Association for Information Systems

AIS Electronic Library (AISeL)

ICEB 2004 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-5-2004

A Mediator to Integrate Databases and Legacy Systems: The Mediated Data Integration (MeDInt) Mediator

Chaiyaporn Chirathamjaree

Follow this and additional works at: https://aisel.aisnet.org/iceb2004

A Mediator to Integrate Databases and Legacy Systems: The Mediated Data Integration (MeDInt) Mediator

Chaiyaporn Chirathamjaree

School of Computer and Information Science, Edith Cowan University, Perth, 6050, Australia
c.chirathamjaree@ecu.edu.au

ABSTRACT

To interoperate data sources which differ structurally and semantically, particular problems occur, for example, problems of changing schema in data sources will affect the integrated schema. In this paper, conflicts between heterogeneous systems are investigated and existing approaches to integration are reviewed. We propose a new mediated approach employing the Mediated Data Integration Mediator (MeDInt), and wrapping techniques as the main components for the integration of databases and legacy systems. The MeDInt mediator acts as an intermediate medium transforming queries to sub-queries, integrating result data and resolving conflicts. Wrappers then transform sub-queries to specific local queries so that each local system is able to understand the queries. This framework is currently being developed to make the integration process more widely accessible by using standard tools. A prototype is implemented to demonstrate the model.

Keywords: Conflict Resolution, Heterogeneous Databases, Integration, Legacy Systems, Mediation, Wrappers

1. INTRODUCTION

The information required for decision making by executives in organizations is normally scattered across disparate data sources including databases and legacy systems. To gain a competitive advantage, it is extremely important for executives to be able to obtain one unique view of information in an accurate and timely manner. To do this, it is necessary to interoperate multiple data sources, which differ structurally and semantically. In the process of interoperating any two or more database systems, there are critical problems that need to be solved, for instance, some databases are designed from different models, objects which have the same meaning in different databases might have different names, and objects which have the same meaning in different systems might be measured by different units. Furthermore, there are identity conflicts, representation conflicts, scope conflicts, etc [1; 2; 4; 8; 9]. Although several researchers have studied the conflicts and integration of heterogeneous database systems [1; 9; 11; 13; 14; 17], there is still no common methodology for resolving conflicts and integrating such databases. Particularly, few studies have focused on the integration of databases and legacy systems. In legacy systems, the semantics are hidden and hard to determine. In fact, some legacy systems store data to flat files, which are completely different in schematic design from database management systems (DBMSs).

Another significant issue is that almost all research on database integration presents pre-integration approaches using global schema techniques, which require complete integration. All local views are mapped by one global view. This method is convenient for users but it does not operate in the real-time manner because the global view must be created before query processing. As a result when only one object of a local system is modified, it affects the global schema requiring huge changes [4]. Furthermore, schema and semantic conflicts must be solved in the process of the global schema creation. The more data sources involved, the more difficult such conflicts are to be solved. This research focuses on the database and legacy integrating solution that avoids using the global schema pre-integration approach.

The Mediated Data Integration (MeDInt) Mediator 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, this approach does not only solve schema and semantic heterogeneities, but also conflicts from different query languages and data models, namely data model heterogeneity.

2. RELATED WORKS

2.1 Conflicts and Resolution

Information from different sources can not be presented to users if it has not passed the process of conflict resolution. In terms of database integration, conflicts are differences of relevant data between component local database systems. The taxonomy of conflicts in this paper is divided into Schema conflicts and Semantic conflicts.

Schema conflicts are discrepancies in the structures or models of heterogeneous database management systems. Naming conflicts [8], Structural conflicts [4; 8; 9], and Identity conflicts fall into this conflict category. Naming conflicts are the synonyms or homonyms of objects in local systems. Structural conflicts are the different uses of data models to represent the same object. Identity

conflicts occurs when the different attributes, as a key, are used to access the same meaning information.

Semantic conflicts are discrepancies in the meaning of related data among heterogeneous systems such as Naming conflicts, Representation conflicts [3; 4], Scaling conflicts [2], Granularity conflicts, Precision conflicts [1], Missing data, Scope conflicts, and Computational conflicts [2]. Naming conflicts are able to occur in data itself as well as in the structure of data. Representation conflicts or Format heterogeneities are the different uses of formats or data types to represent the same meaning objects. The different units of measurement generate Scaling conflicts.

From a survey of the literature, several methods to resolve conflicts have been found. In the case of Naming conflicts, a catalog [7], tables [4], or meta-data repository [1] can be used for maintaining these correspondences. An Object Exchange model [12] is able to transform semantics into simple structures that are powerful enough to represent complex information by using meaningful tags or labels. Kim [7] suggests three ways to resolve different representations of equivalent data: static lookup tables, arithmetic expressions, and mappings. In addition, a formulae has been suggested by Holowczak & Li [4] for converting values in one system to correspond with units in another system. They also introduce Superclasses to encapsulate each component database to create their relationships. Differences in attribute naming are solved by aliases [1; 4]. By using benefits of functions, Hongjun [5] proposes a data mining approach to discover data value conversion rules. Furthermore, independent views can be constructed to solve Structural conflicts. A view neither depends on any specific names nor on changes when schemas are modified [9].

2.2 Integration Approaches

Numerous integration approaches have been introduced throughout the last twenty years to bring about the interoperability among heterogeneous systems. Missier, Rusinkiewicz, & Jin [10] categorise heterogeneity resolution methodologies into four main broad approaches: Translation, Integrated, Decentralised, and Broker based.

Translation approach needs highly specialised translation for each pair of local database systems. Therefore, the number of translators grows up exponentially especially when local systems increase. The development of these ad hoc programs is expensive in terms of both time and money.

In Tight-coupling approach or fully integrated approach, individual schema from multiple data sources is merged by one or more schemas. If only one schema is prepared, it is called a global schema approach.

Otherwise, it is called a federated database approach. The global schema approach allows access of multiple data sources by providing the conceptual global schema as a logically centralised database [6]. Multiple local schemas are consolidated to create the global schema. Users are able to use one database language to query the global schema without understanding any local schemas. Generally, problems of heterogeneity must be resolved in the process of creating the global schema. A major difficulty is the process of creating global schema which thoroughly understands the differences between the independently-designed heterogeneous local schemas, and homogenises such differences [7]. This approach is more difficult when the number of databases increases. Another approach, the federated database, also allows users to query more than one federated schema without knowledge of local data sources. This approach still requires complete pre-integration. The federated schema must be developed before issuing any queries, so any changes in local schemas would affect the federated schema.

Loose-coupling approach [2] or decentralized approach has been introduced in an attempt to resolve the problems arising from tight-coupling approaches by discarding either pre- or partial-integrated global schema. This approach allows users to query local database systems directly without any global schemas by placing the integration responsibility on users. Multi-database manipulation languages, which are capable of managing semantic conflicts through their specification, are provided as query language tools that are able to communicate with the local databases. Users can see all the local schemas and create their own logical export schema from selected schemas relevant to the information they need [3]. However, it requires users to have semantic understanding and to be able to resolve conflicts in creating their schema, which will be numerous with large numbers of data sources. In Broker-based approach, the crucial part is the conflict detector module using shared ontologies, but the process of doing those ontologies is not completely automated.

The limitations of the above integration approaches have led integration technologies towards a new variety of solutions. Various theories have been applied to solve integration problems such as the object-oriented model, knowledge base [11; 14; 16], ontology [13], and modeling [4].

3. THE MeDInt MEDIATOR

The research has introduced a heterogeneous database integration model incorporating a mediator and wrappers as intermediate layers between the application and data sources. The mediator, MeDInt (Figure 1), serves as an information integrator, between the application and wrappers. Generally, mediators are responsible for retrieving information from data sources, for transforming received data into a common

representation, and for integrating the homogenised data   [15]. In this model, the MeDInt Mediator acts as an
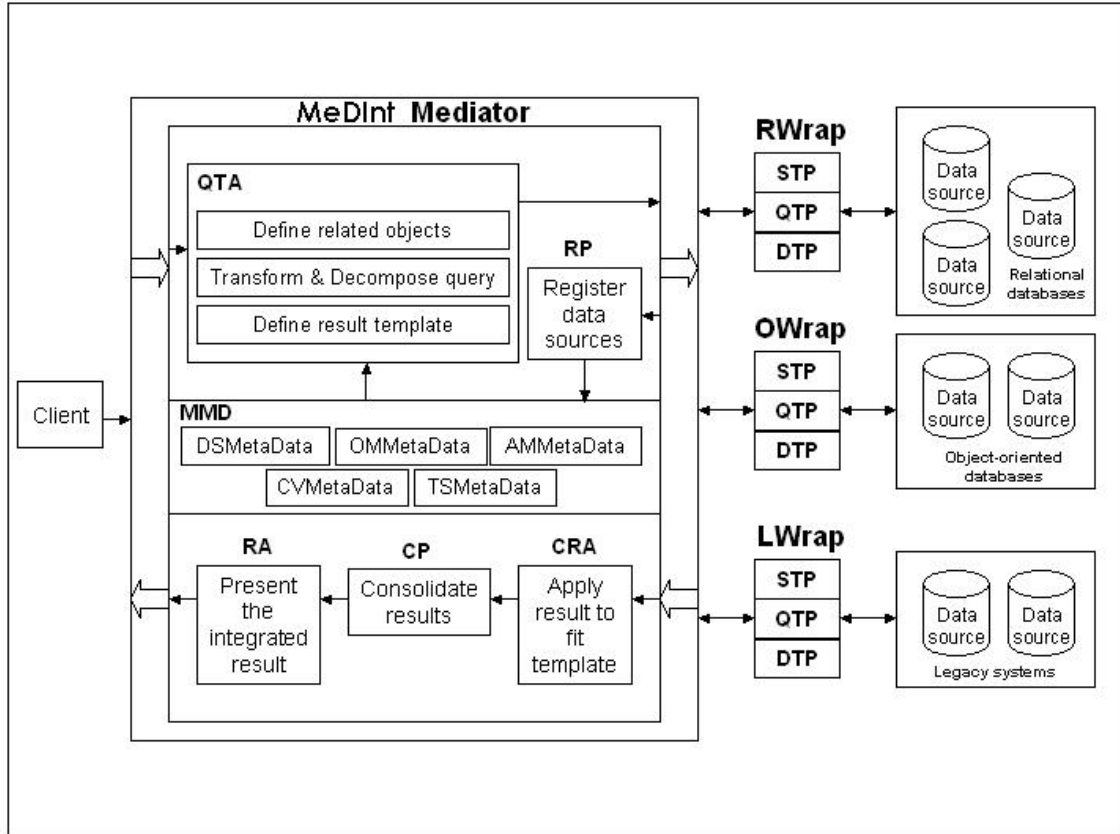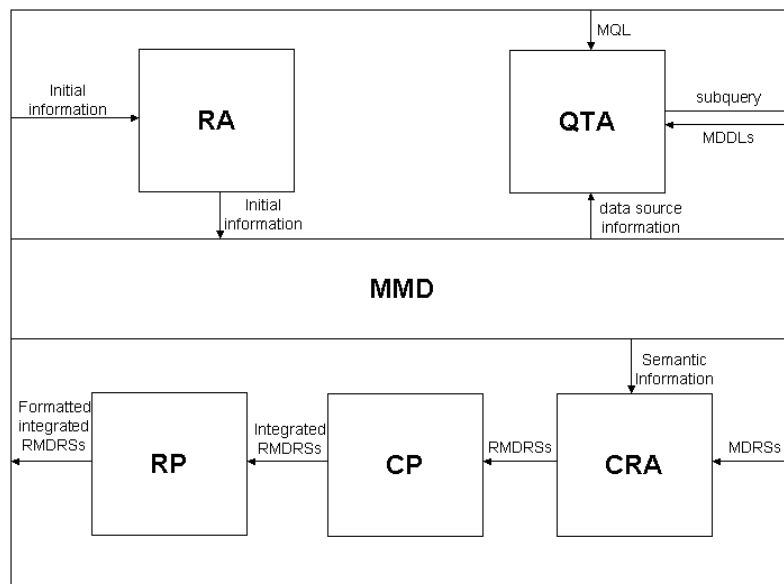

Figure 1. The MeDInt mediator

Figure 2. The six components of the MeDInt mediator

interchangeable agent and facilitator for wrappers and clients. It consists of six components working together transparently to facilitate clients and data sources to achieve the following tasks:

- transforming and decomposing the submitted query into subqueries and then distribute them to associated wrappers;
- providing both schematic and semantic knowledge which is critical for query transformation and conflict resolutions;

- resolving conflicts; and
- consolidating query results.

All the functions above are served by six components (Figure 2), which are the Registering Processor (RP), the Query Transformation Agent (QTA), the Mediated MetaData (MMD), the Conflict Resolution Agent (CRA), the Consolidation Processor (CP) and the Rendering Agent (RA).

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 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 fro the common data model acknowledgeable by components in the MeDInt Mediator.

4. Medint PROCESSES

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 approach 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 mediator and to evaluate its correctness.

Test problem 1 is a Hotel Reservation Information System which provides information for travel agencies. The information systems of contacted hotels need to be interoperated. Heterogeneities have been found when integrating them. The 2nd test problem is a university information system which is composed of a relational system and an object-oriented system.

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 mentioned 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 (the Mediated Data Model has been developed in this study specifically for describing and representing heterogeneous data both schematically and semantically) 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.

Table 1 Summary of the heterogeneities resolved by the MeDInt mediator

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

6. CONCLUSION AND FUTURE WORKS

The research proposes the MeDInt Mediator 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] Abdulla, K., "A new approach to the integration of heterogeneous databases and information systems", *Unpublished doctoral dissertation*, University of Miami, Florida, 1998.

[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] Heimbigner, D., and Mcleod, D., "A federated architecture for information management", in A. Gupta (Ed.), *Integration of information systems: bridging heterogeneous databases*, New York: IEEE Press, 1989.

[4] Holowczak, R. D., and 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/pa per.html, 1996.

[5] Hongjun, L., *A data mining approach for resolving conflicts during data integration,* Department of Computer Science, The Hong Kong University of Science and Technology, Available: http://www.comp.polyu.edu.hk/News/Seminars/sem 980917.html, 1998.

[6] Hughes, J.G., *Object-oriented databases*, New York: Prentice-Hall, 1991.

[7] Kim, W., *Modern database systems: the object model, interoperability, and beyond*, New York: ACM Press, 1995.

[8] Kim, W., Choi, I., Gala, I., and Scheevel, M., "On resolving schematic heterogeneity in multidatabase systems", *Journal of Distributed and Parallel Database*, Volume 1, No. 3, p251, 1993.

[9] Miller, R.J., "Using schematically heterogeneous structures", *SIGMOD'98*, pp189-200, 1998.

[10] Missier, P., Rusinkiewicz, M., and Jin, W., "Multidatabase languages", in A. Elmagarmid & M. Rusinkiewicz and A. Sheth (Eds.), *Management of heterogeneous and autonomous database systems*, CA: Morgan Kaufmann Publishers, Inc., 1999.

[11] Neild, T. H., "The virtual data integrator: an object-oriented mediator for heterogeneous database integration", *Unpublished doctoral dissertation*, Northwestern University, 1999.

[12] Papakonstantinou, Y., Molina, H. G., and Widon, J., "Object exchange across heterogeneous information sources", *ICDE '95 proceedings*, 1995.

[13] Phijaisanit, W., "Dynamic meta-data support for information integration and sharing across heterogeneous databases (federated database)", *Unpublished doctoral dissertation*, George Mason University, 1997.

[14] Srinivasan, U., "A framework for conceptual integration of heterogeneous databases", *Unpublished doctoral dissertation*, University of New South Wales, 1997.

[15] Wiederhold, G., & Genesereth, M., "The conceptual basis for mediation services", *IEEE Expert*, Vol. 12, No. 5, pp38-47, 1997.

[16] Woelk, D., Bohrer, B., Brice, R., Huhns, M., Jacobs, N., Ksieyzk, T., Ong, K., Singh, M., Singh, M., & Tomlinson, C., *Carnot,* Microelectronics and computer technology corporation(MCC*)*, Available: http://www.mcc.com/projects/infosleuth/archives/carnot/,n.d.

[17] Yu, T. F., "Information modeling and mediation languages and techniques for information sharing among heterogeneous information systems", *Unpublished doctoral dissertation*, University of Florida, 1997..