Chapter 13

Automatic Mapping of XML Documents into Relational Database: Introduction

Ibrahim Dweib
Sultan Qaboos University, Oman

Joan Lu
University of Huddersfield, UK

ABSTRACT

Extensible Markup Language (XML) nowadays is one of the most important standard media used for exchanging and representing data through the Internet. Storing, updating, and retrieving the huge amount of web services data such as XML is an attractive area of research for researchers and database vendors. In this chapter, the authors propose and develop a new mapping model, called MAXDOR, for storing, rebuilding, updating, and querying XML documents using a relational database without making use of any XML schemas in the mapping process. The model addressed the problem of solving the structural hole between ordered hierarchical XML and unordered tabular relational database to enable us to use relational database systems for storing, updating, and querying XML data. A multiple link list is used to maintain XML document structure, manage the process of updating document contents, and retrieve document contents efficiently. Experiments are done to evaluate MAXDOR model. MAXDOR will be compared with other well-known models available in the literature (Tatarinov et al., 2002) and (Torsten et al., 2004) using total expected value of rebuilding XML document execution time and insertion of token execution time.

DOI: 10.4018/978-1-4666-1975-3.ch013
INTRODUCTION

The World Wide Web (WWW) nowadays is an important medium used by many people for many activities in their daily life (i.e.; e-management, e-learning, e-mail, e-library and e-business). Many enterprises are working together using XML technologies for exchanging their web services data. Exchanging, sorting, updating and retrieving these huge data has become a source of concern for researchers and database vendors.

At present, storing and retrieving of XML documents can be done using mainly three approaches, i.e., native XML database (Jagadish et al., 2003; M. Grinev et al., 2004), Object Oriented Database (Chung and Jesurajaiah, 2005) and Relational Database (Zhang and Tompa, 2004a; Shanmugasundaram et al., 1999); (Fujimoto et al., 2005; O’Neil et al., 2004) (Tan et al., 2005) (Leonardi and Bhowmick, 2005; Atay, 2006; Atay et al., 2007a; Min et al., 2008; Yun and Chung, 2008; Ahlgren and Collander, 2009).

The most important factor in choosing the target database is the type of XML documents to be stored, data-centric (e.g., bank transaction, airlines transactions) or document-centric (e.g., emails, books, manual).

Using a hybrid approach of relational database to store and retrieve data and XML to exchange and represent it. This will solve most of the data issues of integrity, multi-user access, retrieving, exchanging, concurrency control, crash recovery, indexing, security, storing semi-structure data, and reliability. The previous studies of this approach can also be studied. These are: Loss of information, difficulties in updating its contents and difficulties in rebuilding of original document. The mapping techniques of this approach can generally be classified into two tracks: Schemaless-centric technique and schema-centric (Dweib et al., 2008). Schemaless-centric technique is used to make use of XML document structure to manage mapping process (Zhang & Tompa, 2004; Yoshikawa et al., 2001; Jiang et al., 2002; Tatarinov et al., 2002; Soltan and Rahgozar, 2006). In schema-centric, XML schema information is used to develop a relational storage for XML documents (Shanmugasundaram et al., 1999; Atay et al., 2005; Yahia et al., 2004; Lee et al., 2006; Knudsen et al., 2005; Fujimoto et al., 2005; Xing et al., 2007). Unfortunately, relational storages constructed from schema-centric approach need database reconstruction as any change in the XML schema is very expensive. Each approach introduced some solutions for the mapping process but failed to solve others.

In this thesis we will concentrate on a new approach for mapping XML documents into relational database which is called MAXDOR (i.e. Mapping XML Document into Relational database). The model does not make use of any XML schemas to manage mapping process. In this model, the document structure and document contents are stored in relational database tables. It uses multi-links to reserve document structure and elements relations within the document as parent-child, ancestor-descendant, left-sibling and right-sibling. The use of multi-links will make the insertion process cost for new elements and attributes anywhere in the document close to constant value, since there is no need to relabel the elements and the attributes following the inserted element or attribute. Other models (Tatarinov et al., 2002) (Torsten et al., 2004) which consider the element or attribute label as an identifier to reserve document structure, the cost of insertion in this case will vary depending on the position of insertion, since relabeling is needed after each insertion to maintain the document order.

The proposed model uses a process of four steps: (1) Mapping XML document into relational database. To achieve this objective, a fixed relational schema is presented and used to maintain document contents relations and manage the contents. (2) Building XML document from relational database without a need to the original document.