A Methodology and Tool for Rapid Prototyping of Data Warehouses using Data Mining: Application to Birds Biodiversity

Lucile Sautot, Sandro Bimonte, Ludovic Journaux, Bruno Faivre

A Methodology and Tool for Rapid Prototyping of Data Warehouses using Data Mining: Application to Birds Biodiversity

Lucile Sautot[1,2], Sandro Bimonte[3], Ludovic Journaux[4], Bruno Faivre[1]

[1] Biogéosciences UMR CNRS-uB 6282, University of Burgundy, Dijon, France
l.sautot@agrosupdijon.fr

[2] AgroParisTech, Paris, France

[3] Irstea, TSCF, 9 avenue Blaise Pascal CS20085, 63178 Aubière France

sandro.bimonte@irstea.fr

[4] LE2I, UMR CNRS 6306, University of Burgundy, Dijon, France
ludovic.journaux@agrosupdijon.fr

Abstract. Data Warehouses (DWs) are large repositories of data aimed at supporting the decision-making process by enabling flexible and interactive analyses via OLAP systems. Rapid prototyping of DWs is necessary when OLAP applications are complex. Some work about the integration of Data Mining and OLAP systems has been done to enhance OLAP operators with mined indicators, and/or to define the DW schema. However, to best of our knowledge, prototyping methods for DWs do not support this kind of integration. Then, in this paper we present a new prototyping methodology for DWs, extending [3], where DM methods are used to define the DW schema. We validate our approach on a real data set concerning bird biodiversity.

Keywords: Data Warehouse design, OLAMining, Rapid prototyping

1. Introduction

Data Warehouses (DWs) are huge data repositories aimed at supporting the decision-making process by enabling flexible and interactive analysis on data [10].

A distinction is made on DW design methodologies depending on the role given to user requirements [10,16]: in requirement-driven approaches, a conceptual schema of the DW is designed starting from the user requirements; in source-driven approaches, a conceptual schema is (semi-automatically) derived starting from the schemata of the data sources that will be integrated in the DW; in mixed approaches, the two processes are carried out in parallel. Rapid prototyping of DW is crucial when dealing with complex application and it has been investigated in some work [3,6,9]. In [3], authors presented an agile requirement-driven design methodology and tool, called ProtOLAP. ProtOLAP is based on using DW conceptual models and automatic implementation of DW and OLAP models. Then, decision-makers must manually feed sam-

ple 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, we have noted that feeding DW with sample data can be not a simple task. Furthermore, in some cases, the dimensional data have not a hierarchical structure that can be predefined according to user's requirements.

Some work about the integration of Data Mining (DM) and OLAP systems has been done to enhance OLAP operators with mined indicators [8], and/or to define the DW schema [17,18]. In [8] classical OLAP operators (drill-down and roll-up operations, slicing, dicing, pivoting) are completed by analysis operators based on DM algorithms. [13] presents an OLAP aggregation operator, named OpAC, performing clustering on data with an Agglomerative Hierarchical Clustering. The goal of this new operator is to group facts that are significantly similar. Thus the integration of OLAP and DM can be achieved by enhancing OLAP operators with DM algorithms (i.e. *DM over OLAP*), but DM can be also used in the DW design's physical and conceptual phases (i.e. *OLAP design by DM*). For example, [4,18] uses DM clustering algorithms to define hierarchies, and [12] to define physical models. In the context of conceptual modeling a lot of effort has been done for the DW design. Indeed, several work propose conceptual models for DW using ad-hoc formalism, ER based models or standards such as UML (see [1] for a review). Some works propose conceptual models for DM (e.g. [15]). On the other hand, only [8] presents an integrated framework, based on UML, to define conceptual models for DM algorithms on warehoused data according to the DM over OLAP approach.

Finally, rapid prototyping DW methodologies are based on interactive and iterative multidimensional schemata defined by users [3], where statistical methods [9] are only used to select a subset of data to feed fact and dimensions data. To conclude to best of our knowledge no rapid prototyping methodology for *OLAP design by DM* has been addressed yet.

Thereby, in this paper we present a new prototyping methodology for DWs, extending [3], where DM methods are used to define the DW schema. In particular, (i) we present an extension of the UML profile for spatial DW integrating the Hierarchical Agglomerative Clustering for defining dimension hierarchies [17], we (ii) extend the prototyping methodology and (iii) tool to handle with DM setting, and (iii) we validate our approach on a real data set concerning bird biodiversity.

The paper is organized in the following way. Section 2 describes the case study of the paper; the ProtOLAPMining methodology and its supporting tools are presented in Sections 3 and 4; Section 5 gives some hints to future work and concludes the paper.

2. Motivations

We present an example from an ecological study: a bird census program along the Loire River (France) [5]. This program aims to detect temporal and spatial changes of bird communities. One hundred ninety eight points were located each 5 Km along the river, and at each point birds were numbered using a point count census method: Punctual Abundance Index. Birds have been censused in four occasions during the last 20 years (1990, 1996, 2002 and 2011). Decision-makers of that project are unskilled OLAP users, and then they need DW prototypes to validate their analysis

needs in terms of dimensions and facts. However, they identify a numerical value as analysis subject representing the abundance and three dimensions that characterize it: time, space and the species. The dimensions that describe species and time are easy to design. However, the design of the spatial dimension is more complex. Environment has been described around each point in the years chosen for bird census. To explain bird abundances and their variations, abundances were correlated with environmental variables (such as altitude, etc.).

However, environmental variables belong to different categories: continuous, discrete, ordinal and qualitative variables. In this context, the design of a spatial hierarchy is not obvious, because the description of each point along the river consists of a mixed data set, with no evident hierarchical structure (the French administrative division does not make sense). In other terms, this dimension has not a well defined hierarchical schema, but for example a hierarchical clustering algorithm can be used to derive groups.

In this kind of context, the design methodology of such DW should be based on a particular methodology that allows:

1. Include data mining at conceptual level to create the hierarchy schema and instance. Indeed has been widely recognized that conceptual models are useful in complex application to provide a ridge between users and information technology experts [1].
2. The data mining algorithm should:
    1. Generate a strict, onto and covering hierarchy [14] since they are easily handled by all existing OLAP server.
    2. Generate a hierarchy with several levels.
    3. Generate labels for each level and each member of each level, since hierarchical levels represent a semantic concept.
    4. Control the number of levels of the calculated hierarchy since too much levels are not useful in a classical OLAP exploration.
3. Adopt an agile prototyping paradigm [3]: our design tool must offer the possibility to go back to some of the key steps of the design in order to revise the choices made and refine the DW modeling and DM setting.
4. Being a mixed methodology [16]: our methodology should allow decision-makers to define their functional requirements and at the same time analyze existing data sources to be mined during the hierarchy creation process.

3. ProtOLAPMining

In this section we present our methodology (Sec 3.1), the DM algorithm used (Sec 3.2) and the extension of the ICSOLAP profile [2] (Sec 3.3).

3.1 The methodology

The classical DW development has been extended with agile steps (from 3 to 8), and integrates DM functional requirements in steps 1 and 2, as described in the following (Figure 1):

1. Decision-makers informally define the functional multidimensional needs (i.e. analysis axes and subjects). In our case study, the decision-makers want to analyze the bird abundance according to three analysis axes: time, space and species.
2. Decision-makers informally define the functional data mining needs (i.e. data mining parameters). In particular, the decision-makers choose a variable set for each automatically hierarchical dimension, specify the variable type (qualitative or quantitative) and choose metric and linkage. In our case study, a decision-maker can choose the spatial dimension and three variables: altitude, stream and geology. The altitude is a quantitative variable while the stream and the geology are qualitative variables. The selected variables are qualitative and quantitative, so the metric and the linkage must be adapted. The only metric that we propose for the mixed data set, is the Gower index that is detailed in section 3.2. Several linkage methods are available for mixed data set, so the decision-maker can choose one of them such as the Unweighted Pair Group Method with Arithmetic mean (UPGMA) (c.f. Sec 3.2)
3. Designers create a *conceptual multidimensional-DM schema*, meaning starting from the users' analysis needs defined in step 2. We note *conceptual multidimensional-DM schema* a classical conceptual multidimensional schema enriched with DM methods for hierarchy creation. In our case study, the designers create two classical dimensions (the time and the species dimensions) and a dimension with an automatically generated hierarchy (the spatial dimension) (c.f. Sec 3.3).
4. Decision-makers set a data sample for the DM algorithm.
5. The system automatically creates the DW hierarchy using the DM algorithm with data of step 4 and parameters of step 3.
6. The system automatically create the DBMS and the OLAP server models
7. Decision-makers feed classical dimensions with sample data.
8. Decision-makers explore the DW with an OLAP client. If hierarchy created using the DM algorithm is not suitable go to Step 2. If multidimensional structures (dimensions and facts) are not adapted go to Step 1.
9. Implementing the prototype (ETL and DM running) on all the data set.

Our methodology satisfies the requirements 1, 3 and 4 described in Section 2. Indeed, it allows decision-makers and DW experts to easily define and validate their DW prototypes enriched with DM algorithms for hierarchical design in an incremental way. Let us now describe what DM algorithm we use that satisfies requirement 2 of Section 2.

3.2 The DM algorithm: Agglomerative Hierarchical Clustering

In the implementation of our methodology we have chosen as DM algorithm the Agglomerative Hierarchical Clustering (AHC). Main steps of this algorithm are: (1) Calculation of distances between individuals; (2) Choice of the two nearest individuals. (3) Aggregation of the two nearest individuals in a cluster. The cluster is considered an individual. (4) Go back to the step 1 and loop while there is more than one individual.

For steps 1 and 3, we need to define a metric in order to measure the distance between individuals (distance) and a method to aggregate individuals in different clusters (linkage). Our data set contains qualitative and quantitative variables (mixed data set). With qualitative variables we cannot define a cluster as the centroïd of these members. In this context, several linkage methods can be used. As it shown in [11], we choose the unweighted average distance (UPGMA), because, without knowledge on the data structure, this linkage appears to be the best summary of the distance between two clusters. The distance between two individuals must mix quantitative and qualitative variables. We suggest measuring the distances between individuals with an index that comes from biology: the Gower similarity index [7].

The calculated hierarchy contains numerous levels with numerous clusters. But the users of AHC do not traditionally use the complete hierarchy. In an OLAP context, we cut the calculated hierarchy according to a desired number of levels.

However, our algorithm is based on an unsupervised clustering algorithm. Thereby this algorithm cannot generate labels for levels or clusters.

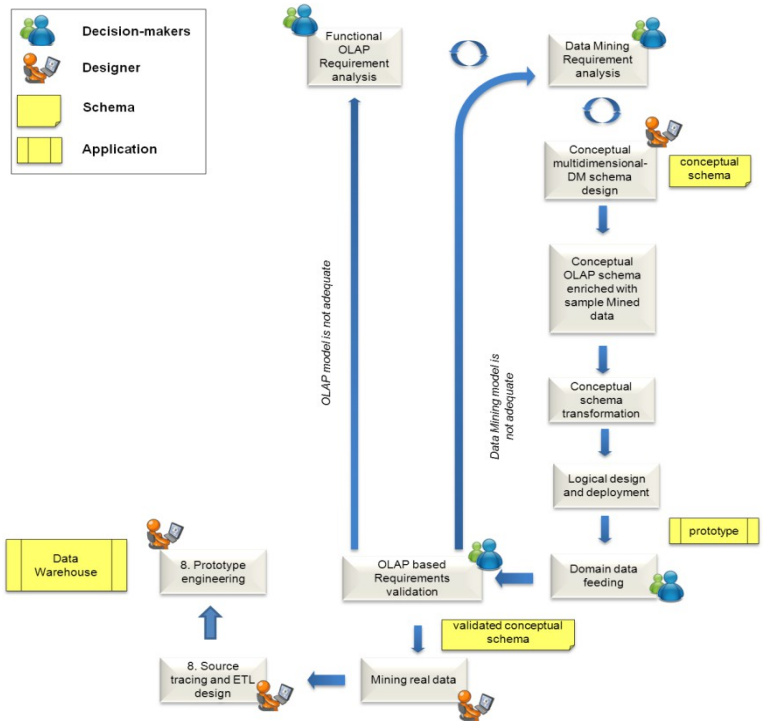

Fig 1. OLAPMining protyping methodology.

3.3 DMICSOLAP UML Profile

As previously described our methodology is based on the formalization of data mining and multidimensional requirements using a conceptual multidimensional-DM model. Then, we extend the ICSOLAP UML profile [3] to include DM parameters (DMICSOLAP UML Profile). Our approach is based on the ProtOLAP methodology

where the conceptual multidimensional schema is defined using the UML Profile for spatial data warehouses ICSOLAP. In ICSOLAP model, for each multidimensional element, a stereotype or a tagged value is defined. In particular dimensions represented as packages are composed of hierarchies that hierarchically organize levels. In particular, a level (“AggLevel” stereotype) is a class composed of a set of descriptive attributes (“DescriptiveAttribute” stereotype) and an identifying attribute. “SpatialAggLevel” designs spatial dimension levels whose geometries are represented with geometric attributes stereotyped “LevelGeometry”. A fact is represented using the stereotype “Fact”, which is a class with attributes that are measures (“NumericalMeasure”).

Our extension of ICSOLAP defines a new stereotype <<AHClevel>> that extends a level with set of attributes with the <<variable>> stereotype. The <<variable>> stereotype represents the variables used by the AHC algorithm, for example the substratum. A <<Variable>> can be Quantitative or Qualitative. Moreover, we define a tagged value Linkage representing the linkage parameter of the algorithm such as UPGMA as in our case study. In the same way we have defined three tagged values representing the distance used when only quantitative variables are used (DistanceQuantitatve), only qualitative variables (DistanceQualitatve), and DistanceMix when qualitative and quantitative variables are used. In our case study DistanceMix has the value Gower. Finally, the number of levels needed by decision-makers is represented with the LevelsNb tagged value.Figure 2 is shown the conceptual multidimensional-DM model of our case study. We can note three dimensions, where two dimensions are classical dimensions (temporal and thematic) and one dimension is composed of hierarchy LocationH with a <<AHClevel>> level Station. This hierarchy has a most detailed spatial level representing the stations, which are grouped in at most 9 coarser levels. The hierarchical relationships are created using the AHC algorithm using Gower index and WPGMA on data representing the stations, which are clustered using the substratum and the valley variables. The fact represents the abundance.

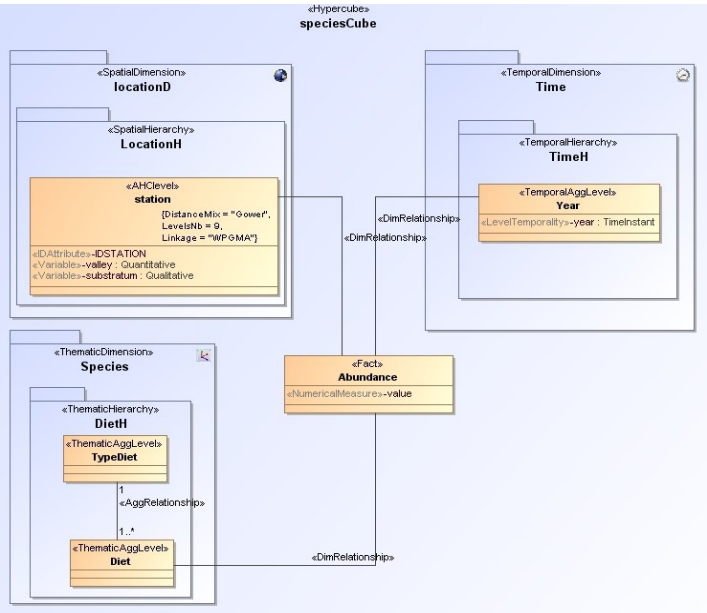

Fig 2. Conceptual multidimensional-DM schema

4. ProtOLAPMining tool

ProtOLAPMining is the system implementing our methodology. It extends ProtOLAP with the DM deployment tier. ProtOLAP is based on a Relational architecture with PostgreSQL as DBMS for storing warehoused data, Mondrian as OLAP server and JRubik as OLAP client. ProtOLAP takes as inputs the UML file representing the multidimensional model and automatically generates the SQL and Mondrian schemas. Moreover, it allows feeding the DW with same sample data using the Feeding tier.

The new tier, DM deployment tier, implements the AHC algorithm. It allows in particular to indicate the database or the file that contains the data representing the <<AHCLevel>> (mined data) and setting the inputs parameters. The tier runs the algorithm, and then creates the SQL and Mondrian schemes for the new created hierarchy. Finally, also other dimensions and facts are automatically created and decision-makers can analyze data with the OLAP client.

5. Conclusion and Future work

In this work, we have presented a prototyping methodology and the associated tool, to design a multidimensional schema. This methodology and this tool are extensions of [3]. The methodology is based on agile method. Thus, it encourages the exchange of views between the decision-makers (the final users of the OLAP cube) and the designers. In fact, the main principle of the methodology is the validation, by the decisions-makers, of a prototype built by the designers. Moreover, the methodology is augmented with a DM algorithm. Our tool uses a clustering algorithm to create automatically a hierarchy in a dimension of the prototype of OLAP cube. This algorithm is based on AHC and is able to build hierarchy with mixed data, that contain quantitative and qualitative variables.

As future work, we expect to complete our methodology and our tool with other data mining methods. The integration of other data mining methods, as classification algorithm, can offer to the decision-makers and to the designers new strategies to build the most suitable OLAP cube. Moreover, the integration of other algorithms can permit the automation of a greater part of the tool and offer a more efficient tool. In addition, we will evaluate the methodology on a panel of users.

6. References

1. Abelló, A., Samos, J., and Saltor, F.: YAM2: a multidimensional conceptual model extending UML. Information Systems 31(6), 541-567 (2006)
2. Bimonte, S., Boulil, K., Pinet, F., Kang, M.-A.: Design of Complex Spatio-multidimensional Models with the ICSOLAP UML Profile – An Implementation in MagicDraw. In Proceedings of the 15th International Conference on Enterprise Information Systems (ICEIS) 1, 310-315 (2013)
3. Bimonte, S., Edoh-Alove, E., Nazih, H., Kang, M.-A., and Rizzi, S.: ProtOLAP: rapid OLAP prototyping with on-demand data supply. In Proceedings of the ACM Sixteenth International Workshop On Data Warehousing and OLAP (DOLAP), 61-66 (2013)

4. Favre C., Bentayeb F., and Boussaid O.: A knowledge-driven data warehouse model for analysis evolution. Frontiers in Artificial Intelligence and Applications 143, 271-278 (2006)
5. Frochot, B., Eybert, M.C., Journaux, L., Roché, J., and Faivre, B.: Nesting birds assemblages along the river Loire: result from a 12 years-study. Alauda 71(2), 179-190 (2003)
6. Golfarelli, M., and Rizzi, S.: Data warehouse testing: A prototype-based methodology. Information and Software Technology 53, 1183-1198 (2011)
7. Gower, J.C.: A general coefficient of similarity and some of its properties. Biometrics 27(4), 857-871 (1971)
8. Han, J.: Olap mining: An integration of OLAP with data mining. In Proceedings of the 7th IFIP, volume 2, pages 1-9 (1997)
9. Huynh, N., and Schiefer, J.: Prototyping Data Warehouse Systems. In Proc. DaWaK, 195-207 (2001)
10. Kimball, R.: The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley (1996) ISBN 0-471-15337-0
11. Kojadinovic, I.: Agglomerative hierarchical clustering of continuous variables based on mutual information. Computational Statistics & Data Analysis 46(2), 269-294 (2004)
12. Mahboubi, H., and Darmont, J.: Data mining-based fragmentation of XML data warehouses. In Proceedings of the ACM Eleventh International Workshop on Data Warehousing and OLAP (DOLAP), 9-16 (2008)
13. Messaoud, R.B., Boussaid, O., and Loudcher Rabaséda, S.: A data mining-based OLAP aggregation of complex data: Application on XML documents. International Journal of Data Warehousing and Mining (IJDWM) 2(4), 1-26 (2006)
14. Pedersen, T.B., Jensen, C.S., and Dyreson, C.E.: A foundation for capturing and querying complex multidimensional data. Information Systems 26(5), 383-423 (2001)
15. Rizzi, S.: UML-based conceptual modeling of pattern-bases. In Proceedings of International Workshop on Pattern Representation and Management (PaRMa) (2004)
16. Romero, O., and Abelló, A.: A Survey of Multidimensional Modeling Methodologies. International Journal of Data Warehousing and Mining 5(2), 1-23, (2009)
17. Sautot, L.; Faivre, B.; Journaux, L., and Molin, P.: The Hierarchical Agglomerative Clustering with Gower Index: a methdology for automatic design of OLAP cube in ecological data processing context. Ecological Informatics (2014) *to appear*
18. Usman, M., and Pears, R.: A methodology for integrating and exploiting data mining techniques in the design of data warehouses. In Proceedings of the 6th International Conference on Advanced Information Management and Service (IMS), pages 361-367. IEEE (2010)