基于形式概念分析构建本体的方法

蔡永嘉

（大连海事大学信息科学技术学院 大连 116026 ）

摘 要 随着语义网的发展，语义数据集成成为研究中的热点。本体作为共享概念的形式化说明，用于表达数据源的语义识别和建立概念间的语义关联达成一致，提供了语义异构问题的解决途径。有效地构建本体成为本体应用的关键问题，本文采用形式概念分析构造本体，由于关系数据只表达了“属性/值”二元关系，不具备语义描述的能力。如何在关系数据中直接抽取数据的语义，构建本体是一个重要的研究方向。文中探讨了运用形式概念分析的技术，结合数据库模式构建本体的方法。

关键词 本体；形式概念分析；概念格

中图法分类号 TP393 **文献标识码** A

Ontology Construction Method Based on Formal Concept Analysis

Cai YongJia

(Department of Information Science and Technology, Dalian Maritime University, Dalian, 116026*, China)*

Abstract As the developing of Semantic Web of Things, Semantic data integration is becoming the hot spot in the research of database technology. Ontology as formal specification of share concepts, provides a method to resolve the semantic heterogeneous, since it can be used to express the semantics of data sources, identify and construct the semantic relationship of the concepts. and ontology construction become the key problems. This paper discusses the technique of using formal concept analysis for ontology construction. However, relational database only express the “attribute/value” relation, it does not have the semantic describing ability. Therefore, it is an important research orientation that how to take out the semantics in the relational data directly. This paper discusses the technique of using formal concept analysis in combination with database mode the data information for ontology construction.

Key words Ontology ,Formal Concept Analysis, Concept lattice

1 引言

互操作问题是分布、异构数据源的访问、集成中的经典问题。结构异构和语义异构是这一问题的两大方面。本体(ontology)作为共享概念的形式化说明,就象一部规范的字典,给通信各方提供公共的概念解释, 用于表达息源的语义、识别和建立概念间的语义关联,达成语义一致,便于解决信息集成中语义异构的难题。本体的开发已受到人们的广泛重视,目前已经有了许多决策、方法上的指导。然而,在本体开发的研究和实施中,有很多具体问题确实很难解决[1],比如:不同的人构造本体的目的不同 ,侧重点有所不同,理解也有所不同,而且本体的表示方法具有多样性,不容易达成共识和重用,所以至今依然没有统一的驱动本体开发的方法

收稿日期：2016-10-15；最终修改稿收到日期：2016-10-19 . 本课题得到国家自然科学基金(70471003)资助.蔡永嘉，男，1995年生，硕士，主要研究方向为智能信息处理，E-mail: 986818694@qq.com.

学和相应技术。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[2]。

2 本体构建

在决定构建本体之前,必须决定构建什么类型的本体,怎么构建本体。从描述范围来看,本体包括领域本体和公共本体[3]。领域本体和特定的应用相关,描述了现实世界内小范围的一个模型;相反,公共本体包含公共的概念和关系,可用于不同的应用之中。公共本体作为本体构建的基石,便于扩展、添加新的概念和关系。更进一步,要清楚是采用人工构建还是自动构建的方式。目前已经有一些自动从文本信息构建本体的文献发表[4]。其基本思想是:从文本中分离出本体概念;使用自然语言处理的办法,分析句子中两个概念的关系。事实上,在分离概念的时候,需要人工专家的参与,因而也只是实现了半自动化。还要决定是用自底向上的方式构建本体,还是使用自顶向下的方式。自顶向下的方式,从“is-a”继承关系的顶端开始往下扩展。许多人工构造就是采用这种办法。而在自底向上的方式中,概念和关系是在发现概念、关系时逐步加入的。这种方式更适于自动构建。

从方法学的角度来看,本体的构建需要 3 个阶段[5]:

(1) 本体获取,即知识获取的过程。这需要识别目标和范围,发现领域内关键的概念及关系,给出定义描述并用准确的词汇表达出来。

(2) 本体译码,决定领域知识在概念模型中的结构。即把上一阶段获取的概念用形式化的语言明确表示出来。先要确立本体中用到的基本词汇,比如类(classes)、属性(properties)、字面值(facets)等,然后选择合适的知识模型来表达这些词汇。

(3) 本体集成,重用原有本体,加快本体开发。这一步比较困难,因为针对不同的本体甚至是结构完全不同、概念抽取方法迥异的本体,很难识别出公共的概念。如何提供指导或者工具辅助本体集成,是开发本体遇到的一个极大挑战。

3 形式概念分析

形式概念分析(Formal Concept Analysis)是应用数学的一个分支,它建立在概念和概念层次的数学化基础之上,根据用二元关系表达的形式背景,从中提取概念层次结构,即概念格[6]。概念格的每个节点就是一个概念,由两部分组成:外延(Extension),即概念所覆盖的实例;内涵(Intension),即概念的描述,该概念覆盖实例的共同特征。

定义 1:一个形式背景(formal context)是一个三元组 K=(G,M,I), 其中 G 是对象的集合, M 是属性的集合,I 是 G 和 M 之间的二元关系。对于坌 g∈G, m∈M 若(g, m)∈I 表示对象 g 与属性 m 的关系,读作“对象 g 具有属性 m”。记做 gIm。

定义 2:对于一个对象集 A,定义 A'={m∈M|gIm, 对所有 g∈A}(A 中所有对象共有的属性集合)。相应地,对于一个属性集 B,定义 B'={g∈G|gIm,对所有的 m∈B}(即包含所有 B 中属性的对象集合)。

定义 3:语境的形式概念(formal concept)是一个集合对(A,B)。其中:A⊆G,B⊆M 并且 A'=B,B'=A。A、B 分别称做概念的外延和内涵。β(G,M,I)表示语境 k=(G,M,I)中的所有概念集合。

定义 4:概念格(concept lattice)对于形式概念 (A1 ,B1),(A2 ,B2)均是语境中的概念,并且 A1⊆A2 ,那么(A1,B1) 被称做(A2 ,B2)的子概念,(A2 ,B2) 则是(A1 ,B1)的超概念,记为(A1 ,B1)≤(A2 ,B2),“≤”反映了概念间的层次关系。这种形式背景中所有形成概念的子概念-超概念的偏序关系所诱导出的格就是概念格(Concept lattice)。

4 用形式概念分析构造本体

考虑在给定的数据库模式及其数据信息的基础使用 FCA 的方法构建本体。传统的数据表是二维结构,这和表达形式概念的语境很相似,但数据表中的属性通常是数值的,不是简单的选择标记。数据表是一种多值的语境,为了运用上一段中介绍的方法,需要把多值语境(multi-valued context)转换成单值语境(one-valued context)[7]。在转换的时候 ,可以把属性中多值的部分,按照数值的不同,分为几类。然后,把这些类别值作为新的属

性 ,添加到原来的数据表中去, 并把原先数据表中的属性删除。这样,对应的新的对象/属性值就全是选择标记了。下面结合具体的例子,来说明这个构建过程(见图 2)。

对象/属性	属性 1	属性 2	身高	属性 4	属性 5
对象 1	√				
对象 2		√			
对象 3	√	√		√	√

图 2　多值语境数据表

这是一张具有 5 个属性的数据表,有 3 条对应的数据信息。其中, 属性 3 (身高) 具有 3 个不同的字面值(facets) ,其余的属性都是单值的。根据 3 种不同的形状,增加 3 种属性 ,并把原来的形状属性去掉。转换成下面的单值语境(见图 3)。

对象/属性	属性 1	属性 2	高	中	矮	属性 4	属性 5
对象 1	√		√				
对象 2		√		√			
对象 3	√	√			√	√	√

图 3　单值语境数据表

第一步：计算出单条记录对应的对象最大数目 N =2,因而 ,格的层次数为 2; A ={ 属性 1, 属性 2, 高,中,矮, 属性 4,属性 5}。

第二步：这步是反复迭代和枚举的过程。以 b ={ 属性 1}为例,这时 a=b'={对象 1,对象 3} ,a'={对象 1 和 3 共有的属性}={属性 1}=b。 因而,(a ,b)构成了一个形式概念, 加入到第一层次中去。根据这个算法,最后构造出来的概念格见图 4。

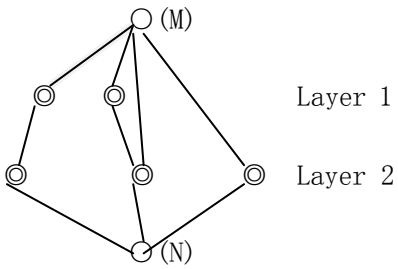

图 4　概念格

为了表达清晰起见,图 4 中没有给出层次上概念的对象和属性。其中第一层上的概念是({对象 1,对象 3} ,属性 1); ({对象 1,对象 3} ,属性 2)。第二层上的概念从左到右分别是({对象 1},{ 属性 1, 高})、({对象 2} ,{ 属性 2,中})、({对象 3} ,{ 属性 4,属性 5,矮})。这些概念之间的关系见上图 4。

5　结束语

在介绍了 FCA 的方法之后,在演示给定数据库模式和数据信息的基础上 ,运用 FCA 来构建本体————概念格。从概念格的线路图上,可以清楚地浏览概念之间的层次关系。现回到本体构建的起点。要构建本体,应先分离出概念,建立概念之间的关系(使用数据库模式时其实已经跳过了这一步),这里离不开领域专家的参与; 然后可以构建数据库模式。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; 最后算法,自底向上地自动构建概念格。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。

参 考 文 献

[1]. Cui Zhan, Jones D M , O' Brien P.Issues in Ontology–based Applications [J] . SIGMOD Record, 2002, 31(1) : 43-48.

[2]. Vol. N. Ontology Learning from Text: Methods, Evaluation and Applications[J]. Computational Linguistics, 2005, 32(4):pägs. 569-572.

[3]. Ruotsalo T. Methods and applications for ontology-based recommender systems[J]. Aalto-yliopiston teknillinen korkeakoulu, 2010.

[4]. Wei Xu, Wenjie Li, Mingli Wu,等. Deriving Event Relevance from the Ontology Constructed with Formal Concept Analysis[M]// Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2006:480-489.

[5]. 黄伟, 金远平. 形式概念分析在本体构建中的应用[J]. 计算机技术与发展, 2005, 15(2):28-31.

[6]. Johannes Wollbold1, Ren Huber, Raimund Kinne, and Karl ErichWolff Conceptual Representation of Gene Expression Processes[C]// Knowledge Processing and Data Analysis,2011:79-100

[7]. HUANG MeiLi, LIU ZongTian, 黄美丽,等. Research on Domain Ontology Building Methods Based on Formal Concept Analysis 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, 33(01):210-212.

Cai yongjia born in 1995
E-mail:986818694@qq.com, The main research direction is intelligent information processing.

Background

During the last three decades, formal concept analysis (FCA) became awell-known formalism in data analysis and knowledge discovery because of its usefulness in important domains of knowledge discovery in databases (KDD) such as ontology engineering, association rule mining, machine learning, as well as relation to other established theories for representing knowledge processing, like description logics, conceptual graphs, and rough sets. In early days, FCA was sometimes misconceived as a static crisp hardly scalable formalism for binary data tables. In this paper, we will try to show that FCA actually provides support for processing