BUILDING SCALABLE MEDIATOR SYSTEMS

Chantal Reynaud
University of Paris-Sud, CNRS (L.R.I.) & INRIA (Fiuturs), orsay, France; University of Paris X-Nanterre, France.

Abstract: This paper deals with PICSEL mediator systems which integrate services. We propose a scalable approach which exploits standardized specifications provided by normalization organisms. The paper focuses on the use of such specifications to automate the construction of the mediator. An illustration in the tourism domain with OTA specifications is given.

Key words: Service integration, mediation, scalability.

1. INTRODUCTION

In the recent years, considerable research work has been done on data mediation systems between users and multiple data sources leading to integration systems related to a same domain. A mediator system provides a uniform interface for querying collections of pre-existing data sources that were created independently. Several data mediation systems have been implemented. They have proved to be suitable for building specialized data servers over a reasonable number of data sources (Chawathe et al., 1994; Genesereth et al., 1997; Kirk et al. 1995).

A mediator system is composed of two parts, a part which contains knowledge corresponding to the application domain of the system and a query engine which is generic. The knowledge part is composed first of a single mediated schema, also called ontology, which is a description of the application domain. Second, it is composed of a set of source descriptions expressed as views over the ontology. They model the correspondence between the ontology and the schemas of the data sources.

Research works have developed languages for describing the mediated schema, the content of data sources and the users' queries (Etzioni et al.,

1994; Papakonstantinou et al., 1995). Other research works addressed the issue of providing sound and complete algorithms for rewriting queries using views over data sources. The information integration context is typical of the need of rewriting queries using views for answering queries because users of data integration systems do not pose queries directly to the sources in which data are stored but to a set of virtual relations that have been designed to provide a uniform and homogeneous access to a domain of interest. The rewriting problem that has been extensively studied concerns the pure relational setting (Halevy, 2001).

Improving scalability is a problem which is now studied in the setting of peer-to-peer computing (Milojicic et al., 2002) but it had not received a lot of attention in the setting of centralized mediator systems. This paper deals with this problem in the setting of the PICSEL project (Goasdoue et al., 2000), a collaboration with France Telecom R&D. A first project, PICSEL1, was dedicated to the construction of a declarative development platform of a mediator system. A very rich knowledge representation language, CARIN, has been proposed to model application domains and the content of the sources. Algorithms have been designed for rewriting queries over data sources in an efficient way. We also proposed a representation of an ontology of a real application domain, the tourism domain, in CARIN. We defined methodological guidelines to represent the ontology given this particular language. However, in spite of these results, building the ontology remains a difficult and very time consuming task preventing the deployment and the scalability of mediator systems.

Consequently we aimed, in a new project, PICSEL2, at automating the construction of the ontology in the PICSEL setting. We considered that resources were XML documents and we used PICSEL to build a mediator in the tourism domain by automating the construction of the ontology from XML documents corresponding to standardized messages specifications provided by Open Travel Alliance, OTA (www.opentravel.com), in the travel industry. This application was very interesting because it has allowed to go beyond the initial objective of the project. Domain standards reuse not only allows the automation of the construction of the ontology but of all the knowledge part. Moreover, suitable user interfaces have been automatically generated from the ontology and XML messages have been automatically written from query plans. As a result, the approach improves scalability avoiding strong dependency between the system and the available resources.

The paper is organized as follows. Section 2 provides the general description of the PICSEL2 approach. In section 3, we present the automation of the construction of the mediator. Section 4 focuses on the interface part.

2. GENERAL DESCRIPTION OF THE PICSEL2 APPROACH

PICSEL2 approach aims at building a scalable mediator-based integration system PICSEL with available resources restricted to XML documents. Improving scalability needs automation. As the query engine is generic, the part of the system concerned by automation is the knowledge-based part of the system, mat is the construction of the ontology and the description of the contents of the available resources.

The specificity of the approach is that automation is based on XML standardized specifications provided by organizations for the exchange of structured e-business messages. Thus, the approach allows services integration instead of data sources integration. Consequently, the ontology whose construction is automated (cf. 3.1) describes the services of a domain. The second part of the knowledge part of the mediator describes the functionalities of available services, this part being also automatically generated (cf. 3.2).

Moreover, the goal of the approach is to increase independency between available service providers and the system. Thus, available services are decoupled into two parts. Decoupling of services includes information hiding based on the difference on internal business and public message exchange protocol interface descriptions. Coupling of the mediator systems with services is achieved via interfaces. The names of the performed messages are mentioned in the public part of these interfaces using a language of description of services (for example WSDL).

In this setting, a user query, formulated on the ontology thanks to an interface dynamically generated from the ontology (cf. 4.1), is relative to a service. It is translated into a set of queries using their descriptions. Then the queries are performed by available service providers. Wrappers are usually needed at this step in data sources integration systems. In our approach, they are replaced by a generic module which automatically translates the set of queries into XML documents corresponding to standardized messages (cf. 4.2).

In respect to this approach, we built a mediator system integrating services (air booking, hotel reservation, ...) from standardized specifications defined by the Open Travel Alliance. OTA is a non-profit organization working to define industry-wide, e-business specifications. Our application exploited 115 OTA XML-Schemas defining the elements to be used in messages when searching for availability and booking in the travel industry.

3. AUTOMATISATION OF THE CONSTRUCTION OF THE MEDIATOR

In a first sub-section we describe the building of the ontology, then the generation of the service descriptions.

3.1 Semi-automated construction of the ontology

The building of the ontology is a two-step process. First, a very simple version, guided by standardized specifications, is manually built. We consider all the messages (e.g. AirAvailabilityRequirement) for which a content is defined by the normalization organism that is considered. Then, we group the messages into categories (e.g. AirBookingService). As a consequence, we obtain the two first levels of a class hierarchy. Names of the classes in level 2 are names of standardized messages. Names of the classes in level 1 are names of categories. The name of the root class denotes the domain of interest (e.g. tourism service).

In a second step, the initial hierarchy is enriched from standardized specifications in a semi-automatic way. Indeed we ask the ontology designer to valid and modify, if needed, the enrichment. A set of classes, a set of properties which characterize classes and a set of relations among classes are extracted from XML-schemas associated to XML documents provided by the standardization organism. More details about heuristics used in this process are given in (Giraldo and Reynaud, 2002). Then, the extracted elements are structured. That means connecting the two-level initial class hierarchy with the output of the extraction phase. Thanks to common classes (terms of level 2 in the initial hierarchy correspond to names of the messages and are root elements in the XML-schemas), the connection can be automated. Finally, the model is automatically represented in CARIN.

3.2 Automated generation of service descriptions

In PICSEL the description of the content of each data source is given in terms of a set of logical implications $v_i(X) \Rightarrow p(X)$ establishing a link between a view v_i and the domain relation p whose instances can be found in the source. The use of PICSEL to integrate services leads to define for each service as many views as messages the service is able to perform. The logical implications are generated in an automatic way from the names of the messages extracted from the public part of the services. For example, S_1-v_1 $\Rightarrow$ AirAvailabilityRequirement could be one implication defining a view with v_1 the name of the view that is associated with

AirAvailabilityRequirement in the ontology, a class corresponding to a name of a message the service S_1 is able to perform. This implication says that the service S_1 fulfills searching for availability of flights.

These view definitions are not sufficient. A user must be able to precisely define the service he is looking for. For example, he must be able to express that he wishes to have a flight from the airport CDG. Thus, a view associated with the relation corresponding to the departure airport in the ontology must be defined. More generally, we have to define views for all the elements composing the messages services providers can perform. As all data composing the messages are already represented in the ontology, such views are automatically generated.

4. INTERFACING PICSEL2 MEDIATOR SYSTEM

This section deals both with end-users and service providers interface.

4.1 End-users interfaces for querying

End-users interfaces are automatically generated from the ontology. Their design is based on the fact that there are optional terms in the ontology, indicated by the cardinalities in the XML-Schemas, not all of the same technical level. Three ordered levels of difficulties are distinguished: low, medium and high. When a user wants to query the mediator system, he has to precise his level to access the system through a suitable interface which only presents terms in the ontology he is able to understand, i.e. terms with a level lower than his level. By reducing the number of terms that a low-level user can used in his query, the system becomes quite usable by non-expert users. The terms of the ontology are visualized in a graphical form and each user navigates the ontology as he likes. Moreover, the graphically specified query is automatically translated in CARIN.

4.2 Providers interfaces for translating query plans

We designed an interface to generate XML documents corresponding to the messages that must be performed by the service providers mentioned in the query plans of the PICSEL mediator. The interface is the equivalent of a wrapper. In our setting, it is a generic interface usable for all the service providers. The components and the structure of the messages come from the XML-Schemas. The rewritings given by the PICSEL query engine provide

the values of the elements to insert into the various XML documents understood by the service providers mentioned in the query plans.

5. CONCLUSION

We presented a scalable approach based on standards reuse to integrate multiple services using the PICSEL mediator system. The main point is to increase automation. We shown that PICSEL2 approach allows to automate the knowledge part of PICSEL. Moreover, the building of interfaces with end-users and service providers can also be automated, which is, in addition to the fact that the approach allows an easy and fast construction of such systems, an important point to increase their deployment.

REFERENCES

Chawathe S., Garcia-Molina H., Hammer J., Ireland K, Papakonstantinou Y., Ullman J, D., Widom J., 1994, The TSIMMIS project: Integration of heterogeneous information sources, In 16th Meeting of the Information Processing Society of Japan, Tokyo, p 7-18.

Etzioni O., Weld D., 1994, "A Softbot-Based Interface to the Internet", Communications of the ACM, v 37, n 7, p 72-76.

Genesereth M. R., Keller A. M., Duschka O. M., 1997, Infomaster: an information integration system, In Joan M. Peckman (ed), proceedings, ACM SIGMOD International Conference on Management of Data: SIGMOD 1997: May 13-15, 1997, Tucson, Arizona, USA, v 26, n 2, SIGMOD Record (ACM Special Interest Group on Management of Data), p 539-542, New-York, NY 10036, USA. ACM Press.

Giraldo G., Reynaud C., 2002, Construction semi-automatique d'ontologies à partir de DTDs relatives à un même domaine, Journées Francophones d'Ingénierie des Connaissances.

Goasdoue F., Lattes V., Rousset M.-C., 2000, The use of CARIN language and algorithms for Integration Information: the PICSEL system, International Journal of Cooperative Information Systems, v 9, n 4, p 383-401.

Halevy A., 2001, Answering queries using views: A Survey, VLDB Journal.

Kirk T., Levy A. Y., Sagiv Y., Srivastava D., 1995, "The Information Manifold", In C. Knoblock & A. Levy (eds), Information Gathering from Heterogeneous, Distributed Environments, AAAI Spring Symposium Series, Stanford University, California.

Milojicic D. J., Kalogeraki V., Lukose R., Nagaraja K., Pruyne J., Richard B., Rollins S., Xu Z., 2002, Peer-to-Peer Computing, Technical report HPL, 57, HP Labs.

Papakonstantinou Y., Garcia-Molina H., Widom J., 1995, Object exchange across heterogeneous information sources, In P. S. Yu and A. L. Chen (eds.), 11th Conference on Data Engineering, Taipei, Taiwan, IEEE Computer Society, p 251-260.