Conceptual and Systematic Design Approach for XML Document Warehouses

Vicky Nassis, La Trobe University, Australia
Rajugan Rajagopalapillai, University of Technology, Australia
Tharam S. Dillon, University of Technology, Australia
Wenny Rahayu, La Trobe University, Australia

ABSTRACT

EXtensible Markup Language (XML) has emerged as the dominant standard in describing and exchanging data among heterogeneous data sources. The increasing presence of large volumes of data appearing creates the need to investigate XML Document Warehouses as a means of handling the data. In this paper our focus is twofold. First we utilise Object Oriented (OO) concepts to develop and propose a conceptual design formalism to build meaningful XML Document Warehouses (XDW). This includes: (1) XML (warehouse) repository (xFACT) using OO concepts followed by the transformation of XML Schema constructs and (2) Conceptual Virtual Dimensions (VDims) using Conceptual views (Rajugan, Chang, Dillon, & Feng, 2003, 2004). Secondly we address several important outstanding issues related to our proposed design of an XML Document Warehouse. Specifically we note that the xFACT portion is now a complex structure, involving several entities and relationships as opposed to being a simple FACT table as was the case in relational data warehouses, and the notion of Virtual Dimensions (VDims) has considerably greater complexity.

Keywords: conceptual design; data modeling; data warehouse; database conceptual design; database design; database logical design; database requirements analysis; database views; information engineering; information requirements analysis; structural modeling; UML; XML

INTRODUCTION

Data Warehousing (DW) has been an approach adopted for handling large volumes of historical data for detailed analysis and management support. Transactional data in different databases is cleaned, aligned and combined to produce data warehouses. Since its introduction in 1996, eXtensible Markup Language (XML) has become the de facto standard for storing and manipulating self-describing information, which creates vocabularies in assisting information exchange between heterogeneous data sources over the Web (Pokorny, 2002). Amongst the purposes of
XML is to carry out electronic document handling, electronic storage, retrieval and exchange. It is envisaged that XML will also be used for logically encoding documents for many domains. Hence, it is likely that a large number of XML documents will populate the would-be repository and include several disparate transactional databases.

The need for managing large amounts of XML document data raises the necessity to explore the data warehouse approach through the use of XML document marts and XML document warehouses.

Since the introduction of dimensional modeling (which revolves around facts and dimensions), several design techniques have been proposed to capture multidimensional data (MD) at the conceptual level. Ralph Kimball’s Star Schema (Kimball & Ross, 2002) proved very popular, from which the well-known conceptual models SnowFlake and StarFlake were derived. More recent comprehensive data warehouse design models are built using Object-Oriented concepts on the foundations of the Star Schema. In Trujillo, Palomar, Gomez and Song (2001), Lujan-Mora, Trujillo, and Song (2002), and Abello, Samos, and Saltor (2001), two different OO modeling approaches are demonstrated where a data cube is transformed into an OO model integrating class hierarchies. The Object-Relational Star Schema (O-R Star) model (Rahayu, Dillon, Mohammad, & Taniar, 2001) aims to envisage data models and their object features, focusing on hierarchical dimension presentation, differentiation and their different sorts of embedded hierarchies.

These models, both object and relational, have a number of drawbacks if one wishes to use them for XML document warehouses, namely: (1) data-oriented without sufficient emphasis or capturing user requirements, (2) extensions of semantically poor relational models (star, snowflake models), (3) original conceptual semantics are lost before building data warehouses as the operational data source is relational, (4) further loss of semantics results from oversimplified dimensional modeling, (5) time consuming if additional data semantics are required to satisfy evolving user requirements, and (6) complex query design and processing is needed, therefore maintenance is troublesome. Traditional design models lack the ability to utilise or represent XML design level constructs in a well-defined abstract and implementation-independent form.

One of the early XML data warehouse implementations includes the Xyleme Project (Xyleme, 2001). The Xyleme project was successful and it was made into a commercial product in 2002. It has a well-defined implementation architecture and proven techniques to collect and archive Web XML documents into an XML warehouse for further analysis. Another approach by Fankhauser and Klements (2003) explores some of the changes and challenges of a document centric XML warehouse. Coupling these approaches with a well defined conceptual and logical design methodology will help future design of such XML warehouse for large-scale XML systems. In this article we are concerned with the design of a data warehouse rather than the implementation of the XML document warehouse. We propose a conceptual modelling approach for the development of an XML Document Warehouse (XDW), while emphasising the design techniques to build the XDW conceptual model in UML. We also carry out a systematic transformation of this conceptual model into an XML Schema. An in-depth analysis for the design of the xFACT and the Virtual Dimension structures is illustrated using a walk-through with a case study example.
Mining Statistically Significant Substrings Based on the Chi-Square Measure
Sourav Dutta and Arnab Bhattacharya (2012). Pattern Discovery Using Sequence Data Mining: Applications and Studies (pp. 73-82).
www.igi-global.com/chapter/mining-statistically-significant-substrings-based/58673?camid=4v1a

Discovering Patterns in Order to Detect Weak Signals and Define New Strategies
www.igi-global.com/chapter/discovering-patterns-order-detect-weak/58681?camid=4v1a