2012

Analytical Processing Over XML and XLink

INTERNATIONAL JOURNAL OF DATA WAREHOUSING AND MINING, HERSHEY, v.8, n.1, pp.52-92, JAN-MAR, 2012
http://www.producao.usp.br/handle/BDPI/42766

Analytical Processing Over XML and XLink

Paulo Caetano da Silva, Salvador University and Federal University of Pernambuco, Brazil

Valéria Cesário Times, Federal University of Pernambuco, Brazil

Ricardo Rodrigues Ciferri, Federal University of São Carlos, Brazil

Cristina Dutra de Aguiar Ciferri, University of São Paulo, Brazil

ABSTRACT

Current commercial and academic OLAP tools do not process XML data that contains XLink. Aiming at overcoming this issue, this paper proposes an analytical system composed by LMDQL, an analytical query language. Also, the XLDM metamodel is given to model cubes of XML documents with XLink and to deal with syntactic, semantic and structural heterogeneities commonly found in XML documents. As current W3C query languages for navigating in XML documents do not support XLink, XLPath is discussed in this article to provide features for the LMDQL query processing. A prototype system enabling the analytical processing of XML documents that use XLink is also detailed. This prototype includes a driver, named sql2xquery, which performs the mapping of SQL queries into XQuery. 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.

Keywords: *eXtensible Markup Language (XML), Link-Based and Multidimensional Query Language (LMDQL), On-Line Analytical Processing (OLAP), XLink-Based Multidimensional Metamodel (XLDM), XLPath, XML Linking Language (XLink)*

INTRODUCTION

XML (eXtensible Markup Language) documents are a rich source of information for organizational decision making. 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

DOI: 10.4018/jdwm.2012010103

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) (DeRose, Maler, & Orchard, 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., Boag, Chamberlin, Fernández, Florescu, Robie, & Siméon, 2007; Berglund, Boag, Chamberlin, Fernández, Kay, Robie, & Siméon, 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; XBRL International, 2006). However, these proposals do not take the use of XLink in XML documents into account.

This paper presents an analytical processing system for XML documents supported by XLink. This system is based on XLDM (XLink-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, 2009), a multidimensional query language, and a process to convert SQL into XQuery queries, implemented by the *sql2xquery* driver. 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 development of an OLAP system for XML, which uses XLink, consists in the motivation to perform this research.

In this next section, some of the most important OLAP proposals for XML data are discussed. The metamodel XLDM, XLPath language and LMDQL query language specification are then presented. We present the analytical system for XML and XLink is detailed together with a driver to convert SQL queries into XQuery. We approach a case study and discuss performance results. Conclusions and future work are given last.

OLAP FOR XML DOCUMENTS

In this section, some proposals for analytical processing in XML data are discussed. In order to make easier the selection of an appropriated model when designing applications that combine OLAP and XML technologies, Ravat et al. (2010) discuss different proposals that use these technologies. This discussion covers since the use of XML data incorporated in traditional data warehouses until those that use XML as a complete solution for warehousing processes. In this proposal, Ravat et al. (2010) consider two types of documents, those based on data and those based on textual information (data-centric XML documents and document-centric XML documents, respectively). The documents based on data have a regular structure, as occurs in relational data. The text centered documents, however, are less structured than the first ones. They are rich text documents and are not adapted to the information exchange among applications. Based on this document classification, the distinction between two types of data warehouses based on XML is made: (i)