2007 IEEE International Conference on Granular Computing

# Interpretations of Discovered Knowledge in Multidimensional Databases

Yuefeng Li
*Faculty of Information Technology*
*Queensland University of Technology, Brisbane, QLD 4001, Australia*
*E-mail: y2.li@qut.edu.au*

## Abstract

*It is a big challenge to guarantee the quality of discovered knowledge in multidimensional databases because of the huge amount of patterns and noises. The essential issue is to provide efficient methods for interpreting meaningful discovered knowledge in databases. This research presents a new technique called granule mining to improve the performance of data mining. Rather than using patterns, it uses granules in different tiers to generalize knowledge in databases. It also provides a mechanism to formally discuss meaningless discovered rules based on relationships between granules in different tiers.*

## 1. Introduction

Traditionally, knowledge engineers spend much time in the acquisition of knowledge from domain experts. This time consuming activity is often referred to the "*knowledge acquisition bottleneck*" [3]. Although there is still a long journey to travel towards solving the bottleneck problem, the capabilities of data mining have been recognized for solving this problem. Currently there are two main research directions in data mining. One direction is concerned with the design *of efficient algorithms* for discovering knowledge [2], and another is concerned with *the interpretations* of discovered knowledge [7] [15] [22].

Several approaches have been used to interpret discovered knowledge: Closed Patterns [16] [17] [18], Constraint-based Association Rules [4] [13], Non-redundant Rules [21] [19] and Decision Tables [1] [7] [11]. Usually not all strong association rules are interesting to users [2] [10], and that becomes particularly difficult in cases of multidimensional databases.

The above approaches have significant performance advantages in decreasing the number of association rules for transaction databases (also called single dimensional databases). However, they are not very efficient for representation of associations in very large multidimensional databases because multidimensional rule mining has to be transformed into single dimensional rule mining when we use these approaches.

We also encountered other difficulties when we tried to use the existing association mining techniques for solving some real problems [8] [9] [18]. It is not feasible to utilize and maintain the discovered knowledge using the traditional knowledge engineering techniques because of the combinatorial explosion in the number of discovered patterns, the existence of noisy discovered patterns and having useful patterns which contain uncertainties. These difficulties motivate us to search for an appropriate interpretation method for discovered knowledge.

Normally, expert knowledge is uncertain (and/or imprecise) and captures generalization; therefore, discussing discovered knowledge at a higher level would be significant for interpreting their meaning. In this research we discuss the concept of granule mining that was recently presented in [6] for interpretations of discovered knowledge in databases. Informally a granule is a clump of expected elements that are drawn together by indistinguishableness, similarity or proximity [20]. Granules describe common features of sets of transactions for selected attributes. They contain more semantic information than patterns. Another advantage of using granules in multidimensional databases is that they can explicitly describe selected dimensions.

In this paper, we study a multi-tier structure for granule mining in multidimensional databases in order to interpret discovery knowledge reasonably. In this structure, attributes are split into some tiers and the large multidimensional database is compressed into granules in different tiers. The antecedent and consequent of an association rule are also both granules. In addition, people can discuss association rules and

0-7695-3032-X/07 $25.00 © 2007 IEEE
DOI 10.1109/GrC.2007.92

their general rules (a rule with a shorter antecedent) in the multi-tier structure. In this way, the meaningless association rules will be justified according to the relationships between association rules and their general rules.

## 2. Pattern Mining to Granule Mining

We show the differences between pattern mining (also called association mining) and granule mining by using some straightforward examples. Formally, a transaction database can be described as a simple information table ($\mathcal{T}$, $V^{\mathcal{T}}$) if we do not consider the time dimension, where $\mathcal{T}$ is the set of objects (transactions) in which each record is a sequences of items, and $V^{\mathcal{T}} = \{a_1, a_2, \ldots, a_n\}$ is a set of selected items (or called attributes in decision tables) for all objects in $\mathcal{T}$, where each item can be a tuple (e.g., <*name*, *cost*, *price*> is a product item).

Table I illustrates an information table, where $V^{\mathcal{T}} = \{a_1, a_2, \ldots, a_7\}$, $\mathcal{T} = \{t_1, t_2, \ldots, t_6\}$. It has 10 frequent patterns but only three closed patterns if *min_sup* = 50%. They are $\{a_3, a_4, a_6\}$, $\{a_1, a_2\}$, and $\{a_6\}$. We can also generate association rules from these closed patterns, for example, from closed pattern $\{a_3, a_4, a_6\}$; we have the following three association rules with the longest antecedents and 100% confidence:

$$a_3 \wedge a_6 \rightarrow a_4;\ a_4 \wedge a_6 \rightarrow a_3;\ a_3 \wedge a_4 \rightarrow a_6 \quad (2.1)$$

The information table can also be compressed into granules according to user constraints, where the simplest case is to group products into two categories, for example, high profit products (also called condition contributes) and low profit products (also called decision attributes). Let $a_1$, $a_2$, $a_3$, $a_4$ and $a_5$ be the condition attributes that are used to form antecedents of rules and $a_6$ and $a_7$ be the decision attributes that are used to form consequents of rules. Table II shows a decision table of Table I, where the set of granules is $\{g_1, g_2, g_3, g_4\}$, and *coverset* is the set of objects that are used to produce a granule.

TABLE I
AN INFORMATION TABLE

| Object | Items |
|---|---|
| $t_1$ | $a_1\ a_2$ |
| $t_2$ | $a_3\ a_4\ a_6$ |
| $t_3$ | $a_3\ a_4\ a_5\ a_6$ |
| $t_4$ | $a_3\ a_4\ a_5\ a_6$ |
| $t_5$ | $a_1\ a_2\ a_6\ a_7$ |
| $t_6$ | $a_1\ a_2\ a_6\ a_7$ |

Every granule in the decision table can be mapped into a decision rule, where we treat the presence and absence of items as the same position if we view the decision table as a multidimensional database. Therefore, we can obtain four decision rules from Table II, and the second granule, $g_2$, can be read as the following decision rule:

$$(a1 = 0 \wedge a2 = 0 \wedge a3 = 1 \wedge a4 = 1 \wedge a5 = 0)$$

TABLE II
A DECISION TABLE

| Granule | Products | | | | | | | coverset |
|---|---|---|---|---|---|---|---|---|
| | High profit | | | | | Low profit | | |
| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | |
| $g_1$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | $\{t_1\}$ |
| $g_2$ | 0 | 0 | 1 | 1 | 0 | 1 | 0 | $\{t_2\}$ |
| $g_3$ | 0 | 0 | 1 | 1 | 1 | 1 | 0 | $\{t_3, t_4\}$ |
| $g_4$ | 1 | 1 | 0 | 0 | 0 | 1 | 1 | $\{t_5, t_6\}$ |

$$\rightarrow (a6 = 1 \wedge a7 = 0) \quad (2.2)$$

We can further deploy the decision rules (large granules) into two tiers: *C-granules* (condition granules) and *D-granules* (decision granules). Table III illustrates a 2-tier structure for the decision table in Table II, where both (A) and (B) include three small granules, and the links describes the association between condition granules and decision granules.

Notice that the decision rules as shown in Eq. (2.2) now can be simply described as follows:

$$cg_2 \rightarrow dg_2 \quad (2.3)$$

where the antecedent and consequent are described as some small granules. This rule shows the association between high profit products and low profit products. It includes much more semantic meaning than the rules in Eq. (2.1) that only show associations between items. In addition, only one rule "$a_3 \wedge a_4 \rightarrow a_6$" in Eq. (2.1) is useful based on the above user constraints; however, it is impossible to identify it before the phase of rule generations in pattern mining.

TABLE III
A 2-TIER STRUCTURE

| ***Condition Granule*** | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | ***coverset*** |
|---|---|---|---|---|---|---|
| $cg_1$ | 1 | 1 | 0 | 0 | 0 | $\{t_1, t_5, t_6\}$ |
| $cg_2$ | 0 | 0 | 1 | 1 | 0 | $\{t_2\}$ |
| $cg_3$ | 0 | 0 | 1 | 1 | 1 | $\{t_1, t_4\}$ |

(A) C-GRANULES

| ***Decision Granule*** | $a_6$ | $a_7$ | ***coverset*** |
|---|---|---|---|
| $dg_1$ | 0 | 0 | $\{t_1\}$ |
| $dg_2$ | 1 | 0 | $\{t_2, t_3, t_4\}$ |
| $dg_3$ | 1 | 1 | $\{t_5, t_6\}$ |

(B) D-GRANULES

Based on the discussion in the last section, we can derive the following advantages of using granules:

1. (1) A granule describes the feature of a set of objects, but a pattern is a part of an object;
2. (2) The number of granules is much smaller than the numbers of patterns;
3. (3) Granules can directly describe multiple values of items; and

(4) It provides a user-oriented approach to determine the antecedent (or called premise) and consequence (conclusion) of association rules.

There are also several disadvantages when we discuss granules based on decision tables. The first problem is that we do not understand the relation between association rules (or patterns) and decision rules (or granules).

Although decision tables can provide a straightforward way to represent discovered knowledge, in cases of large number of attributes, decision tables lose their advantages because they can not be used to efficiently organize granules with different sizes. They also have not provided a mechanism to discuss meaningless rules.

## 3. Decision Rules and Association Rules

**Definition 1.** A set of items $X$ is referred to as an *itemset* if $X \subseteq V^{\mathcal{T}}$. Let $X$ be an itemset, we use *coverset*($X$) to denote the set of all objects $t$ such that $X \subseteq t$, i.e., $coverset(X) = \{t \mid t \in \mathcal{T}, X \subseteq t\}$.

Given an *itemset* $X$, its occurrence frequency is the number of objects that contain the *itemset*, that is $|coverset(X)|$; and its support is $|coverset(X)|/|\mathcal{T}|$. An itemset $X$ is called *frequent pattern* if its support $\geq$ *min_sup*, a minimum support.

**Definition 2.** Given a set of objects $Y$, its *itemset* which satisfies

$$itemset(Y) = \{a \mid a \in V^{\mathcal{T}}, \forall t \in Y \Rightarrow a \in t\}.$$

**Definition 3.** Given a frequent pattern $X$, its *closure* $Closure(X) = itemset(coverset(X))$.

From the above definitions, we have the following theorem (see [20]).

**Theorem 1.** Let $X$ and $Y$ be frequent patterns. We have

(1) $Closure(X) \supseteq X$ for all frequent patterns $X$;

(2) $X \subseteq Y \Rightarrow Closure(X) \subseteq Closure(Y)$.

**Definition 4.** A frequent pattern $X$ is *closed* if and only if $X = Closure(X)$.

We call the tuple $(\mathcal{T}, V^{\mathcal{T}}, C, D)$ a *decision table* of $(\mathcal{T}, V^{\mathcal{T}})$ if $C \cap D = \varnothing$ and $C \cup D \subseteq V^{\mathcal{T}}$.

We usually assume that there is a function for every attribute $a \in V^{\mathcal{T}}$ such that $a: \mathcal{T} \to V_a$, where $V_a$ is the set of all values of $a$. We call $V_a$ the domain of a, for example, $V_a = \{1, 0\}$ in TABLE II. $C$ (or $D$) determines a binary relation $I(C)$ (or $I(D)$) on $\mathcal{T}$ such that $(t_1, t_2) \in I(C)$ if and only if $a(t_1) = a(t_2)$ for every $a \in C$, where $a(t)$ denotes the value of attribute $a$ for object $t \in \mathcal{T}$. It is easy to prove that $I(C)$ is an equivalence relation, and the family of all equivalence classes of $I(C)$, that is a partition determined by $C$, is denoted by $\mathcal{T}/C$.

The classes in $\mathcal{T}/C$ (or $\mathcal{T}/D$) are referred to *C-granules* (or *D-granule*). The class which contains $t$ is called *C-granule* induced by $t$, and is denoted by $C(t)$.

**Definition 5.** Given a granule (e.g., a *C-granule* $cg = C(t)$), its covering set $coverset(cg) = \{t' \mid t' \in \mathcal{T}, (t', t) \in I(C)\}$.

It is also easy to have the following theorem based on the above definitions.

**Theorem 2.** Let granule $g = cg \wedge dg$, where $cg$ is a *C-granule* and $dg$ be a *D-granule*. We have $coverset(g) = coverset(cg) \cap coverset(dg)$.

For example, $g_1 = (a1 = 1 \wedge a2 = 1 \wedge a3 = 0 \wedge a4 = 0 \wedge a5 = 0 \wedge a6 = 0 \wedge a7 = 0) = C(g_1) \wedge D(g_1) = cg_1 \wedge dg_1$ using TABLE II and TABLE III; therefore

$$coverset(g_1) = coverset(cg_1 \wedge dg_1)$$
$$= coverset(cg_1) \cap coverset(dg_1)$$
$$= \{t_1, t_5, t_6\} \cap \{t_1\} = \{t_1\}.$$

**Definition 6.** Let $X$ be an itemset. We call it a decision *pattern* if $\exists g \in \mathcal{T}/C \cup D$ such that $X = \{a_i \in C \cup D \mid a_i(g) = 1\}$. We call $X$ the derived decision pattern of $g$.

From the above definitions, we have the following theorem (see [10]).

**Theorem 3.** Let $(\mathcal{T}, V^{\mathcal{T}}, C, D)$ be a *decision table* and $C \cup D = V^{\mathcal{T}}$. We say that the derived decision pattern of every granule $g \in \mathcal{T}/C \cup D$ is a closed pattern.

## 4. Multi-tier Structures

To solve the drawbacks of using decision tables, in this section, we discuss multi-tier structures. We also clarify the meaning of meaningless in this section.

TABLE IV
SMALLER GRANULES OF CONDITION GRANULES

| $C_i$ Granule | $a_1$ | $a_2$ | coverset |
|---|---|---|---|
| $cg_{i,1}$ | 1 | 1 | $\{t_1, t_5, t_6\}$ |
| $cg_{i,2}$ | 0 | 0 | $\{t_1, t_2, t_4\}$ |

(A) $C_i$-GRANULES

| $C_j$ Granule | $a_3$ | $a_4$ | $a_5$ | coverset |
|---|---|---|---|---|
| $cg_{j,1}$ | 0 | 0 | 0 | $\{t_1, t_5, t_6\}$ |
| $cg_{j,2}$ | 1 | 1 | 0 | $\{t_2\}$ |
| $cg_{j,3}$ | 1 | 1 | 1 | $\{t_1, t_4\}$ |

(B) $C_j$-GRANULES

We assume that $C_i$ and $C_j$ are two subsets of condition attributes. TABLE IV illustrates the smaller granules of the condition granules: $C_i$-granules and $C_j$-granules.

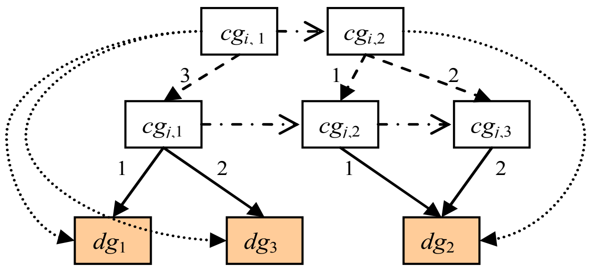


Fig. 1. The hierarchy of the multi-tiers.

A multi-tier structure can be describes as a pair ($\mathcal{H}$, $\mathcal{A}$), where $\mathcal{H}$ is a set of granule tiers and $\mathcal{A}$ is a set of association mappings between granule tiers.

Figure 1 illustrates a 3-tier structure for the possible granules in TABLE II and TABLE III, where $\mathcal{H} = \{C_i, C_j, D\}$ and $\mathcal{A} = \{\Gamma_{cd}, \Gamma_{ij}, \Gamma_{id}\}$. In Figure 1, *C-granules* are divided into *$C_i$-granules* and *$C_j$-granules*. The $C_i$ tier includes *$C_i$-granules* = $\{cg_{i,1}, cg_{i,2}, \ldots, cg_{i,k}\}$, the $C_j$ tier includes *$C_j$-granules* = $\{cg_{j,1}, cg_{j,2}, \ldots, cg_{j,m}\}$, and the $D$ tier includes *D-granules* = $\{dg_1, dg_2, \ldots, dg_s\}$, where $k = 2$, $m = 3$ and $s = 3$.

The 3-tier structure also includes three association mappings $\Gamma_{cd}$, $\Gamma_{ij}$, and $\Gamma_{id}$ which show the linkages between *C-granules* and *D-granules*, *$C_i$-granules* and *$C_j$-granules*, and *$C_i$-granules* and *D-granules*, respectively. They can be used to generate decision rules. For, instance, let $cg_k$ be a *C-granule* and $dg_z$ be a *D-granule*, association mapping $\Gamma_{cd}$ includes all possible links between *C-granules* and *D-granules*. The *link strength* between $cg_k$ and $dg_z$ is

$$|coverset(cg_k \wedge dg_z)|.$$

We call "$cg_k \rightarrow dg_z$" a decision rule. Its support is

$$sup = link\_strength(cg_k, dg_z) / N$$

and confidence is

$$conf = link\_strength(cg_k, dg_z) / |coverset(cg_k)|$$

where $N = |\mathcal{T}|$, the total number of transactions.

**Definition 7.** Let $cg_k$ be a *C-granule* and $cg_k = cg_{i,x} \wedge cg_{j,y}$. We call "$cg_{i,x} \rightarrow dg_z$" (or "$cg_{i,y} \rightarrow dg_z$") a *general rule* of "$cg_k \rightarrow dg_z$".

Especially in the multi-tier structure, we may define the term meaningless for a decision rule.

**Definition 8.** Let $cg_k$ be a *C-granule* and $cg_k = cg_{i,x} \wedge cg_{j,y}$. We call "$cg_k \rightarrow dg_z$" *meaningless* if its confidence is less than or equal to the confidence of its a general rule.

The rationale of this definition is analogous to the definition of interesting association rules. If we add a piece of extra evidence to a premise and obtain a weak conclusion, we can say the piece of evidence is meaningless.

## 5. Association Mappings

Association mappings are used to describe the association relations between granules in different tiers. It is desirable to derive these association mappings in order to generate user requested association rules efficiently based on the multi-tier structure.

The very important regulation for obtaining association mappings is the completeness, that is, the association rules we discover from the association mappings should be the same as we find from the original databases under users constraints. For this purpose, the first step is to formalize a basic association mapping in a decision table, and then we develop methods to derive other association mappings between granules in different tiers based on the basic association mapping.

The associations between *C-granules* and *D-granules* can be described as a basic association mapping $\Gamma_{cd}$ such that $\Gamma_{cd}(cg_k) = \Gamma_{cd}(cg_{i,x} \wedge cg_{j,y})$ is a set of *D-granule* integer pairs. For example, using the granules in Figure 1, we have

$$\Gamma_{cd}(cg_1) = \Gamma_{cd}(cg_{i,1} \wedge cg_{j,1}) = \{(dg_1, 3), (dg_2, 1)\}.$$

From it we can have two decision rules:

"$cg_{i,1} \wedge cg_{j,1} \rightarrow dg_1$" and "$cg_{i,1} \wedge cg_{j,1} \rightarrow dg_2$".

The next step is to derive association mapping $\Gamma_{ij}$ between *$C_i$-granules* and *$C_j$-granules* based on the basic association $\Gamma_{cd}$, where $\Gamma_{ij}(cg_{i,x})$ is a set of *$C_j$-granule* integer pairs.

At last, we will derive the association $\Gamma_{id}$ between *$C_i$-granules* and *D-granules* based on the association mappings $\Gamma_{ij}$ and $\Gamma_{cd}$, where $\Gamma_{id}(cg_{i,x})$ is a set of *D-granule* integer pairs.

It is more complicated to derive the association $\Gamma_{id}$ between $\mathcal{T}/C_i$ and $\mathcal{T}/D$ based on association $\Gamma_{cd}$ and $\Gamma_{ij}$. To simplify this process, we first review the composition operation that defined in [8].

Let $P_1$ and $P_2$ be sets of *D-granule* integer pairs. We call $P_1 \oplus P_2$ the *composition* of $P_1$ and $P_2$ which satisfies:

$$P_1 \oplus P_2 = \{(dg, f_1 + f_2) \mid (dg, f_1) \in P_1, (dg, f_2) \in P_2\} \cup \{(dg, f) \mid dg \in (gname(P_1) \cup gname(P_2)) - (gname(P_1) \cap gname(P_2)), (dg, f) \in P_1 \cup P_2\},$$

where $gname(P_i) = \{dg \mid (dg, f) \in P_i\}$. The operand of $\oplus$ is interchangeable; therefore, we can use $\oplus\{P_1, P_2, P_3\}$ to be the short form of $(P_1 \oplus P_2) \oplus P_3$. The result of the composition is still a set of *D-granule* integer pairs.

**Theorem 4.** Let $\Gamma_{cd}$ be an association between *C-*

*granules* and *D-granules* and $C_i \cap C_j = \varnothing$ and $C_i \cup C_j = C$. We have

(1) $\Gamma_{ij}(cg_{i,x}) = \{(cg_{j,y}, f1) \mid \Gamma_{cd}(cg_k) \neq \varnothing, f1 = \sum_{(dg,f)\in\Gamma_{cd}(cg_k)} f\}$;

(2) $\Gamma_{id}(cg_{i,x}) = \oplus\{\Gamma_{cd}(cg_k) \mid (cg_{j,y}, f1) \in \Gamma_{ij}(cg_{i,x})\}$;

where $cg_k$ be a *C-granule* and $cg_k = cg_{i,x} \wedge cg_{j,y}$.

**Proof:** For (1), based on the definition of the basic association, we have

$$|coverset(cg_{i,x} \wedge cg_{j,y})| = |coverset(cg_k)| = \sum_{(dg,f)\in\Gamma_{cd}(cg_k)} f.$$

Therefore $cg_{i,x}$ should be mapped to $cg_{j,y}$ in the multi-tier structure with link strength $\sum_{(dg,f)\in\Gamma_{cd}(cg_k)} f$.

For (2), there are multiple ways from $cg_{i,x}$ to *D-granules* in the multi-tier structure. The possible ways can be enumerated through $\Gamma_{ij}(cg_{i,x})$ and then $\Gamma_{cd}$, that is,

$$\{cg_{i,x} \rightarrow dg \mid (cg_{j,y}, f1) \in \Gamma_{ij}(cg_{i,x}), (dg, f) \in \Gamma_{cd}(cg_{i,x} \wedge cg_{j,y})\}.$$

It is the same result as using the composition in (2) if we sum all link strengths that converge to the *D-granules*. □

## 6. Evaluation

We first generate a large multidivisional database based on small store environments in a financial year, which includes 26,590 transactions. We first select 300 of most frequent products as attributes. It is also assumed that the user selection of condition attributes are the most frequent products with profits > 50% and selection of decision attributes with profits ≤ 20%. For each transaction, if an item/attribute appears it is set to one, otherwise it is set to zero.

The 2-tier structure is built up firstly based on these granules in the decision table by splitting these granules into condition granule (tier *C*) and decision granule (tier *D*). In this experiment, a 3-tier and 4-tier structures are constructed respectively to examine the actual performance of granule mining. The division is executed on the condition granule *C* for both structures in the experiment. For the 3-tier structure, *C* is partition into $C_i$ tier and $C_j$ tier under the assumption that products in $C_i$ have profit more than 90% and products in $C_j$ have profit between 90% and 50%. In the 4-tier structure, $C_i$ tier is further divided into $C_{i,1}$ and $C_{i,2}$ such that there are 5 attributes in $C_{i,1}$ and $C_{i,2}$, while there are 10 attributes in tier $C_j$ and 15 attributes in decision tier *D*.

Figure 2 shows the numbers of granules in each tier, where *All* means the decision table. Comparing with the decision table, there is much less granules in the multi-tier structure because there is only a small number of granules are useful for rule generation. Further, when the granules are divided to shorter granules for building more tiers, the number of granules became smaller. The reason is that when the original granule is split to smaller granules, more granules with same attributes are compressed together.

Figure 3 depicts the number of the meaningless rules found in the 3-tier structure and the 4-tier structure. The column 1 illustrates the percentage of meaningless rules in the 3-tier structure, where we compare decision rules with their one kind of general rules ($C_i \rightarrow D$). While the column 2 and 3 represents the percentages of meaningless rules in the 4-tier structure, where the column 2 uses ($C_{i,1} \rightarrow D$) only and the column 3 uses both general rules ($C_{i,1} \rightarrow D$) and ($C_i \rightarrow D$).

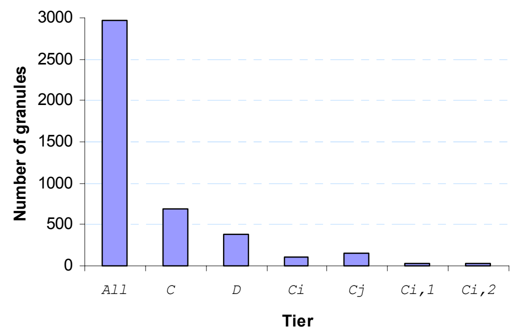


Fig. 2. Tiers and their Granules.

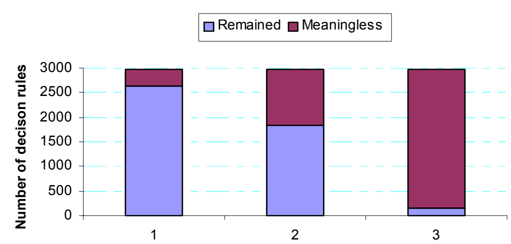


Fig. 3. *The percentage of meaningless decision rules.*

The experimental results demonstrate that the multi-tier structure only uses a small space to store meaningful multidimensional association rules. It can save the memory in the system and also improve the quality of association mining as well.

## 7. Discussions and Conclusion

It has been well recognized that pattern based mining techniques play an important role for interpretations of discovered knowledge. However, pattern based mining techniques often take long times to find patterns that also include much meaningless knowledge as well [13].

People have found some smaller patterns that can make their super-patterns inefficient. The concept of closed patterns has been used to prune these smaller patterns [26]. Although closed patterns can improve the performance of data mining, they still contain many patterns that are not what users really want [17] [18]. Constraint-based techniques attempt to find patterns that meet some sorts of constraints. The research issue is to reduce search spaces based on the properties of constraints and try to find possible useful patterns [4] [5] [13] [14]. These approaches can only be efficient for transaction databases.

Pattern mining is the first important step of traditional association mining. The second step is the rule generation, which is also a time consuming activity and can generate many meaningless rules [2]. Currently, people only have formal methods to eliminate redundant association rules [21] [19] rather than meaningless association rules.

It is painful when we review the above two steps: both take long time and contain uncertain information for determining meaningful knowledge. Now the big question is that is it possible to generate rules without these two steps? One possible solution is using decision rules in decision tables [11] [7] [1] [9]. Although there are many research works for decision tables and rough sets [12] [23], we have not found an appropriate formalization for describing meaningful association rules. There are three problems when we try to directly use decision tables for this question. The first problem is that we do not understand the relationship between decision rules and association rules. The second problem is that decision tables can only represent a small proportion of associations in databases. The last one is that decision tables have not provided a formal method to identify meaningless rules.

The aim of this research is to extend the scope of association rule mining and present an alternative technique, *granule mining*, for efficiently representing discovered knowledge in multidimensional databases.

Aattributes in granule mining are divided into tiers and multidimensional databases are efficiently compressed into granules in tiers. In addition, the very exciting point is that granule mining enables people to discuss association rules and their general rules, and hence the concept of meaningless association rules can be formally defined according to the relationships between association rules and their general rules.

# References


[1] J. W. Guan, D. A. Bell, D. Y. Liu, "The rough set approach to association rules", *ICDM* 2003, Florida, USA, 2003, 529-532.

[2] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques"*, Morgan Kaufmann Publishers, 2000.

[3] J. Hendler and E. A. Feigenbaum, "Knowledge is power: the semantic Web vision", in N. Zhong et al., eds., *Web Intelligence: research and development*, LNAI 2198, Springer-Verlag, 2001.

[4] C. K.-S. Leung, L.V.S. Lakshmanan, R. T. Ng, "Exploiting succinct constraints using FP-trees," *SIGKDD Explorations*, 2002, **4(1):** 40-49.

[5] A. J. T. Lee, W. Lin, C. Wang, "Mining association rules with multi-dimensional constraints," *Journal of Systems and Software*, 2006, **79(1):** 79-92.

[6] Y. Li, W. Yang and Y. Xu, "Multi-tier granule mining for representations of multidimensional association rules", 6th *IEEE International Conference on Data Mining* (*ICDM*), Hong Kong, 2006, 953-958.

[7] Y. Li and N. Zhong, "Interpretations of association rules by granular computing", 3rd *IEEE International Conference on Data Mining* (*ICDM*), Florida, USA, 2003, 593-596.

[8] Y. Li and N. Zhong, "Mining ontology for automatically acquiring Web user information needs," *IEEE Transactions on Knowledge and Data Engineering*, 2006, **18(4):** 554-568.

[9] Y. Li and N. Zhong, "Mining rough association from text documents", *5th International Conference on Rough Sets and Current Trends in Computing* (*RSCTC*), Kobe, Japan, 2006, 368-377.

[10] T. Y. Lin, The lattice structure of database and mining multiple level rules, *Bulletin of International Rough Set Society,* 2002, **6**: 11-16.

[11] Z. Pawlak, "In pursuit of patterns in data reasoning from data, the rough set way," *RSCTC* 2002, USA, 1-9.

[12] Z. Pawlak and A. Skowron, "Rough sets and Boolean reasoning", Information Sciences, 2007, **177:** 41–73.

[13] J. Pei, J. Han, "Constrained frequent pattern mining: a pattern-growth view", *SIGKDD Exploration*, 2002, **4(1):** 31-39.

[14] C. Perng, H. Wang, S. Ma, J. L. Hellerstein, "Discovery in multi-attribute data with user-defined constraints", *SIGKDD Exploration*, 2002, **4(1):** 56-64.

[15] S. Tsumoto and S. Hirano, "Visualization of rule's similarity using multidimensional scaling", *ICDM* 2003, Florida, USA, 2003, 339-346.

[16] P. Tzvetkov, X. Yan and J. Han, "TSP: Mining top-K closed sequential patterns," *ICDM* 2003, Florida, USA, 2003, 347-354.

[17] S.T. Wu, Y. Li and Y. Xu, "Deploying approaches for pattern refinement in text mining", *ICDM* 2006, Hong Kong, 2006, 1157-1161.

[18] S.T. Wu, Y. Li, Y. Xu, B. Pham and P. Chen, "Automatic pattern taxonomy extraction for Web mining," *IEEE/WIC/ACM International Conference on Web Intelligence*, Beijing, China, 2004, 242-248.

[19] Y. Xu, and Y. Li, "Mining non-redundant association rule based on concise bases", *International Journal of Pattern Recognition and Artificial Intelligence*, 2007, **21(4):** 659-675.

[20] L. A. Zadeh, "Granular computing – the concept of generalized constraint-based computing", *RSCTC* 2006, Kobe, Japan, 12-14.

[21] M. J. Zaki, "Mining non-redundant association rules," *Data Mining and Knowledge Discovery*, 2004, **9:** 223 -248.

[22] S. Zhang, J. Zhang, X. Zhu and Z. Huang, "Identifying follow-correlation intemset-pairs", *ICDM* 2006, Hong Kong, 2006, 765-774.

[23] Y. Y. Yao, "A Comparative Study of Formal Concept Analysis and Rough Set Theory in Data Analysis", *RSCTC* 2004, LNCS 3066, 2004, pp. 59-68.