OLAP Analysis Operators for Multi-State Data Warehouses

Franck Ravat, Toulouse Institute of Computer Science Research - Toulouse I Capitole University, Toulouse, France
Jiefu Song, Toulouse Institute of Computer Science Research - Toulouse I Capitole University, Toulouse, France
Olivier Teste, Toulouse Institute of Computer Science Research - University of Toulouse-Jean Jaurès, Toulouse, France

ABSTRACT

Data reduction in Multidimensional Data Warehouses (MDWs) allows increasing the efficiency of analysis and facilitating decision-makers’ tasks. In this paper, the authors model a MDW containing reduced data through a set of states. Each state is valid for a certain period of time; it contains only useful information according to decision-makers’ needs. In order to carry out analyses in a MDW composed of multiple states, an extension of traditional OLAP analysis operators is required. In this paper, the authors define a set of OLAP operators compatible with reduced MDWs. For each operator, they propose a user-oriented definition along with an algorithmic translation. To show the feasibility and the efficiency of the proposed concepts, they implement the analysis operators in an R-OLAP framework.

KEYWORDS

Data Reduction, Multidimensional Design, OLAP Analysis Operators

1. INTRODUCTION

Multidimensional Data Warehouses (MDWs) organize data in a multidimensional way in order to support On-Line Analytical Processing (OLAP) analyses. A MDW schema is based on facts (analysis subjects) and dimensions (analysis axes). The facts contain analysis indicators, while a dimension includes analysis parameters organized in hierarchies from the lower granularity (most detailed) to the higher granularity (most general). In a classical MDW, all data are permanently stored and new data are periodically added. The increasing volume of MDWs makes the tasks of decision-makers more difficult since they may be lost during their analyses. On the other hand, information is usually timely sensitive; most of detailed information loses its value over time (Ravat & Teste, 2000). Nevertheless, data at high granularity levels are more stable, and it can generally fulfill decision-makers’ needs when analyses are carried out over older data (Skyt, Jensen, & Pedersen, 2008). For instance, an analyst may have interest in analyzing sale amounts by product’s brand for the last five years. However, as many of today’s brands did not exist before, the brand granularity level may be useless for an older period. As a result, the analyst may have no more interest in analyzing sale amounts by brand over the last ten years but by a higher and more stable granularity level, such as product’s category.

Facing large volumes of data among which a great amount of inadequate detailed data are found, our aim is to both increase the analysis efficiency and facilitate the analysts’ tasks. To this end, we
propose a solution for supporting OLAP analyses over only relevant data over time. Firstly, we describe a conceptual MDW model based on data reduction to aggregate and then remove useless detailed data. Secondly, we propose a set of algebraic OLAP operators to support analyses over reduced data. The execution process of each operator is illustrated with the help of an algorithm. At last, we implement the reduced MDW model and the set of analysis operators in an R-OLAP framework. The framework aims at proving the feasibility of proposed concepts and evaluating the efficiency of carrying out analyses over reduced data.

This paper is composed as follows. Section 2 studies related works. Section 3 presents preliminary concepts of a reduced MDW illustrated with a case study. Section 4 describes our modeling solution for OLAP analysis operators. Section 5 presents a multi-state analysis framework showing the feasibility of our solution. Section 6 provides results of experimental assessments on the efficiency of the proposed OLAP operators.

2. RELATED WORK

2.1. Model of Reduced Data

Reducing data allows both decreasing the quantity of irrelevant data in decision-making process and increasing future analysis quality (Udo & Afolabi, 2011). In the field of MDWs, the authors of (Skyt et al., 2008) present techniques for a progressive data aggregation of fact. This study specifies data aggregation criteria of a fact according to granularity levels of dimensions. As mentioned in (Iftikhar & Pedersen, 2011), this work is theoretical and it fails to provide us a concrete implementation strategy. In (Iftikhar & Pedersen, 2011), a gradual data aggregation solution based on conception, implementation and evaluation steps is provided. This solution is based on a table containing different temporal granularities, such as second, minute, hour, month and year. However, the work of (Iftikhar & Pedersen, 2011; Skyt et al., 2008) is limited to the reduction of a single fact. Moreover, no formal model for reduced multidimensional data has been proposed.

Other related work concerning the management of data evolution in MDWs is also proposed. The objective of this work is different from the data reduction. In general, its objective is to manage changes in data sources. (Kimball, 1996) put forwards the concept of Slowly Changing Dimensions (SCDs). SCDs imply that data may change within a dimension, even though it occurs less frequently than in a fact. Three basic modeling solutions have been defined for managing dimensional data changes, namely overwrite old data, create new record for each change and keep data changes as alternative values in MDWs. The authors of (Golfarelli & Rizzi, 2009) point out that not only data but also MDW schema can change over time according to user’s requirements. SCDs, however, do not provide any solution to handle schema changes.

Two approaches allowing managing changes of data and schema have been proposed, namely schema evolution, and schema versioning (Golfarelli & Rizzi, 2009). The schema evolution approach supports only the lasted MDW version, while the schema versioning approach allows keeping different MDW versions by stamping each MDW version with a validation period.

In the schema versioning approach, (Eder, Koncilia, & Morzy, 2002) were the first to address questions on multiversion MDWs to manage both schema changes and instance changes. A metamodel is defined to stamp each element in an OLAP cube with a validity interval, so that multiple MDW versions can be managed and queried through the metamodel. However, no discussion about how a new version can be derived from a previous one is made in this work. (Body, Miquel, Bédard, & Tchounikine, 2002) propose a conceptual model allowing building a multiversion fact table. Several structural evolution operators are proposed in this work in order to manage the mapping between
Feature Selection in Data Mining
Yong Seong Kim, W. Nick Street and Filippo Menczer (2003). Data Mining: Opportunities and Challenges (pp. 80-105).
www.igi-global.com/chapter/feature-selection-data-mining/7597?camid=4v1a

Discovering Patterns in Order to Detect Weak Signals and Define New Strategies
www.igi-global.com/chapter/discovering-patterns-order-detect-weak/73516?camid=4v1a