A New Framework based on Cognitive Psychology for Knowledge Discovery

Jun Yang
School of Software, Jiangxi Agricultural University, Nanchang, China
Email: newcustomer@yeah.net

Yinglong Wang
School of Software, Jiangxi Agricultural University, Nanchang, China
Email: wangyinglongdl@sohoo.com

***Abstract*—The mainstream of knowledge discovery encompasses mining algorithms with high-performance and high-scalability in kinds of databases and application backgrounds. Some research were carried out in another shortcut which focuses on research to high-level theory framework in order to reveal the potential essence, principles and complexity of knowledge discovery and then reacts to the mainstream development. The contributions are as follows: 1) the process and system of KD is regards as complex cognitive process and system. 2) Two coordinators, which are used to simulate cognitive psychology characters so as to make system discover knowledge independently and maintain knowledge base in real-time, are proposed. 3) Double bases cooperating mechanism, which can reveal the equivalence relationship between two categories (structure corresponding theorem) in knowledge discovery process under the conditions that database and knowledge base are specially constructed, is proposed. It paves the way for the implementation of directional searching and directional mining of two coordinators. 4) A new process model based on double bases cooperating mechanism is presented. 5) A new mining algorithm is brought forward consequently, which embodies the driving role to mainstream development.**

***Index Terms*—knowledge discovery, data sub-class structure, knowledge node, heuristic coordinator, maintenance coordinator, double bases cooperating mechanism**

I. INTRODUCTION

The fast progress in data acquisition and storage technology has led to the amount of data stored in database, data mart and data warehouse grow rapidly. How to derive useful knowledge from massive data? At present, KDD (Knowledge Discovery in Database) has become a hot academic research [1, 8] and has been applied to many fields of science and engineering [2, 9].

Due to its cross-disciplinary character, KDD has received contributions from various perspectives by researchers in different fields. It mainly includes the following aspects: Some researchers study KDD from the perspective of database, and their emphasis is efficiency [3, 4, 10, 11]. Some researchers study KDD from the perspective of machine learning, and their emphasis is effectiveness [5, 6, 12]. Other researchers study KDD from the view point of statistics, and their emphasis is valid [7, 8]. Still other researchers study KDD from the perspective of microeconomics, and their emphasis is the maximum utility. At present, the mainstream development of KDD has mostly concentrated on developing high-performance and high-scalability mining algorithms in various kinds of databases and application backgrounds [10]. Almost no one researches on the high-level framework or theoretical foundations of KDD. A general data mining framework based on evidence theory is proposed in [11]. It provides a common method for representing knowledge and supports parallel computing and has specific operators. A complete autonomic knowledge discovery framework based on justification is put forward in [12]. Its kernel is a reasoning component. It realizes autonomic knowledge discovery by firstly computing the possibility of every mining task based on interestingness and justification intensity, and then reordering these mining tasks according to these possibilities. In addition, it has heuristic function which is used to propose new mining task. It brought forward an idea of studying data mining theory by combining microeconomics and inductive database [13, 14]. Advanced an "information paradigm" from the perspective of statistics and discussed its application in association rules and classification theoretically.

Unfortunately, the above researches have not been homely discussed the theoretical basis, or failed to provide the specific methods. So they can't obviously improve the performance of knowledge discovery. In this paper, we regard knowledge discovery as a cognitive system, study the process of knowledge discovery from the perspective of cognitive psychology, and our

Corresponding author: Wang Yinglong.

doi:10.4304/jsw.8.1.47-54

emphasis is self-cognition. We make use of two important features of cognitive psychology, i.e. "creating intent" and "psychology information maintenance", to deal with two important issues of knowledge discovery and reconstruct the process of knowledge discovery. Specifically, (1) Making the system find knowledge shortage automatically by simulating "creating intent" in order to realize heuristic focus (in addition to focuses which are interested by users). (2) Performing the function of real-time maintenance of knowledge base by simulating "psychology information maintenance". So we used database and knowledge base at the same time, and justified 1-1 mapping relation between them under the condition that they are specifically constructed. Meanwhile we improved the process model of knowledge discovery and derived some new mining methods. This paper is the discussion of these inherent principles, processing model and technical methods.

Encompassing the above two important problems, we make the following work.

Firstly, we conducted studies of coordinators (algorithms) and constructed two coordinators to solve the above two problems respectively. Namely simulating "creating intent" by use of heuristic coordinator so as to find knowledge shortage automatically, this coordinator was achieved by heuristic coordinating algorithm; and simulating "psychology information maintenance" by use of maintenance coordinator in order to realize real-time maintenance of knowledge base, this coordinator was carried out by maintenance coordinator algorithm.

Secondly, the key technologies of the above two coordinators is "directional searching" and "directional mining" which can reduce the searching space and complexity of algorithms. So we can construct certain mapping between database and knowledge base. This kind of mapping is named as double bases cooperating mechanism. It can discover potential essence, principles and complexity of knowledge discovery from a special perspective.

Thirdly, the research to coordinators (algorithms) and theoretical foundation drives the mainstream development. One manifestation is inducing new processing model of knowledge discovery, such as KDD*.

Finally, the research to coordinators (algorithms) and theoretical foundation drives the mainstream development. Another manifestation is inducing new algorithms of knowledge discovery, such as Maradbcm which was used to mining association rule.

II. Double Bases Cooperating Mechanism: Theoretical Foundation

In order to reduce the complexity of algorithm and improve its efficiency under the new framework, the double bases cooperating mechanism is proposed. That is, under the condition of specific construction of both database and knowledge base, the 1-1 mapping between database and knowledge base (structure corresponding theorem) is constructed. This enables massive database will be mined directionally when knowledge shortage is discovered by heuristic coordinator. Thus the search space and the complexity are both reduced.

A. Basic conceptions

For the convenience of discussion, we firstly give the following basic conceptions.

a. Numerical Domain Boolean Algebra

Each attribute in discourse universe X corresponds to a numerical domain. The numerical domain of each attribute is required to be an ordered set (totally ordered set).

The numerical domain of attribute Xi is denoted by Di, i=1, 2, …, s. The Cartesian Product of the corresponding numerical domains of all attributes in discourse universe X is called the numerical domain of X, which is denoted as D. Namely, $D = D_1 \times D_2 \times \ldots \times D_s$

We need to divide numerical domain. After the division, the subset of the numerical domain is called numerical sub-domain.

Definition 1 The numerical domain D_i is divided into t_i number of numerical sub-domain $D_{i1}, D_{i2}, \cdots, D_{it_i}$, When the division satisfies the following conditions, it is called the regular division of numerical domain D_i.

1) The union of all the numerical sub-domain in some numerical D_i is the numerical domain D_i itself, i.e., $\bigcup_{j=1}^{t_i} D_{ij} = D_i, i = 1, 2, \cdots, s$;
2) The intersection of two arbitrary numerical sub-domains of a numerical domain is empty, i.e., $D_{ij} \cap D_{ik} = \varnothing (j \neq k)$;
3) The order of the elements in numerical domain D_i and the order of corresponding numerical sub-domains $D_{i1}, D_{i2}, \cdots, D_{it_i}$ must meet one of the following conditions:

The numerical sub-domains are sorted according to the ascending (descending) order of the elements in the numerical domain. That is to say, for any v_1, $v_2 \in D_i$, $v_1 \leq v_2$, if $v_1 \in D_{ij_1}$, $v_2 \in D_{ij_2}$, then $j_1 \leq j_2$ (for any v_1, $v_2 \in D_i$, $v_1 \geq v_2$, if $v_1 \in D_{ij_1}$, $v_2 \in D_{ij_2}$, then $j_1 \geq j_2$).

Definition 2 The regular division of each numerical domain can form a topological space. Suppose E_i = $\{ D_{i1}, D_{i2}, \cdots, D_{it_i} \}$, i=1, 2, …, s, then E_i is a set that is composed of some numerical sub-domains and is called the attribute sub-domain family of attribute X_I. Any subset of E_i is called the sub-domain family.

All subsets of attribute sub-domain family E_i of attribute X_i constitutes its power set γ_I. Obviously, $<E_i, \gamma_i>$ is a topological space which is called the numerical sub-domain division topological space of attribute X_I. For any $y \in \gamma_i$, $y = \{D_{ij} | D_{ij} \in E_i, j \in J \subseteq \{1, 2, \ldots, t_i\}\}$, we define the proper set of y as follows:

$$\Psi(y) = \bigcup_{j \in J} D_{i,j}$$

Since the product of topological spaces is still a topological space, $<E_1, \gamma_1> \times <E_2, \gamma_2> \times \ldots \times <E_s, \gamma_s>$is also a topological space which is denoted as $<E, \gamma>$ and is called numerical domain division topological space, numerical domain topological space of discourse universe X for short.

Definition 3 There is such an element in the numerical domain topological space of discourse universe X: it is an ordered tuple composed of s number of numerical sub-domains, each numerical sub-domain is chosen from the corresponding attribute sub-domain family. This kind of ordered tuple is called the Basic Element of numerical domain topological space. The set composed of all basic elements is called the basic element set, which is shown as follows:

$$\{ (d_1, d_2, \ldots, d_s) \mid d_1 \in E_1, d_2 \in E_2, ..., d_s \in E_s\} \quad (1)$$

Suppose the basic element set is F. The number of elements in F is $v=t_1 \times t_2 \times \ldots \times t_s$. Where the meaning of v remains unchanged and we denote $\{1, 2, \ldots, v\}$ byΔ. We can then arrange the basic element set in a row and denote it as $F_1, F_2, \ldots, F_v$.

According to the property of product topological space, γ is the power set of F. Therefore, the number of elements in γ is 2^V. For any $y \in \gamma$, y is depicted as follows:

$$y=\{F_i \mid F_i \in F, i \in I \subseteq \Delta\} \quad (2)$$

Similar to the method defining the proper set in numerical sub-domain topological space, we can define the proper set of the elements of power set γ in numerical domain topological space. First, for any basic element $y=(d_1, d_2, \ldots, d_s)$, the proper set of y is defined as:

$$\psi(y) = d_1 \times d_2 \times \ldots \times d_s$$

Where "$\times$" is Euclidean Product. Therefore, we can define the proper set of any element in γ. For any $y \in \gamma$, $y=\{F_j \mid F_j \in F, j \in J\}$, the proper set of y is: $\Psi(y) = \bigcup_{j \in J} F_j$

According to the definition of proper set, for any y, y_1, $y_2 \in \gamma$,

$$\psi(y_1 \cap y_2) = \psi(y_1) \cap \psi(y_2); \quad (3)$$

$$\psi(y_1 \cup y_2) = \psi(y_1) \cup \psi(y_2); \quad (4)$$

$$\psi(\sim y) = \sim \psi(y). \quad (5)$$

For any y_1, $y_2 \in \gamma$, according to (2), y_1 can be represented as $\{F_i \mid F_i \in F, i \in I_1 \subseteq \Delta\}$, y_2 can be represented as $\{F_i \mid F_i \in F, i \in I_2 \subseteq \Delta\}$. Therefore,

$$y_1 \cap y_2 = \{ F_i \mid F_i \in F, i \in I_1 \cap I_2 \subseteq \Delta\}$$

i.e., $y_1 \cap y_2 \in \gamma$, and similarly, $y_1 \cup y_2 \in \gamma$, $\sim y_2 \in \gamma$. Therefore, in set familyγ, the intersection and union between elements as well as complement of an element is still the element of γ, i.e. these operations are closed.

Theorem 1 Set family γ and the intersection and union operation between its elements as well as the complement operation of its elements constitute a algebra system $<\gamma, \cap, \cup, \sim>$ and this algebra system is a Boolean Algebra system, the zero element of which is Φ; The identity element is the basic element set F.

Definition 4 We call the above Boolean Algebra constituted by the power set γ and the intersection operation, union operation and the complement operation of its elements, which are in numerical domain topological space of the discourse universe X as the numerical domain Boolean Algebra.

b. Knowledge Node Boolean Algebra

Degree word is used to describe the state of attributes in discourse universe, namely language value. Attribute degree word is composed of some attribute and one of its degree words, and it describes a certain state of this attribute. For i=1, 2, …, s, attribute X_i (i=1, 2, …, s) has t_i ($t_i \geq 2$) degree words. These degree words, sorted according to ascending order or descending order, are A_{i1}, A_{i2}, …, A_{it_i}.

The set of all attribute degree words of attribute X_i is denoted as B_i; the set of all attribute degree words in discourse universe X is denoted as B. Obviously, $B=\bigcup_{i=1}^{s} B_i$.

Putting an appropriate relation (disjunction or conjunction) between two attribute degree words will create a new meaning. Suppose a represent "high temperature" and b represent "strong pressure", then $a \wedge b$ represents "high temperature and strong pressure" and $a \vee b$ represents "high temperature or strong pressure".

For any attribute degree word A_{ij}, its negation means its opposite meaning. We define the negation of A_{ij} as:

$$\neg A_{ij} = A_{i1} \vee \ldots \vee A_{i,j-1} \vee A_{i,j+1} \vee \ldots \vee A_{it_i} .. \quad (6)$$

Obviously, this definition is totally different from that defined in common two-valued logic.

Definition 5 A knowledge node of discourse universe X is the following well-formed formula which does not include negation operation:

$$\theta_0 a_1 \theta_1 a_2 \ldots \theta_{m-1} a_m \theta_m \quad (7)$$

where, $a_i \in B$, i=1, 2, …, m; $\theta_i \in J$, i=0, 1, …, m. J is the set that includes 4 symbols---"$\wedge$", "$\vee$", " (" , ")", their corresponding combinations, and NOP; but the values of θ_i in (7) must make sense; only θ_0 and θ_m can be NOP.

When m=0, (7) is empty and is denoted as Φ, which means that the knowledge node is empty. Knowledge node composed of single attribute degree word is named as primitive knowledge node. Knowledge node composed of many attribute degree words named as combined knowledge node. The disjunction of all attribute degree words of attribute X_i is a very important knowledge node and is called the disjunctive min-term of X_i. It means the unrestricted state of this attribute. And this knowledge node is marked as U_i, i=1, 2, …, s i.e., $U_i = A_{i1} \vee A_{i2} \vee \ldots \vee A_{it_i}$ .Contrary to the empty knowledge node, the knowledge node that is the disjunction of all attribute

degree words of discourse universe X is called the complete knowledge node, which is marked as Ω, i.e.:

$$\Omega = \vee_{a \in B} a$$

The set of all the knowledge nodes of discourse universe X is called the knowledge node set of discourse universe X, which is denoted as N.

In set N of knowledge node, we can define the negation and the conjunction and the disjunction between knowledge nodes. For any n_1, $n_2 \in N$, suppose $n_1=F_1$, $n_2=F_2$. Naturally, we can define the conjunction and the disjunction between knowledge nodes n_1 and n_2 as:

$$n_1 \wedge n_2 = F_1 \wedge F_2;\ n_1 \vee n_2 = F_1 \vee F_2$$

And the negation of knowledge node $n= \theta_0 a_1 \theta_1 a_2 \ldots \theta_{m-1} a_m \theta_m$ as:

$\neg n = \neg\ (\theta_0 a_1 \theta_1 a_2 \ldots \theta_{m-1} a_m \theta_m) = \theta_0' a_1' \theta_1' a_2' \ldots \theta_{m-1}' a_m' \theta_m'$, where $\theta_i' \in J$, i=0, 1, …, m. If a_i' is the positive attribute degree word, we won't change it; but if it is the negation of the attribute degree word, we should replace it with the right side of (6). After this replacement, $\neg n$ is still a wff and does not include negation operation.

From the definitions of the above three kinds of operation of knowledge nodes, we can draw the conclusion that the conjunction and the disjunction of two knowledge nodes is still a knowledge node, and also, the negation of a knowledge node is still a knowledge node. Thus the operation of conjunction, disjunction and negation are closed in the set of knowledge nodes.

The normal form theorem: knowledge node $n=\theta_0 a_1 \theta_1 a_2 \ldots \theta_{m-1} a_m \theta_m$ can be exclusively represented as a major disjunctive normal form:

$$n = L_1 \vee L_2 \vee \ldots \vee L_k \tag{8}$$

$$n = \bigvee_{i=1}^{y} \left[\bigwedge_{j=1}^{s} u_{ij} \right] \tag{9}$$

where i=1, 2, …, y, $L_i = u_{i1} \wedge u_{i2} \wedge \ldots \wedge u_{is_i}$, they are all the conjunction of single attribute degree word.

Definition 6 We call the form in (9) as the standard form of knowledge node and the corresponding node is called standard knowledge node.

We can see from (9) that the standard knowledge node is composed of the disjunction of several simple conjunction formulas. Every simple conjunctive formula is composed of s terms, and every term is chosen from its corresponding set of attribute degree words. This kind of simple conjunctive formula is called the basic simple conjunctive formula of discourse universe X. The set of all basic simple conjunctive formulas is called the basic simple conjunctive formula set and is marked as H.

Suppose all elements of H are marked as H_1, H_2, …, H_V, so for any $n \in N$,

$$n = \bigvee_{j \in J} H_j \tag{10}$$

where, J is the subset of Δ.

Theorem 2 $<N, \wedge, \vee, \neg>$ is a Boolean Algebra and the zero element of which is an empty knowledge node; the identity element is the complete knowledge node Ω.

Definition 7 We call the Boolean Algebra formed by the set of knowledge nodes and the operation of disjunction, conjunction, and negation in the set as knowledge node Boolean Algebra.

c. Data Sub-Class Structure Boolean Algebra

Definition 8 For a given discourse universe X, we can establish the relational database in the following pattern:

$$\Re(NO, X_1, X_2, \ldots, X_s). \tag{11}$$

where, NO is the primary key and is chosen from natural number set, and it can exclusively identify a tuple; x_i (i=1, 2, …, s) is the attribute of discourse Universe X. The relational database built on pattern (11) is called the database in discourse universe X and is marked as $\Re(X)$. Any tuple u in the database $\Re(X)$ is a vector of s+1 dimensions (num, x_1, x_2, …, x_s).

For any open set $y \in \gamma$ in topological space $<F, \gamma>$, we can define a set of tuple $\{u |\ u \in \Re(X),\ a(u) \in \psi(y)\}$, and mark it as $<y, \Re(y)>$. This kind of set of tuple is called the data sub-class structure of discourse universe X. Y is called the data part of data sub-class structure $<y, \Re(y)>$. $\Re(y)$ is the tuple part of it. The data sub-class structures of discourse universe constitute the data sub-class structure set $<\gamma, \Re(\gamma)>$.

In the data sub-class structure set, we can define the equivalence relation between two data sub-class structures. i.e. $<y_1, \Re(y_1)> = <y_2, \Re(y_2)>$ iff $y_1=y_2$ and $\Re(y_1)=\Re(y_2)$.

The operating relations among elements of the data sub-class structure set can be constructed as follows. Given any three elements: $<y_1, \Re(y_1)>$, $<y_2, \Re(y_2)>$and $<y, \Re(y)>$ (which can be the same) in data sub-class structure set $<\gamma, \Re(\gamma)>$, we give the following definitions:

$<y_1, \Re(y_1)> \cap <y_2, \Re(y_2)> = <y_1 \cap y_2, \Re(y_1) \cap \Re(y_2)>$ (12)

$<y_1, \Re(y_1)> \cup <y_2, \Re(y_2)> = <y_1 \cup y_2, \Re(y_1) \cup \Re(y_2)>$ (13)

$$\sim <y, \Re(y)> = <\sim y, \sim \Re(y)> \tag{14}$$

So the "union", "intersection" and "complement" of data sub-class structures are closed in data sub-class structure set.

Theorem 3 $<<\gamma, \Re(\gamma)>, \cap, \cup, \sim>$ is a Boolean Algebra in which the zero element is $<\Phi, \Re(\Phi)>$, and the identity element is $<\Omega, \Re(\Omega)>$.

Definition 9 We call the Boolean algebra formed by the data sub-class structure set and the "intersection", "union" and "complement" among its elements as data sub-class structure Boolean Algebra.

B. The Relation among Three Boolean Algebras

From the definition of attribute degree words, we can see that any group of attribute degree words can solely determine a regular partition of numerical domain. Through the relation of attribute degree words and the regular partition of numerical domain, we can establish the corresponding relation between knowledge node and numerical domain topological space. We firstly give the following lemmas:

Theorem 4 [15] the relationship among numerical domain Boolean algebra, knowledge node Boolean algebra and data sub-class structure Boolean algebra is isomorphic.

Theorem 4 demonstrates that the power set γ of numerical domain topological space, knowledge node set N and data sub-class structure set <γ, $\mathcal{R}$ (γ)> have the same algebra structure. It is significant for the creation of the knowledge node and the reproduction of data sub-class structure.

C. Two Categories and Their Relation

Given discourse universe X, it has a set N of knowledge nodes. For example, in the discourse universe X, knowledge node n_1= "high temperature", n_2= "strong pressure". If there is an inherent rule in discourse universe X---"If the temperature is high, then the pressure is strong", then there exists the reasoning relation from knowledge node n_1 to n_2: $n_1 \rightarrow n_2$ or $r(n_1, n_2)=r$.

Definition 10 We call the above reasoning relation from knowledge node n_1 to n_2: $n_1 \rightarrow n_2$ as a positive rule, and $n_1 \times \rightarrow n_2$ as a negative rule. Both the positive rules and the negative rules construct the rule set of discourse universe X.

Theorem 5 The set N of knowledge node in discourse universe X and the reasoning relation r among its elements form a category.

We call the category made up of N and the reasoning relation r among its elements as the reasoning category and denote as $Cr(N)$.

Definition 11 For data sub-class structure set <γ, $\mathcal{R}$ (γ)> of discourse universe X, we can establish, between elements, the accessibility relation "$\propto$": $<y_1, \mathcal{R}(y_1)> \propto <y_2, \mathcal{R}(y_2)>$ iff $\mathcal{R}(y_1) \subseteq \mathcal{R}(y_2)$. If in <γ, $\mathcal{R}$ (γ)>, there is no accessibility relation from the element $<y_1, \mathcal{R}(y_1)>$ to the element $<y_2, \mathcal{R}(y_2)>$, we say that there is inaccessibility relation from $<y_1, \mathcal{R}(y_1)>$ to $<y_2, \mathcal{R}(y_2)>$. All accessibility relations constitute the set of accessibility relation; all inaccessibility relations constitute the set of inaccessibility relation.

The data sub-class structure base of discourse universe X is composed by the data sub-class structure set <γ, $\mathcal{R}$ (γ)> which is created by its relational database $\mathcal{R}(X)$ and the accessibility and inaccessibility relations (between the elements of data sub-class structure set) set.

Theorem 6 the data sub-class structure set <γ, $\mathcal{R}$ (γ)> of discourse universe X and the accessibility relations "$\propto$" among its elements can create a category.

The category composed of data sub-class structure set <γ, $\mathcal{R}$ (γ)> and the accessibility relation among its elements is called accessibility category of data sub-class structure of discourse universe X, denotes as $C_{\propto}<\gamma, \mathcal{R}(\gamma)>$.

Theorem 7 (Structure Corresponding Theorem): The reasoning category $Cr(N)$ of discourse universe X and the accessibility category $C\propto<\gamma, \mathcal{R}^c(\gamma)>$ of complete data sub-class structure are equal.

So far, we have discussed five algebra systems induced by discourse universe X and their relations (as shown in Fig1). Where, three Boolean Algebras are determined by the attributes of discourse universe and the partitions of numerical domain, which is formal; two categories are determined by the inherent relative regulations among each attribute of discourse universe, which is content.

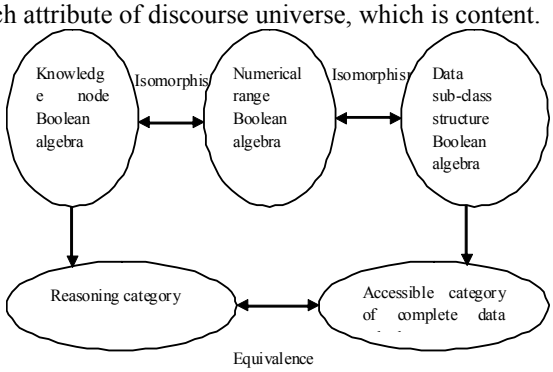

Figure 1. Corresponding relation among 5 structures of discourse X

Based on the above structure corresponding theorem, the 1-1 mapping between the layers of data sub-class structure in mining database and primitive knowledge node in mining knowledge base is constructed (as shown in Fig. 2). This greatly reduced the search space and the complexity of algorithms and built foundations for directional mining and directional searching.

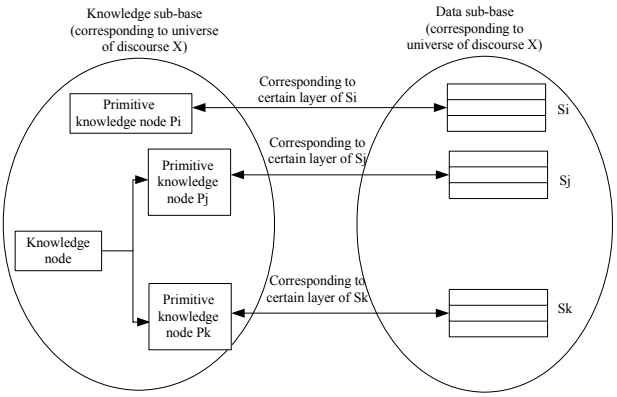

Figure 2. Corresponding relationship between layers of data sub-structure and primitive knowledge nodes

III. RESEARCH ON COORDINATOR AND COORDINATING ALGORITHM

A. Cognitive Psychology: A New Perspective of Knowledge Discovery

Cognitive psychology spring up in the mid-1950s, then in 1967, Neisser delivered a monograph which was first named by cognitive psychology. In this book, cognitive psychology was defined as "processes by which the information achieved by sensory is transformed, reduced, elaborated, stored, recovered and used". Hence, there was a rapid development of the cognitive psychology, and its influence extended from time to time. The key idea of cognitive psychology is information processing, which is the analogy between human brain and computer and look upon human brain as an information processing system

similar to computer. Cognitive psychology focuses on discovering the interior psychology mechanism of cognitive process, namely how the information is acquired, reserved, processed and used. In the knowledge discovery system, two cognitive psychology characteristics: "creating intent" and "psychology information maintenance" was simulated in order to improve the cognitive autonomy of the system, which is the motivation of our research.

B. Heuristic Coordinating Algorithm and Heuristic Coordinator

The function of heuristic coordinator is simulating "creating intent" which is a cognitive psychology character, therefore the system will discover knowledge shortage by itself (The shortage of knowledge here is the knowledge that has not appear in knowledge base so far). In the process of classical KDD, the focus of the system is usually the interesting direction provided by the user, so the potentially useful information in large volume of data is usually ignored. To discover more potentially useful information in order to make up the limitations of both users and experts as well as to improve the cognitive autonomy, we constructed heuristic coordinator. In this way, besides the user-driven focus, knowledge discovery is equipped with a new function of self-focus direction.

a. Find Knowledge Shortage

The implement techniques of heuristic coordinator is mainly on searching the non-association state of knowledge nodes in knowledge base, in order to discover "knowledge shortage", then activate the corresponding data sub-class structure in real database, thus the directional mining process is produced.

What is "knowledge shortage". The following conditions should be satisfied:

1) Shortage knowledge only take the rule with single consequence into account;

2) The attribute degree words of the same attribute cannot appear in both premise and consequence of the same rule;

3) According to specific problem to determine maximum number of premise which is shortage knowledge, because superabundance premise certainly make the rule difficult to understand;

4) For some rule $e_1 \wedge e_2 \wedge \ldots \wedge e_m \rightarrow h$, the length of it is m+1;

5) If there is a rule $A \rightarrow B$ and a rule $B \rightarrow C$ in knowledge base, then rule $A \rightarrow C$ is not shortage knowledge.

How to discover "knowledge shortage": In knowledge base, if only consider the knowledge with single premise and consequence, we can see the premise and consequence as graph vertex, using the method of finding accessibility relation in graph theory, to find "knowledge shortage". But, in knowledge base, many rules have many conditions, so we define directional hypergraph [16] to solve this problem.

Definition 12 A hypergraph is a two-tuple $<V, E>$, where $V=\{p_1, p_2, \ldots, p_n\}$is nonempty set, its elements are vertexes of directional graph; $E=\{e_1, e_2, \ldots, e_m\}$is a set of hyper-edges, where any e_i (i=1, 2, …, m) is a subset of V.

Definition 13: A directional hypergraph is a two-tuple $<V, E>$, where $V=\{p_1, p_2, \ldots, p_n\}$ is a set of primitive knowledge node. These primitive knowledge nodes are vertexes of it. $E=\{e_1, e_2, \ldots, e_m\}$ is a directional edge which is corresponding to the rule in knowledge base. For example, a rule $r_i=p_1 \wedge p_2 \wedge \ldots \wedge p_k \rightarrow p_j$, the directional edge $e_i=<(p_1, p_2, \ldots, p_k), p_j>$ is an order tuple, the first element is a subset of V, corresponding to the premise of the rule, the second element is a element of V, corresponding to the consequence of the rule.

Definition 14: We say that the vertexes which are associated with a hyper-edge are adjacent each other. If there is one common vertex of two hyper edges, we say that the two hyper-edges are adjacent.

We propose the following algorithm which is used to calculate the adjacent matrix P (H) of directional hypergraph based on the Warshall algorithm calculating adjacent of directional graph.

Function calculate_reach_matrix

Form a matrix $Pn \times n$ whose elements are ID of all primitive knowledge nodes in knowledge base, and express it use a 2-D array, all elements are 0, that is $P(i, j)$=0, i, j=1, 2, …, n;

1) e: =1;
2) Read the eth rule with length 2 in knowledge base r_e: $p_i \rightarrow p_j$;
3) Let $P(i, j)$=1;
4) Calculate_matrix1 (j, i, n); //call the procedure of Calculate_matrix1 (see below)
5) Have all the rules with length 2 in knowledge base been read. If not, then e: =e+1, turn to step 3); else turn to step 7);
6) e: =1;
7) Read the eth rule whose length is longer than 2 in knowledge base r_e: $p_{f1} \wedge p_{f2} \wedge \ldots p_{fj} \rightarrow p_i$;
8) Calculate_matrix2 ((f_1, f_2, …, f_j), i); //call the procedure of Calculate_matrix2 (see below)
9) Have all the rules whose length is longer than 2 of knowledge base been read. If not, then e: =e+1, turn to step 8); else end;

Procedure Calculate_matrix1 (j, i, n: integer)

1) for k: =1 to n
2) $P(j, k)$: = $P(j, k) \vee P(i, k)$
3) for m: =1 to n
4) If $P(m, j)$=1 then
5) for k: =1 to n
6) $P(m, k)$: =$P(m, k) \vee P(j, k)$

Procedure Calculate_matrix2 ((f_1,f_2, …,f_j), i)// (j>1)

1) If the virtual knowledge node $pf_1 \wedge pf_2 \wedge \ldots \wedge pf_j$ doesn't exist, add a row to reachable matrix so as to represent it ;
2) P ($pf_1 \wedge pf_2 \wedge \ldots pf_j$, i)=1;
3) for s: =1 to n
4) P ($pf_1 \wedge pf_2 \wedge \ldots pf_j$, s): = P ($pf_1 \wedge pf_2 \wedge \ldots pf_j$, s) $\vee$ $P(i, s)$

Theorem 8 The matrix gained by the above new algorithm is a reachable matrix.

Proof For the elements of matrix whose length is equal to 2, the reachable matrix can be calculated by Warshall algorithm.

For the elements of matrix whose length is longer than 2, denote $i_1 \wedge i_2 \wedge \ldots \wedge i_h \rightarrow t$ as the added edge, according to the specification that the post-condition of rules corresponding to the element of reachable matrix is the single primitive knowledge node, the reachable matrix whose new added value is 1 can only be: (1) p ((i_1, i_2, …, i_h), t); (2) g, which is an arbitrary node reachable from t, is also reachable by hypernode (i_1, i_2, …, i_h). That is if p_{tg}=1 (g=1, 2, …, n), then p ((i_1, i_2, …, i_h), g)=1. There aren't other elements among them existing reachable relation except the two above-mentioned cases.

In summary, we can prove theorem 8.

Now we have picked out shortage knowledge whose length is shorter than 2. But all shortage knowledge whose length is longer than 2 cannot be found in the above reachable matrix P (H). The reason is that this matrix only includes the conjunctive knowledge nodes appeared in the knowledge base. So we defined rule intensity to find out shortage knowledge whose length is longer than 2.

We describe the objectivity of rule intensity using the concept of support of association rule. In other words, the support of rule $A \rightarrow B$ in some transaction set is the percentage of transactions that contain both A and B in this transaction set.

Definition 15 Interestingness is the degree of how the user interested in different attributes or attribute degree words in database. In another word, it is the degree of how the user interested in the primitive knowledge nodes in knowledge base. In the preprocessing, user set the interestingness of each attribute degree word which is the interestingness in primitive knowledge node e_k, denoted by *Interestingness* (e_k) whose domain is. Higher of the *Interestingness* (e_k) is, more interesting the primitive knowledge node is for the user. As for the conjunctive knowledge node $F=e_1 \wedge e_2 \wedge \ldots \wedge e_m$, the interestingness of it is the average value of the interestingness of each primitive knowledge node, i.e.,

$$Interestingness\ (F)=\sum_{i=1}^{m} Interestingness(e_i)/m \qquad (15)$$

For some rule r_i: $F \rightarrow h$, the interestingness of it is:

$$Interestingness(F)=\left[\sum_{i=1}^{m} Interesting(e_i)+Interesting(H)\right] /Len(r_i) \qquad (16)$$

Where *Len* (r_i) is the length of rule r_i.

Definition 16 Intensity considers both objective support and subjective interestingness. As for the rule r_i: $F \rightarrow h$, the intensity of it is:

$$Intensity\ (r_i)=[Interestingness\ (r_i)+support\ (r_i)]/2 \qquad (17)$$

Intensity considers both subjectivity and objectivity. On the one hand, as long as user is very interested in the rule even if the support is smaller, intensity will not be very small. In this case, this knowledge will be focused; On the other hand, the rule would be focused only when it has very high support if the user is not very interested in it.

As intensity contains support, we can focus on shortage knowledge in different hierarchy using support. Namely firstly focusing on shortage knowledge K_2 whose length is 2, then focusing on shortage knowledge K_3 whose length is 3, until the shortage knowledge whose length is L is empty, i.e., $K_L=\phi$ or the length is longer than the predefined maximal length M, i.e., $L>M$. K_2 can be produced directly from reachable matrix P (H). K_2 and the knowledge in knowledge base compose set K_2' ($\forall r_j \in K_2'$, *support* (r_j)>*min_sup*, where *min_sup* is threshold of the support threshold), and K_3 will be produced from K_2' using support. Since $\forall r_3 \in K_3$, the support of r_3 will not greater than that of the subset of r_3 certainly, i.e., *support* (r_3)≤*support* (r_2), where r_2 is the rule which is made up of two random primitive knowledge nodes in r_3. Since *support* (r_3)>*min_sup*, so *support* (r_2)> *min_sup*, as a result, $r_2 \in K_2'$. This character has the same qualities with maximum frequent itemset introduced in [17]. Therefore we can achieve shortage knowledge with this property.

Next, heuristic coordinator will automatically formed new focus in order to find new knowledge, in other words, producing "creating intent".

C. Maintenance Coordinating Algorithm and Maintenance Coordinator

The function of maintenance coordinator is simulating "psychology information maintenance" which is a character of cognitive psychology, in order to implement the real-time maintenance of knowledge base. Based on the exactly defined repeat, contradiction, redundancy, the maintenance coordinator participated in the process of KDD and handled repeat, contradict and redundant knowledge as early as possible in order to only evaluated the assumptions which possibly become new knowledge by making use of theoretical tools, such as hypergraph, so as to minimize the workload of evaluation; At the same time, the knowledge base can be maintained in real-time. In the actual expert system, the percentage of assumptions which maybe eventually become new knowledge to all original knowledge is very low and large number of assumption are repeat and redundant, so the introduced maintenance coordinator will improve the efficiency of KDD to a large extent. Here, we firstly give the definitions of the repeat, contradiction, redundancy of knowledge, then provide the maintenance coordinator algorithm.

Procedure Maintenance_Coordinator (R: $f_{i1} \wedge f_{i2} \wedge \ldots \wedge f_{is} \rightarrow$j)//len (R)=x

1) if R is repeat { $K_x = K_x - R$; return 0; }
2) if R is inconsistent {$K_x = K_x - R$; return 0; }
3) if R is redundant {$K_x = K_x - R$; return 0; }
4) return 1;

IV. CONCLUSION

In this paper, we propose a new knowledge discovery framework based on cognitive psychology features. We introduce what is the double bases cooperating mechanism and describes the knowledge base and its structure, database and its structure with this mechanism.

We also discuss mapping relationship between these two bases. Just based on this kind of mapping, we realize directional mining and directional searching, which lay foundation for the construction of the two coordinators and the new process model KDD*. Our paper shows our new frame will speed KDD mainstream development [18] in the following aspects:

1) This paper proposes the relation between knowledge base and database which are two important factors in the KDD process, constructs new structure model KDD* which reduces search and mining space, makes classical KDD development in an open and intelligent method.

2) A real-time maintenance mechanism of knowledge base is proposed. With the new knowledge increasing, system will check the repetition, redundancy, contradiction, subordination and circulation of the knowledge base

3) This paper represents the cognitive independence of KDD which is the key concept and research keynote, enhances the independence and intelligence of discovery.

4) With this mechanism, we can get highly scalable and efficient algorithm such as Maradbcm algorithm.

5) Bring many new thoughts in philosophy, and these thoughts will improve the development of KDD.

Finally, our work has been applied to agriculture, modern long-distance education, short-term meteorology, competitive intelligent system etc. under the supported by Natural Science Foundation and has made good effectiveness.

ACKNOWLEDGMENT

The authors would like to thank professor Yang Bingru, Zhou Zhun and Hou Wei. This work was supported in part by a grant from Jiangxi Higher school agricultural information technology key laboratories.

REFERENCES

[1] Yang Jun, Yang Bingru, Tang Zhigang, Fang Weiwei, "Coordinator Based on Cognitive Psychology Features and the Corrsponding KDD Process Model", 2008 Chinese Control and Decision Conference, vol.4, 2008, pp. 216-217.

[2] Jun Yang, YingLong Wang, "A New Outlier Detection Algorithms based on Markov chain*", *Advanced Materials Research*, vol. 366, 2012, pp. 456-459.

[3] Zhi-gang Tang, Jun Yang, Bingru Yang, "Spatial Outlier Dection Multiple Attributes Weighted", *PIAGENG. Image. Proce. Photonics*, vol.7489, 2010, pp. 748901-1-748901-9.

[4] Yang Jun, Xu Zhangyan, Yang Bingru, "Different Core Attributes Comparison and Analysis", *IEEE Intern. Confer. Granular Computing.* P676-681, 2009.

[5] Yang Bingru, Hou Wei, Zhou Zhun, Quan Huabin, "KAAPRO: An approach of protein secondary structure prediction based on KDD* in the compound pyramid prediction model", *Expert Systems with Applications*, vol. 36 (5), 2009, pp.9000-9006.

[6] Bingru Yang, Wei Hou, Yonghong Xie, "The research of protein secondary structure prediction system based on KDTICM", *Intern. Confer. Comput Biology*, 2009.

[7] D. Hand, H. Mannila, P. Smyth, "Principles of Data Mining", *MIT Press, Cambridge, CA*, 2001.

[8] G. Piatetsky-Shapiro, "Knowledge Discovery in Database: 10 Years After", *SIGKDD Explore*, vol.1 (2), 2000, pp. 59-61.

[9] J. Han, R. B. Altman, V. Kumar, "Emerging Scientific Applications in Data Mining", *Comm. ACM*, vol. 45 (8), 2002, pp.54-58.

[10] M. S. Chen, J. Han, P. S. Yu, "Data Mining: An Overview from a Database Perspective", *IEEE Trans. Know and Data Engi*, vol. 8 (6), 1996, pp.866-883.

[11] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[12] I. Bose, R. K. Mahapatra, "Business Data Mining-Machine Learning Perspective", *J. Inform .Manag*, vol. 39 (3), 2001, pp. 211-225.

[13] H. Mannila, "Theoretical Frameworks for Data Mining", *SIGKDD Explo,* vol. 1 (2), 2000, pp. 30-32.

[14] R. Coppi, "A Theoretical Framework for Data Mining: the "Informational Paradigm", *J. Comput Stati & Data Analysis*, vol. 38 (4), 2002, pp. 501-515.

[15] B. R. Yang, *Knowledge Discovery Based on Inner Mechanism: Construction, Realization and Application,* Elliott & Fitzpatrick Inc., USA, 2004.

[16] T. Eiter, G. Gottlob, K. Makino, "New Results on Monotone Dualization and Generating Hypergraph Transversals", *SIAM. J. Computing*, vol. 32 (2), 2003, pp.514-537.

[17] B. R. Yang, H. H. Sun, F. L. Xiong, "Mining Quantitative Association Rules With Standard SQL Queries and Its Evaluation", *J. Computer Research and Development*, vol. 39 (3), 2002, pp.307-312.

[18] B. R. Yang, J. Tang, "Research of Discovery Feature Sub-Space Model (DFSSM) Based on Complex Type Data", *Proce. Inter. Conf. Machine Learning and Cybernetics*, 2002.

Jun Yang 1970, He got PHD degree of data mining in University of Science and Technology Beijing, 2011, Beijing, China. The major field of study is data mining and knowledge engineering.

He works in School of Software, Jiangxi Agricultural University, Nanchang, China and do research in Jiangxi Higher school agricultural information technology key laboratories.

Yinglong Wang 1970. He got PHD degree of data mining in University of Science and Technology Beijing, 2011, Beijing, China. The major field of study is data mining and knowledge engineering.

He works in School of Software, Jiangxi Agricultural University, Nanchang, China.