# 基于文本挖掘的本体学习方法研究

于　娟/著



科学出版社

# 基于文本挖掘的本体学习方法研究

于 娟 著

科学出版社
北 京

## 内 容 简 介

本体将客观世界的知识抽象和形式化为人机都可读的符号，使人机通过认知符号世界实现对客观世界的共同理解。本体是知识图谱的核心，为其提供逻辑推理基础；知识图谱是基于本体所构建的应用知识库，是语义服务的基础技术。本书针对手工构建本体和知识图谱的困难，首先系统地介绍了本体学习的理论基础，然后讨论了一系列基于文本挖掘的本体学习方法的研究和应用，包括：从文本数据源自动提取术语的方法，基于术语集合自动获取本体概念集合和本体关系集合的方法。这些方法尤其适合轻量级本体的自动构建和知识图谱的知识抽取。

本书可以作为本体学习和知识图谱研究的基础读物，也可供从事文本挖掘和知识管理的科技人员阅读参考。

**图书在版编目（CIP）数据**

基于文本挖掘的本体学习方法研究 / 于娟著. —北京：科学出版社，2019.12

ISBN 978-7-03-059881-3

Ⅰ. ①基… Ⅱ. ①于… Ⅲ. ①学习方法–研究 Ⅳ. ①G442

中国版本图书馆 CIP 数据核字（2018）第 273559 号

责任编辑：杭 玫/ 责任校对：贾娜娜

责任印制：张 伟/ 封面设计：无极书装

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http：//www.sciencep.com

北京盛通商印快线网络科技有限公司 印刷

科学出版社发行 各地新华书店经销

*

2019 年 12 月第 一 版 开本：720×1000 1/16

2019 年 12 月第一次印刷 印张：10 1/2

字数：212 000

**定价：84.00 元**

（如有印装质量问题，我社负责调换）

# 作 者 简 介

于娟，福州大学副教授、硕士生导师。大连理工大学管理学院工学博士，信息管理与信息系统专业，长期从事知识管理中的文本挖掘和本体研究。主持和参与多项国家级和省部级科研项目。在国内外核心学术期刊发表学术论文二十余篇。获“福建省高校杰出青年科研人才”、福州大学“旗山学者”等荣誉称号。

# 前　言

本体（ontology）的概念起源于哲学领域的本体论（Ontology）。在信息相关领域中，本体是指共享概念模型的明确的形式化说明。本体的目的是将客观世界的知识抽象和形式化为人机都可读的符号，使人机通过认知符号世界实现对客观世界的共同理解。2000 年开始，本体被引入到 Web，为网页添加语义标注，将原 Web 改造为语义 Web。语义 Web 的目标是实现机器自动处理信息，提供如信息代理、语义搜索等智能服务。从此，本体技术得到了人工智能、信息检索、知识管理等领域学者的广泛关注和研究。

近年来，本体有了一个新变种——知识图谱（knowledge graph）。知识图谱旨在描述客观世界的概念、实体、事件及其之间的关系，一经提出即得到了广泛关注和重点研究，被用于语义搜索、智能问答、个性化推荐等。知识图谱与本体密切相关：本体是知识图谱的核心，为其提供逻辑推理基础；知识图谱是本体的实例化，是基于本体所构建的应用知识库，是语义服务的基础技术。当然，本体和知识图谱之间并没有明显的区分界线。本体研究成果是知识图谱研究和应用的一个良好基础；知识图谱的兴起也使本体技术得到更多关注和应用研究。

构建和维护本体其实是一项烦琐而困难的任务，若仅靠领域专家人工构建和维护本体的话，耗时耗力且可能随概念和关系的增多而出错。为了加快本体构建，保证基于本体的应用顺利实施，出现了从已有的数据资源中自动获取本体的思路，即本体学习（ontology learning）。本体学习是指采用机器学习技术，从数据源中（半）自动地获取本体对象，以支持本体构建、提高构建速度并优化结果的过程。其中，本体对象主要是指本体概念和本体关系；数据源一般采用大量显性化的知识源，通常是领域文本或 Web 网页；所采用的机器学习技术通常是文本数据挖掘技术。

基于文本挖掘的本体学习方法和技术，能够实现本体构建过程的一定程度的自动化，以人机结合的方式构建本体，降低本体构建和维护的困难程度，提升速

度和效果。本体学习技术可减少因概念和关系等本体对象增多使得人工无法全面掌握所导致的不一致。再者，通过本体学习技术量化一些主观的指标（如一个概念是否领域专有，两个概念间是否存在相关关系等）能使所构建的本体更客观，能够减少因领域专家知识结构等主观因素所造成的争议，提高本体的通用性。可见，本体学习技术能够有效地支持本体构建。因此，本体学习技术对基于本体的人工智能、语义 Web、知识管理等领域的发展具有重要意义。

本书定位为本体学习和知识图谱研究的基础读物，也可供从事文本挖掘和知识管理的科技人员阅读参考。本书系统地梳理和总结基于文本挖掘的本体学习方法，介绍其基本原理及应用。第 1 章概述本体和知识图谱的定义、构成、分类、应用及二者之间的关系，列举现有的本体和知识图谱。第 2 章系统介绍本体构建与本体学习的理论基础，以及与之相关的本体维护、本体集成和本体管理系统（ontology management system，OMS）。第 3 章到第 5 章分三步介绍从文本数据源自动获取本体对象的主要方法技术：文本术语提取、本体概念学习、本体关系学习。其中，第 3 章综述了从自然语言文本中自动提取术语的方法和技术，涵盖了当前国际上几种主要语言的文本分词与词形规范化工具。第 4 章的本体概念学习介绍基于术语集合自动获取本体概念集合的方法和技术，分为领域专有术语识别和同义词识别两部分。第 5 章的本体关系学习介绍基于本体概念集合自动获取本体关系集合的方法和技术，主要从语义层面构建概念的模型，然后计算概念相关关系。第 6 章介绍本体的两方面应用：基于本体学习的领域术语标准化和基于本体的人物关系推理。

本书所介绍的基于文本挖掘的本体学习方法和技术尤其适合轻量级本体的自动构建和知识图谱的知识抽取。依据需求分析确定需要构建本体的领域范围后，在给定合适的文本数据源的情况下，按照顺序执行以下三个步骤就能以人机结合的方式构建面向应用的本体：①采用本书第 3 章所述的术语提取方法技术提取得到文本数据源中的术语集合；②基于该术语集合，采用本书第 4 章所述的本体概念学习方法技术构建本体的概念集合；③基于该本体概念集合，采用本书第 5 章所述的本体关系学习方法技术构建本体关系集合。上述步骤所生成的概念模型，经领域专家人工修正就得到了所需的本体。用户即可基于该本体研发相关应用系统。

在本书的撰写过程中，大连理工大学系统工程研究所的党延忠教授给予了无私的指导和重要的建议。在此，向党教授表示诚挚的敬意和衷心的感谢！同时，福州大学经济与管理学院的研究生曹晓、施文洁、王建文、黄恒琪等做了大量辅助工作。在此，对这些同学的辛勤工作表示衷心的感谢。

笔者在研究本体学习以及撰写本书的过程中，阅读和参考了大量国内外教材、专著、论文等资料，并有选择地把一些重要知识纳入本书。在此，向这些资

料的写作者表示衷心的感谢。

本书是笔者从事本体研究工作十余年的总结。将以此书作为今后工作的基础和新起点。由于笔者能力有限，本书难免存在不足之处，恳请读者批评指正。

# 目　录

# 第1章 概 述

本体的概念起源于哲学领域的本体论。本体论可追溯至亚里士多德提出的分类学（category）。分类学是研究世界上存在的万事万物的分类。19世纪，德国的哲学家首先提出了ontology的概念，用于区分自然科学中“怎么样”和“为什么这样”的研究，也就是万事万物的存在和存在本质的学问。ontology由两个希腊词ontos和logos构成，ontos是存在的意思，logos是词的意思。

20世纪80年代，本体被研究人工智能的学者引入信息领域，作为一种形式化客观世界的知识表示方法。2000年的XML[①]国际会议上，万维网的发明人Tim Berners-Lee提出将本体作为语义Web的基础技术，为Web网页添加语义标注，以支持语义服务。从此，本体受到了广泛关注，在人工智能、信息检索、知识管理等领域得到越来越多的实际应用。在信息领域，本体的目的是将客观世界的知识抽象和形式化为人机都可读的符号，使人机通过认知符号世界实现对客观世界的共同理解。

近年来，新兴的知识图谱技术进一步促进了本体的应用研究。知识图谱的命名源于2005年在美国成立的Metaweb公司。该公司基于语义Web的理念，从维基百科、美国证券交易委员会等公开数据集中，自动获取现实世界中的实体（人或事物）及其之间的关系，然后以图结构存储在计算机中，为Web语义服务提供开放共享的知识库（knowledge base，KB）。2010年，谷歌收购Metaweb，获得了其语义搜索技术，并于2012年提出了知识图谱的概念。知识图谱与本体密切相关。本体是知识图谱的核心，知识图谱是本体的实例化。自此，本体技术开始了飞速发展并得到了广泛应用。

① XML：extensible markup language，可扩展标记语言。

# 1.1 本　体

## 1.1.1 本体的定义

1980年，人工智能领域的学者、图灵奖的获得者John McCarthy认识到哲学本体论与人工智能的逻辑理论构建活动之间的交叉，提出以逻辑概念为基础的智能系统必须列出所有存在的事物，并构建一个本体描述我们的世界。1984年，John F. Sowa提出了概念图知识表示方法，用于描述客观世界的事物、事物间的联系以及相互影响的方式。概念图是一种有限连接二部图，图中结点表示概念，概念间关系，具体对象（如猫、电话、饭店）或者是抽象对象（如爱、美丽、忠诚）。每个概念图表示一个命题，基于概念图建立的知识库中包含大量这样的图，可以说，概念图就是本体的雏形。

1993年，Gruber（1993）给出了本体的定义：An ontology is a formal, explicit specification of a shared conceptualization，即本体是共享概念模型的明确的形式化说明，该定义得到了广泛的认可。Guarino和Studer等深入分析了此定义，给出了本体的解读（Guarino and Giaretta，1995；Studer et al.，1998）：

（1）概念模型（conceptualization）：是指通过识别客观世界中某些现象或事实的相关概念而得到关于该现象或事定的抽象模型。

（2）明确的（explicit）：是指概念的类型及其使用约束都有明确的定义。例如，定义医学领域的两个概念——疾病和症状，须定义二者之间的关系是引发关系，并且疾病不能引发自身。

（3）形式化的（formal）：是指本体是机器可读的，不是自然语言描述的。

（4）共享（shared）：本体中体现的是共同认可的知识，它面向团体使用而非个人私有。

2000年，本体被引入Web中，为Web网页添加语义。由于第一代Web存在先天不足，只能按照某种格式进行信息显示，无法表达语义，导致Web成为非结构化的庞大的信息容器，计算机无法将Web信息自动生成为知识，也无法进行精确的Web信息检索，语义识别还需要人工来判断。因此，2000年的XML会议提出了语义Web的理念。语义Web以文本为基础，为Web网页添加语义标注，以支持机器自动处理信息，提供诸如信息代理、搜索代理、信息过滤等智能服务。随后，W3C（World Wide Web Consortium，万维网联盟）在2001年2月推出Semantic Web Activity。从此，本体得到了信息与知识管理等领域的广泛研

究，发展迅速。

本体在不同领域中的定义和应用不尽相同。在信息技术领域，本体的目的是一致的：将客观世界的知识形式化为机器可读的符号世界，实现人机对客观世界的共同理解。为此，本体确定客观世界中共同认可的概念词语（即术语），并明确定义这些概念之间的相互关系。

### 1.1.2　本体的构成

本体的主要构成对象（objects）有概念（concept）、概念间类属关系（hierarchical relation 或 taxonomic relation）、非类属关系（non-taxonomic relation 或 non-hierarchical relation）、规则（rule）［也称公理（axiom）］。一个完整的本体可形式化为一个四元组：$O:=\{C, R_h, R_c, A\}$，其中，$C$ 是概念的集合，每一概念由表示概念的语词（一个术语或一组同义术语）和概念内涵的自然语言定义组成；$R_h$ 是定义在 $C$ 上的概念间类属关系的集合，描述概念间的上下位关系，即“subClassOf”或“is-a”关系；$R_c$ 是定义在 $C$ 上的非类属关系的集合，有时会表现为概念的属性；$A$ 是规则（或称公理）集合，对概念和关系的定义施加约束，可用于推理出本体中没有显性说明的知识。

那么，本体是否还应包含实例呢？笔者认为，包含了实例的本体更适合称为知识库或知识图谱。知识库使用本体中的概念和关系来描述领域中的现象和事实。例如，医药本体可能包含“白血病”“皮肤病”等术语的定义，但不应包含对某一具体病人患某一疾病的诊断声明，而这是知识库所要表达的内容。当然，本体也可以包含一部分通用的实例，本体与知识库或知识图谱之间并没有清晰的界限，二者的区别仅在于知识库的哪一部分是可共享和重用的，哪一部分是针对特定应用的，而这种区别还会随具体应用而改变。

### 1.1.3　本体的分类

按照本体的通用性可以将本体划分为通用本体（generic ontology）、核心本体（core ontology）和领域本体（domain ontology）。

通用本体又称上层本体或者顶层本体，定义多领域通用的概念，如状态、事件、过程、动作、组件等。

核心本体是定义在几个领域中通用的概念，介于通用本体和领域本体之间，用于帮助领域本体互操作。核心本体与通用本体之间的界限没有明确的定义。

领域本体是某一领域（或某一应用）的共享概念模型的明确的形式化说明。其中，概念模型指的是领域概念以及概念之间的关系；领域指的是对某一科目分

类划分后的范围，如科学研究的学科范围、商业活动的范围、技术开发的范围、业务工作的范围等。一个领域本体描述了该领域的知识模型，提供了对该领域知识的共同理解；确定了领域内共同认可的词汇，统一了领域概念的术语并确定了概念之间的关系；捕获并形式化了领域知识，澄清了领域的知识结构。因此，领域本体是领域知识的知识分类器，能够统一地标识和分类显性知识。领域本体中的概念是通用本体和核心本体中的概念的特殊化。因为核心本体也能够在某一个领域中达到通用，所以领域本体与核心本体之间的界限没有明确的定义。

按照表示能力可以将本体分为轻量级本体（lightweight ontology）和重量级本体（heavyweight ontology）。

轻量级本体主要是指仅包含概念和概念间类属关系的概念层次树，有些轻量级本体还包含部分的非类属关系。轻量级本体几乎不定义规则，对于概念间的关系也仅定义与结构相关的关系。轻量级本体与重量级本体之间并没有明确界限。

重量级本体包含大量本体规则，能显性表示本体中的约束，支持强推理能力。每一个重量级本体都可以对应生成一个轻量级本体。

每一个本体都可以根据其通用性和表示能力分属于上述两种分类。例如，WordNet 是一种轻量级的通用本体。

### 1.1.4　本体的应用

本体的功能主要有三方面（Uschold and Gruninger，1996；Uschold and Jasper，1999）：①为人与人之间或组织与组织之间的交流提供共同的词汇，帮助不同知识背景的人们进行沟通。②将客观世界的知识抽象和形式化为人机可读的符号，使人机通过认知符号世界实现对客观世界的共同理解。③在不同的建模方法、范式、语言和软件工具之间进行翻译和映射，实现不同应用系统之间的交互。

本体的这些功能吸引了包括人工智能、语义 Web 和知识管理等领域的广泛关注，在这些领域发挥着越来越重要的作用。不同领域对本体的需求不尽相同，应用本体的方法也不尽相同。表 1.1 总结了本体的主要应用。

**表 1.1　本体的主要应用**

| 应用领域 | 人工智能 | 语义 Web | 知识管理 |
| --- | --- | --- | --- |
| 本体功能 | 知识表示与知识推理 | 网页的语义标注 | 知识索引 |
| 本体实例 | 概念和关系的实例 | 网页等各种网络资源 | 文本、数据库模式、多媒体文件、领域专家等 |
| 应用方式 | 知识推理 | 网络资源管理 | 知识检索与共享 |
| 研究侧重 | 智能问答、多智能主体系统 | 语义搜索、电子商务 | 知识推送 |

人工智能领域从20世纪60年代出现以来，瓶颈问题就是没有一种适合多方面应用的知识表示方法。支持较强推理的谓词演算（Chang and Lee，1973）的表示能力较差；表示能力较强的语义网络（Quillian，1967）、框架（Minsky，1975）和概念图（Sowa，1984）等非逻辑的知识表示方法的推理能力较差，且基于这类方法的应用系统间的差异较大，几乎无法进行交互。在人工智能领域，本体是一种兼具丰富知识表示机制和强推理机制的知识表示方法，被认为是解决这一瓶颈问题的有效途径。

语义 Web 引入本体的目的是对网络资源添加语义注释以实现语义搜索等服务。语义 Web 针对上一代 Web 中网页的无结构性问题，基于本体对网页内容添加标注和对网页进行分类，以实现 Web 服务的智能化。在语义 Web 中，本体的实例是带有 RDF（resource description framework，资源描述框架）的 XML 网页。本体的引入将原本非结构化的 Web 改造成为能够实现机器自动处理信息，提供诸如信息代理、搜索代理、信息过滤等智能服务的语义 Web。正是由于语义 Web 将本体作为基础技术，才使得本体得到广泛关注和深入研究，也因此引发了知识图谱的研究和应用。

在知识管理领域，本体是知识库的构建基础。知识库包含大量显性知识，存储为异构数据，包括文本、数据库、多媒体文件等。传播与共享最普遍的形式是通过人的交流和纸面上的交流进行，但是效率较低；通过信息网络，利用电子媒体进行传播与共享是最快且最容易进行的方式（王众托，2016）。因此，首先需要对这些不同类型的知识采用统一的规范化的表示机制进行分类。本体在知识管理领域扮演着知识分类器的角色，能够对这些显性知识进行统一管理。基于本体的知识库能够方便地对知识按内容检索，从而促进知识的共享和重用。

## 1.2 知 识 图 谱

知识图谱是以图的形式表现客观世界中的实体（人、物、事件）及其之间关系的知识库。2012 年，谷歌提出了知识图谱的概念。自此，知识图谱得到了广泛关注和应用研究，现已成为语义搜索、智能问答、管理决策支持等智能服务的基础技术之一。知识图谱与本体密切相关，本体研究成果是知识图谱研究和应用的一个良好基础，本体技术也因知识图谱的兴起而得到了更多关注和应用研发。

### 1.2.1 知识图谱的定义

知识图谱是以图的形式表现客观世界中的实体（人、物、事件）及其之间关系的知识库。知识图谱的研究起源于语义 Web。在 2000 年的 XML 大会上，Tim Berners-Lee 提出了基于本体技术的语义 Web 的理念，目标是为 Web 网页添加语义，以提供诸如信息代理、搜索代理、信息过滤等语义服务。之后，Metaweb 公司成立，致力于研发用于 Web 语义服务的开放共享的世界知识库。2010 年，谷歌收购 Metaweb 并获得了其语义搜索技术，于 2012 年提出了知识图谱的概念。

本体是语义 Web 的底层技术，用于为 Web 网页添加语义。语义 Web 理念中的本体与知识图谱二者密切相关。本体描述概念及概念间的关系，是大多数知识图谱的模式层，是知识图谱的概念模型和逻辑基础。知识图谱与本体的相同之处在于：二者都通过定义元数据以支持语义服务。不同之处在于：知识图谱更灵活，支持通过添加自定义标签划分事物的类别；本体侧重概念模型的说明，能对知识表示进行概括性、抽象性的描述，强调的是概念与概念之间的关系。大部分本体不包含过多的实例，本体实例的填充通常是在本体构建完成以后进行的。知识图谱更侧重描述实体关系，在实体层面对本体进行大量的丰富与扩充。可以认为，本体是知识图谱的抽象表达，描述知识图谱的上层模式；知识图谱是本体的实例化，是基于本体的知识库。

知识图谱采用三元组描述事实，所使用的描述语言（description language）大多是本体语言，如 RDFS（resource description framework schema）、OWL（Web ontology language，Web 本体语言）等。知识图谱也可以通过 RDFS 或 OWL 定义规则用于知识推理。知识图谱的关键技术也与本体很相似，涉及：①知识图谱构建阶段的实体抽取、关系抽取、语义解析等机器学习和自然语言处理方法技术；②用于知识图谱存储的知识表示、数据库、知识融合等方法技术；③知识图谱应用阶段的数据集成、知识推理等。

除了本体之外，与知识图谱相关的概念还有知识地图和科学知识图谱。知识地图将特定组织内的知识索引通过“地图”的形式串联在一起，揭示相关知识资源的类型、特征以及相互关系（Brookes，1980；马费成，1983）。知识地图的主要功能在于实现知识的快速检索、共享和再重用，充分有效地利用知识资源。知识地图是关于知识的来源的知识（Vail，1999）。知识并非存储在知识地图中，而是存储在知识地图所指向的知识源中。知识地图指向的知识源包含数据库、文件以及拥有丰富隐性知识的专家或员工。有的企业应用知识地图来揭示知识的结构，实现对知识及其相关知识的检索。另外，知识地图在文献学中也有应用，即科学知识图谱。

科学知识图谱是用来显示知识演化进程和知识结构的图形化与序列化的知识谱系（陈悦等，2015）。1955 年，加菲尔德提出可以将引文索引应用于检索文献的思想（Garfield，1955）。1965 年，普赖斯（Price，1965）指出，引证网络类似于当代科学发展的“地形图”，并提出使用引文网络来研究当代科学发展脉络的方法。从此，分析引文网络成为一种研究科学发展脉络的常用方法。2005 年，陈悦和刘则渊（2005）引入了“科学知识图谱”的概念。科学知识图谱结合应用计量学引文分析和共现分析、图形学、可视化技术、信息科学等学科的理论与方法（秦长江和侯汉清，2009），图形化地展示各领域的学科结构、各学科的研究内容、学科间的关系、识别和分析学科的发展新趋势以及预测前沿等。

表 1.2 比较分析了知识图谱、本体、知识地图以及科学知识图谱的异同。

**表 1.2　知识图谱相关概念**

| 类别 | 知识图谱 | 本体 | 知识地图 | 科学知识图谱 |
|---|---|---|---|---|
| 学科范围 | 计算机科学、人工智能 | 知识管理、人工智能、语义 Web | 图情学、知识管理 | 图情学、科学计量学 |
| 定义 | 描述客观世界的实体及其之间的关系 | 共享概念模型的明确的形式化说明 | 知识指南、知识管理工具 | 科学知识结构、演化的可视化 |
| 表现形式 | 图、OWL | OWL | 图 | 图 |
| 知识来源 | 开放的数据源（网页、知识库等） | 专家、开放的数据源 | 特定组织结构、科学文献 | 学科知识、科学文献 |
| 应用 | 语义搜索、智能问答等 | 知识管理、语义服务、自动推理 | 组织内知识索引、文献知识分析 | 科学社会网络分析、学科知识趋势分析 |

### 1.2.2　知识图谱的构成

知识图谱由数据层（也称实体层）和模式层（也称本体层）两部分构成（徐增林等，2016）。

在数据层中，事实以“实体—关系—实体”或“实体—属性—属性值”的三元组模式存储。这些三元组数据根据语义关系互相联系，构成一个庞大的关系图。其中，实体是知识图谱的最基本元素。实体是指具体的人名、组织机构名、地名、日期、时间等。关系是指两个实体之间的语义关系，通常以三元组的结构来表示，如实体—关系—实体。属性是对实体的说明，知识图谱认为实体与属性之间是一种名称性关系，也可用三元组的结构表示，如实体—属性—属性值。本体描述语言 OWL 将实体的属性视为实体与属性值之间的关系，使用 hasValue 描述属性值。

模式层建立在数据层之上，对数据层进行规范约束。模式层是结构化的知识库模板，是知识图谱的核心。大多知识图谱采用本体作为模式层，借助本体定义

的公理、规则和约束条件来规范知识图谱中的数据层。知识图谱的模式层是知识图谱的概念模型和逻辑基础。也可以将知识图谱视为实例化了的本体，知识图谱的数据层是本体的实例。有的知识图谱（大多是自底向上构造的）允许只有实体而没有上层本体的约束。

在构建知识图谱时，若采用自底向上的构建方法，在融合来自不同数据源的知识时，需要本体提供统一的术语和结构，规范各个术语间关系，将从各个数据源获取的知识融合成一个庞大的知识库。若采用自顶向下的构建方法，首先需要一个本体作为知识图谱的模式层，定义概念、关系及约束，然后才能添加概念及关系的实例。

### 1.2.3　知识图谱的分类

从构建过程是否依赖自动抽取技术来看，知识图谱大致可分为两类。一类是早期的本体，如 WordNet、Cyc、HowNet 等。这类知识图谱大多由专业人士手工构建，规模较小，但其知识质量高，能够确保准确性与完整性。另一类是从开放的互联网信息中自动抽取实体与关系构建的，如 YAGO、DBpedia 等。这类知识图谱规模大，但因其数据源的复杂多样及自动抽取算法的不完全准确，往往会有大量不完整信息、噪声等。近年来，随着知识图谱成为学界及商界的热点，国内也开始重视知识图谱的研究发展，中文的知识图谱纷纷涌现，如 CN-DBpedia、Zhishi.me 等。

按通用性，本体可分为通用本体与领域本体，类似的，知识图谱也可分为通用知识图谱与行业知识图谱。通用知识图谱描述全面的常识性的知识，主要应用于语义搜索，对知识的准确度要求不高，如百科类的 DBpedia、Zhishi.me 和语言学类的 WordNet、大词林等。通用知识图谱强调知识的广度，大多采用自底向上的方式构建，侧重实体层的扩充，因此也导致其大部分较难构建规范的本体层（胡芳槐，2015）。行业知识图谱面向特定领域，能够进行知识推理，实现辅助分析及决策支持等功能，如 GeoNames、中医医案知识图谱等。行业知识图谱对专业性与准确度的要求高，这也要求其必须有严格的本体层模式，通常采用自底向上与自顶向下结合的方式进行构建。通用知识图谱可作为行业知识图谱的构建基础，行业知识图谱也可在构建完成后补充融合至通用知识图谱中。一般来说，通用知识图谱的使用率更高，是现有知识图谱的基础；而行业知识图谱则推进了知识图谱技术融入生活，服务于民。

表 1.3 总结了主要中文通用本体与知识图谱；表 1.4 总结了主要外文通用本体与知识图谱；表 1.5 总结了主要中文领域本体与行业知识图谱；表 1.6 总结了主要外文领域本体与行业知识图谱。