《智能信息处理》课程考试

# 本体构建研究


**学　　院：** 信息科学技术学院

**专　　业：** 电子信息

**姓　　名：** 张献超

**学　　号：** 1120201446

**授 课 老 师：** 李冠宇


| 考核 | 课程成绩 |
| --- | --- |
| 得分 | |


**大 连 海 事 大 学**

**2020 年 11 月 27 日**

# 本体构建研究

张献超

(大连海事大学 信息科学技术学院，大连 116026)

**摘 要** 对于本体的研究在计算机科学领域越来越广泛，本体学习技术近年来逐渐成为计算机科学领域的一个研究热点。通过对本体构建方法和工具的综述总结，指出了这些方法与工具存在的问题。同时对研究现状中存在的瓶颈问题进行了总结，并提出了未来的研究方向。

**关键词** 本体；本体构建方法；本体构建工具

## Research on Ontology Construction

Zhang Xianchao

(School of Information Science and Technology, Dalian maritime university, Dalian 116026)

**Abstract** The research of ontology is becoming more and more extensive in the field of computer science. Ontology learning technology has gradually become a hot research topic in the field of computer science in recent years. By summarizing the methods and tools of ontology construction, the problems of these methods and tools are pointed out. At the same time, the bottleneck problems in the current research situation are summarized, and the future research direction is put forward.

**Keywords** Ontology; Ontology Construction method; Ontology Building tool

## 1 引言

本体(Ontology)是一个源于哲学的概念，指的是对客观存在物的解释和说明，即“存在论”[1]。本体是人工智能、语义网与知识工程等领域研究的热点，最被广泛认可的定义为“共享概念模型的明确形式化规范说明”。本体可以实现某种程度的知识共享和重用，使得计算机对信息和对语言的理解上升到语义层次，并在一定程度上解决语义异构问题，在信息互操作、知识理解和信息集成等领域具有很大的应用前景，它的研究成果对军队信息化建设也会产生一定的指导意义。本体作为语义 Web 中核心基础元素，已经广泛应用于语义数据集成、Web 服务等领域。然而，由于本体定义的自发性，本体领域的多样性，导致互联网上本体具有分布性和异构性，极大限制了本体数据共享与集成。为了实现基于本体的语义互操作，就必须建立异构本体中元素之间的映射关系,这一过程称为本体映射。但是，由于本体引人人工智能、语义网和信息系统等领域的时间较短，本体的建模方法也初步确立，本体构建方法仍然不成熟。本文主要对目前的各种本体构建方法和构建工具进行了分析比较，最后对这些本体构建方法和工具进行了总结和展望。

## 2 相关概念

关于本体的形式化定义很多,本文采用文献[2]提出的定义形式.

定义 1：本体主要包括概念(Concepts)、关系(Relations)、实例(Instances)以及公理(Axioms),可表示为 $O=(C,R,I,A^O)$；其中,C 表示概念集合;R 表示关系集合;I 分别表示实例集合; $A^O$ 表示公理的

集合。

我们采用[3]的描述来给出本体映射的形式化定义。

定义 2：设 $M=(O_1, O_2, F_{Map})$为本体 $O_1$ 到本体 $O_2$ 的映射，$O_1$ 为源本体，$O_2$ 为目标本体。$F_{Map}$ 为建立 $O_1$ 到 $O_2$ 间本体映射的映射函数。定义为

$$F_{Map} : \{e_{i1}\} \rightarrow \{e_{i2}\}. \quad (1)$$

其中：$\{e_{i1}\}$、$\{e_{i2}\}$ 分别为 $O_1$，$O_2$ 中元素集合。不失一般性，本文只考虑本体概念之间的映射，这是多数本体映射任务所针对的问题。

# 3 本体的构建

本体构建是本体应用的基础，是实现信息交换、共享，解决语义冲突的基础，通过构建统一的术语和概念，实现知识共享，为异构系统间的通讯提供共同的词汇，便于它们之间的互操作和集成。本体构建是一项庞大的系统工程，需要各领域的专家（领域专家、本体工程师等）按照一定的本体构建原则，在合理方法论的指导下，采用合适的关键技术或使用便捷的本体开发工具加以实现。

## 3.1 本体构建的原则

研究人员从实践出发提出了许多指导本体构建的原则，然而，目前仍没有构造本体的统一标准，一般采用 1995 年 T．R．Gruber 提出的指导本体构造的五条原则[4]，具体如下：

1)清晰：领域本体必须能有效地说明所定义术语的含义。定义应该是客观的，与背景独立的；当定义可以用逻辑公理表达时，它应是形式化的，应尽力用逻辑公理表达；定义应该尽可能的完整；所有定义应该用自然语言加以详细说明。

2)一致：领域本体应该是前后一致的，也就是说，它应该支持与其定义相一致的推理。领域本体所定义的公理以及用自然语言进行说明的文档都应该具有一致性。假如从一组公理中推导出来的一个句子与一个非形式化的定义或者实例矛盾，则这个领域本体是不一致的。

3)可扩展性：领域本体的可扩展性是指其提供一个共享的词汇，这个共享的词汇应该为预期的任务提供概念基础。它应该可以支持在已有的概念基础上定义新的术语，以满足特殊的需求，而无须修改已有的概念定义。也就是说，人们应该能够在不改变原有定义的前提下，以这组存在的词汇为基础定义新的术语。

4)编码偏好程度最小：领域本体与特定的符号即编码无关。也就是说，领域本体的表示形式的选择不应该只考虑表示或实现上的方便，概念的描述不应该依赖于某一种特殊的符号层的表示方法，不能依赖于某种确定的语言，因为实际的系统可能采用不同的知识表示方法。

5) 本体约定最小：本体约定应该最小，只要能够满足特定的知识共享需求即可。也就是说，本体应该对所模拟的事物产生尽可能少的推断，而让共享者自由地按照他们的需要去专门化和实例化这个本体。Gruber 还指出，由于本体承诺是以词汇的使用为基础的，因此可以通过定义约束最弱的公理以及只定义应用所需的基本词汇来保证。

上述五条原则给出了构造领域本体的基本思路和框架，但是明显的不足之处就是它们所反映的内容较模糊且难于把握。实际本体构建过程中，以上五原则甚至可能有不一致的情况，本体开发者需要权衡各原则，必要时可能还要参照其他原则，需要灵活运用本体构建原则才能构建高质量的本体。

## 3.2 本体构建的知识来源

目前的本体构建方法主要分为本体论工程方法和将叙词表转化为本体的方法两大类[5]。显然，后者的知识来源是该专业领域的叙词表，丽本体论工程的知识来源则相对丰富，主要有表格、主题词表、数据库、WordNet、WEB 网、领域专家、文本和 W 憾等等。依据知识的结构化程度可以将它们分为两类：结构化知识来源，包括主题词表、表格、WordNet、数据库；非结构化知识，包括 WEB 网、领域专家、文本和 Wiki 等.运用结构化知识最大的优点就是便于半自

动或自动化构建本体，大大提高了构建速度。而运用非结构化知识构建本体往往需要大量人工参与，虽然本体构建质量较好，但是耗时较长。虽然运用自然语言处理学科知识可以对非机构化知识进行一定程度的自动处理，貌似可以提高本体构建的速度，但同时该学科的不成熟导致了本体质量的下降。

# 4 本体构建方法

## 4.1 本体构建方法

本体构建的方法学还没有成熟的理论作指导，而目前的本体构建方法都是针对具体的项目提出的，这就导致各种本体构建方法的出现。

国外主要的构建方法有骨架法、IDEF5 法、TOVE 法、METHONLOGY 法、KACTUS 法、七步法和 SENSUS 法等，其成熟度依次为七步法>METHONLOGY 法>IDEF5 法>TOVE 法>骨架法>SENSUS 法>KACTUS 法[5]，下面主要介绍前面五种相对成熟的方法。

七步法[5]是斯坦福大学医学院提出的基于 Protege 本体构建工具的一种领域本体构建方法。一共包括 7 个步骤，因此被称为七步法：1)确定知识本体的专业领域和范畴；2)考查复用现有知识本体的可能性；3)列出本体中的重要术语；4)定义类(Class)和类的等级(层次)体系；5)定义类的属性；6)定义属性的分面(Facets)；7)创建实例。

METHONLOGY 方法[1,5]是由西班牙马德里理工大学 AI 实验室提出的。该方法结合了骨架法和 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。METHONLOGY 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。

IDEF5[5,6] 法是美国 KBSI(Knowledge BasedSystems Inc．)公司开发用于描述和获取企业本体时所采用的一种结构化的本体开发方法。IDEF5 通过使用图表语言和细节说明语言，获取关于客观存在的概念、属性和概念间关系，并将它们形式化，作为知识本体的主要架构。IDEF5 的本体构建方法流程如下：1)组织和范围：确定本体项目的目标、观点和语境，组织课题队伍并为组员分配角色；2)数据收集：收集本体建设需要的原始数据；3)数据分析：分析数据，为抽取本体做准备；4)知识本体的初步开发：从收集的数据当中建立一个初步的本体；5)本体的精炼与验证：完成本体建设过程。

TOVE 法[5]，也称为评价法，是 Gruninger 和 Fox 等开发 TOVE 工程本体(关于商业过程和活动建模的本体)的经验总结。这种方法并非直接构建以本体形式描述的知识的逻辑模型，而是先建立本体的非形式化描述说明，然后将这种描述形式化。这种方法的本体构建基本流程如下：1)激励情节的获取。Gruninger 和 Fox 认为本体开发是由应用中的具体情节所驱动的。获取激励情节就是定义直接可能的应用和所有解决方案，提供潜在的非形式化的对象和关系的语义表示；2)非形式化能力问题的明确表达。将系统能力问题(能够回答)作为约束条件，包括能解决什么问题和如何解决，这里的问题用术语表示，答案用公理和形式化定义回答。由于是在没有形式化本体之前进行的，所以叫非形式化的能力问题；3)术语的规范化。从非形式化能力问题中抽取非形式化的术语，然后用本体形式化语言进行规范化定义；4)形式化能力问题的明确描述。一旦本体内的概念得到了定义，能力问题就脱离了非形式化，演变为形式化的能力问题；5)将规则形式化为公理。术语定义所遵循的公理用一阶谓词逻辑表示；6)调整能力问题解决方案的条件，从而使知识本体趋于完备。

骨架法[1,5]，也称为 EO 工程法，是 Uschold 和 King 在 1995 年开发 EO(Enterprise Ontology，关于企业建模过程的本体，是相

关商业企业间术语和定义的集合)中的经验总结，它提出了一种本体开发的具体步骤，其基本步骤如下：1)明确本体应用的目的和范围；2)构建本体；3)本体评价；4)本体成文。使用骨架法开发的最重要的本体就是EO，该本体在爱丁堡大学的人工智能应用研究所以及IBM、Lloyd’S Register，Logica UK Limited，和Unilever等合作单位共同开发完成。骨架法清晰地描述了本体开发的具体实现步骤，对于当前本体开发实践具有重要指导意义。

此外，我国研究学者，如李景[7,8]、董慧、刘柏嵩、唐爱民等，在借鉴国外本体构建方法的基础上，根据中文汉语本体构建的实际情况，也提出一些具有影响的本体构建方法。

### 4.2 本体构建方法存在的问题

尽管国内外一些本体构建方法在相应的项目中比较适用，但通过对各方法的熟悉与对比之后可以发现这些方法仍然存在许多问题，例如：

1) 大多数方法不是通用的领域本体构建方法，仅适用于较小专业范围的本体构建，如骨架法是在企业本体开发中总结出来的，它对通用本体开发的指导作用就很有限。

2) 动化程度不高，大多数方法还是运用人工开发，耗费大量人力、物力和财力，开发效率不高。

3) 建设过程缺乏规范性，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。

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。

5) 大部分都是从各自的实践经验出发，勾勒出了本体建模的过程、方法和步骤的轮廓，很多都只提供了建模过程的指导原则，但是却缺少对本体建模的指导原则等进行落实的、可操作性强的方法。

6) 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。

## 5 体构建工具评述

### 5.1 本体构建工具

随着本机机制研究的逐渐深入，越来越多的本体开发活动在国内外陆续开展。然而，本体开发是一项庞大的知识工程，研究人员在采用上述方法构建本体的过程中遇到了各种问题，如一致性检查、本体展示等等，人们迫切希望产生一些工具帮助其完成本体开发任务。在这种情况下，本体构建工具应运而生，各研究单位都试图开发适合特定领域本体构建的环境，以支持本体开发过程中的多个环节。借助这些工具，本体构建者可以把精力集中在本体内容的组织上，而不必了解本体描述语言和描述方式等细节，极大地方便了本体的构建。目前，在国外已经出现了众多的本体构建工具，典型的包括WebOnto、WebODE、KAON和Protege等。

WebOnto[6]起源于英国Open University开始于1997年的KMI项目，目的是开发一个基于Web的本体编辑器。它能提供比Ontolingua更为复杂的浏览、可视化和编辑能力；基于OCML推理引擎的知识模型，提供多重继承、锁机制，支持用户合作地浏览、构建和编辑本体；但是WebOnto没有提供源代码。

WebODE[6]是西班牙马德里技术大学开发的一个综合性的本体建模工具，它集成了本体开发过程中的大多数行为，支持METHONLOGY本体构建方法论，目前只有WebODE和OntoEdit能够将本体开发环境和实际的本体构建方法相对应。WebODE支持构建知识层次的本体，并可以将其转化为不同的本体语言加以描述。它不同于OntoEdit和Protege的插件结构体系，而是采用客户机/服务器模式的体系结构，通过Java、RMI、COBRA、XML等技术实现，具有较高的可扩展性和可用性，允许添加新的服务；使用webODE构建的本体以sQL数据库的形式存

储，对于大规模本体来说具有较高的执行效率；通过定义实例集来提高概念模型的可重用性；支持多重继承、类型一致性、数值一致性、集合基一致性检查，并且提供了分类一致性验证机制。

KAON 是德国 Karlsruhe 大学编制的一套用于语义网和本体研究的工具，包含各种模块用于本体的构建、存储、检索、维护以及应用，其中 OI-Modeler 是 KAON 模块集中的本体建模工具，可便捷的实现本体的创建和维护。

Protege 是斯坦福大学为知识获取而开发的一个工具，主要应用于知识的获取以及现存本体合并和排列，可以免费下载并公开源代码，再加上其支持中文，Protege 已经成为目前国内使用最为广泛的本体编辑工具和基于知识的框架 Protege 主要具有以下特征：

1)可扩展的知识模型能够使用户重新定义原始知识集合；

2)友好的本体导人导出功能，可以从 RDFS、带 DTD 的 XMI．文件、XMI。Schema 等文件中导入本体，也可以将本体转化为多种形式化语言描述，如 RDF(S)、OWL 等。

3)具有强大的功能插件体系和开放的模块化风格。基于开放式组件的体系结构使系统开发者可以通过生成恰当的插件以增加新的功能。

4)提供一个半自动化工具 PROMPT 用于自

动地执行本体的合并和排列。

5)有友好的开发界面。

6)Protege 平台支持两种类型的本体建模嘲：(1)Protege-Frames 编辑器用于构建基于框架的本体，目前最新的版本是 Protege3．4．5，发布于 2011 年 3 月 18 日。在这种模型中，本体是由具有层次结构的类集合组成，类的槽(slots)集合表示概念的属性和关系；类的实例集合则表示概念的具有特定属性值的个体；(2)Protege-OWL 编辑器则用于构建应用于语义网的本体。目前最新版本为 Protege4．1，该版本全面支持 OWL2．0 语言，专门使用 W3C 的 OWL 语言描述，一个 OWL 本体包含类的描述、属性以及实例。

除此之外，还有 Apollo、LinkFactory、OILEd、Ontolingua、OntoSaurus、OpenKnoME 等本体构建与管理工具。这些本体开发工具功能各不相同，对于本体语言的支持能力、表达能力、逻辑支持能力以及可扩展性、灵活性、易用性等都相差甚远。就目前而言，在国内 Protege 和 KAON 的使用最为广泛。

### 5.2 体构建工具存在的问题

尽管目前本体构建研究炙手可热，本体构建工具也多种多样，但是对比之后可以发现，这些工具存在如下问题：

1) 本体工具的多样化和差异性阻碍了不同领域知识的联通和异构系统的互操作。

2) 构建工具不为用户提供通用概念/类的体系，可能使得用户大量时间花在通用概念的构建上，大大降低了本体构建效率。

3) 每种构建工具都有不同的导入/输出格式，缺乏统一的标准和规范，使得不同工具构建的本体之间无法兼容，在异构系统中无法被复用。

4) 大多缺乏对中文的支持，使得国内研究人员在中文本体构建上进展缓慢。

5) 许多工具不支持协作开发，这使得目前构建的本体中掺杂许多个人主观意见，降低了本体的质量。

6) 一些本体工具界面不够友好，降低了本体开发效率。

## 6 基于数据场的本体映射

基于数据场的本体映射算法的基本思想是：首先确定本体中的相关区域，然后再相关区域间建立映射。具体过程为：使用高效的相似度计算方法计算任 2 个本体概念的相似度，并在此基础上得到本体概念与另一个本体的相关度。由于，高效的相似度方法在处理较复杂的本体时准确性往往不足。因此，一般需要使用基于结构的方法，通过周围的概念修正这种错误。当映射本体的规模较大时，这种基于结构的方法必须比较节省资源，本文中使用基于数据场的方法，通

过临近概念间相关度的传播修正初始相关度中的错误。完成最终相关度的计算后，跟据相关度的值选择相关概念并抽取相关的子本体。最后，利用更有针对性的映射方法（此类方法往往比较消耗时间）实现抽取后相关子本体间的映射。

映射算法描述

**算法 1.** OntologyMatching ($O_1$ ,$O_2$):
输入: 参与映射的本体 $O_1$ 和 $O_2$.
输出: 映射结果 $R_M$.
1. Relevance ($O_1$, $O_2$);
2. Choose($O_1$)→$O_1$'； Choose($O_1$)→$O_1$'；
3. Match($O_1$', $O_2$')→$R_M$。

其中，Relevance ($O_1$，$O_2$ )方法用于计算概念 $c_{1i}\in O_1$ 与 $O_2$ 和 $c_{2j}\in O_2$ 与 $O_1$ 的相关度 $r_{1i}$ 和 $r_{2j}$；Choose(O 1 )和 Choose(O 2 )方法利用第 1 步得到的相关度及数据场势函数，计算并抽取 $O_1$ 中与 $O_2$ 和 $O_2$ 中与 $O_1$ 的相关区域本体 $O_1'$ 和 $O_2'$；Match($O_1'$ , $O_2'$ )针对性选择本体映射算法建立 $O_1'$ 与 $O_2'$ 的映射关系，并将结果存入 $R_M$。

## 7 结语

本文首先对目前几种主要的本体构建方法和构建工具进行了介绍与比较，分别指出了存在的问题。针对这些问题，笔者认为：在目前本体评价标准尚不成熟的情况下，为了高效地开发高质量的本体，在方法上，未来本体构建方法发展的方向是自动化、半自动化并且具有严格可操作性的本体构建方法；在本体描述语言上，将由现在的百花争鸣朝着规范化标准化的方向发展；在构建工具上，应该具备以下特点：界面友好，易于使用；能够跨平台，支持多人协作开发；支持功能插件的扩展；提供统一的通用概念常识库；支持多语种；具备强大的推理能力；能支持本体生命周期的大部分开发过程。
本体构建的方法学还没有成熟的理论作指导，现有的本体构建方法参差不齐。对于本体构建方法的使用者来说，应根据现存的方法及其适用范围，选择适合特定本体构建的方法；或借鉴这些方法的框架和步骤，总结出适合自己的方法。在本体评估方面，对本体没有一个具体的评估标准是本体构建的一个瓶颈问题，这也是本体构建方法以后要重点研究的方面。

**参考文献：**


[1]冯志勇，李文杰，李晓红．本体论工程及其应用[M]．北京：清华大学出版社，2007.5

[2]Dean M., Schreiber G., Bechhofer S., Harmelen F.v., Hendler J., Horrocks I., McGuinness D. L., Patel-Schneider P. F. and Stein L. A. OWL web ontology language reference. 2004.

[3]Tang J. , Li J. , Liang B. , Huang X. , Li Y. and Wang K.Using Bayesian Decision for Ontology Mapping . Web Semantics , 2006, 4(4): p. 243-262

[4]T．R.Gruber．Towards Principles for the Design of Ontologies Used for Knowledge Sharing[J]．International Journal of Human Computer Studies，1995

[5]李勇，张志刚．领域本体构建方法研究[J]．计算机工程与科学，2008(5)：129～131

[6]杜文华．本体构建方法比较研究口]．情报方法，2005(10)：24～25

[7]刘宇松．本体构建方法与开发工具研究[J]．现代情报，2009，29(9)：17～24

[8]景．主要本体构建工具比较研究-上[J]．信息系统，2006(1)：109～111