International Journal of Data Warehousing and Mining
Volume 13 • Issue 1 • January-March 2017

Multidimensional Model Design using Data Mining:
A Rapid Prototyping Methodology

Sandro Bimonte, IRSTEA, Clermont Ferrand, France
Lucile Sautot, TETIS, AgroParisTech, Montpellier, France
Ludovic Journaux, LE21, AgroSupDijon, Dijon, France
Bruno Faivre, University of Burgundy Franche-Comté, Dijon, France

ABSTRACT

Designing and building a Data Warehouse (DW), and associated OLAP cubes, are long processes, during which decision-maker requirements play an important role. But decision-makers are not OLAP experts and can find it difficult to deal with the concepts behind DW and OLAP. To support DW design in this context, we propose: (i) a new rapid prototyping methodology, integrating two different DM algorithms, to define dimension hierarchies according to decision-maker knowledge; (ii) a complete UML Profile, to define a DW schema that integrates both the DM algorithms; (iii) a mapping process to transform multidimensional schemata according to the results of the DM algorithms; (iv) a tool implementing the proposed methodology; (v) a full validation, based on a real case study concerning bird biodiversity. In conclusion, we confirm the rapidity and efficacy of our methodology and tool in providing a multidimensional schema to satisfy decision-maker analytical needs.

KEYWORDS

Data Mining, Data Warehouse, Methodologies and Tools, OLAP

1. INTRODUCTION

Business Intelligence technology provides tools, such as Data Warehouses (DWs), On-Line Analytical Processing (OLAP), and Data Mining (DM), that allow decision-makers to explore huge volumes of data, in order to discover patterns and knowledge, and thus confirm their hypotheses.

DWs are large data repositories that support the decision-making process through flexible, interactive data analysis (Kimbal, 1996). Warehoused data are built according to a multidimensional model that defines concepts of facts and dimensions. Facts represent objects and are described by numerical attributes, called measures. Facts are analyzed along dimensions representing the axes of analysis. Dimensions are organized in hierarchies. Measures are aggregated with classical SQL aggregation functions (e.g. SUM, MIN, MAX, etc.) along hierarchical levels, using OLAP operators (Inmon, 2005). These OLAP systems allow decision-makers to visualize and explore facts during query sessions by applying OLAP operators: Slice selects a subset of warehoused data; Roll-Up aggregates measures by moving up through the hierarchy; Drill-Down is the opposite of Roll-Up, etc. A basic Relational OLAP (ROLAP) system architecture consists of: (i) a relational Data Base Management System (DBMS), which stores data in accordance with a multidimensional paradigm; (ii) an OLAP server, which implements the multidimensional model and OLAP operators on top of the DBMS; (iii) an OLAP client, which combines and synchronizes tabular and graphical displays,

DOI: 10.4018/IJDWM.2017010101

and allows DW queries; (iv) an ETL tool, which extracts data from multiple heterogeneous sources, then transforms and loads them into the DW. The classic development cycle of DWs includes several steps, among which ETL design is typically the most time-consuming (Bimonte, Edoh-Alove, et al. 2013). Several DW design methodologies can be characterized by the relative importance of user requirements (Romero & Abelló, 2009; Kimbal, 1996): in requirement-driven approaches, the conceptual DW schema is based primarily on user requirements; in source-driven approaches, the conceptual DW schema is (semi-automatically) derived from the schemata of the data sources; in mixed approaches, these two processes are carried out in parallel. Rapid DW prototyping is crucial when dealing with complex applications, and has therefore been the object of several studies (Bimonte, Edoh-Alove et al., 2013; Golfarelli & Rizzi, 2011; Huynh & Schiefer, 2001). The Bimonte et al. study presented a rapid, requirement-driven design methodology and tool, called ProtOLAP. Their methodology is based on conceptual DW models, which are then implemented automatically. After DW implementation, decision-makers must manually feed sample data into the prototype, dimension by dimension and level by level, for each hierarchy, to simulate an ETL process in the context of a requirement-driven methodology. However, feeding DWs with sample data is not always easy and, in some cases, dimensional data lack the hierarchical structure necessary to fit the user's requirements.

Data Mining (DM) is a data exploration phase of a Knowledge Discovery in Databases (KDD) process (Fayyad et al., 1996). DM is a set of descriptive and predictive methods that aim to explore data by discovering *a priori* unknown links between data attributes (Tufféry, 2011). DM is at the interface between machine learning and statistics, and includes automatic and semi-automatic approaches. DM offers three main techniques:

1. Clustering, or unsupervised classification: this approach corresponds to organizing a data collection (represented by a vector or a point in a multidimensional space) into classes (groups or clusters), based on similarity between group members according to a mathematical indicator (Jain et al. 1999). Classes are not defined by analysts but discovered during the clustering process.
2. Supervised classification: this approach includes an item in a class, within a set of classes predetermined by analysts.
3. Association rule learning, which discovers rules from data.

The integration of OLAP and DM can be achieved by enhancing OLAP operators with DM algorithms (i.e. DM over OLAP; Han, 1997), but DM can be also used in physical and conceptual phases of DW design (i.e. OLAP design by DM; Liu & Luo, 2005). In the field of conceptual modeling, Abelló et al. (2006) focused on DW design, while Torlone (2003) sought to facilitate interaction between decision-makers and DW experts (Torlone, 2003). Only Zubcoff et al. (2009) have presented an integrated framework, based on UML, to define conceptual models for DM algorithms on warehoused data using the DM over OLAP approach.

As yet, however, no rapid prototyping methodology has integrated DM into DW design.

Therefore, in a preliminary study (Sautot et al., 2014), we briefly presented a new prototyping methodology for DWs, using clustering methods to define the DW schema. Building upon our previous study, we now include more advanced DM methods, thus proposing the following improvements:

1. A new rapid prototyping methodology, integrating two different DM algorithms, to define dimension hierarchies according to decision-maker knowledge.
2. A complete UML Profile, to define a DW schema that integrates both DM algorithms.
3. A mapping process to transform multidimensional schemata according to the results of the DM algorithms.
4. A tool implementing the proposed methodology.
5. A full validation, based on a real case study, concerning bird biodiversity.

PaKDD-2007: A Near-Linear Model for the Cross-Selling Problem

Thierry V. de Merckt and Jean-Francois Chevalier (2008). *International Journal of Data Warehousing and Mining (pp. 46-54).*

www.igi-global.com/article/pakdd-2007-near-linear-model/1806?camid=4v1a

Integrating Star and Snowflake Schemas in Data Warehouses

Georgia Garani and Sven Helmer (2012). *International Journal of Data Warehousing and Mining (pp. 22-40).*

www.igi-global.com/article/integrating-star-snowflake-schemas-data/74754?camid=4v1a