《智能信息处理》课程作业

概念格理论的研究及其应用分析

王芳铭

作业	分数[20]
得分	

2021年11月28日

概念格理论的研究及其应用分析

王芳铭

（大连海事大学 信息科学技术学院,辽宁 大连 116026）

摘 要 形式概念分析是由德国的数学家Wille教授于1982年提出的，它是一种对形式背景中的数据进行分析和规则提取的强有力工具,形式概念分析的核心数据结构是概念格.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。随着研究的深入，概念格理论越来越多地被应用到机器学习、模式识别、数据挖掘、信息检索、软件工程等领域，成为组织和处理大规模数据的有效工具。本文主要介绍了概念格基本概念及其相关的算法，对概念格在各个方面的应用进行了简要分析。

关键词 形式概念分析；概念格；应用；

Research and Application Analysis of Concept Lattice Theory

Wang Fangming

(school of Information Science and Technology, Dalian Maritime University, Liaoning Dalian 116026)

Abstract Formal concept analysis was proposed in 1982 by Professor Wille, a German mathematician. It is a powerful tool for analyzing and extracting data in formal contexts. The core data structure of formal concept analysis is concept lattice. Also known as Galois lattice, which accurately and concisely describes the hierarchical relationship between concepts, has become an important knowledge representation method. With the deepening of research, concept lattice theory has been increasingly applied to machine learning, pattern recognition, data mining, information retrieval, software engineering and other fields, becoming an effective tool for organizing and processing large-scale data. This article mainly introduces the basic concepts of the concept lattice and its related algorithms, and briefly analyzes the application value of the concept lattice in various fields.

Keywords formal concept analysis; concept lattice; application

1 引言

概念格，是根据二元关系提出的概念层次结构， 用于数据的分析和规则的提取[1]。从数据集中生成概念格的过程实质上是一种概念聚类的过程， 通过 Hasse 图可以清楚的反应出概念间的层次结构， 以及相互之间泛化与特化的关系，从而实现数据的可视化。它提出的初衷是希望通过形式化的方式刻画现实中的实体对象或抽象概念， 并建立相应的层次知识结构,描述概念之间的泛化与特化关系。近年来，信息技术飞速发展，各个领域的信息与数据急剧增加，并且由于人类的参与使数据与信息中的不确定性更加显著，信息与数据中的关系更加复杂。如何从大量杂乱无章和强干扰的数据中挖掘潜在的、新颖的、正确的、有利的价值知识，

这给智能信息处理提出了严峻的挑战，由此产生了人工智能领域研究的一个崭新领域——数据挖掘(DM)[3]和数据库知识发现(KDD) 。目前已有许多的数据挖掘工具，比如神经网络、遗传算法、支撑向量机、决策树、粗糙集、形式概念分析等等。在 DM 和KDD 诸多方法中形式概念分析(Formal ConcepAnalysis , FCA) 对于处理复杂的信息不失为一种有效的方法。

而概念格(Concept Lattices)[6], 又称伽罗瓦格(Galois Lattices)由 Wille 于 1982 年首次提出,它是形式概念分析中的核心数据单元。概念格理论是应用数学的分支，它来源于哲学相关领域内对概念的理解。在哲学层面上，概念被理解为由外延和内涵两个基本单元组成。在这一思想的基础之上，概念格将哲学中的概念形式化，认为概念格的每一个节点都是一个形式概念，它包含两

个组成部分：外延，即概念所包含的所有实例的集合；内涵是概念的描述，即概念所覆盖实例所共有特征(属性)的集合。此外，概念格是一种由偏序集诱导的格结构，显示出了概念间的泛化和例化关系，而形式概念连同它们之间的泛化和例化关系共同构成了一个概念格。一个概念格的图形由称为哈斯图(Hasse Diagram)的线图表示，哈斯图是一个偏序集的图形表示，是通过向上的箭头来反映偏序集的包含关系。通过哈斯图能够生动展现概念格中概念与概念间的关系，从而实现对数据的可视化。

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。

本文首先介绍了形势概念分析的基本概念，又着重介绍了概念格的相关构造算法与运算规则，是人们更加了解概念格是如何构造的，最后分析了近年来概念格主要的用用领域，为后续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

在这个客观世界中，单独的事物并不能够描述一个具体的系统。同样地，单独的形式概念并不足以描述一个形式概念集，于是我们引入形式背景来解释这个问题。

定义 **1** 一个形式背景 K 是一个三元组：K=(G，M，I)，其中 G 为所有对象的集合，M 为所有属性的集合，I 是 G 与 M 之间的二元关系。设(G，M，I)为形式背景，如果一个二元组(A，B)满足 A'=B'且 B'=A'，刚称(A，B)是一个概念。其中，A 称为概念的外延，B 称为概念的内涵。

定义 **2** 一个形式背景可称为一个语境，能够用一个矩形表来表示，表的每一行是一个对象，每一列是一个属性。若g行m列的交叉处是X，则表示对象g具有属性m，如表1所示。

表1 定义2的表格显示

	属性1	属性2	……	属性n
对象1	X			
对象2	X	X		
……				
对象m				X

2.2 概念格

概念格的每个节点是一个形式概念， 由两部分组成：外延，即概念所覆盖的实例；内涵，即概念的描述，该概念覆盖实例的共同特征。另外， 概念格通过哈希[1]图生动和简洁地体现了这些概念之间的泛化和特化关系。从数据集中(概念格中称为形式背景)中生成概念格的过程实质上是一种概念聚类过程；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。概念格是一种具有完备性的结构，它作为知识表示的一种形式在表现概念之间关系的规则方面有其独特的优势。

定义 **3** 若 C 1 =(A 1 ,B 1)，C 2 =(A 2 ,B 2) 是某个背景上的两个概念，而且 A 1 Í A 2 (等价于 B 2 Í B 1)，则我们称 C 1 是 C 2 的子概念(也称为广义子概念)，C 2 是 C 1 的超概念(也称为广义超概念)，并记作 C 1<C 2 ，关系<称为是概念的“层次序”，简称“序”。(G, M, R)的所有概念用这种序组成的集合用 C(G,M, R)表示，称它为背景(G, M, R)上的概念格。

定义 **4** C 1 =(A 1 ，B 1), C 2 =(A 2 ,B 2)是某个背景上的两个概念，C 1 <C 2 。如果 C 1 不存在某个子结点 C 3 =(A 3 ,B 3)，满足 A 3 Í A 2 ，则称 C 1 是 C 2 的直接父结点(直接父概念)，C 2 是 C 1 的直接子结点(直接子概念)。

定义 **5** 对于形式背景 K=(O，A，R)，存在唯一的一个偏序集<H≤，≤>与之对应，并且该偏序集存在一个唯一的下确界和一个唯一的上确界，这个偏序集产生的格结构称为概念格(concept lattice)，记为 L(O，A，R)。由以上定义可知，概念格中概念的外延集合和内涵集合之间存在对偶关系，一个概念格可看作是相互联系的两个概念格。

概念格可以图形化形式表示为有标号的线图，概念格的每个节点表示一个形式概念，由外延和内涵两部分组成。概念的外延是指此概念所覆盖的对象的集合；概念的内涵则是外延所具有的共同属性的集合。这种线图也称为 Hasse 图，它是概念格的可视化表示。

3 概念格的构造算法

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

使用批生成算法构造概念格要完成两项任务：① 生成所有的格节点，即所有概念的集合；② 建立这些格节点间直接前趋直接后继关系。按这两项任务完成的次序不同，我们可以将批处理算法分为任务分割生成模型和任务交叉生成模型。任务分割生成模型是首先生成全部的概念集合，然后再找出这些概念之间的直接前驱/直接后继关系；任务交叉生成模型是在生成概念的过程中同时确定概念之间的关系。

在已提出多种构造概念格的批生成算法[4]中，只有少数同时生成 Hasse图。根据构造格的不同方式，可以把批处理算法分为三类：自顶向下算法、自底向上算法、枚举算法。

（1）自顶向下算法：先构造全概念，也就是最上层的节点，然后依次生成该节点的所有可能的子节点，并且对每个子节点做上述操作，最后将所有存在父子关系的节点相连。算法的关键在于如何生成子节点。该算法虽简洁直观且较易实现，但存在生成许多冗余节点的问题。

（2）自底向上算法：先构造概念格的最低部的节点，然后再逐层向上进行扩展。算法的关键在于如何完成下一个层次的序对到上一个层次的合并，并且要对生成的节点进行重复性判断。如果在上层出现过，要予以标记并在完成此层操作之前删除该节点。该算法的缺点在于合并过程中会产生大量的重复性节点，效率不高，不能产生相应的 Hasse图，不具备直观性。

（3）枚举算法：按照一定的顺序枚举出格内的节点，在生成 Hasse图的同时，表达出各个节点之间的关系。

3.1 渐进生成算法

渐进生成算法的基本思想是将当前要插入的对象与格中所有的概念求交，根据交的结果进行不同的操作。典型的算法有:Godin 算法、Capineto算法, Addintent 算法。Ho等也提出了一个渐进算法，但和 Godin 算法的思想基本相同。下面简单介绍几个这类算法。

Godin 算法在插入一个新实例时，格中的节点被分为三类:一类是不变节点，这些节点的内涵和要插入实例的特征集没有交集，它们将新格保持不变;第二类是更新节点，这些节点的内涵包含于要插入实例的特征集，因此只需将其外延更新， 包括要插入实例即可;第三类是新增节点，当所有要插入的实例的特征集与原来格中某个节点的内涵交集在格中没有出现过时，需要增加新的节点，该新节点的内涵即为该交集。可以证明，新增节点的父节点必然是某个新增节点或者更新节点， 这使得连接过程很容易实现。Godin 还给出了一个改进算法，当一个新实例插入时，不必对格中所有节点进行检查，只需检查那些和新实例有共同属性的节点。可以用通过维护记录每个属性首次在格中出现的指针来实现。

Capinet算法与Godin算法的基本思想类似，它将生成新概念的条件分为:“交为空”、“交已经存在”和“交包含在已有概念中”。主要的不同处出现在连接过程中。Capineto算法是找到该新节点的最小上界和最大下界，删除它们之间的边， 并将其连接到新概念。

Addintent算法不但生成概念集，也生成概念的格结构。算法渐进地将下一个对象合并到前面对象已生成的图结构中。因此，该算法更适合于既需要得到概念集又需要得到格结构的相关应用。实验结果表明，在某些时候，Addintent算法构造整个格结构的时间与其它算法从概念集中构造格结构的时间相比，所用时间要少很多。

3.3 并行算法

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。该算法思想的提出依赖于高性能计算机与网络并行计算的能力，综合了批处理算法的并行性和渐进式算法的高性能性。

4 概念格的应用

目前，国内外已经有很多专家学者从不同领域研究了概念格理论及在知识发现中的应用，取得了许多令人满意的成果。本文简要分析了概念格在知识发现、数据挖掘、软件工程等领域的应用。

4.1 概念格在知识发现中的应用

当前，基于概念格的知识发现体系已日臻完善，并且已经演变出很多研究范式。知识发现领域中，分类规则和关联规则本身就是具有价值的知识。人

们在进行规则知识挖掘分析时，概念格内涵集之间的关系可以描述规则知识，非常有利于知识提取。概念格外延集之间包含和近似包含关系。可以充分体现规则知识。概念格的结点又反映了内涵和外延的统一，结点关系体现了概念泛化和例化关系，因而非常适合作为知识发现的基础数据结构。

概念格每个结点的内涵本质上就是最大项目集，如同决策树和粗糙集，概念格也成为当前国内外数据分析和知识提取的有效工具之一。近年来，国内外很多专家学者都致力于通过概念格进行聚类知识、分类知识、离群知识和关联知识等算法研究。面对目前数据海量、模糊、粗糙和不确定等特点，为了更好地利用概念格进行知识提取和知识表示，专家学者对概念格进行了扩展研究，大致可分为：扩展概念格、粗糙概念格、模糊概念格、约束概念格、量化概念格、加权概念格和多维概念格。概念格的扩展将使得概念构造效率更高，更加实用。

4.2 概念格在数据挖掘中的应用

随着计算机技术的不断发展，计算机应用领域不断扩大。人们收集和处理数据的能力和数量在不断变大，直接从大量的数据中找到用户感兴趣或对用户有指导意义的知识的难度也在不断变大，从而出现了“数据丰富、知识贫乏”的窘境。因此，数据挖掘技术得到了广泛的研究。关联规则是从数据库中提取知识的主要表现形式，也是数据挖掘研究的核心内容之一。形式概念分析以概念格形式把数据有机地组织起来，数据之间的关系通过概念格节点的特化一例化关系体现出来，体现了概念的内涵和外延的统一，所以，概念格非常适合作为规则发现的基础性数据结构用来发现规则型的知识。概念格应用于关联规则提取，是概念格在数据挖掘中应用的最广、取得成果最丰的一个领域[2]。

4.3 概念格在web中的应用

一方面，随着Web应用领域的不断扩大，Web应用的质量问题也不断受到人们的关注。在信息网络不断发展的今天，Web的应用和构造已经成为软件测试研究的重要内容。在Web的构建过程中，存在差异性、分布性、平台性等特性。[6]这些特性在Web应用软件发展过程中都有着至关重要的影响，另一方面，由于Web软件和互联网技术一样一般存在着开发周期短、更新速度快的特点。如何在这种特点的情况下进行相关的调试和应用测试给当代互联网发展带来了新的挑战。但是新兴形式概念分析法的应用解决了这一难题，形式概念分析方法的应用突破了Web网页在差异性和分布性上的局限。

另一方面，在传统的Web应用环境下虽然取得不少的成果，但是与实际的生活和应用需求还有很大的差距。在传统的自动化的测试中，填充表单问题还没有得到有效的解决。Web网页越来越融入到人们的生活中，很多人都通过Web应用实现各种需求，在电子商务、电子教育和安全性测试中都会出现Web应用的相关概念。形式概念分析方法在Web环境下应用下，突破了原有测试环境下的局限，实现了技术和现代人民生活需求的统一。

5 结语

概念格理论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。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。形式概念分析在计算机技术方面的应用愈加重要，随着科技的快速发展，形式概念分析也在迅速发展，随着技术的不断更新，形式概念分析在各个领域的应用也越来越广泛，然而这仍是一个年轻并在高速发展的领域。现在对概念格的研究还有许多有意义的方面，比如概念格规则提取或属性约简的启发式算法；寻找快速的模糊概念格的建格算法；概念格的规则提取问题；基于概念格的数据挖掘模型的实现等等这些都是我们以后重点研究的方向。

参考文献

[1] WILLE R.Restructuring lattice theory:Anapproach based on hier-archies of concepts[C].Ordered Sets.Dordrecht:Reidel, 1982

[2] 王甦菁. 概念格在数据挖掘中应用的研究[D]. 吉林大学, 2007.

[3] 陈朝晖. 基于概念格的数据挖掘研究及应用[D]. 西安电子科技大学, 2014.

[4] 杨强. 基于概念格的数据挖掘方法研究[D]. 山东科技大学, 2008.

[5] 何淑贤,刘桂枝.形式概念分析及其应用进展[J].应用技术, 2007(5):77-79

[6] 梁高明, 张忠磊. 基于概念格的数据挖掘方法研究[J]. 科技信息, 2007, 2007(8):55-56.

[7] 白剑. 概念格构造算法及数据挖掘应用研究[D]. 吉林大学, 2006.

[8] Wille R. Formal concept analysis as mathematical theory of concepts and concept hierarchies [M] //Formal Concept Analysis. Berlin Heidelberg: Springer-verlag，2005: 1－33.