Enhanced Mining of Association Rules from Data Cubes

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

Laboratoire ERIC – Université Lumière Lyon 2 – France

***Laboratoire LARIM – Université du Québec en Outaouais – Canada**

OLAP capabilities

- Visual exploration of multidimensional data.
- Navigation through hierarchical levels of dimensions.
- Extraction of relevant information for decision-making.

OLAP limitations

- Limitation to exploratory tasks.
- Automatic explanation of associations within data.

OLAP capabilities

- Visual exploration of multidimensional data.
- Navigation through hierarchical levels of dimensions.
- Extraction of relevant information for decision-making.

OLAP limitations

- Limitation to exploratory tasks.
- Automatic explanation of associations within data.

An example: a sales data cube

	Quarter 1	Quarter 2	Quarter 3	Quarter 4
Soccer shoes	$ 9,400	$ 10,000	$ 12,600	$ 10,600
Sleeping bag	$ 20,500	$ 13,700	$ 52,400	$ 21,000
Tennis racket	$ 13,100	$ 14,600	$ 15,200	$ 12,300
Bicycle	$ 11,400	$ 12,000	$ 28,000	$ 10,000

An example: a sales data cube

	Quarter 1	Quarter 2	Quarter 3	Quarter 4
Soccer shoes	$ 9,400	$ 10,000	$ 12,600	$ 10,600
Sleeping bag	$ 20,500	$ 13,700	$ 52,400	$ 21,000
Tennis racket	$ 13,100	$ 14,600	$ 15,200	$ 12,300
Bicycle	$ 11,400	$ 12,000	$ 28,000	$ 10,000

Sales of **sleeping bags** are particulary high in the **third quarter**?

An example: a sales data cube

		Quarter 3		
		June	July	August
Sleeping bag	Young	$ 9,300	$ 24,300	$ 19,100
	Adult	$ 1,200	$ 600	$ 1,600
	Old		$ 300	

Explanation

- **Summer season** and **young customers** are associated with high sales of **sleeping bags**
- **Young** $\wedge$ **July** $\Rightarrow$ **Sleeping bag**

An example: a sales data cube

		Quarter 3		
		June	July	August
Sleeping bag	Young	$ 9,300	$ 24,300	$ 19,100
	Adult	$ 1,200	$ 600	$ 1,600
	Old		$ 300	

Explanation

- **Summer season** and **young customers** are associated with high sales of **sleeping bags**
- Young $\wedge$ July $\Rightarrow$ Sleeping bag

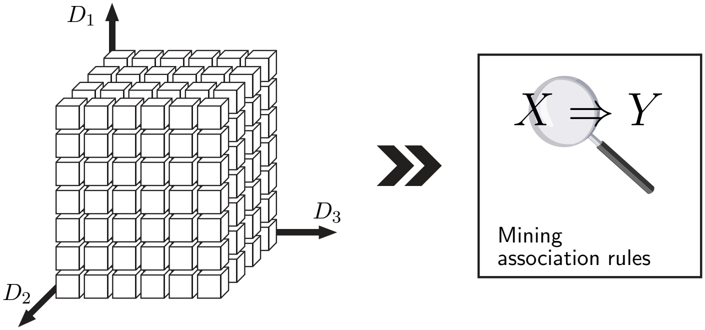
An example: a sales data cube

		Quarter 3		
		June	July	August
Sleeping bag	Young	$ 9,300	$ 24,300	$ 19,100
	Adult	$ 1,200	$ 600	$ 1,600
	Old		$ 300	

Explanation

- **Summer season** and **young customers** are associated with high sales of **sleeping bags**
- **Young** $\wedge$ **July** $\Rightarrow$ **Sleeping bag**

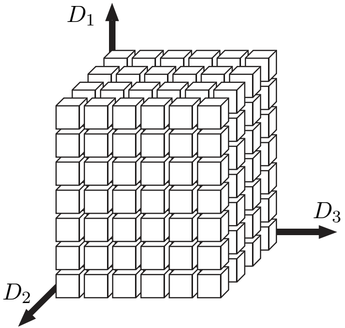
Objectives

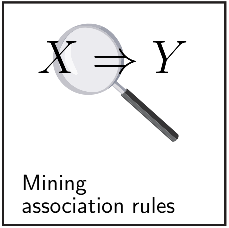


Key idea

Mine **association rules** in **data cubes** in order to **explain** relationships within **multidimensional data**.

Objectives





Key idea

Mine **association rules** in **data cubes** in order to **explain** relationships within **multidimensional data**.

Outline

Related Work

Traditional association rules

- Agrawal *et al.* (1993): the mining of association rules.
- Srikant and Agrawal (1995): *categorical data*.
- Han and Fu (1995): *multilevel* association rules.
- Srikant and Agrawal (1996): *quantitative* association rules.
- ...

Related Work
Traditional association rules

- Agrawal *et al.* (1993): the mining of association rules.
- Srikant and Agrawal (1995): *categorical data*.
- Han and Fu (1995): *multilevel* association rules.
- Srikant and Agrawal (1996): *quantitative* association rules.
- ...

Mining association rules in multidimensional data?

Related Work
Association rules in multidimensional data

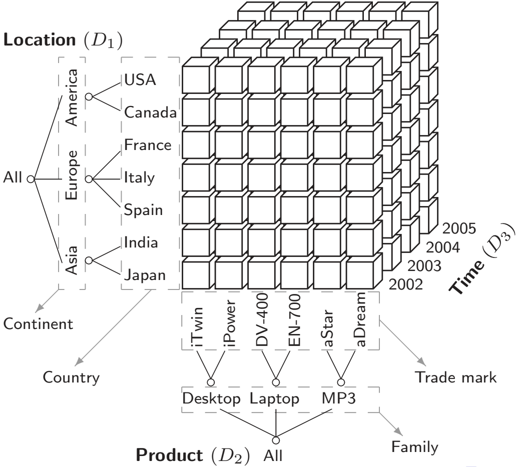
	Dimension		Level		Predicate		Measure		Application domain	
	Intra-dimensional	Inter-dimensional	Single level	Multiple levels	Repetitive	Non-repetitive	COUNT	All measures	Market basket analysis	General
Kamber *et al.* (1997)		•	•			•	•			•
Zhu (1998)	•	•	•		•	•	•		•	
Imieliński *et al.* (2002)		•		•	•			•		•
Dong *et al.* (2004)		•		•	•			•		•
Chen *et al.* (2000)	•			•	•		•		•	
Nestorov & Jukić (2003)	•		•		•		•		•	
Tjioe & Taniar (2005)	•	•		•	•	•	•			•
Our proposal (2006)										

Related Work

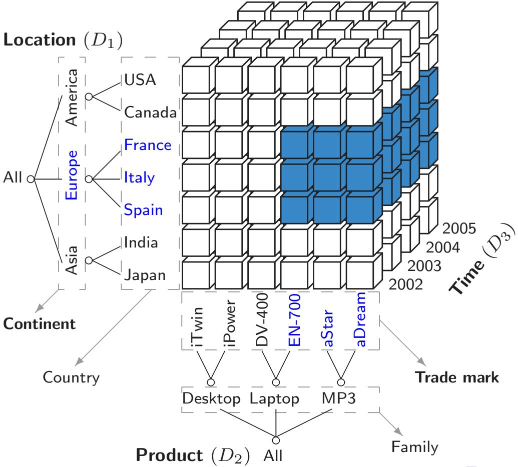
Association rules in multidimensional data

	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.* (1997)		•	•			•	•			•
Zhu (1998)	•	•	•		•	•	•		•	
Imieliński *et al.* (2002)		•		•	•			•		•
Dong *et al.* (2004)		•		•	•			•		•
Chen *et al.* (2000)	•			•	•		•		•	
Nestorov & Jukić (2003)	•		•		•		•		•	
Tjioe & Taniar (2005)	•	•		•	•	•	•			•
Our proposal (2006)		•	•			•		•		•

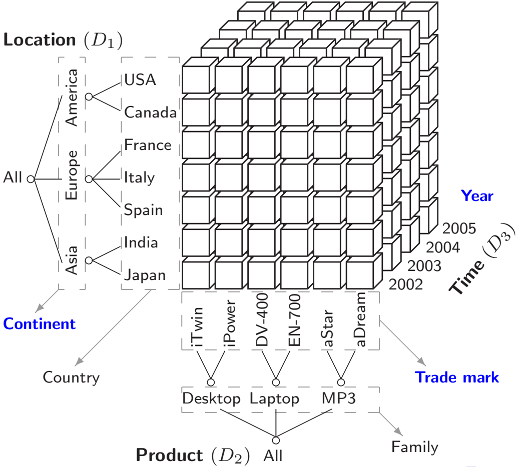
Our Framework
Sub-cube (example)



(Europe, {EN-700, aStar, aDream})

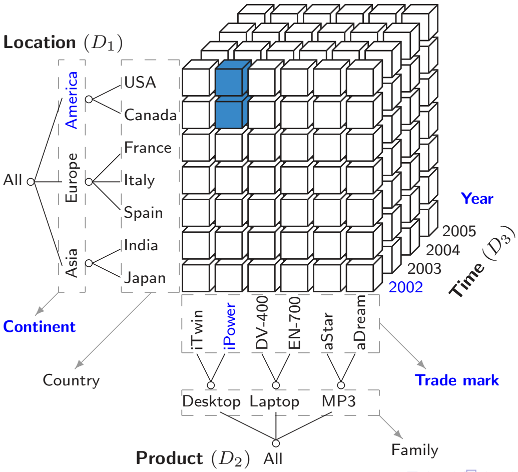


$$\langle a_1 \in \textbf{Continent} \rangle \wedge \langle a_2 \in \textbf{Trade mark} \rangle \wedge \langle a_3 \in \textbf{Year} \rangle$$



Our Framework
Inter-dimensional predicate (example)

$\langle$**America**$\rangle \wedge \langle$**iPower**$\rangle \wedge \langle$**2002**$\rangle$



Our Framework

Inter-dimensional meta-rules

We consider two distinct subsets of dimensions in the original data cube:

- $\mathcal{D}_\mathcal{C}$ is a subset of **context dimensions**
- $\mathcal{D}_\mathcal{A}$ is a subset of **analysis dimensions**

Inter-dimensional meta-rules

In the context **of a sub-cube** accoding to $\mathcal{D}_\mathcal{C}$
Head $\Rightarrow$ **Body**

- **Head** $\wedge$ **Body** is an inter-dimensional predicate in $\mathcal{D}_\mathcal{A}$

Our Framework
Inter-dimensional meta-rules

Example of an inter-dimensional meta-rule

- $\mathcal{D}_{\mathcal{C}}$ ={Profession, Gender}
- $\mathcal{D}_{\mathcal{A}}$ ={Location, Product, Time}

In the context $(\textbf{Student}, \textbf{Female})$
$\langle a_1 \in \textbf{Continent}\rangle \wedge \langle a_2 \in \textbf{Year}\rangle \Rightarrow \langle a_3 \in \textbf{Trade mark}\rangle$

Example of an inter-dimensional rule

R_1 In the context $(\textbf{Student}, \textbf{Female})$
$\textbf{America} \wedge \textbf{2004} \Rightarrow \textbf{Laptop}$

Example of an inter-dimensional meta-rule

- $\mathcal{D}_{\mathcal{C}}$ ={Profession, Gender}
- $\mathcal{D}_{\mathcal{A}}$ ={Location, Product, Time}

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

Example of an inter-dimensional rule

$$R_1 \left| \begin{array}{l} \text{In the context } (\textbf{Student}, \textbf{Female}) \\ \textbf{America} \wedge \textbf{2004} \Rightarrow \textbf{Laptop} \end{array} \right.$$

Example of an inter-dimensional meta-rule

- $\mathcal{D}_\mathcal{C}$ ={Profession, Gender}
- $\mathcal{D}_\mathcal{A}$ ={Location, Product, Time}

In the context (**Student**, **Female**)
$\langle a_1 \in \textbf{Continent}\rangle \wedge \langle a_2 \in \textbf{Year}\rangle \Rightarrow \langle a_3 \in \textbf{Trade mark}\rangle$

Example of an inter-dimensional rule

R_1 | In the context (**Student**, **Female**)
America $\wedge$ **2004** $\Rightarrow$ **Laptop**

▷ **How** to **compute** the **support** and the **confidence** of an inter-dimensional rule**?**

With the **COUNT** measure :

- the **support** and the **confidence** are computed according to the **frequency** of **units of facts** ;
- only the **number of facts** is taken into account to decide whether a rule is **large**, or **strong**, or not.

In OLAP context ...

- Users are usually interested in observing facts according to **summarized values** of **measures more expressive** than their simple **number of occurrences**.

With the **COUNT** measure :

- the **support** and the **confidence** are computed according to the **frequency** of **units of facts** ;
- only the **number of facts** is taken into account to decide whether a rule is **large**, or **strong**, or not.

In OLAP context ...

- Users are usually interested in observing facts according to **summarized values** of **measures more expressive** than their simple **number of occurrences**.

Our Framework
Support and confidence

With the **COUNT** measure :

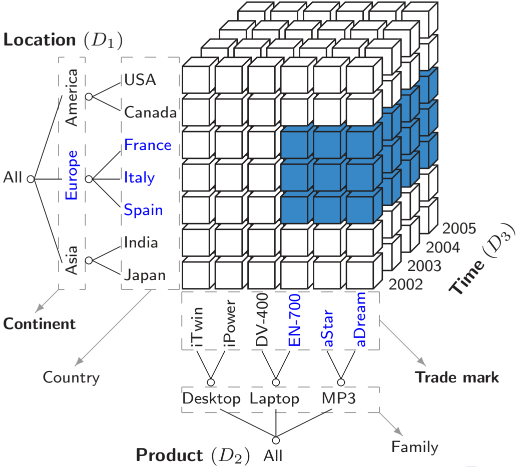
- the **support** and the **confidence** are computed according to the **frequency** of **units of facts** ;
- only the **number of facts** is taken into account to decide whether a rule is **large**, or **strong**, or not.

In OLAP context ...

- Users are usually interested in observing facts according to **summarized values** of **measures more expressive** than their simple **number of occurrences**.

▷ It is **more significant** to compute **support** and **confidence** according to the **SUM** of fact measures supporting the rule.

Profit(Europe, {EN-700, aStar, aDream})



Key idea

With the **sum-based aggregate measure**:

- The rule mining process can handle **any measure** in order to evaluate the **interestingness** of extracted association rules.
- A rule is evaluated according to the **quantity of measures** of its corresponding facts.
- Studied associations concern the population of **units of measures** of these facts.
- The choice of the measure closely depends on the **analysis objectives**.

Only support and confidence ...

Only support and confidence ...

- Support and confidence usually produce a **large number** of association rules.
- Some of extracted association rules **may not** be **interesting**.

Only support and confidence ...

- Support and confidence usually produce a **large number** of association rules.
- Some of extracted association rules **may not** be **interesting**.

▷ Use additional **interestingness criteria** of association rules.

Descriptive Vs. Statistical criteria

- A statistical criterion:
  - **depends** on the **size** of the **mined population**;
  - **loses its discriminating power** and tends to take a value close to one for large number of examples;
  - **requires** a **probabilistic approach** to model the mined population.
- A descriptive criterion:
  - is **easy to use** and **express interestingness** of association rules in a more **natural manner**.

Our Framework
Advanced evaluation of association rules

Descriptive Vs. Statistical criteria

- A statistical criterion:
    - **depends** on the **size** of the **mined population**;
    - **loses its discriminating power** and tends to take a value close to one for large number of examples;
    - **requires** a **probabilistic approach** to model the mined population.
- A descriptive criterion:
    - is **easy to use** and **express interestingness** of association rules in a more **natural manner**.

▷ We use **two descriptive criteria** : the **Lift** criterion (LIFT) and the **Loevinger** criterion (LOEV).

Our Framework

Advanced evaluation of association rules

For a rule $\mathbf{X} \Rightarrow \mathbf{Y}$:

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

Interpretation (Lift)

- **Deviation** of the support of the rule from the support expected under the independence hypothesis of the head and the body.
- **Scale coefficient** of having the body when head occurs.
- Greater **Lift** values indicate **stronger** associations.

For a rule **X** $\Rightarrow$ **Y**:

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

Interpretation (Loevinger)

- **Linear transformation** of the confidence in order to enhance it.
- **Expresses** the **confidence** according to the probability of not satisfying its head.
- Greater **Loevinger** values indicate **stronger** associations.

Search for large itemsets

- The **top-down** approach:
  - starts with k-itemsets and steps down to 1-itemsets;
  - if a k-itemset is frequent, then all sub-itemsets are frequent.
- The **bottom-up** approach:
  - starts from 1-itemsets to longer itemsets;
  - complies with the Apriori property "*for each non frequent itemset, all its super-itemsets are definitely not frequent*";
  - enables the **reduction** of the **search space**, especially when it deals with **large** and **sparse** data sets.

Search for large itemsets

- The **top-down** approach:
  - starts with k-itemsets and steps down to 1-itemsets;
  - if a k-itemset is frequent, then all sub-itemsets are frequent.
- The **bottom-up** approach:
  - starts from 1-itemsets to longer itemsets;
  - complies with the Apriori property "*for each non frequent itemset, all its super-itemsets are definitely not frequent*";
  - enables the **reduction** of the **search space**, especially when it deals with **large** and **sparse** data sets.

▷ We use the **bottom-up** approach **adapted** for the **context of data cubes**.

Our algorithm

- An **adaptation** of the Apriori algorithm for the **multidimensional data structure**.
- Directly extracts **inter-dimensional association** rules from data cubes.
- Enables a **guided-mining** process according to an inter-dimensional meta-rule **defined by users**.
- Extracts **significant** rules, for **OLAP users**, by taking into account **any measure** in the cube.
- Provides **advanced evaluation** of extracted associations by using **Lift** and **Loevinger**.

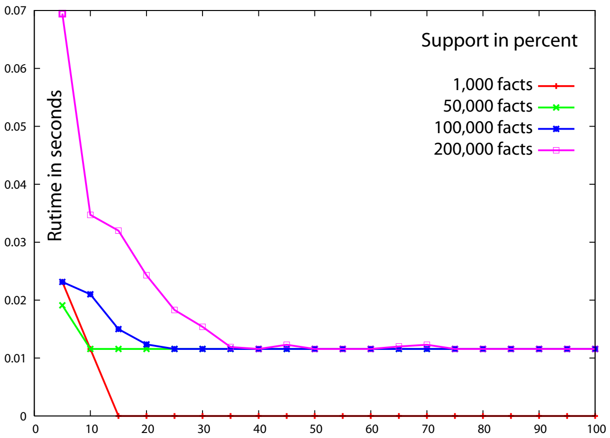
Proposed Algorithm
Implementation

Performance Evaluation
Configuration

- **Food Mart** data cube from **Analysis Services** of MS SQL Server 2000
- **System**: Windows XP
- **Processor**: Intel Pentium 4 (1.60GHz)
- **Main memory**: 480MB

Performance Evaluation

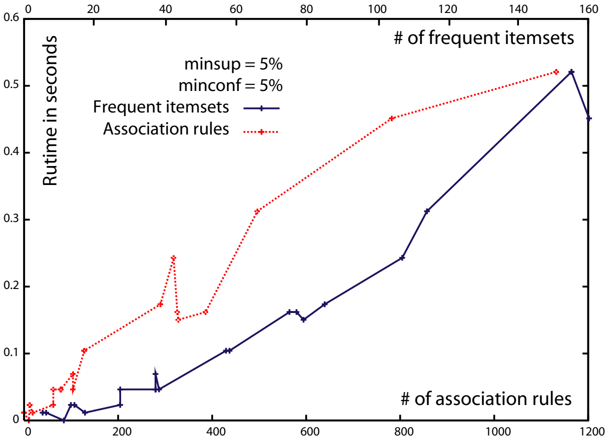
Runtime according to minsupp for different sizes



▷ For **large** *minsupp*, the mining process has already equal response times **independently** from the **number of mined facts**.

Performance Evaluation

Runtime according to # of frequent itemsets and # of association rules



▷ The generation of association rules from frequent itemsets is **more time consuming** than the extraction of frequent itemsets themselves.

1. A general framework for a **guided mining** of **inter-dimensional** association rules from data cubes.
2. **Inter-dimensional meta-rule** which allows users to limit the mining process to **specific contexts**.
3. A **general computation** of support and confidence that can be based on **any measure** from the data cube.
4. **Wide** analysis objectives **not restricted** to associations only driven by the COUNT measure.
5. **Interestingness** of mined rules according to two **additional** descriptive criteria (Lift and Loevinger).
6. An **adaptation** of the Apriori algorithm in order to **handle** multidimensional data.

Conclusion and Perspectives

Perspectives

1. Extension to handle inter-dimensional association rules **with repetitive predicates**.
2. Extension to handle **intra-dimensional** association rules.
3. Embedding the **measure** in the **expression** of mined association rules.
4. Profit from the **hierarchical aspect** of cube dimensions to mine **multi-level** association rules.
5. Cope with the **visualization** for an easier **interpretation** of mined associations by OLAP users.
6. Explore other approaches for association rule mining : **closed itemset generation** and **non-redundant rule generation**.

The end

Thank you for your attention!

Feel free to ask questions...