Chapter 9
Multidimensional Data Analysis Based on Links: Models and Languages

Paulo Caetano da Silva
Salvador University, Brazil

ABSTRACT

Analytical processing (OLAP) tools typically only deal with relational data. Hence, the analytical processing systems on XML data do not have all the functionality provided by OLAP tools to traditional data (i.e. relational). In addition, current commercial and academic OLAP tools do not process XML data that contains XLink. Therefore, there is a need to develop a solution for OLAP systems in order to assist in the strategic analysis of the organizational data represented in XML format. Aiming at overcoming this issue, this chapter proposes an analytical system composed by LMDQL (Link-Based Multidimensional Query Language), an analytical query language; XLDM (XLink Data Metamodel), a metamodel given to model cubes of XML documents with XLink and to deal with syntactic, semantic, and structural heterogeneities commonly found in XML documents; and XLPath (XLink Path Language), a navigation language for XML documents connected by XLink. As current W3C query languages for navigating in XML documents do not support XLink, XLPath is discussed in this chapter to provide features for the LMDQL query processing and a prototype system enabling OLAP queries over XML documents linked by XLink and XML schema. This prototype includes a driver, named sql2xquery, which performs the mapping of SQL queries into XQuery in a relational OLAP server. In order to validate the proposed system, a case study and its performance evaluation are presented to analyze the impact of analytical processing over XML/XLink documents.

INTRODUCTION

The current software applications usually have to deal with multiple sources and data formats. To minimize this problem, XML - eXtensible Markup Language (XML, 1998) was adopted as a way to integrate the data into a standard format. The use of XML as an alternative to integration of heterogeneous data sources has become this technology a de facto standard for data exchange on the Internet. XML documents are a rich source of information for organizational decision mak-
Multidimensional Data Analysis Based on Links

Similarly, the use of Data Warehouses (DW) (Kimball, 2002) and OLAP (On-Line Analytical Processing) tools (Chaudhuri, 1997) allows the identification of tendencies and standards, in order to conduct better strategic decisions for companies businesses. However, the use of these technologies, together, is still in development process.

In XML, it is possible to represent information semantically similar in different ways. This leads to three kinds of data heterogeneity: (i) semantic, where similar information is represented through different names, e.g. enterprise and company, or dissimilar information through equal names, e.g. virus in the informatics field and in the health field; (ii) syntactic, where the semantically equal content is represented in several ways. For example, in different languages or in diverse measure units, e.g. meters and feet; and (iii) structural, in which data is organized in different structures, e.g. in different kinds of hierarchies, in attributes, or in elements (Näppilä, 2008). This representation flexibility is important, however, it makes the use of OLAP concepts in XML data a complex task. Applications and technologies, derived from XML, use XLink (XML Linking Language) (XLink, 2001) as an alternative for representing the information semantic and structure, expressing relationships between concepts. An example of how the data semantic is represented using XLink is XBRL (eXtensible Business Reporting Language) (Hernández-Ros, 2006), an international standard for representing financial reports that uses extended links for modeling financial concepts. A problem that occurs when processing documents, which have XLink and correspond to chains of links, is that the W3C (World Wide Web Consortium) available query languages (i.e. XQuery (XQuery, 2007) and XPath (XPath, 2007)) do not provide support for navigating on them. Although XPath has been widely adopted as query standard in XML documents, it does not provide such navigation functionality. Several proposals have been developed for performing the analytical queries (OLAP) over XML data (Beyer, 2005; Bordawekar, 2005; Näppilä, 2008; Wang, 2007; Jian, 2007). However, these proposals do not take the use of XLink in XML documents into account.

In conventional OLAP server, the data cube is mapped from a relational DW. Queries on the data cube, which are usually expressed through MDX - Multidimensional Expressions (MDX, 2008), an industry standard, are converted into SQL queries that are executed directly in relational DBMS (Databases Management System). However, for information extraction in XML documents, OLAP solutions or for transaction processing (OLTP - Online Transaction Processing) are still poorly developed.

This chapter presents an analytical processing system for XML documents supported by XLink. This system is based on XLDM (XML-based Multidimensional Metamodel), a metamodel suitable for solving XML heterogeneities, XLPath (Silva, 2010), a navigational language, an XPath extension for navigation over links, LMDQL (Link-based and Multidimensional Query Language) (Silva, 2009b), a multidimensional query language, and a process to convert SQL into XQuery queries, implemented by the sql2xquery driver. With the purpose of providing analytical processing of XML data, the mondrain OLAP server (Mondrian, 2008) has been extended by incorporating this driver. The choice of this tool is due to the fact of being an open source OLAP server used in various solutions for Business Intelligence (BI) and be extended and developed in Java, which makes it independent of computer platform. Another reason for choosing this server is that it is compatible with MDX, an industry standard for OLAP queries.

From the analysis of the related proposals discussed in this article, the existence of a query language to XML data, which allows the following, has not been verified: (i) uses a collection of XML documents; (ii) considers the existence of linkbases, as groups of links and source of information; (iii) is based on a data model, defined by XML Schema and XLink, to solve heterogeneity conflicts in XML. For these reasons, the develop-