基于启发式规则的自动化本体扩充

李伊潇[1,2] 李宏伟[1,2,3] 沈立炜[1,2] 赵文耘[1,2]

(复旦大学计算机科学与技术学院 上海 201203)[1] (上海市数据科学重点实验室(复旦大学) 上海 201203)[2]

(江西师范大学计算机信息工程学院 南昌 330022)[3]

摘 要 自动化地获取网络资源中的领域本体可以缩短本体的构建周期,但自动化的本体扩充还是本体工程中的一个挑战,其难点主要在于如何抽取术语并在新术语和已有本体之间建立映射关系。为此,提出了一个基于启发式规则的本体自动化扩充方法。该方法从网络资源中抽取自然语言文本,结合自然语言处理技术进行文本预处理,采用优先匹配对象属性的方式挖掘领域知识术语,然后通过启发式规则匹配术语的方式进行本体扩充,最后进行一致性检测。采用上述方法实现了一个基于Web的本体扩充工具。以城市景观信息核心本体作为研究案例进行了实验,结果显示本方法在扩充实例时具有较高的查准率和查全率,表明其具有有效性和可行性。

关键词 本体扩充,领域本体,术语抽取,启发式规则

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.3.039

Automatic Ontology Population Based on Heuristic Rules

LI Yi-xiao[1,2] LI Hong-wei[1,2,3] SHEN Li-wei[1,2] ZHAO Wen-yun[1,2]

(School of Computer Science, Fudan University, Shanghai 201203, China)[1]

(Shanghai Key Laboratory of Data Science, Fudan University, Shanghai 201203, China)[2]

(School of Computer Information and Engineering, Jiangxi Normal University, Nanchang 330022, China)[3]

Abstract The cycle of building ontology can be shortened by means of automatically extracting domain ontology in Internet resources, but automatic ontology population is still a challenge in ontology engineering. There are two difficulties in this area, which are how to extract terms and how to construct the mapping relationship between the new terms and the existed ontology. Therefore, this paper proposed a method for automatic ontology population based on the proposed heuristic rules. This method extracts natural language texts from the Internet, combines traditional natural language processing methods for text preprocessing, discovers domain terms by preferentially matching object properties, enriches the ontology by matching these terms using heuristic rules, and finally checks the consistency of the enriched ontology. On the base of the proposed method, this paper also implemented a Web-based tool for ontology population. Using an urban landscape information core ontology as a case study, the experimental results show that the method for enriching ontology individuals has a high precision and recall. The results also prove that the proposed method is effective and feasible.

Keywords Ontology population, Domain ontology, Term extraction, Heuristic rule

1 引言

本体旨在描述现实世界中的概念以及它们之间的对应关系。针对一个特定的领域,领域本体涵盖了隶属于该领域的概念、概念间关系、概念属性以及概念的实例。若使用OWL(Web Ontology Language)[1]作为本体的描述语言,那么本体元素可分别使用类(Class)、对象属性(ObjectProperty)、值属性(DatatypeProperty)和个体(Individual)进行表示。

基于本体的应用系统依赖于本体的概念与实例实现知识的共享与推理。一般而言,本体中概念及其关联较为稳定,而概念的实例则会源源不断地扩展。例如,在一个城市中,属于概念"Road"(道路)的新实例会随着市政建设而不断涌现。因此,如何基于一个已有的本体来不断挖掘新的本体实例成为了本体研究领域的一个重要课题。

手工识别新的本体实例是费时费力、易于出错的方式。与之相比,自动化的本体扩充一类是用于实现以上需求的技术,能够将挖掘本体实例的时间控制在可接受的范围内。Petasis等人认为本体扩充是将实例插入到已有本体的过

到稿日期:2015-02-01 返修日期:2015-06-06 本文受国家"863"高技术研究发展计划项目(2013AA01A605),国家自然科学基金项目(61402113)资助。

李伊潇(1990—),女,硕士生,主要研究方向为自动化本体建模技术,E-mail:352741689@qq.com;李宏伟(1975—),男,博士,讲师,主要研究方向为软件再工程;沈立炜(1982—),男,博士,副教授,主要研究方向为软件产品线、自适应软件系统,E-mail:shenliwei@fudan.edu.cn(通信作者);赵文耘(1964—),男,教授,博士生导师,CCF高级会员,主要研究方向为软件工程、基于构件的软件开发等。

程[6]，该过程一般是将已有本体中的概念、关系和属性作为基础术语，再使用抽取工具抽取实例术语，最后使用匹配算法将已抽取的实例匹配到本体。现有的将实例术语添加到本体中的方法主要包括以下 3 种：1)使用已有语料库查询被抽取术语的上下位关系；2)通过计算抽取实例信息中所获得的术语与本体中概念的文本之间的相似度，建立实例与概念间的映射关系；3)使用统计学方法计算与概念共同出现频率较高的术语。这些方法要么需要大量数据源，要么缺少对概念间复杂关系的把握，而且它们很大程度上依赖于已有的外部知识库。

目前本体自动化扩充的数据源主要是已有的本体、网页、自然语言文本和数据库。在这些数据源中，来自网络的资源量和信息量最为巨大。针对网络数据源，自动化本体扩充需要解决以下两个问题：

(1)如何从网络资源中识别和抽取候选的待扩充术语？

(2)如何建立待扩充的术语与概念之间的映射关系，从而实现本体实例的扩充？

本文针对以上问题，提出了一个基于启发式规则的本体自动化扩充方法。该方法从网络资源中抽取自然语言文本，结合自然语言处理技术进行文本预处理，优先识别其中与本体对象属性相对应的动词成分，然后识别候选的名词成分，随后基于预定义的启发式规则建立名词成分与本体概念之间的映射关系，进而通过启发式规则匹配术语的方式进行本体扩充，最后进行一致性检测。基于上述方法，实现了一个本体知识自动化扩充工具，并以一个城市景观信息核心本体为例，使用该工具进行了针对该本体的自动化扩充实验，结果表明所提出的方法具有有效性和可行性。本方法可以抽取出常用词库中没有的实例，并大幅度减少本体扩充的时间。

本文第 2 节简要介绍了本研究所基于的一个简单的领域本体示例；第 3 节详细介绍了基于启发式规则的自动化本体扩充方法；第 4 节介绍了开发的一个基于 Web 的本体扩充工具，使用它从相关 Web 页面中抽取信息来扩充城市景观信息核心本体，以此作为案例验证了本方法的实际效果；第 5 节列举了相关的工作并与本文工作进行了对比；最后是对所提方法的总结和对后续工作的展望。

2 领域本体示例

为方便讨论、实验和验证本文结论，先给出一个表示城市景观信息的核心本体作为领域本体示例。该本体一共包括 13 个本体类以及 32 个对象属性和 4 个值属性，其中部分的类层次关系和对象属性如图 1 所示。在图 1 中，圆角矩形代表概念，带有空心箭头的线代表父类-子类关系，而普通箭头代表对象属性，箭头上的词汇代表对象属性的名称。例如，概念“Attraction”(景观)是概念“Location”(地点)的子概念，概念“Organization”(组织机构)和“Attraction”之间具有“design”(设计)关系，这表示“Attraction”是“Location”的一种，“Organization”“design”了“Attraction”。

万维网上的很多网页包含了城市景观信息。本文选用示例网站[1)]的“上海”主题页面中含有介绍上海各个景观的共计 80 篇英文网页链接，这些网页中包含了城市景观的领域知识信息。本文通过指定这些网页的统一资源定位符(URL)获取这些网页的内容，将其中有关联的实例对象加入到城市景观信息核心本体中。这些网页上的部分自然语言实例如表 1 所列，表中使用下划线的术语代表与本体概念对应的实例，黑色加粗的术语代表和本体对象属性对应的词汇。以第三条语句为例，若使用本体扩充的技术，城市景观信息本体所期望的扩充结果应为：“Century Park”是“Attraction”的实例，“British LUC Company”是“Organization”的实例。

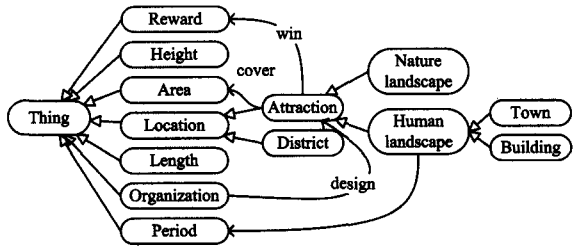

图 1 核心本体中的概念与部分关系实例

表 1 待获取网页上的自然语言文本示例

序号	文本内容
1	Shanghai Wild Animal Park is the first national wild animal park **established** by Shanghai City Government. Shanghai Wild Animal Park is **located in** Sanzhao Town, Nanhui District, Pudong New Area, Shanghai, 35km from Shanghai downtown and **covering** an area of 205 hectares(3100Mo).
2	Shanghai Yinqixing Indoor Skiing Site **locates** at 1835 Qixin Road, Minhang District. It **covers** an area of 100,800 meters.
3	Century Park is **designed** by British LUC Company.

本文提出的方法与实验均以上述本体作为方法示例与实验对象。方法与工具的输入是城市景观信息核心本体以及与上海主题相关的英文版网页；输出是扩充了城市景观术语之后的领域本体。

3 基于启发式规则的自动化本体扩充方法

3.1 方法概述

本文提出了基于启发式规则的自动化本体扩充方法，方法的流程包括 5 个步骤，如图 2 所示。

为了便于理解本节内容，下面先给出几个术语的解释。名词成分：名词或名词短语；候选实例：被加入到三元组中的名词成分；待确认实例：经过候选实例映射后，仍需用户判断是否加入本体中的候选实例；实例：最终被加入到本体中的候选实例或待确认实例。其中，三元组的定义见定义 1。

本体扩充过程的 5 个步骤解释如下：

(1)文本抓取。输入一个领域核心本体和作为数据源的网页 URL，提取出网页中与领域核心本体相关的自然语言文本。

(2)文本预处理。对已抓取的自然语言文本进行处理，包括分词、分句、词性标注、名词成分识别等，为后续的术语抽取步骤做准备。

(3)术语抽取。采用优先匹配对象属性，随后以识别关联对象的方式从文本中抽取类似包含(候选实例 1，关系，候选实例 2)的三元组。

(4)实例映射与本体扩充。通过启发式规则将(3)中所抽

1) http://english.51766.com/detail/city_attractions.jsp?prov_id=10031

取出的三元组与核心本体中的概念、关系进行匹配，将三元组的候选实例映射到相应的概念中，并将确认的候选实例作为概念的一个新实例。由于可能有多个三元组中含有同一个候选实例，导致添加以后同一个候选实例存在于多个概念中，因此在增加新实例的同时，需要运用启发式规则识别并删除冗余的实例。经过上述步骤后，由用户决定待确认实例是否应该加入到核心本体中。

(5)一致性检测。使用已有的本体推理机对扩充后的本体进行一致性检测，检测本体是否存在冲突或不一致信息。

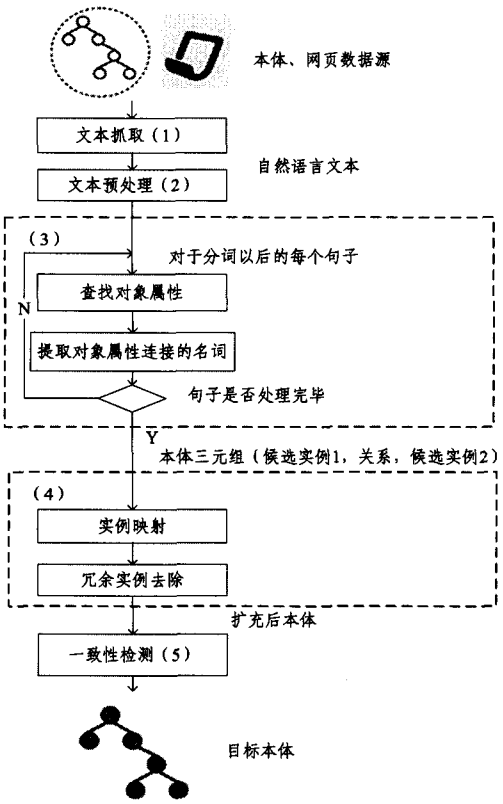

图 2　本体扩充的过程

3.2　文本抓取

“网络爬虫”[13]是一种用于自动化提取网页以及网页内容的技术。为了获取网页中的自然语言成分，本文采用“网络爬虫”从用户指定的 URL 中提取直接链接网页中的纯文本信息，剔除掉网页信息中存在的与本体实例不相关的成分，如某些链接、HTML 标签和脚本等。

3.3　文本预处理

文本预处理涉及以下内容。

• 分句和分词：分别把复合句型划分成单句和单词。

• 去除停词：去除文档中某些如“the”、“a”、“an”等对于本体扩充无意义的词汇。

• 词干提取和词形归并：其目的是减小不同时态或词性对术语抽取的影响。例如将词的单复数形式变换为词干或词根，或者将词的不同词形归并为原形词或词根（如“is”、“was”、“are”、“were”等转换为“be”）。

• 词性标注：用来识别每个词在上下文中的作用，例如词性信息（名词、动词、形容词等），便于接下来识别其作为主语、谓语或宾语的成分以及作用。

• 名词性成分识别：识别自然语言中表示本体实例的名词或名词短语，如名词“Shanghai”和名词短语“Hongmei Road”。

• 指代消解：识别文本中的代词（人称代词、指示代词等）所指示的对象。

经过以上预处理的自然语言文本用于本体中实例术语的抽取。本文采用 NLTK(Natural Language Toolkit)[10]、Stanford CoreNLP[11]和 Stanford Parser[12]等工具对自然语言成分进行预处理。NTLK 是使用 Python 语言开发的用于自然语言处理的工具包，它包含大部分对自然语言进行处理的操作。本文工具采用 NLTK 进行分句、分词、去除停词、词形归并、词性标注、名词性成分识别。Stanford CoreNLP 是一个集成框架，它集成了斯坦福自然语言处理小组的一系列自然语言分析工具。本文工具用它来处理指代消解的问题。Stanford Parser 是斯坦福自然语言处理小组提供的一系列工具之一，能够用来完成语法分析任务，并分析相应的短语和词之间的依赖关系。本文工具用它完成被动语态的识别。

3.4　术语抽取

根据领域核心本体的结构将分词的结果转换为符合主谓宾的语法结构三元组（候选实例 1，关系，候选实例 2）。该三元组中包含了需要被扩充到本体中的实例以及本体模型中可能包含的概念间的关系。

定义 1　三元组定义为（I_1，*Verb*，I_2）。其中，*Verb* 为 I_1 和 I_2 的关系描述，通常由有可能对应领域核心本体中某个对象属性(ObjectProperty)的谓词或谓词短语来表示。I_1 和 I_2 是候选实例，经过扩充后，它们有可能成为领域核心本体中概念的实例，它们是关系所连接的名词或名词短语。

本文使用以下步骤来抽取三元组：

(1)使用核心本体中对象属性作为关键字，在预处理后的每个句子中进行查询，查看这个句子中是否包含关键字。如果不包含关键字，则将此句过滤，取下一句，直到包含关键字为止；如果包含关键字，则执行步骤(2)。

(2)对于每个句子中包含的关键字，提取句子中对应关键字的词或短语作为主谓宾三元组中的 *Verb*，然后执行步骤(3)。

(3)获取 *Verb* 在同一句子中连接的名词成分，将其作为候选实例 I_1 和 I_2。如果 I_1 或 I_2 对应多个名词成分，则生成多个三元组，并将多个名词成分分别作为待确认实例加入到三元组中。

(4)句子经过 Stanford Parser 的解析，可以分析出一系列的依赖关系。若这些依赖关系中含有被动词关系，且该被动词关系中包含 *Verb*，则认为 *Verb* 在自然语言文本中是被动语态，交换 I_1 和 I_2。例如关键字为“border”，依赖关系中包含形如“auxpass(bordered-5,is-4)”的结果（在 Stanford Parser 中代表被动依赖关系），表示该关键字提取的三元组中包含被动关系。

例如，将以上步骤应用于语句“Changfeng Park was established by Shanghai government”。由于在核心本体中有“establish”对象属性，因此“established”对应 *Verb*，“Changfeng Park”和“Shanghai government”是“established”连接的两个名词。由于本句中“establish”是被动语态，因此将“Shanghai government”作为 I_1，将“Changfeng Park”作为 I_2。这个

短句的内容最终对应三元组(Shanghai government, establish, Changfeng Park)。又如"Gongqing Forest Park boasts Cedar Grove and Jungle",则生成(Gongqing Forest Park, boast, Cedar Grove)和(Gongqing Forest Park, boast, Jungle)两个三元组。其中,"Cedar Grove"和"Jungle"是待确认实例。

并非句子中所有的名词成分对提取三元组都有意义。一个名词成分有可能对应目标本体中的一个实例,例如"Hongmei Road";也可能对应一个概念,例如"Road"。而像"Road"这样的概念名词不应作为一个候选实例被抽取到三元组中。本文使用如下方法判断一个名词成分是否属于一个概念。

(1)使用一个数据库来记录概念名词。若名词成分存在于数据库中,则去除该名词成分,不将其加入到三元组中;若名词成分不存在于数据库中,则执行步骤(2)。

(2)在 WordNet 中查询该名词成分。由于 WordNet 中已经包含大部分概念名词,因此若名词成分不存在于 WordNet 中,则将其作为一个候选实例加入到三元组中。如果名词成分存在于 WordNet,表明它可能是概念名词,也可能是实例名词,则将其作为待确认实例加入到三元组中。

(3)将待确认实例经过 3.5 节中的方法建立与核心本体的映射关系,如果这个待确认实例被用户指定为一个概念,则将其加入到步骤(1)所述的数据库中;其他情况不做任何操作。

在一个三元组(I_1, *Verb*, I_2)中,I_1 或 I_2 可能为空值。例如,针对语句"East Sheshan Mountain National Forest Park, which boasts beautiful mountains",通过以上步骤发现 mountains 是概念而非实例。因此,该语句生成的三元组是("East Sheshan Mountain National Forest Park", "boast", null)。

3.5 实例映射与本体扩充

为了实现候选实例映射与本体扩充,本文提出了 2 条启发式规则:增量式规则和删除式规则。

实例映射的过程是逐一遍历所挖掘出的三元组,并使用启发式规则建立核心本体与三元组中候选实例间的映射关系,这个启发式规则被称为增量式规则。在增量式规则运用过程中会出现相同实例被添加到不同本体类的情况,如果这些类中包含父子关系,这时需要使用删除式规则删除冗余实例。

增量式规则:

设有一个三元组(I_1, *Verb*, I_2),*Verb* 代表三元组中的一个关系,其与核心本体中某个对象属性 V 相对应。C_1 和 C_2 是本体中的两个概念类。I_1 和 I_2 是三元组中的候选实例。如果对象属性 V 对应的定义域(domain)是概念类 C_1,对象属性 V 对应的值域(range)是 C_2,则建立 I_1 和 C_1、I_2 和 C_2 之间的映射关系。这个增量式启发式规则的形式化表达如下:

$$\{add(C_1, I_1), add(C_2, I_2) \mid \exists V \in ObjectProperty, (I_1, Verb, I_2) \wedge V = Verb \wedge (V.domain = C_1 \wedge V.range = C_2)\}$$

其中函数 $add(C, I)$ 表示建立候选实例 I 与概念类 C 之间的映射关系。

基于增量式规则建立映射关系后,对于 I_1 和 I_2 中的任一实例 I,如果概念 C 中已经存在与实例 I 相似度过高的实例,则不将实例 I 加入本体;如果实例 I 是一个待确认实例,则保留 I 和本体间的映射关系,但不将其加入本体;否则把 I 直接添加到目标本体中成为实例。

经过初步扩充后,可能有多个三元组中含有同一个实例 I,同一个实例 I 有可能被添加到多个不同的概念类中。因此,当实例 I 同时被添加到概念 C 和概念类组 $CSUi$ 中,且 $CSUi$ 是 C 的子类,任意一个 $CSUi$ 的范围都比 C 的范围小时,只保留 $CSUi$ 中的实例 I 可使扩充结果更加精确,而且 I 是 $CSUi$ 的实例也隐含了 I 是 C 的实例的语义。

下面提出删除式规则用于删除概念类 C 中的实例 I,只保留概念类组 $CSUi$ 中的实例 I。

删除式规则:

设 $CSUj$ 是概念类 C 的一组子类(j 是 0 到 $+\infty$ 之间的非负整数),I 是实例。如果通过增量式规则扩充本体之后,$CSUj$ 和概念类 C 都包含同一个实例 I,则删除父类 C 中的实例 I。这个删除式规则的形式化表达如下:

$$\{remove(C, I) \mid \exists CSUj \in Class, (C.subClass = CSUj) \wedge I \in C \wedge I \in CSUj\}$$

其中,函数 $remove(C, I)$ 表示将实例 I 从概念类 C 中删除。

在城市景观本体中,假设类"Attraction"和类"Location"之间存在对象属性"locate in"和"near"。其中,"Attraction"是"Location"的子类。"locate in"和"near"的定义域和值域都分别是"Attraction"和"Location"。假设有三元组("Yu Gardens", "near", "City God's Temple"),通过增量式规则,可以判断"City God's Temple"是"Location"的一个实例。同时,通过增量式规则和三元组("City God's Temple", "locate in", "Yan'an Road")扩充核心本体,可以判断"City God's Temple"是"Attraction"。因为"Attraction"是"Location"的子类,"Attraction"表示的范围更小,将"City God's Temple"作为"Attraction"的实例更加精确,所以应当删除"Location"的实例"City God's Temple"。

在自动化扩充过程完结后,将待确认实例交由用户决定是否作为一个实例添加到本体中。用户可以选择将待确认实例标注为一个概念名词、当前概念的一个本体实例、其他概念的本体实例或不确定为实例。

3.6 一致性检测

经过自动化扩充的本体可能存在不一致,这是由于数据源(网页中的信息)的多样性而引发的不能保证扩充后本体的一致性。例如,在本体建模语言 OWL 中,若一个实例通过一个关系与另一个实例一一对应,则称该关系为函数型属性(Functional Property)。如果一个属性 P 被标记为函数型属性,那么对于所有的实例(如 x, y 和 z),$P(x, y)$ 与 $P(x, z)$ 蕴含 $y = z$。

例如在示例本体中,"build in"属性是函数型属性,一个景观建于一个特定的年份,这个年份是独一无二的。由于数据源的多样性及时序性,并不能保证一个景观只对应一个年份。因此,经过自动化扩充,可能导致一个景观在"build in"上对应多个值。如图 3 所示,经过扩充后,存在类"Attraction"的实例"Yu Garden"在"build in"关系下同时对应"1559 year"和"Ming Dynasty"。此时表明有冲突,需要由用户决定是保留"1559 year"还是"Ming Dynasty"。

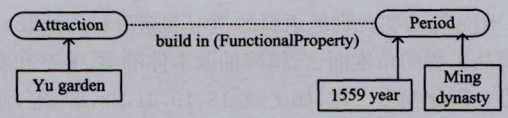

图 3　扩充后本体的冲突与不一致示例

本文使用 Pellet 本体推理机[20]检测被扩充后的本体是否存在冲突和不一致。Pellet 使用逻辑表算法来支持本体中实例、自定义值属性的推理，以此检测找出扩充后本体的不一致性。如果存在冲突或不一致，则提醒用户冲突和不一致的存在，用户据此作相应的修改和调整，直到本体中不存在冲突和不一致为止。

4　工具与实验

为了验证上述方法的有效性，本文实现了一个基于 Web 的本体扩充工具，称为“本体自动化扩充平台”。利用第 2 节讨论的领域本体示例，从示例网站中获取了 80 个英文网页，采用该工具进行了实验，自动化地进行了领域本体的扩充。为了验证结果，本文还使用手工方式构建了上述示例的领域本体，用来对比检查本文提出的自动本体扩充方法的查全率和查准率。

4.1　工具介绍

该工具由一个网络爬虫模块、一个自然语言处理模块、一个语义映射与知识扩充模块组成。其界面与部分结果展示如图 4 所示。

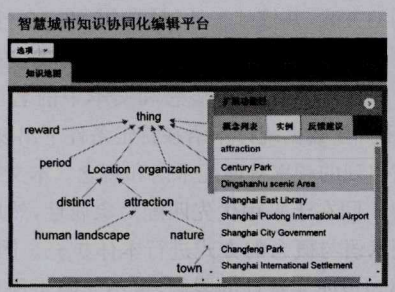

图 4　本体自动化扩充工具与部分结果展示

首先，用户通过输入编辑界面编辑需要被扩充的领域核心本体以及作为数据源的网页的 URL。根据用户输入的 URL，网络爬虫模块负责从网页上抽取自然语言文本。本文采用 BeautifulSoup[9]对网页中的自然语言成分进行抽取。BeautifulSoup 是 Python 中的一个库，可以用于解析 DOM 树，并支持多种编码格式。

自然语言处理模块负责对自然语言进行预处理，并根据核心本体的结构从分词的结果中抽取三元组。如 3.3 节所述，本文采用 NLTK(Natural Language Toolkit)完成分词分句、词性标注和名词成分识别的任务；使用 Stanford CoreNLP 处理指代消解的问题；运用 Stanford Parser 来识别被动关系。

语义映射与知识扩充模块以多个三元组为数据源，根据本文提出的本体扩充规则将三元组中确定的候选实例添加到核心本体中。再通过图形化界面为用户展示待确认实例，由用户决定待确认实例是一个概念名词、当前概念的一个本体实例、其他概念的本体实例还是不确定实例。

工具利用 Pellet 检测本体是否存在冲突和不一致。若存在冲突和不一致，则展示给用户，并由用户负责处理。此后，工具展示和保存扩充后的目标本体，并记录本体的变更情况。

4.2　实验结果

实验对象：城市景观信息核心本体和案例网站中“上海”主题页面下直接相连的所有页面。

为对本文提出的本体扩充方法进行验证，需要评价从给定网页中识别和获得可能的术语的能力，以及识别这些网页中的实例的查准率和查全率。本文采用对象属性识别率、名词成分识别率作为度量实验中间结果的实验指标。对象属性识别率显示了方法对实例对象的识别准确度。名词成分识别率用于衡量通过本方法从文本中获得概念或本体名称的能力。实例查准率用于衡量在方法识别到的实例对象中有多少是正确的，它用于衡量方法的有效性。实例查全率用于检验方法获得的实例对象占所有可能的实例对象的比例，用于衡量方法的能力。

定义 2　对象属性识别率：$Rverb$=(已识别的对象属性/应被识别出的对象属性)×100%。

定义 3　名词成分识别率：Rn=(识别正确的名词成分/应被识别出的名词成分)×100%。

在进行了自然语言文本抓取、预处理和术语抽取后，对象属性识别率和名词成分识别率如图 5 所示，其中对象属性识别率为 94.32%。多数情况下，导致对象属性识别有误的原因是在文本中谓词和对象属性的名词相似度过低。例如，在文本“The building has a perpendicular height of 42 meters”中查找对象属性“has height”，在文本的“has” 和“height”中间隔有“a perpendicular”。因此，尽管表达的是相近的含义，但是“has height”不会与“has a perpendicular height”互相匹配进而提取相应三元组。另外，可以通过调整领域核心本体中对象属性的名称来提高对象属性识别率。但是在本例中，不能将“has height”调整为“height”。因为在建立本体时，用户一般不使用名词作为本体的对象属性，而且将“height”作为关键字在文本中查找可能会找到像“a building height of 40 meters”中的“height”，而此时提取出的三元组不是所期望的三元组。

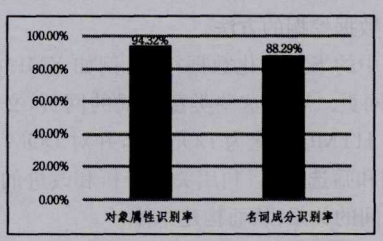

图 5　对象属性识别率和名词成分识别率

在对象属性识别正确的情况下，名词成分识别率为 88.29%。工具可以识别句子中像“Jinjiang Park”和“7000 square meters”这类名词短语，因为有其中的“park”和“square meters”等常见词汇，可以被自然语言处理模块作词性标注，从而被识别为名词成分。但是对于句子中仅有的名词短语“Jinjiang”和“7000m^2”，工具则无法识别出这类专有地名和量词单位等信息。

下面将工具最终扩充的结果与手工构建的领域本体的结果进行对比。在自动扩充的本体和手工构建的本体中都存在同一个实例 I 时，则称这个实例为一个有效实例。若手工构建的本体将一个实例 I 划分在概念类 C 中，而自动扩充的结

果将实例 I 划分在概念类 C 的父类或者子类中，也认为这个实例是有效实例，因为该实例可能根据删除式规则做了自动调整，且由于手工构建本体时的不同理解和实例 I 相关的对象属性(ObjectProperty)本身就不与概念类 C 相关。

定义 4 实例查全率：*Recall*=(扩充的有效实例/应该扩充的实例总量) ×100%。

定义 5 实例查准率：*Precision*=(扩充的有效实例/应该扩充的有效实例) ×100%。

实验最终结果的查全率和查准率如图 6 所示，其中实例的查全率为 79.10%，查准率为 76.81%。查全率主要受两个因素影响：1)领域本体中对象属性的识别率；2)自然语言处理复杂句和长、难句的局限性。

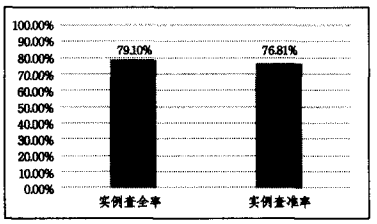

图 6 扩充实例最终结果的查全率和查准率

影响查全率示例：假设有自然语言文本“The Oriental Pearl Radio and Television Tower is 468 meters”，使用本文方法无法提取三元组(“The Oriental Pearl Radio and Television Tower”，“height”，“468 meters”)，这将会影响结果的查全率。而影响查准率的主要因素是名词成分的识别率。

综上所述，本方法对实例对象的查准率和查全率都较高，通过简单的调整或者补充还能得到更好的结果。本方法的结果大大减少了手工编辑本体的工作量，注重于通过对象属性抽取实例术语，有利于提取特定领域的专有词汇。

5 相关工作

迄今为止，很多研究工作从文本中获取本体所需术语的方法，如统计学的方法、自然语言处理方法以及对半结构化的数据源进行数据挖掘的方法。

万维网中的半结构化数据很多，例如 XML 格式和 HTML 格式的网页。对于这些类型格式的网页，文献[15]中的方法通过将 HTML 转化为 DOM 树，并对 DOM 中抽取的文献进行分词和筛选，然后利用关联分析和改进的层次聚类发现领域概念间的关系，从而构建本体。

对于大多数领域知识，它们的知识大多是以非结构化自然语言的形式描述的，所以从自然语言中提取术语并构建本体也是当前的研究热点。文献[5,16,18,19,23,24]都使用了自然语言处理技术获取领域知识。文献[5]将文本按事件方式进行表示，把事件作为基本语义单元来构建事件本体，根据事件间的关系提取文本的语义信息；文献[16]根据文本中词语的词性如名词、动词、形容词等语言学信息抽取概念术语，然后利用术语的领域聚合性和词性特征，采用领域词频比较的方法抽取术语，并利用上下文关联信息、语境信息从候选术语集中筛选出本体术语；文献[23]首先使用语法分析的方法分析自然语言文本，然后计算抽取术语在领域中出现的频率以判断一个术语是否属于该领域；文献[24]对文本的句法格式进行分析，抽取相关子句，将手工标注的结果作为训练集，利用 CRFs 算法实现训练语料的学习和新语料的抽取。

而从工程的角度而言，精确抽取本体的类、关系和实例尤为困难。因此很多工作，如文献[18,19,21,22,25]等，注重于扩充已有本体。一般而言，这类方法首先对自然语言进行分词、词性标注，然后对文本中的术语实体进行识别；再利用外部知识库(如 GATE[8] 和 WordNet[7])、基于规则和机器学习中的一种或者多种方式扩充本体[17]。文献[18]提出的方法在处理自然语言文本后，将文本匹配到预定义的一组英文的句法结构模板，根据不同的句法结构模板(如“是一种”、“和”、“或”、“包括”等常见词汇对应不同模板)建立本体和文本中术语的映射关系；文献[19]提出的方法首先对自然语言文本进行处理，然后通过对象属性匹配候选术语和本体中已有实例扩充本体；文献[21]使用结合基于统计和基于规则的方法，计算术语在文档中的影响程度，计算文档及词语的领域相关度，以此扩充本体；文献[22]提出的框架在抽取出术语之后，利用 WordNet 和已有的网络本体作为背景知识，用背景知识判断术语间的关系；文献[25]通过关联规则挖掘非结构化数据中术语共同出现的规律，以挖掘已有本体中概念间的关系；文献[26]使用自然语言处理方法抽取实例术语，并提出了两个不局限于特定领域的通用规则，对已有术语进行分类，并根据分类将实例术语扩展到本体中。

综上所述，以上方法合理使用了自然语言或统计学的方法抽取构建或扩充本体，它们有的使用外部知识库直接匹配待输入名词的方式，其依赖于外部知识库的完备性，但是这些知识库中鲜有此类特定领域的实例词汇；有的使用预定义语法模板的方式匹配本体中的概念和文本中的名词或名词短语；有的使用大量数据统计术语在特定条件下出现的频率、与概念类同时出现的频率等匹配术语和概念。本文方法与上述方法的主要不同在于采用优先匹配对象属性，然后通过启发式规则匹配术语与概念的方式进行本体扩充。因此，本文方法不受外部知识库完备性的限制，也不受限于预先定义的实例术语、模板类型，且有利于处理冷僻的实例术语。

结束语 本文提出了一种基于启发式规则的本体扩充方法，它利用传统的网络爬虫技术和自然语言处理方法获取新本体术语，并通过启发式规则扩充领域本体知识。基于该方法，本文实现了一个基于 Web 的本体扩充工具，使用该工具，以一个城市景观信息核心本体为例对结果进行了验证。实验结果显示，本方法在扩充实例时查准率为 76.81%，查全率为 79.10%，这个结果表明本方法具有有效性和可行性。

当前，为了获取领域知识，本文方法使用本体中的对象属性作为关键字检索自然语言文本。这种方式依赖于文本术语的组织结构，对于具有更为复杂结构的语句而言，本方法可能无法识别出其中包含的概念实例。针对这一局限，计划在下一步工作中引入更多的自然语言处理机制来应对具有复杂语法结构的自然语言文本。

参 考 文 献

[1] OWL Overview Recommendation [EB/OL]. [2014-12]. http://www.w3.org/TR/2004/REC-owl-features-20040210

[2] Hazman M, El-Beltagy S R, Rafea A. A Survey of Ontology Learning Approaches[J]. International Journal of Computer Ap-

plications,2011,22(9):36-43

[3] Santoso H A,Haw S C,Abdul-Mehdi Z T. Ontology Extraction from Relational Database: Concept Hierarchy as Background Knowledge[J]. Knowledge-Based Systems,2011,24(3):457-464

[4] Wong W,Liu W,Bennamoun M. Ontology Learning from Text: A Look back and into the Future[J]. ACM Computing Surveys (CSUR),2012,44(4):1-36

[5] Yang Jun-hui,Liu Zong-tian,Liu Wei,et al. Extraction Method of Text Summarization Based on Event Network [J]. Computer Science,2015,42(3):210-213(in Chinese)
杨俊辉,刘宗田,刘炜,等.基于文本事件网络自动摘要的抽取方法[J].计算机科学,2015,42(3):210-213

[6] Petasis G,Karkaletsis V,Paliouras G,et al. Ontology Population and Enrichment:State of the Art[C]//Knowledge-Driven Multimedia Information Extraction and Ontology Evolution,2011. Berlin:Springer-Verlag,2011:134-166

[7] WordNet[EB/OL]. [2014-12]. http://WordNet. princeton. edu

[8] Cunningham H,Maynard D,Bontcheva K,et al. A Framework and Graphical Development Environment for Robust NLP Tools and Applications[C]//ACL,2002. ACM Press,2002:168-175

[9] Beautiful Soup [EB/OL]. [2014-12]. http://www. crummy. com/software/BeautifulSoup

[10] NLTK [EB/OL]. [2014-12]. http://www. nltk. org

[11] Stanford CoreNLP [EB/OL]. [2014-12]. http://nlp. stanford. edu/ software/corenlp. shtml

[12] Stanford Parser [EB/OL]. [Dec. 2014]. http://nlp. stanford. edu /software/ lex-parser. shtml

[13] Zhou De-mao, Li Zhou-jun. Survey of High-Performance Web Crawler[J]. Computer Science,2009,36(8):26-29(in Chinese)
周德懋,李舟军.高性能网络爬虫:研究综述[J].计算机科学,2009,36(8):26-29

[14] Davulcu H,Vadrevu S,Nagarajan S. OntoMiner:Bootstrapping Ontologies from Overlapping Domain Specific Web Sites[C]// Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters,2004. ACM Press,2004:500-501

[15] Wang Chao, Li Shu-qin, Xiao Hong. Research on Literature-based Automatic Ontology Construction Method for Agricultural Domain[J]. Computer Applications and Software,2014,31(8):71-74(in Chinese)
王超,李书琴,肖红.基于文献的农业领域本体自动构建方法研究[J].计算机应用与软件,2014,31(8):71-74

[16] Tang Qing,Lv Xue-qiang,Li Zhuo,et al. Research on Term Extraction for Domain Ontology[J]. New Technology of Library and Information Service,2014,30(1):43-50(in Chinese)
汤青,吕学强,李卓,等.领域本体术语抽取研究[J].现代图书情报技术,2014,30(1):43-50

[17] Maynard D,Li Y,Peters W. NLP Techniques for Term Extraction and Ontology Population[C]//Proceeding of the 2008 Conference on Ontology Learning and Population:Bridging the Gap between Text and Knowledge,2008. IEEE Press,2008:107-127

[18] Maynard D,Funk A,Peters W. SPRAT:A Tool for Automatic Semantic Pattern-Based Ontology Population[C]//International Conference for Digital Libraries and the Semantic Web. 2009

[19] Wu Y, Zhang S, Zhao W. Towards Learning Domain Ontology from Legacy Documents:Digital Society,2010[C]//Fourth International Conference on ICDS'10. IEEE Press,2010:164-171

[20] Sirin E,Parsia B,Grau B C,et al. Pellet:A Practical Owl-Dl Reasoner[J]. Web Semantics:Science,Services and Agents on the World Wide Web,2007,5(2):51-53

[21] Chen Yu,Zhu Jian-feng,Wu Yi-jian,et al. New Term Expansion Method Based on Domain Ontology[J]. Computer Engineering,2011,37(7):24-27(in Chinese)
陈宇,朱建锋,吴毅坚,等.一种基于领域本体的新术语扩充方法[J].计算机工程,2011,37(7):24-27

[22] Zablith F. Evolva:A Comprehensive Approach to Ontology Evolution[C]// The Semantic Web:Research and Applications,2009. Berlin:Springer,2009:944-948

[23] Li Jiang-hua, Shi Peng, Hu Chang-jun. Ontology Concept Learning Method for Compound Terms[J]. Computer Science,2013,40(5):168-172(in Chinese)
李江华,时鹏,胡长军.一种适用于复合术语的本体概念学习方法[J].计算机科学,2013,40(5):168-172

[24] Gu Jun,Xu Xin. Study on Ontology Relation Extraction in Chinese Patent Documents[J]. Computer Engineering,2013(10):73-78(in Chinese)
谷俊,许鑫.中文专利中本体关系获取研究[J].现代图书情报技术,2013(10):73-78

[25] Paiva L,Costa R,Figueiras P,et al. Discovering semantic relations from unstructured data for ontology enrichment:Asssociation rules based approach[C]//2014 9th Iberian Conference on Information Systems and Technologies (CISTI). IEEE,2014:1-6

[26] Faria C,Serra I,Girardi R. A domain-independent process for automatic ontology population from text[J]. Science of Computer Programming,2014,95(1):26-43

(上接第 205 页)

[16] Pandita R,Xiao X,Yang W,et al. WHYPER:Towards Automating Risk Assessment of Mobile Applications[C]//Proceedings of the 22nd Conference on USENIX Security Symposium,2013. 2013:527-542

[17] Gorla A, Tavecchia I, Gross F, et al. Checking app behavior against app descriptions[C]// Proceedings of the 36th International Conference on Software Engineering,2014. ACM,2014:1025-1035

[18] Hanna S,Huang L,Wu E,et al. Juxtapp:A scalable system for detecting code reuse among android applications[M]//Detection of Intrusions and Malware, and Vulnerability Assessment. Springer Berlin Heidelberg,2013:62-81

[19] Sun X,Zhongyang Y,Xin Z,et al. Detecting Code Reuse in Android Applications Using Component-Based Control Flow Graph [M]//ICT Systems Security and Privacy Protection. Springer Berlin Heidelberg,2014:142-155