基于范畴论的形式化本体研究*)

章 远 李师贤
(中山大学计算机系 广州 510275)

摘 要 本体技术是语义 Web 的核心。现有的本体方法都是基于集合论的,本文从范畴论的层次分析了集合论数学的局限性,以及这种局限性对现有本体方法的影响,并探讨基于范畴论的本体方法,通过实例系统说明了范畴论本体的构建方法,并指出范畴论本体所克服的现有本体的不足。

关键词 形式化本体,范畴论,略图,图逻辑

Research on Formal Categorical Ontologies

ZHANG Yuan LI Shi-Xian
(Department of Computer, Sun yat-sen University, Guangzhou 510275)

Abstract Formal ontologies are very important for semantic Web. While all formal ontologies used today are based on set theory, we provide a new kind of formal ontology based on category theory. We dissertate why categorical ontologies are better from the point of view of category theory. Some examples show how to build categorical ontologies and that they are more expressive than others.

Keywords Formal ontology, Category theory, Sketch, Graph based logic

1 引言

目前已经有许多种本体方法和本体语言,大体分为逻辑方法和代数方法[1,2]。由于逻辑语言的形式语义以集合论为基础,代数系统的定义也必须基于集合论,因此当前的本体描述方法都归结于集合论。本文把以集合论为基础的数学称为集合论数学。从范畴论的层次分析,集合论数学是有局限性的,该局限性是指其抽象程度不足,难以表达范畴论的概念;相反,大量集合论数学的概念可用范畴论表达,例如高阶逻辑的语义可以用范畴论的概念表达[3]。这种局限性使得现有本体方法不适合表述范畴论语义,它表现为两种情形,其一,某些问题域语义用范畴论表述明显会更简捷和明确;其二,某些语义已经无法用现有本体方法表示,只能采用范畴论方法。针对应用领域语义表述的需求,本文建立了范畴论的本体表示方法。范畴论已经用于数据库的概念模型[4],概念模型和本体虽然使用目的不同,但两者的内容很相似,所以可以理解为本文的相关工作。

在软件工程领域,一些基于图的语义表述方法获得成功,例如实体关系模型和 UML 模型,而基于形式语言的软件规约方法比较而言却未得到广泛应用,这些说明图形是更容易被接受的表达方式,范畴论的表达方式就是一种基于图形的略图语言,这是范畴论本体的一个优点。研究还发现,由于范畴论概念有良好的重用性,这使得基于范畴论的本体普遍适用于各类问题域。

范畴论在语义描述方面的优势主要得益于范畴论比集合论数学更抽象,本文第 2 部分讨论此问题。第 3 部分说明范畴论本体是什么以及如何构建和表达。第 4 部分通过比较,体现出范畴论本体的具体优势。最后是本文的结论。

2 范畴论与集合论数学

范畴论和集合论数学分析问题和提出概念的方法有以下不同:

1)分析问题的起点不同

集合论数学是从元素及其归属开始的,进而定义关系和函数等概念。而范畴论首先确定与问题相关的对象和关系,使之构成范畴,或者确定该问题属于何范畴,然后再用统一的范畴论方法分析问题。例如集合论数学中,群的概念起源于某些具体的集合 G 和满足一定条件的二元运算,而范畴论认为群是 SET 范畴(由集合和函数构成的范畴)中的一个局部结构[5]。

2)范畴论和集合论数学抽象的角度不同

集合论数学的抽象是针对集合和元素与元素之间的关系,例如群的概念抽象出了一类由集合和二元运算构成的系统,而范畴论是从范畴的箭头关系的角度进行抽象,例如广义的群概念是在任意范畴中满足特定关系的一组对象和箭头[5]。由于这种差异,用集合论方法表述范畴论的概念会很困难,例如在范畴论中,"图 1 是某范畴中的积(product)"这句话通过"积"的概念,很简捷地描述了图 1 的数学特征;如果用一阶语言表述相同的语义,则是一个很长的公式[6]:

$$\forall f \forall g((D(f)=D(g) \wedge C(f)=A \wedge C(g)=B) \rightarrow (\exists ! z)(zx=f \wedge zy=g))$$

式中:$D(f)$和 $C(f)$分别指 f 箭头的起点和终点,z 是唯一的满足对易条件的箭头。

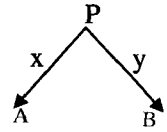

图 1 范畴论中的积结构

3)范畴论和集合论数学抽象的程度不同

由于抽象角度不同,所定义的概念的抽象程度也不同。

*)基金项目:广东省科技计划工业攻关项目(2003A1030403)资助。章 远 博士研究生;李师贤 博士研究生导师。

集合论数学各个分支的概念往往局限于某个具体范畴中，以几个重要数学分支为例。拓扑空间的概念只是集合与⊆关系构成的范畴中的结构，该结构要求拓扑空间的开集对积(product)和共积运算(coproduct)封闭。逻辑研究针对如下范畴：对于一种逻辑，它的所有句子构成集合 S，以 S 的子集为对象，以句子集合之间的逻辑蕴涵关系为箭头构成的范畴。代数研究是针对 SET 范畴。与集合论不同，范畴论的概念是基于范畴中箭头关系加以定义，不论这些箭头的具体内容是什么，所以更抽象，例如代数的群就是范畴论的群概念的特例。这就是集合论研究方法相对于范畴论的局限性。

4)范畴论和集合论数学的概念的重用性不同

由于范畴论概念更抽象，因此它的概念具有良好的重用性。首先，由于范畴是一种抽象和普遍的数学结构[7,8]，大量的数学问题体现为范畴，使得有些学者认为范畴论是沟通各个数学分支的有效语言[9]，这是范畴论概念可重用的基础；其次，范畴论概念能普遍重用于各种具体范畴，例如积(product)。如前所述，集合论数学各分支分别研究不同的具体范畴，各自的概念与这些具体范畴的对象和箭头的定义紧密相关，所以集合论数学的概念都只能在个别具体范畴中使用，它的各个分支之间重用概念很困难，例如考察某个代数系统是否是“量度空间”或某个拓扑空间是否是“交换的”，这类问题在绝大多数情况下没有意义；在范畴论中，“一个 topos 是否是一个 monoidal category”却是意义很清楚的问题，所以集合论数学相比而言不适合重用。

为什么在应用领域仍然应该选择更抽象和更具重用性的概念？下面的两种表述可以体现区别：

a)1 是〈自然数，×〉的单位元

b)任何自然数乘以 1，值不变

表述 a)中重用了单位元的概念，隐藏了运算的细节；b)表述含细节，所以繁琐。对于同样的语义，范畴论和集合论表述的区别与此类似。

在上述比较中 2)和 3)是范畴论与集合论数学的根本区别，1)和 4)是范畴论适合构建本体的直接原因。

为某个问题域构建形式化本体的实质过程就是：首先将问题域抽象为某种数学对象，然后用相应的数学语言描述该对象。例如代数方法将问题域抽象为代数系统，然后描述。好的本体分析方法必须是对各类问题域一致的，在语言方面应选择重用强的概念。集合论数学的局限性决定了现有本体方法的局限性，它们都局限于 SET 范畴，难以表述其它范畴，而计算机应用领域已经存在其它范畴语义需要表达，所以研究范畴论本体是此研究领域的一个重要的发展方向，只有范畴论本体才能提供适合各种范畴、对各类问题域一致的表示方法。关于现有本体的局限性，后面结合实例还有更具体的论述。

范畴论方法不仅具备分析和表述方面的优势，作为一种逻辑，它也已经具备了良好的理论基础。范畴论方法有充分的表达能力，一阶理论都可以用 topos 中的结构表示[10]，即由一个一阶理论 T 可以构造一个针对 topos 结构的略图 G，使得 G 略图在 SET 中的模型与理论 T 的模型一一对应。topos 的略图逻辑的合理性和完备性已经由文[11]论证，所以略图逻辑具备了一阶逻辑的所有能力。topos 的略图包含的结构较多，作为对范畴论本体方法的初步尝试，本文用到的略图只涉及对易图式和极限概念，这二者的表达能力已经相当强，这种略图逻辑的合理性和完备性已被文[11,12]解决。

3 范畴论本体的构建方法

3.1 基本概念

定义 3.1.1 图式(diagram) 设 C 和 G 为有向图，图射 $D: C \to G$ 称为 C 在 G 中的图式。在不混淆的情况下，本文将 D 形成的 C 在 G 中的像也称为图式。

定义 3.1.2 对易图式(commutative diagram) 把一个范畴的对象和箭头分别作为结点和有向边会得到的一个有向图 G，称 G 为该范畴的底图，在不混淆的情况下，本文把 G 范畴的底图也用 G 表示。对于范畴 G 和图式 $D: C \to G$，如果 C 的任意两点 i, j 之间的任意两条路径(参阅有向图理论中路径的定义)如图 2，在范畴 G 都有 $Da_n \circ Da_3 \circ Da_2 \circ Da_1 = Db_n \circ Db_3 \circ Db_2 \circ Db_1$，则称 D 是对易图式。

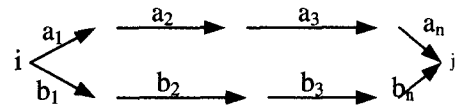

图 2 对易图式

定义 3.1.3 对易锥(commutative cone) 设 D 是范畴 C 的图式，L 是 C 的对象，由 L 以及 L 到 D 的各结点 D_i 的箭头 f_i 构成以 L 为顶点，以 D 为底的锥，如果对于 D 中任意箭头 g，图 3 对易，则称此锥是对易锥，记作$\{f_i: L \to D_i\}$。

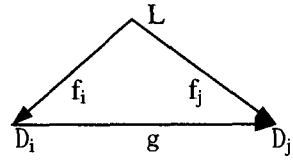

图 3 对易锥中的对易关系

定义 3.1.4 极限(limit) 如果范畴 C 中有一个以图式 D 为底，以 L 为顶点的对易锥$\{f_i: L \to D_i\}$满足如下的条件，即对于任意一个以 D 为底的对易锥$\{g_i: K \to D_i\}$，范畴 C 中有且仅有一个箭头 h 使得对于 D 的任意结点 D_i，图 4 对易，则称对易锥$\{f_i: L \to D_i\}$为 D 的极限。

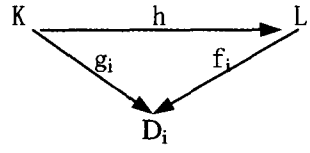

图 4 极限中的对易关系

定义 3.1.5 带极限略图及其模型(finite limit sketch) 带极限略图(简称 FL 略图)S 是一个三元组$\langle G, D, C\rangle$，其中 G 是一个有限有向图(graph)，D 是 G 的图式(diagram)的集合，C 是锥(cone)的集合。FL 略图 S 在范畴 B 中的模型是 S 到 B 的底图的图映射，该图映射将 D 中每个元素映射到 B 的对易图式，将 C 中每个元素映射到 B 中的极限。

FL 略图是数学家 Ehresmann 于上世纪 60 年代末提出的，它是描述数学结构的有效工具。上述内容是关于 FL 略图的基本概念，范畴论本体是用 FL 略图表示的。

3.2 范畴论本体的构建方法

用范畴论的观点分析问题域时，首先确定问题域中有哪

些对象和关系需要表示，其次确定这些对象和关系属于什么范畴或者它们本身构成什么范畴，最后用略图描述这些对象和关系。

例 1 在一次文学作品研讨会上，每个作品的作者都已经到会，会议规定每个作品都必须由其作者作个发言。建立此问题的本体。

此问题域有三类对象：作品、作者和发言。对象之间的关系有：发言→作者，该关系表明每个发言都由一个作者作；发言→作品，该关系表明每个发言都针对一个作品；作品→作者，该关系表明每个作品都有一个作者。上述对象都是集合，关系都是函数，所以该问题域属于 SET 范畴。最后根据问题域的要求，用对易图式和极限指出这些对象和箭头包含的特征结构，该问题域的本体完整表述为下面的略图。

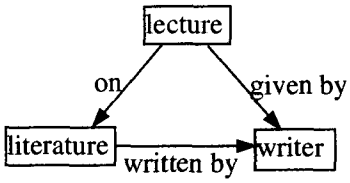

图 5 问题域的基本内容

对易图式集合含一个元素：

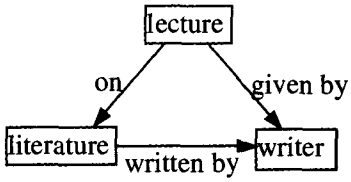

图 6 问题域的对易关系

此对易图式反映了相等关系，即由作者就其作品发言。

锥(cone)的集合含一个元素：

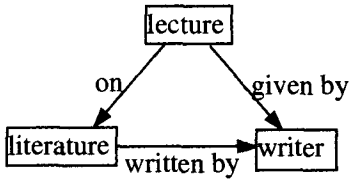

图 7 例一的锥

此锥保证了 on 函数是满射函数和单射函数(根据极限和 SET 范畴的定义，请读者验证)。

同一个图在此略图中多次出现，请读者注意各处的不同意义。

例 2 下面构建数据挖掘中聚类问题的本体。

定义 3.2.1 对于数据库 $D=\{t_1,t_2,\cdots,t_n\}$，不同数据项之间定义了距离 $dis(t_i,t_j)$，现要求将 D 分成不相交的 k 类，对于一种分类方法，每类中数据项之间最大距离称为该类的直径，所有类的直径最大值称为该分类方法的效果，求出所有分类方法中使效果最小的一种。

聚类问题的研究对象是分类方法，对象之间的关系是分类效果的比较关系，这些对象和关系构成一个范畴，求出的理想分类方法是整个范畴的极限。

略图的图(图 8)：f、f_i 和 f_j 表示分类方法，箭头 k 表示“效果小于或等于”，省略号表示省略了其它分类方法。

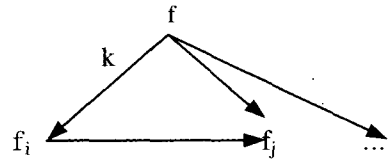

图 8 聚类问题的本体图

略图的对易图式集合有许多元素：

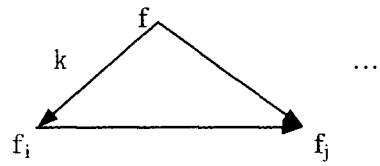

图 9 分类效果的对易关系

这些对易图式表示效果比较关系是传递的，省略号表示还有其它的类似的对易图式。

略图的锥集合有一个元素：

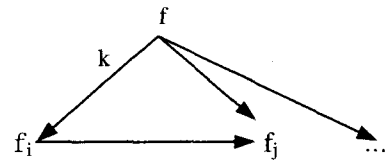

图 10 最佳分类方法体现为极限

此锥表示 f 是效果最佳的分类方法，省略号表示略去了其它分类方法。

上述两个实例说明范畴论分析方法是一致的和普遍适用的，同时体现了极限的概念的重用性，两个属于不同范畴的问题域，却都用极限的概念得到表达，这正是范畴论概念的一般特点。

范畴论本体已经被表示成有向图，它们最终在计算机中的表示属于数据结构的典型问题，本文不深入讨论。为了说明范畴论本体能方便地与现行 Web 融合，附录中给出了例 1 本体的基于 XML 的表示(略)。

4 范畴论本体与现有本体的比较

各种本体语言的优劣通过以下两个方面比较：1)表达能力；2)表达效果，即本体是否简捷直观。这一部分通过一个实例对各类本体方法进行比较，并且从范畴论的角度审视它们各自主要表述了哪些语义。

如果用文[13]的分类方法构建例 1 的本体，由于该方法只能分析和表示分类关系，因此该本体只包含作品、作者和发言三个独立的类，不能再表示其它信息。这种本体只是选取了 SET 范畴中的包含映射作为规范并建立描述系统，所以表达能力较差。

用描述逻辑描述例 1，得到如下本体，它包含作品、作者和发言三个类和三个角色关系(请注意角色是有向关系，given by 角色是由发言指向作者，角色 R^- 表示 R 的反向关系)：

(下转第 117 页)

设初始时刻不同虚路径对于单位带宽所提供的报酬相同，然后根据算法进行求解。其中关于网络中基点评价函数的权值，主要遵循最短路优先的原则，即到达目的地的虚路径中包含的链路数越少，其权值越大。而对于交互行为，主要考虑各链路之间以及各虚路径之间的相互竞争。由于在本例中链路的总带宽供大于求，随着算法叠代过程的进行，各虚路径都不同程度地降低了其所提供的报酬，表现为内部的请求节点不断向外扩张。对于大部分链路，其整体收益能基本维持不便。但是对于个别较特殊的链路，例如 L_5，因为它只能给虚路径 p_1 提供带宽，与其它链路竞争的结果，使得其收益的变化较明显。叠代稳定后的结果为

$P_1=\{0.9,13.3,1.1,7.2,2.5\}$，$P_2=\{5.7,4.3\}$，

$P_3=\{11.2,0.0,0.8\}$，$P4=\{10.4,4.6\}$，$P_5=\{8.0\}$。

结论 本文提出的网络模型及其算法，将优化分配计算机网络资源问题转变为计算多维弹性网络空间的形变过程。分析表明该模型和方法能够刻画多维弹性网络空间中的基类节点之间的复杂的社会交互行为，并能描述基类节点随着局势的变化各自采取的动态策略和自治行为。

参 考 文 献

1 操龙兵，戴汝为. 集智慧之大成的信息系统--Internet [J]. 模式识别与人工智能，2001，14(1)：1～8

2 Gupta A, Stahl D, Whinston A. Priority pricing of integrated services networks, In: Mcknight W, Bailey J. eds. Internet Economics. Cambridge, MA: MIT press, 1997

3 Shenker S, Clark D, Estrin D, et al. Pricing in computer network: Reshaping the research agenda. Computer Communications Review, 1996, 26(2): 19～43

4 Shehory O, Kraus S. Methods for task allocation via agent coalition formation. Artificial Intelligence, 1998, 101(1): 165～200

5 帅典勋，王亮. 一种新的基于复合弹簧网络的多 Agent 系统分布式问题求解方法[J]. 计算机学报，2002，25(8)：853～859

(上接第 3 页)

Concepts:
lecture, literature, writer
Roles:
given by, on, written by
Tbox:
$\exists^{=1}$ given by. writer $\equiv$lecture (表示每个发言都有唯一一个发言人)
$\exists^{=1}$ written by. writer $\equiv$literature
$\exists^{=1}$ on. literature $\equiv$lecture
$\exists^{=1}$ on^{-}. lecture $\equiv$literature(对每个作品都有一个专门发言)

这种本体无法表示元素相等的语义，即无法表示“由其作者针对作品发言”。这类本体重点反映 SET 范畴中的包含映射、角色关系和角色关系所确定的函数(例如一个父亲有几个孩子，此关系确定了一个以父亲为自变量，以自然数为值域的函数)。

如果用多类代数的方法分析和描述该问题域，可表达 SET 范畴中函数和相等关系，所得本体包含三个类型、三个函数和一个相等关系，用 OBJ 语言表示为：

```
theory meeting is
    Sorts lecture literature writer.
    Operator on: lecture → literature.
    Operator givenby: lecture → writer.
    Operator writtenby: literature → writer.
    Var f : lecture.
    Equation givenby (f) = writtenby (on (f)).
End theory
```

这种本体无法表示语义：“每个作品都必须由其作者作个发言”。

范畴论本体完整表述了此问题域语义，而且范畴论本体中各类概念和关系非常直观，所用的范畴论概念很少。相反，现有本体方法都不能完全表示该问题的语义，而且现有本体为了适应某类问题的需要，将 SET 范畴中的关系提取出来加以约定，这样它的表述能力虽然增强，但是也导致了更严重的局限性，即它们都更专属于 SET 范畴。

结论 计算机科学的内容如此丰富，而且还在不断发展，面对这样的需求，现有的本体方法必然是捉襟见肘，例如模糊集和模糊逻辑已经在计算机科学中有广泛应用，与模糊集范畴相关的语义已经无法用现有本体法表达。

略图作为一种形式化方法在文[14]已经有论述，本文从针对不同范畴的重用性的角度进一步澄清了该方法在描述问题域语义方面的优势，这一点对于范畴论在计算机应用领域的推广是很重要的。范畴论的重用结构是对集合论数学中的结构的归纳和抽象，范畴论是对集合论的发展，范畴论本体的优势正是范畴论在数学理论中的这种地位的体现。

本文论述了范畴论本体的必要性、普适性和直观性，建立了范畴论本体的表示方法。

这种本体的机器推理则是下一步研究的内容。图逻辑和传统的逻辑存在着对应关系[11]，它使得图逻辑命题证明的计算复杂性不会超出相应的传统逻辑，并且本体匹配所涉及的命题空间要小于逻辑命题空间，所以在计算性方面范畴论本体有良好的预期。

参 考 文 献

1 Bench-Capon T, Malcolm G, Shave M. Semantics for Interoperability: relating ontologies and schemata, LNCS 2376, 2003

2 Baader F, Horrocks I. Description logics as ontology languages for the semantic web. LNAI, Festschrift in honor of Jorg Siekmann, 2003

3 Barwise J. Handbook of Mathematical Logic. NORTH-HOLLAND PUBLISHING COMPANY, 1977

4 Diskin Z. Formalizing Graphical Schemas for Conceptual Modeling: Sketch-based Logic vs. Heuristic Pictures. http://citeseer.ist.psu.edu, 1995 (Symposium “knowledge Retrieval, use and Storage for Efficiency”)

5 Barr M, Wells C. Toposes, Triples and Theories. www.cwru.edu, 2001

6 Hatcher W S. The Logic Foundations of Mathematics. Pergamon Press, 1982

7 Hillman C. A Categorical Primer. University of Washington, 2001

8 Goguen J. A Categorical Manifesto. Mathematical Structure in Computer Science, 1991, 1

9 Landry E. Category Theory: The Language of Mathematics. Philosophy of Science, 1999, 66

10 Novak V, Perfilieve I. Mathematical Principles of Fuzzy Logic. Kluwer Academic Publishers, 1999

11 Makkai M. Generalized Sketches as a Framework for Completeness Theorems. Journal of Pure and Applied Algebra, 1997

12 Bagchi A, Wells C. Graph-based Logic and Sketches I: The General Framework. www.cwru.edu, 1997

13 Visser U, Stuckenschmidt H. Ontologies for geographic information processing. Compu-ters and Geoscience, 2000

14 Barr M, Wells C. Category Theory for Computing Science The electronic supplement. www.cwru.edu, 1995