How well does Best in Context reflect ad hoc XML retrieval?

James A. Thom[1] and Jovan Pehcevski[2]

[1] RMIT University, Melbourne, Australia
james.thom@rmit.edu.au
[2] MIT – Faculty of Information Technologies, Skopje, Macedonia
jovan.pehcevski@acm.org

Extended Abstract

This extended abstract describes the participation of the RMIT group in the Initiative for the Evaluation of XML retrieval (INEX) ad hoc track in 2007. Of the three tasks in the INEX 2007 XML ad hoc track: Focused, Relevant in Context (RiC), Best in Context (BiC), the RMIT system performed surprisingly well on the last task.

Our Approach

Our approach is limited to retrieval of articles using the Zettair[3] search engine. Zettair is an open source search engine developed at RMIT, which we used to index the full text of Wikipedia articles and return complete articles ranked by their similarity score to the query. Zettair is "one of the most complete engines" according to a recent comparison of open source search engines [3]. Within Zettair we used the Okapi BM25 similarity measure which worked well on the INEX 2006 Wikipedia test collection [1].

For each of the Focused, RiC, and BiC tasks, we simply return the same ranked list of whole documents. Thus these Zettair runs can be seen as a baseline against which element or passage retrieval would be expected to do better.

Results

We present our results that investigate the effectiveness of document retrieval when applied to the three tasks in the INEX 2007 ad hoc track.

For the Focused retrieval task the RMIT system had an interpolated average precision at 0.01 recall of 0.3788 (compared with 0.4259 for the best performing system on this task) and was ranked 17 out of the 79 runs.

For the RiC task the RMIT system had a non-interpolated mean average precision (MAgP) of 0.0884 (compared with 0.1013 for the best performing system on this task) and was ranked 10 out of 66 runs.

For the BiC task the RMIT system had a non-interpolated mean average precision (MAgP) of 0.1951 and was surprisingly the top ranked run (out of 71 runs) for this task.

[3] http://www.seg.rmit.edu.au/zettair/

Discussion

Looking at the results (as compared with other systems), document retrieval (using Zettair) seems to work well on the INEX Wikipedia XML collection. Only relatively small gains are made by the best systems using element or passage retrieval for the Focused and the RiC tasks. For the BiC task, it seems difficult to do better than returning the start of the document as the best entry point.

Why is this the case? Firstly, from the definition of the BiC task we are looking for retrieving relevant documents in the first place. Obviously, Zettair does a good job here (but we already know this from our INEX 2006 ad hoc experiments). Secondly, after locating a relevant document, the task asks systems to find the best entry point (BEP) to start reading the document. In their analysis of the INEX 2006 relevance assessments, Kamps et al. [2] observed that assessors would mainly choose the best entry point to be "some distance" from the start of the document; specifically, they observed the following:

> "What we see is that the BEP is a fair distance into the article (median distance 556 [characters], mean distance 3,090 [characters]). The difference between median and mean distance signals that the distribution is skewed toward the start of the article. Comparing the BEP distance and the length of the article, we find a significant correlation of 0.66."

Judging from the way Zettair performed, we suspect that this skew towards the start of articles is at least as great in the case of INEX 2007 relevance assessments as it was in the case of INEX 2006 relevance assessments. As we retrieve only articles with Zettair, it is therefore of no great surprise that we perform better than any of the other element or passage retrieval systems.

Acknowledgements

Most of this work was completed while James Thom was visiting INRIA and Jovan Pehcevski was working at INRIA.

References

1. D. Awang Iskandar, J. Pehcevski, J. A. Thom, and S. M. M. Tahaghoghi. Social media retrieval using image features and structured text. In *Comparative Evaluation of XML Information Retrieval Systems: Fifth Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2006*, volume 4518 of *Lecture Notes in Computer Science*, pages 358–372, 2007.
2. J. Kamps, M. Koolen, and M. Lalmas. Where to start reading a textual xml document? In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 723–724, New York, NY, USA, 2007. ACM.
3. C. Middleton and R. Baeza-Yates. A comparison of open source search engines. Technical report, Universitat Pompeu Fabra, Barcelona, Spain, 2007. http://wrg.upf.edu/WRG/dctos/Middleton-Baeza.pdf.