《智能信息处理》课程作业

**研究论文题目研究论文题目研究论文题目**

蒋志新

|  |  |
| --- | --- |
| 作业 | 分数[20] |
| 得分 |  |

2016年10月1日

基于形式概念分析的本体构建与展现

蒋志新（大连海事大学 大连 116026）

**摘 要** 作为语义网基础的本体是共享概念模型的明确的形式化规范说明，它的作用是获取、描述和表达相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇和词汇间相互关系的明确定义。本文的主要贡献如下，提出了形式概念分析在基于非结构化数据的本体学习中的应用方法。提出两种展现本体的方法，一、结合几率模式来表达本体中非层次关系概念之间的相似性以及属性概念与对象之间的相关度；二、基于概念格的三维可视化形式来展现，为了简化显示，提出子概念格的判定方法及其折叠与打开算法。提出了基于形式概念分析的本体合并方法。

**关键词** 本体；形式概念分析；概念格；本体构建

**Building and Presenting of Ontology Based on Formal Concept Analysis**

**Jiang Zhixin**

**(Dalian Maritime University ,Dalian 116026,China)**

**Abstract** As the foundation of the semantic web, ontology is a formal, explicit specification of a hared conceptual model. The aim of ontology is to obtain, describe and express the knowledge of related domain. Ontology provides common understanding of the domain knowledge and confirms common approbatory vocabulary in the domain, as well as gives specific definition of the relation between these vocabularies from formal model of different levels. The main contributions are as follows: (1) Putting forward formal concept analysis method applied in ontology learning based on non-structured of source data. (2) Proposing two means of ontology presenting: 1. expressing similar relation of concepts in non-hierarchy and the relevancy of concepts and objects in combination with probabilistic model in ontology; 2. Presenting ontology based on the form of concept lattices visualizing in three-dimensional space, for the sake of reducing display, proposing the determination of sub-concept lattice and the algorithm for folding and unfolding it. (3) Presenting the method of ontology merging based on formal concept analysis.

**Key words** ontology ;formal concept analysis; concept lattice; concept lattice union

1概念

1.1本体的概念

本体的概念来源于哲学领域，在计算机界，本体的定义不断变化。1993年，Gruber给出了本体的一个最为流行的定义，即“本体是概念模型的明确规范说明[1]”，后来，Borst对此稍作修改，提出：“本体是共享概念模型的明确的形式化规范说明”。这个定义主要包含4层含义：概念化、明确、形式化和共享。其中“概念化”指本体是客观世界现象的抽象模型；“明确”指所使用的概念及使用这些概念的约束都有明确的定义；“形式化”指本体是计算机可读的（即能被计算机处理）；“共享”指本体中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，即本体针对的是社会范畴而非个体之间的共识[2]。本体的目的是：用于获取、描述和表达相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇和词汇间相互关系的明确定义。通俗地讲，本体就是用来描述某个领域甚至更广范围内的概念以及概念之间的联系，使得这些概念和联系在共享的范围内有着明确唯一的定义，达成一种共识，便于人、机器进行交流。

1.2形式概念分析

形式概念分析FCA（Formal ConceptAnalysis）是应用数学的一个分支[3]，它来源于哲学领域对概念的理解。根据用二元关系来表达领域中的形式背景，从中提取概念层次结构，概念格，从数据集中生成概念格的过程实际上是一种概念聚类的过程，它的每个节点被称为一个形式概念，形式概念由外延和内涵两部分组成。概念的外延指属于这个概念所有对象的集合，内涵则是指所有这些对象所共有的特征（或属性）集。概念格作为形式概念分析中核心的数据结构，从外延和内涵两方面对概念进行符号形式化描述，实现计算机可以理解的语义信息。本质上描述了对象和特征之间的联系，表明了概念之间的泛化与例化关系。由于概念格为数据分析提供了一种有效的工具，目前该理论已被应用于软件工程[4,5,6,7]，知识发现[8,9]，信息检索[10,11,12]等方面。对于概念格的显示，概念格 Hasse 图虽然能清晰和简洁地表达概念与概念之间的层次结构，但终究不太直观生动。文献[13]提出了概念格的三维自动布局策略，主要解决二维布局中的横向过度扩张和线段交叉的问题，实现了概念格在三维空间的自动布局、动态显示及与用户的互操作。概念格的合并[14]主要源于概念格的分布式存储与并行处理的需求，其基本思想是：通过形式背景的拆分，形成分布存储的多个子背景，然后构造相应的子概念格，再由子概念格的合并得到所需的概念格，也就是说可以通过多个概念格的合并来构造所需的概念格。而多个概念格的合并有横向合并和纵向合并两种。

2形式概念分析用于本体的构建

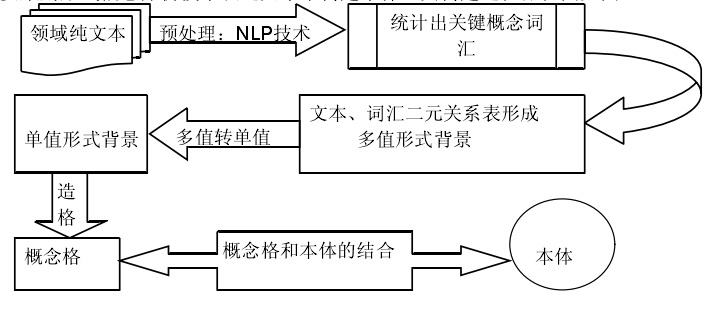
上节探讨了现有本体构建的各种方法，从中发现构建本体的关键问题是（1）、为了使构建的本体具有通用性，且不受开发者的主观影响，需要采用一种成熟、统一的构建方法；（2）、寻找到领域中所有的隐含概念以及概念间的分类关系；（3）、本体中的知识概念以及概念间的关系必须以明确地形式化方式来表达，以便于计算机的处理。而形式概念分析就是从给定数据中自动提取出所有的隐含概念以及概念之间的分类关系，形成概念模型（概念格）。概念格作为形式概念分析的核心数据结构，从外延和内涵两方面对概念进行符号形式化描述，具有明确的层次关系，含有丰富的语义信息，实现了计算机可以理解的语义信息。同时概念格Hasse图能清晰地表达概念与概念之间的层次结构，而且概念格的三维可视化形式更加直观生动地展现所形成的格模型。另外，本体和形式概念分析都来源于哲学领域，概念格由概念的层次关系组成，内涵、外延构成了概念，而本体也是用来体现概念与概念间关系的。因此采用形式概念分析的方法可以帮助构建本体，并为本体的构建提供了一种统一、成熟的方法指导。对于从结构化数据中进行本体构建，有人已经提出结合给定的数据库模式及其数据信息，运用形式概念分析的方法构建本体[15]。但目前网络上大量存在的是类似纯文本的非结构化数据和类似XML、HTML格式网页的半结构化数据，所以这些非结构、半结构化数据应该作为本体构建的主要数据来源。下面主要针对从非结构化数据中进行本体的构建。纯文本是Web中大量存在的一类非结构化数据，它依据一定的造句法表达语义信息，读者可以基于一些背景知识来理解其中的含义。由于缺乏一定的结构，要使机器能够自动的理解纯文本并从中抽取出需要的知识，必须利用自然语言处理（NLP）技术对其预处理。本文主要采用形式概念分析技术从纯文本中构建本体，其构建过程流程图如图

图2-1 基于FCA本体构建流程图

该本体构建流程大致过程分为以下几步：

（1）、首先利用自然语言理解技术（NLP）对收集来的纯文本进行预处理，取得文本中的字词集合。

（2）、利用概率统计的方法，一般采用TF-IDF方法[16]获得能代表文本的关键概念词汇，具体的方法可以是：计算概念词汇在文本集中出现的频率，如果该频率大于指定的阀值，则将其作为领域本体中的概念。然后针对于所找出的概念词汇，结合相应的文本集合形成词汇、文件的二元关系表。

（3）、由于该二元关系表里有多值存在，为了便于造格，必须转换成单值的二元关系表。再由单值形式背景按照造格方法来构造概念格，这时所构造的概念格还不是最终所要的本体。

（4）、最后，还要探讨如何将概念格和本体进行结合[17]，即从概念格转换成相应的本体。这里所用的方法是：采用简化的方法用属性来代表所形成的形式概念，并且在标注时只让属性在概念格中出现一次，由于这里的属性都是词汇，而本体所描述的重点元素也都是词汇概念，因此可以用概念格中的属性来表示本体概念。这就最终形成了本体。另外可以采用概念格中的形式概念直接等同于本体概念，以及概念格中的对象来代表本体概念或者实例。

3 本体的展现

对于本体的展现，有两种表现形式。一种是根据其对应的概念格展现，将本体在格中表现出来。关于概念格的展现，本文采用基于三维布局的概念格展现方式。那么对于复杂的格结构，为了简化概念格的三维显示，本文针对子概念格的可视化互操作，提出关于子概念格的判定定理以及折叠与打开算法。因为以概念格结构为基础的本体主要体现的是概念之间的层次关系，而不具有层次关系的概念只能借助其他模式来表达相似关系。那么就引出了第二种方法：结合概念间的相关性几率模式来展现本体[18,19]，主要是针对本体中具有非层次关系概念之间的相似性以及属性概念与对象之间的相关度。

参考文献

[1] Gruber T R.A Translation Approach to Portable Ontology Specification.Technical Report of Knowledge System Laboratory(KSL).1993:92-71

[2] 宋炜，张铭.语义网简明教程[M].北京:高等教育出版社,2004:108-131

[3] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations [M].

Berlin: Springer Verlag, 1999

[4] Godin R, Mineau G, Missaoui R, St-Germain M, Faraj N. APPlying conce Pt formation methods to software reuse[J]. International Journal of Software Engineering and Knowledge Engineering,1996, 5(l):119-142

[5] Godin R, Mili H, Mineau GW, Missaui R, Arfi A.Design of c 1 ass hierarchies based on conce Pt(Galoi S)lattices[J].Theory and a PPlication of object systems,1997, 4(2): 117-134

[6] ISnelting G. Reengineering of configurations based on mathematic conce Pt analysis.Technical re Port TR95-02，Software Technology De Partment，Technical University of Braunschweig,1995

[7] Schmittl, Saake G, (1998). Merging Inheritance hierarchies for scheme integration based on conce Pt.lattices, htt P://www.tudarmstadt.de/

[8] Sahami M. Learning classification rules using lattices (Extended Abstract)[C]. In ECML-95:Proceedings of the Eighth European Conference on Machine Learning, Berlin, Germany:Springer-Verlag, 1995:343-346

[9] Njiwoua P, Mephu Nguifo E. Forwarding the choice of bias, LEGAL-F: using feature selection to reduce the complexity of LEGAL[C]. In Proceedings of ENELEARN-97, ILK and INFOLAB, Tilburg University, Netherlands, 1997:89-98

[10] U .Krohn, N.J.Davies&R.Weeks, Concept lattices for knowledge management.BT Technol J VOI 17 No 4 October 1999

[11] Godin R, Missaoui R, &April A. Experimental comparison of navigation in a Galois lattice with conventional information retrieval methods[J]. International

Journal of Man-Machine Studies, 1993, 38:747-767

[12] Neuss C, Kent R E.(1999). Conceptual Analysis of Resourse Meta-Information. http://www.igd.fhg.De/.

[13] 马骏. 概念格及其可视化研究. 河南大学硕士学位论文：2005.

[14] 李云,刘宗田,陈崚等.多概念格的横向合并算法[J].电子学报,2004,32(11):1849～1854

[15] Salton G. Introduction to Modern Information Retrieval[M]. Mc Graw Hill Book Co., 1983.

[16] 黄伟，金远平.形式概念分析在本体构建中的应用[J].微机发展，2005,15(2):28-31

[17] Cimiano P,Staab S,Tane J.Automatic acquisition of taxonomies from text:FCA meets NLP[C].In: Proc.of the Intl.Workshop on Adaptive Text Extraction and Mining,2003.

[18] Kang S H, Huh W, Lee S, Kim Y. Automatic Classification of WWW Documents

Using a Neural Network[C]. In: Intl. Conf. On Production Research, Bangkok,

Thailand, 2000.

[19] Neto J L , Santo A D. Document clustering and text summarization[C]. In: Proc. Of 4th Intl. Conf. on Practical Applications of Knowledge Discovery and data Mining

(PADD-2000), 2000:41～55.