An Approach to Define Flexible Structural Constraints in XQuery

Emanuele Panzeri and Gabriella Pasi
University of Milano-Bicocca
Viale Sarca 336, 20126 Milano, Italy
{panzeri,pasi}@disco.unimib.it

Abstract. This paper presents a formal definition of an extension of the XQuery Full-Text language: the proposed extension consists in adding two new flexible axes, named below and near, which express structural constraints that can be specified by the user. Both constraints are evaluated in an approximate way with respect to a considered path, and their evaluation produces a path relevance score for each retrieved element. The formal syntax and the semantics of the two new axis are presented and discussed.

1 Introduction

The increasing number of huge collections of highly structured XML documents has stimulated in recent years a wealth of research aimed at improving XML querying to both increase query languages expressiveness, and to provide an approximate matching of queries with the consequent ranking of the retrieved elements [3,18]. The first XML query languages were designed based on a data-centric view of XML repositories to allow an efficient access to complex data structures; these languages were finalized to the specification of structural constraints as well as content-related constraints (specification of exact values for XML element contents) in a Data Base style: the results produced by such constraints evaluation is a set of relevant elements. Later, several proposals appeared based on a document-centric view of XML repositories; such approaches have been classified by the information retrieval (IR) community as content-only search (CO) and content and structure search (CAS) [14]. CO approaches were mainly aimed at allowing the specification of keyword based queries in an IR style, where query evaluation produces a ranking of the retrieved XML elements [5]. CAS approaches were defined to allow the formulation of constraints on both documents content and structure [14]. CAS approaches that were based on the syntax of XPath [15], constituted a first attempt to merge the IR and DB search paradigms. Since then, the importance of merging the IR and the DB search perspectives has been widely recognized, and it has recently culminated in the W3C standard XQuery Full-Text (XQ-FT) [17] extension. The evaluation of XQ-FT queries produces a set of weighted elements, where scoring is based on a keyword based matching in textual elements. The problem of providing a ranking of XML
elements retrieved by a query based on both content and structural constraints has been addressed in [3], where a query relaxation technique that provides an approximate structural matching was introduced.

None of the above approaches allows users to directly specify the structural relaxations via ad-hoc predicates with a score computation. More recently, in [6], an approach to structural relaxation in XPath via new user specified constraints is proposed; a RDBMS is extended to evaluate relaxed structure matching. However the query evaluation does not provide any ranking of the retrieved fragments. As outlined in [18] and [6], querying highly structured databases or document repositories via structured query models (as XQuery is) forces the users to be well aware of the underlying structure, which is not trivial. In the above cases, users could benefit of a query language that allows a direct specification of flexible structural constraints that easily allow to require the relative position of important nodes, independently of an exact knowledge of the underlying structure(s). To achieve this aim, in this paper, we propose a formal extension of XQuery Full-Text, where two new flexible structural axes, specified by the predicates below and near are defined. The work reported in this paper was originated by a previous research where a flexible extension of the XQuery language was advocated and informally sketched in [9,7]. This is the first work that proposes the full syntax and semantics of the formal extension.

The proposed extension allows to obtain: (1) a ranking based on content predicates evaluation only (as in the original XQ-FT), (2) a ranking based on the flexible structural constraints evaluation (based on our proposal), or (3) a ranking based on a linear combination of the two above scores, which the user may also specify via the order-by clause, as it will be explained in the paper.

In summary, the main contributions of this paper are: (1) to define a formal extension of XQ-FT with two new flexible axes, thus allowing users to explicitly specify their tolerance to an approximate structural matching, while not forcing them to be aware of all the possible structural variations of the data/document structure; (2) to define an ad-hoc approximate matching of the flexible structural constraints thus allowing both a ranking only based on approximate structure matching, and a ranking based on a combination of content predicates and the new flexible structural predicates (while preserving a ranking based only on content predicates).

The paper is organized as follows: Section 2 reviews the research work related to introducing flexibility in XML query evaluation. Section 3 presents the proposed extension of the XQuery Full-Text language with the new flexible structural constraints: both the syntax and the semantics of the new constraints are formalized as well as some usage examples. Section 4 concludes the paper.

2 Related Work

As outlined in Section 1, several approaches to introduce some flexibility in XML retrieval have been proposed in last years, by both the database and the IR communities [8,10,12,16]. In IR, the approaches to inquiry XML documents