less discernable (or less meaningful) in this context. By definition, an element should be assessed fairly relevant if it contains more information than the topic description, but this information is not exhaustive. It is not clear, however, when the content of, for example, an author element would match such a criterion. It is hence more likely that in these cases assessors assigned either marginally or highly relevant degrees. To further investigate this issue, Table 3 shows the averaged relevance distributions for the different categories of CAS topics: those that do not specify target elements; those where the requested result elements are of factual type (e.g. author, title, bibliographic entry); or content-oriented (e.g. article, sec, para). As it can be seen, the distribution of the assessments for CAS topics with no target element closely follows the distribution of CO assessments, while the assessments of CAS topics with target elements demonstrates a very different behaviour. For factual result elements, we find that a dominant percentage were assessed as marginally relevant (52%), while for content-oriented results the ratio of highly relevant assessments is the dominant (51.1%). A reason for the former finding is the high percentage of small result components (e.g. title), whose “exhaustiveness”-level was assessed to satisfy only the minimum criteria of topical relevance. Although no definitive reasoning can be given for the latter finding, we suspect that judges were partly influenced by a match regarding the target element specification (i.e. were partial to judge an element highly relevant if it matched the target element). This is further supported when looking at the breakdown of the assessments for the different element types in Table 2. Looking at the CAS columns, we can see that the majority of highly relevant components (57.7%) consist of articles and sub-components of fm and bm, which also represent over 60% of the target element types of CAS topics.

Table 2. Distribution of topical relevance assessments for CAS and CO topics

	CAS					CO				
rel =	3	2	1	0	Total	3	2	1	0	Total
Total	3102 **9**%	2703 **8**%	3980 **12**%	24345 **71**%	34130 100%	3026 **8**%	8191 **22**%	7844 **21**%	17895 **48**%	36956 100%
Rel. only	**32**%	**27**%	**41**%	-	9785	**16**%	**43**%	**41**%	-	19061
article (%)	**11.6**	8.0	9.3	**38.7**	30.4	**23.2**	10.8	13.8	**54.0**	33.4
bdy (%)	3.5	3.7	2.1	1.0	1.5	14.6	6.9	7.5	3.5	6.0
sec (%)	9.1	12.8	8.2	4.3	5.9	**19.8**	**15.7**	12.6	6.3	10.8
ss1 (%)	3.2	4.4	1.7	0.9	1.4	6.8	5.9	5.4	1.8	3.9
para (%)	8.7	**22.4**	**14.5**	9.0	10.7	**18.7**	**30.7**	**28.0**	12.4	20.3
d.o. para (%)	2.0	5.3	9.4	7.2	6.8	0.4	3.3	3.6	3.7	3.3
fm (%)	1.1	2.2	2.0	4.7	3.9	1.6	1.8	1.5	3.1	2.3
d.o. fm (%)	**28.2**	11.4	7.5	11.5	12.5	3.1	3.5	2.9	2.4	2.8
bm (%)	0.9	1.2	0.8	1.1	1.1	1.5	1.8	2.4	0.7	1.4
d.o. bm (%)	**17.9**	**23.5**	**36.3**	**15.7**	18.9	5.1	12.8	**15.3**	6.9	9.8

Table 3. Distribution of topical relevance assessments for CAS with different types of target elements

rel=	3	2	1
Factual target element (%)	27.9	20.1	**52.0**
Content target element (%)	**51.1**	20.3	28.6
No target element (%)	22.9	40.0	37.1

In contrast, for CO topics, the majority of highly relevant elements are articles, sections and paragraphs (61.7%), and fairly and marginally relevant elements are mostly paragraphs, sub-components of bm, and sections (59.2% and 55.9%). At first glance this would suggest a clear preference for larger components for CO topics. This is, however, not the case, but these findings simply show the propagation effect of topical relevance and confirm its cumulative nature.

4.2 Distribution of component coverage

Table 4 summarises the distribution of assessments with respect to the component coverage dimension. Looking at the totals, a noticeable difference is the relatively high ratio of exact coverage (16%) and the relatively low ratio of too large (7%) and too small (6%) assessments for CAS topics, compared with CO (10%, 22% and 20%, respectively). Looking at the distribution of relevant elements only, we can observe a very high ratio of exact coverage for CAS (57%) compared with CO (19%). A reason for this is that possibly relevant, but non-target elements, which may otherwise had been assessed as too large or too small, were judged irrelevant due to the strict CAS relevance criterion. However, we suspect that another reason for the domination of exact coverage assessments is that assessors incorrectly assigned exact coverage to elements that matched the target element of the topic (instead of assessing components according to the extent to which they focus on the topic of the request). This concern was also verbally confirmed by some of the assessors at the INEX’02 workshop. The root of the problem is that the “too small” and “too large” coverage categories were incorrectly interpreted as the relation between the actual size of the result component and the size of the target element (instead of the relation between the relevant and irrelevant contents of the component). Further evidence of this can be seen in Table 5, which shows the averaged distribution of coverage assessments for CAS topics with and without target elements. Although for factual result elements it is reasonable to expect higher levels of exact coverage assessments, the distribution for content-oriented results is expected to closer follow the distribution of CAS topics with no target element. This is because while it is plausible that an author element, for example, contains only relevant information regarding a query asking for names of experts in a given field, it is less likely that target elements, such as sections or articles, contain no irrelevant information to the topic of the request.

Table 4. Distribution of component coverage assessments for CAS and CO topics

		CAS					CO				
	cov=	E	L	S	N	Total	E	L	S	N	Total
Total		5530 **16%**	2214 **7%**	2041 **6%**	24345 **71%**	34130 100%	3611 **10%**	8219 **22%**	7231 **20%**	17895 **48%**	36956 100%
Rel. only		**57%**	**22%**	**21%**	-	9785	**19%**	**43%**	**38%**	-	19061
article	(%)	5.9	**24.7**	3.7	**38.7**	30.4	**16.3**	**22.5**	3.2	**54.0**	33.4
bdy	(%)	0.5	11.4	0.6	1.0	1.5	4.6	**16.5**	1.0	3.5	6.0
sec	(%)	6.6	**22.3**	4.7	4.3	5.9	**16.7**	**18.3**	10.6	6.3	10.8
ss1	(%)	1.7	6.2	2.4	0.9	1.4	9.0	6.3	3.7	1.8	3.9
para	(%)	11.3	7.8	**32.2**	9.0	10.7	**35.5**	9.8	**44.2**	12.4	20.3
d.o. para	(%)	3.1	0.1	**20.0**	7.2	6.8	1.6	0.8	6.2	3.7	3.3
fm	(%)	0.5	3.5	3.3	4.7	3.9	0.8	2.4	1.3	3.1	2.3
d.o. fm	(%)	**22.4**	5.2	6.3	11.5	12.5	5.0	2.1	3.4	2.4	2.8
bm	(%)	0	3.5	0.7	1.1	1.1	0.2	4.1	0.5	0.7	1.4
d.o. bm	(%)	**40.1**	10.2	9.5	**15.7**	18.9	9.3	10.1	**17.2**	6.9	9.8

Table 5. Distribution of component coverage assessments for CAS with different types of target elements

	rel=	E	L	S
Factual target element	(%)	**76.4**	9.7	13.9
Content target element	(%)	**56.0**	21.8	22.2
No target element	(%)	27.1	39.7	33.2

The distribution of the coverage assessments according to element types (Table 4) shows that for CAS topics the most common elements with exact coverage (73.8%) were paragraphs and sub-components of bm and fm, while articles and sections were mostly judged as too large (47%). For CO topics, we can observe a clear preference towards paragraphs being judged with exact coverage (35.5%), while a large proportion of articles and sections were assessed as too large (40.8%). Although this finding may not be so surprising, it clearly demonstrates that judges preferred more specific elements for both topic types. It also correlates with expectations that purely relevant information is likely to be contained in smaller units and that larger nodes, such as articles, are more likely to include other, irrelevant information. Combined with the relevance distribution data ("Rel. only" row), these results suggest that for CO topics, assessors were able to interpret the coverage criterion correctly and assess components accordingly. Looking at the too small assessments, we find that the majority of element types for CAS are paragraphs and sub-components of paragraphs (52.2%). An interesting observation here is that although a high percentage of paragraphs were found to be too small, their container components, such as sec, bdy and article elements were generally found too large, leaving no elements in between with exact coverage. A reason for this is that container elements, which may otherwise had exact coverage of the topic, were implicitly regarded as irrelevant if they did not match the target element. For CO topics, the distribution of too small assessments is more clear-cut. Here, small elements, e.g. paragraphs and sub-components of bm make up 61.4% of the too small assessments. It should be noted that here the large proportion of too small assessments is complemented with high proportions of exact coverage article, sec and ss1 elements (42%).

4.3 Distribution of the combined relevance assessments

In this section we investigate the correlation between the relevance dimensions. Although the two dimensions are not completely independent (i.e. combinations like 0E, 0L, 0S, 1N, 2N and 3N are not permitted and assessments of 3S are not reasonable), strong correlations in the following analysis would indicate the presence of common factors that influenced the assessment of both dimensions.

Table 6 shows the distribution and correlation of the combined assessments for CAS and CO topics. For each possible combination of topical relevance (columns), *rel*, and component coverage (rows), *cov*, the number of assessed components is shown in the "Total" rows (the percentage values reflect the ratio of these elements to all relevant CAS and CO components, respectively). The "cov|rel correlation" rows indicate the correlation of coverage categories with given relevance degrees, where the percentage values represent the likelihood of an element being assessed with coverage level *cov*, given that it has a relevance degree *rel*. Similarly, the "rel|cov correlation" rows indicate the correlation of topical relevance degrees with given coverage levels, where the percentage values represent the likelihood of an element being assessed with relevance degree *rel*, given that it has a coverage level *cov*. As it can be seen, for CAS topics there is a very high likelihood of highly relevant elements being assessed with exact coverage (80.3%). This implies that the assessment of both dimensions was influenced by a common aspect. As we saw in the previous sections, this influencing factor was whether the result element matched the target element of the query. A similar, but less dominating, tendency can be observed for fairly and marginally relevant components (43.9% and 46.5%). Looking at the correlation of coverage given topical relevance, it appears more likely that an element with exact coverage would be judged highly relevant (45%) than fairly (21.5%) or marginally relevant (33.5%). For too large coverage the dominant relevance degree is marginally relevant (43.1%), while for too small coverage an element is almost equally likely to be assessed as fairly or marginally relevant.

For CO topics, no significant correlations are visible. Highly relevant components are just as likely to be assessed to have exact coverage (46.1%) as being judged too large (53.9%). Fairly and marginally relevant components are slightly more likely to be assessed as too small (48.2% and 41.8%) or too large (36.2% and 46.2%) than with exact coverage (15.6% and 12%). These tendencies, however, are mainly due to reasonable correlations between the relevance dimensions and component size. For example, small elements are less likely to be exhaustive or act as meaningful units of information, while large components are more likely to be exhaustive, but also likely to cover multiple topics. The same can be said when looking at the correlation of component coverage given topical relevance.

Table 6. Distribution and correlation of combined assessments for CAS and CO topics

rel:		3		2		1	
cov:		CAS	CO	CAS	CO	CAS	CO
E	Total	2491(25%)	1396(7%)	1187(12%)	1274(7%)	1852(19%)	941(5%)
	cov|rel %	**80.3**	46.1	**43.9**	15.6	**46.5**	12.0
	rel|cov %	**45.0**	38.6	21.5	35.3	33.5	26.1
L	Total	611 (6%)	1630 (9%)	648(7%)	2963(16%)	955(10%)	3626(19%)
	cov|rel %	19.7	53.9	24.0	36.2	24.0	**46.2**
	rel|cov %	27.6	19.8	29.3	36.1	**43.1**	44.1
S	Total	-	-	868(9%)	3954(21%)	1173(12%)	3277(17%)
	cov|rel %	-	-	32.1	**48.2**	29.5	41.8
	rel|cov %	-	-	42.5	54.7	57.5	45.3

5. Investigating the assessment of related components

Since document components returned by an XML retrieval system may be related we cannot regard their relevance as independent from one another. In this section we examine the relevance values assigned to related components, such as parent, child, sibling, ancestor and descendant nodes. Our aim is to discover possible correlations and identify factors, if any, that influence the assessment of related components.

5.1 Topical relevance

Table 7 lists the occurrence frequency of assessment pairs for each possible combination of element and related element assessments of topical relevance (e.g. 0,0; 0,1; 0,2; 0,3; etc.) for each relation type (element,parent; element,child; etc) and for both CO and CAS topic types.

We first look at the results for CO topics. Due to the definition of topical relevance, we expect to find certain regularities in the assessed relationships, such as the equivalence or increase of the assigned relevance degree for parent and ascending nodes. This is clearly visible, apart from some noise due to assessment mistakes (see section 6), at all levels of topical relevance. In addition, we can observe a tendency to assign equal, instead of higher, degrees of relevance (approx. 60% to 40%) to parent/ancestor nodes. This tendency is particularly strong for the element-parent relations and for the assessment pairs 1,1 (73.09%) and 2,2 (80.43%). An explanation for this is the propagation effect of topical relevance. Other factors include situations where the available relevance degrees are insufficient in reflecting a change in the amount of relevant information contained by a parent/ancestor node. On the other hand, when the relevance level of parent or ascendant nodes is increased, the increase tends to be as minimal as possible, e.g. it is more likely that the parent of a marginally relevant node is assessed as fairly relevant (1,2: 21.53%) rather than highly relevant (1,3: 5.34%). More or less the same observations apply, as expected, to the element-child and element-descendant relations with only slight differences in the actual percentage values. Looking at the element-sibling relations, we are not able to find such clear patterns in the assessments. The trend of assigning related components the same relevance degree seems to hold, but is less prominent (e.g. 3,3: is only 48.31%). Also, although there is still evidence that the relevance level of sibling nodes is more likely to differ from the relevance of a given node by only one degree, this trend does not hold for highly relevant elements (i.e. 3,0: 12.92%; 3,1: 9.18%).

Table 7. Topical relevance assessments of related components for CAS and CO topics

Relation=	parent		child		sibling		ascendant		descendant	
(%)	CAS	CO	CAS	CO	CAS	CO	CAS	CO	CAS	CO
0,0	68.52	46.18	94.48	99.87	93.87	68.07	67.58	38.90	94.90	99.83
0,1	7.34	26.23	1.56	0.13	2.24	19.49	4.40	24.47	2.34	0.17
0,2	11.02	18.21	0.73	0	2.24	9.35	11.23	19.47	1.52	0
0,3	13.12	9.38	3.24	0	1.65	3.10	16.80	17.16	1.25	0
Subtotal	65.30	23.51	47.36	10.87	69.16	24.44	71.28	24.69	50.76	9.62
1,0	5.63	0.05	40.42	21.05	10.13	16.76	9.85	0.05	40.05	24.31
1,1	53.52	**73.09**	59.08	78.70	73.26	62.47	37.84	58.69	58.26	75.32
1,2	31.24	**21.53**	0.42	0.23	14.62	18.88	32.11	26.69	1.36	0.35
1,3	9.60	**5.34**	0.07	0.02	1.98	1.89	20.20	15.01	0.33	0.02
Subtotal	13.09	31.53	11.85	29.29	15.31	28.42	12.05	31.90	7.82	24.86
2,0	2.69	0	33.69	11.07	13.37	5.54	7.21	0	**42.58**	13.25
2,1	0.39	0.20	19.14	17.56	19.29	13.00	0.99	0.25	20.57	23.08
2,2	78.62	**80.43**	47.02	71.30	63.48	77.27	64.24	66.10	36.56	63.57
2,3	18.30	19.37	0.16	0.07	3.86	4.20	27.57	33.65	0.29	0.11
Subtotal	12.78	34.28	21.37	38.67	11.60	41.28	10.70	34.89	18.80	36.28
3,0	17.36	0	**44.11**	7.95	29.01	**12.92**	10.62	0	**52.94**	14.49
3,1	0.09	0.04	6.47	10.42	7.72	**9.18**	0.43	0.06	10.76	16.37
3,2	0.38	0.27	12.04	31.36	11.39	29.59	0.91	0.45	13.04	29.00
3,3	82.16	99.69	37.38	50.27	51.89	**48.31**	88.04	99.49	23.26	40.14
Subtotal	8.83	10.68	19.42	21.17	3.93	5.86	5.97	8.52	22.62	29.24

Looking at the results for CAS topics, none of the above trends can be observed clearly. For example, although the tendency to assign the same relevance degree to related components is still present, there are several cases where this pattern does not apply (e.g. element-child of 3,0: 44.11%; element-descendant of 2,0: 42.58% and 3,0: 52.94%,). The pattern that the relevance degree of related nodes usually only differs by one degree does not apply at all, but the assessment pairs appear more random, although there is a noticeable increase in the 1,0; 2,0 and 3,0 assessment pairs for all relation types. This is again related to the strict relevance criterion for CAS topics.

5.2 Component coverage

Table 8 shows the occurrence frequency of assessment pairs for each possible combination of element and related element assessments of component coverage (e.g. N,N; N,L; N,S; etc.) for each relation type and for both topic types.

Several general trends can be observed for CO topics. The most obvious perhaps is that 90.62% of parent components of too large elements are also assessed as too large. Although this is largely expected, it cannot be applied as a general rule since the ratio of relevant information contained in a parent node is also dependent on the sibling elements' contents. This is reflected in the finding that 6.39% of parent nodes of too large elements have been judged to have exact coverage. The fact that the coverage of ascendant nodes of too large components cannot be inferred with 100% accuracy highlights a false assumption in the assessment procedure. According to the instructions, assessors were allowed to stop the assessment of ascendant nodes once a too large component was reached, assuming that all ascendants would also then be too large. The same applies regarding the stopping rule of assessing descendant elements. Although most child nodes of too small elements are also assessed as too small (84.48%), this is not a certainty, e.g. child nodes may also be irrelevant (8.31%) or too large (6.77%). Other assessment patterns in Table 8 include the high correlation of sibling nodes assessed with the same coverage degree (N,N 68.07%, L,L 55.38%, S,S 86.57% and E,E 49.24%).

Similarly to CO, for CAS topics the parent nodes of too large elements are also assessed as too large (91.47%). However, these assessment pairs comprise only 9.78% of all CAS assessment pairs (i.e. CAS topics with no target element). Another pattern is the relative increase in the number of L,N; S,N and E,N element-child and element-descendant relations due to the already mentioned strict relevance criterion.

Table 8. Component coverage assessments of related components for CAS and CO topics

Relation=	parent		child		sibling		ancestor		descendant	
(%)	CAS	CO	CAS	CO	CAS	CO	CAS	CO	CAS	CO
N,N	68.52	46.18	94.48	99.87	93.87	**68.07**	67.58	38.90	94.90	99.83
N,L	15.75	40.71	1.17	0.13	1.00	8.51	21.08	50.32	1.28	0.13
N,S	3.81	4.63	0.46	0	2.10	15.58	1.59	2.30	1.73	0.03
N,E	11.92	8.48	3.89	0	3.03	7.85	9.75	8.48	2.09	0
Subtotal	65.66	23.51	47.36	10.87	69.16	24.44	71.28	24.69	50.76	9.62
L,N	0	0	**37.59**	17.85	35.10	18.41	10.23	0.06	**47.43**	19.77
L,L	**91.47**	**90.62**	32.54	50.17	43.10	**55.38**	81.39	86.96	16.34	31.05
L,S	3.19	2.99	11.37	16.01	4.92	12.84	3.78	3.66	19.52	32.01
L,E	5.34	6.39	18.50	15.97	16.87	13.37	4.60	9.32	16.71	17.16
Subtotal	**9.78**	29.70	27.35	53.63	1.98	11.30	6.36	22.44	31.68	62.84
S,N	1.65	0	**48.61**	**8.31**	16.59	7.09	6.57	0.01	**43.74**	6.85
S,L	23.51	26.11	6.06	**6.77**	1.11	2.70	46.30	51.76	9.26	9.92
S,S	17.36	33.65	44.84	**84.48**	75.48	**86.57**	8.70	17.66	44.74	82.81
S,E	57.48	40.24	0.49	0.44	6.82	3.63	38.43	30.57	2.27	0.43
Subtotal	13.30	32.89	5.12	13.10	8.75	53.67	13.36	38.86	2.60	8.29
E,N	16.47	0	**38.60**	8.90	10.42	18.10	11.81	0	**46.45**	10.88
E,L	45.21	61.61	2.58	8.47	1.66	14.26	58.80	76.98	1.95	10.86
E,S	0.22	0.41	37.68	59.06	2.97	18.39	0.65	0.25	34.31	61.70
E,E	38.10	37.98	21.15	23.56	84.95	**49.24**	28.73	22.77	17.29	16.57
Subtotal	11.26	13.90	20.17	22.40	20.12	10.59	9.00	14.01	14.96	19.26

6. Exhaustiveness and consistency of the assessments

The assessment procedure stated that for topics without target elements all ascendants and descendants of each relevant component should be assessed until a too large ascendant or a too small or irrelevant descendant element is reached. The assessment of too large nodes was then propagated to ascendants, while all other non-assessed elements were implicitly regarded as irrelevant. Due to these implicit rules, we have only limited means by which to estimate the exhaustiveness of the assessments. For topics with target elements, we have no way of checking if all necessary elements have been assessed due to the strict relevance constraint and the implicitly irrelevant assumption. For the remaining topics, we calculated:

- The number of relevant elements that only have irrelevant descendants: we found only 15 elements (8 2S and 7 1E);
- The number of relevant elements with exact or too small coverage that do not have an ancestor with too large coverage: we obtained 29 highly relevant, 62 fairly relevant and 86 marginally relevant elements;
- The number of elements whose siblings have not been assessed, but whose parent node has a higher level of relevance: there are 229 fairly relevant and 501 marginally relevant elements.

These figures appear very comforting, however with no exact way to confirm other possibly missing assessments we have to consider these only as very rough estimates of assessment exhaustiveness. In addition, circumstantial evidence, in the form of lack of overlap within the submission results (see Table 1), suggests that further relevant components may not have been retrieved and hence assessed.

Table 9. Inconsistent assessment pairs for CO topics

	0,1	0,2	0,3	1,0	1,2	1,3	2,0	2,1	2,3	3,0	3,1	3,2
parent	-	-	-	3	-	-	-	14	-	-	1	6
child	3	-	-	-	14	1	-	-	6	-	-	-
ancestor	-	-	-	10	-	-	-	54	-	-	3	24
descendant	10	-	-	-	54	3	-	-	24	-	-	-

Next, we examine the consistency of the collected assessments. Since the on-line assessment tool already included consistency checking for the individual assessments (such as checking for invalid assessments, e.g. 0E, 0L, 0S, 1N, 2N, 3N and 3S), here we concentrate on the consistency of related assessments. We again deal only with topics that do not specify target elements. In this context, we consider an assessment pair inconsistent when an ancestor of a relevant node is assessed less relevant or irrelevant; or vice-versa when a descendant is assessed as more relevant. Table 9 shows the total number of inconsistencies found. Note that here we only show data based on 4 topics that were assessed before a version of the on-line assessment tool included rules to test the consistency of related assessments. Although our observations here are not statistically significant, we can still conclude that the number of inconsistencies (91) compared with the total number of assessed components (7340, including 3327 relevant) is largely insignificant. As it can be seen, most of the inconsistencies occur as a result of ancestors of fairly relevant node’s being assessed only as marginally relevant (59%), where of the 54 cases 14 are direct element-parent relations. In general, 97% of the inconsistencies are assessments where the ancestor node of the element is judged one degree less relevant than the node itself. A possible reason for the inconsistencies is that in the XML view of the on-line assessment tool large articles required a lot of scrolling up and down to assess parent and ancestor nodes, where assessors could have easily missed some relevance values assigned to sub-components.

7. Conclusions

In this paper we provided an analysis of the relevance assessments of the INEX'02 test collection; examined their distributions over the two dimensions at different levels of granularity; explored the assessments of related components; and investigated assessment consistency and exhaustiveness.

Our findings, in general, showed no surprises, but confirmed expected effects of the relevance criterion (such as the strict criterion for CAS topics, the cumulative nature of topical relevance for CO topics and its propagation effect to ancestor nodes). We found that the combination of the two dimensions worked well for CO topics allowing assessors to identify elements both exhaustive and specific to the topic of request. For topical relevance, the number of relevance degrees was found suitable, although we also found indications that both less and more degrees would have been useful in different situations. An issue regarding the degree of fairly relevant was highlighted by the finding that it appeared less measurable for small factual results. This is because its definition is based on the notion of exhaustivity, which is more suited for content-oriented, and in particular multifaceted, topics. In general, however, the dimension of topical relevance was found appropriate for XML evaluation, even though it does not consider aspects of usefulness or novelty (which are important factors given that retrieved elements may be nested components). Our main criticism regards the coverage dimension and the assessment of CAS topics with target elements. We found evidence to show that a match on a target element type influenced the assessment of both relevance dimensions, and especially the dimension of component coverage. Furthermore, we reported on the concern that the categories of too large and too small were also misinterpreted when target elements were assessed. A general issue regarding the too small coverage was that it combined criteria regarding both topicality and unit size, which actually encouraged its misinterpretation. This issue has since been addressed in INEX'03[4], where the component coverage dimension was redefined to avoid direct association with component size. A solution regarding the assessment of coverage of target elements, where assessors are instructed to ignore the structural constraints within the query, is also being tested in INEX'03.

Regarding the consistency of the assessments we found no reason to warrant concern, even without the consistency checking tools of the on-line assessment system. On the other hand, there is cause for concern regarding the exhaustiveness of the assessments especially as the overlap between retrieval submissions is low. This is

[4] http://inex.is.informatik.uni-duisburg.de:2003/

addressed in INEX’03 by means of rigorous checks being implemented in the assessment tool, where assessors are required to assess all relevant ascendant and descendant components.

Regarding the future of XML evaluation, it is clear that as XML is becoming more popular, the need for test collections also increases. The study of the construction of such test collections, however, is still in its infancy. The first year of the INEX initiative developed and tested a new criterion for relevance based on two dimensions. Our findings in this paper highlighted some issues and concerns regarding the definition of both this criterion and the assessment procedure. Possible solutions to these problems are currently put to the test in INEX’03.

References

[1] Voorhees, E.M. and Harman, D.K., eds. (2002): The tenth Text Retrieval Conference (TREC-2001), Gaithersburg, MD, USA, 2002. NIST.

[2] Harter, S.P. (1996): Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness. Journal of the American Society for Information Science, 47(1):37-47.

[3] Saracevic, T.: www.scils.rutgers.edu/~tefko/Courses/610/Lectures/.

[4] Vakkari, P. and Hakala, N. (2000): Changes in Relevance Criteria and Problem Stages in Task Performance. Journal of Documentation. 56(5): 540-562.

[5] Schamber, L. (1994): Relevance and Information Behaviour. Annual Review of Information Science and Technology (ARIST), 29:3-48.

[6] Borlund, P. (2003): The Concept of Relevance in IR. Journal of the American Society for Information Science, 54(10):913-925.

[7] Cosijn, E. and Ingwersen, P. (2000): Dimensions of Relevance. Information Processing and Management, 36:533-550.

[8] Saracevic, T. (1996): Relevance Reconsidered. Proceedings of the 2nd International Conference on Conceptions of Library and Information Science (COLIS), Copenhagen, pp. 201-218.

[9] Mizzaro, S. (1997): Relevance: the whole history. Journal of the American Society for Information Science, 48(9):810-832.

[10] Tang, R., Shaw, W.M., and Vevea, J.L. (1999): Towards the identification of the optimal numbers of relevance categories. Journal of the American Society for Information Science, 50(3):254-264.

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

[12] Voorhees, E. (2001): Evaluation by highly relevant documents. Proceedings of the 24th ACM-SIGIR conference on research and development in information retrieval, New York, pp. 74-82.

[13] Vorhees, E.M. (1998): Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. Proceedings of the 21st ACM-SIGIR conference on research and development in information retrieval, Melbourne pp. 315-323.

[14] Fuhr, N., Lalmas, M., Kazai, G., Gövert, N. eds (2003): Proceedings of the first workshop of the INitiative for the Evaluation of XML Retrieval (INEX), Dagstuhl. ERCIM workshop proceedings.

[15] Järvelin, K., Kekäläinen, J. (2000): IR evaluation methods for retrieving highly relevant documents. Proceedings of the 23rd ACM-SIGIR conference on research and development in information retrieval, Athens, pp. 41-48.