文章编号:1007-5321(2006)增-0065-05

领域 Ontology 自动构建研究

刘 耀, 穗志方, 胡永伟, 冀铁亮
(北京大学 计算语言学研究所, 北京 100871)

摘要: 利用自然语言处理(NLP)理论和技术方法对已有公认领域知识,如专业叙词表、专业辞典、专业教材或权威著作等进行重构利用;借助领域专家知识,实现了基于网络的知识采集与加工;建立起受限文本的 Ontology 自学习机制, 从而实现领域 Ontology 概念描述体系的自动构建. 最终有效地解决了 Ontology 的自动构建这一瓶颈问题,成功地探索出了一种较为理想、实用的理论与方法,为专业领域 Ontology 的自动构建提供了理论依据及技术支持.

关 键 词: 领域本体; 自然语言处理; 叙词表; 知识工程
中图分类号: TP929.53 **文献标识码**: A

Automatic Construction on of Domain Ontology

LIU Yao, SUI Zhi-fang, HU Yong-wei, JI Tie-liang
(Institute of Computational Linguistics, Peking University, Beijing 100871, China)

Abstract: By employing natural language processing (NLP) theory and technology, the recognized Ontology knowledge as professional thesauruses, professional dictionaries and textbooks are reconstructed. Knowledge collection and manipulation are realized. Based on constraint texts, a learning system is build to fulfill automatic construction of domain Ontology's concept description with the help of domain experts. Also a more ideal and pragmatic method is provided for resolving problems of automatic construction of Ontology effectively.

Key words: domain Ontology; natural language processing; thesaurus; knowledge engineering

Ontology 是一种能在语义和知识层次上描述系统的概念模型,其目的在于以一种通用的方式描述领域中的知识,提供对领域中概念的共同一致的理解,从而实现知识在不同的应用程序和组织之间的共享和重用.

Ontology 作为一种新的知识组织方式, 自 20 世纪 90 年代提出以来, 受到了国内外越来越多的关注; 但 Ontology 研究实际上还处于初步阶段,其理论和方法都有待于进一步完善. 特别是现阶段的 Ontology 很多都是人工开发的,这样需要耗费大量的人力、物力和财力,时间周期也很长. 由于缺少比较理想或实用的领域 Ontology 或通用 Ontology 作为基础,Ontology 的应用研究举步维艰, 正如文献[1]所述:由于 Ontology 构建的困难,以及构建技术不成熟等原因,现在真正能对 Ontology 进行的应用还很少. 在各个应用点方面也只是设想, 或者是在小型的 Ontology 上进行实验,大型的实际应用仍然依赖于构建技术的突破. 因此,Ontology 的有效构建成为 Ontology 研究乃至语义网研究的瓶颈. 探讨构建领域 Ontology 的有效途径, 特别是领域

收稿日期: 2006-09-06
基金项目: 国家"973 计划"项目(2004CB318102); 国家自然科学基金项目(60503071); 北京市自然科学基金项目(4052019)
作者简介: 刘耀(1972—), 男, 讲师, 博士后, E-mail: sdliuy@pku.edu.cn.

Ontology 自动构建的有效途径，成为了一个无法回避的问题.

目前，在 Ontology 构建途径方面，虽然提出了多种方式，但多集中在叙词表与 Ontology 融合、转换等方面[2-3]. 在自动构建技术方面，除少数针对英文资料利用自然语言处理(NLP)技术进行构建的一些设想或小型实验外，国内未见针对性研究报道，相关技术多散见于 NLP 技术及其他领域内.

1 领域知识分析

任何一种新的组织方法，都是在传统方法的基础上发展而来的. 实现领域 Ontology 的自动构建，务必是建立在大量公认领域知识的基础之上. 因此，公认领域知识的有效选择，也就成为了领域 Ontology 自动构建的前提.

1.1 专业叙词表

通过对 Ontology 与传统信息组织方式的关系分析不难发现，Ontology 与以叙词表为主体的主题法极为相似. 那么，主题法所描述的知识，能否作为公认的领域知识引入 Ontology 呢?

首先还应从叙词表的构建阐述. 以叙词法为主的主题法形成于上世纪 50 年代末，是在吸取元词法、标题法及分面组配式分类法等知识组织方法优点的基础上发展起来的. 主题法以研究特定事物为中心，揭示与特定事物有关的全部或部分问题，以表达事物主题概念的规范化词语字顺的先后次序排列. 主题法所使用的规范化语言是被有关的权威机构控制、承认并使用的，其词表中的术语含义明确、清晰、精练、直观、易记，能及时反映新学科、新技术的发展. 词表的优劣依赖于管理机构对术语选择的严格程度，一般而言，词表的选词要遵守以下规则：①如同样的术语在不同的上下文中有不同的概念含义，则必须在名称中对其模糊语义予以限制；②如有多个术语表达同样的含义，则其中的一个词作为词表的首选词，其他则列为同义词或别称. 从选词规则可看出，词表是一个术语的集合，这些术语是被该学科领域公认的，具有明确的含义[4].

另外，专业叙词表不但包含了本学科领域中相对完整的术语，而且都经过了该领域专家多年的有序组织；不仅可以为领域 Ontology 中概念的创建提供指导，而且叙词表中的限义词、含义注释、等级关系、词间关系也为领域 Ontology 概念中的属性、实例，以及关系的创建提供了线索及指导，这将为领域 Ontology 的创建者节省大量的时间及精力.

再者，主题法资源极为丰富. 从 1959 年美国杜邦公司编制的第一部叙词表到 2002 年，国外叙词表已超过 2 000 种，我国叙词表也超过 130 种[5]，基本上覆盖了所有领域，为迅速创建各领域 Ontology 提供了坚实基础.

因此，把叙词表作为公认的领域知识引入 Ontology 的构建中，颇具合理性.

1.2 专业辞典

专业辞典又称为专科辞典，一般具有：由权威机构组织领域专家编写，并经多次修订；准确、全面收集该领域的相关词汇或术语；及时覆盖新出现专业词汇，充分体现专业词典的“新”、“专”等特点.

另外，辞典与 Ontology 也具有一定的相似性，即两者均由概念或词条构成、均对概念或词条有不同程度的解释或说明、均是以提高检索效率与知识的共享为目的. 因此，将专业辞典引入领域 Ontology 的构建，具有一定的优势.

1.3 专业教材或权威著作

教材作为人类文明的结晶和传承与发展人类文明的载体，凝聚了人类文明和人类知识的精华，具有权威性、学术性和知识性的特征，特别是以培养专业技能为核心目的的、以自然科学各学科为代表的高等教育专业教材，更具备这些特性. 其主要体现在 2 个方面.

一是科学性. 教材在符合学科专业培养目标的基础上，在结构安排方面，由浅入深，符合学生的认知规律，并注重与本学科和其他相关学科体系教材之间的衔接；在内容设置和表达方面，概念的说明、原理的推导、观点的表达等应正确、严谨和符合语法规范，并体现学科发展的新内容.

二是先进性. 教材不仅能吸收已有的科技文化成果，更能在适合我国科技和文化水平的基础上吐故纳新，不断吸纳科学文化和本学科的最新成果.

另外，教材形式的专业书籍或著作多以涉及范围全面、系统、内容详尽为特点，常被专业学者作为具有保留价值的参考书，用于疑难问题的查询.

综上所述，一套权威的专业教材或著作，不但能全面地涵盖该领域的基本知识，而且能系统地反映该学科的体系结构. 因此，将其作为公认领域知识，引进领域 Ontology 的构建中，是完全可行的.

2　设计思想

基于以上对公认领域知识的分析，提出利用 NLP 理论和技术方法，对已有公认领域知识进行重构利用；借助领域专家知识，建立受限文本的 Ontology 自学习机制，最终实现领域 Ontology 概念描述体系自动构建的理论与方法.

2.1　基础流程

实现领域 Ontology 的自动构建，务必是建立在大量公认领域知识的基础之上，因此，如何成功地将其引入到 Ontology 的构建中来，也就成为了首要任务. 其流程如图 1 所示.

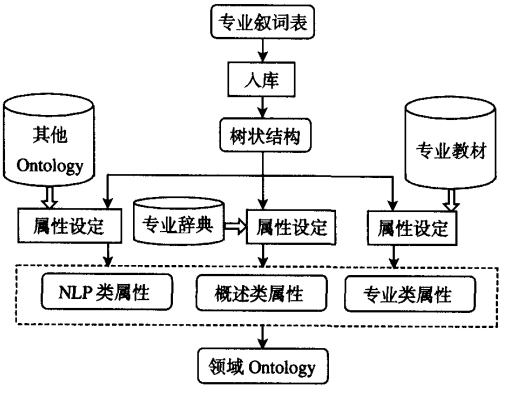

图 1　基础流程图

2.2　扩充流程

基于网络资源，进行知识采集与加工；进而实现受限文本的 Ontology 自学习机制.

Ontology 是一个开放集成的体系. 底层知识库与概念集应该随着学科领域的更新和发展随时进行修正和更新，因此，针对权威机构网站发布的更新信息，进行定期采集与获取，可以有效地解决这一问题. 其流程如图 2 所示.

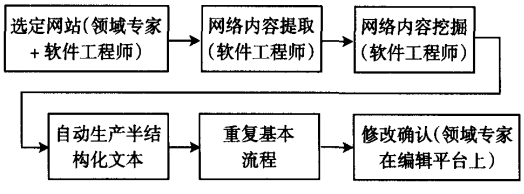

图 2　扩充流程图

3　实现步骤与方法

3.1　专业叙词表的重构与利用

专业叙词表的自动导入实现了描述语言上的一种转换，但是，领域 Ontology 中概念的设计，应最大限度地贴近研究者要研究的专业领域中客观对象和对象间的关系法则. 因此，专业叙词表虽然是该专业公认的领域知识，但叙词表多是由图书情报人员编写而成，有较强的文献标注覆盖面，不能深层次反映学科内在联系，必须对其进行知识重构，使其具备更强的学术性、专业性.

3.1.1　从树状结构到立体网状结构的转变

叙词表为了文献标引的方便，多把主题词平行分布在多个树状结构内，背离了专业知识体系描述习惯与方法. 因此，必须在专业叙词表中寻找关键类，以此作为知识描述的基本单元，建立层次结构体系. 该设计不但可以实现概念描述体系从树状结构到多层嵌套的网状结构的转变；同时，也有效地实现了领域 Ontology 最大单向可扩展性.

例如，MeSH 表为了方便文献标引，把医学领域中主要主题词平行分布在 15 个大类中，如 A 解剖、B 有机体、C 疾病、D 化学制品和药物、E 分析诊断和治疗的技术及设备、N 卫生保健等，类与类之间并无主次之分，背离了医学知识体系描述习惯与方法. 因此，这里在 MeSH 表所分的 15 个大类中，以“疾病类(C)”主题词作为知识描述的基本单元，建立疾病类的层次结构体系，以此建立知识的纵向关联；以其他类作为对“疾病类”知识元的描述属性，以此建立疾病类知识的横向关联.

通过这种转变，不但实现了概念描述体系从树状结构到多层嵌套的网状结构的转变；同时，也预留了大量接口，如“人文科学”、“信息科学”、“社会学和社会现象”等，它们既是多层嵌套网状结构的有机组成部分，又以独立的树状结构而存在，从而有效地实现了领域 Ontology 最大单向可扩展性.

3.1.2　从文献检索到专家系统双重功能的转变

从树状结构到多层嵌套的立体网状结构的转变，虽然可以改变概念体系的描述结构，却没有改变对知识深层的描述方式，必须依据专业知识进行再次重构. 如医学领域可以“临床”为核心组织疾病类知识，即根据临床医学的知识描述框架，将疾病类知识框架中其他类(A 解剖、D 化学制品和药物、N 卫生保健)合并、拆分，得到疾病类属性包含症状与体征、治疗与护理等；同时将其他类也根据专业知识进行进一步的描述，如 D 化学制品和药物的描述属性为作用与用途、剂型规格、性状、用法用量、不良反应、注意事项、贮藏等. 以此分别建立其他 14

类知识的描述框架.

通过这次重构,实现从主要服务于文献检索与标注,到既服务于文献检索与标注又服务于临床诊断与治疗的双重功能的转变.

3.2 基于NLP技术知识描述体系的构建与获取

通过对专业叙词表的重构与利用,也就获得了领域Ontology的基本架构,但这远远不够,还需要集成NLP技术,实现从传统的知识描述到NLP智能分析描述的功能转变.

3.2.1 概念属性的深化描述

为了获得广泛意义上的构建方法与技术,本文突破学科界限,从自然语言分析和知识挖掘的高度出发,将每个概念的属性描述都分为3种方式,即概述类描述、专业类描述和NLP语义类描述.

1) 概述类描述

概述类主要包括名称、英文名、释义、代码与约束,其中名称、英文名、代码等,由叙词表等所带信息自动生成;释义是利用概念词(主题词)与专业词典词条匹配后,实现概念定义文本的自动填充.

2) 专业类描述

每个概念的专业类属性分为自然语言文本描述和知识元描述(NLP主题自动标引)2种描述形式. 如疾病类专业类属性可以描述为"症状与体征"、"发病部位"及"症状与体征2","发病部位2","症状与体征"、"发病部位"的属性值是利用自然语言文本进行描述的,即填充的属性值是自然文本;而"症状与体征2"、"发病部位2"的属性值则是利用自然语言文本描述属性中的文本内容,进行NLP主题自动标引后,进行映射关联形成的,即填充的属性值是相关结点(概念)属性的集成与关联(关联概念携带其固有关系及结构). 如图3所示.

3) NLP语义类描述

NLP语义类主要包括自由词(NLP自动切分)、同义词、相关词、中文概念词典(CCD)词等,其中自由词是由系统对其相应自然文本进行自动切分标注,并利用所得术语与已有概念集(叙词表)进行匹配后,没有相应匹配的术语组成,这种方法既可以有效集成新术语(即新概念的扩充),又可以有效控制概念的冗余度.

领域Ontology应该是该领域绝大部分知识重点的1个最少量的概念集合,同时这些概念应具有最小化的概念冗余. 概念的冗余度是指2个概念相似的程度,2个概念的冗余度大则表示这2个概念具有相近的意义. 当冗余度达到一定的域值时,就认为这2个概念可以只取其中1个[6]. 而叙词表的构建规则中明确规定词与概念之间一一对应,即1个概念只能用1个词表达,1个词只能表达1个概念. 词义规范为,对同义词、准同义词、近义词、不同译名、学名与俗名等加以规范,只能用1个规范化的词作叙词. 利用这一规则及现有成果,能有效降低概念的冗余度.

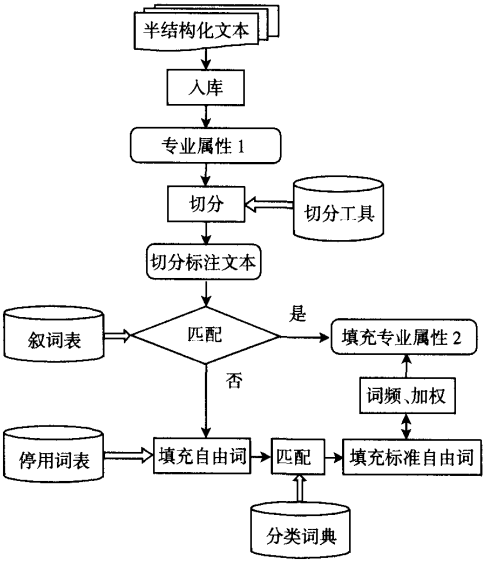

图3 专业类属性生成流程图

3.2.2 其他Ontology的集成方法研究

《中文概念词典(CCD)》是WordNet框架下的现代汉英双语概念词典,同时提供汉英双语概念的语义知识表达. 在词典的设计上,用同义词集合描述概念,用概念间的关系(relation)描述语义. 即针对中文的特点,CCD也对概念的内容和概念间的关系进行了一定的调整和发展;具有方便的语义关系表示和检索手段;同义词集合(同义关系)、上下位关系、整体部分关系等的描述,有利于实现概念的分级扩展和语义距离的计算. 作为基于概念的语义知识库,CCD在信息提取、文本分类等方面是不可或缺的基础资源,为其中的语义理解任务提供宝贵的语义知识库资源.

因此,将其相关概念进行匹配,作为NLP语义类描述属性的一部分,引入到系统的构建中,并对二者做了相关映射,从而有机地实现了领域Ontology与通用Ontology的有效衔接.

通过上述方法,实现了从传统的知识描述到

NLP 智能分析描述的功能转变，从而为领域 Ontology 的自动构建奠定了物质基础.

4　构建平台的研制与开发

将多种公认领域知识自动导入，是实现快速构建领域 Ontology 的必备条件之一，本文在系统实现之初，编制了多种针对性工具，将多种医学领域知识，如 Mesh、国际疾病分类、英汉医学辞典等自动导入到由 Protégé3.1 改进的 Ontology 编辑器（如图 4 所示），并成功保存其原有结构，节省了大量的人力、物力和财力，使项目在较短的时间内快速启动. 其主要特点为：①多样化的导入、导出方式（RTF/XML/OWL 等）. 方便与国际上相关的 Ontology 之间的知识交流、知识共享和知识重用. ②强大的编辑功能. 能够进行层次结构的调整、属性关系的调整、属性值的增删改等. ③强大的检索功能. 可以对知识元或属性进行精确查找和模糊查找. ④多层次网络的知识互联. ⑤多层次知识网络的可视化. ⑥NLP 自动分析. ⑦网络内容提取与挖掘.

图 4　领域 Ontology 自动构建平台界面之一

5　应用研究

在成功构建的基础上，进行了多种应用研究，主要体现在：基于知识元数据库，自动生成医学知识，引证和补充百科知识库；基于知识元数据库，从互联网中搜索相关文献，提高网络搜索的查准率；在搜索文献基础上分析文献内容，基于知识元数据库整理相关数据，形成对当前最新研究现状的总结、述评及趋势预测，运行结果如图 5 所示.

通过以上分析，不难看出本应用示范系统不但可以利用网络资源来辅助更新百科全书，而且也可以利用百科全书的权威内容指导、引领网络资源的开发和利用.

图 5　应用研究界面

6　结束语

因为任何一种新的组织方法都是在传统方法的基础上发展而来，所以，实现领域 Ontology 的自动构建，务必是建立在大量公认领域知识的基础之上的. 本文利用 NLP 理论和技术方法，并借助领域专家知识，对已有公认领域知识进行重构利用，建立起受限文本的 Ontology 自学习机制，实现了领域 Ontology 概念描述体系的自动构建，有效地解决了领域 Ontology 自动构建这一瓶颈问题.

参考文献：

[1] 何海芸，袁春风. 基于 Ontology 的领域知识构建技术综述[J]. 计算机应用研究, 2005(3):14-25.
He Haiyun, Yuan Chunfeng. Overview of technology of building domain knowledge based on Ontology[J]. Application Research of Computers, 2005 (3):14-25.

[2] 唐静. 叙词表转换为 Ontology 的研究[J]. 情报理论与实践, 2004, 27(6):642-645.
Tang Jing. Research on transforming thesauri into Ontology[J]. Information Studies:Theory & Application, 2004, 27(6):642-645.

[3] 唐爱民，真溱，樊静. 基于叙词表的领域本体构建研究[J]. 现代图书情报技术, 2005(4):1-5.
Tang Aimin, Zhen Zhen, Fan Jing. Thesaurus-based approach to build domain Ontology[J]. New Technology of Library and Information Service, 2005(4):1-5.

[4] 高凡，李景. Ontology 及其与分类法、主题法的关系[J]. 图书馆理论与实践, 2005(2):44-46.

[5] 常春，卢文林. 叙词表编制历史、现状与发展[J]. 农业图书情报学刊, 2002(5):25-28.

[6] 李景，孟连生. 构建知识本体方法体系的比较研究[J]. 现代图书情报技术, 2004(7):17-22.
Li Jing, Meng Liansheng. Comparison of seven approaches in constructing Ontology [J]. New Technology of Library and Information Service, 2004(7):17-22.