基于多Agent的大规模中文领域本体的自动化构建方法与应用研究

支丽平 ◎ 著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

河南省高等学校重点科研项目计划“基于多Agent的大规模中文领域本体的自动化构建方法与应用研究”资助出版

基于多Agent的大规模中文领域本体的自动化构建方法与应用研究

支丽平 著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS
·北京·

图书在版编目（CIP）数据

基于多Agent的大规模中文领域本体的自动化构建方法与应用研究 / 支丽平著. —北京：科学技术文献出版社，2017. 11
ISBN 978-7-5189-3543-7

Ⅰ.①基… Ⅱ.①支… Ⅲ.①中文—知识本体—检索系统—研究 Ⅳ.①G254.92

中国版本图书馆 CIP 数据核字（2017）第 267798 号

基于多Agent的大规模中文领域本体的自动化构建方法与应用研究

策划编辑：周国臻　　责任编辑：李　鑫　　责任校对：文　浩　　责任出版：张志平

出 版 者　科学技术文献出版社
地　　址　北京市复兴路15号　邮编 100038
编 务 部　(010) 58882938，58882087（传真）
发 行 部　(010) 58882868，58882874（传真）
邮 购 部　(010) 58882873
官方网址　www.stdp.com.cn
发 行 者　科学技术文献出版社发行　全国各地新华书店经销
印 刷 者　虎彩印艺股份有限公司
版　　次　2017 年 11 月第 1 版　2017 年 11 月第 1 次印刷
开　　本　710×1000　1/16
字　　数　200千
印　　张　12
书　　号　ISBN 978-7-5189-3543-7
定　　价　52.00元

前　言

本书是笔者在安阳师范学院承担的河南省高等学校重点科研项目计划“基于多 Agent 的大规模中文领域本体的自动化构建方法与应用研究(16A520040)”的支持下，对近年来在本体学习研究方面取得的一些成果进行的梳理和总结。

本体是一种从语义层次上对概念及相关关系的规范化说明，广泛应用于自然语言处理、机器翻译、数字图书馆等众多领域。随着领域本体逐渐进入商业应用，领域本体的规模越来越大，很难由个人或者小规模团队构建，这需要大量的领域专家、工程技术人员和项目管理人员，耗费很长时间才能协作完成。考虑到人工构建费时、费力，那么如何减少人员的参与，并采用自动化方式来协作构建领域本体，将是一个极有意义的研究。因此，大规模领域本体的自动化构建是当前学术研究的热点，但是针对中文大规模领域本体的构建，目前仍没有有效的方法。

近年来，随着分布式人工智能的提出，基于多智能体的系统在各个领域得到了越来越广泛的重视。多 Agent 技术适合于具有高度局部性和分布式处理特征的研究领域，这也为大规模领域本体的自动化构建研究提供了新的研究思路。

本研究将多 Agent 理论引入到大规模中文领域本体的构建中，旨在为自动化构建领域本体提供一种新的方法论，该方法采用多 Agent 技术自动进行知识源获取、本体创建、本体测试、本体融合、映射失效的检测与修正等本体构建工作。本书共分七个部分：第一章主要阐述研究背景和意义，设计研究思路方法；第二章从文献角度阐述本体相关理论研究现状与关键技术；第三章回顾了 Agent、多 Agent 系统的研究进展与重要理论，对其分布式特征进行了阐述，为多 Agent 建模研究提供理论基础；第四章通过分析大规模领域本体的构建过程，提出一种基于多

Agent 的自动化构建的方法，继而给出多 Agent 系统的基于 FIPA 的体系结构、消息传递机制、存储方式、各类 Agent 的具体任务及部分算法；第五章研究了主要 Agent 的关键算法及部分代码；第六章选择专利领域进行实证研究，进行基于专利本体的语义检索系统和语义推理系统的研究；第七章为全书总结与展望，总结了本书开展的主要工作及取得的主要研究成果，指出了进一步研究的方向。

本书的创新之处主要体现在以下 5 个方面：

（1）将多 Agent 理论引入到大规模中文领域本体的构建中，为自动化构建本体提供一种新的方法论。采用多 Agent 技术自动进行知识源获取、本体创建、本体测试、本体融合、映射失效的检测与修正等领域本体构建工作，研究了该多 Agent 系统的体系结构、消息传递机制、关系存储方式、各类 Agent 的具体任务及关键算法。因每个 Agent 的运作受限于局部和不完整的信息（如局部目标、局部规划等），所以很难实现全局一致的行为。为了克服多 Agent 系统的缺点（即不一致性），本文在设计体系结构时增加了项目管理 Agent 和调解 Agent，进行各类 Agent 之间的任务分配、协商和冲突消解等工作，以实现全局性的协同目标。

（2）针对中文语料库的特点，对现有的各类抽取算法进行了改进，给出了十类主要 Agent 的算法设计思路与部分实现代码，并设计实验验证这些算法有很好的性能。包括术语抽取 Agent、术语过滤 Agent、概念抽取 Agent、实例抽取 Agent、分类关系抽取 Agent、非分类关系抽取 Agent、关系修剪 Agent、集成 Agent、编辑 Agent 等主要 Agent。从实验结果来看，本文提出的抽取算法有较高的查全率和查准率。

（3）本文研究设计了一种应用于关系数据库和本体之间语义映射的维护方法。由于关系数据库模式和本体都在不断演化，可能造成语义映射的失效。本文对语义映射的维护方法进行了研究，进而设计了检测 Agent 和修正 Agent，其中检测 Agent 又可以进一步分为模式检测 Agent 和本体检测 Agent 两类，分别用于关系数据库模式演化和本体演化条件下的变化检测，假如检测出来失效的语义映射，则将该模式映射和已有的语义映射进行适当组合就能够得到新的语义映射，即通过映射组合的方法实现对失效映射的维护。实验结果表明，本方法能够较好地检测出

失效的语义映射，并能够自动修正。

（4）针对本研究提出的构建大规模领域本体的新方法，本文将其应用到专利领域本体的自动化构建中，并编程实现了基于该专利本体的语义检索系统，最后设计实验从实例、属性和关系3个层次来检验该专利本体的应用效果。相比传统的基于关键词的单一检索方式，能有效提高领域信息的查全率和查准率。

（5）将专利本体和描述逻辑（DL）结合起来对语义推理系统进行了研究，并设计了专利领域规则库。实验结果表明，不仅可以检索出显性知识，而且可以检索出隐性知识。另外，本文提出的自动化构建领域本体的方法，不但理念先进，而且具有很强的可操作性，能极大地提高构建效率，这为大规模本体进入商业应用领域提供了有效的技术手段，具有良好的应用前景。

本书能够顺利完成，首先感谢王恒山教授，他严谨的科学作风、渊博的学识、敏锐的科研洞察力，以及谦逊和蔼、淡泊名利的处世态度都给我留下了深刻的印象，他虚怀若谷、遇事泰然处之的生活态度也对我产生了极大的影响，尤其是王老师“做事先做人”的君子风范将使我终身受益。感谢多年来在领域本体构建研究方面合作的博士同窗和同事；我还要感谢我的家人，感谢他们多年辛苦无怨的付出，解决我的后顾之忧，让我安心研究；感谢丈夫对我的悉心照顾，在研究做得最不顺心的时候，是爱人开导我，给我鼓励并让我重新获得了信心；感谢本书的编辑科学技术文献出版社周国臻、李鑫老师，正是他们的积极鼓励和协调才促成了本书的出版。

我期望本书的出版能为本体学习研究方向做出应有的贡献。由于笔者水平有限，书中难免有不足之处，恳请读者批评指正。

支丽平

2017年10月

目　录

第一章 绪 论

1.1 选题背景

本体（Ontology）是近年来计算机及相关领域普遍关注的一个研究热点，作为一种能在语义和知识层次上描述信息系统的概念模型建模工具，已被广泛应用于知识工程、系统建模、信息处理、数字图书馆、自然语言理解、语义 Web 等领域。

虽然 20 世纪 90 年代以来，研究人员从各自的专业角度出发对本体的理论和应用进行了深入研究，取得了丰富的研究成果，本体理论与技术也随之日趋成熟，但是随着本体逐渐进入商业应用领域，本体的规模越来越大，本体的复杂度也越来越高。例如，美国国立癌症研究所（NCI）的知识库覆盖近 8 万个概念，仅凭个人或者小团队很难高效地开发。再如，根据国际 PCI 分类法，专利可以分运输、化学、冶金、纺织、固定建筑物、机械工程、物理、电学 8 大类，30 个小类，涉及计算机服务业、食品制造业、烟草制品业、医药业、生物化工、海洋渔业等 22 个行业领域，其本体库覆盖概念多达近 60 万个。可见，大规模领域本体的构建将是一项浩大的工程，这需要大量的领域专家、工程技术人员、项目管理人员的参与，来并行、协作开发。

目前已有一些本体开发工具支持部分协同开发本体的功能。例如，斯坦福大学 Tania 等人于 2008 年在 ISWC 国际会议上提出了 Protégé Web 版，然而，它仍然是以手工方式来协作开发本体，并且只能处理英文，不支持中文。Onto Wiki 是一个基于 Web 的手工本体编辑工具，优点是可以为实例数据提供不同的视图。例如，为地理数据提供地图视图，为含有日期的数据提供日历视图。然而，Onto Wiki 集中于实例的获取，

对本体编辑仅仅提供入门级的功能，尚不能满足开发复杂功能的大规模本体的需要。Hozo 本体编辑器将本体划分为多重互联模块，从而允许异步地手工开发本体。当开发者检测并且锁定一个特定的模块，可以对该模块进行本地编辑，然后检查并将其放回去。然而，如果该本体不是模块化的，那么开发人员必须把整个本体锁住以防止他人对该本体进行编辑。所以该方法在许多情况下并不实用。另外，还有一些基于维基的协同本体编辑工具。例如，Biomed GT，Cicero 和 Coefficient Makna 等工具。这些基于维基的工具具有一个简单的界面，是最适合对本体进行简单的改变。通常特别适用开发某种特定的编辑工作流，而且使采用该工作流的项目工作得很好。维基为讨论、交流信息提供了一个天然的论坛，这些建议很容易被存档。然而，这些工具本身并不能满足本体编辑的需求，因为它们只遵循预先设计的特定工作流而非其他的工作流，而且不提供结构化的存取 - 控制机制。

另外，还有一些由德国卡尔斯鲁厄大学 AIFB 研究所等知名机构提出了用于领域本体（半）自动化构建的方法，如 OntoLearn 方法、Kietz 方法、Doan 方法、DODDLE 和 SEISD 等。其中，OntoLearn 方法、DODDLE 方法和 SEISD 方法都是基于 WordNet 词典，考虑到 WordNet 是基于英文的词典，所以此类方法不适用于本著作中文领域本体的自动化构建；而 Kietz 方法和 Doan 方法则需要用户参与构建领域本体，所以它们是一种半自动的构建方法，也不适用于本研究。在领域本体的（半）自动构建工具方面，国外的研究也相对成熟，如 Text2Onto、Hasti、OntoLearn、OntoBuilde 和 OntoLiFT 等工具。不过它们都仅能支持基于某种特定类型源数据，如 Hasti 和 OntoLearn 仅支持纯文本，而且大部分工具仅支持西文（主要是英语），并且要求领域专家的参与，所以不能适用于中文领域本体的自动化构建研究。

国内方面，周耀等人提出了一种开发大规模本体的架构，主要是针对并发系统的协同问题进行了讨论，给出了一个理论框架，并没有形成一个可以构建本体的实用系统；国家标准化研究院国家标准馆的李景研究员提出了采用类似软件工程方法的多人协作开发知识本体的思路，开发了 LODE 原型系统，尝试从大量文档中自动提取知识与操作人员手工

修改与审查相结合建立大规模知识本体的方法，但是该方法仍然需要大量的技术人员和管理人员参与；浙江大学刘柏嵩博士开发的 GOLF 系统是一种支持中英文的本体学习系统，并且提出了大量抽取算法，然而由于其设计的算法多为基于模式匹配的算法，即需要大量专家进行人工阅读语料库模式规则的提取，所以也是一种半自动的构建工具。

可见，现有的支持协同开发的本体工具仍然主要以手工方式进行开发，已有的一些支持半自动化构建工具也仍需要领域专家的参与，而且绝大多数工具都仅支持西文，目前缺少面向中文领域本体的自动化构建方法和实用工具。对于已有的大型本体知识库，如 WordNet、HowNet，是需要大量的领域专家、工程技术人员和项目管理人员，耗费很长时间才能协作完成。大规模本体自动化构建方法和工具的缺乏，已经成为本体商用化的一个急需解决的“瓶颈”问题。那么如何减少人员的参与，并采用自动化方式来协作构建中文领域本体，这将是一个极有意义的研究。

1.2 研究的目的、意义及方法

1.2.1 研究目的

现有的支持多人协作构建大规模领域本体的方法需要依赖大量的领域专家、工程技术人员和管理人员，存在构建代价过高、开发时间过长的难题，严重限制了大规模领域本体的构建。

为解决这一难题，本著作旨在研究一种通用的、自动化的大规模领域本体构建的新方法；通过对领域本体的（半）自动化构建和多 Agent 理论的现有文献进行研究，提出一种基于多 Agent 的大规模领域本体的自动化构建方法，探讨采用多 Agent 技术自动进行知识源获取、本体创建、本体测试、本体融合、映射失效的检测与修正等工作的关键内容，最后选择专利领域进行实证研究（语义检索系统和语义推理系统），这给领域本体的自动化构建提供了可行的新方法论。

1.2.2 研究意义

本体已经成为人工智能、知识工程和图书情报等领域的研究工具，

在知识的获取、表示、分析和应用等方面都具有重要的应用。目前，本研究（自动化构建本体）对语义 Web、知识管理、基于语义的数字图书馆、企业建模、智能信息检索、机器翻译等多个领域都有极为重要的研究意义。

（1）语义 Web

语义 Web（Semantic Web）被称为第三代互联网，由 Tim Berners-Lee 于 2000 年 12 月 18 日在 XML 国际会议上正式提出；语义 Web 以实现 Web 中的信息可以被计算机处理并理解为目标，以实现不同计算机之间数据语义互操为特点，是一个由机器可理解的大量数据所构成的一个分布式体系结构。

2001 年，Tim Berners-Lee 给出了语义网结构图（图 1-1），其中，本体层位于整个语义网结构图的中心位置，用于定义各类概念及概念之间的关系，在语义 Web 的交流和通信中，本体担当着语义沟通的重要角色，是语义 Web 的关键技术。

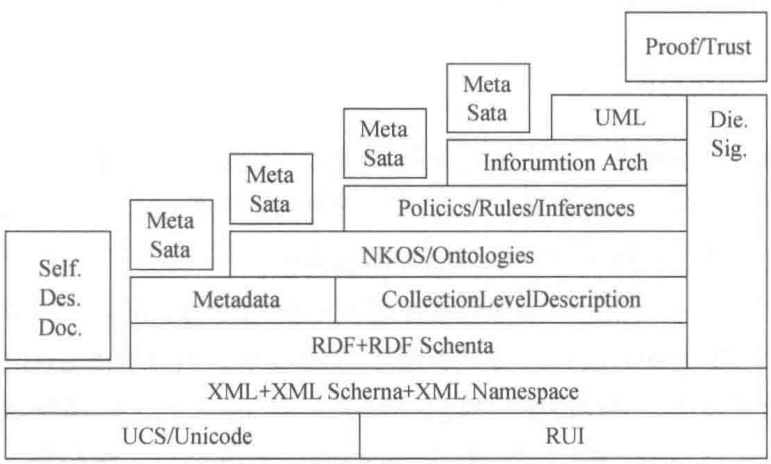

图 1-1 Tim Berners-Lee 的语义网结构图

（2）知识管理

知识管理是指一个组织整体上对知识的获取、存储、学习、共享、创新的管理过程，目的是提高组织中知识工作者的生产力，提高组织的应变能力和反应速度，创新商业模式，增强核心竞争力。全球最大的知

识管理网站创始人王德禄，在《知识管理的 IT 实现——朴素的知识管理》一书中给出知识管理的架构图（图 1-2）。

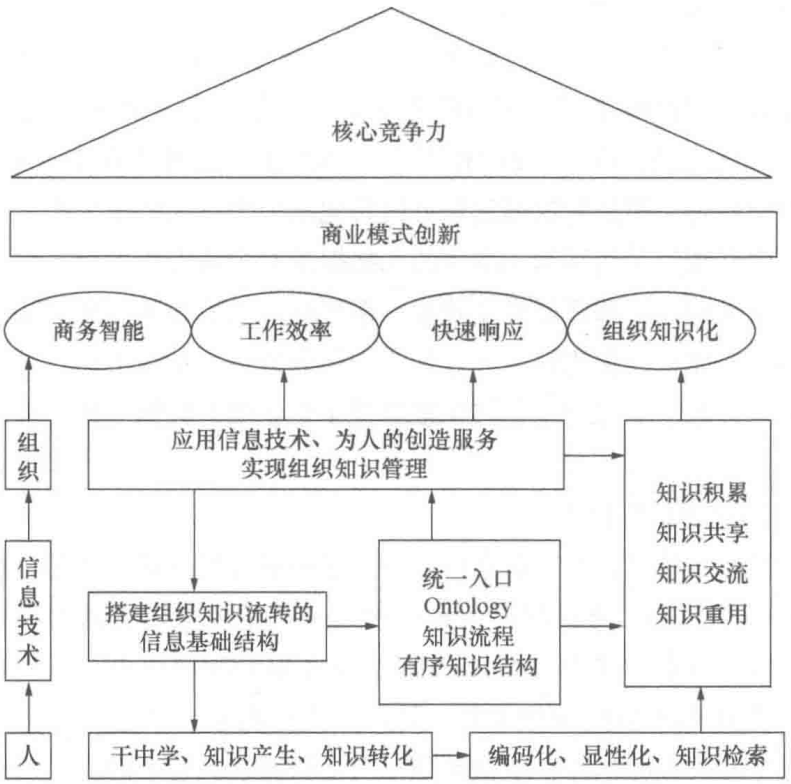

图 1-2 知识管理的架构

如图 1-2 所示，知识管理的主体（人）利用知识管理的工具（IT 技术）完成了知识的生产和转化，经过编码、显性化的处理，将隐性知识转化为显性知识，从而成为组织的知识。可见，在组织的知识管理中，Ontology（本体）对组织的知识进行分类和编码，是将隐性知识有序化处理的工具；显然，本体在隐性知识向显性知识的转化过程中，起到了关键作用。

（3）基于语义的数字图书馆

随着 Web 上多媒体数据的日益增加，对它们的管理和检索也变得越来越重要。传统的多媒体检索技术使用颜色、纹理和形状等特征来描述视频或图像；基于语义的数字图书馆则需要使用本体来描述各种多媒

体信息和图书信息，从而支持基于语义的检索和导航。可见，本体的构建对于基于语义的数字图书馆至关重要。

（4）企业间的数据交换

企业间的数据交换一直是基于 Web 的电子商务和 ERP 系统的重要组成部分，现有很多项目都围绕着企业间数据交换而展开，这些项目（如 OntoWeb）的目标是激励和支持本体技术从学术界向工业界转化，它们假设各企业提供的数据信息可以转化成一个巨大的知识库。这种转化的重要基础就是利用基于本体的元数据来对企业发布的信息或内部文档进行语义标注，那么就需要开发一系列相关技术和工具来支持，如标注工具、本体自动化构建工具、基于本体的推理工具等工具。

可见，基于本体的企业知识建模对于企业间的数据交换起着关键的作用。

（5）智能信息检索

面对海量信息，智能信息检索一直是科研人员的重要课题。然而传统的 Web 信息表示方法使得信息检索面临了很多难以逾越的障碍，因此改进信息检索的重要方法之一就是整理和重新规范 Web 上的信息，即从传统的 Web 页面（如 HTTP 网页）提取出语义信息，并构建出能够描述这些页面的本体；也就是赋予网络资源及其各个内容元素以相应的语义标注，然后利用本体、其他元数据和网络资源中的语义信息进行智能检索和推理。

如果手工实现这一过程需要耗费大量的人力和时间，所以本体的自动化构建是实现智能信息检索的关键因素。

（6）机器翻译

高质量的机器翻译（Machine Translation，MT）系统需要结合语言学知识及语言中的普遍知识（即常识）。本体是客观世界的概念模型，由概念及概念间丰富的语义关系构成。通过把源语言中的词汇映射到本体中的概念，以可以支持在源语言分析时进行歧义消解和目标语言生成时的词汇为选择，同时，本体也可以作为源语言和目标语言之间中介表示的概念来源。王小捷等给出了基于本体的机器翻译体系结构（图 1-3），可以看出，Ontology（本体）作为英语（源语言）和汉语