HiXEval: Highlighting XML Retrieval Evaluation

Jovan Pehcevski and James A. Thom

School of Computer Science and Information Technology, RMIT University
Melbourne, Australia
{jovanp, jat}@cs.rmit.edu.au

Abstract. This paper describes our proposal for an evaluation metric for XML retrieval that is solely based on the highlighted text. We support our decision of ignoring the exhaustivity dimension by undertaking a critical investigation of the two INEX 2005 relevance dimensions. We present a fine grained empirical analysis of the level of assessor agreement of the five topics double-judged at INEX 2005, and show that the agreement is higher for specificity than for exhaustivity. We use the proposed metric to evaluate the INEX 2005 runs for each retrieval strategy of the `CO` and `CAS` retrieval tasks. A correlation analysis of the rank orderings obtained by the new metric and two `XCG` metrics shows that the orderings are strongly correlated, which demonstrates the usefulness of the proposed metric for evaluation of XML retrieval performance.

1 Introduction

How to properly evaluate XML retrieval effectiveness is a much-debated question among the XML retrieval research community. Over the past four years INEX has been used as an arena to investigate the behaviour of a variety of evaluation metrics. However, unlike in previous years a new set of official metrics was adopted at INEX 2005, which belong to the eXtended Cumulated Gain (`XCG`) family of metrics [2, 4]. Two official INEX 2005 metrics are `nxCG` (with the `nxCG[r]` measure), which for a rank `r` measures the relative retrieval gain a user has accumulated up to that rank, compared to the gain they could have accumulated if the system had produced the optimal ranking; and `ep/gr` (with the `MAep` measure), which for a cumulated gain level measures the amount of relative effort (as the number of visited ranks) a user is required to spend compared to the effort they could have spent while inspecting an optimal ranking [3].

Since 2003, two relevance dimensions — exhaustivity and specificity — have been used in INEX to measure the extent to which an element is *relevant* to the information need expressed by an INEX topic. A highlighting assessment approach was used at INEX 2005 to gather relevance judgements for the retrieval topics [7]; here, the specificity of an element is automatically computed as the ratio of highlighted to fully contained text, while the assessor is asked to explicitly judge the exhaustivity of that element. Figure 1 shows a sample of relevance judgements obtained for INEX 2005 Content Only (`CO`) topic 203. For each

```
<file collection="ieee" name="co/2000/r7108">
 <element path="/article[1]" E="1" size="13,556" rsize="5,494"/>
 <element path="/article[1]/bdy[1]" E="1" size="9,797" rsize="4,594"/>
 <element path="/article[1]/bdy[1]/sec[2]" E="1" size="2,064" rsize="2,064"/>
 <element path="/article[1]/bdy[1]/sec[2]/st[1]" E="?" size="30" rsize="30"/>
</file>
```

Fig. 1. A sample from the INEX 2005 CO topic 203 relevance judgements for article `co/2000/r7108`. For each judged element, `E` shows the value for exhaustivity (with possible values `?`, 1 and 2), `size` denotes the element size (measured as total number of contained characters), while `rsize` shows the actual number of characters highlighted as relevant by the assessor

judged element, `E` shows the exhaustivity 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 highlighted characters by the assessor.

To measure the *relevance* of an element, the official INEX 2005 metrics combine the values obtained from the two INEX relevance dimensions. For example, if the observed value for `E` is 1 and both values for `size` and `rsize` are the same, the element is deemed as *highly specific* but only *partially exhaustive* [7]. A quantisation function is then used to combine these two values into a number that is subsequently used to reflect the relevance of the element [3]. However, in previous work we have shown that finding the best way to combine the values from exhaustivity and specificity to reflect the relevance of an element is too difficult [8]; moreover, recent analysis by Trotman has also shown that the element level agreement between two assessors across the twelve double-judged topics at INEX 2004 is very low, suggesting that "quantization functions based on relevance levels will prove unsound" [10]. In Section 2 we revisit and validate the above claim by analysing the level of assessor agreement across the five topics that were double-judged at INEX 2005.

Another criticism of the official INEX metrics is their lack of simplicity, and their slight departure from the well-established information retrieval norms [1]. Further to this, to consider the level of overlap among retrieved elements, the `XCG` family of metrics use a rather ad hoc methodology in constructing the so-called *ideal recall-base* [3], where a dependency normalisation function is also used to adjust the scores of the descendents of the ideal elements. To date, critical analysis of whether the reliance on these or alternative choices has a positive or negative impact on the XML retrieval evaluation has not been provided.

We contend that the purpose of the XML retrieval task is to find elements that contain *as much relevant information as possible*, without also containing a significant amount of non-relevant information. To measure the extent to which an XML retrieval system returns relevant information, we follow an evaluation

methodology that only takes into account the amount of highlighted text in a retrieved element, without considering the E value of that element. In Section 3 we introduce HiXEval (pronounced hi–ex–eval) – an Evaluation metric for XML retrieval that extends the traditional definitions of precision and recall to include the knowledge obtained from the INEX 2005 Highlighting assessment task.

We recognise that there are no absolute criteria for the choice of a metric for XML retrieval. However, we argue that HiXEval meets all the requirements needed for an unbiased XML retrieval evaluation, and show in Section 4 that, given the strong correlations of its rank orderings to the ones obtained by the XCG metrics, it can and should be used to evaluate XML retrieval effectiveness.

2 Analysis of INEX 2005 CO and VVCAS Relevance Judgements

In this section, we analyse the INEX 2005 relevance judgements obtained for the CO and Vague Content And Structure (VVCAS) topics. First, we analyse the distribution of the E?, E1, and E2 judged elements across the INEX 2005 topics. Then, we analyse the level of assessor agreement obtained from the five topics that have been double-judged at INEX 2005 (four CO, and one VVCAS).

2.1 Distribution of Relevant Elements

The INEX 2005 IEEE document collection comprises 16,820 scientific articles published between 1995–2004, with an average article length of 44,030 characters. Currently, there are 29 CO and 34 VVCAS topics that have corresponding relevance judgements available[1]. We use relevance judgements for both parent and child VVCAS topics in our analysis.

By analysing the INEX 2005 relevance judgements, we aim to discover whether the average number, size, and proportion of contained relevant information in judged elements differ depending on the *exhaustivity* value given to these elements. For example, we expect to find many relevant elements whose exhaustiveness is judged as "?", making them too small. The proportion of relevant information found in these elements (their *specificity* value) is expected to be very high, reflecting the fact that most of the contained information is relevant. Conversely, it is reasonable to expect that the distribution of other relevant elements (such as E1 or E2) is likely to differ from the distribution of the too small elements, both in terms of their average number, size, and proportion of contained relevant information.

Table 1 shows our analysis of the INEX 2005 CO and VVCAS relevance judgements. As expected, for both types of topics the assessment trends are clear: The too small (E?) elements are the most common, the smallest in size, and contain the highest proportion of relevant information. In contrast, the highly exhaustive (E2) elements are the least common, the largest in size, and contain

[1] We use version 7.0 of the INEX 2005 ad hoc relevance judgements.

Table 1. Statistical analysis of the distribution of E? (too-small), E1, and E2 relevant elements across the INEX 2005 CO and VVCAS topics. Numbers for Size and RSize represent averages obtained from each of the 29 CO and 34 VVCAS topics, respectively. Mean average values (calculated across all topics) are shown in bold

Value	CO			VVCAS		
	Total (elements)	Size (bytes)	RSize (%)	Total (elements)	Size (bytes)	RSize (%)
E? (too-small)						
Mean Average	**1706**	**190**	**97**	**5710**	**101**	**99**
Minimum	2	4	59	2	7	91
Maximum	14,543	1,497	100	44,628	497	100
Median	392	72	100	2,422	74	100
Standard Deviation	3,281	359	8	9,118	104	2
E1						
Mean Average	**389**	**7,508**	**60**	**439**	**9,359**	**64**
Minimum	14	497	20	8	1,738	21
Maximum	1,519	13,477	100	1,876	20,236	100
Median	251	7,177	59	365	7,835	71
Standard Deviation	378	3,379	19	415	5,156	20
E2						
Mean Average	**143**	**18,039**	**55**	**174**	**21,575**	**58**
Minimum	2	2,686	16	14	3,746	19
Maximum	1,203	45,909	100	839	55,028	94
Median	46	17,297	50	53	16,832	54
Standard Deviation	237	10,961	20	222	12,550	19

the smallest proportion of relevant information. The partially exhaustive (E1) elements lie in between.

These statistics show that — on average, at least — the assignment of the three exhaustivity grades seems to properly reflect their initial relevance definitions [7]. However, a closer look at the too small element distribution reveals some inconsistencies in connection to the E? relevance grade. For example, Table 1 shows that the maximum average size of the too small elements is 1,497 characters, which is found for CO topic 207. On the other hand, the minimum value for the proportion of contained relevant information is 59%, found for CO topic 222. A closer inspection of the relevance judgements for these two topics reveals many cases where an article body is judged to be too small, while the whole article is judged to be either E1 or E2, despite the fact that the sizes of the article and its body are nearly the same. Given that the average size of an article in the INEX 2005 document collection is 44,030 characters, we should ask the question: How can a 40KB article body be so incomplete that it is judged to be too small?

These and similar examples suggest that assessors seem to have their own interpretations of what *too small* means; arguably, these interpretations could have

Table 2. Overall article and element level agreement between two assessors for the five topics double-judged at INEX 2005. Agreements are calculated on all relevant (non-zero) items, and separately on items that belong to a relevance grade of an INEX relevance dimension. For an INEX 2005 topic, the value of ∪ represents the total number of unique relevant items judged by the two assessors, while ∩ shows the number of mutually agreed relevant items. The ∩/∪ values reflect the level of agreement between the two assessors. Mean average ∩/∪ values are shown in bold

	Non-zero			E?	E1	E2	S1	S2	S3
Topic (Type)	(∪)	(∩)	(∩/∪)	(∩/∪)	(∩/∪)	(∩/∪)	(∩/∪)	(∩/∪)	(∩/∪)
Article level									
209 (CO)	133	48	0.36	—	0.05	0.33	0.19	0.06	0.00
217 (CO)	58	19	0.33	0.00	0.10	0.17	0.00	0.00	0.19
234 (CO)	254	193	0.76	—	0.14	0.22	0.71	0.58	0.00
237 (CO)	134	25	0.19	—	0.09	0.13	0.19	—	—
261 (VV)	38	11	0.29	—	0.03	0.70	0.14	0.50	0.00
Mean	123	59	**0.39**	**0.00**	**0.08**	**0.31**	**0.25**	**0.28**	**0.05**
Element level									
209 (CO)	17,599	2,122	0.12	0.08	0.12	0.07	0.08	0.03	0.10
217 (CO)	10,441	1,911	0.18	0.17	0.01	0.06	0.00	0.01	0.18
234 (CO)	5,785	2,824	0.49	0.01	0.15	0.15	0.62	0.22	0.43
237 (CO)	1,630	220	0.13	0.02	0.10	0.11	0.14	0.05	0.09
261 (VV)	5,470	1,657	0.30	0.30	0.12	0.29	0.12	0.23	0.30
Mean	8,185	1,747	**0.24**	**0.12**	**0.10**	**0.14**	**0.19**	**0.11**	**0.22**

an adverse effect on retrieval evaluation, especially in cases where *exhaustivity* is given a high weight by the evaluation metric.

Next, for each grade of the two INEX relevance dimensions, we undertake an analysis of the level of agreement between the two assessors of the five double-judged topics at INEX 2005, to find whether there is indeed a reason for ignoring the exhaustivity dimension during evaluation.

2.2 Level of Assessor Agreement

Four of the five topics double-judged at INEX 2005 are `CO` topics (numbers 209, 217, 234, and 237), while one is a `VVCAS` topic (number 261). As shown in Table 2, we calculated two separate assessor agreements: one at article level, and another at element level. The ∪ values represent the number of unique relevant items judged by the two assessors, while ∩ values are the number of mutually agreed relevant items. The level of assessor agreement is shown by the ∩/∪ values.

The assessor agreements shown in Table 2 are calculated for seven different cases: once for all relevant (non-zero) items, and for six other cases when relevant items belong to each of the six relevance grades of the two INEX relevance dimensions. Since the specificity dimension at INEX 2005 is measured on a continuous [0-1] scale, we decided to divide this scale to three equal relevance

Table 3. Fine grained element level agreement between two assessors for the five topics double-judged at INEX 2005. For an INEX 2005 topic, `MA` represents the number of mutually agreed relevant (non-zero) elements. For a relevance grade of an INEX relevance dimension, the value of $\cup$ represents the number of relevant elements judged by the assessor of the official INEX 2005 topic, which are part of the mutually agreed relevant elements for that topic. The value of $\cap/\cup$ reflects the fraction of elements confirmed to belong to the same relevance grade by the additional assessor of the INEX 2005 topic. Mean average $\cap/\cup$ values are shown in bold

		E?		**E1**		**E2**		**S1**		**S2**		**S3**	
Topic	`MA`	($\cup$)	($\cap/\cup$)	($\cup$)	($\cap/\cup$)	($\cup$)	($\cap/\cup$)	($\cup$)	($\cap/\cup$)	($\cup$)	($\cap/\cup$)	($\cup$)	($\cap/\cup$)
`209`	2,122	1,629	0.73	424	0.50	69	0.70	94	0.84	59	0.25	1,969	0.83
`217`	1,911	1,889	0.88	15	0.33	7	0.86	1	0.00	1	1.00	1,909	0.97
`234`	2,824	878	0.01	810	0.81	1,136	0.19	782	0.96	145	0.49	1,897	0.99
`237`	220	29	0.28	145	0.86	46	0.26	129	0.89	29	0.34	62	0.90
`261`	1,657	1,545	0.98	72	0.54	40	0.70	19	0.58	25	0.48	1,613	1.0
`Mean`	1,747	1,194	**0.58**	293	**0.61**	260	**0.54**	205	**0.65**	52	**0.51**	1,490	**0.94**

sub-scales, and to assign the marginally specific (`S1`) items to the (0-0.33] scale, the fairly specific (`S2`) items to (0.33-0.67] scale, and the highly specific (`S3`) items to the (0.67-1.0] scale. We have also experimented with different (three- and four-graded) variations of relevance sub-scales, and found that the choice of the sub-scale does not influence the validity of the reported results.

At article level, the assessor agreement for non-zero articles (those articles considered relevant by both assessors, irrespective of their relevance grades) varies from 0.19 on topic 237, to 0.76 on topic 234. The mean article-level agreement between the two assessors is 0.39, which is greater than the value of 0.27 reported by Trotman on the INEX 2004 topics [10], but still lower than the three values — 0.49, 0.43, and 0.42 — reported by Voorhees on the TREC-4 topics [11]. When considering article-level agreements on individual relevance grades, we observe that the highest level of agreement between the two assessors is 0.31 (on highly exhaustive `E2` articles).

At element level, the assessor agreement when all the non-zero elements are considered varies from 0.12 on topic 209, to 0.49 on topic 234. The mean element-level agreement between the two assessors is 0.24, which is (again) greater than the value of 0.16 reported by Trotman on the INEX 2004 topics [10]. Unlike for the article-level agreements, the agreement between the two assessors on individual relevance grades seems to be higher for specificity rather than for exhaustivity, with the highest level of agreement (0.22) on highly specific `S3` elements. We realise, however, that these values should be treated with care, since results from only five topics — the only ones known to be double-judged at INEX 2005 — are used in our analysis.

Although this analysis provides a useful insight as to how the concept of *relevance* is understood by the INEX assessors, it still does not provide enough

evidence to answer the following question: Is it easier for the assessor to be consistent while highlighting relevant content, or while choosing an exhaustivity value using a three-graded relevance scale? We believe that the first activity is a series of *independent* relevant-or-not decisions, whereas the second activity additionally involves comparison with other *dependent* decisions, given that the exhaustivity value for a parent element is always equal or greater than the value of any of its children. In Table 3 we present a fine-grained analysis of the element-level agreement on each of the six relevance grades, by only considering those elements that were mutually agreed to be relevant by both assessors.

The methodology is as follows. First, we take all the judgements obtained from each assessor of the five *official* INEX 2005 topics, and then for each topic we select only those relevant (non-zero) elements that are also confirmed to be relevant by the *additional* assessor of the INEX 2005 topic. We refer to these elements as *mutually agreed* (`MA`) elements. Next, for both exhaustivity and specificity, we count how many of the `MA` elements belong to a particular relevance grade. For example, Table 3 shows that the distribution of the 2,824 `MA` elements for topic 234 is as follows. For exhaustivity, 878 are too small, 810 are `E1`, and 1,136 are `E2` elements. For specificity, 782 are `S1` elements, 145 are `S2`, and 1,897 are `S3` elements. Last, for each relevance grade, we calculate the proportion of `MA` elements that are also confirmed to belong to the same relevance grade by the additional assessor of the INEX 2005 topic. These numbers are then averaged across the five INEX topics. For example, for topic 234 the `E1` relevance grade has the highest level (0.81) of `MA` element agreement for exhaustivity (but almost zero agreement for `E?`), while two relevance grades for specificity, `S1` and `S3`, have almost perfect `MA` element agreement.

From the average numbers, we identify two cases where conclusions can be drawn: the case of the `E?` relevance grade, with the average of 1,194 `MA` elements, and the case of `S3` relevance grade, with the average number of 1,490 `MA` elements. We observe that (on average) only 58% of the `E?` `MA` elements are also confirmed to be `E?` by the additional assessors of the INEX 2005 topics. This confirms our previous conjecture that the assessors do not agree on the exact interpretation of too small. Conversely, on average 94% of the `S3` `MA` elements are also confirmed to be `S3`, indicating that assessors clearly agree on the highlighted relevant content. The agreements for the other four relevance grades (all above 50%) are more or less similar, however no conclusions can be drawn due to the relatively small average number of `MA` elements.

The results obtained from the analysis of the level of assessor agreement suggest that there is good reason for ignoring the exhaustivity dimension during evaluation, since it appears to be easier for the assessor to be consistent when highlighting relevant content than when choosing one of the three exhaustivity values. In the next section, we present an evaluation metric for XML retrieval that solely uses *specificity* to evaluate the XML retrieval effectiveness.

3 HiXEval — Highlighting XML Retrieval Evaluation

Our proposal for an alternative metric for XML retrieval is mainly motivated by the need to simplify the XML retrieval evaluation, as well as the need to use a metric that is conformant to the well-established evaluation measures used in traditional information retrieval. The `HiXEval` metric credits systems for retrieving elements that contain as much highlighted (relevant) textual information as possible, without also containing a significant amount of non-relevant information. To measure the extent to which an XML retrieval system returns relevant information, we only take into account the amount of highlighted text in a retrieved element, without considering the value of exhaustivity for that element. We propose to extend the traditional definitions of precision and recall as follows.

$$Precision = \frac{amount\ of\ relevant\ information\ retrieved}{total\ amount\ of\ information\ retrieved}$$

$$Recall = \frac{amount\ of\ relevant\ information\ retrieved}{total\ amount\ of\ relevant\ information}$$

Let $\mathbf{e}$ be an element that belongs to a ranked list of elements returned by an XML retrieval system. Three distinct scenarios are possible for this element:

1. $\mathbf{e}$ is a not-yet-seen element (NS);
2. $\mathbf{e}$ has previously been fully seen (FS), and
3. $\mathbf{e}$ is an element-part, that has been in part seen previously (PS).

Let $rsize(\mathbf{e})$ be the number of highlighted (relevant) characters. To measure the value of retrieving relevant information from $\mathbf{e}$ at rank $\mathbf{r}$, we define the relevance value function $\mathbf{rval_r}(\mathbf{e})$ as:

$$\mathbf{rval_r}(\mathbf{e}) = \begin{cases} rsize(\mathbf{e}) & if\ \mathbf{e}\ is\ NS \\ rsize(\mathbf{e}) - \alpha \cdot rsize(\mathbf{e}) & if\ \mathbf{e}\ is\ FS \\ rsize(\mathbf{e}) - \alpha \cdot \sum_{\mathbf{e}'} rsize(\mathbf{e}') & if\ \mathbf{e}\ is\ PS \end{cases}$$

where $\mathbf{e}'$ represents a previously retrieved element that at the same time is descendant of $\mathbf{e}$, which appears at rank higher than $\mathbf{r}$ (if any). The parameter α is a weighting factor that represents the importance of retrieving non-overlapping elements in the ranked list. By introducing α in the $\mathbf{rval_r}(\mathbf{e})$ function, different models of user behaviour can be represented. For example, setting α to 1 (`overlap=on`) models users that do not tolerate overlap, and ensures that the system will only be credited for retrieving relevant information that has not been previously retrieved by other overlapping elements. Conversely, setting α to 0 (`overlap=off`) models tolerant users and ensures that the system is always

credited for retrieving relevant information, regardless of whether the same information has previously been retrieved.

Let $size(\mathbf{e})$ be the total number of characters contained by $\mathbf{e}$, and **Trel** the total amount of relevant information for an INEX topic (if $\alpha = 1$, then **Trel** is the number of highlighted characters across all documents; if $\alpha \in [0, 1)$, then **Trel** is the number of highlighted characters across all elements). Let $\mathbf{i}$ be an integer that reflects the rank of an element, and $\mathbf{i} \in [1, r]$.

We measure the fraction of *retrieved relevant information* as:

$$\texttt{P@r} = \frac{1}{\mathbf{r}} \cdot \sum_{\mathbf{i=1}}^{\mathbf{r}} \frac{\mathbf{rval_i}(\mathbf{e})}{size(\mathbf{e})}$$

The `P@r` measure ensures that, to achieve a *precision* gain at rank $\mathbf{r}$, the retrieved element $\mathbf{e}$ needs to contain *as little non-relevant information as possible*.

We measure the fraction of *relevant information retrieved* as:

$$\texttt{R@r} = \frac{1}{\mathbf{Trel}} \cdot \sum_{\mathbf{i=1}}^{\mathbf{r}} \mathbf{rval_i}(\mathbf{e})$$

The `R@r` measure ensures that, to achieve a *recall* gain at rank $\mathbf{r}$, the retrieved element $\mathbf{e}$ needs to contain *as much relevant information as possible*.

In addition to the above measures, we also calculate values for `MAP` and `iMAP`, which represent mean average precision (calculated at natural recall levels), and interpolated mean average precision (calculated at standard 11 recall levels), respectively.

The two precision and recall values could be combined in a single value for a given rank $\mathbf{r}$ using the F-measure (the *harmonic mean*) as follows.

$$\texttt{F@r} = \frac{2 \cdot \texttt{P@r} \cdot \texttt{R@r}}{\texttt{P@r} + \texttt{R@r}}$$

By comparing the `F@r` values obtained from different systems, it would be possible to see which system is more capable of retrieving as much relevant information as possible, without also retrieving a significant amount of non-relevant information.

4 HiXEval versus XCG in XML Retrieval Experiments

In this section, we demonstrate the usefulness of `HiXEval` compared to the `XCG`-based metrics in XML retrieval experiments. More specifically, we make direct use of the INEX evaluation methodology — its desire to order XML retrieval runs to understand which retrieval techniques work well and which do not — to find how the run orderings generated by the `HiXEval` measures compare to the run ordering obtained when using measures from the `XCG` family of metrics.

Table 4. Spearman correlation coefficients calculated from the run orderings obtained from pairs of evaluation measures using the 55 submitted runs for the INEX 2005 `CO.Thorough` (upper part) and 44 runs for the `CO.Focussed` (lower part) retrieval strategies. Correlation scores between an evaluation measure from the `XCG` family of metrics and its corresponding measure from `HiXEval` are shown in bold

	nxCG			**ep/gr**	**HiXEval**						
Metric (measure)	10	25	50	MAep	P@10	R@10	P@25	R@25	P@50	R@50	MAP
`CO.Thorough` (`overlap=off`)											
nxCG (`nxCG[10]`)	1.00	0.94	0.91	0.82	**0.96**	0.31	0.91	0.28	0.88	0.34	0.83
nxCG (`nxCG[25]`)	0.94	1.00	0.97	0.86	0.90	0.35	**0.95**	0.33	0.94	0.39	0.85
nxCG (`nxCG[50]`)	0.91	0.97	1.00	0.92	0.85	0.40	0.92	0.37	**0.96**	0.43	0.89
ep/gr (`MAep`)	0.82	0.86	0.92	1.00	0.74	0.52	0.80	0.49	0.87	0.54	**0.94**
`CO.Focussed` (`overlap=on`)											
nxCG (`nxCG[10]`)	1.00	0.98	0.96	0.95	**0.92**	0.17	0.89	0.23	0.88	0.22	0.73
nxCG (`nxCG[25]`)	0.98	1.00	0.98	0.96	0.90	0.17	**0.91**	0.24	0.89	0.23	0.73
nxCG (`nxCG[50]`)	0.96	0.98	1.00	0.98	0.92	0.17	0.92	0.24	**0.93**	0.23	0.76
ep/gr (`MAep`)	0.95	0.96	0.98	1.00	0.91	0.18	0.92	0.25	0.93	0.24	**0.82**

We present results for all the retrieval strategies explored in the two INEX 2005 sub-tasks (`CO` and `CAS`).[2] From the `XCG` family of metrics, we use `nxCG` and `ep/gr` in our experiments. The `genLifted` quantisation function is used with both metrics, which means that the `E?` elements are included during evaluation [3]. We use the rank (Spearman) correlation coefficient to measure the extent to which the rank orderings obtained from a pair of measures correlate. We choose this primarily because, with non-parametric correlation using the Spearman coefficient, there is no need to assume that data in the pairs come from normal distributions. Values of the Spearman coefficient range from +1 (perfect positive correlation), through 0 (no correlation), to -1 (perfect negative correlation). All reported values are statistically significant ($p < 0.01$).

4.1 INEX 2005 `CO` Sub-Task

Three retrieval strategies are explored in the `CO` sub-task: `Thorough`, `Focussed`, and `FetchBrowse` [6]. We use different settings for each evaluation measure of the `nxCG`, `ep/gr` and `HiXEval` metrics depending on the strategy used. For example, for the `Focussed` strategy we use a setting which penalises runs that retrieve overlapping elements (`overlap=on`), whereas for the `Thorough` strategy a setting that ignores the overlapping retrieved elements is used (`overlap=off`).

[2] Results for retrieval strategies in the `+S` sub-task are not included, since they are similar to the results presented for the `CO` sub-task.

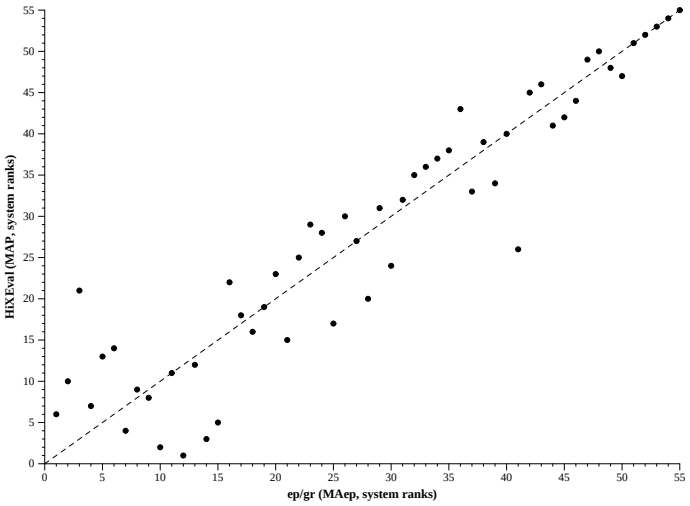

Fig. 2. Correlation between run orderings generated by `MAep` (`ep/gr`) and `MAP` (`HiXEval`) using the 55 submitted runs for the `CO.Thorough` retrieval strategy. The Spearman correlation coefficient is 0.94

`Thorough` **Retrieval Strategy** The upper part of Table 4 shows Spearman correlation coefficients calculated from the run orderings using the 55 submitted runs for the INEX 2005 `CO.Thorough` retrieval strategy. We observe that each of the three `nxCG` measures is strongly correlated to the corresponding precision measure of `HiXEval`. Interestingly, there is low correlation between the three `nxCG` measures and their corresponding recall measures in `HiXEval`. The table also shows coefficients when the measures from the `nxCG` and `ep/gr` metrics are compared with each other. The overall trend observed when pairs of these measures are compared (in terms of how well they correlate) is also observed when comparing the corresponding pairs of `XCG` and `HiXEval` measures. The Spearman coefficient shows that there is a strong correlation (0.94) between run orderings generated by `MAep` (`ep/gr`) and `MAP` (`HiXEval`) when comparing mean average precision. The graph of Fig. 2 provides a detailed overview of the observed correlation between these run orderings, showing that the biggest differences in rankings occur with the best performing systems.

The observed correlations between the corresponding measures of `HiXEval` and `XCG` (all greater than 0.9) show that similar run orderings are generated by the two metrics.

`Focussed` **Retrieval Strategy** The lower part of Table 4 shows Spearman correlation coefficients calculated from the run orderings using the 44 submitted runs for the INEX 2005 `CO.Focussed` retrieval strategy. The calculated correlation numbers for the three `nxCG` measures and their corresponding `HiXEval` precision measures are again greater than 0.9, with a similar trend to that ob-

Table 5. Spearman correlation coefficients calculated from the run orderings obtained from pairs of evaluation measures using the 31 correctly submitted runs for the INEX 2005 `CO.FetchBrowse-Article` (upper part) and `CO.FetchBrowse-Element` (middle and lower parts) retrieval strategies. Correlation scores between the `MAep` measure (`ep/gr`) and the `MAP` measure (`HiXEval`) are shown in bold

	ep/gr	**HiXEval**		
Metric (measure)	MAep	Prec	Rec	MAP
CO.FetchBrowse-Article (overlap=off,on)				
ep/gr (MAep)	1.00	0.69	0.70	**0.85**
CO.FetchBrowse-Element (overlap=off)				
ep/gr (MAep)	1.00	0.90	0.88	**0.95**
CO.FetchBrowse-Element (overlap=on)				
ep/gr (MAep)	1.00	0.80	0.92	**0.67**

served for the `CO.Thorough` strategy. However, the correlation coefficient is lower for this strategy (0.82) when comparing mean average precision. Unlike for the `CO.Thorough` strategy, there are strong correlations between `MAep` and each of the three *precision* measures of `HiXEval`, whereas there is almost no correlation between `MAep` and each of the three *recall* measures. This suggests that, for the `CO.Focussed` retrieval strategy, the methodology used in creating the *ideal recall-base* has an adverse effect on the overall recall, which seems to dramatically influence the run orderings obtained from `MAep` measure of the `ep/gr` metric.

`FetchBrowse` **Retrieval Strategy** The evaluation methodology for this retrieval strategy is different from those for the other two `CO` strategies in that two separate evaluation results are reported: an article-level result and an element-level result, the latter calculated by using both (`off` and `on`) overlap settings [3]. To obtain element-level results, in addition to `MAP` we report values obtained by the following two `HiXEval` measures: `Prec`, which measures precision at final rank for each article cluster, averaged over all clusters and then over all topics; and `Rec`, which measures recall at final rank for each article cluster, also averaged over all clusters and topics. To obtain article-level results with `HiXEval`, we used the article-derived runs along with their corresponding relevance judgements, which means that values for `Prec` and `Rec` refer to those for precision and recall at final cut-offs (1500), respectively.

Table 5 shows Spearman correlation coefficients calculated from the run orderings using the 31 correctly submitted runs for the INEX 2005 `CO.FetchBrowse` retrieval strategy. For article-level results, the calculated value for the Spearman coefficient between `MAep` and `MAP` is 0.85. The probable cause for this behaviour is that different methodologies are used by the two metrics to determine the pre-

Table 6. Spearman correlation coefficients calculated from the run orderings obtained from pairs of evaluation measures using 25 submitted runs for SSCAS, 23 runs for SVCAS and VSCAS, and 28 runs for VVCAS retrieval strategies. Correlation scores between an evaluation measure from the XCG family of metrics and its corresponding measure from HiXEval are shown in bold

	nxCG			**ep/gr**	**HiXEval**						
Metric (measure)	10	25	50	MAep	P@10	R@10	P@25	R@25	P@50	R@50	MAP
SSCAS (overlap=off)											
nxCG (nxCG[10])	1.00	0.97	0.75	0.69	0.82	**0.95**	0.62	0.98	0.60	0.92	0.58
nxCG (nxCG[25])	0.97	1.00	0.84	0.66	0.81	0.94	0.69	**0.97**	0.68	0.95	0.55
nxCG (nxCG[50])	0.75	0.84	1.00	0.57	0.80	0.79	0.92	0.77	**0.91**	0.90	0.53
ep/gr (MAep)	0.69	0.66	0.57	1.00	0.74	0.64	0.64	0.62	0.66	0.70	**0.96**
SVCAS (overlap=off)											
nxCG (nxCG[10])	1.00	0.98	0.98	0.94	**0.98**	0.94	0.91	0.92	0.94	0.92	0.93
nxCG (nxCG[25])	0.98	1.00	0.99	0.94	0.97	0.93	**0.94**	0.93	0.96	0.94	0.93
nxCG (nxCG[50])	0.98	0.99	1.00	0.95	0.97	0.92	0.93	0.93	**0.96**	0.94	0.92
ep/gr (MAep)	0.94	0.94	0.95	1.00	0.93	0.86	0.84	0.88	0.89	0.91	**0.95**
VSCAS (overlap=off)											
nxCG (nxCG[10])	1.00	0.97	0.96	0.83	**0.98**	0.76	0.95	0.74	0.94	0.71	0.88
nxCG (nxCG[25])	0.97	1.00	0.99	0.86	0.96	0.73	**0.98**	0.72	0.97	0.71	0.91
nxCG (nxCG[50])	0.96	0.99	1.00	0.86	0.94	0.75	0.97	0.75	**0.97**	0.74	0.92
ep/gr (MAep)	0.83	0.86	0.86	1.00	0.81	0.59	0.88	0.64	0.86	0.63	**0.90**
VVCAS (overlap=off)											
nxCG (nxCG[10])	1.00	0.93	0.90	0.75	**0.96**	0.57	0.93	0.56	0.91	0.59	0.73
nxCG (nxCG[25])	0.93	1.00	0.97	0.85	0.92	0.62	**0.95**	0.65	0.98	0.67	0.84
nxCG (nxCG[50])	0.90	0.97	1.00	0.90	0.87	0.72	0.90	0.74	**0.95**	0.76	0.91
ep/gr (MAep)	0.75	0.85	0.90	1.00	0.72	0.72	0.75	0.76	0.84	0.75	**0.91**

ferred article answers; indeed, the ep/gr metric uses knowledge of the highest scoring element within an article to obtain the ordering of the ideal article gain vector, whereas articles are inspected on their own merit by HiXEval.

Table 5 also shows that, for element-level results, the overlap setting dramatically changes the observed level of correlation between the rank orderings of the two metrics. With overlap set to off, there is a strong correlation between the two mean average precision measures. With overlap set to on there is little correlation between MAep and MAP; however, we observe that in this case MAep is better correlated with recall (0.92) than with precision (0.80). We believe that the probable cause for this behaviour is that, unlike the case of the CO.Focussed retrieval strategy where overlap is also set to on, here the number of relevant elements that comprise the ideal recall-base for each *article cluster* is much smaller,

which in turn makes it easier for runs to achieve perfect recall for a given cluster. The small correlation value between MAep and MAP, however, suggests that the two metrics could have differently implemented the mean average precision measure for this overlap setting.

4.2 CAS Sub-Task

Four retrieval strategies are explored in the CAS sub-task: SSCAS, SVCAS, VSCAS, and VVCAS; these differ in the way the target and support elements of a CAS topic are interpreted [6]. We use the overlap=off setting for each evaluation measure of nxCG, ep/gr, and HiXEval.

Table 6 shows Spearman correlation coefficients calculated from the run orderings using different numbers of submitted runs for each of the four INEX 2005 CAS retrieval strategies. We observe that there is a strong correlation between the two metrics for the CAS sub-task, irrespective of the retrieval strategy used. However, the observed correlation between each of the three measures of nxCG with the two precision and recall measures of HiXEval changes depending on the way the target element is interpreted. For the two *strict* CAS retrieval strategies (SS and SV), nxCG seems to be more recall- than precision-oriented, whereas for the *vague* CAS retrieval strategies the reverse is true. We suspect that, as with CO.FetchBrowse-Element retrieval strategy, the fewer relevant elements comprising the recall-base for the two strict CAS strategies may have an impact on the evaluation methodology of the nxCG metric.

5 Conclusions and Future Work

HiXEval addresses many of the concerns that have been raised in connection with the XCG-based metrics. Its main features are simplicity, compatibility with the well-understood measures used in traditional information retrieval, ability to model different user behaviours, and most importantly, minimal reliance on subjective decisions during evaluation. Indeed, if there was broad acceptance of HiXEval, there would be no need for assessors to judge exhaustivity, as only highlighting of relevant passages would be required. This would substantially reduce the time taken to undertake the relevance judgements.

The HiXEval metric is based solely on the specificity dimension, which, as we have shown through our analysis of the level of assessor agreement of the five INEX 2005 topics, is much better interpreted by assessors than the definition of the exhaustivity dimension. Moreover, our correlation analysis of the rank orderings between HiXEval and the two XCG-based metrics has confirmed that both metrics perform broadly the same task, and thus measure the same (or similar) retrieval behaviour.

The correlation analysis has also identified the different *orientations* of the XCG-based metrics; indeed, regardless of whether the level of overlap among retrieved elements is considered or not, in the case where the number of the so-called *ideal* retrieval elements is rather small, the XCG metrics seem to be more

recall- than precision-oriented. Conversely, with a sufficient number of ideal retrieval elements in the recall-base, the two metrics are clearly precision-oriented.

In the future we intend to check the reliability and stability of `HiXEval` and the two `XCG` metrics. We plan to undertake reliability tests for the `HiXEval` metric, similar to the ones performed for `XCG` and the INEX-2002 metrics [5]. To test stability, we plan to measure significance and error rates by pursuing a simplification of the methodology used by Sanderson and Zobel [9]. We also plan to further investigate the observed differences on the best performing systems.

Acknowledgements

We thank Saied Tahaghoghi and the anonymous reviewer for their comments.

References

1. D. Hiemstra and V. Mihajlovic. The Simplest Evaluation Measures for XML Information Retrieval that Could Possibly Work. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 6–13, Glasgow, UK, 2005.
2. K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20:422–446, 2002.
3. G. Kazai and M. Lalmas. INEX 2005 evaluation metrics. In *INEX 2005 Workshop Pre-Proceedings, Dagstuhl, Germany, November 28–30, 2005*, pages 401–406, 2005.
4. G. Kazai and M. Lalmas. Notes on what to measure in INEX. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 22–38, Glasgow, UK, 2005.
5. G. Kazai, M. Lalmas, and A. P. de Vries. Reliability tests for the XCG and INEX-2002 metrics. In *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 *LNCS*, pages 60–72, May 2005.
6. M. Lalmas. INEX 2005 retrieval task and result submission specification. In *INEX 2005 Workshop Pre-Proceedings, Dagstuhl, Germany, November 28–30, 2005*, pages 385–390, 2005.
7. 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.
8. 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.
9. M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 162–169, Salvador, Brazil, 2005.
10. A. Trotman. Wanted: Element Retrieval Users. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 63–69, Glasgow, UK, 2005.
11. E. M. Voorhees. Variations in relevance judgements and the measurment of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.