17th International CODATA Conference-
"Data and Information for the Coming Knowledge Millennium"
CODATA 2000
Theme I-3, Integration of Heterogeneous Databases and Data Warehousing

A Logic-Agent based System for Semantic Integration

F.Sadri, F. Toni, I.Xanthakos
Department of Computing, Imperial College, London
fs, ft, ix98 @doc.ic.ac.uk

Introduction

The heterogeneity of information sources can syntactic or semantic. The former can be due, for example to platform mismatches and different Database Management Systems, and, the latter to different database schemas and ontological mismatches [1]. Syntactic heterogeneity is catered for by middle-ware components known as wrappers, which provide a relational view of the data stored in the sources. The semantic integration of information involves providing a uniform interface to a number of semantically heterogeneous information sources. This is achieved via a global schema (a set of views) and a global ontology, in the spirit of the mediator architecture introduced by Wiederhold [2]. The users can then form queries using the global schema, independently of the schema (and ontology) of the sources.
If the views that consist the global schema are materialized, the information integration task reduces to Datawarehousing. Data Warehouses provide users with very fast answers, since all the available information is stored locally, but the cost of updating them when the underlying sources change can be very high [3]. However, if the global schema is virtual, then the user query has to be translated into another query or possibly a set of disjunctive (alternative) queries (known as query-plans) that can be posed to the information sources "on the fly".
There are two main approaches to the virtual semantic integration task, namely the query-centric approach and the source-centric one. In the first, the views of the global schema are defined in terms of the information sources, reducing the query-translation to unfolding the definition of each view of the global schema occurring in the user-query [4]. The disadvantage of this approach is that if an information source becomes available, is altered or is no longer available, the definition of the global schema has to change. One possible way to overcome this is to "reverse" the definition of the global schema, which is the idea behind the source-centric approach. In this approach, each source provides the mediator with a description of itself, in terms of global schema views. These descriptions are then used to find the relevant sources, when a user-query is posed [5]. The architecture of the system in this case is much more flexible, but the query-translation task is much more complicated; the two most well-known algorithms for this task are the inverse-rules and the bucket algorithm [6].
Many information integration systems use AI ideas, mainly knowledge representation techniques, planning and learning. Most of the recent information systems in particular follow the agent paradigm, in order to have a more dynamic architecture.

Methods

We have developed a logic agent (namely the KS-agent [8]) based framework for virtual integration of semantically heterogeneous information sources. The KS-agent framework consists of an agent cycle, which in its heart has the IFF abductive proof procedure [9]. The interleaved execution of the IFF proof procedure is responsible for the transformation of the user-queries to query-plans that can be

understood by the information sources. All the information about the global schema and the information sources is given in an abductive logic program, where the sources are represented by abductive predicates. Other integration systems have also used abduction, such as [10].

Results

Our current implementation, developed in Java, follows the query-centric approach and can handle a number of issues that arise during the task of semantic integration, such as harmonizing ontological mismatches, semantically optimizing query-plans using known integrity constraints about the sources, allowing cross-source joins in the definition of global schema views, and supporting the conditional prioritization of sources.
We have further extended our framework, using the inverse-rules algorithm, to accommodate both the query-centric and the source-centric approach for disjoint sets of sources. Additionally, our system overcomes a major restriction of the source centric-approach, where cross-source joins were not supported.

Conclusions

We have developed a logic agent based approach to information integration that combines and extends the query and source centric approaches.
In the future, we intend to develop a query-scheduling component that will be embedded in the agent cycle, along with the actual execution of the query-plans, providing an adaptive behavior that can cope with unexpected results or failures. Other possible extensions of our work involve multi-agent communication (using KQML) and co-operation for optimized information-flow.

References

[1] Hull, R., "Managing Semantic Heterogeneity in Databases: Atheoretical Perspective", in PODS '97, Tuscon, Arizona, USA, 1997
[2] Wiederhold, G. "Mediators in the architecture of Future Information Systems", in Readings in Agents, edited by M.N. Huhns & Munindar P.Singh, Morgan Kaufmann
[3] Widom, J., "Research Problems in Data Warehousing", in the Proc. of the 4th Int. Conf. on Information and Knowledge Management, 1995
[4] Ullman, J., D., "Information Integration using Logical Views", in the Proc. of the Int. Conf. On Database Theory, Delphi, Greece, 1997
[5] Alon, Y., Levy, Anand Rajaraman and Joann J. Ordille, "Querying heterogeneous information sources using source descriptions", in the Proc. of the Int. Conf. On Very Large Data Bases (VLDB), Bombay, India, 1996
[6] Alon, Y., Levy, "Logic-Based Techniques in Data Integration", University of Washington, May 1999
[7] Duschka, O., M., Genesereth, M.,R., "Infomaster – an Information Integration Tool", in the Proc. of the Int. Workshop on Inetlligent Information Integration, Freiburg, Germany, 1997
[8] Kowalski, R.A., Sadri, F., "From Logic Programming to Multi-agent Systems", In Annals of Mathematics and Artificial Intelligence, 1999
[9] T.H Fung, R. Kowalski, "The IFF Proof Procedure for Abductive Logic Programming", In: Journal of Logic Programming (1997)
[10] Bressan, S., et al., "The Context Interchange Mediator Prototype", in the Proc. of ACM-SIGMOD, 1997
[11] Duschka, O., M., Genesereth, M., R., and Levy, A., "Recursive query plans for data integration", Journal of Logic Programming, special issue on Logic Based Heterogeneous Information Systems, 1999