《智能信息处理》课程作业

# 形式概念分析及其应用进展

栾利广

(大连海事大学 信息科学技术学院 辽宁 大连 1120200326)

| 作业 | 分数[20] |
| --- | --- |
| 得分 | |

2020 年 11 月 24 日

# 形式概念分析及其应用进展

栾利广

(大连海事大学 信息科学技术学院 辽宁 大连 1120200326)

**摘 要** 形式概念分析(Formal Concept Analysis，FCA)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。

**关键词** 形式概念分析，本体研究，知识发现，人工智能

# Formal Concept Analysis and Its Application Progress

**Luan Liguang**

**Abstract** Formal Concept Analysis (FCA) is a powerful tool for data Analysis and rule extraction from Formal context, which is based on mathematics, the concept lat-tice, I. E. Ontology, is constructed according to the context, which can express the structure of ontology clearly. The process of ontology construction is semi-automatic. In the stage of concept formation, domain experts are needed to identify the objects and attributes in the domain and construct the relationships between them, then, the ontology is generated automatically by using CLCA algorithm to generate concept lattice. Formal Concept Analysis, which emphasizes human cognition as the center and provides a completely different method from traditional statistical data analysis and knowledge representation, has become an important research object of artificial intelligence, it is widely used in machine learning, data mining, information retrieval and so on.

**Keywords** Formal concept analysis, Ontology research, Software engineering, Artificial intelligence

## 1. 绪论

人类在认识过程中，把所感觉具有共同特点的事物抽出来，加以概括，就称为概念。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理解，德国的 Wille 教授提出了形式概念分析(Formal Concept Analysis),用于概念的发现、排序和显示。在形式概念分析中，概念的外延被理解为属于这个概念的所有对象的集合，而内涵则被认为是所有这些对象所共有的特征(或属性)集，这实现了对概念哲学理解的形式化。所有的概念连同它们之间的泛化与例化关系则构成。

近年来我国的计算机技术得到了飞速的发展，并且成为了人们日常生活以及工作中不可或缺的一部分，人类已经进入了大数据时代，网络上各个领域各种错综复杂的信息越来越多，研究学者们发现，大规模的数据和信息蕴含了大量的信息和知识，而获取知识是人工智能领域的瓶颈问题。学术界有许多学者致力于研究数学问题实用化，以提升数学研究成果实际应用价值。为了提升应用领域实用价值，学者

们提出了形式概念分析（Formal Concept Analysis，FCA）。FCA 是一种工具，可以用来分析数据，提取规则。目前在国内，形式概念分析理论已被广泛研究，并应用到机器学习，粗糙集理论和知识获取等领域。

人类在认识过程中，把所感觉具有共同特点的事物抽出来，加以概括，就称为概念。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理解，德国的 Wille 教授提出了形式概念分析(Formal Concept Analysis), 用于概念的发现、排序和显示。在形式概念分析中，概念的外延被理解为属于这个概念的所有对象的集合，而内涵则被认为是所有这些对象所共有的特征(或属性)集，这实现了对概念哲学理解的形式化。所有的概念连同它们之间的泛化与例化关系则构成

## 2. 形式概念分析概念

### 2.1 基本概念

形式概念分析理论的主要思想源于哲学中对概念的定义。在哲学体系中概念是由外延和内涵两个部分组成的思想单元，外延被定义为属于这个概念的所有对象的集合，而内涵被定义为属于这个概念的所有对象所共同具有的属性的集合。形式概念分析理论从概念的哲学定义中得到启发，根据数据集中对象与属性之间的二元关系来获取数据中隐含的概念及其结构。所有的概念连同他们之前的泛化与例化关系可以构成一个概念格。概念格是形式概念分析的核心数据机构，其本质上描述了对象和特征之间的联系。一般认为外延是概念覆盖的实例，而内涵则是对于概念的描述，概念进一步可以通过 Hasse 图来实现可视化，通常 Hasse 图的每一个节点就代表一个概念。

早在 1940 年 Birkhoff 就已为该方法提供较好的数学理论基础；之后 Ganter 等人将其作为一个较好的数据分析方法，深化、完善该理论基础，并将它们扩展到各种现实应用中。形式概念分析提供了一种较好的层次化（形式）对象的分析方法，它能够识别那些具有共同（形式）属性的一组（形式）对象的组合。在应用形式概念分析方法的过程中，线路图的制定是非常重要的一个环节，其本身也是对于概念化的图形化表示。通过线路图能够对语境中所包含的对象和属性关系进行展示，在一些特定的语境下还包含有继承以及发展的关系，因此说形式概念分析其本质是一种准确性高以及使用范围广泛的分析模式。

我们已经知道概念的内涵与外延是关于概念的对象与属性的两个基本特征，但是它们同对象的属性和对象本身既有联系又有区别，。对于内涵来说，对象的各种特有属性或者本质属性都可以反映在特定的概念中而成为该概念的内涵，任何概念的内涵也都是反映特定兑现一定方面的特定属性或本质属性。但是，并非对象的特有属性或本质属性就是概念的内涵，而是只有当对虾干的特有属性或本质属性被反映到概念之中时，才转化为概念的内涵。对外延来说，任何事情都可以反映在特定概念中而成为概念的外延，概念的外延就是指适用于该概念的对象。同理，与概念的内涵一样，并非一般客观事物都是概念的外延，而是只有当客观事物被反映到概念之中成为其对象时，才转化为概念的外延。

#### 2.1.1 概念格的定义

概念格是 FCA 的核心数据结构。概念格的每个节点是一个概念，由外延和内涵组成。外延是概念所覆盖的实例;而内涵是概念的描述，是该概念所覆盖实例的共同特征。概念格可以通过其 Hasse 图生动简洁地体现概念之间的泛化和例化关系。概念格结构模型是形式概念分析理论中的核心数据结构。其本质上描述了对象和特征之间的联系，表明了概念之间的泛化与例化关系。这种概念格构建的过程是半自动化的。

概念格结构模型是形式概念分析理论中的核心数据结构。它本质上描述了对象

和特征之间的联系，表明了概念之间的泛化与例化关系。

形式概念分析中的概念格具有清晰的概念层次结构，是进行概念提升的有效工具，同时它非常适合对本体进行操作。作为数据分析和知识处理的形式化工具，概念格理论已被广泛地应用于软件工程知识工程等领域。

概念格(Concept Lattices)的每个节点是一个形式概念。概念格结构模型是形式概念分析中的核心数据结构，它本质上描述了对象和属性(特征)之间的关系

若$(A_1,B_1)$，$(A_2,B_2)$是某个背景下的两个概念，且$A_1 \subseteq A_2 (B_2 \subseteq B_1)$，则称$(A_1,B_1)$是$(A_2,B_2)$的子概念，$(A_2,B_2)$是$(A_1,B_1)$的超概念，并记作$(A_1,B_1) \le (A_2,B_2)$，其中关系$\le$称为概念的层次序，$(G,M,I)$上的所有概念用这种序组成的集合用$\underline{B}(G,M,I)$来表示，称它为背景$(G,M,I)$上的概念格。

2.1.2 概念格合并算法

概念格的合并可以转化为概念格的纵向合并和横向合并，本文重点讨论概念格的纵向合并，然后根据概念格的对偶原理，直接给出概念格的横向合并算法。

2.1.3 概念格构造形式背景

我们也可以从所有的概念构造出背景来。首先，确定出最大的形式概念，它的外延应是所有对象的集合，即是$G$，即是最大的形式概念应是$(G, f(G))$：其次，再确定出最小的形式概念，他的内涵应是所有属性的集合，即是$M$，也即最小的形式概念应是$(g(M),M)$，最后再由$I = U\{A \times B \mid (A,B) \in B(G,M,I)\}$，给出相关的关系$I$。

2.1.4 概念格的生成与运算

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.般可分为两类:批生成算法和渐进式生成算法(Incremental Algorithm )。

2.1.5 概念格的同构与净化背景

不同背景下的不同概念格有可能是同构的，具有相同内涵的对象和具有相同外延的属性的各自合并，并不改变该概念格的结构。

如果一个背景$(G,M,I)$是净化的，就是对于任意两个元素$g$和$h(g,h \in G)$，满足$f(g) = f(h)$时，当且仅当$g = h$；对偶地，对于任意两个元素$m$和$n(m,n \in M)$，满足$g(m) = g(n)$时，当且仅当$m = n$。在净化背景中，将概念格中同一对象的某几个属性合并为一个属性也不回影响该概念格的内容，意义与结构。

## 3. 形式概念分析在本体研究领域的应用研究

实际应用中 FCA 与本体两种形式化方法差别不大，他们都强调概念主体间一致的重要性， 都强调模式形式说明的必要。不同之处在于本体的目标在于提供一种共识，以支持知识密集型的应用，而 FCA 是在给定数据的基础上，对领域知识进行分析和结构化，是人造产物。FCA 主要依赖于所给定的对象和数据集合，而本体在没有数据的情况下

也可以建立。因此在 FCA 中，概念的外延和内涵是两个同等重要的方面，而本体则强调概念格的内涵部分。由于 FCA 与本体各有特点，目前研究主要从两个方向上进行结合:

一方面 FCA 作为一种技术应用于本体工程，FCA 以概念格给定化的数据用于提取概念层次作为本体应用的基础，用于手工或者半自动生成本体。将 FCA 引入本体生成过程中可以解决寻找概念之间的关系非常困难，手工将概念组织到本体中去费时费力和易受开发者的主观影响等问题。它以概念格来表示从给定的数据中获得的概念，帮助找到所有可能存在的概念以及概念之间的关系。

在利用 FCA 构建本体时候，如何使形式背景和本体二者对应起来是最为关键的问题，对于不同的应用一般有两种结合方式:将两者的概念等同起来，或者将本体中的概念和 FCA 中的属性相匹配。Obitko 和 Haav 采用的是第一种方式，而 Cimiano 采用的则是第二种方式。

近年来 ,本体 (Ontology)作为领域内共享概念模型的明确的形式化规范说明，凭借其在知识共享和知识重用方面的优势，人们越来越重视人工智能和知识工程领域的相关研究。以往多采用手工方法进行传统的本体构建，尽管研究人员一直致力于探寻自动或半自动构建本体的方法，但效果并不理想，直到 Wille 以形式概念分析重构 概念格理论使得 这一状况得到改变。

在本体构建相关研究中，人们普遍认为，以自然语言为基础的缺少形式化的本体尽管易于开发，但形式化的本体能够以“自动”的方式更好地被重用和共享。Obitko 等认为，定义一个好的本体不仅需要语法结构，更需要语义描述。由于形式语义是实现基于本体的自动推理的重要环节，因此使得具有良好语义特征的本体更容易在不同领域中被共享和重用。在此基础上，Obitko 等提出，利用形式概念分析能够通过构建概念格探寻潜在的对象和属性，并将现有的和潜在的实体以可视化的方式自动呈现。从而使得基于形式概念分析构建的本体在知识重用和知识共享等应用层面上比单纯的分类法具有更大的优势。Formica 以本体重用为目的，将形式概念分析和概念格应用于对现有本体的分析改造。基于给定的形式背景，通过形式概念分析构建概念格，对同一背景下的概念进行比较，以及对不同背景的概念格中的概念进行相似性评估 ,提出了概念的相似性推理，从而实现对现有本体的重用。Bendaoud 等则在形式概念分析基础上扩展出关联概念分(Relational Concept Analysis ,RCA)，将其应用于设计给定领域的真实世界本体，并以 OWL 编码，实现基于分类的推理。

在本体研究领域，以形式概念分析为基础构建的概念格凭借其在潜在对象和属性的探索功能、概念及概念间关系的呈现功能以及可视化方面的优势，被广泛地应用于本体构建、本体映射和本体合并等方面，促进了本体的重用和知识的共享。特别是将概念格理论引入数字图书馆知识组织体系构建，不仅拓宽了数字图书馆知识组织相关研究领域的视野，还开辟“概念格与本体的互补融合”等新的研究途径。从“分类与主题”到“概念格与本体的互补融合”，促使数字图书馆知识组织沿着“文献 →信息 →知识”不断向前数字图书馆递进，从表象向本质不断深入。同时，“概念格与本体的互补融合”也使实现真正意义上的知识构建成为可能，有助于突破数字图书馆知识组织研究领域的瓶颈。

## 4. 形式概念分析在知识发现领域的应用研究

国际学术界中，将形式概念分析和概念格理论应用于知识发现研究领域的杰出代表人物是德国学者 Gerd Stumme。在最近 10 余年间，这一领域的核心成果与文献几乎都出自 Stumme 及其合作者。鉴于其在该领域研究的代表性和典型性，这一部分将以 Stumme 及其合作者的相关研究为主线进行阐述。

在知识发现领域，形式概念分析与概念格不但能够提高知识挖掘的响应效率，还能在没有信息损失的前提下以直观的视图呈现规则，因此适合于大型及特大型数据库的知识发现。

## 5. 形式概念分析在软件工程领域的应用研究

最近 10 年来, 软件工程成为国外形式概念分析与概念格应用研究中一个新的热点方向。相比其他方面的应用而言，软件工程更贴近于人们的社会经济生活，因此国际上关于形式概念分析与概念格在这一领域中应用的相关研究发展迅速而突出。Tilley 等基于 IS012207 软件工程标准对形式概念分析和概念格在软件工程领域中的应用进行分析，重点梳理形式概念分析和概念格在软件工程前期和后期两个阶段的应用情况。形式概念分析和概念格在软件编码之前的应用主要包括：需求分析、组件重用、形式规范等；形式概念分析和概念格在软件编码之后的应用包括：需求矫正、环境适应、完善功能等。

通过形式概念分析，将由用户及用户行为触发的代码特征映射集合进行重构，然后在给定特征集中识别出全局和局部计算单元，通过动态与静态分析结合的方法迅速聚焦于相关的特定特征集，其效果优于早期的特征与场景的一对一的对应关系。此外，形式概念分析还被应用于软件工程不同阶段的需求连续性问题的处理上，研究表明，用形式概念分析处理软件工程不同阶段的需求连续性问题，可以取得对软件需求连续性的系统化的精确验证。

事实上，国外关于形式概念分析和概念格理论在软件工程领域中的应用远不止于上述范围。在软件工程领域，形式概念分析与概念格还被应用于组件重构、故障跟踪、设计帮助等诸多方面。这些应用使得在软件开发过程中，客观成分逐渐加大，形式化的特征逐步增强，标准化趋势更加明显。另一方面，这些应用使得软件工程对个体认知的依赖程度逐步减弱，受研发人员个体素质的影响逐渐消失，进而大大提高了软件设计、开发、维护、重用的效率和效益。

## 6. 总结

事实上，形式概念分析与概念格理论在以上 3 个方面的应用并非是绝对孤立的，而是相互交叉、相互融通的。本体的构建、映射与合并，在本质上正是借助于形式概念分析与概念格理论在概念化知识呈现与处理方面的优势，而这一优势同时又为其在知识发现领域中的应用奠定了基础，软件工程中的代码定位、组件重构等应用一定程度上体现的就是面向软件开发与维护领域的知识重用。

随着对形式概念分析与概念格理论应用研究的不断深入，其发展前沿和研究热点也不会一成不变，本文所列的 3 个方面只是形式概念分析与概念格理论应用研究中诸多分支中的一部分，并不足以囊括国际上这一领域研究成果的全貌。在今后的研究中，仍然需要
断跟踪和把握国际学术界的发展前沿和研究热点。

**参考文献**

[1]Wille R. Restructuring Lattice Theory :An Approach Based on Hi-erarchies of Concept[ C] . In:Proceedings of the 7th InternationalConference on Formal Concept Analysis. Berlin:Springer-Verlag ,2009 :314 - 339.

[2]何丹丹.形式概念分析在软件工程中的应用 [J]. 计算机光盘软件与应用， 2014（2）： 138~139.

[3]胡鑫.形式概念分析在软件工程中的应用 [J]. 电脑迷， 2016（5）：39.

[4] Ganter B, Wille R. Applied Lattice Theory :Formal Concept Anal-ysis[EB/OL ]. [ 2010 - 09 -29 ] . http:/ /citeseerx. ist. psu. edu/viewdoc /download ！doi =10. 1. 1. 42. 9907＆rep =rep1＆type =pdf.

[5]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,(11):17-23.

[6]张文修，魏玲，祁建军.概念格的属性约简理论与方法 [M]. 中国科学 E 辑：信息科学，2005,

35(6):Pages: 628-639.
[7]盛秋艳，刘群，一种基于本体的叙词语义描述方法 情报科学 第 25 卷第 9 期，2007 年 9 月
[8]韩婕，向阳.本体构建研究综述.计算机应用与软件，2007.9 Vol.24 No.9．
[9]贺晓丽，刘华丽，刘瑶瑶.多粒度数据的区间形式概念分析方法[J].计算机工程与应用，2019,55(19)，52-57.