Multidimensional Modeling of Complex Data

Omar Boussaid
University of Lyon, France

Doulkifli Boukraa
University of Jijel, Algeria

INTRODUCTION

While the classical databases aimed in data managing within enterprises, data warehouses help them to analyze data in order to drive their activities (Inmon, 2005).

The data warehouses have proven their usefulness in the decision making process by presenting valuable data to the user and allowing him/her to analyze them online (Rafanelli, 2003). Current data warehouse and OLAP tools deal, for their most part, with numerical data which is structured usually using the relational model. Therefore, considerable amounts of unstructured or semi-structured data are left unexploited. We qualify such data as "complex data" because they originate in different sources; have multiple forms, and have complex relationships amongst them.

Warehousing and exploiting such data raise many issues. In particular, modeling a complex data warehouse using the traditional star schema is no longer adequate because of many reasons (Boussaïd, Ben Messaoud, Choquet, & Anthoard, 2006; Ravat, Teste, Tournier, & Zurfluh, 2007b). First, the complex structure of data needs to be preserved rather than to be structured linearly as a set of attributes. Secondly, we need to preserve and exploit the relationships that exist between data when performing the analysis. Finally, a need may occur to operate new aggregation modes (Ben Messaoud, Boussaïd, & Loudcher, 2006; Ravat, Teste, Tournier, & Zurfluh, 2007a) that are based on textual rather than on numerical data.

The design and modeling of decision support systems based on complex data is a very exciting scientific challenge (Pedersen & Jensen, 1999; Jones & Song, 2005; Luján-Mora, Trujillo, & Song; 2006). Particularly, modeling a complex data warehouse at the conceptual level then at a logical level are not straightforward activities. Little work has been done regarding these activities.

At the conceptual level, most of the proposed models are object-oriented (Ravat et al, 2007a; Nassis, Rajugan, Dillon, & Rahayu 2004) and some of them make use of UML as a notation language. At the logical level, XML has been used in many models because of its adequacy for modeling both structured and semi structured data (Pokorný, 2001; Baril & Bellahsène, 2003; Boussaïd et al., 2006).

In this chapter, we propose an approach of multidimensional modeling of complex data at both the conceptual and logical levels. Our conceptual model answers some modeling requirements that we believe not fulfilled by the current models. These modeling requirements are exemplified by the Digital Bibliography & Library Project case study (DBLP)[1].

BACKGROUND

The DBLP (Ley & Reuther, 2006) represents a huge tank of data whose usefulness can be extended from simple reference searching to publication analysis, for instance:

- Listing the publications ordered by the number of their authors, number of citations or other classification criteria;
- Listing the minimum, maximum or average number of an author's co-authors according to a given publication type or year;
- Listing the number of publications where a given author is the main author, by publication type, by subject or by year;
- For a given author, knowing his/her publishing frequency by year, and knowing where he/she publishes the most (conferences, journals, books).

Currently, the DBLP database is not structured in such a way that allows data analysis. Along with this chapter, we propose a new structure for DBLP making further and richer data analysis possible. The DBLP case study raises many requirements that are worth

considering when modeling the data warehouse. Here follows the main requirements.

Complex Objects to Analyze

The objects to be analyzed may be characterized by simple linear attributes such as numerical measures or dimension attributes we find in the classical data warehouse (Kimball & Ross, 2002). However, in real life, an object may have a more complex structure (tree-like or graph-like). For instance, authors may be characterized by their names and affiliations, but publications are rather semi-structured and composed of sections, paragraphs, internal and external links, etc.

Complex Objects as Analysis Axes

Like the facts, the analysis axes may also be complex. For instance, an academic institution may be interested in evaluating authors according to their contributions, which may have a complex structure.

Objects Being Simultaneously Facts and Dimensions

In classical data warehouse modeling, facts and dimensions are treated separately. Even in symmetric models, they remain distinct at a given time. However, a need may occur to analyze an object according to objects of the same nature, and thus, one object may occur in dimension objects and facts objects simultaneously. For instance, it may be interesting to analyze the authors according to their co-authors in publications. Another example is the citation relationship between publications when we want to evaluate a publication according to its referencing publications.

Explicit and Richer Relationships Between Objects

In the classical star schema, the relationships between facts and dimensions are implicit. For instance, when relating "sales" as measures to "departments", "products" and "time" as dimensions, we know implicitly that our schema models product sales for each department during periods of time. However, in real-life applications, the relationships need to be explicit. Moreover, there may be more than one relationship between two objects. For example, we can distinguish two relationships between authors and publications: authoring and reviewing.

Complex Aggregations

Traditionally, aggregation functions such as SUM and AVERAGE deal with numerical data, but these are not the only aggregation needs we face. For example, Ravat et al (2007a) propose to aggregate documents using a function TOP_KEYWORDS that returns the most used keywords of some analyzed documents.

MAIN FOCUS

Three concepts compose the core of our model: the complex object, the relationship, and the hierarchy. In addition, we separate the definition of the (multi)dimensional model from that of the cube in one hand and the description of metadata from that of data in the other hand. In the following, we present our conceptual and logical models.

Conceptual Model

The model we define is composed of the following elements:

1. **Complex Object:** A complex object is a focus of analysis either as a subject (fact) or as an axis (dimension). An object can be unstructured, semi-structured, or structured. It can hold numerical or textual data relating to each other in different ways (linear, tree-like and graph-like).
   Formally, we refer to an object using the abbreviation "obj" and to a class of objects using "Obj". The set of object instances is noted E_{obj} where $E_{obj} = \{obj_i, i=1, n$ *where n is the number of instances of Obj*$\}$.
2. **Relationships between objects:** A relationship is a three-tuple "R" where:
   $R = (Obj_1,$ *relationship_name,* $Obj_2)$. Similarly, a relationship instance is a three-tuple "*r*" where $r = (obj_1,$ *relationship_name,* $obj_2)$. The set of relationship instances is noted E_r where
   $E_r = \{r_i, i=1, m$ *where m is the number of instances of R*$\}$.
3. **Hierarchies:** A hierarchy is an n-tuple of objects *H* where $H= (Obj_1, Obj_2, ..., Obj_n)$ *where*