Length Normalization in XML Retrieval
(Extended abstract)

Jaap Kamps Maarten de Rijke Börkur Sigurbjörnsson

Informatics Institute, University of Amsterdam
kamps,mdr,borkur@science.uva.nl

Abstract

The full paper appeared as: J. Kamps, M. de Rijke, and B. Sigurbjörnsson, "Length Normalization in XML Retrieval," In: *Proceedings 27th Annual International ACM SIGIR Conference (SIGIR 2004)*, pages 80-87, 2004.

Introduction. The importance of document length normalization is a recurring theme in information retrieval (IR). The advent of the Text Retrieval Conferences (TREC) in 1992 introduced large-scale test collections with full-text documents. Documents in these collections were much longer than documents in collections previously used for evaluation purposes (mostly based on abstracts), and had more length variety. In particular, full-text retrieval necessitated a revision of document length normalization. The introduction of XML retrieval marks a similar revolution in IR. Although a text collection of XML documents may have a similar number of articles as standard TREC-sized collections, the number of XML elements in the collection takes us to quite a different scale. There are millions of XML elements that can potentially be retrieved as an answer to a query, having a great variety in length (ranging from single words or phrases put in italics or in titles, to full-blown articles). XML retrieval prompts us to revisit the issue of length normalization.

XML retrieval. In XML *element* retrieval, each of the text elements into which XML documents are divided, is an object that can in principle be returned in response to a query. The INitiative for the Evaluation of XML retrieval (INEX) was launched in 2002 to assess the effectiveness of retrieval methods for XML document and element retrieval. We focus on so-called *content-only* (CO) topics, which are traditional IR topics written in natural language. Length-wise there are several noteworthy aspects of the INEX test collection. First, the collection has over 12,000 articles, but nearly 7,000,000 XML elements. Second, the XML element length distribution is much more skewed than normal document length distributions. Third, in XML element retrieval the assessors have a strong bias toward retrieval of long elements. By accounting for these length aspects of XML elements during retrieval, systems can improve performance.

Aims. Main components that affect the importance of a term in a text are the term frequency, the inverse document frequency, and document length. In the generative language modeling approach that we adopt in this paper, these three aspects are captured by the model(s), smoothing procedures, and priors. Our overall motivation is to identify effective XML retrieval methods that are highly portable across XML collections in the sense that they only exploit statistical aspects (both content and non-content) of XML documents, and do not depend on specific schemas or tag sets. Specifically, in this paper we aim to understand how *priors* and *smoothing* affect XML element retrieval performance.

Priors. For the *priors* aspect, we need to bridge the gap between the average element length and average *relevant* element length. Since we want to balance the "pinpointing" nature of the XML element retrieval task with the (apparent) importance of long elements, we want to do something more intelligent than only returning the longest possible elements (i.e., articles) in the collection. Priors allow one to import "non-content" features of documents (or elements) into the scoring mechanism. Document length is a good example of information about a document that is not directly related to its contents, but might still be related to the possible relevance of the document. For ad hoc document retrieval, there is a correlation between document length and a priori probability of relevance.

Smoothing. Our other main issue in this paper is *smoothing* for XML element retrieval. Since document (and element) language models may suffer from inaccuracy due to data sparseness, a core issue in language modeling is *smoothing*, which refers to adjusting the maximum likelihood estimator for the document (or element) language model by combining it with a background language model. Two things are at stake: first, since element scores are constructed from very short amounts of text, improving the probability estimates is very important. Second, smoothing facilitates the generation of common terms (a $tf \cdot idf$ like function). Smoothing plays a special role in XML retrieval: With smoothing, short elements containing only one or a few of the query terms will receive a high relevance score.

Main findings. We perform a comparative analysis of the length of arbitrary elements versus that of relevant elements, and highlight the importance of length as a parameter for XML retrieval. Within the language modeling framework, we investigate techniques that deal with length either directly or indirectly: length priors, index cut-off, and the amount of smoothing. We observe a length bias introduced by the amount of smoothing, and show the importance of extreme length priors for XML retrieval. When used with extreme length priors, the smoothing parameter regains its normal function of controlling term importance. Furthermore, we show that simply removing shorter elements from the index (by introducing a cut-off value) does not create an appropriate document length normalization. After restricting the minimal size of XML elements occurring in the index, the importance of an extreme length bias remains. The combination of length priors with index cut-off does lead to a slight further improvement.

The value of the approach has been demonstrated by the top ranking results of a system implementing the approach at the INEX 2003 workshop.