领域本体的构建方法研究

（马文虎，南京理工大学信息管理系）

目　　录

摘 要：

领域本体的构建方法是当前本体研究的热点问题之一。但是，目前领域本体的创建还缺乏系统的、针对所有领域的、工程化的方法。本文介绍了本体的相关理论，并结合领域本体一般构建原则，分析了手工建立本体的六种常见本体构建工程思想,归纳总结出了构建领域本体的一般步骤。此外本文还重点分析了现有的领域本体构建方法以它们及存在的问题。

关键词：

本体 领域本体 构建方法

引言

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 等领域之中[1]。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，但是领域本体的建设问题仍然制约这些应用的发展。本文将详细研究和分析目前领域本体构建的各种方法以及存在的问题，为寻找新的构建方法提供参考。

1 本体的相关理论

本体又称为实体，源自于形而上学的哲学分支，它对客观世界的事物进行分解，发现其基本的组成部分，进而研究客观事物的抽象本质[10]。

1.1 本体的概述

本体最早是一个源于哲学的概念，是一种对“存在”的系统化解释，用于描述事务的本质。后来知识工程学者借用了这个概念，在开发知识系统时用于领域知识的获取[8]。

1.1.1 本体的定义

近年来，本体的概念被越来越多的应用于计算机知识工程领域，用于对客观世界的存在进行系统化描述，方便知识的重用和交互。人们已经从不同的角度和方面为本体论概念进行了定义。

虽然不同研究者对本体有不同的描述，但是从内涵上来看，他们都是把本体当作某个领域内不同主体(人、代理、机器等)之间进行交流的一种语义基础，即由本体提供明确定义的词汇表，描述概念和概念之间的关系，作为使用者之间达成的共识[10]。因此，本体的用途包括交流、共享、互操作、重用等。

本体是用于描述一个领域的术语集合，其组织结构是层次结构化的，可以作为一个知识库的骨架和基础。一般认为本体就是 Gruber 提出的“本体是概念模型的明确的规范说明”。Fensel 时这个定义进行分析后认为本体的概念包括概念化、明确、形式化和共享四个主要方面。

总而言之，本体的目标是获取、描述和表示相关领域的知识，提供对该领域知识的共同理解，确定领域内共同认可的词汇，并从不同层次的形式化模式上给出了这些词汇(术语)和词汇间相互关系的明确定义[6]。从而能够描述领域内部甚至更广范围内的一些概念和概念之间的联系，使得这些概念和联系在共享的范围内有着明确唯一的解释，这样人、系统之间就可以进行交流[11]。

一般来说，本体具有两个特征：静态性和动态性—静态性指的是它反映的概念模型，没有涉及动态的行为；动态性指的是它的内容和服务对象是不断变化的，针对不同的领域，可以定义和构造不同的本体[6]。

1.1.2 本体的构成

本体的体系结构应该包括 3 个要素：核心元素集、元素间的交互作用以及这些元素到规范语义间的映射关系。ISO 704 标准和 OKBC 模型是现有的有关本体体系结构的规定。ISO 704 认为本体的体系结构应含概念、定义和术语 3 部分。ISO 704 建议，一个概念应该用一个自然语言的术语得到理想的表达。

对于本体的具体构造过程，可以用以下公式（1-1）形象地表示:

本体 = 概念(Concept) + 属性(Property) + 公理(Axiom) + 取值(Value) + 名义(Nominal)　　（1-1）[6]

Perez 等人用分类法组织了 Ontology，并归纳出本体的五个基本构成元素（建模元语），即：①类（Classes）或概念（Concepts）；② 关系（Relations）；③ 函数（Functions）；④ 公理（Axioms）；⑤ 实例（Instances）。

从语义上讲，基本的关系共有 4 种，如表 1 所示：

表 1 基本的关系种类[9]

关系名	关系描述
part-of	表达概念之间部分与整体的关系。
kind-of	表达概念之间的继承关系，类似于面向对象中的父类与子类之间的关系。给出两个概念 C 和 D，记 $C'=\{x \mid x$ 是 C 的

	实例}，D′={x｜x 是 D 的实例}，如果对任意的 x 属于 D′，X 都属于 C′，则称 C 为 D 的父概念，D 为 C 的子概念
instance-of	表达概念的实例与概念之间的关系，类似于面向对象中的对象和类之间的关系。
attribute-of	表达某个概念是另一个概念的属性。如概念“颜色”是概念“玫瑰花”的一个属性。

在实际建模过程中，不一定要严格地按照上述 5 类基本建模元语来创建 Ontology，概念之间的关系不限于上面列出的 4 种基本关系，可以根据领域的具体情况定义相应的关系，以满足应用的需要，案例如图 1 所示。

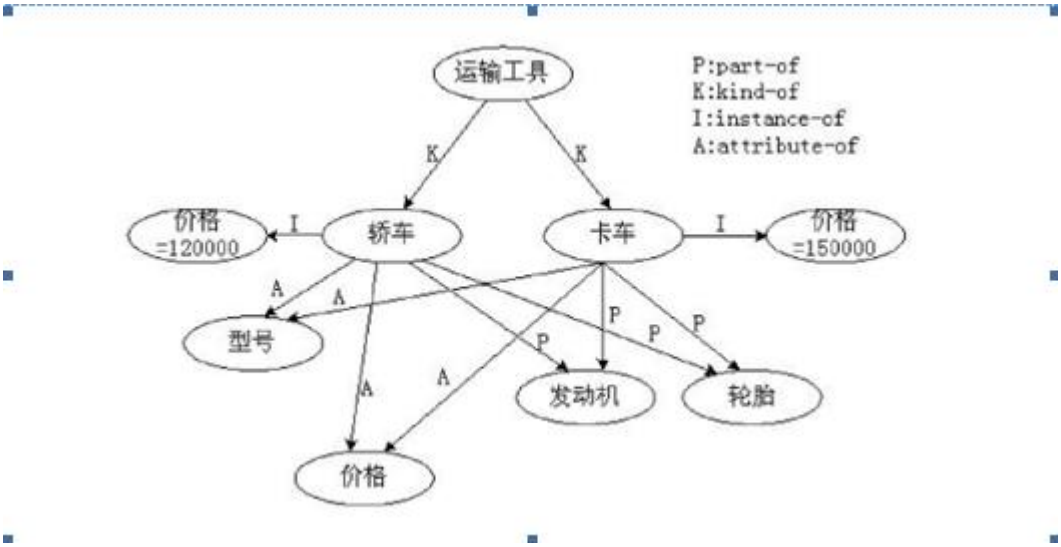

图 1 本体的构成案例[8]

1.1.3 本体的分类

目前关于本体的研究非常广泛，尤其是在国外，许多研究组织和机构都研究建立了各种各具特色的本体。针对目前出现的各种各样的本体，也出现了不同的分类方法，最为广泛的分类方法是根据本体应用主题，将这些为数众多的本体划分为五种类型:领域本体、通用或常识本体、知识本体、语言学本体和任务本体。

其中，领域本体在一个特定的领域中可重用，它们提供该领域特定的概念定义和概念之间的关系，提供该领域中发生的活动以及该领域的主要理论和基本原理等。对特定领域的本体研究和开发目前已涉及许多领域，包括企业本体、医学概念本体、酶催化生物学本体、陶瓷材料机械属性本体等。

领域本体主要有以下作用: 可以明确专业术语、关系及其领域公理, 使其形式化；在人与人之间、人与机器之间达到共享；实现一定程度的领域知识复用[10]。

此外，Guarin 也提出以详细程度和领域依赖度两个方面对本体进行划分。其中，根据本体对领域的依赖程度由高到低可分为四个类别：顶级本体(top-level

Ontologies)、领域本体(domain Ontologies)、任务本体(task Ontologies)和应用本体(application Ontologies)[12]，如图 2。

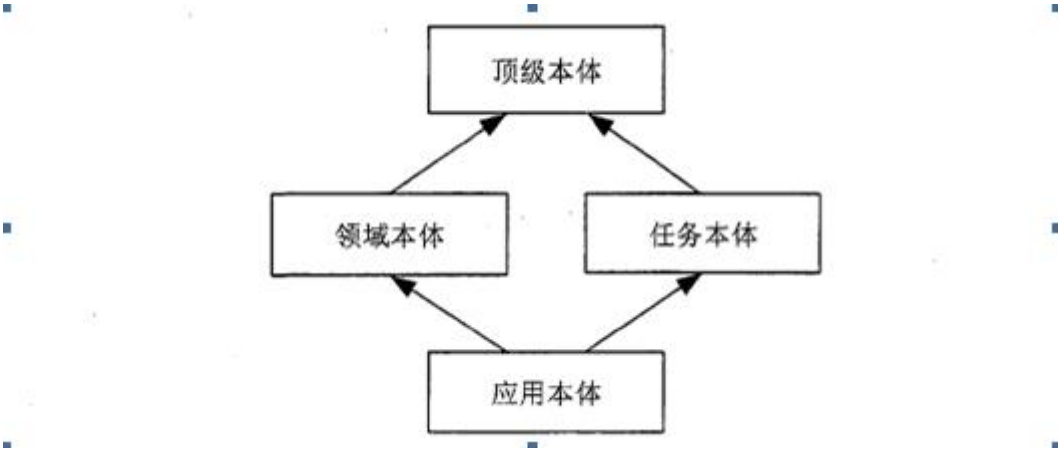

图 2 依照领域依赖程度的本体分类[5]

1.1.4 本体的应用领域

目前，本体已经被广泛应用于知识工程、自然语言处理、数字图书馆、信息检索和 Web 异构信息的处理、软件复用、面向对象技术和语义 Web 等领域。典型的应用有：

(l)基于语义的信息检索，特别是网络搜索引擎和数字化图书馆。

(2)基于本体的数据集成、机器学习等。

(3)领域本体的应用。比如，在生物信息学中已建成的 GeneOntology，尽管只包括了 part-of 等简单的关系，但是对生物信息学界已经有巨大的影响。

(4)语义 Web 服务。

(5)在线元数据管理和自动信息发布。[10]

1.2 本体的描述语言

关于本体的标记语言，可称之为置标语言，又称本体的构建语言或者是表示语言。作为表示本体的语言工具，应该具有如下的基本功能：

1) 为本体的构建提供建模元语(Modeling Primitives)。
2) 为本体从自然语言的表示格式转化成为机器可读的逻辑表达格式提供标引工具。
3) 为本体在不同系统之间的导入和导出提供标准的机读格式。
4) 形式化语言表示，利用机器可读的形式化表示语言表示本体，可以直接被计算机存储、加工、利用，或在不同的系统之间进行互操作。[9]

本体语言使得用户为领域模型编写清晰的、形式化的概念描述成为可能，因此它应该具备良好定义的语法、语义，有效的推理支持，充分、方便的表达能力。

自上个世纪 90 年代以来，大量的研究工作者活跃在该领域，因此诞生了许多本体描述语言，有 RDF 和 RDF-S，OWL（注:DAML+OIL 认为它是 OWL 的一个过渡)、KIF，SHOE，XOL，OCML，Ontolingua，Cycl，Loom。这里简单把它们归类如下：

(l)基于 Web 的本体语言(也叫做本体标记语言)有：RDF 和 RDF-S，OWL，SHOE，XOL。其中 RDF 和 RDF-S，OWL，XOL 之间有着密切的联系，是 W3C 的本体语言栈中的不同层次，也都是基于 XML 的。而 SHOE 是基于 HTML 的，是 HTML 的一个扩展。

(2)基于 AI(Artificial Intelligence)的本体实现语言有:KIF，Ontolingua，Cycl，Loom，OCML，Flogic。KIF 已经是美国国家标准，但是它并没有被广泛应用于互联网，作为一种交换格式更多的应用于企业级。[10]

1.3 本体的编辑工具

到目前为止，已经出现了许多本体编写工具。根据这些工具所支持的本体描述语言，大致可以分为两类。

第一类包括 Ontolingua、OntoSaurus、WebOnto 等。这三个工具的共同点是，都基于某种特定的语言，并在一定程度上支持多种基于 AI 的本体描述语言。

第二类包括 Protégé 系列、WebODE，OntoEdit，OliEd 等。这些工具最大的特点是独立于特定的语言，可以导入/导出多种基于 Web 的本体描述语言格式(如 XML，RDF(S)，OWL 等)。其中，除了 OliEd 是一个单独的本体编辑工具外，其他都是一个整合的本体开发环境或一组工具。它们支持本体开发生命周期中的大多数活动，并且因为都是基于组件的结构，很容易通过添加新的模块来提供更多的功能，具有良好的可扩展性[10]。

1.4 建立本体的难点

本体的构造过程是个费时费力的过程，需要完整的工程化、系统化的方法来支持，目前特定的领域本体还需要专家进行参与。通用的大规模本体很少，大多本体只是针对某个具体应用领域或应用而构造的，在实际应用中，不同本体之间常常需要进行映射、扩充与合并处理，以及根据特定的需要从一个大的本体中提取满足要求的小的本体等操作，此外，当现实的知识体系发生变化时，先前构造的本体必须作出相应的演化以保持本体与现实的一致性，这都是本体工程所需研究的问题。

本体工程已成为现阶段研究中的一个热点问题。如何才能大规模的构造本体?如何集成现有的不同本体?如何维护本体及其进化过程?这一系列的问题都需要方法论作为指导，目前该领域研究还处于探索阶段，没有形成成熟的方法论，是一个有价值的研究方向。

此外，本体构造不仅需要理论上的探讨和研究，还必须实实在在的构造出本体。如何能利用软件系统辅助人们构造本体?这些软件能在哪些方面自动化或者半自动化的发挥作用?本体开发过程中如何支持协同工作?不同软件开发的本体如何集成?构造好的本体如何管理和维护?这些也成为该领域函待解决的问题。[4]

1.5 本体研究的现状

对本体的研究和应用近年来发展很快。在 1998 年 6 月，第一届“信息系统中的形式化本体论国际会议”的召开标志着这一领域在逐渐走向成熟。

从国外的研究情况来看，20 世纪 80 年代末至 90 年代初，哲学领域的概念“Ontology”被 AI 领域所借鉴，本体的建模方法也初步确立，本体论把知识工程中的知识向更深入的方向推进。近年来，国外对本体建模作了大量研究并将其运用于知识工程领域。主要代表为：① 万维网联盟 W3C（World Wide Web Consortium）的研究；② 德国卡尔斯鲁厄大学的 Rudi Studer，Alexander Maeche 和以他们为首的 AIFB 研究所从事的创建基于本体的知识门户和语义门户的研究；③ 美国斯坦福大学的知识系统实验室(KSL)对本体建模工具和本体应用层面的研究[9]。

与国外相比，国内无论是在理论研究、实证研究还是在技术手段的实现和应用方面都相对落后，与国外高水平的研究相比存在很大差距。国内对于本体的研究大约始于 20 世纪 90 年代初。

目前，国内进行本体研究的主要有三支科研力量。一是中国科学院计算所、数学所、自动化所的若干实验室，代表人物是陆汝钤院士、金芝博士、武成岗、曹存根等人。二是哈尔滨工业大学计算机系，代表人物是王念滨博士。三是浙江大学人工智能研究所，代表人物是博士生导师高济教授。

国内外重要的本体系统典型代表有：WordNet、FrameNet、SENSUS 、OntoSeek 、Cyc、GUM 通用上层模型(Generalized Upper Model)、HowNet、Mikrokmos 等。

2 领域本体的构建研究

领域本体(Domain ontology) 是用于描述指定领域知识的一种专门本体,它给出了领域实体概念及相互关系领域活动以及该领域所具有的特性和规律的一种形式化描述[16] 。目前本体构建主要有手工构建、复用已有本体（半自动构建）以及自动构建本体三种方法[17]。本节主要介绍手工构建本体的方法，并归纳出构建领域本体的一般步骤。

2.1 领域本体构建遵循的原则

目前已有的本体很多，出于对各自问题域和具体工程的考虑，构造本体的过程也是各不相同的。由于没有一个标准的本体构造方法，不少研究人员出于指导人们构造本体的目的，从实践出发，提出了不少有益于构造本体的标准。通过分析总结，本体的设计原则可以概括如下[10]：

① 明确性和客观性：即本体应该用自然语言对所定义术语给出明确的、客观的语义定义。

② 完全性：即所给出的定义是完整的，完全能表达所描述术语的含义。

③ 一致性：即由术语得出的推论与术语本身含义是相容的，不会产生矛盾。

④ 最大单调可扩展性：即向本体中添加通用或专用的术语时，不需要修改其已有的内容。

⑤ 最小承诺：即对待建模对象给出尽可能少的约束。

⑥ 最小编码偏差：本体的建立应尽可能独立于具体的编码语言。

⑦ 兄弟概念间的语义差别应尽可能小。

⑧ 使用多样的概念层次结构实现多继承机制。

⑨ 尽可能使用标准化的术语名称。

2.2 本体的构建工程思想

当前，建立本体大部分还是采用手工编辑方式，还远远没有成为一种工程性的活动，每个本体开发组都有自己的原则、设计标准和定义方法。为了减少本体构建过程中的人为参与，现在出现很多基于人工智能的半自动化及自动化本体构建方法。较纯手工的本体构建方法相比，这些方法虽然节省了效率，但遗憾的是也没有达到本体方法学的标准[12]。比较有名的本体构建工程思想有：

2.2.1 IDEF-5 方法

IDEF 的概念是在 70 年代提出的，是在结构化分析方法的基础上发展起来的。在 1981 年美国空军公布的 ICAM(integrated computer aided manufacturing)工程中首次用了名为“IDEF”的方法。IDEF 是 ICAM Definition method 的缩写，到目前为止它已经发展成了一个系列。IDEF5 是 KBSI(Knowledge Based Systems Inc.)开发的一套用于描述和获取企业本体的方法。IDEF5 通过使用图表语言和细化说明语言，获取关于客观存在的概念、属性和关系，并将它们形式化成本体。

IDEF5 创建本体的 5 个主要步骤是：① 定义课题、组织队伍；② 收集数据；③ 分析数据；④ 本体初步开发；⑤ 本体优化与验证。

2.2.2 Skeletal Methodolody 骨架法（Uschold 方法）

Mike Uschold & Micheal Gruninger 的骨架法（Skeletal Methodology），又称 Enterprise 法，专门用来创建企业本体(Enterprise ontology，是有关企业建模过程的本体)。“骨架法”流程见图 3。

不符合
符合
确定只是本体应用的目的和范围
本体分析
本体表示
本体的建立
本体的评价
评价

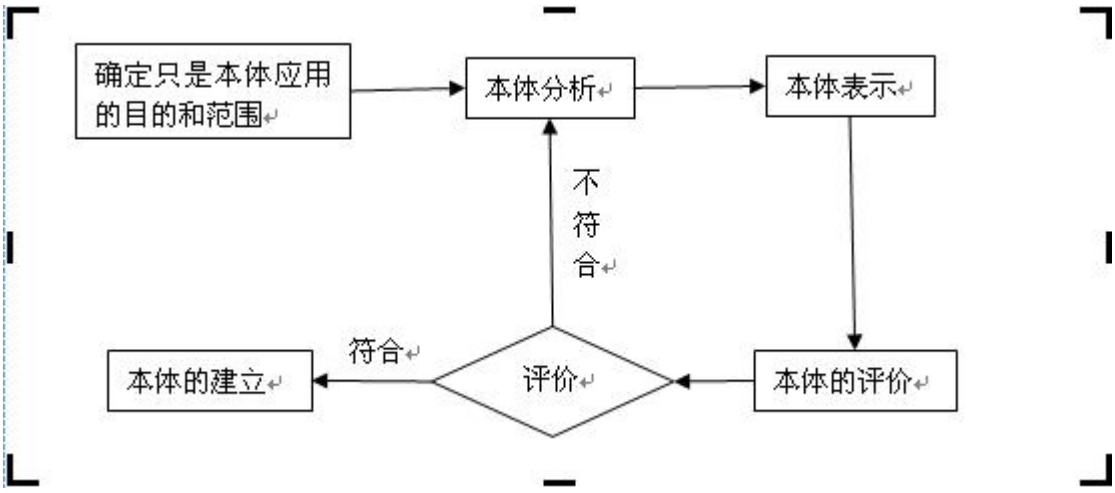

图 3 骨架法流程图[1]

2.2.3 TOVE 企业建模法

TOVE 法，又称 Gruninger & Fox“评价法”是加拿大 Toronto 大学企业集成实验室基于在商业过程和活动建模领域内开发 TOVE 项目本体的经验，通过本体建立指定知识的逻辑模型。用一阶逻辑构造了形式化的集成模型，包含企业设计本体、项目本体、调度本体或服务本体。

TOVE 流程见下图。

设计动机
非形式化的系统能力问题
术语的形式化
形式化的系统能力问题
使知识本体趋于完备
将规则形式化为公理

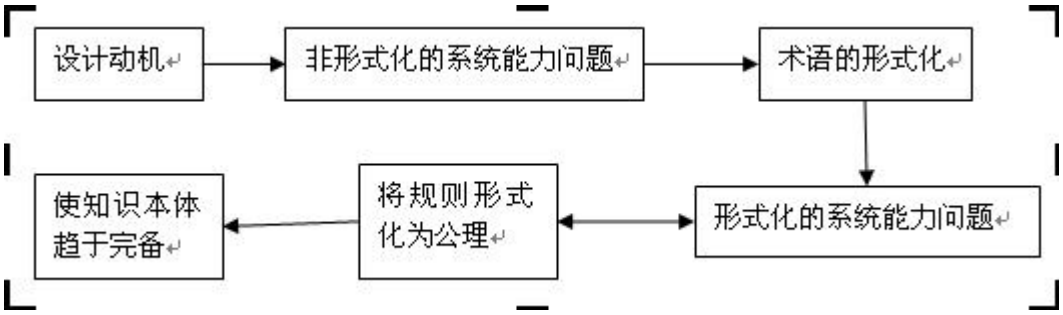

图 4 TOVE 流程图[1]

2.2.4 Methontology 方法

Mariano Fernandez & GOMEZ-PEREZ 等的 Methontology 方法是由西班牙 Madrid 理工大学 AI 实验室提出的。该方法是在结合了骨架法和 GOMEZ-PEREZ 方法后，提出的一种更为通用的本体建设方法。这个本体开发方法更接近软件工程开发方法。它将本体开发进程和本体生命周期两个方面区别开来，并使用不同的技术予以支持。

Methontology 法，专用于创建化学本体(有关化学元素周期表的本体)，该方法已被马德里大学理工分校人工智能图书馆采用。它的流程包括：

(1)管理阶段：这一阶段的系统规划包括任务的进展情况、需要的资源、如何保证质量等问题。

(2)开发阶段：分为规范说明、概念化、形式化、执行以及维护五个步骤。

(3)维护阶段：包括知识获取、系统集成、评价、文档说明、配置管理五个步骤。

2.2.5 循环获取法

Alexander Maedche 等的 Cyclic Acquisition Process，是一种环状的结构。基本流程如下：

(1)资源选取：这是环形的起点，是一个通用的核心本体的选择。任何大型的通用本体(像 Cyc、Dahlgren 的本体)、词汇-语义网(像 WordNet,GermaNet)、或者领域相关的本体(像 TOVE)都可以作为这个过程的开始。选定基础本体后，用户必须确定用于抽取领域相关实体的文本。

(2)概念学习：从选择的文本中获取领域相关的概念，并建立概念之间的分类关系。

(3)领域集中：除去领域无关的概念，只留下和领域相关的。这时，建立起了目标本体的概念结构。

(4)关系学习：除了从基础本体中继承的一些关系，其它的关系需要通过学习的方法从文本中抽取。

(5)评价：对得到的领域相关的本体进行评价，接着还可以进一步地重复上述过程。

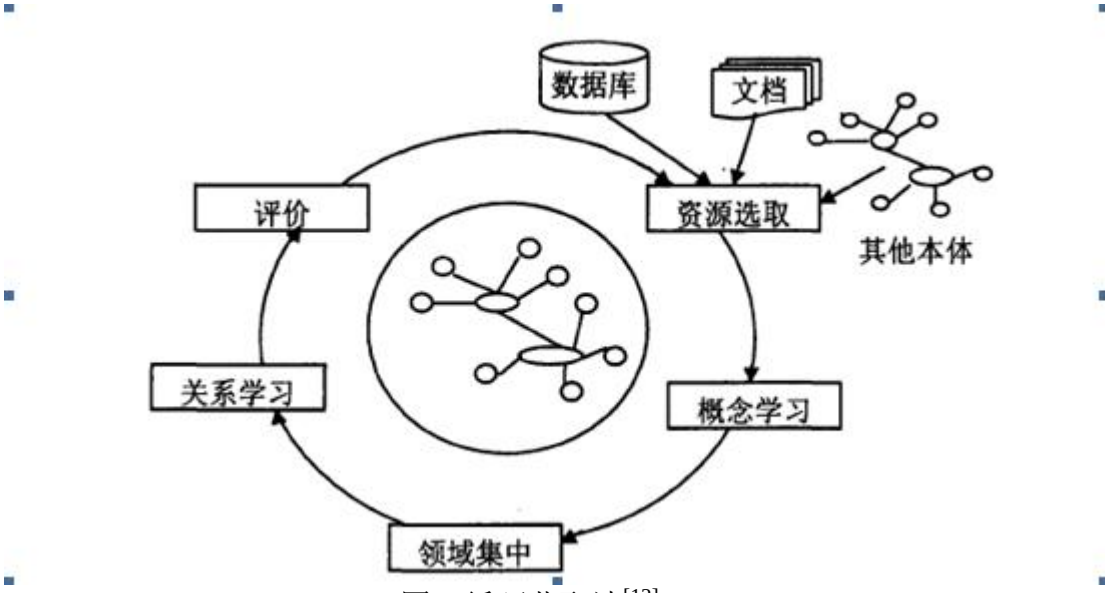

图 5 循环获取法[12]

2.2.6 七步法

斯坦福大学医学院开发的七步法,主要用于领域本体的构建。七个步骤分别是: ① 确定本体的专业领域和范畴; ② 考查复用现有本体的可能性; ③ 列出本体中的重要术语; ④ 定义类和类的等级体系（完善等级体系可行的方法有：自顶向下法、自低向上法和综合法[7]）; ⑤ 定义类的属性; ⑥定义属性的分面; ⑦ 创建实例[1]。

2.3 构建领域本体的步骤

本体的开发和完善是一个反反复复不断补充的迭代过程。领域本体中的概念应该贴近于要研究的专业领域中的客观实体和关系法则。综合上节几种本体构建的工程思想，归纳并总结出构建领域本体的几个步骤[11]:

2.3.1 确定领域本体的专业领域和范畴

领域知识往往十分庞大的，本体不可能包括所有的概念，因此，在建立本体前必须先确定本体将覆盖的专业领域、范围和应用目标，本体应该在哪些方面发挥作用以及它的系统维护者与应用对象。不同的应用领域，领域概念肯定是不同的，即使是同一个领域，由于应用的不同，本体表示的概念的侧重点肯定也会有所不同。因此，建立本体之前一定要明确本体建立的领域和应用目标。本体是一个复杂的知识体系，确定每个阶段的范围和目标有助于对本体模型的范围作一个限定，有利于复杂系统的实现。

2.3.2 考虑复用现有的本体

本体的主要作用就是解决知识的共享和重用问题。所以在设计和建立自己的领域本体之前，应该考虑重用已经存在的本体。如果系统需要和其它的应用平台

进行互操作，而这个应用平台又与特定的领域本体或相关概念联系在一起，那么复用现有的本体是行之有效的方法。例如 Ontolingua 的本体文库可以导入到本体开发系统中，并且本体的格式转换也并不困难。

2.3.3 列出本体涉及领域中的重要术语

领域本体是描述概念以及概念与概念之间的关系，首先要列举出该领域中的所有概念以及对该概念的详细解释。在特定领域，这些概念就是与领域相关的专业术语。把领域中一些重要术语列举出来，有利于知识工程师更好地理解本体建立的目标，明确方向。除此之外，针对每个概念，要列出它所有可能的属性，每个属性都有对应的属性值。

2.3.4 定义分类概念和概念分类层次

概念分类层次将领域概念进行分类组织，用于描述领域概念间的类属关系，并将本体中的概念模块化。建立一个分类概念的层次结构有 3 种可行的方法: 自顶向下法、自底向上法和综合法。

一般领域概念分类层次对应着一棵树，树中的节点体现了领域概念间的层次结构关系。树有四类元素组成：根节点，枝节点，树枝，叶节点。

建立领域概念的分类关系后，将分类概念的属性值添加到分类概念中，这样就把领域概念通过树形结构形象地描述出来，并且通过树结构清晰地体现了领域概念间的类属关系。每一个子树都对应着领域中独立的、模块化的知识模型。

领域分类概念应该包括:概念名称，语义描述，该概念可能的同义词、缩略语。定义分类概念，就是对这些信息进行描述。同时，要对所建立的概念分类层次进行检验，保证没有重复的概念，防止冗余定义。

2.3.5 定义概念之间的关系

概念的分类层次结构体现了分类概念之间的一种继承关系(kind-of)，但是在领域本体中，概念和概念之间通过关系来交互，除了继承关系，在我们构建的领域本体中还可以根据需要，定义其他的关系。

3 现有的领域本体构建方法及存在的问题

目前, 领域本体主要依赖手工构建,需要耗费大量的人力, 因此本体的构建成为第二代互联网发展的瓶颈。如何自动或半自动构建领域本体成为研究的热点。

国内外在本体构建方法上，研究最多的是以下两种方式：一种是从知识工程的角度，探讨本体的构建方法，可称为本体工程；一种是探讨利用现有的词表资源，直接向本体转化的半自动构建方法。此外，丁晟春、李岳盟等在综合二者的基础上提出了基于顶层本体的综合（半自动）本体构建方法[13]。

3.1 构建领域本体的知识工程方法

知识工程方法的主要特点是强调构建本体时要按照一定的规范和标准。相对于一般的系统，本体更强调共享、重用，可以为不同系统提供一种统一的语言，因此本体构建的工程性更为明显。目前为止，本体工程中比较有名的几种方法包括 TOVE 法、Methontology 方法、骨架法、IDEF-5 法和七步法等。这些方法大多是手工构建领域本体，具体过程已在上文中介绍，这里不再赘述。

由于本体工程到目前为止仍处于相对不成熟的阶段，领域本体的建设还处于探索期，因此构建过程中还存在着很多问题。与标准软件开发生命周期法 IEEE1074-1995[IEEE96]相比，还没有一种本体建设方法体系完全成熟。以上几种常用方法的成熟度依次为:七步法、Methontology 方法、IDEF-5 法、TOVE 法、骨架法[13]。

3.2 基于叙词表的领域本体构建

叙词表又称为主题词表，它是一种语义词典，由术语及术语之间的各种关系组成，能反映某学科领域的语义相关概念[15]。叙词表收录了某一领域的所有叙词和非叙词，按照一定顺序排列。叙词表的语义关系包括“用、代、分、属、参”，分别用来表示叙词款目之间的等同、等级、相关等语义关系。由于叙词表包含丰富的领域概念和一定的语义关系，在表达知识结构上与本体有着天然联系，包含了本学科领域中相对比较完整的术语，因此，国内外很多学术团体都在尝试着基于叙词表进行本体的构建，研究重点在于叙词表向本体转换的方法。

目前由叙词表进行转换的思路主要有两种: ① 直接用某种本体表示语言表示叙词表中的词汇和关系；② 仅将叙词表作为本体中概念的来源。这两种方式都需要对转换得到的本体进行属性、关系的添加和修正, 并添加公理和函数。

国外已经有 10 多种叙词表用各种方法转换为本体，如由联合国粮农组织转换为农业本体的 Agrovoc 叙词表，教育资料网关(GEM) 中的受控词表，艺术和建筑叙词表(AAT) 等。国外在这方面研究得比较成熟的是通过何种本体表示语言对叙词表的词语和关系进行转换，总结起来有以下几种:① 用 XML Schema 构建叙词标记语言。如澳大利亚 CSIRO 的 M. Lee 等所开发的叙词标记语言(TML)，构建了叙词描述本体的框架。② 用 RDF Schema 关系表示叙词内容。典型的如 AAT 一类的分面形式的叙词表，可以将叙词表某个子集作为本体某一类属性的值直接引入。③ 用 RDF Schema 表示叙词关系。大多数叙词表采用的是这种方式转换，如 L IMER 和 ELSST 社会科学叙词表等。④ 用 DAML + OIL 关系表示叙词关系。DRC 提出了一个用 DAML +OIL 表示叙词关系的建议。[13]

国内对叙词表转化的研究正处于热点阶段，目前已转化为本体原型的主要有《国防科学技术叙词表》和《中国农业科学叙词表》的一部分。中国农业科学院科技文献信息中心的常春博士基于《中国农业科学叙词表》的“作物大类”，构建了一个有关食物安全的本体原型。目前本体原型还正在进一步的完善研究中, 主要是解决核心本体与转化来的本体概念重复问题以及对叙词表原有关系细化等问题。中国国防科技信息中心的唐爱民等则对如何基于国防叙词表来构建国防领域本体进行了研究, 他们结合 Enterprise 方法、Methontology 方法与软件开发模型——“瀑布模型”提出了一种基于叙词表的领域本体构建方法。他们通过基于《国防科学技术叙词表》成功构建了军用飞机领域本体的原型，构建模型如图 6：

某学科领域叙词表

确定领域本体的应用目的
领域本体的整体设计
领域本体的详细设计
领域本体的表示
领域本体的评价
领域本体

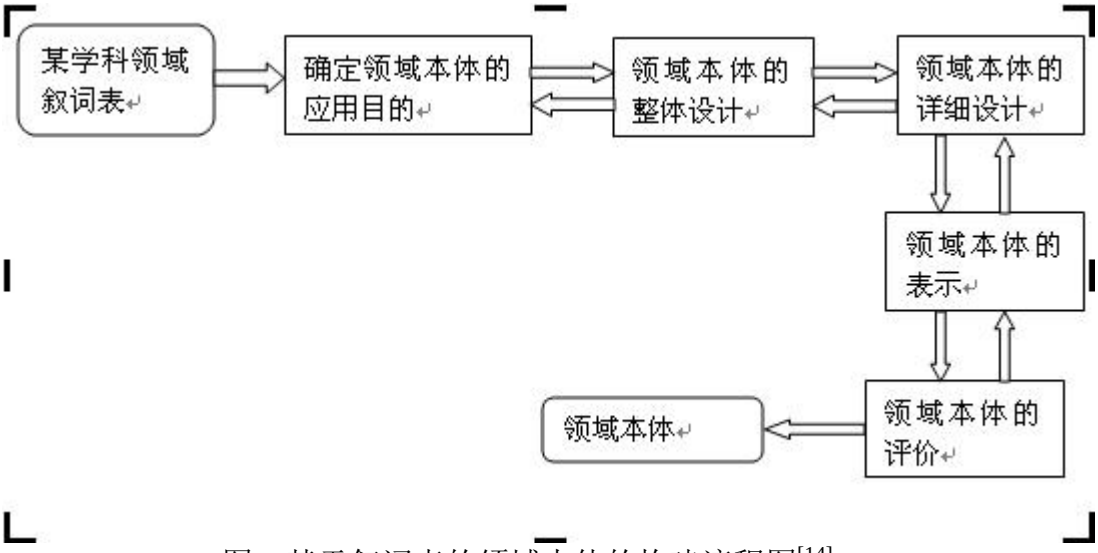

图 6 基于叙词表的领域本体的构建流程图[14]

其中，领域本体的详细设计过程也可称为领域本体的具体构建过程，详细设计是本方法中最核心、最关键的步骤，流程如图 7：

把叙词转换成领域本体中的概念
根据叙词间的层次关系，确定所对应的
领域本体中概念间的等级关系
参考叙词的限义词、注释为领域本体中的
概念添加属性
参照叙词间的关系为领域本体中的概念添加关系
为领域本体中的概念添加实例

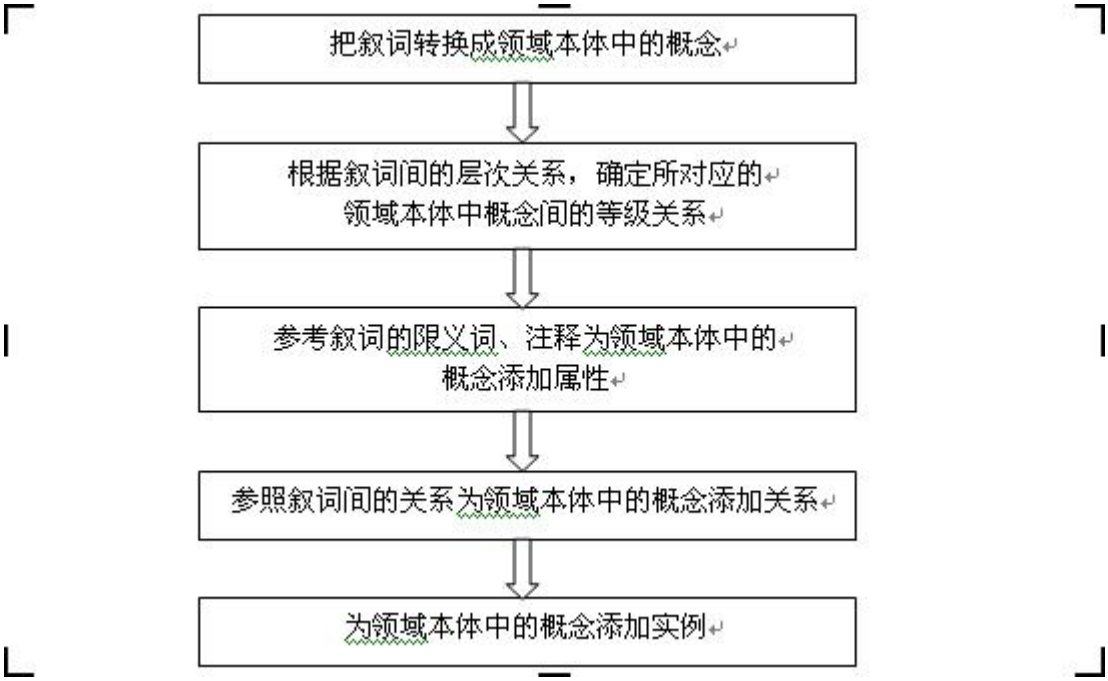

图 7 详细设计流程图[14]

3.3 基于顶层本体构建领域本体的构建方法

丁晟春、李岳盟等认为，本体构建的理论探讨已经比较成熟，但当将构建完的本体与实际应用联系起来的时候，就会浮现出本体构建过程中所存在的一些问题：① 领域本体构建与应用脱节；②领域本体难以复用和集成；③ 由叙词表难以转化成真正的本体；④ 本体构建的概念体系不够规范[13]。

针对本体构建与应用中出现的问题，她们深入考察了现有的本体构建方法和国外重要的三大顶层本体（Cyc、SENSUS 和 SUMO），并与中科院文献情报中心和中国农科院科技信息文献中心的专家学者就存在的问题和解决方案进行了深入探讨，提出了基于顶层本体开发领域本体的指导方法。该方法从本体工程方法论的成熟度和领域本体构建的特点出发，借鉴了骨架法和七步法，并融合了叙词表和顶层本体资源，对概念体系的规范化校验和本体的标准化处理提出了具体的方法和步骤。

研究方法的核心思想是，从本体工程的基本思想出发，借助词表法对选词进行规范化处理，并选择合适的顶层本体，对领域本体构建进行标准化处理，最后将领域本体嫁接入顶层本体中。基于顶层本体的领域本体构建框架如图 8 所示。

修正和进化
标准化处理
确定本体的领域和范围
考虑复用现有本体
定义类及类的等级体系
定义类的属性
创建实例
概念的规范化处理

顶层本体
本体表示
本体评价
合并入顶层本体
概念
体系
构建

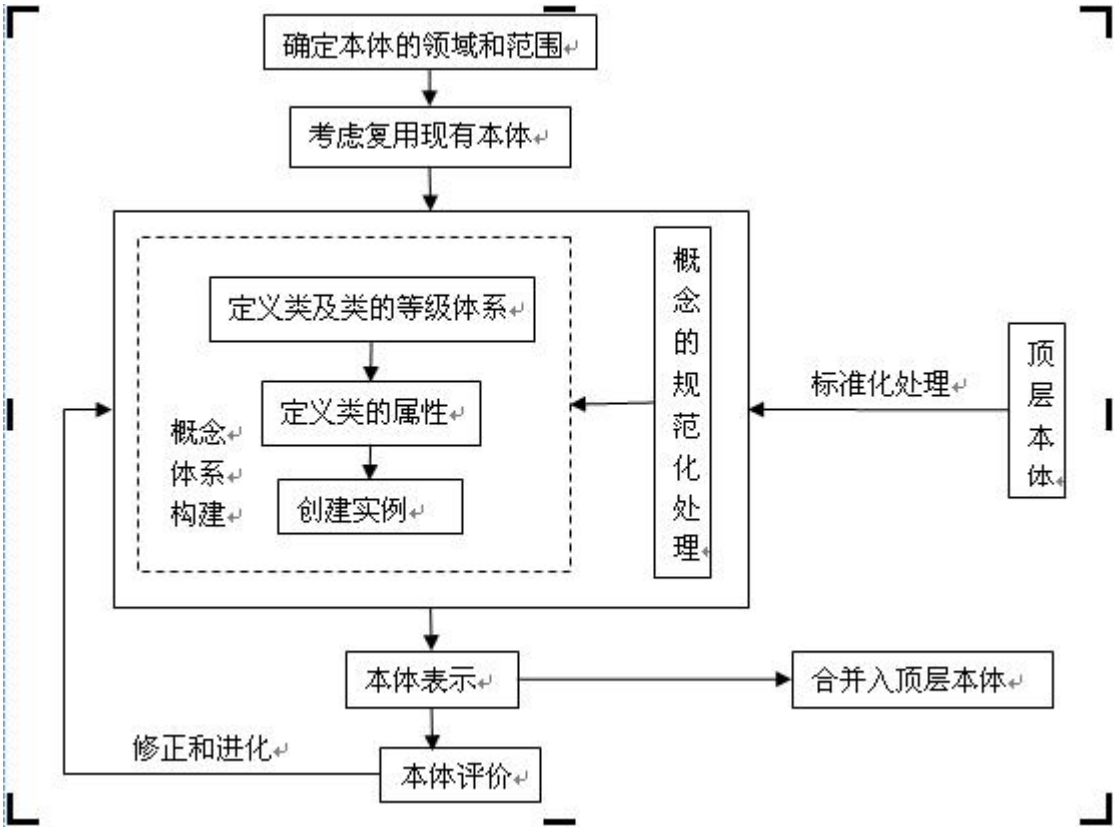

图 8 基于顶层本体的构建方法框架[13]

根据上述提出的基于顶层本体的领域本体的综合构建方法, 她们参照了《世界飞机手册》和《航空工业科技词典》等资料，在使用《国防科学技术叙词表》对概念进行规范化处理的基础上,构建出初具规模的、能面向实用的军用飞机领域本体(包含 300 多个类、70 多个属性、近 900 个实例) ,最后通过分析上层通用本体 SUMO 的构建标准和体系结构, 尝试着将该领域本体嫁接入 SUMO 中, 以实现本体的可共享和可集成。

3.4 领域本体构建过程中存在的问题

目前领域本体构建的主要三种方法：手工构建、复用已有本体以及自动构建本体，其中前两种方法最为常用。目前, 领域本体主要依赖手工构建，需要耗费大量的人力。但综合来看，三种构建方法都各自存在着不同程度的弊端。

3.4.1 手工构建

尽管本体编辑工具在近 10 年已经比较成熟，然而手工构建本体费时、费力且花费巨大，已经成为不争的事实。目前的手工构建本体主要方法有 TOVE 法、骨架法、IDEF-5 法、METHONTOLOGY 法、SENSUS 法、KACTUS 工程法、七步法等综合性方法。本体手工构建过程尚缺少一套工程化的科学管理流程作为支撑, 这使得本体的构建主观性太强, 且比较随意, 缺少科学管理和评价机制。

以上本体的建设方法存在主要问题有以下几点[7]：

1) 需求描述不充分和建设过程的无计划性

对于某个领域的本体建设，它的具体需求是很难描述清楚的。所以在没有充分明确的需求情况下去建设本体，会直接导致本体建设过程的无计划性，这样在建设过程中就有可能要重新计划。

2) 建设过程缺少规范性

领域本体建设还没有成熟的方法论作为指导，更不用说对建设过程的规范管理。从软件开发过程的管理中，可以看出文档的重要作用。因此，在领域木体建设过程中同样也得关注文档，从文档的编写中总结出规范。

3) 成果没有评价标准

本体的评价方法没有统一的标准，更没有标准的测试集。不能对本体的建设成果进行合理评价，必然影响到下一个周期中的进化过程。

4) 忽视本体的共享和重用

领域本体建设的目的不能仅为某一个系统提供服务，而是为不同系统提供交流的语义基础。本体建设的过程，也是人类知识机器化积累的过程。所以共享和重用是本体的本质要求，这也是领域本体建设中很重要的问题。

3.4.2 复用已有的本体

上文介绍的基于叙词表和基于顶层本体的构建方法均属于复用已有本体的半自动构建方法。复用已有的本体，可以获得领域知识以及概念关系，使得本体构建有一个很好的起点。

目前可复用的本体资源主要有：① 叙词表资源，如中国农业科学叙词表、国防科学技术叙词表等；② 顶级本体，如 Cyc、SUMO、WordNet、FrameNet 等；③ 数据库资源；④ 在线本体库，如 Ontology Engineering Group 和 DAML。

但是，目前很少有现存的不经修改就能被复用的本体，况且有不少领域没有可供利用的本体资源。同时本体复用带来了不同本体匹配的问题，本体映射目前仍然是第二代互联网研究中亟待解决的难题之一。此外有些本体资源改造起来需要大量的投入，改造已有本体的代价是否值得，也是目前正在研究的课题。

3.4.3 自动构建本体[17]

自动构建本体是目前的一个研究热点。研究者借鉴知识获取的相关技术，有基于自然语言规则的方法和基于统计分析的机器学习方法。目前这种构建方法还处于研究阶段，利用机器学习会产生大量的噪音数据，缺乏必要的语义逻辑基础， 抽取的概念关系松散且可信度无法得到很好的保障。利用自然语言处理技术，概念间潜在关系的分析则需要依赖复杂的语言处理模型。尽管机器学习应用于本体自动构建有巨大的潜力，但是距离良好的可理解性尚有很大的距离，随着研究的深入这种状况应该有望得到改善。

结 论

本体是某一领域共享的、概念化(conceptualization) 、形式化表示的知识体系。第二代互联网的发展需要大量的领域本体作为支撑。目前，领域本体主要依赖手工构建，需要耗费大量的人力, 因此本体的构建成为第二代互联网发展的瓶颈。

本文在笔者查阅、研究大量期刊和学位论文等资料的基础上形成的，论文首先对本体的相关理论（包括本体的定义、描述语言、建设工具等）进行介绍，结合领域本体一般构建原则，对各种领域本体构建方法以及存在的问题进行了详细分析。

创建领域本体的起点可产生自不同情况。可以是从抓取开始，也可以从已存在本体开始，还可从数据源文集开始，或者是后两个方法的组合。创建本体的自动化程度也是不同的，从完全的人工、半自动化到全自动化。当前, 全自动化的方法只能实现受限条件下的轻量级本体的构建。领域主体的构建是一项极其艰巨的任务, 如何应用知识获取技术来降低本体构建的开销目前也是一个很有意义的研究方向。

参 考 文 献

[1] 刘仁宁，李禹生. 领域本体构建方法[J].武汉工业学院学报，2008，27(1)：73-77.
[2] 李景，苏晓鹭，钱平. 构建领域本体的方法[J]. 计算机与农业，2003 (7)： 7-10.

[3] 顾芳. 多学科领域本体设计方法的研究[D]. 北京：中国科学院计算机研究所，2004.
[4] 张小鹏. 汉语特定领域本体的自动构造研究[D]. 武汉：华中师范大学，2007.
[5] 吴正超. 基于关系数据库的领域本体自动构建方法研究[D]. 大连：大连海事大学，2007.
[6] 廖军. 基于领域本体的信息检索研究[D]. 长沙：中南大学，2007.
[7] 刘爱军. 基于领域本体的语义信息检索及相关技术研究[D]. 西安：西北大学，2008.
[8] 翟林. 领域本体的半自动构建方法研究与实现[D]. 南京：东南大学，2005.
[9] 陈建. 领域本体的创建和应用研究[D]. 北京：对外经济贸易大学，2006.
[10] 郭嘉琦. 领域本体的构建及其在信息检索中的应用研究[D]. 北京：北京邮电大学，2007.
[11] 张志刚. 领域本体构建方法的研究与应用[D]. 大连：大连海事大学，2008.
[12] 张囡囡. 面向语义网的领域本体半自动构建方法的研究[D]. 大连：大连海事大学，2008.
[13] 丁晟春，李岳盟，甘利人. 基于顶层本体的领域本体综合构建方法研究[J].情报理论与实践，2007，30(2)：236-240.
[14] 唐爱民，真溱. 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.
[15] 孙倩，万建成. 基于叙词表的领域本体构建方法研究[J]. 计算机工程与设计，2007，28(20)：5054-5056.
[16] 肖敏. 领域本体的构建方法研究[J]. 情报杂志，2006(2)：70-72.
[17] 何琳，杜慧平，侯汉清. 领域本体的半自动构建方法研究[J]. 图书馆理论与实践，2007(5)：26-28.