第 35 卷 第 6 期 计 算 机 工 程 2009 年 3 月
Vol.35 No.6 Computer Engineering March 2009
·软件技术与数据库· 文章编号：1000—3428(2009)06—0076—03 文献标识码：A 中图分类号：TP311

基于范畴论的本体集成描述

杨先娣[1,2]，何 宁[1,2]，吴黎兵[1]
(1. 武汉大学计算机学院，武汉 430072；2. 武汉大学计算中心，武汉 430072)

摘 要：针对语义 Web 中的本体异构问题，提出一种基于范畴论的本体集成描述方法。与集合论相比，范畴论具有更高的抽象性和更强、更直观的表达力，是本体集成形式化嘎理想工具。把本体结构作为对象，范畴论中的“态射”可描述本体映射，“外推”可描述本体合并，运用图例进行说明并给出本体合并算法。

关键词：范畴论；本体集成；本体映射；本体合并

Ontology Integration Description Based on Category Theory

YANG Xian-di[1,2], HE Ning[1,2], WU Li-bing[1]
(1. Computer School, Wuhan University, Wuhan 430072; 2. Computer Center, Wuhan University, Wuhan 430072)

【Abstract】 In order to solve the problem of ontology heterogeneity in semantic Web, an ontology integration description based on category theory is proposed. Compared with set theory, category theory is more abstract and its representation is stronger, so it is perfect tool to formalize ontology integration. Taken the ontology structure as object, the concept of “morphism” captures the idea of ontology mapping and “pushout” is used to merge ontology. Examples are used to illustrate the problem and the algorithm of ontology merging is presented.

【Key words】 category theory; ontology integration; ontology mapping; ontology merging

1 概述

本体是使 Web 具有语义性的关键技术，在语义 Web 中起着重要的作用：它提供了一套对特定领域知识的共享和共同认识，帮助人们在语法和语义上与机器实现准确的交流；它是对领域的形式化和结构化的描述，是人和机器、程序间知识交流的语义基础。构造本体的目的是知识共享和重用。然而，语义 Web 由多种信息源组成，每个信息源都以自身本体的形式表示，由于时间、地点、目的、知识以及构造者的不同，因此即使对同一问题，本体的构造也有很大差异。这必然造成本体之间的冲突，很难实现真正的共享和重用，为了解决该问题，提出了本体集成技术。本体集成是在 2 个或多个本体之间建立映射关系，据此提供一个集成的、一致的本体，使它们可以互操作。

本体集成的形式化研究有助于本体集成应用系统的开发，是实现系统可重用性，提高系统可比性、可靠性的重要手段。目前的本体描述方法大体分为逻辑方法和代数方法，它们都归于集合论，因为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，本文把以集合论为基础的数学称为集合论数学。传统的基于集合论的本体描述方法有一定的局限性，主要存在 2 个缺点：(1)集合论数学的抽象程度不够，导致本体的重用性降低；(2)对某些问题域语义的表述复杂晦涩，不够直观，甚至无法用传统方法表达。

范畴论作为一门新兴的数学理论，为计算机理论科学提供了一种工具、思维方法和研究手段，与集合论相比，范畴论具有更高的抽象性和更强、更直观的表达力。范畴是群、环、域等抽象数学结构的进一步抽象，其研究重点在于对象之间的关系而非对象的内部结构，因此，比集合论更适合建立较高抽象层次的模型。此外，范畴论的表达方式是一种基于图形的略图语言，这种基于图的语义表达方法更直观易懂。本体集成的主要特性也很适合用范畴论进行形式化研究，主要体现在 3 个方面：(1)本体集成过程中更关注实体之间的联系而非单个实体的内部结构；(2)在表达形式上，本体需要更高的抽象级别，而且重用性好；(3)在进行本体映射时，一些基于图的表达方式更易于被接受。

本文提出一种基于范畴论的本体集成描述方法，定义了本体的范畴，结合图例对本体集成过程中的 2 个主要操作——本体映射和本体合并进行了探讨，用范畴论的相关性质对其形式化并给出了相应算法。这一研究能够较好地指导本体集成系统的开发实践工作，具有较高的理论和实际意义。

2 基本概念

2.1 范畴论

范畴论产生于 20 世纪 40 年代，其特点是观察各种数学对象的普遍特征和相似性，强调各种数学对象之间的联系，而不是孤立地分开研究。这种反映数学各分支共性的理论，即研究各种数学结构之间联系的一般理论就发展成为范畴论。在理论计算机科学研究中，范畴论在函数程序指令、程序语义学和程序逻辑学等领域有着广泛应用，被视为计算机科学中强有力的数学工具。

定义 1 一个范畴(category) C 包括：

(1)一组对象(object)的集合 O；

(2)一组态射(morphism)的集合 M，其中，态射 $f: A\rightarrow B$，$A,B\in O$ 称 A 是 f 的论域(domain)，B 为 f 的余论域(codomain)，记作 $dom(f)=A$, $cod(f)=B$。

作者简介：杨先娣(1974－)，女，讲师、博士研究生，主研方向：语义信息集成，本体集成；何 宁，副教授、博士研究生；吴黎兵，副教授、博士

收稿日期：2008-08-15 **E-mail：**yxiandi@126.com

满足下列条件：

1)复合运算律：若 $A,B,C\in O$；态射 $f: A\to B$, $g: B\to C$，则存在唯一的复合态射 $g\circ f: A\to C$，称为 f 与 g 的复合；

2)结合律：若 $A,B,C,D\in O$，态射 $f: A\to B$, $g: B\to C$, $h: C\to D$，则有

$(h\circ g)\circ f = h\circ(g\circ f)$；

3)单位态射：每一个对象 A，存在一个单位态射 $ID_A: A\to A$，使得对任意的态射 $f: A\to B$，有

$f\circ ID_A = f$, $ID_B\circ f = f$

范畴是基于图的，可以把一个范畴看成一个有向图，以上定义在图 1 中可以更形象地表示，图中的节点表示对象，箭头表示态射，每个箭头有一个源(source)节点和目标(target)节点，分别表示论域和余论域。

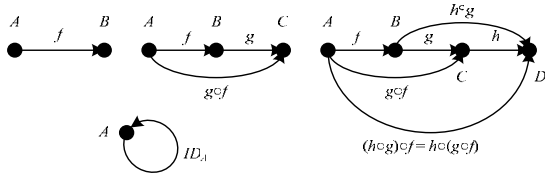

图 1　范畴的有向图表示

2.2　本体集成

随着语义 Web 中信息量的增多，本体的数量也越来越多。由于本体的构造一直没有一个统一的规范和标准，势必造成本体的冗余重复，影响本体间的重用和互操作，导致这些本体所表示的知识间存在冲突。异构的本体之间不能进行互操作，这是本体共享的主要障碍，因此，需要进行本体集成，要实现本体集成首先需要建立本体间的联系，即对本体进行映射，最终实现本体合并。本体映射和本体合并是实现本体集成的 2 个主要过程，本文的研究主要围绕这 2 个过程展开。

下面，对本体的结构及其相关操作进行说明，以便后文利用范畴论对它们进行形式化描述。

定义 2 本体结构 O 可以用一个四元组(C, R, H_C, rel)表示，其中，C 表示概念集合；c 表示概念($c\in C$)，概念是特定领域中的一组或一类实体或者事物，每个概念可以由属性分别描述其不同方面的特点；R 表示关系集合，r 表示关系($r\in R$)，关系描述了概念与概念之间或者属性与属性之间的关系。关系可以分为 2 类：分类关系(taxonomies)和连接关系(associative relationships)；H_C 表示分类关系，是概念与概念之间的父类、子类等上下位的层次关系，$H_C\subseteq C\times C$；rel 表示连接关系，是除了上下位层次关系以外的其他关系，$rel: R\to C\times C$。

定义 3 本体映射指在 2 个本体的实体(概念或关系)之间发现语义对应关系的过程。文献[1]给出了一个形式化的本体映射函数：

$map: O_1\to O_2$

$map(e_{1i}) = e_{2j}$，如果 $sim(e_{1i}, e_{2j})>t$，t 作为阈值，当 e_{1i} 与 e_{2j} 的相似度大于 t 时，便认为它们在语义上是完全相等的，将 e_{1i} 映射到 e_{2j}。

定义 4 本体合并建立在本体映射基础上，是将 $n(n\geq 2)$个相关的本体统一合并成一个新的本体的过程，新本体是这 n 个本体的并集，不仅包括原来 n 个本体的语义相似部分，也包含了语义不同的部分。

3　基于范畴论的本体集成

考虑到本体的实质就是概念以及概念之间的关系所构成的有向图，可以把这种有向图看作一个本体范畴。本体集成的 2 个关键问题——本体映射和本体合并可以由范畴论的相关性质来描述。把本体结构作为对象，范畴论中的“态射”可实现本体映射，同样，可以通过“外推”实现本体合并也是很自然的。

3.1　本体映射

定义 5 本体范畴：把本体结构作为对象，本体映射看成态射，则本体的范畴 Ont 可定义为态射函数$(f, g): O\to O'$，其中，$O= (C, R, H_C, rel)$和 $O'= (C', R', H_C', rel')$是本体结构；$f: C\to C'$和 $g: R\to R'$满足条件：(1)如果$(C_1, C_2)\in H_C$，则$(f(C_1), f(C_2))\in H_C'$；(2) 如果 $(C_1, C_2)\in rel(R)$，则 $(g(C_1), g(C_2))\in rel'(g(R))$。

上述条件(1)的态射保持了概念的层次结构，条件(2)的态射保持了概念之间的关系。下面通过例子来说明利用范畴论进行本体映射的具体情况。

例 1 本体 O_1和 O_2如图 2 所示，$O_1=(\{x_0, x_1, x_2, x_3\}, \{r_1, r_2\}, \{(x_1, x_0), (x_2, x_0), (x_3, x_2)\}, \{(r_1, x_0, x_3), (r_2, x_1, x_2)\})$, $O_2=(\{y_0, y_1, y_2\}, \{s_1, s_2\}, \{(y_1, y_0), (y_2, y_1)\}, \{(s_1, y_0, y_2), (s_2, y_1, y_1)\})$。从态射$(f, g)$: $\{x_0, x_1, x_2, x_3\}\times\{r_1, r_2\}\to\{y_0, y_1, y_2\}\times\{s_1, s_2\}$可得本体映射：$x_0\ \ y_0, x_1\ \ y_1, x_2\ \ y_1, x_3\ \ y_2, r_1\ \ s_1, r_2\ \ s_2$。这些映射保持了本体的层次结构，例如，在本体 O_1 中，x_1 和 x_2 是 x_0 的子概念，相应地在 O_2 中，$f(x_1), f(x_2)= y_1$ 是 $f(x_0)= y_0$ 的子概念。概念之间的关系也是如此，例如，本体 O_1 中的 r_1 (x_0, x_3)与 O_2 中的 $g(r_1)(f(x_0), f(x_3))=(s_1, y_0, y_2)$形成映射。

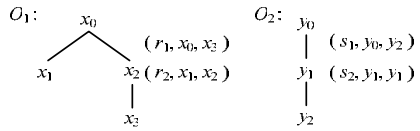

图 2　本体映射

3.2　本体合并

范畴论中外推的作用是形成所谓的融合和(amalgamated sum)[2]，这里，通过范畴论的外推性质实现本体合并是很自然、形象的。

定义 6 本体的外推：如图 3 所示，设有本体 O_1, O_2和 O，本体态射$(f_1, g_1): O\to O_1$和$(f_2, g_2): O\to O_2$，则称 O'及态射$(f_1, g_1)': O_1\to O'$和$(f_2, g_2)': O_2\to O'$是 O_1和 O_2的外推，并满足：(1)$(f_1, g_1)'\circ(f_1, g_1) = (f_2, g_2)'\circ(f_2, g_2)$，对任一其他本体 O''的态射$(f_1, g_1)'': O_1\to O''$和$(f_2, g_2)'': O_2\to O''$，存在唯一的射$(f, g): O'\to O''$使$(f, g)\circ(f_1, g_1)' = (f_1, g_1)''$，$(f, g)\circ(f_2, g_2)' = (f_2, g_2)''$。

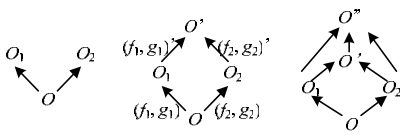

图 3　本体外推

下面，基于外推的概念给出本体合并的算法。设 $O_i= (C_i, R_i, H_{Ci}, rel_i)$ (其中，i=1,2)，$O'= (C', R', H_C', rel')$是 O_1和 O_2的外推(即合并本体)，$O= (C, R, H_C, rel)$是 O_1和 O_2的语义交集。初始条件 $O'=O_e=(\varnothing, \varnothing, \varnothing, \perp)$是一个空本体，$\varnothing$ 表示集合为空，$\perp$表示函数(这里是 f_i' , g_i')在所有的域内无定义。步骤(1)中，对 C 的每一个概念 c 在 C'中相应增添一个新的概念 c'，c'由 $f_i(c)$定义，因此，态射 f_i'被相应地定义，即 $f_i(c)$和新增的概念 c'之间建立了映射，这是由步骤(1)的第 3 行，第 4 行实现的。f_i'与 f_i 的复合 $f_i'\circ f_i$ 保证了 C'中的每一个概念都是

C_1 和 C_2 的交集，这样，通过步骤(1)就将 C_1 和 C_2 中相同的 2 个概念作为 C'中的一个新概念添加到 C'中。接着，步骤(2)和步骤(3)将 C_1 和 C_2 中其余的概念也添加到 C'中。最后，步骤(4)和步骤(5)将层次关系 H_{C1} 和 H_{C2} 增添到 H_C'中。由于态射是保持结构的，f_i'∘ f_i 保证了 H_C'是包含 H_{Ci} 的，它也保证了 C'是概念的外推且集合最小，任何其他能替代 C'的集合均较大。集合 R'和函数 rel'的定义方法同理类推。

外推算法

输入 (f_1, g_1): $O \to O_1$, (f_2, g_2): $O \to O_2$

输出 (f_1', g_1'): $O_1 \to O'$, (f_2', g_2'): $O_2 \to O'$

初始条件：O'=O_ε, $f_i'=\perp$, $g_i'=\perp$

(1) for all c∈C

C':=C'∪c', where c'∉(C_1∪C_2)

f_1':=f_1'∪(f_1(c)　c')

f_2':=f_2'∪(f_2(c)　c')

(2) for all c∈C_1 that is not in the image of f_1

C':=C'∪c

f_1':=f_1'∪(f_1(c)　c)

(3) for all c∈C_2 that is not in the image of f_2

C':=C'∪c

f_2':=f_2'∪(f_2(c)　c)

(4) for all (c, d)∈H_{C1}

$H_{C'}$:=$H_{C'}$∪(f_1'(c), f_1'(d))

(5) for all (c, d)∈H_{C2}

$H_{C'}$:=$H_{C'}$∪(f_2'(c), f_2'(d))

例 2 仍以上例中的本体 O_1 和 O_2 为例，说明怎样从 O_1 和 O_2 中创建一个新的本体 O'来集成它们的概念和关系。先定义一个新本体 O，态射(f_1, g_1): $O \to O_1$ 和(f_2, g_2): $O \to O_2$ 仅包含它们的语义交集。设本体 O=($\{w_0, w_1\}$, $\{t_1\}$, $\{(w_1, w_0)\}$, $\{(t_1, w_0, w_1)\}$)，态射(f_1, g_1): $O \to O_1$ 和(f_2, g_2): $O \to O_2$ 定义为：f_1: w_0　x_0, w_1　x_3, g_1: t_1　r_1, f_2: w_0　y_0, w_1　y_2, g_2: t_1　s_1。由上述算法，先将 O_1 和 O_2 中的相同部分(即 O)加入到 O'中，再逐一添加其余的概念及它们的关系，最终得出外推本体 O'=($\{z_0, z_1, x_1, x_2, y_1\}$, $\{u_1, r_2, s_2\}$, $\{(x_1, z_0), (y_1, z_0), (x_2, z_0), (z_1, y_1), (z_1, x_2)\}$, $\{(u_1, z_0, z_1), (r_2, x_1, x_2), (s_2, y_1, y_1)\}$)即为 O_1 和 O_2 的合并本体，其中，态射(f_1', g_1'): $O_1 \to O'$和(f_2', g_2'): $O_2 \to O'$定义为：f_1': x_0　z_0, x_1　x_1, x_2　x_2, x_3　z_1, g_1': r_1　u_1, r_2　r_2, f_2': y_0　z_0, y_1　y_1, y_2　z_1, g_2': s_1　u_1, s_2　s_2。O_1, O_2, O 和 O'的情况由图 4 来说明，图中的箭头表示它们之间的映射。

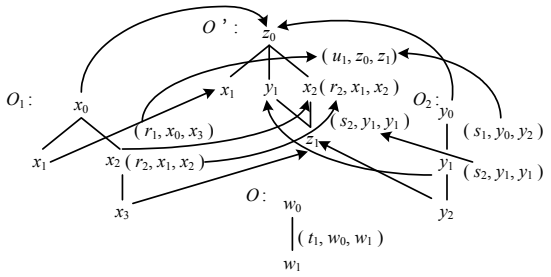

图 4　本体外推示例

4　结束语

现有的本体集成描述大多是基于集合论的，本文从范畴论的角度分析了集合论本体的局限性，对本体集成的范畴论描述方法进行了初步探索。范畴论具有很好的数学理论基础，在其指导下，可在一定程度上实现自动的本体集成，前提是要通过适当的方法找出本体之间的映射关系，目前已经有许多本体映射方法，文献[3]对现有的本体映射方法进行了全面的综述。本体及本体之间的映射构成了范畴，利用范畴论的“态射”实现本体映射，“外推”实现本体合并，保证了结果的正确性。关于范畴论在本体演化、本体排列，尤其是本体推理等应用方面还有许多挑战性的工作尚待解决，需要进一步研究。

参考文献

[1] Ehrig M, Staab S. QOM-Quick Ontology Mapping[C]//Proc. of International Semantic Web Conference. Hiroshima, Japan: Springer, 2004.

[2] Barr M, Wells C. Category Theory for Computing Science[M]. [S. l.]: Prentice Hall, 1990.

[3] Giunchiglia F, Yatskevich M, Shvaiko P. Semantic Matching: Algorithms and Implementation[C]//Proc. of the 10th Conf. on Data Semantics. [S. l.]: Springer, 2007.

编辑　张正兴

(上接第 75 页)

研究热点。管理信息模型映射是异构网络管理体系综合集成的关键问题。针对 SNMP 协议的广泛应用，本文研究了 SNMP 到 WSDM 的管理信息模型映射问题，提出了一种基于层次的信息模型映射框架和机制，设计并实现了 WSDM- SNMP 桥接器，为基于 Web 服务的综合网络管理信息系统的实现奠定了基础。在后续工作中，将研究利用 Web 服务可编排特性提高大规模网络管理系统的自适应能力。

参考文献

[1] Schoenwaelder J, Pras A, Martin-Flatin J P. On the Future of Internet Management Technologies[J]. IEEE Communications Magazine, 2003, 41(10): 90-97.

[2] Oh L J. Enabling Network Management Using Java Technologies[J]. IEEE Communications Magazine, 2000, 38(1): 116-123.

[3] OASIS. Web Services Distributed Management TC[EB/OL]. (2007-10-30). http://www.oasis-open.org/committees/wsdm/.

[4] Sun Microsystems, Ins.. Web Services for Management[EB/OL]. (2007-11-20). http://www.dmtf.org/standards/.

[5] Soldatos J, Alexopoulos D. Web Services-based Network Management: Approaches and the WSNET System[J]. ACM International Journal of Network Management, 2007, 17(1): 33-50.

[6] Goddard T. NETCONF over SOAP[EB/OL]. (2008-05-10). http://www.ietf.org/internet-drafts/draft-ietf-netconf-soap-01.txt.

[7] IBM. Proposal to WSDM to CIM[EB/OL]. (2008-06-10). http://www. ibm.com/developerworks/library/specification/ws-wsdm/.

编辑　索书志