面向语义网的本体学习技术和系统研究*

王 昊[1] 刘建华[2] 苏新宁[1] 杨建林[1]
[1](南京大学信息管理系 南京 210093)
[2](南京政治学院基础部 南京 210003)

【摘要】针对网络中存在的不同结构化程度的数据,探讨目前用于实现语义网的各种常见本体学习技术、方法及其可获得的本体元素、存在的问题,比较当前融合多种本体学习技术的本体学习系统,分析所采用的关键技术、适用的处理对象以及生成的结果描述。

【关键词】语义网 本体学习 本体学习技术 本体学习系统 结构化程度
【分类号】TP182

Research on Techniques and Systems of Ontology Learning for Semantic Web

Wang Hao[1] Liu Jianhua[2] Su Xinning[1] Yang Jianlin[1]
[1](Information Management Department of Nanjing University, Nanjing 210093, China)
[2](Basis Department of PLA Nanjing Institute of Politics, Nanjing 210003, China)

【Abstract】According to structured degree of data on Web, this paper discusses all kinds of techniques, methods and gettable ontology elements, existing problems in ontology learning, which always used to achieve semantic web. It also introduces and compares existing ontology learning systems which integrate with multi - techniques, and analyzes their adopted key techniques, objects in point and result description.

【Keywords】Semantic Web Ontology learning Ontology learning technique Ontology learning system Structural degree

1 引 言

互联网创始人 Tim Berners - Lee 在 1998 年首次提出了语义网(Semantic Web, SWeb)的概念[1],20 世纪 90 年代被引入知识工程领域的本体机制作为 SWeb 知识组织层面上的核心技术开始被人们广泛关注。实现 SWeb 的两个重要环节,知识本体的构建及网络数据的语义标引研究和实践逐渐成为了各大领域的探讨热点。

本体构建是网络数据语义标引的基础,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,采用支持完整开发过程的便捷的本体开发工具加以实现。然而,现有的各种本体开发工具(如 Protégé、KAON 等)提供的仅仅是本体元素的编辑和管理功能,支持的是手工构建本体的方式,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。

在这种背景下,利用知识自动获取技术来降低本体构建的开销成为了一个很有意义的研究方向,该研究称为

收稿日期:2008 - 10 - 12
收修改稿日期:2008 - 11 - 09
*本文系"江苏省研究生培养创新工程 - 科技创新计划"人文社科项目"基于本体的学术资源网络模型研究"(项目编号:CX07B - 252r)的研究成果之一。

本体学习(Ontology Learning),即利用语言分析、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间的关系,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。本文主要对目前常用的针对不同结构化程度数据资源的本体学习技术,以及现存的并取得较好实用效果的本体学习系统进行系统探讨和综合比较。

2 本体学习的关键技术

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。

2.1 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。

(1)对于概念的获取,现有的方法可以分为三类:

①基于语言学的方法[2]。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,也可称为基于规则匹配的概念抽取,Hasti系统的自然语言处理采用了这种方式。由于语法模板在大多数情况下与具体语言相关,因此这类方法要求针对具体的语言作相应的处理。鉴于中文语法的复杂性和随意性,往往需要构建大规模的模板进行概念匹配,目前使用具有一定的难度。

②基于数学统计的方法[3-4]。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,以机器学习的方式鉴别领域概念。目前大多数基于统计的方法关注于多字词汇(Multi-Word Unit,MWU)的抽取,主要方式是利用互信息(Mutual Information)[5]计算MWU各组成部分之间的联系程度。

③语言学和统计混合方法[6-7]。结合语言学和统计学的技术获取领域概念。有的是在统计处理之后采用语法过滤器,抽取符合统计意义且与给定词法模板匹配的词汇;有的首先采用语言规则选出候选项,然后再计算候选项的统计意义[8]。

(2)对于概念间关系的识别,常用的方法有以下6种:

①基于模板驱动的方法[9-10]。通过分析领域相关文本,总结出频繁出现的语言模式作为规则,然后判断文本中词的序列是否匹配某个模式。如果匹配,则识别相应的关系,如满足"It is a kind of"模板的概念之间可能具有"IS_A"关系等。这些模式可以是手工定义[11]的,也可以是从某些样本句子或百科辞典中学习得到的[12]。

②基于概念聚类的方法。利用概念之间的语义距离,对概念进行聚类,同簇中的概念具有语义近似关系,如层次聚类,其结果就是概念间的分类关系。基于概念层次聚类构建本体层次结构的研究较多,例如Maedche[13]、Aguirre[14]、Khan[15]、Bisson[16]等都提出采用层次聚类实现概念层次结构的自动生成。然而,基于层次聚类得到的层次结构多为非循环的严格的层次关系,这与本体中一个概念有多个父概念的事实不符,为此,Faure等[17]尝试采用宽度优先的方法对概念进行逐层聚类,试图获得概念的所有父概念。

③基于关联规则的方法。常用于获取概念间的非分类语义关系,其基本思想是:如果两个概念经常出现在同一文档(或段落或句子)中,则这两个概念之间必定存在关系。Maedche等首先描述了在浅层文本分析的基础上使用关联规则挖掘概念间关系的具体过程[18],随后将其置入本体学习系统Text-To-Onto中,在给定类层次结构作为背景知识的基础上探讨句法结构上相关概念的非分类语义关系。

④基于词典的方法。根据一些现有的词汇词典或本体中定义的同义词、近义词和反义词等知识以及基于词典的启发式的模式匹配[13]来获取概念间的关系。

⑤其他数学方法。除了上面介绍的方法外,形式概念分析(FCA)[19-20]、基于潜在语义索引的奇异矩阵分解(LSI/SVD)[21]、机器学习(如贝叶斯分类、决策树学习等)[22]等数学方法正被尝试应用于本体概念关系自动识别的研究中,并取得了一定的成果。

⑥基于多策略混合的方法。在实际构建本体的活动中,上述关系识别策略往往被混合使用。特别是在各种本体学习系统如Text-To-Onto、OntoLearn、Hasti、OntoBuilder等中,都是融合了各种方法,试图将关系识别的准确率和召回率达到最高。

(3)公理的抽取

目前有关公理抽取研究成果很少,对此最早研究公理抽取的是 Lin 等[23],他们设计了一个 DIRT 模型,提出将 Harris[24]的分布假设理论应用到熟语料的依赖树路径中,认为在一定上下文范围内同时出现的路径具有相似性;在各种本体学习系统中,也只有 Hasti 系统支持从自然语言文本中自动获取公理。该系统采用了基于模板驱动的公理抽取方法,即在对句子结构分析的基础上,应用人工预先定义的模板,获取匹配模板的本体公理;Dagan 等[25]则提出要关注词汇蕴涵规则的识别,促使大量可用于基于文本语料学习本体公理的方法产生。

2.2 基于半结构化数据的本体构建技术

半结构化数据是介于非结构化和结构化数据之间的一种具有隐含结构,但不具备固定或严格结构的数据格式,如 HTML、XML、词典、知识库等都可以认为是半结构化数据。

(1)XML、HTML、RDF 等半结构化文档

可以先使用基于纯本文的本体自动获取技术,然后基于半结构化文档隐含的结构来改善或修正抽取结果,如根据结构设置权重等。Papatheodorou 等人[26]提出了一种从 XML 或 RDF 文档中获取概念间分类关系的方法:

①数据收集并预处理:抽取出表征每篇文档内容的关键词;

②模式发现:基于关键词聚类,将文档集分成簇,保证同簇内的文档内容相似;

③模式后处理并评价:基于统计的方法抽取最能描述每簇文档内容的关键词,将这些关键词作为本体中的概念,并根据先前聚类的结果给出概念间的分类关系。

此外,Doan 等人[27]提出了一种基于模式识别技术从半结构化学习映射图(学习阶段),再根据映射图对新文档资源实行文档匹配(分类阶段)从而构建本体的机器学习方法;Volz 等[28]则提出将 XML 文档转化为一个具有非终端有限集、终端有限集、开始符号集合、生成规则的有限集等四元组的规则树文法,然后将非终端和终端翻译成本体中的概念和角色从而获得 XML schema 中的语义。

(2)机器可读的词典(MRD)

主要采用句法分析、模式匹配和图形映射的方法抽取概念关系。Litkowski[29]通过对词典中每个定义的句法分析,获取概念之间的分类关系;Hearst[30]和 Rigau 等[31]都是使用一组预定义的词汇 - 句法模式自动地从词典中发现词与词之间的新的上下位关系;Jannink 等[32]提出将字典术语转化成图形结构,然后使用统计方法和 PageRank 算法构建层次分类结构。

(3)具有半结构化特征的知识库

根据知识库的具体结构采用一定算法将其转化为本体结构。Suryanto 等[33]提出了一种基于知识库的本体抽取方法,其核心技术是把知识库中所有规则分组成多个类,为每一对类之间的每一个关系计算定量的测量值,这个定量测量值提供了包含、排斥、相似等关系存在的置信度。

2.3 基于结构化数据的本体构建技术

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 Engineering),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。Johannesson 等[34]在 1994 年提出了一种使用映射技术将关系模型转换为概念模型的方法,然后由用户对该概念模型进行反复修正生成最终的本体;2002 年,Stojanovic 等[35]试图使用映射技术从概念数据库模型中构建轻量型本体,他们采用了 5 个步骤来完成移植,包括使用逆向工程从关系型模型中获取关系、属性、属性类型、主关键字、外部关键字和内部依赖等信息,分解上述信息并应用一系列的映射规则(包括本体创建规则、继承规则、关系规则等)构建候选本体,实现模型转化,然后对构建的本体进行评价、验证和精细,最后将关系数据库中的元组数据转化为本体实例并建立实例间关系。Stojanovic 方法是 OntoLiFT 系统的一部分,该系统是目前唯一支持间关系型数据库转化为本体的本体学习系统。

Johannesson 和 Stojanovic 等的研究都是将关系型数据库直接转化为本体,从内容、蕴涵的语义而言没有

发生变化，只是在知识组织方式上发生了变化，构建的是轻量型本体，结构简单，概念间关系不复杂。1999年，Kashyap[36]提出首先将数据库模型分解成主关键字、外部关键字以及内含依赖，再根据逆向工程生成一个初步的本体；然后基于用户提问添加或删除对应实体的属性和实体之间关系，或者创建新的实体作为在数据库中已经存在实体下位类或上位类，进一步丰富本体概念和关系。然而，通过用户提问的方式具有一定的随机性和随意性，质量无法得到保证。2004 年，Astrova[37]通过对关系数据库元组的分析，得到了概念间的“继承”关系，从元组中挖掘隐含语义的思路为基于关系型数据库建立更复杂的重量型本体指明了一个方向。2005 年，Astrova 等[38]提出在无法获得数据库模式信息的情况下，可以通过分析网页中的 HTML 表单（是用户和数据库交互的常用界面）来推断关系数据库的语义，然后利用关系模式和用户的领域知识对获得的语义进行补充。余霞等[39]提出了一种基于规则全自动的实现关系数据库到本体转换的方法，即通过分析关系模式的主键、属性、引用关系、完整性约束和引用关联，针对各类数据库对象分别设置一组关系模式到本体的映射规则，完成模式转换后，再将数据迁移到本体实例上，形成知识库。

3 典型的本体学习系统及其比较

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。

3.1 典型的本体学习系统

(1) Text - To - Onto

Text - To - Onto[13]是 University of Karlsruhe 的 Maedche 等人在 2001 年开发的一个整合性的本体学习工具，在 2005 年被重新设计为 Text2Onto[40]。其基本思路是：收集与目标本体相关的应用领域文档；使用语言技术工具分析标识这些文档；从被标识的文档集合中抽取术语形成术语集合；使用机器学习算法确定这些术语之间可能的关系；最后在一个正式的本体中将确定的术语和关系用类和属性表示出来。

Text - To - Onto 支持从英文和德文的多种数据源中获取本体。它能够采用加权的词频统计、层次聚类、关联规则挖掘、基于模板的学习等方法，从非结构化数据（纯文本）和半结构化数据（如 HTML、XML、词典等）中获取概念及其关系。

Text - To - Onto 的体系结构包含 4 个主要组成部分[41]：

①本体管理组件：用于选择来源数据以及合适的资源处理方法和算法；

②资源处理组件：选择不同的处理策略发现、导入、分析和转换输入各种类型的数据；

③算法库组件：提供关联规则挖掘、形式概念分析以及聚类方法等学习技术和算法支持从不同类型的数据中抽取本体元素，该组件采用多策略学习和结果合并方法，以规范化的统一结构输出本体综合结果；

④协调组件：用于协调和控制各个本体学习组件之间的交互，资源处理和算法库使用。

(2) OntoLearn

OntoLearn[42-43]是 University of Rome 的 Missikoff 等人在 2002 年开发的一个基于非结构化文本的本体学习工具，其主要特点是将语义解释的方法应用到了本体自动构建中。

OntoLearn 的体系结构如图 1 所示[42]，可分为三个过程：

①首先使用基于语言学和统计的方法从专业文本集中抽取出领域相关的专用术语；

②使用 WordNet 和 SemCor 语料库中明确定义的概念对抽取的术语进行语义解释，确定术语之间的分类和其他语义关系；

③获取概念间的分类和其他语义关系，生成领域本体森林，最后使用 SymOntoX 执行本体匹配，将生成的领域本体整合到现存的上层本体中。

以 OntoLearn 系统为中心，Missikoff 等人开发了一个软件环境，能够建立和评估领域本体以支持虚拟用户社区的智能信息集成。此外，在两个欧洲项目中测试了 OntoLearn，在项目中它作为语义交互平台的基础用于小中规模的旅游企业。

(3) Hasti

Hasti[2] 是 Amirkabir University of Technology 的

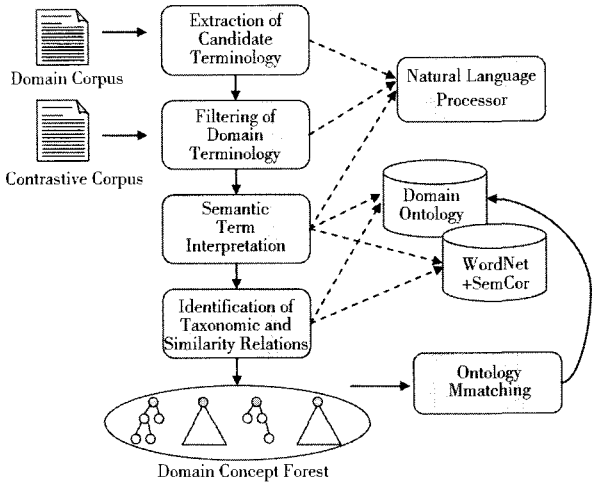

图 1　OntoLearn 的体系结构[42]

Shamsfard 等人在 2004 年开发的一个本体学习工具，它是为数不多的能够学习本体公理的系统。其基本思路是：从初始的核心本体出发，基于文本理解自动地从纯文本中获取新的概念、关系和公理，不断扩充核心本体。其中，核心本体包含少量手工定义的基本概念、分类和非分类关系、推断公理和操作符等基本元知识。使用核心本体的目的是便于对新获取的概念、关系和公理在本体中进行预定位。

Hasti 支持从波斯语文本中学习本体元素。该系统体系结构如图 2 所示，包括自然语言处理器、工作存储器、知识抽取器、知识库、词典管理器和本体管理器 6 个组件，其工作过程如下：

①使用基于语言学方法和句法模板的自然语言处理器对输入文本进行语素构造分析，抽取案例角色，并将获得的术语

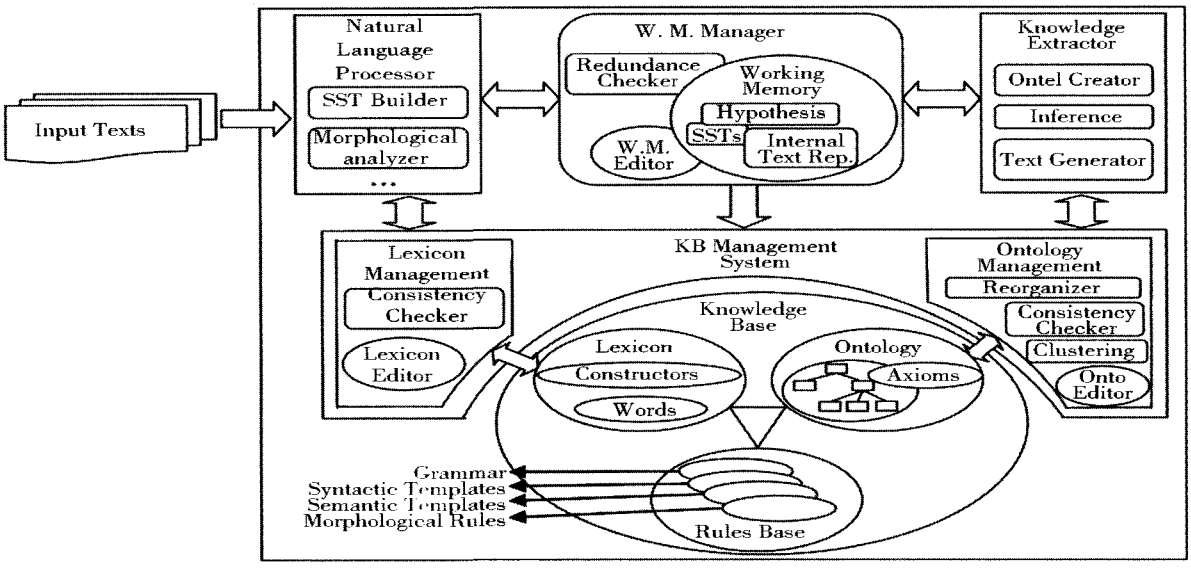

图 2　Hasti 本体学习系统的体系结构[2]

和概念存放在工作存储器和词典管理器中；

②使用知识抽取器获取概念间关系，并将获得的新知识存放在工作存储器和本体管理器中；

③将工作存储器中的概念和概念关系等知识转入知识管理系统中；

④在本体管理器中放置本体元素，根据语义分析方法和聚类方法重新组织并细化本体。

Hasti 系统使用了启发式的学习方法，即在本体学习过程中，当同时出现多个可能的候选结果时，利用一些启发式的规则来减少假设空间，消除不确定性。

(4) OntoBuilder

OntoBuilder[44-45] 是 Mississippi State University 开发的一个从 XML 和 HTML 等半结构化数据中完全自动化获取本体概念及其关系的工具。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。OntoBuilder 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 支

持英文网页，但在实际使用中，并不能适用于所有网站，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。

(5) OntoLiFT

OntoLiFT[28]是 University of Karlsruhe 开发的一个采用基于规则映射的方法从 XML schema 或 DTDs 等半结构化数据和结构化关系数据库中获取本体概念及其关系的工具。

①从 XML Schema 和 DTDs 中获取本体，其结构如图 3 所示[28]。从 XML Schema 到 RDFS 本体的映射由基于 EU - funded Harmonise 工程中开发的 hMarfra 工具加以实现；OntoLiFT 提供了一个中介工具 DTD2XS，实现从 DTDs 到 XML Schema 的映射，将两个工具合并，可实现从 DTDs 到 RDFS 本体的转化。

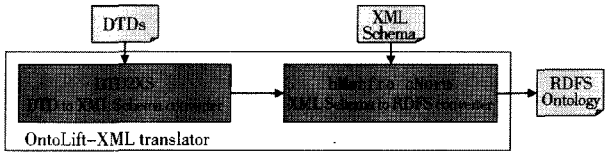

图 3　OntoLiFT 系统中基于 XML(XML Schema 和 DTDs)的本体转化器

②从关系数据库中获取本体。基于 JDBC 标准提供的接口，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预先设定的映射规则转换为本体元素，实现在结构化数据中抽取本体。

(6) GOLF

GOLF[46]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。

3.2　本体学习系统的比较分析

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[47-48]，如表 1 所示。

(1)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；OntoLearn 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。

(2)各种学习方法被综合使用：Text - To - Onto、Hasti 和 GOLF 都集成了多种本体学习技术，包括语言学分析、层次聚类、关联规则挖掘、基于模板匹配以及自然语言处理等，而其他三个系统采用的技术相对比

表 1　对 6 个本体学习系统的比较分析

系统 指标	Onto - to - Text	OntoLearn	Hasti	OntoBuilder	OntoLiFT	GOLF
支持的数据格式	文本、HTML、XML、词典	文本	文本	HTML、XML	XML、HTML、关系数据库	文本、HTML、XML
采用的学习方法	词频统计、层次聚类、关联规则分析、模板匹配	语言学分析、词频统计、语义解释	语言学分析、启发式规则、层次聚类、模板匹配	词频统计、模板匹配	规则映射	语言学分析、概念聚类、关联规则、自然语言处理
生成的本体元素	概念、关系	概念、关系	概念、关系和公理	概念、关系	概念、关系	概念、关系
支持的语种类型	英文、德文	英文	波斯语	英文	英文	中文、英文
系统开发单位	University of Karlsruhe	University of Rome	Amirkabir University of Technology	Mississippi State University	University of Karlsruhe	浙江大学

较单一;在对不同本体元素的抽取中,各系统也都采用了不同的学习方法。以 Hasti 系统为例,在概念抽取中,主要采用语言学分析和启发式学习技术,对关系的抽取则采用了层次聚类、模板匹配和启发式学习混合方法,对于公理的学习,则采用模板匹配和启发式学习结合方式;语言学方法和数据挖掘的方法在本体学习系统中被广泛使用,语义解释、启发式学习等方法还有待进一步研究和推广。

(3)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。

(4)中文本体学习的研究相对匮乏。自然语言处理是本体学习中的基础技术,特别是从非结构化文本中抽取本体,自然语言处理的合理性直接影响本体抽取的成功率和可靠性,因此,本体学习工具一般具有很强的语言特征。由于中文语法的复杂性和中文自然语言处理技术的不成熟,使得当前中文本体学习系统主要停留在实验室阶段(如 GOLF)。支持英文的实用系统较多,也不乏存在对特殊语种进行处理的系统,典型的如 Hasti。

(5)目前,本体学习研究主要集中在欧洲各国。如 Onto - to - Text 和 OntoLiFT 都由德国卡尔斯鲁厄大学 AIFB 研究所开发,OntoLearn 由意大利罗马大学开发完成。特别是 AIFB 研究所是目前世界上该领域研究最为深入、成果最多、影响最大的研究团体。国内在本体学习以及系统的开发研究中,目前还处于起步阶段。

除了上述具有代表性的系统外,国内外其他多种实用性或实验性的本体学习系统还包括基于文本的 LTG[49]、ASIUM[50]、Mo' K Workbench[16]、Welkin[51]、HOLA[52]、OntoGen[53]、OntoLancs[54] 等;基于词典的 SEISD、DODDLE 等[55];面向中文的有 OntoSphere[56]、SOAT[57]等。

4 结 语

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。完全自动化的本体构建还有赖于本体学习技术的进一步成熟和更多语言分析技术和数学模型的引入。

由于汉语文法结构及其自然语言处理技术的复杂性,基于汉语资源的本体学习研究目前处于起步阶段。国内部分学者开始逐步把形式概念分析、奇异矩阵分解、关联规则挖掘、文本聚类以及自然语言处理等方法和技术应用于本体学习的研究中,虽然目前还没有实用性的本体学习系统产生,但是随着研究的进一步深入以及汉语自然语言处理技术的逐渐成熟,基于汉语资料的本体学习,无论在技术上还是系统开发上都将有所突破。

参考文献:

[1] Berners - Lee T. Semantic Web Road Map[EB/OL]. [2008 - 11 - 08]. http://www.w3.org/DesignIssues/Semantic.html.

[2] Shamsfard M, Barforoush A A. Learning Ontologies from Natural Language Texts [J]. *Int' l Journal Human - Computer Studies*, 2004, 60(1):17 - 63.

[3] Xu F Y, Kurz D, Piskorski J, et al. An Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and Their Relations with Bootstrapping[C]. In: *Proceeding of the LREC*, 2002.

[4] Zhou L N. Ontology Learning: State of the Art and Open Issues [J]. *Information Technology and Management*, 2008, 8(3): 241 - 252.

[5] 王红滨, 刘大昕, 王念滨, 等. 基于非结构化数据的本体学习研究[J]. *计算机工程与应用*, 2008, 44(26):30 - 33.

[6] Velardi P, Fabriani P, Missikoff M. Using Text Processing Techniques to Automatically Enrich a Domain Ontology[C]. In: *Proceeding of the FOIS*. New York: ACM Press, 2001: 270 - 284.

[7] 方卫东, 袁 华, 刘卫红. 基于 Web 挖掘的领域本体自动学习[J]. *清华大学学报:自然科学版*, 2005, 45(S1): 1729 -

1733.

[8] Daille B. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology [C]. In: *Proceeding of the ACL'94 Workshop*, 1994.

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

[10] Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora[C]. In: *Bourigault D., ed. Proceeding of the COLING*, 1999:539 - 545.

[11] 梁 健, 王惠临. 基于文本的本体学习方法研究[J]. 情报理论与实践, 2007(1): 112 - 115, 17.

[12] Casado Ruiz M, Alfonseca E, Castells P. Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia[C]. In: *Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems*, NLDB 2005, Alicante, Spain, June 15 - 17, 2005:67 - 79.

[13] Maedche A, Staab S. Ontology Learning for the Semantic Web[J]. *IEEE INTELLIGENT SYSTEMS*, 2001, 16(2):72 - 79.

[14] Agirre E, Ansa O, Hovy E et al. Enriching Very Large Ontologies Using the WWW [C]. In: *Proceedings of the Ontology Learning Workshop, ECAI* 2000, *Berlin, Germany*, 2000.

[15] Khan L, Luo F. Ontology Construction for Information Selection [C]. In: *Proceeding of 14th IEEE International Conference on Tools with Artificial Intelligence, Washington DC*, 2002: 122 - 127.

[16] Bisson G, Nedellec C, Canamero L. Designing Clustering Methods for Ontology Building - The Mo'K Workbench[C]. In: *Proceedings of the ECAI Ontology Learning Workshop*, 2000: 13 - 19.

[17] Faure D, Nedellec C. A Corpus - based Conceptual Clustering Method for Verb Frames and Ontology Acquisition[C]. In: *Velardi P. Proceeding of the LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications. Granada: LREC*, 1998: 5 - 12.

[18] Maedche A, Staab S. Discovering Conceptual Relations from Text [C]. In: *Horn W. Proceeding of the ECAI* 2000. *Amsterdam: IOS Press*, 2000: 321 - 325.

[19] Weng S S, Tsai H J, Liu S C, et al. Ontology Construction for Information Classification[J]. *Expert Systems with Applications*, 2006, 31(1):1 - 12.

[20] Cimiano P, Hotho A, Staab S. Learning Concept Hierarchies from Text Corpora Using Formal Concept Analysis[J]. *Journal of Artificial Intelligence Research*, 2005.

[21] Maddi G R, Velvadapu C S. Ontology Extraction from Text Documents by Singular Value Decomposition[D]. Bowie: Bowie State University, 2001.

[22] Omelayenko B. Learning of Ontologies for the Web: The Analysis of Existent Approaches[C]. In: *Proceedings of the International Workshop on Web Dynamics*, 2001.

[23] Lin D, Pantel P. DIRT - Discovery Of Inference Rules From Text [C]. In: *Proceedings of ACMSIGKDD Conference on Knowledge Discovery and Data Mining*, 2001: 323 - 328.

[24] Harris Z. Distributional Structure[A] // Katz J. J. The Philosophy of Linguistics. New York: Oxford University Press, 1985: 26 - 47.

[25] Dagan I, Glickman O, Magnini B. The Pascal Recognizing Textual Entailment Challenge[C]. In: *Proceedings of the PASCAL worksho Pon Recognizing Textual Entailment, Southampton, UK, April* 11 - 13, 2005.

[26] Papatheodorou C, Vassiliou A, Simon B. Discovery of Ontologies for Learning Resources Using Word - based Clustering[A]// ED - MEDIA 2002. Copyright by AACE. Reprinted from the EDMEDIA 2002 Proceedings[C], *August* 2002 *with Permission of AACE, Denver, USA, August*, 2002.

[27] Doan A, Domingos P, Levy A. Learning Source Descriptions for Data Integration [C]. In: *Proceedings of the Third International Worksho Pon the Web and Databases*, 2000:81 - 86.

[28] Volz R, Oberle D, Staab S. OntoLiFT Prototype[A]// IST Project 2001 - 33052 WonderWeb Deliverable 11[R]. 2003.

[29] Litkowski K. Models of the Semantic Structure of Dictionaries[J]. *Journal of Computational Liguistics*, 1978,15(81):25 - 74.

[30] Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora[C]. In: *Proceedings of the Fourteenth International Conference on Computational Linguistic, Nantes, France*, July 1992.

[31] Rigau G, Rodrigues H, Agirre E. Building Accurate Semantic Taxonomies From Monolingual MRDs[C/OL]. In: *Proceeding of the COLING - ACL. San Francisco: Morgan Kaufmann Publishers*, 1998:1103 - 1109. http://acl. ldc. upenn. edu/P/P98/P98 - 2181. pdf.

[32] Jannink J, Wiederhold G. Ontology Maintenance with an Algebraic Methodology: A Case Study[C]. In: *Proceedings of AAAI worksho Pon Ontology Management*, 1999.

[33] Suryanto H, Compton P. Discovery of Ontologies from Knowledge Bases[C]. In: *Gil Y., Musen M., Shavlik J. Proceedings of the First International Conference on Knowledge Capture, British Columbia Canada*, 21 - 23 *Oct.* 2001, *New York: The Association for Computing Machinery*, 2001: 171 - 178.

[34] Johannesson P. A Method for Transforming Relational Schemas into Conceptual Schemas[C]. In:*10th International Conference on Data Engineering, Ed. M. Rusinkiewicz, Houston: IEEE Press*, 1994: 115 - 122.

[35] Stojanovic L, Stojanovic N, Volz R. Migrating Data - intensive Web Sites into the Semantic Web[C]. In: *Proceedings of the 17th ACM*

Symposium on Applied Computing (*SAC*), *ACM Press*, 2002: 1100 - 1107.

[36] Kashya PV. Design and Creation of Ontologies for Environmental Information Retrieval[C]. *Twelfth Worksho Pon Knowledge Acquisition, Modelling and Management Voyager Inn, Banff, Alberta, Canada. October*, 1999.

[37] Astrova I. Reverse Engineering of Relational Database to Ontologies [C]. In: *Proceeding of the ESWC* 2004. *Heidelberg: Springer - Verlag*, 2004: 327 - 341.

[38] Astrova I, Stantic B. An HTML - Forms - Driven Approach to Reverse Engineering of Relational Databases to Ontologies[C]. In: *Proceeding of the* 23*rd IASTED International Conference on Databases and Applications*(*DBA*), *Innsbruck, Austria*, 2005: 246 - 251.

[39] 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.

[40] Cimiano P, Volker J. Text2onto - A Framework for Ontology Learning and Data - driven Change Discovery[C]. In: *Proceeding of NLDB* 2005, *Lecture Notes in Computer Science, vol.* 3513, *Springer, Alicante*, 2005: 227 - 238.

[41] Maedche A, Staab S. Ontology Learning from Text[C]. *NLDB* 2000: 364.

[42] Missikoff M, Navigli R, Velardi P. Integrated Approach to Web Ontology Learning and Engineering[J]. *IEEE Computer*, 2002, 35 (11): 60 - 63.

[43] Roberto N, Paola V. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites[J]. *Computational Linguistics*, 2004, 30(2): 151 - 179.

[44] Gal A, Modica G, Jamil H M, et al. Automatic Ontology Matching Using Application Semantics[J]. *AI Magazine*, 26(1), 2005.

[45] Roitman H, Gal A. OntoBuilder: Fully Automatic Extraction and Consolidation of Ontologies from Web Sources Using Sequence Semantics[C]. *ICSNW'06*, 2006: 573 - 576.

[46] 刘柏嵩. 基于 Web 的通用本体学习研究[D]. 杭州: 浙江大学, 2007(1).

[47] 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.

[48] 张囡囡, 李冠宇, 曲丽宁. 主要本体学习工具的比较分析[J]. *微计算机信息*, 2008(12): 189 - 190, 182.

[49] Mikheev A, Finch S. A Workbench for Finding Structure in Texts [C]. In: *Proceedings of ANLP* - 97 (*Washington D. C.*). *ACL March*, 1997: 8.

[50] Faure D, Nédelle c(C). Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM[C]. In: *Fensel*(*D*) *and Studer R. editors, Proceeding of the* 11*th European Workshop* (*EKAW'99*), *LNAI* 1621, 1999: 329 - 334.

[51] Alfonseca E, Rodríguez P. Automatically Generating Hypermedia Documents Depending on User Goals[R], Workshop on Document Compression and Synthesis in Adaptive Hypermedia Systems, AH - 2002, Málaga, Spain, 2002.

[52] Manzano D, Gómez - Pérez A, Borrajo D. HOLA: A Hybrid Ontology Learning Architecture[A] *//The* 15*th International Conference on Knowledge Engineering and Knowledge Management*[C], Podebrady, 2006.

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

[54] Gacitua R, Sawyer P, Rayson P. A Flexible Framework to Experiment with Ontology Learning Techniques[J]. *Knowledge - Based Systems*, 2008, 21(3): 192 - 199.

[55] Gómez - Pérez A, Manzano - Macho D. A Survey of Ontology Learning Methods and Techniques[A]// OntoWeb: Ontology - based Information Exchange for Knowledge Management and Electronic Commerce[R], IST Project IST - 2000 - 29243, 2003: 63 - 65.

[56] 孔敬. 本体学习: 原理、方法与相关进展[J]. *情报学报*, 2006 (6): 657 - 665.

[57] WU Sh H, HSU W L. SOAT: A Semi - automatic Domain Ontology Acquisition Tool from Chinese Corpus[C]. In: *Proceedings of the* 19*th International Conference on Computational Linguistics*(*COLING* 2002), *Taipei*, 2002: 1313 - 1317.

(作者 E - mail: ywhaowang810710@ sina. com)