基于形式概念分析的构造算法研究

侯雅静

作业	分数
得分	

2020 年 11 月 13 日

第 41 卷 第 12 期 计 算 机 学 报 Vol. 37 No. 12
2020 年 11 月 CHINESE JOURNAL OF COMPUTERS Nov. 2020

基于形式概念分析的构造算法研究

侯雅静

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘 要 形式概念分析是由德国的 Wille 教授于 20 世纪 80 年代初提出的，它反映了概念的哲学理解，其核心数据结构概念格，也称 Galois 格，准确而简洁地描述了概念之间的层次关系，已成为一种重要的知识表示方法。概念格通过 Hasse 图生动和简洁地体现了这些概念之间的泛化和特化关系。因此，概念格被认为是进行数据分析的有力工具。目前，已经有了一些建造概念格的算法，并且概念格在信息检索、数字图书馆、软件工程和知识发现等方面得到应用。本文讨论了概念格的基本原理，介绍了概念格的相关构造算法，并对各种建格算法加以论述，并对概念格的后继发展研究应用做了相应的讨论。

关键词 形式概念分析 概念格 构造算法

中图法分类号 TP311.20 **DOI 号** 10.3970/j.issn.1001-3695.2019.11.031

Research on Construction Algorithm Based on Formal Concept Analysis

Hou Yajing

(Computer science and technology，Dalian maritime university, LiaoningDalian,116026,China)

Abstract * Formal concept analysis was proposed by Professor Wille in Germany in the early 1980s. It reflects the philosophical understanding of concepts. Its core data structure concept lattice, also known as Galois lattice, accurately and concisely describes the hierarchical relationship between concepts and has become an important method of knowledge representation. Concept lattices vividly and succinctly represent the generalization and specialization relationships among these concepts through Hasse diagrams. Therefore, concept lattice is considered as a powerful tool for data analysis. At present, there have been some algorithms for constructing concept lattice, and concept lattice has been applied in information retrieval, digital library, software engineering and knowledge discovery. In this paper, the basic principle of concept lattice is discussed, the correlative construction algorithm of concept lattice is introduced, and various construction algorithms are discussed, and the subsequent development and application of concept lattice are also discussed.

Key words Formal concept analysis; Concept lattice; Construction algorithm;

1 引言

人类在认知过程中，把所感觉到的具有共同特点的事物抽取出来，加以概括，称为概念。在哲学中，概念被理解为由外延和内涵两个部分所组成的思想单元。基于概念的这一哲学思想，德国的 R.Wille 教授于 1982 年首先提出了形式概念分析理论。 目前形式概念分析已被广泛研究并应用到机器学习、软件工程和信息检索等领域。基于形式概念分析的粗糙集模型、基于概念格的多示例集成学习模型、基于概念格的不同粒度下的领域本体模型及形式概念分析在不同粒度下知识获取模型，这些模型不仅在理论上拓展形式概念分析方法，而且对形式概念分析的应用起到积极的推动作用。

概念格是 FCA 的核心数据结构。概念格理论最早由 Wille R 等提出，是应用数学的分支，它来源于哲学相关领域内对概念的理解。随着研究的深入，很多学者逐渐认识到概念格自身结构的巨大优势，研究从开始的单纯理论扩展发展到理论与实际

应用相结合，并且融合交叉多个相关理论，成为许多专家学者关注的热点。作为数据分析和知识处理的形式化研究方法，概念格在知识发现、信息检索等方面均得到了广泛的应用。概念格理论的研究不仅能用于解决知识发现领域中所涉及的关联规则、蕴含规则、分类规则的提取，还能够实现对信息的有机组织，减少冗余度，简化信息表，所以对于概念格理论及其构造方法的研究具有十分重要的意义。

本文首先介绍了形式概念分析的基本概念，又着重介绍了概念格的构造算法与运算规则，使人们更加了解构造概念格的算法，最后介绍了概念格近年来的主要应用领域，为后续的研究方向与应用领域提供了思路。

2 基本概念

2.1 形式背景和概念格

定义 2.1：（形式背景）一个形式背景 K=(G,M,I)由集合 G、M 以及它们之间的关系组成 I，G 的元素称为对象(Objects)，M 的元素称为属性(Attributes)。为了表示一个对象 o 和一个属性 m 在关系 I 中，可以写成 oIm 或(o,m)∈I，读成“对象 o 有属性 m”。

表 1 形式背景举例

G \ M	a	b	c	d
1	×	×		×
2	×		×	
3		×	×	
4	×	×		×
5	×			

定义 2.2：（概念）对于形式背景 K，在 G 的幂集和 M 的幂集之间可以定义两个映射日 f 和 g 如下：

- ∀O⊆G: f(O)={d|∀x∈O:(xId)}
- ∀D⊆M: g(D)={x|∀d∈D:(xId)}

来自 P(G)×P（M）的二元组(O，D)如果满足两个条件：O=g(D)及 D=f(O)，则它被称为是形式背景 K 的一个形式概念，简称概念，记为 C=(O,D)，其中 D 和 O 分别被称为概念 C 的内涵和外延。K 的所有形式概念的集合被标记为 CS(K)。

定义 3：（概念格）对于概念$(O_1，D_1)$和$(O_2，D_2)$。如果$D_2⊆D_1$，则形式概念$(O_1，D_1)$是形式概念$(O_2，D_2)$的亚概念，记为$(O_1，D_1) ≤(O_2，D_2)$。通过这个关系，我们得到一个有序集 CS(K)=(CS(K)，≤)，这是一个完全格，被称为形式背景 K 的概念格，记为 L(K)。

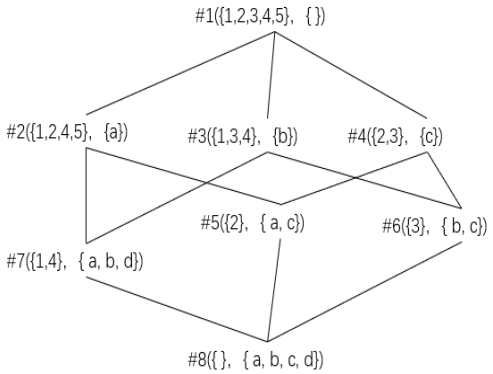

图 1　表 1 的形式背景所对应的概念格

3 概念格的构造与运算

概念格的构造问题是形式概念分析应用的前提。由于概念格的时空复杂度随着形式背景的增大而可能指数性的增大，有关概念格的生成问题一直是形式概念分析应用研究的一个重点。国内外的学者和研究人员对此进行了深入的研究，提出了一些有效的算法来生成概念格，这些算法一般可分为批生成算法 (Batch Algorithm)、渐进式生成算法(Incremental Algorithm）和并行算法（Parallel algorithm）。

3.1 批生成算法

现有的批处理概念格生成算法大多都是先生成形式背景所对应的所有概念，然后再决定概念之间的亚概念—超概念连接关系。有的算法只生成所有的概念 有的算法用来产生其图，也有的算法既生成所有的概念，又同时形成其 Hasse 图。

概念格的批生成算法：

- 输入：形式背景
- 输出：概念格 L
  1) 初始化格；L = {(G, f(G))}
  2) 初始化队列；F = {(G, f(G))}
  3) 取出队列 F 中的一个概念 C，产生出

它的每个子概念 C_C；
4) 如果某个子概念 C_C 以前没有产生过，则加入到 L 中，加入队列 F；
5) 增加概念 c 和其子概念 cc 的链接关系；
6) 反复步骤 3~步骤 5，直至队列 F 为空
7) 输出概念格 L。

3.2 渐进式生成算法

Godin R 等在 1995 年提出的概念格生成算法是最经典的一个渐进式生成算法，通常称为 Godin 算法。该算法从空概念格开始，通过将形式背景中的对象逐个插入概念格来实现对概念格的渐进式构造。

对于每次新增一个对象，都需和已生成概念格中的概念进行比较，这时已有的概念节点和新增的对象之间可以存在三种关系：无关概念（Old Concept)、更新概念(Modified Concept)和新增概念的产生子概念(Generator Concept)。渐进式构造主要是对更新概念和新增概念进行不同处理后，再调整概念之间的相互关系。

概念格的渐进式生成算法：

- 输入：形式背景
- 输出：概念格 L

1) 初始化格 L 为{({ }, M)};
2) 从 G 中取一个对象 g;
3) 对于格 L 中的每个概念 $C_1=(A_1,B_1)$, 如果 $B_1 \subseteq f(g)$，则把 g 并到 A_l 中 ；
4) 如果同时满足：$B_1 \cap f(g) \neq \varnothing$; $B_1 \cap f(g) \neq B_1$ 和不存在(A_l, B_l) 的某个父节点满足 $B_2 \supseteq B_1 \cap f(g)$，则要产生一个新节点$(A_1 \cup \{g\}, B_1 \cap f(g))$;
5) 对新产生的节点加入到 L 中，同时调整节点之间的链接关系;
6) 反复步骤 2 到步骤 5，直至形式背景中的对象处理结束
7) 输出概念格 L。

3.3 并行生成算法

并行算法是针对数据规模较大时， 概念格求解在时间复杂度和空间复杂度上计算量日益突出而提出的，问题的主要矛盾在于如何协调集中式的数据存储方式与串行式的算法设计。并行算法思想的提出依赖于高性能计算机与网格并行计算的能力，综合了批处理算法的并行性和渐进式算法的高性能性。 国内对于此类算法的研究并不是很多，主要是论述如何将不一致的形式背景转化为独立背景或是一致性背景，从而解决了概念格并行构造算法的基础性问题。其中主要的算法思想是在构建概念格之前，先将形式背景拆分成诸多个分布存储的子形式背景，进而并行的构造每个子形式背景所对应的子概念格，最后将所有的子概念格合并得到最终的概念格。随着形式背景的日益庞大，此类算法具有很好的发展空间，是今后概念格构造类算法发展的主要趋势。

3.4 概念格的运算

定义 3.1: 如果形式背景 $K_1=(U_1, A_1, I_1)$和 $K_2=(U_2, A_2, I_2)$满足 $U_1 \subseteq U$，$U_2 \subseteq U$，$A_1 \subseteq A$，$A_2 \subseteq A$，则称 K_1 和 K_2 是同域形式背景，$L(K_1)$和 $L(K_2)$是同域概念格，如果 $U_2 \cap U_2=\varnothing$，则称 K_2 和 K_2、$L(K_2)$和 $L(K_2)$分别是外延独立的，简称独立的。

定义 3.2: 如果 $K_{1=}(U_1, A_1, I_1)$和 $K_2=(U_2, A_2, I_2)$是同域且独立的，则 $K_1+K_2=(U_1 \cup U_2, A_1 \cup A_2, I_1 \cup I_2)$。

定义 3.3: 对于 $C_1=(O_1, D_1)$和 $C_2=(O_2, D_2)$，如果 $D_1=D_2$，则称 C_1 内涵等于 C_2，简称 $C_1=C_2$。

定义 3.4: 对于 $C_1=(O_1, D_1)$和 $C_2=(O_2, D_2)$，如果 $D_1 \subset D_2$，则称 C_1 内涵小于 C_2，简称 C_1 大于 C_2，或称 C_2 小于 C_1。

定义 3.5: 对于 $C_1=(O_1, D_1)$，$C_2=(O_2, D_2)$和 $C_3=(O_3, D_3)$，定义 C_1+C_2 等于 C_3，如果 $O_3=O_1 \cap O_2$，$D_3=D_1 \cap D_2$。

4 概念格的应用研究

概念格主要用于机器学习，模式识别，专家系统，计算机网络，数据分析，决策分析，数据挖掘，信息检索等领域。研究概念格的价值在于解决知识发现领域中所涉及的关联规则、蕴含规则、分类规则的提取，和实现对信息的有机组织，减少冗余度，简化信息表等。

概念格理论的研究主要集中在以下几个方面：

(1) 概念格的建造

从数据集（在概念格中称为形式背景）中生成概念格的过程实质是一种概念聚类过程。对于同一批数据，所生成的格是唯一的。建格算法可以分为：批处理算法、渐进式算法(或称增量算法)、并行算法。

(2) 概念格的约简

概念格的约简能够有效地提高概念格的维护效率。使形式背景中所蕴含的知识易于发现，简化知识的表示方式。约简概念格实际上是在保持对象集

不变的条件下，如何求得最小的属性集的过程。国内的研究主要是以张文修等提出的理论为基础。给出概念格属性约简的判定定理，引入形式背景的可辨识属性矩阵。并依此为基础求得属性约简的方法。

(3) 规则提取

概念格上的规则提取具有广泛的应用前景。规则挖掘是近年来数据挖掘的研究课题，每个概念格节点本质上就是一个最大项目集，为关联规则挖掘提供了平台，体现了概念之间的包含与分类关系。更加易于理解和表示。由于规则本身是由内涵间的关系来描述的。而表现的却是外延之间的包含与被包含关系，正是由于概念节点统一了内涵与外延之间的关系，基于概念格的分类规则的提取在知识发现等方面有着广泛的应用。目前。对于概念格上分类规则的研究主要集中在优化概念格的构建和求解算法上。

(4) 模糊概念格和基于神经网络的概念格

由于各个应用领域中存在的信息具有复杂性和不确定性，在处理以上问题时。传统的形式概念分析很难们将模糊理论与形式概念分析结合起来，由此产生了模糊形式概念分析。

粗糙集理论是一种新的处理模糊和不确定性知识的数学工具。其理论的主要思想是在保持基本分类能力不变的前提下。利用不可分辨关系来描述等价关系上不可定义的知识。即粗糙集(Rough set)。该理论能够利用已有的知识库，对知识进行近似的或者不确定的描述。最大的特点在于不需要提供处理该问题所需的数据集合之外的任何先验信息，对问题处理的不确定性比较客观。

(5) 现实应用

概念格已成功的应用于数字图书馆及文献检索，软件工程，知识发现等领域，而且已取得了良好的经济效益和社会效益。如，Cole R.等将概念格方法应用于分析和可视化具有 1962 个属性和 4000 个处方摘要的医药数据库；Eklund P. w.等展示了概念格层次进行 B 文档索引和导航的能力；Cole R.等的 CEM 电子邮件管理系统通过将 Email 存储在概念格中，而不是常用的树状结构中，从而在检索电子邮件时获得了更大的灵活性。

Y. Y. Yao 提出了面向对象概念格，Duntsch 和 Gediga 构造了另外一种新的概念格——面向属性概念格。得到了两种新的概念格：面向对象概念格和面向属性概念格。

随着概念格理论与方法的进一步完善和发展，以及与其他知识发现理论与方法的交叉与融合，概念格理论与方法将成为一种知识发现的有力工具。

结束语

随着数据库系统的广泛应用和网络技术的高速发展，数据库技术也进入一个全新的阶段。我们面临着称为的“信息丰富而知识贫乏”窘境。数据库在给我们提供丰富信息的同时，也体现出明显的海量信息特征。信息爆炸时代，海量信息给人们带来许多负面影响，最主要的就是有效信息难以提炼，过多无用的信息必然会产生信息距离。因此，人们迫切希望能对海量数据进行深入分析，发现并提取隐藏在其中的信息，以更好地利用这些数据，即是数据挖掘。概念格理论是一种基于概念由外延和内涵两部分所组成的思想单元这一哲学理解提出的。它是知识的一种表现模型,依据知识体在内涵和外延上的依赖或因果关系,建立概念层次结构。概念格的图体现了一种概念层次结构,实现了对数据的可视化。因此，概念格作为一种具有极大潜力和有效的数据挖掘工具,备受人工智能工作者的广泛关注。目前,概念格正在广泛应用于机器学习、模式识别、专家系统、计算机网络、数据分析、决策分析等领域。

参 考 文 献

[1]. 李金海,闫梦宇,徐伟华,折延宏,张文修. 概念认知学习的若干问题与思考[J]. 西北大学学报(自然科学版), 2020(04)

[2]. 李金海, 魏玲, 张卓,等. 概念格理论与方法及其研究展望[J]. 模式识别与人工智能, 2020(7).

[3]. Engineering - Reliability Engineering; Investigators from Central University of Venezuela Have Reported New Data on Reliability Engineering (Introduction To Formal Concept Analysis and Its Applications In Reliability Engineering)[J]. Journal of Technology & Science,2020.

[4]. 陈锦坤. 基于图论的概念格属性约简方法及其应用[D]. 2020.

[5]. 谢小贤,李进金,陈东晓,林荣德.基于布尔矩阵的保持二元关系不变的概念约简[J].山东大学学报(理学版),2020,55(05):32-45.

[6]. 张云中, 柳迪, 张原铭. 基于形式概念分析的知识发现研究态势[J]. 情报科学, 2018, 036(009):153-158.

[7]. 毕强, 滕广青. 国外形式概念分析与概念格理论应用研究的前沿进展及热点分析[J]. 现代图书情报技术, 2010, 26(011):17-23.

[8]. 胡可云, 陆玉昌, 石纯一. 概念格及其应用进展[J]. 清华大学学报:

自然科学版, 2000, 040(009):77-81.

[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]. 王甦菁，陈震.基于概念格的数据挖掘方法研究[J]. 计算机应用,2005,25(4):157-161

[11]. 徐清泉，朱玉文，刘万春，基于概念格的关联规则算法阴.计算机应用 2005, 25(8): 1856-1860

[12]. Hu Xuegang,Wang DeYing，Liu Xiaoping，Guo Jun，Wang Hao，The Analysis on Model of Association Rules Mining Based on Concept Lattice and Apriori Algorithm[C].In: IEEE The Third International Conference on Machine Learning and Cybernetices， Shanghai，2002，8:1620-1624.