UNIVERSITI PUTRA MALAYSIA

GENERATING NESTED XML DOCUMENT WITH DTD FROM RELATIONAL VIEWS

MOHAMMED NASSER AHMED

FSKTM 2008 12

GENERATING NESTED XML DOCUMENT WITH DTD FROM RELATIONAL VIEWS

By

MOHAMMED NASSER AHMED

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALYSIA

2008

GENERATING NESTED XML DOCUMENTS WITH DTD FROM RELATIONAL VIEWS

By

MOHAMMED NASSER AHMED

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

September 2008

Dedicated to my Parents; Nasser and Mokhlas,

to my wife and

my kids; Zinab, Nada, Aidah , Yasir, and Sarah

to my family.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

GENERATING NESTED XML DOCUMENTS WITH DTD FROM RELATIONAL VIEWS

By

MOHAMMED NASSER AHMED

September 2008

Chairman : Hamidah Ibrahim, PhD

Faculty : Computer Science and Information Technology

Converting relational database into XML is increasing daily for publishing and exchanging data on the web. Most of the current approaches and tools for generating XML documents from relational database generate flat XML documents that contain data redundancy which leads to produce a massive data on the web. Other approaches assume that the relational database for generating nested XML documents is normalized. In addition, these approaches have problem that lies in the difficult of how to specify the parent elements from the children elements in the nested XML document. Moreover, most of the current approaches and tools do not generate nested XML documents automatically. They require the user to specify the constraints and the schema of the target document.

This research proposes an approach to automatically generate nested XML documents from flat relational database views that are unnormalized. The research aims to reduce

data redundancy and storage sizes for the generated XML documents. The proposed approach consists of three steps. The first step is converting flat relational view into nested relational view. The second is generating DTD from the nested relational view. The third is generating nested XML document from the nested relational view.

The proposed approach is evaluated and compared to other approaches such as NeT, CoT, and Cost-Based and tools such as Allora, Altova, and DbToXml with respect to two measurements: data redundancy and storage size of the document. The first measurement includes several parameters that are number of data values, elements, attributes, and tags.

Based on the results of comparing the proposed approach to several other approaches and tools, the proposed approach is more efficient for reducing data redundancy and storage size of XML documents. It can reduce data redundancy and storage size by approximately 50% and 55%, respectively.

Abstrak tesis dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

MENJANA DOKUMEN BERSARANG XML DENGAN DTD DARIPADA PANGKALAN

Oleh

MOHAMMAD NASSER AHMED

Jun 2008

Pengerusi : Hamidah Ibrahim, PhD

Fakulti : Sains Komputer dan Teknologi Maklumat

Penukaran pangkalan data hubungan kepada XML meningkat setiap hari untuk penerbitan dan penukaran data dalam web. Kebanyakan pendekatan dan produk semasa untuk menjana dokumen XML daripada pangkalan data hubungan, menjana dokumen XML tidak bersarang yang mengandungi lewahan data yang membawa kepada penghasilan sejumlah data yang banyak di web. Pendekatan lain mengandai bahawa pangkalan data hubungan telah dinormalkan. Tambahan pula, pendekatan ini mempunyai masalah dalam menentukan bagaimana menentukan elemen induk daripada elemen anak? Seterusnya, kebanyakan pendekatan dan produk semasa tidak menjana dokumen XML secara automatik. Mereka memerlukan pengguna untuk menentukan kengkangan dan skema bagi target dokumen.

Penyelidikan ini mencadangkan satu pendekatan untuk menjana secara automatik dokumen XML bersarang daripada gambaran pangkalan data hubungan yang tidak bersarang yang belum dinormalkan. Matlamat penyelidikan adalah untuk mengurangkan lewahan data dan saiz storan bagi dokumen XML yang dijana. Pendekatan yang

dicadangkan ini terdiri daripada tiga langkah. Langkah pertama adalah menukar pangkalan data hubungan yang tidak bersarang kepada gambaran hubungan bersarang. Kedua adalah menjana DTD daripada pangkalan data hubungan bersarang. Ketiga, menjana dokumen bersarang XML daripada pangkalan data hubungan bersarang.

Pendekatan yang dicadangkan dinilai dan dibandingkan dengan pendekatan lain seperti NeT, CoT dan Cost-Based dan alat seperti Allora, Altova dan DbToXml berdasarkan dua ukuran: lewahan data dan saiz storan dokumen. Ukuran pertama melibatkan beberapa parameter iaitu bilangan nilai data, elemen, attribut, dan tag. Berdasarkan kepada keputusan membandingkan pendekatan yang dicadangkan ini dengan beberapa pendekatan yang lain dan alat pendekatan yang dicadangkan lebih efisien bagi mengurangkan lewahan data dan saiz storan bagi dokumen XML. Ia boleh mengurangkan lewahan data dan saiz storan masing-masing hampir 50% dan 55%.

ACKNOWLEDGEMENTS

In the name of ALLAH, the Beneficent, the Compassionate and who gave me strength, patience, and motivation to complete this research work. I would like to express my deep appreciation to my advisor, Associate Professor Dr. Hamidah Ibrahim. She guided me not only on numerous research and technique problems, but also on how to do research. Her profound knowledge, her dedication to research and science, her kindness and optimistic attitude toward the world, will always inspire me. I am also very thankful to my committee members: Associate Professor Dr. Ali Mamat and Associate Professor Dr. Md Nasir Sulaiman continuous help on my research and taking the time to guide me through my thesis. I would like to thank all my colleagues especially Moneer Al-qbatee, Ali alsharafee, Abdurahman A.Mothana, Yahya Al-wadhaf, and Mohammed Radee for helping me. Finally, I am grateful to my family from the bottom of my heart for their unconditional support. Especially, I would like to thank my father, my mother, my wife and my daughters for the love and encouragement.

A great thanks to the Faculty of Computer Science and Information Technology, the university library and Universiti Putra Malaysia that provided the working environment for performing this work.

I certify that an Examination Committee has met on …… 2007 to conduct the final examination of Mohammed Nasser Ahmed on his Doctor of Philosophy thesis entitles "Generating Nested XML Document with DTD from Relational View" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Professor
(External Examiner)

_____

HASANAH MOHD. GHAZALI, PhD
Professor / Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Hamidah Ibrahim, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Ali Mamat, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Md. Nasir Sulaiman, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

_____
AINI IDERIS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 13 November 2008

DECLARATION

I declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously and is not concurrently submitted for any other degree at UPM or at any other institution.

_____

MOHAMMED NASSER AHMED

Date:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

.43

LIST OF ABBREVIATIONS

ATGs	Attribute Translation Grammars
ConvRel	Conversion to XML nested Structure
CoT	Constraints-based Translation
DOM	Data Object Model
DTD	Document Type Definition
FD	Functional Dependency
FRV	Flat Relational View
FT	Flat-based Translation
HTML	Hiper Text Markup Langauge
IND	Inclusion Dependency
MVD	Multi-valued Dependency
NeT	Nesting-based Translation
NPJ	Nest Project Join
NRV	Nested Relational View
ODBC	Open Database Connectivity
PA	The proposed Approach
ROX	Relation Over XML
RXL	Relational to XML transformation Language
SGML	Standard Generlization Markup Language
W3C	World Wide Web Constrium
XML	eXtensible Markup Language
XNF	XML Normal Form
XSD	XML Schema Definition

CHAPTER 1

INTRODUCTION

1.1 Overview

The relational database model and eXtensible Markup Language (XML) model are closely related in most web applications. However, these two models have different structure of schema. The relational database model is based on a two dimensional table that has neither hierarchy nor significant order. XML is based on trees in which order is significant. The hierarchy and sequence features are not used to model information in relational database model but for XML, these features are the main ways to represent information in XML.

XML is useful because of its flexible structure where it closely matches the structure used to display the same information in HTML. Most of data on the web comes from relational database that needs to be converted to XML. Converting relational databases into XML documents includes translating relational schema into XML schemas (DTD or XML Schema) or translating query results (views) with schema into XML documents. Translating query results (views) into XML documents has many problems as follows: view is considered as single relation, non-normalized, and it may contain over millions records that may have duplicates values. If the relational view is converted into a flat XML document, then there will be an amount of useless space because of data

redundancy. To reduce the huge space and data redundancy of XML documents, the view is normalized by the nested relational model. There is a significant advantage in nesting a relation formed by using functional and multi-valued dependencies. The nested relational view can give a great benefit for reducing the data redundancy when it contains at least two groups of attributes from tables that are connected using inclusion dependencies (Lawrence and Ramon, 2005).

1.2 Problem Statement

Although XML has become the prime standard for data exchange on the web, and is increasingly used to represent data currently resides in databases, XML documents have a little semantic and take a large space to store data. Mapping relational database views into XML documents occurs frequently. Thus, there will be massive data on the web because of data redundancy that takes up unnecessary storage, inflate data transfer cost, and lead to update anomalies (Cong and H. Jagadish, 2006). Furthermore, such data redundancies can lead to rendering the database inconsistent (Cong and H. Jagadish, 2008).

Most of the current approaches and tools for generating XML documents from relational database such as Allora (EBIZQ, 2005 and Ronald, 2007), Altova (ALTOVA, 2007 and Ronald, 2007) and DB2XML (Ronald, 2007) generate flat XML documents. The flat XML document contains data redundancy that leads to increase in the cost of storage size. Unlike, Nested XML document has less data redundancy and small storage size. Furthermore, it represents the nature of XML model which is hierarchy. There are some

approaches such as CoT (Dongwon and Murali, 2002), ConvRel (Angela C. et al, 2004), ROX (Alan H. et al, 2004), and Cost-Based (Lawrence and Ramon, 2005) which assume that the relational database tables for generating nested XML documents are normalized with appropriate degree (typically the third degree). However, these approaches required the user to specify the constraints and define the target XML document manually that leads to consume time more than the automatic mapping. Furthermore, the approaches that deal with normalized relational database tables as input to generate nested XML documents have problem to identify and determine which elements are the parents and which elements are the children in the nested XML document. The posed question is how nested XML documents with their schemas can be automatically generated from unnormalized flat relational views with redundancy reduced and less storage size?

1.3 Objectives

The main objective of this research is to propose an automatic approach for generating nested XML documents from relational database views (unnormalized) with less data redundancy and less storage size. There are several sub-objectives derived from the main objective as follows:

1. To propose an approach for converting relational database view into nested relational view.

2. To propose an algorithm for automatically generating nested document type definition (DTD) from nested relational view.