Peculiarity Oriented Multidatabase Mining

Ning Zhong, *Member*, *IEEE*, Yiyu (Y.Y.) Yao, *Member*, *IEEE*, and Muneaki Ohshima

Abstract—*Peculiarity rules* are a new class of rules which can be discovered by searching relevance among a relatively small number of peculiar data. Peculiarity oriented mining in multiple data sources is different from, and complementary to, existing approaches for discovering new, surprising, and interesting patterns hidden in data. A theoretical framework for peculiarity oriented mining is presented. Within the proposed framework, we give a formal interpretation and comparison of three classes of rules, namely, *association rules*, *exception rules*, and *peculiarity rules*, as well as describe how to mine interesting peculiarity rules in multiple databases.

Index Terms—Peculiarity oriented mining, interestingness, multidatabase mining.

1 Introduction

The main idea of this work can be summarized in a phrase: *peculiarity oriented discovery of interesting patterns from multiple databases*. **Peculiarity** represents a new interpretation of interestingness, an important notion long identified in data mining [10], [25], [26]. Peculiarity, unexpected relationships/rules, may be hidden in a relatively small number of data. *Peculiarity rules* are a typical regularity hidden in many scientific, statistical, and transaction databases. They may be difficult to find by applying the standard association rule mining method due to the requirement of large support. In contrast, peculiarity oriented mining focuses on some interesting data (peculiar data) in order to find novel and interesting rules (peculiarity rules). The second keyword is **multiple databases**, which are the objects of discovery and learning. Mainstream KDD (Knowledge Discovery and Data Mining) research is limited to rule discovery in a single universal relation or an information table [1], [11]. Multidatabase mining is to mine knowledge in multiple related information sources.

By considering the two related issues of peculiarity and multiple databases, we propose a framework of peculiarity oriented mining in multidatabases. The identification of peculiarity rules, as well as algorithms of mining peculiarity rules, will enhance the effectiveness of data mining and extend its domain of applications.

Studies on peculiarity oriented mining can be divided into three phases:

1. developing methods of peculiarity oriented mining,
2. extending peculiarity oriented approaches to multiple data sources, and
3. enabling peculiarity oriented mining in a distributed and cooperative mode.

In the paper, we investigate the first two phases by concentrating on the theoretical development of a framework for peculiarity oriented mining. Detailed experimental evaluations in several different domains can be found in related papers [20], [26], [27], [28].

The paper is organized as follows: The rest of Section 1 gives the background and motivations for peculiarity oriented mining. Section 2 gives a formal interpretation and comparison of three classes of rules: association rules, exception rules, and peculiarity rules. Section 3 presents a method of peculiarity oriented mining. Section 4 extends the peculiarity oriented mining to multiple databases. Finally, Section 5 gives concluding remarks.

1.1 Interestingness and Peculiarity

The purpose of data mining is to discover interesting knowledge hidden in databases. The evaluation of interestingness, such as peculiarity, surprisingness, unexpectedness, usefulness, and novelty, can be done in preprocessing and/or postprocessing of the knowledge discovery process [5], [6], [10], [25]. Evaluating in preprocessing is to select interesting data before the knowledge discovery process; evaluating in postprocessing is to select interesting rules after the knowledge discovery process. Interestingness evaluation may be either *subjective* or *objective* [10]. Subjective evaluation is user-driven. The user is asked to explicitly specify what types of data (or rules) are interesting and uninteresting and the system then discovers those rules satisfying the user requirements. Objective evaluation is data-driven. The system analyzes structures of data and discovers rules based on certain criteria such as predictive performance, statistical significance, and so forth.

We study a new class of rules called *peculiarity rules* [25]. Roughly speaking, data are *peculiar* if they represent a relatively small number of objects and, furthermore, those objects are very different from other objects in a data set. Peculiarity rules are discovered by searching relevance among peculiar data. A peculiarity rule has a low support value.

- *N. Zhong is with the Department of Information Engineering, Maebashi Institute of Technology, 460-1, Kamisadori-Cho, Maebashi-City 371-0816, Japan. E-mail: zhong@maebashi-it.ac.jp.*
- *Y.Y. Yao is with the Department of Computer Science, University of Regina, Saskatchewan, Canada. E-mail: yyao@cs.uregina.ca.*
- *M. Ohshima is with the Graduate School, Maebashi Institute of Technology, 460-1, Kamisadori-Cho, Maebashi-City 371-0816, Japan. E-mail: ohshima@wi-lab.com.*

Manuscript received 7 Mar. 2002; revised 3 Oct. 2002; accepted 21 Feb. 2003. For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 116050.

Patterns (rules) with low support have been considered by many researchers. Two examples are the studies of exception rules [13] and emerging patterns [4]. They share a common feature with peculiarity rules in the sense that all describe a relatively small number of objects. They differ in the way in which peculiar data are collected, interpreted, and used, as well as the interpretation of corresponding rules. An exception rule is an amendment to, or a clarification of, a more general rule. The peculiar data covered by an exception rule is obtained from the subset of data covered by the general rule. On the other hand, a peculiarity rule stands on its own and has a common-sense interpretation, like ordinary association rules [1], [13]. Dong and Li proposed a framework for discovering emerging patterns [4]. Their method is essentially a study of the change of supports in different data sets. A large change suggests an interesting emerging pattern. Since emerging patterns with large supports are perhaps well-known facts, they concentrated on emerging patterns with small supports. In particular, they suggested that patterns with low support, such as 1 to 20 percent, can give useful new insights about data. Unfortunately, such patterns are difficult to discover by traditional association rule mining methods. Their approach provides another use of peculiar data, which is different from our consideration of peculiarity rules. In general, it may be desirable to have a unified framework within which many different uses of peculiar data can be studied.

A notion related to peculiarity is noise, which is an unavoidable problem in real-world databases. Although noise may appear as peculiar data, one may identify noise based on domain knowledge or metaknowledge of the database. In this paper, we concentrate on syntactically defined peculiar data, which are characterized by attribute values and the distribution of values in a database. The task of differentiating actual peculiar data and noise is left to domain experts.

The success of any algorithm at identifying peculiar data depends on the quality of the database, as well as attributes used for such a purpose. For simplicity, we will not consider the problem of feature selection in this paper. It is assumed that some well-known features selection algorithms have been applied to the data set. Our task is thus restricted to the identification of peculiar data in a cleaned and preprocessed data set.

1.2 Multidatabase Mining

Tasks of multidatabase mining can be divided roughly into three levels:

1. **Mining from multiple relations in a database**. Although, theoretically, any relational database with multiple relations can be transformed into a single universal relation, practically this can lead to many issues such as universal relations of unmanageable sizes, infiltration of uninteresting attributes, loss of useful relation names, unnecessary join operations, and inconveniences for distributed processing.
2. **Mining from multiple relational databases**. Some regularities, relationships, and rules cannot be discovered if we just search a single database simply because useful knowledge often hides in multiple databases [25].
3. **Mining from multiple mixed-media databases**. Many real-world data sets contain more than just a single type of data [14], [25]. How to handle such mixed-media, multiple data sources is a new, challenging research issue.

Multidatabase mining involves many related topics including interestingness and relevance checking, granular computing, and distributed data mining.

Liu et al. proposed a relevance measure for identifying relevant databases as the first step for multidatabase mining [9]. Ribeiro et al. described a way of extending the INLEN system for multidatabase mining by the incorporation of primary and foreign keys, as well as the development and processing of knowledge segments [12]. Wrobel extended the concept of foreign keys into foreign links, as multidatabase mining is also interested in getting to nonkey attributes [16]. Aronis et al. introduced a system called WoRLD that uses spreading activation to enable inductive learning from multiple tables in multiple databases across the network [2]. Dong and Li studied the change of supports of itemsets in different databases to discover emerging patterns [4]. Zhong et al. proposed a way for peculiarity oriented multidatabase mining [25].

A challenge to multidatabase mining is heterogeneity. Different databases may use different terminology and conceptual level to define their schemes. Explicit foreign key relationships may not exist among databases. A key issue is therefore to find/create the relevance among different databases. Granular computing techniques provide a useful tool for this purpose through changing information granularity [8], [18], [21], [25].

Multidatabase mining is particularly relevant to peculiarity oriented knowledge discovery. A peculiarity rule is useful and meaningful if one can provide some explanation or justification. In many cases, the justification of a peculiarity rule cannot be obtained from a single database as the explanation lies in another database [25].

1.3 An Illustrative Example

The basic ideas of peculiarity oriented mining can be illustrated by a simple example. Given a *Supermarket-Sales* relation in Table 1, one can observe the following peculiarity rule:

$$rule_1 : meat\text{-}sale(low) \wedge vegetable\text{-}sale(low) \wedge fruits\text{-}sale(low) \rightarrow turnover(very\text{-}low).$$

This rule covers data in one tuple on 30 July and it can be easily interpreted by common-sense. Algorithms for mining association rules and exception rules may fail to find such useful rules. On the other hand, a manager of the supermarket may be interested in such a rule because it shows that the turnover was a marked drop.

To discover peculiarity rules, we first search peculiar data in the relation *Supermarket-Sales*. In Table 1, the values of the attributes *meat-sale*, *vegetable-sale*, and *fruits-sale* on 30 July are very different from other values. Those values are regarded as peculiar data. Furthermore, $rule_1$ is generated by searching relevance among peculiar data.

TABLE 1
Supermarket-Sales

Addr.	Date	meat-sale	vegetable-sale	fruits-sale	...	turnover
Ube	July-1	400	300	450	...	2000
...	July-2	420	290	460	...	2200
...	...	...	...	...	...	...
...	**July-30**	**12**	**10**	**15**	...	**100**
...	July-31	430	320	470	...	2500
...	...	...	...	...	...	...

We use the qualitative representation for the quantitative values in $rule_1$. The transformation of quantitative to qualitative values is done by using the following background knowledge about information granularity:

Basic granules:
$bg_1 = \{high, low, very\text{-}low\}$;
$bg_2 = \{large, small, very\text{-}small\}$;
$bg_3 = \{many, little, very\text{-}little\}$;
.....

Specific granules:
$kanto\text{-}area = \{Tokyo, Tiba, Saitama, \ldots\}$;
$chugoku\text{-}area = \{Yamaguchi, Hiroshima, Shimane, \ldots\}$;
$yamaguchi\text{-}prefecture = \{Ube, Shimonoseki, \ldots\}$;
.....

Through granulation, quantitative conditions

$$meat\text{-}sale = 12, vegetable\text{-}sale = 10, fruits\text{-}sale = 15,$$

and $turnover = 100$ are replaced by the granules, "*low*" and "*very-low*," respectively.

2 Interpretation of Rules

This section discusses a formal interpretation and comparison of three classes of rules: *association rules*, *exception rules*, and *peculiarity rules*.

2.1 A Framework for the Interpretation of Rules

Typically, a rule can be expressed in the form $\phi \Rightarrow \psi$, where ϕ and ψ are formulas of certain language used to describe objects (tuples) in the database. We adopt the decision logic language (DL-language) studied by Pawlak [11], in Tarski's style through the notions of a model and satisfiability. The model is a database S consisting of a finite set of objects U. An object $x \in U$ either satisfies a formula ϕ, written $x \models_S \phi$ or, in short, $x \models \phi$, or does not satisfy the formula, written $\neg x \models \phi$. The satisfiability depends on the semantic interpretation of expressions and must be defined by a particular rule mining method. In general, it should satisfy the following conditions [11]:

1. $x \models \neg\phi$ iff not $x \models \phi$,
2. $x \models \phi \wedge \psi$ iff $x \models \phi$ and $x \models \psi$,
3. $x \models \phi \vee \psi$ iff $x \models \phi$ or $x \models \psi$,
4. $x \models \phi \rightarrow \psi$ iff $x \models \neg\phi \vee \psi$, and
5. $x \models \phi \equiv \psi$ iff $x \models \phi \rightarrow \psi$ and $x \models \psi \rightarrow \phi$,

where $\neg$, $\wedge$, $\vee$, $\rightarrow$, and $\equiv$ are standard logical connectives. If ϕ is a formula, the set $m_S(\phi)$ defined by:

$$m_S(\phi) = \{x \in U \mid x \models \phi\} \tag{1}$$

is called the meaning of the formula ϕ in S. If S is understood, we simply write $m(\phi)$. Obviously, the following properties hold [11]:

1. $m(\neg\phi) = -m(\phi)$,
2. $m(\phi \wedge \psi) = m(\phi) \cap m(\psi)$,
3. $m(\phi \vee \psi) = m(\phi) \cup m(\psi)$,
4. $m(\phi \rightarrow \psi) = -m(\phi) \cup m(\psi)$,
5. $m(\phi \equiv \psi) = (m(\phi) \cap m(\psi)) \cup (-m(\phi) \cap -m(\psi))$.

The meaning of a formula ϕ is the set of all objects having the property expressed by the formula ϕ. Conversely, ϕ can be viewed as a description of the set of objects $m(\phi)$. Thus, a connection between formulas and subsets of U is established.

A formula ϕ is said to be true in a database S, written $\models_S \phi$, if and only if $m(\phi) = U$, namely, ϕ is satisfied by all objects in the universe. Two formulas ϕ and ψ are equivalent in S if and only if $m(\phi) = m(\psi)$. By definition, the following properties hold [11]:

1. $\models_S \phi$ iff $m(\phi) = U$,
2. $\models_S \neg\phi$ iff $m(\phi) = \emptyset$,
3. $\models_S \phi \rightarrow \psi$ iff $m(\phi) \subseteq m(\psi)$, and
4. $\models_S \phi \equiv \psi$ iff $m(\phi) = m(\psi)$.

Thus, we can study the relationships between concepts described by formulas based on the relationships between their corresponding sets of objects.

A rule $\phi \Rightarrow \psi$ can be interpreted by logical implication, namely, the symbol $\Rightarrow$ is interpreted as the logical implication $\rightarrow$. In most cases, the expression $\phi \rightarrow \psi$ may not be true in a database. Only certain objects satisfy the expression $\phi \rightarrow \psi$. The ratio of objects satisfying $\phi \rightarrow \psi$ can be used to define a quantitative measure of the strength of the rule:

$$T(\phi \Rightarrow \psi) = \frac{|m(\phi \rightarrow \psi)|}{|U|}, \tag{2}$$

where $|\cdot|$ denotes the cardinality of a set. It measures the degree of truth of the expression $\phi \rightarrow \psi$ in a database. A problem with the logical implication interpretation can be seen as follows: For an object, if it does not satisfy ϕ, by definition, it satisfies $\phi \rightarrow \psi$. Thus, even if the degree of truth of $\phi \rightarrow \psi$ is very high, we may not conclude too much on the satisfiability of ψ given the object satisfies ϕ. In reality, we want to know the satisfiability of ψ under the

TABLE 2
Contingency Table

	ψ	$\neg\psi$	Totals
ϕ	a	b	$a+b$
$\neg\phi$	c	d	$c+d$
Totals	$a+c$	$b+d$	$a+b+c+d=n$

condition that ϕ is satisfied. In other words, our main concern is the satisfiability of ψ in the subset $m(\phi)$. Obviously, logical implication is inappropriate in this case.

2.2 Probabilistic Interpretations of Rules

In data mining, rules are typically interpreted in terms of probability. A detailed analysis of probability related measures associated with rules has been given by Yao and Zhong [19]. The characteristics of a rule $\phi \Rightarrow \psi$ can be summarized by Table 2. The contingency table has been used by many researchers in data mining [15], [19], [22]. From the contingency table, different measures can be defined to reflect various aspects of rules.

The *generality* of ϕ is defined by:

$$G(\phi) = \frac{|m(\phi)|}{|U|} = \frac{a+b}{n}, \tag{3}$$

which indicates the relative size of the concept ϕ. A concept is more general if it covers more instances of the universe. If $G(\phi) = \alpha$, then (100α) percent of objects in U satisfy ϕ. The quantity may be viewed as the probability of a randomly selected object satisfying ϕ. Obviously, we have $0 \leq G(\phi) \leq 1$.

The *absolute support* of ψ provided by ϕ is the quantity:

$$AS(\phi \Rightarrow \psi) = AS(\psi|\phi) = \frac{|m(\psi) \cap m(\phi)|}{|m(\phi)|} = \frac{a}{a+b}. \tag{4}$$

The quantity, $0 \leq AS(\psi|\phi) \leq 1$, shows the degree to which ϕ implies ψ. If $AS(\psi|\phi) = \alpha$, then (100α) percent of objects satisfying ϕ also satisfy ψ. It may be viewed as the conditional probability of a randomly selected object satisfying ψ given that the object satisfies ϕ. The *change of support* of ψ provided by ϕ is defined by

$$CS(\phi \Rightarrow \psi) = CS(\psi|\phi) = AS(\psi|\phi) - G(\psi) = \frac{a}{a+b} - \frac{a+c}{n}. \tag{5}$$

The change of support varies from -1 to 1. One may consider $G(\psi)$ to be the prior probability of ψ and $AS(\psi|\phi)$ the posterior probability of ψ after knowing ϕ. The difference of posterior and prior probabilities represents the change of our confidence regarding whether ϕ actually relates to ψ. For a positive value, one may say that ϕ is positively related to ψ; for a negative value, one may say that ϕ is negatively related to ψ.

The generality $G(\psi)$ is related to the satisfiability of ψ by all objects in the database and $AS(\phi \Rightarrow \psi)$ is related to the satisfiability of ψ in the subset $m(\phi)$. A high $AS(\phi \Rightarrow \psi)$ does not necessarily suggest a strong association between ϕ and ψ as a concept ψ with a large $G(\psi)$ value tends to have a large $AS(\phi \Rightarrow \psi)$ value. The change of support $CS(\phi \Rightarrow \psi)$ may be more accurate.

2.3 Comparison of Association Rules, Exception Rules, and Peculiarity Rules

Within the proposed framework, we can easily analyze the ordinary association rules by a slightly different formulation. Let I denote a set of items and T denote a set of transactions. For each item $i \in I$, we define an atomic expression $F_{\{i\}} = (i = 1)$ with the satisfiability given by $t \in T$,

$$t \models F_{\{i\}} \text{ iff } t \text{ contains } i, \tag{6}$$

and

$$m(F_{\{i\}}) = \{t \in T \mid t \text{ contains } i\}. \tag{7}$$

For each subset $A \subseteq I$, we define a formula $F_A = \bigwedge_{i \in A} F_{\{i\}}$. A transaction satisfies the formula F_A if it contains *all* items in A. For two disjoint subsets of items A and B, an association rule can be expressed as $F_A \Rightarrow F_B$. It is interpreted as saying that a customer who purchases *all* items in A tends to purchase *all* items in B.

Two measures, called the support and the confidence, are used to mine association rules. They are indeed the generality and absolute support:

$$\begin{aligned} &supp(F_A \Rightarrow F_B) = G(F_A \wedge F_B) = G(F_{A \cup B}), \\ &conf(F_A \Rightarrow F_B) = AS(F_A \Rightarrow F_B). \end{aligned} \tag{8}$$

By specifying threshold values of support and confidence, one can obtain all association rules whose support and confidence are above the thresholds. Association rules can be extended to nontransaction databases so that both the left-hand and right-hand sides are formulas expressing properties of objects in a database.

With an association rule, it is very tempting to relate a large confidence with a strong association between two concepts. However, such a connection may not exist. Suppose we have $conf(\phi \Rightarrow \psi) = 0.90$. If we also have $G(\psi) = 0.95$, we can conclude that ϕ is in fact negatively associated with ψ. This suggests that an association rule may not reflect the true association. An association rule with low confidence may have a relatively large change of support. In mining association rules, concepts with low support are not considered in the search for association. On the other hand, two concepts with low supports may have either large confidence or a large change of support. In summary, algorithms for mining association rules may fail to find such useful rules. Other mining algorithms are needed.

Exception rules have been studied as an extension of association rules to resolve some of the above problems [13]. For an association rule $\phi \Rightarrow \psi$ with high confidence, one may associates an exception rule $\phi \wedge \phi' \Rightarrow \neg\psi$. Roughly speaking, ϕ' can be viewed as the condition for exception to rule $\phi \Rightarrow \psi$. To be consistent with the intended interpretation of the exception rule, it is reasonable to assume that $\phi \wedge \phi' \Rightarrow \neg\psi$ have a high confidence and low support. More

TABLE 3
Qualitative Characterization of Association Rules, Exception Rules, and Peculiarity Rules

Rule	G (supp)	AS (conf)	CS	*semantic*
Association rule:				
$\phi \Rightarrow \psi$	High	High	Unknown	*common-sense*
Exception rule:				
$\phi \Rightarrow \psi$	High	High	Unknown	
$\phi \wedge \phi' \Rightarrow \neg\psi$	Low	High	High	*exception*
Peculiarity rule:				
$\phi \Rightarrow \psi$	Low	High	High	*common-sense*

specifically, we would expect a low generality of $\phi \wedge \phi'$. Otherwise, the rule cannot be viewed as describing exceptional situations. Consequently, exception rules cannot be discovered by association rule mining algorithms.

Recently, we identified and studied a new class of rules called *peculiarity rules* [25], [26], [27]. In mining peculiarity rules, one considers the distribution of attribute values. More specifically, attention is paid to objects whose attribute values are quite different from that of other objects. This is referred to as peculiar data identification. After the isolation of peculiar data, peculiarity rules with low support and high confidence and high change of support are searched for. Although a peculiarity rule may share the same properties with an exception rule, as expressed in terms of support and confidence, it does not express exception to another rule. Semantically, they are very different. Algorithms for mining peculiarity rules are different from mining association rules and exception rules. It should be realized that peculiarity rules only represent a subset of all rules with high change of support.

Based on the above discussion, we can qualitatively characterize association rules, exception rules, and peculiarity rules as shown in Table 3. From the viewpoint of support, both exception rules and peculiarity rules attempt to find rules that are missed by association rule mining methods. While exception rules and peculiarity rules have a high change of support values, indicating a strong association between two concepts, association rules do not necessarily have this property. All three classes of rules are focused on rules with a high level of absolute support. For exception rules, it is also expected that the generality of $\phi \wedge \phi'$ is low. For peculiarity, the generalities of both ϕ and ψ are expected to be low. In contrast, the generality of the right hand of an exception rule does not have to be low.

It may be argued that rules with high absolute support and high change of support are of interest. The use of generality (support) in association rule mining is mainly for the sake of computational cost, rather than semantics consideration. Exception rules and peculiarity rules are two subsets of rules with high absolute support and high change of support. It may be interesting to design an algorithm to find *all* rules with high absolute support and high change of support.

3 Peculiarity Oriented Mining

Peculiarity rules are discovered from peculiar data evaluated using unified knowledge-based statistical criteria. The main task of mining peculiarity rules is the identification of peculiar data. Peculiar data are a subset of objects in the database and are characterized by two features: 1) very different from other objects in a data set and 2) consisting of a relatively small number of objects [25], [26].

There are many ways of finding peculiar data. We describe an attribute-oriented method which analyzes data from a new view and is different from traditional statistical methods.

3.1 Finding Peculiar Data

Table 4 shows a relation with attributes $A_1, A_2, \ldots, A_m$. Let x_{ij} be the value of A_j of the ith tuple and n the number of tuples. The peculiarity of x_{ij} can be evaluated by the *Peculiarity Factor*, $PF(x_{ij})$,

$$PF(x_{ij}) = \sum_{k=1}^{n} N(x_{ij}, x_{kj})^{\alpha}, \tag{9}$$

TABLE 4
A Sample Table (Relation)

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

TABLE 5
An Example of Peculiarity Factors for a Continuous Attribute

Region	ArableLand		PF
Hokkaido	**1209**		*134.1*
Tokyo	12	$\Longrightarrow$	60.9
Osaka	18		60.3
Yamaguchi	162		60.5
Okinawa	147		59.4

where N denotes the conceptual distance, α is a parameter which can be adjusted by a user, and $\alpha = 0.5$ is used as default. Equation (9) evaluates whether x_{ij} has a low frequency and is very different from other values x_{kj}.

There are several advantages to the proposed method. One can handle both continuous and symbolic attributes based on a unified semantic interpretation. Background knowledge represented by binary neighborhoods can be used to evaluate the peculiarity if such background knowledge is provided by a user. If X is a continuous attribute and no background knowledge is available, in (9) we use the distance,

$$N(x_{ij}, x_{kj}) = |x_{ij} - x_{kj}|. \tag{10}$$

Table 5 shows peculiarity factor values for the attribute **ArableLand**. If X is a symbolic attribute and the background knowledge for representing the conceptual distances between x_{ij} and x_{kj} is provided by a user, the peculiarity factor is calculated by the conceptual distances, $N(x_{ij}, x_{kj})$ [8], [18], [25], [26]. The conceptual distances are assigned to one if no background knowledge is available.

There are two major methods for testing if peculiar data exist or not (it is called *selection of peculiar data*) after calculating peculiarity factors. The first is based on a threshold value:

$$\begin{aligned} threshold = {} & mean\ of\ PF(x_{ij}) + \\ & \beta \times\ standard\ deviation\ of\ PF(x_{ij}), \end{aligned} \tag{11}$$

where β can be adjusted by a user and $\beta = 1$ is used as default. The threshold indicates that a data is a peculiar one if its PF value is much larger than the mean of the PF set. In other words, if $PF(x_{ij})$ is over the threshold value, x_{ij} is a peculiar data. By adjusting the parameter β, a user can define suitable threshold value. The other method for *selection of peculiar data* uses the chi-square test when the data size is sufficiently large [3].

3.2 Attribute Oriented Clustering

Searching the data for a structure of natural clusters is an important exploratory technique. Clusters can provide an informal means of assessing interesting and meaningful groups of peculiar data.

In the real world, there are many real-valued attributes and symbolic-valued attributes. In order to discover interesting knowledge, conceptual abstraction and generalization are necessary. *Attribute oriented clustering* is a useful technique to quantize continuous values and eventually perform conceptual abstraction [23]. It serves as an important step of the peculiarity oriented mining process.

A key issue in clustering is the incorporation of background knowledge about information granularity. Our approach is to provide various methods in the mining process so that data with different features can be handled effectively. If background knowledge about information granularity (e.g., domain-specific ontologies) is available, it is used for semantic-based attribute-oriented clustering. Otherwise, the *nearest-neighbor method* is used for clustering of continuous-valued attributes [3], [7].

3.3 An Algorithm

An algorithm for finding peculiar data is outlined as follows:

Step 1. Execute attribute oriented clustering for each attribute.

Step 2. For attributes 1 to m do

- *Step 2.1*. Calculate the peculiarity factor $PF(x_{ij})$ in (9) for all values of an attribute.
- *Step 2.2*. Calculate the threshold value in (11) based on the peculiarity factor obtained in *Step 2.1*.
- *Step 2.3*. Select the data that are over the threshold value as peculiar data.
- *Step 2.4*. If the current peculiarity level is enough, then go to *Step 3*.
- *Step 2.5*. Remove peculiar data from the attribute and, thus, we get a new data set. Then, go back to *Step 2.1*.

Step 3. Change the granularity of peculiar data by using background knowledge on information granularity if the background knowledge is available.

The algorithm can be done in a parallel-distributed mode for multiple attributes, relations, and databases because this is an attribute-oriented finding method.

3.4 Relevance among Peculiar Data

A peculiarity rule is discovered by searching the relevance among peculiar data. Let $X(x)$ and $Y(y)$ be peculiar data found in two attributes X and Y, respectively. We deal with the following two cases:

- If both $X(x)$ and $Y(y)$ are symbolic data, the relevance between $X(x)$ and $Y(y)$ is evaluated by:

$$R_1 = P(X(x)|Y(y))P(Y(y)|X(x)), \tag{12}$$

  that is, the larger the product of the probabilities, the stronger the relevance between $X(x)$ and $Y(y)$ is.
- If both $X(x)$ and $Y(y)$ are continuous attributes, the relevance between $X(x)$ and $Y(y)$ is evaluated by using the method developed in the KOSI system that finds functional relationships [24].

Equation (12) is suitable for handling more than two peculiar data found in more than two attributes if $X(x)$ (or $Y(y)$) is a granule of peculiar data.

TABLE 6
Weather

Region	Date	...	Weather
Yamaguchi	July-1	...	sunny
...	July-2	...	cloud
...	...	...	...
...	**July-30**	...	**typhoon (no. 2)**
...	July-31	...	cloud
...	...	...	...

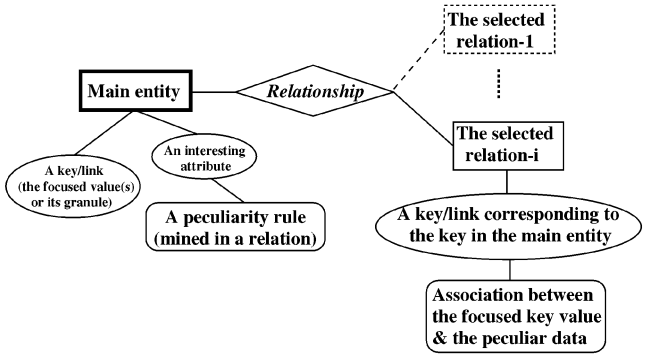

Fig. 1. The graphic description of the RVER model.

4 Peculiarity Oriented Mining in Multiple Databases

This section extends the peculiarity oriented approach for mining multiple data sources (i.e., multidatabase mining).

4.1 Mining in Multiple Databases

Generally speaking, the tasks of multidatabase mining for the first two levels stated in Section 1.2 can be described as follows:

The concept of a foreign key in the relational databases needs to be extended into a foreign link because we are also interested in getting to nonkey attributes for data mining from multiple relations in a database [16]. A major work is to find peculiar data in multiple relations for a given discovery task when foreign link relationships exist. In other words, our task is to select n relations, which contain peculiar data, among m relations $(m \geq n)$ with foreign links.

The method for selecting n relations among m relations can be divided into the following steps:

Step 1. Focus on a relation as the *main table* and find peculiar data from this table. Then, elicit peculiarity rules from peculiar data by using the methods stated in Section 3.

Step 2. Find the value(s) of the focused key corresponding to the mined peculiarity rule (or peculiar data) in *Step 1* and change its granularity of the value(s) of the focused key if the background knowledge on information granularity is available.

Step 3. Find peculiar data in the other relations (or databases) corresponding to the value (or its granule) of the focused key.

Step 4. Select n relations that contain peculiar data, among m relations $(m \geq n)$. In other words, we just select the relations that contain peculiar data relevant to the peculiarity rule mined from the main table.

A peculiarity rule can be discovered from peculiar data hidden in multiple relations by searching relevance among peculiar data. If peculiar data, $X(x)$ 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 by:

$$R_2 = P_1(K(k)|X(x))P_2(K(k)|Y(y)). \quad (13)$$

The above-stated methodology can be extended for mining from multiple databases. A challenge in multidatabase mining is a semantic heterogeneity. Usually, no explicit foreign key/link relationships exist among different databases. The key issue of the extension is to find/create the relevance among databases. We use *granular computing* techniques, such as approximation and abstraction, for solving the issue [8], [21].

Consider again the illustrative example of Section 1.3. After finding that turnover experienced a marked drop in one day from a supermarket-sale database, the manager of the supermarket needs to know the reason for such a peculiar case. The rule itself does not provide an answer. On the other hand, if we search several related data sources, such as a weather database as shown in Table 6, we can find that there was a violent typhoon that day. This explains why the turnover was a marked drop. For this case, the granule of $addr. = Ube$ in Table 1 needs to be changed into $region = yamaguchi$ for creating an explicit foreign link between the supermarket-sales database and the weather database.

4.2 Representation and Relearning

We use the RVER (Reverse Variant Entity-Relationship) model to represent peculiar data and the conceptual relationships among peculiar data discovered from multiple relations (databases) [25]. Fig. 1 shows a graphic description of the RVER model. In this figure, the "main entity" is the main table/database specified by a user and the "selected relation" is the table/database with peculiar data corresponding to the mined peculiarity rule (or peculiar data) in the main table/database.

Fig. 2 shows the framework of peculiarity oriented mining in multiple data sources as well as a result mined from two databases on supermarket sales at Yamaguchi prefecture and the weather of Japan. First, 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. If the data in the main table/database is not sufficient for finding interesting rules, search the peculiar data in the related databases. More interesting peculiarity rules can be discovered from peculiar data hidden in multiple relations/databases by searching the relevance among the peculiar data.

The RVER model is different from an ordinary ER model in that we just represent the attributes relevant to peculiar data (or their granules) in the RVER model. The RVER

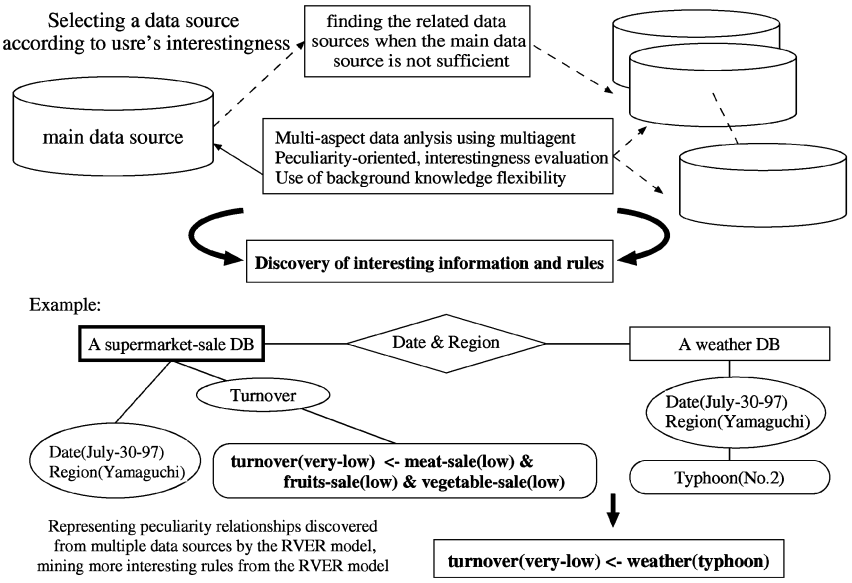

Fig. 2. The framework of peculiarity oriented mining in multiple data sources as well as a result mined from two databases on supermarket-sales and weather.

model provides all interesting information that is relevant to learn more useful rules among multiple relations (databases) by focusing on certain attributes. For example, in the supermarket-sale database, we only consider the conditions $turnover = very\text{-}low$, $region = yamaguchi$, and $date = July\text{-}30$.

In the RVER model, one can discover useful rules through a *relearning* process. For example, the following rule can be learned from the RVER model shown in Fig. 2:

$$rule_2 : weather(typhoon) \rightarrow turnover(very\text{-}low).$$

We can see that a manager of the supermarket may be more interested in $rule_2$ (rather than $rule_1$) because $rule_2$ shows the reason why the turnover was a marked drop.

5 Concluding Remarks

A method of mining peculiarity rules from multiple data sources was presented. It is an effective technique of information fusion in data mining to improve the performance of mining results. This paper showed that peculiarity rules represent a typically unexpected, interesting regularity hidden in databases.

The subjective interestingness of association rules has been systematically investigated by many researchers. For example, Liu et al. studied subjective evaluation of interestingness in postprocessing, i.e., evaluating the mined rules [10]. In contrast, our work is about objective evaluation of interestingness in preprocessing, i.e., selecting interesting (peculiar) data before rule generation. Our approach can mine a new class of patterns, called *peculiarity rules*, in *multiple* data sources.

With respect to two levels of the multidatabase mining tasks, i.e., mining from multiple relations in a single database and mining from multiple relational databases, we have used many databases, such as Japan-survey, amino-acid data, weather, supermarket, and hepatitis, to test our approach [20], [26], [28]. The results are very encouraging and clearly show the usefulness and effectiveness of the proposed approach. Currently, we are also working on the third level of the multidatabase mining task, that is, mining from multiple mixed-media databases, such as fMRI brain data [17], [28], and tracking multiple people in image sequences.

In many real-world problems, it may be more effective to combine peculiarity oriented mining with other approaches such as ordinary association rule mining, classification rule mining for multiaspect analysis. Our future work includes developing a systematic method to mine the rules from multiple data sources where there are no explicitly foreign key (link) relationships and to induce interesting rules from the RVER model discovered from multiple data sources, as well as extending our system for multiaspect analysis.

Acknowledgments

The authors are grateful for the constructive comments and suggestions from referees of this paper. This work was partially supported by the grant-in-aid for scientific research on priority area "Active Mining" from the Japanese Ministry of Education, Culture, Sports, Science, and Technology.

References

[1] R. Agrawal et al., "Fast Discovery of Association Rules," *Advances in Knowledge Discovery and Data Mining*, pp. 307-328, 1996.
[2] J.M. Aronis et al., "The WoRLD: Knowledge Discovery from Multiple Distributed Databases," *Proc. 10th Ann. Conf. Florida AI Research Society (FLAIRS '97)*, pp. 337-341, 1997.
[3] G.K. Bhattacharyya and R.A. Johnson, *Statistical Concepts and Methods*. John Wiley & Sons, 1977.
[4] G. Dong and J. Li, "Efficient Mining of Emerging Patterns: Discovering Trends and Differences," *Proc. Fifth Int'l Conf. Knowledge Discovery in Databases (KDD '99)*, pp. 43-52, 1999.
[5] A.A. Freitas, "On Objective Measures of Rule Surprisingness" *Proc. Second European Symp. Principles of Data Mining and Knowledge Discovery (PKDD '98)*, pp. 1-9, 1998.

[6] R.J. Hilderman and H.J. Hamilton, "Evaluation of Interestingness Measures for Ranking Discovered Knowledge," *Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD '01),* pp. 247-259, 2001.
[7] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis.* Prentice Hall, 1998.
[8] T.Y. Lin, "Granular Computing on Binary Relations 1: Data Mining and Neighborhood Systems," *Rough Sets in Knowledge Discovery,* L. Polkowski and A. Skowron, eds., vol. 1, pp. 107-121, Physica-Verlag, 1998.
[9] H. Liu, H. Lu, and J. Yao, "Identifying Relevant Databases for Multidatabase Mining," *Proc. Second Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD '98),* pp. 210-221, 1998.
[10] B. Liu, W. Hsu, S. Chen, and Y. Ma, "Analyzing the Subjective Interestingness of Association Rules," *IEEE Intelligent Systems,* vol. 15, no. 5, pp. 47-55, Sept./Oct. 2000.
[11] Z. Pawlak, *Rough Sets, Theoretical Aspects of Reasoning about Data.* Kluwer, 1991.
[12] J.S. Ribeiro, K.A. Kaufman, and L. Kerschberg, "Knowledge Discovery from Multiple Databases," *Proc. First Int'l Conf. Knowledge Discovery and Data Mining (KDD '95),* pp. 240-245, 1995.
[13] E. Suzuki, "Autonomous Discovery of Reliable Exception Rules," *Proc. Third Int'l Conf. Knowledge Discovery and Data Mining (KDD '97),* pp. 259-262, 1997.
[14] S. Thrun et al., "Automated Learning and Discovery," *AI Magazine,* pp. 78-82, Fall 1999.
[15] S. Tsumoto, "Statistical Test for Rough Set Approximation Based on Fisher's Exact Test," *Proc. Conf. Rough Sets and Current Trends in Computing (RSCTC '02),* pp. 381-388, 2002.
[16] S. Wrobel, "An Algorithm for Multi-Relational Discovery of Subgroups," *Proc. First European Symp. Principles of Data Mining and Knowledge Discovery (PKDD '97),* pp. 367-375, 1997.
[17] J. Wu and N. Zhong, "An Investigation on Human Multi-Perception Mechanism by Cooperatively Using Psychometrics and Data Mining Techniques," *Proc. Fifth World Multi-Conf. Systemics, Cybernetics, and Informatics (SCI '01),* vol. X, pp. 285-290, 2001.
[18] Y.Y. Yao, "Granular Computing using Neighborhood Systems," *Advances in Soft Computing: Eng. Design and Manufacturing,* R. Roy, T. Furuhashi, and P.K. Chawdhry, eds., pp. 539-553, Springer, 1999.
[19] Y.Y. Yao and N. Zhong, "An Analysis of Quantitative Measures Associated with Rules," *Proc. Third Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD '99),* pp. 479-488, 1999.
[20] H. Yokoi, S. Hirano, K. Takabayashi, S. Tsumoto, and Y. Satomura, "Active Mining in Medicine: A Chronic Hepatitis Case—Towards Knowledge Discovery in Hospital Information Systems," *J. Japanese Soc. Artificial Intelligence,* vol. 17, no. 5, pp. 622-628, 2002.
[21] L.A. Zadeh, "Toward a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic," *Fuzzy Sets and Systems,* vol. 90, pp. 111-127, 1997.
[22] R. Zembowicz and J.M. Zytkow, "From Contingency Tables to Various Forms of Knowledge in Databases," *Advances in Knowledge Discovery and Data Mining,* U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., pp. 329-349, AAAI/MIT Press, 1996.
[23] N. Zhong and S. Ohsuga, "Discovering Concept Clusters by Decomposing Databases," *Data Knowledge Eng.,* vol. 12, no. 2, pp. 223-244, 1994.
[24] N. Zhong and S. Ohsuga, "KOSI—An Integrated System for Discovering Functional Relations from Databases," *J. Intelligent Information Systems,* vol. 5, no. 1, pp. 25-50, 1995.
[25] N. Zhong, Y.Y. Yao, and S. Ohsuga, "Peculiarity Oriented Multi-Database Mining," *Proc. Third European Symp. Principles of Data Mining and Knowledge Discovery (PKDD '99),* pp. 136-146, 1999.
[26] N. Zhong, M. Ohshima, and S. Ohsuga, "Peculiarity Oriented Mining and Its Application for Knowledge Discovery in Amino-acid Data," *Proc. Fifth Pacific- Asia Conf. Knowledge Discovery and Data Mining (PAKDD '01),* pp. 260-269, 2001.
[27] N. Zhong, Y.Y. Yao, M. Ohshima, and S. Ohsuga, "Interestingness, Peculiarity, and Multi-Database Mining," *Proc. IEEE Int'l Conf. Data Mining (ICDM '01),* pp. 566-573 2001.
[28] N. Zhong, A. Nakamaru, M. Ohshima, J.L. Wu, and H. Mizuhara, "Peculiarity Oriented Mining in Multiple Human Brain Data," *Proc. 2003 Int'l Conf. Intelligent Data Eng. and Automated Learning (IDEAL '03),* 2003.

Ning Zhong received the PhD degree in the interdisciplinary course on advanced science and technology from the University of Tokyo. He is currently head of the Knowledge Information Systems Laboratory, and a professor in the Department of Systems and Information Engineering at Maebashi Institute of Technology, Japan. He is also CEO of Web Intelligence Laboratory, Inc., a new type of venture intelligent IT business company, as well as a guest professor at the Beijing University of Technology. He has conducted research in the areas of knowledge discovery and data mining, rough sets and granular-soft computing, Web intelligence, intelligent agents, and knowledge information systems, with more than 100 journal and conference publications and seven books. He is the editor-in-chief of the *Web Intelligence and Agent Systems* journal (IOS Press), regional editor of the *Knowledge and Information Systems* journal (Springer), editor-in-chief of the *Annual Review of Intelligent Informatics* (World Scientific), editor (the area of intelligent systems) of the *Encyclopedia of Computer Science and Engineering* (Wiley), and a member of the editorial board of *Advanced Information and Knowledge Processing* (AI&KP) book series (Springer). He is the cochair of the Web Intelligence Consortium (WIC), vice chair of the IEEE Computer Society Technical Committee on Computational Intelligence, a member of the steering committee of the IEEE International Conferences on Data Mining (ICDM), and a member the advisory board of the International Rough Set Society. He has served or is currently serving on the program committees of more than 60 international conferences and workshops, including IEEE ICDM '02 (conference chair), IEEE/WIC WI '03 (conference chair), IEEE/WIC IAT '03 (conference chair), and IJCAI '03 (advisory committee member). He is a member of the IEEE, the ACM, and the AAAI.

Yiyu (Y.Y.) Yao received the PhD degree in computer science from the University of Regina, Canada. He is currently a professor of computer science in the Department of Computer Science, University of Regina, Canada. His research interests are information retrieval, Web intelligence, data mining, fuzzy sets, rough sets, and granular computing. He has published more than 120 journal and conference papers. He is a member of the editorial boards of the *Web Intelligence and Agent Systems* journal (IOS Press), technical committee of the Web Intelligence Consortium, the advisory board of the International Rough Set Society, and the advisory board of a Special Interest Group on Granular Computing in Berkeley Initiative in Soft Computing (BISC/SIG-GrC). He has also served and is serving as a program cochair of three international conferences and as a program committee member of more than 20 international conferences. He is a member of the IEEE and the ACM.

Muneaki Ohshima is currently a PhD student in the Department of Systems and Information Engineering at Maebashi Institute of Technology, Japan. His research interests include knowledge discovery and data mining, Web intelligence, intelligent agents, and knowledge information systems.