小 型 微 型 计 算 机 系 统 2014 年 1 月 第 1 期
Journal of Chinese Computer Systems Vol. 35 No. 1 2014

一种基于树图的属性约简算法

张 涛,路 静,任宏雷

(燕山大学 信息科学与工程学院,河北 秦皇岛 066004)
E-mail: lujinggc@163.com

摘 要: 针对基于概念格对数据的规则挖掘中,概念格结构的复杂度随着形式背景的复杂化呈指数递增的问题,本文从树图的角度研究决策形式背景的属性约简问题,以树型结构为基础,首先提出了新的强弱背景的判定方法;进一步定义了条件树对象集与决策外延的相关函数,并以此作为启发信息,设计了基于树图的逐层属性约简算法,避免了计算区分矩阵这个既消耗时间又消耗空间的过程;最后在属性约简树的基础上,给出了约简后的概念树生成算法,以简洁、直观的概念树的代替错综复杂的概念格.理论分析和实验结果表明,该算法是有效可行的.

关 键 词: 协调决策背景;属性约简;树图;决策外延;概念树
中图分类号: TP18 **文献标识码**: A **文 章 编 号**: 1000-1220(2014)01-0177-04

An Algorithm Based on Tree Graph for Calculation for Attribute Reduction

ZHANG Tao,LU Jing,REN Hong-lei
(College of Information Science and Engineering, Yanshan University,Qinhuangdao 066004, China)

Abstract: Based on lattice in rules mining, to address the problem that the complexity of lattice structure increase exponentially with the complicated of data. Attribute reduction in formal context is proposed in terms of tree graph. Firstly, based on the structure of tree, a new method of judging the context that strong and weak is proposed, on the basis of that, define a related function between the node of condition tree and the decision concept, then the formula measuring node importance is used as heuristic information to design an efficient attribute reduction algorithm based on the tree graph, and avoid the process of computing discernibility matrix with time-consuming and space-consuming; finally, an algorithm to generate concept tree after reduction is presented, so the concise, intuitive concept tree replaces the concept lattice. Theoretical analysis and experimental results show that the algorithms of this paper are efficient and effective.

Key words: consistent decision formal context; attribute reduction; tree graph; decision extension; concept trees

1 引 言

形式概念分析,是由 Wille 提出来用于分析和处理数据的重要数学工具[1].目前已经成功应用到信息获取[2,3]、机器学习[4]、信息检索[5,6]、软件工程[7]等领域.

属性约简是形式概念分析的一个重要研究内容,属性约简的目的就是删除其中不相关或不重要的属性,在保证知识表示简化的同时而又不丢失基本信息.对决策形式背景进行属性约简可以简化决策形式背景的信息量,从而提高决策效率,这在理论和应用上具有重要意义.因此针对属性约简,目前已经得到很多研究成果.王虹等提出了协调决策形式背景的属性约简[8];张文修等提出的决策形式背景的概念格属性约简[9],分析并提出了有关强、弱协调决策形式背景的属性约简理论;针对区间类型的形式背景,Jinhai Li,et 提出的 Knowledge reduction in real decision formal contexts[10],给出了非典型(非布尔式)决策背景下属性约简的一种方案.以上方法都是围绕着区分矩阵和布尔函数进行的,然而区分矩阵是一个以概念数目为阶数的矩阵,因此对于大数据集,其区分矩阵庞大,可辨识函数过于繁杂,难以化简.很难实现对大型数据库的处理;另外,这些算法都是依赖于概念格,而概念格结构的复杂度随着形式背景的复杂化呈指数递增[11],直接影响到了大规模数据下的应用.

针对以上问题,文献[11]给出了形式背景的属性树表示方法,将错综复杂的概念格转化为属性树,即仅对形式背景的树状分析,并未考虑形式背景中的冗余属性,不能实现属性的约简.本文针对决策形式背景,将属性树延伸成概念树,提出了一种基于树图的属性约简算法.该方法在属性约简的过程中逐步生成形式背景的概念,并在属性约简树的基础上,得到约简后的概念树.一方面,属性约简过程与概念生成过程几乎同步进行,另一方面,属性约简不需要区分矩阵,大大节省了存储空间.因此提高了形式概念分析处理数据的效率.

2 基本理论知识

形式背景是形式概念的研究对象,也是形式概念分析的

收稿日期:2012-07-16 基金项目:国家自然科学基金项目(60904100)资助;河北省自然科学基金项目(F2011203073)资助 作者简介:张涛,男,1979 年生,讲师,研究方向为网格计算、形式概念分析;路 静,女,1985 年生,硕士,研究方向为粗糙集、形式概念分析;任宏雷,女,1989 年生,硕士,研究方向为粗糙集、形式概念分析.

数据表示方式. 形式背景可以表示为(U,A,I)，其中，U是有限对象集，A是有限属性集，$I \subseteq U \times A$表示两者之间的一个关系集合. 对于集合$P \subseteq A$，记$M^* = \{u \in U \mid (m,u) \in I, \forall m \in P\}$，相应的，对于集合$Q \subseteq U$，记$Q^* = \{a \in A \mid (q,a) \in I, \forall q \in Q\}$.

定义1. [8]：如果还存在一个背景(U,T,J)，与(U,A,I)具有相同的论域，则称(U,A,I,T,J)是决策形式背景.

定义2. [8]：在决策形式背景中，若存在条件形式背景概念格$L(U,A,I)$到目标形式背景概念格$L(U,T,J)$的一个蕴含关系f，即$\forall (Y^T,M) \in L(U,T,J)$总存在$Y \subseteq Y_T$，使得$(Y,B) \in L^f(U,A,I)$，则为协调背景. 协调背景分为两类：如果满足所有$Y = Y_T$时为强协调；否则为弱协调. Y为决策外延.

对于一个形式背景，可能包含重复信息，为了简化运算，首先进行背景净化，即合并具有相同内涵的对象，合并具有相同内涵的对象.

3 逐层约简分析

3.1 树图及相关性质

属性约简是形式概念分析研究的基本内容. 与传统的区分矩阵属性约简算法不同，本文基于树图提出了逐层属性约简算法. 在以上基本理论的基础上，从树图结构出发，对树图进行了如下定义：

定义3. 树图的根节点为(U,U^*)，树图的第一层节点为(a^*,a)，$\forall a \in A$.

性质1. 如果根节点$U^* \neq \varnothing$，则属性U^*必为冗余属性.

性质2. 树图的第一层节点不一定为概念.

证明：根据概念的含义以及树图定义，显而易见.

定义4. 已知$\forall (Y_T,M) \in L(U,T,J)$，若$Y = a_i^* \cap a_j^*$，$a_i, a_j \in A$，且$Y \subseteq Y^T$，$Y$为决策外延.

定义5. 已知条件背景树图的节点(X,B)，决策外延为Y，则X，Y二者的相关函数为：

$$R(X_k,Y_j) = \sum_{j=1}^{\#\{Y\}} \frac{\#\{Y_j\}}{\#\{Y\}} \quad X_k \supseteq Y_j \tag{1}$$

式(1)中$\#\{Y\}$表示决策外延Y的个数

定义6. 在树图第一层节点中(X_i,a)，当$\exists X_i = Y_j$时，即$R_1(X_k,Y_j) = 1$那么a为辅助属性.

定义7. 设节点(X_k,B)，当$X_k \supset Y_j$时，且$R(X_i,Y_j)_{\max} = R(X_k,Y_j)$，则该节点为父节点.

性质3. 节点(X_k,B)为父节点，则$\#\{Y_j \mid X_k \supset Y_j\}$表示该节点下的子节点的个数，且子节点为$(Y_j,Y_j^*)$，$\{Y_j^*\} \subseteq A$.

性质4. 在树图中，除第一层节点外，其他层的节点均为概念.

证明：由定义1知：根节点为(U,U^*)，且$\{U^*\} = U$，综合性质2和性质3，可证.

定义8. 包含除了$(\varnothing,T)$全部概念的树图为概念树.

性质5. 在树图中，同层任意两节点(X_i,B_i)，(X_j,B_j)满足$X_i \subset X_j = \varnothing$，即不再产生新的子节点的树图称为概念树.

3.2 强弱协调背景判定

定理1. 假设条件背景的树图第一层节点(X,B)，$\forall (Y^T,M) \in L(U,T,J)$，若$\exists Y_k^T \neq X_i \cap X_j (i,j = 1,2,\cdots,\#\{X\})$，则为弱协调，反之为强协调.

证明：依据协调背景定义及树图定义1，显然，如果第一层节点中$\exists X_i \cap X_j \neq Y_k^T$，那么该背景必定是弱协调的；否则，$\forall Y = X$，必为强协调的.

对于强协调背景，根据Y^T易得决策外延$Y = Y^T$.

3.3 逐层属性约简

所谓属性约简，就是在保持决策背景的分类能力不变的情况下，删除冗余属性，且保证约简中的任一属性都是必要属性，所以约简的关键在于删除冗余属性，那么基于树图如何判定属性，以下定理和推论给出了属性逐层判定的依据：

定义9. [8]：在协调决策背景(U,A,I,T,J)中，f：$L(U,T,J) \to L(U,A,I)$为蕴含映射，若存在属性集$R \subseteq A$，使得$L_U^f(U,A,I) \subseteq L_U(U,R,I_R)$，如果对于$\forall r \in R$，不满足$L_U^f(U,A,I) \subseteq L_U(U,R-\{r\},I_{R-\{r\}})$，即$(U,R-\{r\},I_{R-\{r\}}T,J)$为不协调的，则称$R$为$(U,A,I,T,J)$关于$f$的属性约简.

定理2. 辅助属性必为必要属性.

证明：反证法，假设辅助属性节点为(X,a)，a为冗余属性，即$a \notin R$，根据定义9，满足$L_U^f(U,A,I) \subseteq L_U(U,R,I_R)$，根据定义4：$X = Y$，则有$(X,B) \in L_U^f(U,A,I)$，则必须存在$(X,B) \in L_U(U,R,I_R)$且$(B \in R)$，由概念含义得出$B^* = X$，针对树图节点定义，$a^* = X$，因此$a \subseteq B$，则$a \in R$，与假设矛盾，假设不成立，所以$a$必不可少.

推论1. 不是所有决策形式背景都存在辅助属性，且辅助属性一定存在于第一层节点中.

定理3. 已知一个父节点(X_1,B_1)，同层节点(X_2,B_2)，以及决策外延Y，其中，$X_2 \not\subset Y_j$，下一层节点$(X_1 \cap X_2, B)$，且$X_1 \cap X_2 = Y_k$

i：若$B_1 \cup B_2 = B$，则B为必要属性.

ii：若$\forall (Y_j,M_j) \subseteq L^f(U,A,I)(j \neq k)$，$B_1 \cup B_2 \subset B$，$B_3 = C_B(B_1 \cup B_2)$，$B_3 \cap M_j = \varnothing$，则$B_3$为冗余属性.

i：证明：由于$B_1 \cup B_2 = B$，那么$(B_1 \cup B_2)^* = X_1 \cap X_2$，$(X_1 \cap X_2)^* = B_1 \cup B_2$，且$X_1 \cap X_2 = Y_k$，$X_1$，$X_2 \not\subset Y_j$，即$(Y_k,B) \subseteq L_U^f(U,A,I)$，如果去掉$B$不满足$L_U^f(U,A,I) \subseteq L_U(U,R-\{B\},I_{R-\{B\}})$，因此，$B$为必要属性.

ii：证明：假设B_3为必要属性，即$B_3 \in R$，满足$(U,R-\{B_3\},I_{R-\{B3\}},T,J)$为不协调的，由于$\forall (Y_j,M_j)(j \neq k)$，$B_3 \cap M_j = \varnothing$，这就意味着$B_3$仅对$Y_k$类的划分起到了作用，即$B_3$的存在与否不影响$Y_k$以外类别的正确划分，即$\forall (Y_j,M_j) \in L_U^f(U,A-\{B_3\},I_{A-\{B3\}})$，此外，由已知条件$X_1 \cap X_2 = Y_k$，$(B-B_3)^* = \{X_1 \cap X_2\}$，则$(X_1 \cap X_2, B-B_3) \in L_U^f(U,A-\{B_3\},I_{A-\{B3\}})$因此，由定义9得，$(U,R-\{B_3\},I_{R-\{B3\}},T,J)$为协调的. 与假设相悖，因此假设不成立，$B_3$为冗余属性.

4 属性约简算法

基于树图结构的特征，本文将属性约简过程以树图的形式直观展现，该过程从根节点开始划分层次，依据条件背景树图节点的对象集X与决策外延Y的相关性，不断地添加属性对节点进行划分，直到满足(Y,B)均为树图的节点为止.

已知一致协调决策背景(U,A,I,T,J)，$A = \{a_1,a_2,\cdots,a_v\}$，$\forall (Y_T,M) \in L(U,T,J)$首先依据强弱背景判定方法，判

断背景的强弱性，然后得出决策外延 Y. 图 1 表示树图生成的流程图.

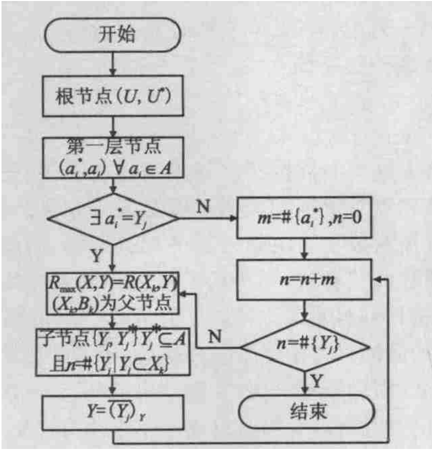


图 1　树图流程图

Fig. 1　Flowcharts of tree graph

算法 1. 属性约简算法

输入: 一致协调决策背景(U,A,I,T,J)

输出: 属性约简 R 和约简后的树图

Step 1. 净化背景，依据图 1 生成形式背景的树图;

Step 2. 判断第 0 层节点(U,U^*)，如果 $U^* \neq \varnothing$，则 U^* 为冗余属性，该节点改为$(U,\varnothing)$.

Step 3. 判断第一层的节点，如果 $\exists R_1(a^*,Y_j) > 1$，则 a 为辅助属性;

Step 4. 判断下一层节点，若子节点为$(X_k \cap a^*, B_k \cup a)$，$B_k$，$a$ 必要属性; 若为$(X_k \cap X_p, B)$，且 $B = B_p \cup B_k$，则 $B_p \cup B_k$ 为必要属性，否则转入 Step 5;

Step 5. 若 $B \supset B_p \cup B_k$，则 $\overline{(B_p \cup B_k)}_B$ 为冗余属性，若节点 $\exists(X_r, \overline{(B_p \cup B_k)}_B \cup B')$，当 $B' = \varnothing$时，删除该节点，否则该节点仅删除冗余属性，即为(X_r, B');

Step 6. 输出属性约简 R，属性约简树图 T.

算法 2. 生成约简后的概念树

输入: 由算法 1 生成的属性约简树图 T

输出: 约简后的概念树 $Tree(U,T,J)$

Step 1. 树图第一层中的父节点 $\forall(X_j, B_j)$，对应子节点(X_j', B_j')，如果满足$(\cup X_j')_{X_j} = X_m$ 且 $X_m \neq \varnothing$，则父节点继续划分节点为(X_m, X_m^*). 否则转入 Step 3;

Step 2. 判断树图中的同层其他节点，任意两节点(X_i, B_I)，(X_j, B_j)，若 $X_i \cap X_j = \varnothing$，输出树图 T'，转入 Step 4; 否则转入 Step 5;

Step 3. 任选 X_i，X_k 节点之一作为父节点继续划分树图，且$(X_i \cap X_k, \{X_i \cap X_k\}^*)$ 作为其子节点，转入 Step 3;

Step 4. $\forall(X_z, B_z)$ 为新生成的终端节点，向上遍历，父节点(X_f, B_f)，若 $\exists X_z = X_f, B_z \supset B_f$，删除父节点; 若 $\exists B_z = B_f, X_z \subset X_f$，删除子节点; 输出概念树 $Tree(U,T,J)$.

5　实验分析

为了说明该算法的有效性，本节分别设计了两组决策形式背景进行试验: 一组为强协调决策背景; 另一组为弱协调决策背景. 从实验角度验证该方法适用于协调决策背景. 表 1 为强协调决策背景.

表 1　强协调决策背景

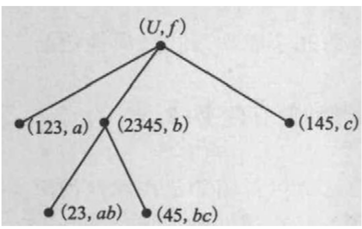
Table 1　Strong consistent decision context

U	a	b	c	d	e	f	g	h	i
1	1	0	1	0	1	1	0	0	1
2	1	1	0	1	1	1	1	0	0
3	1	1	0	0	1	1	1	0	0
4	0	1	1	1	0	1	0	1	1
5	0	1	1	1	0	1	0	1	1

根据强弱协调背景判定方法，该决策背景为强协调，即 $Y = Y_T$，则蕴含关系如下:

$\{((145, h, (145, 145^*)); ((23, f)(23, 23^*)); ((45, gh), (45, 45^*))\}$

由图 1 可以得出，根节点(U,f)，根据性质 1，属性 f 为冗余属性; 第一层中的节点$(145,c)$的外延与决策外延$\{145\}$等价，根据定理 1，可得 c 为辅助属性，即 $c \in R$; 而对于第二层节点$(2345,b)$，$R(X,Y)_{max} = R(\{2345\},Y) = 0.5$，首先，$\{2345\} \cap \{145\} = \{45\}$，$\{b \cup c\} \in \{bcd\}$，且 $d \cap \{bc\} = \varnothing$，则根据定理 2(i)可得，$d$ 为冗余属性; 另外$\{2345\} \cap \{23\} = \{23\}$，$(23, ab) = (123 \cap 2345, a \cup b)$，根据定理 2(ii)可得 ab 为必要属性. 此时，满足 $\forall(Y,B) = L^f(U,A,I)$，$(Y,B)$ 均为树图的节点. 因此该决策背景的属性约简集为 $R = \{a,b,c\}$.



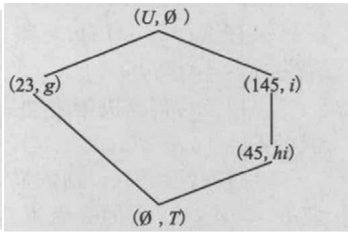


图 2　表 1 的约简树　　图 3　表 1 的 $L(U,T,J)$

Fig. 2　Reduction tree of table 1　Fig. 3　$L(U,T,J)$ about table 1

经过属性约简后，则依据算法 2 可得图 3，包含该条件背景的所有概念:

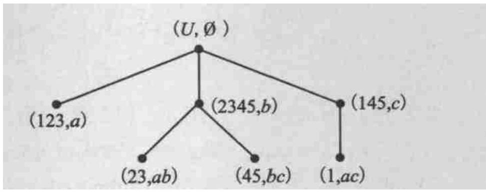


图 4　表 1 约简后的概念树

Fig. 4　Concept tree of table 1 after reduction

表 2　弱协调决策背景

Table 2　Weak consistent decision context

U	a	b	c	d	e	f	g
1	1	0	0	0	1	0	0
2	1	0	0	1	1	1	0
3	1	1	0	0	1	1	0
4	0	1	1	1	0	1	1
5	0	0	1	1	0	1	1

表 2 为弱协调决策背景.

根据强弱协调背景判定方法，该决策背景为强协调，即 Y

$\subset Y_T$，则蕴含关系如下：

$\{((123,e),(123,123^*)),((2345,f),(34,34^*)),((23,ef),(3,3^*)),((45,g),(4,4^*))\}$

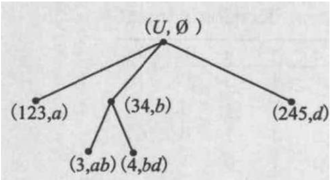


图 5 表 2 的约简树

Fig. 5 Reduction tree of table 2

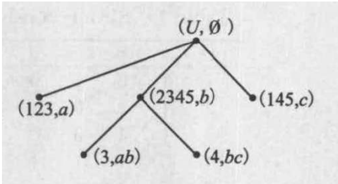


图 6 表 2 的另一约简树

Fig. 6 Another reduction tree of table 2

由图 2 可得，第一层中的节点 $(34,b)$ 与决策背景中概念 $(2345,f)$ 为映射关系，$(123,a)$ 与 $L(U,T,J)$ 中概念 $(123,e)$ 也为映射关系，根据定理 1，可得 a,b 均为辅助属性，即 $a,b \in R$；在第二层中节点中，$R(X,Y)_{\max} = R(\{34,Y\}) = 0.5 + 0.5 = 1(Y_j \subset X_i)$，首先，$\{34\} \cap \{123\} = \{3\}$，$(3,ab) = (34 \cap 123, a \cup b)$，根据定理 2(ii) 可得 ab 为必要属性；

由于 $R(45,c) = R(145,d) = 0.25$，当选节点 $(45,c)$ 时：$\{34\} \cap \{45\} = \{4\}$，$4^* = bcd$，而 $\{(34)^* \cup (45)^*\} = \{b \cup c\} \subseteq \{bcd\}$ 且 $d \cap \{ab\} = \varnothing$，则根据定理 2(i) 可得，$d$ 为冗余属性，此时，满足 $\forall (Y,B) \in L^f(U,A,I)$，即 (Y,B) 均为树图的节点，属性约简集为 $R_1 = \{a,b,c\}$；

同理选 $(145,d)$ 时，可得 c 为冗余属性，此时，属性约简集为 $R_2 = \{a,b,d\}$.

由以上可得该决策表的属性约简有两个：$R_1 = \{a,b,c\}$ 或者 $R_2 = \{a,b,d\}$.

经过属性约简后，则依据算法 2 可得图 7 是在属性约简集 $R_1 = \{a,b,c\}$ 下的概念树；图 8 是在属性约简 $R_2 = \{a,b,d\}$ 下的概念树.

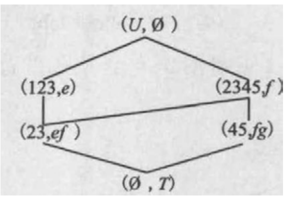


图 7 表 2 的 $L(U,T,J)$

Fig. 7 $L(U,T,J)$ about table 2

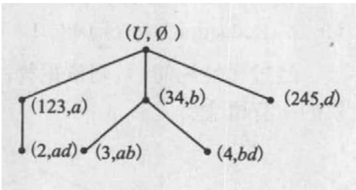


图 8 表 2 约简后的概念树

Fig. 8 Concept tree of table 2 after reduction

算法正确性分析：

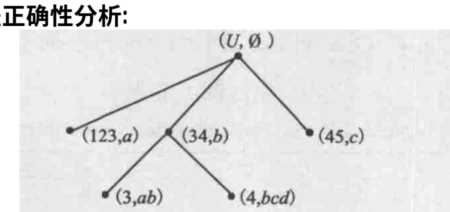


图 9 表 2 约简后的概念树

Fig. 9 Another concept tree of table 2 after reduction

树图生成算法中选择相关函数最大的作为父节点，如果 $R(X_k,Y_j)$ 越大，说明 X_k 包含 Y(决策类别) 越多，意味着在该节点的基础上可区分的 Y 越多，即实现用最少属性区分最多类别，因此是可行的；而在属性约简过程中，依据逐层属性约简算法，删除了树图节点 (X,B) 的属性集 B 中的冗余属性，同时保证了 $L_U^f(U,A,I) \subseteq L_U(U,R,I_R)$，因此，依据本文算法得到的 R 即为属性约简集.

6 结 论

在进行大规模的数据挖掘与规则提取中，决策形式背景中大量冗余属性常常制约着形式概念分析方法的应用. 本文引入了条件树对象集与决策外延的相关函数，提出了基于树图的快速属性约简算法. 与传统的区分矩阵属性约简方法比较，如果只进行属性约简，本文方法省去了求取概念格的过程；并且在约简的过程中逐步生成概念，最后在属性约简树的基础上得到约简后的概念树. 实验结果表明，新算法一般都能得到比较理想的属性约简集，且比布尔推理方法更有效.

References:

[1] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts [M]. Dordrecht: Reidel, 1982: 445-470.

[2] Belohlavek R. A note on variable threshold concept lattices: threshold-based operators are reducible to classical concept-forming operators [J]. Information Sciences, 2007, 177(15): 3186-3191.

[3] Wu W Z, Leung Y, Mi J S. Granular computing and knowledge discovery in formal contexts [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(10): 1461-1474.

[4] Liu Xue-long, Hong Wen-xue. Using formal concept analysis to visualize relationships of syndromes in traditional Chinese medicine [J]. Lecture Notes in Computer Science, 2010, 61(65): 315-324.

[5] Chen Rung-Ching, Cho-Tscan Bau, Chun-Ju Yeh. Merging domain ontologies based on the word net system and fuzzy formal concept analysis techniques [J]. Applied Soft Computing, 2011, 11(2): 1908-1923.

[6] Tam T Nguyen, Siu Cheung Hui, Kuiyu Chang. A lattice based approach for mathematical search using formal concept analysis [J]. Expert Systems with Applications, 2012, 39(5): 5820-5828.

[7] Wang Li-dong, Liu Xiao-dong, Cao Jian-nong. A new algebraic structure for formal concept analysis [J]. Information Sciences, 2010, 180(24): 4865-4876.

[8] Wang Hong, Wan Jin-feng. Attribute reduction in consistent decision formal context [J]. Chinese Journal of Engineering Mathematics, 2006, 23(3): 455-460.

[9] Wei Ling, Qi Jian-jun, Zhang Wen-xiu. Attribute reduction theory of concept lattice based on decision formal contexts [J]. Science in China: Series F-Information Sciences, 2008, 38(2): 195-208.

[10] Li Jin-hai, Mei Chang-lin, Lv Yue-jin. Knowledge reduction in real decision forma lcontexts [J]. Knowledge-Based Systems, 2011, 24(5): 709-715.

[11] Zhang Tao, Hong Wen-xue, Lu Jing. Attribute tree representation for formal context [J]. System Engineering Theory and Practice, 2011, 31(2): 197-202.

附中文参考文献：

[8] 王 虹，万金凤. 协调决策形式背景的属性约简 [J]. 工程数学学报，2006，23(3)：455-460.

[9] 魏 玲，祁建军，张文修. 决策形式背景的概念格属性约简 [J]. 中国科学 E 辑：信息科学，2008，38(2)：195-208.

[11] 张 涛，洪文学，路 静. 形式背景的属性树表示 [J]. 系统工程理论与实践，2011，31(2)：197-202.