基于形式概念分析的本体研究

商迪

作业	分数
得分	

2020 年 12 月 10 日

基于形式概念分析的本体研究

商迪

（大连海事大学 计算机技术 辽宁省大连市 中国 116026）

摘　要：随着信息共享和数据交换的不断扩大,本体作为共享化概念的形式说明，其应用不断被采纳。本体作为一种有效的、表现概念层次结构和语义的模型,被越来越多的领域所应用，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。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，形式概念分析的方法弥补了已有本体构建的缺点。本文对现有的本体构建方法做了总体性介绍。并在此基础上详细描述了几种基于形式概念分析的领域本体构建方法. 最后对形式概念分析用于领城本体构建方法做了总结。

关键词：概念格；本体；形式概念分析；本体构建；

中图法分类号：TP311.20　DOI 号：　10.3969/j.issn.1001-3695.2020.01.030

Ontology research based on formal concept analysis

Di Shang

Department of computer, Dalian Maritime University, City Dalian

Abstract： With the expansion of information sharing and data exchange, ontology has been adopted as a form of sharing concept. Ontology, as an effective model to express conceptual hierarchy and semantics, has been applied in more and more fields. The appearance of ontology can solve some difficulties in computer application field. Aiming at the problems of traditional ontology construction methods, such as relying on manual labor, heavy subjective interference and insufficient extraction of implicit concepts and relations, a method of ontology construction based on formal concept analysis is proposed, which makes up for the shortcomings of existing ontology construction. This paper gives a general introduction to the existing ontology construction methods. Several domain ontology construction methods based on formal concept analysis are described in detail. At last, the paper summarizes the method of formal concept analysis used to construct the city ontology.

Key words：Concept lattice; Ontology; Formal concept analysis; Ontology construct；

1 概述

本体作为能在语义和知识层次上描述信息系统的概念模型,自从被提出来后就得到了广泛的关注。本体是现实世界的模型,所建立的本体必须能够客观反映现实世界。如何快速构建本体，特别是构建一个反映特定领域的领域本体，是人们研究的一个热点。

研究人员提供了许多经典本体构建方法，如 Tove 法、Idef-5 方法、Kactus 工程法、Methontology、Sensus 法、骨架法、七步法等。这些方法都有自己的特点和适用领域，再加上本体构建本身也没有统一标准，因此难以在不同领域本体的构建中保持一致。客观上,本体构建是一件复杂且费时的过程。而对领域专家来说，从给定的数据和文本中发现本体十分困难，需要一种能够半自动获取本体的方法，降低本体构建的复杂度和成本。

2 FCA 和本体中的概念

2.1 形式概念分析

形式概念分析，由 Wille R 于 1982 年首先提出，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。因此，概念格被认为是进行数据分析的有力工具。概念格是形式概念分析理论中的核心数据结构,它利用二元关系建立一种概念间的层次关系,概念格是应用数学的分支，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，形式概念分析越来越多地被应用到数据挖掘,信息检索,软件工程等方面，成为许多专家学者关注的热点

形式背景（formal context）可以表示为三元组 T=(O，D，R)，其中 O 是事例（对象）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与之对应，并且这个偏序集产生一种格结构，这种由背景（O，D，R）所诱导的格 L 称为概念格。格 L 中的每个节点是一个序偶(称为概念),记为(X,Y),其中 X∈P(O)称为概念的外延；其中 Y∈P(D)称为概念的内涵；每一个序偶关于关系 R 是完备的，即有性质：

1)X={x∈O|∀y∈Y,xRy}

2)Y={y∈D|∀x∈X,xRy}

在概念格节点间能够建立起一种偏序关系。具体地，给定概念 H1=(X1,Y1) 和 H2=(X2,Y2) ，则 H1<H2<==>Y1⊂Y2,领先次序意味着 H1 是 H2 的父节点或称直接泛化。根据偏序关系可生成格的 Hasse 图：如果 H1<H2 并且不存在另一个元素 H3 使得 H1<H3<H2，则从 H1 到 H2 就存在一条边。

概念格将每一个节点表示为一个形式概念，每个形式概念包含概念的外延(extent)和内涵(intent)两部分内容。外延表示此概念所包含的所有对象的集合、即此概念所涵盖的实例，内涵则表示概念中所有对象的共有特征。对于给定的形式背景 K=(G.M.1)(其中 C 为对象集合，M 为属性集合，I 是 G 与 M 之间的一个二元关系)，存在惟一一个偏序集合与之相对应。由偏序集构成一种格结构，并且此偏序集满足自反性、反对称性和传递性。若 g∈C.m∈M.glm 表示对象 g 具有 m 属性。格中的每个节点称之为概念,记作 C(X,Y),X∈G 是概念 C(X,Y)的外延。

Y∈M 是概念中对象的共有属性(内涵)。节点概念与节点概念之间存在着偏序关系，若 C1=(X1,Y1),C2=(X2,Y2),并且 X1<X2 <=>Y1<Y2，称 CI 为 C2 的父节点。

2.2 本体

本体最早是一个哲学上的概念。从哲学的范畴来说,本体是客观存在的一个系统的解释或说明.关心的是客观现实的抽象本质。随着人工智能的发展,本体被人工智能界给予了新的定义。然而,最初人们对本体的理解并不完善,对本体的定义也在不断地发展变化中。其中最为大家普遍接受的本体定义: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。

2.3 形式概念分析与本体

概念格由概念的层次关系组成，内涵和外延构成了概念;而我们的本体又是用来体现概念和概念之间的关系的。FCA 与本体之间既有联系又有区别。

本体的目的是对人能感觉到的现实世界建立共享的概念模型,从而支持有丰富知识的应用领域。FCA 不是为现实建模,而是为人工世界建模,目的是支持用户在给定数据的基础上进行领域分析和建模。

我们可以在没有任何数据的前提下为领城构建本体,但是 FCA 必须建立在给定数据集(形式背景)的基础上。

观察到本体和形式概念分析(FormalConceptAnalysis,FCA)都是对概念的形式化表达，并且其表现形式都是概念和关系组成的层级结构，所以基于 FCA 构建本体具有可行性，并且具有以下特点:

(1)概念格算法的研究已经较为成熟，在基于 FCA 构建本体的过程中,原本依赖人工的初始本体构造可以转化为概念格构造,实现了本体构建的半自动化;

(2)概念格中的概念是算法自动从形式背景中获取,并按照序关系形成格结构,避免了传统本体构建中人工主观因素的干扰;

(3)FCA 同时关注对象和属性，而本体只注重属性,将 FCA 引人本体构建,丰富了本体概念关系提取方法，发现更多隐含概念关系;

(4)本体在视觉上像“树”，而概念格则像“网”，树中的节点非此即彼，网中的节点四通八达,通过使用概念格表示本体,可以使本体更像一张“网”，增加节点知识的互联性。

虽然 FCA 与本体不同,但是它们同样源于哲学，都是对概念以及概念之间关系的描述，所以可以将两者结合。在本体的提炼、合并和映射等本体的一些相关工作中都有研究作者在做 FCA 与本体结合的工作.如 FCA-Merge 是到目前为止比较有名的 FCA 在本体合并中的应用。

3 形式概念分析用于本体构建的方法

3.1 Philipp Cimiano 的方法

这是 AIFB 研究机构在 IST-Dot Kom 项目中应用的方法，总的思想是使用一个自然语言的解析器.通过该解析器从领域文本中的每一个句子可以得到一颗语法树,由语法树可以直接得到动词/对象间的依赖关系;进-步通过词典查询，对提取的动词和对象用词的原形来规范化表示。如 bought/ buys 转换成原形 buy,并给动词加上后缀_ able,使得它们看起来更像是属性;最后,将 FCA 中的概念和本体中的概念直接等同,得到概念格，由概念格得领域本体。

3.2 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 GuTao 开发的 Protege2000 的插件。其功能是通过本体与 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 要求的形式背景输人格式。

4)ConExp 建立概念格

通过 ConExp 从 FeaTab 输出的形式背景建立与形式背景同构的概念格。领城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 3)和 4)直到满意为止。

3.3 Marek Obitko 的方法

Marek Obitko 等人在 GACR 项目中提出如下方法:

●概念由属性来描述;

●属性决定概念的层次;

●当不同概念的属性相同时.认为这些概念是同一个概念;

●直接由修改过的概念格作为个体表示。

具体步骤是:

1)从空的对象和属性集合开始。

2)由使用者根据需要把对象和属性添加到形式背景中。

3)显示形式背景对应的概念格。

4)用户可以在显式化的概念格的基础。上做如下操作：

a)根据本体使用的需要直接编辑：

i)添加或移除对象;

ii)添加或移除属性;

iii)给对象添加属性或从对象移走某一属性。

b)由程序提示编辑本体

i)当两个对象有相同的属性时.要么合并成

一个对象，要么给对象添加属性以区别对象。

ii)FCA 能产生新的对象.这些对象直接由属性构成。

5)整个过程可以不断地循环重复.直到设计者满意为止。

3.4 Hele-Mai Hav 的方法

HeleMai Hav 提出了基于概念格的本体表示方法.主要适用于领域文本内容比较短的情况.而且假设领城文本描述了某一实体.里面包含了描述领域的术语。做法如下:

1)从领域文本或数据中抽取形式背景。

●形式背景的对象(objecls):用自然语言表示、描述领域实体的文本(对文本编号 A1...),假设文本中使用了领域词汇。而且文本都很短。如广告、产品描述、组件的技术描述都可以:

●对象的属性:在描述领域实体的领域文本中出现的名词短语:

●对象和属性的关系:在对文本做 NIL.P 的过程中获得。名词短语集合和文本的编号存储到数据库表中，这样得到的数据库表表示了领域向用的形式背景。其中的元关系是文本和名词短语之间的关系,

2)通过 FCA 和概念格缩减。从形式背景计算得初始本体,在该方法中.FCA 是作用在存储了形式背景的数据库表当中的，通过 FCA 得到概念格，为了获得基于概念恪的本体表示。对获得的慨念格进行缩减和对部分节点命名。

3)将初始本体移植成用一价谓词逻辑表示的集合，这一步提供初始本体到霍恩逻辑的方法。该过程产生初始本体的逻辑表述.并用一阶谓词逻辑表示语义描述。

为了这个目的.上述文献介绍了初始本体的基于霍恩逻辑的公式。根据公式.初始本体自动转化成事实集合(a sct offacts),初始集合中包含了偏序关系的规则和为了进行本体推理而提供的格公理及格操作。由逻辑描述。可以使用一种自动定理证明方法进行自动推理。该方法使用了一阶逻辑语言.所以在实际应用中可以结合不同的本体推理引擎(通过将本体描述翻泽成任何推理引擎规则语言)。这也是选择基于霍恩逻辑的规则语言的原因。

4)通过增加规则和事实扩充初始本体。

正如我们已经看到的概念之间的分类关系能通过在本体描述上使用基于逻辑的概念格公式自动产生。为了定义非分类关系需要定义相应的谓词和规则，如概念属性、属性的继承等的表示。

5)本体推理

推理是保证本体设计质最的很重要的一项内容。推理能发现相互矛盾的概念.能得到隐含的关系等等能用格公理和格操作的推理规则来确定概念之间的分类关系由于添加了额外的规则.推理非分类关系成为可能。

4 结语

基于 FCA 构建本体的方法研究呈现出应用环境和技术手段的多样化，注重与当今信息网络环境发展趋势结合，并针对具体应用具体构建。有借助云环境下的技术理念探索基于 FCA 的领域本体协作构建模式，提高本体构建的效率和质量;也有给出基于 FCA 和 Folksonomy 的本体构建方法，为网络社区环境下通过社群分类法实现及时、灵活和人本的本体构建过程提供新的思路;还有的是结合情报学领域本体构建实例说明 FCA 在本体构建中的应用；这些研究丰富了本体与其他相关技术的结合,拓宽了本体的应用领域。

本体的构建过程离不开领域专家的参与，FCA 能帮助结构化和构建本体.能将本体在格中表示出来.用格来表示概念相比树更易于理解且可作为构建本体的指南。但是有一点要注意,它只是提供了指南.最终选择的本体仍与开发者有很大关系。虽然 FCA 有着很强的数学基础,从格中能很方便地给出一些隐含的概念供选择,但是,当本体的应用领域非常复杂时,相应地建立的概念格必将很复杂。这样复杂的格结构将淹没有用信息,从而又为新概念和关系的选择带来难度。随着 FCA 中的概念更合理地同本体中的概念联系起来，且更好地同自然语言理解、机器学习等领域的方法相结合,更完善的本体开发工具的出现,我们有理由相信领域本体的

构建将不再困难,构建的领域本体定能更好地表达领域并为之服务。

5 参考文献

[1] 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.

[2] 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.

[3] 程利涛, 邢欣, 李伟. 基于形式概念的本体学习方法[J]. 电脑开发与应用, 2014, v.27;No.242(010):55-58.

[4] 陆佳莹, 袁勤俭, 黄奇,等. 基于概念格理论的产品领域本体构建研究[J]. 现代图书情报技术, 2016(5):38-46.

[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] 李志珂, 李俊. 一种基于形式概念分析的术语定义方法[J]. 数码设计.CG WORLD, 2019, 008(009):P.27-30.