

Peculiarity Oriented Multi-Database Mining

Ning Zhong

The International WIC Institute (WICI)/BJUT
Maebashi Institute of Technology, Japan





Acknowledgements

- n 周志华教授, 王珏教授, 高阳博士, 费翔林博士
Organizers of the MLA 2005
- n Profs. Y.Y. Yao, C. Liu, J.L. Wu, S. Tsumoto,
T.Y. Lin etc.
- n Students and Post-doc/Visitors at:
 - n Maebashi Institute of Technology
 - n Beijing University of Technology



The Main Objective

Developing a framework for

(Attribute-Value/Relational)

*peculiarity oriented mining in multiple
(statistical / transaction / scientific)
databases for discovering interesting
rules/patterns.*

Keywords:

- 2 *Interestingness and Peculiarity (Oriented Mining)*
- 2 *Multi-Database Mining (MDM)*
- 2 *Relational Peculiarity Oriented Mining*



Interestingness and (Attribute-Value Level) Peculiarity Oriented Mining (POM)

**N. Zhong., Y.Y. Yao, and M. Ohshima:
Peculiarity Oriented Multi-Database Mining,
IEEE Transactions on Knowledge and Data Engineering,
Vol. 15, No. 4 (2003) 952-960.**



The Purpose of Data Mining

- | Hypotheses (knowledge) generated from DBs:
 - | Incorrect hypotheses
 - | Useless hypotheses
 - | *New, surprising, interesting hypotheses*
- | The purpose of DM is to discover *new, surprising, interesting* knowledge hidden in data.
- | The evaluation of interestingness (peculiarity, surprisingness, unexpectedness, usefulness, novelty) should be done in pre-processing and/or post-processing of the discovery process.



Interestingness Evaluation

Interestingness evaluation can be done in

- 2 **Pre-processing**: select interesting data before hypotheses (knowledge) generation
- 2 **Post-processing**: select interesting rules after hypotheses (knowledge) generation



Interestingness Evaluation (2)

Interestingness evaluation may be

- ² **Subjective** (user-driven):
 - ² asking the user to explicitly specify what type of rules are interesting and uninteresting.
 - ² the system then generates or retrieves those matching rules.
- ² **Objective** (data-driven):
 - ² Analyzing a rule's structure, predictive performance, statistical significance, and so forth.



Interestingness vs. Peculiarity

- ² **Peculiarity** is a kind of interestingness.
- ² Peculiarity relationships/rules (with common sense) may be hidden in a relatively **low** number of data.



Related Work

- 2 Bing Liu systematically investigated how to analyze the *subjective* interestingness of association rules in *post-processing* (IEEE Intelligent Systems, Vol.15, No. 5, 2001)
- 2 Freitas discussed on objective measures of rule surprisingness (PKDD'98, Springer LNAI 1510)
- 2 Hilderman and Hamilton discussed on interestingness measures for ranking discovered knowledge (PAKDD'01, Springer LNAI 2035)



Related Work (con.)

- | Exception rules (Suzuki, 1997)
- | “Rule + Exception” strategies/learning (J. Wang, Y.Y. Yao, F.Y. Wang, 03-05)
 - | Describing a regularity for a relatively small number of objects.
 - | Representing exceptions to the well-known facts with common sense (as a rule pair).



Peculiarity Rules

Peculiarity rules (Zhong et al, 99):

- Describing a relatively **low number** of objects.
- Representing well-known facts with **common sense**.
- Discovering from *peculiar data* by searching the relevance among the peculiar data.

$meatSale(low) \wedge vegetableSale(low) \wedge fruitsSale(low)$

$\rightarrow turnover(veryLow)$

$weather(typhoon) \rightarrow turnover(veryLow)$



Peculiar Data

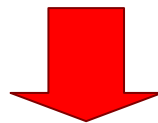
Peculiar data are a subset of objects in the database and are characterized by two features:

1. consisting of a relatively low number of objects (**frequency**).
2. very different from other objects in a dataset (**distance**).



A Supermarket Sales DB

Addr.	Date	MeatSale	VegetableSale	FruitsSale	...	Turnover
Ube	7/1	400	300	450	...	2000
...	7/2	420	290	460	...	2200
...
...	7/30	10	11	12		100
...	7/31	430	320	470		2500
...



*If meat-Sale(low) & vegetable-Sale(low) & fruits-Sale(low)
Then turnover(very-low)*

Mining Association Rules in a Supermarket DB



Ordinary Association Rule

*If mealSale(high) &
vegetableSale(high) &
fruitsSale(high)
Then turnover(veryHigh)*

Peculiarity Rule

*If mealSale(low) &
vegetableSale(low) &
fruitsSale(low)
Then turnover(veryLow)*

Why the peculiarity rule is an interesting one?



Background Knowledge on Information Granularity

- 2 Basic granules:
 - 2 $\{high, low\}, \{large, small\}, \{many, few\},$
 - 2 $\{far, close\}, \{long, short\}, \dots$
- 2 Specific granules:
 - 2 $biggest-cities = \{Tokyo, Osaka\}$
 - 2 $kanto-area = \{Tokyo, Tiba, Saitama, \dots\}$
 - 2 $kansei-area = \{Osaka, Kyoto, Nara, \dots\}$
 - 2 \dots

Qualitative Characterization of Three Classes of Rules

Rule	G (supp)	AS (conf)	CS	Semantic
Association rule: $f \Rightarrow y$	high	high	unknown	common-sense
Exception rule: $f \Rightarrow y$ $f \wedge f' \Rightarrow \neg y$	high low	high high	unknown high	exception
Peculiarity rule: $f \Rightarrow y$	<i>low</i>	high	high	<i>common-sense</i>



Observations

- 2 *Peculiarity rules* are a typical regularity hidden in a lot of scientific, statistical, and transaction DBs.
- 2 Sometimes, the standard association rules that represent the well-known fact with common sense cannot be found from numerous DBs, or although they can be found, the rules may be *uninteresting* ones to the user.
- 2 We focus on some *interesting (peculiar) data*, and then we can find more *novel / interesting (peculiarity) rules* from the data.



Peculiarity Oriented Mining

- ∅ The main task of peculiarity oriented mining (POM) is the **identification of peculiar data**.
- ∅ **Peculiarity Factor** (PF) evaluates whether x_{ij} occurs in relatively small number and is very different from other data x_{kj} .

$$PF(x_{ij}) = \sum_{k=1}^n D(x_{ij}, x_{kj})^a \quad (1)$$

$$D(x_{ij}, x_{kj}) = |x_{ij} - x_{kj}|$$

$$a = 0.5 \quad (\text{default})$$

(Attribute-Value Level) Peculiar Data Identification

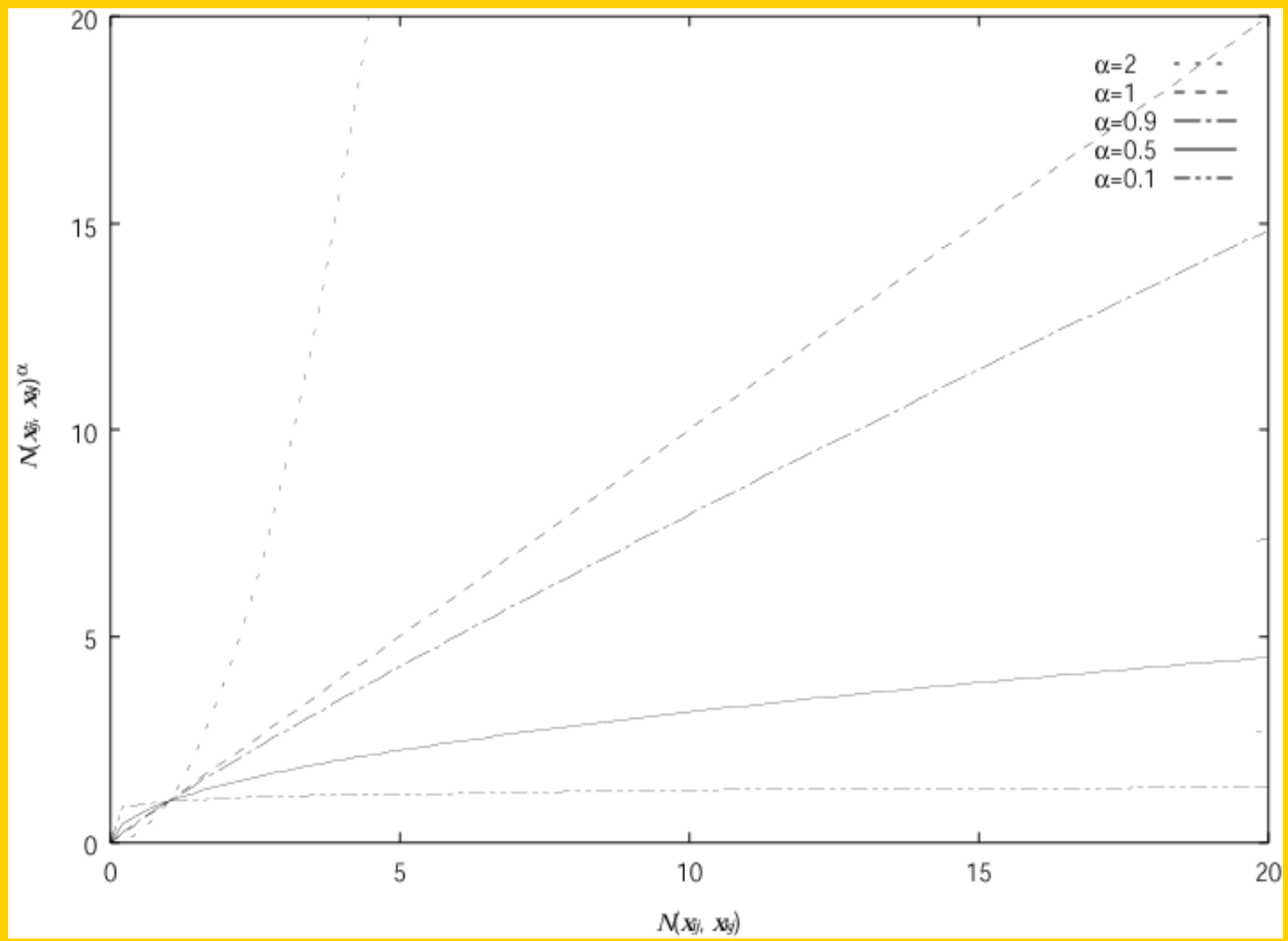
A_1	A_2	\dots	A_j	\dots	A_m
x_{11}	x_{12}	\dots	x_{1j}	\dots	x_{1m}
x_{21}	x_{22}	\dots	x_{2j}	\dots	x_{2m}
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots
x_{i1}	x_{i2}	\dots	x_{ij}	\dots	x_{im}
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots
x_{n1}	x_{n2}	\dots	x_{nj}	\dots	x_{nm}

$$\begin{aligned} PF(x_{ij}) &= \sum_{k=1}^n D(x_{ij}, x_{kj})^a \\ &= D(x_{ij}, x_{1j})^a + D(x_{ij}, x_{2j})^a + \dots + D(x_{ij}, x_{nj})^a \end{aligned}$$



Measure of Peculiarity

- 2 D denotes the conceptual distance.
- 2 The PF evaluates whether x_{ij} has a low **frequency** and is very different from other values x_{kj} (**distance**).
- 2 By adjusting the parameter a , a user can control and adjust the degree of PF that depends on both the **frequency** and the **distance**.
- 2 $a = 0.5$, as default, will get a good balance.



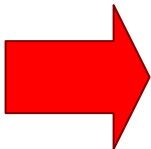


Measure of Peculiarity (2)

- ≈ Major merits using the **Peculiar Factor**:
 - It can handle both the **continuous** and **symbolic** attributes based on a unified semantic interpretation.
 - **Background knowledge** (BK) represented by binary neighborhoods can be used to evaluate the peculiarity if such BK is provided by a user.
- ≈ If X is a data set of a **continuous** attribute and no BK is available, in Eq. (1),

$$D(x_{ij}, x_{kj}) = |x_{ij} - x_{kj}| \quad (2)$$

An Example of the Peculiarity Factor for a Continuous Attribute

Region	ArableLand		<i>PF</i>
<i>Hokkaido</i>	<i>1209</i>		<i>131.4</i>
Tokyo	12		72.0
Niigata	196		58.2
Yamaguchi	162		54.3
Okinawa	147		55.1

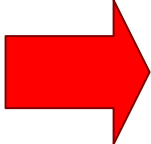


Measure of Peculiarity (3)

- 2 If X is a data set of a *symbolic* attribute and/or the BK for representing the conceptual distances between x_{ij} and x_{kj} is provided by a user, the peculiarity factor, $PF(x_{ij})$, is calculated by the **conceptual distances**, $D(x_{ij}, x_{kj})$.
- 2 If no background knowledge is available, the conceptual distances are assigned to 1, as default.

An Example of the Peculiarity Factor for a Symbolic Attribute

Restaurant	Type	
Wendy	<i>American</i>	2.2
Tokyo	<i>French</i>	2.6
Osaka	<i>Chinese</i>	1.6
Yamaguchi	<i>Japanese</i>	1.6
Okinawa	<i>Chinese</i>	1.6



<i>PF</i>
2.2
2.6
1.6
1.6
1.6



The Binary Neighborhoods for a Symbolic Attribute

Type	Type	N
<i>Chinese</i>	<i>Japanese</i>	1
<i>Chinese</i>	<i>American</i>	3
<i>Chinese</i>	<i>French</i>	4
<i>American</i>	<i>French</i>	2
<i>American</i>	<i>Japanese</i>	3
<i>French</i>	<i>Japanese</i>	3



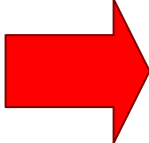
Measure of Peculiarity (4)

- ∅ After the PF evaluation, the peculiar data are selected by using a **threshold value**

$$p = \text{mean of } PF(\bullet) + g \times \text{standard deviation of } PF(\bullet) \quad (3)$$

where g can be specified by a user. That is, if $PF(x_{ij})$ is over the threshold value, x_{ij} is a peculiar data.

An Example of the Peculiarity Factor and the Threshold Value

Region	ArableLand		<i>PF</i>
<i>Hokkaido</i>	1209		131.4
Tokyo	12		72.0
Niigata	196		58.2
Yamaguchi	162		54.3
Okinawa	147		55.1

The mean of *PF* = 74.2

The standard deviation of *PF* = 29.3

The threshold value = 103.5

$$\gamma = 1$$



The Process of Finding the Peculiar Data

Step1: Execute *attribute oriented clustering* for each attribute respectively.

Step2: Calculate the peculiarity factor in Eq. (1) for all values in a data set (an attribute).

Step3: Calculate the threshold value in Eq. (3) based on the peculiarity factor obtained in **Step2**.

Step4: Select the data that is over the threshold value as the *peculiar data*.

Step5: If the current peculiarity level is enough, then go to **Step7**.



The Process of Finding the Peculiar Data (2)

Step6: Remove the *peculiar data* from the data set and thus, we get a new data set. Then go back to **Step2**.

Step7: Change the granularity of the *peculiar data* by using BK on information granularity if the BK is available.

Note: the process can be done in a parallel-distributed mode for multiple attributes, relations, and DBs.



Relevance Among the Peculiar Data

A *peculiarity rule* is discovered from the *peculiar data*, which belong to a cluster, by searching the relevance among the *peculiar data*.

Let $X(x)$ and $Y(y)$ be the *peculiar data* found in two attributes X and Y respectively. We deal with the following two cases:



Relevance Among the Peculiar Data (2)

- 2 If $X(x)$ and $Y(y)$ are symbolic data, the relevance between $X(x)$ and $Y(y)$ is evaluated in

$$R_1 = P_1(X(x) | Y(y))P_2(Y(y) | X(x)) \quad (4)$$

- 2 If both $X(x)$ and $Y(y)$ are continuous attributes, the relevance between $X(x)$ and $Y(y)$ is evaluated by using our KOSI system (Zhong-95).

Furthermore, Eq. (4) is suitable for handling more than two peculiar data found in more than two attributes if $X(x)$ (or $Y(y)$) is a granular of the peculiar data.



Application in Amino-acid Data Mining

- 2 Two groups of data:
 - 2 amino-acid matrix (VH and VL) and
 - 2 experimental data (combining coefficients and coefficients related thermodynamics etc.)
- 2 The main features of the data set:
 - 2 Too many attributes (230 + 7)
 - 2 Relatively small number of instances (35)



The Goal of Amino-acid Data Mining

- 2 Find the association between the amino-acid matrix and experimental data.
- 2 How experimental data change when amino-acid data are changed.

*See our paper (zhong et al, PAKDD-2001)
published at Springer LNAI 2035 for detail*



Multi-Database Mining (MDM) = Peculiarity Oriented Mining in Multiple Data Sources

J. Hu and N. Zhong: Organizing Multiple Data Sources for Developing Intelligent e-Business Portals, *Data Mining and Knowledge Discovery, An International Journal*, Springer (in press).

N. Zhong., Y.Y. Yao, M. Ohshima: Peculiarity Oriented Multi-Database Mining, *IEEE TKDE*, 15(4) (2003) 952-960.



Three Levels of MDM

➤ Mining from multiple relations (tables)

Although theoretically, any multi-RDB can be transformed into a single universal relation, practically this can lead to many issues such as

- universal relations of unmanageable size
- infiltration of uninteresting attributes
- losing of useful relation names
- unnecessary join operation
- inconvenience for distributed processing



Three Levels of MDM (2)

- 2 Mining from multiple relational DBs
 - In most organizations, data are rarely specially collected/stored in a database for the purpose of mining knowledge.
 - Some interesting patterns (concepts, regularities, causal relationships, and rules) cannot be discovered if we just search a single DB **since the knowledge hides in multiple DBs basically.**



Three Levels of MDM (3)

2 Mining from multiple mixed-media DBs

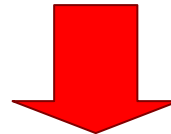
Many datasets in the real world contain more than just a single type of data.

For example, medical datasets often contain numeric data (e.g., test results), images (e.g., X-rays), nominal data (e.g., person smokes /does not smoke), and acoustic data (e.g., the recording of a doctor's voice).



A Supermarket Sales DB

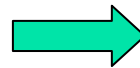
Addr.	Date	MeatSale	VegetableSale	FruitsSale	...	Turnover
Ube	7/1	400	300	450	...	2000
...	7/2	420	290	460	...	2200
...
...	7/30	10	11	12	...	100
...	7/31	430	320	470	...	2500
...



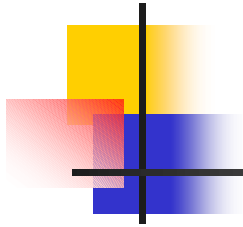
*If meat-Sale(low) & vegetable-Sale(low) & fruits-Sale(low)
Then turnover(very-low)*

Mining an Association Rule

(Peculiarity Rule) in a Transaction DB



*If mealSale(low) &
vegetableSale(low) &
fruitsSale(low)
Then turnover(veryLow)*



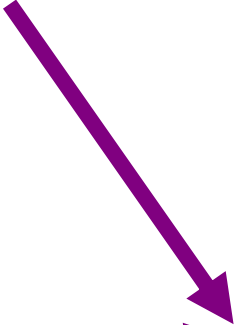

A Weather DB

Region	Date	...	Weather
yamaguchi	7/1	...	sunny
...	7/2	...	cloud
...
yamaguchi	7/30	...	<i>typhoon (no 2)</i>
...	7/31	...	cloud
...



Peculiarity Rules Mining

Peculiarity Rules:



*If mealSale(low) &
vegetableSale(low) &
fruitsSale(low)
Then turnover(veryLow)*



*If weather(typhoon)
Then turnover(veryLow)*



Related Topics

- 2 Interestingness/Peculiarity Evaluation
- 2 Relevance Analysis
- 2 ***Database Reverse Engineering (DRE)***
- 2 ***Granular Computing (GrC)***
- 2 Distributed Data Mining



Related Work on MDM

- 2 Ribeiro et al. (at KDD'95) described a way for extending the INLEN system for MDM by the incorporation of primary and foreign keys as well as the processing of knowledge segments.
- 2 Wrobel (at PKDD'97) extended the concept of foreign keys into foreign links because MDM is also interested in getting to non-key attributes.
- 2 Huan Liu et al. proposed an interesting method for relevance measure and an efficient implementation for identifying relevant DBs as the first step for MDM (Proc. PAKDD-98).



Granular Computing (GrC)

- GrC provides a useful way to find/create the relevance among different DBs by changing information granularity, to group the attribute values, and to learn the different concept hierarchy.



Mining Peculiarity Rules in MD

- 2 MDM in Different Levels:
 - 2 *Mining from multiple relations (tables)*
 - 2 *Mining from multiple relational DBs*
 - 2 Mining from multiple mixed-media DBs.
- 2 Mining from Multiple Relations:
 - 2 Selecting n relations, which contain the *peculiar data*, among m relations ($m \geq n$) with foreign links.



Relevance among Peculiar Data in MD

If the $X(x)$ and $Y(y)$ are found in two different relations, we need to use a value (or its granule) in a key (or foreign key/link) as the relevance factor, $K(k)$, to find the relevance between $X(x)$ and $Y(y)$. Thus, the relevance between $X(x)$ and $Y(y)$ is evaluated in:

$$R_2 = P_1(K(k) | X(x))P_2(K(k) | Y(y))$$



Process of Selecting n Relations from m Relations

Step1: Focus on a relation as the *main table* and find the *peculiar data* from this table. Then elicit the *peculiarity rules* from the *peculiar data*.

Step2: Find the value(s) of the focused key corresponding to the mined *peculiarity rule* in *Step1* and change its granularity of the value(s) of the focused key if the BK on its granularity is available.



Process of Selecting n Relations from m Relations (2)

Step3: Find the *peculiar data* in other relations (or DBs) corresponding to the value (its granule) of the focused key.

Step4: Select n relations that contain the *peculiar data*, among m relations $m \geq n$.

We just select the relations that contain the peculiar data, and that are relevant to the peculiarity rules/data mined from the main table.

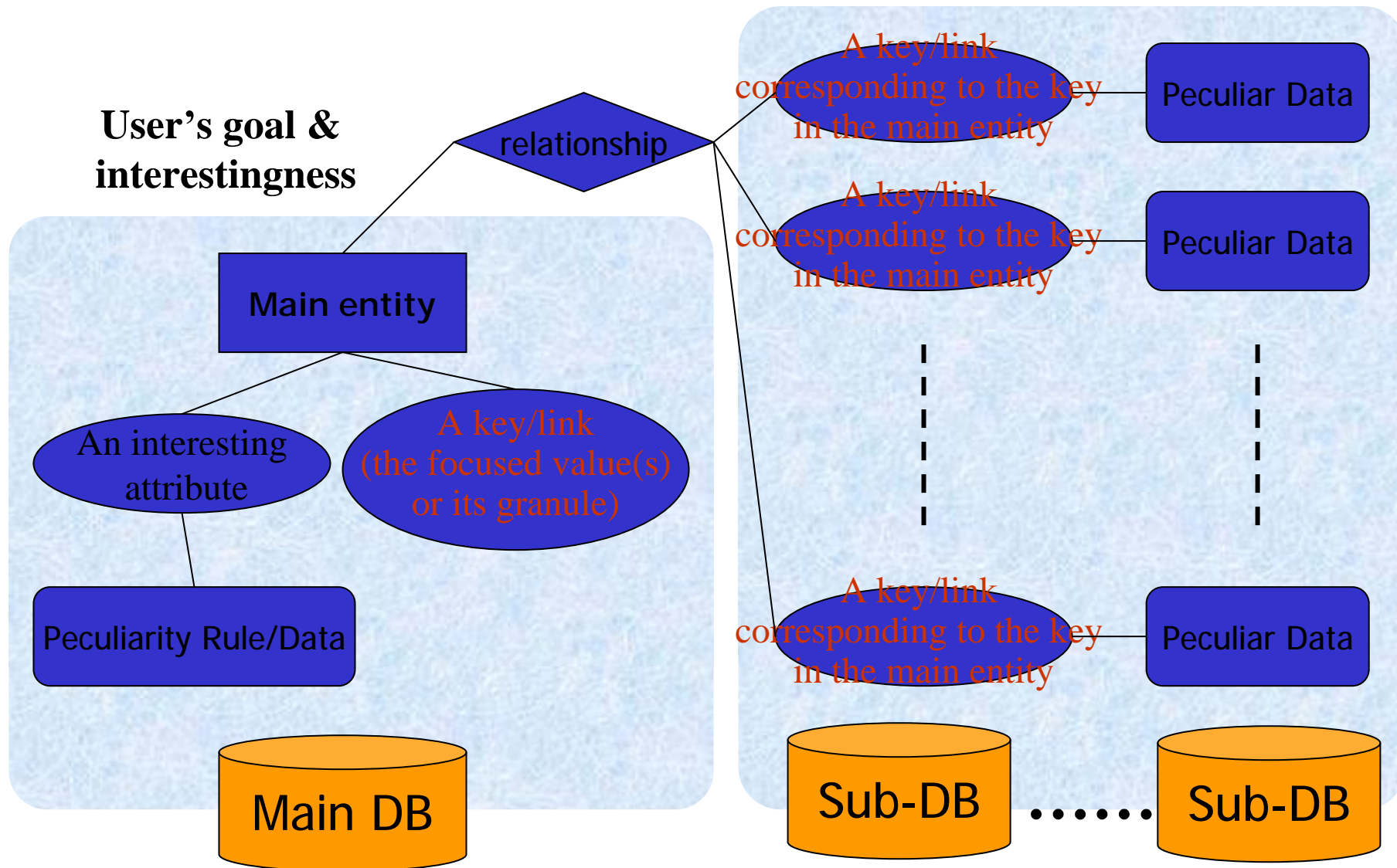


How to Represent/Learn in a Cluster of Multiple Data Sources?

- Represent the conceptual relationships among *various types of interesting data* (including *peculiar data*) mined from multiple data sources by using the **RVER model**.
- Learn the **advanced rules** (hidden **inter-multiple data sources**) from the RVER model in a **parallel-distributed cooperative** mode.

RVER: Reverse Variant Entity-Relationship

The RVER Model





Multi-Relation in the Japan-Survey DB

Japan-Geography

Region	Area	...	PopDensity	PeasantsN	ArableLand	Forest
<i>Hokkaido</i>	82410.58	...	<i>67.8</i>	93	<i>1209</i>	<i>5355</i>
Aomori	9605.45	...	156.8	87	169	623
...
Tiba	5155.64	...	1100.3	116	148	168
<i>Tokyo</i>	2183.42	...	<i>5317.2</i>	21	<i>12</i>	<i>80</i>
...
<i>Osaka</i>	1886.49	...	<i>4531.6</i>	39	<i>18</i>	<i>59</i>
...



Multi-Relation in the Japan-Survey DB (2)

Economy

Region	PrimaryInd.	TertiaryInd.	...
<i>Hokkaido</i>	<i>9057</i>	<i>96853</i>	...
Aomori	2597	22722	...
...
Tiba	3389	76277	...
Tokyo	839	484294	...
...
Osaka	397	209492	...
...



Multi-Relation in the Japan-Survey DB (3)

Alcoholic-Sales

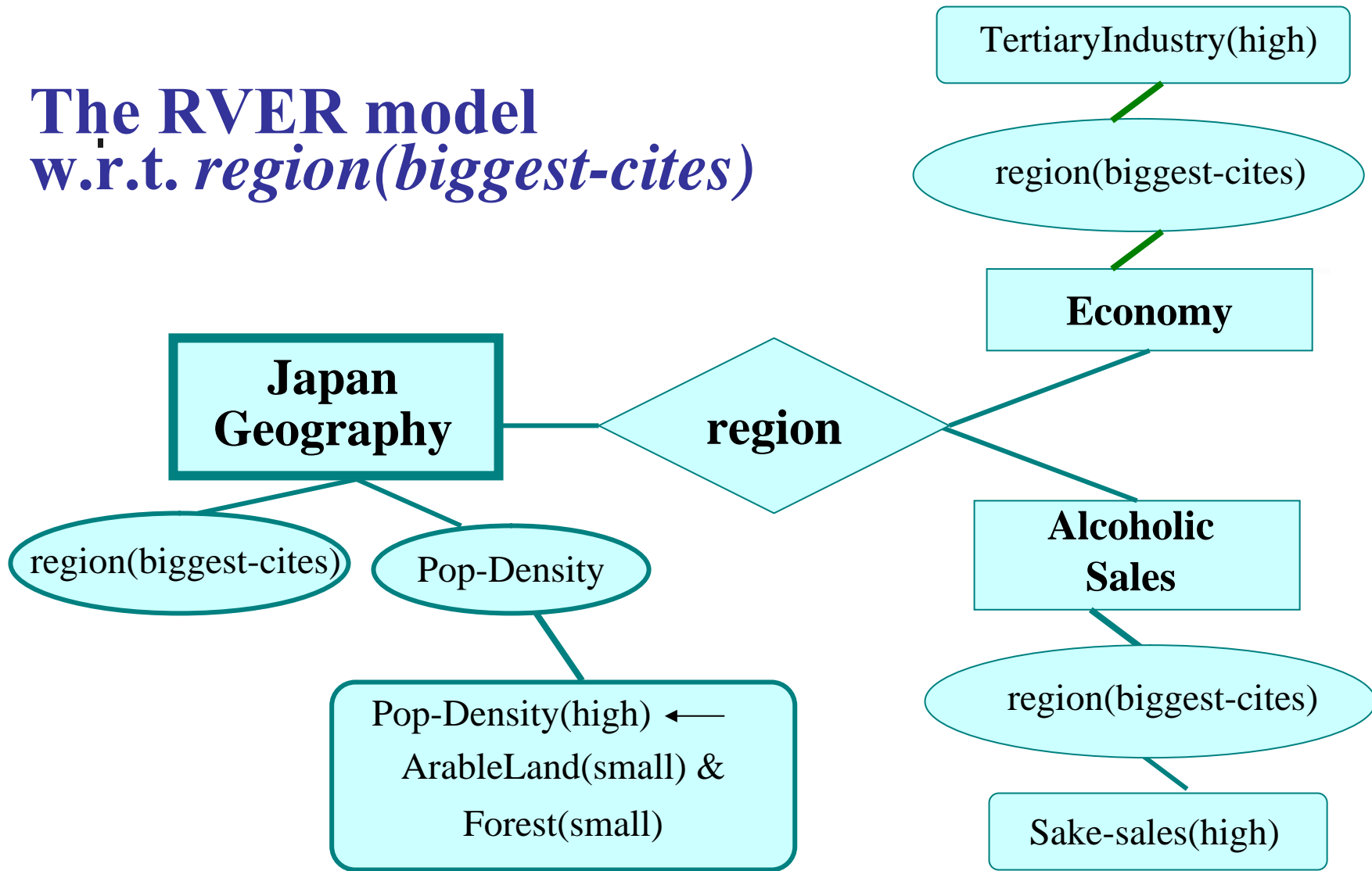
Region	Sake	Beer	...
Hokkaido	42560	257125	...
Aomori	18527	60425	...
...
Tiba	47753	205168	...
Tokyo	150767	838581	...
...
Osaka	100080	577790	...
...



Multi-Relation in the Japan-Survey DB (4)

- 2 Other relations:
 - 2 Crops, Livestock-Poultry, Forestry, Industry, etc.
- 2 Basic granules:
 - 2 *{high, low}, {large, small}, {many, few},*
 - 2 *{far, close}, {long, short}, ...*
- 2 Specific granules:
 - 2 *biggest-cities = {Tokyo, Osaka}*
 - 2 ...

The RVER model w.r.t. *region(biggest-cites)*



mining in inter-multiple data sources

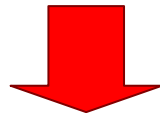


$arableLand(small) \wedge forest(small) \rightarrow tertiaryIndustry(high)$

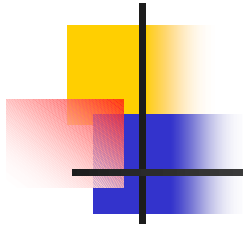


A Supermarket Sales DB

Addr.	Date	MeatSale	VegetableSale	FruitsSale	...	Turnover
Ube	7/1	400	300	450	...	2000
...	7/2	420	290	460	...	2200
...
...	7/30	10	11	12	...	100
...	7/31	430	320	470	...	2500
...



*If meat-Sale(low) & vegetable-Sale(low) & fruits-Sale(low)
Then turnover(very-low)*



A Weather DB

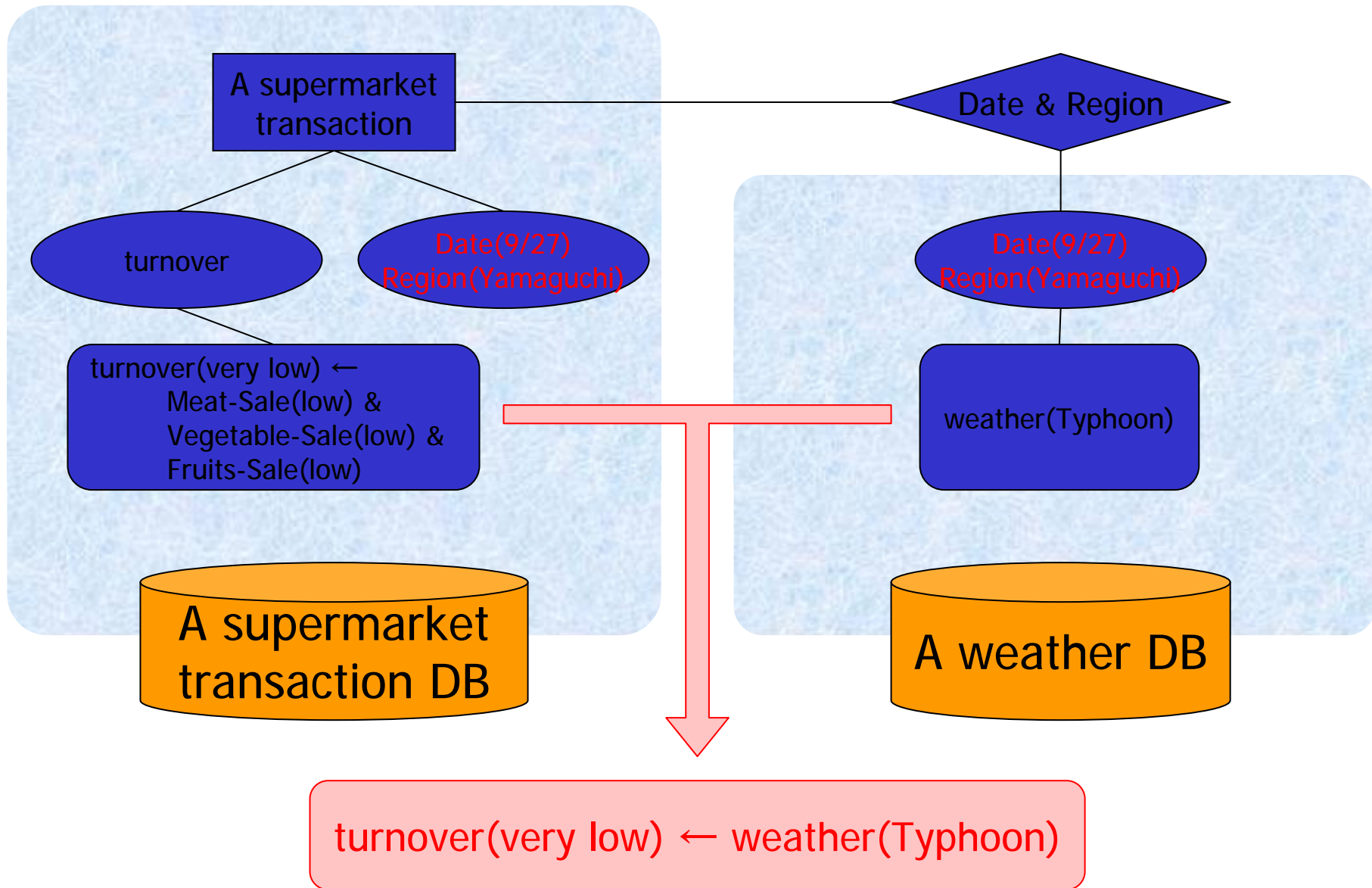
Region	Date	...	Weather
yamaguchi	7/1	...	sunny
...	7/2	...	cloud
...	
yamaguchi	7/30	...	<i>typhoon (no 2)</i>
...	7/31	...	cloud
...



BK for Granular Changing

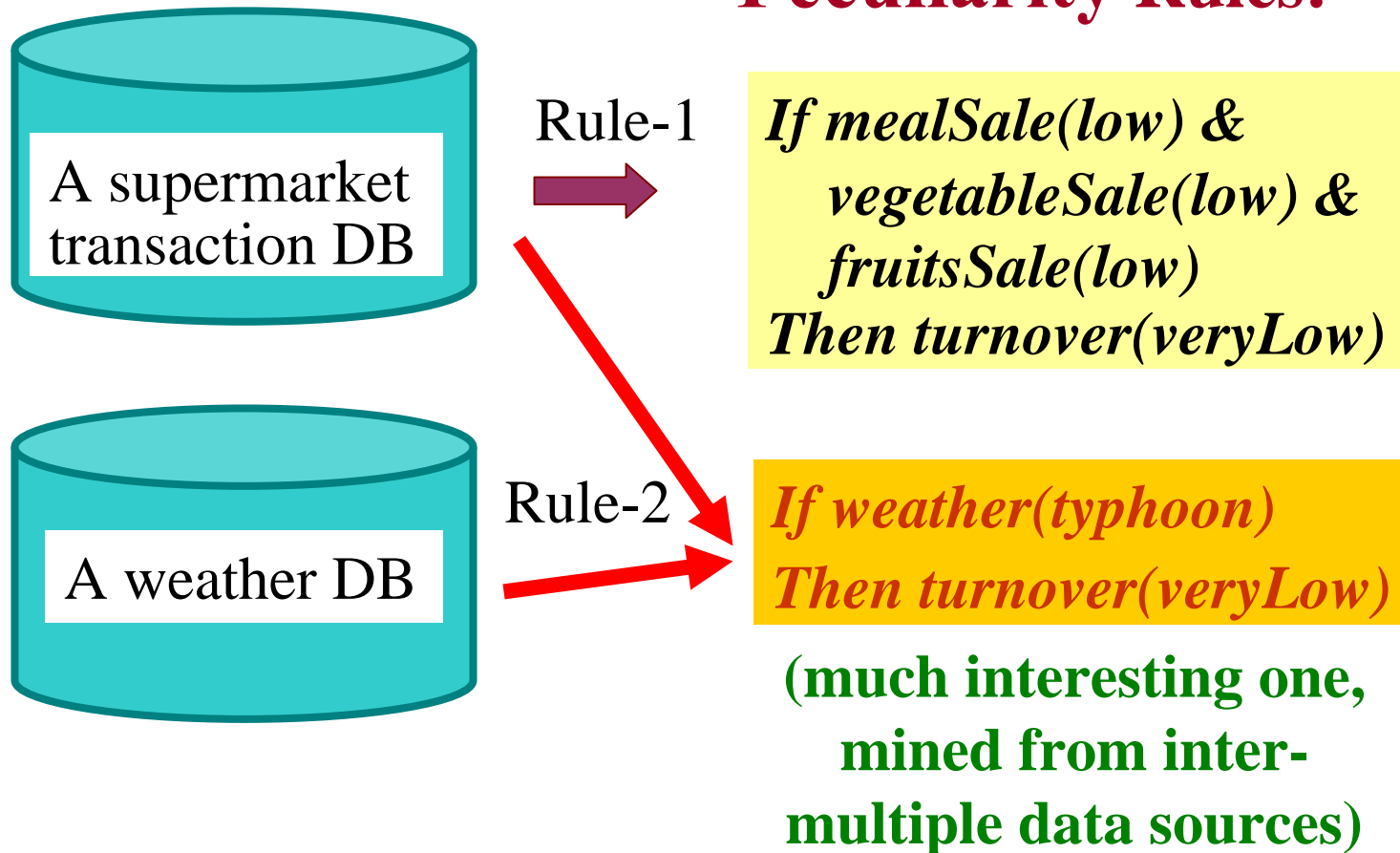
- 2 *Yamaguchi* = {*Ube, Shimonoseki, Hagi, Tokuyama, Yamaguchi-city*}
- 2 *Hiroshima* =

RVER Representation

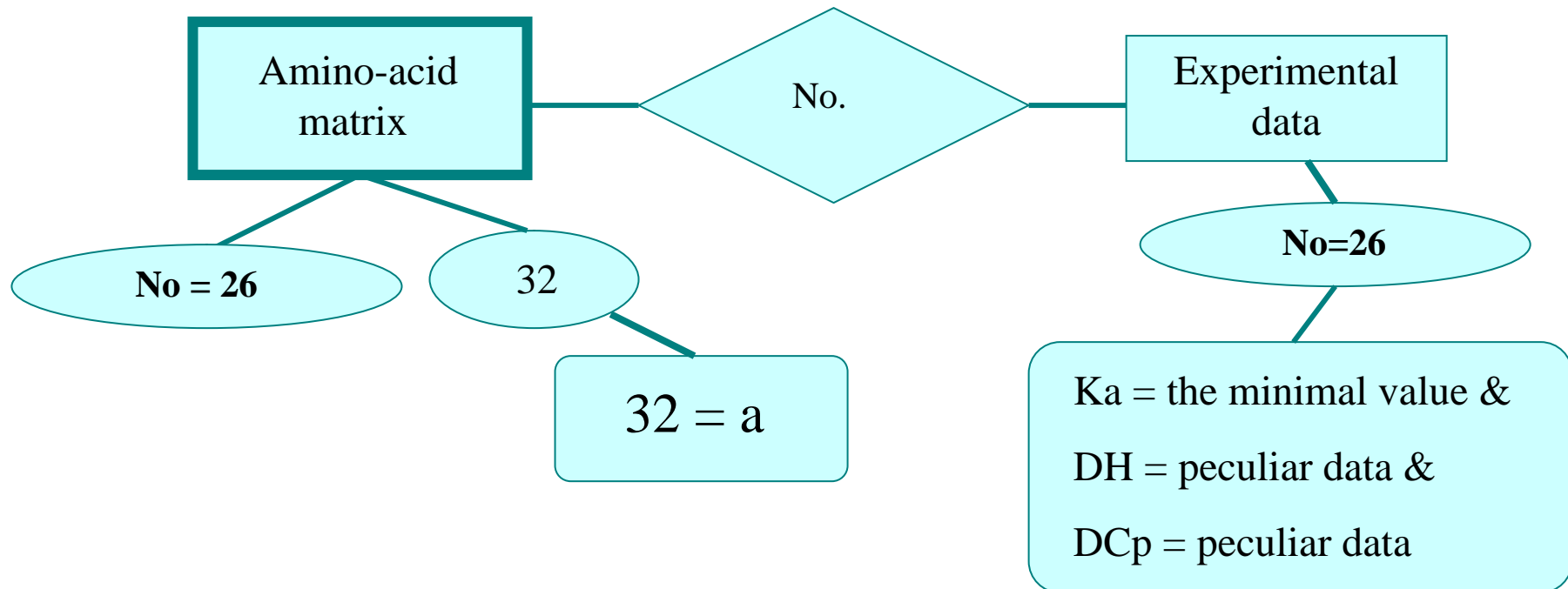


Peculiarity Rules Mined from Multiple Data Sources

Peculiarity Rules:



A RVER Model in Amino-acid Data Mining





Discovered Rules in the RVER Model from the Amino-acid Data

If the value in 32 of VL amino-acid matrix is
changed to *a*,

Then the value of *Ka* is the minimum one and
the values of *DH* and *DCp* are peculiar ones

If the value of *Ka* is the minimum one and
the values of *DH* and *DCp* are peculiar ones,
Then the value in 32 of VL amino-acid matrix
is changed to *a*



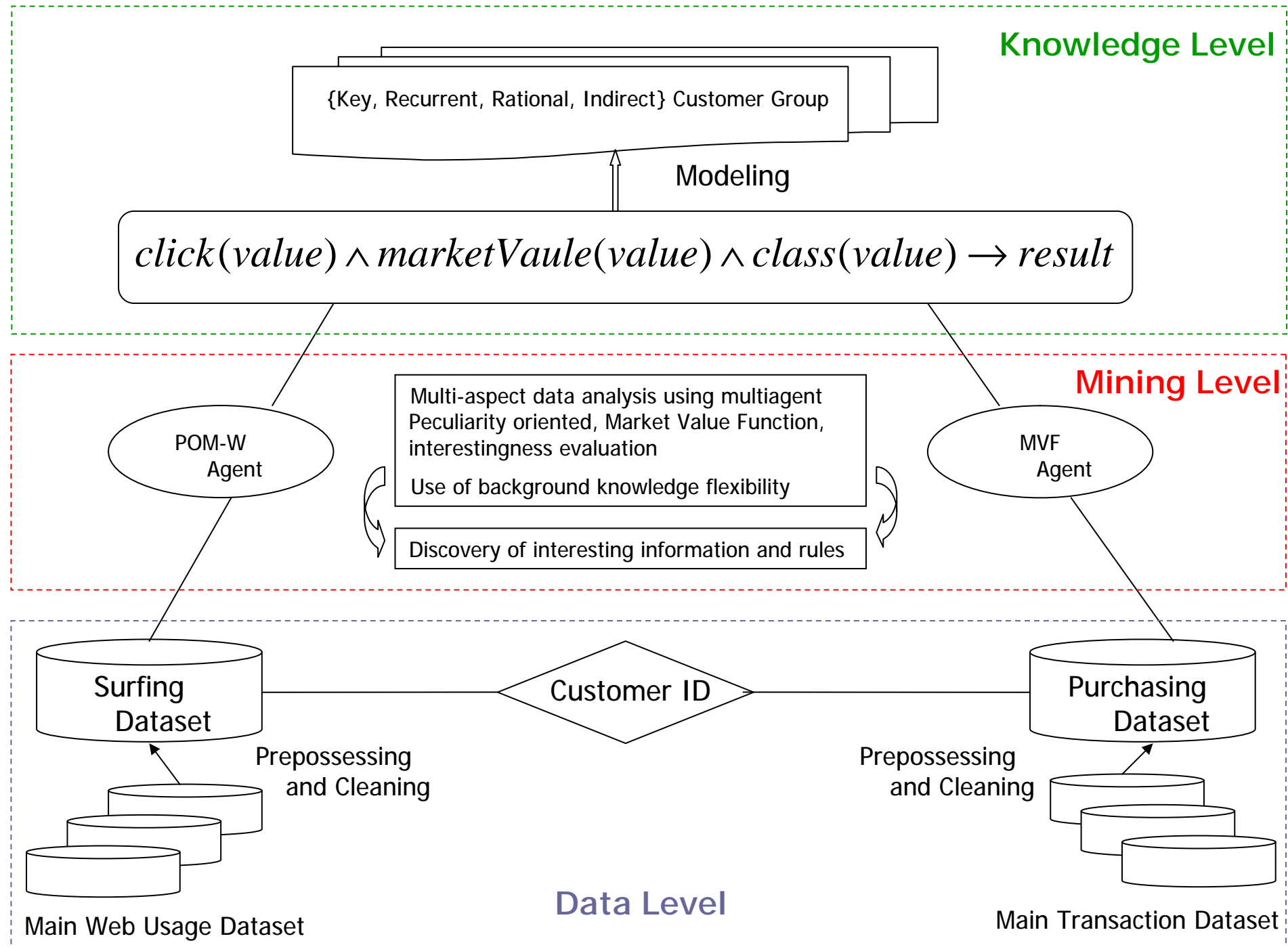
MDM in Different Levels

- 2 ***Mining from multiple relations (tables)***
- 2 ***Mining from multiple relational DBs***
- 2 **Mining from multiple mixed-media DBs.**



Example of Mining from Multiple Mixed-media DBs

- 2 Mining in Web based multiple information sources
 - 2 Usage Mining
 - 2 Content Mining
 - 2 Structure Mining
- 2 Mining in tracked image sequences of multiple people
- 2 Mining in fMRI brain images and EEG brain waves





Relational Peculiarity Oriented Mining

**M. Ohshima, N. Zhong, Y.Y. Yao, C. Liu:
Relational Peculiarity Oriented Mining,
Data Mining and Knowledge Discovery,
An International Journal, Springer (in press)**



Two Levels of Peculiarity

2 The **attribute-value** level

The **attribute-oriented methods** focus on finding peculiar attribute rules.

2 The **record** level

The **record-oriented methods** look for surprising, interesting patterns by analyzing the relationship among peculiarity entities.



Types of Inductive Learning vs. POM

- n Attribute-Value Learning (Propositional Logic)
(e.g. GDT-RS, C4.5) – **Attribute-Based POM**
- n Relation Learning (Predicate Logic)
(or called ILP: Inductive Logic Programming,
First-Order Learning) – **Record-Based POM**



Advantages of ILP

(Compared with Attribute-Value Learning)

- n It can learn knowledge which is **more expressive** because it is in predicate logic.
- n It can utilize **background knowledge** more naturally and effectively because in ILP the examples, the background knowledge, as well as the learned knowledge are all expressed within the same logic framework.



The Main Objective

- 2 Propose a framework for **record**-oriented peculiarity analysis.
- 2 By drawing results from **relational mining**, we present a model for **mining relational peculiarity rules** (RPR).
- 2 The relational peculiarity oriented mining (RPM) deals with two tasks, namely, **description** and **explanation**.



Peculiarity Oriented Relational Mining Based on ILP

- 2 One of task of mining relational peculiarity rules is the identification of peculiar records.
 - 1' . Peculiar records represent a relatively small number of records (**frequency**) and
 - 2' . Those records are very different from other records in the database (**distance**).

- 2 Relational peculiarity rule mining is a kind of multi-database mining.



Peculiar Record Identification

Let X denote a record set in a relation A , that is, $X = \{X_1, X_2, \dots, X_n\}$. A record X_i is represented by $\{x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{im}\}$, where x_{ij} denotes the value of the X_i on attribute a_j .

Record Peculiarity Factor (RPF)

$$RPF(x_i) = \sum_{k=1}^n \sqrt{\sum_{j=1}^m \beta_j \times (PF(x_{ij}) - PF(x_{kj}))^2} \quad (5)$$

where β_j is the weight of attribute β_j which depends on the knowledge provided by a user. $\beta_j=1$ is used as default.



Merits of the Peculiar Factors

- 2 It can handle both the **continuous** and **symbolic** attributes based on a **unified semantic interpretation**.
- 2 **Background knowledge** represented by **binary neighborhoods** can be used to evaluate the peculiarity if such BK is provided by a user.



Calculating $D(x_{ij}, x_{kj})$ in the RPF

Let **prior knowledge** denote the knowledge provided by a user. In Eqs. (1) and (2), $D(x_{ij}, x_{kj})$ can be calculated as follows:

(1) a_j is a **non-key** attribute.

If a_j is a **numerical attribute** and no BK is available,

$$D(x_{ij}, x_{kj}) = |x_{ij} - x_{kj}| \quad (6)$$



The Standardization of Attributes

- ≈ It is obvious that the measurement unit of attribute a_i will affect the result of Eq. (6).
In turn, the result of Eq. (6) will affect the result of Eq. (5).
- ≈ Standardizing measurements attempt to make all variables to have an **equal contribution**.
- ≈ To convert the original measurements to unitless variables.



The Standardization of Attributes (2)

- 2 Let x_{1j}, \dots, x_{nj} be values of attribute a_j , m_j be their mean value, and σ_j be the standard deviation. Then, the values x_{ij} and x_{kj} can be standardized respectively.

$$x'_{ij} = \frac{x_{ij} - m_j}{S_j}, \quad x'_{kj} = \frac{x_{kj} - m_j}{S_j}.$$

- 2 Eq. (6) is transformed into

$$D(x_{ij}, x_{kj}) = |x'_{ij} - x'_{kj}|.$$

- 2 This method is valid to standardize **interval-scaled variables**.



Issue of Variable Standardization

- 2 A difficulty with variable standardization is that we may not have the information of the variable type.
- 2 Instead of standardization variables, we can scale **the peculiarity factor**,

$$PF'(x_{ij}) = \frac{|PF(x_{ij}) - m'_j|}{s'_j}.$$



Issue of Variable Standardization (2)

\approx a_j is a symbolic or continuous attribute and **prior knowledge is available**. $D(x_{ij}, x_{kj})$ is defined by the prior knowledge.

\approx a_j is a **symbolic attribute** (or other type of attribute) and no prior knowledge is available.

$$D(x_{ri}, x_{ji}) = 1.$$

\approx a_j is a **date attribute**. $D(x_{ij}, x_{kj})$ denotes the interval between x_{ij} and x_{kj} .



Issue of Variable Standardization (3)

(2) a_j is a **key** attribute.

\approx a_j is a **main key** of relation A . In this case, a_j is not a factor for the distance because the main key is the identification of a record.

\approx a_j is a **foreign key** of relation A . Relations A and B are linked by both attribute a_j in A and attribute b_k in B . Let $Y_i = \{y_{i1}, y_{i2}, \dots, y_{im}\}$ be the i th record in relation B , where y_{ij} denotes the value of the i th record on attribute b_j in relation B . For a record X_i , we can find a record $Y_{i'}$ such that $x_{ij} = y_{i'j}$. Thus, we first obtain x'_{ij} by $x'_{ij} = RPF(Y_{i'})$ and then the peculiarity of x_{ij} can be computed by $PF(x_{ij}) = PF(x'_{ij})$ after replacing the values of a_j by x'_{ij} .



Measure of Peculiarity

- ² After the evaluation for the peculiarity, we use a **threshold** value pr to test if a *peculiar record* exist or not,

$$pr = \text{mean of } RPF(\bullet) + \gamma \times \text{standard deviation of } RPF(\bullet) \quad (7)$$

where γ can be specified by a user.

That is, if $RPF(X_i)$ is over the threshold value, X_i is a *peculiar record*.



Finding Peculiar Records

Step1: Calculate the peculiarity factor $RPF(X_j)$ for all records in a relation.

Step2: Calculate the threshold value based on the result obtained in ***Step1***.

Step3: Select the records that is over the threshold value as *peculiar records*.

Step4: If the current peculiarity level is enough, then go to ***Step6***.

Step5: Remove *peculiar records* from the current records and thus, we get a new record set. Then go back to ***Step1***.

Step6: End.



Generating Peculiarity Rules

Step1: Select a main relation as a targeted relation and identify other relations related to the main relation in a database.

Step2: Find peculiar records in the main relation and other related relations.

Step3: Identify the relations related to the main relation in a database.

Step4: Obtain the input document for FOIL through the target predicate and BK.

Step5: Get the learned result by using FOIL.

Step6: Evaluate the result.



Experiments

- ∅ The relational peculiarity mining algorithms are applied to analyze the China Statistics Yearbook database (www.efair.gov.cn) with 46 relations/tables such as *EstatePrice* and *income*.
- ∅ Relation *EstatePrice* reflects the sale prices of residence, villa, apartment, economic house, office building/business building, and so on, in cities of China.
- ∅ Relation *income* gives the annual income per head of families in various areas

Target: *Why the sale price of house in Beijing and Shanghai is different from that in other areas?*



Experiments (2)

- 2 We choose relation *EstatePrice* as the main table and information in relation *income* as BK.
- 2 We analyzed the target attribute and peculiarity in related tables. The discretization and attribute selection are also carried out as steps of preprocessing, respectively.

estateprice-building(A,B) :-

income-nationalcorp(A,B),

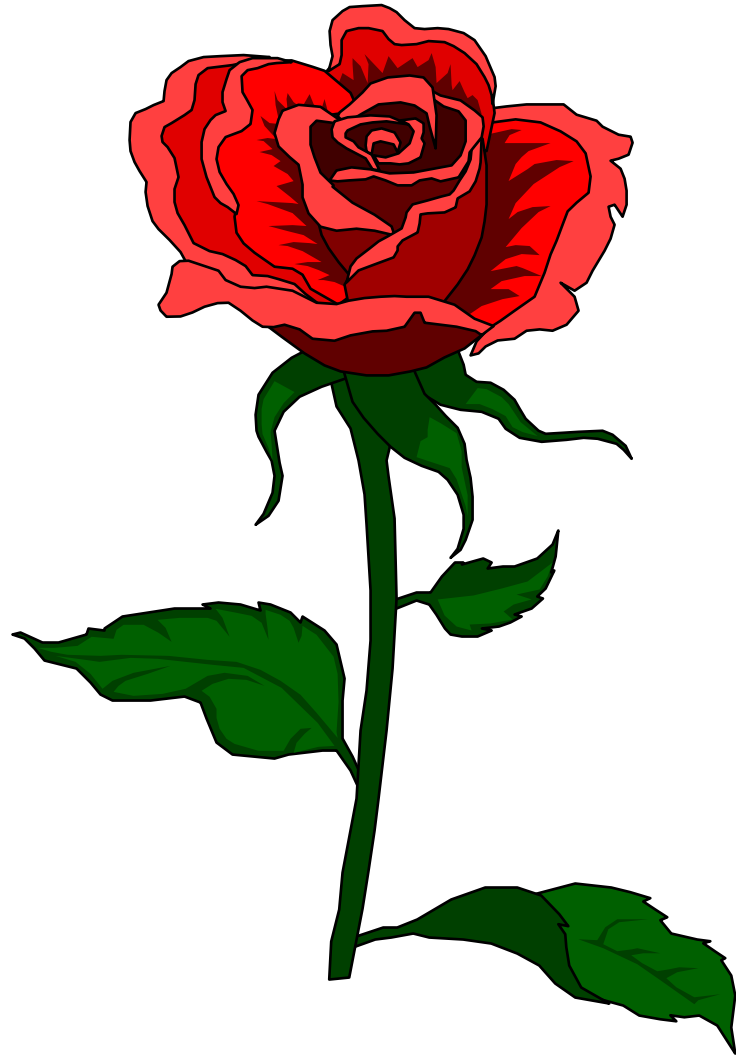
income-disposal(A,B).



Some Interesting Rules Obtained

- 2 If both disposable-income and income of the state-owned unit employee are high in a certain area, the price of building of this area is high.
- 2 If both fact-income and income of the state-owned unit employee are high in a certain area, the price of building of this area is high.
- 2 If income of the state-owned unit employee and other income are high, the price of building of this area is high.

The rules verify the notion that if the income of one area is high then the price of house is high.





Applications

Japan-survey, weather, supermarket, amino-acid, hepatitis, tracked multi-people images, fMRI brain images and EEG brain waves etc.

See our papers at PKDD'99, ICDM'01, PAKDD'01, IDEAL'03, PAKDD'04, ATELS'04, ICDM'04, SIANT'05

TKDE-15(4) 2003, CSR-5(3) 2004, CI-21(2) 2005, DMKD 2005 etc.



Peculiarity Oriented Mining in Tracked Multi-People Images

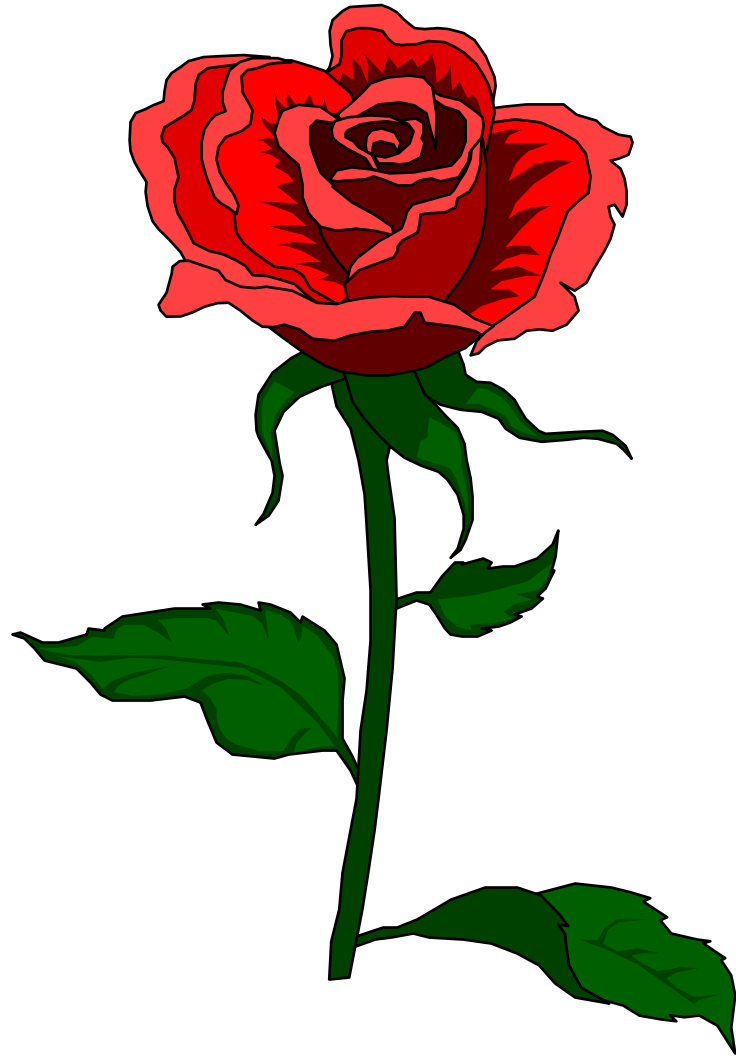
**M. Ohshima, N. Zhong, Y.Y. Yao, C. Liu:
Relational Peculiarity Oriented Mining,
Data Mining and Knowledge Discovery,
An International Journal, Springer (in press)**



Peculiarity Oriented Multi-aspect Brain Data Analysis

N. Zhong, J.L. Wu, A. Nakamaru, M. Ohshima, H. Mizuhara: Peculiarity Oriented fMRI Brain Data Analysis for Studying Human Multi-Perception Mechanism, *Cognitive Systems Research*, An International Journal, Elsevier (2004) 241-256.

N. Zhong, J. Hu, S. Motomura, J. Wu, C. Liu: Building a Data Mining Grid for Multiple Human Brain Data Analysis, *Computational Intelligence*, An International Journal, Blackwell, 21(2), (2005) 177-196.





Concluding Remarks

- 2 We proposed a framework of **POM**.
- 2 There are at least **two levels** of peculiarity in DBs.
 - The **attribute-value** level peculiarity reveals the local characteristics of a record.
 - The **record** level peculiarity reflects the overall characteristics of a record.

Both levels of peculiarity are potentially useful for understanding knowledge hidden in DBs.

- 2 The experimental results demonstrate that POM is potentially useful and effective.



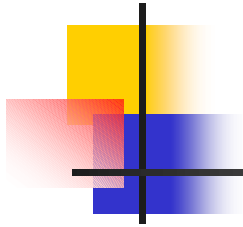
Major Publications on POM

- n Ohshima, M., Zhong, N., Yao, Y.Y. Yao, C. Liu., "Relational Peculiarity Oriented Multi-Database Mining", *Data Mining and Knowledge Discovery*, An International Journal, Springer (in press).
- n Hu, J. and Zhong, N., "Organizing Multiple Data Sources for Developing Intelligent e-Business Portals, *Data Mining and Knowledge Discovery*, An International Journal, Springer (in press)
- n N. Zhong, J. Hu, S. Motomura, J. Wu, C. Liu: Building a Data Mining Grid for Multiple Human Brain Data Analysis, *Computational Intelligence*, An International Journal, Blackwell, 21(2), (2005) 177-196.
- n Zhong, N., Motomura, S., Wu, J.L., "Peculiarity Oriented Multi-Aspect Brain Data Analysis for Studying Human Multi-Perception Mechanism", *Proc. SAINT 2005 Workshops* (Workshop 8: Computer Intelligence for Exabyte Scale Data Explosion), IEEE Press, (2005) 306-309.
- n Zhong, N., Wu, J.L., Nakamaru, A., Ohshima, M., Mizuhara, H. "Peculiarity Oriented fMRI Brain Data Analysis for Studying Human Multi-Perception Mechanism", *Cognitive Systems Research*, An International Journal, Elsevier (2004) 241-256.
- n Zhong, N., Liu, C., Yao, Y.Y., Ohshima, M., Huang, M., Huang, J.J. "Relational Peculiarity Oriented Mining", *Proc. 4th IEEE International Conference on Data Mining (ICDM'04)*, IEEE Press (2004) 575-578.
- n Ohshima, M., Zhong, N., Yao, Y.Y., Murata. S. "Peculiarity Oriented Analysis in Multi-people Tracking Images", *Proc. PAKDD'04*, LNAI 3056, Springer (2004) 508-518.



Major Publications on POM (2)

- n Zhong, N., Yao, Y.Y., Ohshima, M. "Peculiarity Oriented Multi-Database Mining", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 4 (2003) 952-960.
- n Zhong, N., Nakamaru, A., Ohshima, M., Wu, J.L., Mizuhara, H. "Peculiarity Oriented Mining in Multiple Human Brain Data", *Proc. IDEAL'03*, LNCS 2690, Springer (2003) 742-750.
- n Zhong, N. "Mining Interesting Patterns in Multiple Data Sources", Vicenc Torra (ed.) *Information Fusion in Data Mining*, in the Studies in Fuzziness and Soft Computing Series, Vol. 123, Springer (2003) 61-77.
- n Zhong, N., Yao, Y.Y., Ohshima, M., and Ohsuga, S. "Interestingness, Peculiarity, and Multi-Database Mining", *Proc. 2001 IEEE International Conference on Data Mining (ICDM'01)*, IEEE Press (2001) 566-573.
- n Zhong, N., Ohshima, M., Ohsuga, S. "Peculiarity Oriented Mining and Its Applications for Knowledge Discovery in Amino-acid Data", *Proc. PAKDD'01*, LNAI 2035, Springer (2001) 260-269.
- n Zhong, N., Yao, Y.Y., Ohsuga, S. "Peculiarity Oriented Data Mining", *Proc. PKDD'99*, LNAI 1704, Springer (1999) 136-146.



Thank You !