基于形式概念分析的领域本体构建方法研究*)

黄美丽[1,2] 刘宗田[1]

(上海大学计算机工程与科学学院 上海 200072)[1] (浙江林学院信息工程学院 临安 311300)[2]

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

Research on Domain Ontology Building Methods Based on Formal Concept Analysis

HUANG Mei-Li[1,2] LIU Zong-Tian[1]

(School of Computer Engineering and Science, Shanghai University, Shanghai 200072)[1]

(School of Information Science and Technology, Zhejiang Forestry University, Linan 311300)[2]

Abstract Recently, ontology as an effective model of representing concept hierarchy and semantic is being applied to increasing domains. It can be said that ontology can perfectly solve some problems existing in the computer application domains today, including human-computer interaction, automatic reasoning, knowledge representation and reuse. However, ontology building appears a new difficult problem before ontology is effectively applied. In this paper, firstly, the existing methods of domain ontology building are briefly presented, then several methods of ontology building based on formal concept analysis are described in detail. Finally, this paper analyses and compares every method mentioned above and makes conclusion of them.

Keywords Formal concept analysis, Ontology, Ontology building

1 引言

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。

那么,在确定了领域的情况下,从领域中找到概念以及概念之间的关系是问题的关键。事实上,对于现实生活中的某一领域,与该领域相关的概念以及概念之间的关系是隐含在人们头脑中的,或者是存在领域文档中的。从这个角度讲,构建本体的过程中,一个很大的挑战是如何才能将这些隐含的领域知识用本体显式地表示出来。

本文主要介绍基于形式概念分析的领域本体的构建方法。文中第2部分将系统介绍一般的本体构建方法,第3部分详细介绍形式概念分析用于领域本体的构建,最后给出分析和总结。

2 本体构建方法

目前,构建本体大多采用手工方式,远远没有成为一种工程性的活动。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。两种主要的用来构建本体的方法主要有两种:1)在领域专家的帮助下用本体描述语言将本体描述出来;2)从结构化的数据或文本中抽取或学习或发现领域本体。

用第一种方法构建本体,是完全手工构建本体。对于一些复杂的应用领域而言,这将是一项费时费力的任务,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,构建出来的本体都将是千差万别的。这样,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。

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:

· 从词典进行本体学习。将构建本体建立在已有的机器

*)本文受国家自然科学基金(60275022)、上海市科委项目(035115028)和上海市高等学校青年发展基金(03AQ99)资助。黄美丽 硕士、助教,研究方向为人工智能;刘宗田 教授,博导,从事人工智能、软件工程等方面的研究。

可读的词典的基础上，从中抽取相关的概念和概念间的关系；

• 从知识库中进行学习。通过从已有知识库中的学习来构建本体；

• 从关系数据库中抽取本体；

• 从半结构化的数据学习。从类似于 XML Schema 这样的半结构化的数据源提取概念和概念之间的关系，以构建本体；

• 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。

典型例子有：

• Woelk and Tomlinson1994 年在 Infosleuth 系统中的做法：从文本数据库（textual databases）中半自动地构建本体。他们使用了本体学习方法，在学习的过程中，领域专家提供一系列的从高层本体中获得表征概念的初始词，系统处理文档，抽取包含初始词的短语，产生相应的概念术语，并将它们分类到本体中去。这是一个不断的迭代的过程，每一步的结果都由领域专家进行验证。

• Preece 等 1999 年在 KRAFT 中的做法：从共享本体获得本地本体。

• Maedche 和 Staab 2000 年使用浅文本处理方法（shallow text processing methods）从文本中发现非分类的关系，他们的技术集成到了本体学习工具 TextToOnto 中[4]。

3 形式概念分析用于本体构建

形式概念分析，即 FCA（Formal Concept Analysis），其核心的数据结构为概念格，提供了一种支持数据分析的有效工具。概念格的每个节点是一个形式概念，它由两部分组成：外延和内涵。外延是概念所覆盖的所有对象；内涵是概念的描述，是该概念覆盖对象的共同特征。概念格通过 Hasse 图生动简洁的体现概念之间的泛化和特化关系。由于概念格是进行数据分析的有力工具，所以在信息检索、数字图书馆、软件工程和知识发现等方面有广泛的应用。近几年，FCA 在本体中的应用越来越引起人们的重视。下面我们将详细描述目前已有的将 FCA 用于本体构建的一些工作。

3.1 形式概念分析与本体

正如上面所述，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域构建本体，但是 FCA 必须建立在给定数据集（形式背景）的基础上。

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工作者在做 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。尽管有许多的研究工作者在做 FCA 与本体的结合工作，但是 FCA 在半自动构建本体中的应用是最近提出来的想法，处于刚起步的阶段。3.2 节将详细描述目前已有的这方面工作。

3.2 形式概念分析用于本体构建的方法

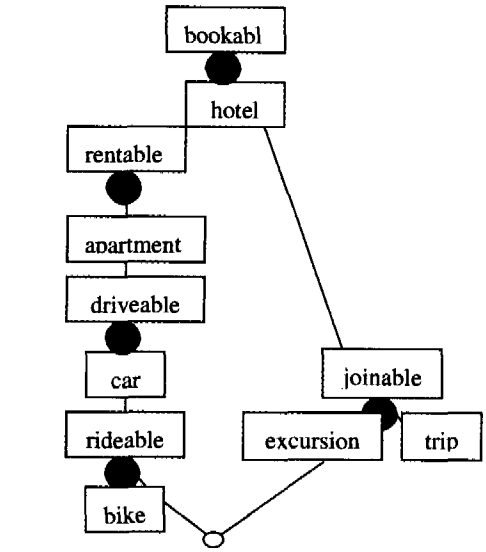
3.2.1 Philipp Cimiano 的方法

这是 AIFB 研究机构在 IST-Dot Kom 项目（http://www.dat-kom.org）中应用的方法[5,6]，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；进一步通过词典查询，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。

从概念格到领域本体的实现如图 1。

1）直接将概念节点的底下元素从概念节点中移走；

2）为每个移走底下元素的概念节点添加子概念节点。



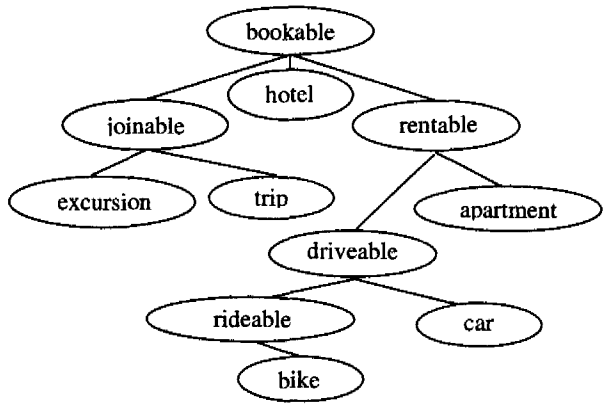


图 1 从概念格转换到本体

3.2.2 GuTao 的方法

他提出的形式概念分析用于本体构建的方法[7]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[8]，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[9]

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。

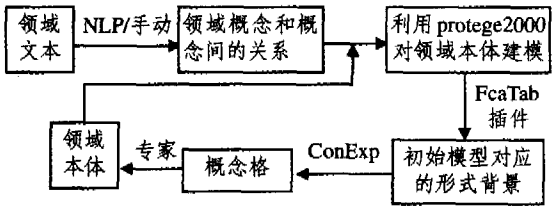


图2 GuTao的基于FCA的本体构建过程

3.2.3 Marek Obitko的方法

Marek Obitko等人在GACR项目中提出如下方法[10]：

·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本体表示。

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。

3.2.4 Hele-Mai Haav的方法

Hele-Mai Haav提出了基于概念格的本体表示方法[11,12]，主要适用于领域文本内容比较短的情况，而且假设领域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(objects)：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NLP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应用的形式背景，其中的二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。为了获得基于概念格的本体表示，对获得的概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 set of facts)，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(通过将本体描述翻译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量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 方法的分析比较与小结

根据以上对各种方法的详细描述，我们对4种方法做分析比较得表2。

我们可以看出，要将FCA应用到本体的构建中来，需要注意以下几点：

1)对于不同的应用目的，FCA与本体的结合方法是不相同的。

怎样将FCA与本体结合将是一个很重要的研究课题。

2)结合NLP与FCA来完成本体的构建任务。

通过FCA固然可以帮助自动获得一些隐含的本体概念和概念间关系，但是初始的本体概念和关系无法通过FCA获得，这一步我们可以借助于NLP或手动方法获得。

3)离不开领域专家的参与。

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。

4)处理复杂领域时有待改进。

(下转第239页)

latent semantic indexing. In: Proc. of SIGIR-94, 17th ACM Intl. Conf. on Research and Development in Information Retrieval (Dublin, Ireland, 1994), 282~289
7 Lam S L, Lee D L. Feature reduction for neural network based text categorization. In: Proc. of DASFAA-99, 6th IEEE Intl. Conf. on Database Advanced Systems for Advanced Application (Hsinchu, Taiwan, 1999), 195~202
8 Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Proc. of ECML-98, 10th European Conf. on Machine Learning (Chemnitz, Germany, 1998), 137~142
9 Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules. In: Proc. of VLDB, 1994. 487~499
10 Antonie M, Zaiane O R. Text Document Categorization by Term Association. In ICDM, 2002. 19~26
11 Han J, Pei J, Yin Y. Mining Frequent Patterns without Candidate Generation. In SIGMOD, 2000. 1~12
12 Brutlag J D, Meek C. Challenges of the Email Domain for Text Classification. In: Proc. of ICML, 2000. 103~110
13 Agrawal P, Srikant R. Mining Sequencial Patterns: [Research Report]. 1995
14 Ayres J, Gehrke J. Sequential PAttern Mining using A Bitmap Representation, SIGKDD 02
15 Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 2002, 34(1): 1~47
16 Grahne G, Zhu J. Efficiently Using Prefix-trees in Mining Frequent Itemsets. In: Proc. FIMI, 2003
17 Dumais S, Platt J, Heckerman D, Sahami M. Inductive Learning Algorithms and Representations for Text Categorization. In CIKM98. 148~155

(上接第 212 页)

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。

结束语 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。

表 2 四种方法的比较分析

		Philipp Cimiano	GuTao	Marek Obitko	Hele-Mai Haav
概念及其关系获取		NLP	NLP/手动	NLP/手动	NLP
背景形式	对象	领域词汇（名词）	类	领域实体	领域文本的编号
	属性	领域词汇（动词）	槽	实体属性	领域词汇集
本体表示		格	Protégé 模型	三元组	格
优点		实现了本体的自动构建；易于实现本体的更新；通过把相同上下文的名词处理为相同的词来解决同义词问题；提供了一套本体的评价方法。	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的领域本体构建；可消去分类结构中概念的冗余，并得需要的概念。	提供了分布式的本体编辑环境；按属性进行分类，克服了当前分类方法存在的问题；这种方法构建的本体适用于知识交换。	自动构建领域的形式化本体；实现了本体的逻辑表述，从而易于本体的推理和验证；易于实例化本体和检索本体实例；适用于领域文本内容比较短的情况。
缺点		没有考虑词的多义情况；概念分类相对单一，只考虑了 is-a 关系。	对于属性值是多值的情况，必须先转换成单值才可以使用 FcaTab。	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的迭代的过程，所以到底需要添加或删除哪些内容难以把握，而且具体到什么时候结束也不容易确定。	需要对缩减后的格中的概念命名，并且将概念和具体的文档做映射，这两项任务都是不容易做到的。

参 考 文 献

1 Uschold M. Ontologies Principles, Methods and Applications. Knowledge Engineering Review, 1996, 11(2)
2 Gruninger M, Fox M S. Methodology for the Design and Evaluation of Ontologies. Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Montreal, 1995
3 Fernandez M, Gomez-perez A, Juristo N. Methontology: From Ontological Art Towards Ontological Engineering. AAAI-97 Spring Symposium on Ontological Engineering, Stanford University, 1997
4 http://cui.unige.ch/~hilario/icdm-01/DM-KM-Final/Volz.pdf
5 Cimiano P, Staab S, Tane J. Automatic acquisition of taxonomies form text: FCA meets NLP. In: Proceeding of the International Workshop on Adaptive Text Extraction and Mining, 2003
6 Cimiano P, Staab S, Tane J. Deriving concept hierarchies from text by smooth formal concept analysis. In: Proc. of the GI Workshop "Lehren Lernem-Wissen-Adaptivitat" (LLWA), 2003
7 Gu Tao. Using Formal Concept Analysis for Ontology Structuring and Building. ICIS, Nanyang Technological University, 2003
8 http://protege.stanford.edu
9 http://sourceforge.net/projects/conexp
10 Marek O, et al. Ontology Design with Formal Concept Analysis In: Snasel V, Belohlavek R, eds. Concept Lattices and their Applications, Proceedings of the 2nd International CLA Workshop, TU of Ostrava, 2004
11 Haav H M. An Application of Inductive Concept Analysis to Construction of Domain-specific Ontologies. In: Thalheim B, Fiedler G, eds. Emerging Database Research in East Europe. Proceedings of the Pre-conference Workshop of VLDB 2003, Computer Science Reports, Brandenburg University of Technology at Cottbus, 2003. 63~67
12 Haav H M. A Semi-automatic Method to Ontology Design by Using FCA. In: Snasel V, Belohlavek R, eds. Concept Lattices and their Applications. Proceedings of the 2nd International CLA Workshop, TU of Ostrava, 2004. 13~25
13 Noy F N, McGuinness D L. Ontology Development 101: A Guide to Creating Your First Ontology: [Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880]. 2001