本体集成:概念、过程、工具与方法综述*

王效岳 胡泽文 白如江 李玉平
山东理工大学科技信息研究所 淄博 255049

〔摘要〕针对目前本体集成领域相关概念和集成工具较多、集成过程不一、集成方法介绍过少等问题,简要辨析本体集成的概念,给出本体集成的基本过程。对比分析国内外流行的本体集成工具,对目前本体集成领域新出现的方法,如形式化概念分析法、范畴论法、RDFS 闭包图法等理论基础和实现过程进行详细分析,以便为国内学者在该领域的研究提供启发和指导。

〔关键词〕本体集成 本体映射 本体异构 形式化概念分析 范畴论 RDFS 闭包图
〔分类号〕TP311 G350

Review on Concepts, Processes, Tools and Methods of Ontology Integration

Wang Xiaoyue Hu Zewen Bai Rujiang Li Yuping
Institute of Scientific and Technical Information, Shandong University of Technology, Zibo 255049

〔Abstract〕Ontology integration domain has emerged many related concepts and integration tools. Among these concepts and tools, there isn't a definite definition and differentiation. At present, in ontology integration domain, there isn't a uniform ontology integration process, and the introduce to ontology integration methods is few. So to solve these problems, the authors simply differentiate and analyse concepts of ontology integration, detailedly describe its basic processes, carry out a contrast analysis to prevalent tools of ontology integration at home and abroad, and deeply analyse theory basis and implementation processes of ontology integration methods, such as formal concept analysis, category theory, RDFS graph closure methods, in order to provide some enlightenments and instructions to this domain research.

〔Keywords〕ontology integration ontology mapping ontology heterogeneity formal concept analysis category theory RDFS graph closure

1 引 言

随着新一代语义 Web 的迅速发展,作为语义 Web 基础的本体成为国内外学者、机构的研究热点,出现很多语义丰富的、具有实用价值的本体库,如 SUMO[1]、WordNet[2]、DBpedia[3]、生物医学本体[4]等,这些本体库由于开发者、应用目的和应用领域不同而存在巨大差异,相互之间无法进行有效通信和互操作,极大地阻碍了本体在各领域的广泛应用。本体集成是解决本体异构、推动语义 Web 广泛应用与深度发展、实现服务与数据集成的最有效途径。

国外学者和机构最早提出本体集成的问题,对其研究已有近 10 年的历史,出现了一大批理论成果,并面向实践,设计了一批功能丰富、技术先进、可操作性强的本体集成工具,如 PROMPT[5]、OntoMerge[6]、GLUE[7]、OntoMap[8]、COMA + +[9]。在本体集成理论研究方面,Pinto H S 等人在 *Some Issues on Ontology Integration* 一文中辨析了本体集成过程中涉及的三个概念:integration、merge 和 use[10]。德国卡尔斯鲁厄大学 AIFB 研究所 Stumme G 和 Maedehe A 提出基于形式化概念分析(formal concept analysis, FCA)的本体集成方法[11]。

国内学者和机构对本体集成的研究虽然起步较晚,但紧跟国外本体集成领域的最新研究前沿和动态,对本体集成领域的基本理论和方法进行深入研究和创新,出现了一批具有较高理论价值的成果,如以篇名为"本体集成"在中国知网上进行模糊检索,发现有关本

* 本文系国家社会科学基金项目"海量网络学术文献自动分类研究"(项目编号:10BTQ047)和教育部人文社会科学研究项目"基于本体集成的文本分类关键技术研究"(项目编号:09YJA870019)研究成果之一。
收稿日期:2011 - 01 - 26 修回日期:2011 - 03 - 30 本文起止页码:119 - 125 本文责任编辑:高 丹

体集成方面的研究论文达到147篇，其中核心期刊论文达到60篇，占全部论文的40.82%。国防科技大学信息中心的卢胜军等人在《本体集成相关的基本概念研究》一文中对本体匹配、本体映射、本体联结、本体融合、本体集成以及本体协同等相关概念进行了分析[12]。武汉大学计算中心杨先娣、何宁、吴黎兵提出基于范畴论的本体集成方法[13]。燕山大学信息科学与工程学院张忠平、赵海亮、田淑霞提出基于RDFS的本体集成方法[14]。唯一不足的是，国内学者和机构在本体集成的实践领域研究成果过少，目前国内比较权威的本体集成工具有南京大学瞿裕忠和胡伟等人研发的一款本体匹配工具Falcon-AO[15]和天津大学魏哲雄等人研发的一款本体合并工具的雏形OnMerge[16]。

2 本体集成概念辨析

由于国内外学者对本体集成领域的研究才刚刚兴起，因此在对本体集成概念名称界定方面，出现很多语义相似、词形不同的概念名称，如本体集成(ontology integration)、本体融合或本体合并(ontology merging)、本体调解、本体对齐(ontology alignment)等。国内学者于娟、党延忠对上述不同概念名称进行了综合，提出本体集成的概念[17]，认为本体集成是指，当某一本体任务中应用到多个本体，而这些本体之间存在多方面的不一致(也称为本体异质)时，为了使得这些异质本体能够交互，在这些本体的实体间建立映射、处理映射，以达到本体对齐或者是本体合并的过程。在于娟和党延忠所定义的本体集成概念中，涉及到本体集成过程中的四个相关概念即本体异质、本体映射、本体对齐和本体合并。笔者认为本体异质是进行本体集成的原因，本体映射是不同本体的实体之间语义关系的形式化表示，是实现本体集成的基础性任务，本体对齐和本体合并是本体集成的最终目标。本体对齐和本体合并的涵义虽相似，但并不相同，本体对齐是指根据本体实体之间的语义关系，在需要对齐的本体实体之间建立一个映射集合以达成本体之间的互操作，并不生成新的本体[18]。本体合并是指在原有本体的基础之上根据应用的需要生成一个新的本体。

3 本体集成的基本过程

本体集成的基本过程主要为以下6步：① 本体选择；② 本体预处理；③ 本体概念及其语境抽取；④ 语义相似度计算；⑤ 映射表示；⑥ 集成操作，如图1所示：

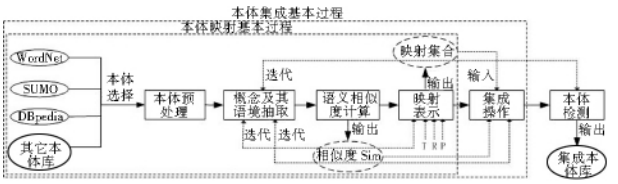

图1 本体集成基本过程[19]

3.1 本体选择

由于目前国内外本体较多、质量不一、语义和内容丰富程度也不相同，因此在具体本体集成项目中，应该根据集成的需求、目的、程度和内容，利用科学的评价方法对相关本体进行选择。本体选择时，需要具有较高的领域知识背景，才能使本体评估选择更加科学合理，因此笔者认为可采用德尔菲法(又称专家调查法)[20]对相关本体进行评价，计算各本体的综合得分，然后选择质量较高、语义和内容较丰富的本体进行集成。

3.2 本体预处理

首先利用当前一些流行的本体编辑工具如Protégé、OntoEdit等[21]并选择一种本体描述语言如OWL将涉及到的本体表述成统一的格式，然后对本体中的词汇进行标准化处理，也就是消除词汇在语义和表示形式上的差异。最后将预处理后的本体导入新构建的本体库NontoLib中，以便本体及其语境的抽取。

3.3 概念及其语境抽取

设定一定规则，利用抽取算法从新构建的本体库NontoLib中抽取出相关本体的概念及其语境信息如概念实例、属性和关系结构，并加上本体库名称标识符，重新组合成一个含有本体库标识符、概念名称、实例、属性和结构等属性的本体数据库，以便本体概念语义相似度的计算。

3.4 语义相似度计算

语义相似度的计算是本体映射的关键。利用国内学者提出的一种综合的语义相似度计算方法[22]，分别计算概念的名称相似度、实例相似度、属性相似度和结构相似度，然后进行加权综合得到最终的概念语义相似度，并设定一定阈值T，以区分相似概念和非相似概念，通常假定本体概念语义相似度大于T，则相似；小于T，不相似。

3.5 映射表示

映射表示方法通常有两种：① 人工映射方法。该

方法通常由映射分析专家如语言学、心理学、自然语言处理等领域专家根据自己的专业知识分析确定概念之间的映射关系。② 自动映射方法。该方法主要依据概念语义相似度 Sim、表示阈值 T、概念关系 R、其他匹配参数 P 等相关参数，对相关本体中的相关概念及其之间的语义关系进行映射表示，映射的表示形式可根据具体工程应用中对映射表示形式的要求进行定义。

3.6 集成操作

依据已确定的本体概念间的映射关系，结合本体集成工具的功能，对本体概念执行添加、修改、删除、合并、声名新规则等类型的集成操作，最终实现本体集成。

3.7 本体检测

选择一种推理机如 Jena[23] 对本体进行一致性检验、并借助领域知识对本体进行比较、验证和修改，同时邀请一些领域专家对本体进行评估，对本体的层次结构和概念间的关系进行校验。

通常来说，在步骤 5(3.5)、步骤 6(3.6)、步骤 7(3.7) 与步骤 3(3.3) 之间会存在一个反馈环，依据用户对部分映射关系的调整情况或本体检测时出现不一致的情况，迭代进行映射表示和集成操作，从而得到更好的、更完整一致的集成结果。

4 本体集成工具

4.1 工具介绍

PROMPT 是美国斯坦福大学医药信息研究小组于 2000 年开发的可用于多本体管理、本体合并和交互的工具包，其可以作为 Protégé[24] 的插件，在 Protégé 本体编辑环境实现它的功能。AnchorPROMPT、PROMPTDiff、IPROMPT 和 PROMPTFactor 是 PROMPT 工具中四个相互关联的组件。AnchorPROMPT 的主要功能是从本体结构图中发现映射；PROMPTDiff 主要从不同版本的本体中发现它们之间的区别；IPROMPT 主要实现基于交互式的本体合并功能；PROMPTFactor 能够从一个本体中抽取一部分。同时 PROMPT 能够识别诸如类、属性和关系等的合并操作及合并操作后所产生的命名冲突，从而实现本体的半自动化维护功能。

OntoMerge 是美国耶鲁大学在 2002 年研发的，可以通过本体合并实现本体翻译。OntoMerge 实现本体合并的基本流程是：首先基于名称空间对两个本体的公理进行合并，然后利用 OntoMerge 中的半自动化“桥公理(Bridging Axioms)”产生工具产生相应桥公理对两个本体中的术语进行连接。为了获得更准确的桥公理，在其产生过程中，需要专家的参与。

GLUE 系统是美国华盛顿大学 AnHai Doan、Jayant Madhavan 等人于 2004 年研发的。其将本体视为一个概念层次结构，基于实例数据，使用机器学习技术和统计方法自动计算本体概念之间的相似度，实现本体映射。GLUE 实现本体对齐的基本流程是：首先利用基于机器学习的分类器来区分本体 O_1 中概念 A 的实例与本体 O_2 中概念 B 的实例是否相同；然后，基于概念 A 和概念 B 的实例集合，利用联合概率分布统计分析方法计算概念 A 和概念 B 的联合概率分布，基于概率生成一个相似度矩阵；最后基于相似度矩阵，使用启发式规则选择最有可能的一致关系对本体 O_1 和 O_2 进行对齐。

OntoMap 是德国卡尔斯鲁厄大学 AIFB 研究所在 2005 研发的一款本体对齐工具，可以作为 OntoStudio[25] 的一个插件来使用。OntoStudio 是一款专业的本体开发工具，在 OntoStudio 中，OntoMap 的主要功能是进行本体映射的创建和管理。OntoMap 支持概念间映射、属性间映射、关系映射及属性与概念间的映射，主要采用可视化的方式对映射进行语义表示，比如两个概念之间的映射用一条边进行连接，支持拖放操作和简单的属性一致性检验。

COMA ++ 是由德国莱比锡大学在 2005 年所研究的一款模式和本体匹配工具，其功能丰富的用户界面为用户提供很多交互功能，用户可以通过很多方式对匹配过程进行干预和指导，并且能够合并不同的匹配算法对一些模式如数据库、表、XML 格式信息等之间及不同本体间的语义对应关系进行识别和匹配。COMA ++ 工具主要由以下 5 个相互关联的组件构成：外部存储库(repository)、模式池(schema pool)、映射池(mapping pools)、匹配定制器(match customizer)和执行引擎(execution engine)。其中 repository 主要用于存储匹配相关的数据；schema pool 和 mapping pools 主要用于管理模式、本体和内存中的映射；match customizer 主要用于配置匹配器和匹配策略；execution engine 主要用于执行匹配操作。

Falcon-AO 是南京大学计算机科学与技术系万维网软件研究组瞿裕忠和胡伟等人开发的一个本体匹配系统，它的体系结构基本上与 COMA ++ 的体系结构类似，主要包括 5 个模块：本体模型池(model pool)，用于解析输入到内存中的本体；匹配结果集(alignment set)，用于产生本体匹配的结果并对匹配结果进行评

估;匹配器库(matcher library) ,用于管理初始匹配器库;中央控制器(central controller) ,用于匹配策略的人工匹配、执行匹配器和合并相似度;外部存储数据库(repository stores) ,用于存储匹配过程中可重复使用的数据。

OnMerge 是天津大学的魏哲雄等人在 2006 年研发的一款本体合并工具。其主要基于编辑距离计算概念与概念之间、属性与属性之间的相似度,形成本体合并的建议,并将合并建议以可视化的方式提供给专家,以备专家根据合并建议对相关本体进行合并。

4.2 工具比较与分析

笔者从本体集成的方式、映射范围、映射方法、自动化程度、应用领域、易用性、支持语言、人机交互方式和可获取性 9 个方面对 4.1 节中的 7 个国内外权威本体集成工具进行分析与比较,希望能够藉此为国内学者和机构在本体集成领域的实践提供一定的指导。具体分析与比较结果如表 1 所示:

表 1 本体集成工具的比较分析结果

工具名称	PROMPT	OntoMerge	GLUE	OntoMap	COMA ++	Falcon-AO	OnMerge
集成方式	对齐、合并	合并	对齐	合并	对齐	对齐	合并
映射范围	概念间映射	术语和公理	概念	概念、属性和关系	本体片段数据库及表	本体片段数据库	概念和属性
映射方法	基于本体结构	基于名称空间	统计学和机器学习	语义匹配	集成匹配算法	各类匹配算法	基于编辑距离
自动化程度	半自动	半自动	自动	半自动	自动	自动	半自动
应用领域	本体集成	本体集成	本体集成	本体集成	服务互操作数据集成	语义 Web 应用之间的互操作	本体集成
易用性	中	低	低	中	中	中	中
支持语言	不限	DAML	不限	OWL, F-Logic	XSD/OWL	RDF(S) /OWL	OWL
人机交互方式	GUI	自然语言	无	GUI	GUI	GUI	GUI
可获取性	易	易	难	易	易	易	难

从表 1 可以得出如下结论:

4.2.1 集成方式 目前本体集成方式主要有对齐和合并两种,大部分集成工具只有一种集成方式。由于本体映射方法大部分不涉及到语言层面,因此,国内学者在开发本体集成工具时可以在借鉴国外本体集成工具体系结构和源码的基础上,加入自己的算法或者直接对国外本体集成工具进行改进和汉化。

4.2.2 映射范围 目前存在的本体集成工具总体映射范围比较广,小到本体概念、属性和公理等,大到本体片段、数据库、表等模式之间映射关系的建立。不过模式之间的映射关系,大部分是一对一的,同类模式之间,很少有一对多、多对多、异构模式之间的映射。

4.2.3 映射方法 大部分本体集成工具提供的映射方法过于单一,不能更高程度地提高本体映射的召回率和准确率。因此在未来的研究中,有必要对各类映射方法进行融合、改进和提高,以便创造性能更优的算法,如 COMA + + 提供的集成匹配算法。

4.2.4 自动化程度 由于本体集成的过程过于复杂,缺乏性能优越的算法,因此目前大部分本体集成工具仍采用半自动化的集成方式,通过人工干预,提高集成效果。因此未来集成算法研究过程中,要充分考虑算法的智能性,以减少人工干预,实现自动化本体集成。

4.2.5 应用领域 目前本体集成工具的应用领域主要分两种:①为解决本体异质和单一本体过小,概念语义过于贫乏而进行的本体集成;② 为解决各类数据、信息、知识和服务等中存在的异构,无法进行通信,阻碍新一代语义 Web 的应用而进行的本体集成。

4.2.6 易用性 由于本体集成的技术门槛及其对专业知识的要求比较高,加上目前本体集成工具的语义交互能力和智能性较低,因此大部分本体集成工具的易用性都在中低层。

4.2.7 支持语言 从表 1 中可以看出,本体集成工具支持的语言类型较多,不够标准化,笔者认为导致该问题的主要原因是目前本体描述语言过多,没有统一的标准。

4.2.8 人机交互方式 为降低本体集成的操作难度,目前大部本体集成工具支持图形用户界面(graphic user interface) 。

4.2.9 可获取性 所调查的 7 种本体集成工具中,除了 GLUE 和 OnMerge 在网上难以获取之外,其他几种都是开源的,网上有源代码和系统原型。

5 本体集成方法

随着本体集成领域研究的拓展和延伸,国内外学者提出一些新的本体集成方法,如基于形式概念分析(FCA) 的本体集成方法、基于范畴论的本体集成方法[26]、基于 RDFS 闭包的本体集成方法。其中,FCA 集成方法主要利用集合理论对本体概念及其之间的语义关系进行描述, 形成形式背景,然后在此基础上进行本体集成。相对传统基于语法和语义匹配的启发式规则方法,该方法能够对集成过程提供一个全局性的结构化描述,更加直观易懂。后两种方法主要将本体看成一个范畴或一个有向图,借助数学理论中范畴论和图论中的技术、方法与原理对不同本体进行图形化表达和集成。这类基于图的本体表达和集成方法抽象程

度较高，本体重用性高，具有更强、更直观的表达力，更易于理解，不过其对语义推理能力的要求较高，集成过程较为复杂。

5.1 基于形式概念分析(FCA)的本体集成方法

FCA 主要用于对象与属性之间关系以及概念泛化与例化关系的形式化描述，能够形成表达概念泛化与例化关系的概念格[27]。它的数学表述是一个形式背景(formal context)，由一个三元组 FC = (G, M, I)，其中 G、M 是非空有限集合，I 是 G 和 M 之间的二元关系，G 为对象集合，M 为属性集合。若(g, m) ∈ I，则称对象 g 中有属性 m。形式概念分析一般分为三个步骤：①形式背景的生成：一个形式背景可以用一个关系表来表示，行表示对象，列表示属性；②概念格的构建：根据形成的形式背景，利用概念格构建算法将形式背景转换成概念格，即将形式背景中的属性转换成概念格中的概念节点；③概念格的转换：将概念格的顶端节点作为根概念，底端节点删除，节点之间的关系转化为概念之间的语义关系，从而将概念格转换成层次清晰的本体概念语义图。基于形式概念分析(FCA)的本体集成方法如图 2 所示：

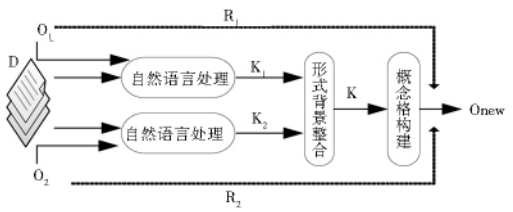

图 2 基于形式概念分析的本体集成方法

基于形式概念分析的本体集成方法分为实例文档集和本体的输入、形式背景的形成、形式背景的融合、概念格构建、目标本体的形成 5 个部分。方法实现的基本流程：① 文档集和本体的输入。输入要集成的两本体 O_1、O_2 和覆盖两本体中所有概念的实例文档集 D。需要注意的是，输入的实例文档集要与本体涵盖领域密切相关，且能够很好地将本体中的概念分隔开。② 形式背景的形成。首先利用自然语言处理方法，从两源本体中抽取相关概念，并判断其在实例文档集 D 中是否存在，如果不存在，需要手工添加；如果存在，需要判断相同文档中的相关概念是否相同，如果不同，需要将它们合并成一个概念。然后分别以本体 O_1 和 O_2 中的所有概念作为属性列，文档集 D 中的实例文档作为对象行，以概念是否在文档集中某个实例文档中存在作为对象与属性之间的二元关系，从而构建两本体的形式背景 K_1 和 K_2。即本体 O_1 中的所有概念在文档集 D 中存在的二元关系表和本体 O_2 中的所有概念在文档集 D 中存在的二元关系表。③ 形式背景的融合。依据表征源本体概念与实例文档之间二元关系的形式背景 K_1 和 K_2，对两本体的形式背景进行融合，即同一实例文档中的相同概念的去重、不同概念的合并，同一概念的不同实例文档之间的融合，形成统一的，能够表征两本体中共同概念、新增概念与实例文档之间二元关系的形式背景 K，K = (D, C, R)，其中，D 为文档对象集合，C 为概念集合。若(d, c) ∈ R，则称文档对象 d 中有概念属性 c。④ 概念格构建。根据形式背景 K，利用 TITANIC 算法[28]将形式背景 K 转换成概念格。⑤ 目标本体的形成。在领域专家参与下，将概念格的顶端节点作为根概念，底端节点删除，并依据源本体中概念之间的语义关系，将节点之间的关系转化为概念之间的语义关系，从而将概念格转换成目标本体的概念语义图，实现本体之间的集成。

5.2 基于范畴论的本体集成方法

范畴是基于图的，一个范畴可以看成一个有向图，包括一组对象(object)集合和一组态射(morphism)集合，对于对象集合中的每对对象 a 和 b，如果两者之间存在态射 $f: a \to b$ or $a \xrightarrow{f} b$，称 a 是 f 的论域(domain)，b 为 f 的余论域(codomain)，记作 dom(f) = a，cod(f) = b。一个范畴满足如下定律[29-30]：①复合运算律：若 a、b、c 属于对象集合，并且存在态射 $f: a \to b$ 和 $g: b \to c$，则存在唯一的复合态射：$g \circ f: a$ 时 c，称为 f 与 g 的复合；②结合律：若 a、b、c、d 属于对象集合，并且存在态射 $f: a \to b$，$g: b \to c$，$h: c \to d$，则有 $(f \circ g) \circ h = f \circ (g \circ h)$；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 $f \circ ID_a = f$，$ID_b \circ f = f$，如图 3 所示(图中的节点表示对象，箭头表示态射，每个箭头有一个源节点和目标节点，分别表示论域和余论域)：

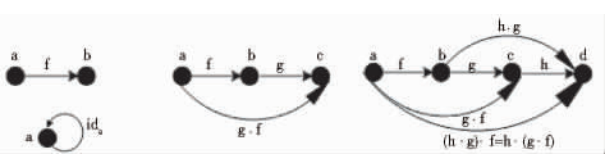

图 3 范畴的有向图表示

基于范畴论的本体集成方法就是将本体看作一种范畴，利用范畴论中的“态射”实现本体映射，利用范畴论中的“外推”方法实现本体合并。本体映射指在两个本体的实体(概念或关系)之间发现语义对应关系的过程。本体合并建立在本体映射基础上，是将 n

(n≥2) 个相关的本体统一合并成一个新的本体的过程。

5.2.1 本体映射的实现 首先定义一个本体结构,目前流行的本体结构是四元组(C,R,Hc,rel) 表示的本体结构。其中,C 表示概念集合; R 表示关系集合; Hc 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; rel 表示连接关系,即除了上下位层次关系以外的其他关系。然后以本体结构为对象,定义本体映射的态射函数(f, g) : O→O',其中,O 和 O'是本体结构,O = (C,R,Hc,rel) 、O' = (C',R',Hc',rel'); f: C→C' 和 g: R→R' 满足条件: ① if(C1,C2) ∈ Hc,(f(C1),f(C2)) ∈ Hc'; ② if(C1,C2) ∈ rel(R),(g(C1),g(C2)) ∈ rel'(g(R))。条件①的态射保持了概念之间的层次结构,即如果本体 O 的概念与本体 O'的概念属于相同的分类关系,则它们之间存在映射关系; 条件②的态射保持了概念之间的关系,如果本体 O 的关系与本体 O'的关系属于相同的连接关系,则它们之间存在映射关系。

5.2.2 本体合并的实现 主要利用范畴论的外推性质实现本体的合并,如图 4 所示:

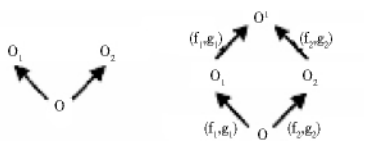

图 4 本体合并

设有本体 O_1、O_2 和 O,本体态射: $(f_1,g_1): O\to O_1$ 和 $(f_2,g_2): O\to O_2$,其中 O = (C,R,Hc,re1) 是 O_1 和 O_2 的语义交集。首先将 O_1 和 O_2 的共同部分(即 O) 加入到新本体 O',形成 O_1 和 O_2 的外推本体 O'。O_1 和 O_2 与外推本体 O'之间的本体态射: $(f_1,g_1)': O_1\to O'$ 和 $(f_2,g_2)': O_2\to O'$,态射满足条件: $(f_1,g_1)'\circ(f_1,g_1) = (f_2,g_2)'\circ(f_2,g_2)$。根据该条件,将 O_1 和 O_2 中不同部分的概念及关系逐一添加到 O'中,形成最终的外推本体 O',即 O_1 和 O_2 的合并本体。

5.3 基于 RDFS 的本体集成方法

RDF(resource description framework) 是一种描述 Web 资源及其之间关系的数据模型,不足的是该模型不能描述资源类及其之间的层次、继承的关系。RDFS (resource description framework schema) 是对 RDF 的一种补充。RDFS 定义了类和性质,这些类和性质可以用来描述其他的类和性质,从而增强了 RDF 对资源的描述能力。并且 RDFS 提供了一些建模原语,用来定义一个描述类及其之间关系的简单模型,由资源(resource) 、属性(property) 和声明(statement) 三部分构成,其中,statements 由资源及其属性、属性值构成。RDFS 本体可用一种有向图来表示,有向图的节点是资源及其属性,边表示资源与其属性之间的语义关系,有向边的箭头由资源指向它的属性[31]。

基于 RDFS 图闭包本体集成方法的基本过程: ①利用 Jena 中的 ARP(another RDF parser) 工具将不同描述语言(如 OWL、XOL 等) 的本体转换成易于推理和检索的 RDFS 三元组形式。②利用 RDFS 本体图闭包生成算法[14]生成 RDFS 图闭包。RDFS 闭包中包含所有显示和隐含的声明,因此基于 RDFS 图闭包的本体集成,可以保留更多的领域知识。③ 生成图闭包后,利用文献[14]提出的综合相似度算法计算本体间实体的相似度。④ 在相似度计算的基础上,对源本体实体间的映射关系如类同义、类包含、关系等价、关系蕴含等进行推理,形成源本体的集成 RDFS 有向图。⑤对集成 RDFS 有向图进行剪枝,即删除源本体中没有用到的类、实例、关系等,形成最终的集成 RDFS 本体。

6 结 语

随着新一代语义 Web 的兴起,作为语义网基础的本体成为国内外学者研究的热点,然而本体之间的异构问题极大地阻碍了语义 Web 应用的深度和广度,本体集成是消除这一障碍的最有效途径。然而目前国内学者对本体集成工具和方法的研究还不够深入和具体,构建的本体集成工具过少。在对本体集成方法的研究方面,仅停留在概念、系统、框架和模型方面,没有深入到实现层面。对目前新出现的一些本体集成方法如形式概念分析法、范畴论法、RDFS 闭包图法等研究较少,这些方法具有非常好的数学理论基础和可视化能力,能够在一定程度上提高本体集成的自动化程度和集成过程的可视化程度,是今后国内外学者研究的热点领域,同时也是一个难点。

参考文献:

[1] Suggested Upper Merged Ontology (SUMO). [2010-05-10]. http://www.ontologyportal.org/.

[2] About WordNet. [2010-05-10]. http://wordnet.princeton.edu/.

[3] DBpedia. [2010-05-10]. http://blog.dbpedia.org/.

[4] 吴正荆,黄薇,牟冬梅,等. 生物医学领域本体开发项目比较研究. 中华医学图书情报杂志,2010,19(5):16-19.

[5] Prompt. [2010-06-01]. http://protege.stanford.edu/plugins/

prompt/prompt. html.
[6] OntoMerge. [2010 - 06 - 01]. http://cs-www. cs. yale. edu/homes/dvm/daml/ ontology-translation. html.
[7] Doan A, Madhavan J, Domingos P, et a1. Ontology matching: A machine learning approach//Staab S, Studer R. Handbook on Ontologies in Information Systems. Berlin, Germany: Springer-Verlag, 2004: 397 - 416.
[8] Kiryakov A, Kiril lv. S, Dimitrov M. OntoMap: Portal for upper-level ontologies//Proceedings of Formal Ontology in Information Systems: Collected Papers from the Second International Conference. Ogunquit: ACM, 2001: 47 - 58.
[9] COMA + +. [2010 - 06 - 03]. http://dbs. uni-leipzig. de/Research/coma. html.
[10] Pinto H S, Gomez-Perez J P. Martins. Some issues on ontology integration. [2010 - 06 - 03]. http://citeseerx. ist. psu. edu/viewdoc/download? doi = 10. 1. 1. 67. 7066&rep = rep1&type = pdf.
[11] Stumme G, Maedehe A. FCA-Merge: Bottom-up merging of ontologies//Proceedings of the Seventeenth International Conference on Artificial Intelligence (IJCA101). Seattle: Morgan Kaufmann, 2001: 225 - 230.
[12] 卢胜军, 真溱. 本体集成相关的基本概念研究. 情报理论与实践, 2008, 31(3): 443 - 446.
[13] 杨先娣, 何宁, 吴黎兵. 基于范畴论的本体集成描述. 计算机工程, 2009, 35(6): 76 - 78.
[14] 张忠平, 赵海亮, 田淑霞. 基于 RDFS 的本体集成方法. 计算机工程与应用, 2008, 44(15): 131 - 135.
[15] Hu Wei, Qu Yuzhong. Falcon-AO: A practical ontology matching system. Journal of Web Semantics, 2008, 6(3): 237 - 239.
[16] 魏哲雄, 冯志勇. 基于字典技术的本体整合系统. 计算机应用, 2007, 27(2): 428 - 430.
[17] 于娟, 党延忠. 本体集成研究综述. 计算机科学, 2008, 35(7): 9 - 13.
[18] Euzenat J, Bach T L, Barrasa J, et a1. D2. 2. 3: State of the art on ontology alignment. [2010 - 06 - 10] http://starlab. vub. ac. be/research/projects/knowledgeweb/kweb-223. pdf.
[19] 卢胜军, 李法勇, 钱建军, 等. WCONS +: 一种基于 WCONS 的本体集成方法. 现代图书情报技术, 2009(2): 18 - 22.
[20] 林晓华. 运用德尔菲法建立高校文献招标评价体系的研究. 图书与情报, 2010(2): 111 - 115.
[21] 徐国虎. 本体构建工具的分析与比较. 图书情报工作, 2006, 50(1): 44 - 48.
[22] 张忠平, 田淑霞, 刘洪强. 一种综合的本体相似度计算方法. 计算机科学, 2008, 35(12): 142 - 145.
[23] Jena-A Semantic Web Framework for Java. [2010 - 07 - 01]. http://jena. sourceforge. net/.
[24] Welcome to protégé. [2010 - 07 - 02]. http://protege. stanford. edu/.
[25] OntoStudio. [2010 - 07 - 03]. http://semanticweb. org/wiki/OntoStudio.
[26] Hitzler P, krotzsch M, Ehrig M, et al. What is ontology merging-a category-theoretical perspective using pushouts//Proceedings of 2005 AAAI Workshop. Pittsburgh, United States: American Association for Artificial Intelligence, 2005: 104 - 107.
[27] Salton G. Introduction to modern information retrieval. New York: McGraw Hill Book Co., 1983: 1 - 40.
[28] Stumme G, Taouil R, Bastide Y, et al. Fast computation of concept lattices using data mining techniques//Proceedings of 7th Intl. Workshop on Knowledge Representation Meets Databases. Berlin: Springer-Verlag, 2000: 21 - 22.
[29] Healy M J, Olinger R D, Young R J. Applying category theory to improve the performance of a neural architecture. Neurocomputing, 2009, 72(13): 3158 - 3173.
[30] Barr M, Wells C. Category Theory for Computing Science. Upper Saddle River: Prentice Hall, 1990.
[31] Hayes J. A graph model for RDF. [2010 - 07 - 05]. http://www. dcc. uchile. cl/cgutierr/papers/rdfgraphmodel. pdf.

〔作者简介〕王效岳, 男, 1961 年生, 教授, 馆长, 发表论文 70 余篇; 胡泽文, 男, 1985 年生, 硕士研究生, 发表论文 8 篇; 白如江, 男, 1979 年生, 馆员, 发表论文 18 篇; 李玉平, 女, 1984 年生, 硕士研究生, 发表论文 6 篇。

下 期 要 目