Report on The INEX 2005 Workshop on Element Retrieval Methodology

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

Mounia Lalmas
Department of Computer Science
Queen Mary University of London
London, UK
mounia@dcs.qmul.ac.uk

Abstract

On the 30th July 2005, a workshop on XML element retrieval methodology was held as part of the Information Retrieval Festival at the University of Glasgow. Ten papers were presented in four sessions. Each session addressed one aspect of the methodology of XML IR: metrics, users, interactive / heterogeneous, and judging (relevance ranking was excluded). This report outlines the events of the workshop and the major outcomes.

1 Introduction

In XML element retrieval the retrieved result is a part of a document (an XML element) and not a whole document. This leads to many complex methodological problems. If, for example, a given element is relevant, then all parent elements must also be relevant – so what do precision and recall mean in this environment? Of course, if a user is looking for a particular element then they must specify this in their query – but queries are usually short and do not ordinarily contain structural targets.

Many of these problems have been identified in working groups at INEX [2]. Nonetheless, it has become necessary to address these issues separately and in an open forum. The INEX 2005 Workshop on Element Retrieval Methodology provided such a forum.

2 Sessions

The workshop was divided into four sessions. Each speaker was allotted 45 minutes, during which they were expected to give a 20 minute presentation and direct a discussion of 25 minutes.

2.1 Session 1: Metrics

In the first presentation [5] Gabriella Kazai reminded us that in the debate over metrics, we must not loose sight of the aim of the metric. The metrics must rank systems according to how well they satisfy the user given a retrieval task. For element retrieval she provided a list of requirements that any metric must satisfy: it must consider element size, score near-misses, neither reward nor punish overlap, consider presentation method, handle the 2-dimensional graded relevance judgments, and work from the overlapping judgments. Working through four of the six published metrics, she identified the strengths and weaknesses of each. No single one was identified as ideal.

Djoerd Hiemstra presented an opposing view [3]. Comparing the precision / recall graphs from TREC and from INEX he showed a methodological difference. Using Robertson's compatibility argument (standard measures should be used unless there is good reason for doing otherwise) he suggested there should not be. He advised bringing INEX closer to TREC by reporting precision at fixed document cut-offs and, as there is (currently) no evidence showing quantization of different relevance degrees has any effect on the metrics, binary quantization should be used. Including element overlap scores, a verbal reading of his metric might be "at 10 documents, 5 documents are relevant, however of those, 4 overlap with each other".

Discussion during the session was lively. David Hawking reminded us that the metrics should reflect the use of the system. Identifying the "use cases" of element retrieval will shed light on what should be measured. At present the INEX tasks are identified as: i) thorough retrieval of all relevant elements, ii) focused retrieval of non-overlapping elements, and iii) Fetch & Browse of documents. Each of these must be measured appropriately. Shlomo Geva asked that in future the metric used for each task is stated as part of the task. Keith van Rijsbergen asked if the existing metrics correlate – they do not.

David Hawking asked why the semantic structure of XML is not taken into consideration when computing performance. Simply put, an error of one word across a paragraph boundary has little effect on the semantics of a retrieved unit, whereas an error of one word from author to title has a catastrophic effect. He argued that this should be accounted for in the metrics.

2.2 Session 2: Users

Roelof van Zwol presented a graphical user interface for querying structured documents called Bricks [11]. This interface was designed to reduce the complexity of query formulation, while at the same time maximising the expressive power of the language. Usability experiments were run with 54 students who were trained in using the INEX query language NEXI [10]. 27 topics divided into three complexity groups were used and participants were surveyed with a questionnaire after the experiment. Overall, using queries without structure (NEXI CO) was quicker than using either Bricks or content and structure (NEXI CAS) queries. User behaviour was different for each approach. With CO queries users were happy to work iteratively with the search engine. With CAS topics, there was less iteration, but a great deal of syntax checking. With Bricks it took longer to build a query, however less iteration was seen.

In discussion Anastasios Tombros identified the difference between performance experiments and usability experiments. This experiment is trying to close the gap. Ross Wilkinson sees languages like NEXI fitting between the user and the search engine – and stressed the importance of experiment like that of van Zwol.

Gabriella Kazai asked why no iteration was needed when an interface like Bricks was used. Ross Wilkinson asked if this was affected by the complexity of the query. In response, the users did not need to concern themselves with the syntax of the query, only with the semantics – consequently they tended to write good queries. With complex queries there remained confusion over specifying paths; some users preferred to specify paths from root to leaf (a//b//c), however others preferred the other way around (c\\b\\a).

In Börkur Sigurbjörnsson's presentation [4] two types of user were identified: ignorant users who only know the names of some of the elements; and semi-ignorant users who additionally have a limited knowledge of the hierarchical structure. He showed that NEXI is an appropriate language for semi-ignorant users, while a proper sub-language of NEXI is more suitable for ignorant users. Finally, he suggested that there exists no query in which structure was an inherent part of the information need.

This final point drew much discussion. One side argued that structure can never be a part of the information need otherwise the information need would not be satisfied if the structure were removed. The other side argued that structure was implicit and the removal of explicit structure was not the removal of all structure – and therefore structure was necessary to satisfy all information needs.

Ross Wilkinson asked whether Sigurbjörnsson's research might be used to identify the ideal user of an element retrieval search engine.

In the next talk Trotman claimed that the IEEE documents are atomic and therefore unsuitable for element retrieval [9]. He suggested that a more chaotic collection is needed as element retrieval is about "plucking" elements from a chaotic mix. He argued that the INEX topics do not display much structural use as the collection is not open to structural querying. After briefly discussing the lack of correlation between metrics he moved on to examine the judgments. He showed the agreement levels of INEX judges are in line with TREC when considering document-level binary relevance decisions. When looking at binary relevance of elements the agreement levels are very low. Considering exact (10 point) agreement levels between different judges of the same topic, he said they show "nearly complete disagreement". Trotman concluded that the most important next step for element retrieval is the identification of a user base on which studies can be conducted.

2.3 Session 3: Heterogeneous / Interactive Search

Ray Larson reminded us that to specify a structural query a user must be, at least in part, familiar with the structure of a document [7]. Learning the DTD of a document collection is known to be a burden for a user. In a heterogeneous environment that burden is too large for any ordinary

user. Quite simply, learning hundreds of DTDs is impractical. He proposed an XML postulate of impotence "You can either have heterogeneous retrieval, or precise element specifications in queries, but you cannot have both simultaneously". This problem might be tackled by moving from explicit specification of elements in a query to abstract specification. In this way a user no longer needs to know the multitude of different tags used for "author", but can instead abstractly specify that the search be limited to author names. Using Z39.50 and the Dublin Core as examples of successful systems, he identified some obstacles that need to be overcome in XML.

Birger Larsen gave an overview of the INEX 2004 interactive experiment done jointly with Queen Mary University of London and the University of Duisburg-Essen [6]. 10 groups participated with at least 8 users per site. Users entered content only (CO) queries into an element retrieval engine and were asked to judge the relevance of retrieved results. Unfortunately, the experiment was highly obtrusive; it required a high cognitive load to make the judgments, which affected the natural browsing behaviour. At the conclusion of the experiment only 60% of the viewed elements were judged by users. Alternative experimental methods were proposed such as measuring browse time, eye tracking, book-marking, and talk-aloud studies.

Discussion on both talks focused on agreement with the speakers, with particular emphasis on using these experiments to investigate user requirements and preferences rather than simply to verify laboratory experiments. The interactive experiments could, for example, be used to determine if elements returned in-context is preferred over returning elements out of context.

Gabriella Kazai asked about plans to compare different interfaces to the same XML search engine.

Mounia Lalmas asked if users changed their mind once they had made an assessment. Users do reassess the same element, but to what extent is unknown.

2.4 Session 4: Judgments

Charlie Clarke made the case that single elements may not make sense on their own as retrieval results [1]. He proposed including ranges of elements, and provided syntax for doing so. Drawing a similarity to passage retrieval he said the relevance judgments should also be made as ranges. This work requires new metrics to measure the performance of a retrieval run, and Clarke provided two such metrics.

Shlomo Geva presented an analysis of the novelty of the relevance judgments [12]. In this study he showed that the performance of a meta-search engine is better than that of the judgment pool. Examining the effect of out-of-pool judgments he showed that these judgments have only a minor effect on the relative performance of search engines. From this he suggested a new method of pooling, and to drop forced (out-of-pool) judging. By comparing the relative order of XML search engines computed using graded verses binary relevance judgments, significantly different performance was shown, suggesting that graded relevance judgments remain important.

Heated debate followed Geva's presentation when he suggested that overlap has no effect on the relative performance of search engines. He presented (contentious) results that showed the performance of each of the INEX 2004 runs computed with overlapping elements removed.

Anne-Marie Vercoustre presented work done jointly with RMIT University in which the agreement levels between judges and interactive users were analysed [8]. The judgments of the topic judges and the 88 users from the interactive experiment were compared for two of the interactive topics. They found that agreement levels for the extremes of the relevance dimensions (not-relevant, E1S1, and E3S3) were high, whereas at other points the agreement levels were low. They also found users did not provide relevance assessments for overlapping elements. She proposed a simpler four point relevance scale in which relevant information is either: too broad, too narrow, or just right (the fourth point being "not relevant"). As, in their definition, only one element in a path can be "just right", the multiple E3S3 element problem will no longer exist. Vercoustre provided a partial mapping from the old judgment scale to the proposed scale and noted that no change to the existing metrics was necessary. She also supported moving to a "yellow highlight" approach for judging, showing that the proposed judgment scale is particularly adapted to grading in this way.

Ross Wilkinson asked if it was possible to infer the judgments from the yellow highlighting – if this is possible then the judgment process could be simplified reducing the time spent assessing.

3 Major Outcomes

The outcomes of the workshop come not only from the presented papers, but from on-going email exchanges on the INEX organisers email list, the INEX participants email list, discussion during the sessions, during lunch, leading up to the workshop, and immediately following the workshop. From these several changes have been made to the element retrieval methodology used at INEX.

Several presentations questioned the use of the inex_2002 metric. The metric rewards overlapping elements, in the interactive experiments users were shown to disfavour this. The metric has been deprecated because it is not obvious how to extend it so that it can both consider near misses when evaluating performance, and not reward overlapping elements. The INEX 2005 official metric will be XCG.

Discussion on methods to tighten the task definitions suggested more information was needed up-front. In particular the metric used for each task should be published with the task definition. From 2006 this will occur.

There was disagreement about the suitability of the IEEE document collection for element retrieval. INEX has recently gained permission to use the Lonely Planet Guide for experiments and has made this collection available to participants.

The 10 point two dimensional relevance scale has been deprecated in favour of a simpler "yellow highlighter" judgment method. Specificity is computed as the ratio of relevant to irrelevant continent. Exhaustivity will be specified manually on a three point scale: not, partly, and very.

4 Acknowledgements

We would like to thank the University of Glasgow for hosing the workshop. Thanks also go to the program committee, the paper authors, and the participants. INEX is an activity of the DELOS network of excellence in digital libraries. The University of Otago is hosting the workshop proceedings which are online at the workshop website and available for download along with slides of the presentations (http://www.cs.otago.ac.nz/inexmw/).

5 References

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

[2] Fuhr, N., & Lalmas, M. (2004). Report on the INEX 2003 workshop, Schloss Dagstuhl, 15-17 December 2003. *SIGIR Forum,* 38(1), 42-47.

[3] Hiemstra, D., & Mihajlovic, V. (2005). The simplest evaluation measures for XML information retrieval that could possibly work. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 6-13).

[4] Kamps, J., Marx, M., de Rijke, M., & Sigurbjörnsson, B. (2005). Understanding content-and-structure. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 14-21).

[5] Kazai, G., & Lalmas, M. (2005). Notes on what to measure in INEX. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 22-38).

[6] Larsen, B., Tombros, A., & Malik, S. (2005). Obtrusiveness and relevance assessment in interactive XML IR experiments. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 39-42).

[7] Larson, R. (2005). XML element retrieval and heterogeneous retrieval: In pursuit of the impossible? In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 43-46).

[8] 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).

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

[10] Trotman, A., & Sigurbjörnsson, B. (2004). Narrowed Extended XPath I (NEXI). In *Proceedings of the INEX 2004 Workshop*, (pp. 16-40).

[11] van Zwol, R., Baas, J., van Oostendorp, H., & Wiering, F. (2005). Query formulation for XML retrieval with bricks. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 80-88).

[12] Woodley, A., & Geva, S. (2005). Fine tuning INEX. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 70-79).