增量式广义概念格结构的生成算法研究与实现

胡 健[1] 杨炳儒[2]
(江西理工大学信息工程学院 赣州 341000)[1] (北京科技大学信息工程学院 北京 100083)[2]

摘 要 介绍分析了概念格的研究现状,给出了基于规则的广义概念格的基本定义。通过构建树结构,缩小产生子格节点的范围,产生增量式广义概念格算法。最后,通过实例验证了所提出的算法的时空有效性,并给出了对几种概念格的生成算法有效性的比较结果。

关键词 广义概念格,概念格,增量式生成算法,知识构建,产生式规则
中图法分类号 TP311 **文献标识码** A

Research and Implenentation of Incremental Generalized Concept Lattice Construction Algorithm

HU Jian[1] YANG Bing-ru[2]
(Faculty of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)[1]
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)[2]

Abstract This paper gave a rules based fundamental definition of generalized concept lattice, by analyzing the research status on concept lattices. Then we presented an incremental algorithm for generalized lattice construction, after a tree structure was built and the ranges of sub-lattice nodes were reduced. The efficiency of this algorithm was verified by an illustration, and the comparative result of several algorithms of generating concept lattice was given.

Keywords Generalized concept lattice, Concept lattice, Incremental construction algorithm, Knowledge construction, Production rule

概念格理论自 Wille 提出以来,其内在的优势得到越来越多的科研工作者的注意,并迅速在多个领域得到发展,如:Cole 和 Eklund 使用概念缩放来制作嵌套的 Hasse 图,并将概念格方法应用于分析和可视化医药数据库[1]。Eklund 和 Martin 借助一个基于概念格的工具 Web KB,通过使用一种形式语言来描述概念的义及概念间的层次关系,从而提高了 Web 文档索引和导航的能力[2]。Godin 和 Mineau 等人描述了使用概念格方法从现存软件系中生成和检索摘要的方法[3],他们通过两个重要的软件重用过程演示了该方法的用性,该研究是包含许多商业学术伙伴的一项软件工程研究项目的一部分。Neuss 和 Kent 使用概念格进行 Internet 文档元信息的自动分类和分析[4]。Corbett 和 Burrow 提出使用概念格表示建筑早期设计软件支持环境 (SEED) 中的状态常图,使得设计中获得的知识可以重用[5]等等。

规则知识本身是用概念内涵的集合(或属性集合)间的关系来描述的,而体现于相应的外延集之间的包含(或近似包含)关系。由于概念格结点很好地体现了概念外延和内涵的统一,因此概念格作为规则知识发现的基础结构,成为表示规则知识的一个天然的平台。同时,结点间的关系体现了概念之间的泛化和例化关系,因此基于概念格进行规则知识的发现的过程变得十分方便。

1 产生式规则的广义概念格表示

1.1 广义概念格的基本定义

根据概念格的基本定义和产生式规则表示的需求特点,给出广义概念格表示的定义。

定义 1 若有形式背景 $K=(U,D,R,S)$,其中 U 是规则集合,D 是规则特征属性的集合,R 是 U 与 D 之间的一个二元关系,即 $R\subseteq U\times D$,S 是规则支持度的集合,则在此形式背景下,存在偏序集合与之相对应,并且这个偏序集合产成唯一的格结构。

这里,规则集合 U 可仅以规则序列号的集合来表示。规则特征属性集合 D 就是知识素节点的集合。R 是此规则所具有的特征属性,或说是本规则的条件和决策属性的包含。格中的每一个元素是一个三元组 (X,Y,S),$X\in P(U)$,$Y\in P(D)$,$P(A)$ 表示 A 的幂集,S 为 X 上支持度的均值。对于 xRy,表示 $x\in U$ 与 $y\in D$ 间存在关系 R,读作"规则 x 具有特征 y",或"规则 x 是由节点特征 y 所组成的"。

定义 2 当且仅当三元组 (X,Y,S) 满足性质:

$X=g(Y)$,其中 $g(Y)=\{x\in U\mid \forall y\subseteq D, xRy\}$

$Y=f(X)$,其中 $f(Y)=\{y\in D\mid \forall x\subseteq R, xRy\}$

时,称三元组 (X,Y,S) 关于 R 是完备的,且有 $f(\Phi)=D$,$g(\Phi)=U$。

到稿日期:2008-07-01 本文受国家自然科学基金资助项目(60675030)资助。
胡 健(1967-),男,博士,副教授,主要研究方向为数据挖掘与智能信息检索,E-mail:euguenehu@yahoo.com.cn;杨炳儒 教授,博士生导师,主要研究方向为知识发现与智能系统等。

定义 3 由定义 1 和定义 2 所诱导的格 L 就称为广义概念格。

定理 1 所有广义概念格中的节点都是最大扩展序偶。

格中每个节点(X,Y,S)是一个序偶，且对于关系 R 都是完备的，因此显然可见所有节点都是最大扩展序偶。即只有最大扩展的序偶才出现在广义概念格的层次结构中。

定理 2 这种最大扩展是偏序集中的一种闭包。对于偏序集$(U,=)$中的闭包，有 $h:U\to U$，性质如下：

ⅰ) $\forall x\forall y, x=y\Rightarrow h(x)=h(y)$

ⅱ) $\forall x, h(x)=x$

ⅲ) $\forall x, h(h(x))=h(x)$

$h(x)$称为 x 的 h 闭包。若 $x=h(x)$，则称 x 是 h 近似。

定理 3 在广义概念格节点 $C_1(X_1,Y_1,S_1)$ 和 $C_2(X_2,Y_2,S_1)$，若 $Y_1<Y_2\Leftrightarrow X_2\subset X_1$，则有 $C_1<C_2\Leftrightarrow X_2\subset X_1$。

根据广义概念格上的这种偏序关系，根据如下原则得到广义概念格图(或哈斯图)：若 $C_1<C_2$，且不存在元素 C_3，使得 $C_1<C_3<C_2$，则存在边$(C_1<C_2)$。称 C_1 为 C_2 的父节点，C_2 为 C_1 的子节点。

1.2 产生式规则的广义概念格表示

由广义概念格定义，可构造匹配的形式背景(U,D,R,S)，如表 1 所列：$U=\{1,2,\cdots\}$为规则序号，$D=\{a_1,a_2,a_3,a_4,a_5,b_1,b_2,\cdots\}$为知识素节点的集合，$R$ 为每条规则的条件知识素节点和决策知识素节点，S 为此条规则的支持度。可以此形式背景得到相应的广义概念格。图 1 为其所对应的广义概念格的哈斯图。

表 1 产生式规则知识的形式背景

R	a	b	c	d	e	…	sup
1	2	3	4	3	1	…	s_1
2	4	4	2	1	5	…	s_2
3	1	0	0	1	0	…	s_3
4	0	1	1	0	0	…	s_4
5	0	1	0	0	1	…	s_5
…	…	…	…	…	…	…	…

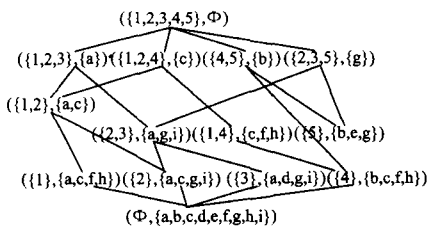

图 1 表 1 所对应的广义概念格哈斯图

2 基于广义概念格的知识构建算法

针对广义概念格的知识构建，下面给出了一种快速的广义概念格的增量式生成算法。实验表明算法是快速而有效的，同时这种算法在简单修改后均可以有效适用于通常意义下的概念格的生成。对于规则，尤其是经过 KDD 过程形成的挖掘规则来说，支持度和可信度是衡量规则的两个很重要的客观性指标。在构造基于规则的广义概念格时，应该要考虑到此评价指标。同时将评价指标引入生成算法，可以有效地对过程进行剪枝操作，从而降低算法的复杂性。在文献[6]中，提出了将支持度的概念引入概念格的生成，但是此概念格是简单的基于数据的，而且对支持度的使用也只是记数操作，不能满足基于规则概念格的构造需要。

通过构建树结构，缩小产生子格节点的范围，而无需对每个格节点都测试其成为产生子格节点的可能性，产生增量式广义概念格算法。其思路是对索引树节点进行分类，方法如下：

第一类是更新树节点。如果从树根到节点 T_1 的路径集合是 $f(x^*)$的子集，则 T_1 被称 为是更新树节点。也就是说，如果 T_1 的父节点 T_2 是更新树节点且 $T_1=T_2$.lchildren，则 T_1 是一个更新树节点。显然，有效的更新树节点集合和更新格节点之间是一一对应的，任何一个有效的更新树节点所对应的格节点都是一个更新格节点。

第二类被称为非候选树节点。存在两种情况使得 T_1 成为一个非候选树节点：一种是它的父节点 T_2，是一个有效的更新树节点，而它与其父节点之间的边不属于 $f(x^*)$，此时显然有 $Intent(C_1)\cap f(x^*)=Intent(C_2)$，显然 C_1 必然不是产生子；另一种是它的父节点，是一个有效的非更新树节点，而它与其父节点之间的边不属于 $f(x^*)$，对于这两种情况，如果 T_1 是有效的，则它所对应的格节点一定不是产生子格节点。

第三类是候选树节点。存在 3 种情况使得 T_1 称为一个候选树节点。一种是，它的父节点是一个无效的更新树节点，而它与其父节点之间的边不属于 $f(x^*)$；第二种是，它的父节点是一个无效的候选树节点；第三种是，它的父节点是一个有效的非候选树节点，而它与其父节点之间的边属于 $f(x^*)$。

基于上述考虑，我们设计了一个基于树的快速增量式广义概念格生成算法，如算法 1 所示。

算法 1 广义概念格的快速增量式更新算法

Step1 将索引树的根节点了 root 加入至 ojIndexNodeArray 中；

Step2 将 T_root 标注为更新树节点；

Step3 如果 LEIndexNodeArray 非空，则转 step4；否则，结束；

Step4 从 Index. Node. Array 的头部移出一个树节点并置于 IndexNode；

Step5 对每个 d∈D 按升序排列，并令：IndexSubNode:= IndexNode. children[d]；

Step6 如果 IndexSubNode 非空，则将 IndexSubNode 插入到 IndexNodeArray 的头部；

Step7 如果 IndexNode 是有效候选树节点或者是非候选树节点，则将 IndexSubNode 标注为候选节点，否则，将 IndexSubNode 标注为非候选节点；

Step8 若未结束，继续执行 step5，否则，转 step9；

Step9 如果 IndexNode 是一个有效的更新树节点，则将 x* 加入到 IndexNode. lbtticenonode 的外延集中，并加入到 newnodes_index[Intent (IndexNode. latticenode)]中；

Step10 如果 IndexNode 是一个有效的候选树节点，则：intersection=Intent()，创建一个新的节点 C 加入到格的节点集合中；

Step11 转 step3。

3 实验及算法评价

采用对象数目为 2000 个，特征数目为 30，最小支持度为

(下转第 228 页)

③ 在评分介于 0.7 到 0.8 的实验者中，半数人认为如果给出的群体整体兴趣中虽不含权值最低的，但也不含权值最高的，不会提升他们的满意度。这说明了权值介于中间的关键词并不能有效提高用户满意度，提高用户满意度还需要依靠用户最感兴趣的东西。

④ 群体整体兴趣，即 GP2 中关键词的排列顺序会影响满意度。约 35%的人希望按由低到高的兴趣观看，35%的人希望按由高到低的兴趣观看，20%的人希望以起伏式如由高到低再到高的兴趣观看，剩下的人持无所谓态度。但是在对于最不感兴趣物品的态度上，大多数人认为最好不要将其放在最开始的位置。

结束语 个性化的主动式信息服务已经成为人们日益关注的一个热点，群体兴趣是进行群体个性化服务的重要数据支撑。本文对个体兴趣模型和群体兴趣模型进行了分析与定义，提出了一种由群体成员的个体兴趣得到群体整体兴趣的融合算法，并通过实验证明了此算法的合理性和有效性。

参 考 文 献

[1] Petrelli, et al. A user centered approach to user modeling // Proceedings of the 7th Int. Conference on User Modeling (UM99). Springer Wien, New York, 1999: 255-264

[2] McCarthy J F, Anagnost T D. MusicFX: An Arbiter of Group Preferences for Computer Supported Collaborative Workouts [C] // Proceedings of the 1998 Conference on Computer-Supported Cooperative Work. 1998: 363-372

[3] Ardissono L, Goy A, Petrone G, et al. INTRIGUE: Personalized recommendation of tourist attractions for desktop and handset devices[J]. Applied Artificial Intelligence, 2003, 17: 687-714

[4] Kobsa A. Generic User Modeling Systems[J]. User Modeling and User-Adapted Interaction Journal, 2001, 11(1/2): 49-63

[5] Kay J, Kummerfeld B, Lauder P. Personis: A server for user models // AH'02: Proceedings of Adaptive Hypermedia and Adaptive Web-based Systems. London, UK: Springer-Verlag, 2002

[6] Jameson A. More than the sum of its members: challenges for group recommender systems // AVI '04: Proceedings of the Working Conference on Advanced Visual Interfaces. Gallipoli, Italy: ACM Press, 2004

[7] Masthoff J. Group modeling: Selecting a sequence of television items to suit a group of viewers[J]. User Modeling and User-Adapted Interaction, 2004, 14(1): 37-85

(上接第 224 页)

0.4，最小可信度为 0.3，图 2 所示为算法的 CPU 时间随库中节点元素的变化情况。其中□为 Bordat 算法，○为 Chein 算法，△为批量式建造算法，◇为增量式建造算法。

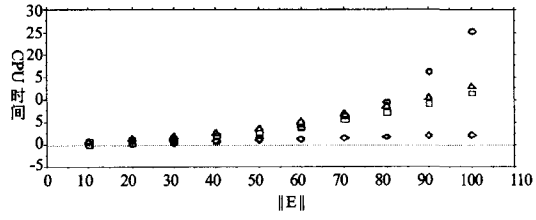

图 2 四种算法的 CPU 时间比较

采用文献[9]所提供的 UCI 知识库验证本算法，UCI 知识库是一个开放的平台，包括有 17 个大的集合，13000 个数据，43 个属性，791 个规则，对其做预处理以使它的格式适用于本文对知识的要求。结果表明，采用基于广义概念格的规则生成，从规则集合中得到了 298 个假设，由于原有的算法不对规则知识进行处理，因此这一点上没有可比性。

结束语 本文在已经提出的概念格生成算法的基础上，给出了广义概念格结构的增量式生成算法。由于增加了支持度和实时的可信度信息，大大提高了知识发现的效率。同时通过构建树结构，缩小产生子格节点的范围，这样就无需对每个格节点都测试其成为产生子的可能性。这种算法在简单修改后均可以有效适用于通常意义下的概念格的生成。

参 考 文 献

[1] Cole R, Eklund P. Scalability in formal concept analysis [J]. Computational Intelligence, 1999, 15 (1): 11-27

[2] Martin P, Eklund P W. Knowledge Retrieval and the World Wide Web[J]. IEEE Intelligent Systems, 2000, 15(3): 18-25

[3] Godin R, Mineau G, Missaoui R, et al. Applying concept formation methods to software reuse [J]. International Journal of Knowledge Engineering and Software Engineering, 1995, 5(1): 119-142

[4] Kent R E, Neuss C. Creating a Web Analysis and Visualization Environment[J]. Computer Networks and ISDN Systems, 1995, 28(1/2): 109-117

[5] Corbett D, Burrow A L. Knowledge reuse in SEED exploiting conceptual graphs[C] // International Conference on Conceptual Graphs. Sydney, 1996: 56-60

[6] Krajci, Stanislav. A Generalized Concept Lattice[J]. Logic Journal of the IGPL, 2005, 13(5): 543-550

[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] Bordat J-P, Berry A, Sigayret A. A local approach to concept generation[J]. Source Annals of Mathematics and Artificial Intelligence archive, 2007, 49 (1): 117-136

[9] Newman D J, Hettich S, Blake C L, et al. UCI Repository of Machine Learning Databases [DB/OL]. Available at: http://www.ics.uci.edu/~mlearn/MLRepository.html

[10] 胡可云，陆玉昌，石纯一. 粗糙集理论及其应用进展[J]. 清华大学学报：自然科学版，2001，41(1)：64-68

[11] 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(2)：1415-1421

[12] 王志海，胡可云，胡学钢，等. 概念格上规则提取的一般算法与渐进式算法[J]. 计算机学报，1999，22(1)：66-70