

关系粒度的概念格增量维护与关联规则更新

智慧来 智东杰

(河南理工大学计算机科学与技术学院 焦作 454000)

摘要 为了适应动态环境下概念格应用的实践需要,有必要研究关系粒度的概念格维护和关联规则更新。首先,给出了父子概念对的概念,并在此基础上设计了基于关系粒度的概念格增量维护算法。其次,发现了一个概念的内涵缩减由这个概念和其父概念的内涵差集决定,一个概念的内涵缩减决定这个概念蕴涵的关联规则。最后,根据这一观点提出了基于关系粒度的概念格关联规则更新的方法。

关键词 概念格,增量维护,关联规则,关系粒度

中图分类号 TP18 **文献标识码** A

Incremental Maintenance of Concept Lattice and Association Rules under Granularity of Relation

ZHI Hui-lai ZHI Dong-jie

(School of Computer Science and Technology, He'nan Polytechnic University, Jiaozuo 454000, China)

Abstract In order to meet the needs of concept lattice application in dynamic environment, it is necessary to find a way to maintain concept lattice based on the granularity of relation. Firstly, we put forward the term father-son concept pair and gave its definition, and brought about an incremental maintenance algorithm. Secondary, we discovered that association rules can be calculated by intent reduction of concepts which is determined by the intersection of its father concept's intent and its own intent. Finally, we put forward a method for intent updating.

Keywords Concept lattice, Incremental maintenance, Association rule, Granularity of relation

1 引言

形式概念分析是 R. Wille^[1]提出的、以序理论和完备格理论为基础、依据数据库中提供的基本信息建立的一种刻画对象与属性之间关系的数学结构。基于概念格的数据挖掘和数据维护的应用十分普遍,例如基于概念格的服务选择^[2]、基于概念格的数学公式搜索和组织^[3]、基于概念格的多粒度本体组织^[4]、基于概念格和关联规则的个人检索^[5]等。这些研究都采用了概念格模型。随着描述客体的变化,概念格的结构将发生变化。这些变化不仅包括对象或者属性的添加和删除,还包括对象和属性之间关系的变化。

在应用中,概念格的维护是一个瓶颈问题。国内外的学者和研究人员对此进行了深入的研究。使用渐进式生成算法可以实现以对象或者属性为粒度的概念格的维护,业已提出的算法有:Nourine^[6]提出了一个构造算法,算法中利用一棵词典树,增长地生成概念集并输出一棵概念树;Godin^[7]提出了一个增长算法,其在生成概念格的同时也生成了一些蕴涵规则;近代 FCA 理论的创始人之一 Ganter^[1]提出了一个 Next Closure 算法来构造概念格,该算法将概念的内涵视为属性集的闭包,每次使用已存在的闭包,基于预定义的序产生下一个闭包;Stumme^[8]在 Next Closure 算法的基础上提出了 Titanic 算法,其在闭包产生和概念候选集产生这两个环节上

做了改进。

以上这些算法都是以对象或者属性作为增量更新的粒度,都没有涉及关系这一粒度。本文研究当对象和属性之间存在的关系变动时概念格的维护与关联规则更新,也就是关系粒度的概念格维护与关联规则更新。

2 概念格的相关概念

下面将逐一给出概念格的基本概念,这些内容均出自于 Ganter B 的著作“Formal Concept Analysis”^[1],但表述形式可能有所不同。

定义 1 一个形式背景 $K := (G, M, I)$ 是由两个集合 G 和 M 以及 G 与 M 之间的关系 I 组成。 G 的元素称为对象, M 的元素称为属性。 $(g, m) \in I$ 或 gIm 表示对象 g 具有属性 m 。

定义 2 设 A 是对象集合 G 的一个子集,定义 $f(A) = \{m \in M \mid \forall g \in A, gIm\}$,相应地设 B 是属性集合 M 的一个子集, $g(B) = \{g \in G \mid \forall m \in B, gIm\}$ 。如果 A, B 满足条件 $f(A) = B$ 且 $g(B) = A$,则称序对 (A, B) 为形式背景 K 的一个概念, A 称为概念 (A, B) 的外延, B 称为概念 (A, B) 的内涵。

定义 3 若 $(A_1, B_1), (A_2, B_2)$ 是某个形式背景的两个概念,而且 $A_1 \subseteq A_2$ (等价于 $B_2 \subseteq B_1$),则称 (A_1, B_1) 是 (A_2, B_2) 的子概念, (A_2, B_2) 是 (A_1, B_1) 的父概念,并记作 $(A_1, B_1) \leq (A_2, B_2)$,关系 \leq 称为是概念的“层次序”。 (G, M, I) 的所有

到稿日期:2012-06-15 返修日期:2012-10-20 本文受国家自然科学基金(60975033),河南理工大学博士基金(B2011-102)资助。

智慧来(1981-),男,讲师,主要研究领域为粗糙集合、本体、形式概念分析等,E-mail:zhihuilai@126.com;智东杰(1952-),男,高级实验师,主要研究领域为形式概念分析、符号计算等。

概念用这种序组成的集合称为概念格,记作 $L(G, M, D)$ 。

3 基于关系的概念格增量维护

定义 4 如果概念 $(A_1, B_1) > (A_2, B_2)$, 并且不存在 (A_3, B_3) , 使得 $(A_1, B_1) > (A_3, B_3) > (A_2, B_2)$, 那么 (A_1, B_1) 和 (A_2, B_2) 称为是父子概念对, 记作 $[(A_1, B_1), (A_2, B_2)]$ 。

定理 1 增加关系 aIb 进行概念格的维护时, 如果父子概念对 $[(A_1, B_1), (A_2, B_2)]$ 满足: A_1 包含 a, B_2 包含 b , 那么生成新概念 $(A_2 \cup a, B_1 \cup b)$ 。

证明: 因为 (A_2, B_2) 是 (A_1, B_1) 的子概念, 所以 A_2 具有属性 B_1 ; 又因为 B_2 包含 b , 那么 A_2 既具有属性 B_1 , 又具有属性 b 。因为 A_1 包含 a , 同时 A_1 具有属性 B_1 , 那么 a 具有属性 $B_1 \cup b$, 所以 $A_2 \cup a$ 具有属性 $B_1 \cup b$, 具有属性 $B_1 \cup b$ 的对象只有 $A_2 \cup a$ 。故定理成立。

算法 1 增加关系 aIb 时概念格的维护算法

步骤 1 确定概念格的维护位置; 寻找这样的父子概念对 $[(A_1, B_1), (A_2, B_2)]$, 其中 A_1 包含 a, B_2 包含 b ;
对每一对找到的父子概念对执行步骤 2—步骤 4;

步骤 2 生成新概念: 生成概念 $(A_2 \cup a, B_1 \cup b)$;

步骤 3 父子概念对中概念的更新: 判断并执行其中的一个动作

动作 1 若 $A_1 \subseteq A_2 \cup a, B_1 \cup b \supseteq B_2$, 则由生成概念 $(A_2 \cup a, B_1 \cup b)$ 代替父子概念对 $[(A_1, B_1), (A_2, B_2)]$;

动作 2 若 $A_1 \subseteq A_2 \cup a$ 成立, $B_1 \cup b \supseteq B_2$ 不成立, 则由生成结点 $(A_2 \cup a, B_1 \cup b)$ 代替父概念 (A_1, B_1) ; 若 $A_1 \subseteq A_2 \cup a$ 不成立, $B_1 \cup b \supseteq B_2$ 成立, 则由生成结点 $(A_2 \cup a, B_1 \cup b)$ 代替子概念 (A_2, B_2) ;

动作 3 若 $A_1 \subseteq A_2 \cup a$ 不成立, $B_1 \cup b \supseteq B_2$ 不成立, 则不执行任何动作;

步骤 4 修改父子概念对与邻接的父子概念对的关系;

如果步骤 3 执行动作 1, 则: 若更新父子概念对后指向小的父子概念对, 只保留父子概念对更新前的父概念的指针; 若更新父子概念对后指向大的父子概念对, 只保留父子概念对更新前的子概念的指针;

如果步骤 3 执行动作 2, 则: 若更新后的父子概念对中两个概念同时指向一个小的概念, 只保留子概念的指针; 若更新后的父子概念对中两个概念同时指向一个大的概念, 只保留父概念的指针;

如果步骤 3 执行动作 3, 则: (A_1, B_1) 和 (A_2, B_2) 成为生成概念 $(A_2 \cup a, B_1 \cup b)$ 的父概念和子概念;

步骤 5 算法结束, 返回。

4 基于关系的概念格关联关系增量更新

关联规则一直是表达关联关系的主要方式, 是数据挖掘研究的核心内容^[9,10]。采用概念格的内涵缩减是得到关联规则的一个有效的途径^[9]。

定义 5 对于给定的概念 $C=(A, B)$, 如果属性集合 R 满足下述两个条件: (1) $g(R)=g(B)=A$; (2) 对于任意的 $T \subset R$ 有 $g(T) \supset g(R)$; 则称 R 是 C 的一个内涵缩减^[9,10]。

对于内涵缩减, 假设概念 $C=(A, B)$ 的内涵缩减为 R , 则提取关联规则 $R \rightarrow B - R$, 置信度 $conf(R \rightarrow B - R) = 1$ 。其物理意义是, 如果 R 能表示概念 C , 那么就能由 R 得到概念 C 的其它属性。

定义 6 对于一个给定的族集 $\mathcal{F} = \{F_1, F_2, \dots, F_n\}$, 集合 F 如果满足 $\forall F_i \in \mathcal{F} (F \cap F_i \neq \{\})$, 且 $\forall S \subset F (\exists F_i \in \mathcal{F} (S \cap$

$F_i = \{\}))$, 则 F 被称为族集 \mathcal{F} 的一个极小覆盖^[9]。

定理 2 对于一个概念结点 $c=(A, B)$ 和一个子集 $P \subseteq B$, P 是 c 的一个内涵缩减当且仅当 P 是族集 $\{B - B_p \mid c_p = (A_p, B_p) \text{ 是 } c \text{ 的一个父结点的}\}$ 的一个极小覆盖^[9]。

由定理 2 可知, 概念的内涵缩减由这个概念以及它的父概念决定, 更准确地讲是由这个概念与它的父概念的内涵差集决定。分析概念以及其父概念的更新规律就可以得到这个概念的内涵缩减的更新规律, 进而得到概念蕴涵的关联规则的更新规律。

根据算法 1, 概念的更新分为 3 种情况: 生成概念代替父子概念对中的子概念、生成概念代替父子概念对中的父概念、生成概念代替父子概念对。因此, 由增加关系诱导的关联规则的更新也需要分为 3 种情况进行研究。

① 生成概念代替父子概念对中的子概念时关联规则的更新

定理 3 形式背景中增加关系, 概念格更新后, 生成概念代替父子概念对中的子概念, 被代替的子概念(被更新概念, 也就是生成概念)的内涵缩减不变。

证明: 根据算法 1, 被更新概念的外延扩大, 内涵不变, 又由于被更新概念的父概念不变, 因此被更新概念与它的父概念的内涵差集不变, 所以被更新概念的内涵缩减不变, 故定理成立。

② 生成概念代替父子概念对中的父概念时关联规则的更新

定理 4 形式背景增加关系 aIb , 概念格更新后, 生成概念 (A_{new}, B_{new}) 代替父子概念对中的父概念, $C_f(A_f, B_f)$ 为 (A_{new}, B_{new}) 的父概念且 $b \notin B_f$, 假设这个被代替的父概念(被更新概念)的内涵缩减在更新前为 R , 那么更新后其内涵缩减为 R 和 T , T 由 R 中的 c 用 b 替换得到, $c \in Z, Z = B_{new} - B_f - b$ 。

证明: 根据算法 1 可知概念格更新前, 被更新概念的内涵为 $B_{new} - b$ 。

容易证明概念格更新后 (A_{new}, B_{new}) 的父概念中有且仅有一个内涵不包含属性 b , 记作 $C_f(A_f, B_f)$, 也就是说 $C_f(A_f, B_f)$ 在概念格更新前后不变。

假设 (A_{new}, B_{new}) 的其它父概念的内涵为 B_{f1}, \dots, B_{fm} , 那么根据概念格中属性的继承关系, 概念格更新后这些概念的内涵增加属性 b , 改变为 $B_{f1} \cup b, \dots, B_{fm} \cup b$ 。

那么更新前被更新概念与父概念的差集为 $\{(B_{new} - b) - B_f, (B_{new} - b) - B_{f1}, \dots, (B_{new} - b) - B_{fm}\}$, 更新后为 $\{(B_{new} - b) - B_f, B_{new} - (B_{f1} \cup b), \dots, B_{new} - (B_{fm} \cup b)\}$, 经过比较更新前后仅差集的第一项增加属性 b 。

根据最小覆盖的计算方法可知, 若原来的内涵缩减为 R , 那么更新后内涵缩减为 R 和 T , T 由 R 中的 c 用 b 替换得到, $c \in Z, Z = B_{new} - B_f - b$ 。故定理成立。

③ 生成概念代替父子概念对时关联规则的更新

定理 5 形式背景增加关系 aIb , 概念格更新后, 生成概念代替父子概念对, 那么, 生成概念的内涵缩减为更新前父子概念对中子概念的内涵缩减。

证明: 增加关系 aIb , 假设有 $[(A_1, B_1), (A_2, B_2)]$, 其中 A_1 包含 a, B_2 包含 b , 生成概念 $(A_2 \cup a, B_1 \cup b)$, 生成概念代

替父子概念对,则 $A_2 \cup a = A_1, B_1 \cup b = B_2$ 。

假设更新前 (A_1, B_1) 的父概念为 (A_{1f}, B_{1f}) , 更新前 (A_2, B_2) 的父概念为 (A_{2f}, B_{2f}) , $B_2 - B_1 = b$, 易知 $B_{2f} - B_{1f} = b$ 。

根据算法 1, 生成概念代替父子概念对, 若更新父子概念对后指向大的父子概念对, 则只保留父子概念对更新前的子概念的指针, 是相当于 $(A_2 \cup a, B_1 \cup b)$ 占据了 (A_2, B_2) 的位置, 其父概念为 (A_{2f}, B_{2f}) 。因为 (A_{2f}, B_{2f}) 内涵包含属性 b , 所以更新后内涵不变。从而 $(A_2 \cup a, B_1 \cup b)$ 内涵与其父概念内涵的差集为 $(B_1 \cup b) - B_{2f} = (B_1 \cup b) - (B_{1f} \cup b) = B_1 - B_{1f}$, 也就是更新后生成概念与其父概念的内涵差集等于更新前子概念与其父概念的内涵差集。根据定理 2 易知, 生成概念的内涵缩减等于更新前父子概念对中子概念的内涵缩减, 故定理成立。

例 1 在图 1 的概念格 L_1 中, 对象 3 获得属性 7, 增加关系 3I7, 研究概念格的维护与关联关系的更新。

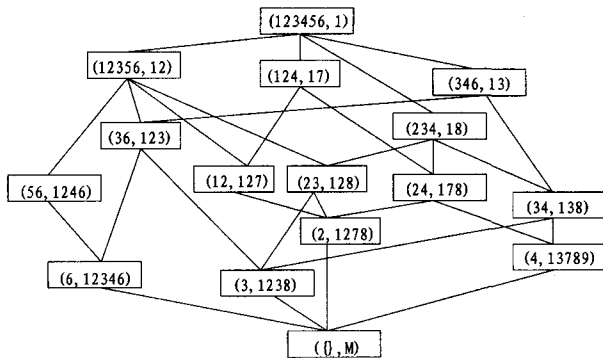


图 1 概念格 L_1

增加关系 3I7 前部分概念的内涵缩减如下:

$(124, 17)$ 内涵缩减为 $\{7\}$; $(12356, 12)$ 内涵缩减为 $\{2\}$; $(12, 127)$ 内涵缩减为 $\{2, 7\}$; $(234, 18)$ 内涵缩减为 $\{8\}$; $(24, 178)$ 内涵缩减为 $\{7, 8\}$; $(23, 128)$ 内涵缩减为 $\{2, 8\}$; $(2, 1278)$ 内涵缩减为 $\{2, 7, 8\}$; $(34, 138)$ 内涵缩减为 $\{3, 8\}$; $(4, 13789)$ 内涵缩减为 $\{3, 7\}$; $(3, 1238)$ 内涵缩减为 $\{2, 3, 8\}$ 。

增加关系 3I7 后, 执行算法 1 并更新内涵缩减:

步骤 1 确定概念格的维护位置: $[(123456, 1), (124, 17)], [(12356, 12), (12, 127)], [(234, 18), (24, 178)], [(23, 128), (2, 1278)], [(34, 138), (4, 13789)], [(3, 1238), (\emptyset, M)]$;

步骤 2 生成新概念: $(1234, 17), (123, 127), (234, 178), (23, 1278), (34, 1378), (3, 12378)$;

步骤 3、步骤 4 概念的更新与关系的更新:

$[(123456, 1), (124, 17)]$ 中的 $(124, 17)$ 更新为 $(1234, 17)$, 定理 3, 内涵缩减不变, 为 $\{7\}$;

$[(12356, 12), (12, 127)]$ 中的 $(12, 127)$ 更新为 $(123, 127)$, 定理 3, 内涵缩减不变, 为 $\{2, 7\}$;

$[(234, 18), (24, 178)]$ 更新为 $(23, 1278)$, 定理 5, $(23, 1278)$ 的内涵缩减为 $(234, 18)$ 的内涵缩减, 即 $\{8\}$;

$[(23, 128), (2, 1278)]$ 更新为 $(23, 1278)$, 定理 5, $(23, 1278)$ 的内涵缩减为 $(23, 128)$ 的内涵缩减, 即 $\{2, 8\}$;

$[(34, 138), (4, 13789)]$ 中的 $(34, 138)$ 更新为 $(34, 1378)$, 定理 4, $(34, 138)$ 的内涵缩减为 $38, Z = 1378 - 13 - 7 = 8$, 所以 $(34, 1378)$ 的内涵缩减为 $\{3, 8\}$ 或 $\{3, 7\}$;

$[(3, 1238), (\emptyset, M)]$ 中的 $(3, 1238)$ 更新为 $(3, 12378)$, 定理 4, $(3, 1238)$ 的内涵缩减为 $238, Z = 12378 - 123 - 7 = 8$, 所以 $(3, 12378)$ 的内涵缩减为 $\{2, 3, 8\}$ 或 $\{2, 3, 7\}$ 。

步骤 5 返回概念格 L_2 (见图 2)。

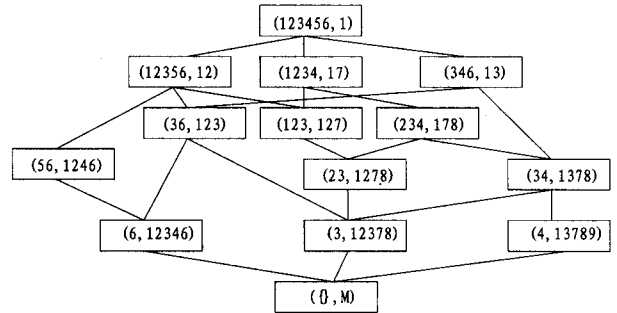


图 2 概念格 L_2

经过计算检验, 上述计算结果与使用基于极小覆盖的方法完全相同。依据更新后的内涵缩减, 就可以得到关联规则, 这里不再赘述。

结束语 在现实的环境下, 事物始终处在变化之中。形式背景由于是对客观事物的抽象, 因此也处在不断变化之中, 相应地, 其上提取的关联规则也随之改变。本文研究了增加关系时概念格的增量维护与关联规则更新, 还需要研究的问题是删除关系时的概念格维护和关联规则更新问题。

参考文献

- [1] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundation[M]. New York: Springer-Verlag, 1999
- [2] Aversano L, Bruno M, Canfora G, et al. Using Concept Lattices to Support Service Selection[J]. International Journal of Web Services Research, 2006, 3(4): 32-51
- [3] Nguyen, Tam T, Cheung H S, et al. A lattice-based approach for mathematical search using Formal Concept Analysis[J]. Expert Systems with Applications, 2012, 39(5): 5820-5828
- [4] Kang Xiang-ping, Li De-ya, Wang Lu-ge. Research on domain ontology in different granulations based on concept lattice[J]. Knowledge-Based Systems, 2012, 27(5): 152-161
- [5] Rita K, Chang-Kai H, Chang Mai-ga, et al. A personalized webpage reconstructor based on concept lattice and association rules[J]. Journal of Internet Technology, 2011, 12(6): 1015-1024
- [6] Nourine L, Raynaud O. A fast algorithm for building lattices[J]. Information Processing Letters, 1999, 71(1): 199-204
- [7] Godin R, Missaoui R, Alaoui H. Incremental concept formation algorithms based on Galois (concept) lattices[J]. Computational Intelligence, 1995, 11(2): 246-267
- [8] Stumme G, Taouil R, Bastide Y, et al. Fast computation of concept lattices using data mining techniques [C]//Proceedings of the 7th International Workshop on Knowledge Representation Meets Databases. Berlin: Technical University of Aachen, 2000: 129-139
- [9] 谢志鹏, 刘宗田. 概念格与关联规则发现[J]. 计算机研究与发展, 2000, 37(12): 1415-1421
- [10] Pasquier N. Closed set based discovery of small covers for association rules[J]. Networking and Information Systems Journal, 2010