Sub-document level Information Retrieval: Retrieval and Evaluation

Sukomal Pal
Department of Computer Science & Engineering
Indian School of Mines, Dhanbad, India
sukomalpal@gmail.com

September 10, 2012

Abstract

XML is increasingly used to mark up content in present day information repositories. Over the last decade or so, retrieval from XML document collections has emerged as an area of active research. For the Information Retrieval community, XML retrieval poses a two-fold problem:

1. finding effective techniques to retrieve appropriate or the most useful XML elements in response to a user query; and
2. devising an appropriate evaluation methodology to measure the effectivity of such retrieval techniques.

This study examines both these issues. First, we revisited the pivoted length normalization scheme in the Vector Space Model using standard benchmark collections for XML retrieval. We reduced two parameters used in pivoted length normalization to a single combined parameter and experimentally found its optimum value, which works well at both the element and document levels for XML retrieval.

We observed that a substantial number of focused queries used in XML retrieval clearly state, besides the information need, what the user *does not* want. We demonstrated that this negative information, if not handled properly, degrades retrieval performance. We proposed a solution for automatically removing negative information from XML queries. This led to significant improvements in retrieval results.

On the evaluation of XML retrieval, we first studied the sensitivity & robustness of various evaluation metrics and reliability and reusability of the assessment pool that has been used at INEX Ad Hoc track since 2007. Specifically we investigated the behaviour of the metrics when assessments are incomplete, or when query sets are small. We observed that early precision metrics are more error-prone and less stable under both these conditions. Average metric, however, performs comparatively better in this respect. System rankings remain largely unaffected even when assessment effort is substantially (but systematically) reduced. We also found that the INEX collections remain usable when evaluating non-participating systems.

For a fixed amount of assessment effort, judging shallow pools for many queries is found to be better than judging deep pools for a smaller set of queries. However, when judging only a random sample of a pool, it is better to completely judge fewer topics than to partially judge many topics.

We also proposed a simple and pragmatic approach of creating assessment pool for evaluation of retrieval systems. Instead of using an apriori-fixed pool depth for all topics, the pool is incrementally built and judged interactively on a per query basis. When no new relevant document is found for a reasonably long run of pool-depths, pooling was stopped for the topic. Our proposed approach offers a trade-off between the available effort and required level of performance. Moreover, it is flexible to *deep pooling* for potential topics in order to ensure *better* estimate of recall. We demonstrated the effectivity of the technique by substantially reducing the assessment effort without seriously compromising on the reliability of evaluation. The approach provided good results in the evaluation of XML retrieval as well as traditional document retrieval.

This doctoral work was done at and submitted to the Indian Statistical Institute, Kolkata, under the supervision of Dr. Mandar Mitra. Dr. Ellen Voorhees and Dr. Tetsuya Sakai served as reviewers; Dr. Soumen Chakrabarti was the examiner at the defence of the thesis. Available online at `http://www.isical.ac.in/~mandar/sukomal-thesis.pdf`