On the Applicability of Schema Integration Techniques to Database Interoperation

Mark W.W. Vermeer and Peter M.G. Apers

Centre for Telematics and Information Technology
University of Twente, Enschede, The Netherlands
{vermeer,apers}@cs.utwente.nl

Abstract. We discuss the applicability of schema integration techniques developed for tightly-coupled database interoperation to interoperation of databases stemming from different modelling contexts. We illustrate that in such an environment, it is typically quite difficult to infer the real-world semantics of remote classes from their definition in remote databases. However, defining relationships between the real-world semantics of schema elements is essential in existing schema integration techniques. We propose to base database interoperation in such environments on instance-level semantic relationships, to be defined using what we call object comparison rules. Both the local and the remote classifications of the appropriately merged instances are maintained, allowing for the derivation of a global class hierarchy if desired.

1 Introduction

Interoperation among pre-existing, heterogeneous, and autonomous databases has been an important research topic in the last few years. Recently, the trend in database interoperability research is moving towards architectures for inter-operation of databases on a scale that goes beyond the context of a single organisation, exploiting the communication facilities offered by world-wide networks. The canonical model used is often an object-oriented one [1]. It has been recognised that such an environment requires flexible and scalable architectures, where users of the component databases are provided with tools to establish importation of information from remote data sources [2,3]. Tightly-coupled approaches, where the schemata of all component databases are unified into a single global schema by a central modelling authority possessing a helicopter view of all component databases [4], are generally agreed to be infeasible in such situations, if only because the component databases may be quite diverse, and no-one can be expected to grasp all information available in the interoperation environment.

Two main approaches towards a more loosely-coupled style of database interoperation can be distinguished. In the multidatabase approach [2], users are expected to define their information needs using a powerful query language with constructs for on-the-fly semantic reconciliation of heterogeneous data. It has been argued, however, that this puts an unacceptable burden on the user, to whom a single logical view of the interoperable databases is no longer presented.
An alternative is the federated approach [5], where a locally integrated schema is composed out of the schema of the local database and the import schema, which is a selection on export schemata of remote databases, by defining relationships between local and imported data. In this paper, we address the question to what extent schema integration techniques developed primarily for tightly-coupled environments are applicable in such environments.

One of the central problems with this approach is that the definition of relationships between local and imported data is far from trivial in a situation where information on the meaning of a remote schema is limited. In tightly-coupled architectures, the schema integrator is supposed to have a helicopter view of all databases, possibly obtained from intensive communication with the DBA's of the participating databases, but in our context a schema integrator must typically thrive on the remote class definitions and some limited additional information. Although many authors have advocated a more loosely-coupled approach to database interoperation, this issue has not been dealt with satisfactorily. Integrated data definition techniques used in federated architectures are usually directly based on schema integration techniques employed in tightly-coupled approaches [6], which in turn are often strongly similar to view integration techniques [7]. All these schema integration techniques require either explicitly or implicitly that (the relationship between) the real-world semantics of the classes to be integrated is known. This is a reasonable assumption in tightly-coupled approaches, but as we will illustrate in this paper, in a federation of databases from multiple modelling contexts this may be surprisingly difficult.

Instance integration [8] has been considered to logically succeed schema integration; i.e. once the relationship between classes defined in different schemata has been determined, the integration of the database instances becomes an issue. More recently, some work has emerged that explicitly considers instances in determining schematic relationships [9,10]. In this paper, we argue that in absence of full knowledge on the semantics of remotely defined classes, instance level semantic relationships form an appropriate basis for database interoperation. In essence, we maintain both the local and the remote classifications on a set of appropriately merged objects. Relationships between local and remote classes may then be derived from relationships between the objects they classify.

The remainder of this paper is organised as follows. In Section 2, we review some basic ideas of schema integration, illustrating that the semantic knowledge required by these techniques may not be available in loosely-coupled environments. In Section 3, we propose an instance-based approach to database integration. In Section 4, we discuss the derivation of integrated objects and classes from instance level relationships. Section 5 then illustrates that several schema-integration techniques are still applicable in our approach. Section 6 presents our conclusions.