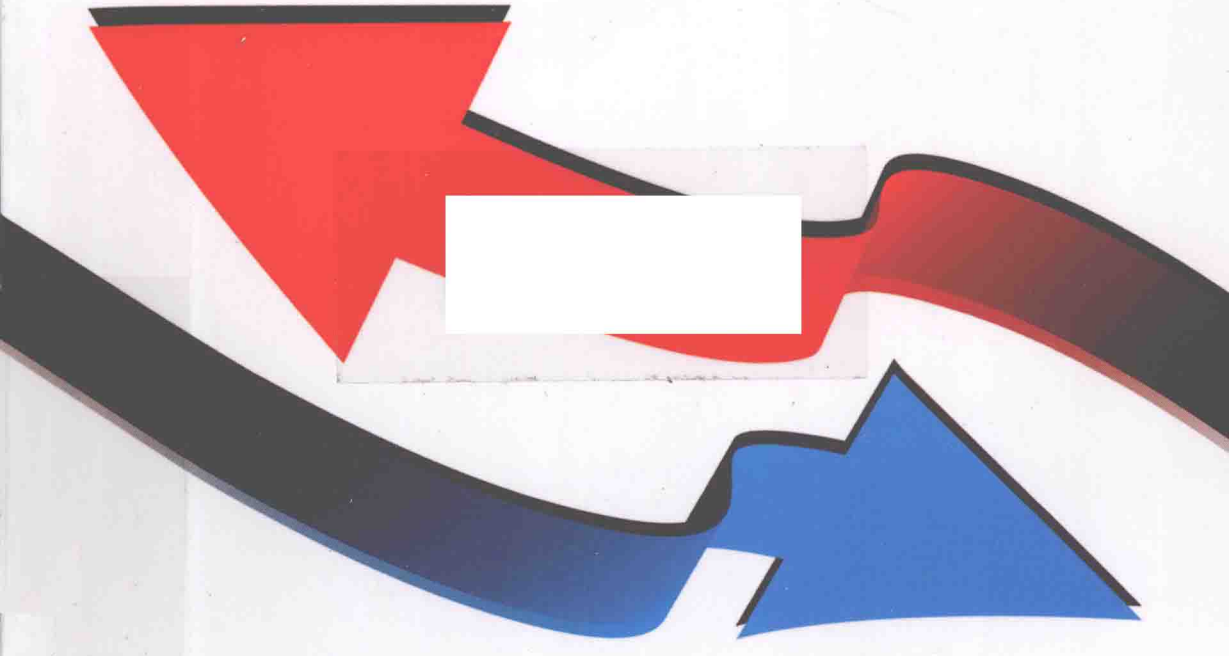
领域本体构建方法及实证研究

——以测绘学领域为例

余凡 著

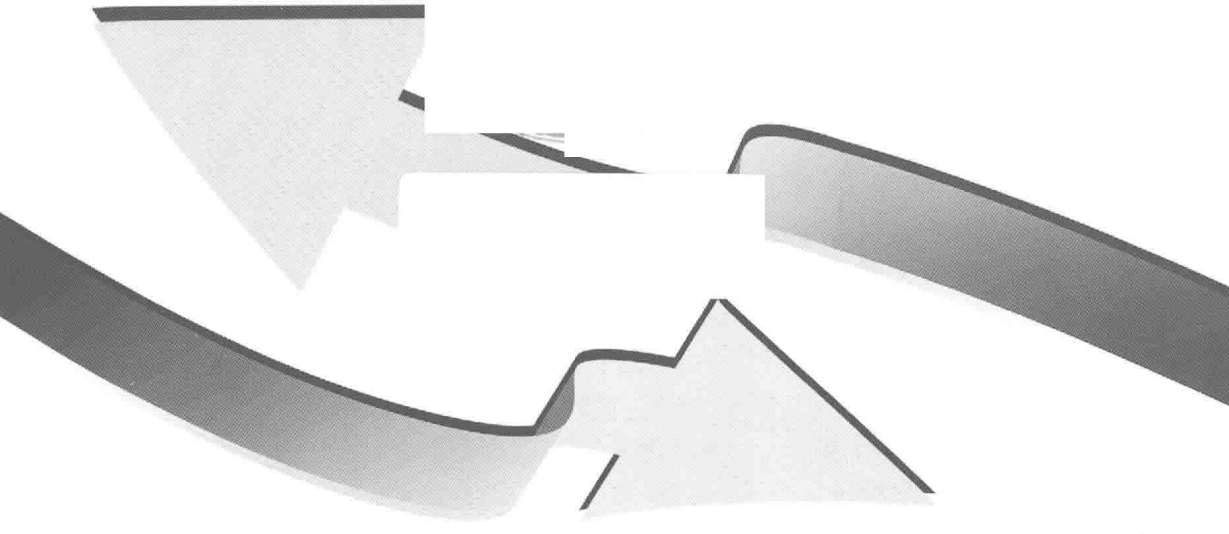


WUHAN UNIVERSITY PRESS
武汉大学出版社

领域本体构建方法及实证研究

——以测绘学领域为例

余凡 著



WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

领域本体构建方法及实证研究:以测绘学领域为例/余凡著. —武汉:武汉大学出版社,2015.8

ISBN 978-7-307-16246-4

Ⅰ.领… Ⅱ.余… Ⅲ.计算机应用—研究 Ⅳ.TP39

中国版本图书馆 CIP 数据核字(2015)第 155401 号

责任编辑:陈 红 责任校对:汪欣怡 整体设计:马 佳

出版发行:**武汉大学出版社** (430072 武昌 珞珈山)

(电子邮件:cbs22@whu.edu.cn 网址:www.wdp.com.cn)

印刷:武汉中远印务有限公司

开本:720×1000 1/16 印张:16.5 字数:328 千字 插页:2

版次:2015 年 8 月第 1 版 2015 年 8 月第 1 次印刷

ISBN 978-7-307-16246-4 定价:38.00 元

序

本体被引入信息科学、人工智能领域后，其在知识组织方面显现出了独特的优势。随着各个学科的研究者对本体产生兴趣，本体逐渐被引入医学、军事学、地理科学、农学等多个学科，本体的研究呈现百花齐放之势，形成了多个领域本体。纵观不同学科的领域本体构建方法，存在许多重复之处，但又不是完全相同，很难实现本体的重用和共享。如果能够规范领域构建方法，就可以使后来的研究者能够遵照规范的方法构建领域本体，不必再重新制定方法，避免重复浪费。基于此想法，笔者开始撰写这本书。

到目前为止，运用领域本体构建知识库的方法非常多，比如基于数理统计的N-gram算法、关联规则、相似度等，又比如基于语言学的规则提取算法。本书在现有本体构建抽象方法的基础上，以测绘学的领域知识为数据来源，检验了现有应用领域本体的方法，并对现有方法进行了部分优化，比如将叙词表和文本两种不同的数据源进行本体构建，又比如扩展了现有关联规则以及综合运用加权算法和信息熵筛选概念等，在这些方法的综合运用和优化的基础上，笔者提出了一套构建本体的具体规范方法。

由于本书是以方法研究为主，数据模型的构建、数据检验贯穿全书，为避免读者读起来乏味，笔者尽量运用通俗易懂的非专业词汇描述方法，即使是非本专业的读者，想必也能一口气读下来。同时，对领域本体感兴趣的读者，可以把这本书看作一本工具书，书里的方法非常详尽，每一种具体的方法也列举了具体的步骤。读者可以一边做数据检验，一边查看本书。

从开始写这本书到最终完成，前后差不多花费了一年的时间，而真正花在写作上的时间并不多，主要时间花费在数据实验上。书稿完成之际，想起在写作过程中得到过许多人的帮助，在此一定要感谢。首先要感谢的是我的妻子楼雯，在数据实验的过程中，难免会遇到实验失败的时候，心情沮丧时，妻子总能及时给予安慰，这是我执著地将这本书完成的最大动力。当然，还必须感谢在数据实验上给了我极大帮助的张聪，没有张聪的帮助，我的书稿是无法完成的。还要感谢写作过程中给予我帮助的所有人。最后感谢武汉大学出版社对本书的编辑和整理。

余　凡

2015年6月于珞珈山枫园13舍

摘　要

知识是世界经济蓬勃发展的重要因素，是人类社会文明传承和发扬的源泉。随着全球步入知识经济时代，知识已经成为国家经济发展、社会进步的可循环、利润回报非常丰厚的资源。互联网为信息的传播提供了广阔的平台，但却为广大用户搜索准确的信息制造了障碍。网络正在飞速地蔓延到世界的各个角落，如何在浩瀚的信息资源中快速地摄取到最准确的知识是我们不得不面对的一个共同问题。检索效率随着检索技术的更新不断提升，但是检索结果始终达不到用户的预期。如何规范地组织知识成为人们关注的焦点。本体的出现为知识组织带来了契机。本体被引入信息科学、人工智能领域后，其在知识组织方面显现出了独特的优势。随着各个学科的研究者对本体产生兴趣，本体逐渐被引入医学、军事学、地理科学、农学等多个学科，本体的研究呈现百花齐放之势。经过 10 多年的不懈努力，本体的理论、方法和应用都得到了极大的丰富。但是，本体构建方法的多样性、领域区分性为本体的重用、共享带来了困难。只有规范本体构建方法，才能保证本体构建流程的顺利进行以及实现大规模本体构建。对本体构建方法的对比、总结并加以改进，能够在提高现有方法构建效率的基础上规范方法的执行，从而使得本体在知识组织方面的优势最大化，为知识的存储、分析、检索提供强有力的保障。

本书以本体构建抽象方法为指导，把本体构建工作划分成概念获取、关系获取和形式化三部分，在对叙词表和文本两种不同数据源进行综合运用的基础上，对文本中的信息进行了基于组词规则和 N-gram 算法的概念提取、基于扩展互信息和上下文信息的概念过滤、基于加权算法和信息熵的核心领域词汇的筛选、基于空间向量相似度的等级关系提取、基于语法规则和扩展关联规则的非等级关系提取和基于 Jena 的形式化处理，最后以测绘学叙词表和文献为例，基于以上方法构建了测绘学领域本体，对方法的可用性进行了实证研究。通过对基于语言学和统计学的概念提取方法以及基于字和词的相似度方法的对比，对互信息和关联规则方法的扩展以及对加权算法和信息熵的综合，本书提供了一套构建本体的方法，并对现有的方法进行了改进，不仅丰富了本体构建的方法，而且为形成本体构建的一般方法提供了参考。

本书包括七章，除去引言和结论与展望外，剩下的五章主要分为三个部分：

第一部分（第 1 章）探讨本体及相关理论。首先对信息科学领域中本体的概

念进行描述和界定，讨论了本体在知识描述、知识共享方面具备的特征；列举并描述了通用本体、顶级本体和领域本体等九种不同类型的本体；阐述并解释了概念、关系、函数、公理和实例五个本体的基本元素；详细描述了 XML、RDF 和 OWL 三种本体描述语言的规范、标签以及三者之间的联系；描述并评价了 IDEF5 法、TOVE 法、骨架法和 METHONTOLOGY 法四种常见的本体构建抽象方法和规则匹配、N-gram 算法、互信息、信息熵、关联规则和相似度六种常见的本体构建具体方法；最后对 Protégé 和 Jena 两种构建工具及其优缺点进行了阐述。

第二部分（第 2、3、4 章）分别对本体构建的概念提取、关系提取和形式化三大块进行方法探讨和实验分析。其中：

第 2 章利用字符串函数和关系二维表的数据结构匹配和存储叙词，利用叙词表的编码规则实现映射，完成叙词由文本到数据库的结构转换。通过对叙词表切词和词性标注，提取最常用的叙词组词规则，利用叙词组词规则和 N-gram 算法提取文本概念，并描述了两种方法的算法，分析了两种方法计算的结果，将两种结果综合起来作为下一阶段的数据；对提取的概念进行了上下文和互信息过滤，并把两词互信息扩展到三、四词；最后对信息熵进行扩展，加入邻近词汇平均值后与加权算法一起筛选领域核心概念。

第 3 章利用关系二维表将叙词表中的属、分和族等级关系进行结构转换。在叙词表等级关系的基础上，通过相似度算法邻近词汇的筛选以及基于字和词两种相似度的计算结果对比，将相似度阈值分成同级类平均相似度、父子类平均相似度和同父类的子类平均相似度三种，文本中的概念以这三种阈值为标准添加进本体层次模型。将用和代两种非等级添加进关系二维表。利用中文造句的语法规则分别把主语、谓语和宾语提取出来，并在关联规则的基础上加入平均值对三元组进行筛选，最后得到本体所有的三元组。

第 4 章深入探讨了本体、OWL 和语义之间的关系，指出具有语义的数据是指能够减少用户参与，增加数据内容自动分析的数据；论述了选择本体描述语言的方法；分析了手工和自动两种本体形式化方法；最后利用 Jena 对测绘学领域本体进行形式化处理。

第三部分（第 5 章）构建了本体构建系统，提出了系统在分词、概念获取、关系获取和形式化方面的具体需求；对系统进行了总体设计和详细设计，总体设计中把系统分为概念提取模块、概念筛选模块、等级关系提取模块、非等级关系提取模块和领域本体形式化模块五大模块；详细设计中对每一模块的系统界面和功能进行了详细的论述。

目　录

0 引　言

0.1 选题背景与研究意义

0.1.1 选题背景

计算机和互联网的普及使得知识以前所未有的速度在全球范围内传播。利用知识创造经济价值的例子层出不穷。知识已经不再作为附属品，而是作为一种可持续利用的生态资源成为个人、企业、国家生存和发展的基本要素。

（1）知识已经成为经济增长最基本的因素之一

20 世纪 90 年代以来，知识已逐步成为经济发展的主要因素。1983 年，“新经济增长理论”的诞生标志着知识经济理论进入萌芽阶段。美国经济学家保罗·罗默认为知识是经济增长的主要原动力。罗默认为知识也应该成为基本的生产要素，知识能够直接对经济的增长产生积极影响①。1993 年，美国的“管理之父”彼得·德鲁克在《后资本主义社会》一书中指出，后资本主义社会，也被称为知识社会作为一种新的社会形态已经在经济社会的全球化发展中崭露头角。在知识社会里，资本、土地和劳动力这些传统的生产要素已经无法发挥决定性作用，知识才是生产力发展、经济发展的主导因素②。2005 年，联合国教科文组织在《迈向知识社会》的报告中再次提到“知识社会”一词，联合国教科文组织强调知识在知识社会中的重要性，指出要形成一个知识社会，必须以知识充分共享为前提③。《国家中长期科学和技术发展规划纲要 2006—2020 年》也提到要重点发展信息产业。知识服务是信息产业的核心。不同于其他生产要素，知识是完全可持续发展的。知识产生的经济效应已经广泛地体现在以高科技为主导产品的公司中。微软、百度等

---

① 胡炳志．罗默的内生经济增长理论述评［J］．经济学动态，1996（5）：60-63.

② Drucker P. F. *The Post-Capitalist World: Toward a Knowledge-Based Society* ［D］. Current: Washington, 1993.

③ 罗晖，程如烟．建设知识社会是人类可持续发展的必由之路［J］．中国软科学，2006（6）：156-160.

公司将知识充分地融合到产品服务中，它们的盈利完全可以和传统企业相媲美。知识作为经济增长的基本因素，其优越性正逐步体现并被放大。

（2）作为知识组织的概念模型，本体已成为研究热点

怎样用合适的方法表示知识以及将知识进行合理的组织是应用知识的前提。本体的提出为知识表示和知识组织带来了契机。通过对世间万物进行概念抽象，本体对概念以及概念之间进行明确的、无歧义的定义，最后利用一种人、机可理解的形式化网络语言按照一定规则将概念以及关系描述出来①。这种构造思想与其他数据组织方法截然不同。本体构建的知识库不仅能够涵盖传统关系数据库的所有信息，而且是对信息组织的一次升级，能够在信息的基础上提供结构清晰的语义层面的信息，将信息升级成知识②。本体能够对抽象的知识进行准确的描述，同时，W3C推荐的本体形式化语言与计算机网络紧密结合起来，这为本体的应用和推广提供了保障。本体在知识组织方面体现出良好的适用性，使得医学③④⑤、农业⑥⑦、地理⑧、工业⑨⑩、军事⑪等领域纷纷把本体引入本学科领域中，本体的研究呈现百花齐放之势。

---

① Studer R., Benjamins V. R., & Fensel D.. Knowledge Engineering Principles and Methods [J]. *Data and Knowledge Engineering*, 1988, 25 (1-2): 161-197.

② 谭富强. 本体知识库集成的方法研究 [J]. 科学技术与工程, 2012, 20 (6): 1416-1420.

③ Juijen C., Peiwen P., &Yungcheng H., et. al. Applying Ontology Techniques to Develop a Medication History Search and Alert System in Department of Nuclear Medicine [J]. *Journal of medical systems*, 2012, 36 (2): 737-746.

④ 付强. 基于主题词表的医学领域本体的构建研究 [D]. 长春: 吉林大学, 2011: 38-58.

⑤ 谢琪. 基于本体方法构建中医药概念信息模型的方法学示范研究 [D]. 北京: 中国中医科学院, 2011: 51-69.

⑥ Xiaolu S., Jing L., & Yunpeng C. Review on the Work of Agriculture Ontology Research Group [J]. *Journal of Integrative Agriculture*, 2012, 11 (5): 720-730.

⑦ 李景. 本体理论及在农业文献检索系统中的应用研究——以花卉学本体建模为例 [D]. 北京: 中国科学院研究生院, 2004: 145-181.

⑧ Alexander M., Olga P., &Nils D. Geography of Social Ontologies: Testing a Variant of the Sapir-Whorf Hypothesis in the Context of Wikipedia [J]. *Computer Speech and Language*, 2011, 25 (3): 716-740.

⑨ 杜振兴. 面向爆破行业的领域本体自动提取技术研究 [D]. 广州: 华南理工大学, 2011: 28-41.

⑩ 李丽亚, 李春梅, 薛中, 等. 工业自动化仪表领域本体的构建研究 [J]. 图书情报工作, 2009 (10): 111-115.

⑪ 张飞. 基于认知语言学的一般性军事本体构建 [D]. 长沙: 国防科学技术大学, 2009: 57-94.

（3）本体没有形成统一的构建方法

虽然本体已经有了一个明确的定义，但是没有一套成熟的、规范的构建方法。本体库的构建成为一项费时费力的浩大工程。本体的构建一般分为两大部分。第一部分是确定本体的概念及其关系。第二部分是利用各种工具、方法使概念及其关系形式化。本体概念及其关系构建的方法有很多种。利用分词工具对文档进行切词，然后定制抽取规则获取领域概念和关系是比较常见的一种方法。该方法无法自动获取顶层本体，顶层本体需要领域专家确定。利用叙词表本身的层次构建映射成本体相对简单易行，不过生成的本体库揭示的语义也相对肤浅，需要人工增加内容丰富本体库。本体学习主要是利用基于语言学的知识抽取算法、自然语言分词算法以及聚类、统计等方法对文档内容进行自动处理，抽取概念、关系的一种方法。相对于前两种方法，本体学习提高了本体库的自动构建程度。本体的形式化的目标是将概念、关系生成 W3C 推荐的 RDF、OWL 等本体描述语言。利用 Protégé 可视化工具能够手工形式化本体，但是只能针对规模比较小的本体。利用关系数据库的特性与本体库的特性进行映射，能够完成大规模的本体形式化。基于主题词表和 SKOS 的本体形式化方案可行性较高，因为它们生成的关系数量很少，能够快速地制定映射规则。从本体构建方法来看，无论是本体概念、关系的确定方法，还是本体形式化的方法，都存在多种构建方法。如何有效地整合这些方法，并形成规范指导本体构建是本体研究者需要解决的问题。

（4）良好的知识组织是当前测绘学领域的迫切需求

测绘学是以多个学科的理论技术为基础的，包括计算机科学里的图像处理技术，物理学里的力学理论以及信息处理、信息传输等技术①。不同的学科都包含着大量学科知识，将这些知识聚合在一起作为测绘学的理论技术基础，无疑需要对这些交叉学科知识进行有序的组织。杂乱无章的知识的简单堆砌显然很难让我们对测绘学进行全面深入的研究。测绘学已经广泛地应用到各个领域，比如：对土地资源的监测，掌握土地资源的利用和分布情况；对城市建设的调查，分析城市交通、污染等各方面的情况；对考古遗址的探测，全面掌握遗址的信息。作为一种信息获取技术，随着高新科技的出现，测绘学能够获取的信息越来越多，如何在各个领域获取的大量信息中提炼出知识并对其进行科学合理的组织，以供研究人员认识、吸收、消化已经成为当前迫切需要解决的问题②。

（5）计算机科学、信息计量学等多学科的发展为本体的构建奠定了坚实的

① 黄茂军，杜清运，杜晓初．地理本体空间特征的形式化表达机制研究［J］．武汉大学学报：信息科学版，2005，30（4）：337-340.

② 刘伟，顾和和．基于语义的地理信息空间关系检索［J/OL］．测绘科学［2013-01-17］．http：//www. cnki. net/kcms/detail /11. 4415. P. 20130118. 1414. 002. html.

基础

从数据源的分析到概念、关系的确定和抽取再到概念层次模型的生成，无一不需要多个学科的方法作为基础。语言学对于中文结构的研究能够对数据源进行句子结构剖析。计算机科学中文分词的研究以及开发的分词软件能够提高数据处理的效率。信息计量学的齐普夫定律以及与文献有关的统计、聚类方法有助于本体概念、层次的获取。本体描述语言以一阶谓词逻辑为基础，并且借用了面向对象语言的思想。本体的发布和应用还需要网络技术的支撑。可以说，多个学科方法的交叉使用才能实现本体的构建。反过来，随着本体研究的深入，多个学科的相关方法同样会得到更大的发展。

0.1.2 研究意义

领域本体的构建方法研究为实现领域知识序化和知识组织提供方法参考，同时也可以作为信息检索方法的有益补充。本研究不仅丰富了领域本体构建的方法，为形成领域本体构建方法规范做出一定贡献，而且为不同领域的知识管理、知识服务提供数据基础。

（1）规范领域本体构建的方法

本体已经应用到多个领域，各领域的本体构建方法仍处于摸索阶段，还没有走向成熟规范，不同领域往往会采取不同的本体构建方法。目前本体还没有形成一套完整的方法。只有对其进行广泛、深入的研究，才能找到领域本体构建的一般方法和规律。本研究提出本体构建的方法，并将其应用于测绘学领域，有利于为形成领域本体构建的一般方法提供新的思路。

（2）提供科学的知识组织方法

结构良好的知识组织能够充分反映出本领域的总体知识体系和局部知识结构以及各个知识点之间的关联。这对于了解本学科领域的知识，掌握学科热点，推测学科发展趋势起着不可替代的作用。相反，错综复杂的知识组织不但不利于掌握本领域的知识，还会加重学习知识的负担。知识组织也是知识服务、知识共享的基础和前提。没有好的知识组织结构，很难实现完全的知识共享，更难提供优质的知识服务。本体实质上是一种知识组织，是对知识的序化。通过构建测绘学领域本体，能够为领域知识建模提供新的思路。

（3）提高领域本体构建的效率

本体的构建分为几个阶段，从源数据的分词到概念、关系的抽取再到形式化表示，每一个阶段都会涉及人工干预问题。本体的终极目标是构建一个全球语义网，在该本体库中能够找到所有知识，因此，本体的研究从小规模的本体构建研究逐步扩大，向中大规模发展。随着本体规模的不断增大，手工构建本体已经很难满足当前的需求。如何提高本体构建的速度自然而然地成为研究热点。本研究在构建测绘

学领域本体的各个阶段都会首先尽量使用机器处理，然后对结果进行适当的人工干预。这样不仅可以提高本体构建的速度，将语料收集的繁琐过程从本体构建中剥离出来，减少人工干预引发的错误，而且还能够尽量避免仅仅依靠机器自动处理产生的噪音数据，提高本体库的准确性。

（4）提高信息检索结果的质量

传统的基于关键词的检索方式存在一些问题：检索者很难用简单的几个关键词来表达想要检索的内容；自然语言在知识表达上并不科学严谨，有些概念可能比较宏观，有些相对微观，最严重的是有些概念存在一词多义的现象，这些都会导致自然语言与索引关键词产生差异，从而影响检索的结果；检索者在检索一个概念时，可能想知道与此概念关系紧密的其他概念的相关信息。这些问题是基于关键词的检索方式无法解决的。其根本原因是基于关键词的信息检索是在语法层次上的检索，而检索者往往需要语义层次上的检索结果。领域本体的构建能够为语义检索提供数据支撑，提高返回结果的质量，为该领域的研究者提供更加准确的知识反馈。

（5）便于进行知识管理和提供知识服务

本体库是将领域知识进行规范化、统一化的过程。本体中定义的概念和关系都是本领域达成共识的知识，不存在歧义，同时，概念之间的层次结构非常清晰，因此，以本体库为基础，能够方便地对某一领域进行统一的知识管理。知识管理只是手段，并不是最终目的，最终目的是为广大用户提供知识服务，比如：知识导航，知识地图等。知识管理是知识服务的前提和保证。以本体库为基础建立一套完整的知识管理系统平台，能够实现语义信息检索和领域的知识共享，从而提供知识服务。

0.2 国内外研究综述

本体研究初期，本体构建的理论方法并不成熟，不同的研究者在不同的领域运用不同的方法构建本体。随着研究的深入，一些类似的方法被抽象出来，形成了相对通用的构建方法。应用相对普遍的方法有：应用于化学领域的 METHONTOLOGY 方法①，应用于知识建模的 KACTUS 方法②，利用抽象图形和具体描述语言进行本体构建的 IDEF5 方

① Fernandez M., Gomez-perez A., & Juristo N. METHONTOLOGY: From Ontological Art towards Ontological Engineering [C]. AAAI-97 Spring Symposium on Ontological Engineering, 1997: 33-40.

② Bernaras A., et al. Building and Reusing Ontologies for Electrical Network Applica-tion [C]. *Proc of the European Conf on Artificial Intelligence*, 1996: 298-302.

法①，对自然语言进行自动处理的 SENSUS 方法②，应用于企业建模的 TOVE 方法③以及为本体构建提供抽象指导的骨架法④。这些方法的提出为本体构建方法的发展奠定了基础，但是这些方法是在以完成本体构建为目标的前提下提出的，研究的重点并不是如何提高本体构建的效率。随着本体构建的规模不断扩大，构建领域不断丰富，研究者发现依靠纯手工构建大规模本体是不现实的。如何利用计算机技术替换手工操作，从而缩短本体构建周期成为研究的重点。以上提出的通用方法过于抽象，如何在这些抽象方法的指导下，快速高效地完成本体的构建是我们必须解决的问题。虽然通过不断的探索本体构建的抽象方法不断地更新，不同领域的研究者根据本领域的特点对抽象方法进行了部分修改，但是总体来说，本体构建方法大同小异，大体上可以分为以下几个步骤：确定领域范围，确定领域概念，确定领域关系，确定概念层次模型，模型形式化，本体评价⑤。针对不同的构建步骤，产生了很多自动化的构建方法。本体自动构建根据构建的不同步骤，可以划分为：本体概念获取研究、本体关系获取研究和本体形式化研究⑥。值得注意的是，这里提到的自动化仅针对本体构建中的一个步骤，并不是本体构建所有步骤的自动化。本体的构建方法从手工转变到自动是一个大的发展趋势，本研究将针对以上几个方面对本体的构建方法进行国内外研究综述。

0.2.1 基于不同数据源的本体半自动构建方法研究

在构建本体之前，需要确定获取概念、关系的数据源，不同的数据源对于本体的构建工作影响非常大。本体构建的数据源可以分为三大类：叙词表、文本、关系数据表。有的研究者同时利用两种数据源进行本体构建。下面分别对以上几种情况进行综述。

（1）基于叙词表的本体半自动构建方法

联合国粮农组织的研究项目中有一个关于本体构建的项目，该项目负责将

---

① IDEF Family of Methods [EB/OL]. [2012-11-20]. http://www.idef.com/.

② Ontology Creation and Use: SENSUS [EB/OL]. [2012-12-20]. http://www.isi.edu/natural-language/resources/sensus.html.

③ Gruninger M., & Fox M. S. Methodology for the Design and Evaluation of Ontologie [C]. *Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995: 1-10.

④ Uschold M. Ontologies Principles, Methods and Applications [J]. *Knowldege Engineering Review*, 1996, 11 (2): 93-136.

⑤ 董慧，余传明，杨宁，等. 基于本体的数字图书馆检索模型研究（Ⅲ）——历史领域资源本体构建 [J]. 情报学报，2006，25 (5)：564-574.

⑥ 刘柏嵩. 面向数字图书馆的本体自动构建 [J]. 中国图书馆学报，2006 (5)：47-51.

AGROVOC 映射成本体①。首先把 AGROVOC 叙词表中的词以及词间的关系与本体中概念以及概念之间的关系进行一一对应，然后编写计算机映射规则，最后将叙词表中的可用知识抽取出来。Markvan Assem 等以 MeSH 和 WordNet 为数据源，将 MeSH 和 WordNet 叙词表转换成本体②。Markvan Assem 等提出的半自动构建方法包括四个步骤，每一步都会根据叙词表的语法、语义来制定规则，实现叙词表到本体的转换。Wielinga 利用艺术与建筑叙词表构建了一个家具本体③。Wielinga 通过建立家具属性及艺术与建筑叙词表的联系来描述家具本体的概念和关系。这种对应的联系实际上也是利用规则进行映射。Hahn 利用规则将 UMLS 转换成严格的描述逻辑格式，也就是本体格式，从而实现大规模知识库的更新和维护④⑤。Kang 通过事先构建一个叙词表转换成本体的机器转换词典实现本体概念层次结构和关系的半自动化转换⑥。Assem 还提出了一种将叙词表转换成 SKOSRDF/OWL 标准的方法⑦。SKOS 是以 RDF 为基础的，因此也属于本体的范畴。Assem 利用规则映射实现了 IPSV，GTAA 和 MeSH 的映射工作。

张继东等运用 AGROVOC 叙词表本体构建的方法，构建了一个历史领域本体，进一步证明了该方法的实用性⑧。付佳佳开发了一个小系统，用于叙词表到本体的转换⑨。该系统包括几个模块：利用领域专家的经验和本体实例获取转换规则，并对规则进行检验，然后对叙词表进行规则的转换。仓定兰同样利用叙词表中含有的所有规则对叙词表进行映射，得到概念及关系，形成初级本体。然后对初级本体的

---

① Dagobert S. Building a Rich Ontology from AGROVOC [EB/OL]. [2012-12-20]. http://www.dsoergel.com/cv/B93.ppt.

② Assem M., Menken M., & Schreiber G., et al. A Method for Converting Thesauri to RDF/OWL [C]. Proceedings of the Third International Semantic Web Conference, 2004: 17-31.

③ Wielinga B., Schreiber A., & Wielemaker J., et al. From Thesaurus to Ontology [C]. Proceedings 1st International Conference on Knowledge Capture, 2001: 194-201.

④ Hahn V. Turning Informal Thesauri Into Formal Ontologies: a Feasibilit Study on Biomedical Knowledge Re-use [J]. Comparative and Functional Genomics, 2003 (4): 94-97.

⑤ Hahn U, & Schulz S. Towards a Broad-coverage Biomedical Ontology Based on Description Logics [J]. *Pac Symp Biocomput*. 2003 (8): 577-588.

⑥ Kang S., & Lee J. Semai-Automatic Practical Ontology Construction by Using a Thesaurus, Computational Dictionaries, and Large Corpora [J]. *Div of Electrical and Computer Engineering*, 2001 (6): 784-790.

⑦ Assem M., Malaise V., & Miles A., et al. A Method to Convert Thesauri to SKOS [J]. *The Semantic Web: Research and Application*, 2006: 95-109.

⑧ 张继东，余以胜. 利用叙词表构建本体的方法研究 [J]. 图书情报知识，2006 (4): 82-85.

⑨ 付佳佳. 基于叙词表的领域本体建模研究 [D]. 上海：华东师范大学，2006.

关系进行修改、添加和删除，得到最终的可用本体①。曾新红等在对中文叙词表进行仔细研究的基础上，制定了中文叙词表转换成能够用 OWL 本体描述语言进行形式化的映射规则，并对构建的本体结果进行了一致性检测，将该本体应用于知识共享②③。

基于叙词表的本体构建方法研究中，出现最多的两个词是规则和映射，说明叙词表转换成本体方法中最核心的部分就是制定规则和完成映射。通常情况下，规则的制定会在领域专家的协助下完成，并且由于叙词表的结构并不复杂，记录的关系也相对简单，所以制定的规则都很简单。确定了规则之后，就可以利用各种工具或者代码进行映射。完成映射之后，叙词表中的词及词间关系能够直接转换成本体中的概念和关系④。在本体的构建过程中，初始概念和关系的获取通常会耗费大量时间，基于叙词表的本体构建方法往往会直接跳过获取初始概念和关系的步骤，因此，此方法能够大幅度提高本体构建效率。

（2）基于文本的本体半自动构建方法

Tang Jie 提出了一种半自动语义标注的方法，该方法是以规则学习为基础的，主要针对中文的半结构化文本。Tang Jie 提出的方法中最核心的步骤是规则学习。规则学习利用相似规则学习算法对初始的规则进行相似性检验，相似度高的规则将被合并，从而解决了随机合并的现象。实证表明该方法具有较高的语义标注精度⑤。在研究中文文本语义标注的过程中，中文分词通常是进行语义标注的基础和前提。最大熵框架和条件随机场框架均采用字标注，利用字标注对中文进行分词，与词标注相比，分词的效果更加理想⑥。Pazienza 等提出的方法在字标注方法的基

① 仓定兰．基于叙词表的领域本体半自动构建的研究和实现［J］．科学技术与工程，2009，9（24）：7588-7593.

② 曾新红，明仲．中文叙词表本体共建共享系统研究［J］．情报学报，2008，27（3）：386-394.

③ 曾新红．中文叙词表本体——叙词表与本体的融合［J］．现代图书情报技术，2009（1）：34-43.

④ 张新，党延忠．基于规则与统计的本体概念自动获取方法研究［J］．情报学报，2007，26（6）：813-820.

⑤ Jie T.，Mingcai H.，& Juanzi L.，et al. Tree-structured Conditional Random Fields for Semantic Annotation［C］. Proceedings of the 5th International Semantic Web Conference，2006：640-653.

⑥ 张瑞霞，庄晋林，杨国增．基于《知网》的中文信息结构消歧研究［J］．中文信息学报，2012，26（4）：43-49.