Chapter 9

Pattern-Based Schema Mapping and Query Answering in Peer-to-Peer XML Data Integration System

Tadeusz Pankowski
Poznan University of Technology, Poland

ABSTRACT

This chapter addresses the problem of data integration in a P2P environment, where each peer stores schema of its local data, mappings between the schemas, and some schema constraints. The goal of the integration is to answer queries formulated against a chosen peer. The answer must consist of data stored in the queried peer as well as data of its direct and indirect partners. The chapter focuses on defining and using mappings, schema constraints, query propagation across the P2P system, and query answering in such scenario. Schemas, mappings, constraints (functional dependencies) and queries are all expressed using a unified approach based on tree-pattern formulas. The chapter discusses how functional dependencies can be exploited to increase information content of answers (by discovering missing values) and to control merging operations and propagation strategies. The chapter proposes algorithms for translating high-level specifications of mappings and queries into XQuery programs, and it shows how the discussed method has been implemented in SixP2P (or 6P2P) system.

DOI: 10.4018/978-1-60960-475-2.ch009
INTRODUCTION

The goal of data integration is to enable rapid development of new applications requiring information from multiple sources (Haas, 2007). Data integration consists in combining data from different sources into a unified format (Bernstein & Haas, 2008). There is a number of different research fields relevant to data integration. Among them we can distinguish: identification of the best data sources to use, cleansing and standardizing data coming from these sources, dealing with uncertainty and tracing data provenance, the way of querying diverse sources and optimizing queries and execution plans. Integration activities cover any form of data reuse, such as exchanging data between different application’s databases, translating data for business-to-business e-commerce, and providing access to structured data and documents via a Web portal (Bernstein & Haas, 2008).

A variety of architectural approaches can be used to deal with the problem of data integration. The most popular is the materialized integration realized by means of data warehouse that consolidates data from multiple sources. Other approaches use paradigm of virtual integration. While warehouses materialize the integrated data, virtual data integration offers a mediated schema against which users can pose queries. The query is translated into queries on the data sources and results of those queries are merged so that it appears to have come from a single integrated database (Miller et al., 2000, Pankowski & Hunt, 2005). In a peer-to-peer (P2P) data integration the role of the mediated schema can play schema of any peer database. Then the user issues a query against an arbitrarily chosen peer and expects that the answer will include relevant data stored in all P2P connected data sources. The data sources are related by means of XML schema mappings. A query must be propagated to all peers in the system along semantic paths of mappings and reformulated accordingly. The partial answers must be merged and sent back to the user’s peer (Madhavan & Halevy, 2003; Pankowski, 2008c; Tatarinov & Halevy, 2004).

Much work has been done on data integration systems both with a mediated (global) schema and in P2P architecture, where the schema of any peer can play the role of the mediated schema (Arenas & Libkin, 2005; Madhavan & Halevy, 2003; Melnik et al., 2005, Yu & Popa, 2004). There is also a number of systems built in P2P data integration paradigm (Koloniari & Pitoura, 2005), notably Piazza (Tatarinov et al., 2003), PeerDB (Ooi et al., 2003). In these works the focus was on overcoming syntactic heterogeneity and schema mappings were used to specify how data structured under one schema (the source schema) can be transformed into data structured under another schema (the target schema) (Fagin et al., 2004; Fuxman et al., 2006). Some attention has been paid to the question of how schema constraints influence the query propagation.

This chapter describes formal foundations and some algorithms used for XML data integration in P2P system. Schemas of XML data are described by means of a class of tree-pattern formulas, like in (Arenas & Libkin, 2005). These formulas are used to define both schema mappings and queries. In contrast to (Arenas & Libkin, 2005), except for schemas we use tree-pattern formulas also to specify constraints (functional dependencies) over schemas. Schemas, mappings, queries and constraints are specified in a uniform way as a class of tree-pattern formulas. Thanks to this, we are able to translate high-level specifications into XQuery programs. We also discuss the problem of query propagation between peers. We show how mutual relationships between schema constraints and queries can influence both propagation of queries and merging of answers. Taking into account such interrelationships may improve both, efficiency of the system and information content included in answers. We show in brief how the issues under consideration have been implemented in 6P2P system (SixP2P, Semantic Integration of XML data in P2P environment).