Hybrid XML Retrieval Revisited

Jovan Pehcevski
RMIT University
Melbourne, Australia
jovanp@cs.rmit.edu.au

James A. Thom
RMIT University
Melbourne, Australia
jat@cs.rmit.edu.au

Anne-Marie Vercoustre
INRIA
Rocquencourt, France
anne-marie.vercoustre@inria.fr

ABSTRACT

In this paper, we report on the participation of the RMIT University group in the INEX 2004 ad-hoc track. Our preliminary analysis of CO and VCAS relevance assessments identifies two complementary cases of modified relevance assessments: General *and* Specific. *Further analysis of the General relevance assessments reveal two categories of retrieval topics:* Broad *and* Narrow. *We design runs that follow a hybrid XML approach and implement two retrieval heuristics with different level of overlap among the result elements. We show that for the initial INEX 2004 test collection the overlap CO runs outperform the non-overlap runs, and the heuristic which favours less specific over more specific result elements performs best. Importantly, we present results which show that, in a scenario where users prefer compound and non-overlapping answers to their queries, the choice of using a plain full-text search engine is still a very effective choice for XML retrieval.*

Keywords

XML Search & Retrieval, eXist, Zettair, INEX

1. INTRODUCTION

INEX 2004 explores two types of ad-hoc retrieval topics: Content-Only (CO) topics and Vague Content-And-Structure (VCAS) topics. Forty CO topics are used in the CO ad-hoc sub-track, while thirty-five VCAS topics are investigated in the VCAS ad-hoc 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" contains a substantial number of mutually overlapping elements [6].

VCAS topics enforce restrictions on the existing document structure and explicitly specify the target element (such as article, section or paragraph). However, the structural constraints in a VCAS topic need not be strictly matched. This means that not only are the restrictions on document structure *vague* restrictions, but also that the target element could also represent any element considered *likely to be 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.

The system we use for the ad-hoc track in INEX 2004 follows a *hybrid XML approach*, utilising the best features from Zettair[1] (a full-text search engine) and eXist[2] (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) [9].

The above approach however resulted in rather poor system performance for INEX 2003 CO topics, 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) [8]. We show elsewhere that this hybrid-CRE system produces performance improvements for the (V)CAS topics [7]. Different heuristic combinations 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 heuristic combination yields the best retrieval performance, and to investigate whether having non-overlapping result elements in the answer list 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 a more effective VCAS retrieval.

The remainder of this paper is organised as follows. In Section 2 we undertake a preliminary analysis of the INEX 2004 relevance assessments to identify the types of highly relevant elements. By analysing the relevance assessments for the CO and VCAS topics, we aim to understand what users — or the topic authors who later assess the relevance of returned answer elements — consider to be the most useful. In Section 3 we provide a detailed description of the runs we consider for the CO and the VCAS sub-tracks. In Section 4 we

[1] `http://www.seg.rmit.edu.au/zettair/`
[2] `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>
```

Figure 1: An extract from the INEX 2004 CO relevance assessments

present results of our CO and VCAS runs. These results reflect different retrieval scenarios based on our analysis of the INEX 2004 relevance assessments. Finally, we conclude in Section 5.

2. ANALYSIS OF INEX 2004 RELEVANCE ASSESSMENTS

Analysing the INEX 2004 CO and VCAS relevance assessments we observe that since neither topic restricts the answer elements, the final answer list may contain elements of different types with varying sizes and granularity. The names of some element types in the XML document collection correspond as follows: `article` to a full article, `abs` and `bdy` to article abstract and article body, `sec`, `ss1` and `ss2` to section and subsection elements, and `p` and `ip1` to paragraph elements. 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.

2.1 CO relevance assessments

Figure 1 shows an extract from the INEX 2004 CO relevance assessments. Values for the two INEX relevance dimensions, *exhaustivity*[3] (how many aspects of the topic are covered in the element), and *specificity*[4] (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.

The focus of our analysis is on *highly relevant* elements. These are elements that — for a given topic — have been assessed as both highly exhaustive and highly specific (E3S3) elements. In Figure 1 there are 8 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*. Note that, unlike the INEX definitions for exhaustivity and specificity, the definitions for General and Specific (highly relevant) elements result from our analysis as follows.

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

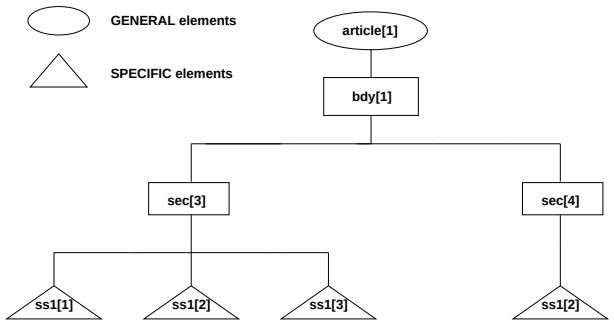

Figure 2: A tree-view example of GENERAL versus SPECIFIC elements.

General:

> "For a particular article in the collection, a *General* element is the least-specific highly relevant element containing other highly relevant elements" [8].
>
> Based on the above definition, `article[1]` is the only General element in the example in Figure 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 Figure 1. The General element is the element shown in ellipse.

Specific:

> "For a particular article in the collection, a *Specific* element is the most-specific highly relevant element contained by other highly relevant elements" [8]. In Figure 2, the Specific elements are the highly relevant 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 our analysis.

Figure 3 shows the overall distribution of the most frequent highly relevant elements (including full articles) that appear in more than half the CO topics. The figure shows three distinct cases when relevance assessments consider all highly relevant elements (*Original* relevance assessments), General highly relevant elements only (*General* relevance assessments) and Specific highly relevant elements only (*Specific* relevance assessments), respectively. The x-axis contains the names of the six highly relevant elements that appear in more than half the CO topics (in the case of Original relevance assessments). The y-axis contains the number of overall occurrences of each element.

In the case of Original relevance assessments, `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 overall

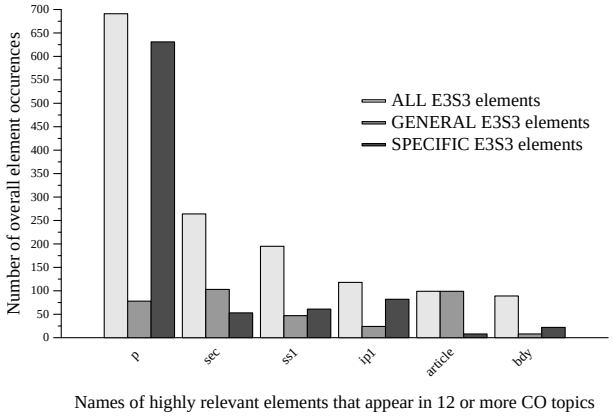

Figure 3: Overall distribution of highly relevant elements that appear in more than half the INEX 2004 CO topics, for three distinct cases of relevance assessments.

occurrences. The latter implies that in most cases when a `bdy` was assessed as highly relevant, the parent `article` is also likely to have been assessed as highly relevant too.

For General relevance assessments, one may expect that the situation should change in favour of the least specific and highly relevant elements. However, in this case `sec` elements are most frequent with 103 overall occurrences, followed by `article` elements with 99 occurrences (however the `article` occurrences are distributed across 16 topics, whereas there are 15 topics where `sec` elements occur). Surprisingly, `p`, `ss1` and `ip1` follow next, with 78, 47 and 24 overall occurrences, respectively. By looking at the number of `bdy` elements, we notice that there are 8 occurrences (distributed across 6 topics) where, when a `bdy` was assessed as highly relevant, the parent `article` has *not* been assessed as highly relevant.

The last case shown in Figure 3 is for Specific relevance assessments. As expected, the situation changes here in favour of the most specific elements, with `p` elements being most frequent. The `ip1`, `ss1`, `sec` and `bdy` 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 two distinct cases of relevance assessments, General and Specific, typically model different user (retrieval) behaviours. Indeed, in the absence of empirically-based models for expected user behaviour, the former case reflects users that prefer compound and more informative answers for their queries, whereas the latter case reflects users that prefer specific, more focused answers for their queries. The knowledge obtained from the above statistics may therefore be appropriately utilised by an XML retrieval system, particularly because distinct cases of relevance assessments favour different types of highly relevant elements.

Topic categories

In the following analysis we consider the case of General relevance assessments. Our aim is to distinguish those CO retrieval topics that seek to mostly retrieve less specific elements (such as `article` and `bdy`), from those that mostly retrieve other, more specific elements. Consider Figure 4: a point on this graph represents a CO topic. The x-axis shows the total number of General `article` and `bdy` elements contained by a CO topic, whereas the y-axis shows the total number of General elements other than `article` and `bdy` contained by the same topic. For example, the CO topic depicted at coordinates (23,11) contains 23 highly relevant `article`/`bdy` elements and 11 highly relevant elements other than `article`/`bdy`.

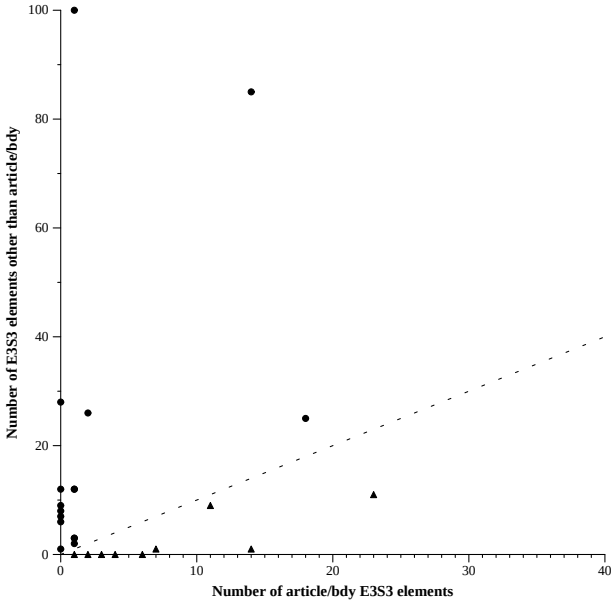

Figure 4: Categories of INEX 2004 CO topics when relevance assessments consider General (highly relevant) elements only.

We use this graph to identify two different categories of INEX 2004 CO topics. The first category, shown as full triangles on the graph and located below the dashed line, favours larger, 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.

The second category, shown as full circles on the graph, favours smaller, 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 in the other two assessment cases, that is, for either the Original or the Specific relevance assessments. However, further analysis shows that four topics (numbers 168, 178, 190 and 198) clearly belong to the Broad category even in this two cases. We observed in our previous work a varying behaviour of an XML retrieval system when its performance is measured against different categories of CO topics [8]. Indeed, it has also been experimentally shown to be a valid observation for a fragment-based XML retrieval system [3]. Thus, it is likely to be useful to distinguish between different categories of CO topics.

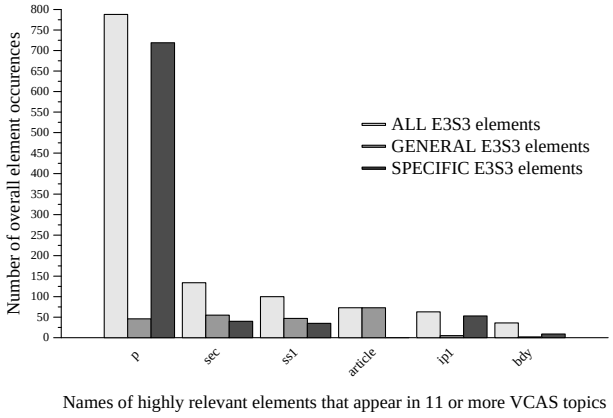

Figure 5: Overall distribution of highly relevant elements that appear in more than half the number of VCAS topics, for three distinct cases of relevance assessments.

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 5 shows the overall distribution of the most frequent highly relevant elements that appear in more than half the VCAS topics. The figure also shows three distinct cases of Original, General and Specific relevance assessments.

Since the VCAS relevance assessments have been done in much the same way as those for the CO topics, it is not surprising that graphs in Figures 5 and 3 show similar statistics. In both cases, the number of overall occurrences of `p` elements (in the case of Original and Specific relevance assessments) is far greater than the numbers of all the other elements. Nevertheless, there are some differences for the number of overall occurrences of `article` and `bdy` elements. In the case of VCAS Original relevance assessments, the number of `article` elements is much greater than that of the `bdy` elements (73 `article` occurrences across 12 topics, compared to 36 `bdy` occurrences across 11 topics). For VCAS General relevance assessments, `article` elements are the most frequent among all the other more specific elements. In the case of VCAS Specific relevance assessments, the number of `article` elements is zero, whereas there are 9 highly relevant `bdy` elements, which are distributed across 3 topics.

Topic categories

As for the CO topics, we use the case of General relevance assessments to identify two different categories of INEX 2004 VCAS topics. The first category of topics favours 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. The second category of topics favours more specific elements as highly relevant answers. There are 16 such topics, referred to as *Narrow* topics.

An interesting observation is that only two 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). Further analysis also shows that surprisingly, the last topic belongs to the Broad category even in the other two cases of Original and Specific relevance assessments.

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, similar to the case for the CO topics, important information that an XML retrieval system should use.

3. RUNS DESCRIPTION

3.1 Background

All the runs we consider for the INEX 2004 ad-hoc track are based on the hybrid XML retrieval approach. To determine the ranks of CREs in the final list of answer elements, the CRE module in our hybrid system uses a combination of the following XML-specific heuristics:

1. the number of times a CRE appears in the absolute path of each extracted 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 beginning (`B`) or nearer to end (`E`).

There are 16 possible CRE heuristic combinations, since 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 and more general elements are ranked higher than more specific and less general elements.

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

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

3.2 CO sub-track

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

- `Zettair` – using the full-text information retrieval system as a baseline run;

		Different quantisation functions (Original assessments)									
		strict		s3_e321		s3_e32		e3_s321		e3_s32	
CO run	%Ovp	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
Zettair	0	0.049	0.073	0.008	0.097	0.020	0.088	0.088	0.206	0.071	0.132
Hybrid_MpE	82.2	**0.124**	**0.103**	**0.041**	0.194	**0.072**	0.174	**0.178**	**0.218**	**0.155**	**0.182**
Hybrid_MpE_NO	0	0.051	0.076	0.008	0.100	0.020	0.091	0.089	0.209	0.073	0.138
Hybrid_PME	82.1	0.081	0.100	0.038	**0.206**	0.052	**0.182**	0.089	0.141	0.083	0.121
Hybrid_PME_NO	0	0.047	0.088	0.023	0.197	0.027	0.165	0.031	0.123	0.034	0.109

Table 1: Performance results of INEX 2004 CO runs when using different quantisation functions and across 25 CO topics. The case of Original relevance assessments is used. For each run, an overlap indicator shows the percentage of overlapping elements in the answer list. Values for the best runs (for each measure and under each function) are shown in bold.

- Hybrid_MpE – using the hybrid system with MpE heuristic combination in the CRE module;
- Hybrid_MpE_NO – using the hybrid system, with MpE heuristic combination, and no overlap among the elements in the final answer list;
- Hybrid_PME – using the hybrid system with PME heuristic combination in the CRE module;
- Hybrid_PME_NO – using the hybrid system, with 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 best performance for the hybrid system under different retrieval scenarios. Second, we aim to investigate the impact of overlapping result 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 and more general elements to remain in the list and removes all the other (contained) elements, whereas the latter retains more specific and less general elements, and removes all the other elements that contain them. 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 information retrieval system as a baseline run;
- Hybrid_CO_MpE – using the hybrid system with MpE heuristic combination in the CRE module. The structural constraints and the target element in the Title part of each VCAS topic are removed, leaving query terms only.
- Hybrid_CO_PME – using the hybrid system with PME heuristic combination in the CRE module. As with the previous run, each VCAS topic is treated as being a CO topic;
- Hybrid_VCAS_MpE – using the hybrid system with MpE heuristic combination in the CRE module. The target element in the Title part 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 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 in the Title part of each VCAS topic are strictly matched.

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

4. EXPERIMENTS AND RESULTS

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: *Mean Average Precision* (MAP), which measures the ability of a system to return relevant elements, and *Precision at 10* (P@10), which measures the number of relevant elements within the first 10 elements returned by a system.

In INEX 2004, an evaluation metric with different quantisation functions is used to evaluate the retrieval effectiveness of XML systems [5]. Thus, the exhaustivity and specificity values for *relevant* elements may vary depending on the choice of quantisation function. For example, if the strict quantisation function (e3_s3) is used, MAP will measure the ability of a system to return *highly relevant* (E3S3) elements, whereas if the e3_s321 or s3_e321 functions are used, MAP will measure the ability of a system to return *highly exhaustive* (E3S3, E3S2, E3S1) or *highly specific* (E3S3, E2S3, E1S3) elements. In the following we describe results obtained from evaluating the retrieval effectiveness of our runs for each CO and VCAS sub-track.

		Strict quantisation function (General assessments)					
		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 case of General relevance assessments is used. For each run, an overlap indicator shows the percentage of overlapping elements in the answer list. Values for the best runs (for each measure and under each topic category) are shown in bold.

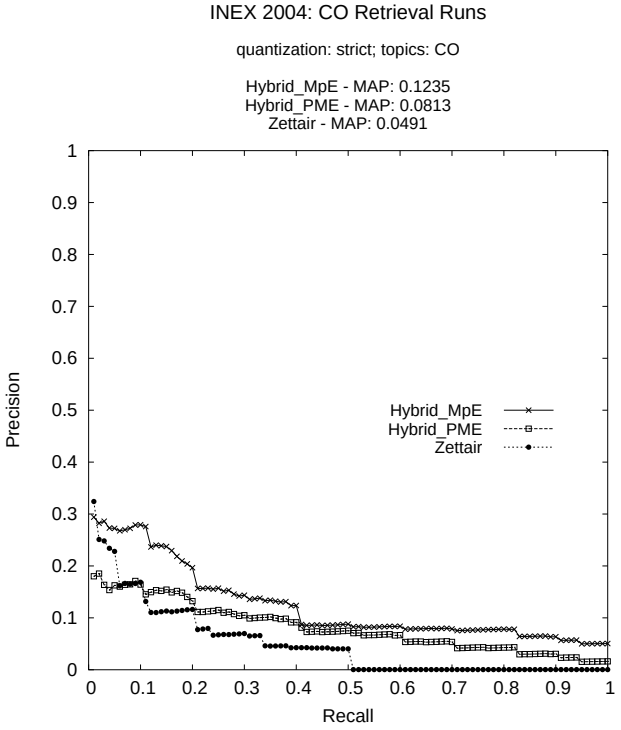

Figure 6: Evaluation of three INEX 2004 CO retrieval runs using strict quantisation function and the case of Original relevance assessments.

4.1 CO sub-track

Table 1 shows evaluation results for the CO retrieval runs when the case of Original relevance assessments is considered. Different quantisation functions are used to evaluate the retrieval effectiveness, and values for the best runs (for each measure under each function) 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, except for the *highly specific* quantisation functions (`s3_e321` and `s3_e32`), where the number of relevant elements in the first 10 returned elements is on average higher when using the `PME` heuristic.

Second, the non-overlap hybrid runs perform worse than the corresponding overlap hybrid runs. This is very likely to be a result of the varying CO recall base, which, as previously discussed in Section 2.1, is as apparent in INEX 2004 as it was in INEX 2003 [8]. We revisit the latter comparison in the next section, where a non-varying recall base is considered for evaluation (the case of General relevance assessments).

Last, the hybrid runs perform better on average than the baseline run, except for the *highly exhaustive* quantisation functions (`e3_s321` and `e3_s32`), where the baseline run is competitive, and, with the `P@10` measure, performs even better than both the overlap and the non-overlap `Hybrid_PME` runs. These results show that when highly exhaustive elements are the target of retrieval, a full-text search engine could still be used to satisfy the information need almost as well.

Figure 6 shows recall/precision curves for the two overlap hybrid runs (`Hybrid_MpE` and `Hybrid_PME`) and the baseline run (`Zettair`). The runs are evaluated by using the strict quantisation function and the case of Original relevance assessments. 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 strict quantisation function and the case of General relevance assessments to compare the performance of the two `Hybrid_MpE` runs (overlap and non-overlap) with `Zettair`. When a run is evaluated with the strict quantisation function, the case of General relevance assessments reflects a non-overlapping recall base, since an article is allowed to only contain General, non-overlapping (highly relevant) elements (see Section 2.1 for definition of General elements). Moreover, our previous analysis has distinguished two different categories of CO topics. Thus, in this General retrieval scenario, 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 varying recall base (the case of Origi-

VCAS run	%Ovp	Different quantisation functions (Original assessments)									
		strict		s3_e321		s3_e32		e3_s321		e3_s32	
		MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
Zettair	0	0.052	0.119	0.012	0.146	0.021	0.146	0.063	0.296	0.033	0.154
Hybrid_CO_MpE	78.3	0.101	0.104	**0.037**	0.200	**0.056**	0.158	**0.110**	0.262	**0.084**	0.162
Hybrid_CO_PME	78.2	0.034	0.096	0.029	0.189	0.036	0.135	0.068	0.204	0.051	0.123
Hybrid_VCAS_MpE	67.8	**0.103**	**0.154**	0.027	**0.235**	0.047	**0.192**	0.107	**0.323**	0.078	**0.227**
Hybrid_VCAS_PME	67.8	0.045	0.142	0.021	0.227	0.029	0.187	0.072	0.258	0.059	0.196
Hybrid_CAS	5.4	0.032	0.142	0.018	0.212	0.026	0.173	0.030	0.200	0.034	0.189

Table 3: Performance results of INEX 2004 VCAS runs when using different quantisation functions and across 22 VCAS topics. The case of Original relevance assessments is used. For each run, an overlap indicator shows the percentage of overlapping elements in the answer list. Values for the best runs (for each measure and under each function) are shown in bold.

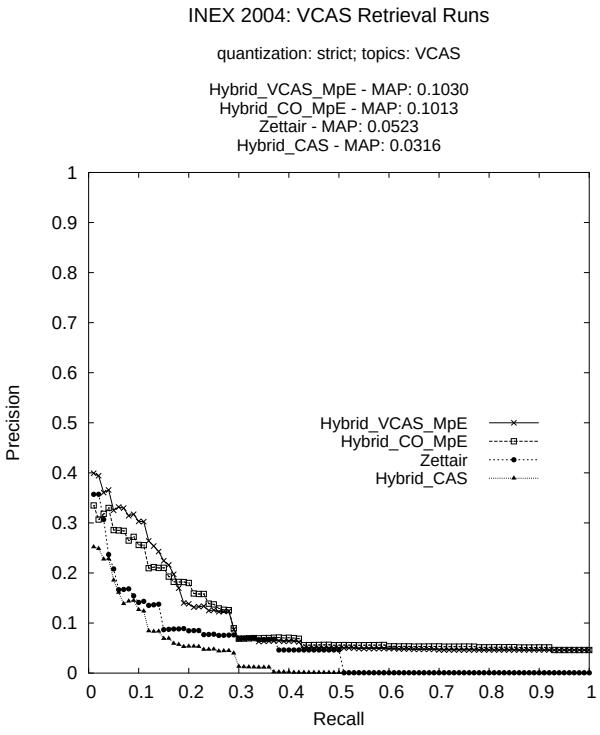

Figure 7: Evaluation of four INEX 2004 VCAS retrieval runs using strict quantisation function and the case of Original relevance assessments.

nal relevance assessments), the non-overlap hybrid run substantially outperforms the overlap hybrid run. In the case of *Narrow* topics, the overlap hybrid run performs best, whereas the performance of the other two runs is the same.

4.2 VCAS sub-track

Table 3 shows evaluation results for the VCAS retrieval runs and the case of Original relevance assessments. Different quantisation functions are used to evaluate the retrieval effectiveness, and values for the best runs (for each measure under each function) are shown in bold. Several observations can be drawn from the results of Table 3.

First, the strict hybrid run (`Hybrid_CAS`) (where structural constraints and the target element of a VCAS topic are strictly matched) performs worse than the other hybrid runs. This is not surprising, since the relevance assessments for VCAS topics have been done in the same way as those for CO topics. Moreover, when using strict quantisation function (with both `MAP` and `P@10`) `Hybrid_VCAS` runs (the choice of strict structural constraints and no explicit target element) perform better than `Hybrid_CO` runs (the choice where plain CO queries are used).

Second, as with CO topics the `MpE` heuristic in hybrid runs yields better performance than the `PME` heuristic.

Last, the hybrid runs perform better on average than the baseline run, except when using strict quantisation function (with `MAP`), where `Zettair` performs better than the strict hybrid run and both the hybrid-`PME` runs.

Figure 7 shows recall/precision curves for the three hybrid runs that use different retrieval choices (`CO`, `VCAS` and `CAS`) and the baseline run (`Zettair`). The runs are evaluated using the strict quantisation function and the case of Original relevance assessments. The `VCAS` run performs best, particularly for low recall (0.2 and less), however its performance is almost identical with that of the `CO` run for 0.3 (and higher) recall. When highly relevant elements are target of retrieval, `Zettair` clearly outperforms the strict (`CAS`) hybrid run.

General VCAS retrieval scenario

In this scenario we use the strict quantisation function and the case of General relevance assessments to compare the performance of three hybrid VCAS runs (with retrieval 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 4 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` by far 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. For each VCAS topic category (with `P@10` measure), the strict hybrid run also outperforms the other two hybrid runs. Of these, the

VCAS run	%Ovp	Strict quantisation function (General assessments)					
		All topics		Broad topics		Narrow topics	
		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 4: Performance results of four INEX 2004 VCAS runs when using strict quantisation function and different CO topic categories. The case of General relevance assessments is used. For each run, an overlap indicator shows the percentage of overlapping elements in the answer list. Values for the best runs (for each measure and under each topic category) are shown in bold.

VCAS run again performs better (overall) than the CO run.

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 two different cases of INEX 2004 relevance assessments, which were identified as a result of our analysis, model different user behaviours; we have shown that the preferred retrieval aspects vary on the user model used. Moreover, distinguishing between existing topic categories can, in some assessment cases, influence the choice of these aspects.

For the CO sub-track, we have shown that where users prefer less specific and non-overlapping answers, a full-text search engine alone can satisfy user information needs. Our hybrid system, which is also capable of retrieving non-overlapping compound answers, is another effective alternative. However, our results also show that a system should also distinguish between different categories of CO retrieval topics. For a particular topic category, an XML system capable of retrieving more focused — and possibly overlapping — answers is a better choice.

For the VCAS sub-track, in the same retrieval scenario where users prefer less specific and non-overlapping answers to their queries, the same choice of using a full-text search engine, which ignores all the structural constraints and target elements, is very effective. Distinguishing between different topic categories in this case does not appear to make any difference on performance.

We have also used Zettair and the hybrid system in the INEX 2004 Heterogeneous track. However, since the relevance assessments for the heterogeneous XML collections are not yet available, we do not report their performance results in this paper.

The performance values for our INEX 2004 runs, generated with MAP and P@10, are much lower comparing to the same values for information retrieval systems retrieving whole documents. 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.

6. ADDITIONAL AUTHORS

S.M.M. Tahaghoghi, RMIT University, Melbourne, Australia. E-mail: saied@cs.rmit.edu.au.

7. 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*, 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*, 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*, 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 (accepted for publication)*, 2004.

[9] J. Pehcevski, J. A. Thom, and A.-M. Vercoustre. RMIT INEX Experiments: XML Retrieval using Lucy/eXist. In Fuhr et al. [2], pages 134–141.