《智能信息处理》课程作业

基于形式概念分析的本体研究

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[80]
得分	

2020 年 11 月 24 日

基于形式概念分析的本体研究

栾利广

(大连海事大学 信息科学技术学院 辽宁 大连 1120200326)

摘 要 近年来，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。应该说，本体的出现能很好地解决目前计算机应用领域中存在的一些困难，如人机交互或机器与机器之间的通信、自动推理、知识表示和重用等。但是，在能很好地应用本体之前，我们面临一个新的难题：本体的构建。本文对现有的领域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域本体构建方法做了分析、比较和总结。

关键词 形式概念分析，本体，本体构建

Ontology Research Based on Formal Concept Analysis

Luan Liguang

Abstract In recent years, ontology, as an effective model representing conceptual hierarchy and semantics, has been applied in more and more fields. It should be said that the emergence of ontology can well solve some difficulties existing in the current computer application field, such as human-computer interaction or machine-to-machine communication, automatic reasoning, knowledge representation and reuse. However, before we can apply ontology well, we are faced with a new problem: ontology construction. This paper gives a general introduction to the existing methods of domain ontology construction, and on this basis, describes several methods of domain ontology construction based on formal concept analysis in detail. Finally, it analyzes, compares and summarizes the methods of domain ontology construction based on formal concept analysis.

Keywords Formal concept Analysis, Ontology, Ontology construction

1 引言

本体原本是一个哲学概念。在哲学意义上，本体论是对客观存在的一种系统的解释或解释，涉及客观现实的抽象本质。随着人工智能的发展，人工智能给出了本体的新定义。然而在开始的时候，人们对本体论的理解并不完善，这个词的定义是不断演变的。最广泛接受的本体定义之一:本体是共享概念模型的明确的形式化规范说明。

本体论反映的是被所有人接受的共同知识，反映的是相关领域的共同知识。本体的目标是捕获相关领域的知识，提供对领域知识的共同理解，并确定领域内共同识别的概念是由不同层次的形式化模型给出的，这些概念(术语)以及它们之间关系的明确定义。

现有的本体设计原则非常笼统，抽象的本体没有明确的语义和可操作的语义，

甚至有些原始的本体设计原则之间也存在不一致的地方。目前,仍然没有接受本体设计、评价标准和质量保证标准的原因，也是做深入的研究的原因。知识体系建设是知识的来源通常来自用户的需求。本文将系统介绍一般的本体构建方法和形式概念分析用于领域本体的构建，也将这作为一个重要的切入点，知识系统需求通过对知识系统的需求分解，将知识分为动态知识和静态知识，从而更全面、有效地描述知识系统的本体知识。

2 FCA 和本体中的概念

2.1 形式概念分析

形式概念分析理论是德国数学家 Wille 教授在 1982 年提出的，用于概念的发现、排序和显示，并且在 1999 年 Ganter 对形式概念分析理论的早期成果作了总结。FCA 不会像其他数据分析方法那样粗粒度减少给定的信息，并且能够包含所有数据细节。其在本体构建过程中的概念提取和关系提取（分类关系和非分类关系）部分的应用被许多学者研究。尤其是对非分类关系的提取，FCA 的表现尤为突出。

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。

在形式概念分析中，概念的外延被理解为属于这个概念的对象的集合，而内涵则被认为是所有这些对象所共有的特征或属性集，这实现了对概念的哲学理解的形式化。所有的概念连同它们之间的泛化/例化关系构成一个概念格。

关于形式概念和概念的主要定义如下：

定义 1 一个形式背景 $K=(G,M,I)$ 是由 2 个集合 G 和 M 以及 G 和 M 之间的关系 I 组成。G 的元素成为对象，M 的元素称为属性。$(g,m)\in I$ 或 $g\,\mathrm{Im}$ 表示对象 g 具有属性 m 。

定义 2 设 A 是对象集合 G 的一个子集，定义 A 中对象共有属性的集合：

$$f(A)=\{m\in M\mid \forall g\in A, g\,\mathrm{Im}\}$$

相应地设 B 是属性集合的一个子集，定义具有中所有属性的对象的集合：

$$f(B)=\{\mathrm{g}\in M\mid \forall m\in B, g\,\mathrm{Im}\}$$

定义 3 若（A_1,B_1），（A_2,B_2）是某个形式背景的 2 个概念，而且 $A_1\subseteq A_2$，则称（A_1,B_1）是（A_2,B_2）的子概念，（A_2,B_2）是（A_1,B_1）的父概念，并记作 $(A_1,B_1)\le(A_2,B_2)$，关系≤称为是概念的“序层次”（简称“序”）。形式背景中的所有概念用这种序组成的集合称为概念格，记作：$L(G,M,I)$ 。

2.2 本体

Gruber 于 1993 年给出了 Ontology 的定义，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。

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

Ontology 的 5 个基本建模元语。这些元语分别为：类，关系，函数，公理和实例，通常也把 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

$O:=(C,\leq C,\mathrm{R},\delta,\leq \mathrm{R})$,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 $\leq C$ 叫做概念层级；函数

$\delta:\mathrm{R}\rightarrow C^{+}$,定义域是 R，值域是 $C\times C$； R 上的偏序集 $\leq R$ 是关系层级。

3 本体构建方法

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。根据本体构建数据源的结构化程度.将这些构建方法分为 3 类，即基于结构化资源、基于非结构化资源和异构资源的合并本体构建方法。针对这 3 种类别，分析和阐述代表性的本体构建方法的优缺点，在比较结果中发现基于形式概念分析构建本体具有较大的改进空间，结合具体应用领城构建时需要在对象和属性的取舍、针对不同语言特点构建形式背景等问题上作进一步研究。

目前，本体构建大多采用手工方法，远远没有成为一种工程性活动。大家在构建本体的过程中，每个人都有自己的原则、标准和本体定义，缺乏公认的建模方法，影响了本体的重用、共享和互操作性。本体构建主要有两种方法:1)在领域本体描述语言中，借助专家的帮助描述本体。2)从结构中从数据或文本中提取、学习或发现领域本体。

第一种方法中的本体构建完全是手工构建的本体。对于一些复杂的应用领域来说，这将是一项费时费力的任务，同时也有很多主观性。由不同的人，甚至是领域专家来构建本体，所构建的本体将是非常不同的。这样，本体论的构建就违背了引入本体论的初衷。

为了解决完全手工构建本体所带来的一些问题，提出了第两种本体构建方法:自动构建本体和半自动构建本体。这样可以简化手工本体构建的工作量，提高本体构建的质量。

Alexander Maedche 和 Steffen Staab 根据本体学习的知识源不同，对采用自学习的方法半自动地构建本体的方法做了如下分类：

- 通过字典进行本体学习。该本体是在现有机器可读词典的基础上构建的，用于抽取相关概念及其关系；
- 从知识库中学习；
- 从现有的知识基础结构中学习，从关系数据库中提取本体；
- 从半结构化数据中学习。比如 XMI。样本半结构化数据源的模式提取概念和概念之间的关系来构建牙科学。
- 从文本中学习

构建方法有：

- 基于模板的提取方法；
- 关联规则；
- 概念聚类；
- 形式概念分析。

4 基于形式概念的本体构建

4.1 技术路线

1)使用自然语言对处理后的纯文本文档进行预处理，得到字词组合。

2)使用统计方法获取能够代表文本的关键字。本步骤一般采用 TF—IDF 方法。首先，计算概念词在文本集中出现的频率。如果它大于阈值，则确定该概念为领域本体中的概念。最后，找出所有的概念和文本集，形成一个二元关系表。

3)将前一步形成的二元表的多个值转化为单值，形成单值二元关系表，用于构造概念格。

4)将概念格转换为相应的本体。使用属性而不是形式概念，并确保它们只在注释中出现一次。因为属性都是文字，主体的焦点也是文字，所以这里我们使用形式概念的属性来代替主体中的概念，完成从形式概念到本体的转换。

5)自动更新生成的概念格。

4.2 本体构建的方法

GuTao 的方法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 L8_，用 classes(领域概念) 、slots(概念约束)、facets(对属性的约束)来表示领域本体。

3)FcaTab 插件

FcaTab 是由 GuTao 开发 Protege2000 的插件。其功能是通过表 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 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 的形式背景建立与形式背景同结构的概念格。领域专家或本体开发者在得到的概念格中可选择需要的而原先没有的这种概念和关系，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)

4)，直到满意为止。

整个过程如图 2 所示。

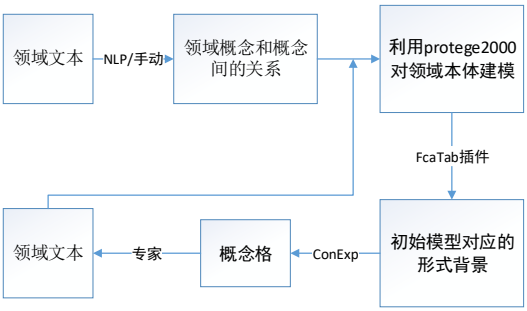

图 2 GuTao 的基于 FCA 的本体构建过程

5 小结

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，相立地建立的概念格必将很复杂。这样复杂的格结构将淹没某些信息，从而又为新概念和关系的选择带来难度。

目前，FCA 用于本体的构建处于刚刚起步的阶段，在实际应用过程中还存在许多的问题，但是 FCA 为构建领域本体这一难题提供了新的解决思路。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。

参考文献

[1] Haav H.M A Semi-automatic Method to Ontology Design by Using FCA. In：Snasel V. Balohlavek R，eds. Concept Lattices and their Applications. Proceedings of the 2nd International CLA Workshop．TU of Ostrava．2004,13～25.

[2] 颜时彦，王胜清，罗云川，等，云环境下基于 FCA 的领域本体写作构建模式初探[J].现代图书情报技术，2014,30（3）：49-56.

[3] 王立政. 基于本体的知识检索模型优化研究[D].吉林大学,2011.

[4] 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.

[5] 陈壮生,瞿裕忠. 基于本体的信息处理系统的设计与实现[J]. 计算机工程,2005,(11).

[6] 常春. Ontology 在信息管理领域的研究背景[J]. 现代图书情报技术,2003,(6).

[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] 黄美丽，刘宗田. 基于形式概念分析的领域本体构建方法研究[J].《计算机科学》2006 年 第 1 期 210-212.