《智能信息处理》课程考试

基于形式概念分析本体构建方法研究

王芳铭

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 10 日

基于形式概念分析本体构建方法研究

王芳铭

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘 要：随着信息共享和数据交换的不断扩大,作为一种表现概念层次结构和语义的模型,本体越来越多地被应用到计算机科学的众多领域中。如何自动或半自动构建有效且实用的领域本体,即本体学习,成为目前的研究热点问题。本体的出现能很好地解决计算机应用领城中存在的一些困难，例如人机交互、自动推理、知识表示等。针对传统本体构建方法依靠人工费时费力、主观干扰较大、对隐含概念和关系提取不足等问题，提出基于形式概念分析构建本体的方法。本文对现有的本体构建方法进行了初步总结，并在此基础上详细描述了几种基于形式概念分析的本体构建方法.对形式概念分析用于本体构建方法及相关的概念格算法做了分析、比较和总结，最后对本体在各个方面的应用进行了简要分析。

关键词:本体；形式概念；本体构建

Ontology Research Based on Formal Concept Analysis

Wang Fangming

（Dalian Maritime University　Dalian, Liaoning Department of Computer Technology）

Abstract: With the continuous expansion of information sharing and data exchange, as a model of expressing conceptual hierarchical structure and semantics, ontology has been increasingly used in many fields of computer science. How to automatically or semi-automatically construct an effective and practical domain ontology, namely ontology learning, has become a current research hotspot. The emergence of ontology can well solve some of the difficulties that exist in the domain of computer applications, such as human-computer interaction, automatic reasoning, and knowledge representation. Aiming at the problems of traditional ontology construction methods that rely on labor, time and effort, subjective interference, and insufficient extraction of implicit concepts and relationships, a method of constructing ontology based on formal concept analysis is proposed. This article briefly summarizes the existing ontology construction methods, and on this basis, describes in detail several ontology construction methods based on formal concept analysis. The formal concept analysis used in ontology construction methods and related concept lattice algorithms are analyzed. , Comparison and summary, and finally a brief analysis of the application of ontology in various aspects.

Keyword: Ontology; formal concept; ontology construction

1 概述

本体最早是一个哲学上的概念。从哲学的范畴来说，本体是客观存在一个系统的解释或说明，关注的是客观的抽象本质。本体定义:本体是共享概念模型的明确的形式化规范说明。本体体现的是共同认可的知识，反映的是相关领域中公认的概念集，它所针对的是团体而不是个体。本体的目标是捕获相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的概念，并从不同层次的形式化模式上给出这些概念(术

语)和概念之间相互关系的明确定义。

形式概念分析(formal concept analysis)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 lattice)，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，自动产生本体。

本文首先介绍了有关形式概念分析及本体的基本概念，又介绍了用形式概念分析来进行本体构建的方法，最后介绍了概念格生成的相关算法及基于形式概念分析本体的应用。

2 形式概念分析的相关定义

形式概念分析作为一种数学理论被提出的，是人们组织和分析数据的一种方法，将数据及其结构、本质以及依赖关系进行形象化的一种描述。那么，对现实世界中的概念和背景在形式概念分析时就会形成形式概念和形式背景。

定义 **1** 设形式对象集:U(X属于U)，形式属性集：A（B 属于 A）二元关系 R 属于（U X A）。若 X={x|x 属于 U，对任意的 a 属于 B，x R a}，B={a|a 属于 A，对任意的 x 属于 X，x R a}，则 二元组（X，B）被称为形式概念。X 中 x 每个 x 都有全部属性，B 中 a 每个 x 都有的属性。

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 **2** 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 **3** 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 3 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

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。

定义 **4** 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 **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 本体的定义

Gruber 于 1993 年给出了 Ontology 的定义，本体是概念模型的明确的规范说明，是共享概念模型的形式化规范说明，概念可以被理解为对世界或领域的抽象描述。类，关系，函数，公理和实例是 Ontology 的 5 个基本建模元语。通常也 classes 写成 concepts；概念可以指任何事物；关系表示概念间的相互作用；函数是一种特殊的关系，表示前 n-1 个元素唯一确定第 n 个元素；公理表示永真断言；实例表示元素。

本体的结构可以表示为 O:（C, ≤C, R, β, ≤R) 其中，C 和 R 分别表示概念集合和关系集合；C 上的偏序关系≤C 叫做概念层级；函数: β:R→C+，定义域是 R ，值域是 C ×C ； R 上的偏序集 ≤R 是关系层级。

本体的目标是获取、描述和表示相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并且在不同层次的形式化模式上给出这些词语和词语之间相互关系的明确定义，本体论是反省客观存在的概念模型。用一个公式表示：本体=概念+关系+函数+公理+实例。

尽管本体的功能已经被认识，并逐渐得到应用，但目前本体构建的研究还处于探索阶段，没有形成成熟，统一的方法作为指导，不能满足本体应用的需求。目前仅有的一些方法也是从个案的开发过 程中通过逆向工程总结出来的，并存在着很多不足和固有的局限性，这就产生了一对矛盾：本体技术在信息科学中的广泛应用在客观上要求必须有先进的本体构建技术来对其支撑，而当前的构建方法由于各自种种的缺陷并不能很好的满足这种客观需求， 因此，要解决这对矛盾，就必须在本体构建方法方面做出积极的探索，领域本体的构建方法，必须向标准化，工程化，规范化，细节化的方向来发展。

就国内外当前的情况来看，将形式概念分析理论应用于领域本体的构建理论都尚在起步阶段，形式概念分析与领域本体结合的方式不尽相同，结合的深度各异，尽管有些理论存在着一定的可取之处，但现有的结合方式仍存在着很多不足，将形式概念分析引入到领域本体构建方法当中，还有大量的工作要做，还需深入地进行研究。

4 本体的构建方法

本体表示的是现实世界的模型，因此建立的本体必须能够客观反映现实。因此本体的构建应该是一个反复迭代的过程，这个过程将贯穿于本体的整个生命周期。

首先要明确构建的本体将覆盖的专业领域、应将本体的目的、作用以及它的系统开发，维护和应用对象，这些对于领域本体的建立过程中有着很大的关系，所以应当在开发本体前注意，对于特定的专业领域的一些特殊的表达法和特定的详细内容的注释，应当明确。

1）在领域本体创建的初始阶段，尽可能列举出系统想要陈述的或要向用户解释的所有概念。这上面的概念和术语是需要声明或解释的。而不必在意所要表达的概念之间的意思是否重叠，也不要考虑这些概念到底用何种方式（类、属性还是实例）来表达。对领域文档的预处理，依据“对象-属性”关系，生成领域形式背景，依据单值背景，构造概念格（Hasse 图），将概念格进行转换，处理，得到领域本体概念层次模型。

2）设计元本体，定义领域中概念与概念的关系。尽量做到领域无关性，并且包含的元概念数量尽可能少。概念的定义可以使用元本体中定义的元概念，也可重用已有本体。

3）定义类和类层次，领域中的概念主要用类来描述。而对类层次的定义有以下 3 种方法：自上向下法、自下向上法、混合法

4）定义类的属性和属性约束，仅仅定义类还不足以描述整个领域，还要进一步描述类的内部结构。本体中用属性来描述类的内部结构。属性类也有属性，即属性约束。

5）对领域本体编码

6）对领域本体模型进行扩充。对领域本体原型的扩充是在领域专家的参与下完成的。包括：对领域本体属性、实例、公理、非分类关系的扩充。

7）对领域本体的复用。通过本体的映射，发现领域本体间是否存在着关系。若存在，就可以在本体映射的基础上，进行同域本体的合并或者结盟，形成新的领域本体；对概念格的复用，概念格是在构建领域本体的过程中形成的中间产物，在二次开发过程中，可直接应用已有的领域概念格，将之与新的开发过程中的概念格进行合并；对形式背景和领域关键概念的复用

8）本体的检验评价，本体形式化以后，是否满足了我们刚开始提出的需求、是否满足本体的建立准则、本体中的术语是否被清晰定义、本体中的概念及其关系是否完整等问题都需要我们在本体建立过程后进行检验和评估。

5. 基于 FCA 概念格的相关算法

为了使本体构建更加的贴近实际且实用性强，构建本体有许多的规则：清晰性；完全性；一致性；可扩展性；最小承诺（只定义必要的术语，只定义约束最弱的关系）；最小编码偏好（不指定术语形式化用何编码）。

构建本体的层次可以划分为数据源/技术层，处理层，输出层，并且这 5 种方法中所使用的数据源和技术，在数据源/技术层-处理层以及处理层-输出层之间建立联系。

5.1 Uschold 和 King 方法

建立在企业本体基础之上，是相关企业间术语和定义的集合，该方法只提供开发本体的指导方针，目前企业本体在爱丁堡大学人工智能研究所及他的合作伙伴，具体的步骤：

1）确定本体应用的目的和范围：根据所研究的领域和任务，建立相应的领域本体或过程本体，领域越大，所建立本体越大，因此需限制研究的范围。

2）本体分析：定义本体所有术语的意义及其之间的关系，该步骤需领域专家的参与，对该领域越了解，所建的本体就越完善。

3）本体表示：一般用语义模型表示本体。

4）本体评价：建立本体的评价标准是清晰性、一致性、完善性、可扩展性。清晰性就是本体中的术语应该无二义性；一致性指的是本体中的关系逻辑上应当一致；完整性指的是本体中的概念以及关系应当是完整的，应该包括该领域的所有概念，但是很难达到，需要不断完善；可扩展性本体应用能够扩展，在该领域不断发展时能加入新的概念。

5）本体的建立：对所有本体按照以上标准进行检验，符合要求的文件的形式存放，否则转到（2）。

5.2 Gruninger 和 Fox 方法

定义直接可能的应用和所有解决方案，并且提供潜在非形式化的对象和关系的语义表示。

1）能力问题作为约束的条件，把能够解决什么样的问题和这个问题应当怎样去解

决，这里的问题用属于表示，答案用公理和形式化定义回答，由于在没有形式化Ontology之前进行的，所以叫非形式化的能力问题。

2）术语的规范化：从非形式化能力问题中提取非形式化的术语，然后用Ontology形式化语言进行定义。

3）形式化的能力问题：一旦能力问题脱离了非形式化，Ontology术语已经定义，则能力问题自然形式化了。

4）形式化公理：术语定义所遵循的公理用一阶谓词逻辑表示，其中包括定义和语义或者解释。

5）说明问题的解决方案必须是完全的。

5.3 Berneras 方法

这种方法开发本体由应用开发控制。所以每一个应用都有相应的表示该应用所需的 Ontology。这些本体既能重用其他的Ontology，也能被后继应用集成，应用于电子网络的开发。开发过程如下：

1）应用的说明：提供应用的上下文和应用模型所需要的组件。

2）相关本体论范畴的初步设计：首先搜索已经存在的Ontologies,然后进行抽象、提炼、扩充。

3）Ontology 的构造：最小关联原则用来确保模型既能相互依赖，有尽可能一致，一直得到最大同构。

5.4 Methontology 方法

这种方法由马德里大学工艺分校开发人工智能图书馆使用。它分为三个阶段：

1）管理阶段：这一阶段的系统规划规则包括任务的进展情况、需要的资源、如何保证质量等问题。

2）开发阶段：规范说明 y 概念化 y 形式化 y 执行 y 维护。

3）维护阶段：包括知识的获取、系统的集成、评价、文档的说明与解释、配置的管理与改善。

5.5 基于 Sensus 方法

这个 Ontology 用于自然语言程序，有ISI 自然语言组企图为机器翻译提供广泛的概念结构，共有 5 万多个概念，为了能在Sensus 基础上构造特定领域的 Ontology，必须把不相关的术语从 Sensus 中剪出掉，具体过程如下：

1） 定义/叶子术语；

2）把叶子属于手工的和 Senses 术语相连；

3）找出叶子节点到 Sensus 根的路；

4）增加和域相关并且没有出现的概念；

5）用启发式思维找出全部的特定的域的术语：对于某些有两条以上路经过的结点必须是一颗子树的父节点，那么这颗子树以上的所有结点都和该域相关，是要增加的术语。对于高层结点通常有多条路经过，则很难判断。

6 FCA 用于本体构建的方法分析

国外再将基于形式概念分析用于本体构建方面已经取得一些的成果，具有代表性的方法主要：Cimiano 的方法 ，GuTao的方法，Haav的方法，Marck Obitko方法，将形式概念分析用于领域本体构建现在仍处在一个探索的阶段。

6.1 Cimiano 的方法

Cimiano等提出了一种采用形式概念分析分析词语在文本中的使用方式来获得相应的背景知识进而生成本体的领域本体构建方法。该方法的基本思想是：首先，使用一个自然语言的解析器，通过该解析器从领域本体中的每一个句子都可以得到一颗语 法树；其次，由语法树直接得到动词对象间的依赖关系；再次，通过进一步的词典查询，对提取的动词和对象用词进行规范化表示。最后，将形式概念分析中的概念和本体中的概念直接等同，得到概念格，由概念格得到领域本体。

6.2 GuTao 的方法

GuTao 提出的形式概念分析用于本体构建的方法如图2所示：

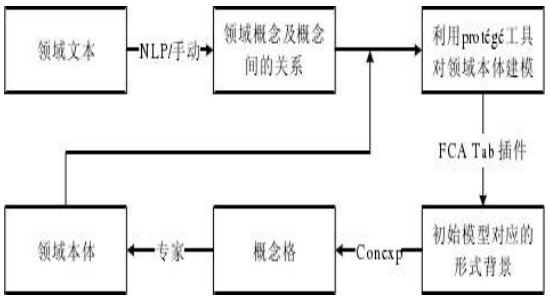

图2 方法

6.3 Haav 的方法

Haav 提出了将形式概念分析与基于规则的语言相结合，进行半自动的领域本体构建的方法，该方法适应的两个前提是：领域文本内容比较短的情况；假设领域本体描述了某一实体，里面包含了描述领域的术语。该方法将概念格节点的属性作为本体中的概念，对本体的属性表达也就成为空缺，不利于表达领域本体属性间的关系。该方法在形式背景的建立上存在不合理性。

6.4 Marek Obitko 的方法

Marek Obitko 等人在GACR 项目中提出了一种将形式概念分析用于领域本体构建 的方法，该方法约定：概念由属性来描述；属性决定概念的层次结构；当两个概念的属性相同时，这两个概念也相同；可以直接有修改过的概念格作为本体表示。基于以上规则，该方法的具体步骤如下：

1）方法的使用从空的对象和属性开始；

2）由领域本体开发人员根据需要把对象和属性添加到形式背景当中；

3）显示形式背景所对应的概念格；

在可视化基础上，按需要做如下操作：一是直接编辑，包含增加或删除对象，增加或删除属性，给对象增加属性或从对象中删除属性；二是根据提示编辑概念格，例如两个对象 在概念格上同一个位置，要么被对象添加属性，以区别对象。另外，FCA 可以产生新的对象，它们不在形式背景中，可以增加这些概念。最后重复这一过程，直到领域本体开发人员满意为止。

Marek Obitko的方法只适合小领域本体的构建，总的来说，该方法的自动化程度较低。

7.本体的应用研究

7.1 语义 Web

本体技术对语义 Web 研究、应用起重要推动作用。李曼等用领域本体及其推理能力生成优化的服务组合图，提出 Web 服务动态组合方法；吴健等基于词语间距离度量、相似度两种词汇语义相似度算法，提出基于本体论和词汇语义相似度的 Web 服务发现方法；彭晖等提出基于本体概念相似度（通过概念间语义距离计算）描述服务请求方和发布方的 Web 服务匹配算法；邓志鸿等分析本体在 Web 信息集成中应用；周明建等提出基于 Web 页面的信息项本体和结构项本体的信息提取规则以有效提取 Web 页面信息；袁柳等提出基于领域本体语义标注 Web 数据库查询结果以便机器处理和用户理解。

7.2 信息检索

本体技术可有效提高信息检索系统性能。廖明宏提出基于本体信息检索方法；徐振宁提出基于本体的智能化、个性化语义信息检索系统体系结构；武成岗等提出基于本体和多智能主体、对用户需求及 Internet 信息进行领域分类的信息检索服务框架以提高检索结果针对性；万捷等基于本体将用户检索需求扩充成语义集并通过文档分析器过滤检索结果以提高检索质量；丁晟春等分析 Jena 在语义检索中作用与应用；廖乐健等从知识表示与推理角度提出基于本体与模板规则混合技术，混合本体和树形模板以

增强模板语义表达能力，提高信息抽取智能性。

7.3 知识工程

顾芳概述了知识工程中本体构建准则、方法、表示语言、代表项目、主要应用等；王英林提出基于本体重构知识管理系统框架，解决知识管理中知识类型不可扩充局限性；郭鸣提出基于本体和语义 Web、支持知识处理的结构层次化产品信息模型并给出从 EXPRESS 模式到 DAML+OIL 映射方法；胡玉杰等基于产品知识表达模型构建流本体和功能本体，进而定义特定领域产品共享、通用知识并提出基于本体的产品知识表达应用模型和集成框架。

7.4 其他研究

主要是本体在关系数据库与本体库中不同存储格式间转换问题。李曼提出将常用本体查询信息按类分别存于不同表以减少本体查询时表连接代价；徐振宁将知识表示和处理引入到 Web 信息处理，为半结构化 Web 数据和关系数据库提供统一语义模型，实现基于数据库的 Web 信息动态发布与多数据源集成。

结束语

由于本体提供通信双方的公共理解，类似于网络协议在通信双方的地位，只不过本体是人工智能角度出发构造的软件。正由于此，国外研究本体异常活跃，国内则处于起步阶段。本体的范围从形成所有领域的知识表达的基础的非常通用的术语到限定特定知识领域的专业术语，本体可以应用于许多领域，如：电子工程、化学、远程教育、电子商务等。信息检索系统、数字图书馆、易购信息的集成、以及 Internet 搜索引擎都需要领域本体来组织信息和指导搜索过程。

现有的教学系统缺乏知识的工程化，针对教学系统进行设计的常用词汇表和框架，在恰当的抽象层次上给智能教育的任务加以形式化。使学习者通过教学系统发现自己的不足，实现异构的自治系统间的互操作，设计本体，充分利用现有的教学资源使形式概念分析更加实用，将其作用尽可能最大化。

参考文献

[1][德]B. 甘特尔, R. 威尔. 形式概念分析[M]. 马垣, 张学东等译. 北京：科学出版社, 2007.

[2] R.Wille, "Methods of conceptual knowledge processing", Formal Concept Analysis 4th International Conference ICFCA 2006, vol.3874, pp. 1-29, February 13-17, 2006.

[3] 王亚慧, 李端明, 王萝娜, 等. 基于 FCA 与概念格属性约简的本体合并方法研究[J]. 情报科学, 2018.

[4] 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(2):300-306.

[5] 龚雪. 基于形式概念分析的本体学习方法研究[D]. 吉林大学.

[6] Mike Uschold,Michael Gruninget The Knowledge Engineering Review[J]1 1996,11(2):93

[7] Cimiano P, Staab S, Tane J. Automatic Acquisition of Taxonomies from Text:FCA meets NLP. In:Proc.of the Intl.Workshop on Adaptive TextExtraction and Mining, 2003, 10一17

[8] Gu Tao. Using formal concept analysis for ontology structuring and building.ICIS, Nanyang Technological University, 2003

[9]Haav H M. A semi-automatic method to ontology design by using FCA.In:Snasel V, Belohlavek R, eds. Concept Lattices and their Appfications.Proceedingsof the 2nd International CLA Workshop, TU of Ostrava, 2004, 13-2_5

[10]刘树鹏，李冠宇. 基于形式概念分析的本体合并方法[J]. 计算机工程与设计，2011(04):09-121.

[11] 张文秀，朱庆华. 领域本体的构建方法研究[J]. 图书与情报，2011，（1）：16-19, 40.

[12]滕广青，毕强. 国外本体协调研究前沿进展及热点分析[J]. 中国图书馆学报，2012，（1）：113-120.

[13]张云中. 一种基于本体构建方法[J]. 现代图书馆情报技术，2011，27（12）：15-23

[14]刘萍，高慧琴，胡月红. 基于形式概念分析的情报学领域本体构建[J]