Chapter 8.6
Integrating Heterogeneous Data Sources in the Web

Angelo Brayner
University of Fortaleza, Brazil

Marcelo Meirelles
University of Fortaleza, Brazil

José de Aguiar Moraes Filho
University of Fortaleza, Brazil

ABSTRACT

Integrating data sources published on the Web requires an integration strategy that guarantees the local data sources' autonomy. A multidatabase system (MDBS) has been consolidated as an approach to integrate multiple heterogeneous and distributed data sources in flexible and dynamic environments such as the Web. A key property of MDBSs is to guarantee a higher degree of local autonomy. In order to adopt the MDBS strategy, it is necessary to use a query language, called the MultiDatabase Language (MDL), which provides the necessary constructs for jointly manipulating and accessing data in heterogeneous data sources. In other words, the MDL is responsible for solving integration conflicts. This chapter describes an extension to the XQuery Language, called MXQuery, which supports queries over several data sources and solves such integration problems as semantic heterogeneity and incomplete information.

INTRODUCTION

The Web (World Wide Web) can be seen as a wide network consisting of the union of several local area networks (LANs) spread over the entire world. However, the local networks that constitute the Web are autonomous and capable of plugging or unplugging themselves into and from the Web at any time.

Over the last few years, the Web has been used to publish several databases. Of course, databases available on the Web are heterogeneous since they might be defined by using different data models (e.g., relational or object data model), managed by different database systems (DBSs), or running in different computational environments (regarding operating system and hardware). Furthermore, the integration of databases on the Web should be realized without interfering in the management and processing of local data. In other words, databases should be integrated preserving the local autonomy of each database. Despite the fact that heterogeneity and the autonomy of multiple databases on the Web is a reality nowadays, users (and applications) need shared access to those databases. Thus, it is possible to submit queries against several heterogeneous databases located in distinct local networks throughout the Web.

Consequently, integrating databases published on the Web has become a challenge to the database technology. Several approaches for integrating heterogeneous and autonomous data sources have been proposed since the late '80s. In this chapter, we propose a strategy based on the multidatabase approach for integrating heterogeneous databases published on the Web. For that reason, we describe a new MultiDatabase Language (MDL), called MXQuery, since the proposed strategy uses XML (extensible markup language) as the common data model (CDM; conceptual schema) to represent the multiple data sources' schemas. The MXQuery, which is an extension to the XQuery Language, provides constructors and operators for supporting queries over multiple heterogeneous data sources. The MXQuery solves integration problems such as semantic heterogeneity and incomplete information. Furthermore, this chapter presents an architecture to process MXQuery queries.

This chapter is organized as follows. Approaches for integrating heterogeneous data sources are studied next. Then, related work is discussed, followed by a description of the MXQuery MultiDatabase Language. Next we present in detail the features of the proposed integration strategy, and then give an overview of the query-processor architecture for the MXQuery Language. Finally, we conclude the chapter.

DATA-INTEGRATION APPROACHES

Federated Databases

Federated database is an approach for integrating heterogeneous databases. In a federation of databases, there is a federated schema global to all local component databases. With a federated schema, users make use of an external schema to submit data queries and updates. The federated schema suffers from the (local to global) schema-evolution drawback. In other words, an evolution (modification) of a local schema demands a corresponding modification of the federated (global) schema and can bring about a consistency-loss risk. Federated schema is static in the sense that its maintenance is up to a database administrator (DBA). In general, the federated schema is stored as part of each component database.

Mediators

Motivated by the Web, several research works (Bougamin, Fabret, Mohan, & Valduriez, 2000; Chen, DeWitt, Tian, & Wang, 2000; Das, Shuster, & Wu, 2002; Goldman & Widom, 2000; Manolescu, Florescu, & Kossman, 2001) have focused on the issue of using mediators for integrating databases available on the Web. Mediators are usually specialized software components (and/or engines) for integrating data. A mediator provides a set of virtual views over different sources, called mediated schema, so that it does not change any local database (LDB) schemas. It provides a service for hiding all data characteristics from its users, allowing them to get data in a uniform way.

With respect to the task of building an integrated view (federated or mediated schema), there are three different strategies. The first is