Introduction to 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

1. INTRODUCTION

With a wealth of documents originating in markup languages such as XML, it is appropriate to ask how this markup might be used in information retrieval. One answer is to change the focus of retrieval from whole documents to document elements.

In document-centric IR the user searches whole documents and is returned a ranked list of documents that match their queries. By contrast, in element retrieval document elements are returned – perhaps a chapter of a book, or a section of an academic paper.

Since 2002 the annual INEX workshop [2] has been examining element ranking algorithms for XML documents. Most specifically, the IEEE collection of 12,107 documents. Arguably progress has been made.

It is this "arguably" that has become the center of attention. On the outset it would appear as though element retrieval is a simple derivation of document retrieval – but experience at INEX has shown this to be far from the truth.

A document centric search engine makes a binary decision about the relevance of a given document – either it will appear in a result list or it will not. It cannot "partly appear".

An element centric search engine having decided a piece of text is relevant is faced with how to return that information. Perhaps only a paragraph is relevant, or perhaps the sub-section, or the section, or it may be the entire document. The same piece of text can be returned in many different ways.

When humans are making judgment decisions, they too, are faced with similar problems. If a given paragraph is relevant, then surely a containing section is also relevant. How much more so, or less so?

Combining these, how can the performance of a search engine be measured?

There are clearly methodological issues in element retrieval, and these need addressing. It is these issues that are of interest at this workshop.

For many the most pressing issues is this: when there is no community accepted methodology it is not possible to claim any one system is better than any other.

2. FOCUS OF THE WORKSHOP

The workshop was organized to address some of the methodological issues in element retrieval. Specifically six areas requiring attention were identified: theory, application, measurement, judgment, experience, and other. Two areas were excluded: ranking algorithms and existing software.

2.1 Theory

A sound theoretical basis for element retrieval is yet to be established, both in terms of the document collection, and the interaction model.

It is not clear what properties of an XML document collection make it more suitable for element retrieval than for document retrieval. It is also not clear what properties make that collection either "heterogeneous" or "multimedia".

As yet there is no established theoretic basis of interaction with element retrieval – although the INEX interactive track is investigating this [14]. It is not clear when an element is a better answer than a document, or if elements must be bound by context when returned to the user.

2.2 Application

It is entirely possible that many of the methodological issues can be resolved if there existed an application of element retrieval (outside the research community). It is not clear where to look for such an application, or if such an application will ever exist.

2.3 Measurement

One of the methodological issues addressed from the outset is that of performance measures. Kazai [7] identifies five different metrics that have been proposed, and there are more besides. It is clear that these metrics measure different things, however what is not clear is what should be measured – or how to measure it. With no single community accepted performance metric, it is impossible to identify one algorithm as any better than any other. Consequently, progress on ranking algorithms is impossible to make.

2.4 Judgment

At present INEX judgments are made on two separate dimensions, one is a measure of how specific the element is, and the other how exhaustive the element is. Each is on a four point scale (not, marginally, fairly, highly), giving a total of ten grades (if an element is not specific it cannot be exhaustive, and *vice versa*).

Prior investigations into the judgments (such as that of Pehcevski [11]) have raised questions as to whether or not the assessors understand this scale – and it is not clear. It is entirely possible that the complexities of grading a "near miss" element (that encloses relevant information but is not itself entirely relevant) are beyond the capabilities of a subjective assessor.

What is clear is that different assessors have different marking conventions. Some will mark references as valid, where others

may not. Under investigation is the judgment process and the judgments. Are they, or are they not sound?

2.5 Experience

Drawing on the experience of other evaluation workshops (including TREC [3], NTCIR [6], and CLEF [13]) may provide answers to some of the methodological issues facing element retrieval. Parallels, for example, can be drawn between element retrieval and passage retrieval.

2.6 Other

By including an "other" category the workshop remained open to discussion of any additional issued not discussed above.

2.7 Exclusions

Ranking algorithms were specifically excluded from focus primarily to ensure the workshop would not act as a "half-INEX". That is., by allowing submissions on the topic of relevance ranking there was a perceived danger that the workshop would turn into an evaluation forum. The end of year INEX workshop fulfills this purpose admirably so accepting contributions on this topic would only blur the boundaries between the two workshops.

The existing software was excluded for two reasons. First, the mammoth efforts of those who build it should not go unnoticed, and attracting criticism of this effort was perceived as departmental to both the individuals involved and to the community as a whole. Second, the software should not dictate methodology, but should reflect methodology – as such the focus of the workshop was shifted from what the community currently does to what it should do.

3. CONTRIBUTIONS

A general call for papers was widely distributed. Interested parties were asked to contribute opinion papers for the purpose of promoting discussion. A total of eleven contributions were received, of which ten were accepted. Originally only four were to be accepted; however the papers were unexpectedly broad and workshop was reorganized to accommodate this.

3.1 Short Review of Submissions

Clarke [1] attacks individual elements as a suitable search engine result. He provides evidence that relevant information lies in sequences of tags (e.g. two consecutive paragraphs) and identifies a mismatch between returned results and relevant information. He suggests results should be returned as element ranges and provides a syntax for doing so. He suggests judgments should be done in the same manner and proposes using text-highlighting as a method of achieving this.

Hiemstra and Mihajlovic [4], apply the "simplest possible" approach to evaluation metrics and argue that precision-at-n elements reported with overlap scores provides a wealth of information for comparing two systems. They provide scores for several runs from INEX 2004 and explain how to read their scores and what, exactly, the scores mean.

Kamps *et al*. [5] examine what can (in principle) be expressed in a query language, then examine how users actually use such languages. From this they suggest formulating a set of topics with CO, CAS, and NLP expressions of the same information need (i.e. sharing a narrative). Judging against the narrative makes it possible to compare the performance of each of the queries and to directly compare each type of query. This will provide evidence of the superiority (or not) of using structural hints in a query.

Kazai and Lalmas [8] examine the requirements for an element retrieval precision metric. They classify each of the existing metrics against a list showing that they all fall short on some account.

Larsen *et al*. [9] identify the obtrusiveness of the relevance scale in user interaction experiments. By removing this imposition a true investigation of the element-centric searching behavior of users could be conducted. They provide several suggestions of non-obtrusive ways to examine user interaction.

Larson [10] focuses on heterogeneous searching. Identifying with the user, he notes that as the number of document collections increases, the cognitive load of the user increases. Whereas a user might have the ability to intimately know one DTD, there is little chance they will intimately know hundreds of DTDs. He identifies content and structural heterogeneous search as a possibly impossible user task. He suggests the issues might be addressed with reference to prior work in IR including embracing the principles of the Dublin Core.

Pehcevski *et al*. [12] examine the different judging behaviors between topic assessors and users (from the INEX interactive track). They identify patterns in judging behavior which demonstrate that the 10 point relevance scale is not well understood. They recommend changing the judgment scale.

Trotman [15] claims that the methodological issues in element retrieval stem from a lack of user grounding. If an application existed it could be examined and issues resolved with respect to the application. Identifying the IEEE collection as not suitable for element retrieval he calls for a shift to an audio or video collection, and metrics that do not reward element milking.

Woodley and Geva [18], frustrated at the judgment process, investigate ways to generate a more reliable set of judgments, while at the same requiring less work on the part of the judge. They provide evidence that to remain stable the judgment pool must be made from all retrieval runs and that the judgments must continue to be graded. However, they also identify out-of-pool judgments (those not in the pool, but forced by element context) as unnecessary. Secondly, they discuss ways to annotate the document collection. Finally they propose several possible future tracks.

Van Zwol *et al*. [17] suggest the complex structures of NEXI [16] are beyond the abilities of end users. They propose a visual query language called Bricks. This method of searching, they suggest, is more successful at completing the end user task than keyword search, while being faster (for the same purpose) then NEXI.

4. CONCLUSIONS

The INEX 2005 Workshop on Element Retrieval Methodology aims to provide a forum for discussion of element retrieval issues (other than relevance ranking).

Collected in this volume are papers on a broad set of issues ranging from user interaction through to performance metrics. These opinion papers were solicited with the aim of promoting discussion, and they no doubt will. The collection forms a discussion document for the workshop.

It is the combination of the discussion document and the face to face debate at the workshop that will enable progress on the many raised issues. When reading these papers, remember the object was to raise issues for discussion, not to solve the problems.

5. ACKNOWLEDGEMENTS

The organizers would like to thank the University of Glasgow for hosing the workshop. INEX is an activity of the DELOS network of excellence in digital libraries.

Without the discussions of the element retrieval community, including INEX and the various discussion lists, element retrieval would never have developed to where it is – it is the work of these others that makes the work of us possible.

It is always necessary to thank the program committee, the paper authors, and the participants, for without these people there would be no workshop.

6. REFERENCES

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

[2] Fuhr, N., Gövert, N., Kazai, G., & Lalmas, M. (2002). INEX: Initiative for the evaluation of XML retrieval. In *Proceedings of the ACM SIGIR 2000 Workshop on XML and Information Retrieval*.

[3] Harman, D. (1993). Overview of the first TREC conference. In *Proceedings of the 16th ACM SIGIR Conference on Information Retrieval*, (pp. 36-47).

[4] 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*.

[5] Kamps, J., Marx, M., Rijke, M. d., & Sigurbjörnsson, B. (2005). Understanding content-and-structure. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*.

[6] Kando, N. (2001). Overview of the second NTCIR workshop. In *Proceedings of the 2nd NTCIR Workshop*.

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

[8] Kazai, G., & Lalmas, M. (2005). Notes on what to measure in INEX. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*.

[9] 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*.

[10] 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*.

[11] Pehcevski, J., Thom, J. A., Tahaghoghi, S. M. M., & Vercoustre, A.-M. (2004). Hybrid XML retrieval revisited. In *Proceedings of the INEX 2004 Workshop*, (pp. 153-167).

[12] 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*.

[13] Peters, C. (2004). What happened in CLEF 2004? Introduction to the working notes. In *Proceedings of the CLEF 2004*.

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

[15] Trotman, A. (2005). Wanted: Element retrieval users. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*.

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

[17] 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*.

[18] Woodley, A., & Geva, S. (2005). Fine tuning INEX. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*.