Chapter 7.4
A Deterministic Approach to XML Query Processing with Efficient Support for Pure and Negated Containments

Dunren Che
Illinois University at Carbondale, USA

ABSTRACT

This article reports the result of the author's recent work on XML query processing/optimization, which is a very important issue in XML data management. In this work, in order to more effectively and efficiently handle XML queries involving pure and/or negated containments, a previously proposed deterministic optimization approach is largely adapted. This approach resorts to heuristic-based deterministic transformations on algebraic query expressions in order to achieve the best possible optimization efficiency. Specialized transformation rules are thus developed, and efficient implementation algorithms for pure and negated containments are presented as well. Experimental study confirms the validity and effectiveness of the presented approach and algorithms in processing of XML queries involving pure and/or negated containments.

INTRODUCTION

XML has become the de facto standard for information, data exchange, and representation on the Internet and elsewhere. As a result, more and more data sources have been adopting the XML standard. The rapid accumulation of XML data calls for specialized solutions for managing and querying XML data resources. Among the many challenges related to the XML database management technology, XML query optimization is very interesting because it not only is a critical issue for XML DBMS but it also provides a key infrastructure for the future semantic Web and applications (especially, the semantic-based Web search engines).

Query optimization typically applies the cost-based approach (Selinger, Astrahan, Chamberlin, Lorie, & Price, 1979) and aims at obtaining the least expensive—the optimal—evaluation plan for each input query. Heuristic knowledge may be exploited in addition so that a reduced number of (only highly potential) candidate plans need to be examined and from which the best plan is to be identified.

XML data has the semi-structured nature and XML queries needed to check not only the *contents* but also the *structural patterns* of the source XML data. Comparing to relational data, XML data has higher complexity (as it has to additionally deal with the structural part), and this complexity trivially translates to an enlarged search space for the "optimal" plan during query optimization (assuming the cost-based approach is adopted). Consequently, a plain application of the cost-based optimization approach does not usually yield the same good efficiency for XML queries as it does for relational queries. On the other hand, apart from adding extra complexity and causing inefficient (cost-based) query optimization, the structural part of XML data implies a rich resource of knowledge that can be used in favor of efficient optimization of XML queries. We are thus motivated to develop a comprehensive optimization framework for XML queries. This framework consists of two separate yet collaborative optimization stages. The first stage performs logical-level optimization. This stage is unique to XML as it explores the specific features (e.g., semantic knowledge) of XML data for query optimization. By its nature, this stage is strongly heuristic-based because it does not rely on any particular knowledge about the underlying storage structure. The second stage—physical optimization—typically applies specialized cost-based optimization techniques. In such an optimization framework, these two stages need to collaborate in a way that the first stage provides a reduced (or pre-screened) set of logical plans to the second stage and the latter shall conduct specialized techniques for cost-based optimization by adequately considering the optimization that has already been exerted at the first stage.

In previous research (Che, 2003, 2005, 2006), we studied the query equivalence issue in the context of XML, which forms the basis of our transformation-based optimization approach to XML queries. A comprehensive methodology for fast XML query optimization at the logic level was proposed (Che, Aberer, & Özsu, 2006) based on exclusive application of deterministic transformations on XML queries represented as PAT algebraic expressions (Böhm, Aberer, Özsu, & Gayer, 1998; Salminen & Tompa, 1994). The utmost benefit of this unique approach is the great potential for superb optimization efficiency. More recently, we substantially extended the PAT algebra, which leads to ePAT (extended PAT), endowed with more expressive power. Based on ePAT, we redefined the query equivalences and the deterministic transformation rules, and adapted the prior optimization strategy in order to efficiently support XML queries with pure and/or negated containments. Containment is a core operation in XML queries, and negated containment is as important as the regular containment for XML queries (for example, "find all employees who do *not* have a homepage," though simple, can be a common (sub-)query pattern). However, little result on efficient support for pure and negated containments has been reported.

In this article, we make the following important contributions:

- An adapted deterministic optimization approach for XML queries with pure and/or negated containments is presented.
- A group of specialized join algorithms dedicated for pure and negated containment operations are proposed (realizing the known structural join algorithms (Srivastava et al., 2002; Zhang, Naughton, DeWitt, Luo, & Lohman, 2001) cannot provide efficient support for these special containment operations).
- An experimental study is conducted, and the obtained result confirms the validity and effectiveness of this new approach and the specialized supporting algorithms.