Chapter 2.15
Web Data Warehousing Convergence:
From Schematic to Systematic

D. Xuan Le
La Trobe University, Australia

J. Wenny Rahayu
La Trobe University, Australia

David Taniar
Monash University, Australia

ABSTRACT

This article proposes a data warehouse integration technique that combines data and documents from different underlying documents and database design approaches. The well-defined and structured data such as relational, object-oriented and object relational data, semi-structured data such as XML, and unstructured data such as HTML documents are integrated into a Web data warehouse system. The user specified requirements and data sources are combined to assist with the definitions of the hierarchical structures, which serve specific requirements and represent a certain type of data semantics using object-oriented features including inheritance, aggregation, association, and collection. A conceptual integrated data warehouse model is then specified based on a combination of user requirements and data source structure, which creates the need for a logical integrated data warehouse model. A case study is then developed into a prototype in a Web-based environment that enables the evaluation. The evaluation of the proposed integration Web data warehouse methodology includes the verification of correctness of the integrated data, and the overall benefits of utilizing this proposed integration technique.

INTRODUCTION AND MOTIVATION

Currently, there are more and more techniques being provided to accommodate the high demand for exchanging and storing business information including Web and operational data. While the

well-defined structured data are operated and stored in relational, object-oriented (Buzydlowski, 1998), object relational database environments, semi-structured data in XML or unstructured documents are stored in HTML. The problem of related information being separated and stored in multiple places happens quite often within an organization. Information from these applications is extracted and further developed into business analysis tools such as OLAP and data warehousing, which aim to support data analysis, business requirements, and management decisions.

Relevant business Web data have rapidly increased in significant amounts. Recently, XML has increased in popularity and has become a standard technique for storing and exchanging information over the Internet. The data integration (Breitbart, Olson, & Thompson, 1986) in the data warehousing has certainly received a lot of attention. There are three particular articles that are very close to the work in this article. Jensen, Moller and Pedersen (2001) allow an integration of XML and relational data. Even though the object-oriented concept is used in this model, the semantic contribution in this work lacks object-oriented features. Therefore, the semantics of data have been only partially supported. Other systems (Golfarelli, Rizzi, & Birdoljak, 1998, 2001; Huang & Su, 2001) focus on supporting Web data at the schematic level. While their initial focus is to incorporate XML data, Relational data have also been mentioned but not yet been incorporated. They mostly concentrate on the creation of a logical model.

Hence, it is clear that there is yet to be developed a standard integration technique that provides a means of handling multiple data sources being integrated into a data warehouse system (Bonifati, Cattaneo, Ceri, Fuggetta, & Paraboschi, 2001), and allowing a full capture of semantics of data in the data source models.

The purpose of this article can be summarized as follows:

- To ensure the integration technique allows a meaningful uniformed integrated object-oriented data warehouse structure.
- To ensure the integrated data and their semantics are explicitly and fully represented.
- To ensure a proposed integrated data warehouse system with consistency and high quality.
- To ensure the correctness of integrated data and benefits such as usefulness of the proposed integrated data warehouse system.

Figure 1 shows an overview of the proposed works in this article. The integration technique starts with a conceptual integrated data warehouse model (Ezeife & Ohanekwu, 2005) where the user requirement and underlying data source structures are used to assist with the design. The integrated Web data warehouse conceptual model deals with class formalization and hierarchical structures. The specified conceptual integrated Web data warehouse model has created a need for an integrated Web data warehouse logical model where underlying source structures are then absorbed and specified onto the existing conceptual Web integrated Web data warehouse model. The proposed Web integrated data warehouse models are then translated into a suitable

Figure 1. Integration Web data warehouse overview