Chapter XII

Temporal Semistructured Data Models and Data Warehouses

Carlo Combi
University of Verona, Italy

Barbara Oliboni
University of Verona, Italy

Abstract

This chapter describes a graph-based approach to represent information stored in a data warehouse, by means of a temporal semistructured data model. We consider issues related to the representation of semistructured data warehouses, and discuss the set of constraints needed to manage in a correct way the warehouse time, that is the time dimension considered storing data in the data warehouse itself. We use a temporal semistructured data model because a data warehouse can contain data coming from different and heterogeneous data sources. This means that data stored in a data warehouse are semistructured in nature; that is, in different documents the same information can be represented in different ways, and the document schemata can be available or not. Moreover, information stored in a data warehouse is often time varying, thus as for semistructured data, also in the data warehouse context, it could be useful to consider time.
In recent years the database community has proposed flexible data models to repre-
sent semistructured information. Semistructured data have no absolute schema fixed in
advance. The structure may be irregular or incomplete (Abiteboul, 1997).

In the literature there are a number of approaches in which labeled graphs are
used to represent semistructured data (Comai, Damiani, Posenato, & Tanca, 1998;
Damiani, Oliboni, Tanca, & Veronese, 1999; Papakonstantinou, Garcia-Molina, &
Widom, 1995). These models organize data in graphs where nodes denote objects
or values, and edges represent relationships between them.

In the semistructured data context, the eXtensible Markup Language (XML) (World
Wide Web Consortium, 1998) is spreading out as a standard for representing, ex-
changing, and publishing semistructured information (Abiteboul, Buneman, &
Suciu, 2000), making information “self-describing,” that is it is possible there is
no separate description of the type or structure of data.

A data warehouse is a repository of data coming from different and heterogeneous
data sources. This means that data stored in a data warehouse are semistructured
in nature, because in different documents the same information can be represented in
different ways, and moreover, the document schemata can be available or not.
Furthermore, data warehouses can be used to store XML documents and WWW
data. A data warehouse storing information represented by means of XML is called
XML data warehouse (Marian, Abiteboul, Cobena, & Mignet, 2001), and a data
warehouse collecting information from the Web is called Web data warehouse
(Bhowmick, Madria, Ng, & Lim, 1998). In the literature are also considered XML

A dynamic warehouse for XML data was proposed and implemented in the Xyleme
project (Xyleme, 2001). The prototype was then turned into a product by a startup
company also called Xyleme.

Information stored into a data warehouse is often time varying, thus as for semi-
structured data, also in the data warehouse context, it could be useful to consider
time. The classical time dimensions, considered in the literature, are transaction
time and valid time. The transaction time is the time when a fact is current in the
database and may be retrieved, while the valid time is the time when a fact is true
in the considered domain (Jensen, Dyreson, Bohlen, et al., 1998).

In the semistructured data context, graph-based data models have been extended to
represent the time dimension of information, and issues related to the representation
of transaction and valid times have been studied (Chawathe, Abiteboul, & Widom,
1998; Combi, Oliboni, & Quintarelli 2004; Oliboni, Quintarelli, & Tanca, 2001). In
the data warehouse context, proposals in the literature focus on the representation

---

Copyright © 2007, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of
Idea Group Inc. is prohibited.