Using Relevance Feedback in XML Retrieval

Roger Weber

9.1 Introduction

Information retrieval has a long tradition: in the early days, the main focus was on the retrieval of plain text documents and on search systems for books and structured documents in (digital) libraries. Often, users were assisted by well-trained librarians or specialists to retrieve documents fitting their information need. With the proliferation of the internet, retrieval systems for further media types like images, video, audio and semi-structured documents have emerged. But more importantly, an ever increasing number of untrained users deploy retrieval systems to seek for information. Since most users lack a profound understanding of how retrieval engines work and of how to properly describe an information need, the retrieval quality is often not satisfactory due to bad query formulations. As an illustration of this, Jansen et al. [177] reported that 62% of queries submitted to the Excite web search engine consisted of less than three query terms. Obviously, this is by far insufficient to accurately describe an information need. But search systems often do not support users (or only rudimentary) to adjust their queries to improve retrieval effectiveness.

As a countermeasure for the query refinement problem, relevance feedback was introduced in the late 1960’s [169, 251]. The basic idea is to model the search as an iterative and interactive process (cf. Figure 9.1) during which the system assists users with the task of query refinement. To this end, the user has to assign relevance values to the retrieved documents. This feedback together with the original query is processed according to a feedback model and yields a new query which, hopefully, returns new and more relevant documents. This iteration can continue until the user is satisfied or the retrieval process is aborted. The feedback process bears a number of design options: 1) capturing of feedback (implicit vs. explicit, granularity, feedback values), 2) reformulation of a query given the feedback (feedback model), and 3) provision of methods for users to accept/reject parts of the refined query.

In this chapter, we focus on relevance feedback techniques for XML-retrieval. In this context, we describe and deploy a retrieval model that
comprises both structural as well as content related query parts following the principles of augmentation [127, 128] and query specific retrieval statistics (cf. Chapter 6 in this book). We describe how relevance feedback methods may refine such queries: for that purpose, we have identified five dimensions for query adaptation, namely: 1) query expansion, i.e., the addition and removal of query terms, 2) query term weighting, 3) recalculation of the discrimination power of terms, 4) structural weighting, and 5) selection of the retrieval model or parts of the model. Along each of these dimensions, we describe feedback methods adopted from approaches for text retrieval. We further summarize some of the cognitive results from earlier work [191] with respect to the design of feedback interaction. Finally, this chapter does not contain an evaluation of the different feedback techniques due to the lack of an appropriate benchmark with relevance assessments on the retrieved documents. However, in light of the recent development of INEX [125], we plan to implement, to tune, and to evaluate the different feedback models with this benchmark.

This chapter is organized as follows: Section 9.2 describes a generic retrieval model for XML documents. Section 9.3 adapts previous work on relevance feedback to the context of XML retrieval. Section 9.4 concludes this chapter.

9.2 XML Retrieval Model

Relevance feedback techniques are always tightly bound to the underlying retrieval model. In the following, we shortly summarize an XML retrieval model which provides flexible means to query XML documents by structure.