XML Warehousing and OLAP

Hadj Mahboubi
University of Lyon (ERIC Lyon 2), France

Marouane Hachicha
University of Lyon (ERIC Lyon 2), France

Jérôme Darmont
University of Lyon (ERIC Lyon 2), France

INTRODUCTION

With the eXtensible Markup Language (XML) becoming a standard for representing business data (Beyer et al., 2005), a new trend toward XML data warehousing has been emerging for a couple of years, as well as efforts for extending the XQuery language with near On-Line Analytical Processing (OLAP) capabilities (grouping, aggregation, etc.). Though this is not an easy task, these new approaches, techniques and architectures aim at taking specificities of XML into account (e.g., heterogeneous number and order of dimensions or complex measures in facts, ragged dimension hierarchies…) that would be intricate to handle in a relational environment.

The aim of this article is to present an overview of the major XML warehousing approaches from the literature, as well as the existing approaches for performing OLAP analyses over XML data (which is termed XML-OLAP or XOLAP; Wang et al., 2005). We also discuss the issues and future trends in this area and illustrate this topic by presenting the design of a unified, XML data warehouse architecture and a set of XOLAP operators expressed in an XML algebra.

BACKGROUND

XML warehousing research may be subdivided into three families. The first family focuses on Web data integration for decision-support purposes. However, actual XML warehouse models are not very elaborate. The second family of approaches is explicitly based on classical warehouse logical models (star-like schemas). The third family we identify relates to document warehousing. In addition, recent efforts aim at performing OLAP analyses over XML data.

XML Web Warehouses

The objective of these approaches is to gather XML Web sources and integrate them into a data warehouse. For instance, Xyleme (2001) is a dynamic warehouse for XML data from the Web that supports query evaluation, change control and data integration. No particular warehouse model is proposed, though.

Golfarelli et al. (2001) propose a semi-automatic approach for building a data mart's conceptual schema from XML sources. The authors show how multidimensional design may be carried out starting directly from XML sources and propose an algorithm for correctly inferring the information needed for data warehousing.

Finally, Vrdoljak et al. (2003) introduce the design of a Web warehouse that originates from XML Schemas describing operational sources. This method consists in preprocessing XML Schemas, in creating and transforming the schema graph, in selecting facts and in creating a logical schema that validates a data warehouse.

XML Data Warehouses

In his XML-star schema, Pokorný (2002) models a star schema in XML by defining dimension hierarchies as sets of logically connected collections of XML data, and facts as XML data elements.

Hümmer et al. (2003) propose a family of templates enabling the description of a multidimensional structure for integrating several data warehouses into a virtual

or federated warehouse. These templates, collectively named XCube, consist of three kinds of XML documents with respect to specific schemas: XCubeSchema stores metadata; XCubeDimension describes dimensions and their hierarchy levels; and XCubeFact stores facts, i.e., measures and the corresponding dimensions.

Rusu et al. (2005) propose a methodology, based on the XQuery technology, for building XML data warehouses, which covers processes such as data cleaning, summarization, intermediating XML documents, updating/linking existing documents and creating fact tables. Facts and dimensions are represented by XML documents built with XQueries.

Park et al. (2005) introduce an XML warehousing framework where every fact and dimension is stored as an XML document. The proposed model features a single repository of XML documents for facts and multiple repositories for dimensions (one per dimension).

Eventually, Boussaïd et al. (2006) propose an XML-based methodology, X-Warehousing, for warehousing complex data (Darmont et al., 2005). They use XML Schema as a modeling language to represent users' analysis needs, which are compared to complex data stored in heterogeneous XML sources. Information needed for building an XML cube is then extracted from these sources.

XML DOCUMENT WAREHOUSES

Baril and Bellahsène (2003) envisage XML data warehouses as collections of materialized views represented by XML documents. Views allow filtering and restructuring XML sources, and provide a mediated schema that constitutes a uniform interface for querying the XML data warehouse. Following this approach, the authors have developed the DAWAX system.

Nassis et al. (2005) propose a conceptual approach for designing and building an XML repository, named xFACT. They exploit object-oriented concepts and propose to select dimensions based on user requirements. To enhance the XML data warehouse's expressiveness, these dimensions are represented by XML virtual views. In this approach, the authors assume that all dimensions are part of fact data and that each fact is described in a single XML document.

Rajugan et al. (2005) also propose a view-driven approach for modeling and designing an XML fact repository, named GxFact. GxFact gathers xFACTs (distributed XML warehouses and datamarts) in a global company setting. The authors also provide three design strategies for building and managing GxFact to model further hierarchical dimensions and/or global document warehouses.

Finally, Zhang et al. (2005) propose an approach to materialize XML data warehouses based on frequent query patterns discovered from historical queries. The authors apply a hierarchical clustering technique to merge queries and therefore build the warehouse.

OLAP ANALYSES OVER XML DATA

Chronologically, the first proposals for performing OLAP analyses over XML data mainly rely on the power of relational implementations of OLAP, while more recent research directly relates to XOLAP.

Jensen et al. (2001) propose an integration architecture and a multidimensional UML model for relational and XML data. They also discuss the design of XML databases supporting OLAP analyses. The output of this process is a unified relational representation of data, which are queried with the OQL language. Niemi et al. (2002) follow the same logic, but propose a dedicated language named MDX. In both these approaches, XML data are mapped into a relational database and exploited with relational query languages. Hence, no XML-specific OLAP operator is defined.

Pedersen et al. (2004) also advocate for federating XML data and existing OLAP cubes, but in addition, they propose an algebra composed of three operators. The most fundamental operator in an OLAP-XML federation, decoration, attaches a new dimension to a cube with respect to linked XML elements. Selection and generalized projection help filter and aggregate fact measures, respectively. They more or less correspond to the classical OLAP slice and dice operators. These three operators are implemented with an extension of the SQL_M language, SQL_{XM}, which helps associate XPath queries to SQL_M queries. SQL_M is itself an extension of SQL for processing multidimensional data.

Eventually, Park et al. (2005) propose an OLAP framework for XML documents called XML-OLAP and introduce the notion of XML cube (XQ-Cube). A specific multidimensional language (XML-MDX) is applied on XQ-Cubes. Wang et al. (2005) also propose a general aggregation operator for XML, GXaggrega-