Data Integration Based WWW with XML and CORBA

Zhengding Lu [1] and Suzhi Zhang [1, 2]

[1] College of Computer science and Technology, Huazhong University of Science and technology, Wuhan, Hubei, China, 430074
zhsuzhi@sohu.com
[2] Department of Computer science and Technology, ZhengZhou Institute of Light Industry, Zhengzhou, Henan, China, 450002
zhsuzhi@sina.com

Abstract. At the present time, WWW is becoming important and potential resources for delivering and sharing information over the world, and more and more Web applications require querying and integrating the data from multiple, distributed, heterogeneous sources. Web data integration is a challenging research project. In this paper, we propose the architecture of Web data integration with XML and CORBA, which is based on our developing project, *Panorama.* The architecture can integrate the data from multiple data sources, including relational database, object-oriented database, HTML and XML documents, and structured text files. In the architecture, we view Web as a huge database, and take CORBA as object model and XML as mediated data model, and use XML-QL to accomplish data query and integration on the Web. We also elaborate and analyze some implementing methods, such as integrator and wrappers corresponding to various data sources.

1 Introduction

With the increasing of application requirements and development of network technology, more and more users expect to access and manipulate the data from multiple data sources[3][4]. WWW is becoming important and potential resources for delivering and sharing information over the world. The resources on the Web involve not only conventional database, such as relational database and object-oriented database, which have well-form data model, but also unstructured and semi-structured data. It is difficult to store and manage all the Web data with conventional database technology, because of its irregularity and various forms of the web data. How to integrate the data from distributed, heterogeneous, and multiple data sources on the Web into an available whole is a full of challenges and urgent work to many application and enterprises[4][6].

In this paper, we propose our architecture of Web data integration with XML and CORBA, which is based on our developing project, *Panorama.* In the architecture, we view Web as a huge database, and take CORBA as object model, and XML as common data model (CDM), and use XML-QL[3] to accomplish data query and integration on the Web.

J.M. Cueva Lovelle et al. (Eds.): ICWE 2003, LNCS 2722, pp. 455–458, 2003.

2 XML Data Management Basis

2.1 Classification of XML Documents

We can look at XML documents in two ways[1]. One is data-centric document, which are regular in structure and homogeneous in content. The other is document-centric document, which is more irregular and data are heterogeneous. Specially, we consider a data-centric document to be a database, called XML database, In which the structured data is called XML data, and its conformed DTD is regarded as schema in this paper[2].

2.2 XML Data Model

XML data are similar to semi-structured data. We can say XML data are semi-structured data on the Web. XML data model is a graph/tree model, and divided into an unordered data model and an ordered data model [3].

2.3 Querying and Transforming XML Data

XML data is fundamentally different from conventional relational and object-oriented data, and therefore, neither SQL nor OQL is appropriate for XML data. XML-QL is a kind of querying language for XML developed by AT&T lab, which can be used for data extraction, transformation and integration.

Due to take XML data model as Common Data Model (CDM) in our architecture, it arose the transformation to XML data model from others. These transformations are implemented by wrappers respectively in our architecture.

3 System Architecture

We proposed our architecture of Web data integration based on the discussion and analysis above, depicted in Fig. 1.

In the architecture, we take XML data model as mediated model and XML-QL as query language to complete data query, schema transformation and data integration against various data sources on the Web. The main program modules involve: *Application and Visual User Interface, XML-QL Querying Processor, XML Data Integrator, Metadata Dictionary* and *Wrappers* corresponding to various data sources.

4 Implementation and Method

4.1 Implementation of Integrator

Integrator is used to combine the multiple mediated results (XML data) which come from the wrappers of local data sources into a whole in the global schema, which is formed a virtual database. Because the inputs of the integrator are all XML data, we

can implement integrator by XML-QL. Filter function in integrator is implemented by regular-path expression, join operator in XML-QL.

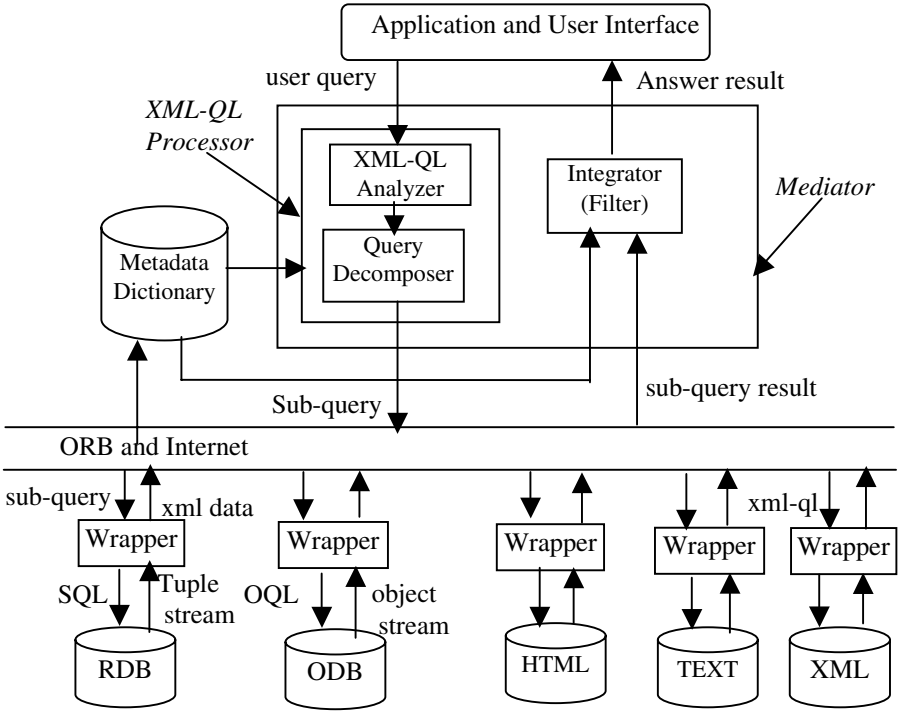

Fig. 1. Architecture of Integrated Web Data

4.2 Metadata Dictionary

Metadata management is important for the data integration of heterogeneous data sources because each source has its own interfaces and attributes. It is necessary to know the interfaces and attributes of local sources when processing a query and collecting query results. In the architecture, metadata dictionary is modeled by the DTD for data representation in order to collaborate with mediator and wrappers.

4.3 Implementation of Wrappers

Wrappers accomplish mainly the transformations from global sub-query (XML-QL) to local data query (such as SQL or OQL), and the transformations from the local query results to mediated data (XML data). There are different wrappers against various data sources, so their implementing methods are also different. (1)XML data source: only need to complete the transformation from XML data to global schema, mainly process the problem of schema integration. (2) HTML data source and structured text are semi-structured and unstructured data sources. The wrapper mainly extracts structured data (XML data) from these files. (3) To speed up the development of RDB and ODB wrappers, it is far more preferable to have the DBMS provider

rather than application developer implement data transformation. The wrapper receives XML-QL query and translates it into SQL or OQL in order to query local DBMS, and query result in the form of XML data return to integrator directly.

5 Conclusion

It is significant objective for researcher to integrate data on the Web. To achieve the objective completely, there is a lot of hard work required to be resolved. In the paper, we propose our architecture of Web data integration with XML and CORBA. In the architecture, we take CORBA as object model, XML as data model, and XML-QL as query language. There are many advantages in the architecture to integrate multiple heterogeneous data on the Web.

(1) We make use of CORBA technology in the architecture. CORBA is an open system model that supports communication between the software components in distributed environment as well as locating data sources dynamically. CORBA offers an IIOP for interoperating with other object-oriented model.

(2) XML data model as CDM and XML-QL as query and integration language. XML data model as CDM is not only consistent with data characteristics on the Web, but also resolves the problem of selecting CDM.

(3) Metadata dictionary modeled by DTD is consistent with data model in mediator. Thus, it is benefit to integrator and XML-QL processor for cooperating with metadata dictionary.

(4) As possible as making use of existing technologies. Specially, to conventional proprietary database, we can make use of the transformation functions provided in the DBMS. It facilitates the design of the wrapper.

The architecture is a straightforward and practicable, but there exists still some embarrassment needed to be overcome. We are going for it in order to make it more robust and perfect in *Panorama* project.

References

1. Bertino,E., Catania,B.: Integrating XML and Databases. IEEE Internet Computing. Julu-August(2001) 84–88
2. Salminen,A., Tompa,F. W.: System Desiderata for XML Databases. Proceedings of the 27th VLDB Conference, Roma, Italy(2001)
3. Deutsch, A., Fernandez, M., Florescu, D. et al.: A query language for XML. Computer Networks. 31 (1999) 1155–1169
4. Chang, Y. S., Ho, M. H., Yuan, S. M.: A unified interface for integrating information retrieval. Computer Standards & Interfaces. 23 (2001) 325–340
5. Garcia-Molina, H., Hammer, J. ,Ireland, K.et al.: Integrating and accessing heterogeneous information sources in TSIMMIS. Proceedings of the AAAI Symposium on Information Gathering, Stanford, California, USA(1995) 61–64
6. Bouguettaya, A., Benatallah, B., Ouzzani, M., Hendra, L.: WEBFINDIT: An Architecture and System for Querying Web Databases. IEEE Internet Computing. July-August(1999) 30–41