Relevance in XML Retrieval: The User Perspective

Jovan Pehcevski
School of CS & IT
RMIT University
Melbourne, Australia
jovanp@cs.rmit.edu.au

ABSTRACT

A realistic measure of relevance is necessary for meaningful comparison of alternative XML retrieval approaches. Previous studies have shown that the current INEX relevance definition, comprising two dimensions based on topical relevance, is too hard for users to understand. In this paper, we propose and evaluate a new relevance definition that uses five-point scale to assess the relevance of returned elements. We perform a comparative analysis of the judgements obtained from interactive user experiments and the INEX 2005 relevance assessments to demonstrate the usefulness of the new relevance definition for XML retrieval.

1. INTRODUCTION

It is a commonly held view that *relevance* is one of the most important concepts for the fields of documentation, information science, and information retrieval [8, 14]. Indeed, the main purpose of a retrieval system is to retrieve units of information estimated as *likely to be relevant* to a user information need. To build and evaluate effective information retrieval systems, the concept of relevance needs to be clearly defined.

In traditional information retrieval, a binary relevance scale is often used to assess the relevance of an information unit (usually a whole document) to a user request (usually a query). The relevance value of the information unit is restricted to either zero (when the unit is not relevant to the request) or one (when the unit is relevant to the request). However, binary relevance is not deemed to be sufficient in XML retrieval, primarily due to the hierarchical relationships among the units of retrieval [13].

Each year since 2002, a new set of retrieval topics has been proposed and assessed by participants in INEX.[1] Analysing the behaviour of *assessors* when judging the relevance of returned elements may provide insight into possible trends within the relevance assessments [4, 13]. An interactive track was established for the first time in INEX 2004 [2] to investigate the behaviour of *users* when elements of XML documents (rather than whole documents) are presented as answers [15].

At INEX 2003 and 2004, two relevance dimensions — *Exhaustivity* and *Specificity* — were used to measure the extent that an element respectively *covers* and is *focused on* an information need. Each dimension used four grades to reflect how exhaustive or specific an element was: "none", "marginally", "fairly", and "highly". To assess the relevance of an element, the grades from each dimension were combined into a single 10-point relevance scale. In our previous work we have performed an empirical analysis of the two INEX 2004 relevance dimensions, where we have demonstrated that the highest level of agreement between the assessor and the users was at the end points of the relevance scale (representing highly relevant and non-relevant elements, respectively), and that the two INEX 2004 relevance dimensions were perceived as one (mostly because the two INEX dimensions are based on topical relevance) [11]. When the two INEX 2004 relevance dimensions were separately analysed, we observed that there was more overall agreement for *Exhaustivity* than for *Specificity*. The most likely reason for this was that both assessors and users seemed to have less understood an important property of the INEX 2004 *Specificity* dimension: an element should be judged as *highly specific* if it *does not* contain *non-relevant* information.

At INEX 2005 the relevance definition was slightly changed, and a highlighting assessment approach was used to gather the relevance assessments [1, 5]. A second interactive track was also established, comprising three tasks and two different XML document collections [6]. In Section 2 we briefly describe the INEX 2005 relevance definition, and present some findings about the assessor understanding of the two relevance dimensions. In Section 3 we propose a new definition of relevance for XML retrieval that uses a five-point scale to assess the relevance of returned elements. In Section 4 we demonstrate the usefulness of the new relevance scale through a comparative analysis of the judgements obtained from the INEX 2005 relevance assessments and those from users in the INEX 2005 Interactive track. We show that users perceive the new five-point relevance scale to be relatively simple, and that the grades of the new relevance

[1] INEX, INitiative for the Evaluation of XML Retrieval. `http://inex.is.informatik.uni-duisburg.de/`

SIGIR 2006 Workshop on XML Element Retrieval Methodology
August 10, 2006, Seattle, Washington, USA.

```
<file collection="ieee" name="co/2000/r7108">
<element path="/article[1]" exhaustivity="1" size="13556" rsize="5494"/>
<element path="/article[1]/bdy[1]" exhaustivity="1" size="9797" rsize="4594"/>
<element path="/article[1]/bdy[1]/sec[1]" exhaustivity="1" size="1301" rsize="409"/>
<element path="/article[1]/bdy[1]/sec[1]/p[1]" exhaustivity="1" size="531" rsize="408"/>
<element path="/article[1]/bdy[1]/sec[2]" exhaustivity="1" size="2064" rsize="2064"/>
<element path="/article[1]/bdy[1]/sec[2]/st[1]" exhaustivity="?" size="30" rsize="30"/>
<element path="/article[1]/bdy[1]/sec[2]/p[2]" exhaustivity="1" size="738" rsize="738"/>
<element path="/article[1]/bm[1]" exhaustivity="1" size="3267" rsize="900"/>
<element path="/article[1]/bm[1]/app[1]" exhaustivity="1" size="2085" rsize="900"/>
<element path="/article[1]/bm[1]/app[1]/p[3]" exhaustivity="1" size="438" rsize="438"/>
</file>
```

Figure 1: A sample from the INEX 2005 CO topic 203 relevance assessments for the relevant file `co/2000/r7108`. For each judged element, `exhaustivity` shows values for *Exhaustivity* (possible values `?`, `1`, or `2`), `size` denotes the element size (measured as total number of contained characters), while `rsize` shows the actual number of highlighted characters.

scale can easily be deduced from the amount of highlighted text in the relevant elements.

2. INEX 2005 RELEVANCE

The highlighting assessment task used at the INEX 2005 ad hoc track to gather relevance assessments for the retrieval topics had three main steps [5]. The assessor was first required to highlight the relevant content in each returned article. The assessment tool automatically identified the elements that enclosed the highlighted content, and the assessor was then asked to judge the *Exhaustivity* of these elements, and of all their ancestors and descendants. Last, the tool automatically computed the *Specificity* as the ratio of highlighted to fully contained text. The highlighting assessment task was also used at the INEX 2005 multimedia (MM) track, with the difference that the assessor was not asked to judge the *Exhaustivity* of the elements that contained highlighted content [17].

Figure 1 shows a sample of the relevance assessments obtained for the INEX 2005 Content Only (CO) topic 203. For each judged element, `exhaustivity` shows the *Exhaustivity* value of the element, with possible values of `?` (too small), 1 (partially exhaustive), and 2 (highly exhaustive); `size` denotes the total number of characters contained by the element; and `rsize` shows the actual number of characters highlighted by the assessor.

To measure the *relevance* of an element, a quantisation function is used to normalise the values obtained from the two INEX 2005 relevance dimensions [3]. For example, if the observed `exhaustivity` value is 1 and both values for `size` and `rsize` are the same (see Figure 1), the element is deemed as *highly specific* but only *partially exhaustive* [5].

To examine the extent to which the assessors understand the two INEX 2005 relevance dimensions, we have performed an analysis of the level of assessor agreement on the five topics that were double-judged at INEX 2005 [10]. The results show that there is good reason to ignore the *Exhaustivity* dimension during evaluation, since it appears to be easier for assessors to be consistent when highlighting relevant content than when choosing one of the three exhaustivity values [10].

This suggests that a much simpler relevance scale would be a better choice for evaluation in INEX and XML retrieval in general. Indeed, in their analysis of relevance judgements obtained from the users of the INEX 2004 Interactive track, Pharo and Nordlie [12] also observed the following: "A combined measure of relevance with so many alternatives as the one used in this experiment proves difficult for the searchers to relate to. In further experiments it might be fruitful to use another scale and resort to two separate assessments". In the next section we propose one such relevance scale.

3. A NEW DEFINITION OF RELEVANCE FOR XML RETRIEVAL

In this section we present a new relevance definition for XML retrieval. We describe the aspects and the two dimensions of the new relevance definition, and its five-point relevance scale. To demonstrate the simplicity of the new relevance scale, we also analyse user feedback gathered from the INEX 2005 Interactive track.

3.1 Aspects and dimensions

We base our new relevance definition on three aspects:

- There should be only *one* dimension of relevance based on *topical relevance*;
- The first relevance dimension should use a *three-graded* relevance scale, which will determine whether an XML element is either *highly relevant*, *relevant*, or *not relevant* to an information need; and
- There should be a *second* dimension of relevance, based only on the intrinsic hierarchical relationships among the XML elements.

Using only one topical relevance dimension allows the new relevance definition to be more intuitive than the INEX 2004

and INEX 2005 relevance definitions, which have two relevance dimensions based on topical relevance.

The first relevance dimension is inspired by our analysis of the level of agreement between the assessor and the users on the INEX 2004 CO topics, where the highest level of agreement was shown to be on *highly relevant* and on *non-relevant* elements [11]. However, in addition to the above two grades we also allow for a third relevance grade, *relevant*, to be incorporated in our first relevance dimension. This is supported by the fact that — to explore the effect of incorporating only highly relevant documents in the retrieval evaluation — most recent web tracks in TREC have adopted a similar three-point scale based on topical relevance [18].

The second dimension of relevance, as introduced in the third aspect above, is based only on the hierarchical relationships which are intrinsic to XML documents. O'Keefe [9] analyses some properties of the INEX 2004 IEEE document collection, and finds that elements that are highly coupled to their context are more difficult to judge than elements with low coupling. In this scenario, what matters most is "not how big the fragments are but how tightly they are coupled to their context" [9]. O'Keefe also argues that the usefulness of the XML retrieval task would also depend on the size of the retrieved information units; indeed, the appropriate units of retrieval should be self-contained, with a reasonable size, and at the same time with some coupling to their containing documents. Trotman [16] also examines these properties in detail.

We follow the above reasoning and allow three grades for our second relevance dimension: *just right*, *too large*, and *too small*. An XML element is *just right* if it is reasonably self-contained, and at the same time has enough coupling to be bound to its containing XML document. Alternatively, the element can be either *too large* or *too small*. An XML element is *too large* if it is reasonably self-contained, but it is either too big to be examined as an answer, or its coupling is so low that it can represent a free-standing XML document. An XML element is *too small* if it is not self-contained and its content is highly dependent on the context (high coupling), which makes it too small to be examined as an answer.

This second dimension of relevance is similar to *document coverage* used in INEX 2002 [4]. Indeed, document (or component) coverage was used as a relevance dimension in INEX 2002 to measure how specific (or focused) the unit of retrieval is to the information need. In a similar way to our second dimension, some aspects of document coverage depend on the context of the element; indeed, for a *too small* element Kazai et al. state that "the component is too small to act as a meaningful unit of information when retrieved by itself" [4]. However, the other two relevance grades, *too large* and *just right*, were not explicitly captured by the document coverage relevance dimension.

3.2 Relevance scale

As described above, our new relevance definition uses two *dimensions* to calculate the assessment score of an XML element.

	Questions	
Value	Q4.5	Q4.6
Mean	**2.51**	**2.96**
Minimum	1	1
Maximum	5	5
Median	2	3
StDev	1.27	1.29

Table 1: Analysis of responses on questions Q4.5 and Q4.6 gathered from 29 users that participated in Task C of the INEX 2005 Interactive track. For both questions, users were required to choose from five available answers, ranging from 1 ("Not at all") to 5 ("Extremely"). Mean average values obtained for each question are shown in bold.

The first relevance dimension determines the extent to which an XML element *contains relevant information* for the search task. It can take one of the following three values: *highly relevant*, *relevant*, or *not relevant*. The second relevance dimension determines the extent to which an XML element *needs the context* of its containing XML document to make full sense as an answer. It can take one of the following three values: *just right*, *too large*, or *too small*.

Thus, the final assessment score of an XML element can take one of the following five nominal values:

- **Exact Answer** (EA), if-and-only-if the XML element is *just right* and *highly relevant*;
- **Partial Answer** (PA), if-and-only-if the XML element is *just right* and *relevant*;
- **Broad Answer** (BA), if-and-only-if the XML element is *too large* and either *relevant* or *highly relevant*;
- **Narrow Answer** (NA), if-and-only-if the XML element is *too small* and *highly relevant*; and
- **Not Relevant** (NR), if the XML element does not cover any of the aspects of the information need.

To demonstrate that the above scale is not hard for users to understand, next we present analysis of the user responses obtained from the questionnaires collected for Task C of the INEX 2005 Interactive track.

3.3 User satisfaction

To measure the user satisfaction while using the new five-point relevance scale, users were asked to provide answers to the following two questions:

- Was it hard to understand and use the five-point relevance scale? (question Q4.5)
- Would it have been better if a simpler relevance scale was used instead? (question Q4.6)

For both questions, users were required to choose from five available answers, ranging from 1 ("Not at all") to 5 ("Extremely"). Table 1 shows an analysis of the responses gathered from 29 users for the two questions. The relatively

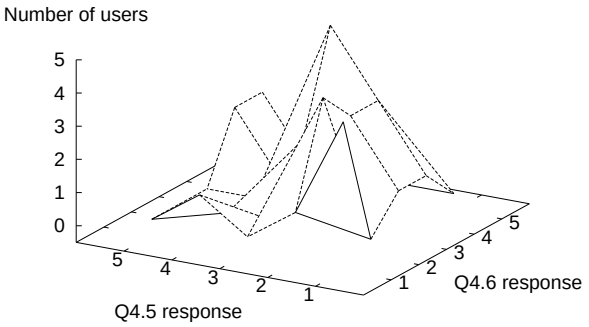

Figure 2: A 3D histogram of user responses on Q4.5 and Q4.6.

low mean average value (2.51) of responses to question `Q4.5` shows that users had little difficulty in understanding the new five-point relevance scale. At the same time, the mean average value of responses to question `Q4.6` (2.96) indicates that it was not really necessary to have a simpler relevance scale than the one used.

Figure 2 shows a more detailed analysis of the user responses to questions `Q4.5` and `Q4.6`, allowing us to explore whether there is any correlation between the responses to the two questions. We find that for question `Q4.5`, around 83% of the users chose one of the first three answers (1, 2, or 3). Of these, the largest number of users (38%) chose answer 2, while 24% and 21% of the users chose answers 3 and 1, respectively. Around 67% of the users who chose answer 1 for question `Q4.5` also chose the same answer for question `Q4.6`. The correlation is similar for answer 2, where the highest percentage of users who chose this answer for question `Q4.5` also chose the same answer for question `Q4.6`. These statistics show that users participating in Task C of the INEX 2005 Interactive track did not perceive the new five-point relevance scale to be hard to use.

4. EXPERIMENTS WITH THE NEW RELEVANCE DEFINITION

In this section, we present experiments that demonstrate the usefulness of our new relevance definition for XML retrieval. We first compare the new relevance scale to the one used in INEX 2005, and design a mapping between their respective relevance grades. We then present a performance analysis of simulated runs that use this mapping to construct their answer elements.

4.1 Comparison to the INEX 2005 relevance

Three tasks were explored in the INEX 2005 Interactive track [6]:

- Task A, where users searched three topics using a common baseline system with the INEX 2005 IEEE XML document collection (which was also used in the INEX 2005 ad hoc track);
- Task B, where groups with a working interactive XML retrieval system could test their system against the baseline system; and
- Task C, where users searched four topics using alternative system with the Lonely Planet XML document collection (which was also used in the INEX 2005 MM track).

In the following analysis, we focus on results obtained from Tasks A and C.

Task A judgements

For Task A, six topics grouped in two categories (General and Challenging) were selected for users, who were required to choose and search on only one topic per category. The six topics were derived from selected topics used in the INEX 2005 ad hoc track. We analyse relevance judgements obtained from a number of users for topics G1 (21 users) and G2 (18 users) of the General topic category, and relevance judgements for topics C2 (17) and C3 (26) of the Challenging category. We chose these four topics as all of them have corresponding assessor judgements available,[2] which makes it possible to analyse and compare the extent to which *both* assessors and users perceived the relevant answers for those topics. A simple three-point relevance scale was used by users of Task A, with the following values: `Relevant` (2), `Partial` (1), and `Not Relevant` (0). This relevance scale closely reflects the one used for the INEX 2005 *Exhaustivity* dimension. Our aim in the following analysis is to deduce a relationship between the two points of this scale that are assigned to *relevant* elements by users and the actual judgements assigned to the same elements by assessors.

Table 2 shows a statistical analysis of the overall distribution of user and assessor judgements across the four topics. For a relevance grade (`Relevant` or `Partial`), the `Total` values show the total number of (non-zero) elements judged by users across the four topics. Of these elements, the `MA` values show the number of those elements that were also mutually agreed to be non-zero by the assessor. The `E2`, `E1`, and `E?` values show the actual distribution of assessor judgements on the `MA` elements. For example, of the total 486 elements judged as `Relevant` by users, 352 were also judged as having non-zero relevance by assessors (denoted as `MA`). However, assessors did not always agree that these elements were `Relevant` (denoted as `E2` in the assessor judgements). In fact, 256 of the 352 `MA Relevant` elements were judged by assessors as `E2`, 96 were judged as `E1`, while none were judged as `E?` (too small). The `Agreement` values show the actual agreement between users and assessors on a relevance grade (for example, the overall agreement for the `Relevant` grade is $256/352 = 73\%$). As shown in the table, for a relevance grade we also measure the proportion of the relevant information contained by the agreed `MA` elements (`av_prel`) along with the corresponding standard deviation (`StDev`).

From the numbers shown in Table 2 we observe that, first, the overall agreement between assessors and users seems

[2]We used the relevance assessments that belong to the INEX 2005 CO topics 235 and 241 for topics G1 and C2, and those that belong to the INEX 2005 VVCAS topics 256 and 257 for topics C3 and G2, respectively.

User judgements	Non-zero		Assessor judgements					
	Total	MA	E2	E1	E?	av_prel	StDev	Agreement
Relevant	486	352	256	96	0	**0.57**	**0.32**	**0.73**
Partial	388	202	142	60	0	**0.49**	**0.27**	**0.30**

Table 2: Statistical analysis of the overall distribution of user and assessor judgements calculated across the two General (G1 and G2) and the two Challenging (C2 and C3) topics used in Task A of the INEX 2005 Interactive track.

to be higher for `Relevant` than for `Partial` *relevant* elements (73% compared to 30%); and second, the proportion of relevant information contained by the `Relevant` elements seems to be larger than for `Partial` elements (57% compared to 49%). However, these observations should be treated with care, since results from only four topics are used in this analysis.

The first observation seems to be in line with our previous finding on the INEX 2004 topics, where highly relevant answers were perceived better than partially relevant answers [11]. The second observation allows for a mapping to be established between the proportion of relevant information contained by a relevant element and the two grades, exact (`EA`) and partial (`PA`), that can be assigned to the relevant element using our five-point relevance scale. However, this does not provide any indication as to how broad (`BA`) and narrow (`NA`) elements should be mapped. Intuitively, from their definition we expect the `NA` elements to be the smallest in size and to contain the highest proportion of relevant information. Likewise, the `BA` elements should be the largest in size, and should contain the smallest proportion of relevant information.

We now explain how these expectations are validated by comparing the relevance judgements provided by users in Task C of the INEX 2005 Interactive track to the relevance assessments obtained from the INEX 2005 MM track.

Task C judgements

For Task C, eight topics — some derived from the INEX 2005 MM track topics — were arbitrarily grouped in two categories. Users were asked to choose and search on two topics in each category, and assess relevance using our five-point relevance scale. We analyse relevance judgements obtained from a number of users for topics LP1 (11) and LP2 (18) of the first topic category, and relevance judgements for topics LP5 (22) and LP7 (13) of the second category. These four topics also have assessor judgements available.[3]

Table 3 shows a statistical analysis of the overall distribution of user and assessor judgements calculated across the four topics. We observe that the number of user judgements is highest for the broad (`BA`) elements, and that these elements also have the highest number of mutually agreed relevant (`MA`) elements. As expected, on average the `BA` elements contain a very small proportion of relevant information (9%), and, for most of the mutually agreed `BA` elements, the proportion of found relevant information falls in the range 0%–32%. For the `EA` elements, the average proportion of relevant information is similar to that observed for Task A (Table 2), whereas for `PA` and `NA` elements we observe a different proportion of relevant information than that reported (and expected) previously. This can be attributed to the very low number of mutually agreed relevant elements.

[3]We used the relevance assessments for INEX 2005 MM topics 4 and 21 for topics LP1 and LP2, and for INEX 2005 MM topics 6 and 25 for topics LP5 and LP7, respectively.

User judgements	Non-zero		Assessor judgements	
	Total	MA	av_prel	StDev
Exact (EA)	59	17	**0.59**	**0.40**
Partial (PA)	93	9	**0.22**	**0.37**
Broad (BA)	120	39	**0.09**	**0.23**
Narrow (NA)	66	5	**0.55**	**0.50**

Table 3: Statistical analysis of the overall distribution of the user and assessor judgements calculated across four topics (LP1, LP2, LP5, and LP7) used in Task C of the INEX 2005 Interactive track.

In light of these statistics, a *reasonable* mapping between the continuous relevance scale of the INEX 2005 *Specificity* dimension and our five-point relevance scale would be as follows:

1. `EA` $\in (0.66, 1.00]$
2. `PA` $\in [0.33, 0.66]$
3. `BA` $\in (0.00, 0.33)$
4. `NA` $= 1.00$
5. `NR` $= 0.00$

In this mapping, there may be cases where both `EA` and `NA` elements are mapped as highly specific (1.00) elements. This property — illustrated in Figure 3 — is an important property of the above mapping, which as we discuss next primarily ensures to correctly identify the `NA` elements.

Figure 3 shows how the proposed mapping can be used to identify the four types of answer elements from the sample of relevance assessments for document `co/2000/r7108` of the INEX 2005 CO topic 203 (previously shown in Figure 1). The figure shows 10 relevant elements, and for each element the number in parentheses shows the proportion of contained relevant information. An element is identified as a `NA` element if it contains only relevant information (1.00) *and* at the same time its parent also contains only relevant information. There are two such elements shown in Figure 3 (`st[1]` and `p[2]`). However, although two elements, `sec[2]`

Value	CO			VVCAS		
	Total (elements)	av_size (chars)	av_prel	Total (elements)	av_size (chars)	av_prel
EA						
Mean	**332**	**1 145**	**0.98**	**572**	**1 960**	**0.98**
Minimum	17	155	0.95	23	29	0.90
Maximum	1 568	7 250	1.00	3 440	9 329	0.99
Median	269	800	0.98	375	965	0.98
StDev	355	1 318	0.01	693	2 191	0.02
PA						
Mean	**61**	**6 369**	**0.48**	**70**	**10 556**	**0.48**
Minimum	1	489	0.43	3	81	0.44
Maximum	271	26 379	0.55	295	40 798	0.59
Median	32	2 969	0.47	48	5 636	0.48
StDev	73	7 374	0.02	64	10 161	0.03
BA						
Mean	**204**	**19 367**	**0.11**	**186**	**25 351**	**0.13**
Minimum	13	10 225	0.08	16	8 371	0.03
Maximum	995	39 345	0.17	615	47 955	0.19
Median	105	17 054	0.11	130	23 303	0.12
StDev	238	6 933	0.02	150	10 789	0.04
NA						
Mean	**1 635**	**92**	**1.00**	**5 493**	**97**	**1.00**
Minimum	13	9	1.00	1	9	1.00
Maximum	13 994	272	1.00	44 600	283	1.00
Median	234	75	1.00	2 318	85	1.00
StDev	3 252	59	0.00	9 056	70	0.00

Table 4: Statistical analysis of the distribution of EA, PA, BA and NA relevant elements across the 29 CO and 34 VVCAS topics at INEX 2005. For a relevance grade, the Total values show the actual number of relevant elements that belong to that grade, while av_size and av_prel represent averages for the size of the relevant elements (in characters) and the proportion of relevant information contained by the relevant elements, respectively. Mean average values (calculated across all the CO or VVCAS topics) are shown in bold.

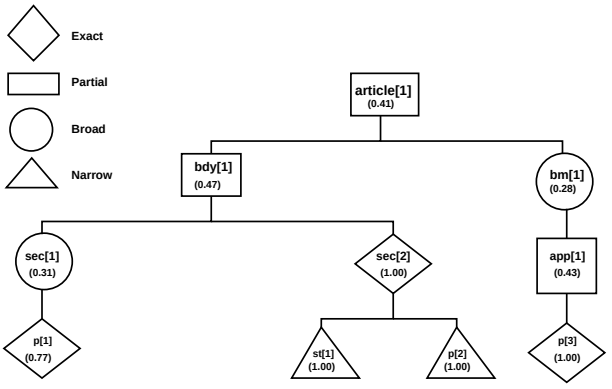

Figure 3: Identifying Exact, Partial, Broad, and Narrow answer elements from the relevance assessments sample that belongs to file co/2000/r7108 of the INEX 2005 CO topic 203. For each element, the number in parentheses shows the proportion of contained relevant information.

and p[3], also contain only relevant information, both are nevertheless identified as EA elements. The above example also shows that full article elements need not always be identified as BA elements; indeed, it is the proportion of contained relevant information in an element that determines its element type. Next, we use the proposed mapping and the INEX 2005 relevance assessments to find the actual distribution of the four element types across the INEX 2005 CO and Vague Content And Structure (VVCAS) topics.

INEX 2005 CO and VVCAS judgements

Table 4 shows a statistical analysis of the distribution of EA, PA, BA and NA relevant elements across the 29 CO and 34 VVCAS[4] topics at INEX 2005, when using the proposed mapping. As expected, the assessment trends are clear for both types of topics: the NA elements are the most common, the smallest in size, and contain only relevant information. The PA elements are the least common elements, while the BA elements are the largest in size, and contain the smallest proportion of relevant information. The EA elements are smaller in size than the PA elements, but contain higher proportion of relevant information.

To investigate the relationship between the four relevance grades and the three values of the INEX 2005 *Exhaustivity* dimension, we also analyse the distribution of the three *Exhaustivity* values across the four types of relevant elements. Table 5 shows this distribution, which is calculated separately for the INEX 2005 CO and the VVCAS topics. We observe that for the INEX 2005 CO topics the majority of EA elements were judged as partially exhaustive (E1), while for the INEX 2005 VVCAS topics most of the EA elements were judged as too small. This is somewhat surprising, showing that (on average) INEX 2005 assessors considered the

[4]We analyse relevance assessments for both parent and child VVCAS topics.

Value	CO				VVCAS			
		Exhaustivity				Exhaustivity		
	Total	E2	E1	E?	Total	E2	E1	E?
EA								
Mean	332	0.16	**0.48**	0.36	571	0.19	0.35	**0.46**
PA								
Mean	61	0.32	**0.63**	0.05	70	0.35	**0.57**	0.08
BA								
Mean	204	0.27	**0.69**	0.04	186	0.28	**0.68**	0.04
NA								
Mean	1 635	0.08	0.11	**0.81**	5 493	0.02	0.07	**0.91**

Table 5: Distribution of the three Exhaustivity values across the EA, PA, BA and NA relevant elements found for the 29 CO and 34 VVCAS topics at INEX 2005. For each of the four types of relevant elements, the Total values show the actual number of relevant elements, while E2, E1 and E? represent values for the proportion of those relevant elements that were assigned a corresponding Exhaustivity value. The highest values are shown in bold.

elements that contain most of the highlighted content to either discuss only some aspects of the underlying information need or to be too small. The partially exhaustive elements also represent the majority in both cases of PA and BA elements, while not surprisingly, most of the NA elements were correctly judged to be too small.

4.2 Performance analysis

In the following, we aim at investigating which of the four element types yields the best value in retrieving (non-overlapping) relevant information, which we believe could represent valuable knowledge in tuning the XML retrieval system parameters for optimal performance. We use the INEX 2005 CO topics to evaluate the performance of six simulated runs, four of which were created by only considering relevant elements that belong to the corresponding four element types (EA, PA, BA, and NA). The fifth run contains all the (overlapping) relevant elements found for the INEX 2005 CO topics (FullRB). To also investigate the XML retrieval performance when only the highlighted passages are units of retrieval, the sixth simulated run was created such that it contains (provisional) elements with sizes that strictly match the sizes of the corresponding passages.

For each run and an INEX 2005 topic, at most 1 500 elements were considered in the final answer list, where retrieved units were ranked in descending order according to the harmonic mean between precision (the proportion of relevant information to all the information retrieved from the element) and recall (the proportion of relevant information retrieved from the element to all the relevant information found for the topic). Overlapping answer elements were allowed in the answer lists of the five element runs. We use the HiXEval evaluation metric to measure the retrieval performance [10], with a parameter setting that penalises the retrieved overlapping relevant information among elements. A *system*-oriented retrieval task is considered for this performance analysis, where runs are rewarded if they retrieve as much non-overlapping relevant information as possible (high recall), without also retrieving a substantial amount of non-relevant information (high precision).

The graph in Figure 4 shows the retrieval performance of the six simulated runs. Perfect retrieval performance is achieved

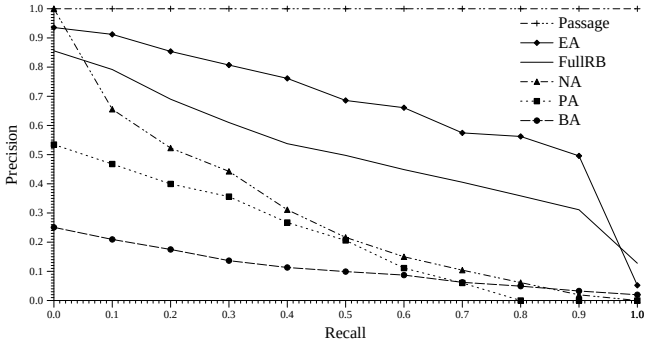

Figure 4: Performance evaluation of the six simulated runs on the 29 INEX 2005 CO topics using the HiXEval evaluation metric.

with the passage run, the EA run performs the best among the five element runs, while the BA run, which only contains broad answer elements, performs the worst. When the performance of the FullRB run is compared to that of the other element runs, we observe that the EA run performs better than FullRB. This shows that, when overlap is considered by HiXEval, better value in retrieving relevant information is achieved by identifying the (overlapping) exact answers, and not by retrieving all the (overlapping) relevant elements. Of the other two simulated runs, the NA run performs better than the PA run. Two factors influence this performance behaviour: first, as shown in Table 4 the average number of NA elements across the INEX 2005 CO topics is approximately 27 times that of PA elements, which allows for the NA simulated run to achieve higher overall recall than that achieved by the PA run; and second, the proportion of retrieved relevant information from the NA elements is always higher than that retrieved from the PA elements, which also leads to higher overall precision for the NA run.

The system-oriented retrieval task highlights the importance of identifying the *exact* answer elements. Indeed, the above knowledge that — of all the relevant elements retrieved for this task — the EA elements bring the best value in retrieving relevant information could influence the choice of tuning the XML retrieval system parameters for optimal performance.

5. CONCLUSIONS

In this paper we have presented an empirical analysis of what the experience of assessors and users suggests about how *relevance* should be defined and measured in XML retrieval. We have proposed a new relevance definition that is founded on results obtained from interactive XML retrieval experiments, and which uses a five-point relevance scale to assign an assessment score for an answer element.

There is a recent argument that a complex relevance scale may lead to an increased level of obtrusiveness in interactive user environments [7]. We have demonstrated that the new relevance scale was successfully used in Task C of the INEX 2005 Interactive track, where users did not find it to be very hard to use.

By analysing results from the topics judged by the assessors in INEX 2005 and by the users participating in the INEX 2005 Interactive track, we have been able to empirically establish a mapping between the continuous scale used by the *Specificity* dimension at INEX 2005 and our new five-point relevance scale. This mapping has allowed us to analyse the distribution of the four types of relevant elements in the INEX 2005 relevance assessments. We have presented an analysis of the performance of four simulated runs, each containing elements that belong to one of the four element types, and have shown that identifying and retrieving exact answer elements yields the best value in retrieving relevant information.

The performance evaluation shown in the last section is a *system*-oriented than a *user*-oriented evaluation. We plan to experiment with different types of relevance assessments, which may reflect different models of user behaviour, to more closely investigate whether or not the user model influences the best value in retrieving relevant information.

Acknowledgements

We thank James Thom, Saied Tahaghoghi, and anonymous reviewers for their comments on earlier drafts of this paper.

6. REFERENCES

[1] N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors. *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Dagstuhl 28-30 November 2005*, volume 3977 of *Lecture Notes in Computer Science*. Springer-Verlag, January 2006.

[2] N. Fuhr, M. Lalmas, S. Malik, and Z. Szlavik, editors. *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Revised Selected Papers*, volume 3493 of *Lecture Notes in Computer Science*. Springer-Verlag, May 2005.

[3] G. Kazai and M. Lalmas. INEX 2005 evaluation measures. In Fuhr et al. [1], pages 16–29.

[4] G. Kazai, S. Masood, and M. Lalmas. A study of the assessment of relevance for the INEX 2002 test collection. In *Proceedings of the 26th European Conference on IR Research (ECIR)*, pages 296–310, Sunderland, UK, 2004.

[5] M. Lalmas and B. Piwowarski. INEX 2005 relevance assessment guide. In *INEX 2005 Workshop Pre-Proceedings, Dagstuhl, Germany, November 28–30, 2005*, pages 391–400, 2005.

[6] B. Larsen, S. Malik, and A. Tombros. The Interactive track at INEX 2005. In Fuhr et al. [1], pages 398–410.

[7] B. Larsen, A. Tombros, and S. Malik. Obtrusiveness and relevance assessment in interactive XML IR experiments. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 39–42, Glasgow, UK, 2005.

[8] S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science and Technology*, 48(9):810–832, 1997.

[9] R. A. O'Keefe. If INEX is the answer, what is the question? In Fuhr et al. [2], pages 54–59.

[10] J. Pehcevski and J. A. Thom. HiXEval: Highlighting XML retrieval evaluation. In Fuhr et al. [1], pages 43–57.

[11] J. Pehcevski, J. A. Thom, and A.-M. Vercoustre. Users and assessors in the context of INEX: Are relevance dimensions relevant? In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 47–62, Glasgow, UK, 30 July 2005.

[12] N. Pharo and R. Nordlie. Context matters – an analysis of assessments of XML documents. In *Proceedings of 5th International Conference on Conceptions of Library and Information Sciences, CoLIS 2005*, pages 238–248, Glasgow, UK, 2005.

[13] B. Piwowarski and M. Lalmas. Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management (CIKM '04)*, pages 361–370, Washington DC, USA, 2004.

[14] T. Saracevic. Relevance reconsidered. In *Proceedings of 2nd International Conference on Conceptions of Library and Information Sciences, CoLIS 1996*, pages 201–218, Copenhagen, Denmark, 1996.

[15] A. Tombros, B. Larsen, and S. Malik. The Interactive track at INEX 2004. In Fuhr et al. [2], pages 410–423.

[16] A. Trotman. Wanted: Element retrieval users. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 63–69, Glasgow, UK, 2005.

[17] R. van Zwol, G. Kazai, and M. Lalmas. INEX 2005 multimedia track. In Fuhr et al. [1], pages 497–510.

[18] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 74–82, New Orleans, USA, 2001.