《智能信息处理》课程作业

语义网环境下的本体学习技术研究

王麒博

作业	分数[20]
得分	

2020 年 11 月 13 日

语义网环境下的本体学习技术研究

王麒博

（大连海事大学 信息科学技术学院，大连 116026）

摘　要：本体论是一门重要的新兴学科,现已成为知识工程、自然语言处理、信息系统、智能系统集成和知识管理等多个领域的热门研究方向。它为人们及广泛异构的应用系统提供共同的领域知识理解,并为第三代互联网(语义网,Semantic Web)中基于内容的知识获取、互用和交流提供高质量的保证。而目前来说本体构造的主流手段仍然是效率和准确率都非常低的手工构造，这很容易导致知识获取的瓶颈。近年来，因为领域本体自动创建可以克服手工方法的不足，成为当前的研究热点之一；而本体学习是自动或半自动构建本体的一系列方法和技术。本文探讨目前用于实现语义网的各种常见本体学习技术、方法，比较当前融合多种本体学习技术的本体学习系统。

关键词：语义网;本体;本体学习;本体学习技术

中图法分类号：TP311　　**文献标识号：**A

Research on Ontology Learning Technology in Semantic Web Environment

Wang Qibo

(College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China)

Abstract: The creation of semantic web needs a common standard concept system, namely ontology. At present, the main method of ontology construction is still manual construction with very low efficiency and accuracy, which easily leads to the bottleneck of knowledge acquisition. In recent years, because the automatic creation of domain ontology can overcome the shortcomings of manual methods, it has become one of the current research hotspots; and ontology learning is a series of methods and technologies for automatic or semi-automatic ontology construction.This paper discusses a variety of common ontology learning technologies and methods used to realize semantic web, and compares the current ontology learning systems integrating various ontology learning technologies.

Key words: Semantic Web; ontology; ontology learning; Ontology learning technique

1 引言

本体是对客观的事物以一种形式化的、客观的并且系统化的方式进行描述。本体由哲学领域发起，对现实世界的客观事物进行本质化的描述。它在哲学中的定义为“对世界上客观存在物的系统地描述,即存在论”,是客观存在的一个系统的解释或说明,关心的是客观现实的抽象本质，现在较多的翻译为本体论。互联网创始人 Tim Beruers-Lee 于 1998 年提出了语义互联网的概念。由于本体机制是语义网知识组织的核心技术，因此 20 世纪 90 年代就开始了发展，并且受到了很多人的注意。知识的构建以及网络、数据和语义索引的研究和实践已经成为多个领域的研究焦点。

本体构建是网络数据语义至关重要的基础，它是一项庞大的工程。这需要不同领域的专家(专家、工程师等)在适当方法指导下，按照一定的本体构建原则并采用本体开发工具来实现。但是，现在被广泛使用的本体开发工具仅仅能够提供本体元素的实体编辑和管理功能，还必须手工构建本体。我们需要逐个输入和编辑领域中每个概念以及概念的名字、约束、属性以及关系等内容以生成本体。 手工方式构建本体的方法费时、费力，构建过程带有片面性。在此背景下，本体学习成为了一个非常有用的研究方向，它专门用于利用自动知识识别技术来降低本体构建的成本。本体学习利用语言分析、机器学习、数学统计、数据挖掘等技术通过计算机自动或半自动地从大量数据源(包括非结构化、半结构化、结构化数据)中发现潜在的概念和概念问的关

系，再由领域专家适当进行辅助修正和评价，以获取期望的知识本体。

目前，本体学习方面的研究在全世界非常活跃。虽然提出了许多本体学习方法，但大多数方法并不理想。由于缺乏统一的本体学习体系结构和方法，虽然开发了一些本体学习方法，但都难以被其他系统重用。

2 本体学习方法

本体学习的关键在于概念抽取和概念关系抽取的方法。目前本体学习的抽取算法包括概念抽取、概念定义抽取、概念实例抽取、概念层次(类与子类)关系抽取、部分与整体关系抽取、同义关系抽取等算法。

2.1 概念抽取

概念术语的自动提取和识别方法包括浅层解析技术、互信息（MI）、TFIDF、RTF、熵 H 和 C/NC 方法。首先，我们从文本中提取概念候选。接下来，使用 tfiidif、RTF、信息熵或 C/NC 值来确定主题或术语概念。目前，大多数本体学习系统都是基于浅层分析技术和统计规则方法来提取概念的。例如，text on、on、dogma 和 Liu-baisiong 本体学习系统使用浅层解析技术从文本中提取候选词。浅层句法分析法是一种发现语篇中单词之间的语法关系（宾语谓语和动宾关系）的方法，它检测句子中单词的边界并标记词性。

2.2 概念定义

除了直接从术语词典或知识库中检索外，模式匹配通常用于获取概念定义。Velardi 等在评估报告中描述了这两种概念定义获取方法。一方面通过在线术语表，另一方面，通过在线术语、模式匹配规则和语法分析器，从文档中提取概念术语的定义，定义与过滤规则剪枝无关的概念规则或非正式术语的定义。此外，在我国也有相应研究，如徐勇使用扩展的 BNF 来表示术语的定义，以及使用这些模式从互联网语料库中定义的术语，以及未定义的的排除模式，对相关术语进行了定义和人工总结。

2.3 概念实例

在 text2on 上使用相似性计算方法。首先，我们从文本集中提取实例和概念的向量空间表示，然后使用 sketch diversity 提出的相似度索引和最高向量作为例子计算实例和概念的向量相似度。赋予具有相似性的概念。因子为 32.6%。

2.4 部分与整体的关系

概念间部分与整体关系的获取主要采用模式匹配方法。Chamiak 等描述了在大量语料中发现部分 与整体关系的模式，Cimiano 在 Text2Onto 系统中开发的 JAPE 模式引用了这一模式，通过计算部分与整体关系的模式共现率指示概念术语之间部分与整体关系的概率。也可借助通用本体如 WordNet 中的语义关系推理概念术语间的部分与整体关系。

2.5 同义关系

概念间的同义关系，一方面可利用通用本体或其他义类辞典来推理获取；另一方面可假设同义概念在文本中具有相同的上下文句法结构，因此可使用浅层解析器等抽取方法比较概念在上下文中的特征，然后计算两者的相似性，以指示两者同义关系的概率。此方法已在 Text2Onto 中运用。

综上所述，目前本体学习算法主要应用了自然语言处理的浅层解析技术、模式匹配以及计算语言学统计方法和机器学习方法。

3 本体学习关键技术

本体学习技术是针对语义网的实现提出来的。语义网试图使用知识本体来描述网络中的海量数据,从而达到解析信息语义的目的。面向语义网的知识本体来源于网络中各种类型数据，如文本、网页、XML 文档、数据库文件等。针对不同结构化程度的数据源,适用的本体学习技术各有差异,能够获得的本体元素(如概念、概念间关系和公理等)也不尽相同。针对不同类型的数据源需要采用不同的本体学习技术，所以本文根据数据源的结构化程度，将本体学习技术分为 3 大类：基于结构化数据的本体学习技术、基于非结构化数据的本体学习技术和基于半结构化数据的本体学习技术。

3.1 基于非结构化数据的本体学习技术

非结构化数据没有固定结构。纯文本就是一类典型的非结构化数据，是 目前基于非结构化数据

自动抽取本体的主要研究对象。纯文本依据一定的造句法表达特殊丰富的语义，使得读者可以基于一些背景知识来理解其中的含义。但是，由于纯文本缺乏一定的结构，要使机器能够理解并从中抽取出知识，则必须利用自然语言处理(NLP)技术对其进行预处理，然后利用数据挖掘、机器学习等手段从中获取知识。本文把基于纯文本的本体构建过程分为三个部分，即：概念抽取、概念关系识别和公理生成。

对于概念的获取，现有的方法可以分为 3 类：基于语言学的方法、基于统计的方法和混合方法。

(1)基于语言学的方法主要根据领域概念的特殊词法结构或模板，寻找和抽取结构符合这些特定模板的字符串。由于这些 模板在大多数情况下是与具体语言相关的，因此，这类方法要求针对具体的语言作相应的处理。

(2)基于统计的方法主要根据领域概念与普通词汇拥有不同的统计特征(例如，领域相关性和领域通用性)，以鉴别出领域概念。大多数基于统计的方法关注于多字词汇(multi word unit)的抽取，主要方式是计算各组成部分之间的联系程度。

(3)混合方法往往是结合语言学和统计学的技术，有的是在统计处理之后采用语法过滤器，以便抽取出经过统计计算有意义的、与给定词法模板匹配的词汇组合；有的则是首先采用语言技术选出候选项，然后再用统计方法对这些候选项进行计算。

3.2 基于半结构化数据的本体学习技术

半结构化数据是指具有隐含结构，但缺乏固定或严格结构的数据。Web 中的半结构化数据很多，例如大量的 XML 格式和 HTML 格式的网页，以及它们遵循的文档类型定义(XML schema 或 DTD)，还有越来越多的用 RDF 标注的网页，都可以作为本体学习的数据源。

对于 XML，HTML 和 RDF 等格式的网页，可以直接使用那些从纯文本中获取本体的方法。实际上，机器可读的词典(MRD)也是一种特殊的半结构化数据。作为一种通过手工方式认真组织的可靠的领域知识资源，它们也是一种非常好的本体学习数据源。这类数据源的 内部结构虽然在很大程度上也是一种纯文本，但对于领域概念及其关系的抽取来说，仍有很多规律可循。另外，随着语义 Web 的发展，Web 中会出现越来越多的用 OWL，RDF(S)等语言描述的本体，它们也是一种半结构化的数据。

3.3 基于结构化数据的本体学习技术

结构化数据是指具有固定的严格结构的数据格式，目前主要指关系型数据库中的数据。关系数据库采用 关系模型，实体以及实体间的联系都用表来表示。因此，无论是概念的获取还是概念间关系的获取，首先必须区分出描述实体(或实体关系)的表，然后才能将实体(或实体关系)映射为本体中的概念(或概念间关系)。

基于关系型数据库获取本体的基本方法是采用关系数据库逆向工程(Relational Data—base Reverse E．gi—neering)，即获取关系模型的物理语义结构，并将其重新 设计成更复杂的逻辑模型或概念模型的一系列技术的总称。逆向工程虽然并不是针对本体构建提出来的，但是它的指导思想以及一些方法都可以应用到将关系型 数据库转化为本体的开发中。

4 本体学习系统

目前，国外对本体学习的研究很活跃，已经尝试将各种自动化技术和数学方法融合到一个系统中，完成对不同结构化程度数据源的充分而准确地本体学习,并且已经取得了一定进展,出现了如 Text-To-Onto、OntoLearn、Hasti、OntoBuilder、OntoLiFT 等具有一定实用价值的本体学习系统；在中文本体学习研究中，也出现了 GOLF 等实验性系统。

4.1 典型的本体学习系统

（1）Text-To-Onto

Tex-To-Onto 是 University of Karlsruhe 开发的一个整合的本体学习工具。其主要特点是可以支持从多种数据源中获取本体。目前，它已经可以做到从非结构化数据(纯文本)和半结构化数据(HTML，词典)中获取概念及其关系。对于从非结构化数据中学习本体，它使用加权的词频统计方法来获取概念，使用基于概念层次聚类法来获取分类关系，使用基于关联规则的方法来获取非分类关系；对于 HTML 数据，它将其预处理成纯文本，然后利用基于非结构化数据的本体学习方法从中获取本体；对于词典，它使用基于模板的学习方法。该系统能够

处理德文和英文的数据源。

（2）OntoBuilder

OntoBuilder 是 Mississippi State University 开发的一个从 XML 和 HTML 中获取本体(包括概念及其关系)的工具。它看起来像一个 Web 浏览器、当使用它来获取本体之前，需要手工构建一个初始的领域本体；然后，在用户浏览包含相关领域信息的网站的过程中，该工具会为每个网站生成一个候选本体；最后，在用户的参与下将这些候选本体与初始本体合并。其中，使用的本体学习方法主要是词频统计和模式匹配(包括子串匹配、内容匹配、词典匹配)。OntoBuilder 可以支持英文的网页，但在实际中，它并不能适用于所有的网站，因为有些网站包含了它不支持的技术，例如带有脚本(scripting)的网页。

（3）OntoLiF

OntoLiF 是 University of Karlsruhe 开发的一个从半结构化数据(XML schema，DTD)和结构化数据(关系数据库)中获取本体(包括概念及其关系)的工具．对于这两种类型的数据源，它都采用基于映射规则的方法来获取本体、在系统实现中，从 XML Schema 和 DTD 中获取本体的部分是基于一个已有的工具(hMarfra)。HMarfra 能够实现从 XML Schema 到本体的映射。然后，OntoLift 开发了一个从 DTD 到 XML Schema 映射的中间工具。这样，将这两个工具合并起来，就实现了从 XML Schema 和 DTD 中获取本体。从关系数据库中获取本体的部分是基于 Java JDBC 标准提供的接口，然后按照一定的命名规范将数据库中的表名和属性名等信息，按照映射规则转换为本体中的元素。

（4）GOLF

GOLF 是浙江大学刘柏嵩博士开发的一个基于分层循环技术的通用多策略的实验性本体学习系统。其基本思想是：针对 Web 中存在的半结构化数据(包括 HTML 和 XML 文档)，基于“术语→概念和实例→概念分类体系→概念间非分类语义关系→规则和公理”的分层学习技术路线，采用模式匹配、关联规则、层次聚类等技术，自动构建本体。

GOLF 系统的体系结构主要分为 4 个部分：

①通用本体学习模块：包括文档的收集和预处理、领域术语及概念抽取、语义关系获取及优化、构建分类层次体系 等子模块及相关的算法库和词典库；

②通用本体库：存放本体的基本概念及其分类关系和非分类语义关系；

③本体修剪和评价模块：在本体工程师和领域专家的参与下，评价本体学习算法的性能，推断系统对本体构建的作用，并通过与标准本体的比较评定基于 GOLF 获得的本体的领域覆盖度；

④本体的形式化表示模块：将本体概念及其关系以 OWL 形式加以描述。在与 Text-to-Onto 的比较中，GOLF 系统呈现出较好的实验结果。

4.2 本体学习系统比较

综上所述，笔者发现：不同的本体学习系统，其支持的数据源格式、采用的本体学习技术、本体学习过程、自动生成的本体内容以及支持的语种处理等都有所差异。

支持非结构化和半结构化数据源的本体学习系统较多。OntoBuilder 和 OntoLiFT 不支持文本数据源，而 GOLF 虽然面向网页，但从处理过程来看，该系统采用了自然语言处理技术同样适用于文本数据；On- toLearn 和 Hasti 目前只能处理文本数据；6 个系统中只有 OntoLiFT 支持结构化的关系数据库。到目前为止，还没有一种本体学习系统能够支持所有格式的数据源，目前研究主要集中在对非结构化和半结构化格式的数据上。

自动生成公理的系统很少。在上述 6 个系统中，只有 Hasti 系统支持从自由文本中获得公理。公理是本体逻辑推理的基础。从开发时间上看，Hasti 开发较晚，表明公理的识别目前已经逐渐受到重视。Hasti 系统中的公理识别还存在较大缺陷，只针对语法比较特殊的波斯语，但它为今后研究提供了一条思路。

5 结论

随着本体学习技术的提出，本体的构建不再停留在繁杂而耗时、耗力的手工行为阶段上，自动化构建本体成为了本体开发的主要发展方向。研究实践表明，采用各种数学模型、自然语言处理技术和数据挖掘方法，基于不同结构化程度的来源数据，可以自动获得各种本体元素，从而在一定程度上实现知识本体的自动构建。

目前国外已经出现将各种本体学习技术集成在一起，具有一定成熟度的本体学习系统。然而，系统自动构建的不确定性使得获得的本体的准确

性和可靠性往往需要通过领域专家进行人为的调整和修正。于是，以本体学习技术为主体，以领域专家参与为辅助的半自动化的本体构建成为目前本体开发的主流。

参 考 文 献

[1] 林龙成.语义网中 OWL 本体概述及其构建方法研究[J].电脑知识与技术,2020,16(12):203-204.

[2] Jian-Jun Qi. Attribute reduction in formal contexts based on a new discernibility matrix. Journal of Applied Mathematics and Computing. 2009, 30(1-2): 305-314.

[3] 李京杰.基于语义本体的个性化学习推荐研究[J].软件导刊(教育技术),2016,15(09):77-78.

[4] 余霞,刘强,叶丹.基于规则的关系数据库到本体的转换方法[J].计算机应用研究,2008(3):767 -770,785.

[5] Fortuna B , Grobelnik M, Mladenic D. OntoGen :Semi - automati c Ontology [A] / / Smith M J, Salvendy G (Eds.). Human Interface[C],Part II , HCII 2007 , LNCS 4558 , 2007:309 -318 .

[6] 刘小乐,马捷.语义网环境下基于本体的知识集成研究进展[J].现代情报,2015,35(01):159-163+169.

[7]]Hearst M A． Automated Discovery of WordNet Relations[A] Christiane F． WordNet：An Electronic Lexical Database[M]， MIT Press，1992：132—152．

[8] 杜小勇,李曼,王珊.本体学习研究综述[J].软件学报,2006,17(9),1837-1847 .

[9] Deng ZH，Tang SW，Zhang M，Yang DQ，Chen J．Overview of ontology． Acta Scientiarum Naturalium Universitatis Pekinensis，2002，38(5)：730-738(in Chinese with English abstract