Hybrid XML Retrieval Revisited

Jovan Pehcevski[1], James A. Thom[1], S. M. M. Tahaghoghi[1], and Anne-Marie Vercoustre[2]

[1] School of CS and IT, RMIT University, Melbourne, Australia
{jovanp, jat, saied}@cs.rmit.edu.au
[2] INRIA, Rocquencourt, France
anne-marie.vercoustre@inria.fr

Abstract. The widespread adoption of XML necessitates structure-aware systems that can effectively retrieve information from XML document collections. This paper reports on the participation of the RMIT group in the INEX 2004 ad hoc track, where we investigate different aspects of the XML retrieval task. Our preliminary analysis of CO and VCAS relevance assessments identifies three XML retrieval scenarios: *Original*, *General* and *Specific*. Further analysis of the relevance assessments under the General retrieval scenario reveals two categories of CO and VCAS topics: *Broad* and *Narrow*. We design runs that follow a hybrid XML approach and implement two retrieval heuristics with different levels of overlap among the answer elements. For the Original retrieval scenario we show that the overlap CO runs outperform the non-overlap CO runs, and the VCAS run that uses queries with structural constraints and no explicitly specified target element performs best. In both CO and VCAS cases, runs that implement the retrieval heuristic that favours less specific over more specific answer elements produce most effective retrieval. Importantly, we present results which show that, for the General retrieval scenario where users prefer less specific and non-overlapping answers to their queries, the choice of using a plain full-text search engine is a very effective choice for XML retrieval.

1 Introduction

Two types of retrieval topics are explored in the INEX 2004 ad hoc track: Content-Only (CO) topics and Vague Content-And-Structure (VCAS) topics. Forty CO topics are used in the CO sub-track, while thirty-five VCAS topics are investigated in the VCAS sub-track.

CO topics do not refer to the existing document structure. An XML retrieval system using these topics may return elements with varying sizes and granularity, prompting a revisit of the issue of *length normalisation* for XML retrieval [4]. Moreover, a large proportion of overlapping result elements may be expected, since the same textual information in an XML document is often contained by more than one element. This *overlap problem* is particularly apparent during evaluation, where the "overpopulated and varying recall base" [6] contains a substantial number of mutually overlapping elements.

Strict Content-And-Structure (SCAS) topics enforce restrictions on the existing document structure and explicitly specify the target element (such as article, section, or paragraph). The structural conditions in a VCAS topic, however, need not be strictly matched. This means that not only are the restrictions on document structure *vague*, but also that the target element could be any element considered *relevant* to the information need. Thus, the same retrieval strategies for CO topics may also be used for VCAS topics, since CO topics may be considered as *loosely restricted* VCAS topics.

We undertake a preliminary analysis of the INEX 2004 CO and VCAS relevance assessments to identify the types of highly relevant elements. Arising from our analysis we identify many cases where, for a particular CO/VCAS topic and an XML document, several layers of elements in the document hierarchy (such as `article`, `bdy`, `sec` and `ss1`) have all been assessed as highly relevant. It then follows that this overlap problem is not only an evaluation problem, but it is also a serious retrieval problem, since the choice of the *preferable units of retrieval* for a CO/VCAS topic becomes a non-trivial one. For instance, given an overlapping recall base, an XML retrieval system that returns overlapping answer elements is likely to exhibit better performance than a system that returns non-overlapping answers. However, the former system will obviously retrieve and present a substantial amount of redundant information, which raises the question: is this what users really want?

Different evaluation metrics — which typically aim at modelling different user behaviours — have been proposed for XML retrieval, but only some of them attempt to address the overlap problem [6]. To investigate this and other similar aspects of XML retrieval, from the above analysis we distinguish between three *scenarios* of XML retrieval: the *Original* retrieval scenario, where all the highly relevant (and possibly mutually overlapping) elements are considered; the *Specific* retrieval scenario, where only the *most specific* highly relevant elements are considered; and the *General* retrieval scenario, where only the *least specific* highly relevant elements are considered. Unlike the Original retrieval scenario, the latter two scenarios allow for non-overlapping recall base. Indeed, in the absence of more realistic user models for XML retrieval, the Specific retrieval scenario reflects users that prefer specific, more focused answers for their queries, whereas the General retrieval scenario models users that prefer more encompassing answers for their queries.

Further analysis of the CO and VCAS relevance assessments under the General retrieval scenario reveals two categories of retrieval topics, which we call *Broad* and *Narrow*. We observed in our previous work that an XML retrieval system appears to behave differently when its performance is measured against different categories of CO topics [8]. Indeed, this has also been experimentally shown to be true for a fragment-based XML retrieval system [3]. Thus, distinguishing between different categories of topics — whether it applies to CO or VCAS — is likely to be useful information during retrieval.

The system we use for the ad hoc track in INEX 2004 follows a *hybrid XML approach*, utilising the best features of Zettair[3] (a full-text search engine) and eXist[4] (a native XML database). The hybrid approach is a "fetch and browse" [1] retrieval approach, where full articles considered likely to be relevant to a topic are first retrieved by Zettair (the *fetch* phase), and then the most specific elements within these articles are extracted by eXist (the *browse* phase) [8].

The above approach resulted in rather poor system performance for the CO topics in INEX 2003, where Zettair performed better than our initial hybrid system. We have since developed a retrieval module that utilises the structural information in the eXist list of answer elements, and identifies and ranks *Coherent Retrieval Elements* (CREs). The hybrid system with the CRE module more than doubles the retrieval effectiveness of Zettair [8]. We show elsewhere that this hybrid-CRE system also produces performance improvements for the INEX 2003 SCAS topics [7]. Different retrieval heuristics may be used by the CRE module, mainly to determine the final rank of each CRE.

For the INEX 2004 CO sub-track, we use our hybrid system to explore which CRE retrieval heuristic yields the best retrieval performance, and to investigate whether — under different retrieval scenarios and topic categories — having non-overlapping answer elements has an impact on system performance.

For the INEX 2004 VCAS sub-track, we also investigate which retrieval choice — plain queries; queries with structural constraints and no explicitly specified target element; or queries with both structural constraints and a target element — results in more effective VCAS retrieval. Different retrieval scenarios and topic categories are also used for this investigation.

The remainder of this paper is organised as follows. In Section 2 we present our analysis of the INEX 2004 relevance assessments, both for the CO and the VCAS retrieval topics. A detailed description of the runs we consider for the CO and the VCAS sub-tracks is provided in Section 3. In Section 4 we present the evaluation results of our CO and VCAS runs. These results reflect different retrieval scenarios, which are based on our analysis of the INEX 2004 relevance assessments. We conclude in Section 5 with a brief discussion of our findings.

2 Analysis of INEX 2004 relevance assessments

Some names in the XML document collection include: `article` for a full article; `abs` and `bdy` for article abstract and article body; `sec`, `ss1` and `ss2` for section and subsection elements; and `p` and `ip1` for paragraph elements. Analysing the INEX 2004 CO and VCAS relevance assessments, we observe that since neither case restricts the answer elements, the final answer list may contain elements of different types and of varying sizes and granularity. We expect that `article` elements may represent preferable answers for some topics, while for other topics more specific elements may be preferable over `article` elements.

[3] `http://www.seg.rmit.edu.au/zettair/`
[4] `http://exist-db.org/`

```
<file file="ic/2000/w4036">
 <path path="/article[1]" E="3" S="3"/>
 . . . . .
 <path path="/article[1]/bdy[1]" E="3" S="3"/>
 . . . . .
 <path path="/article[1]/bdy[1]/sec[3]" E="3" S="3"/>
 <path path="/article[1]/bdy[1]/sec[3]/ss1[1]" E="3" S="3"/>
 <path path="/article[1]/bdy[1]/sec[3]/ss1[2]" E="3" S="3"/>
 <path path="/article[1]/bdy[1]/sec[3]/ss1[3]" E="3" S="3"/>
 . . . . .
 <path path="/article[1]/bdy[1]/sec[4]" E="3" S="3"/>
 <path path="/article[1]/bdy[1]/sec[4]/ss1[2]" E="3" S="3"/>
 . . . . .
</file>
```

Fig. 1. An extract from the INEX 2004 CO relevance assessments (CO topic 176)

2.1 CO relevance assessments

Figure 1 shows an extract from the INEX 2004 CO relevance assessments for the CO topic 176. Values for the two INEX relevance dimensions, *exhaustivity*[5] (how many aspects of the topic are covered in the element), and *specificity*[6] (how specific to the topic is the element), are assigned to an `article` and elements within `article` for assessing their relevance to a CO topic.

In our analysis we focus on *highly relevant* elements. For a given topic, these are elements that have been assessed as both highly exhaustive and highly specific (E3S3) elements. In Fig. 1 there are eight such elements, including the article itself. These answer elements represent the most useful retrieval elements, even though there is a substantial amount of overlap between them. Following our previous analysis of INEX 2003 relevance assessments [8], we identify two distinct types of highly relevant elements: *General* and *Specific*. Unlike the INEX definitions for exhaustivity and specificity, the definitions for General and Specific elements result from our analysis as follows [8].

General:

> "For a particular article in the collection, a *General* element is the least-specific highly relevant element containing other highly relevant elements". Based on this definition, `article[1]` is the only General element in the example of Fig. 1. However, an article may contain several General elements if the article as a whole is not highly relevant. Figure 2 shows a tree representation of all the highly relevant elements shown in Fig. 1. The General element is the element shown in the ellipse.

[5] E represents the level of exhaustivity (values between 0-3)
[6] S represents the level of specificity (values between 0-3)

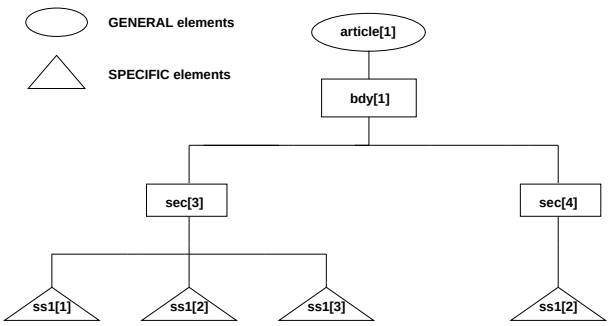


Fig. 2. A tree-view example of GENERAL versus SPECIFIC elements.

Specific:

"For a particular article in the collection, a *Specific* element is the most-specific highly relevant element contained by other highly relevant elements". In Fig. 2, the Specific elements are the elements shown in triangles.

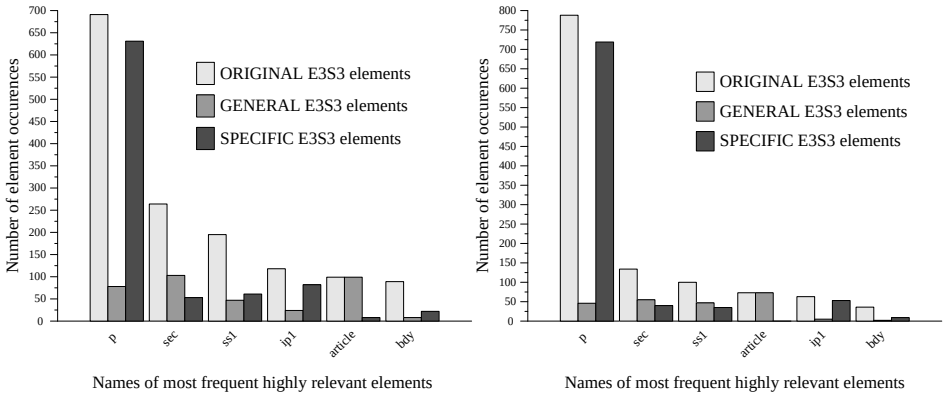
When there is only one highly relevant element in an article, that element is both a General and a Specific element.

There are 40 CO topics in INEX 2004 (numbers 162–201). We use version 3.0 of the INEX 2004 relevance assessments, where 34 of the 40 CO topics have their relevance assessments available. Of these, 9 topics do not contain highly relevant (E3S3) elements. Consequently, a total of 25 CO topics are used.

In the following analysis, we focus on those highly relevant elements that appear in more than half the CO topics (that is, elements that appear in 12 or more CO topics). We choose this because we want to eliminate the outlier elements that may occur very frequently, but these occurrences are distributed across a few CO topics (such as 297 occurrences distributed across 6 topics for the `it` element). Figure 3(a) shows the frequency of highly relevant elements (including full articles) that appear in more than half the CO topics. The figure shows three distinct scenarios: the *Original* retrieval scenario, where all highly relevant elements are considered; the *General* retrieval scenario, where only General elements are considered, and the *Specific* retrieval scenario, where only Specific elements are considered. The x-axis contains the names of the six most frequent highly relevant elements (under the Original retrieval scenario). The y-axis contains the number of occurrences of each element.

Under the Original retrieval scenario, `p` and `sec` elements occur most frequently, with 691 and 264 overall occurrences, respectively. The `ss1` and `ip1` elements come next, followed by `article` and `bdy` with 99 and 89 occurrences. The latter suggests that in most cases when a `bdy` element was assessed as highly relevant, the parent `article` is also likely to have been assessed as highly relevant too.

Under the General retrieval scenario, `sec` elements are most frequent with 103 overall occurrences, followed by `article` elements with 99 occurrences; however,



(a) CO topics (b) VCAS topics

Fig. 3. Number of occurrences of highly relevant elements that appear in more than half the INEX 2004 CO and VCAS topics, for three distinct retrieval scenarios. In each CO/VCAS case, the Original retrieval scenario is used to determine the element ordering.

the `article` occurrences are distributed across 16 topics, whereas there are 15 topics where `sec` elements occur. There are 8 cases (occurring across 6 topics) where a `bdy` was assessed as highly relevant, but the parent `article` was *not* assessed as highly relevant.

The last scenario shown in Fig. 3(a) is the Specific retrieval scenario. As expected, the situation changes here in favour of the more specific elements, with `p` elements being most frequent. The `ip1`, `ss1`, `sec` and `bdy` elements come next, followed by only 8 occurrences of `article` elements. The 8 occurrences are distributed across 4 topics, where these `article` elements were the most specific elements assessed as highly relevant.

The above statistics provide an interesting insight of what might happen when the performance of an XML retrieval system is evaluated against three distinct XML retrieval scenarios. For instance, under the General retrieval scenario one would expect that a full-text search engine could solely be used for effective XML retrieval, given that the full article is the second most frequent highly relevant element. The above information may therefore be appropriately utilised by XML retrieval systems, particularly because distinct retrieval scenarios favour different types of highly relevant elements.

Topic categories In the following analysis we consider the General retrieval scenario. Our aim is to distinguish those CO topics that are mostly about less specific highly relevant elements (such as `article` and `bdy`), from those that are mostly about more specific highly relevant elements (such as `sec` and `p`). Consider Fig. 4: a point on this graph represents a CO topic. The x-axis shows the total number of General `article` and `bdy` elements for a CO topic, whereas the y-axis shows the total number of General elements other than `article` and `bdy`.

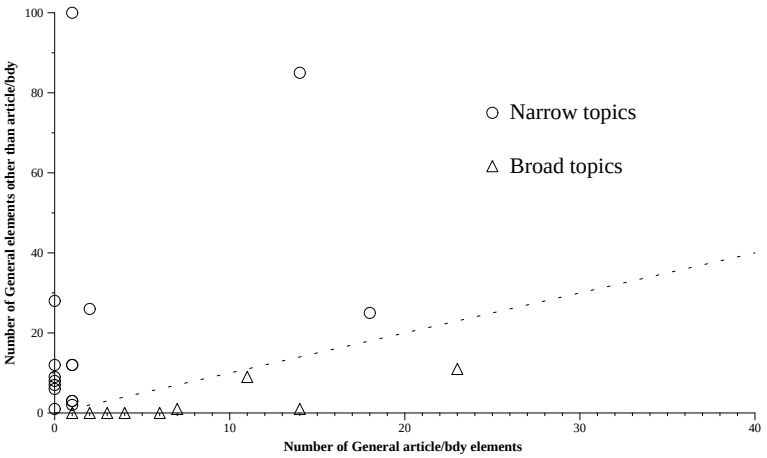


Fig. 4. Categories of INEX 2004 CO topics under the General retrieval scenario.

For example, the CO topic 183 depicted at coordinates (23,11) contains 23 General `article/bdy` elements and 11 General elements other than `article/bdy`.

We use this graph to identify two categories of INEX 2004 CO topics. Topics in the first category, shown as triangles on the graph and located below the dashed line, favour less specific elements as highly relevant answers. There are 9 such topics (numbers 164, 168, 175, 178, 183, 190, 192, 197 and 198). We refer to these as *Broad* topics.

Topics in the second category, shown as circles on the graph, favour more specific elements as highly relevant answers. There are 16 such topics. We refer to these as *Narrow* topics.

The above topic categorisation cannot easily be derived under the other two scenarios, that is, under either the Original or the Specific retrieval scenario. However, we also observed that four Broad topics (numbers 168, 178, 190 and 198) clearly belong to the Broad category even under these two scenarios.

2.2 VCAS relevance assessments

There are 35 VCAS topics in INEX 2004 (numbers 127-161). We use version 3.0 of the INEX 2004 relevance assessments, where 26 (out of 35) VCAS topics have their relevance assessments available. Of these, 4 topics do not contain highly relevant (E3S3) elements, and so we limit our analysis to a total of 22 VCAS topics.

Figure 3(b) shows the frequency of highly relevant elements that appear in more than half the VCAS topics. The figure also shows the three distinct scenarios: Original, General and Specific.

Since the VCAS relevance assessments have been done in much the same way as those for the CO topics, it is not surprising that Fig. 3(a) and Fig. 3(b) are similar. In both assessment cases, the frequency of `p` elements (under the

Original and Specific retrieval scenarios) is far greater than that of all the other elements. However, the frequencies of `article` and `bdy` elements in the VCAS case differ from the frequencies of the same elements in the CO case. Under the VCAS Original retrieval scenario, the number of `article` elements is much greater than that of the `bdy` elements (73 `article` occurrences compared to 36 `bdy` occurrences). Under the VCAS General retrieval scenario, the `article` elements are the most frequent highly relevant elements. Under the VCAS Specific retrieval scenario, the number of `article` elements is zero, whereas there are 9 highly relevant `bdy` elements (distributed across 3 topics).

Topic categories As for the CO topics, we use the General retrieval scenario to identify two categories of INEX 2004 VCAS topics. Topics in the first category favour less specific elements as highly relevant answers. There are 6 such topics (numbers 130, 131, 134, 137, 139 and 150), which we refer to as *Broad* topics. Topics in the second category favour more specific elements as highly relevant answers. There are 16 such topics, referred to as *Narrow* topics.

An interesting observation is that only two out of six VCAS topics of the Broad category (137 and 139) explicitly ask for retrieving `article` or `bdy` elements in their titles (that is, these elements represent their *target* elements). This is not the case with the other four Broad topics, where two topics ask for `sec` (134 and 150), one asks for `abs` (131), and one asks for `p` (130).

The above analysis clearly shows that highly relevant elements for VCAS topics do not necessarily represent target elements. We believe that distinguishing between categories of VCAS topics is, similarly as for the CO topics, important information that an XML retrieval system should utilise.

3 Runs description

The following sections provide a detailed description of our runs for each (CO and VCAS) sub-track.

3.1 Background

Most of the runs we consider for the INEX 2004 ad hoc track use a system that follows a hybrid XML retrieval approach. The system implements the best retrieval features from Zettair and eXist [8]. To further increase the system's retrieval effectiveness, an additional module that identifies and ranks Coherent Retrieval Elements (CREs) is used.

A CRE is defined as follows. The list of matching elements, extracted by eXist, is an article-ordered list. This list is processed by considering a pair of matching elements, starting from the first element down to the second last. In each step, a CRE is identified as the most specific ancestor of the two matching elements that constitute the pair [8]. To determine the ranks of CREs in the final answer list, the CRE module in our system uses a combination of the following XML-specific retrieval heuristics:

1. The number of times a CRE appears in the absolute path of each matching element in the eXist answer list — more matches (`M`) or fewer matches (`m`);
2. The length of the absolute path of the CRE, taken from the root element — longer path (`P`) or shorter path (`p`); and
3. The ordering of the XPath sequence in the absolute path of the CRE — nearer to the beginning (`B`) or nearer to the end (`E`).

There are 16 possible CRE heuristic combinations, since the first two heuristics can be applied in any order, and the third heuristic is complementary to the other two and is always applied at the end. We have found that for the INEX 2003 test set, the best results are obtained when using the `MpE` heuristic combination [8]. With `MpE`, less specific elements are ranked higher than more specific elements.

However, we have also observed that different CRE heuristic combinations may be more suitable for different XML retrieval scenarios, where retrieving more specific elements early in the ranking (such as with using the `PME` heuristic) produces better results. We implement and compare these two retrieval heuristics in different runs for the ad hoc track in INEX 2004.

3.2 CO sub-track

For the CO sub-track we consider the following runs.

- `Zettair` – using the full-text search engine as a baseline run.
- `Hybrid_MpE` – using the hybrid system with the `MpE` heuristic combination in the CRE module.
- `Hybrid_MpE_NO` – using the hybrid system, with the `MpE` heuristic combination, and no overlap among the elements in the final answer list.
- `Hybrid_PME` – using the hybrid system with the `PME` heuristic combination in the CRE module.
- `Hybrid_PME_NO` – using the hybrid system, with the `PME` heuristic combination, and no overlap among the elements in the final answer list.

Our goals are threefold. First, we aim to explore which heuristic combination yields the best performance for the hybrid system under different retrieval scenarios. Second, we aim to investigate the impact of overlapping answer elements on system performance. Thus, the two cases of non-overlap runs, `Hybrid_MpE_NO` and `Hybrid_PME_NO`, implement different non-overlap strategies: the former allows less specific elements to remain in the list and removes all the other (contained) elements, whereas the latter retains more specific elements, and removes all the other (encompassing) elements. Finally, by comparing the hybrid runs with the baseline run, we aim to better understand the issues surrounding the CO retrieval task.

3.3 VCAS sub-track

For the VCAS sub-track we consider the following runs.

- `Zettair` – using the full-text search engine as a baseline run.
- `Hybrid_CO_MpE` – using the hybrid system with the `MpE` heuristic combination in the CRE module. The structural constraints and the target element of each VCAS topic are removed, leaving only plain query terms.
- `Hybrid_CO_PME` – using the hybrid system with the `PME` heuristic combination in the CRE module. As with the previous run, each VCAS topic is treated as a CO topic.
- `Hybrid_VCAS_MpE` – using the hybrid system with the `MpE` heuristic combination in the CRE module. The target element of each VCAS topic is not explicitly specified (that is, it is allowed to have any granularity), while the structural constraints are strictly matched.
- `Hybrid_VCAS_PME` – using the hybrid system with the `PME` heuristic combination in the CRE module. As with the previous run, the structural constraints remain, while the target element is allowed to represent any element.
- `Hybrid_CAS` – using the initial hybrid system (without the CRE module), where the structural constraints and the target element of each VCAS topic are strictly matched.

We aim to achieve several goals through these VCAS runs. First, we aim to investigate which query choice (`CO`, `VCAS` or `CAS`) results in more effective VCAS retrieval. Second, for the hybrid runs using the CRE module and a particular query choice, we aim to identify the best choice of retrieval heuristic. Finally, by comparing the hybrid runs with the baseline run, we wish to empirically determine whether we can justify using a plain full-text search engine in the VCAS retrieval task.

4 Experiments and results

In INEX 2004, an evaluation metric with different quantisation functions is used to evaluate the retrieval effectiveness of XML systems [5]. Since our focus is on highly relevant elements, we use the strict quantisation function (E3S3) in our experiments.

For each of the retrieval runs, the resulting answer list for a CO/VCAS topic comprises up to 1500 articles or elements within articles. To measure the overall performance of each run, two standard information retrieval measures are used with the strict quantisation function: *Mean Average Precision* (MAP), which measures the ability of a system to return highly relevant (E3S3) elements, and *Precision at 10* (P@10), which measures the number of highly relevant (E3S3) elements within the first 10 elements returned by a system. In the following we describe results obtained by evaluating the retrieval effectiveness of our runs — under different retrieval scenarios — for each CO and VCAS sub-track.

CO run	%Ovp	Original MAP	Original P@10
Zettair	0	0.049	0.073
Hybrid_MpE	82.2	**0.124**	**0.103**
Hybrid_MpE_NO	0	0.051	0.076
Hybrid_PME	82.1	0.081	0.100
Hybrid_PME_NO	0	0.047	0.088

(a) CO runs

VCAS run	%Ovp	Original MAP	Original P@10
Zettair	0	0.052	0.119
Hybrid_CO_MpE	78.3	0.101	0.104
Hybrid_CO_PME	78.2	0.034	0.096
Hybrid_VCAS_MpE	67.8	**0.103**	**0.154**
Hybrid_VCAS_PME	67.8	0.045	0.142
Hybrid_CAS	5.4	0.032	0.142

(b) VCAS runs

Table 1. Performance results of INEX 2004 CO and VCAS runs when using the strict quantisation function and the Original retrieval scenario. For each run, an overlap indicator shows the percentage of overlapping elements in the answer list. Values for the best runs are shown in bold.

4.1 CO sub-track

Original CO retrieval scenario Table 1(a) shows evaluation results for the CO retrieval runs under the Original retrieval scenario. Values for the best runs are shown in bold. Several observations can be drawn from these results.

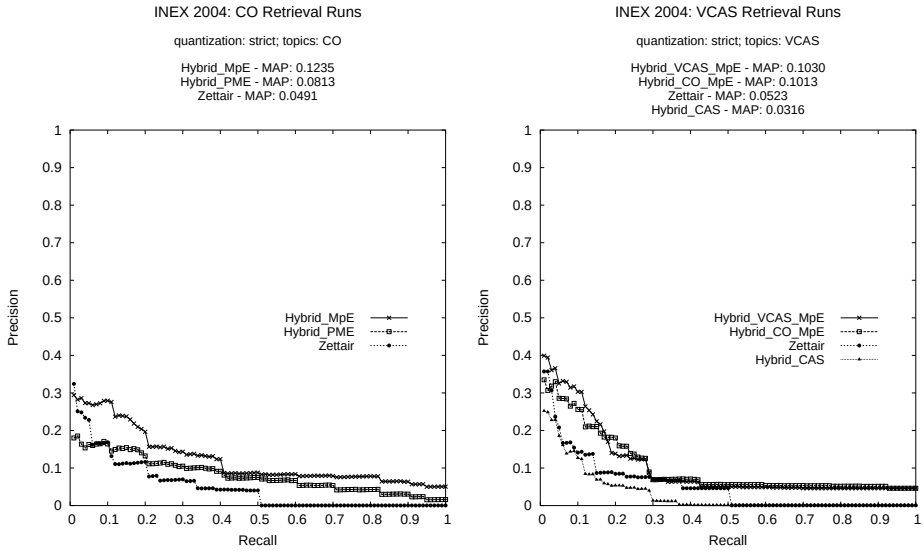
First, for overlap runs using the hybrid system, the MpE heuristic yields better performance than the PME heuristic. This result shows that under the Original CO retrieval scenario, systems that prefer retrieving less specific over more specific answer elements yield better performance.

Second, the non-overlap hybrid runs perform worse than the corresponding overlap hybrid runs. This is very likely to be a result of the "overpopulated" CO recall base, and reflects the inability of the strict quantisation function to cope with the overlap problem. We revisit the latter comparison in the next section (the General CO retrieval scenario), where a non-overlapping recall base is considered for evaluation.

Last, all the hybrid runs perform better on average than the baseline run. However, we observe that the baseline run is very competitive with the non-overlap hybrid runs, and, when the MAP measure is used for evaluation it even performs better than the non-overlap Hybrid_PME run. Since the answer list of the non-overlap Hybrid_PME run contains more specific (and non-overlapping) elements, the last result again confirms that retrieving more specific answer elements leads to poor system performance under this retrieval scenario.

The graph in Fig. 5(a) shows recall/precision curves for the two overlap hybrid runs and the baseline run. For low recall (0.1 and less), Zettair outperforms Hybrid_PME, although its performance gradually decreases and reaches zero for 0.5 (and higher) recall. Overall, Hybrid_MpE performs best and is substantially better than Hybrid_PME.

General CO retrieval scenario In the following analysis, we use the General CO retrieval scenario to compare the performance of the two Hybrid_MpE runs (overlap and non-overlap) with Zettair (the baseline run).



(a) CO runs (b) VCAS runs

Fig. 5. Evaluation of the INEX 2004 CO and VCAS retrieval runs when using the strict quantisation function and the Original retrieval scenario.

The General retrieval scenario reflects a non-overlapping recall base, since the relevance assessments allow an article to only contain General elements. Moreover, our previous analysis has also distinguished different categories of CO topics. Thus, the performance of the above runs are also compared across three topic categories: the *All topics* category, with all the 25 CO topics, and the *Broad* and the *Narrow* categories, with 9 and 16 CO topics, respectively.

Table 2 shows the evaluation results for each run. Two observations are clear in the cases of *All* and *Broad* topic categories: first, with both `MAP` and `P@10` measures `Zettair` performs best, although with `P@10` the non-overlap hybrid run (`MpE_NO`) performs the same as `Zettair`; and second, unlike for the case of overpopulated recall base (the Original retrieval scenario), the non-overlap hybrid run substantially outperforms the overlap hybrid run. These results show that, when a non-overlapping CO recall base is used for evaluation, the strict quantisation function can safely be used to reliably evaluate XML retrieval systems. Thus, systems that return overlapping answer elements (or redundant information) perform worse than systems that return non-overlapping answer elements. More specifically, the choice of using a full-text search engine results in very effective XML retrieval under this scenario.

In the case of *Narrow* topic category, the overlap hybrid run performs best, whereas the performance of the other two runs is the same. The latter result shows that a different topic category needs a different choice of optimal retrieval parameters.

		General					
		All topics		Broad topics		Narrow topics	
CO run	%Ovp	MAP	P@10	MAP	P@10	MAP	P@10
Zettair	0	**0.154**	**0.073**	**0.364**	**0.211**	0.036	0.024
Hybrid_MpE	82.2	0.126	0.050	0.240	0.056	**0.062**	**0.048**
Hybrid_MpE_NO	0	0.152	**0.073**	0.359	**0.211**	0.036	0.024

Table 2. Performance results of three INEX 2004 CO runs when using the strict quantisation function and different CO topic categories. The General retrieval scenario is used. For each run, an overlap indicator shows the percentage of overlapping elements in the answer list. Values for the best runs are shown in bold.

4.2 VCAS sub-track

Original VCAS retrieval scenario Table 1(b) shows evaluation results for the VCAS retrieval runs under the Original retrieval scenario. Values for the best runs are shown in bold. Several observations can be drawn from these results.

First, the Hybrid_CAS run (where structural constraints and the target element of a VCAS topic are strictly matched) performs worse than the other hybrid runs. Of these, the Hybrid_VCAS runs (the choice of strict structural constraints and no explicit target element) perform better than the Hybrid_CO runs (the choice where plain text queries are used). The former results can partly be explained from our analysis of the VCAS relevance assessments (see Section 2.2), which showed that highly relevant elements for VCAS topics do not necessarily represent their target elements. The latter results, however, show that the choice to strictly follow the structural constraints in the VCAS topics results in more effective retrieval than the choice of using only plain text queries.

Second, as with CO topics the MpE heuristic in the hybrid runs yields better performance than the PME heuristic. This shows that even with VCAS topics retrieving less specific over more specific answer elements is better.

Last, the hybrid runs perform better overall than the baseline run, except when using the MAP measure, where Zettair performs better than Hybrid_CAS and the two hybrid-PME runs. These results again confirm that, under the Original VCAS retrieval scenario, systems that prefer retrieving more specific answer elements and explicitly specify the target element in their queries exhibit poor performance.

The graph in Fig. 5(b) shows recall/precision curves for the three hybrid runs (CO, VCAS and CAS) and the baseline run (Zettair). The VCAS run performs best, particularly for low recall (0.2 and less), however its performance is almost identical to that of the CO run for 0.3 (and higher) recall. Figure 5(b) also shows that, when highly relevant elements are the target of retrieval, Zettair clearly outperforms the Hybrid_CAS run.

General VCAS retrieval scenario In the following analysis, we use the General retrieval scenario to compare the performance of the three hybrid VCAS

		General					
		All topics		Broad topics		Narrow topics	
VCAS run	**%Ovp**	MAP	P@10	MAP	P@10	MAP	P@10
`Zettair`	0	**0.192**	**0.119**	**0.625**	**0.367**	**0.029**	**0.045**
`Hybrid_CO_MpE`	78.3	0.128	0.035	0.417	0.100	0.020	0.015
`Hybrid_VCAS_MpE`	67.8	0.128	0.046	0.412	0.100	0.021	0.030
`Hybrid_CAS`	5.4	0.061	0.085	0.162	0.233	0.023	0.040

Table 3. Performance results of four INEX 2004 VCAS runs when using strict quantisation function and different VCAS topic categories. The General retrieval scenario is used. For each run, an overlap indicator shows the percentage of overlapping elements in the answer list. Values for the best runs are shown in bold.

runs (query choices `CO`, `VCAS` and `CAS`) with `Zettair`. The three VCAS topic categories are also used in this analysis: the *All* category, with all the 22 VCAS topics, and the *Broad* and *Narrow* categories, with 6 and 16 VCAS topics, respectively.

Table 3 shows the evaluation results for each run. One observation is very clear: for each VCAS topic category (with both `MAP` and `P@10` measures), `Zettair` outperforms all the other runs. This is a very interesting observation, since the unit of retrieval in Zettair is a full article, and queries used are plain content-only queries. These results show that under this retrieval scenario, applying a full-text search engine may be a better choice than an XML-specific retrieval approach.

5 Conclusions

In this paper we have reported on our participation in the ad-hoc track of INEX 2004. We have designed and submitted different runs for each CO and VCAS sub-track to investigate different aspects of the XML retrieval task.

The results of our preliminary analysis of the INEX 2004 CO and VCAS relevance assessments have identified many cases of mutually overlapping elements in the recall base. This finding, which is also known as the *overlap problem*, turns out to be not only an evaluation problem, but also a serious retrieval problem. Indeed, we have shown that in what we call the Original retrieval scenario, the strict quantisation function is not capable of dealing with the overlap problem. Efforts are being made, however, in the direction of unifying existing INEX metrics into a robust evaluation metric which aims at addressing this problem [6].

The two different XML retrieval scenarios, General and Specific, which were identified as a result of our analysis, model different user behaviours; we have shown that the preferred retrieval parameters — such as the choice of retrieval heuristic, level of element overlap or query type — vary depending on which user model is used. Moreover, distinguishing between existing topic categories can, in some retrieval scenarios, influence the choice of these parameters.

For the CO sub-track, we have shown that under the General retrieval scenario where users prefer less specific and non-overlapping answers, a full-text search engine alone can satisfy users' information needs. Our hybrid system, which is also capable of retrieving less specific and non-overlapping answers, is another effective alternative. However, our results have also shown that distinguishing between different categories of retrieval topics is very useful for the General CO retrieval scenario. Indeed, depending on the topic category, using a retrieval heuristic capable of retrieving more focused — and possibly overlapping — answers may be a better choice.

For the VCAS sub-track, we have shown that under the same General retrieval scenario, the same choice of using a full-text search engine — which ignores all the structural constraints and target elements — is very effective. Unlike for the General CO retrieval scenario, the choice of optimal retrieval parameters is not affected by a VCAS topic category.

It is our hope that this work will aid better understanding of the different aspects of the XML retrieval task, and ultimately lead to more effective XML retrieval.

References

1. Y. Chiaramella, P. Mulhem, and F. Fourel. A Model for Multimedia Information Retrieval. Technical report, FERMI ESPRIT BRA 8134, University of Glasgow, April 1996.
2. N. Fuhr, M. Lalmas, and S. Malik, editors. *INitiative for the Evaluation of XML Retrieval (INEX). Proceedings of the Second INEX Workshop. Dagstuhl, Germany, December 15–17, 2003*, March 2004.
3. K. Hatano, H. Kinutan, M. Watanabe, Y. Mori, M. Yoshikawa, and S. Uemura. Keyword-based XML Fragment Retrieval: Experimental Evaluation based on INEX 2003 Relevance Assessments. In Fuhr et al. [2], pages 81–88.
4. J. Kamps, M. de Rijke, and B. Sigurbjoernsson. Length Normalization in XML Retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK, July 25–29, 2004*, pages 80–87, 2004.
5. G. Kazai. Report on the INEX2003 Metrics working group. In Fuhr et al. [2], pages 184–190.
6. G. Kazai, M. Lalmas, and A. P. de Vries. The Overlap Problem in Content-Oriented XML Retrieval Evaluation. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK, July 25–29, 2004*, pages 72–79, 2004.
7. J. Pehcevski, J. A. Thom, and A.-M. Vercoustre. Enhancing Content-And-Structure Information Retrieval using a Native XML Database. In *Proceedings of The First Twente Data Management Workshop (TDM'04) on XML Databases and Information Retrieval. Enschede, The Netherlands, June 21, 2004*, pages 24–31, 2004.
8. J. Pehcevski, J. A. Thom, and A.-M. Vercoustre. Hybrid XML Retrieval: Combining Information Retrieval and a Native XML Database. *Journal of Information Retrieval: Special Issue on INEX (to appear)*, 2004.