Report On The SIGIR 2006 Workshop on XML Element Retrieval Methodology

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

On the 10th of August 2006 the SIGIR 2006 Workshop on XML Element Retrieval Methodology was held as part of SIGIR in Seattle, Washington, USA. Six papers were presented in four sessions. This report outlines the events of the workshop and summarized the major outcomes.

1 Introduction

In recent years XML Information Retrieval researchers have turned their attention to research questions beyond improving the performance of a search engine. In element retrieval the search engine is required to identify not only relevant documents, but relevant document parts (XML elements) that satisfy the user's information need. But is this an artificial task for which no users exist? Do XML elements provide the best granularity for results? What kinds of questions can users ask in their queries? There are a multitude of questions, and at this workshop many were raised and discussed in an open forum.

2 Sessions

The workshop was divided into 4 sessions with 6 papers presented. Each speaker was allotted 45 minutes during which they gave a 20 minute presentation and lead a discussion for 25 further minutes.

2.1 Session 1

Kamps presented work conducted jointly with Larsen [3], the analysis of a questionnaire conducted as part of topic creation at INEX 2006. Topic authors were first asked to submit a topic then asked to answer 19 short questions about that topic. These questions were formulated to answer a small number of research questions about XML element retrieval, specifically: what do users expect from an element retrieval system? What kind of information needs do they have? What sort of results do they expect?

In total 195 topics were submitted by 81 authors. The topics were formulated against the INEX Wikipedia collection, presented in XML format. The sample of users was taken from the topic creators (INEX participants) and so the results may be indicative only for that very special group of users, although this is not necessarily the case.

Several questions were asked to verify the topic set. Most authors were, indeed, familiar with the topic and almost all topics were real queries that would have been typed into a web search engine. About half

the topics were specific while the others were (presumably) general purpose. A few more authors were interested in reading "a lot" of relevant information than not. As expected, topic familiarity and specificity were correlated while topic specificity and reading "a lot" were inversely correlated.

Questions relating to the nature of the topic set were also asked. An answer was expected to come from combining different parts of documents, and was expected to be found in more than one document (even though, in general, the topic was not based on a previously seen part of a document). Relevant material was expected in elements of different sizes ranging from sentences through to complete articles, topics were not expected to be satisfied by a single relevant answer. Most authors stated that it was important to read several related articles, and important to know all relevant results, and that they would prefer to see all relevant results. This tells us that recall oriented metrics are important for element retrieval.

Most authors did not assume perfect knowledge of the DTD, nor did they know the structure of a single relevant result. They (presumably consequently) assumed references to structure were vague and imprecise. Most would prefer results presented in the context of the document rather than out of context.

Kamps and Larsen are further investigating how their results might be used to categorize topics, and to help understand the differences between the INEX tasks.

The survey replies have been included in the INEX 2006 collection.

2.2 Session 2

Kazai and Ashoori conducted an analysis of assessments data from the (pre-INEX) Focus project into structured information retrieval and from INEX 2005 [4]. Kazai's presentation focused on the analysis of best entry points (BEPs) collected during the construction of the Shakespeare XML test collection, and how those results might broaden our understanding of BEPs within INEX.

The Focus project conducted experiments with 11 English and Drama students and 12 Shakespeare plays. Participants created queries and provided both relevance and BEP assessments, where BEPs were defined as optimal starting points for browsing in order to access relevant information.

Participants submitted 215 queries (43 were chosen) of which 43% were Content And Structure (CAS) and 57% were Content Only (CO). Of the CAS queries the most commonly used structural condition was the unit of a <play> (80%). This provides support for the INEX Fetch & Browse task which presents results within document context.

Assessment was done by hand on paper with a highlighter pen (binary relevance). An on-screen yellow highlighting method had been adopted by INEX in 2005. Experimentation with several alternative methods has shown this to be the best method to date.

Topics were assessed by multiple assessors and the analysis of cross judge agreement levels suggests assessors agree on the general area containing relevant information but not on the exact location. Similar results have been found at INEX.

The BEPs were obtained via interviews. 521 BEPs were acquired with an average of 12 BEPs per query, 94% of which were leaf nodes of the document tree or <speech> nodes. This means that more specific answers were preferred. Three types of BEPs were identified: Start Reading Here, Container, and Combined BEPs.

A Start Reading Here BEP (45%) is characterized by being one node in a sequence of relevant nodes. Most (62%) were the first leaf node in a sequence, but some were the last - providing evidence of the importance of presenting results in context. Container BEPs (31%) are characterized by being the parent nodes of a set of relevant elements. Most (80%) were speeches with the remainder being almost exclusively scenes. Combined BEPs (one of a sequence of parent nodes) occurred 24% of the time with 94% of them being the first speech.

In conclusion Kazai suggested that both CO and CAS topics are natural; that the whole document was the natural semantic unit of thought; that assessors agree on the general area of relevance but not on the exact location; that agreement levels in BEPs was high and that Start Reading Here BEPs were preferred.

Kazai's presentation sparked considerable debate on BEPs at INEX. Geva put the case for search engines identifying a single BEP but assessors being permitted to identify multiple BEPs in each document. That is, if there are several "best" points from which to start reading a document then the assessor should identify them all even if the search engine can present the user with only one in a results list. Others argued contrariwise, that assessors should identify the one true "best" entry point and the search engine should identify this. For 2006 INEX has adopted the latter.

Kazai's presentation also challenged the community to agree on and adopt standard performance metrics for measuring the performance of a search engine with respect to BEPs. A suitable metric might, for example, use a distance measure. An open question, however, concerns the issue of normalisation by document length. Trotman suggested a Levenshtein distance based on traversing from an identified entry point to a BEP.

The second presentation was by Lehtonen [5]. He identified three approaches to XML retrieval, the IR view, the DB view, and the Document Engineering view. Of the former, most systems are experimental, XDBMS systems are both experimental and commercial; of the latter, XML document management systems are not only available commercially but are common. It is ironic (and perhaps problematic) that user studies conducted by the XML research community are on experimental systems and not on commercial systems.

Lehtonen identified two problems with prior user studies. First, the user studies are usually conducted on experimental systems. Second, the experimental environment leads to experimental results. The experiments are often sidetracked into interesting research questions rather than focusing on important user issues, the consequence of which is that the results often remain in the research community and have little penetration outside that community.

As yet no study has compared user performance on an element retrieval system to one on a document retrieval system, perhaps one of the most fundamental questions facing XML-Retrieval is: does the XML help the user satisfy their information need more efficiently?

Results on the IEEE collection are not yet known to generalize to any other collection. The IEEE documents tend to be large and sub-document retrieval may well be of immense value, but in a collection in which the documents are much smaller combining the results from multiple documents may be a better approach.

Arguments for changing the relevance assessment method in interactive experiments were also given. A system like that also proposed by Pehcevski [7] was suggested with the importance of presenting

elements in context being stressed – how can a judge identify if an element is *just right* in size if the larger context is not given?

Examining the most recent INEX topic set, Lehtonen identified the importance of the experiment comparing the performance of CO to CAS queries for the same information need. However the structural hints in that experiment were not about content so the queries could easily be answered without the structural hints. Experiments in which the structure is a vital part of the answer are needed. Such experiments might be conducted on a collection in which the tags delineate semantic units within the document (such as medical literature separating <diagnosis> from <treatment>). There is a growing interest in building user interfaces to formulate queries in such environments and Lehtonen further stressed the importance of this.

Are there any users of XML element retrieval systems? Lehtonen drew our attention to Scopus which holds over 12,000 documents and has tens of thousands of users. Although the XML is not visible to the user, this does not prevent it from being an XML retrieval system. A second system was also discussed, the MarkLogic XML content server, it uses XQuery and provides element granularity search. Future academic user studies and future academic investigations into XML element retrieval systems should consider commercial systems and build on experiences already learned from deployed search engines.

2.3 Session 3

Dopichaj [1] answered many of the questions that have been asked about commercial element retrieval systems. He identified three commercial web sites believed to rely on structured information retrieval: Books27x7, Safari, and Google Book Search. Each provides access to books and chapters of books. All have search facilities

Google book search provides results including highlighting of text in-context. This provides evidence of the utility of the Fetch & Browse task; if a commercial online system relies on it, it is reasonable to assume it is useful.

Both Books24x7 and Safari provide search over whole books but return parts of books as results (sub-document results). Both also provide results of varying granularity mixed in together; overlapping results in the results list. They are commercial Thorough retrieval systems.

Both provide “advanced” search facilities with functionality for searching specific document structures (such as code fragments or titles). Evidence of a need for supplying structural hints in queries, and further evidence for the kinds of graphical query builder discussed by Lehtonen [5].

Dopichaj investigated the history of the Safari user interface and showed that it had hardly chanced since 2002. This, he suggested, is evidence of the success of the system, and evidence that XML element retrieval systems have a real world application.

The second presenter in the session, Geva, discussed work (conducted with Hassler and Tannier [2]) on the natural language track at INEX and proposed extensions to the INEX query language (NEXI) needed for the track. His proposed extensions known as NE^{2}XI (and also as XOR) are a strict extension to NEXI and continue with the philosophy that the query is a collection of hints to the search engine.

The first proposed extension allows the user to provide multiple Boolean separated clauses in a query. NEXI already allows the user to specify multiple target elements, but only a single clause. NE^{2}XI by

comparison allows the user to ask questions like “sections about relativity or references about Einstein”. In addition the Boolean operator AND NOT is now supported.

Qualifiers have been added to the language and examples of several were given. A structural hint might be strict or vague, and the language now allows the user to specify which interpretation to use (as a qualifier). For natural language applications part of speech qualifiers were added. Qualifiers for case sensitive searching were also added allowing the search to differentiate between `AJAR` and `ajar`.

The introduction of the Wikipedia collection has prompted the introduction of language facilities for specifying hypertext links. The new LinkTo() and LinkFrom() functions have a similar syntax to about(). NE^{2}XI also further extends the wildcard support in path specifications.

Geva’s presentation initiated a discussion on XQuery. Westerveld put the case that should NEXI become sufficiently complex then XQuery Full-Text might serve the same purpose.

2.4 Session 4

In the final session Trotman (in work with Geva [8]) argued that element retrieval as done at INEX relies on the assumption that an element has the correct granularity of result to return to the user. If the query has a semantic meaning and the XML corpus is based on semantic mark-up then the assumption is that the two match. This assumption can be tested and drawing on prior work Trotman concluded that sequences of consecutive paragraphs are better suited as answers – at least for the INEX IEEE document collection.

To date the agreement levels between multiple judges at INEX have been measured only twice (2004 and 2005). Although the sample in both cases was very small, the agreement level increased substantially between 2004 and 2005. Trotman argued that this was due to the introduction of the yellow highlighting method used to identify relevant content. That change was from judging elements to judging passages, and the increased agreement level between assessors lends support to the idea that passages match user preferences better than elements.

When looking at which elements were most likely to be relevant, paragraphs prevailed. This was not unexpected as the number of paragraphs in the collection vastly outnumbers the number of sections and articles. When comparing the specificity of different relevant elements, paragraphs were almost always completely specific (94% of the text was deemed relevant on average) but only about half (51%) of a sections content was deemed relevant on average. Trotman concluded that assessors were identifying sequences of consecutive paragraphs. He went on to show that of these sequences of relevant paragraphs, only about one third (36%) were themselves elements.

Recent work by Ogilvie and Lalmas [6] suggests that assessment using specificity alone is as stable as using the two relevance dimensions – that is, highlighting passages is as good as judging elements. Moreover, the cross-judge agreement level of passages appears to be reasonable.

Trotman claims the case for passage retrieval is compelling and there are natural analogues of the existing Focused, Relevant in Context, and Best in Context tasks for passage retrieval. Metrics exist both within INEX (HiXEval) and in TREC, and the assessment tool already used at INEX could be used unchanged.

Additional new tasks were suggested, with an emphasis on two: first the comparison of XML-Retrieval systems on the same corpus, but with varying levels of structural granularity in the documents (from

XML to plain text). The second was the task of identifying similar documents. The Wikipedia was identified as a good collection for this second task because documents already contain links to other documents and are therefore an ideal topic set.

2.5 Other Included Work

It has been believed that binary relevance is not appropriate for XML element retrieval due to the hierarchical nature of XML documents. If a paragraph is relevant then so too is any section containing that paragraph, but (perhaps) to a lesser extent. This facilitated the two-dimensional relevance scale used at INEX. In Pehcevski's paper [7] he provides evidence that this scale is just too complex for an assessor. Specifically, agreement levels are high only at the ends of the assessment scale (highly relevant and not relevant) and assessors perceived the two dimensions as one.

Pehcevski suggests a new two dimensional relevance scale. In the first an element is identified as either *highly relevant*, *relevant*, or *not relevant*. In the other an element is *just right*, *too large*, or *too small*. He demonstrates that this two dimensional relevance scale categorises into five possible grades that are easily understood by a judge: *Exact Answer*, *Partial Answer*, *Broad Answer*, *Narrow Answer*, or *Not Relevant*. Evidence from the INEX 2005 Interactive Track is given in support of the ability of assessors to understand the categorization. Also given is a mapping to the proposed relevance categories from the continuous specificity scale of before. Pehcevski demonstrates that the best performance is, indeed, achieved when a search engine identifies *Exact Answer* elements.

3 Major Outcomes

The outcome of a workshop comes not only from the presented papers, but also from the discussions before, during, and after the workshop. From all of these several changes have been proposed.

The assessment method used at INEX 2006 results in a collection of relevant passages for each topic. At the workshop Geva and Kazai started working on a method of scoring elements from these passages. This difference in granularity had not, to this point, been tackled and must be so before the 2006 runs can be scored.

Discussion on BEPs has opened the possibility for multiple BEPs within a single document. Pushing strongly for allowing the assessor to do this were Geva and Trotman. It is likely INEX will examine the feasibility of doing this in future rounds.

Trotman put a strong case for passage retrieval. At INEX 2007 a passage retrieval task is anticipated. Although details are not yet clear, it is likely to be a heat-map based task; the identification of relevant passages within relevant documents sorted first by document, then by relative passage importance within document.

Some groups have already started working on identifying related documents, and it is not a new task to IR. Future INEX rounds will likely adopt such a task as the cost of assessment is very low.

4 Acknowledgements

We would like to thank ACM and SIGIR for hosting this workshop. We would also like to thank the program committee, the paper authors and the participants for a great workshop. Some workshop paper authors contributed to this paper prior to submission (thanks). The University of Otago is hosting the workshop proceedings which are online (http://www.cs.otago.ac.nz/sigirmw/).

5 References

[1] Dopichaj, P. (2006). Element retrieval in digital libraries: reality check. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, (pp. 1-4).

[2] Geva, S., Hassler, M., & Tannier, X. (2006). XOR- XML oriented retrieval language. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, (pp. 5-12).

[3] Kamps, J., & Larsen, B. (2006). Understanding differences between search requests in XML element retrieval. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, (pp. 13-19).

[4] Kazai, G., & Ashoori, E. (2006). What does Shakespeare have to do with INEX? In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, (pp. 20-27).

[5] Lehtonen, M. (2006). Designing user studies for XML retrieval. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, (pp. 28-34).

[6] Ogilvie, P., & Lalmas, M. (2006). Investigating the exhaustivity dimension in content oriented XML element retrieval evaluation. In *Proceedings of the 15th ACM Conference on Information and Knowledge Management (CIKM 2006).*

[7] Pehcevski, J. (2006). Relevance in XML retrieval: The user perspective. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, (pp. 35-42).

[8] Trotman, A., & Geva, S. (2006). Passage retrieval and other XML-retrieval tasks. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, (pp. 43-50).