Locating Relevant Text within XML Documents

Jaap Kamps[1,2] Marijn Koolen[1] Mounia Lalmas[3]

[1] Archives and Information Studies, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam
[3] Department of Computer Science, Queen Mary, University of London

ABSTRACT

Traditional document retrieval has shown to be a competitive approach in XML element retrieval, which is counter-intuitive since the element retrieval task requests all and only relevant document parts to be retrieved. This paper conducts a comparative analysis of document and element retrieval, highlights the relative strengths and weaknesses of both approaches, and explains the relative effectiveness of document retrieval approaches at element retrieval tasks.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries
General Terms: Measurement, Experimentation, Performance
Keywords: XML retrieval, focused retrieval, passage retrieval

1. INTRODUCTION

Focused retrieval, as it is practiced at the Initiative for the Evaluation of XML retrieval [2], requires not only to locate relevant documents but also to locate exactly the relevant information inside these documents. Hence, focused retrieval combines two different aspects: i) the retrieval of the relevant documents similar to traditional document retrieval, and ii) the retrieval of the relevant text within these documents. A focused retrieval approach need not distinguish these two steps and may target the relevant information directly irrespective of their document context. A typical example is a pure XML element-based approach, in which every possible element is treated as a "document." The other extreme is to ignore the second aspect by always returning the entire article—effectively backing-off to document retrieval. This latter strategy has repeatedly proven to be a non-trivial baseline for focused retrieval approaches [e.g., 3, 6]. In 2006, the Relevant In Context task was introduced at INEX, which combines both these aspects explicitly, requiring that the elements retrieved for a topic have to be grouped per document. In this paper, we look in detail at how document and element retrieval approaches fare at both aspects of focused retrieval: i) in terms of locating relevant documents, and ii) in terms of locating the relevant text within documents.

Throughout this paper we use two base runs: a *Document* run, based on index at the article level, and an *Element* run, where each XML element is treated as a document and indexed as such. We use a language model with default setting for smoothing and length prior. We grouped elements belonging to the same document, while removing overlapping elements, and ranked documents on the best scoring element. The runs are based on 221 INEX Ad hoc topics of 2006 and 2007 and the Wikipedia collection [1] consisting of over 650,000 documents and more than 52 million XML elements.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

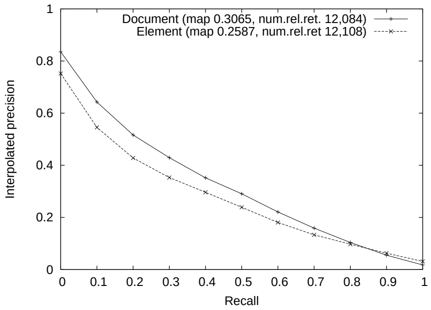

Figure 1: Precision/Recall of Document and Element runs

2. LOCATING RELEVANT DOCUMENTS

We investigate the first aspect of focused retrieval: how good are document and element retrieval approaches in finding and ranking relevant documents? We derive a document ranking of the *Element* run by selecting their document context on a first-come-first-served basis. We mapped the INEX Ad hoc assessments to TREC style Qrels, where each document containing at least some relevant text is considered a relevant document. Figure 1 plots the precision/recall curves for the *Document* and *Element* runs, from which it is clear that the document ranking of the *Document* run is superior. Up to 60% recall, the precision of the *Document* run is substantially higher. This is to be expected, as most of the elements in the element index are small and an off-topic element with only a few words and one or two keywords can get a high score. The difference in MAP between the *Document* run (0.3065) and the *Element* run (0.2587) is highly significant ($p < 0.001$, t-test, one-tailed). Although the element run could be improved by further tuning or using a mixture model [5], it should come as no surprise that document-retrieval approaches are effective for the first aspect of focused retrieval—document retrieval.

3. LOCATING RELEVANT TEXT WITHIN DOCUMENTS

We investigate the second aspect of focused retrieval: how good are document and element retrieval approaches in finding the relevant text *inside* those documents? We look at the ratio of relevant text in *relevant* and *retrieved relevant* documents (1,000 results per topic). In Figure 2, the distribution of relevant documents over the ratio of relevant text is given. The relevant documents are distributed over the ratio of relevant text: the first bin contains documents with 0-5% relevant text, the second bin 5-10%, etc. By far the most relevant documents are in the bin with less than 5% relevant text. There is a small increase again at the last bins, where almost all text is relevant. What is interesting to see is that the *Document* run has more relevant documents in most of the bins, but the

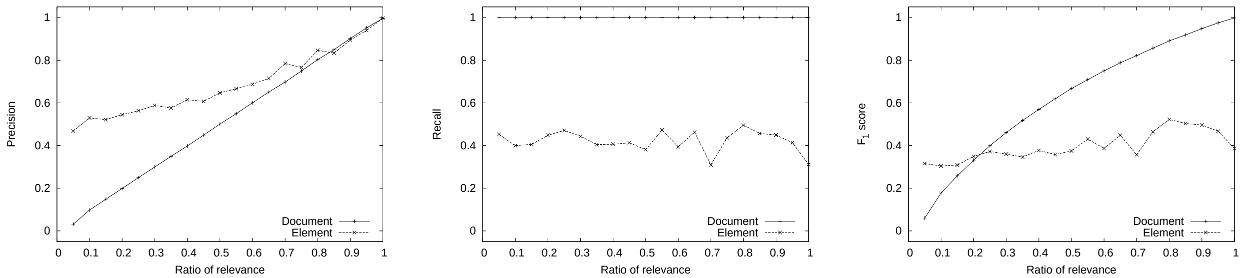

Figure 3: Precision (left), Recall (middle) F_1-score (right) over ratio of relevant text for Document and Element runs

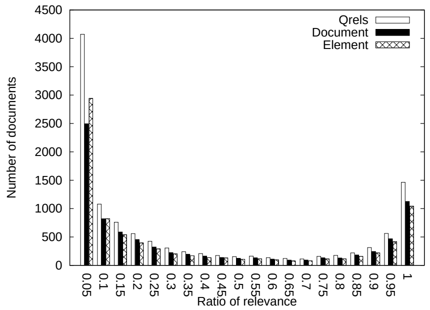

Figure 2: Distribution over ratio of relevant text

Table 1: Element level evaluation over INEX 2006+2007 topics

Index	Relevant in Context		Focused
	gP[5]	**MAgP**	**iP[0.01]**
Document index run	0.2293	0.1157	0.4706
Element index run	0.1996	0.0968	0.5368

Element run has more documents in the first bin.

We now look at precision (how much of the retrieved text is relevant) and recall (how much of the relevant text in the document is retrieved) over the ratio of relevant text. Figure 3 shows the per document Precision, Recall, and F_1-score over the ratio of relevant text in documents. The F_1 score combines precision and recall as $F_\alpha = \frac{(1+\alpha)\cdot(\text{precision}\cdot\text{recall})}{(\alpha\cdot\text{precision}+\text{recall})}$. As expected, the *Element* run has much higher precision at the lower ratios. At the high ratios, both runs have high precision scores. In terms of recall, the *Document* run gets the perfect score of 1 since the whole document is retrieved. The recall curve of the *Element* run shows that recall is not much affected by ratio of relevant text, as all bins have a recall of around 0.4. From the F_1-score curves it is clear that for documents where 25% or more of the text is relevant, the *Document* retrieval approach is superior, whereas for the many documents with low fractions of relevant text, the *Element* retrieval approach is more effective. The document retrieval approach is, obviously, a sensible approach if a large fraction of the document is relevant, and much less attractive if only a small part is relevant.

4. RELEVANT IN CONTEXT

The Relevant in Context task combines the two aspects of focused retrieval. The measure used for this task calculates a score per document based on how well the retrieved elements match the relevant text, and combines these per document scores using average generalized precision [4]. Since documents are elements, we can also evaluate both runs against the element level judgments. Table 1 shows the gP[5] and MAgP scores for the *Document* and *Element* runs. We see that after five documents, the generalized precision of the *Document* run is higher (not significantly) than of the *Element* run. Obviously both runs rank initially the documents with large fractions of relevant text, and here document retrieval makes sense. Also the overall mean average generalized precision of the document run is higher (again, not significantly) than of the standard run on the element index. If we regard the runs as "ranked lists of non-overlapping element" we can evaluate them using the measure of the Focused task, resulting in a higher interpolated precision at 1% recall for the element run (significant for $p < 0.05$), showing that the initial results of the element run are well on target. Again, the element run may be improved by further tuning or by bringing in the document context, but the main point is that document retrieval is competitive.

5. DISCUSSION AND CONCLUSIONS

The analysis above has shown how document retrieval is a competitive approach for focused retrieval. Its document ranking is superior to the element retrieval approach, possibly because it has to pick from fewer and on average bigger "documents," and with enough documents with a large fractions of relevant text it also gets high F-scores per document. Several ideas have been proposed to make document retrieval a less attractive option in focused retrieval evaluation. How to combine per-document recall and precision has a great impact, and this should reflect the underlying task well. The F_1 score equally weights precision and recall, and hence the document run is predominantly determined by recall. Arguably, focused retrieval tasks may intuitively require more emphasis on precision, e.g., by using an $F_{0.5}$ score weighting precision twice as much as recall. Given the relative differences in Figure 3, a more radical weighting may be needed to make document retrieval less attractive. It has also been suggested to use a measure that rewards documents proportional to their length or ratio of relevant text (for example by using a DCG measure over graded article scores). This will, however, further promote document retrieval approaches, and further reduce the impact of documents with little relevant text.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO, under project # 640.001.501).

REFERENCES

[1] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40:64–69, 2006.

[2] INEX. INitiative for the Evaluation of XML Retrieval, 2007. `http://inex.is.informatik.uni-duisburg.de/`.

[3] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. The importance of morphological normalization for XML retrieval. In *Proceedings of the First INEX Workshop*, pages 41–48. ERCIM, 2003.

[4] J. Kamps, M. Lalmas, and J. Pehcevski. Evaluating relevant in context: Document retrieval with a twist. In *Proceedings of SIGIR 2007*, pages 723–724. ACM Press, New York NY, USA, 2007.

[5] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. An Element-Based Approach to XML Retrieval. In *INEX 2003 Workshop Proceedings*, pages 19–26, 2004.

[6] J. A. Thom and J. Pehcevski. How well does best in context reflect ad hoc XML retrieval. In *Pre-Proceedings of INEX 2007*, pages 124–125, 2007.