ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 1, January 2013

"Integrating Heterogeneous Data Sources Using XML"

[1] Yogesh R.Rochlani , [2] Prof. A.R. Itkikar

[1]Department of Computer Science & Engineering Sipna COET, SGBAU, Amravati (MH), India
[2]Department of Computer Science & Engineering Sipna COET, SGBAU, Amravati (MH), India

***Abstract*-**Nowadays organizations not only are increasing the data volume, but also they have to work with a large variety of data sources with different types of data. The central problem of information sources integration resides on their heterogeneity on the level of their format, the design and/or the semantic aspect. It would be necessary to use some kind of systems like wrappers or mediators in order to get a common format for the information. In this paper we propose a new system to solve this problem, which uses XML as common format. It describes the system architecture and focuses on the distributed query processing technology in this component. Proposed System allows querying, retrieving data from different sources integrate them and translates the results to XML. The goal of such a system is to intercept the user queries and to find the more adequate data from several heterogeneous resources, to answer the queries of the user, and to turn over the result in a transparent way to the users.

Keywords—Heterogeneous Database Integration, Relational, XML.

I. INTRODUCTION

Modern organizations are suffering numerous changes. They have to manage large volumes of data in addition to data from heterogeneous sources. Mainly when organizations get into the web it has to manage new types of web data such as XML or HTML, for example to exchange information with other organizations or to publish its information on the web. However, the organization information also continues to be stored in the traditional databases such as the relational one.

On the other hand XML is becoming the standard format to exchange information over the internet. The advantages of XML as an exchange model, such as rich expressiveness, clear notation, and extensibility, make it the best candidate for supporting the integrated data model. Tools and infrastructures for data integration are required due to the increasing number of distributed heterogeneous data sources on-line. In this paper, we are going to study how in business, and especially when it is converted to e-business, it is necessary to integrate information from different sources and manage it together. We are going to centre on two main data sources: the conventional relational database systems, because as we have just said most of the business data continues to be stored in them, and XML data, because it is the most extended data format for exchanging and publishing information on the web. The aim would be to integrate these two types of data to obtain at the end, all the necessary data in the same format, in XML.

II. RELATED WORK

Data integration has received significant attention since the early days of databases. In the recent years, there have been several works focusing on heterogeneous information integration. Most of them are based on common mediator architecture [12]. In this architecture, mediators provide a uniform user interface to views of heterogeneous data sources. They resolve queries over global concepts into sub queries over data sources. Mainly, they can be classified into structural approaches and semantic approaches.

In structural approaches, local data sources are assumed as crucial. The integration is done by providing or automatically generating a global unified schema that characterizes the underlying data sources. On the other hand, in *semantic approaches*, integration is obtained by sharing a common ontology among the data sources. According to the mapping direction, the approaches are classified into two categories: global-as-view and local-as-view [13]. In *global-as-view* approaches, each item in the global schema is defined as a view over the source schemas. In *local-as-view* approaches, each item in each source schema is defined as a view over the global schema. The local-as-view approach better supports a dynamic environment, where data sources

All Rights Reserved © 2013 IJARCET

can be added to the data integration system without the need to restructure the global schema.

There are several well-known research projects and prototypes such as Garlic [2], Tsimmis [3], MedMaker [9], and Mix [10] are structural approaches and take a global-as-view approach. A common data model is used, e.g., OEM (Object Exchange Model) in Tsimmis and MedMaker. Mix uses XML as the data model; an XML query language XMAS was developed and used as the view definition language there. DDXMI (for Distributed Database XML Metadata Interface) builds on XML Metadata Interchange. DDXMI is a master file including database information, XML path information (a path for each node starting from the root), and semantic information about XML elements and attributes. A system prototype has been built that generates a tool to do the metadata integration, producing a master DDXMI file, which is then used to generate queries to local databases from master queries. In this approach local sources were designed according to DTD definitions. Therefore, the integration process is started from the DTD parsing that is associated to each source.

Many efforts are being made to develop semantic approaches, based on RDF (Resource Description Framework) and knowledge-based integration [3]. Several ontology languages have been developed for data and knowledge representation to assist data integration from a semantic perspective, such as Ontolingua [1]. F-logic [11] is employed to represent knowledge in the form of a domain map to integrate data sources at the conceptual level. An ontology based approach [5] is one from many other researches which use ontology to create a global schema.

We classify our system as a structural approach and differ from the others by following the local-as-view approach. The XML Schema language is adopted in our work instead of DTD grammar language, which has limited applicability. While only simple cases of heterogeneity conflicts among elements were handled in the paper [2], this work involves more features of XML schema components; we handle more mapping cardinality cases involving attributes in which the core purpose is to provide more information about the elements.

III. ANALYSIS OF PROBLEM

To integrate or link the data stored in heterogeneous data sources, a critical problem includes entity matching, i.e., matching records representing semantically corresponding entities in the real world, across the sources.

Data integration involves schematic conflicts and semantic conflicts:

Schema integration is the process of merging autonomously developed DB schema into a unified, global schema to provide transparency through a unified view. Schema integration has been variously described as a 3, 4 or 5-step process [16] and involves the tasks of pre-integration (schema translation into common data model form), comparison (process of semantic conflict identification), conformance (making conflicts compatible for merging by similar representation), and merging (integrating schemas) including restructuring (refining schema)[16].

Semantic integration, implicit within schema integration, resolves differences in conceptual representation of data by determining equivalence between schema constructs and removing ambiguity among component DBs.

IV. PROPOSED WORK

A. Functional objectives of the system:

Our software tool uses XML as a mediator for integrating and querying disparate heterogeneous information sources. The main objectives of our tool can be defined in the following way:

Query Analysis: It carries out the syntactic analysis in accordance with grammar and semantic analysis in accordance with the referred view or with the query schema.

Query Translation: This phase translates the user's query under the XML query language.

Sub query generation: In this phase our tool divides the xml query in several sub-queries meant for the different sources, according to global Schema.

Translation of the Result to the User's Format: To reformulate the answer to be validated in accordance with the user's query language.

B. Architecture of the System:

This mediator system is the result of a detailed study of the advantages and disadvantages of several existing mediators. The implementation of the core is based on the management technology of the objects

All Rights Reserved © 2013 IJARCET

distributed around the two data models: the relational and the XML model.

The generic architecture of this mediator system is illustrated in Figure 1

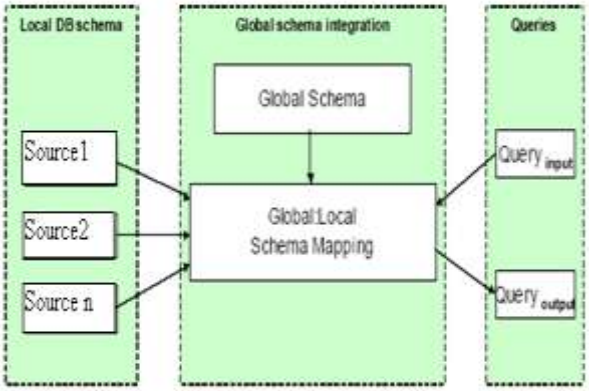

Figure 1. Architecture of the system

Local DB Schema contains the already existing data sources which are to be integrated. Global DB Schema integration involves definition of global schema and mapping between local schema and global schema. Third block in fig.1 includes user system interface which gives query in form of Global Schema as an input and receives the output in the form of XML.

C. Global Schema and Query Processing

In this part, we are interested in the definition of the system's global Schema and the necessary stages to process the users queries and to generate sub-queries to adapt them to the different sources integrated by the system [9]-[11].

Definition of the Global Schema

The Global Schema definition is based essentially on the identification of all the domains which model the whole of the data and services case study. These domains are modeled by a hierarchical structure, where each node represents a domain grouping sub domains defined by the children of this node. Consequently, each node of the Global Schema integration tree is characterized by: a name and a description of the domain, a list of its attributes, a list of the integrated sources, a list of the integrated tools and a list of sub-domains generated by the father domain. The main reports of the Global Schema integration definition is based on the process of successive refinement by specialization starting from a basic federator domain (i.e. Root of the integrating tree). Moreover we suppose that each source represents a view on a sub-domain in the integration tree hierarchical structure (L.A.V Approach). This Global Schema can be described by the following integration tree:

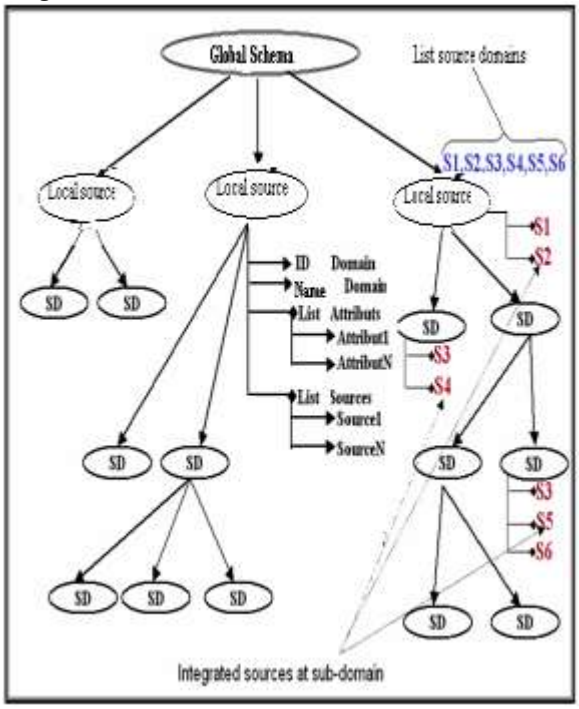

Figure 2: Integration Tree structure.

The improvement of these domains structuring allows to facilitate and optimize, at the same time, the phase of the mediator's query by the users and to generate the queries execution plan. Indeed, the user can easily explore the integration Global Schema tree to determine an optimum list of sources which can be queried by the mediator. Indeed, after having to carry out the rewriting of requests and affecting each sub-request a specific Domain in the tree of Global Schema, the mediator generates a plan of execution preestablishes, following an in-depth course of the tree. The results of execution of each sub-request are stored in the temporary memories associated the domain of the tree. At the time of the customer request evaluation and to generate the finale result required by the customer, these data will be amalgamated starting from the answers partial recorded to the level of the sheets to the root of the tree of Global Schema. Consequently, in the mapping phase only the necessary sources are treated. This structure will also enable us to define an execution plan of sub-queries generated by the mediator. After the rewriting phase of a query, the order of execution starts with sub queries generated for the sources integrated into the low level domains (possibly

All Rights Reserved © 2013 IJARCET

sheets) until the federator domain For the Global Schema integration definition the following basic constraints have been proposed:

1. A source can be integrated by several domains.
2. The list of the sources integrated by a domain is the list of all the sources integrated by all sub domains of this domain.
3. Sub-domains are disjoined: sources can be affected only to one and one sub-domain of the same domain.
4. If two domains (or several) of the same levels (even depth in the integration tree) integrate same sources, the level of integration of this sources moves on the level of the father domain of these domains.

Definition of the Mapping Schema

One of the main problems arising from the data integration consists in carrying out the correspondence between a data source schema and the Global Schema. Generally, it is a question of laying down the rules which make it possible to bind the elements of the Schema of a source to those of the Global Schema (inter-Schema s correspondence). This makes it possible to the mediator to answer the queries of the user which are submitted on the Global Schema.

Correspondences Identification:

When the Global Schema reaches the desired level of conformity, the following stage consists in identifying the common correspondence rules. With each time that is possible, the correspondences are defined in intention. The integration process consists in finding these correspondences between the elements of the sources and those of the Global Schema. These rules form part of the integration process result. In our model, the elements in correspondences can be entities, attributes or many access paths to the attributes and methods signature, etc. These elements are varied according to the sources model (i.e. Relational, XML.). The correspondence elements can be summarized in the following table

	source Element	global Schema Element
Relational	Relation	Entity
	Attribute	Attribute
	Trigger	Service
XML	Entity	Entity
	Attribute	Attribute
	Path	Path

Query Processing:

Firstly, System receives a query formulated in terms of the unified global Schema. The query is decomposed by the rewriter component into sub-queries and addressed to specific data sources. This decomposition is based on source descriptions by global Schema and mapping Schema, which play an important role in sub-queries execution plan optimization. Finally, the sub-queries are sent to the wrappers of the individual sources, which transform them into queries over the sources. The results of these sub queries are sent back to the mediator system. At this point answers of these sub queries are merged and returned to the user. Besides the possibility of making queries, the mediator has no control over the individual sources.

V .CONCLUSION:

The proposed architecture satisfies almost all requirements for a mediator allowing an efficient integration of heterogeneous information systems. Besides the integration of different kinds of data sources it offers now a more flexible way of extending the system. Our Mediation system currently provides following features:

- With such mediation architecture for information systems it is possible to make several information systems with different designs and architectures cooperate.
- Several heterogeneous data sources can be easily integrated, updated or just removed from the system by simply changing the global Schema.
- A large amount of available databases, structured text files and Web Services are supported due to already available wrappers.
- The mapping process is carried out by certain simple mapping actions.

All Rights Reserved © 2013 IJARCET

REFERENCES:

[1] Wiederhold G.: "Intelligent Integration of Information", ACM SIGMOD Conf. On Management of data, pp. 434-437, Washington D.C., USA, May 1993.

[2] Haas L., Kossman D., Wimmers E., Yang J.: "Optimizing Queries across Diverse Data Sources", 23rd Very Large Data Bases, August 1998, Athens, Greece, 1997.

[3] Chawathe S., Garcia-Molina H., Hammer J., Ireland K., Papakonstantinou Y., Ullman J., and Widom J.: "The TSIMMIS Project : Integration of Heterogeneous Information Sources", IPSJ Conference, pp. 7-18, Tokyo, Japan, October 1994.

[4] Fankhauser P., Gardarin G., Lopez M., Muñoz J., Tomasic A.: "Experiences in Federated Databases: From IRO-DB to MIRO-Web", 24rd Very Large Data Bases, pp. 655-658, August 24-27, 1998, New York City, New York, USA, 1998

[5] Cluet S., Delobel C., Siméon J., Smaga K.: "Your Mediators Need Data Conversion", ACM SIGMOD Intl. Conf. on Management of Data, pp. 177-188, Seattle, Washington, USA, 1998.

[6] Christophides V., Cluet S., Siméon J.: "On Wrapping Query Languages and Efficient XML Integration", ACM SIGMOD 2000, pp. 141-152, May 16-18, 2000, Dallas, Texas, USA. SIGMOD Record 29(2) ACM 2000.

[7] Manolescu I., Florescu D., Kossmann D.: "Answering XML Queries over Heterogeneous Data Sources", 27th Very Large Data Bases, pp. 241-250, Roma, Italy, Sept. 2001.

[8] Shanmugasundaram J., Kiernan J., Shekita E., Fan C., Funderburk J.: "Querying XML Views of Relational Data",Proc. Of the 27th International Conference on Very Large Data Bases, pp. 261-270, Roma, Ital., Sept. 2001.

[9] PAPAKONSTANTINOU, Y., GARCIA-MOLINA, H., ULLMAN, J., MedMaker: A Mediation System Based on Declarative Specifications, in: Proc. of the IEEE Int. Conf. on Data Engineering, New Orleans, LA, February 1996, pp. 132-141.

[10] BARU, C., GUPTA, A., LUDASCHER, B, MARCIANO, R., PAPAKONSTANTINU, Y., VELIKHOV, P., and CHU, V., XML-Based Information Mediation with MIX, in: Proc. of the ACM SIGMOD Int. Conf. on Management of Data, 1999, pp. 597-599.

[11] NAM Y., GOGUEN, J., WANG, G., A Metadata Integration Assistant Generator for Heterogeneous Distributed Databases, in: Proc. of the Confederated International Conferences DOA, CoopIS and ODBASE, Irvine CA, October 2002, LNCS 2519, Springer, pp. 1332-1344.

[12] WIEDERHOLD, G., Mediators in the Architecture of Future Information System, in: IEEE Computer Magazine, Vol. 25, No. 3, March 1992, pp. 38-49.

[13] LENZERINI, M., Data Integration: A Theoretical Perspective, in: Proc. of the ACM Symposium on Principles of Database Systems, Madison, Wisconsin, USA, June 2002, pp. 233-246.

[14] MAY, W., A Rule-Based Querying and Updating Language for XML, in: Proc. of the Workshop on Databases and Programming Languages, Springer LNCS 2397, 2001, pp. 165-181.

[15] S. Cluet, C. Delobel, J. Siméon, K. Smaga, "*Your* Mediators Need Data Conversion", ACM SIGMOD Intl. Conf. on Management of Data, pp. 177-188, Seattle, Washington, USA, 1998.

[16] Roberto Berjón Gallinas, Ana María Fermoso García , María José Gil Larrea "USING XML TO INTEGRATE INFORMATION FROM HETEROGENEOUS DATABASES" IADIS International Conference Applied Computing 2006.

[17] Huimin Zhao, Sudha Ram" Entity matching across heterogeneous data sources: An approach based on constrained cascade generalization" Data & Knowledge Engineering 66 (2008).

[18] A.Sathish, A.Noormohamed, S.G.Shahidha Taj" XQUERY PROCESSING IN VERY LARGE XML DATABASES" International Journal of Communications and Engineering Volume 03– No.3, Issue: 03 March 2012

All Rights Reserved © 2013 IJARCET