Passage Retrieval and other XML-Retrieval Tasks

Andrew Trotman
Department of Computer Science
University of Otago
Dunedin, New Zealand
andrew@cs.otago.ac.nz

Shlomo Geva
Faculty of Information Technology
Queensland University of Technology
Brisbane, Australia
s.geva@qut.edu.au

ABSTRACT

At INEX there is an underlying assumption that XML-retrieval and element retrieval are one and the same. This is, in fact, not the case. The hypothesis at INEX is that XML markup is useful for information retrieval. We firmly believe this, but no longer in element retrieval. In this contribution we examine in detail the evidence collected in support of element retrieval and suggest that, contrary to expectation, it in fact supports passage retrieval and not element retrieval. Particularly, we draw on other studies that collectively show that INEX assessors are identifying relevant passages (not elements), they agree on where in a document those passages lie, that there already exists suitable metrics in the XML-retrieval community for evaluating passage retrieval algorithms, and that the tasks make more sense as passage retrieval tasks. Finally we show that future tasks of XML-retrieval also fit well with passage retrieval.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models, Search process*.

General Terms

Human Factors, Theory

Keywords

Element retrieval, XML-retrieval, passage retrieval

1. INTRODUCTION

The IEEE document collection used at INEX [5] between 2002 and 2005 has been replaced in 2006 by the Wikipedia collection. On initial inspection, structurally this new collection does not appear to be as versatile as the previous, the DTD does not appear to be semantically as rich, and the applicability of the content itself to element retrieval does not appear to be strong.

These "weaknesses" are only of concern if the underlying assumption is that element retrieval is the most appropriate way to search the collection – and this does not appear to be the case.

In this investigation we examine the methodological evidence for passage retrieval as a replacement for element retrieval in XML-retrieval. What we find is that assessors are highlighting passages; these highlighted passages are not typically elements; and that methodology is already in place for measuring the performance of passage retrieval within INEX.

After presenting the evidence for passage retrieval, we show that some of the problems facing element retrieval do not exist if passages are used. The problems associated with identifying focused results are problems of elements, and not problems of XML-retrieval. The problem of "too small" elements does not exist if the natural relevant unit is a passage and not an element.

Information retrieval is user-centered task; the purpose is to identify relevant information and to present it to a user. We show that, in fact, some of the current element retrieval tasks are a consequence of elements and not users – specifically we ask: what are the natural tasks for a passage-retrieval system? We show that *focused* retrieval and *thorough* retrieval are equivalent under passage retrieval.

Finally we examine some possible future tasks for XML-retrieval and show that passages are the natural unit in which to specify them.

We do not suggest the XML markup is of no benefit – such markup might be used for identifying good passages. Elements might also be good answers to question answering topics.

In conclusion we propose parallel element retrieval and passage retrieval tasks at INEX 2007 with the possibility of passage only tasks at INEX 2008 and onwards.

2. Element Retrieval and Passage Retrieval

In this section we examine element retrieval and passage retrieval, then put the case that evidence collected to support element retrieval in fact supports passage retrieval.

2.1 Element Retrieval

If a document is marked up in a semantic mark-up language such as XML, it is possible for a search engine to take advantage of the structure. It could, for example, return a more focused result than a whole document. In element retrieval the search engine is tasked to identify not only which documents are relevant, but also which semantic structures (or elements) within those documents are relevant to an information need.

On initial inspection element retrieval appears to be a reasonable technology. Considering the INEX IEEE document collection, instead of returning a whole (say 10-page) document, the search engine might return a document section, subsection, or just a paragraph to the user. This far more focused result is clearly of benefit to our user. Several algorithms have been proposed and tested within [12; 29] (and without [6]) INEX

The benefit to the user is obvious. Whereas a document-centric search engine would return 10 pages, the element-centric search engine returns, perhaps, a single relevant page filtered from, perhaps, 9 other pages of irrelevant content. This machine filtering reduces the cognitive load on the users by increasing the ratio of relevant to irrelevant content presented to them.

2.2 Passage Retrieval

An alternative (and earlier) technology exists for identifying relevant parts of documents – passage retrieval. Should a

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

document be long, say 10 pages, but not contain semantic markup, then element retrieval is inappropriate. Considering the same IEEE collection, but this time as a collection of PDF formatted documents, the search engine is again tasked to identify the relevant parts of the document but has no semantic markup to use. This time it must use the document content itself and not rely on explicit markup.

Several approaches have been suggested. Harper and Lee [8], for example, suggest sliding a fixed sized window over the text and computing a window score for each and every word – resulting in a relevance profile for a document. Such approaches are generally based on one variation or another of the proximity heuristic and are hence language model free. More sophisticated approaches such as natural language processing (NLP) have also been used in passage retrieval. NLP techniques appear to be successful at question answering but not yet at *ad hoc* retrieval where other than very simple techniques have yet to succeed. In question answering a more refined context analysis approach, beyond simplistic proximity heuristics, is advantageous [1; 13].

As with element retrieval the aim of passage retrieval is to reduce the cognitive load on the user. This, again, is by filtering relevant from irrelevant content within a document. Both technologies aim to increase precision.

2.3 Element Assessments

Along with the increased understanding of element retrieval came changes to the assessment methodology. At INEX 2004 assessors were presented with documents and asked to judge (pooled) elements from those. At INEX 2005 the assessors were presented with documents and asked to first identify relevant passages, then to apply exhaustivity values to elements within those passages [19]. Critically, this change allowed the analysis of passages in relation to XML documents. The evidence is in favor of passages.

2.4 Applicability

If we assume that XML markup adequately takes care of fine grained semantics, it is then a reasonable hypothesis that element retrieval is the most appropriate technology for XML and that passage retrieval is not necessary for XML documents. This, however, does not appear to be the case.

Extensive analysis of the judgments collected at INEX 2004 was done by Trotman [27] and by Pehcevski *et al*. [17]. Trotman focused his discussion on the agreement levels between judges on 12 topics assessed by two independent judges. He presents the binary document-centric agreement level as 0.27 which is low by comparison to TREC (between 0.33 and 0.49), but in line. Exact 10 relevance-point agreement of elements was 0.16, very low. Pehcevski *et al.* examined the agreement levels between the judges and participants in interactive experiments. They show agreement only at the extreme ends of the relevance scale, that is, E3S3 and E0S0 only. This end-only agreement is also seen in the cystic fibrosis collection [22]. In an effort to increase cross judge agreement the assessment method was changed from judging elements to highlighting passages – on the hypothesis that this might reduce the cognitive load on the judge resulting in an increase in agreement levels.

There has also been extensive analysis of the INEX 2005 passage and element results.

Trotman and Lalmas [28] examine which elements were identified as relevant. They found that regardless of the query specific target element there were more relevant paragraph elements than any other element. Even when the judgments were filtered for focused retrieval (with the exception of queries targeting whole articles), paragraphs prevailed in the judgments. They suggest that this might be because the assessors are identifying relevant and consecutive passages of text, and not elements, when identifying relevant content in a document.

Piwowarski *et al*. [19] examine the average specificity of paragraph elements and report a value of 0.94. For comparison, the average specificity of a section element is 0.51. They conclude that paragraphs are, in general, either completely relevant to an information need, or not at all relevant.

Piwowarski *et al*. go on to examine the correlation between passages and elements in the judgments. They define two types of passages: elemental passages and non-elemental passages. An elemental passage is a passage that is also a whole element whereas a non-elemental passage is a subset of the content of the smallest fully encompassing element. They report that only 36% of passages are elemental (therefore 64% are not). The conclusion is that assessors are not, in general, highlighting relevant elements, but are identifying relevant passages.

Ogilvie and Lalmas [14] examine the stability of the metrics under different conditions. They conclude that the exhaustivity dimension can be dropped from the assessment procedure without unduly affecting the relative performance of search engines. They suggest assessment by specificity only, or in other words highlighting passages of text and performing element retrieval based solely on these highlighted passages (as do Pehcevski and Thom [16]).

Finally, Pehcevski and Thom [16] examined the agreement levels between judges at INEX 2005 (using highlighting). They report a non-zero document level agreement of 0.39 and an exact element agreement of 0.24. Piwowarski *et al*. measured the agreement level of whole passages and report a value of 0.23. Although only 5 topics were used in this comparison, a large improvement is seen. An improvement indicating that a passage is a more natural unit than an element.

In summary, assessors are highlighting passages of text and not elements, these passages consist mostly of whole paragraphs. The judges agree not only on which documents are relevant, but on the passages within those documents. The obvious conclusion is that passage retrieval is a more appropriate technology for the INEX IEEE document collection than element retrieval.

2.5 The Case For Passage Retrieval

The INEX focused retrieval task aims to identify document elements of just the right size, however *right size* is not a well defined concept. There is scope for disagreement between assessors, and they do disagree. Furthermore, while systems are required to return XML elements of optimal granularity, the assessors as asked perform relevant passage identification. This discrepancy means that the elements of the optimal granularity (in the judgments) must somehow be derived from the relevant passages identified by the judges.

Several ways to do this have been proposed and opinions on effectiveness differ. There was, for example, much discussion and disagreement at INEX 2005 about the automatic derivation of “too small” elements. A too small element is part of a relevant passage, while at the same time insufficient in itself at fulfilling any of the information need. Such an element might

be a citation number in flowing text – relevant in context but on its own just a number.

There are two ways such difficulties might be overcome. Either ask systems to return passages instead of elements, or ask assessors to identify focused elements and too small elements and not passages. In either case there must be a direct correspondence between the retrieval task and the assessment task. It seems that passage retrieval is the obvious option from the assessment point of view, and hence probably the more reasonable approach – particularly if it more accurately matches the user needs.

But does moving to passage retrieval mean that element retrieval is unnecessary? The hypothesis being tested at INEX is that XML markup is useful in retrieval. INEX is not an element retrieval evaluation forum; it is an XML-retrieval evaluation forum. In past workshops the hypothesis was tested by comparing results that were obtained by content only (CO) queries and content and structure (CAS) queries. For some systems the hypothesis holds and for other it does not [28], but it is still an open question whether *markup* is useful. The nature of the broad concept of *ad hoc* querying, and the semantically weak markup of the INEX IEEE collection did not allow this hypothesis to be vigorously tested. By moving to passage retrieval (and perhaps with it also moving to more focused tasks such as question answering) the usefulness of exploiting XML *markup* may come to the fore. We believe this is a compelling argument for moving to passage retrieval and to more sophisticated tasks and challenges.

Can passage retrieval be assisted by XML markup? In the context of question answering, summarization, or even known entity searching it is reasonable to believe so, especially in the case of a collection with semantically strong markup and strongly typed elements. Therefore, it is necessary not only to move to passage retrieval, but to also change the kind of tasks under study and the type of collections that we use. Some of these issues are addressed in the later part of this paper, where we discuss potential future tasks for XML-retrieval systems.

2.6 Transition

Passage retrieval and element retrieval are not mutually exclusive technologies and a transition from one to the other is possible. Specifically, the transition from elements to passages is of interest for two reasons. First, this is the transition which INEX is facing. Second, it is likely to result in an increase in precision as further irrelevant content can be removed from a user's result list (that content in an element, but at the same time not relevant to the user's information need).

2.6.1 From Elements to Passages

Given a ranked set of elements from an element retrieval search engine, it is trivially possible to convert these into a set of passages. The start and end of an element become the start and end points of a passage. Additionally, immediately adjacent passages may need to be merged into a single passage.

2.6.2 From Passages to Elements

The conversion from passages to a thorough set of elements is straightforward; all elements containing any part of a passage are relevant.

The conversion to focused elements is not trivial. A passage could start mid-way through an element, cross several element boundaries and finish midway through another element. Conversion to a single element is straightforward; the smallest element fully enclosing the passage would be selected. Unfortunately it is not clear that this element is the best focused result as such an element may not be fully specific. An alternative approach might be to identify the largest elements fully enclosed by the passage. These elements would be fully specific; however there remains the potential for some relevant content to be lost, that content jutting-in to an adjacent element.

2.6.3 Passage Specification

Several methods for specifying passages have already been proposed. Previous investigations into passage retrieval such as TREC HARD have used byte offset into document and length in characters. Such a method is not suitable for XML-retrieval as mid-way through a tag might be specified.

Clarke [3] suggests element range results at INEX and recommends an XPath syntax for doing so. We note that the INEX 2005 judgments already specify passages and suggest this convention also be used for specifying passages in runs.

2.7 Passage Assessments

The transition to assessing passages has already started, albeit not for passage retrieval purposes. At INEX 2005 the assessors first identified relevant passages, then exhaustivity values were assigned to any element intersecting the passage [19]. The extensions necessary to change to passage retrieval could be done in one of two possible ways. Either the assignment of exhaustivity would be to a passage and not an element, or alternatively the assessment of exhaustivity could be dropped. The latter has been suggested already by Ogilvie and Lalmas [14] and is already under consideration for INEX 2006. Should this be adopted then everything, except the task definitions, are in place for passage retrieval.

3. Passage Retrieval Tasks

Passage retrieval is well suited to XML documents. Additionally, passages can be more accurate as there is no requirement for a passage to start (or end) on a tag boundary. But what of the element retrieval tasks currently under investigation? It is important to look at user needs before task definition, but it turns out there are direct analogies between the existing element retrieval tasks and those one might expect for passage retrieval.

We initially envisage three tasks: the first is it the identification of relevant passages of text which are presented to the user in passage-relative order of relevance – this turns out to be a combination of the existing *focused* task and *thorough* task. The second is the identification of relevant passages of text which are presented to the user in document-relative order of relevance – essentially the *relevant in context* task. Finally, the identification of relevant documents presented to the user with and entry point identified – the *best in context* task essentially unchanged.

The retrieval task specification for INEX 2006 [4] discusses these 4 tasks with respect to element retrieval. In this section we discuss transitioning them to passage retrieval.

3.1 Focused Retrieval

In the existing *focused* task, a search engine must identify only those relevant elements that are most focused on the information need. A list of focused results may not contain any overlapping elements. For the search engine there are two problems at hand: the first is the identification of a relevant piece of text (where); and the second is the identification of the appropriate size of the text.

This task would change only subtly. Whereas at present the task is to identify non-overlapping elements (essentially passages), it would be changed to the identification of non-overlapping passages. A transitional requirement might be that passages must start and end on a tag boundary. This transition would allow the continued use of the current metrics. Alternatively, the introduction of a metric such as HiXEval [16] would alleviate this transitional need.

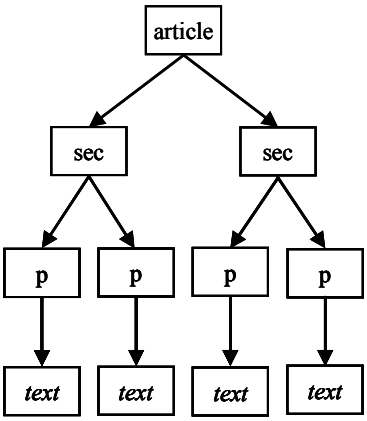

Figure 1: A simple document tree with text at the leaves

3.2 Thorough Retrieval

In the existing *thorough* task, a search engine must identify each and every relevant element in the document collection, and it must rank these relative to each other. This task is the only task that has continued in INEX since the first workshop.

This task has been criticized as it, by its very definition, requires the search engine to return overlapping elements in the results list [27]. Examining the document tree in Figure 1, and relevant text in a <p> element, and that inside a <sec> element, and that inside an <article> element. A thorough retrieving search engine will identify all three, and rank them relative to each other.

A natural consequence of this task is that the same text could be identified multiple times. To be *thorough* the search engine must identify *all* overlapping elements. In an interactive environment in which these overlapping results are displayed on-screen for a user, that user could potentially be presented with the very same element of text, and only that element of text, for the entire first page of results. Experiments conducted as part of the interactive track at INEX 2004 show that users do not want overlapping elements in results lists [11; 24]. This makes it a target for criticism on the basis of having no user-model, and it has been criticized for this [27].

We believe these criticisms are short-sighted, not because they are wrong but because the conversion to thorough results list from a passage is straightforward. This task could, therefore, act as a sanity check during the transition from elements to passages.

Of course, under the definition of passages, the thorough and focused task are equivalent[1] – the identification of documents, start points and the end points of all passages of text that satisfy the user's information need. These passages are sorted relative to each other.

[1] Until the use-case, we avoid discussing tasks with overlapping passages

3.3 Relevant In Context

In the existing *relevance in context* task, a search engine must first identify which documents are relevant, and then identify which elements within those documents are relevant. Results are grouped first by document, and then presented in document order. Overlapping elements are forbidden. This task is based on the experimental *Fetch Browse* task of INEX 2005 but the older task was thorough. This tasks is already (essentially) a passage retrieval tasks.

The change to passage retrieval is just a change in granularity. Whereas an element retrieval search engine is restricted to identifying elements, a passage retrieval search engine might identify passages that do not start or end on element boundaries (perhaps sentences).

By switching this task to a passage retrieval task, it is brought inline with the focused task. The difference between them being the order passages are returned. Relevance in context results lists would be in document order whereas focused results lists would be relative to other passages.

3.4 Best In Context

In the existing *best in context* task, a search engine must first identify relevant documents and then a single best point (BEP). The BEP is used to direct the user to relevant content within the document. At present this entry point is specified as an element start point. Only one best entry point into a document may be given and results are ranked on document topical relevance.

There may not be one best entry point in a document. Piwowarski *et al*. [19] examine the number of relevant passages per relevant document in the INEX 2005 judgments. They report that fewer than 50% of relevant documents contain only one relevant passage, while over 85% of relevant documents contain 5 or fewer relevant passages. As many as 49 passages are seen in one relevant document. When there are multiple passages in a single document it is not clear that one particular passage must necessarily be any better than all the others. This leads to questions about cross-judge agreement levels – which remain to be computed (this task is new for INEX 2006).

Conversion of this task to passage retrieval requires one subtle change; the entry point would no longer be required to lie on a tag boundary.

With passage retrieval this task is very close in definition to both focused and relevant in context. In relevant in context, documents are sorted relative to each other. Focused results are sorted relative to each other. Best in context results are first sorted on documents and then within document they are sorted relative to each other.

3.5 Passage Retrieval At TREC

The TREC HARD track [25] examined passage retrieval in 2003 and 2004. There the granularity of a query result was specified in metadata attached to the query. A query could target a document, passage, sentence or phrase sized units. Passages were specified in submissions as byte offset into a document, and length.

The TREC Genomics track is using a collection of scientific articles marked up in HTML for question answering. Results to queries are passages, identified by document identifier, passage offset, and passage length. Several TREC Genomics participants pushed for the collection in XML, including some also active in INEX.

We believe INEX should be looking at passage retrieval in semi-structured (XML) documents. TREC Genomics is already looking at passage retrieval in semi-structured (HTML) documents. This is an ideal opportunity to share results – and document collections.

By sharing document collection the algorithms from INEX and TREC Genomics could be compared head to head, this would imply also sharing metrics.

4. The Performance Task

Thorough retrieval is the only retrieval task that has been at INEX since the start. It could be used to measure the annual performance increase seen in ranking algorithms (as could other tasks, but this task has existed from the start).

A mapping from a passage to a thorough list is mechanical. All elements fully contained by the passage are fully and equally relevant. All those not intersecting with a passage are not relevant. For all others the relevance can be computed in the manner in which specificity is currently computed in the judgments: the ratio of thought-relevant text to the size of the element. A relevance value for elements in all documents can be computed and these ranked relative to each other.

With thorough rankings for search engines from the start of INEX, and a single (appropriately chosen) metric, the performance of the best submitted runs can be computed for each year and the result graphed since the beginning of INEX. Care must be taken when interpreting such a result as differences could reflect the hardness of the topic set and not improvements in search engine performance.

Alternatively, a set of unchanging benchmark topics could be used. These topics would remain the same from year to year and would not form part of evaluation – only new topics would be used for that. However, by analyzing the global performance on benchmark topics we would be able to say with confidence whether, or not, performance across the board was improving. There is still the risk that over-fitting will occur if INEX participants use these benchmark topics to train their systems – as they will no-doubt attempt to do. This might be overcome if neither the topics nor the judgments were released. Only performance statistics would be given.

Introduction of the Wikipedia collection is opportune. 125 topics have already been published for INEX 2006. From those, some suitably large number (say 25), might be used as benchmark topics and the other (say 100) for standard evaluation purposes. The judgments for the benchmarks would be withheld whereas the other judgments would be published.

Informal discussions, currently centered on an efficiency track, have suggested participants should submit their search engines and not runs. Should INEX adopt such an approach then performance changes from year to year could be measured on these submitted search engines. Care must also be taken with this approach as each year some participants re-train their search engines using the results from previous years. Re-running queries on these re-trained search engines is equivalent to measuring the performance of the training set – which should be optimal.

None the less, with INEX in its 5th year it is still not clear that any one relevance ranking algorithm is superior to any other. There are no standard benchmarks to which new algorithms are compared, and no clear evidence that improvements are being made from year to year. The purpose of this track would be to identify the state of the art and to introduce a standard methodology for experimentation.

In whole document retrieval the performance of a new ranking algorithm is compared to that of BM25 [20], pivoted length normalized retrieval [23], or language models [33]. Any differences are checked for statistical significance using either the *t*-test or Wilcoxon test [21]. No such standard methodology exists for XML-retrieval – because it is not clear which algorithms are state of the art.

Part of the cause of this problem has been the shifting metrics. An effective metric should be both stable, and say something useful. For XML-retrieval, something useful has been the cause of much debate. Generalized Precision Recall (inex_2002) [9] was criticized because it rewarded search engines for returning overlapping elements [10] – something shown to be a cause of frustration to users in the interactive experiments [11; 24].

The first alternative, NG [7] was criticized because it treated precision and recall separately and did not combine them into a single metric [32]. Because it assumed relevant content was uniformly distributed in an element, and because it did not address the overpopulated recall base problem [32].

There was very much a need for an appropriate metric when XCG [10] was introduced. Variants of this metric were used at INEX 2005, however there was debate. Woodley and Geva [32] showed that this metric is overlap negative that is, runs including overlapping elements were penalized. Piwowarski and Dupret [18] criticized it for having no user model.

Further metrics have been proposed: PRUM and EPRUM [18] model the behavior of a user in a hypertext environment. Such a user might click on a result in a results list, and then navigate from there to a relevant document through a hypertext link. This metric stochastically models this behavior. The versatility of this metric makes it appropriate for XML-retrieval – however we await the investigation into the behavioral parameters needed before it could be applied without controversy.

If passage retrieval is to take the place of element retrieval then metrics specifically designed for measuring passage-based performance are needed.

Two such metrics have been proposed for the TREC HARD track [25]. The first is the R-Precision of the F measure of individual passage precision and recall scores (passage precision and recall were computed on a character by character basis). This measure was shown to prefer a large number of short and contiguous passages over a small number of non-contiguous passages, that is, it encouraged identifying passages and then splitting them. The second was the bpref [2] of the top 12,000 characters.

TREC 2006 Genomics track [26] is proposing to use mean average passage precision (MAPP) where passage precision is computed as character overlap with relevant passages.

For XML-retrieval, Pehcevski and Thom suggest HiXEval [16], the F measure of the passage precision and passage recall, where passage precision and passage recall are defined with a tuning parameter to compensate for overlapping passages.

In summary, elements can be converted into passages. The performance of each of the runs thus-far submitted to INEX could be computed using a metric such as HiXEval, and the top performing algorithms identified. The performance of these could be graphed identifying if, or not, progress is being made

at XML-retrieval. A standard methodology could be put in place by which new algorithms are compared to old and statistical tests could be used to show the significance of any reported improvements.

5. Multiple Document Formats

XML is one of many semi-structured formats; SGML and HTML are two others. Or a document might be stored in plain unstructured text. The premise of XML-retrieval is that the structure, necessarily present in an XML document, can be used to improve performance. It might be used by a user to state, more specifically, where in a document relevant content might be found (a CAS query). Or it might be used by a search engine to increase the precision by returning only relevant elements (in a CO query). But does this structure help?

Trotman and Lalmas [28] compare the performance of a set of content only (CO) queries to their counterpart with structure added (CO+S queries). They found no statistical difference in performance of the best runs (submitted to INEX 2005) for the two types of queries on the same document collection.

The document collection they used was highly marked up. For both kinds of query (CO and CO+S) the search engines were able to, and did, take advantage of the structure. It is not at all obvious that the result would be the same if the same queries were run on documents not so strongly marked up. For the collection they used (INEX IEEE), such a derivative collection could be constructed by removing XML tags from each document leaving just the pain text. For the INEX Wikipedia collection, HTML, XML and plain text versions could be made available.

It is reasonable to assume that a search engine working without structured documents would not perform as well as one working with structured documents – but there are reasons to believe it might. Without structure the search engine is forced to identify relevant passages; and passages are more likely to be a better fit to the user's information need than are elements. It is reasonable to assume the precision might increase as a result. On the other hand, the element boundaries might help with the identification of passages so precision on the XML collection might be better.

Either way, it is reasonable to assume some queries will be better serviced by XML documents, some by HTML, and others by plain text. Knowing which will help identify the circumstances under which markup is of benefit, of how much benefit, and how much markup is needed for that benefit.

Opening up XML-retrieval to include HTML, plain text, and passage techniques will bring with it techniques from other information retrieval domains. This will provide an opportunity for understanding semi-structured document retrieval without being tied to XML.

6. Related Articles (Mini-Web)

Web retrieval differs considerably from other forms of information retrieval. The web is a dynamic hyperlinked environment where all pages are current and two pages can link to each other. In an academic document collection (such as the INEX IEEE collection) links can only point backwards in time – an academic article cannot be changed (after it has appeared in print) to cite papers published *post facto*.

Wikipedia articles are more like the web than like academic articles (the IEEE collection) in this regard. All articles are current and articles can cite each other, thus forming a mini-web. This leads to two problems: First the maintenance problem of keeping all cross links up-to-date. Second the selection of the mini-web when a new article is added.

In a dynamic environment new articles are constantly being added and old articles deleted, in both cases links must be maintained. Examining article 5001 on "Bathyscaphe Trieste", there is a section entitled "See also" that contains links to three other articles in the collection as well as one yet to be written. But there is no "See also" link to the vehicle's successor, the "Bathyscaphe Trieste II". The person who created (or maintains) the article also had to make the connection – this is tedious and requires extensive knowledge they may not have. Incoming links should also have been added to the collection, but from where? This task is even more tedious, perhaps prohibitively so as it requires updating many documents. The added value of a related articles task is clear.

An automated system would take a written article, find others like it (using XML-retrieval techniques) and suggest a mini-web of bidirectional links that a user may then (fetch) browse, filter, clean, and adopt as a desirable set of mini-web links. This process would both significantly enhance the collection and facilitate an activity that is highly unlikely to occur otherwise.

Creating cross-document links is a document similarity problem. This has already been examined in many domains (such as medicine [31]). But the Wikipedia offers a unique opportunity to examine document similarity in XML-retrieval. This is for one important reason – human generated links between documents are already in the collection. An almost cost-free evaluation method presents itself.

We expect a good concept formation system to return a set of links that is at least partially overlaps those that are already defined by the original contributors to the article.

The links between articles in the collection could be removed. Several articles from the collection could be selected as a test set, and a search engine would be tasked to insert links to relevant articles from the collection. The submitted runs would be compared to the ground truth – the links that were removed from the article in the first place.

If resources are available manual relevance judgments may be performed on those links identified by a search engine, but not already known to be appropriate. This would not be too onerous as it is a simple yes / no question – either the articles are related or they are not.

The performance of a search engine could also be computed in a straightforward manner. The precision with respect to a single article could be measured with mean average precision, and the mean of this might be used over a collection of query articles.

A clear task with a real need has presented itself. Topics already exist and evaluation is inexpensive. Best of all, the task only makes sense in a semi-structured hyperlinked environment – it is an ideal XML-retrieval task.

The task has an analogue for passages of text. In this case the need is not for "See also" links, but for links from the paragraph text to other articles. In this case a test set might be created by removing the links from pre-existing paragraphs. Natural language processing techniques might be used by a search engine to re-insert them. This task might be treated as a known entity searching problem and performance might be measured using mean reciprocal rank (MRR).

7. Question Answering

O'Keefe [15] examined the queries submitted to INEX 2003 and noted the high proportions that did not target elements as

return results. Trotman and Lalmas [28] identify only 13 (68%) of the 19 assessed CAS topics at INEX 2005 targeting elements. In the words of O'Keefe "If INEX is the answer, what is the question?".

Piwowarski *et al*. [19] observed that paragraphs are almost exclusively either fully specific or not specific to an information need. By comparison, only half of a relevant section element was specific on average. It is reasonable to conclude from their investigation that if elements are the right granularity of answer then the queries should be targeting paragraphs, or perhaps paragraphs and elements smaller than paragraphs: sentences, phrases, or single words.

Queries targeting words, sentences and paragraphs are not the usual domain of the *ad hoc* query. They usually target whole documents (or, of course, passages from documents). Words, sentences, or sometimes paragraphs are the granularity of answer expected of a question answering system.

INEX does not, at present, have a question answering track, but it is an obvious extension to both the NLP track and the Entity Ranking track. Questions would be asked in natural language and information (entity) extraction techniques would be used to identify answers. Standard methods such as those used at TREC Question Answering [30] would be used to evaluate performance.

It is reasonable to believe the markup present in an XML document will be of help in this task. The templates present in the INEX Wikipedia collection are of particular interest. One might ask "When was Edmund Burke first made Paymaster of the Forces?" to which the answer (1782) is held in a single template tag of the document on Edmund Burke (document 10030).

8. Conclusions

In this contribution we have examined evidence collected (by others) in favor of element retrieval with XML documents and shown that, in fact, it supports passage retrieval.

Prior studies into the agreement levels between judges show that that when judges are asked to identify relevant passages, and not elements, that the agreement level is very much higher than when asked to identify relevant elements.

Studies into which elements are most likely to be relevant show that paragraphs are essentially an atomic unit of relevance. Studies correlating passages and elements show that relevant passages in the text are not usually elements, but rather collections of consecutive elements (or, indeed, passages).

We discussed some of the problems facing element retrieval. Specifically we note that the problem of automatically identifying "too small" elements does not exist with passage retrieval. The problem of deriving the "ideal recall base" for focused retrieval disappears. We also drew evidence from a study that showed that the two dimensional relevance, itself problematic, is unnecessary if assessors judge passages and not elements. Methods for search engine evaluation, we believe, are already in place if metrics like HiXEval are used.

We examined possible user tasks for passage retrieval and showed that the existing XML-retrieval tasks *focused* and *thorough* are analogous under passage retrieval. We examined the *relevant in context* task, and the *best in context* track and showed they not only do they exist essentially unchanged with passages, but that the differences between all these tasks is easily explained.

Finally we examined possible future XML-retrieval tasks and showed that a paradigm shift to passage retrieval not only has no negative impact on these tasks, but is likely to enhance them.

The future of XML-retrieval is, we believe, with passage retrieval and not element retrieval. We showed that the transition from element to passages can be smooth, and that methods are already in place to make the transition. We now propose that INEX 2007 fully embrace passage retrieval and run parallel passage and element tasks with the intent of moving solely to passages for 2008.

9. References

[1] Bernardi, R., Jijkoun, V., Mishne, G., & de Rijke, M. (2003). Selectively using linguistic resources throughout the question answering pipeline. In *Proceedings of the 2nd CoLogNET-ElsNET Symposium*.

[2] Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th ACM SIGIR Conference on Information Retrieval*, (pp. 25-32).

[3] Clarke, C. (2005). Range results in XML retrieval. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 4-5).

[4] Clarke, C., Kamps, J., & Lalmas, M. (2006, to appear). INEX 2006 retrieval task and result submission specification. In *Proceedings of the INEX 2006 Workshop*.

[5] Fuhr, N., Gövert, N., Kazai, G., & Lalmas, M. (2002). INEX: Initiative for the evaluation of XML retrieval. In *Proceedings of the ACM SIGIR 2002 Workshop on XML and Information Retrieval*.

[6] Fuhr, N., & Großjohann, K. (2000). XIRQL an extension of XQL for information retrieval. In *Proceedings of the ACM SIGIR 2000 Workshop on XML and Information Retrieval*.

[7] Gövert, N., Kazai, G., Fuhr, N., & Lalmas, M. (2003). *Evaluating the effectiveness of content-oriented XML retrieval*: University of Dortmund, Computer Science 6.

[8] Harper, D. J., & Lee, D. (2004). On the effectiveness of relevance profiling. In *Proceedings of the 9th Australasian Document Com-puting Symposium*, (pp. 10-16).

[9] Kazai, G. (2003). Report of the INEX 2003 metrics working group. In *Proceedings of the INEX 2003 Workshop*.

[10] Kazai, G., Lalmas, M., & de Vries, A. P. (2004). The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of the 27th ACM SIGIR Conference on Information Retrieval*, (pp. 72-79).

[11] Kim, H., & Son, H. (2004). Interactive searching behavior with structured XML documents. In *Proceedings of the INEX 2004 Workshop*, (pp. 424-436).

[12] Mass, Y., & Mandelbrod, M. (2004). Component ranking and automatic query refinement for XML retrieval. In *Proceedings of the INEX 2004 Workshop*, (pp. 73-84).

[13] Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lacatusu, F., Novischi, A., Badulescu, A., & Bolohan., O. (2002). LCC tools for question answering. In *Proceedings of the 12th Text REtrieval Conference (TREC-11)*.

[14] Ogilvie, P., & Lalmas, M. (2006 (submitted)). Investigating the exhaustivity dimension in contentoriented XML element retrieval evaluation.

[15] O'Keefe, R. A. (2004). If INEX is the answer, what is the question? In *Proceedings of the INEX 2004 Workshop*, (pp. 54-59).

[16] Pehcevski, J., & Thom, J. A. (2005). HiXEval: Highlighting XML retrieval evaluation. In *Proceedings of the INEX 2005 Workshop*.

[17] Pehcevski, J., Thom, J. A., & Vercoustre, A.-M. (2005). Users and assessors in the context of INEX: Are relevance dimensions relevant? In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 47-62).

[18] Piwowarski, B., & Dupret, G. (2006). Evaluation in (XML) information retrieval: Expected precision recall with user modelling (EPRUM). In *Proceedings of the 29th ACM SIGIR Conference on Information Retrieval*.

[19] Piwowarski, B., Trotman, A., & Lalmas, M. (2006 (submitted)). Sound and complete relevance assessments for XML retrieval.

[20] Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M., & Payne, A. (1995). Okapi at TREC-4. In *Proceedings of the 4th Text REtrieval Conference (TREC-4)*, (pp. 73-96).

[21] Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of the 28th ACM SIGIR Conference on Information Retrieval*, (pp. 162-169).

[22] Shaw, W. M., Wood, J. B., Wood, R. E., & Tibbo, H. R. (1991). The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research,* 13, 347-366.

[23] Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th ACM SIGIR Conference on Information Retrieval*, (pp. 21-29).

[24] Tombros, A., Larsen, B., & Malik, S. (2004). The interactive track at INEX 2004. In *Proceedings of the INEX 2004 Workshop*, (pp. 410-423).

[25] TREC. (2003). Hard, high accuracy retrieval from documents TREC 2003 track guidelines. TREC. Available: http://ciir.cs.umass.edu/research/hard/guidelines2004.html [2006, 16 June].

[26] TREC. (2006). TREC 2006 genomics track draft protocol. TREC. Available: http://ir.ohsu.edu/genomics/2006protocol.html [2006, 16 June].

[27] Trotman, A. (2005). Wanted: Element retrieval users. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 63-69).

[28] Trotman, A., & Lalmas, M. (2006). Why structural hints in queries do not help XML retrieval. In *Proceedings of the 29th ACM SIGIR Conference on Information Retrieval*.

[29] Trotman, A., & O'Keefe, R. A. (2003). Identifying and ranking relevant document elements. In *Proceedings of the 2nd workshop of the initiative for the evaluation of XML retrieval (INEX)*.

[30] Voorhees, E. M., & Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd ACM SIGIR Conference on Information Retrieval*, (pp. 200-207).

[31] Wilbur, W. J., & Coffee, L. (1994). The effectiveness of document neighboring in search enhancement. *Information Processing & Management,* 30(2), 253-266.

[32] Woodley, A., & Geva, S. (2005). XCG overlap at INEX 2004. In *Proceedings of the INEX 2005 Workshop*, (pp. 25-39, pre-proceedings).

[33] Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *Transactions on Information Systems,* 22(2), 179-214.