Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 24 (2012) 240 – 246

Physics Procedia

2012 International Conference on Applied Physics and Industrial Engineering

# Extension Data Mining Knowledge Representation

Xie Guangqiang [1,2],Li Yang [1,2]

*1. Faculty of Computer Guangdong University of Technology*
*Guangzhou 510006 , China*
*2. Faculty of Automation*
*Guangdong University of Technology*

**Abstract**

This paper research on the representation of transformable knowledge from extension data mining. Traditional data mining technology obtain static knowledge, on the contrary, extension data mining obtain transformable knowledge, which widening the source of knowledge needed in extension strategy generating system. Transformable knowledge representation using finite automata provide us a new method of extension data mining knowledge representation.



*Keywords:*Extension data mining; extension knowledge; knowledge representation

## 1. Introduction

The development of extension data mining technology offer a new way to obtain knowledge which extension strategy generating system required. Transformable implication knowledge getting by extension data mining technology is different from the knowledge getting by traditional method. It is dynamic, transformable knowledge. The representation and storage of this kind of knowledge is mainly discussed in this paper.

## 2. Extension transformation is a basic approach of extension strategy generation

Extenics research on the theory and method solving contradiction problem[1].Extension strategy generating is a technology of decision-making for contradiction problem. Extension methods are needed to analysis and solve contradiction problem, in these methods, extension transformation is a basic technology


doi:10.1016/j.phpro.2012.02.036

to generate strategies[2]. Extension transformation base on rule and knowledge, which are commonly derived from expert knowledge and extension data mining.

### *2.1 Basic step of extension strategy generation*

Extension strategy generation technology is based on to rhombus thinking method[3], which combine divergent with convergent. Divergent method base on existing data and knowledge, utilize base-element's feature of extensible, generate a certain amount of contradiction problem's solution transformation; Convergent method use dependent function to judge the validity of transformation above, then select some validate transformation as good strategy by advantageous degree appraisal method. The main step of building Extension Strategy Generating System (ESGS as follows) will be introduced below.

- Express non-compatible problem, establish dependent function
- Make extension analysis to non-compatible problem
- Make extension transformations to goal g and condition l, generating strategies to be selected.
- Calculate the advantageous level of the strategies and select the better ones.

### *2.2 Knowledge acquisition*

It needs a large amount of knowledge relative with contradictory problem to generate extension strategy. Basing on the method of acquiring knowledge and formulae, there are tow ways for the research of ESGS[2], one is based on expert system, the other is based on extension data mining method.

ESGS basing on expert system use expert knowledge to establish "rule warehouse", then make extension transformation to contradictory problem basing on "rule warehouse" in the way of man-machine conversation, strategies which can solve problem are generated.

ESGS basing on extension data mining mines acquire extension knowledge from base database, to contradictory problem, it make extension transformation using extension knowledge to generate strategies which can solve problem.

Comparing two methods of acquiring knowledge above, knowledge from expert system has the shortage of inflexibility, on the contrary, knowledge from extension data mining is closer to problem, more convenient to generate efficient strategy.

## 3. Knowledge acquired from extension data mining is extension knowledge

Data mining refer to the procedure of extracting useful information and knowledge from a large amount of structured and unstructured data, it's an effective means of knowledge discovery[5]. Due to data's static nature which stand for the existing fact, knowledge mined from it also has static nature[6]. On the other side, extension data mining acquire extension knowledge.

### *3.1 Data mining method*

The object of data mining can be structured data, such as relative database and data warehouse, as well as half-structured data and even unstructured text, picture, video, Web data etc.

Data mining consist of following steps[7]: subject mining, data preprocessing, mining algorithm selection, data mining, result showing, evaluation.

The major methods of data mining are[6]: inductive learning, biotechnology imitation, formula detection, statistical analysis, fuzzy mathematics, visualization technology etc.

Knowledge acquired through above methods are static knowledge, however, it needs extension transformation to solute contradictory problem, so variable knowledge are needed, extension data mining provide a new method to mine extension knowledge.

### *3.2 Extension data mining*

Comparing to traditional data mining, extension transformation is added to acquire variable knowledge in extension data mining, providing extension knowledge to solve contradictory problem. There are two kinds of knowledge acquired from extension data mining:

- Interval information of dependent function

Taking the case of primary dependent function which middle point in interval is the most advantageous point, basic formula is[8]:

$$k(x)=\begin{cases}\dfrac{\rho(x,X_0)}{D(x,X_0,X)}-1, \rho(x,X_0)=\rho(x,X\ ) \text{ and } x\notin X_0 \\ \dfrac{\rho(x,X_0)}{D(x,X_0,X)}, \qquad \text{other conditions}\end{cases} \qquad X_0=\langle a,b\rangle \ , \ X=\langle c,d\rangle \ , \text{ and } X_0\subset X$$

inside, call is primary dependent function of point x in and which middle point in interval is the most advantageous point.

To establish appropriate dependent function, it is necessary to determine the appropriate values of four interval parameter: a, b, c, d. To acquire the information of interval parameter through data mining method is an important task of extension data mining.

- Implicit transformation formula mining

Existing knowledge: condition c -> conclusion r, make extension and conductible transformation separately to c and r, acquired variable knowledge, which is extension knowledge:

$T_c \Rightarrow T_r$ .

There are two theorems of implicit transformation knowledge mining and extension data mining as follows[9]:

THEAREM 1. To two sort of rules $A \Rightarrow P$ , $B \Rightarrow N$ , $A=\wedge a_i$ , $B=\wedge b_j$ , under regular situation, if condition extension transformation $T_c$ exist, $T_c(B)=A$ , and conclusion extension transformation $T_r$ ( conductible transformation of $T_c$ ) exist, $T_r(N)=P$ , then extension transformation rule knowledge(variable knowledge) is tenable.

$$(T_c(B)=A \Rightarrow T_r(N)=P) \qquad (1)$$

That means if $T_c(B)=A$ then $T_r(N)=P$ .

THEAREM 2. To two sort of congener rules $A \Rightarrow P$ , $C \wedge B \Rightarrow P$ , if extension transformation $T_c(B)=A$ exist, then extension transformation rule knowledge is tenable.

$$(T_c(B)=A \Rightarrow P) \qquad (2)$$

That means If $T_c(B)=A$ then $P$ .

## 4. Knowledge representation

As knowledge from extension data mining are variable, it is necessary to solve the problem of variable knowledge representation before research deeply on the technology of extension data mining, on the base of knowledge representation, we can research on the corresponding arithmetic and its realization on computer.

Taking one corporation's notebook computer sales data as example, we will describe the representation of static and variable knowledge in TABLE 1.

TABLE I. A CORPORATION'S NOTEBOOK COMPUTER SALES DATA

| Purchase notebook computer | Age | Salary | If teacher | City or country |
|---|---|---|---|---|
| purchase | 26-35 | high | no | country |
| | 31-35 | middle | no | city |
| | 31-35 | high | yes | country |
| | >35 | middle | no | country |
| | >35 | low | yes | country |
| | 31-35 | low | yes | city |
| | ≤25 | low | yes | country |
| | >35 | middle | yes | country |
| | ≤25 | middle | yes | city |
| | >35 | low | yes | city |
| No purchase | ≤25 | middle | no | country |
| | ≤25 | high | no | country |
| | ≤25 | high | no | city |
| | >35 | middle | no | city |

### *4.1 Data mining knowledge representation*

There are mainly six forms of knowledge from data mining[10]:

- Rule

Rule knowledge consist of precondition and conclusion, precondition compose of "and" and "or" operation on the value of field term(property), conclusion compose of the value or type of decision making field term(property).

Taking table 1 as an example, we can obtain the following rule knowledge by the methods of data mining.

IF（age=“26-35”）THEN purchase
IF（age=“ 25”） （teacher=“yes”）
THEN purchase
IF（age=“ 25”） （teacher=“no”）
THEN no purchase
IF（age=“>35”）（city or country=“city”）THEN purchase
IF（age=“>35”）（city or country =“country”）
THEN no purchase

- Decision tree

Decision tree is a kind of treelike graph to indicate people's series of judgment procedure in order to make some decision. Decision tree consists of decision nodes, branches and leaves. The top node is root which is the start of the decision tree. In the course of searching along the decision tree from top to bottom, there will be a problem in each node, different answers will lead to different branches, it will reach a leaf at last. Each leave belong to a different classification[5]. Figure 1 is a decision tree corresponding with TABLE 1.

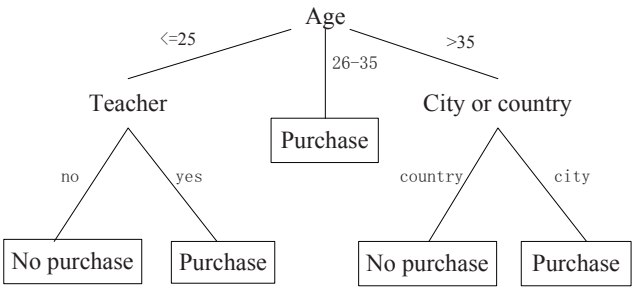

Figure 1. Decision tree

- Knowledge base (concentrated data)

Through calculating important degree of the field term in database, we can delete some unimportant fields. Merge records in database according to definite principle. Acquire concentrated data called knowledge base from condensed database.

- Network weight

Neural network method is through the way of studying training sample to acquire knowledge of network connection weight and node threshold, usually expressed as matrix and vector.

- Formula

In science and engineering database, large amount of experiment data are usually stored, regularity always implied at there. Through formula discovery algorithm, we can find the variables' correlations and express them by formula.

- Case

Case is a complete event people experienced. we can utilize the solutions to problem or results of its processing in past cases as a reference to revise properly to solve some new problem. Case knowledge are expressed as triple:

<problem description, solution description, effects description>

### *4.2 Variable knowledge representation from extension data mining*

Knowledge representation methods introduced above don't suit for representing variable knowledge from extension data mining. Variable knowledge representation will be introduced below. According to theorem 1 and 2, suppose the following extension data mining knowledge exist:

$$\text{IF}\,(A = a_0)\ \text{and}\,T(B = b_1) = (B = b_2)\,\text{THEN}\,T(N) = P \quad (3)$$

$$\text{IF}\,(B = b_0)\,\text{and}\,T(C = c_1) = (C = c_2)\ \text{THEN}\,T(P) = N \quad (4)$$

In formula (3) (4), A, B, C stand for attribute, stand for the corresponding attribute value, P，N stand for different classification. We use finite automata which has one initial state and many terminal states to represent the variable rules above:

- Generating one initial state and many terminal states, each terminal state represents a classification or transition between classifications.
- Each node of finite automata represents an attribute value.
- Each edge of finite automata represent precondition, relation of conditions or transformation. Edges which are from and to the same node represent precondition of rules. "AND"or "OR" represents relation of the conditions. "T" represents transformation.
- In finite automata, a path from initial state to terminal state represents a piece of variable knowledge.Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".

Take formula (3) (4) as example, we will give the variable knowledge representation as Figure 2.

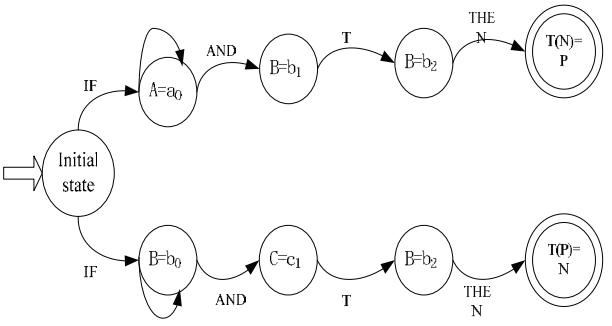

Figure 2. Finite automata represent variable knowledge

## 5. Variable knowledge representation example

Knowledge obtained by data mining method to the data of table 1 can be made extension data mining basing on theorem 1,2, the following variable knowledge is acquired, we only list two transformations which decision makers care: "T(no purchase)=purchase" and "T(purchase)=no purchase". Variable knowledge representation can be seen in figure 3.

IF (age=25) AND T(teacher="yes")=(teacher="no")
THEN T(purchase)=no purchase
IF (age>35) AND T(city or country="city")=(city or country="country")
THEN T(purchase)=no purchase
IF (city or country="country") AND T(age>35)= (age=26-35)
THEN T(no purchase)=purchase
IF (teacher="no") AND T(age=25)=(age=26-35)
THEN T(no purchase)=purchase
IF (age=25) AND T(teacher="no")=(teacher="yes")
THEN T(no purchase)=purchase
IF (age>35) AND T(city or country="country")=(city or country="city")
THEN T(no purchase)=purchase

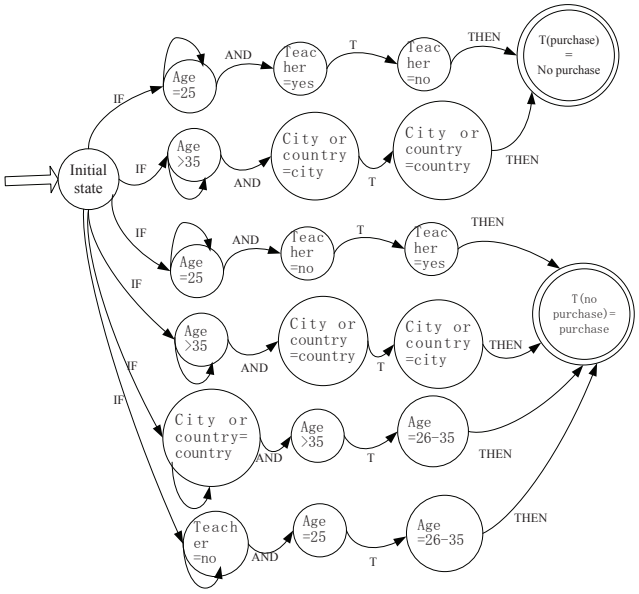


Figure 3. Variable knowledge representation of purchasing notebook computer

Using finite automata for the representation of variable knowledge has the virtue of being flexible and easy to manage with computer. It is easy to construct various extension data mining and extension strategy generating arithmetic in this form.

## Acknowledgment

This paper is supported by the National Natural Science Foundation of China under Grant No.70671031 and the Natural Science Foundation of Guangdong Province under Item No.10151009001000044.

## References

[1] W. Cai, C.Y. Yang and B. He, Foundation of Extension Logic. Beijing: Science Press, 2003, in Chinese: 1

[2] L.X. Li, C.Y. Yang and H.W. Li, Extension Strategy generating System. Beijing: Science Press, 2006, in Chinese: 7,208

[3] W. Cai, C.Y. Yang and W.C. Lin, Extension Engineering Method. Beijing: Science Press, 2000, in Chinese

[4] Y. Li, C.Y. Yang and L.X. Li, Design and implement the analysis and solving of enterprise resources problems. Journal of Harbin Institute of Technology, 2006，38(7)：1195-1198

[5] X.N. Su, J.L. Yang and X. Li, Data Warehouse and Data Mining. Beijing: Tsinghua University Press, in Chinese,2006：115,162

[6] W.W. Chen, Data Warehouse and Data Mining Course. Beijing: Tsinghua University Press, in Chinese,2006：272，126

[7] J.Q. Xu, Data Warehouse and Decision Support System. Beijing: Science Press, in Chinese, 2005:69

[8] C.Y. Yang and W. Cai. Extension Engineering. Beijing: Science Press, 2007, in Chinese: 86

[9] W.W. Chen, C.Y. Yang and J.C.Huang, Extension Knowledge and Extension Knowledge Reasoning. Journal of Harbin Institute of Technology, 2006，38(7)：1094-1096

[10] W.W. Chen, Decision Support System Course. Beijing: Tsinghua University Press, in Chinese,2004：212-214