Enhanced Mining of Association Rules from Data Cubes

Riadh Ben Messaoud⋆
rbenmessaoud@eric.univ-lyon2.fr

Sabine Loudcher Rabaséda⋆
sabine.loudcher@univ-lyon2.fr

Omar Boussaid⋆
omar.boussaid@univ-lyon2.fr

Rokia Missaoui♭
rokia.missaoui@uqo.ca

⋆ Laboratoire ERIC – Université Lumière Lyon 2
5 avenue Pierre Mendès–France,
69676 Bron Cedex, France
♭ Département d'informatique et d'ingénierie – Université du Québec en Outaouais
C.P. 1250, succursale B, Gatineau (Québec)
Canada, J8X 3X7

ABSTRACT

On-line analytical processing (OLAP) provides tools to explore and navigate into data cubes in order to extract interesting information. Nevertheless, OLAP is not capable of explaining relationships that could exist in a data cube. Association rules are one kind of data mining techniques which finds associations among data. In this paper, we propose a framework for mining inter-dimensional association rules from data cubes according to a *sum-based aggregate measure* more general than simple frequencies provided by the traditional COUNT measure. Our mining process is guided by a meta-rule context driven by analysis objectives and exploits aggregate measures to revisit the definition of support and confidence. We also evaluate the interestingness of mined association rules according to *Lift* and *Loevinger* criteria and propose an efficient algorithm for mining inter-dimensional association rules directly from a multidimensional data.

Categories and Subject Descriptors: H.2.8 [Information systems]: Database applications—*Data mining*; H.4.2 [Information systems]: Types of systems—*Decision support*.

General Terms: Algorithms, Experimentation.

Keywords: OLAP, association rules, data cubes.

1. INTRODUCTION

Data warehousing and OLAP technology has known an important progress since the 90s. In addition, with efficient techniques developed for computing data cubes, OLAP users have become widely able to explore multidimensional data, navigate through hierarchical levels of dimensions, and therefore extract interesting information according to multiple levels of granularity in data. Nevertheless, the OLAP technology is quite limited to an exploratory task and does not provide automatic tools to explain relationships and associations within data. For example, we can note from a data cube that sales of *sleeping bags* are particulary high in a given city. Nevertheless, current OLAP tools are not able to automatically explain the causes of this particular fact. Users are usually supposed to explore the data cube according to multiple dimensions in order to manually find an explanation for a given phenomenon (e.g., high sales). For instance, one possible explanation 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 *et al.* state that it is important to explore data cubes by using association rules algorithms since data cube structures make good use of structured warehouse and pre-computed aggregation information [8]. Further, Imieliński *et al.* believe that OLAP is closely intertwined with association rules and shares with them the goal of finding patterns in the data [7]. In [5], Goil and Choudhary argue that automated techniques of data mining can make OLAP more useful and easier to apply in the overall scheme of decision support systems. Indeed, data mining techniques such as association rule mining can be used together with OLAP to discover knowledge from data cubes. Moreover, dimension hierarchies can be exploited to generate multilevel association rules. The aggregate values needed for discovering association rules are already pre-computed and stored in the data cube. A COUNT cell of a cube stores the number of occurrences of the corresponding multidimensional data values. A dimension COUNT cell stores the sum of COUNTs of the whole dimension. With this structure, it is straightforward to calculate the values of the support and the confidence of association rules based on the values in these summary cells.

The COUNT measure corresponds to the frequency of facts. Nevertheless, in an analysis process, users are usually interested in observing multidimensional data and their

DOLAP'06, November 10, 2006, Arlington, Virginia, USA.

associations according to measures more elaborated than simple frequencies. In this paper, we establish a general framework for mining *inter-dimensional* association rules from multidimensional data. We use the concept of *inter-dimensional meta-rule* which allows users to guide the mining process and focus on a specific context from which rules can be extracted. Our framework allows also a redefinition of the support and confidence measures based on the aggregate functions (SUM and COUNT) used as cube indicators (measures). Therefore, the computation of support and confidence according to the COUNT measure becomes a particular case in our proposal. In addition to support and confidence, we use two other descriptive criteria (*Lift* and *Loevinger*) in order to evaluate interestingness of mined associations. These criteria are computed according to a *sum-based aggregate measure* in the data cube and reflect interestingness of associations in a more relevant way than what is offered by support and confidence. We developed our proposal according to a *bottom-up* algorithm for searching association rules. Our algorithm consists in an adaptation of the traditional Apriori algorithm to multidimensional data.

This paper is organized as follows. In Section 2, we give a brief overview of association rule mining from multidimensional data. In Section 3, we develop the formal background, notations, and definitions of our proposal. Section 4 introduces our proposed framework: the concept of *inter-dimensional meta-rule*, the general computation of support and confidence based on measures, and criteria for the advanced evaluation of mined association rules. Section 5 provides an implementation and describes our algorithm for mining inter-dimensional association rules. In Section 6, we conduct some experiments concerning the performance of the developed algorithm. Finally, Section 7 gives a conclusion and future research directions.

2. RELATED WORK

Association rule mining was first introduced by Agrawal *et al.* [1] 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 the minimum support and association rules based on the minimum confidence. Since then, many developments have been performed in order to handle various types and structures of data. For instance, the problem of mining *quantitative association rules* from large relational tables was first addressed in [17]. In [16], Srikant and Agrawal proposed to mine association rules for categorical data. In [6], Han and Fu introduced *multilevel association rules* which cope with multilevel data abstractions.

To the best of our knowledge, Kamber *et al.* [8] were the first who addressed the issue of mining *association rules from multidimensional data*. They introduced the concept of *metarule-guided mining* which consists in using rule templates defined by users in order to guide the mining process. This mining process considers precomputed data cubes and dynamic construction of relevant data cubes. Inter-dimensional association rules with distinct predicates are mined from single levels of dimensions. Support and confidence are computed according to the COUNT measure. Zhu considers the problem of mining association rules from data cubes under three groups: *inter-dimensional*, *intra-dimensional*, and *hybrid* association mining [19]. Intra-dimensional association rule cover repetitive predicates from a single dimension, whereas inter-dimensional association rules are mined from multiple dimensions without repetition of predicates in each dimension. His proposal does not profit from hierarchical levels of dimensions since it flattens data cubes to mine associations. Further, he uses the COUNT measure and does not take into account further OLAP measures to evaluate discovered rules. Cubegrades, proposed by Imieliński *et al.*, are a generalization of association rules. They focus on significant changes that affect measures when a cube is modified through specialization, generalization or mutation [7]. 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 [4], Dong *et al.* study 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 concepts of cubegrades and constrained gradients are quite different from the traditional association rules mining. They discover modifications on OLAP aggregates when moving from a *Source-Cube* to a *Target-Cube* rather than data patterns and associations included in the cube itself. Nevertheless, we can consider a cubegrade or a constrained gradient as an inter-dimensional association rule with repetitive predicates. Chen *et al.* mine intra-dimensional association rules by adding *features* from other dimensions features at multiple levels [3]. Therefore, association rules at different area levels and time levels can be specified. Nevertheless, the use of association rules in this approach closely depends on the specific domain of Web access analysis. Furthermore, it lacks a formal description that enables a concrete generalization to other application domains. *Extended association rules* were proposed in [14] by Nestorov and Jukić. It consists in mining associations from data warehouses by utilizing the SQL processing power of the data warehouse itself without using a separate data mining tool. An *extended association rule* is a repetitive predicate rule which involves attributes of non-item dimensions defined by the user. The authors focus on mining associations from transactional databases and do not take dimension hierarchy and data cube measures into account for computing support and confidence. Tjioe and Taniar [18] 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 this, the authors 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.

In Table 1, we summarize a set of association rule mining proposals according to a set of criteria. Some of these proposals mine intra-dimensional association rules, whereas others deal with inter-dimensional association rules. Almost all intra-dimensional proposals provide associations with repetitive predicates from a single dimension of a data

Proposal	Dimension		Level		Predicate		Measure		Application domain	
	Intra-dimensional	Inter-dimensional	Single level	Multiple levels	Repetitive	Non-repetitive	COUNT	All measures	Market basket analysis	General
Kamber *et al.*		•	•			•	•			•
Zhu	•	•	•		•	•	•		•	
Imieliński *et al.*		•		•	•			•		•
Dong *et al.*		•		•	•			•		•
Chen *et al.*	•			•	•		•		•	
Nestorov & Jukić	•		•		•		•		•	
Tjioe & Taniar	•	•		•	•	•	•			•
Our proposal		•	•			•		•		•

Table 1: Comparison of mining associations proposals from multidimensional data

cube. Inter-dimensional associations are not necessarily restricted to the traditional *market basket analysis* and can be extended to large application domains. The hierarchical aspect of multidimensional data is not widely exploited. It rather concerns intra-dimensional associations with repetitive predicates coming from different levels of the mined dimension. Except for cubegrades [7] where any measure can be used in order to check the effects of changes on the cube structure, all the proposed approaches are restricted to the COUNT measure in the mining process. In this paper, we use the notion of metarule-guided mining proposed by Kamber *et al.* [8] to guide a general process of mining inter-dimensional associations with non-repetitive predicates.

The main contribution of our proposal consists in integrating the measures of a data cube in the computation of the support and the confidence of association rules. We also use advanced criteria in order to evaluate interestingness of mined associations. An Apriori-based algorithm is also adapted in order to handle multidimensional data.

3. FORMAL BACKGROUND AND NOTATIONS

Let $\mathcal{C}$ be a data cube with a non empty set of d dimensions $\mathcal{D} = \{D_i\}_{(1 \leq i \leq d)}$, and a non empty set of measures $\mathcal{M}$. Each dimension $D_i \in \mathcal{D}$ has a non empty set of hierarchical levels. We assume that H^i_j is the j^{th} ($j \geq 0$) hierarchical level in D_i. The coarse level of D_i, denoted H^i_0, 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^1_0, the Continent level is denoted H^1_1, and the Country level is denoted H^1_2. Let $\mathcal{H}_i$ be the set of hierarchical levels of dimension D_i, where each level $H^i_j \in \mathcal{H}_i$ consists of a non empty set of members denoted $\mathcal{A}_{ij}$. For example, in Figure 1, the set of hierarchical levels of D_2 is $\mathcal{H}_2 = \{H^2_0, H^2_1, H^2_2\} = \{\mathsf{All}, \mathsf{Family}, \mathsf{Article}\}$, and the set of members of the Article level of D_2 is $\mathcal{A}_{22} =$ {iTwin, iPower, DV-400, EN-700, aStar, aDream}.

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

As defined above, a sub-cube according to a set of dimensions $\mathcal{D}'$ corresponds to a portion from the initial data cube $\mathcal{C}$. It consists in setting for each dimension from $\mathcal{D}'$ a non empty subset of member values from a single hierarchical level of that dimension. For example, consider $\mathcal{D}' = \{D_1, D_2\}$ a subset of dimensions from the cube of Figure 1. $(\Theta_1,\Theta_2) =$ (Europe, {EN-700, aStar, aDream}) is therefore a possible sub-cube on $\mathcal{C}$ according to $\mathcal{D}'$, 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 $\mathcal{D}=\{D_1,D_2,D_3\}$.

One particular case of the sub-cube definition is when it is defined on $\mathcal{C}$ according to $\mathcal{D}' = \mathcal{D} = \{D_1, \ldots, D_d\}$ 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 $\mathcal{C}$. For example, the black cell in Figure 1 can be considered as the sub-cube (Japan, iTwin, 2002) on $\mathcal{C}$ according to $\mathcal{D} = \{D_1, D_2, D_3\}$. Each cell from the data cube $\mathcal{C}$ represents an OLAP fact which is evaluated in $\mathbb{R}$ according to one measure M from $\mathcal{M}$. 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 $\mathcal{C}$ according to $\mathcal{D}' \subseteq \mathcal{D}$. The sum-based aggregate measure of sub-cube $(\Theta_1, \ldots, \Theta_p)$ according to $M \in \mathcal{M}$, noted $M(\Theta_1, \ldots, \Theta_p)$, is the SUM of measure M of all facts in the sub-cube.

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

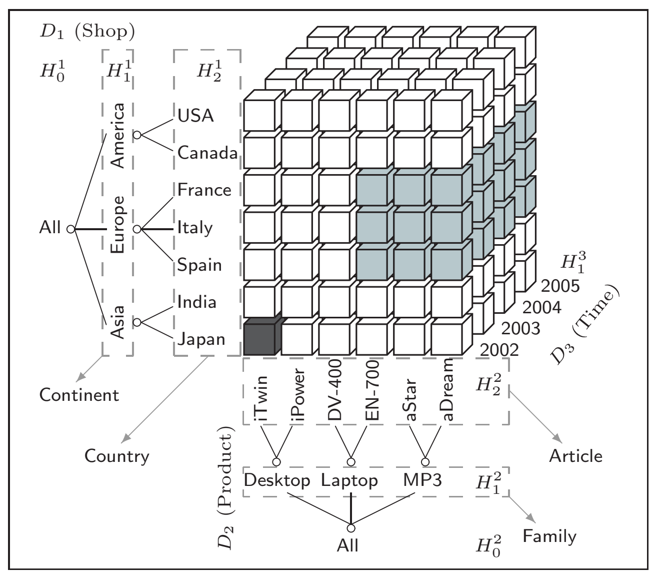


Figure 1: Example of Sales data cube

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 $\langle a \in \mathcal{A}_{ij} \rangle$.

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 $\alpha_1 = \langle a \in \mathcal{A}_{11} \rangle = \langle a \in \{\mathsf{America}, \mathsf{Europe}, \mathsf{Asia}\} \rangle$.

Definition 4. (**Inter-dimensional predicate**)
Let $\mathcal{D}' \subseteq \mathcal{D}$ be a non empty set of p dimensions $\{D_1, \ldots, D_p\}$

from the data cube $\mathcal{C}$ $(2 \leq p \leq d)$. $(\alpha_1 \wedge \cdots \wedge \alpha_p)$ is called an inter-dimensional predicate in $\mathcal{D}'$ iff $\forall i \in \{1, \ldots, p\}$, α_i is a dimension predicate in D_i.

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

4. THE PROPOSED FRAMEWORK

As mentioned earlier, our proposal consists in (i) exploiting metarule 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.

4.1 Inter-dimensional meta-rules

As in [15], we consider two distinct subsets of dimensions in the data cube $\mathcal{C}$: (i) $\mathcal{D}_\mathcal{C} \subset \mathcal{D}$ is a subset of p *context dimensions*. A sub-cube on $\mathcal{C}$ according to $\mathcal{D}_\mathcal{C}$ defines the context of the mining process; and (ii) $\mathcal{D}_\mathcal{A}$ 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:

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

where $(\Theta_1, \ldots, \Theta_p)$ is a sub-cube on $\mathcal{C}$ according to $\mathcal{D}_\mathcal{C}$. It defines the portion of cube $\mathcal{C}$ to be mined. Unlike the meta-rule proposed in [8], 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 $\mathcal{D}_\mathcal{C} = \emptyset$, no particular analysis context is selected. Therefore, the mining process covers the whole cube $\mathcal{C}$.

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 $\mathcal{D}_\mathcal{A}$. Therefore, the conjunction $(\alpha_1 \wedge \cdots \wedge \alpha_s) \wedge (\beta_1 \wedge \cdots \wedge \beta_r)$ is an inter-dimensional predicate in $\mathcal{D}_\mathcal{A}$, where the number of predicates $(s + r)$ in the meta-rule is equal to the number of dimensions in $\mathcal{D}_\mathcal{A}$. 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: $\mathcal{D}_\mathcal{C} = \{D_5, D_6\}$ ={Profession, Gender}, and $\mathcal{D}_\mathcal{A} = \{D_1, D_2, D_3\}$ ={Shop, Product, Time}. One possible inter-dimensional meta-rule scheme is:

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

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.

4.2 Measure-based support and confidence

Traditionally, as it was introduced in [1], the support (SUPP) of an association rule $X \Rightarrow Y$, in a database of transactions $\mathcal{T}$, 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 $\mathcal{C}$, 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 $\mathcal{C}$, 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 precomputed frequencies of OLAP facts. With this structure, it is straightforward to calculate support and confidence of associations in a data cube based on these 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:

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

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

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 profit measure. Table 2(a) and 2(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 2(a), whereas they are not frequent according to the SUM of the profit measure in Table 2(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 2(a)). Whereas these itemsets are not *large* in Table 2(b). The *large* itemsets according to the SUM of the profit measure are rather (⟨ Europe ⟩, ⟨ Laptop ⟩, ⟨ 2004 ⟩) and (⟨ Europe ⟩, ⟨ Laptop ⟩, ⟨ 2005 ⟩).

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

	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)

Table 2: Fragment of the Sales cube according to the (a) COUNT measure and the (b) SUM of the profit measure

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 unit of sales themselves. Therefore, we define a general computation of support and confidence of inter-dimensional association rules according to a user defined measure $M \in \mathcal{M}$ from the mined data cube. Consider a general rule R which complies with the defined inter-dimensional meta-rule (1):

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

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

$$\textsc{Supp}(R) = \frac{M(x_1,\dots,x_s,y_1,\dots,y_r,\Theta_1,\dots,\Theta_p,\mathsf{All},\dots,\mathsf{All})}{M(\mathsf{All},\dots,\mathsf{All},\Theta_1,\dots,\Theta_p,\mathsf{All},\dots,\mathsf{All})} \quad (3)$$

$$\textsc{Conf}(R) = \frac{M(x_1,\dots,x_s,y_1,\dots,y_r,\Theta_1,\dots,\Theta_p,\mathsf{All},\dots,\mathsf{All})}{M(x_1,\dots,x_s,\mathsf{All},\dots,\mathsf{All},\Theta_1,\dots,\Theta_p,\mathsf{All},\dots,\mathsf{All})} \quad (4)$$

where $M(x_1,\dots,x_s,y_1,\dots,y_r,\Theta_1,\dots,\Theta_p,\mathsf{All},\dots,\mathsf{All})$ 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. Nevertheless, in order to simplify notations, we keep on referring to our generalized support and confidence with the usual terms.

4.3 Advanced evaluation of association rules

Support and confidence are the most known measures for the evaluation of association rule interestingness. These measures are key elements of all Apriori-like algorithms [1] 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 [9, 10].

Let consider again the association rule $R : X \Rightarrow Y$ which complies with the inter-dimensional meta-rule (1), where $X = (x_1 \wedge \dots \wedge x_s)$ and $Y = (y_1 \wedge \dots \wedge y_r)$ are conjunctions of dimension predicates. We also consider a user-defined measure $M \in \mathcal{M}$ from data cube $\mathcal{C}$. 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 $(\Theta_1, \dots, \Theta_p)$ in the context dimensions $\mathcal{D}_\mathcal{C}$. 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 $P_{Y/X}$, of matching Y given that X is already matched.

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

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

$$P_{XY} = \textsc{Supp}(R) = \frac{M(x_1,\dots,x_s,y_1,\dots,y_r,\Theta_1,\dots,\Theta_p,\mathsf{All},\dots,\mathsf{All})}{M(\mathsf{All},\dots,\mathsf{All},\Theta_1,\dots,\Theta_p,\mathsf{All},\dots,\mathsf{All})}$$

$$P_{Y/X} = \textsc{Conf}(R) = \frac{M(x_1,\dots,x_s,y_1,\dots,y_r,\Theta_1,\dots,\Theta_p,\mathsf{All},\dots,\mathsf{All})}{M(x_1,\dots,x_s,\mathsf{All},\dots,\mathsf{All},\Theta_1,\dots,\Theta_p,\mathsf{All},\dots,\mathsf{All})}$$

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 [9]. 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 particulary 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) [2] and the Loevinger criterion (Loev) [11]. These two criteria take the independence of itemsets X and Y as a reference, and are defined on rule R as follows:

$$\textsc{Lift}(R) = \frac{P_{YX}}{P_X P_Y} = \frac{\textsc{Supp}(R)}{P_X P_Y} \quad (5)$$

$$\textsc{Loev}(R) = \frac{P_{Y/X} - P_Y}{P_{\overline{Y}}} = \frac{\textsc{Conf}(R) - P_Y}{P_{\overline{Y}}} \quad (6)$$

The Lift of a rule can be interpreted as the deviation of the support of the rule from the support expected under the independence hypothesis between the body X and the head Y [2]. 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, $\textsc{Lift}(R) = 2$ means that facts matching with X have two times more chances to match with Y. By opposition 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 evaluation for association rules [11]. 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 terms, the Loevinger criterion normalizes the centered confidence of a rule according to the probability of not satisfying its head.

5. IMPLEMENTATION AND ALGORITHMS

We developed a Web application to mine association rules from data cubes according to our proposal. This application is a *Mining Association Rule Module* that runs on a Client/Server platform, called MiningCubes, which already includes previous work on mining multidimensional data [12, 13]. The platform is equipped with a *data loader component* that enables connection to multidimensional data cubes stored in the *Analysis Services* of *MS SQL Server 2000*. By employing MDX (MultiDimensional eXpressions) queries, this component loads information about the structure (labels of dimensions, hierarchical levels and measures) and the content of a user selected data cube. The *Mining Association Rule Module* allows the definition of required parameters to run an association rule mining process. In fact, a user is able to define analysis dimensions $\mathcal{D}_\mathcal{A}$, context dimensions $\mathcal{D}_\mathcal{C}$, a meta-rule with its context sub-cube $(\Theta_1, \ldots, \Theta_p)$ and its inter-dimensional predicates scheme $(\alpha_1 \wedge \cdots \wedge \alpha_s) \Rightarrow (\beta_1 \wedge \cdots \wedge \beta_r)$, the measure M used to compute criteria of association rules, and the thresholds *minsupp* and *minconf*.

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: (i) 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* criterion. In addition, it supposes that if a k-itemset is frequent, then all sub-itemsets are frequent, too; (ii) the *bottom-up* approach which goes from 1-itemsets to longer itemsets. It complies with the Apriori property [1] which proves 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 particulary the case of OLAP data cubes. Therefore, we base our algorithm 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 level i is the number of items in the set. 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 $\mathcal{D}_\mathcal{A}$ over the data cube $\mathcal{C}$. These 1-candidates correspond to members of $\mathcal{D}_\mathcal{A}$, 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 $\mathcal{D}_\mathcal{A}$, we capture 1-candidates from $\mathcal{A}_{ij}$, which is the set of members of the j^{th} 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 2. We suppose that, according to a user meta-rule, mined association rules are needed to comply with the

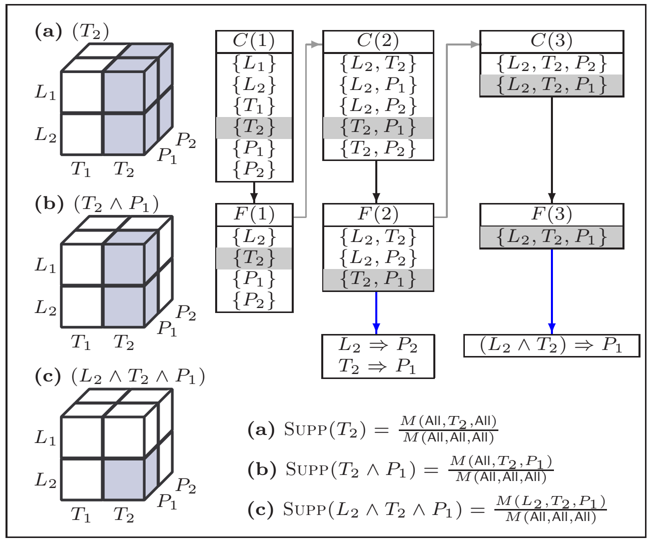


Figure 2: Example of a bottom-up generation of association rules from a data cube

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 $C_1 = \{\{L_1\}, \{L_2\}, \{T_1\}, \{T_2\}, \{P_1\}, \{P_2\}\}$.

Our algorithm is a level wise iterative process. For each level i, if the set $C(i)$ is not empty and i is less than $s + r$, **the first step** of our 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 $\mathcal{D}_\mathcal{A}$, i.e., A must be a conjunction of members from i distinct dimensions from $\mathcal{D}_\mathcal{A}$; 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 by the example of Figure 2, $\text{Supp}(A)$ is a measure-based support computed according to a user selected measure M from the cube.

From each $A \in F(i)$, **the second step** extracts association rules with respect to 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 2, $P_2 \Rightarrow L_2$ 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 formalism defined in Subsection 4.3. 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. One $(i+1)$-candidate is the union of two i-itemsets A and B from $F(i)$ that respects three conditions: (i) A and B must have $i-1$ commun items; (ii) all non empty sub-itemsets from $A \cup B$ must be instances of inter-dimensional predicates in $\mathcal{D}_\mathcal{A}$; and (iii) all non empty sub-itemsets from $A \cup B$ must be frequent itemsets. For ex-

```
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
```

Algorithm 1: Algorithm for mining association rules in a data cube

ample, in Figure 2, itemsets $A = \{L_2, T_2\}$ and $B = \{L_2, P_2\}$ from $F(2)$ have $\{L_2\}$ as a commun 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 $C(3)$.

```
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].[Profit],
       [Profession].[Profession category].[Student],
       [Gender].[Gender].[Female])
```

Figure 3: A MDX query example

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 according to the formalism defined in Subsections 4.2 and 4.3. In our implementation, these functions work with MDX queries which directly pick up required precomputed 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 by considering the Profit measure, the support of R_1 is written as follows:

$$\textsc{Supp}(R_1) = \frac{\text{Profit(America,Laptop,2004,All,Student,Female,All)}}{\text{Profit(All,All,All,All,Student,Female,All)}}$$

The numerator value of $\textsc{Supp}(R_1)$ is therefore returned by the MDX query of Figure 3.

6. PERFORMANCE EVALUATION

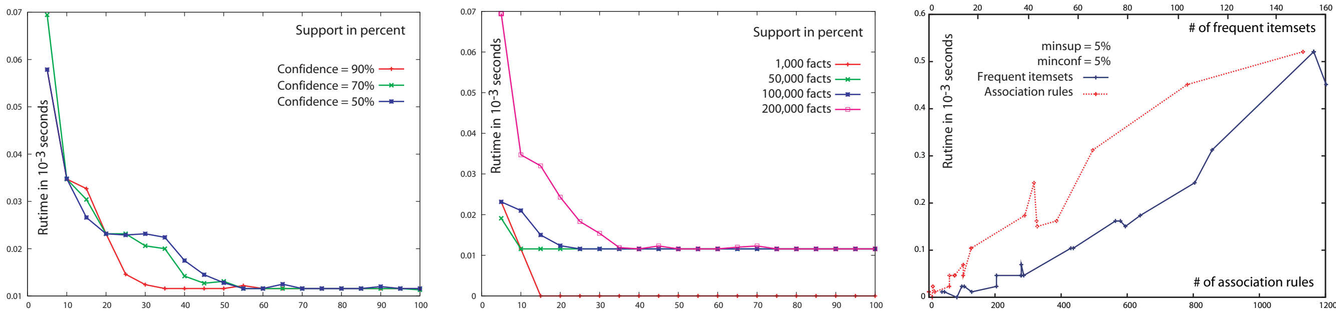
In order to evaluate the performance of our developed application, we conducted a set of experiments under Windows XP on a 1.60GHz PC with 480MB of main memory, and an Intel Pentium 4, and used *Analysis Services* of *MS SQL Server 2000*.

Figure 4(a) shows the relation between the runtime of our algorithm 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 4(b) presents a test of our algorithm on a varying number 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 4(c). In the latter figure, for a support and a confidence thresholds equal to 5%, we can clearly notice that 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. In fact, 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*.

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 precomputed aggregates of the multidimensional cube which helps compute the support and the confidence via MDX queries; and (iii) the use of the Apriori property which is definitely suited to sparse data cubes and considerably reduces the complexity of large itemsets search.

7. CONCLUSION AND PERSPECTIVES

In this paper, we establish a general framework for mining inter-dimensional association rules from data cubes. We use inter-dimensional meta-rule which allows users to limit the mining process to a specific context defined by a particular portion in the mined data cube. In our proposal, we provide a general computation of support and confidence of association rules that can be based on any measure from the data cube. This issue is quite interesting since it expresses associations which consider wide analysis objectives and do not restrict users' analysis to associations only driven by the traditional COUNT measure. We also propose to evaluate



(a) (b) (c)

Figure 4: The running times of our algorithm according to (a) support with different confidences, (b) support with different # of facts, (c) # of frequent itemsets and # of association rules

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. We developed our proposal according to a bottom-up algorithm for searching association rules. Our algorithm consists in an adaptation of the traditional Apriori algorithm in order to handle the multidimensional structure of data. Series of experiments proved the efficiency of our proposal and the acceptable runtime of our algorithm.

Some future directions need to be addressed for this work. First, we plan to extend this proposal in order to handle inter-dimensional association rules with repetitive predicates as well as intra-dimensional association rules. Another possible extension consists in embedding the measure in the expression of mined association rules. In addition, we need to profit from the hierarchical aspect of cube dimensions to mine multi-level association rules. Finally, we also have to cope with the visualization of mined association rules in the space representation of the data cube itself in order to make associations easier to interpret and exploit by OLAP users.

8. REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 1993)*, pages 207–216, Washington, D.C., USA, May 1993. ACM Press.

[2] S. Brin, R. Motwani, and C. Silverstein. Beyond Market Baskets: Generalizing Association Rules to Correlations. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 1997)*, pages 265–276, Tucson, Arizona, USA, May 1997. ACM Press.

[3] Q. Chen, U. Dayal, and M. Hsu. An OLAP-based Scalable Web Access Analysis Engine. In *Proceedings of the* 2^{nd} *International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2000)*, Lecture Notes in Computer Science, pages 210–223, London, UK, September 2000. Springer-Verlag.

[4] G. Dong, J. Han, J. M. W. Lam, J. Pei, and K. Wang. Mining Constrained Gradients in Large Databases. *IEEE Transactions on Knowledge and Data Engineering*, 16(8):922–938, August 2004.

[5] S. Goil and A. Choudhary. High Performance Multidimensional Analysis and Data Mining. In *Proceedings of the High Performance Networking and Computing Conference (SC 1998)*, Orlando, Florida, USA., November 1998.

[6] J. Han and Y. Fu. Discovery of Multiple-Level Association Rules from Large Databases. In *Proceedings of the* 21^{st} *International Conference on Very Large Data Bases (VLDB 1995)*, pages 420–431, Zurich, Switzerland, September 1995. Morgan Kaufmann.

[7] T. Imieliński, L. Khachiyan, and A. Abdulghani. Cubegrades: Generalizing Association Rules. *Data Mining and Knowledge Discovery*, 6(3):219–258, 2002.

[8] M. Kamber, J. Han, and J. Chiang. Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes. In *Proceedings of the* 3^{rd} *International Conference on Knowledge Discovery and Data Mining (KDD 1997)*, pages 207–210, Newport Beach, CA, USA, August 1997. The AAAI Press.

[9] S. Lallich, B. Vaillant, and P. Lenca. Parametrised Measures for the Evaluation of Association Rules Interestingness. In *Proceedings of the* 6^{th} *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005)*, pages 220–229, Brest, France, May 2005.

[10] P. Lenca, B. Vaillant, and S. Lallich. On the Robustness of Association Rules. In *Proceedings of the 2006 IEEE International Conference on Cybernetics and Intelligent Systems (CIS 2006)*, pages 596–601, Bangkok, Thailand, June 2006. IEEE Computer Society.

[11] J. Loevinger. A Systemic Approach to the Construction and Evaluation of Tests of Ability. *Psychological Monographs*, 61(4), 1947.

[12] R. B. Messaoud, O. Boussaid, and S. L. Rabaséda. A Data Mining-Based OLAP Aggregation of Complex Data: Application on XML Documents. *International Journal of Data Warehousing and Mining*, 2(4):1–26, 2006.

[13] R. B. Messaoud, O. Boussaid, and S. L. Rabaséda. Efficient Multidimensional Data Representation Based on Multiple Correspondence Analysis. In *Proceedings of the* 12^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 662–667, Philadelphia, PA, USA, August 2006. ACM Press.

[14] S. Nestorov and N. Jukić. Ad-Hoc Association-Rule Mining within the Data Warehouse. In *Proceedings of the* 36^{th} *Hawaii International Conference on System Sciences (HICSS 2003)*, pages 232–242, Big Island, Hawaii, USA, January 2003. IEEE Computer Society.

[15] M. Plantevit, Y. W. Choong, A. Laurent, D. Laurent, and M. Teisseire. M^2SP: Mining Sequential Patterns Among Several Dimensions. In *Proceedings of the* 9^{th} *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005)*, Lecture Notes in Computer Science, pages 205–216, Porto, Portugal, October 2005. Springer-Verlag.

[16] R. Srikant and R. Agrawal. Mining Generalized Association Rules. In *Proceedings of the* 21^{st} *International Conference on Very Large Data Bases (VLDB 1995)*, pages 407–419, Zurich, Switzerland, September 1995. Morgan Kaufmann.

[17] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 1996)*, pages 1–12, Montreal, Quebec, Canada, June 1996. ACM Press.

[18] H. C. Tjioe and D. Taniar. Mining Association Rules in Data Warehouses. *International Journal of Data Warehousing and Mining*, 1(3):28–62, 2005.

[19] H. Zhu. On-Line Analytical Mining of Association Rules. Master's thesis, Simon Fraser University, Burnaby, British Columbia, Canada, December 1998.