基于形式概念分析的领域本体构建方法

杨姿
(大连海事大学 信息科学技术学院，辽宁 大连 116026)

摘 要：本体作为语义网的基础，越来越受到人们的关注，也被越来越多的领域所应用。本体的出现是计算机领域的一大贡献，解决了目前计算机领域存在的一些问题，例如语义异构问题。领域本体在构建的过程中大多是面向特定的领域，因此很难做到一致性。因此，对领域本体构建方法的研究具有重要的意义。本文对领域本体的构建方法做了系统的概述，并在此基础上又对基于形式概念分析的领域本体构建方法极分别进行了描述，并对几种方法进行了对比。

关键词：形式概念分析；本体；概念格；领域本体构建
中图法分类号：TP305　　文献标识码：A　　文章编号：1000-7024(2016) 09-1-04

Domain ontology construction method based on formal concept analysis

YANG Zi
(College of Information Science and Technology，Dalian Maritime University，Dalian 116026，China)

Abstract：As the basis of semantic web, ontology has attracted more and more attention, and has been used in more and more fields. The emergence of ontology is a great contribution to the field of computer, which solves some problems existing in the computer field, such as the semantic heterogeneity problem. Domain ontology in the process of construction is mostly oriented to specific areas, so it is difficult to achieve consistency. Therefore, it is of great significance to study the construction method of domain ontology. In this paper, the domain ontology construction methods are summarized, and on this basis, based on formal concept analysis of domain ontology construction methods are described, and several methods are compared.

Key words：formal concept analysis；ontology；concept lattice；domain ontology construction

0 引　言

本体[1]的概念最初起源于哲学领域，是共享概念模型的明确形式化规范说明。它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词汇，并从不同层次的形式化模式上给出这些词汇（术语）和词汇之间相互关系的明确定义。本体的概念有四层含义：概念化(conceptualization)、形式化(formal)、明确(explicit)、共享(share)。它是一种能在语义和知识层次上描述信息系统的概念模型。本体应用的基础是本体的构建。

形式概念分析是从形式背景进行数据分析和规则提取的强有力的工具[2]。形式概念分析以数学为基础而建立，对本体中的概念、属性以及相互关系等用形式化的语言表述，然后根据语境构造出概念格，清晰的表达出本体的结构。

1 形式概念分析与本体简介

1.1 形式概念分析

形式概念分析是 20 世纪 90 年代 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 lattice)，从而清楚地表达出概念及概念间关系的结构。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概念格。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，

收稿日期：2016-11-15；修订日期：2016-12-5。
基金项目：国家自然科学基金项目（60972014）。
作者简介：杨姿（1993— ），女，河北石家庄人，硕士研究生，研究方向为智能信息处理。
E-mail：dmuyangz@gmail.com

在机器学习、数据挖掘、本体研究、软件工程、知识发现以及 Web 语义检索等领域得到了广泛的应用[3,4]。

现实世界是由各种各样的对象组成的，每个对象都有自己的一组属性或者特征。概念就是指对象、属性以及它们之间的关系，概念反应了对象的特有属性，分为两部分：一部分是对象，一部分是属性集。因此，概念也可以表示为（对象，属性集）的二元组形式。背景是概念的集合，也就是对象集合及其具有的属性的集合。任何一个概念都是从背景中提取出来的一个子集，通常以对象-属性集的二维表表示一个背景，用 1 表示某个对象具有某个属性，而用 0 表示某个对象不具有某个属性。形式概念分析是做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 1.1 形式概念：设形式对象集 G，形式属性集 M，二元关系 I⊆G×M。若 X⊆G 并且 Y⊆M，X=｛x | x∈G，∀ y∈Y，xIy｝，Y=｛y | y∈G，∀ x∈x，xIy｝，则二元组(X，Y)称为形式概念其中 X 称为形式概念的外延，表示属于这个形式概念的对象的集合；Y 称为形式概念的内涵，属于这个形式概念的属性的集合。

定义 1.2 形式背景：三元组 K=(*G*，*M*，*I*)被称为形式背景，其中 G 为形式对象的集合，M 为形式属性的集合，I 是 G 和 M 之间的二元关系，I⊆G×M。若 g 是 G 中的一个形式对象，m 是 M 中一个形式属性，那么用（g，m）∈I 表达 g 与 m 之间的关系，读作“形式对象 g 具有形式属性 m”[1]。

定义 1.3 概念格：对于形式背景 H=(G，M，I)存在唯一的一个偏序集<H，≤>与之对应，并且该偏序集的子集的上确界与下确界都存在，这个偏序集产生的格结构称为概念格。

1.2 本体

本体的英文解释是 ontology[5]，它的概念最初起源于哲学领域，它在哲学中定义为“对世界上客观存在物的系统的解释或说明”它的英文解释是它的概念最早初起源于哲学领域，它在哲学中的定义为“对世界上客观存在物的系统地描述，即存在论”，是对客观存在的一个系统的解释或说明，关心的是客观现实的抽象本质。

在人工智能界，最早给出本体定义的是 Neches 等人，他们将 Ontology 定义为“本体定义了组成主题领域的词汇表的基本术语和关系，以及组合这些术语和关系来定义词汇表外延的规则”。后来在信息系统、知识系统等领域，越来越多的人研究本体，并给出了许多不同的定义。

目前被大部分人所工人的定义是“本体是关于共享概念的一致约定。共享概念包括用来对领域知识进行建模的概念框架、需要互操作的主体之间用于交互的与内容相关的协议，和用于表示特定领域的理论的共同约定。在知识共享的情况下，本体的形式特化为具有代表性的词汇的定义。一种最简单的形式是一种层次结构，用来详细描述类和它们之间的包含关系。关系数据库的框架（schemata）也是一种本体，它用来描述能共享的数据库之间的关系和集成这些数据库需遵循的约束”。

本体(Ontology)的逻辑结构可以看成是一个五元组，O:={C, R, H, rel, A}，本体逻辑结构为：

⑴C 和 R 为两个交集为空的集合。C 与 R 包含的元素分别被称为概念标示符和关系标示符。

⑵概念层次 H 是一个有向的传递关系 H 是 C×C 子集。

⑶rel 为一个函数，其定义域为 R，值域是 C×C 的一个子集。

⑷公理集 A 包含了本体所需的公理，是用逻辑语言表示的。

1.3 形式概念分析和本体的关系

形式概念分析是从形式背景出发，通过对概念、对象和属性之间的关系形式化出的一种理论，然后通过对象、属性极其二者之间的二元关系组成形式背景，通过形式背景构造出概念格，即本体。二者之间的这种共性从本质上揭示了二者能有紧密的联系的根本原因，即具有相同的代数结构。这种相同的代数结构使得二者很容易产生一种映射关系。综合所述：形式概念分析与本体相联系的原因包括：都是对概念之间关系的描述；都具有相同的格结构；都是形式化的工具和方法；在信息科学、概念知识处理及知识表示等各个层次上的应用都有共同的应用领域等。二者的映射关系如图 1 所示。

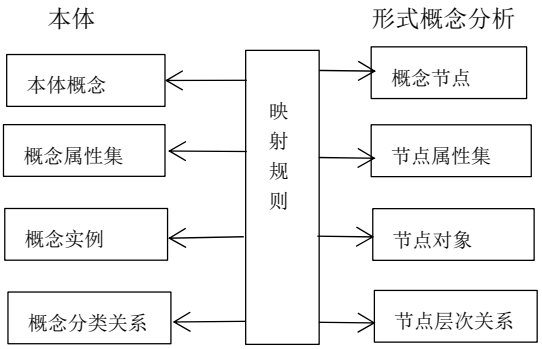

图 1 本体和形式概念分析的映射关系

在实际应用中，形式概念分析(FCA)和本体作为两种形式化方法，具有一定的相似性。它们都强调概念主体之间一致的重要性，也都强调模型形式说明的必要性，因此形式概念分析(FCA)可作为一种学习技术用于支持本体的构建。

2 本体构建方法

目前，领域本体的构建大多采用手工方式.，各领域在建立各自的本体时，都有自己的原则、标准和定义，缺乏公认的建模方法，影响了本体的重用、共享和互操作。目前用来构建领域本体的方法主要有两种：

方法 1 在领域专家的帮助下用本体描述语言将本体描述出来。

方法 2 从结构化的数据或文本中抽取或学习或发现领域本体。

用第一种方法构建领域本体，是完全手工构建的.对于一些复杂的应用领域而言，费时费力，而且具有很大的主观性。由不同的人来构建本体，即使是领域专家，构建出来的本体也有很大的差别，如此一来，构建的本体就违背了引进本体的初衷。为了解决完全手工构建本体带来的一些问题，出现了第二种本体构建方法，即采用自动化的或是半自动的方法来构建本体。这样，可以简化手工构建本体的工作量，提高本体的质量。

3 形式概念分析用于领域本体构建

3.1 形式概念分析用于领域本体的构建

（1）Philipp Cimiano 的方法[6]

Cimiano 等人提出了一种采用 FCA，通过分析词语在文本中的使用方式，获得相应的背景知识进而生成本体的领域本体的方法。该方法的具体方法如下：

①产生形式背景。分析领域文本，查找动词和它们的直接宾语得到领域内的概念，生成语法树，通过自然语言解析器生成文本中动词 / 宾语之间的依赖关系，假设解析出来的关系穷尽了文本中的关系（信息完整性假设）。

②将抽取出来的动词和宾语进行规范化。通过词典将它们接异体形式归类，如动词转换为其动词原型，复数形式转换成单数形式，并将动词加上后缀“-able”，使它们看起来更像是属性，以便自动产生的概念格和概念层次中的概念更易理解。

③分析上下文。对宾语分组，利用 FCA 将其结构化为抽象的概念，由形式背景生成概念格。

④通过直接删除概念格最底层元素，可以将其转换成偏序关系，将生成的形式概念作为本体的概念（以其内涵命名）。

（2）GuTao 的方法[7]

GuTao[54]提出的形式概念分析用于本体构建的方法如下：

① 通过 NLP 的方法或手工地从领域文本获得领域概念和属性。

② 用 Protege2000 进行建模，用 classes（领域概念）、slots（概念的属性）、facets（对属性的约束）来表示领域本体。

③ FcaTab 插件

FcaTab 是由 GuTao 开发的 Protege2000 的插件。其功能是通过表 3.1 所示本体与 FCA 的对应关系自动得到形式背景，并能将形式背景转化成概念格工具 ConExp[43]要求的形式背景输入格式。

表 1 FCA 与本体的对应关系

Ontology	Context
Class	Object
Slot	Attribute
Facets	多值属性值

④ ConExp 建立概念格

通过 ConExp 从 FcaTab 输出的形式背景建立与形式背景同构的概念格。领域专家或本体开发者在得到的概念格中可以选择需要的而原先没有的一些概念和关系，将它们添加到本体中去。这样，原来的形式背景就改变了。可以重复步骤③、④，直到满意为止。

整个过程如图 2 所示。

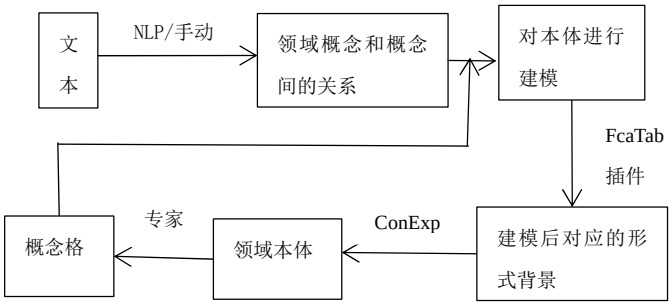

图 2 GuTao 方法流程

（3）MareK Obitko 的方法[8]

MareK Obitko 方法由 MareK Obitko 等人在 GACR 项目中提出，其基本原理为：概念由属性描述；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；本体的生成基于概念格对概念及属性的可视化，本体原型可由概念格映射得到。

其主要步骤为：

①以概念和属性的空集开始。

②按需要向概念表中添加概念和属性。

③用 FCA 对概念及其属性的表格进行可视化。

④基于可视化结果，对其进行规范化，方法包括直接编辑或根据本体构建工具的建议编辑。

⑤重复上述过程直到获得满意的本体。

3.2 方法的比较和分析

根据以上对三种基于 FCA 的本体构建方法详细描述，对三种方法进行综合的研究分析[9]得表 2。

表 2 三种方法的比较

	Cimiano	GuTao	*MareK Obitko*
概念的获取方式	NLP	手动/NLP	手动/NLP
对象	名词	类	实体
属性	动词	槽	实体属性
本体表示	格	Protégé 模型	三元组
构建模式	线性开发，封闭式结构	封闭式结构	线性开发，封闭式结构

在实际应用情况下，三种方法各有自己的优缺点，各有

自己的适用情况。Cimiano 方法实现了本体的自动构建，易于实现本体的更新，提供了一套本体的评价方法。GuTao 方法可以校区分类结构中概念的冗余，并且得到需要的概念；自开发的 fcatab 插件可自动从领域概念和关系得到形式背景；结合领域专家的参与实现半自动化的领域本体构建，提出了领域本体构建过程中应当依靠循环反馈不断完善的开发思想，但是没有充分考虑形式概念分析中属性多值的情况，不舍和处理多值情景。MareK Obitko 方法提供了分布式的本体编辑环境，实现了领域本体的分布式开发，按属性进行分类，倡导将领域本体概念分类关系作为领域本体概念层次关系的重点，克服了当前分类方法存在的问题；提出了一整套对形式背景和概念格的编辑修改机制，值得借鉴；实现了可视化基础上的概念格编辑.但是形式背景的生成需要手动完成，其整个过程是通过添加或删除概念和属性调整这样一个不断迭代的过程，所以到底需要添加和删除哪些内容难以把握，其具体到什么时候结束也不易确定；该方法每次都从空的对象和属性开始，因此对对象和属性的添加是一项及其复杂的过程，而且工作量大，因此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之前两个方法较低.。

4 结束语

将 FCA 应用于领域本体的构建才处于一个刚刚起步的阶段，在实际应用过程中还存在许多问题，但是 FCA 为构建领域本体这一难题提供了新的解决思路。本文重点介绍了几种领域本体的构建方法，认为在本体构建原则的指导下，选择适合特定领域本体构建方法，使用支持该开发构建工具，领域专家可以方便、可视化的开发特定领域的本体。随着 FCA 中的概念更合理的同本体中的概念联系起来，且更好的同自然语言理解等领域结合起来，以及更完善的本体开发工具出现，将在一定程度上解决领域本体的构建过程中遇到的一些问题，我们相信，在未来的领域本体构建中遇到的困难将会大大减少。

参考文献：

[1] Formica A . Ontology-based concept similarity in formal concept analysis , Information Sciences , 176(2006) , 2624~2641

[2] [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.

[3] Baidu . Formal Concept Analysis . http :// baike . Baidu . com / v-Iew / 4660144 . htm[OL],2016,11,3.

[4] 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,2010,15(11):17-23.

[5] 钱杰.基于形式概念分析的本体构建与映射方法研究，国防科技大学硕士学位论文,2006

[6] Cimiano P . Staab S , Tane J . Deriving concept hierarches from text by smooth formal concept analysis . Proc . Of the GI Workshop “ Lehern Lernem-Wissen-Adapttivitat”.2003.

[7] Gu Tao.Using Formal Concept Analysis for Ontology Structuring and Building .ICIS , Nanyang Technological University , 2003

[8] Marek O , et al . Ontology Design with Formal Concept Analysis In: Snasel V , Belohlavek R , eds . Concept Lattices and their Applocationis , Proceeding of the 2nd International CLA Workshop TU of Ostrava,2004