HAL
open science

OLEMAR: An Online Environment for Mining Association Rules in Multidimensional Data

Riadh Ben Messaoud, Sabine Loudcher Rabaseda, Missaoui Rokia, Omar Boussaïd

To cite this version:

Riadh Ben Messaoud, Sabine Loudcher Rabaseda, Missaoui Rokia, Omar Boussaïd. OLEMAR: An Online Environment for Mining Association Rules in Multidimensional Data. Data Mining and Knowledge Discovery Technologies, Idea Group Inc., pp.35, 2007, Advances in Data Warehousing and Mining. halshs-00476503

HAL Id: halshs-00476503

https://shs.hal.science/halshs-00476503v1

Submitted on 28 Apr 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Running head: *OLEMAR*: an On-Line Environment for Mining Association Rules in Multidimensional Data

OLEMAR: an On-Line Environment for Mining Association Rules in Multidimensional Data

Riadh Ben Messaoud*, Sabine Loudcher Rabaséda*, Rokia Missaoui**, Omar Boussaid*

* Laboratory ERIC - University of Lyon 2,

5, avenue Pierre Mendès-France, 69676, Bron Cedex, France.

** Laboratory LARIM – University of Québec in Outaouais,

C.P. 1250, succursale B, Gatineau (Québec), Canada, J8X 3X7.

ABSTRACT

Data warehouses and OLAP (Online Analytical Processing) provide tools to explore and navigate through data cubes in order to extract interesting information under different perspectives and levels of granularity. Nevertheless, OLAP techniques do not allow the identification of relationships, groupings or exceptions that could hold in a data cube. To that end, we propose to enrich OLAP techniques with data mining facilities to benefit from the capabilities they offer.

In this paper, we propose an on-line environment for mining association rules in data cubes. Our environment, called *OLEMAR* (On-Line Environment for Mining Association Rules), is designed to extract associations from multidimensional data. It allows the extraction of *inter-dimensional* association rules from data cubes according to a *sum-based aggregate measure*, a more general indicator than aggregate values provided by the traditional COUNT measure. In our approach, OLAP users are able to drive a mining process guided by a meta-rule which meets their analysis objectives. In addition, the environment is based on a formalization which exploits aggregate measures to revisit the definition of the support and the confidence of discovered rules. This formalization also helps evaluate the interestingness of association rules according to two additional quality measures: *Lift* and *Loevinger*. Furthermore, in order to focus on the discovered associations and validate them, we provide a visual representation based on the *graphic semiology* principles. Such a representation consists in a graphic encoding of frequent patterns and association rules in the same multidimensional space as the one associated with the mined data cube. We have developed our approach as a component in a general on-line analysis platform, called *Miningcubes*, according to an *Apriori*-like algorithm, which helps extract inter-dimensional association rules directly from materialized multidimensional structures of data. In order to illustrate the

effectiveness and the efficiency of our proposal, we analyze a real-life case study about breast cancer data and conduct performance experimentation of the mining process.

Keywords: Data warehouses, OLAP, data cubes, guided mining, meta-rules, association rules, visualization.

INTRODUCTION

Data warehousing and OLAP (Online Analytical Processing) technologies have gained a widespread acceptance since the 90s as a support for decision making. A data warehouse is a collection of subject-oriented, integrated, consolidated, time-varying and non-volatile data (Kimball, 1996; Inmon, 1996). It is manipulated through OLAP tools which offer visualization and navigation mechanisms of multidimensional data views, commonly called *data cubes*.

A data cube is a multidimensional representation used to view data in a warehouse (Chaudhuri & Dayal, 1997). The data cube contains *facts* or *cells* that have *measures* which are values based on a set of dimensions where each dimension usually consists of a set of categorical descriptors, called *attributes* or *members*. Consider for example a *Sales* application where the dimensions of interest may include, *Costumer*, *Product*, *Location*, and *Time*. If the measure of interest in this application is the *sales amount*, then an OLAP fact represents the sales measure corresponding to a single member in the considered dimensions. A dimension may be organized into a hierarchy. For instance, the location dimension may form the hierarchy *city* → *state* → *region*. Such dimension hierarchies allow different levels of granularity in the data warehouse. For example, a *region* corresponds to a high level of granularity whereas a *city* corresponds to a lower level. Classical aggregation in OLAP considers the process of summarizing data values by moving from a hierarchical level of a dimension to a higher one. Typically, additive data are suitable for simple computation according to aggregation functions (SUM, AVERAGE, MAX, MIN and COUNT). For example, according to such a computation, a user may observe the sum of sales of products according to year and region.

Furthermore, with efficient techniques developed for computing data cubes, users have become widely able to explore multidimensional data. Nevertheless, the OLAP technology is quite limited to an exploratory task and does not provide automatic tools to identify and visualize patterns (e.g., clusters, associations) of huge multidimensional data.

In order to enhance its analysis capabilities, we propose to couple OLAP with data mining mechanisms. The two fields are complementary, and associating them can be a solution to cope with their respective limitations. OLAP technology has the ability to query and analyze multidimensional data through exploration, while data mining is known for its ability to discover knowledge from data. The general issue of coupling database systems with data mining was already discussed and motivated by Imieliński and Mannila (1996). The authors state that data mining leads to new challenges in the database area, and to a second generation of database systems for managing KDD (Knowledge Discovery in Databases) applications just as classical ones manage business ones. More generally, the association of OLAP and data mining allows elaborated analysis tasks exceeding the simple exploration of data. Our idea is to exploit the benefits of OLAP and data mining techniques and to integrate them in the same analysis framework. In spite of the fact that both OLAP and data mining were considered two separate fields for a while, several recent studies showed the benefits of coupling them.

In our previous studies, we have shown the potential of coupling OLAP and data mining techniques through two main approaches. Our first approach deals with the reorganization of data cubes for a better representation and exploration of multidimensional data (Ben Messaoud, Boussaid & Loudcher, 2006a). The approach is based on multiple correspondence analysis (MCA) which allows the construction of new arrangements of modalities in each dimension of a data cube. Such a reorganization aims at bringing together cells in a reduced part of the multidimensional space, and hence giving a better view of the

cube. Our second approach constructs a new OLAP operator for data clustering, called *OpAC* (Ben Messaoud, Boussaid & Loudcher, 2006b), which is based on the agglomerative hierarchical clustering (AHC).

In this paper, we present a third approach which also follows the general issue of coupling OLAP with data mining techniques but concerns the mining of association rules in multidimensional data. In (Ben Messaoud, Loudcher, Boussaid & Missaoui, 2006), we have proposed a guided-mining process of association rules in data cubes. Here, we enrich this proposal and establish a complete On-Line Environment for Mining Association Rules (*OLEMAR*). In fact, it consists of a mining and visualization package for the extraction and the representation of associations from data cubes. Traditionally, with OLAP analysis, we are used to observe summarized facts by aggregating their measures according to groups of descriptors (members) from analysis dimensions. Here, with *OLEMAR*, we propose to use association rules in order to better understand these summarized facts according to their descriptors. For instance, we can note from a given data cube that sales of *sleeping bags* are particularly high in a given city. Current OLAP tools do not provide explanations of such particular fact. Users are generally supposed to explore the data cube according to its dimensions in order to manually find an explanation for a given phenomenon. For instance, one possible interpretation of the previous example consists in associating sales of *sleeping bags* with the *summer season* and *young tourist costumers*.

In the recent years, many studies addressed the issue of performing data mining tasks on data warehouses. Some of them were specifically interested in mining patterns and association rules in data cubes. For instance, Kamber, Han and Chiang (1997) state that it is important to explore data cubes by using association rule algorithms. Further, Imieliński, Khachiyan, and Abdulghani (2002) believe that OLAP is closely interlinked with association rules and shares with them the goal of finding patterns in the data. Goil and Choudhary (1998)

argue that automated techniques of data mining can make OLAP more useful and easier to apply in the overall scheme of decision support systems. Moreover, cell frequencies can facilitate the computation of the support and the confidence, while dimension hierarchies can be used to generate multilevel association rules.

OLEMAR is mainly based on a mining process which explains possible relationships of data by extracting *inter-dimensional* association rules from data cubes (i.e., rules mined from multiple dimensions without repetition of predicates in each dimension). This process is guided by the notion of *inter-dimensional meta-rule* which is designed by users according to their analysis needs. Therefore, the search of association rules can focus on particular regions of the mined cube in order to meet specific analysis objectives. Traditionally, the COUNT measure corresponds to the frequency of facts. Nevertheless, in an analysis process, users are usually interested in observing multidimensional data and their associations according to measures more elaborated than simple frequencies. In our approach, we propose a redefinition of the support and the confidence to evaluate the interestingness of mined association rules when *SUM-based* measures are used. Therefore, the support and the confidence according to the COUNT measure become particular cases of our general definition. In addition to support and confidence, we use two other descriptive criteria (*Lift* and *Loevinger*) in order to evaluate the interestingness of mined associations. These criteria are also computed for *sum-based aggregate measures* in the data cube and reflect interestingness of associations in a more relevant way than what is offered by support and confidence.

The mining algorithm works in a *bottom-up* manner and is an adaptation of the *Apriori* algorithm (Agrawal, Imieliński, and Swami, 1993) to multidimensional data. It is also guided by user's needs expressed through the meta-rule, takes into account a user selected measure in the computation of the support and the confidence, and provides further evaluation of extracted association rules by using *Lift* and *Loevinger* criteria.

In addition to the mining process, the environment also integrates a visual tool which aims at representing the mined frequent patterns and the extracted association rules according to an appropriate graphical encoding based on the *graphic semiology* principles of Bertin (Bertin, 1981). The peculiarity of our visualization component lies on the fact that association rules are represented in a multidimensional space in a similar way as facts (cells).

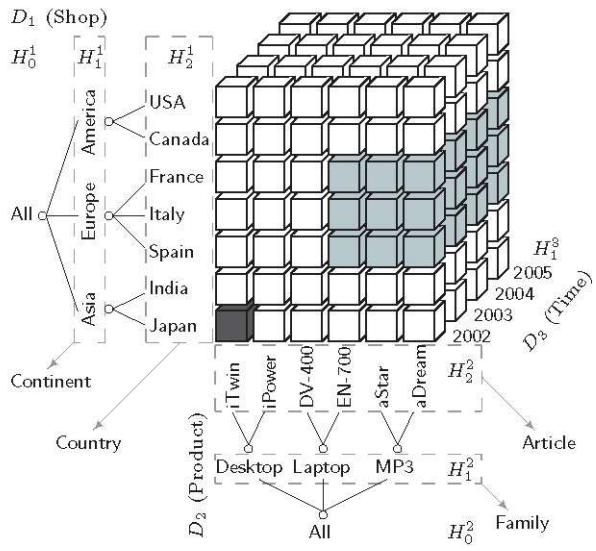
This paper is organized as follows. In the second section, we define the formal background and notions that will be used in the sequel. The third section presents the key concepts of our approach for mining inter-dimensional association rules: the concept of inter-dimensional meta-rule; the general computation of support and confidence based on OLAP measures; and criteria for the advanced evaluation of mined association rules. The fourth section deals with the visualization of the mined inter-dimensional association rules while the fifth section provides the implementation of the on-line mining environment and describes our algorithm for mining inter-dimensional association rules. In the sixth section, we use a case study about mammographies to illustrate our findings while the seventh section concerns the experimental analysis of the developed algorithm. In the eighth section, we present a state of the art about mining association rules in multidimensional data. We also provide a comparative study of existing work and our own proposal. Finally, we conclude this paper and address future research directions.

FORMAL BACKGROUND AND NOTATIONS

In this section, we define preliminary formal concepts and notations we will use to describe our mining process. Let C be a data cube with a non empty set of d dimensions $\mathbf{D} = \{D_1, \ldots, D_i, \ldots, D_d\}$ and a non empty set of measures $\mathbf{M}$. We consider the following notations:

- Each dimension $D_i \in \mathbf{D}$ has a non empty set of hierarchical levels. C;
- H_j^i is the jth ($j \geq 0$) level hierarchical level in D_i. The coarse level of D_i, denoted H_0^i, corresponds to its total aggregation level *All*. For example, in Figure 1, dimension *Shop* (D_1) has three levels: *All*, *Continent*, and *Country*. The *All* level is denoted H_0^1, the *Continent* level is denoted H_1^1, and the *Country* level is denoted H_2^1;
- $\mathbf{H}_i$ is the set of hierarchical levels of dimension D_i, where each level $H_j^i \in \mathbf{H}_i$ consists of a non empty set of members denoted A_{ij}. For example, in Figure 1, the set of hierarchical levels of D_2 is $\mathbf{H}_2 = \{H_0^2, H_1^2, H_2^2\} = \{$*All*, *Family*, *Article*$\}$, and the set of members of the *Article* level of D_2 is $A_{22} = \{$*iTwin*, *iPower*, *DV-400*, *EN-700*, *aStar*, *aDream*$\}$.

Figure 1. Example of Sales data cube



Definition 1. (Sub-cube)

Let $\mathbf{D'} \subseteq \mathbf{D}$ be a non empty set of *p* dimensions $\{D_1, \ldots, D_p\}$ from the data cube C ($p \leq d$). The *p*-tuple ($\Theta_1, \ldots, \Theta_p$) is called a sub-cube on C according to $\mathbf{D'}$ iff $\forall i \in \{1, \ldots, p\}$, $\Theta_i \neq \emptyset$ and there exists a unique *j* such that $\Theta_i \subseteq A_{ij}$.

As defined above, a sub-cube according to a set of dimensions $\mathbf{D'}$ corresponds to a portion from the initial data cube C . It consists in setting for each dimension from $\mathbf{D'}$ a non empty subset of member values from a single hierarchical level of that dimension. For example, consider $\mathbf{D'} = \{D_1, D_2\}$ a subset of dimensions from the cube of Figure 1. (Θ_1, Θ_2) = (*Europe*, {*EN-700*, *aStar*, *aDream*}) is therefore a possible sub-cube on C according to $\mathbf{D'}$, which is displayed by the grayed portion of the cube in the figure. Note that the same portion of the cube can be defined differently by considering the sub-cube ($\Theta_1, \Theta_2, \Theta_3$) = (*Europe*, {*EN-700*, *aStar*, *aDream*}, *All*) according to $\mathbf{D} = \{D_1, D_2, D_3\}$.

One particular case of the sub-cube definition is when it is defined on C according to $\mathbf{D'} = \{D_1, \ldots, D_d\}$ and $\forall i \in \{1, \ldots, d\}$, Θ_i is a single member from the finest hierarchical level of D_i . In this case, the sub-cube corresponds to a cube cell in C . For example, the black cell in Figure 1 can be considered as the sub-cube (*Japan*, *iTwin*, *2002*) on C according to $\mathbf{D} = \{D_1, D_2, D_3\}$. Each cell from the data cube C represents an OLAP fact which is evaluated in $\Re$ according to one measure from $\mathbf{M}$. In our proposal, we evaluate a sub-cube according to its *sum-based aggregate measure* which is defined as follows:

Definition 2. (Sum-based aggregate measure)

Let ($\Theta_1, \ldots, \Theta_p$) be a sub-cube on C according to $\mathbf{D'} \subseteq \mathbf{D}$. The sum-based aggregate measure of sub-cube ($\Theta_1, \ldots, \Theta_p$) according to a measure $M \in \mathbf{M}$, noted M($\Theta_1, \ldots, \Theta_p$), is the SUM of measure *M* of all facts in the sub-cube.

For instance, the *sales turnover* of the grayed sub-cube in Figure 1 can be evaluated by its sum-based aggregate measure according to the expression *Turnover*(*Europe*, {*EN-700*, *aStar*, *aDream*}), which represents the SUM of the *sales turnover* values contained in grayed cells in the *Sales* cube.

Definition 3. (Dimension predicate)

Let D_i be a dimension of a data cube. A dimension predicate α_i in D_i is a predicate of the form $\langle a \in A_{ij} \rangle$.

A dimension predicate is a predicate which takes a dimension member as a value. For example, one dimension predicate in D_1 of Figure 1 can be of the form $\alpha_1 = \langle a \in A_{ij} \rangle = \langle a \in \{America, Europe, Asia\} \rangle$.

Definition 4. (Inter-dimensional predicate)

Let $\mathbf{D'} \subseteq \mathbf{D}$ be a non empty set of p dimensions $\{D_1, \ldots, D_p\}$ from the data cube C ($2 \leq p \leq d$). $(\alpha_1 \wedge \ldots \wedge \alpha_p)$ is called an inter-dimensional predicate in $\mathbf{D'}$ iff $\forall i \in \{1, \ldots, p\}$, α_i is a dimension predicate in D_i.

For instance, let consider $\mathbf{D'} = \{D_1, D_2\}$ a set of dimensions from the cube of Figure 1. An inter-dimensional predicate can be of the form: $(\langle a_1 \in A_{12} \rangle, \langle a_2 \in A_{22} \rangle)$. An inter-dimensional predicate defines a conjunction of non-repetitive predicates, i.e., each dimension has a distinct predicate in the expression.

THE PROPOSED MINING PROCESS

As mentioned earlier, our mining process consists in (i) exploiting meta-rule templates to mine rules from a limited subset of a data cube, (ii) revisiting the definition of support and confidence based on the measure values, (iii) using advanced criteria to evaluate

interestingness of mined associations, and (iv) proposing an *Apriori*-based algorithm for mining multidimensional data.

Inter-Dimensional Meta-Rules

We consider two distinct subsets of dimensions in the data cube C: (i) $D_C \subset \mathbf{D}$ is a subset of p *context dimensions*. A sub-cube on C according to D_C defines the context of the mining process; and (ii) $D_A \subset \mathbf{D}$ is a subset of *analysis dimensions* from which predicates of an inter-dimensional meta-rule are selected. An inter-dimensional meta-rule is an association rule template of the following form:

$$\left| \begin{array}{l} \text{In the context } (\Theta_1,\ldots,\Theta_p) \\ (\alpha_1 \wedge \ldots \wedge \alpha_s) \Rightarrow (\beta_1 \wedge \ldots \wedge \beta_r) \end{array} \right. \tag{1}$$

where $(\Theta_1,\ldots,\Theta_p)$ is a sub-cube on C according to D_C. It defines the portion of cube C to be mined. Unlike the meta-rule proposed in (Kamber, Han & Chiang, 1997), our proposal allows the user to target a mining context by identifying the sub-cube $(\Theta_1,\ldots,\Theta_p)$ to be explored. Note that in the case when $D_C = \emptyset$, no particular analysis context is selected. Therefore, the mining process covers the whole cube C.

We note that $\forall k \in \{1,\ldots,s\}$ (respectively $\forall k \in \{1,\ldots,r\}$), α_k (respectively β_k) is a dimension predicate in a distinct dimension from D_A.

Therefore, the conjunction $(\alpha_1 \wedge \ldots \wedge \alpha_s) \wedge (\beta_1 \wedge \ldots \wedge \beta_r)$ is an inter-dimensional predicate in D_A, where the number of predicates $(s+r)$ in the meta-rule is equal to the number of dimensions in D_A. We also note that our meta-rule defines a non-repetitive predicate association rules since each analysis dimension is associated with a distinct predicate. For instance, suppose that in addition to the three dimensions displayed in Figure 1,

the *Sales* cube contains four other dimensions: *Profile* (D_4), *Profession* (D_5), *Gender* (D_6), and *Promotion* (D_7). Let consider the following subsets from the *Sales* data cube: $D_C = \{D_5, D_6\}$ = {*Profession*, *Gender*}, and $D_A = \{D_1, D_2, D_3\}$ = {*Shop*, *Product*, *Time*}. One possible inter-dimensional meta-rule scheme is:

$$\begin{array}{|l} \text{In the context (}\textit{Student}, \textit{Female}\text{)} \\ \langle a_1 \in \textit{Continent} \rangle \wedge \langle a_3 \in \textit{Year} \rangle \Rightarrow \langle a_2 \in \textit{Article} \rangle \end{array} \tag{2}$$

According to the above inter-dimensional meta-rule, association rules are mined in the sub-cube (*Student*, *Female*) which covers the population of sales concerning female students. The dimensions *Profile* and *Promotion* do not interfere in the mining process. Dimension predicates in D_1 and D_3 are set in the body of the rule whereas the dimension predicate in D_2 is set in the head of the rule. The first dimension predicate is set to the *Continent* level of D_1, the second one is set to the *Year* level of D_3, and the third dimension predicate is set to the *Article* level of D_2.

Measure-Based Support and Confidence

Traditionally, as it was introduced in (Agrawal, Imieliński & Swami, 1993), the support (SUPP) of an association rule $X \Rightarrow Y$, in a database of transactions T, is the probability that the population of transactions contains both X and Y. The confidence (CONF) of $X \Rightarrow Y$ is the conditional probability that a transaction contains Y given that it already contains X. Rules that do not satisfy user provided minimum support (*minsupp*) and minimum confidence (*minconf*) thresholds are considered uninteresting. A rule is said *large*, or *frequent*, if its support is no less than *minsupp*. In addition, a rule is said *strong* if it satisfies both *minsupp* and *minconf*.

In the case of a data cube C, the structure of data facilitates the mining of multidimensional association rules. The aggregate values needed for discovering association rules are already computed and stored in C, which facilitates calculus of the support and the confidence and therefore reduces the testing and the filtering time. In fact, a data cube stores the particular COUNT measure which represents pre-computed frequencies of OLAP facts. With this structure, it is straightforward to calculate support and confidence of associations in a data cube based on this summary information. For instance, suppose that a user needs to discover association rules according to meta-rule (2). In this case one association rule can be $R_1 : America \wedge 2004 \Rightarrow Laptop$. The support and confidence of R_1 are computed as follows:

$$\text{SUPP}(R_1) = \frac{COUNT(America, Laptop, 2004, All, Student, Female, All)}{COUNT(All, All, All, All, Student, Female, All)}$$

$$\text{CONF}(R_1) = \frac{COUNT(America, Laptop, 2004, All, Student, Female, All)}{COUNT(America, All, 2004, All, Student, Female, All)}$$

Note that, in the previous expressions, the support (respectively the confidence) is computed according to the frequency of units of facts based on the COUNT measure. In other words, only the number of facts is taken into account to decide whether a rule is *large* (respectively *strong*) or not. However, in the OLAP context, users are usually interested to observe facts according to summarized values of measures more expressive than their simple number of occurrences. It seems naturally significant to compute the support and the confidence of multidimensional association rules according to the sum of these measures. For example, consider a fragment from the previous *Sales* sub-cube (*Student*, *Female*) by taking once the COUNT measure and then the SUM of the *sales turnover* measure. Table 4 (a) and Table 4 (b) sum-up views of these sub-cube fragments. In this example, for a selected *minsupp*, some itemsets are *large* according to the COUNT measure in Table 4 (a), whereas they are not frequent according to the SUM of the *sales turnover* measure in Table 4 (b), and

vice versa. For instance, with a *minsupp* = 0.2, the itemsets (<America>, <MP3>, <2004>) and (<America >, < MP3>, <2005>) are *large* according to the COUNT measure (grayed cells in Table 4 (a)); whereas, these itemsets are not *large* in Table 4 (b). The *large* itemsets according to the SUM of the profit measure are rather (<Europe>, <Laptop>, <2004>) and (<Europe>, <Laptop>, <2005>).

Table 5. Fragment of the Sales cube according to the (a) COUNT measure and the (b) SUM of the sales turnover measure

	2004		2005	
	America	Europe	America	Europe
Desktop	1,200	800	950	500
Laptop	2,500	2,700	2,800	3,200
MP3	10,600	5,900	11,400	9,100

(a)

	2004		2005	
	America	Europe	America	Europe
Desktop	$ 60,000	$ 33,000	$ 28,000	$ 10,000
Laptop	$ 500,000	$ 567,000	$ 420,000	$ 544,000
MP3	$ 116,000	$ 118,000	$ 57,000	$ 91,000

(b)

In the OLAP context, the rule mining process needs to handle any measure from the data cube in order to evaluate its interestingness. Therefore, a rule is not merely evaluated according to probabilities based on frequencies of facts, but needs to be evaluated according to quantity measures of its corresponding facts. In other words, studied associations do not concern the population of facts, but they rather concern the population of units of measures of these facts. The choice of the measure closely depends on the analysis context according to which a user needs to discover associations within data. For instance, if a firm manager needs to see strong associations of sales covered by achieved profits, it is more suitable to compute the support and the confidence of these associations based on units of profits rather than on units of sales themselves. Therefore, we define a general computation of support and confidence of inter-dimensional association rules according to a user defined (sum-based)

measure *M* from the mined data cube. Consider a general rule *R* which complies with the defined inter-dimensional meta-rule (1):

$$\left| \begin{array}{l} \text{In the context } (\Theta_1, \ldots, \Theta_p) \\ (x_1 \wedge \ldots \wedge x_s) \Rightarrow (y_1 \wedge \ldots \wedge y_r) \end{array} \right.$$

The support and the confidence of this rule are therefore computed according to the following general expressions:

$$\text{SUPP}(R) = \frac{M(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}{M(All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)} \qquad (3)$$

$$\text{CONF}(R) = \frac{M(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}{M(x_1, \ldots, x_s, All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)} \qquad (4)$$

where $M(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, All, \ldots, All)$ is the sum-based aggregate measure of a sub-cube. From a statistical point of view, the collection of facts is not studied according to frequencies but rather with respect to the units of mass evaluated by the OLAP measure *M* of the given facts. Therefore, an association rule $X \Rightarrow Y$ is considered *large* if both *X* and *Y* are supported by a sufficient number of the units of measure *M*. It is important to note that we provide a definition of support and confidence which generalizes the traditional computation of probabilities. In fact, traditional support and confidence are particular cases of the above expressions which can be obtained by the COUNT measure. In the above expressions, in order to insure the validity of our new definition of support and confidence, we suppose that the measure *M* is additive and has positive values.

Advanced Evaluation of Association Rules

Support and confidence are the mostly known measures for the evaluation of association rule interestingness. These measures are key elements of all *Apriori*-like algorithms (Agrawal, Imieliński & Swami, 1993) which mine association rules such that their

support and confidence are greater than user defined thresholds. However, they usually produce a large number of rules which may not be interesting. Various properties of interestingness criteria of association rules have been investigated. For a large list of criteria the reader can refer to (Lallich, Vaillant & Lenca, 2005; Lanca, Vaillant & Lallich, 2006).

Let consider again the association rule $R: X \Rightarrow Y$ which complies with the inter-dimensional meta-rule (1), where $X = (x_1 \wedge \ldots \wedge x_s)$ and $Y = (y_1 \wedge \ldots \wedge y_r)$ are conjunctions of dimension predicates. We also consider a user-defined measure *M* from data cube *C*. We denote by P_X (respectively, P_Y, P_{XY}) the relative measure *M* of facts matching *X* (respectively *Y*, *X* and *Y*) in the sub-cube defined by the instance ($\Theta_1, \ldots, \Theta_p$) in the context dimensions D_C. We also denote by $P_{\overline{X}} = 1 - P_X$ (respectively, $P_{\overline{Y}} = 1 - P_Y$) the relative measure *M* of facts not matching *X* (respectively *Y*), i.e., the probability of not having *X* (respectively *Y*). The support of *R* is equal to P_{XY} and its confidence is defined by the ratio $\frac{P_{XY}}{P_X}$ which is a conditional probability, denoted $P_{X/Y}$, of matching *Y* given that *X* is already matched.

$$P_X = \frac{M(x_1, \ldots, x_s, All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}{M(All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}$$

$$P_Y = \frac{M(All, \ldots, All, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}{M(All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}$$

$$P_{XY} = \text{SUPP}(R) = \frac{M(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}{M(All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}$$

$$P_{Y/X} = \text{CONF}(R) = \frac{M(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}{M(x_1, \ldots, x_s, All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}$$

There are two categories of frequently used evaluation criteria to capture the interestingness of association rules: *descriptive* criteria and *statistical* criteria. In general, one

of the most important drawbacks of a statistical criterion is that it depends on the size of the mined population (Lallich, Vaillant & Lenca, 2005). In fact, when the number of examples in the mined population becomes large, such a criterion loses its discriminating power and tends to take a value close to one. In addition, a statistical criterion requires a *probabilistic* approach to model the mined population of examples. This approach is quite heavy to undertake and assumes advanced statistical knowledge of users, which is not particularly true for OLAP users.

On the other hand, descriptive criteria are easy to use and express interestingness of association rules in a more natural manner. In our approach, in addition to support and confidence, we add two descriptive criteria for the evaluation of mined association rules: the Lift criterion (LIFT) (Brin, Motwani & Silverstein, 1997) and the Loevinger criterion (LOEV) (Loevinger, 1947). These two criteria take the independence of itemsets *X* and *Y* as a reference, and are defined on rule *R* as follows:

$$\text{LIFT}(R) = \frac{P_{XY}}{P_X P_Y} = \frac{\text{SUPP}(R)}{P_X P_Y}$$

$$\text{LOEV}(R) = \frac{P_{Y/X} - P_Y}{P_{\bar{Y}}} = \frac{\text{CONF}(R) - P_Y}{P_{\bar{Y}}}$$

The Lift of a rule can be interpreted as the deviation of the support of the rule from the expected support under the independence hypothesis between the body *X* and the head *Y* (Brin, Motwani & Silverstein, 1997). For the rule *R*, the Lift captures the deviation from the independence of *X* and *Y*. This also means that the Lift criterion represents the probability scale coefficient of having Y when *X* occurs. For example, LIFT(*R*) = 2 means that facts matching with X have twice more chances to match with Y. As opposed to the confidence, which considers directional implication, the Lift directly captures correlation between body *X* and its head *Y*. In general, greater Lift values indicate stronger associations.

In addition to support and confidence, the Loevinger criterion is one of the oldest used interestingness evaluations for association rules (Loevinger, 1947). It consists in a linear transformation of the confidence in order to enhance it. This transformation is achieved by centering the confidence on P_Y and dividing it by the scale coefficient $P_{\overline{Y}}$. In other words, the Loevinger criterion normalizes the centered confidence of a rule according to the probability of not satisfying its head.

THE VISUALIZATION OF INTER-DIMENSIONAL ASSOCIATION RULES

In addition to the previous mining process, our on-line mining environment includes facilities for a graphic representation of the mined inter-dimensional association rules. This representation offers an easier access to the knowledge expressed by a huge number of mined associations. Users can therefore get more insight about rules and easily focus on interesting ones. A particular feature of our visualization solution consists in representing association rules in a multidimensional way so that they can be explored like any part of the data cube.

Traditionally, a user observes the measures associated with facts (cells) in a data cube according to a set of dimensions in a multidimensional space. In our visualization approach, we embed in this space representation a graphic encoding of inter-dimensional association rules. This encoding refers to the principles of *graphic semiology* of Bertin (Bertin, 1981). Such principles consist to organize the visual and perceptual components of graphics according to features and relations between data. They mainly use the visual variables of *position*, *size*, *luminosity*, *texture*, *color*, *orientation* and *form*. The position variable has a particular impact on human retention since it concerns dominant visual information from a perceptual point of view. The other variables have rather a retinal property since it is quite

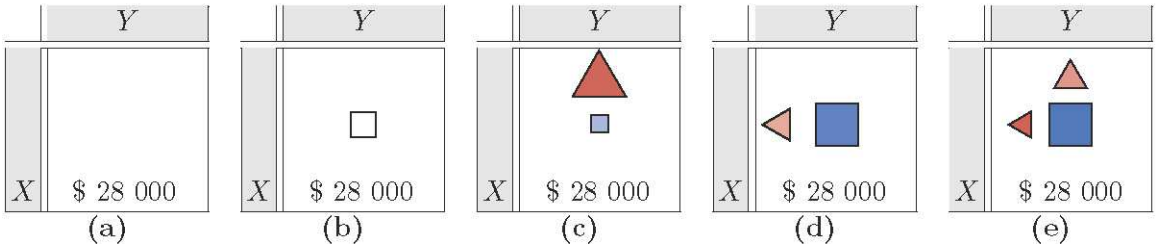
possible to see their variations independently from their positions. The size variable generally concerns surfaces rather than lengths. According to Bertin, the variation of surfaces is a sensible stimulus for the variation of size and more relevant to human cognition than variation of length.

We note that the position of each cell in the space representation of a data cube is important since it represents a conjunction of predicate instances. For instance, let *c* be a cell in the space representation of the data cube *C*. The position of *c* corresponds to the intersection of row *X* with column *Y*. *X* and *Y* are conjunctions of modalities where each modality comes from a distinct dimension. In other words, *X* and *Y* are inter-dimensional instance predicates in the analysis dimensions retained for the visualization. Therefore, cell *c* corresponds to the itemset $\{X,Y\}$. According to the properties of the itemset $\{X,Y\}$, we propose to represent the appropriate graphic encoding as follows (see Figure 2):

- if $\{X,Y\}$ is not frequent, only the value of the measure *M*, if it exists, is represented in cell *c*;
- if $\{X,Y\}$ is frequent and it does not generate association rules, a white square is represented in cell *c*;
- if $\{X,Y\}$ is frequent and generates the association rule $X \Rightarrow Y$, a blue square and a red triangle are displayed in cell *c*. The triangle points to *Y* according to the implication of the rule;
- if $\{X,Y\}$ is frequent and generates the association rule $Y \Rightarrow X$, a blue square and a red triangle are displayed in cell *c*. The triangle points to *X* according to the implication of the rule;
- if $\{X,Y\}$ is frequent and generates the association rules $X \Rightarrow Y$ and $Y \Rightarrow X$, a blue square and two red triangles are displayed in cell *c*. The first triangle

points to Y according to the implication of the rule $X \Rightarrow Y$, and the second triangle points to X according to the implication of the rule $Y \Rightarrow X$.

Figure 2. Examples of association rule representations in a cube cell



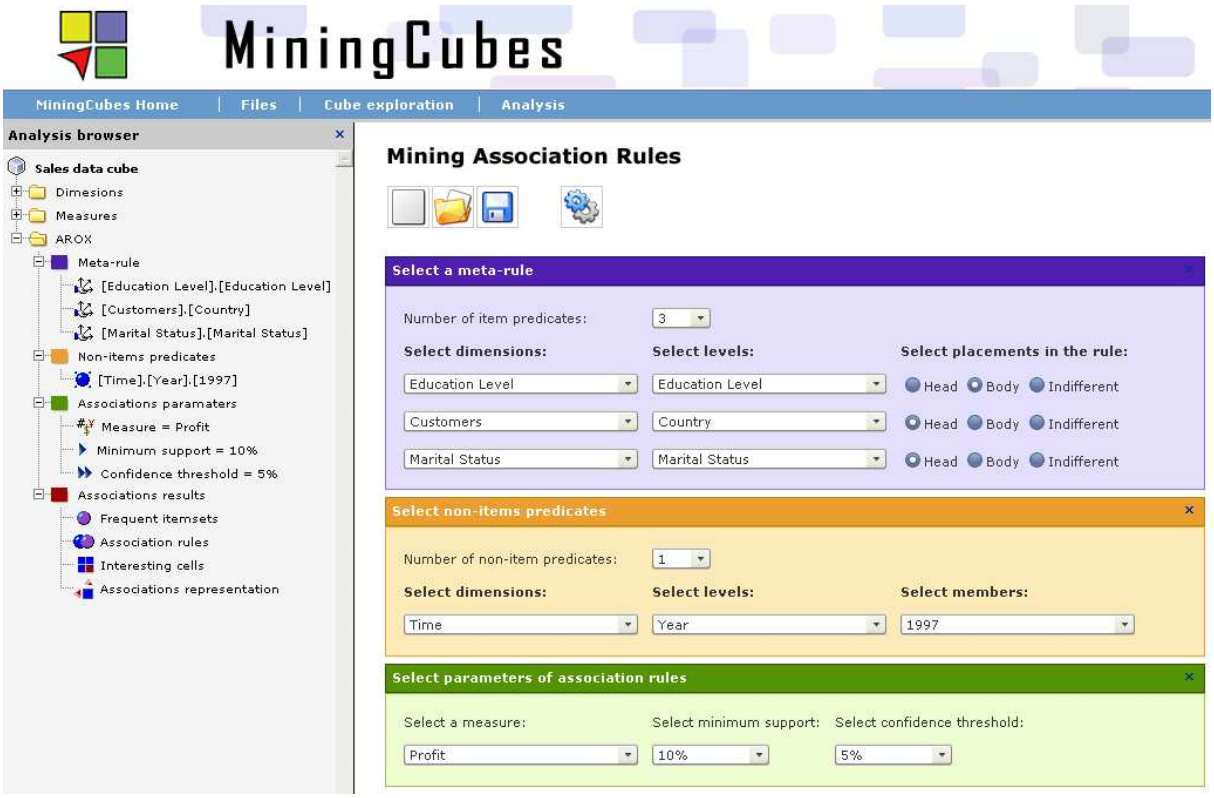
For a given association rule, we use two different forms and colors to distinguish between the itemset of the rule and its implication. In fact, the itemset $\{X, Y\}$ is graphically represented by a blue square and the implication $X \Rightarrow Y$ is represented by a red equilateral triangle. We also use the surface of the previous forms in order to encode the importance of the support and the confidence. The support of the itemset $\{X, Y\}$is represented by the surface of the square and the confidence of the rule $X \Rightarrow Y$ is represented by the surface of the triangle. Since the surface is one of the most relevant variables to human perception, we use it to encode most used criteria to evaluate the importance of an association rule. For high values of the support (respectively, the confidence), the blue square (respectively, the red triangle) has a large surface, while low values correspond to small surfaces of the form. Therefore, the surfaces are proportionally equal to the values of the support and the confidence.

The Lift and the Loevinger criteria are highlighted with the luminosity of their respective forms. We represent high values of the Lift (respectively, the Loevinger criterion) by a low luminosity of the blue square (respectively, the red triangle). We note that a high luminosity of a form corresponds to a pale color, whereas, a low luminosity of a form corresponds to a dark color.

IMPLEMENTATION AND ALGORITHMS

We have developed *OLEMAR* as a module of on a Client/Server analysis platform, called *MiningCubes*, which already includes our previous proposals dealing with coupling OLAP and data mining (Ben Messaoud, Boussaid & Loudcher, 2006a; Ben Messaoud, Boussaid & Loudcher, 2006b). *MiningCubes* is equipped with a *data loader component* that enables connection to multidimensional data cubes stored in *Analysis Services* of *MS SQL Server 2000*. The *OLEMAR module* allows the definition of required parameters to run an association rule mining process. In fact, as shown in the interface of Figure 3, a user is able to define analysis dimensions D_A, context dimensions D_C, a meta-rule with its context sub-cube ($\Theta_1,\ldots,\Theta_p$) and its inter-dimensional predicates scheme $(\alpha_1 \wedge \ldots \wedge \alpha_s) \Rightarrow (\beta_1 \wedge \ldots \wedge \beta_r)$, the measure *M* used to compute quality criteria of association rules, and the thresholds *minsupp* and *minconf*.

Figure 3. Interface of the OLEMAR module in MiningCubes



The generation of association rules from a data cube closely depends on the search for *large* (frequent) itemsets. Traditionally, frequent itemsets can be mined according to two different approaches:

- the *top-down* approach which starts with *k*-itemsets and steps down to 1-itemsets. The decision whether an itemset is frequent or not is directly based on the *minsupp* value. In addition, it assumes that if a *k*-itemset is frequent, then all sub-itemsets are frequent too;
- the *bottom-up* approach which goes from 1-itemsets to larger itemsets. It complies with the *Apriori* property of *anti-monotony* (Agrawal, Imieliński & Swami, 1993) which states that *for each non frequent itemset, all its super-itemsets are definitely not frequent*.

The previous property enables the reduction of the search space, especially when it deals with large and sparse data sets, which is particularly the case of OLAP data cubes. We implemented the mining process by defining an algorithm based on the *Apriori* property

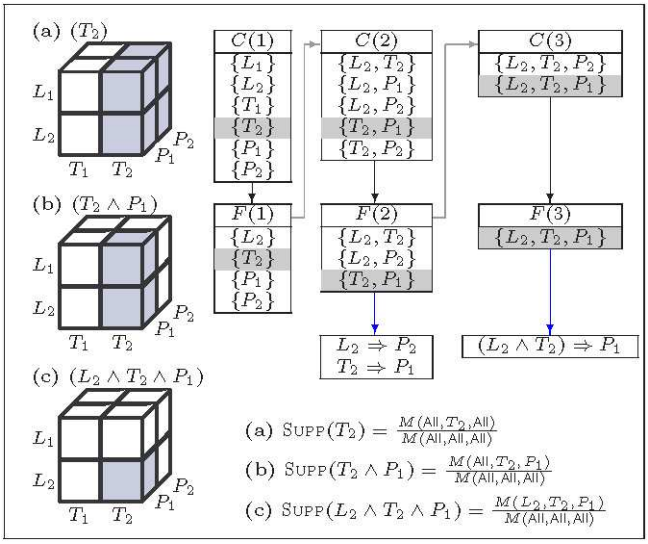
according to a *bottom-up* approach for searching large itemsets. As summarized in Algorithm 1, we proceed by an increasing level wise search for large *i*-itemsets, where *i* is the number of items in the itemset. We denote by $C(i)$ the sets of *i-candidates*, i.e., *i*-itemsets that are potentially frequent, and $F(i)$ the sets of *i-frequents*, i.e., frequent *i*-itemsets.

At **the initialization step**, our algorithm captures the 1-candidates from user defined analysis dimensions D_A over the data cube *C*. These 1-candidates correspond to members of D_A, where each member complies with one dimension predicate α_k or β_k in the meta-rule *R*. In other words, for each dimension D_i of D_A, we capture 1-candidates from A_{ij}, which is the set of members of the jth hierarchical level of D_i selected in its corresponding dimensional predicate in the meta-rule scheme. For example, let consider the data cube of Figure 4. We assume that, according to a user meta-rule, mined association rules need to comply with the meta-rule scheme:

$$\langle a_1 \in \{L_1, L_2\}\rangle \wedge \langle a_2 \in \{T_1, T_2\}\rangle \Rightarrow \langle a_3 \in \{P_1, P_2\}\rangle .$$

Therefore, the set of 1-candidates is: $C(1) = \{\{L_1\},\{L_2\},\{T_1\},\{T_2\},\{P_1\},\{P_2\}\}$.

Figure 4. Example of a bottom-up generation of association rules from a data cube



For each level *i*, if the set $C(i)$ is not empty and *i* is less than $s+r$, **the first step** of the algorithm derives frequent itemsets $F(i)$ from $C(i)$ according to two conditions: (i) an itemset $A \in C(i)$ should be an instance of an inter-dimensional predicates in D_A, i.e., *A* must be a conjunction of members from *i* distinct dimensions from D_A; and (ii) in addition to the previous condition, to be included in $F(i)$, an itemset $A \in C(i)$ must have a support greater than the minimum support threshold *minsupp*. As shown in Figure 4, SUPP(*A*) is a measure-based support computed according to a user selected measure *M* from the cube.

Algorithm 1. The algorithm for mining inter-dimensional association rules from data cubes

```
input  : C, D_C, D_A, D_U, R, M, minsupp, minconf
output: X ⇒ Y, SUPP, CONF, LIFT, LOEV
C(1) ← ∅;
for i ← 1 to (s + r) do
  | C(1) ← C(1) ∪ A_ij;
end
i ← 1;
while C(i) ≠ ∅ and i ≤ (s + r) do
    F(i) ← ∅;
    foreach A ∈ C(i) do
        if A is an inter-dimensional predicates then
            SUPP ← COMPUTESUPPORT(A, M);
            if SUPP ≥ minsupp then F(i) ← F(i) ∪ {A};
        end
    end
    foreach A ∈ F(i) do
        foreach non empty B ∈ A do
            if A\B ⇒ B complies with R then
                CONF ← COMPUTECONFIDENCE(A\B, B, M);
                if CONF ≥ minconf then
                    X ← A\B;
                    Y ← B;
                    LIFT ← COMPUTELIFT(X, Y, M);
                    LOEV ← COMPUTELOEVINGER(X, Y, M);
                    return (X ⇒ Y, SUPP, CONF, LIFT, LOEV);
                end
            end
        end
    end
    C(i + 1) ← ∅;
    foreach A ∈ F(i) do
        foreach B ∈ F(i) that shares i − 1 items with A do
            if All Z ⊂ {A ∪ B} of i items are
            inter-dimensional predicates and frequent then
                C(i + 1) ← C(i + 1) ∪ {A ∪ B};
            end
        end
    end
    i ← i + 1;
end
```

From each $A \in F(i)$, **the second step** extracts association rules based on two conditions: (i) an association rule $X \Rightarrow Y$ must comply with the user defined meta-rule *R*, i.e., items of *X* (respectively, items of *Y*) must be instances of dimension predicates defined in

the body (respectively, in the head) of the meta-rule scheme of *R*. For example, in Figure 4, $P_2 \Rightarrow L_2$ can not be derived from $F(2)$ because, according to the previous meta-rule scheme, instances of $\langle a_1 \in \{L_1, L_2\}\rangle$ must be in the body of mined rules and not in their head; and (ii) an association rule must have a confidence greater than the minimum confidence threshold *minconf*. The computation of confidence is also based on the user defined measure *M*. When an association rule satisfies the two previous conditions, the algorithm computes its Lift and Loevinger criteria according to the formulae we gave earlier. Finally, the rule, its support, confidence, Lift and Loevinger criteria are returned by the algorithm.

Based on the *Apriori* property, **the third step** uses the set $F(i)$ of large *i*-itemsets to derive a new set $C(i+1)$ of $(i+1)$-candidates. A given $(i+1)$-candidate is the union of two *i*-itemsets *A* and *B* from $F(i)$ that verifies three conditions: (i) *A* and *B* must have $i-1$ common items; (ii) all non empty sub-itemsets from $A \cup B$ must be instances of inter-dimensional predicates in D_A; and (iii) all non empty sub-itemsets from $A \cup B$ must be frequent itemsets. For example in Figure 4, itemsets $A = \{L_2, T_2\}$ and $B = \{L_2, P_2\}$ from $F(2)$ have $\{L_2\}$ as a common 1-itemset, all non empty sub-itemsets from $A \cup B = \{L_2, T_2, P_2\}$ are frequents and represent instances of inter-dimensional predicates. Therefore, $\{L_2, T_2, P_2\}$ is a 3-candidate included in $C(3)$.

Note that the computation of support, confidence, Lift, and Loevinger criteria are performed respectively by the functions: COMPUTESUPPORT, COMPUTECONFIDENCE, COMPUTELIFT and COMPUTELOEVINGER. These functions take the measure *M* into account and are implemented using MDX (Multi-Dimensional eXpression language in *MS SQL Server 2000*) that provides required pre-computed aggregates from the data cube. For instance, reconsider the *Sales* data cube of Figure 1, the meta-rule (2), and the rule R_1:

$America \wedge 2004 \Rightarrow Laptop$. According to formula (3) and considering the *sales turnover* measure, the support of R_1 is written as follows:

$$\text{SUPP}(R_1) = \frac{Sales_turnover(America, Laptop, 2004, All, Student, Female, All)}{Sales_turnover(All, All, All, All, Student, Female, All)}$$

The numerator value of SUPP(R_1) is therefore returned by the following MDX query:

```
SELECT
     NON EMPTY {[Shop].[Continent].[America]} ON AXIS(0),
     NON EMPTY {[Time].[Year].[2004]} ON AXIS(1),
     NON EMPTY {[Product].[Family].[Laptop]} ON AXIS(2)
FROM Sales
WHERE     ([Measures].[Sales_turnover],
          [Profession].[Profession category].[Student],
          [Gender].[Gender].[Female])
```

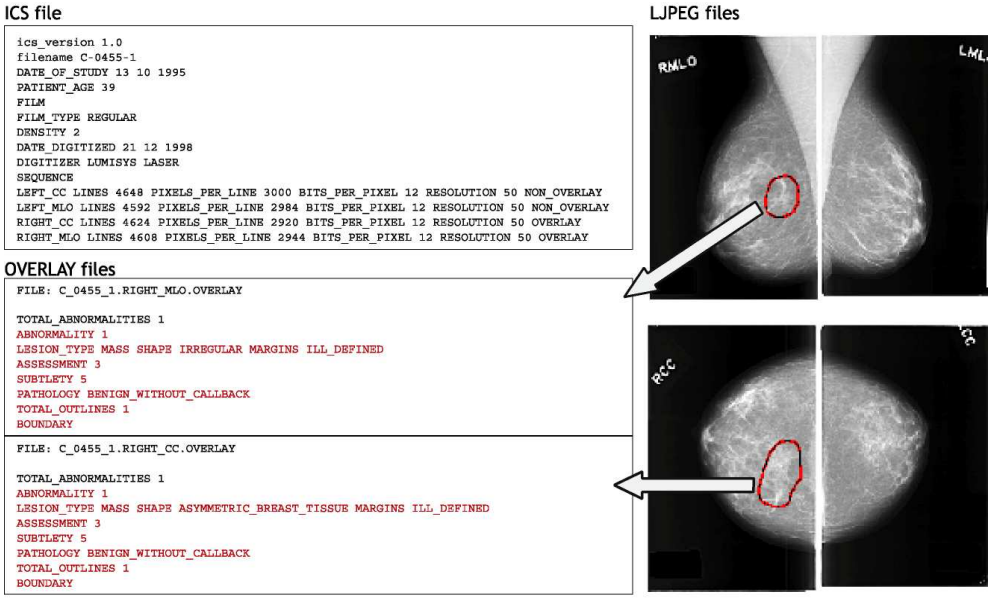
A CASE STUDY

In order to validate our approach, this section presents the results of a case study conducted on clinical data dealing with the breast cancer research domain. More precisely, data refer to suspicious regions extracted from the *Digital Database for Screening Mammography* (DDSM). In the following, we present the DDSM and the generated data cube.

The Digital Database for Screening Mammography (DDSM)

The DDSM is basically a resource used by the mammography image analysis research community in order to facilitate sound research in the development of analysis and learning algorithms (Heath, Bowyer, Kopans, Moore & Jr, 2000). The database contains approximately 2 600 studies, where each study corresponds to a patient case.

Figure 5. An example of a patient case study taken from the DDSM

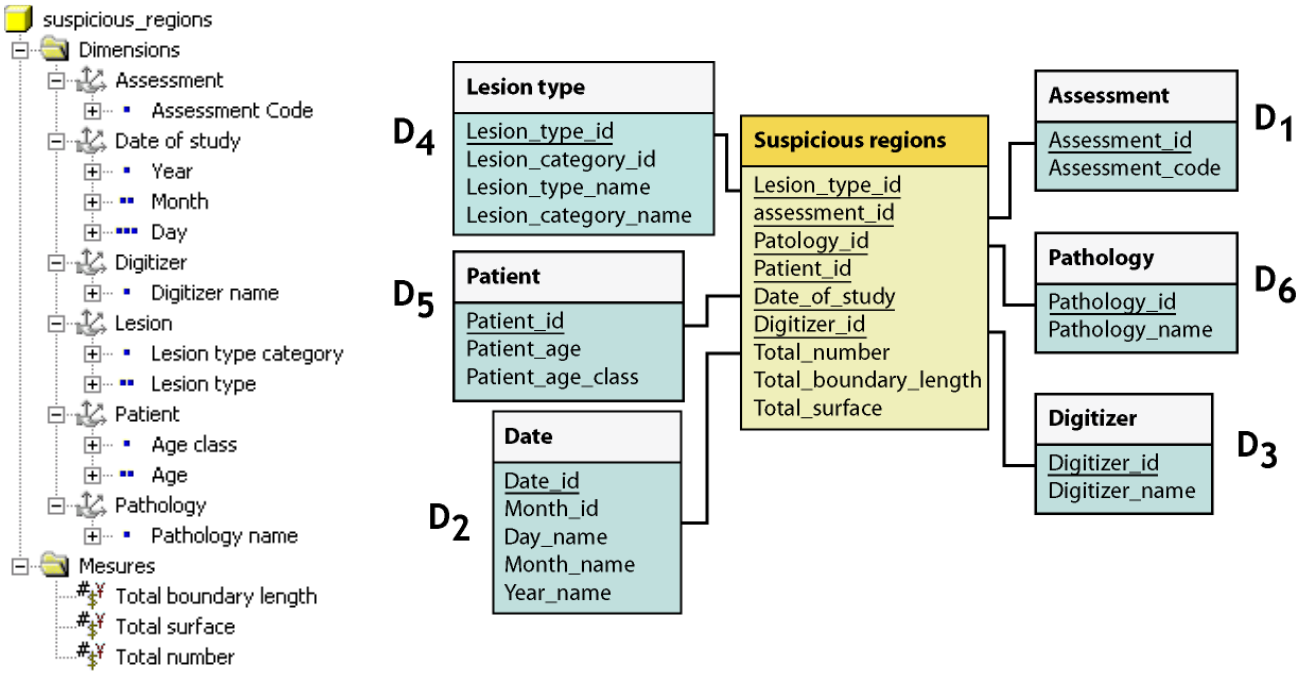


As shown in Figure 5, a patient case is a collection of images and text files containing medical information collected along exams of screening mammography. The DDSM contains four types of patient cases: *Normal*, *Benign without callback*, *Benign*, *Cancer*. *Normal* type covers mammograms from screening exams that were read as normal and had a normal screening exam. *Benign without callback* cases are exams that had an abnormality that was noteworthy but did not require the patient to be recalled for any additional checkup. In *Benign* cases, something suspicious was found and the patient was recalled for some additional checkup that resulted in a benign finding. *Cancer* type corresponds to cases in which a proven cancer was found.

The Suspicious Regions Data Cube

A patient file refers to different data formats and encloses several subjects that may be studied according to various points of view. In our case study, we focus on studying the screening mammography data by considering suspicious regions (abnormalities) detected by an expert as an OLAP fact.

Figure 6. (a) the physical structure, and (b) the conceptual model of the suspicious regions data cube



(a) **(b)**

Under *Analysis Services* of *MS SQL Server 2000*, we have constructed the *suspicious regions data cube* from the DDSM data. Our data cube contains 4 686 OLAP facts. Figure 6 (a) and Figure 6 (b) illustrate, respectively, the physical structure and the conceptual model of the constructed cube as they are presented in the cube editor of *Analysis Services*. According to this data cube, a set of suspicious regions can be analyzed according to several axes: the *lesion*, the *assessment*, the *subtlety*, the *pathology*, the *date of study*, the *digitizer*, the *patient*, etc. The fact is measured by the *total number* of regions, the *total boundary length*, and the *total surface* of the suspicious regions. We note that, in this cube, the set of concerned facts deals only with *Benign*, *Benign without callback*, and *Cancer* patient cases. *Normal* cases are not concerned since they do not contain suspicious regions.

Application on the Suspicious Regions Data Cube

We have applied our on-line mining environment on the suspicious regions data cube *C*. To illustrate this mining process, we suppose that an expert radiologist looks for associations that could explain the reasons of cancer tumors. We assume that the expert restricts his study to suspicious regions found on scanners of mammograms digitized thanks to a *Lumisis Laser* machine. This means that the subset of context dimensions $\mathbf{D}_C$ contains the dimension *Digitizer* (D_3) and the selected context corresponds to the sub-cube (*Lumisis Laser*) according to $\mathbf{D}_C$. We also suppose that the expert needs to explain the different types of pathologies in these mammograms. In order to do so, he chooses to explain the modalities of the *Pathology name* level (H_1^6), included in the dimension *Pathology* (D_6), by both those of the *Assessment code* level (H_1^1), from dimension *Assessment* (D_1), and those of the *Lesion type category* level (H_1^4), from dimension *Lesion* (D_4). In other words, the subset of analysis dimensions $\mathbf{D}_A$ consists of the dimensions *Assessment* (D_1), *Lesion* (D_4) and *Pathology* (D_6). Thus, according to our formalization:

- the subset of context dimensions is $\mathbf{D}_C = \{D_3\} = \{Digitizer\}$;
- the subset of analysis dimension is $\mathbf{D}_A = \{D_1, D_4, D_6\} = \{Assessment, Lesion, Pathology\}$.

Therefore, with respect to the previous subset of dimensions, to guide the mining process of association rules, the expert needs to express the following inter-dimensional meta-rule:

> In the context (*Lumisis Laser*)
> $\langle a_1 \in Assessment\ code \rangle \wedge \langle a_4 \in Lesion\ type\ category \rangle \Rightarrow \langle a_6 \in Pathology\ name \rangle$

Note that, in order to explain the pathologies of suspicious regions, the dimension predicate in D_6 ($\langle a_6 \in Pathology\ name \rangle$) is set to the head of the meta-rule (conclusion) whereas the other dimension predicates ($\langle a_4 \in Lesion\ type\ category \rangle$ and $\langle a_1 \in Assessment\ code \rangle$) are rather set to its body (consequence).

Assume that *minsupp* and *minconf* are set to 5%, and *Surface* of suspicious regions is the measure on which the computation of the support, the confidence, the Lift, and the Loevinger criteria will be based. The guided mining process provides the association rules that we summarize as follows:

	Association rule *R*	**SUPP**	**CONF**	**LIFT**	**LOEV**
1	{*All, Calcification type pleomorphic*} $\Rightarrow$ {*Benign*}	5.03%	24.42%	0.73	-0.14
2	{***3***, *All*} $\Rightarrow$ {***Cancer***}	5.15%	8.50%	0.60	-0.62
3	{*0, All*} $\Rightarrow$ {*Benign*}	5.60%	66.72%	1.99	0.50
4	{***4***, *Calcification type pleomorphic*} $\Rightarrow$ {***Cancer***}	6.10%	61.05%	1.01	0.01
5	{*All, Mass shape lobulated*} $\Rightarrow$ {*Cancer*}	6.14%	48.54%	0.80	-0.31
6	{*All, Mass shape lobulated*} $\Rightarrow$ {*Benign*}	6.21%	49.03%	1.47	0.23
7	{*3, All*} $\Rightarrow$ {*Benign*}	7.09%	49.99%	1.99	0.09
8	{*All, Mass shape oval*} $\Rightarrow$ {*Benign*}	8.59%	65.82%	1.97	0.49
9	{***5***, *Calcification type pleomorphic*} $\Rightarrow$ {***Cancer***}	8.60%	98.92%	1.63	0.97
10	{***5***, *Mass shape irregular*} $\Rightarrow$ {***Cancer***}	14.01%	96.64%	1.60	0.91
11	{*All, Calcification type pleomorphic*} $\Rightarrow$ {*Cancer*}	15.43%	74.97%	1.24	0.36
12	{***4***, *All*} $\Rightarrow$ {***Cancer***}	16.43%	46.06%	0.76	-0.37
13	{*4, All*} $\Rightarrow$ {*Benign*}	18.64%	52.29%	1.56	0.28
14	{*All, Mass shape irregular*} $\Rightarrow$ {*Cancer*}	20.38%	87.09%	1.44	0.67
15	{***5***, *All*} $\Rightarrow$ {***Cancer***}	36.18%	98.25%	1.62	0.96

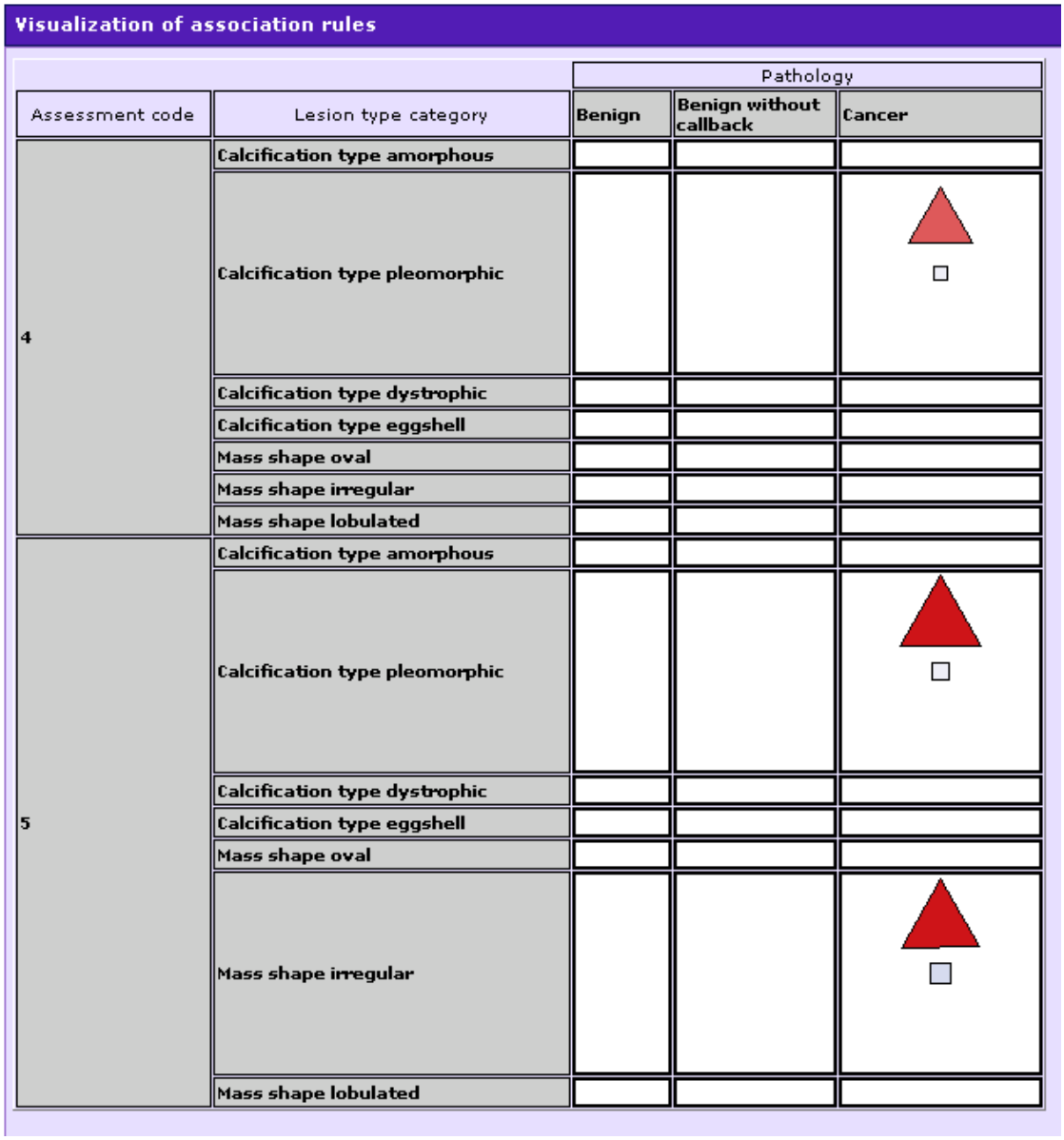
Note that the above association rules comply with the designed inter-dimensional meta-rule, which aims at explaining pathologies according to assessments and lesions. From these associations, an expert radiologist can easily note that cancer cases of suspicious regions are mainly caused by high values of assessment codes. For example, rule R_{15}:{*5, All*} $\Rightarrow$ {*Cancer*} is supported by 36.18% of surface units of suspicious regions. In addition,

its confidence is equal to 98.25%. In other words, knowing that a suspicious region has an assessment code of 5, the region has 98.25% chances to be a cancer tumor. Rule R_{15} has also a Lift equal to 1.62, which means that the total surface of cancer suspicious regions having an assessment code equal to 5 is 1.62 times greater than the expected total surface under the independence situation between the assessment and the pathology type.

The lesion type can also explain pathologies. From the previous results, we note that the *mass shape irregular* and the *calcification type pleomorphic* are the major lesions leading to cancers. In fact, rules R_{11}:{*All, Calcification type pleomorphic*} $\Rightarrow$ {*Cancer*} and R_{14}:{*All, Mass shape irregular*} $\Rightarrow$ {*Cancer*} confirm this observation with supports respectively equal to 15.43% and 20.38%, and confidences respectively equal to 74.97% and 87.09%.

Recall that our on-line mining environment is also able to provide an interactive visualization of its extracted inter-dimensional association rules. Figure 7 shows a part of the data cube where association rules R_4, R_9, and R_{10} are displayed in the visualization interface.

Figure 7. Visualization of extracted association rules in MiningCubes



PERFORMANCE EVALUATION

We have evaluated the performance of our mining process algorithm for the suspicious regions data cube. We conducted a set of experiments to measure time processing for different situations of input data and parameters of the *OLEMAR module* supported by *MiningCubes*. These experiments are achieved under Windows XP on a 1.60GHz PC with

480MB of main memory, and an Intel Pentium 4 processor. We also used *Analysis Services* of *MS SQL Server 2000* as a *local-host* OLAP server.

Figure 8. The running times of the mining process according to support with different confidences

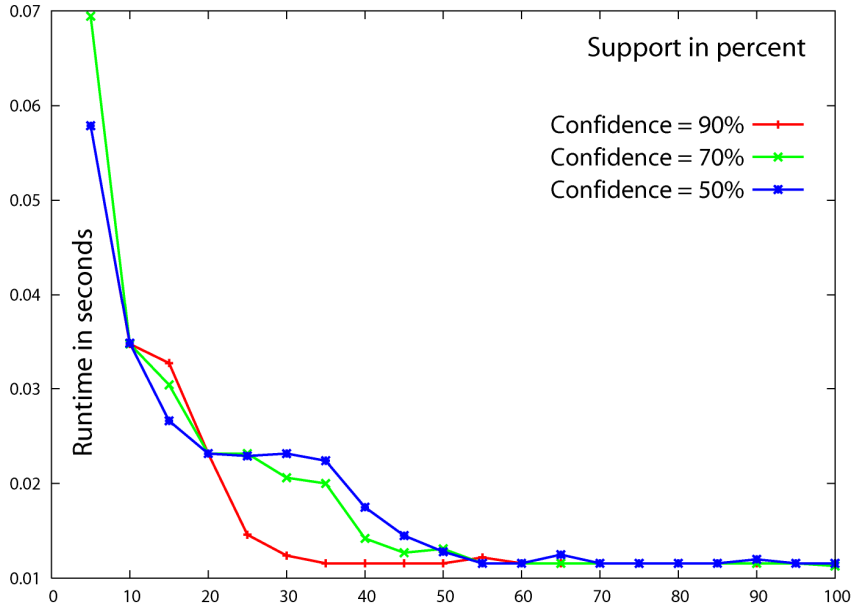


Figure 8 shows the relationship between the runtime of our mining process and the support of mined association rules according to several confidence thresholds. In general, the mining of association rules needs less time when it deals with increasing values of the support.

Figure 9. The running times of the mining process according to support with different numbers of facts

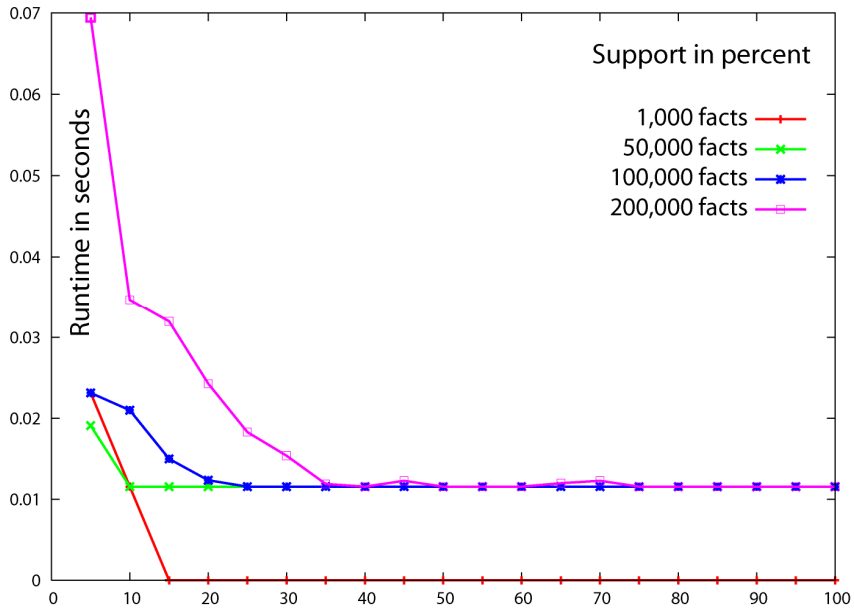
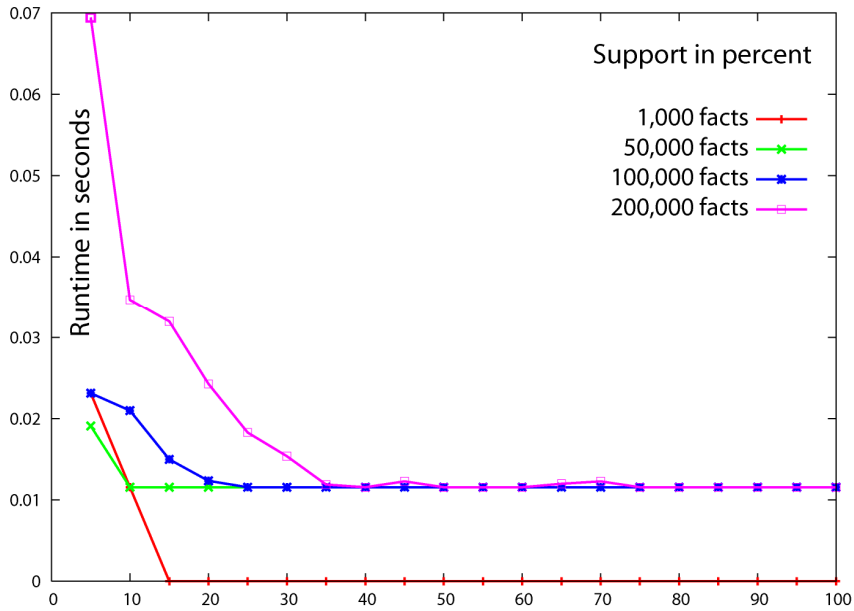


Figure 9 presents a test of our algorithm for several numbers of facts. For small support values, the running time considerably increases with the number of mined facts. However, for large supports, the algorithm has already equal response times independently from the number of mined facts. Another point of view of this phenomenon can be illustrated by Figure 10 which indicates that for a support and a confidence threshold equal to 5%, the efficiency of the algorithm closely depends on the number of extracted frequent itemsets and association rules. The running time obviously increases according to the number of extracted frequent itemsets and association rules. Nevertheless, the generation of association rules from frequent itemsets is more time consuming than the extraction of frequent itemsets themselves.

An *Apriori*-based algorithm is efficient for searching frequent itemsets and has a low complexity level especially in the case of sparse data. Nevertheless, the *Apriori* property does not reduce the running time of extracting association rules from a frequent itemset. For each frequent itemset, the algorithm must generate all possible association rules that comply with the meta-rule scheme and search those having a confidence greater than *minconf*.

Figure 10. The running times the mining process according to the number of frequent itemsets and the number of association rules



In general, these experiments highlight acceptable runtime processing. The efficiency of our algorithm is due to: (i) the use of inter-dimensional meta-rules which reduce the search space of association rules and therefore, considerably decreases the runtime of the mining process; (ii) the use of pre-computed aggregates of the multidimensional cube which helps compute the support and the confidence via MDX queries; and (iii) the use of the anti-monotony property of *Apriori* which is definitely suited to sparse data cubes and considerably reduces the complexity of large itemsets search.

RELATED WORK

Association Rule Mining in Multidimensional Data

Association rule mining was first introduced by Agrawal, Imieliński, and Swami (1993) who were motivated by *market basket analysis* and designed a framework for extracting rules from a set of transactions related to items bought by customers. They also proposed the *Apriori* algorithm that discovers *large* (*frequent*) itemsets satisfying given

minimal support and confidence. Since then, many developments have been performed in order to handle various types and data structures.

To the best of our knowledge, Kamber, Han and Chiang (1997) were the first researchers who addressed the issue of mining *association rules from multidimensional data*. They introduced the concept of *meta-rule-guided mining* which consists in using rule templates defined by users in order to guide the mining process.They provide two kinds of algorithms for extracting association rules from data cubes: (1) algorithms for materialized MOLAP (*Multidimensional OLAP*) data cubes; and (2) algorithms for non-materialized ROLAP (*Relational OLAP*) data cubes. These algorithms can mine *inter-dimensional* association rules, with distinct predicates, from single levels of dimensions. An *inter-dimensional* association rule is mined from multiple dimensions without repetition of predicates in each dimension, while an *intra-dimensional* association rule cover repetitive predicates from a single dimension. The support and the confidence of mined associations are computed according to the COUNT measure.

Zhu considers the problem of mining three types of associations: *inter-dimensional*, *intra-dimensional*, and *hybrid* rules (Zhu, 1998). The latter type consists in combining intra and inter-dimensional association rules. Unlike Kamber, Han and Chiang (1997) - where associations are directly mined from multidimensional data – Zhu (1998) generates a *task-relevant working cube* with desired dimensions, flattens it into a tabular form, extracts frequent itemsets, and finally mines association rules. Therefore, this approach does not profit from hierarchical levels of dimensions since it flattens data cubes in a pre-processing step. In other words, it adapts multidimensional data and prepares them to be handled by classical iterative association mining process. Further, the proposal uses the COUNT measure and does not take into account further aggregate measures to evaluate discovered rules. We also note the lack of a general formalization for the proposed approach.

Cubegrades, proposed in (Imieliński, Khachiyan & Abdulghani, 2002), are a generalization of association rules. They focus on significant changes that affect measures when a cube is modified through specialization (*drill-down*), generalization (*roll-up*) or mutation (*switch*). The authors argue that traditional association rules are restricted to the COUNT aggregate and can only express relative changes from body of the rule to body and head. In a similar way, Dong, Han, Lam, Pei and Wang (2001) proposed an interesting and efficient version of the *cubegrade* problem, called *multidimensional constrained gradients*, which also seeks significant changes in measures when cells are modified through generalization, specialization or mutation. To capture significant changes only and prune the search space, three types of constraints are considered. The concept of *cubegrades* and *constrained gradients* is quite different from classical mining of association rules. It discovers modifications on OLAP aggregates when moving from a *source-cube* to a *target-cube*, but it is not capable of searching patterns and association rules included in the cube itself. We consider a *cubegrade* as an inter-dimensional association rule with repetitive predicates which implicitly takes into account hierarchical levels of dimensions.

Chen, Dayal and Hsu (2000) propose a *distributed OLAP based infrastructure* which combines data warehousing, data mining, and OLAP-based engine for Web access analysis. In the data mining engine, the authors mine intra-dimensional association rules from a single level of a dimension, called *base dimension*, by adding *features* from other dimensions. They also propose to consider the used features at multiple levels of granularity. In addition, the generated association rules can also be materialized by particular cubes, called *volume cubes*. However, in this approach, the use of association rules closely depends on the specific domain of Web access analysis for a *sale* application. Furthermore, it lacks a formal description that enables its generalization to other application domains.

Extended association rules were proposed by Nestorov and Jukić (2003) as an output of a cube mining process. An *extended association rule* is a repetitive predicate rule which involves attributes of *non-item* dimensions (i.e., dimensions not related to items/products). Their proposal deals with an extension of classical association rules since it provides additional information about the precise context of each rule. However, the authors focus on mining associations from transaction databases and do not take dimension hierarchy and data cube measures into account when computing support and confidence.

Tjioe and Taniar (2005) propose a method for mining association rules in data warehouses. Based on the multidimensional data organization, their method is able to extract associations from multiple dimensions at multiple levels of abstraction by focusing on summarized data according to the COUNT measure. In order to do so, they prepare multidimensional data for the mining process according to four algorithms: VAvg, HAvg, WMAvg, and ModusFilter. These algorithms prune all rows in the fact table which have less than the average quantity and provide an *initialized table*. This table is next used for mining both non-repetitive predicate and repetitive predicate association rules.

Discussion and the Position of our Proposal

The previous work on mining association rules in multidimensional data can be studied and compared according to various aspects.

Table 1. Comparison of association rule mining proposals from multidimensional data across application domain, data representation, and measure

Proposal	Application domain		Mined data structure		Measure	
	General	BMA	Tabular	MD	COUNT	All measures
Kamber, Han and Chiang (1997)	•			•	•	
Zhu (1998)	•		•		•	
Imieliński, Khachiyan and Abdulghani (2002)	•		•			•
Dong, Han, Lam, Pei and Wang (2001)	•		•			•
Chen, Dayal and Hsu (2000)		•	•		•	
Nestorov and Jukić (2003)		•	•		•	
Tjioe and Taniar (2005)	•		•		•	
Our proposal	•			•		•

As shown in Table 1, most of the proposals are designed and validated for sales data cubes. Their applications are therefore inspired by the well-known *basket market analysis* problem (*BMA*) driven on transactional databases. Nevertheless, we believe that most of the proposals (except for the proposals of Chen, Dayal and Hsu (2000) and Nestorov and Jukić (2003)) can easily be extended to other application domains. Our approach covers a wide spectrum of application domains. It depends neither on a specific domain nor on a special context of data.

Almost all the previous proposals are based on the frequency of data, by using the COUNT measure, in order to compute the support and the confidence of the discovered association rules. As indicated earlier, Imieliński, Khachiyan and Abdulghani (2002) can exploit any measure to detect *cubegrades*. Nevertheless, the authors do not compute the support and the confidence of the produced *cubegrades*. Tjioe and Taniar (2005) use the average (AVG) of measures in order to prune uninteresting itemsets in a pre-processing step. However, in the mining step, they only exploit the COUNT measure to compute the support and the confidence of association rules. Our approach revisits the support and the confidence of association rules when SUM-based aggregates are used.

Table 2. Comparison of association rule mining proposals from multidimensional data across dimension, level, and predicate

Proposal	Dimension		Hierarchy		Predicate	
	Intra	Inter	Single	Multiple	Repetitive	Non-repetitive
Kamber, Han and Chiang (1997)		•	•			•
Zhu (1998)	•	•	•		•	•
Imieliński, Khachiyan and Abdulghani (2002)		•		•	•	
Dong, Han, Lam, Pei and Wang (2001)		•		•	•	
Chen, Dayal and Hsu (2000)	•			•	•	
Nestorov and Jukić (2003)	•		•		•	
Tjioe and Taniar (2005)	•	•		•	•	•
Our proposal		•		•		•

According to Table 2, some of the proposals mine inter-dimensional association rules, whereas others deal with intra-dimensional rules. In general, an inter-dimensional association rule relies on more than one dimension from the mined data cube and consists of non-repetitive predicates, where the instance of each predicate comes from a distinct dimension. An intra-dimensional rule relies rather on a single dimension. It is constructed within repetitive predicates where their instances represent modalities from the considered dimension. Nevertheless, a *cubegrade* (Imieliński, Khachiyan & Abdulghani, 2002), or a *constrained gradient* (Dong, Han, Lam, Pei & Wang, 2001), can be viewed as an inter-dimensional association rule which has repetitive predicates. The instances of these predicates can be redundant in both the head and the body of the implication. Furthermore, the proposal of Tjioe and Taniar (2005) is mostly the only one which allows the mining of inter-dimensional association rules with either repetitive or non-repetitive predicates. In our proposal, we focus on the mining of inter-dimensional association rules with non-repetitive predicates.

We note that, except for (Kamber, Han and Chiang, 1997; Zhu, 1998), most of the previous proposals try to exploit the hierarchical aspect of multidimensional data by

expressing associations according to multiple levels of abstractions. For example, a *cubegrade* is an association which can be expressed within multiple levels of granularity. Association rules in (Chen, Dayal & Hsu, 2000) also exploit dimension hierarchies. In our case, the definition of the context in the meta-rule can be set to a given level of granularity.

Table 3. Comparison of association rule mining proposals from multidimensional data across user interaction, formalization, and association exploitation

Proposal	User interaction		Formalization		Association exploitation	
	Yes	No	Yes	No	Yes	No
Kamber, Han and Chiang (1997)	•			•		•
Zhu (1998)	•			•	•	
Imieliński, Khachiyan and Abdulghani (2002)	•		•			•
Dong, Han, Lam, Pei and Wang (2001)	•		•			•
Chen, Dayal and Hsu (2000)		•		•		•
Nestorov and Jukić (2003)	•			•		•
Tjioe and Taniar (2005)	•			•		•
Our proposal	•		•		•	

According to Table 3, we note that the proposal of (Chen, Dayal & Hsu, 2000) does not consider any interaction between users and the mining process. In fact, in the proposed *Web infrastructure*, analysis objectives are already predefined over transactional data and therefore users can not interfere with these objectives. In (Kamber, Han and Chiang, 1997) user's needs are expressed through the definition of a meta-rule.

Except for *cubegrades* (Imieliński, Khachiyan & Abdulghani, 2002) and *constrained gradients* (Dong, Han, Lam, Pei & Wang, 2001), almost all proposals miss a theoretical framework which establishes a general formalization of the mining process of association rules in multidimensional data.

In addition, in all these proposals, Zhu (1997) is mostly the only one who proposes association rule visualization. Nevertheless, the proposed graphical representation is similar to the ones commonly used in traditional association rules mining, and hence does not take into account multidimensionality.

OLEMAR is entirely driven by user's needs. It uses meta-rules to meet the analysis objectives. It is also based on a general formalization of the mining process of inter-dimensional association rules. Moreover, we include a visual representation of rules based on the graphic semiology principles.

CONCLUSION, DISCUSSION AND PERSPECTIVES

In this paper, we design an on-line environment for mining inter-dimensional association rules from data cubes as a part of a platform called *CubeMining*. We use a guided rule mining facility which allows users to limit the mining process to a specific context defined by a particular portion in the mined data cube. We also provide a computation of the support and the confidence of association rules when a SUM-based measure is used. This issue is quite interesting since it expresses associations which do not restrict users' analysis to associations driven only by the traditional COUNT measure. The support and the confidence may lead to the generation of large number of association rules. Therefore, we propose to evaluate interestingness of mined rules according to two additional descriptive criteria (Lift and Loevinger). These criteria can express the relevance of rules in a more precise way than what is offered by the support and the confidence. Our association rule mining procedure is an adaptation of the traditional level-wise *Apriori* algorithm to multidimensional data. In order to make extracted knowledge easier to interpret and exploit, we provide a graphical representation for the visualization of inter-dimensional association rules in the multidimensional space of the mined data cube. Empirical analysis showed the efficiency of our proposal and the acceptable runtime of our algorithm.

In the current development of our mining solution, we integrate SUM-based measures in the computation of interestingness criteria of extracted association rules. However, this choice assumes that the selected measure is additive and has only positive values. In the

suspicious regions data cube, the *surface of regions* is an appropriate measure for the computation of the revisited criteria. Nevertheless, the *total boundary length* of regions can not be used for that computation since the SUM of boundary lengths does not make concrete sense. In some cases, an OLAP context may be expressed by facts with non-additive or negative measures. For instance, in the traditional example of a sales data cube, the *average of sales* is typically a non-additive measure. Furthermore, the *profit of sales* is also an OLAP measure that can have negative values. In such situations, we obviously need a more appropriate interestingness estimation of association rule to handle a wider spectrum of measure types and aggregate functions (e.g., AVG, MAX).

Our proposal provides inter-dimensional association rules with non-repetitive predicates. Such rules consist of a set of predicate instances where each one represents a modality coming from a distinct dimension. This kind of association rules helps explain a value of a dimension by other values drawn from other dimensions. Nevertheless, an inter-dimensional association rule does not explain a modality by other ones from the same dimension. For instance, the latter type of rules is not able to explain the sales of a *product* by those of other *products* or even other *product categories*. In order to cope with this issue, we also need to extend our proposal in order to cover the mining of inter-dimensional association rules *with repetitive predicates* as well as *intra-dimensional* association rules. In addition, these new kinds of associations should profit from dimension hierarchies and allow modalities from multiple granularity levels.

The association rule mining process in our environment is based on an adaptation of the traditional level-wise *Apriori* algorithm to multidimensional data. The *anti-monotony* property (Agrawal, Imieliński & Swami, 1993) allows a fast search of frequent itemsets, and the guided mining of association rules we express as a meta-rule limits the search space according to the analysis objectives of users. However, some recent studies have shown the

limitations of *Apriori* and privileged the notion of frequent closed itemsets like in *Close* (Pasquier, Bastide, Taouil & Lakhal, 1999), *Pascal* (Bastide, Taouil, Pasquier, Stumme & Lakhal, 2000), *Closet* (Pei, Han & Mao, 2000), *Charm* (Zaki & Hsiao, 2002), and *Galicia* (Valtchev, Missaoui & Godin, 2004).

Finally, measures are used in our environment for computing interestingness criteria. We plan to study the semantics of association rules when measures appear in the expression of rules.

REFERENCES

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 1993)*, Washington, D.C., USA, May, (pp 207-216).

Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., & Lakhal, L. (2000). Mining Frequent Patterns with Counting Inference. *SIGKDD Explor. Newsl. 2*, 66–75.

Ben Messaoud, R., Boussaid, O., & Loudcher, R. S. (2006a). Efficient Multidimensional Data Representation Based on Multiple Correspondence Analysis. In *Proceedings. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, Philadelphia, USA, August, (pp 662-667).

Ben Messaoud, R., Boussaid, O., & Loudcher, R. S. (2006b). A Data Mining-Based OLAP Aggregation of Complex Data: Application on XML Documents. *International Journal of Data Warehousing and Mining*, *2*(4), 1-26.

Ben Messaoud, R., Loudcher, R. S., Boussaid, O., & Missaoui, R. (2006). Enhanced Mining of Association Rules from Data Cubes, In *Proceedings of the 9th ACM International*

Workshop on Data Warehousing and OLAP (DOLAP 2006), Arlington, VA, USA, November, (pp 11-18).

Bertin, J. (19981). *Graphics and Graphic Information Processing*. de Gruyter.

Brin, S., Motwani, R., & Silverstein, C. (1997). Beyond Market Baskets: Generalizing Association Rules to Correlations. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 1997)*, Tucson, Arizona, USA, May, (pp 265-276).

Chaudhuri. S., & Dayal, U. (1997). An Overview of Data Warehousing and OLAP Technology. *SIGMOD Record*, *26*(1), 65-74.

Chen, Q., Dayal. U., & Hsu. M. (1999). A Distributed OLAP Infrastructure for E-Commerce. In *Proceedings of the 4th IECIS International Conference on Cooperative Information Systems (COOPIS 1999)*, Edinburgh, Scotland, September, (pp 209-220).

Chen, Q., Dayal. U., & Hsu. M. (2000). An OLAP-based Scalable Web Access Analysis Engine. In *Proceedings of the 2nd International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2000)*, London, UK, September, (pp 210-223).

Dong, G., Han, H., Lam, J. M. W., Pei, J., & Wang, K. (2001). Mining Multi-Dimensional Constrained Gradients in Data Cubes. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB 2001)*, Rome, Italy, September, (pp 321-330).

Goil, S., & Choudhary, A. (1998). High Performance Multidimensional Analysis and Data Mining. In *Proceedings of the 1st International Workshop on Data Warehousing and OLAP (DOLAP 1998)*, Bethesda, Maryland, USA, November, (pp 34-39).

Heath, M., Bowyer, K., Kopans, D., Moore, R., & Jr, P.K. (2000). The Digital Database for Screening Mammography. In *Proceedings of the 5th International Workshop on Digital Mammography*, Toronto, Canada, June.

Imieliński, T., Khachiyan, L., & Abdulghani, A. (2002). Cubegrades: Generalizing Association rules. *Data Mining and Knowledge Discovery*, *6*(3), 219-258.

Inmon, W.H. (1996). *Building the Data Warehouse*. John Wiley & Sons.

Kamber, M., Han, J., & Chiang, J. (1997). Multi-Dimensional Association Rules Using Data Cubes. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD 1997)*, Newport Beach, CA, USA, August, (pp 207-210).

Kimball, R. (1996). *The Data Warehouse Toolkit*. John Wiley & Sons.

Lallich, S., Vaillant, B, & Lenca, P. (2005). Parametrised Measures for the Evaluation of Association Rules Interestingness. In *Proceedings of the 6th International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005)*, Brest, France, May, (pp 220-229).

Lenca, P., Vaillant, B, & Lallich, S. (2006). On the Robustness of Association Rules. In *Proceedings of the 2006 IEEE International Conference on Cybernetics and Intelligent Systems (CIS 2006)*, Bangkok, Thailand, June, (pp 596-601).

Loevinger, J. (1974). A Systemic Approach to the Construction and Evaluation of Tests of Ability. *Psychological Monographs*, *61*(4).

Nestorov, S., & Jukić. N. (2003). Ad-Hoc Association-Rule Mining within the Data Warehouse. In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS 2003)*, Big Island, Hawaii, USA, January, (pp 232-242).

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999). Efficient Mining of Association Rules Using Closed Itemset Lattices. *Information Systems*, *24*(1), 25-46.

Pei, J., Han, J., & Mao, R. (2000). Closet : An Efficient Algorithm for Mining Frequent Closed Itemsets. In *Proceedings of the ACM SIGMOD International Workshop on Data Mining and Knowledge Discovery (DMKD 2000)*, Dallas, Texas, USA, May, (p 21-30).

Tjioe, H. C., & Taniar D. (2005). Mining Association Rules in Data Warehouses. *International Journal of Data Warehousing and Mining*, 1(3), 28-62.

Valtchev, P., Missaoui, R., & Godin, R. (2004). Formal Concept Analysis for Knowledge and Data Discovery : New Challenges. In *Proceedings of the 2nd International Conference on Formal Concept Analysis (ICFCA 2004)*, Sydney, Australia, February, (pp 352-371).

Zaki, M. J., & Hsiao, C. J. (2002). CHARM: An Efficient Algorithm for Closed Itemset Mining. In *Proceeding of the 2nd SIAM International Conference on Data Mining (SDM'02)*, Arlington, VA, USA, April.

Zhu, H. (1998). *On-Line Analytical Mining of Association Rules*. Master's thesis, Simon Faster University, Burnaby, British Columbia, Canada, December.