IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITP4554

Heterogeneous Data Mediation Using Meta-models – An Architectural View

Luyin Zhao
School of Management, Rutgers University
luyin@pegasus.rutgers.edu

Keng Siau
College of Business Management, University of Nebraska-Lincoln
ksiau@unl.edu

ABSTRACT

This research-in-progress paper introduces the concept of data mediation and a typical mediation architecture used by previous solutions, points out limitations of this architecture and states why meta-model can be adopted for better interoperability. Based on the latest meta-model standard from Object Management Group - CWM, a new mediation architecture is proposed. Final discussion is on future research in terms of prototyping, toolkit and formalizing meta-model standard for data mediation.

INTRODUCTION

Data mediation is a research area that deals with integrating information from different, usually heterogeneous, data sources, including regular databases, XML source, record files, email systems, etc. The software that handles or masks data heterogeneity from end users is called a *mediator*.

Mediation based interoperability provides users with (probably converted) data view and query language for querying heterogeneous data sources. This type of interoperability is considered at the *data* level in contrast to the *service* level interoperability (CORBA, DCOM). More specifically, mediation solutions provide users a way to send on-demand queries to heterogeneous data sources. In other words, in users' eyes, there is a homogeneous (common) view despite the heterogeneous data sources. User queries issued on this view are intercepted by the mediation system and converted to query formats that can be accepted by heterogeneous data sources[1].

EXISTING DATA MEDIATION ARCHITECTURE

In information mediation research community, there are several research projects that have been completed. Two of the most important ones are TSIMMIS (by Stanford University [3][4]) and MIX (by University of California at San Diego [1][2]). Both of them use a typical mediation architecture with major components shown in Figure 1.

– Common data schema (model) definition
The purpose of defining a common data schema is because heterogeneous data sources have different data schemas (e.g., relational database has relational schema, XML database has tree-like schema, etc.) Common data schema provides users with a common view so that queries can be issued on this view.

– Common query language definition
Common query language provides a single query language for querying different data sources.

– Wrappers
Wrappers sit on top of heterogeneous sources to export data in a uniform format to the mediator. Wrappers provide access to heterogeneous data sources by converting application queries into source specific queries or commands. Wrappers also accept user queries, decide whether they are allowed, translate them into queries that underlying data sources can recognize, and return query results to mediator by converting results into formats as defined by the data schema.

The above mediation architecture has many limitations in the aspects of:

– Handling the diversity of data schemas
Although there might be multiple mediators in a mediation system that handles a set of data sources, the variety of data sources can be significant in terms of different data schemas. Therefore choosing of common mediation data schema is preferably dynamic (depend upon actual data sources) instead of static in order to avoid losing too much information during the conversion to common data schema.

– Meeting different query preferences
Information users usually have different query language preferences. Therefore forcing them to use a query language that is specific to the common data schema may be very time consuming and error-prone.

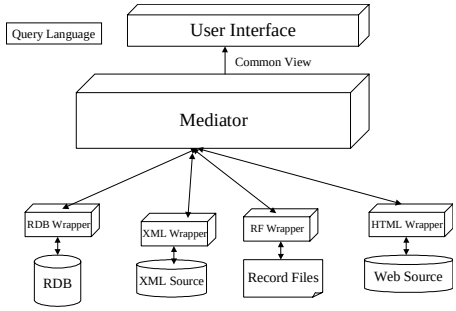
– Providing a true independent middleware that is comparable to CORBA/ DCOM

This mediation architecture does not:
(1) Allow data sources to publish their schema precisely in a neutral way for communication purpose, which is similar to publishing CORBA services using IDL.
(2) Allow data users to choose their preferred data schemas and query languages, which is similar to the flexibility possessed by CORBA clients.

META-MODEL BASED DATA MEDIATION

Simply speaking, meta-data is data that describes data. Correspondingly, meta-model is the meta-data for defining meta-data. Common Warehouse Meta-model (CWM), a new standard adopted and maintained

Figure 1 – A typical mediation architecture



by Object Management Group (OMG) recently, is a typical and relatively complete meta-model standard for defining meta-data for all aspects of data warehousing [5][6].

Database or data warehousing domains can have many meta-models for exchanging meta-data. Similarly, for the purpose of data mediation, we only need meta-models that can be used for exchanging:

(1) Different data schemas. For example, relational data schema, XML schema, record file schema, etc. For this purpose, we may borrow some meta-models from the "Resource" level meta-model from CWM standard.
(2) Different query languages. For example, SQL, XQL, etc. Currently there is no standardized meta-model for this purpose. Although this part may need extensive work, we believe it is viable as long as meta-model for data schemas can be standardized.

Being meta-model enabled, data sources become capable of providing interchangeable data schemas and data queries to other entities using a precise, neutral, standard, and metadata way. The consequence is that a "tighter" common data schema and an appropriate query language could be chosen by end users for querying heterogeneous data sources.

Based on the above discussion, we propose a meta-model based mediation architecture as shown in Figure 2. This architecture assumes underlying data sources are all meta-model aware. Being meta-model aware can eliminate the use of wrappers. In this architecture, the way data sources present themselves to the mediator and receive queries is implementation-independent encoded as mediation meta-model compatible meta-data. The mediator has a meta-model parser for converting between original meta-data (data schemas, queries) and meta-model compatible meta-data. In addition, if necessary, the schema integrator can do semi-automatic schema integration with the help of end users by rendering the integration process using UML. Customized data schema view indicates that end users are able to select their preferred view of integrated schema. For example, when 90% underlying data sources are XML, a user may want to use XML view as the common data schema. Because the mediator receives meta-model encoded data schemas from underlying data sources, it is easy for it to integrate these schemas into different views.

Meta-model based data mediation inherits the data encoding approach of CWM. That is, meta-data is represented at two layers: At the level that needs to interact with end users (we call it presentation layer), for example, data providers and data users, meta-data is represented using UML to take advantage of its visual description and user interaction capability. While at the data transmission layer, XML is used to encode meta-data because XML is Web friendly and self-described.

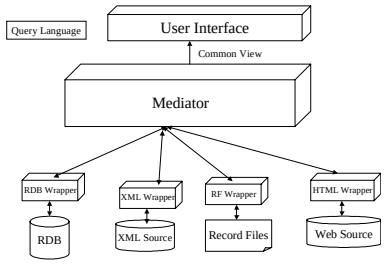
Instead of using one specific data schema (meta-data) as adopted by previous mediation solutions, a mediation meta-model (meta-meta-data) is used as the communication language between mediator layer and data source layer. That is, all meta-data (encoded data schemas) transferred over the network must conform to the mediation meta-model specification. This mediation architecture provides a more independent data schema encoding and exchanging approach while leaving the flexibility of selecting integrated data schema and query language to users.

A sample scenario is that, first, data sources provide mediation meta-model encoded data schema to the mediator. In this way, the mediator can gather data schema information from various data sources. Second, the mediator user chooses what common data schema he (she) would like to use. Because the mediator has model information from all data sources, it is able to transform those data schemas to an available common model chosen by the user. The conversion is based on predefined mapping rules. As an example, [7] proposes a MOF based meta-data solution for database schema integration. Finally, the user issues appropriate queries based on the common data schema. The queries will be encoded again using mediation meta-model and transferred back corresponding data sources.

CONCLUSIONS AND FUTURE RESEARCH

The advent of XML and meta-model standard (CWM) provides new approach to meet Web information interoperability requirements. Different from previous information mediation solutions that used a common mediation model with a single data view and query language, this paper proposes a meta-model based mediation model that connects data source side and mediator side by meta-data rather than specific common data schemas, therefore allowing users to choose their preferred data schemas and query languages. Initial architectural investigation shows positive potentials for further research in this area. Apparently the next step is to build a prototype to validate the architecture and discover hidden problems. Some of the most important issues we envision to work on include:

Figure 2 – Meta-model based mediation architecture



–Design a set of mediation meta-models for the scope of prototyping

Although CWM provides a good startup in terms of offering many standardized meta-models for resources (data sources) such as XML, RDB and Record, there is no standardized meta-model for queries. However, technically this should not be a big issue since the standardization of meta-models for data schema indicates the feasibility of designing meta-models for queries.

–Evaluate and find a toolkit

Because of the huge complexity of handling problems like meta-model processing, certain toolkits are absolutely necessary for prototyping purpose. CWM standard contributors like Oracle, IBM, and Meta Integration Technology Inc. are in the process of developing such toolkits. We will evaluate these toolkits and use them for the prototype.

FOOTNOTES

[1] Please note that semantic interoperability is not the focus of this paper

REFERENCES

[1] Baru, C., (1999). Xviews: XML Views of Relational Schemas. Proceedings of the 10th International Workshop on Database and Expert Systems Application. 700-705.

[2] Baru, C., Gupta A., Ludäscher, B., Marciano, R., Papakonstantinou, Y., Velikhov, P., & Chu, V., (1999). XML-Based Information Mediation with MIX. Proceedings of the 1999 ACM SIGMOD international conference on Management of data. 28 (2), 597-599.

[3] Molina, H.G., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., & Widom, J., (1995). Integrating and Accessing Heterogeneous Information Sources in TSIMMIS. Proceedings of the AAAI Symposium on Information Gathering, 61-64.

[4] Molina, H.G., Papakonstantinou, G., Quass, D., Rajaraman, A, Sagiv, Y., Ullman, J., Vassalos, V., & Widom, J., (1997). The TSIMMIS approach to mediation: Data models and Languages. Journal of Intelligent Information Systems. 2, 117-132

[5] OMG, (2001). Common Warehouse Meta-model standard, http://www.omg.org/technology/cwm

[6] Poole, J., Chang, D., Tolbert, D., & Mellor, D., (2002). Common Warehouse Meta-model - An Introduction to the Standard for Data Warehouse Integration. Wiley Computer Publishing.

[7] Tan, J., Zaslavsky, A., & Bond, A., (2000). Meta Object Approach to Database Schema Integration. Proceedings of International Symposium on Distributed Objects and Applications. 145-154.