OpAC: A New OLAP Operator Based on a Data Mining Method

Riadh Ben Messaoud [*], Sabine Rabaséda [**], Omar Boussaid [**], Fadila Bentayeb [*]

Laboratoire ERIC – Université Lumière Lyon 2
5 avenue Pierre Mendès-France
69676 Bron Cedex – France
http://eric.univ.lyon2.fr
[*]{rbenmessaoud, bentayeb}@eric.univ-lyon2.fr
[**]{sabine.rabaseda, boussaid}@univ-lyon2.fr

Abstract. For a few years, on-line analysis processing (OLAP) and data mining have known parallel and independent evolutions. Some recent studies have shown the interest of the association of these two fields. Currently, we attend the increase of a more elaborated analysis's need. We think that the idea of coupling OLAP and data mining will be able to fulfill this need. We propose to adopt this coupling in order to create a new operator, *OpAC* (Operator for Aggregation by Clustering), for multidimensional on-line analysis. The main idea of *OpAC* consists in using the agglomerative hierarchical clustering to achieve a semantic aggregation on the attributes of a data cube dimension.

Keywords. On-line analysis processing, Data cubes, Data mining, Agglomerative hierarchical clustering, Semantic aggregation.

1. Introduction

Data warehouses provided several solutions to the management of huge amount of data [9]. In fact, a data warehouse is an analysis oriented structure that stores a large collection of subject-oriented, integrated, time variant and non-volatile data. The warehousing process starts by extracting, transforming and loading data from heterogeneous sources (ETL). Some particular models, such as the star schema and the snow-flaked schema, are designed in order to prepare integrated data to analysis using the on-line analytical processing technology (OLAP). These models support decision making tasks by exploring multidimensional data views, commonly called *data cubes* [1]. So far, a data warehouse becomes a large infrastructure for designing efficient decision process through visualization and navigation into large data volumes.

On the other side, data mining uses machine learning methods to discover, describe and predict non trivial patterns from data. These patterns are usually expressed in valid and understandable models. However, data mining is a dependent step in the process of knowledge discovery in databases. In fact, all data mining methods need to work on integrated, consistent and cleaned data, which often requires data cleaning as preprocessing steps [5].

OLAP and data mining have known parallel and independent evolutions. For long, they were considered as two different fields. Currently, we think that their association could allow a more elaborated OLAP task exceeding the simple exploration of a data cube.

In one hand, OLAP is characterized by its aggregation tools, its navigational aspect and its power for visualizing data. In the other hand, data mining is known for the descriptive and predictive power of its results. Moreover, we think that multidimensional data structure can provide a suitable context for applying data mining methods. Our purpose is to take advantage as well from OLAP as from data mining and to integrate them in the same analysis process to provide exploration, explication and prediction capabilities. We look, particularly, for improving the traditional OLAP operators by creating a new form of aggregation based on a data mining method. Taking into account the multidimensional structure of data and the need to integrate them in a more elaborated analysis process, our idea consists in developing a new aggregation operator, called *OpAC* (Operator for Aggregation by Clustering), and based on the AHC (Agglomerative Hierarchical Clustering) [10].

The remaining of this paper is organized as follows. In section 2, we expose the related works to the coupling between OLAP and data mining. In section 3, we present the objectives of our proposed operator. In section 4, we motivate why we choose the AHC as an aggregation method. We develop, in section 5, a theoretical formalization for the *OpAC* operator. In section 6, we propose an implementation of a prototype and in section 7, we conclude our work and propose some future research topics.

2. Related work

A few research studies deal with the coupling of OLAP and data mining. This is due partly to the fact that most of the attention is directed towards separated improvements of the two areas. Nevertheless, we distinguish three principal groups of approaches:

The first approach consists in simulating the data mining methods by extending OLAP operators. Han proposes a system, called *DBMiner*, which can perform some data mining functions including association, classification, prediction, clustering, and sequencing [8]. Chen et al. suggest an approach consisting in mining functional association rules using the distributed OLAP and data mining infrastructure [3]. The purpose of this work is to enhance the expressive power of association rules. The infrastructure mines e-commerce transaction data and generate association rules expressing customer behavior patterns. Goil and Choudhary propose to mine knowledge from data cubes by using the OLAP operators [6]. Their approach consists in discovering association rules from the quantitative summary information contained in a data cube.

The second approach aims to adapt multidimensional data in order to make them understandable by data mining methods. Two strategies are proposed.

One consists in taking advantages from multidimensional database management system (MDBMS) to help the construction of learning models. For instance, Laurent proposes a cooperation between *Oracle Express* and a fuzzy decision tree software (Salammbô) [11]. This cooperation allows transferring learning tasks, storage constraints and data handling to the MDBMS.

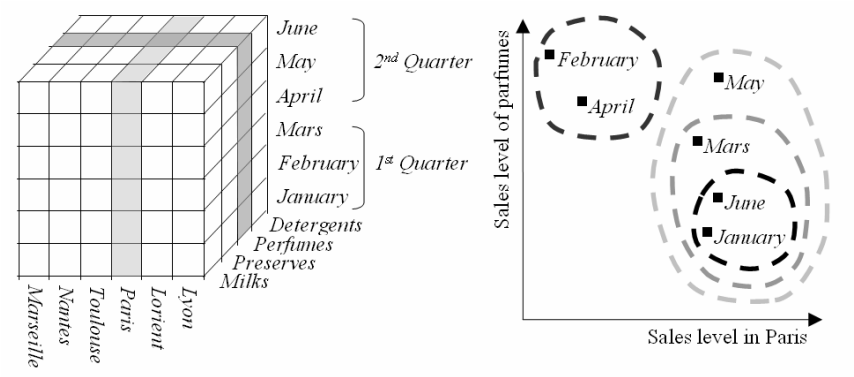
Another strategy transforms the multidimensional data and makes them usable by the data mining methods. Pinto et al. integrate multidimensional information in data sequences and apply on them the discovery of frequent patterns [12]. In order to implement a decision tree on multidimensional data, Goil and Choudhary flatten data cubes to extract contingency matrix for each dimension at each construction step of the tree [7]. Chen et al. propose to adopt OLAP as a preprocessing step of the knowledge discovery process [2]. Transformed data can therefore be exploited by data mining methods.

The third approach aims to adapt data mining algorithms and employs them directly in multidimensional data. Sarawagi et al. propose to integrate a statistical module, based on multidimensional regression, (*Discovery-driven*) in OLAP server. This module guides the user to detect relevant areas at various hierarchical levels of a cube [13]. In [14], Sarawagi proposes a new tool, *iDiff*, based on dynamic programming, which detects both relevant areas in a data cube and the reasons of their presence. Similar works were released in [4] for the generation of natural language from multidimensional data.

Finally, we note that none of the existing approaches employs the coupling between data mining and OLAP in order to enhance the functionalities of OLAP operators. Our approach associates the exploratory tools of OLAP with the descriptive and predictive aspect of data mining. The current work aims to define a new generation of analysis operators based on data mining methods. We propose in this paper a new operator based on a data mining method to fulfill more elaborated analysis.

3. *OpAC* operator objectives

The construction of a data cube targets precise analysis goals. The selection of its dimensions and measures depends on the analysis needs. Usually, a dimension is organized according several hierarchies expressing various levels of granularity. Each hierarchy contains a set of modalities, and each modality of a hierarchy includes modalities from the hierarchy immediately below according to the logical membership order.



(a) OLAP context (b) Clustering context

Figure 1. Principle of the aggregation operator *OpAC*.

In fact, the modalities of a dimension are always organized according to the logical order of membership well known in the natural use of objects and concepts of the real world. Let's consider the data cube presented in Fig.1(a). The cube is made up of three dimensions: *Location*, *Time* and *Product*. The *Time* dimension is organized according to two hierarchical levels: *Months* and *Quarters*. It is natural to say that the modality "*1st Quarter*" of the temporal dimension aggregates the months: "*January*", "*February*" and "*March*".

Unlike the traditional OLAP aggregation, exposed above, our approach takes the cube measures into account in order to provide a *semantic* aggregation over dimension modalities. The goal of our operator *OpAC* is to use a clustering method in order to highlight aggregates semantically richer than those provided by the current OLAP operators.

As shown in Fig.1(b), the new operator enables us to note that "*January*" and "*June*" form a more significant aggregate since they represent periods where sales level of "*Perfumes*" in the city of "*Paris*" are slightly similar.

Existing OLAP tools, like the *Slicing* operator, can also create new modalities' aggregates in a cube dimension. Therefore, these tools always need handmade user assistance, whereas our operator is based on a clustering algorithm that provides automatically relevant aggregates. Furthermore, with classical OLAP tools, aggregates are created in an intuitive way in order to compare some measure values, whereas *OpAC* creates significant aggregates expressing deep relations with the cube's measures. Thus, the construction of this kind of aggregates is very interesting to establish a richer on-line analysis context.

4. The choice of the agglomerative hierarchical clustering

According to our *OpAC* operator objectives, we chose the agglomerative hierarchical clustering (AHC) as an aggregation method for the *OpAC* operator. This choice is motivated by the following points:

- The hierarchical aspect constitutes a relevant analogy between the AHC results and a hierarchical structure of a dimension. Furthermore, the objectives and the results representation expected for *OpAC* match perfectly with the AHC strategy;
- Unlike the DHC (Divisive Hierarchical Clustering), the AHC adopts an agglomerative strategy beginning by the finest partition where each individual is considered like a class. This allows including the finest modalities of a dimension in the results of *OpAC*. Moreover, the ascending strategy is faster than the divisive one;
- The results of the AHC are compatible with exploratory aspect of OLAP and can be reused by its classical operators. The AHC provides several hierarchical partitions of individuals. By moving from a partition level to the higher one, two aggregates are joined together. Conversely, by moving from a partition level to the lower one, an aggregate is divided into two new aggregates. These operations are strongly similar to the classical operators *Roll-up* and *Drill-down*.

5. *OpAC* operator formalization

This formalization defines the individuals and the variables domains for the clustering problem. Let's Ω be the set of individuals and Σ the set of variables. We suppose that:

- C is a data cube having d dimensions and m measures;
- $D_1,\ldots,D_i,\ldots,D_d$ the dimensions of C;
- $M_1,\ldots,M_q,\ldots,M_m$ the dimensions of C;
- $\forall i \in \{1,\ldots,d\}$ the dimension D_i contains n_i hierarchical levels;
- h_{ij} the j^{th} hierarchical level of D_i, where $j \in \{1,\ldots,n_i\}$;
- $\forall j \in \{1,\ldots,n_i\}$ the hierarchical level h_{ij} contains l_{ij} modalities;
- g_{ijt} the t^{th} modality of h_{ij}, where $t \in \{1,\ldots,l_{ij}\}$;
- $G(h_{ij})$ the set of modalities of h_{ij}.

Let's consider the modalities of h_{ij} as the set of individuals. i.e.

$$\Omega = G(h_{ij}) = \{g_{ij1}, \ldots, g_{ijt}, \cdots, g_{ijl_{ij}}\}$$

We adopt now the following notations:

- $*$ a meta-symbol indicating the total aggregate of a dimension;
- $\forall q \in \{1,\ldots,m\}$ we define the measure M_q as the function: $M_q : G \to \Re$;
- G the set of d-tuples of all the hierarchies modalities of the cube C including the total aggregates of dimensions;

$$G = \prod_{i=1}^{d} \left(G(h_{ij}) \underset{j\in\{1,\ldots,n_i\}}{\cup} \{*\} \right)$$
$$= \left(G(h_{1j}) \underset{j\in\{1,\ldots,n_1\}}{\cup} \{*\} \right) \times \cdots \times \left(G(h_{ij}) \underset{j\in\{1,\ldots,n_i\}}{\cup} \{*\} \right) \times \cdots \times \left(G(h_{dj}) \underset{j\in\{1,\ldots,n_d\}}{\cup} \{*\} \right)$$

Reconsider again the cube of Fig.1(a), with the dimensions: D_1 (*Time*), D_2 (*Location*), D_3 (*Product*) and the measure M_1 (*Sales level*). For instance, M_1(*February 1999*, *Lyon*, $*$) indicates the sales level of all products in *February 1999* for the city of *Lyon*.

We adopt the cube measures as quantitative variables describing the population $\Omega = G(h_{ij})$. Nevertheless, in order to insure their statistical and logical validity, it is necessary to respect two fundamental constraints in the choice of these variables.

- **First constraint:** Hierarchical levels belonging to the dimension D_i, retained for the individuals, can not generate variables. In fact, describing an individual by a property which contains it has no logical sense. Conversely, a variable which specifies a property of an individual can only serve for the description of this particular individual;
- **Second constraint:** In order to insure the independence of variables, by dimension, only one hierarchical level can be chosen to generate them. In fact, the value taken by a modality can be obtained by linear combination of modalities belonging to the lower hierarchy.

Therefore, all possible extracted variables belong to the following set:

$$\Sigma \subset \left\{ \begin{array}{l} X / \forall t \in \{1,\ldots,l_{ij}\} \\ \underset{j\in\{1,\ldots,n_j\}}{X(g_{ijt})} = M_q(*,\ldots,*, \underset{j\in\{1,\ldots,n_j\}}{g_{ijt}}, *,\ldots,*, \underset{r\in\{1,\ldots,n_s\}}{g_{srv}}, *,\ldots,*) \\ \text{with } s \neq i,\ r \text{ is unique for each } s,\ v \in \{1,\ldots,l_{sr}\} \text{ and } q \in \{1,\ldots,m\} \end{array} \right\}$$

To enhance the understanding of this formalization, we reconsider the cube of Fig.1(a). Let's suppose that an expert wishes to classify months according to their sales levels by location and/or by product. For this, we retain the modalities of the *Months* level of the dimension D_1 as the set of individuals, i.e. Ω={*January*, *February*, *Mars*, *April*, *May*, *June*}. Thus, we can choose one hierarchical level of D_2 and/or D_3 as generator of variables. For instance, if we choose *Cities* level of D_2, we operate total aggregations (*Roll-up*) on the rest of the cube's dimensions except D_1, the dimension retained for individuals, i.e. we roll-up totally D_3. We obtain a contingency table expressing sales levels by cities at each

month. In the same way, we can generate variables from D_3 by operating a total aggregation on D_2 .

6. Implementation

To illustrate our method, we propose a prototype[1] for the *OpAC* operator. Its implementation was realized with *Visual Basic* under *Windows XP Professional*. The setup of *MS SQL Server* and *MSOLAP driver* is necessary for the running of the prototype. Three principal components constitute our prototype:

- **A parameter setting interface** to assist user in the selection of individuals and variables from a data cube with respect to the above defined constraints. It allows, also, the selection of the clustering's parameters;

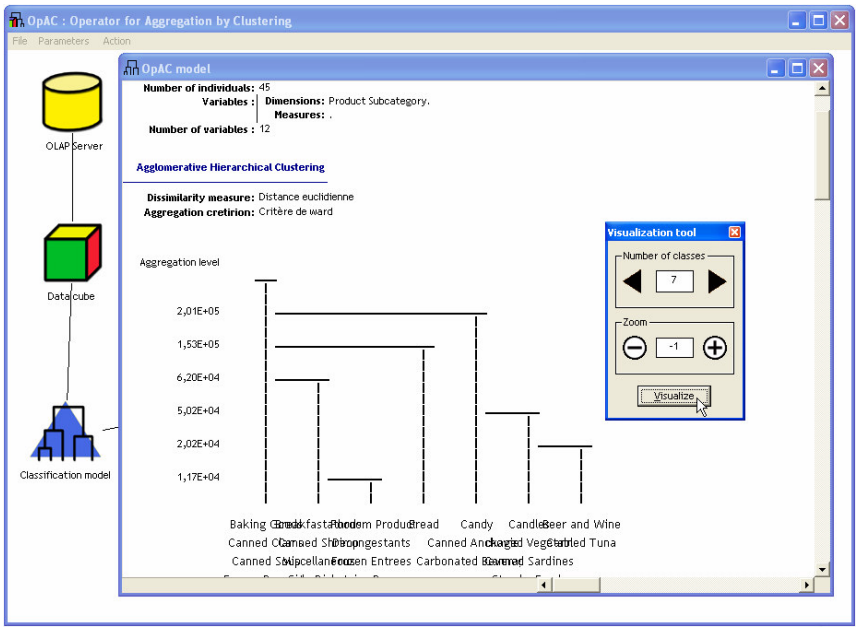


Figure 2. The *OpAC* prototype.

- **A data loading module** that ensures the connection to a data cube via the OLAP server; imports information about the cube's structure (labels of dimensions, hierarchies and measures); and loads data to be analyzed;

[1] http://bdd.univ-lyon2.fr/download/opac.zip

- **A clustering module** to construct the AHC model and plots its results via a dendrogram. The graphic representation of the dendrogram includes a summary of the AHC's parameters and the analyzed data.

As shown in Fig.2, we have provided our prototype with an interactive interface and several visual tools. These tools allow navigation into the dendrogram and a better interpretation of analyzed data.

7. Conclusion

The objective of our study is to satisfy the need of more elaborated on-line analysis. For this, we have created a new aggregation operator which integrates the AHC method into the multidimensional data structure. First, we identified the objectives we plan to *OpAC*. Then, we motivated the choice of AHC as a suited aggregation method. A theoretical formalization was proposed to define the individuals and variables of the clustering problem. We have validated our approach by implementing a prototype.

The *OpAC* operator distinguishes from classical OLAP operators by its ability to aggregate dimension modalities with respect to their semantic bounds. Its aggregates reflect real facts contained in a data cube. Our operator represents a possible way to realize elaborated on-line analysis. Moreover, our choice of the AHC does not exclude the use of other clustering methods. More generally, we think that the use of data mining methods would be suitable to establish new models of on-line learning on multidimensional data.

Finally, the *OpAC* operator can be enhanced in several possible ways. We plan to provide it with an evaluation tool to measure the quality of generated aggregates and to extend it in order to treat as well numerical as complex data cubes.

References

[1] Chaudhuri, S., Dayal, U. (1997) An Overview of Data Warehousing and OLAP Technology. SIGMOD Record. 26(1), 65 – 74.

[2] Chen, M., Zhu, Q.U., Chen, Z.X. (2001) An integrated interactive environment for knowledge discovery from heterogeneous data resources. Information and Software Technology. July 2001, 43(8), 487 – 496.

[3] Chen, Q., Dayal, U., Hsu, M. (2000) An OLAP-based Scalable Web Access Analysis Engine. In: 2nd International Conference on Data Warehousing and Knowledge Discovery (DAWAK'2000). September 2000, London, UK.

[4] Favero, E.L., Robin, J. (2001) Using OLAP and Data Mining for Content Planning in Natural Language Generation. Lecture Notes in Computer Science. 1959, 164 – 175.

[5] Fayyad, U.M., Shapiro, G.P., Smyth, P. et al. (1996) Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

[6] Goil, S., Choudhary, A. (1998) High Performance Multidimensional Analysis and Data Mining. In: High Performance Networking and Computing Conference (SC'98). November 1998, Orlando, USA.

[7] Goil, S., Choudhary, A. (2001) PARSIMONY: An Infrastructure for parallel Multidimensional Analysis and Data Mining. Journal of parallel and distributed computing. 61(3), 285 – 321.
[8] Han, J. (1998) Toward On-line Analytical Mining in Large Databases. In: SIGMOD Record. 27(1), 97 – 107.
[9] Kimball, R. (1996) The Data Warehouse toolkit, John Wiley & Sons.
[10] Lance, G.N. and Williams, W.T. (1967) A general theory of clustering sorting strategies: Clustering systems. The Computer Journal. 10, 271 – 277.
[11] Laurent, A. (2001) De l'OLAP Mining au F-OLAP Mining. Revue Extraction des connaissances et apprentissage (ECA). Hermès (ed.), 1(1-2), 189 – 200.
[12] Pinto, H., Han, J., Dayal, U. et al. (2001) Multi-dimensional Sequential Pattern Mining. In: On Information and Knowledge Management (CIKM'01). November 2001, Atlanta, USA.
[13] Sarawgi, S., Agrawal, R., Megiddo, N. (1998) Discovery-driven Exploration of OLAP Data Cubes. In: Proceeding of the 6th Int'l Conference on Extending Database Technology (EDBT). Mars 1998, Valencia, Spain.
[14] Sarawagi, S. (2001) iDiff: Informative summarization of differences in multidimensional aggregates. Data Mining And Knowledge Discovery. 5(4), 213 – 246.