Abstract. This paper proposes a method of ranking XML documents with respect to an Information Retrieval query by means of fuzzy logic. The proposed method allows imprecise queries to be evaluated against an XML document collection and it provides a model of ranking XML documents. In addition the proposed method enables sophisticated ranking of documents by employing proximity measures and the concept of editing (Levenshtein) distance between terms or XML paths.

1 Introduction

Information Retrieval (IR) techniques have traditionally been applied to search large sets of textual data. The emerge of XML as a standard for data representation and exchange on the Internet poses new challenges to structured document retrieval. Integrating IR and XML search techniques could enable more sophisticated search on the structure as well as the content of the documents. Some of the recent XML-IR proposals [6, 8, 11, 4, 2] focus on indexing for improving the execution of simple IR-like queries. Other approaches [9, 15, 25] are based on soft computing techniques to model uncertainty. This paper is closer to the second approach and it uses fuzzy logic to rank the documents of an XML collection with respect an IR query. Ranking of structured documents is very important and may further boost the quality of the results. Ranking structured documents is mainly based on probabilistic models [24,21,11,26] by mainly using the idf (inverse document frequency) and tf, within document frequency of terms.

Searching effectively large collections of XML documents requires a knowledge of the documents structure. For instance, using XQuery [5], it requires some knowledge on the XML schema (or DTD) [1,3]. On the other hand searching XML documents from the Information Retrieval (IR) point of view requires little knowledge on the structure of the documents. Although knowledge on the XML schema could be desirable to reduce extensive search, it is not required. Another important aspect in retrieving XML documents is the relevance of the retrieved documents to the submitted query. In most cases, we are not interested in finding XML documents that exactly match the query but rather in XML documents, which are relevant to the query to some extent. Our tolerance to accept in the result set XML documents that do not precisely match the query is stemmed from the fact that the query itself is not precise. IR based queries cannot be precise mainly due to the lack of knowledge of the underlying document.
collection structure and the inherent difficulty to formulate a query that precisely reflects what we are looking for. On the other hand, the relevance is by nature a fuzzy concept. So it seems that the employment of fuzzy logic for estimating the relevance of the documents to a given query is more expressive and promising.

Modeling vagueness in information retrieval has been addressed by several researchers [14, 15, 18, 25, 9] and recently in the area of XML documents through the INEX workshops [28, 7]. Weighted boolean models [14] have been proposed to handle constraints imposed by the query terms. Fuzzy set theory has been proposed [15, 16, 18, 25] for modeling flexible information retrieval systems. Models for integrating IR and database systems is presented in [12]. In the area of structured documents [25] fuzzy aggregations have been employed to facilitate retrieval. A survey on information retrieval techniques based on soft computing (mainly fuzzy sets and neural networks) is presented in [18]. In [9] the logical structure among objects have been employed to model structured documents in which the logical structure among objects is captured by means of knowledge augmentation. While in classical retrieval the quality is estimated by means of idf (inverse document frequency) and tf (within document term frequency), [27] proposes a third dimension called accessibility that captures the structure of the document. Vague queries and inexact data matching in databases is proposed in [22] and fuzzy Hamming distance that extends the Hamming concept for measuring dissimilarity between vector objects is proposed in [17].

This paper proposes a fuzzy logic based technique for ranking XML documents against a given query. The employment of fuzzy logic enables more sophisticated ranking by exploiting the imprecise formulation of the queries. To achieve this, a database is used for indexing purposes and for storing all the facts that constitute the XML document collection. Fuzzy measurements based on editing distance are used to handle uncertainty and imprecision. The rest of the paper is organized as follows: Section 2 discusses the IR query format. Fuzzy ranking is discussed in section 3. The realization of the database structure used to facilitate the ranking is presented in section 4. The fuzzy relevance of a document against to a sub-query is discussed in the section 5. Some implementation notes are presented in the section 6 and the section 7 summarizes the contributions and concludes the paper.

2 Query Format

XML Querying mechanisms that facilitate Information Retrieval (IR) should allow the formulation of simple queries that require no knowledge of the underlying structure of the XML document collection. However, the query method should be also flexible enough to allow more complex queries that take into consideration the structure of the underlying documents. In addition the query mechanism should also allow the user to focus the search on particular XML elements that might be more interesting. So, element prioritization that reflects the importance of the element in context and in relation with other elements is equally important.