Approximate Querying of XML Fuzzy Data

Patrice Buche[1], Juliette Dibie-Barthélemy[1,2], and Fanny Wattez[1,2]

[1] Mét@risk MIA INRA
16, rue Claude Bernard,
F-75231 Paris Cedex 05
[2] UMR MIA INA P-G/INRA
16, rue Claude Bernard,
F-75231 Paris Cedex 05
{Patrice.Buche, Juliette.Dibie, Fanny.Wattez}@inapg.inra.fr

Abstract. The MIEL++ system integrates data expressed in two different formalisms: a relational database and an XML database. The XML database is filled with data semi-automatically retrieved from the Web, which have been semantically enriched according to the ontology used in the relational database. These data may be imprecise and represented as possibility distributions. The MIEL++ querying system scans the two databases simultaneously in a transparent way for the end-user. To scan the XML database, the MIEL query is translated into an XML tree query. In this paper, we propose to introduce flexibility into the query processing of the XML database, in order to take into account the imperfections due to the semantic enrichment of its data. This flexibility relies on fuzzy queries and query rewriting which consists in generating a set of approximate queries from an original query using three transformation techniques: deletion, renaming and insertion of query nodes.

1 Introduction

Numerous approaches have been proposed in the bibliography to introduce flexibility in the comparison between an XML tree query and XML data trees. The first one is based on the encoding of the XML data trees [6]. This method only permits the introduction of intermediate nodes in the tree query structure in order to carry out the comparison with the data trees. The second approach [1] is based on the rewriting of the XML tree query. It permits the introduction, renaming and deletion of nodes in the query. The third one [12] is a combination of the previous two: the data are encoded and the query is rewritten. It provides an accurate computation of the transformation cost between the query and the data. But, it is very difficult to use because it requires one to redefine the management of the index and data encoding in the XML Database Management System. In a fourth approach [2], fuzzy predicates are introduced into the query to express flexible selection conditions and to perform fuzzy subtree matching. But it does not take into account suppression and renaming of nodes. In this paper, we propose a new XML querying system. It combines the flexibility provided by the

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 26–38, 2006.

use of fuzzy sets to represent the user's preferences in an XML query and the flexibility of XML query rewriting (including insertion, deletion and renaming of nodes) to perform an approximate comparison between an XML tree query and an XML data tree. Moreover, it supports XML imprecise data expressed as possibility distributions which is also an original contribution because few research has been done in modeling and querying imprecise XML data. To the best of our knowledge (see [10] for a recent synthesis), only imprecise data and probabilistic data modeling in XML have been proposed. Our querying system is fully compatible with XML querying standards since the final rewriting of the queries is performed in XQuery language (http://www.w3.org/XML/Query/).

This work is realised in the framework of a system development whose aim is to integrate heterogeneous data sources. Two approaches are generally considered to solve this problem: the data warehouse approach [14] in which data are transformed to be stored in one global schema and mediated architectures [15] where the data remain stored in the original sources, the mapping between the global integration schema and the schemas of original sources being carried out by wrappers. In our system, we propose data integration based on a *mediated architecture*. More precisely, we use a global schema to integrate data expressed in two different formalisms: a relational database and an XML database. This architecture, called MIEL++ [4], is close to a *Global as Views* approach, in which the global schema is defined in terms of the local schemas to be integrated, as in the TSIMMIS [13] system. An original aspect of our approach is that our XML database is comparable to a data warehouse, since it contains data, semi-automatically retrieved from the Web, which have been modified in order to be expressed in the same vocabulary and semantic relations as the ones used in the relational database [9]. These data, called SML data, may be imprecise [3] and represented as possibility distributions [17]. Moreover, in order to avoid empty answers, MIEL++ querying system proposes to the end-user to express selection criteria by means of fuzzy sets used as expression of preferences [3]. In [5], we have defined a wrapper which translates a MIEL query into an XML tree query to scan the XML database. We made the assumption that the XML data trees retrieved by the MIEL++ system have to fit exactly the structure of the XML tree query. This assumption does not permit the imperfections of the SML data to be taken into account. Therefore, we cannot make the assumption as in [11] that we know precisely the schema of the Web data sources we want to integrate. In this paper, we study the way of introducing more flexibility into the MIEL query processing of an XML database. This work is done in the framework of the development of a real data warehouse in an actual application domain: risk assessment in food safety.

In section 2, we briefly present firstly the fuzzy set framework that we use to represent imprecise data and preferences in the queries, secondly the MIEL query language and thirdly the way imprecise data are represented in an XML database. In section 3, we define a new wrapper which translates a MIEL query into a set of approximate XML queries. In section 4, we present the implementation and preliminary test results of this wrapper in a real application.