基于形式概念分析的本体构建方法研究

安东

（大连海事大学，大连，辽宁 116026）

摘 要：随着今日的万维网快速发展，万维网中的信息变得越来越庞大。随着信息量的快速增大，人们在进行检索这些数据时，产生了很大的问题。万维网中存在这查全率和查准率的问题。形式概念分析（Formal Concept Analysis,FCA）是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的工具。形式概念分析建立在数学的基础之上，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，得到了现今的广泛应用。在万维网中，我们可能会用不同的术语来表达相同的含义，或者一个术语含有多个含义。因此，消除术语差异是很有必要的。目前的解决方法是，建立一个公共的本体。在设计该领域的时候，我们都使用这个公共本体来进行规范。将本体概念以及概念之间关系统一表现在概念格之中。基于形式概念分析的本体构建，避免概念冗余保证本体最大可扩展性的原则。

关键词：形式概念分析；概念格；本体构建；

**Research on Ontology Construction Method Based on Formal Concept Analysis**

An Dong

【absract】With today's rapid development of the World Wide Web, the information on the World Wide Web is becoming more and more extensive. With the rapid increase in the amount of information, people in the search of these data, resulting in a big problem. There is a problem of this recall and precision in the World Wide Web. Formal Concept Analysis (FCA) is a powerful tool for data analysis and rule extraction from formal background. Formal concept analysis is based on mathematics, and the concept, attributes and relations of composition ontology are expressed in formalized context, which is widely used today. In the World Wide Web, we may use different terms to express the same meaning, or a term containing multiple meanings. Therefore, the elimination of terminological differences is necessary. The current solution is to create a common ontology. In the design of the field, we use the common ontology to regulate. The concept of ontology and the concept of the relationship between the unity of expression in the concept lattice.Based on the concept of the concept of ontology construction, to avoid the concept of redundancy to ensure the principle of ontology maximum scalability.

【Key words】Formal concept analysis; concept lattice; ontology construction;

0 引言

随着今日万维网的蓬勃发展，万维网中的信息，如信息共享和信息交互变得越来越庞大。数据量也变得越来越越大。这样用户在万维网中检索这些庞大的数据时就会产生很大的问题。比如，查全率和查准率。目前，万维网中依靠关键字的查找，这样的基于关键字的查找，会对用户产生大量的无关数据，万维网中存在这查全率和查准率低的缺陷。那么如何解决这样的问题？更好的满足用户的对信息检索的需求。那么为因特网中信息提供具有计算机可以理解的语义【[[1]](#endnote-0)】。主要思想是通过在语义Web中引入本体来表达信息，实现语义信息的共享。形式概念分析是建立在数学的基础上，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。如何基于形式概念分析的本体构建，下面将进行详细的介绍。

1 形式概念分析和本体的概念

1.1 形式概念分析的基本概念

形式概念分析（Formal Concept Analysis,FCA）【[[2]](#endnote-1)】是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的工具。形式概念分析建立在数学的基础之上，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。一个概念就是最大限度地收集对集合中共同特点有帮助的元素， 并且运用形式概念分析的方法 ， 可以发现构造和 展示由属性(Attributes )和对象(Objects)构成的概念 ( Concept) 及其之间的关系。因而形式概念分析的方法已经运用在软件开发等众多环节之中。

1.2 本体的概念

本体（Ontology）的概念最早源于亚里士多德（公元前384-322年）对事物存在本质的研究，在哲学上的定义为“对世界上客观存在物的系统地描述，即存在论”，是客观存在的一个系统的解释或说明，关心客观现实的抽象本质【[[3]](#endnote-2)】。近十几年来，这个词被用在了计算机界，并在人工智能和、计算机语言以及数据库理论中起到越来越重要的作用。本体被哲学定义过的范畴，在现今被人工智能赋予了新的定义，这使得它被带入到了信息科学中。信息科学对本体的定义是慢慢走向成熟的。1991年Neches等人最早给出Ontology在信息科学中的定义：“给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延规则的定义。”后来在信息系统、知识系统等领域，随着越来越多的人研究Ontology，产生了不同的定义。1993年Gruber定义Ontology 为“概念模型的明确的规范说明”。1997年Borst进一步完善为“共享概念模型的形式化规范说明”。Studer等人对上述两个定义进行了深入研究，认为Ontology是共享概念模型的明确的形式化规范说明，这也是目前对Ontology概念的统一看法。下面对本体进行举例来更好的理解本体。

1.2.1、本体的举例

拿苹果来举例，中文的“苹果”、英文的“apple”、“苹果的图片”。都可以让人们知道这是表现苹果这个东西。但是它们对苹果的表现却有很多种表现方法，但这些都是对本体的映射。在哲学层面上，苹果这样的东西是亚里士多德口中的实体。

在万维网中，我们可能会用不同的术语来表达相同的含义，或者一个术语含有多个含义。因此，消除术语差异是很有必要的。目前的解决方法是，建立一个公共的本体。在设计该领域的时候，我们都使用这个公共本体来进行规范。

2.形式概念分析的定义

设K=（G,M,I）为一个形势背景，包含两个稽核（G和M）和变现了一种二元关系，在这个语境中,在这个语境中,G中的元素称为对象，M中的元素称为属性。一般用glm，或者（g,m）∈I 来表达对象g和属性m的关系，读作“对象g具有属性m”。

根据定义1，可通过矩阵来表示语境。每行的开头是对象名，每列的开头属性名。行g和列m的交叉表示对象g具有属性m ,T/F 该对象是否具有这些属性。

例如：设有形势背景（G，M,I），G={d,e,f,g} G是对象集。M={a,b,c,} M是属性集。T代表具有该属性，F代表不具有该属性。见表1。

表1 预语境的矩阵表示

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 属性(M)  对象(G) | d | e | f | g |
| a | T | T | F | T |
| b | T | T | F | T |
| c | F | T | F | T |
| d | T | F | T | F |
| e | T | F | F | F |

该形式背景共有6个概念： c5({d},{a,b,d,e})， c2({e,g},{a,b, c}), c3({d,f},{d})， c4({d,e,g},{a,b})， c1(G, Ø)， c6 ( Ø,M)。为了方 便描述，将其简记为(1, abde)， (24, abc)， (13, d)， (124, ab)， (G,Ø )， ( Ø,M)，其概念格如图 1 所示。

(G, Ø)

c1

c1

(13, d) (124, ab)

c2

c3

(1, abde) (24, abc)

c4

c5

( Ø,M)

c6

图1 形势背景概念格

2 基于FCA构建本体的方法

FCA和本体的两种方法，表达的差异不起作用，他们是主体间交互的概念的重要性之一，强调需要对模型进行正式化。不同之处在于，本体旨在提供共识以支持知识密集型应用，而FCA基于给定的数据，知识分析和结构的领域，是人造产品。 FCA依赖于一组给定的对象和数据，并且可以在没有数据的情况下创建本体。因此，在FCA中，扩展和内涵的概念是两个同等重要的方面，而本体强调概念网格

内容。由于FCA和本体具有自己的特征，目前的研究主要结合两个方向：一方面FCA作为一种应用于本体工程的技术，FCA与网格数据的概念相关，用于提取概念层次作为本体论的基础应用，用于手动或半自动生成身体。例如，Gu方法【[[4]](#endnote-3)】，Obitko方法【[[5]](#endnote-4)】，其中Gu方法包含OKBC [22]知识模型。除了上述两种方法之外，还有Haav方法【[[6]](#endnote-5)】，Ciminao方法【[[7]](#endnote-6)】等。将FCA引入到本体生成过程中可以解决发现概念之间的关系的难题。手动将该概念组织到本体中的主题是影响时间和主观的。它表达从概念格中的给定数据导出的概念，帮助找到所有可能的概念和概念之间的关系[[[8]](#endnote-7)]。四种利用FCA构建本体的方法比较见表2。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Gu | Obitko | Haav | Cimiano |
| 初始形势背景 | 领域中获得 | 人工构建 | 利用自然语言处理方法获得 | 从领域文本中获得 |
| 概念对应关系 | 两者中概念等同 | 两者中概念等同 | 两者中概念等同 | 本体中概念对应FCA中的属性 |
| 概念格转换为本体 | 半自动 | 手工 | 半自动 | 手工 |

3 结语

基于形式概念分析的本体构建方法的研究，分析了形式概念分析的概念、定义。概念格的定义。对本体的定义，基于FCA构建本体的方法。基于FCA构建本体的方法可以有效的解决了本体构建的复杂的概念，将本体概念以及概念之间关系统一表现在概念格之中。基于形式概念分析的本体构建，避免概念冗余保证本体最大可扩展性的原则。

4 参考文献

1. Berners-Lee T,James Hendler , Ora Lassila. The Semantic Web[J]. Scientific American,2001,5 [↑](#endnote-ref-0)
2. Wolff K E. A First Course in Formal Concept Analysis- How to understand line diagram[J].Statistical ,1993,14(4):429-438. [↑](#endnote-ref-1)
3. 邓志鸿，唐世渭，张铭等。Ontology研究综述，北京大学学报（自然科学版），2002，38 730-738 [↑](#endnote-ref-2)
4. Haav H M A Semi-automatic Method to Ontology Design by Using FCA. In: Proceedings of the 2nd International CLA Workshop on Concept Lattices and their Applications, 2004:13-25. [↑](#endnote-ref-3)
5. Cimiano P, Staab S, Tane J. Automatic acquisition of taxonomies from text: FCA meets NLP. In: Proceeding of the International Workshop on Adaptive Text Extraction and Mining, 2003:10-17. [↑](#endnote-ref-4)
6. Cimiano P, Staab S, Tane J. Deriving concept hierarchies from text by smooth formal concept analysis.In: Poce. of the GI Workshop on Lehren Lemem-Wissen-Adaptivitat (LLWA), 2003:72-79. [↑](#endnote-ref-5)
7. STUMME FCA-Merge: bottom-up merging of ontologies in proc of the 17th intl. conf. on artificial intelligence, seattle, WA, USA, 2001:225-230. [↑](#endnote-ref-6)
8. Ganter B,Stumme G,Creation and merging of ontology top-levels In Computational Science (ICCS '03), 2002:131-145. [↑](#endnote-ref-7)