Chapter L
Data Integration: Introducing Semantics

Ismael Navas-Delgado
University of Málaga, Spain

Jose F. Aldana-Montes
University of Málaga, Spain

INTRODUCTION

The growth of the Internet has simplified data access, which has involved an increment in the creation of new data sources. Despite this increment, in most cases, these large data repositories are accessed manually. This problem is aggravated by the heterogeneous nature and extreme volatility of the information on the Web. This heterogeneity includes three types: intentional (differences in the contents), semantic (differences in the interpretation), and schematic (data types, labeling, structures, etc.). Thus, the increase of the available information and the complexity of dealing with this amount of information have involved a considerable amount of research into the subject of heterogeneous data integration. The database community, one of the most important groups dealing with data heterogeneity and dispersion, has provided a wide range of solutions to this problem. However, this issue has also been addressed and solutions have been offered by the information retrieval and knowledge representation communities, making this area a connection point between the three communities.

The Web offers a huge amount of structured and unstructured information. The representation mechanisms are simple, and there are no rules on how to represent the information, so accessing it is a fundamental problem. Basically, information can be accessed by browsing texts and graphics, and users can follow links or use search engines (based on keyword searches) to reach Web documents. The query response capability and inference mechanisms of the Web are limited in comparison with relational and deductive databases, which allow concise queries and reasoning mechanisms to facilitate new knowledge discovery.

Using ontologies for data integration has some advantages over keyword-based systems: ontologies provide a common and shared vocabulary (concepts) for representing the information included in the document (contents). In addition, ontologies allow us to define relationships between concepts (roles). Thus, we can make use of these concepts and roles to perform more complex queries and retrieve exactly the information in which the user is interested. In this way, it is possible to obtain not only extensional information but also intentional information.

Currently, data integration based on ontologies is a very active area of research, which is referred to by different names—semantic mediation, conceptual mediation, semantic data integration, and so forth—depending on the goal. Consequently, great advances are being made in the context of the Semantic Web, and some important problems such as semantic interoperability are being analyzed. However, there are many other problems to be solved in this area, and it is necessary to study new proposals and find improvements that will cover current and future needs.

BACKGROUND

Traditional approaches for heterogeneous data integration try to resolve semantic and schematic heterogeneity using solutions based on rich data models. These data models tend to represent the relationships between distributed and heterogeneous data sources. Despite the fact that most traditional systems deal with a small number of structured data sources, more recent approaches deal with a larger number of data sources (both structured and unstructured).

Data integration systems are formally defined as a triple <*G,S,M*>, where *G* is the global (or mediated) schema, *S* is the heterogeneous set of source schemas, and *M* is the mapping that maps queries between the source and the global schemas. Both *G* and *S* are expressed in languages over alphabets comprised of symbols for each of their respective relations. The mapping *M* consists of assertions between queries over *G* and queries over *S*. When users send queries to the data integration system, they describe those queries over *G*, and the mapping then asserts connections between the elements in the global schema and the source schemas.

The most important proposal to solve the data integration problem is the wrapper/mediator architecture (Figure 1). In this architecture, a mediator, which is an intermediate virtual database (with a schema *G* according to a previous definition of data integration system), is established between the data sources (with a set of schemas *S*) and the application using them. A wrapper is an interface to a data source that translates data into a common data model used by the mediator. The user accesses the data sources through one or several mediator systems that present high-level abstractions (views) of combinations of source data. The user does not know where the data come from but is able to retrieve the data by using a common mediator query language.

Mediator-based integration has query translation as its main task. A mediator in our context