# 基于形式概念分析本体构建方法的研究

祁美晶

（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(Formal Concept Analysis，FCA)是由德国教授Wille在1982年提出来的，常用在概念发现、排序以及显示中。形式概念分析的思想主要来源于哲学，而概念格作为形式概念分析的核心数据结构，在规则提取与数据分析方面有着广泛的应用。形式背景是形式概念分析的一个重要元素。利用形式背景生成概念格，再运用概念格的构造算法可自动产生本体。本文采用形式概念分析的方法来构建本体，简述了有关形式概念分析及本体的概念，介绍了概念格的相关构造算法以及本体构造的过程。通过概念格图形的形式来展现本体的研究领域中概念及概念之间的关系，寻找所有隐含概念及概念间的关系，从而清楚的表达出本体的结构。

**关键词** 形式概念分析；本体；形式背景；概念格

**中图法分类号** TP311 **文献标识码** A

## Based On Formal Concept Analysis Research of

## Ontology Construction Method

QI Mei-Jing

(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 in 1982, commonly found in the concept, sort and display. Formal Concept Analysis comes mainly from the philosophical ideas, but as an core data structure of the Formal Concept Analysis, which has been widely used in rules extraction and data analysis. Formal Context is an important element of Formal Concept Analysis.The concept lattice is generated by formal context and using the construction algorithm of concept lattice can generate ontology automatically.This article adopts method of formal concept analysis to construct the ontology and sketches The concepts of formal concept analysis and ontology.The related construction algorithms of concept lattice and the process of ontology construction are introduced.Through the form of concept lattice figure shows the relationship between the concept in the field of ontology research,it is to find all the implicit concept and the relationship between the concept ,so clearly expressed ontology structure.

**Keywords** Formal Concept;Ontology;Analysis;Formal Context;Concept Lattice

## 1 引言

本体的本质是共享概念模型的明确的形式化的规范说明,其目标[1]是通过概念模型对信息作完全的形式化描述，提供对该领域知识的共同理解，并从不同层次的形式化模式上给出这些信息的互相关系，使计算机可以理解并处理网上的信息。因此，构建本体就成为本体应用的关键的问题，目前该领域的研究还处于探索阶段，没有形成成熟、统一的方法。文中探讨了基于形式概念

收稿日期：2016-10-12
作者简介：祁美晶（1994-）女，甘肃敦煌人，硕士生在读。

分析(Formal Concept Analysis，FCA)的理论来构建本体。形式概念分析能从形式背景中发现概念结构，生成概念格，具有明确的层次关系和丰富的语义信息。概念格的 Hassse 图能清晰地表达概念之间的层次关系，即本体的层次，从而清楚地表达出本体的结构。

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 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。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。

本文首先介绍了有关形式概念分析及本体的基本概念，又介绍了概念格的相关算法，然后以形式概念分析的方法来构建本体，根据构建本体的步骤来演示构建的过程。

## 2 形式概念分析和本体中的有关概念

### 2.1 形式概念分析

形式概念分析(FCA)是一种从形式背景进行数据分析和规则提取的强有力工具。它是信息处理的一种理论，是应用数学的一个分支，建立在数学基础上，对组成本体的概念、属性以及关系等用形式化的语境表达出来，然后根据语境，构造出概念格(concept lattice)。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。

在形式概念分析中，数据是用形式背景表示的。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

形式背景是形式概念分析理论中的一个重要元素，也是形式概念分析的基础，是用于表达和记录对象与属性之间二元关系的数据载体。

**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 $B'=A'$，则称(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

概念格是形式概念分析的核心数据结构。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。它本质上描述了对象和属性之间的关系。另外，概念格可以通过 Hasse 图生动简洁地体现概念之间的泛化和例化关系。因此，概念格被认为是进行数据分析的有力工具。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。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了成功的应用。

概念格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。它是一种具有完备性的结构，首先在概念表示方面，每个概念用格节点的形式表示，一个格节点允许有多个父节点，彼此相连的节点间具有某种偏序关系，这种结构非常适合于综合相似信息并用于信息比较与浏览；其次在特征信息综合方面，形势概念

分析本身就是概念聚类技术，它将所有内涵相同的概念聚类并为每个概念类提供内涵描述，这些描述可使分类更加清晰明确。

**定义 3** 对于形式背景 K=(O，D，R)，其中 O 是事例集合，D 是描述符（属性）集合，R 是 O 和 D 之间的一个二元关系，则存在唯一的一个偏序集合与之对应，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，这个偏序集合产生一种格结构，这个格结构就称为概念格(concept lattice)，记为 L(O，D，R)。格 L 中的每个节点是一个序偶（称为概念）。

### 2.4 本体

Gruber 于 1993 年给出了 Ontology 的定义[2]本体是对概念模型明确的形式化说明，概念可以被理解为对世界或领域的抽象描述。文献[3]中总结了 Ontology 的 5 个基本建模元语。这些元语分别为: 类( classes )，关系( relations )，函数 (functions)， 公 理 (axioms) 和 实 例 ( instances )。它的逻辑结构可看成一个五元组，O={C,R,H,rel,A}[4]。其逻辑结构为:

(1) 两个交集为空的集合 C 和 R。它们的元素分别被称为概念标识符和关系标识符。

(2) 概念层次 H 是一个有向的传递关系，H 是 C×C 的子集。H(C1,C2)表示 C1 是 C2 的子概念。

(3) 函数 rel:rel 的定义域是 R,值域是 C×C 的一个子集，即 rel(R)=(C1,C2)。

(4) 公理集 A 包含了本体所需的公理，它用逻辑语言表示。

## 3 概念格构造算法

通过阅读相关文献，了解到概念格的构造过程就是概念聚类的过程。对于具有相同形式背景的数据，可以生成惟一的格结构，不受数据或属性排列次序的影响。国内外已提出很多关于概念格的建格算法，这些算法大致可以分成三大类：批处理算法、渐进式算法（或称增量算法）和并行算法。开发人员可以通过这些算法快速的构造概念格，使之提高平时的开发效率和减少维护成本。这里主要介绍 Godin 算法。

渐进式算法的主要思想是将待插入的对象与格内已存在的概念节点进行交运算，根据结果的不同使用相应的处理办法。对于新插入的实例，对格内的节点会产生以下三种不同的影响：①更新节点，该类节点内涵包含在新对象内涵之中，仅仅需要将新对象的外延加入到外延中即可；②不变节点，这种结点的内涵与新对象的内涵无关（没有任何交集），不做任何修改；③新增节点，新节点对象的内涵与格内节点内涵的交集首次出现，即原格内所没有的新概念需要添加的节点。在渐进式算法中，较经典的算法是 Godin 算法，该算法在新对象插入时，用遍历所有的节点，仅仅检查是否至少和新对象有一个共同属性的节点。该操作通过维护一个可包含每个属性首次在格内出现的指针来实现，该指针能自顶而下进行深度优先搜索。仅仅检查是否至少和新对象有一个共同属性的节点。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。

Godin 算法是 Godin.R 等在 1995 年提出的概念格生成算法，其算法的过程如下：

(1) 初始化格 L 为一个空格；

(2) 从 G 中取一个对象 g；

(3) 对于概念格 L 中的每个概念 C1=(A1,B1)，如果$B_1 \subseteq f(g)$,则把 g 并到 A1 中如果同时满足：$B_1 \cap f(g) \neq \emptyset$，$B_1 \cap f(g) \neq B_1$和不存在(A1,B1)的某个父节点(A2, B2)满足$B_1 \cap f(g) \subseteq B_2$，则要产生一个新节点；

(4) 把新产生的节点加入到 L 中，同时调整节点之间的链接关系；

(5) 重复(2)到(5)，直至形式背景中的对象处理结束；

(6) 输出概念格 L。

概念格的渐进式生成算法在产生所有概念节点的同时，还产生了概念之间的父概念－子概念连接关系，同时它非常适合于处理动态数据库，被认为是一种生命力很强的概念格生成算法。

人们对 Godin 算法的改进也没有停止过。谢志鹏等[5]提出了一种利用字典索引树的快速概念格渐进式构造算法，该算法利用一个辅助索引树来快速判断概念节点的类型，并根据概念节点的类型来决定概念格的渐进修改策略。

## 4 用形式概念分析的方法构建本体

### 4.1 基于形式概念分析的本体构建步骤

(1)利用自然语言理解技术对收集来的纯文本进行预处理，取得文本中的字词集合，在字词

集合中按照一定的方法获得能表达文本的关键概念词汇及其关系。再针对所找出的概念词汇给出定义描述并用准确的词汇表达出来，结合相应的文本集合形成词汇、文件的二元关系表。

(2)把上一阶段获取的概念用形式化的语言明确表示出来。对存在多值的二元关系表，转换成单值的二元关系表即单值形式背景。再由单值形式背景按照造格方法来构造概念格。在实际应用中，属性抽取得到的形式背景往往是多值的，而形式背景的约简和概念格的构造都是从单值形式背景出发的，因此需要将多值形式背景转化成单值形式背景。

(3)探讨如何将概念格转换成相应的本体。由于这里的属性都是词汇，而本体所描述的也都是词汇概念，因此，可以用概念格中的属性来表示本体中的概念，使得本体的构建与展现清晰易懂。如图1 所示。

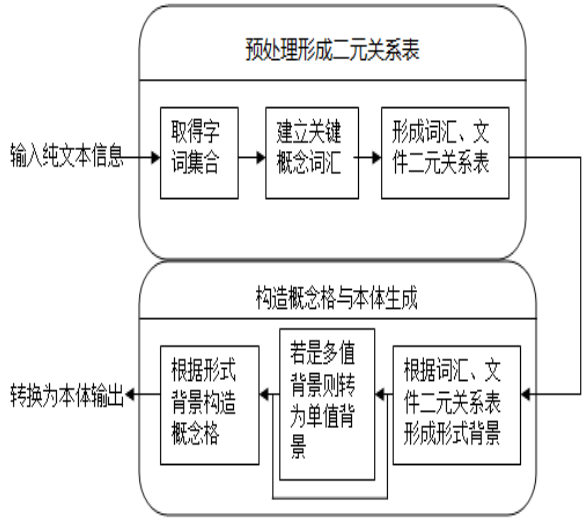


图 1 基于 FCA 的本体构建流程图

### 4.2 基于形式概念分析的本体构建实例过程

假设已经筛选出能代表文件的词汇集合，将文件集合与词汇集合做对应，某文件内含有某词汇，则标注上该文件包含的词汇个数， 如此可以产生带属性值的二元关系矩阵如表 1。本体的形式背景定义为 K，本体的相关文件集合定义为 U，本体的相关词汇集合定义为 D，文件与词汇的相互关系则为 I，因此以上的关系式可以表示为 K=(U,D,I),U={Tea,Cola,Beer,Wine,Coffee,Champagne,Mineralwater},D={ Non-alcoholic,Hot, Alcoholic, Caffine, Sparking}。这里的形式背景是已经转化后的单值形式背景。

表 1 单值形式背景

| 词汇 / 文件 | Non-alcoholic | Hot | Alcoholic | Caffine | Sparking |
|---|---|---|---|---|---|
| Tea | √ | √ | | | |
| Cola | √ | | | √ | √ |
| Beer | | | √ | | √ |
| Wine | | | √ | | |
| Coffee | √ | √ | | √ | |
| Champagne | | | √ | | √ |
| Mineral water | √ | | | | √ |

根据表1 采用经典的Godin 算法进行概念格的构造，生成的概念格如图2所示。

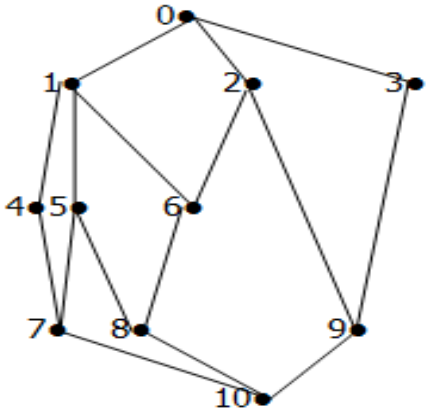


图 2 表 1 对应的概念格

图2中，每个概念节点均具有具体的内容。如果以对偶“(<{外延}、{内涵}>)”的形式来表示每个概念节点的具体内容，那么图2所示的每个概念节点的具体内容如表2 所示。

表 2 图 2 所对应的形式概念

| 标号 | 形式概念 |
|---|---|
| 0 | <{tea, coffee, mineral water, wine, beer, cola, champagne},{}> |
| 1 | <{tea, coffee, mineral water, cola},{non-alcoholic}> |
| 2 | <{mineral water, beer, cola, champagne},{sparking}> |
| 3 | <{wine, beer, champagne},{alcoholic}> |
| 4 | <{tea, coffee},{non-alcoholic, hot}> |
| 5 | <{coffee, cola},{non-alcoholic, caffine}> |
| 6 | <{mineral water, cola},{non-alcoholic, sparking}> |
| 7 | <{coffee},{non-alcoholic, hot, caffine}> |
| 8 | <{cola},{non-alcoholic, caffine, sparking}> |
| 9 | <{beer, champagne},{alcoholic, sparking}> |
| 10 | <{},{non-alcoholic, hot, alcoholic, caffine, sparking}> |

## 5 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。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、计算机网络、数据分析、决策分析、数据挖掘等领域。

要构建本体，应先分离出概念，建立概念之间的关系，这里离不开领域专家的参与；然后可以构建数据库模式(或者建立XML文件，再转换成关系数据表)；其次还要做一个多值语境到单值语境的转换，这一步必须对数据信息的数值做分类处理，可以是人工的，也可以借助工具；最后运用Godin算法构建概念格。可以看出，虽然本体的开发过程依然离不开人的因素，但概念格作为本体的构建方式，清楚表达了概念以及概念之间的关系，而且容易为人们所理解。

研究发现传统的本体构建方法在各方面存在着缺陷,而形式概念分析能很好地突破这些限制，故成为本体构造的可行方法。本文采取 FCA 方法构造本体，清楚地表达了本体的层次构架和概念间相关性等。结合实例阐明了本体从构造到展现的全过程，从一定程度上提高了本体应用在信息检索中的查找效率。但是在处理多值形似背景的时，系统的时间复杂度和空间复杂度将会成成倍增加，因此寻找一种更优化的算法将是今后进一步的研究目标;同时如果能在概念之间加入权重，计算出不同的概念之间的权重关系，得到相关性的排序，这还需要进一步的研究。

## 参 考 文 献


[1] 宋炜，张铭.语义网简明教程.北京:高等教育出版社,2004.108～131.

[2] Gruber T R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing[J]. International Journal of Human-computer Studies，1995，43(5):907-928

[3] Gómez-Pérez A,Benjamins R.Overview of Knowledge Sharing and Reuse Components: Ontologies and Problemsolving Methods[C]//Proceedings of International Joint Conferenceon Artificial Intelligence. Stockholm，Sweden:[s.n.],1999: 1-15.

[4] Salton G. Introduction to modern information retrieval [M].New York : McGraw Hill Book Co. , 1983.

[5] 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

[6] 黄伟，金远平.形式概念分析在本体构建中的应用 [J].微机发展,2005,15(2):28-31

[7] 郑珂,李涵.基于形式概念分析的本体构建方法研究 [J].福建电脑,2011,2:61-62

[8] 韩道军,甘甜,叶曼曼,等.基于形式概念分析的本体构建方法研究[J].计算机工程,2016,42(2): 301-302

[9] Ganter B,Wille R.Formal concept analysis: mathematical foundations[M].Berlin: Springer Verlag, 1999.

[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] 曲立平,刘大昕. 基于属性的概念格快速渐进式构造造算法[J].计算机研究与发展,2007,44(增刊):251-256

[12] 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

[13] 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(10):17-23