Chapter 74
Innovative Approaches for Efficiently Warehousing Complex Data from the Web

Fadila Bentayeb
University of Lyon, France

Nora Maïz
University of Lyon, France

Hadj Mahboubi
CEMAGREF Centre Clermont-Ferrand, France

Cécile Favre
University of Lyon, France

Sabine Loudcher
University of Lyon, France

Nouria Harbi
University of Lyon, France

Omar Boussaïd
University of Lyon, France

Jérôme Darmont
University of Lyon, France

ABSTRACT

Research in data warehousing and OLAP has produced important technologies for the design, management, and use of Information Systems for decision support. With the development of Internet, the availability of various types of data has increased. Thus, users require applications to help them obtaining knowledge from the Web. One possible solution to facilitate this task is to extract information from the Web, transform and load it to a Web Warehouse, which provides uniform access methods for automatic processing of the data. In this chapter, we present three innovative researches recently introduced to extend the capabilities of decision support systems, namely (1) the use of XML as a logical and physical model for complex data warehouses, (2) associating data mining to OLAP to allow elaborated analysis tasks for complex data and (3) schema evolution in complex data warehouses for personalized analyses. Our contributions cover the main phases of the data warehouse design process: data integration and modeling, and user driven-OLAP analysis.

DOI: 10.4018/978-1-4666-2455-9.ch074

INTRODUCTION

Traditional databases aim at data management (i.e., they help organize, structure, and query data). They are transaction processing-oriented and are often qualified as production databases. In opposition, data warehouses have a very different vocation: analyzing data (Kimball & Ross, 2002; Inmon, 2005) by exploiting specific models (star, snowflake and constellation schemas). They are termed as On-Line Analytical Processing (OLAP) databases. Data are then organized around indicators called measures, and analysis axes called dimensions. Dimension attributes either form a hierarchy or are just descriptive. Dimension hierarchies allow for obtaining views of data at different granularities (i.e., summarized or detailed through roll-up and drill-down operations, respectively).

Research in data warehousing and OLAP has produced important technologies for the design, management and use of information systems for decision support. To achieve the value of a data warehouse, incoming data must be transformed into an analysis-ready format. In the case of numerical data, data warehousing systems often provide tools to assist in this process. Unfortunately, standard tools are inadequate for producing relevant analysis when data are complex. Indeed, with the development of Internet, the availability of various types of data (Web data, multimedia data, biomedical data, etc.) has increased. Thus, users require applications to help them obtaining knowledge from the Web. For example, in the context of e-commerce, analyzing the behavior of a customer, a product, or a company consists of monitoring one or several activities (commercial or medical pursuits, patents deposits, etc.). The Web then becomes a real data source with which decision support applications should deal.

Furthermore, many Business Intelligence (BI) applications necessitate external data sources. For instance, performing competitive monitoring for a given company requires the analysis of data available only from its competitors. In this context, the Web is a tremendous source of data, and may be considered as a farming system.

However, the specific characteristics of Web data make it difficult to create such applications. One possible solution to facilitate this task is to extract information from the Web, transform and load it to a Web Warehouse, which provides uniform access methods for automatic processing of the data. Web Warehousing extends the lifetime of Web contents and its reuse by different applications across time.

Moreover, the special nature of complex data poses different and new requirements to data warehousing technologies, over those posed by conventional data warehouse applications. In this case, the data warehousing process should be adapted in response to evolving complex data and information requirements. Tools must be developed to provide the needed analysis. Therefore, the issue that may arise "Can we OLAP complex data?" To address this issue, we need a new generation of data warehousing models that can organize complex data in a multidimensional way and new OLAP operators that can analyze them.

The XML formalism has emerged as a dominant W3C standard for describing and exchanging semistructured data among heterogeneous data sources. Its self-describing hierarchical structure enables a manipulative power to accommodate complex, disconnected and heterogeneous data. It allows describing the structure of a document and constraining its contents. With its vocation for semistructured data exchange, the XML language offers great flexibility for representing heterogeneous data, and great possibilities for structuring, modeling and storing them.

Furthermore, a data cube structure can provide a suitable context for applying data mining methods. More generally, the association of OLAP and data mining allows elaborated analysis tasks exceeding the simple exploration of a data cube. The aim is to take advantage of OLAP, as well as data mining techniques, and to integrate them into the same analysis framework in order to provide