从文本中构建领域本体技术综述

任飞亮 沈继坤 孙宾宾 朱靖波

(东北大学计算机科学与工程学院 沈阳 110819)

(东北大学医学影像计算教育部重点实验室 沈阳 110819)

摘 要 本体是一种重要的知识库,其包含的丰富的语义信息可以为问答系统、信息检索、语义 Web、信息抽取等领域的研究及相关应用提供重要的支持.因而,如何快速有效地构建本体具有非常重要的研究价值.研究者们分别从不同角度提出了大量有效地进行本体构建的方法.一般来讲,这些本体构建方法可以分为手工构建的方法和采用自动、半自动技术构建的方法.手工本体的方法往往需要本体专家参与到构建的整个过程,存在着构建成本高、效率低下、主观性强、移植不便等缺点,因而,此类方法正逐步被大量基于自动、半自动技术的本体构建方法所代替.自动、半自动构建的方法不需要(或仅需少量)人工参与,可以很方便地使用其它研究领域(如机器学习、自然语言处理等)的最新研究成果,也可以方便地使用不同数据源进行本体构建.其中,文本数据源具有数据量大、获取方便的优点.因而,越来越多的研究者开始关注如何有效地使用文本资源进行本体构建.该文全面分析了以文本为数据源,采用自动、半自动技术进行本体构建的国内外最新研究成果.该文首先介绍了当前一些具有代表性的关于以文本为数据源进行本体构建的技术综述文章.在这一部分内容中,该文重点关注了各个综述文章针对本体构建技术研究所得出的结论.接着,该文从"全局"与"局部"两个角度对本体构建方法进行了详细的介绍.在"全局"角度介绍中,该文根据本体构建过程中用到的主导技术,将本体构建方法分为统计主导的方法和语言分析主导的方法两类,分别对各类方法进行了详细的介绍,并分析了各类方法的优缺点.在"局部"角度介绍中,该文把本体构建过程分为以下子任务:术语抽取、概念抽取、关系(包括层次关系和非层次关系)抽取、本体形成.分别从每个任务所使用的技术,从"任务-技术"这一角度,介绍了当前以文本为数据源进行本体构建的国内外最新技术研究进展.第三,该文对当前本体构建技术的常用评价方法以及最新关于本体构建技术评价方法的研究成果进行了介绍.第四,该文选取了几种当前在国际上具有广泛影响力的本体构建系统,对其进行本体构建的具体过程以及生成的本体结果进行了介绍.第五,该文对当前本体构建研究过程中所面临的问题和挑战进行了深入的分析.最后,该文结合当前机器学习及自然语言处理研究领域的最新研究成果,讨论了本体构建未来的研究方向.

关键词 本体构建;本体术语抽取;本体概念抽取;本体关系抽取;深度学习

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.00654

A Review for Domain Ontology Construction from Text

REN Fei-Liang SHEN Ji-Kun SUN Bin-Bin ZHU Jing-Bo

(*School of Computer Science and Engineering, Northeastern University, Shenyang* 110819)

(*Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, Shenyang* 110819)

Abstract Ontology is a kind of important knowledge base. Because of its rich semantic information, it is of great help for improving the performances of applications like question and answering, information retrieval, semantic web, information extraction, and so on. How to construct ontology effectively and quickly is of great research value. Lots of ontology construction methods have been proposed from different perspectives. Generally, these methods can be classified into manual construction methods and (semi-) automatic construction methods. For manual construction

收稿日期:2016-07-18;在线出版日期:2017-05-06.本课题得到国家自然科学基金(61572120,61300097,61432013)资助.**任飞亮**,男,1976 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为自然语言处理、领域本体构建.E-mail:renfeiliang@mail.neu.edu.cn.**沈继坤**,男,1991 年生,硕士研究生,主要研究方向为自然语言处理、知识图谱构建.**孙宾宾**,男,1992 年生,硕士研究生,主要研究方向为自然语言处理、知识图谱构建.**朱靖波**,男,1973 年生,博士,教授,博士生导师,主要研究领域为自然语言处理、机器翻译.

methods, they usually need some dedicated ontology experts participate in the whole process of ontology construction. Due to the shortcomings of high cost, low efficiency, subjectivity, and inconvenient in transplantation, these manual ontology construction methods are being replaced by large numbers of (semi-) automatic ontology construction methods. For (semi-) automatic methods, they don't need much manual effort and can easily use the latest research results in other research fields like machine learning, natural language processing, etc. Meanwhile, they can also construct ontologies by using different kinds of data source. Currently, large amount of text source is easily acquired, thus how to construct ontologies taking text as data source is attracting more and more researchers' attention. A large number of (semi-) automatic construction methods have been proposed. This paper thoroughly reviews the state-of-the-art (semi-) automatic ontology construction methods that take text as data source. Firstly, this paper reviews several existing representative technical survey papers about ontology construction that take text as data source. In this part, the authors focus on the conclusions drawn in these survey papers on ontology construction. Secondly, this paper makes a thorough review about the state-of-the-art construction methods proposed up to now on ontology construction from two aspects. In the first aspect, this paper introduces these methods from a global perspective and classifies the ontology construction methods into two main kinds based on the dominant techniques used in the process of ontology construction: one kind is statistical based methods, and the other is linguistic analysis based methods. This paper introduces these two kinds of methods one by one in detail. For every kind of method, its advantages and disadvantages are also analyzed. In the second aspect, this paper introduces the ontology construction methods from a local perspective and divides the whole ontology construction process into several sub-tasks: term extraction, concept extraction, relation (including hierarchical relations and non-hierarchical relations) extraction, and ontology formalization. Then this paper reviews the techniques used in these sub tasks one by one. Thirdly, this paper reviews the widely used evaluation methods for ontology construction, and introduces the lasted evaluation research results. Fourthly, this paper introduces several representative and well known ontology construction systems. For each system, this paper ignores its technical detail and only focuses on the sub-tasks involved in their ontology construction processes and the outputs generated by them. Fifthly, the challenges and problems in ontology construction are discussed. In the final part, this paper points out several possible research directions for ontology construction based on some latest research results in the fields of machine learning and natural language processing.

Keywords　ontology construction; ontology term extraction; ontology concept extraction; ontology relation extraction; deep learning

1　引　言

本体是一种重要的知识库资源. 根据目前广泛被接受的本体的定义[1],本体是对概念形式及概念间关系的一种规范、明确的定义. 一般来讲,一个本体通常由三部分组成:概念、概念间的关系以及建立在关系之上的公理. 根据一个本体所要描述的目标范围,可将本体分为通用本体和领域本体. 通用本体旨在建立可广泛应用于多种应用场景的本体知识,是对通用类知识的一种规范描述. 领域本体则不同,旨在对某一具体领域建立相应的知识规范描述.

相应地,本体构建(也有文献称为本体挖掘、本体学习等)就是指构建本体的过程. 一般来讲,有两种常见的本体构建方法:一种是依靠本体专家手工构建,另一种则是在一些机器学习方法的帮助下采用自动或半自动的方法进行构建. 显然,手工构建本体的方法缺乏灵活性,构建成本高,而且效率低下.

并且,不同的本体专家对一些概念的认知不同也会导致手工构建本体的方法带有很强的主观性,构建的本体很难被其他专家进行扩展.因而,自动或半自动的本体构建方法逐渐成为当前本体构建的主流方法.

在本体构建过程中,领域本体的构建难度远小于通用本体,在有领域数据的情况下非常适合采用机器学习的方法进行自动或半自动地构建,所以当前本体构建研究基本以构建领域本体为目标.在本体的构建过程中,当前的研究者主要关注从给定数据中挖掘概念以及概念间的关系,很少关注公理的建立.文献[2-3]讨论了一些建立公理的简单方法,但这些方法多是基于规则的简单方法.当前研究者并未针对本体公理的建立展开深入的研究,因而,在本文中,我们将不对本体公理建立的研究进行讨论.

图 1 显示了一个从计算机领域数据中构建的本体中的部分概念及关系.在图 1 中,矩形框中的短语表示概念,矩形框之间的边表示相应概念间的关系,箭头表示概念间关系连接的方向:即起始矩形框表示的概念和结束框表示的概念之间拥有对应边所表示的关系.如图 1 中"*C-E machine translation*"和"*Machine translation*"的关系类型是"*kind-of*",该关系可解释为"*C-E machine translation*"是一类(*kind of*)"*Machine translation*".图 1 中,边上没有标识具体关系类型的,表示对应的概念之间的关系类型为普通的层次关系.

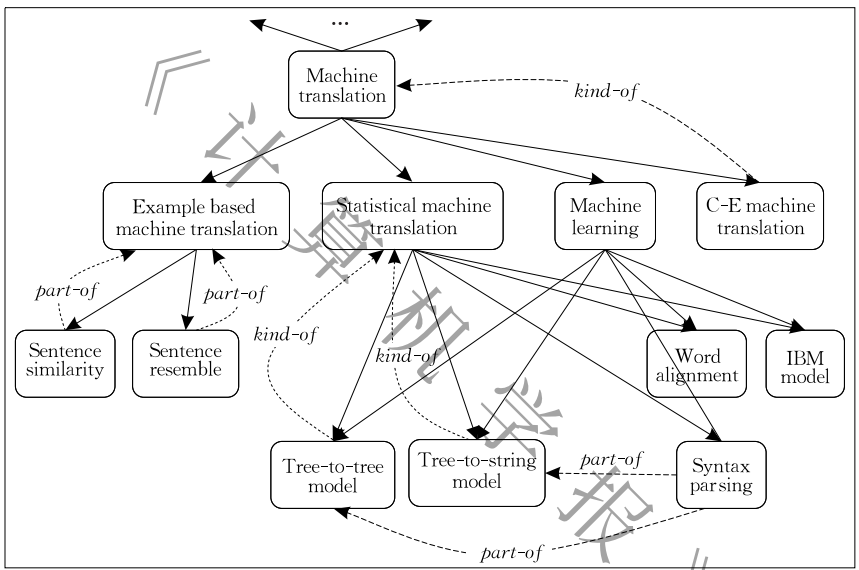

图 1 一个领域本体中概念及关系的部分实例

从图 1 可以看出,本体中含有丰富的语义信息,这些语义信息可以有效地降低概念理解上的歧义,对提升许多文本处理相关任务的性能有重大帮助.因此,目前本体被广泛地应用在许多与文本处理相关的任务中,如信息检索[4-11]、信息抽取[12]、信息整合[13-14]、数据管理[15-17]、信息推荐[18-20]、文本分类与聚类[21-23]、问答系统[24]等,并且均取得了很好的效果.

由于本体在多种应用中发挥的巨大作用,越来越多研究者开始关注如何高效地获取本体知识,并提出了大量的本体构建方法.与一些研究者[25-27]关注使用本体编辑器等工具进行本体构建不同,本文将从不同角度深入地分析当前以文本为数据源、自动或半自动地进行本体构建的主流方法,并探讨这些方法当前面临的主要问题及挑战,以及未来的研究方向.

本文在第 2 节中将介绍当前已有的关于以文本为数据源进行本体构建研究的技术综述文献;第 3 节将对本体构建的研究现状进行详细分析,包括本体构建过程中的主导技术分析、本体构建过程中各个任务模块的常用技术分析、本体的评价方法介绍以及常见的几种本体构建系统分析等内容;在第 4 节中,我们将讨论本体构建中的问题与挑战;第 5 节将讨论本体构建未来的研究方向;最后,在第 6 节我们将对本文的工作进行总结.

2 相关综述分析

在本节中,我们将简单介绍在本文之前的几篇关于从文本中进行本体构建研究的综述文章.

在文献[28]中，作者主要从以下方面对本体构建方法进行了综述：(1)本体构建过程；(2)本体构建过程中的技术分析；(3)主流的本体构建系统分析；(4)技术进展以及面临的挑战.在论文最后的总结中作者指出未来本体研究的热点问题可能包括：(1)如何有效利用社交数据(social data)进行本体构建；(2)如何通过设计新的算法而利用网络数据(web data)中的结构信息进行本体构建；(3)如何进行与语言无关的本体中实体的表示研究；(4)在进行网络规模级(web-scale)本体构建研究中，如何保证算法的有效性以及鲁棒性；(5)如何进行实用化的本体(heavyweight ontologies)构建研究.

2003 年，欧盟信息社会技术方案委员会(Information Society Technologies Programme of the Commission of the European Communities)的研究者们也对本体构建研究进行了综述研究[29].在该综述文章中，研究者们通过研究从文本中进行本体构建的 36 个方法，分析了(1)以文本为数据源进行本体构建的常用方法与工具；(2)以字典为数据源进行本体构建的常用方法与工具；(3)以知识库为数据源进行本体构建的常用方法与工具；(4)以半结构化的图表数据为数据源进行本体构建的常用方法与工具；(5)以关系型的图表数据为数据源进行本体构建的常用方法与工具.通过对以上研究分析，作者认为：(1)针对从文本中进行本体构建研究而言，并不存在一个通用型、指导性的方法体系；(2)当前本体构建过程主要是基于自然语言分析技术，并通过具体使用的语料资源来决定整个本体构建过程；(3)很多方法都会把 WordNet 作为本体构建的初始资源，先通过 WordNet 获取一些初始的概念以及关系，之后，再通过其它技术来进行扩展而得到最后的目标本体；(4)对于本体构建而言，几乎没有完全自动化的系统，多数方法需要用户的参与来从标注语料库中获取相应的概念以及关系；(5)需要进一步研究对本体构建进行评价的方法，以便于不同方法之间进行比较.

在文献[30]中，作者调研了最新的 50 多个本体构建系统及方法，重点从构建什么、从哪里构建、怎么构建这三个角度进行了分析.在该文献中，作者认为，本体构建研究中尚未有效解决的问题包括如下几项：(1)多数研究主要关注层次关系(hierarchical relation，也称 taxonomic relation)抽取，对于非层次关系(non-hierarchical relation，也称 non-taxonomic relation)抽取的研究相对较少；(2)多数研究主要关注领域相关的本体构建，而较少关注采用自动的方法进行通用领域本体构建的研究；(3)多数提出的方法主要在一些规模较小、领域受限的语料中使用，而没有在真实使用环境中验证方法的有效性(文献[31]也指出未来在本体构建研究中应该使用网络规模级的数据进行本体构建算法的验证).最后，文献[30]指出，未来本体构建的研究应该主要集中在以下几个方面：(1)本体公理的学习；(2)找到可以客观评价本体准确率、算法效率、本体完备性的文法；(3)应找到更有效的全自动的本体学习方法；(4)移植性更好的本体构建方法.

文献[32]根据本体构建过程中所用数据的结构化程度(结构化、半结构化、非结构化)以及本体学习对象的层次(概念、关系、公理)，将本体构建问题划分为 9 类子问题，并分别阐述了这 9 类问题的基本特征、常用方法，并进一步比较了一些本体学习工具.最后，作者认为本体构建研究存在如下问题：(1)在本体构建方法上，当前的方法在各类数据上均存在一些需要进一步深入研究的问题.并且，本体构建方法缺乏通用性，构建方法也应向自动化学习方向努力；(2)当前一些本体学习工具需进一步完善；(3)需要统一的评价本体构建结果的标准.

文献[33-34]基于 12 个研究项目对本体构建方法进行了综述分析，并得出了以下结论：(1)本体构建所需的数据资源或多或少是半结构化的，需要一些领域专家提供一些种子概念集合，并基于这些种子数据进行其它的数据资源的收集或是用这些种子数据构建目标本体的基本框架.而以自由文本或是异构数据为数据源进行本体构建，距离实际的应用还有一定距离；(2)对于本体构建中的概念抽取，存在一些比较成熟的技术，如词性标注、词义消歧、词汇化、模式匹配等；(3)关系抽取在本体构建过程中更加复杂，解决难度也更大；因而也是本体构建过程以及本体应用中的主要障碍.作者在该综述文献的结论中指出，对于自动、半自动的本体构建技术而言，还没有显著的技术突破，但却吸引了大量科研人员的注意力，应该是未来本体构建研究的主要研究方向.

文献[35]对本体构建的方法、评价、应用进行了综述研究.该文献指出，本体构建研究要取得突破，必须要对其任务、子任务有清晰且明确的定义，并且相应的定义必须在学术界有广泛的共识，并能以此为依据，设定本体构建各个任务的评价方法.作者依此提出，应该为本体构建研究提供一些基准任务

(benchmarks)以及相应的评测机制,以便于不同的本体构建方法可以在同样的任务下进行比较.同样,文献[31]也指出了建立统一的基准任务的重要性.

文献[36]主要关注从无结构文本中进行本体或是类本体结构(ontology-like structures)的构建.该文献的作者认为,本体构建技术可以分为从头构建本体(constructing ontologies from scratch)或是扩展已有本体(extending existent ontologies)两类,并认为前者通常可定义为一个聚类问题,而后者则通常可定义为一个分类问题.作者认为,(1)目前许多存在的本体并不能履行其对应名称中所暗示的承诺.多数本体仍只能称为本体原型(ontology prototype);(2)采用聚类方法进行本体构建过程中,前几项任务中取得的结果并不会对随后的任务提供很大的帮助.并且,本体构建中从输入数据中提取到的一些显性的本体关系模式在实际文本中很少能够重现;(3)语义 web 领域对本体的需求和从文本中学习到的本体之间存在着很大的差异.

在文献[37]中,作者为本体构建过程定义了一系列任务,并认为可以根据本体构建过程中所采用的数据形式对本体构建方法进行分类:如采用结构化数据的本体构建方法、采用半结构化数据的本体构建方法、采用无结构数据的本体构建方法等.在该文献中,作者认为采用自动方法进行非层次关系抽取的研究已经引起了研究者们的关注,但仍未达到成熟阶段,而对于本体中公理的学习则仍处于最初始阶段.同时,该文献的作者也认为,如何有效地对本体构建过程进行评价仍是一个未解决的问题.

此外,一些研究者也从知识获取的角度对本体的研究工作进行了综述.如在文献[38]中,本体知识被当作是常识性知识(commonsense knowledge)的一个类型.在该文献中,作者对常识性知识获取的任务、所用的技术以及评价方法进行了分析,并对几种如 Cyc、YAGO 等代表性本体的构建进行了介绍.该文献虽然与本文的从文本中进行本体构建的关注点不同,但在其结论中提到的需要客观、公正的本体评价方法,这和本节前面介绍的一些综述文献的观点是一致的.文献[39]也对本体研究进行了综述,但其主要是从方法论角度讨论本体构建.

通过对以上几篇针对本体构建研究的已有综述文献来看,研究者们对以下几点具有共识:(1)对于非层次关系的自动抽取研究仍然是本体构建研究的重点内容;(2)从大规模文本中构建实用化的而不是玩具本体(toy ontology)是未来的研究方向之一;(3)未来的本体构建应该是以自动化的形式完成,或是仅需少量的人工参与.

3 研究现状分析

本文首先用图 2 来显示本体构建的总体结构框图.

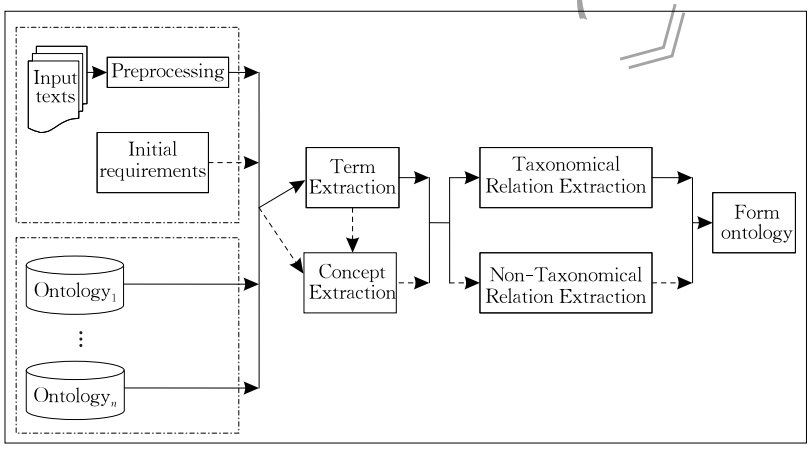

图 2 本体构建的总体结构框图

从图 2 中我们可以看出一个本体构建过程通常包含以下几个任务模块:

(1)预处理模块(preprocessing),主要功能是对输入的文本进行先期处理(如断句、分词、词性标注、短语识别等),以使之符合后续任务对输入的格式要求.

(2)术语抽取模块(term extraction),主要功能是从输入文本中提取那些和目标领域高度相关的领域术语.

(3)概念抽取模块(concept extraction),主要功能是从输入文本中提取那些和目标领域高度相关的

概念.

(4) 层次关系抽取模块(hierarchical relation extraction),主要功能是抽取本体概念之间的层次关系.

(5) 非层次关系抽取(non-hierarchical relation extraction),主要功能是抽取本体概念之间的层次关系以外的其它类型关系.

(6) 形成本体(form ontology)模块,主要功能是将前面任务中抽取出来的术语、概念以及概念之间的各种关系进行整合,形成最终的目标本体.

需要指出的是,上面本体构建的六个任务模块在一个本体构建任务中并不都是必须的,在图 2 中,我们用虚箭头线表示对应的任务是可选的. 例如有些研究者为了简化任务而将术语简单地看作是概念,因而将术语抽取和概念抽取这两个任务模块合二为一.

从图 2 也可以看出,本体构建可以有两种类型的数据输入:一种是文本输入,即从文本中挖掘本体知识;另一种是本体输入,指通过本体合并技术将多个已有的本体重新组织成一个新的目标本体,或是以某些现有本体为基础,通过不断扩展新的概念以及新的关系而形成一个新的目标本体[40-46]. 对于后者,一个先决条件是必须有一些可用的本体资源,在实际情况中这一条件往往很难被满足,因此,我们将重点介绍前一种情况,即从文本数据中挖掘本体,这也是当今本体构建的主流研究方向.

我们进一步用本体构建技术路线图(图 3)来更加清晰地显示本体构建的各个任务模块以及对应的常用技术. 本文接下来的章节将以图 3 为基础,选择了部分有代表性的参考文献所用的技术为例,从以下两个角度由总体到局部地对本体构建技术的研究现状进行全面深入地介绍.

(1) 本体构建过程中采用的主导方法.

(2) 本体构建过程中各个任务模块的常用技术.

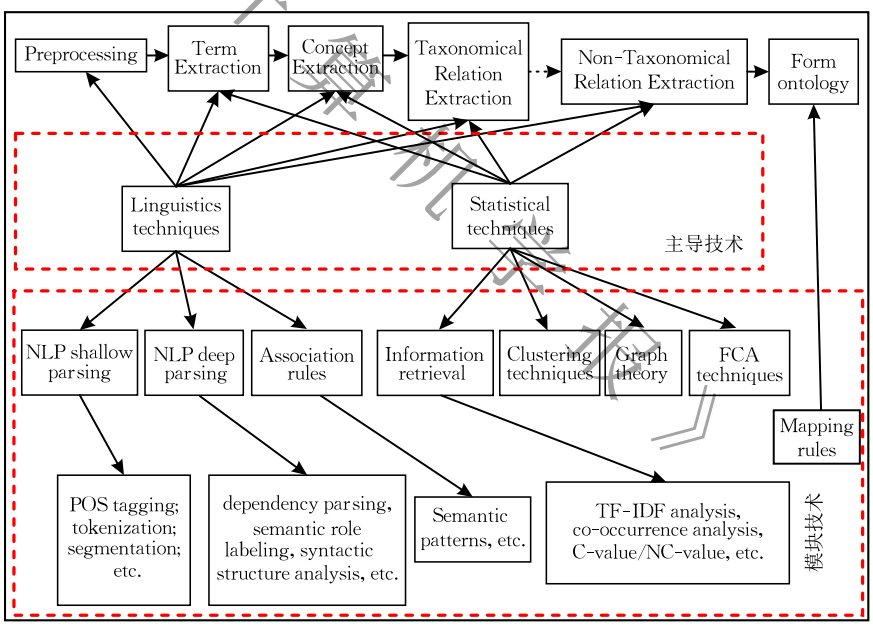

图 3　本体构建技术路线图

3.1　本体构建过程中采用的主导方法

根据使用的主导方法,可将本体构建方法分为统计主导的构建方法和语言学主导的构建方法.

3.1.1　统计主导的方法

在统计主导的本体构建方法中,来源于自然语言处理、信息检索等领域的多种统计技术被广泛应用在本体构建过程的各个模块中. 通常在这一类方法中,研究者们较少关注文本中包含的语义学知识以及文本内容之间的关联,在整个构建过程中也不需要深层次的语言学知识以及额外的领域相关或语种相关的资源. 一些常用的统计方法,包括聚类、词频统计、词共现分析、潜在语义分析、术语搭配、关系规则挖掘、浅层自然语言处理等技术被广泛地应用在这类统计主导的本体构建方法中. 这类方法的主要思想是词汇单元(单个词、词组、短语、词串等)间的共现信息可以为识别它们之间的关系以及描述对应词汇的重要程度提供重要的指示信息,因而可直接应用在本体构建过程中的概念抽取及关系挖掘中. 这类本体构建方法的典型代表包括文献[47-50]中介绍的方法.

在文献[47]中,研究者们提出一种基于图论的领域本体自动构建方法.在他们的方法中,每一个输入文档将首先被表示成一个图,在这个图中,结点表示词汇,边表示词汇间的共现关系.接着,研究者采用基于随机漫步的权重分配算法(random walk term weighting)来评估每个词汇与目标领域的关联程度,这种评估不仅基于一些局部信息,同时也基于一些全局信息.接下来,研究者使用马尔可夫聚类算法(Markov clustering)来对图中的词汇进行词义消解,并将词义相近的词汇分到一组进而形成领域概念.最后,一个改进的 gSpan 算法被应用进来进行有高频子图挖掘,挖掘到的每一个子图都生成一种关系,对应的关系类型由该子图中的核心动词以及与核心动词相关联的两个概念决定.在文献[48]中,研究者们提出一种基于水结晶模型(crystallizing model)的领域本体构建方法.在该方法的概念抽取过程中,研究者首先从某一词汇出发,不断合并与其临近且意义相近的词汇,进而生成更大的词汇组,最终形成可以表示某个概念的词汇群,整个过程就象是水蒸汽的结晶过程一样.具体而言,他们先用一些自然语言处理工具对输入文档进行预处理,抽取一些领域术语.接着,他们执行了一个称为概念铸型(concept casting)的过程来分析术语之间的同义关系.最后,利用水结晶模型来提取概念以及概念间的层次关系(hierarchical relations)及非层次关系(non-hierarchical relations).在他们的预处理模块中,主要用到了一些浅层的自然语言处理技术,比如词性标注、去停用词、词干提取等.在其概念铸型过程中,他们采用有限状态自动机来识别名词短语,并采用共现相似度、语言及语义相似度来判断两个术语间的相似程度;他们采用关联规则挖掘进行层次关系的识别,用自组织映射聚类算法(self-organizing map clustering algorithm)来进行概念聚类,并采用位于质心位置的概念来代表整个词汇类所描述概念的特性.在文献[49]描述的本体构建方法中,作者首先从互联网上搜集大量领域相关的网页作为数据源;接着,他们使用 html 标签来从网页中选择一些有意义的术语词汇;接着,他们使用 TF-IDF 技术来选择重要的术语词汇作为领域概念,并使用 ART 网络(Adaptive Resonance Theory Network)来进行术语聚类.在聚好的每一个类中,都选择一个候选词作为代表该类的概念;并采用奇异值分解(SVD)操作来减少抽取的术语的数量并发现术语之间的潜在语义信息,最后,采用布尔操作来识别概念间的层次关系.文献[50]的工作与文献[49]的工作类似,也是使用网页文本来进行领域本体的构建.但不同的是,文献[50]也进行了非层次关系的挖掘,这种挖掘主要是基于核心动词的方法,也就是先在句子中识别动词,之后分别向左、向右搜寻与该动词最邻近的概念,之后,这两个概念就可建立一个由核心动词所确定的关系.类似的研究还包括文献[51]的工作,他们也是采用自组织映射聚类算法来进行概念聚类并根据聚类结果定义概念间的层次关系.

需要指出的是,目前还有一些研究者们采用形式概念分析(Formal Concept Analysis,FCA)来进行本体构建.这类方法主要是基于概念格(concept lattice)的相关理论,在文本中对于每一个抽取到的术语寻找一定数量的形式上下文(formal context),之后按照〈对象,属性〉二元组间满足偏序关系(partial relations)而从下向上构造概念格.概念格中不同层次上的〈对象集,属性集〉就可以被解释成为一个个本体概念.这类方法的典型代表包括文献[40-41,52-56]中介绍的方法.一般来讲这类基于 FCA 的本体构建方法的特点是构造过程中计算代价大,因而现在的研究者们只是用它进行了小规模本体的构建研究.

3.1.2 语言学主导的方法

与统计方法主导的本体构建方法相对应的是语言学方法主导的本体构建方法.在这类方法中,一些来源于自然语言处理的深层分析技术被广泛使用并在整个本体构建过程中起主导作用.这些自然语言处理技术包括词性标注、句法分析、依存分析、语义角色标注等.此外,一些语言相关的资源,包括语义词典、语义模板、词汇-句法模板等,也被广泛地应用在本体构建的各个过程中.这类本体构建方法的基本思想是:在依据给定文本构建本体的过程中,本体概念以及概念之间的关系隐式地存在于输入文本中,这类知识只有通过全面而深入的文本分析技术才可以获得.因此,这类本体构建方法往往需要通过深入而全面的句法分析技术来揭示文本中各个片段之间的潜在关联.

文献[57]提出一种基于语义角色标注(semantic role labeling)的领域本体构建方法.在他们的方法中,C-value/NC-value 算法首先被用来从文本中抽取可以描述领域特性的多词术语(multiword term),TF-IDF 方法被用来从文本中获取可以描

述领域特性的单词术语(single word term).在这一过程中,他们使用了一个被称为 Freeling(http://nlp.lsi.upc.edu/freeling)的自然语言处理工具来对输入的文档进行预处理;预处理工作包括断句、词性标注、句法分析等.接下来,对于每一个句子都进行语义角色标注.再往后,在语义角色标注的帮助下,为每个句子寻找核心动词.接着,以核心动词为中心,在句子中分别向左、向右寻找与之关联的最近的概念,并由这些概念和核心动词形成一个类似〈$concept_1$,$verb_i$,$concept_2$〉的关系三元组,这个三元组表示 $concept_1$ 与 $concep_2$ 之间存在着由 $verb_i$ 所确定的关系.

在文献[3]的方法中,作者使用了斯坦福大学的自然语言处理工具先对输入文本进行预处理,并得到每个句子的 PCFGs 句法分析结果以及依存句法分析结果.之后,在此基础上通过简单的分析规则来进行概念抽取以及关系抽取.在文献[3]所描述的方法中,只有名词、形容词会被留下来做为进一步概念抽取的候选词.之后,利用句法分析的结果,将临近的名词或是有依存关系的名词合并为更大长度的名词短语(在文献[3]中规定不超过 7 个词)作为术语抽取的初始结果.之后,以动词为核心进行术语间关系的抽取.抽取的方法和文献[57]所描述的方法类似.

类似工作还包括文献[58-59]中介绍的工作.在文献[58]中,作者首先使用一些自然语言处理工具对输入文档进行全文分析,包括词性标注、句法分析、词义消歧等.由于句法分析已经清晰地对一个句子中各个成分间的关系进行了分析,因而,可以以此为基础,进行相应的本体概念抽取及概念间的关系挖掘.在具体的层次关系挖掘中,作者设计了一个启发式的基于规则的抽取算法;而对于非层次关系挖掘,作者也采用了核心动词法,即在句法分析的基础上先为每个句子寻找其相应的核心动词,之后,以核心动词为中心,分别在句子中向左、向右寻找与之相临的概念,并进而组成关系.在文献[58]中,概念抽取和关系抽取为两个独立的模块,也就是说,概念抽取模块和关系抽取模块中的概念集合并不一样.最后,作者使用 WordNet 为参考,将抽取的概念及相应关系进行分类并整合成最终的领域本体.

在文献[59]的工作中,作者提出一种基于深度语义分析技术与图论技术相结合的领域本体构建方法.在他们的方法中,使用了斯坦福大学开发的词性标注工具,并根据词性标注的结果,使用规则进行 chunk 识别、领域术语识别、*is-a* 关系识别.之后,各种在图论中被广泛使用的技术被应用进来进行概念及关系的过滤.这些技术包括 PageRank 算法、Hits 算法、结点度分析等.这些分析技术得到的结果被一些基于规则的投票机制整合到一起,用来作为最终概念及关系过滤的依据.最后,一些基于规则的句法模式(syntactic pattern)被设计出来用来对上面抽取出来的关系进行映射,进而形成最终的本体.

3.1.3　本体构建研究主导方法小结

一般来讲,在统计主导的本体构建方法中,使用的自然语言处理技术都是很浅层次的,这将使其在识别由复合词组成的本体概念以及挖掘非层次关系时往往效果有限.但由于这类方法使用的自然语言处理技术在很多语种中都很容易被满足,因而,这类方法更灵活,可以广泛地被应用于不同领域、不同语种下的本体构建,尤其是对那些缺乏深度自然语言分析技术的语种(如蒙语、藏语、维语等),或现有深度自然语言分析技术效果有限的领域(如医药领域、化工领域、生物领域等).

另一方面,由于使用自然语言处理技术对输入文本进行了深层次的分析,语言学主导的本体构建方法往往更容易获得更高性能的本体知识,在处理由复合词组成的本体概念以及概念间的非层次关系时往往会得到更高的精度.但随之而来的就是这类方法对自然语言处理技术的高度依赖.可以说,高性能的自然语言处理技术是该类方法取得成功的保证.但需注意的是,对于一些小语种语言(如蒙语、藏语、维语等)而言,由于缺乏与英语等语种可比的高性能的自然语言分析工具,因此,这类语言学主导的方法往往就很难被使用.此外,即使是对于像英语这种自然语言处理技术已经在某些领域取得了很高性能的语种而言,在处理一些特殊领域(多数是那些缺乏训练数据、相应自然语言处理模型训练不充分的领域)的文本数据时,也很难获得高质量的自然语言分析结果.因而,语言学主导的本体构建技术在实际使用中会受到许多限制.

3.2　本体构建过程中各个任务模块的常用技术

在本节中,我们将从本体构建中的“任务-技术”这个角度来对本体构建的研究现状进行介绍.需说明的是,这里我们没有对预处理模块进行单独介绍,因为其主要功能大多集成在术语抽取或概念抽取任务中.

3.2.1 术语抽取常用技术

术语(term)抽取是指从输入文本中抽取那些与目标本体所描述的特定领域相关的,且能较好地描述该领域特征的词汇,是本体构建过程中的一项基本任务.对许多当前的本体构建方法而言,术语是一项必不可少的输出.而且在一些当前的本体构建方法中,术语也被直接当作概念使用,尽管二者本质上并不相同.对于术语,它们可能是简单词,也可能是由多个简单词组成的复合词.在本体构建过程中,为了获取术语,输入的文本需进行一些必要的预处理操作,一些浅层的自然语言处理技术经常被应用在这一过程中.比如,噪音数据清洗、标记化(tokenization)、分词、词性标注等.接下来,一些统计或概率的方法被用来评价一些名词序列的搭配强度以及与对应领域的关联程度.那些具有稳定搭配关系的词序列以及与对应领域关联度比较大的词序列将被视为术语.在评价词汇间的搭配强度以及与某一领域的关联程度时,TF-IDF、C-value/NC-value 等技术会被经常用到.比如,在文献[57]的工作中,作者就同时用到了 TF-IDF 技术和 C-value/NC-value 技术来进行领域术语的抽取.在文献[47]中,经过文本预处理后,作者使用随机漫步赋权(random walk term weighting)方式对每个词汇进行权重分配.这些权重从全局以及局部两个角度反映了一个词汇与对应领域间的关联程度,可以作为领域术语抽取的重要依据.在文献[58]的本体构建工作中,作者使用了一个自然语言处理工具集对输入文本进行预处理,包含词性标注、句子分析、词义消歧等.接着,采用一些基于语言学的过滤规则来选取领域术语候选词汇.

实际上,在领域术语的抽取过程中,不仅会用到一些浅层的自然语言分析技术,一些深层次的自然语言分析技术也经常被广泛使用.比如依存句法分析就是一个在术语抽取过程中被广泛使用的技术.在文献[59]的工作中,作者首先对输入文本进行了深层次的句法分析,之后,以句法分析的结果为基础进行了领域术语的抽取.其他相似工作可参见文献[48,51]等.

3.2.2 概念抽取常用技术

概念抽取是本体构建过程中一个必不可少的任务,因为概念是本体的基础,本体中的各类关系就是建立在概念的基础之上的.但需指出的是,并不是所有的本体构建方法都会明确地输出概念,相反,一些研究者在他们的研究工作中,直接将术语看作概念.比如,在文献[57,59-60]等工作中,研究者们就是将抽取到的领域术语直接看作是领域概念.在文献[22-23]中,作者将学术论文中的关键词直接作为领域概念.在文献[61]中,构建本体的数据是有用户文本标注的图片集,作者将用户标注的 tag 文本信息集合中,每一个存在 Wikipedia 页面的 tag 都当作是一个概念.也有一些研究者是对抽取到的术语做进一步的过滤,选择那些更重要的术语作为概念.比如在文献[58]中,作者设计了一种领域相关指标(domain relevance measure)来选择一些领域术语作为领域概念.

另一方面,在许多明确输出本体概念的本体构建方法中,概念通常是通过将相似的术语进行聚类而形成.这一过程往往被进一步分为形成概念(form concept)和标注概念(label concept)两个部分.形成概念的主要任务包括发现术语的各种变形以及判断相似的术语,并将它们进行聚类.而标注概念的主要任务是为每一个术语组选择一个合适的名称来表示其对应的概念.如在文献[62]中,研究者采用 K-Means 聚类方法对输入文本进行聚类,并从每个类中选择出现频率最高的 5 个词做为对应类所表示的概念候选.在文献[47]中,作者使用马尔可夫聚类算法对抽取到的领域术语进行聚类,之后每一个聚好的类都被解释成一个概念.在文献[63]中,作者采用 Chir 统计的方法来计算术语与目标领域的相关度,并根据统计结果来决定其作为目标本体中概念的可能性.在概念标注过程中,文献[63]把每一个概念类中在输入文本中出现次数最多的一个术语作为整个概念类所对应的概念.在文献[51]中,作者采用自组织映射聚类算法把相似的术语聚在一起而形成相应的概念.

此外,还有一些研究者利用一些已有的资源来辅助进行概念的抽取.一般来讲,这类方法往往包含两部分.第一部分是概念抽取.在这一阶段,往往需要一个比较大的标注语料库,并通过一些机器学习方法以及一些自然语言处理方法来从这个标注语料库中识别出本体概念.第二部分是概念属性分配.在这一阶段,针对概念的一些语义解释等属性信息会被分配到相应的概念上.要完成这一阶段的工作,往往需要一个通用性的知识资源,研究者们常常用到的是 WordNet.这类方法的代表研究工作包括著名的领域本体学习工具 OntoLearn[64-67].在 OntoLearn

中，首先基于领域相关性(domain relevance)和领域一致性(domain consensus)从输入文本中提取领域术语.接着，一些复杂的术语将会根据 WordNet 提供的信息进行解释和组织.最后，在 WordNet 的帮助下，对抽取出来的领域概念进行过滤.其他相似工作可参见文献[2,68-70]等.

3.2.3　关系抽取常用技术

概念间的关系抽取是本体构建任务中另一个必不可少的任务.本体中的关系被用来描述概念间的各种联系.在一个本体中，概念间的关系决定了本体的最终结构，也决定了本体的最终质量.因此，概念关系抽取是本体构建过程中最重要的一个步骤.一般来讲，概念间的关系可以被分为两类：层次关系和非层次关系.层次关系抽取的主要任务是在概念之间建立层次，而这种层次多数情况下可解释为一种“*is-a*”关系.层次关系的建立有多种方法，比如依据已有的背景知识或专家经验设计一些预定义的关系模板，利用这些关系模板进行层次关系的抽取；也可基于语言学规则或模式进行抽取，或是采用聚类的方法.如在文献[71]中，作者采用一种称为 CBC (Clustering By Committee)的聚类方法将抽取到的本体概念组织成层次结构.

非层次关系的抽取要比层次关系的抽取复杂得多：首先是非层次关系的类别难以明确确定；其次是非层次关系相对于层次关系而言更具隐蔽性.因此，对于非层次关系的抽取是当前本体构建方法研究中的一项重点内容.现在，发现并标注非层次关系主要是基于句法结构分析与依存关系分析.在这个过程中，动词被认为是一个非常好的可以揭示非层次关系的指示词，并且可以有效地帮助本体构建专家对相应的概念关系进行识别并标注.在非层次关系的挖掘中，一些深层的自然语言处理技术(如句法分析、依存分析等)往往在动词识别、关系确定等方面起着关键性的作用.在这些非层次关系中，最终抽取的结果基本都可表示为一个形如“$concept_m$ $verb_i$ $concept_n$”的三元组，这个三元组表示概念 $concept_m$ 与 $concept_n$ 之间存在着一种由动词 $verb_i$ 决定的关系.例如，在文献[57]中，作者采用基于语义角色标注的本体构建方法.在他们的方法中，输入文本中的每一个句子都会被标注各个成分的语义角色.在此基础上，一些基于语言学的模式被设计出来用以识别层次关系.对于那些非层次类别关系，它们首先对每个句子识别核心动词，之后，以该核心动词为中心，在句子中分别向左、向右寻找与该核心动词紧临的概念.之后，找到的两个概念连同核心动词一起组成一个关系三元组.其他相似工作可参见文献[47,49,72-73].

一些研究者也通过统计对应概念间的共现信息来进行关系抽取.在文献[61]中，作者定义了“*Co-Occurrent*”关系以及“*Subsumption*”两类关系.在对“*Co-Occurrent*”关系的抽取中，作者的考虑因素之一就是两个待分配关系概念的共现属性，并将这些共现属性通过一种相似度公式进行整合.在其方法中，用来计算两个概念间的共现信息以及整合共现信息的相似度公式分别如式(1)、(2)所示.

$$d(c_i,c_j)=\frac{\max(\log N(c_i),\log N(c_j))-\log N(c_i,c_j)}{\log N_{\text{total}}-\min(\log N(c_i),\log N(c_j))} \tag{1}$$

$$\phi_t(c_i,c_j)=(1-\exp(-d(c_i,c_j)))^{-1} \tag{2}$$

上面公式中的 $N(c_i)$ 表示对应概念的出现频率信息(在文献[61]中表示包含对应概念的图文件的个数)，$N(c_i,c_i)$ 表示概念 c_i 和 c_i 的共现信息(在文献[61]中表示同时包含概念 c_i 和 c_i 的图文件的个数)，N_{total} 表示所有概念出现的频率信息.

而对“*Subsumption*”关系的抽取则进一步应用了上面公式的结果，用来判断两个概念之间存在“*Subsumption*”关系的依据如式(3)所示.

$$p(c_i|c_j)=\frac{s(c_i,c_j)}{\sum_z s(c_i,c_j)} \tag{3}$$

上面公式中的 $s(c_i,c_i)$ 是式(2)中整合后的概念共现信息的线性组合，其计算方式如式(4)所示.

$$s(c_i,c_j)=\lambda\phi_v(c_i,c_j)+(1-\lambda)\phi_t(c_i,c_j) \tag{4}$$

上面公式中，ϕ_t 采用式(2)计算后得到，ϕ_v 是文献[61]中另外一种类似式(2)的计算图形视觉相似度的指标.

本质上来讲，文献[61]所采用的方法仍属于一种规则映射的方法，其实质是根据概念间的某种属性而决定其对应的关系类型.类似的方法还包括文献[22-23]所提出的本体构建方法.在文献[22-23]的关系抽取方法中，作者也是通过计算概念间的某种相似度值并根据计算结果对概念分配对应类型的关系.在文献[22-23]的相似度计算过程中，考虑了如概念间的共现信息、字符串匹配信息、上下文信息等因素.

也有一些研究者针对特定类型的关系而设计相应的抽取方法.如文献[74-75]中，为了抽取“部分-整体”关系，作者首先构造了一些基于部分整体关系

的意图查询,之后将意图查询提交给搜索引擎,利用搜索引擎从 web 中获取尽可能多的包含部分整体关系的语料.然后根据网页中的 html 标记和意图查询的格式过滤语料,并从中抽取候选部分整体关系.最后基于部分整体关系在自然语言表述中的特点和汉语的构词规律,选出最终的部分整体关系.采用类似关系抽取思路的还包括文献[76]的工作.

还有一些研究者将本体关系的抽取任务转换为一个分类任务.在这类方法中,研究者提前对本体概念间可能出现的关系类型进行了定义.之后,以每一对可能存在关系的概念对作为分类目标,从输入数据中提取描述它们的各种特征(如词表相似度、词性、共现信息、tf-idf 等),并用这些特征来描述分类目标.最后,选取一个分类器(如 SVM、最大熵等)进行训练,并用训练所得到的模型为新的分类目标分配对应的关系类型.如文献[77-78]就属于此类方法.在给定一定数量的标注训练数据的前提下,该类方法往往可以取得很好的关系抽取结果,但构建相应标注训练数据的代价往往会限制这类方法的应用范围.

3.2.4 本体形成常用技术

一般来讲,本体构建中的各个任务是以一种串行的方式进行的,也就是先进行术语识别,再进行概念抽取,接下来是概念间的层次关系抽取,最后是概念间的非层次关系抽取.在这一串行过程中,一个任务的输出往往是下一任务的输入,比如,抽取到的术语往往是下一步概念抽取的基础,而概念则又是关系抽取的基础.但在一些研究工作中,本体的各个任务并不完全串行,常见的方式是概念抽取与概念间关系抽取相互独立,分别进行.这样得到的概念集合与关系抽取中的概念集合就并不相同.在这种情况下,往往需要一个独立的本体形成(form ontology,如图 2、图 3 中所示)模块来把抽取到的概念和概念间的关系以一种合理的方式重新组织以形成最终的本体.如文献[58]的工作就属于这一类,其概念抽取与关系抽取分别独立运行,并最终以一种简单的基于规则的本体映射(ontology mapping)方式组织成统一的一个本体.文献[61]也使用了一种简单的方法形成最终的目标本体.在其方法中,作者将最终目标定义为生成一个有向无环图.其对每一个本体概念都计算一个熵值,并根据该值来决定把哪些概念及其对应的关系加入到最终的有向无环图.类似的工作还包括文献[48]中用到的方法.

3.3 其它类本体构建分类方法

除上面讨论的本体构建分类方法外,还有一些从其它角度对本体构建方法进行分类的方法.

(1)构建方法对语种的依赖性

从语种是否独立的角度可以将本体构建方法分为独立于语种的方法和依赖语种的方法.

对于独立于语种的方法而言,那些语种相关的分析技术以及资源均不需要.如文献[47]的方法就属于这一类,在该方法中,没有使用任何与特定语种相关的资源或技术.相反,对于那些依赖语种的方法,通常会用到一些与特定语种相关的资源或技术,以期得到对应于相应语种性能更好的本体结果.如文献[40,58]的方法中使用了和英文相关的外部资源 WordNet.在文献[57]的工作中,作者需要额外的西班牙语的语义角色标注技术.

(2)本体构建所需的文本类型

根据文献[79]的观点,本体构建过程所需要的数据资源可以分为以下 3 类:

① 结构化数据:如数据库;

② 半结构化数据:如科技文献;

③ 无结构数据:泛指各类文本资源,如网络数据.

相应地,根据本体构建输入数据的类型,也可将本体构建方法分为从结构化数据中构建本体的方法、从半结构化数据中构建本体的方法和从无结构数据中构建本体的方法.如文献[47-48,57]等工作可归结为使用无结构数据进行本体构建,而文献[22-23,58,60,78]等工作则可归结为使用半结构数据进行本体构建.

(3)目标本体的获取方式

根据目标本体的获取方式,可将本体构建方法分为从零构建目标本体的方法和通过合并已有本体构建目标本体的方法.对于后者,需要预先存在一些可用的本体资源,构建的目标是通过合并或是扩充的方式构建新的更大规模的目标本体.在本文中,没有关注此类方法.

3.4 本体构建评价方法

如何有效地对本体构建方法进行评价一直是许多研究者关注的问题之一.文献[28]对本体常用的评价方法进行了总结.文献[80-81]分析了在设计针对本体构建结果进行评价时所需考虑的基本思路、基本原则等问题.一般来讲,本体评价可以从以下两个角度进行:基于应用的评价和基于本体自身的评价.

基于应用的评价主要从应用效果角度来评价所构建的本体的性能.在这类方法中,首先需选取某一具体应用,之后可以通过以下两种方法来评价某一本体的性能:(1)比较“使用/不使用”对应本体条件下,该应用性能的变化,进而间接地评价对应本体的质量;(2)将不同方法构建的本体用到该应用中,通过比较该应用在使用不同本体条件下系统性能的差异而间接地比较对应本体的性能.如在文献[25]中,作者采用信息检索任务来评价其所构建的本体.具体而言,作者构建了一个中国古建筑领域的本体,之后,将该本体应用于检索中的“查询扩展(query expansion)”中,并比较使用该本体进行查询扩展及不使用该本体进行查询扩展两种条件下,查询结果的平均准确率及平均召回率指标.通过比较不同条件下的查询结果,来评价所构建本体中的概念以及关系是否准确.

基于本体自身的评价主要是针对所构建本体中的概念及关系等要素分别进行评价.对于本体概念评价,常用的评价指标包括准确率、召回率及 $F1$ 值.对于本体关系评价,可以针对不同类型的关系分别评价其对应的准确率、召回率及 $F1$ 值.

由于第一类本体评价方法需依赖某一具体应用,对本体的评价不直观,因而,研究者较少使用.而第二类本体评价方法,由于其可以非常直观地显示本体的性能,也可以非常方便地比较不同的本体构建方法,因而在目前的本体构建研究工作中被广泛使用.如在文献[2,22-23,82]等的研究工作中就使用了准确率、召回率、$F1$ 值作为所构建本体的评价指标.

近期,也有部分研究者使用准确率及召回率的某种变换形式来进行本体性能的评价.如文献[61]使用类似于 $Recall@k$ 和 $Precision@k$ 的方法作为本体构建的评价指标.在文献[61]中,概念抽取的评价指标为 $AP@20$(Average Precision at rank 20),即对于每一个目标概念,统计系统输出的前 20 个结果概念的准确性.当然,这样评价的前提是对于每一个概念,都是由一类词汇来表示.在对概念关系的评价中,文献[61]所采用的评价公式如下:

$$Recall@k=\frac{1}{|Q|}\sum_{c\in Q}\frac{|S_k(c)\bigcap S_{\text{truth}}(c)|}{|S_{\text{truth}}(c)|} \quad (5)$$

$$Precision@k=\frac{1}{|Q|}\sum_{c\in Q}\frac{|S_k(c)\bigcap S_{\text{truth}}(c)|}{k} \quad (6)$$

在上面公式中,$S_k(c)$ 表示概念 c 的 top-k 个输出结果,Q 表示概念集合,而 $S_{\text{truth}}(c)$ 则表示真实的关系集合.在有标准结果集合的情况下,采用上面 $Recall@k$ 和 $Precision@k$ 的评价方法所得到的结果往往更客观,更能反映系统的真实性能.

3.5　常见的本体构建系统分析

在本节中,我们将介绍当前主流的几个本体构建系统.在选择这些本体构建系统时,我们主要考虑了该系统的适用性、知名度、支持的输出结果等因素.

(1)GRAONTO

GRAONTO[47]是一个全自动的基于图的领域本体构建系统,适用于从普通的无结构文本中进行本体构建.该系统采用统计的方法完成本体构建的各项任务,具体如下:

①文档预处理.在该阶段,与目标领域相关的文档将被转化为后续步骤需要的格式.停用词、低频词将被过滤掉,剩余词的词性、词频、邻接词信息等将被统计出来.

②生成文档图.在这一阶段,语料库中的每一个文档都将被表示为一个图.图中的结点对应术语,边则表示术语之间的关系.而图中结点及边上都标注了对应元素在文档中出现的频率信息.为了避免由于文档大小不同而造成的影响,所有的文档图中的结点与边都进行了归一化处理.

③概念抽取.这一阶段包含两个关键步骤:(a)基于随机漫步的术语权重分配,用来评价一个术语相对于目标领域的重要程度;(b)应用马尔可夫聚类算法来把分配好权重的术语聚成不同的小类,每一小类都被用来表示一个本体概念.

④关系抽取.基于重要的关系一定会在对应的语料库中反复出现这一基本直觉,该系统将关系抽取任务转换为文档图中的高频子图挖掘任务.之后,每一个被挖掘到的高频子图都会被解释为关系.具体而言,首先在高频子图中寻找词性为动词的结点,之后以该结点为中心,寻找其左右邻接结点.如果其左右邻接结点均为名词,则一组关系将被确立.如果该中心动词的邻接结点也为动词,则进行动词合并,并以合并后的动词为中心继续寻找其邻接词.直到该中心动词的左右邻接结点均为名词,进而确立相应的关系.在该关系抽取步骤中,中心动词用来解释关系的属性,而其邻接的名词则被表示为关系所关联的概念.

该本体构建系统使用了 TREC-9 数据集以及一个包含 670 篇关于夹具设计(fixture design)的文档集合进行了方法的验证.实验结果显示,概念抽取的准确率可达到 70.8%左右,而关系抽取的 $F1$ 值

约为 50%. 同时,该系统也在更大规模的数据集(10000 篇文档)上进行了测试. 实验结果显示,当数据规模增加时,系统的性能将迅速下降.

(2) CRCTOL

CRCTOL[58]是另外一个旨在从领域相关的文本中自动进行本体构建的系统. 该系统包含的各个处理模块及其对应的处理技术分别介绍如下:

① 文档预处理. 在该阶段,一个自然语言处理工具包(Stanford 的词性标注工具包和 Berkeley 的句法分析器)被用来进行文档预处理. 输入文档的每个单词都被标注词性以及句法标签.

② 概念抽取. 这一阶段主要采用了先抽取候选概念,再进行过滤的方法. 具体过程如下:(a) 从文档中提取所有的标注为名词或名词短语的多词术语(multiword terms),并将修饰这些多词术语的冠词、形容词等去掉;(b) 对于得到的每一个多词术语,使用一种称为"Domain Relevance Value"的评价方法来计算该术语与目标领域的相关度. 相关度高于一定阈值的多词术语就被认为是一个本体概念.

③ 关系抽取. 该系统使用字符串匹配方法以及"词汇-句法(lexico-syntactic)"模板来抽取层次关系. 在该系统中,研究者共设计了 5 个抽取模板. 同时,一些启发式的字符串匹配规则也和抽取模板一起使用,共同进行层次关系的抽取. 对于非层次关系的抽取,该系统主要是基于句法分析以及词性标注的结果,从输入文档中选择动词,之后,以该动词为中心,从其前、后分别选择对应的名词,进而组成关系三元组.

④ 本体映射. 当概念抽取及关系抽取分别完成后,本体映射过程将把二者合并为最终的目标本体. 映射过程相对简单,首先,求取概念抽取阶段得到的概念集、层次关系抽取得到的概念集以及非层次关系抽取得到的概念集合之间的并集,并将该并集作为初始的本体概念集合;之后,使用层次关系建立目标本体的基本结构框架;其次,使用非层次关系进行关系扩展. 没有关系连接的概念将会被过滤.

系统开发者使用了以下两个数据集合进行了系统性能的验证:(1) 美国国务院发布的全球恐怖主义报告(1991~2002)文档集合,该文档集合共包含 104 个 html 文件,每个文件包含约 1500 个单词;(2) 一个关于足球的数据集,共包含 3542 篇英文文档,该数据集中包含一个基准本体概念集,里面包含 608 个概念. 在第一个测试语料集合中,研究者从 11745 个多词术语中挑选了 200 个作为目标本体概念. 此外,从关系抽取过程中得到的 144 个单个名词也作为概念加入到了目标本体概念集合中. 经过最后的本体映射过程,最终得到了 271 个关系. 对于第二个测试数据集,用 CRCTOL 系统共抽取到了 150 个包含在基准本体中的概念. 对于层次关系,该系统的准确率为 74.0%,而对于非层次关系,得到的准确率为 69.4%.

(3) Text2Onto

Text2Onto[70]是一个可以从文本数据中进行本体构建的工具,也是早期 TextToOnto[83-84]工具的升级版. 其主要功能模块以及使用的相应技术介绍如下:

① 文档预处理. 该阶段的处理主要包括标记化(tokenization)、句子切分、词性标注、词干提取等. 经过这些简单的处理之后,输入文本的格式将被转化为后续步骤所需的格式形式.

② 概念形成. Text2Onto 实现了一些用于评价一个术语与目标领域相关的方法,如相对术语频率(relative term frequency)、Tf-Idf、熵、C-Value/NC-Value 等. 这些评价指标将被作为概念抽取的主要过滤指标.

③ 关系抽取. Text2Onto 采用不同的方法抽取不同类型的关系. 具体如下:

(a) Subclass-of 关系. Text2Onto 实现了几个典型的用来进行 subclass-of 关系抽取的算法. 包括使用 WordNet 的上下位结构进行抽取、文献[85]提出的 Hearst 模式匹配方法、文献[67]采用的启发式方法等. 之后,采用文献[86]提出的方法将不同算法所得到的结果进行合并.

(b) Mereological 关系. Text2Onto 采用文献[87]采用的匹配模式进行 Mereological 关系的抽取. 抽取到的相应关系在所在文档中出现的频率也将被统计并作为相应关系过滤的依据之一. 同时,WordNet 也被用来进行最终 Mereological 关系的确认.

(c) Instance-of 关系. Text2Onto 采用基于相似度计算的方法抽取 instance-of 关系. 对于每一个概念,Text2Onto 均抽取其一定的上下文信息,之后采用文献[88]所采用的方法进行相似度计算,相似度高的概念对之间将定义 instance-of 关系.

(d) Equivalence 关系. 和 instance-of 关系一样,这里也对每一个概念抽取一定长度的上下文,并根据两个概念上下文之间的相似度来决定二者之间

是否存在 equivalence 关系.

(e) General 关系. Text2Onto 采用浅层句法分析来提取一种次范畴化框架(subcategorization frames),并辅助以术语在该框架中出现的频率. 之后,这些次范畴化框架将被通过规则映射为 general 关系. 次范畴化框架形式如下:

(i) transitive, e.g. love(subj,obj)

(ii) intransitive + PP-complement,

e.g. walk(subj,pp(to))

(iii) transitive + PP-complement,

e.g. hit(subj,obj,pp(with))

(iv) ……

例如,下面形式的次范畴化框架 hit(subj:person,obj:thing,with:object)将最终生成如下形式的关系:hit(domain:person,range:thing).

(4) 其它的本体构建系统

除了上面介绍的 3 个本体构建系统之外,还存在另外一些本体构建工具,如 OntoLearn[64]、TextStorm/Clouds[89-90]、ASIUM[91]等,这里不再一一介绍.

3.6 小　结

从第 2 节的讨论我们知道,部分研究者对本体构建研究中存在的问题以及未来可能的研究方向进行了分析. 我们结合近几年本体构建研究的最新成果,对第 2 节中研究者们所提到的关于本体构建未来所应关注的研究点与近几年本体研究的实际情况进行比较,比较结果如表 1 所示.

表 1　已有综述文献中对本体构建未来研究方向的预测与实际研究现状比较

综述文献	结果比较 指出存在的问题或可能的研究方向	当前状态
文献[28]	1. 如何有效利用社交数据进行本体构建.	*
	2. 如何通过设计新的算法而利用网络数据中的结构信息进行本体构建.	×
	3. 如何进行与语言无关的本体中实体的表示研究.	×
	4. 在进行网络规模级本体构建研究中,如何保证算法的有效性以及鲁棒性.	*
	5. 如何进行实用化的本体构建研究.	*
文献[30]	1. 本体公理的学习.	×
	2. 找到可以客观评价本体准确率、算法效率、本体完备性的文法.	×
	3. 应找到更有效的全自动的本体学习方法.	*
	4. 移植性更好的本体构建方法.	*
文献[32]	1. 在本体构建方法上,当前的方法在各类数据上均存在一些需要进一步深入研究的问题. 并且,本体构建方法缺乏通用性,学习方法也应向自动化学习方向努力.	*
	2. 当前一些本体学习工具需进一步完善.	√
	3. 需要一个统一的评价本体构建结果的标准.	√
文献[33-34]	1. 自动、半自动的本体构建技术是未来本体构建研究的主要研究方向.	*
文献[29]	1. 对于从文本中进行本体构建研究而言,并不存在一个通用型、指导性的方法体系.	○
	2. 当前本体构建过程主要是基于自然语言分析技术,并通过具体使用的语料资源来决定整个本体构建过程.	○
	3. 很多方法都会把 WordNet 作为本体构建的初始资源,先通过 WordNet 获取一些初始的概念以及关系,之后,再通过其它技术来进行扩展而得到最后的目标本体.	○
	4. 对于本体构建而言,几乎没有完全自动化的系统,多数方法需要用户的参与来从标注语料库中获取相应的概念以及关系.	√
	5. 需要进一步研究对本体构建进行评价的方法,以便于不同方法之间进行比较.	√
文献[31,35]	1. 应该为本体构建研究提供一些基准任务以及相应的评测机制	○
文献[36]	1. 多数本体仍只能称为本体原型.	○
	2. 本体构建中所用的数据资源中提取到的一些显性的本体关系模式在实际文本中很少能够重现.	○
	3. 语义 web 领域对本体的需求和从文本中学习到的本体存在着很大的差异.	○
文献[37]	1. 本体中公理的学习仍处于最初始阶段.	○
	2. 如何有效地对本体构建过程进行评价仍是一个未解决的问题.	√

注:表中符号解释:"*"表示依然为研究热点;"×"表示对应研究很少有研究者关注;"√"表示对应的问题已在很大程度上得到解决;"○"表示该问题依然存在.

4　问题与挑战

从前面的分析中可以看出,最近几年中,研究者们对本体构建任务进行了较为深入的研究,取得了一定的研究成果. 从表 1 也可以看出,一些之前困扰研究者们的问题现在已经取得了很大的研究进展,但也有一些传统问题仍然存在. 比如多数高性能的本体构建方法的移植性普遍较差,多数方法无法被应用在构建大规模、实用化的本体,本体公理的建立依然需要进行深入的研究等. 此外,随着当前构建本体可用的文本数据的快速扩展,现有的本体构建方

法也面临着如下挑战.

首先,本体的更新问题.一般来讲,本体作为一种常识性通用知识,其所描述的内容具有一定的稳定性.但随着现在网络上各种文本数据的快速增加,会导致即使是在同一领域内,新的概念、现有概念新的属性等本体要素也会不断涌现,进而与这些新概念对应的关系也需要进行调整或重要建立.例如,在影视娱乐领域,新的电影作品、新的演员会不断出现.要使构建的对应领域的本体在实际应用(如问答系统、信息检索等)中有效地发挥作用,就需要快速有效地识别这些新出现的概念,并为之分配合适的关系及属性.定期更新数据源并重新进行本体构建是一种效率比较低的方法,并且由于当前许多本体概念识别方法都依赖于该概念在数据中出现的频率信息,而把新概念和老概念放在统一的数据源中进行重新学习,往往会得到带偏的(biased)结果.文献[82]中曾用如下例子说明了这一问题:"情感分析"和"统计机器翻译"这两个概念,实际上分别是概念"自然语言处理"下面的两个研究方向,但由于"统计机器翻译"在学术期刊上出现的时间早于"情感分析",而且,从研究的广度和深度来看,对"统计机器翻译"相关主题的研究也远大于对"情感分析"主题的研究.因而,传统的关系抽取方法研究中,很容易将"情感分析"这个概念看作是"统计机器翻译"下的一个子概念,进而产生错误的概念关系.文献[92]提出了一种基于语义关联的本体完善方法,但该方法主要是以多本体完善为目标,并没有讨论针对单一本体的更新、完善问题.

第二,本体中关系的消歧问题.目前在本体构建研究中,一类方法是提前根据领域专家的参与而预先定义若干种关系类型,而另一类方法中,关系由核心动词所决定.对于后者,由于在自然文本中,一词多义现象普遍存在,因而一个核心动词也往往具有多个语义,如果不对其进行消歧处理,就会造成由之确定的关系在实际应用中出现语义上的不确定性.而这种语义不确定性将极大地影响本体在实际应用中的性能.因而,对于由核心动词所确定的本体关系,必须要考虑消歧问题.然而当前这一问题很少被研究者所关注.

第三,本体概念属性的自动获取以及概念的消歧问题.目前大多数的本体构建方法得到的本体都是一种扁平化的结构(如图1所示).在这些本体中,概念要么以单个词或词组的形式给出,要么以一个相近术语集合的形式给出.这样的描述形式显然无法深入地揭示概念在一个特定领域中所具有的属性.虽然一些基于人工的本体构建方法可以对一个概念进行较为深入的描述,但显然人工获取的方式无法应用在从大规模数据中进行本体构建的任务.在自动本体构建技术越来越成为当前主流本体构建技术的背景下,如何自动有效地获取类似于人工本体构建中对于本体概念丰富的属性描述无疑是一个无法回避的问题.此外,由单个词或词组构成的概念不可避免地会出现歧义现象,即使是以相近术语集合的形式表示本体概念,也需要对集合中的术语进行消歧处理,以确定其代表的真实语义.本体概念的消歧问题还将直接影响随后的关系抽取任务.但目前,在本体构建研究领域,这些问题尚未获得足够的重视.

5 未来研究方向

本体构建研究未来研究方向将以解决上面指出的当前研究中面临的各种问题与挑战为主,即提出可以进行大规模、实用化、高性能、移植性好的本体构建方法,并具有良好的本体概念以及关系的表现形式.为此,更高效的机器学习方法以及依据文本数据特点进行本体构建的研究必然会成为未来本体构建研究的重要研究方向.

首先,本质上来讲,当前在本体构建中存在的移植性差的问题、难以构建大规模实用化本体的问题等在很大程度上揭示了应用机器学习方法进行自然语言处理的两个固有难题:有限的标注数据和无限的标注需求之间的矛盾;以及有限的人工特征构建能力与无限的实际特征之间的矛盾.之前研究者们解决上面两个问题的方法往往带有很大的局限性,很难使问题得到根本性的解决.而近几年引起学术界广泛关注的深度学习(Deep Learning,也称深度神经网(deep neural networks),文献[93-99])技术为我们有效解决上面两个问题提供了新的工具.深度学习技术本质上是通过学习一种深层非线性网络结构而揭示输入信息所具有的丰富的属性特征.在该结构中,每一层都可以看作是对输入层信息的一种特征抽象,模型学习目标是使这些抽象后的特征越来越逼近信息的本质,进而提升分类或预测的性能.深度学习算法已经在图像和音频处理领域取得了很好的成果,研究者们在自然语言处理领域也进行了卓有成效的探索.利用深度学习技术解决自然语言处理中问题的一大优点就是研究者并不需要在

特征选择上投入太多精力，这部分工作可由模型自身依靠强大的特征表现能力完成. 而且可以从大规模无标注文本数据中学习每个单词所具有的高维属性特征(word embedding)，此高维属性特征可以较好地提示词汇本身所具有的语义特性. 其中，在与本体关系抽取任务相关的关系分类任务中，研究者们已经用深度学习技术进行了研究，并取得了很好的结果. 比如文献[100]使用一种递归神经网(Recursive Neural Network)进行关系分类的研究；文献[101]使用卷积深度神经网(Convolutional Deep Neural Network)来进行实体关系的分类研究. 这些研究均取得了较好的实验结果. 深度学习的特点决定了其在解决现阶段本体构建相关问题方面具有先天优势，应用深度学习进行本体构建无疑将会成为未来本体构建的研究方向之一.

第二，最近知识图谱的兴起给本体构建的研究提供了新的参考，尤其在本体关系消歧、概念属性的自动获取及概念消歧方面. 知识图谱研究中的一个重要研究内容就是从文本中发现实体，并为实体间分配合适的关系类型. 和本体构建中的关系抽取任务不同，在知识图谱的关系挖掘研究中，往往存在一定的训练数据，相应的研究任务也往往会被转化为分类任务. 在知识图谱的关系挖掘具体研究过程中，当前的主流研究方法往往是把一个关系三元组中的实体以及关系分别表示成为向量的形式，之后，这些向量被映射到某些高维空间中，通过高维空间中距离的关系来判断两个实体间可能存在的关系类型. 如在文献[102]中，作者为每一个关系 $\boldsymbol{r}$，都设置一个映射函数 $\boldsymbol{M}_r$，用来将该关系所对应的两个实体从实体空间映射到关系空间中，即 $\boldsymbol{h}_r=\boldsymbol{h}\boldsymbol{M}_r$，$\boldsymbol{t}_r=\boldsymbol{t}\boldsymbol{M}_r$. 之后，关系的分类或是关系元素的补全过程都是基于对如下损失函数的训练：$f_r(\boldsymbol{h},\boldsymbol{t})=\|\boldsymbol{h}_r+\boldsymbol{r}-\boldsymbol{t}_r\|_2^2$. 上面的公式中，$\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}$ 分别表示一个关系三元组$\langle\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}\rangle$中的头实体、关系类型以及尾实体. 采用类似方法进行知识图谱相应任务研究的当前主流方法还包括 TransE[103]、TransH[104]、SE[105]、NTN[106]、SME[107] 等. 这些方法的一个共同核心点就是使用 embedding 的技术将实体(可简单地近似认为是本体中的概念)以及关系类型所具有的属性表示为一个高维向量. 在这个高维向量中，个体词汇所具有的各种属性可以被充分体现. 该方法正好可以较好的解决本文第 4 节分析的关系消歧问题以及概念属性自动获取以及概念消歧等问题. 同时，知识图谱研究中的实体发现任务也和本体构建中的概念抽取任务具有一定的关联. 因而，在可预见的未来，利用知识图谱的研究思路来进行本体概念以及本体关系抽取研究，将会成为未来本体构建的研究方向之一.

第三，我们认为未来研究者们将更关注于利用某些特殊类型的文本数据进行本体构建的研究. 这些特殊类型的数据往往具有获取容易、质量高、属性稳定的特点，利用其构造出来的本体的质量也会更高. 比如在科技文献中，均会包含标题、作者、摘要、正文等信息. 这就使研究者们可以有针对性地提出在此类文本数据上构建本体的有效方法. 如文献[22-23,82]利用科技文献为数据源进行本体构建，文献[70]利用医药数据进行本体构建，文献[108]进行了智能交通系统(Inteligent Transportation Systems)数据上的本体构建研究，文献[109]将本体数据应用到了植物学研究中，文献[110]学习几何领域本体等等. 而在将来，一定会有更多的研究者会投入到类似研究任务中.

此外，表 1 中列举的当前尚未解决的问题，如本体公理的构建、构建本体评测的开放平台等，也将成为未来本体构建研究中重要的研究方向.

6　结　论

在本文中，我们详细分析了当前以文本为数据源进行本体构建研究的国内外研究现状，并分别从本体构建过程中所用到的主导性技术以及“任务-技术”这两个不同角度分析了本体构建当前研究的最新成果. 介绍了当前对本体构建结果的常用评价方法，并对当前一些常见的本体构建系统进行了介绍. 在此基础上，我们对本体构建过程中所面临的问题和挑战进行了讨论，并对未来本体构建的研究方向进行了分析.

参　考　文　献

[1] Gruber T. A translation approach to portable ontology specifications. Knowledge Acquisition, 1993, 5(2): 199-220

[2] Rios-Alvarado A B, Lopez-Arevalo I, Tello-Leal E, Sosa-Sosa V J. An approach for learning expressive ontologies in medical domain. Journal of Medical Systems, 2015, 39(8): 1-15

[3] de Azevedo R R, Freitas F, Rocha R G C, et al. An approach for learning and construction of expressive ontology from text in natural language//Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence(WI)

and Intelligent Agent Technologies (IAT). Washington, USA, 2014: 149-156

[4] Dragoni M, da Costa Pereira C, Tettamanzi A G B. A conceptual representation of documents and queries for information retrieval systems by using light ontologies. Expert Systems with Applications, 2012, 39(12): 10376-10388

[5] Ding L, Finin T, Joshi A, et al. Swoogle: A search and metadata engine for the semantic web//Proceedings of the 13th ACM International Conference on Information and Knowledge Management. Washington, USA, 2004: 652-659

[6] Shi L, Setchi R. Ontology-based personalized retrieval in support of reminiscence. Knowledge-Based Systems, 2013, 45(3): 47-61

[7] Kayed A, El-Qawasmeh E, Qawaqneh Z. Ranking web sites using domain ontology concepts. Information & Management, 2010, 47(7): 350-355

[8] Vallet D, Fernandez M, Castells P. An ontology-based information retrieval model//Proceedings of the 2nd European Conference on the Semantic Web: Research and Applications. Heraklion, Greece, 2005: 455-470

[9] Castells P, Fernandez M, Vallet D. An adaptation of the vector-space model for ontology-based information retrieval. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(2): 261-272

[10] Wang Shan, Zhang Jun, Peng Zhao-Hui, et al. Ontology-based semantic search over relational databases. Journal of Frontiers of Computer Science and Technology, 2007, 1(1): 59-78(in Chinese)
(王珊, 张俊, 彭朝晖等. 基于本体的关系数据库语义检索. 计算机科学与探索, 2007, 1(1): 59-78)

[11] Yang Yue-Hua, Du Jun-Ping, Ping Yuan. Ontology-based intelligent information retrieval system. Journal of Software, 2015, 26(7): 1675-1687(in Chinese)
(杨月华, 杜军平, 平源. 基于本体的智能信息检索系统. 软件学报, 2015, 26(7): 1675-1687)

[12] Wimalasuriya D C, Dou Dejing. Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 2009, 36(3): 306-323

[13] Wache H, Vogele T, Visser U, et al. Ontology-based integration of information—A survey of existing approaches //Proceedings of the IJCAI-01 Workshop: Ontologies and Information Sharing. Seattle, USA, 2001: 108-117

[14] Cui Z, Jones D, O'Brien P. Issues in ontology-based information integration//Proceedings of the IJCAI-01 Workshop: Ontologies and Information Sharing. Seattle, USA, 2001: 141-146

[15] Lenzerini M. Ontology-based data management//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Scotland, UK, 2011: 5-6

[16] Calvanese D, De Giacomo G, Lembo D, et al. Tractable reasoning and efficient query answering in description logics: The DL-Lite family. Journal of Automated Reasoning, 2007, 39(3): 385-429

[17] Hajmoosaei A, Abdul-Kareem S. An ontology-based approach for resolving semantic schema conflicts in the extraction and integration of query-based information from heterogeneous web data sources//Proceedings of the 3rd Australasian Workshop on Advances in Ontologies. Gold Coast, Australia, 2007: 35-43

[18] Ozdikis O, Orhan F, Danismaz F. Ontology-based recommendation for points of interest retrieved from multiple data sources//Proceedings of the International Workshop on Semantic Web Information Management. Athens, Greece, 2011: 1-6

[19] Qiao Dong-Chun, Liu Xiao-Yan, Fu Xiao-Dong, Cao Cun-Gen. An ontology-based recommendation system model. Computer Engineering, 2014, 40(11): 282-287(in Chinese)
(乔冬春, 刘晓燕, 付晓东, 曹存根. 一种基于本体的推荐系统模型. 计算机工程, 2014, 40(11): 282-287)

[20] Liang Jun-Jie, Liu Qiong-Ni, Yu Dun-Hui. Personalization recommendation algorithm for Web resources based on ontology. Journal of Computer Applications, 2014, 34(11): 3135-3139(in Chinese)
(梁俊杰, 刘琼妮, 余敦辉. 基于本体的 Web 资源个性化推荐算法. 计算机应用, 2014, 34(11): 3135-3139)

[21] Hotho A, Maedche A, Staab S. Ontology-based text document clustering. Advances in Soft Computing, 1998, 4(6): 48-54

[22] Ren Feiliang. A cheap domain ontology construction method based on graph generation and conversion method. Journal of Information and Computational Science, 2012, 9(18): 5823-5830

[23] Ren Feiliang. A frequency based mining method of complex concept relations for domain ontology. Journal of Information and Computational Science, 2013, 10(9): 2509-2517

[24] Wang Dong-Sheng, Wang Shi, Wang Wei-Min, et al. Interactive question answering based on ontology and semantic grammar. Journal of Chinese Information Processing, 2016, 30(2): 142-159(in Chinese)
(王东升, 王石, 王卫民等. 基于本体和语义文法的上下文相关问答. 中文信息学报, 2016, 30(2): 142-159)

[25] Qian Li-Ping, Yang Xiao-Ping, Song Yu. Design for ontology knowledge base based on structural members. International Journal of Database Theory and Application, 2015, 8(5): 27-38

[26] John S, Shah N, Smalov L. Incremental and iterative agile methodology (IIAM): Hybrid approach for ontology design towards semantic web based educational systems development. International Journal of Knowledge Engineering, 2016, 2(1): 13-19

[27] Sharma M K, Siddiqui T J. An ontology construction approach for retrieval of the museum artifacts using protégé. International Journal of Computer Science Issues, 2016, 13(4): 47-51

[28] Wong W, Liu W, Bennamoun M. Ontology learning from text: A look back and into the future. ACM Computing Surveys, 2012, 44(4): 1-36

[29] Gomez-Perez A, Manzano-Macho D. A survey of ontology learning methods and techniques. Technical Report of the OntoWeb Project: Deliverable 1.5, 2003

[30] Shamsfard M, Barforoush A A. The state of the art in ontology learning: A framework for comparison. Knowledge Engineering Review, 2003, 18(4): 293-316

[31] Zhou L. Ontology learning: State of the art and open issues. Information Technology and Management, 2007, 8(3): 241-252

[32] Du Xiao-Yong, Li Man, Wang Shan. A survey on ontology learning research. Journal of Software, 2006, 17(9): 1837-1847(in Chinese)
(杜小勇，李曼，王珊. 本体学习研究综述. 软件学报，2006，17(9)：1837-1847)

[33] Ding Y, Foo S. Ontology research and development: Part1—A review of ontology generation. Journal of Information Science, 2002, 28(2): 123-136

[34] Ding Y, Foo S. Ontology research and development: Part2—A review of ontology mapping and evolving. Journal of Information Science, 2002, 28(5): 375-388

[35] Buiterlaar P, Cimiano P, Magnini B. Ontology Learning from Text: An Overview. Ontology Learning from Text: Methods, Evaluation and Applications. Amsterdam: IOS Press, 2005: 3-12

[36] Biemann C. Ontology learning from text: A survey of methods. LDV-Forum, 2005, 20(2): 75-93

[37] Drumond L, Girardi R. A survey of ontology learning procedures//Proceedings of the 3rd Workshop on Ontologies and Applications. Salvador, Brazil, 2008: 13-25

[38] Zang Liang-Jun, Cao Cong, Cao Ya-Nan, et al. A survey of commonsense knowledge acquisition. Journal of Computer Science and Technology, 2013, 28(4): 689-719

[39] Wang Xiang-Qian, Zhang Bao-Long, Li Hui-Zong. Overview of ontology research. Journal of Intelligence, 2016, 35(6): 163-170(in Chinese)
(王向前，张宝隆，李慧宗. 本体研究综述. 情报杂志，2016，35(6)：163-170)

[40] Chen Rung-Ching, Bau Cho-Tscan, Yeh Chun-Ju. Merging domain ontologies based on the WordNet system and fuzzy formal concept analysis techniques. Applied Soft Computing, 2011, 11(2): 1908-1923

[41] Stumme G, Maedche A. FCA-merge: Bottom-up merging of ontologies//Proceedings of the 17th International Conference on Artificial Intelligence. Seattle, USA, 2001: 225-230

[42] Fernandez-Breis J T, Chiba H, del Carmen Legaz-Garcir M, Uchiyama I. The orthology ontology: Development and applications. Journal of Biomedical Semantics, 2016, 7(34): 1-11

[43] Mungall C J, Koehler S, Robinson P, et al. k-BOOM: A Bayesian approach to ontology structure inference with applications in disease ontology construction. bioRxiv, 2016: 1-4

[44] Tang Jie, Liang Bang-Yong, Li Juan-Zi, Wang Ke-Hong. Automatic ontology mapping in semantic Web. Chinese Journal of Computers, 2006, 29(11): 1956-1976(in Chinese)
(唐杰，梁邦勇，李涓子，王克宏. 语义 Web 中的本体自动映射. 计算机学报，2006，29(11)：1956-1976)

[45] Ma Liang-Li, Sun Yu-Fei, Liu Qing. Research on ontology matching on semantic Web. Application Research of Computers, 2017, 34(5): 300-308(in Chinese)
(马良荔，孙煜飞，柳青. 语义 Web 中的本体匹配研究. 计算机应用研究，2017，34(5)：300-308)

[46] Zheng Li-Ping. A Study of Ontology Mapping [M. S. dissertation]. Shandong University of Science and Technology, Qingdao, 2005(in Chinese)
(郑丽萍. 本体映射的研究[硕士学位论文]. 山东科技大学，青岛，2005)

[47] Hou Xin, Ong S K, Nee A Y C, et al. GRAONTO: A graph-based approach for automatic construction of domain ontology. Expert Systems with Applications, 2011, 38(9): 11958-11975

[48] Shih Cho-Wei, Chen Ming-Yen, Chu Hui-Chuan, Chen Yuh-Min. Enhancement of domain ontology construction using a crystallizing approach. Experts Systems with Applications, 2011, 38(6): 7544-7557

[49] Chen Rung-Ching, Liang Jui-Yuan, Pan Ren-Hao. Using recursive ART network to construction domain ontology based on term frequency and inverse document frequency. Expert Systems with Applications, 2008, 34(1): 488-501

[50] Sanchez D, Moreno A. Learning non-taxonomic relationships from web documents for domain ontology construction. Data & Knowledge Engineering, 2008, 64(3): 600-623

[51] Lee Chang-Shing, Kao Yuan-Fang, Kuo Yau-Hwang, Wang Mei-Hui. Automated ontology construction for unstructured text documents. Data & Knowledge Engineering, 2007, 60(3): 547-566

[52] Kang Xiangping, Li Deyu, Wang Suge. Research on domain ontology in different granulations based on concept lattice. Knowledge-Based Systems, 2012, 27(3): 152-161

[53] Gu Tao. Using Formal Concept Analysis for Ontology Structuring and Building [Ph. D. dissertation]. Nanyang Technological University, Singapore, 2003

[54] Haav Hele-mai. A semi-automatic method to ontology design by using FCA//Proceedings of the 2nd International Workshop on Concept Lattices and Their Applications. Ostrava, Czech Republic, 2004: 13-25

[55] Haav Hele-mai. An application of inductive concept analysis to construction of domain-specific ontologies//Proceedings of the VLDB Pre-conference Workshop on Emerging Database Research in East Europe. Berlin, Germany, 2003: 63-67

[56] Obitko M, Snasel V, Smid J. Ontology design with formal concept analysis//Proceedings of the 2nd International Workshop on Concept Lattices and Their Applications. Ostrava, Czech Republic, 2004: 111-119

[57] Ochoa J L, Valencia-Garcia R, Perez-Soltero A, Barcelo-Valenzuela M. A semantic role labelling-based framework for learning ontologies from Spanish documents. Experts Systems with Applications, 2013, 40(6): 2058-2068

[58] Jiang Xing, Tan Ah-Hwee. CRCTOL: A semantic-based domain ontology learning system. Journal of the American Society for Information Science and Technology, 2010, 61(1): 150-168

[59] Zouaq A, Gasevic D, Hatala M. Towards open ontology learning and filtering. Information Systems, 2011, 36(7): 1064-1081

[60] Hsieh Shang-Hsien, Lin Hsien-Tang, Chi Nai-Wen, et al. Enabling the development of base domain ontology through extraction of knowledge from engineering domain handbooks. Advanced Engineering Informatics, 2011, 25(2): 288-296

[61] Fang Quan, Xu Changshen, Sang Jitao, et al. Folksonomy-based visual ontology construction and its applications. IEEE Transactions on Multimedia, 2016, 18(4): 702-713

[62] Song Qiuxia, Liu Jin, Wang Xiaofeng, Wang Jin. A novel automatic ontology construction method based on web data// Proceedings of the 10th International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Kitakyushu, Japan, 2014: 762-765

[63] El Idrissi Esserhrouchni O, Frikh B, Ouhbi B. HCHIRSIMEX: An extended method for domain ontology learning based on conditional mutual information//Proceedings of the 3rd IEEE International Colloquium in Information Science and Technology. Tetouan, Morocco, 2014: 91-95

[64] Missikoff M, Navigli R. Integrated approach to web ontology learning and engineering. IEEE Computer, 2002, 35(11): 60-63

[65] Navigli R, Velardi P. Semantic interpretation of terminological strings//Proceedings of the 6th International Conference on Terminology and Knowledge Engineering. Nancy, France, 2002: 95-100

[66] Velardi P, Fabriani P, Missikoff M. Using text processing techniques to automatically enrich a domain ontology// Proceedings of the ACM International Conference on Formal Ontology in Information Systems, Ogunquit, USA, 2001: 270-284

[67] Velardi P, Navigli R, Cucchiarelli A, Neri F. Evaluation of ontolearn a methodology for automatic learning of domain ontologies. Ontology Learning from Text: Methods, Applications and Evaluation. Amsterdam, Netherlands: IOS Press, 2006

[68] Boonchom V, Soonthornphisaj N. ATOB algorithm: An automatic ontology construction for Thai legal sentences retrieval. Journal of Information Science, 2012, 38(1): 37-51

[69] Navigli R, Velardi P. Learning domain ontologies from document warehouses and dedicated web sites. Computational Linguistics, 2004, 30(2): 151-179

[70] Cimiano P, Völker J. Text2Onto—A framework for ontology learning and data-driven change discovery//Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems. Alicante, Spain, 2005: 227-238

[71] Rios-Alvarado A B, Lopez-Arevalo I, Sosa-Sosa V J. Learning concept hierarchies from textual resources for ontologies construction. Expert Systems with Applications, 2013, 40(15): 5907-5915

[72] Buitelaar P, Olejnik D, Sintek M. A protégé plug-in for ontology extraction from text based on linguistic analysis// Proceedings of the 1st European Semantic Web Symposium. Heraklion, Greece, 2004: 31-44

[73] Ciaramita M, Gangemi A, Ratsch E, et al. Unsupervised learning of semantic relations between concepts of molecular biology ontology//Proceedings of the 19th International Joint Conference on Artificial Intelligence. San Francisco, USA, 2005: 659-664

[74] Cao Xin-Yu, Cao Cun-Gen, Wu Yu-Ming. Acquiring part-whole relation from the Web. Journal of Chinese Information Processing, 2013, 27(2): 26-33(in Chinese)
(曹馨宇, 曹存根, 吴昱明. 从 Web 获取部分整体关系. 中文信息学报, 2013, 27(2): 26-33)

[75] Xia Fei, Cao Xin-Yu, Fu Jian-Hui, et al. Extracting part-whole relations based on coordinate structure. Journal of Chinese Information Processing, 2015, 29(1): 88-96(in Chinese)
(夏飞, 曹馨宇, 符建辉等. 基于并列结构的部分整体关系获取方法. 中文信息学报, 2015, 29(1): 88-96)

[76] Wang Na-Na, Huang Yun-You, Tang Su-Qin, et al. Term relationship acquisition and verification based on OMKast. Application Research of Computers, 2015, 32(11): 3319-3323(in Chinese)
(王娜娜, 黄运有, 唐素勤等. 基于 OMKast 的术语关系获取和验证. 计算机应用研究, 2015, 32(11): 3319-3323)

[77] Ren Feiliang. An unsupervised cascade learning scheme for "cluster-theme keywords" structure extraction from scientific papers. Journal of Information Science, 2014, 40(2): 167-179

[78] Hu Fanghuai, Shao Zhiqing, Ruan Tong. Self-supervised Chinese ontology learning from online encyclopedias. The Scientific World Journal, 2014, (1): 1-13

[79] Benz D. Collaborative Ontology Learning [M. S. dissertation]. University of Freiburg, Freiburg, Germany, 2007

[80] Ma Wen-Feng, Du Xiao-Yong. An evaluation framework for domain ontology. Library and Information Service, 2006, 50(10): 68-71, 75(in Chinese)
(马文峰, 杜小勇. 领域本体评价研究. 图书情报工作, 2006, 50(10): 68-71, 75)

[81] Liao Li-Li, Shen Guo-Hua, Huang Zhi-Qiu, Kan Shuang-Long. Survey on ontology evaluation research. Application Research of Computers, 2015, 32(3): 647-651(in Chinese)
(廖莉莉，沈国华，黄志球，阚双龙. 本体评估方法研究综述. 计算机应用研究，2015，32(3)：647-651)

[82] Ren Feiliang. Learning time-sensitive domain ontology from scientific papers with a hybrid learning method. Journal of Information Science, 2014, 40(3): 329-345

[83] Maedche A, Maedche E, Staab S. The text-to-onto ontology learning environment//Proceedings of the 8th International Conference on Conceptual Structures. Darmstadt, Germany, 2000: 14-18

[84] Maedche A, Volz R. The ontology extraction & maintenance framework: Text-to-onto//Proceedings of the 2001 IEEE International Conference on Data Mining. San Jose, USA, 2001: 1-12

[85] Hearst M A. Automatic acquisition of hyponyms from large text corpora//Proceedings of the 14th Conference on Computational Linguistics. Nantes, France, 1992: 539-545

[86] Cimiano P, Pivk A, Schmidt-Thieme L, Staab S. Learning taxonomic relations from heterogeneous evidence//Proceedings of the ECAI Workshop on Ontology Learning and Population. Valencia, Spain, 2004: 59-73

[87] Berland M, Charniak E. Finding parts in very large corpora//Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Maryland, USA, 1999: 57-64

[88] Lee L. Measures of distributional similarity//Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Maryland, USA, 1999: 25-32

[89] Pereira F C, Oliveira A C, Cardoso A. Extracting concept maps with clouds//Proceedings of the Argentine Symposium of Artificial Intelligence. Buenos Aires, Argentina, 2000: 1-11

[90] Oliveira A, Pereira F C, Cardoso A. Automatic reading and learning from text//Proceedings of the International Symposium on Artificial Intelligence. Kolhapur, India, 2001: 302-310

[91] Faure D, Poibeau T. First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX//Proceedings of the 1st Workshop on Ontology Learning. Berlin, Germany, 2000: 7-12

[92] Han Huan, Feng Zhi-Yong, Chen Shi-Zhan, Huang Ke-Man. Multi-ontology renewal based on semantic association relations. Journal of Shanxi University (Natural Science Edition), 2016, 39(4): 1-11(in Chinese)
(韩欢，冯志勇，陈世展，黄科满. 基于语义关联的多本体完善方法. 山西大学学报(自然科学版)，2016，39(4)：1-11)

[93] Huang E, Socher R, Manning C, Ng A. Improving word representations via global context and multiple word prototypes//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea, 2012: 873-882

[94] Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3: 1137-1155

[95] Mnih A, Hinton G. Three new graphical models for statistical language modelling//Proceedings of the 24th International Conference on Machine Learning. Corvallis, USA, 2007: 641-648

[96] Mnih A, Hinton G. A scalable hierarchical distributed language model//Proceedings of the 21st International Conference on Neural Information Processing Systems. Vancouver, Canada, 2008: 1081-1088

[97] Mikolov T. Statistical Language Models Based on Neural Networks[Ph. D. dissertation]. Brno University of Technology, Czech, 2012

[98] Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 2010: 384-394

[99] Mikolov T, Yih Wen-tau, Zweig G. Linguistic regularities in continuous space word representations//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, USA, 2013: 746-751

[100] Zeng Daojian, Liu Kang, Lai Siwei, et al. Relation classification via convolutional deep neural network//Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland, 2014: 2335-2344

[101] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011, 12: 2493-2537

[102] Lin Yankai, Liu Zhiyuan, Sun Maosong, et al. Learning entity and relation embeddings for knowledge graph completion//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2181-2187

[103] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data//Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 2787-2795

[104] Bordes A, Glorot X, Weston J, Bengio Y. A semantic matching energy function for learning with multi-relational data. Machine Learning, 2014, 94(2): 233-259

[105] Bordes A, Weston J, Collobert R, Bengio Y. Learning structured embeddings of knowledge bases//Proceedings of the 25th AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011: 301-306

[106] Socher R, Chen D, Manning C D, Ng A. Reasoning with neural tensor networks for knowledge base completion//Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 926-934

[107] Bordes A, Glorot X, Weston J, Bengio Y. Joint learning of words and meaning representations for open-text semantic parsing//Proceedings of the 15th International Conference on Artificial Intelligence and Statistics. Pensacola Beach, USA, 2012: 127-135

[108] Gregor D, Toral S, Ariza T, et al. A methodology for structured ontology construction applied to intelligent transportation systems. Computer Standards & Interfaces, 2016, 47: 108-119

[109] Daly L, French K, Miller T L, Nic Eoin L. Integrating ontology into ethnobotanical research. Journal of Ethnobiology, 2016, 36(1): 1-9

[110] Zhong Xiu-Qin, Fu Hong-Guang, She Li, Huang Bin. Geometry knowledge acquisition and representation on ontology. Chinese Journal of Computers, 2010, 33(1): 167-174(in Chinese)
(钟秀琴, 符红光, 佘莉, 黄斌. 基于本体的几何学知识获取及知识表示. 计算机学报, 2010, 33(1): 167-174)

[111] Keet C M. Transforming semi-structured life science diagrams into meaningful domain ontologies with DiDOn. Journal of Biomedical Informatics, 2012, 45(3): 482-494

[112] Zhang Xutang, Hou Xin, Chen Xiaofeng, Zhuang Ting. Ontology-based semantic retrieval for engineering domain knowledge. Neurocomputing, 2013, 116(10): 382-391

[113] Santarelli V. Towards efficient and practical solutions for ontology-based data management//Proceedings of the Joint EDBT/ICDT 2013 Workshops. Genoa, Italy, 2013: 23-30

[114] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations. Secaucus, USA: Springer, 1999

[115] Blaz F, Marko G, Dunja M. Semi-automatic data-driven ontology construction system//Proceedings of the 9th International Multi-Conference Information Society. Ljubjana, Slovenia, 2006: 309-318

[116] Miller G A. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39-41

[117] Cimiano P, Hotho A, Stumme G, Tane J. Conceptual knowledge processing with formal concept analysis and ontologies//Proceedings of the 2nd International Conference on Formal Concept Analysis. Sydney, Australia, 2004: 189-207

[118] Cimiano P, Staab S, Tane J. Automatic acquisition of taxonomies from text: FCA meets NLP//Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining. Cavtat-Dubrovnik, Croatia, 2003: 10-17

[119] Ganter B, Stumme G. Creation and merging of ontology top-levels. Conceptual Structures for Knowledge Creation and Communication, 2003, 2746: 131-145

[120] Rezgui Y. Text-based domain ontology building using Tf-idf and metric clusters techniques. Knowledge Engineering Review, 2007, 22(4): 379-403

[121] Philpot A G, Fleischman M. Semi-automatic construction of a general purpose ontology//Proceedings of the International Lisp Conference. New York, USA, 2003: 1-8

[122] Lammari N, Metais E. Building and maintaining ontologies: a set of algorithms. Data & Knowledge Engineering, 2004, 48(2): 155-176

[123] Chen Rung-Ching, Chuang Cheng-Han. Automating construction of a domain ontology using a projective adaptive resonance theory neural network and Bayesian network. Expert Systems, 2008, 25(4): 414-430

[124] Faure D, Nedellec C. A corpus based conceptual clustering method for verb frames and ontology acquisition//Proceedings of the LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications. Granada, Spain, 1998: 5-12

[125] Drymonas E, Zervanou K, Petrakis E G M. Unsupervised ontology acquisition from plaintexts: The OntoGain system //Proceedings of the Natural Language Processing and Information Systems, and the 15th International Conference on Applications of Natural Language to Information Systems. Cardiff, UK, 2010: 277-287

[126] Guarino N. Formal ontology and information system//Proceedings of the International Conference on Formal Ontology in Information Systems. Trento, Italy, 1998: 3-15

[127] Chen Rung-Ching, Huang Yun-Hou, Bau Cho-Tsan, Chen Shyi-Ming. A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. Expert Systems with Applications, 2012, 39(4): 3995-4006

[128] Jiang Xing, Tan Ah-Hwee. Mining ontological knowledge from domain-specific text documents//Proceedings of the 5th IEEE International Conference on Data Mining. Chicago, USA, 2005: 665-668

[129] Brunner Jean-Sebastien, Ma Li, Wang Chen, et al. Explorations in the use of semantic web technologies for product information management//Proceedings of the 16th International Conference on World Wide Web. Banff, Canada, 2007: 747-756

[130] Tao Teng-Yang, Zhao Ming. An ontology based information retrieval model for vegetables e-commerce. Journal of Integrative Agriculture, 2012, 11(5): 800-807

[131] Fernández M, Cantador I, López V, et al. Semantically enhanced information retrieval: An ontology-based approach. Web Semantics: Science, Services and Agents on the World Wide Web, 2011, 9(4): 434-452

[132] Zhang Yi, Vasconcelos W, Sleeman D. OntoSearch: An ontology search engine//Proceedings of the 24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. London, UK, 2004: 256-259

[133] AMDC Moura , MC CavalcantiLachtim F A, Moura A M C, Cavalcanti M C. Ontology matching for dynamic publication in semantic portals. Journal of the Brazilian Computer Society, 2009, 15(1): 27-43

[134] Uschold M, Gruninger M. Ontologies: Principles, methods and applications. Knowledge Engineering Review, 1996, 11(2): 93-155

[135] Gu Fang, Cao Cun-Gen. Ontology research and existing problems in knowledge engineering. Computer Science, 2004, 31(10): 1-10(in Chinese)
(顾芳，曹存根. 知识工程中的本体研究现状与存在问题. 计算机科学，2004，31(10)：1-10)

[136] Wu Xin-Dong, He Jin, Lu Ru-Qian, Zheng Nan-Ning. From big data to big knowledge: HACE+BigKE. Acta Automatica Sinica, 2016, 42(7): 965-982(in Chinese)
(吴信东，何进，陆汝钤，郑南宁. 从大数据到大知识：HACE+BigKE. 自动化学报，2016，42(7)：965-982)

[137] Lu Ru-Qian, Ying Ming-Sheng. A model for knowledge inference. Science in China (Series E), 1998, 28(4): 363-369 (in Chinese)
(陆汝钤，应明生. 知识推理的一个模型. 中国科学(E辑)，1998，28(4)：363-369)

[138] Cao Yu-Juan, Niu Zhen-Dong, Zhao Kun, Peng Xue-Ping. Near duplicated web pages detection based on concept and semantic network. Journal of Software, 2011, 22(8): 1816-1826(in Chinese)
(曹玉娟，牛振东，赵堃，彭学平. 基于概念和语义网络的近似网页检测算法. 软件学报，2011，22(8)：1816-1826)

[139] Yu Shan-Shan, Su Jin-Dian, Yi Fa-Ling. Descriptions for ontologies based on category theory. Computer Science, 2016, 43(5): 42-47(in Chinese)
(余珊珊，苏锦钿，易法令. 基于范畴论的本体描述方法. 计算机科学，2016，43(5)：42-47)

[140] Yan Hong-Can, Zhang Feng, Liu Bao-Xiang. Granular computing learning model of ontology building. Computer Engineering and Applications, 2017, 53(1): 87-91(in Chinese)
(阎红灿，张奉，刘保相. 本体构建的粒计算学习模型. 计算机工程与应用，2017，53(1)：87-91)

[141] Li Wen-Qing, Sun Xin, Zhang Chang-You, Feng Ye. A semantic similarity measure between ontological concepts. Acta Automatica Sinica, 2012, 38(2): 229-235(in Chinese)
(李文清，孙新，张常有，冯烨. 一种本体概念的语义相似度计算方法. 自动化学报，2012，38(2)：229-235)

[142] Liu Bai-Song. A Study on Web-Based Domain Independent Ontology Learning [Ph. D. dissertation]. Zhejiang University, Hangzhou, China, 2007(in Chinese)
(刘柏嵩. 基于 Web 的通用本体学习研究[博士学位论文]. 浙江大学，杭州，2007)

[143] Wang Ru-Juan. Research on Ontology Mapping Methods [Ph. D. dissertation]. Jilin University, Changchun, 2012 (in Chinese)
(王茹娟. 本体映射的若干方法研究[博士学位论文]. 吉林大学，长春，2012)

[144] Chen Hong. The Research on Ontology-Based Knowledge Representation [M. S. dissertation]. Changsha University of Science & Technology, Changsha, 2006(in Chinese)
(陈宏. 基于本体的知识表示研究[硕士学位论文]. 长沙理工大学，长沙，2006)

[145] Chen Jian. Research on Creation and Application for Domain Ontologies [M. S. dissertation]. University of International Business and Economics, Beijing, 2006(in Chinese)
(陈建. 领域本体的创建和应用研究[硕士学位论文]. 对外经济贸易大学，北京，2006)

[146] Zhang Zhi-Gang. The Research and Application of Domain Ontology Construct Methods [M. S. dissertation]. Dalian Maritime University, Dalian, 2008(in Chinese)
(张志刚. 领域本体构建方法的研究与应用[硕士学位论文]. 大连海事大学，大连，2008)

[147] Gong Zi. Research on Ontology Reasoning Based on OWL [M. S. dissertation]. Jilin University, Changchun, 2007(in Chinese)
(龚资. 基于 OWL 描述的本体推理研究[硕士学位论文]. 吉林大学，长春，2007)

REN Fei-Liang, born in 1976, Ph.D., associate professor. His main research interests include natural language processing and ontology construction.

SHEN Ji-Kun, born in 1991, M. S. candidate. His main research interests focus on natural language processing and knowledge graph construction.

SUN Bin-Bin, born in 1992, M. S. candidate. His main research interests focus on natural language processing and knowledge graph construction.

ZHU Jing-Bo, born in 1973, Ph. D., professor, Ph. D. supervisor. His main research interests include natural language processing and machine translation.

Background

As an important knowledge base, ontology is of great help for improving the performances of many information-based applications. Thus constructing large scale ontologies quickly and efficiently is attracting more and more research attentions. Among these research efforts, constructing ontologies using text as data sources is being a hot research recently. Focusing on the (semi-) automatic ontology construction methods, this paper makes a thorough review on

the latest research results of these methods up to now. The main contribution of this paper is to provide researchers following five aspects of comprehensive information about ontology construction. First, what is the state-of-the-art technologies used in ontology construction? Second, what is the evaluation matrices used in ontology construction? Third, are there any representative systems on ontology construction? Fourth, what are the problems and challenges in ontology construction research? Finally, what are the future directions for ontology construction?

To make our review clear and comprehensive, in the main part of this paper, we review the state-of-the-art ontology construction methods from a "global" perspective and a "local" perspective. In the "global" perspective, we classify the ontology construction methods into statistical based methods and linguistic analysis based methods. We introduce these two kinds of methods one by one in detail. And their advantages and disadvantages are also analyzed. In the "local" perspective, we divide the whole ontology construction process into several sub-tasks: term extraction, concept extraction, relation (including hierarchical relations and non-hierarchical relations) extraction, and ontology formalization. Then the techniques used in these sub tasks are reviewed one by one. In the third part of this paper, we review the widely used evaluation methods for ontology construction, and introduce the lasted research results about the evaluation methods for ontology construction. In the fourth part of this paper, we introduce several representative and well known ontology construction systems. For each system, we ignore the technical details and only focus on the sub-tasks involved in their ontology construction processes and the outputs generated by them. In the fifth part of this paper, the challenges and problems in ontology construction are discussed. In the final part, we point out several possible research directions for ontology construction based on some latest research results in the fields of machine learning and natural language processing.

Currently, our research group's research interests mainly focus on artificial intelligence field, specifically, including ontology construction, knowledge representation, knowledge graph construction. Ontology construction is an important research topic in artificial intelligence. This paper's work would be of great help for ontology construction researchers.

Our research work is supported by the National Natural Science Foundation of China (Nos. 61572120, 61300097 and 61432013).