A New Approach to Integrate Heterogeneous Databases: The Mediated Data Integration Architecture

Chaiyaporn Chirathamjaree
School of Computer and Information Science
Edith Cowan University
Australia
c.chirathamjaree@ecu.edu.au

Abstract

Information required for decision making is generally scattered across disparate data sources. To gain competitive advantage, it is extremely important for executives to be able to obtain one unique view of information in an accurate and timely manner. The heterogeneities occur when integrating such data sources which were designed independently. MeDInt – the Mediated Data Integration architecture is introduced to solve this problem. The mediator and wrappers are employed as the middle layer between application and data sources to homogenise heterogeneous data sources. Both the mediator and wrappers are well supported by the Mediated Data Model (MDM), an object-oriented based data model which can describe or represent heterogeneous data schematically and semantically. The MeDInt architecture has been tested and evaluated, and the results are promising.

Keywords: Heterogeneous databases, data sources, integration, mediator, wrappers.

1. Introduction

When interoperation between multiple heterogeneous data sources is required, there would be a number of conflicts arising not only from different database designs, but also from different kinds of data models employed within heterogeneous databases. These conflicts generate the difficulties of homogenisation in terms of data model, schema and semantic [2] [3] [4] [5]. The Mediated Data Integration (MeDInt) architecture for the heterogeneous data integration framework is introduced in an attempt to overcome the above difficulties. It has been developed by focusing on providing a solution to interoperate heterogeneous data sources by transforming both the queries and the data transparently. Furthermore, MeDInt does not only solve schema and semantic heterogeneities, but also conflicts from different query languages and data models, namely data model heterogeneity.

The integration approach proposed in this research incorporates the advantages of both the mediator systems and metadata repository systems. This means that new data sources need to be registered when they are added to the integration system. However, the heterogeneities are resolved at the query time making the mediator system more dynamic. The mediated architecture basically adds a third layer between applications and data sources.

2. Architecture Requirements

The following architecture requirements have been formulated as the framework to develop the integration architecture.

- The schema evolution should not affect the integration. This is to cater for dynamic systems where schemas could be changed frequently. When schema modification is made on data sources, it should not cause large-scale modification to the integration system.
- The integration should cover the major kinds of data sources widely used such as legacy, relational model, and object-oriented model systems.
- The approach should increase automation and reduce amount of work required by end-users. Users should not have to deal with conflict resolutions once they issue queries. The different terminologies used in data sources and the different structures of data sources should not affect users when issuing queries.
- The integration architecture should require minimised modifications when a new data source is added or removed.

3. The MeDInt Architecture

Figure 1 depicts MeDInt (the Mediated Data Integration Architecture) which is based on mediation and wrapping techniques. The two main components are the mediator and wrappers acting as the intermediate agents between clients and multiple data sources to communicate both request queries from clients to data sources and also query results from data sources to clients.

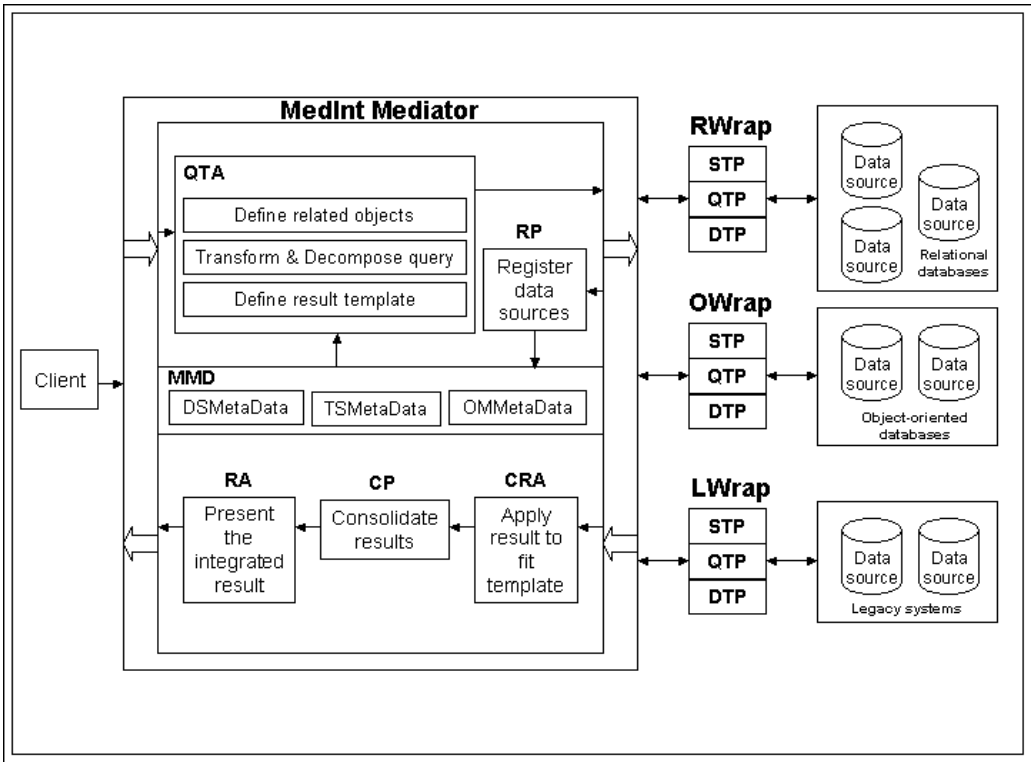

Figure 1: The MeDInt architecture

The MeDInt architecture is represented by four-tiers of components: the application systems which interface to users, the mediator, wrappers and data sources. In addition, the Mediated Data Model (MDM), a data model designed especially for the heterogeneous data integration framework [1], works along with the MeDInt Mediator and wrappers functioning as a central data model and working as the hidden backbone of the integration facilitating the communication between Mediator and wrappers.

3.1 The User Interface

In general, query languages are not capable of utilising and specifying the heterogeneities between heterogeneous systems [6]. Therefore, this approach provides a data model with a query language which captures the heterogeneities on behalf of users so that they can specify their own queries, including semantic contexts.

3.2 The MeDInt Mediator

The MeDInt Mediator provides middle-layer services, as an information integrator does, between the application and wrappers. In general, mediators are responsible for: retrieving information from data sources, transforming received data into a common representation, and integrating homogenised data [7]. In this research, the MeDInt Mediator has been designed to include the following common characteristics of the integration processes:

- registering data sources information,
- defining associate objects and requesting object schemas from wrappers,
- decomposing and transforming a query to subqueries according to data sources,
- generating a result template,
- applying the multiple sets of results to a pre-defined template,
- consolidating the conflict-resolve sets of results, and
- displaying the integrated result to the user.

3.3 Wrappers

Wrappers are in the intermediate layer between the MeDInt Mediator and data sources. A wrapper is invoked when a data source in a difference data model is added to the integration system. Wrappers mainly act as translators providing the MeDInt Mediator with information in the common data model used in the integration system by dealing with the data model heterogeneities of different data sources. The principle objective of wrappers is dealing with data model heterogeneities including the different data definition languages and data manipulation languages by mapping variety data models to the Mediated Data Model. Each MeDInt wrapper is composed of a Schema Translation Processor, a Query Translation Processor and a Data Translation Processor. One novel feature of the architecture is to push unshared characteristics to the wrappers to reduce the amount of middleware modification when a data source is added, removed or modified. In addition, the use of the Mediated Data Model eliminates problems relating to the data model heterogeneity by providing the common data model

acknowledgeable by components in the MeDInt Mediator.

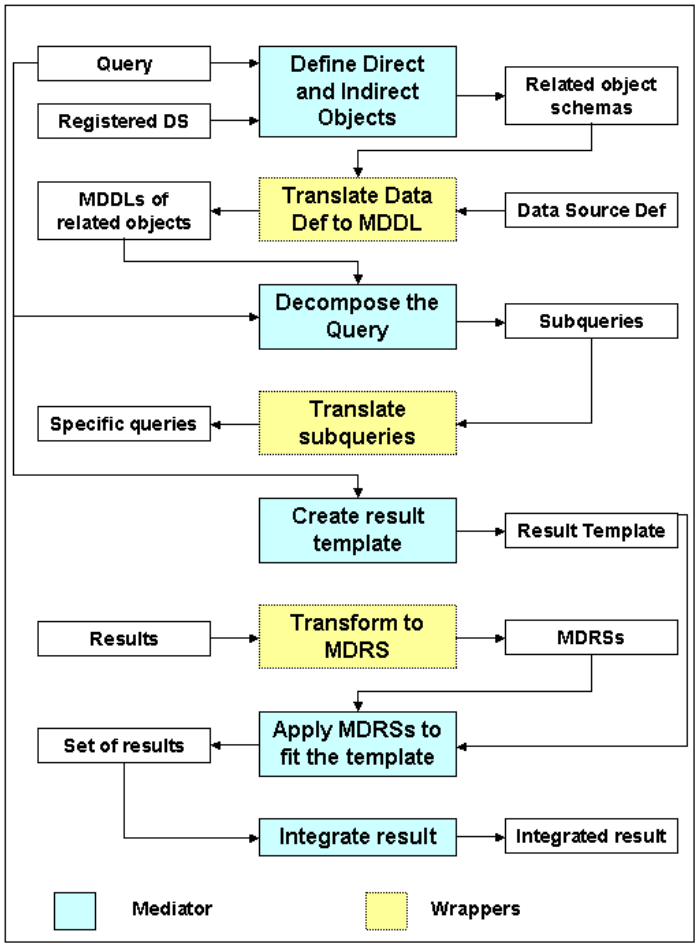

Figure 2: MeDInt Proceses

4. MeDInt Processes

Figure 2 illustrates the processes of the MeDInt architecture.

When a new data source is added to the integration system, it is registered to the Mediated MetaData (MMD). Data source information, for example, assigned name, location, type, description, and constraints relating to its structure and semantics are collected into the Data Source Metadata (DSMetaData), a category of MMD. A query from a user to retrieve the information from heterogeneous data sources is sent to the MeDInt Mediator instead of directly to the data sources. The required objects are determined and a request is submitted to the wrapper to get the related object schema definitions. The submitted query from the user is transformed to a specific query language appropriate to the database management system of the data source. A template for the results is created from the results obtained from multiple data sources. This method does not try to resolve conflicts directly which would be more difficult and complicated.

After getting a response data back from data sources, a component of a wrapper translates the query results into the Mediated Data Representation Structure (MDRS). The conflict resolution is done by applying all MDRSs to fit into the structure of the predefined template. The resultant MDRSs that are structurally equivalent are then integrated and consolidated. Finally the integrated result is sent to the user.

This architecture overcomes the weakness inherent in other approaches that require the physical or logical integration of component schemas. Only the query result from each source, according to the result template, will be integrated instead. The template will be created from the submitted query. The resultant data from each data source will be applied to fit to the template which is the means by which the heterogeneities are resolved.

5. Results and Discussion

A number of example problems of heterogeneities from a number of information systems that require integration have been tested. The objectives are to demonstrate the integration process using the MeDInt architecture and to evaluate its correctness.

Table 1: Summary of the heterogeneities resolved in the MeDInt architecture

Heterogeneities	Conflicts	Test Problem1	Test Problem2	
			Query 1	Query 2
Model		√	√	√
Schema	Naming	√	√	√
	Structural	√	√	
	Specialisation		√	√
	Relationship		√	
Semantic	Naming		√	
	Scaling	√		√
	Abstraction			√
	Representation	√		

The proposed MedInt Architecture and MDM have been tested for functionalities and the outcomes look promising. Results (Table 1) indicate that the objectives in resolving conflicts both structurally and semantically have been achieved. From the table above, the following three categories of heterogeneities have been determined: Model, Schema, and Semantic. All of them have been solved as shown by the MedInt with the support of the MDM which is suitable for homogenising different data models, schemas and semantics of component data sources. Another feature of our proposed model is that it can be implemented in any languages. We have chosen XML as the implementation language in the prototype because it offers a number of advantages. XML is platform independent, provides self-described tags which are easy to understand. It is also suitable for describing schema and semantic of objects in a real world since XML is based on an object-oriented model.

6. Conclusion and Future Works

The research proposes the MeDInt Architecture as the framework based on the mediated approach for the integration of heterogeneous data sources to solve conflicts occurring when interoperability is required. The paper presents a new approach for achieving the interoperability of multiple data sources logically integrated at the time the query is issued. The system is able to describe or represent heterogeneous data both schematically and semantically. No pre-integration is required before users can issue their queries. This avoids the problem of local schema evolution which usually happens in dynamic systems. Further investigations are planned to cover the query performance issues. Another possible future work is to incorporate the write access through the updating of master data sources and the replication of data sources.

References

[1] Chirathamjaree, C. & Mukviboonchai, S. “A mediated data model for heterogeneous data integration,” The 2nd Annual International Conference on Computer and Information Science (ICIS '02), Seoul, Korea, 8-9 August 2002.

[2] Goh, C.H., Madnick, S.E. & and Siegal, M.D. “Context interchange: overcoming the challenges of large-scale interoperable database systems in a dynamic environment,” The third International Conference on Information and Knowledge Management, Gaithersburg, MD, 1994.

[3] Holowczak, R. D. & Li, W. S. “A survey on attribute correspondence and heterogeneity metadata representation,” *Institute of Electrical & Electronics Engineers*, Available: http://church.computer.org/conferences/meta96/li/paper.html, 1996.

[4] Kim, W., Choi, I., Gala, I. & Scheevel, M. “On resolving schematic heterogeneity in multidatabase systems,” *Journal of Distributed and Parallel Database*, 1(3), 251, 1993.

[5] Miller, R.J. “Using schematically heterogeneous structures,” *SIGMOD'98*, 189-200, 1998.

[6] Papakonstantinou, Y., Molina, H. G. & Widon, J. “Object exchange across heterogeneous information sources,” ICDE '95 proceedings, 1995.

[7] Wiederhold, G. & Genesereth, M. “The conceptual basis for mediation services,” *IEEE Expert*, 12(5), 38-47, 1997.