基于形式概念的构造算法分析

张冰

（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(FCA)是由德国教授 Wille 提出来的，Wille 教授对概念进行了形式化描述，提出的形式概念分析可以用于概念的发现、排序和显示。形式概念分析(Formal Concept Analysis)的思想主要来源于哲学，在哲学中,概念被理解为由外延和内涵两部分组成的思想单元。而概念格作为一种研究概念内涵与外延的有效工具，在规则提取与数据分析方面有着广泛的应用。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。

关键词 形式概念分析，形式背景，概念格

中图法分类号 TP311 **文献标识码** A

Based On Formal Concept Construction Algorithm Analysis

Zhang Bing

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026，China)

Abstract Formal Concept Analysis (FCA) is made by a German professor Wille come, Professor Wille formalization description of concept, put forward the concept of formal concept analysis can be used to find, sort and display. Formal Concept Analysis comes mainly from the philosophical ideas, In philosophy, the concept is understood as the ideas of the two parts unit denotation and connotation. but as an effective tools of the research concept lattice connotation and extension, which has been widely used in rules extraction and data analysis. This article discusses the basic principles of the concept lattice and introduced the construction algorithm which related to the concept lattice, and a variety of algorithms to build the grid to be addressed.

Keywords Formal Concept Analysis ,Formal Context, Concept Lattice

1 引言

在计算机与网络信息技术飞速发展的今天,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) 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 Concep Analysis , FCA) 对于处理复杂的信息不失为一种有效的方法。

而概念格则是 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

收稿日期：2016-10-9
作者简介：张冰（1993-）女，黑龙江人，硕士生在读。

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和其形式背景，又介绍了概念格的基本概念，然后以一个例子构建形式背景和概念格，进行简单的概念分析，并从分析中获得关联规则，再通过对构造概念格的算法进行论述，使人们更加了解构造概念格的算法，方便以后的运用。

2 基本概念

2.1 形式概念分析

形式概念是现实世界中各种概念的抽象,通过概念外延与内涵之间的关系形式化地刻画抽象概念。在形式概念分析中,数据是用形式背景表示的。形式概念分析是 Wille 提出的一种从形式背景进行数据分析和规则提取的强有力工具，形式概念分析建立在数学基础之上，对组成本体的概念、属性以及关系等用形式化的语境表述出来，然后根据语境，构造出概念格(concept lat-tice)，即本体，从而清楚地表达出本体的结构。这种本体构建的过程是半自动化的，在概念的形成阶段，需要领域专家的参与，识别出领域内的对象、属性，构建其间的关系，在概念生成之后，可以构造语境，然后利用概念格的生成算法 CLCA，自动产生本体。形式概念分析强调以人的认知为中心，提供了一种与传统的、统计的数据分析和知识表示完全不同的方法，成为了人工智能学科的重要研究对象，在机器学习、数据挖掘、信息检索等领域得到了广泛的应用。

2.2 形式背景

形式概念分析通常由形式背景这一基本概念开始，形式背景是形式概念分析的基础，从现实中正确提取形式背景是能正确类聚关联、推理决策的前提。

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。

对于 $A\subseteq G$，定义 $A'=\{m \mid m\in M，\ \forall g\in A，gIm\}$，对于 $B\subseteq M$，定义 $B'=\{g \mid g\in G，\forall m\in B，gIm\}$。

定义 2 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。

2.3 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 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。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 $C_1=(A_1，B_1)$和 $C_2=(A_2，B_2)$是形式背景（$G，MI$）上的任意两个概念，定义二元关系≦：

$$C_2\leqq C_1\Leftrightarrow B_1\subseteq B_2\Leftrightarrow A_2\subseteq A_1，$$

称 C_1 是 C_2 的父概念，C_2 是 C_1 的子概念，如果不存在另外一个概念 C_3，使得 $C_2\leqq C_3\leqq C_1$，称 C_1 是 C_2 的直接父概念，C_2 是 C_1 的直接子概念，记为 $C_2\prec C_1$。

显然，关系“≤”是集合 $\beta(K)$上的一个偏序，它可诱导出 $\beta(K)$上的一个格结构，可以证明，它是一个完备格，相应的下确界和上确界定义为：

$$\wedge(At，Bt)=(\cap At，(\cup Bt)')$$

$$\vee(At，Bt)=((\cup At)'，\cap Bt)$$

其中$(At，Bt)\in\beta(K)$，T 是指标集，此完备格称为形式背景 K 的概念格。

定义 4 对于形式背景 $K=(O，A，R)$，存在唯一的一个偏序集$<H\leq，\leq>$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 $L(O，A，R)$。

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 建格算法分析

在应用概念格的过程中，概念格的构造效率始终是一大难题，人们对此进行了广泛的研究，提出了各种不同的构造算法,但只有少数的算法能够同时生成相应的 Hasse 图。这些算法主要可以分为 3 大类 ：批处理算法、渐进式算法和并行算法。批处理算法思想是首先生成所有概念，然后根据它们之间的直接驱 - 后继关系生成边，完成概念格的构造。渐进式算法思想是首先初始化概念格为空，将当前要插入的对象和现有格中所有的形式概念作交运算，根据交的结果不同采取不同的行动。概念格并行生成思想就是通过形式背景的拆分，形成分布存储的多个子背景，然后同时并行构造相应的子概念格，再由子概念格的合并得到所需的概念格。

3.1 构成形式背景

概念格的形式背景通常是由如表 1 所示的二维数来表示横向维表示属性，纵向维表示对图 1 所示的 Hasse 图。

表 1 形式背景示例

G	A	B	C	D
1	0	1	1	0
2	1	0	0	1
3	1	0	1	1
4	0	1	1	0
5	1	0	0	0

表 2 所生成的概念

编号	外延	内涵
0	{1,2,3,4,5}	{}
1	{3,5}	A
2	{3,4}	B
3	{1,2,4}	C
4	{3}	A,B
5	{4}	B,C
6	{1,2}	C,D
7	{}	A,B,C,D

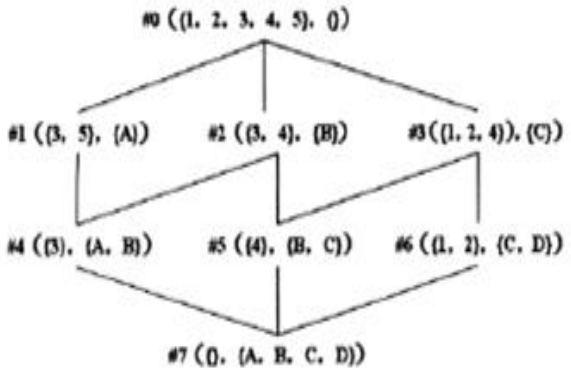

图 1 与表 1 相对的 Hasse 图

3.2 批处理构造算法

批处理构造算法是出现相对较早的一类概念格构造算法，其算法构造的主要思想是首先根据概念对象及属性生成所有的概念，然后再根据他们之间的直接父子关系生成 Hasse 图，以此来完成构造概念格的更个过程。按照其构造概念格方式的不同，可将批处理构造算法分成三类：自顶向下型、自底向上型和枚举方法。自顶向下算法的主要构造过程就是首先把概念格中的顶层节点构造出来，再根据节点间的父子关系逐层的向下建造，Bordat 算法就是典型的这种构造算法；而自底向上构造算法则就恰好相反，它是先构造概念格的最低部的节点,然后再逐层向上进行扩展，比较典型的算法就是 Chein 算法；枚举法主要是根据现有的数据集合，按照一定的规则顺序枚举出所有格中的节点,然后生成各节点之间的关系，最后根据这些节点间的关系生成概念格，代表的算法有 Nourine 算法、Ganter 算法；总的来说，这一类的算法都需要多遍的扫描数据库，因此这类构造算法的性能不是很好。

下面我们将用批处理算法中的自顶向下的构造算法来具体说明格的构造过程：

1)现建立一个只包含格顶节点的队列 *L* 和概念格 *G*;

2）生成队列 *L* 中首节点 *A1* 中所包含的全部子概念；

3）如果在上一步中产生了一个新的子概念 *C2*，之前它没有出现过，则就将 *C2* 添加到队列 *L* 和格 *G* 中去；

4）添加 *C1* 和 *C2* 之间的关系；

5）重复 2）-4）步骤直到队列 *L* 为空停止则概念格构造完成。

3.3 渐进式算法

渐进式算法又可称为是增量式算法，先将概念格初始化为空，然后再把新插入的对象与概念格中现有的所有概念做交运算，最后再通过交运算的不同结果来更新每个概念，这类算法主要是通过在格中插入记录来调整格结构的，其比较经典的算法包括 Godin 算法、Earpinet 算法等。由于这种概念格的构造算法的具有优越的时间性能，现在的大多数的概念格构造算法都是基于它而搭建的渐进式的概念格构造算法又可以分为基于对象的渐进式构造算法和基于属性的渐进式过造算法这两类，下面我们以添加对象的构造算法在介绍概念格的构造算法。下面以基于对象的渐进式构造算法 Godin 算法为例来说明渐进式构造算法的构造过程：

定义：设 *G（K）*是形式背景 *K=（A，B，I）*所对应的概念格，*C1=（A1，B1）*是格 *G* 上的任意一节点，新增对象 *S* 所对应的属性集是 *B1*，把 *S* 插入到格 *G（K）*中：

1）如果属性集 *B1* 与属性集 *B2* 的交为空，则就称 *C1* 为不变节点；

2）如果属性集 *B1* 为属性集 *B2* 的子集，则 *C1* 就是我们所要更新的节点，并将 *C1* 节点更新成 *（A1∪S，B1）*；

3）如果属性集 *B1* 与属性集 *B2* 的交不为空，并且也满足属性集 *B1* 与属性集 *B2* 的交集与格 *G（K）*中的任一节点内涵都不相同，且 *C1* 的任何父节点所对应的内涵与 *B2* 的交集不等于他们的交集，则称 *（A1∪S，B1∩B2）*为新增的节点。具体的构造如下：

步骤 1：建立空格 *G*；

步骤 2：从形式背景中取出一个对象 *O*；

步骤 3：从 *G* 中按照顺序取出概念 *C1=（A1，B1）*；

步骤 4：将 *O* 的属性集与 *A1* 所对应的属性集 *B1* 进行比对，并参照上述的定义中的规则来进行节点的更新；

步骤 5：将第 4 步骤中新生成的概念，插入到格中，并调整节点间的链接；

步骤 6：重复执行 2-5 步，到完整的概念格生成。

渐进式生成概念格的求解过程中，要着重解决三类问题： 如何生成新节点、 如何避免重复节点的产生和如何更新连接节点的边。对于上述三类问题， 谢志鹏等[8]较为详尽的论述了如何快速构造概念格。

3.4 并行构造算法

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，文献论述了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。文献［15～16］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。

4 结束语

形式概念分析现在已被广泛地应用到各个领域，而不同领域又有其特点。为了更好地把概念格应用到具体的领域，通常要对概念格模型进行必要的扩展，在原概念格基础上提出了：量化概念格、约简概念格、加权概念格、规则概念格的概念。概念知识系统与概念信息粒格理论讨论了概念信息粒之间的蕴含关系及由概念生成不确定规则的方法，给出了任意对象集和属性集构成概念信息粒的方法，以及迭代形成概念的方法。

自概念格提出以来，国内外对其理论和方法的研究愈来愈多，算法研究日益成为焦点， 相关理论的交叉应用也十分广泛。对国内外概念格的研究与发展进行了系统地总结，提出如规则提取、属性约简、 子格及商格、维护和建格算法等仍是研究的热点方向。另外，概念格与粗糙集、模糊集、本体、 语义 Web 等相关理论相结合， 发挥多个理论之间交叉融合的优势，也是很好的研究方向。并且概念格以其独特的优势正在赢得越来越多的研究者关注，从产生到现在取得了长足的发展，已经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、数据挖掘等领域。

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；不协调的形式背景，模糊形式背景的属性约简与规则提取问题；与其它知识发现理论与方法的交叉和融合问题；把粗糙集与概念格的约简理论抽象化，提取出一个统一的约简理论；多值形式背景的相关问题等。

参 考 文 献

[1]. 黄天民.格、序引论及其应用[M].成都:西南交通大学出版社,2012.

[2]. Boda J P.Calcul pratique dutreillis de galois dunecorrespon-dence[J].MathSci Humaines,2010,96(2):31-47.

[3]. CheinM.Algorithmde recherche des sous-matrices premieres dunematrice[Z].2013.

[4]. 杨 强, 赵明清.概念格研究进展[J].计算机工程与设计,2011,29(20):5293-5296.

[5]. 何淑贤,刘桂枝.形式概念分析及其应用进展[J].应用技术, 2013(5):77-79

[6]. 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,2014(10):17-23

[7]. 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

[8]. 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

[9]. 曲立平,刘大昕. 基于属性的概念格快速渐进式构造算法[J]. 计算机研究与发展,2011,44(增刊):251-256

[10]. 曲开社，翟岩慧. 偏序集、包含度与形式概念分析[J]. 计算机学报，2006，29（2）：219-226