·计算机工程与应用·

基于WordNet与Wikipedia的平面几何本体的构建

符红光[1]，刘 莉[2]，钟秀琴[1]，蒋 彦[1]，孙媛媛[1]

(1. 电子科技大学计算机科学与工程学院 成都 611731; 2. 西南民族大学计算机科学与技术学院 成都 610041)

【摘要】针对目前本体构建中存在的如手工构建难以确保高效性和可扩展性，且自动构建难度大，可操作性不强等研究现状，提出了一种基于WordNet和Wikipedia的学科领域本体半自动构建方法。首先构建一个领域顶层本体，在此基础上，重用WordNet的结构，从深度上对其进行术语和术语层次的扩展；同时根据Wikipedia中的页面信息，从广度上对其进行术语间关系的扩展和术语的补充；并将该本体构建方法应用于平面几何领域。实验表明该方法能大大提高本体构建的效率，并在一定程度上保证了本体的质量。

关 键 词 领域本体; 半自动构建; 维基百科; WordNet
中图分类号 TP31 文献标志码 A doi:10.3969/j.issn.1001-0548.2014.03.018

Semi-Automatic Construction of Plane Geometry Ontology Based-on WordNet and Wikipedia

FU Hong-guang[1], LIU Li[2], ZHONG Xiu-qin[1], JIANG Yan[1], and SUN Yuan-yuan[1]

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731;
2. School of Computer Science and Technology, Southwest University for Nationalities Chengdu 610041)

Abstract Ontology, as a member of the Semantic Web's hierarchical structure, is located in the central position. Regarding the current research situation of ontology construction, the manual construction is difficult to ensure its efficiency and scalability; and the automatic construction is hard to guarantee its interoperability. This paper presents a semi-automatic domain ontology construction method based on WordNet and Wikipedia. First, we construct the top-level ontology and then reuse WordNet structure to expand the terminology and terminology-level at the depth of the ontology. Furthermore, we expand the relationship and supplement the terminology at the width of the ontology by referring to page information of Wikipedia. Finally, this method of ontology construction is applied in elementary geometry domain. The experiments show that this method can greatly improve the efficiency of ontology construction and ensure the quality of the ontology to some degree.

Key words domain ontology; semi-automatic construction; Wikipedia; WordNet

本体作为语义网的核心，是在语义层和知识层描述信息的建模工具[1]。本体所描述的领域知识，不仅能被人所理解，更重要的是能让机器自动处理。因此，本体对于异构的、面向计算机的海量信息起着举足轻重的作用，它在知识工程、信息检索、信息集成等领域得到了广泛的应用。本体应用价值凸显的同时，该如何高效地构建本体，特别是在各种特定的领域，如何构建一个该领域特定的本体，成为当前研究的热点。

当前构建本体的方法包括人工构建、半自动构建和自动构建。人工构建本体需要大量的领域专家参与，虽然这种方法产生的结果更加准确，但是太费时费力，且构建的效果会受专家主观意识的影响。半自动本体构建，常常对一些本体构建中简单的子任务采用自动化技术，如OntoGen[2]是一个基于文本的半自动本体构建工具，它采用K均值聚类、单值分解及SUM分类的方法从资料库中抽取出领域概念，这些概念再由领域专家通过OntoGen GUI界面建立它们之间的关系。尽管半自动本体构建过程在一定程度上提高了构建效率，但是人工还是要承担大部分工作，也不能满足信息技术的发展。对于本体的自动构建，国内外已经提出了很多方法和工具。

收稿日期：2012 – 09 – 21; 修回日期：2014 – 01 – 14
基金项目：国家自然科学基金面上项目(61073099)；国家自然科学基金(61202257)
作者简介：符红光(1965 –)，男，博士，教授，主要从事计算机代数及人工智能方面的研究.

如本体学习工具OntoLearn[3]，它能抽取特定领域的概念，再借助WordNet区分出领域的专有概念，进而抽取领域概念的关系，形成新的本体。在中文本体自动构建方面，如本体学习框架GOLF[4]，它能利用领域中的Web数据，采用自然语言处理技术进行处理，通过WordNet、HowNet和领域词库识别出领域术语和实例，并利用模糊推理机制进行实例学习。文献[5]提出了中文的本体构建方法，利用中文字典CKIP和SOM，从非结构化文档中抽取词产生术语，然后在抽取出固定的语境内，利用汉字的语形学分析抽取术语间的关系，并利用模糊推理机制进行实例学习。

虽然中文本体半自动或自动构建的方法在国内外有着大量的基础性研究，但仍存在如下一些问题：

1) 语言的障碍。很多成功的本体构建方法基于某一特定语言，由于中西方语言差距较大，故无法直接适用。

2) 术语与关系的模糊。存在对术语的区分不清晰，对术语间的关系提取不明确。

3) 仍需大量人工参与。针对提取出的术语与关系的准确性，仍需领域专家的参与修正。

4) 理论不成熟。中文本体的自动构建方法大多处于理论模型的研究阶段，没有成型的可以应用于实际的系统。

为了更好地解决上述问题，本文在对现有的本体构建方法进行研究的基础上[6]，提出了一种新的基于WordNet[7]和Wikipedia[8]的领域本体半自动构建方法。该方法将通过计算术语匹配度来消除语言的障碍，通过深度和广度扩展来提取术语及其关系，并结合信息抽取技术设计规则系统来自动抽取关系。

1 领域本体半自动构建流程

文献[9]在分析了著名的本体设计项目(包括Cyc、WordNet等)后，结合开发经验，给出了一种构建本体的七步法，其领域本体的构建主要考虑两部分：领域知识层次结构和领域知识点间的关系。领域知识层次结构描述了领域知识的架构，它类似于字典的索引表，而领域知识点间的关系则描述了该领域术语及这些术语间的关系，它类似于字典的内容。结合七步法，本文提出一种基于通用本体WordNet和百科知识库Wikipedia的领域本体半自动构建方法，其构建流程如图 1 所示。

首先构建领域顶层本体。作为本体构建的初始本体，顶层本体在确定领域范围后，只需要构建出最基本的领域术语和层次结构即可。顶层本体的构建是手工的，并基于作者前期的工作[10-11]。

然后获取中文领域术语。该部分将直接从Wikipedia 的领域分类页面和链接页面中获取术语，提取领域术语的流程如下：

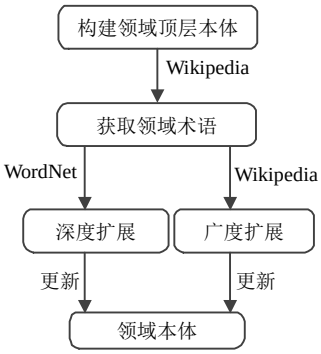

图 1 领域本体半自动构建流程

1) 从 Wikipedia“领域术语”分类页面获取领域术语，存入术语表。

2) 获取术语表中处于未分析状态的术语，访问术语的链接页面获取页面内容中的概念链接，如果不存在未分析的术语，则结束。

3) 访问术语链接页面，查找黑体术语与链接术语，将黑体术语存入术语表；并查看链接术语的链接名与页面名是否一致，如果一致，则将链接名存入术语表；否则将链接名丢弃。

4) 分析完成一个术语的链接页面后，设置该术语的状态为已分析，跳转到步骤 2)。

在此基础上，一方面，从深度上扩展领域本体的层次结构，即利用 WordNet 的继承关系，整体与部分关系等扩展层次结构；另一方面，从广度上扩展术语及其关系，即利用 Wikipedia 描述概念的页面，扩展更多的术语及术语之间的关系。并通过设计规则系统来自动抽取关系。

最后更新领域本体。本体的构建是一个循环迭代的过程，可定期访问 WordNet 和 Wikipedia，并不断进行本体的更新与完善。

总之，该方法将 WordNet 和 Wikipedia 作为领域本体的信息源，从中抽取了领域相关术语、术语的层次关系和术语间通用的语义关系，部分重用了领域相关的知识。接下来将详细地描述领域本体的深度扩展和广度扩展。

2 基于WordNet的深度扩展

2.1 WordNet解析

WordNet 是普林斯顿大学在 1985 年开发的一款轻量级本体，它作为语义词典根据词义而不是词形来组织词汇信息，并在这些词汇间建立关系，主要

包括：同义关系、反义关系、继承关系、整体与部分关系、相似关系、成员关系、领域关系、属性关系和扩展关系等。

WordNet不只是把单词以字母顺序排列，而且按照单词的意义组成一个“单词的网络”。并按照词性如名词、动词、形容词和副词各自被组织成一个同义词的的网络，每个同义词集合都代表一个基本的语义概念，并拥有唯一的索引号，每个同义词集合可以包含数个同义词，并给出了相应的定义和例句。

本文采用的版本是 WordNet 3.0。下面以“car”来说明词义和语义间的关系。将“car”输入 WordNet 中，通过应用程序，可以看到“car”有 5 个词义。

词义 1：{car, auto, automobile, machine, motorcar}；

词义 2：{car, rail car, railway car, railroad car}；

词义 3：{cable, car}；

词义 4：{car, gondola}；

词义 5：{car, elevator}。

可以看出每个词义是包含若干个单词的同义词集合，每个集合表示一个概念。WordNet 中的概念是由概念间的关系联系在一起，并形成一个网络。如图 2 所示为以“car”的词义 1 为例构建出的一个子网络，可以帮助理解 WordNet 的结构。

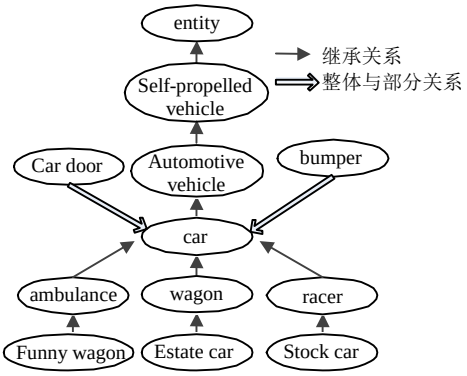

图 2　car 继承关系截图

2.2　深度扩展流程

从 WordNet 的结构可以看出，它可以从深度上对领域本体进行扩展。本文将 WordNet 作为本体层次扩展时使用的数据源，从中抽取适用于某个具体领域的本体。由于本文需要构建的本体是中文的，而 WordNet 是英文的，存在大量一词多义的情况。因此，需先将中文翻译成英文，再进行英文与 WordNet 的映射，最后获取概念的层次关系，其详细扩展流程如图 3 所示。

首先，将中文术语翻译成英文术语。从术语表中取出未翻译的术语，访问金山词霸网站接口，获取该术语的英文单词翻译。由于中文翻译为英文的过程中存在大量一词多义的情况，因此得到该术语的英文单词集合，并以集合为单位存入术语表，直到术语表中不存在未翻译的术语。

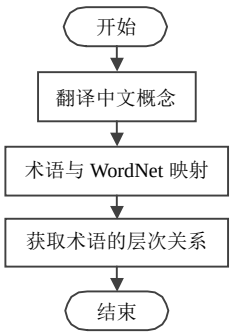

图3　基于WordNet扩展的流程图

然后将术语与 WordNet 进行映射。由于一词多义的情况，因此需要计算术语匹配度。其过程为获取英文单词集合中每个单词在 WordNet 中的所有同义词集合；再计算每个同义词集合的继承关系，并列关系和其自身存在于术语表中的数量，以该数量的多少作为匹配度(数量越多匹配度越大)；在此基础上，以术语匹配度最大的第一个单词为其英文义项。例如对术语“三角形”进行 WordNet 的映射，如图4所示。金山词霸对三角形的翻译有 triangle 和 trigon，在 WordNet 中 triangle 对应了5个义项，其中 SID-13879320-N 同义词集合的匹配度最大，因此，将“三角形”和同义词集合的第一个义项进行匹配。

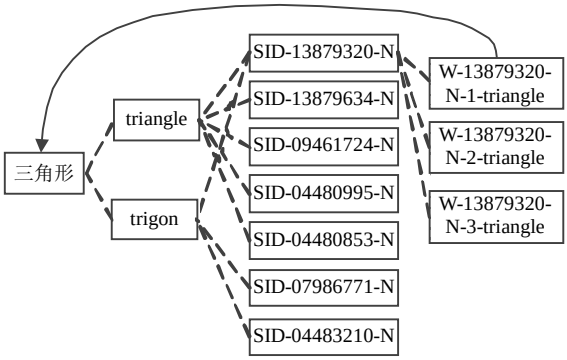

图4　概念“三角形”的映射图

最后获取术语的层次关系。其实现步骤描述如下：

1) 依次从 WordNet 中一词多义最少的单词开始，匹配术语表中未匹配过的英文单词，如果比较完 WordNet 中的所有单词或者匹配完术语表中的所有单词，则执行步骤 3)。

2) 如果找到英文单词，则将其翻译成对应的中文，此中文对应的所有英文单词则设为已匹配，执行步骤 1)。

3) 依次从未处理过的 WordNet 的最底层的已翻译成中文的术语开始，查找已翻译成中文的上位

术语，并标记关系，直至最顶层的“Entity”。例如标记“锐角”的层次关系，如图 5 所示。在 WordNet 中，“acute_angle#0”的上位是“oblique_angle#0”，但 oblique_angle 没有对应的中文解释。因此，在进行层次关系扩展时，将“acute_angle#0”的上位定为“angle#0”。

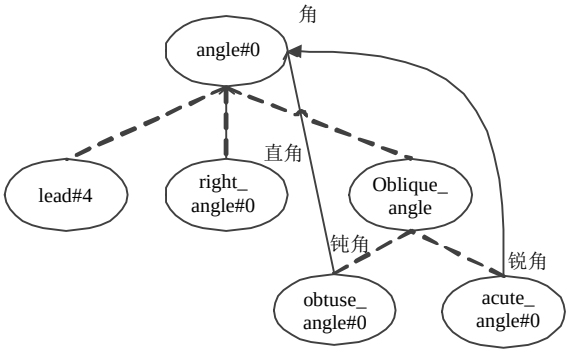

图5 概念“三角形”层次关系图

3 基于Wikipedia的广度扩展

3.1 Wikipedia解析

维基百科(Wikipedia)是基于 Wiki 技术的全球性多语言百科全书，也是一个动态的、可自由访问和编辑的知识库。维基百科自 2001 年 1 月 15 日成立以来，条目数第一的英文维基百科已拥有超过 300 万条条目，中文维基百科也已拥有了 30 万条条目。维基百科中的大部分页面可以由任何人使用浏览器进行阅览和修改。一个维基百科页面的主要标记如图 6 所示。

```
<!-- firstHeading -->页面概念<!-- /firstHeading -->
<!-- subtitle -->重定向概念<!-- /subtitle -->
<!-- bodytext -->页面主内容<!-- /bodytext -->
<!-- catlinks -->页面分类<!-- /catlinks -->
```

图 6 维基百科页面主要标记

根据对维基百科页面结构的分析，可以得出维基百科页面的如下特性：

1) 每个页面描述一个实体，且没有重复。这就意味着对每个概念不存在一词多义的情况，便于对概念进行处理。

2) 对于同一含义或相关含义的实体，维基百科会自动跳转到具体的描述页面。比如“长方形”会跳转到“矩形”，“等腰三角形”会跳转到“三角形”。

3) 页面中属于实体的，维基百科用超链接表示，因此页面之间都以超链接相连，并且能非常容易地获取概念实体。

4) 页面中存在大量的表格数据，这些表格数据主要描述了实体的属性信息，因此可以从中提取概念之间的关系。

5) 页面中常出现固定的语言模式，如“矩形又称长方形”，使得统计语义信息成为可能。

3.2 广度扩展流程

在已经获得的术语和术语间层次关系的基础上，由于 Wikipedia 对于本文已经处理的术语都有一个页面进行描述，因此本文将利用这些描述页面进行术语和关系的进一步扩展。使用 Wikipedia 进行术语和关系扩展的详细流程如图 7 所示。

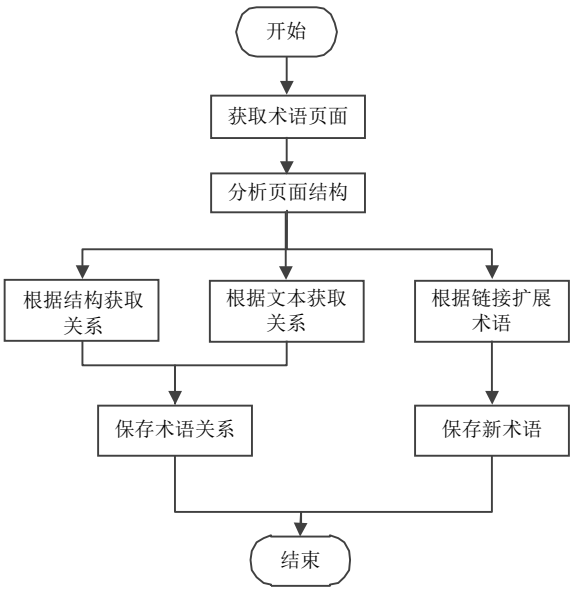

图7 基于Wikipedia扩展的流程图

根据对 Wikipedia 页面的结构分析以及使用自然语言处理技术对相应的文本进行解析，给出以下几种规则以获取术语及其关系。

规则 1：根据页面的目录进行关系抽取。

规则 2：根据页面的表格进行关系抽取。

规则 3：通过术语的定义获取关系。对术语的定义进行分词后，如果得到有效术语，则能确定术语间的依赖关系。

规则 4：通过句子的语法分析，提取术语之间的详细关系描述。

规则 5：根据页面中的黑体术语、链接术语，对比术语表中的术语，将术语表中不存在的术语进行添加。

综上所述，基于 WordNet 的深度扩展主要是从类层次关系等纵深维度对本体进行扩展，而基于 Wikipedia 的广度扩展主要是从其相关关系等方面进行相应的扩展。当然，不排除 WordNet 深度扩展中得到广度数据以及 Wikipedia 广度扩展中得到深度数据。例如三角形与直角三角形通过 WordNet 深度扩展得到其关系为直角三角形与三角形的关系为

继承关系；而通过 Wikipedia 广度扩展得到直角三角形与三角形的关系为下位关系。分析表明这些数据是无矛盾的，可以同时保留。

4　实验分析

根据上述提出的领域本体半自动构建流程，以及对本体的深度和广度扩展流程，本文设计了一个本体构建系统，其总体框架如图 8 所示。

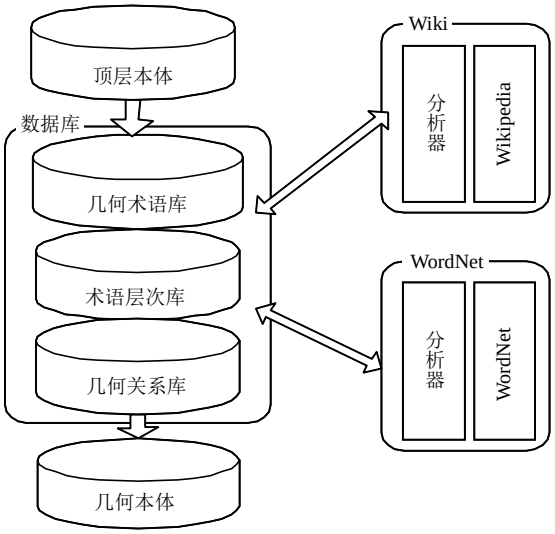

图8　本体构建系统框架图

本系统使用 Protégé 3.4.4 作为本体的编辑工具，Gruff 3.1.8 作为本体展示工具，在作者前期工作[10-11]的基础上，搭建起几何顶层本体。其中，顶层本体建立类包括公式类、公理类、关系类、单位类、定理类、性质类、术语类、运算类等，它们的父类均为 Thing。属性包括继承关系、同义关系、并列关系、整体与部分关系、相关公式、相关定理、相关公理、相关单位、相关性质、相关运算等。

在此基础上直接从 Wikipedia 的“几何术语”分类页面中提取几何术语，共获取术语 112 个，重定向术语 17 个。并根据 2.2 节所描述的基于 WordNet 的扩展流程进行深度扩展。以术语“直角三角形”为例，自下向上，查找已翻译成中文的上位术语，并标记关系，直至顶层术语，具体深度扩展片段如图 9 所示。

其中部分概念的关系需要调整，这是因为 WordNet 的分层和 Wikipedia 的顶层分层不一致，并且由词典翻译得来的英文不能完全与 WordNet 中的英文所匹配。通过 WordNet 的处理，最终获得了更多的术语和更多的层次关系，其中包括继承关系、整体与部分关系、近义关系、反义关系、并列关系等。

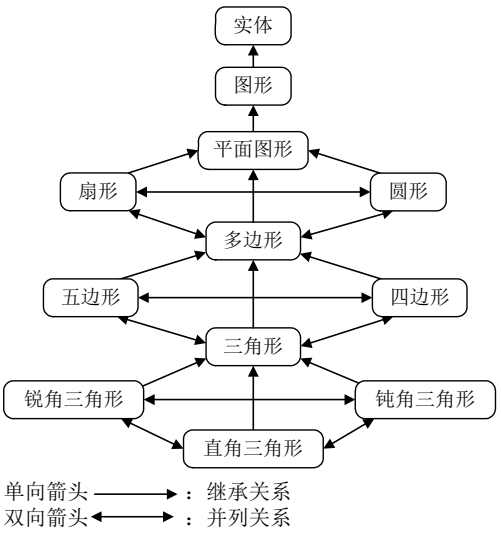

图9　深度扩展本体片段

在深度扩展的基础上，根据 3.2 节所描述的基于 Wikipedia 的扩展流程进行广度扩展。以三角形为例，在 Wikipedia 的搜索页面中输入“三角形”，根据规则 1，对分类页面进行整理，可以提取出<三角形，分类，直角三角形>、<全等三角形，相关定理，边角边定理>等关系。根据规则 2，如对三角形五心这个表格进行分析，可以得到<三角形，三角形五心，内心>等关系。根据规则 3，对术语的定义进行分词后，如果得到有效术语，则能确定术语间的依赖关系。如等边三角形的定义“等边三角形(又称正三角形)，为三边相等的三角形，其三个内角相等，均为 60°”进行分词，可以得到如下三元组：<正三角形，同义关系，等边三角形>、<等边三角形，上位关系，边>等。根据规则 4，对句子进行分析，如“等腰三角形中的两条相等的边被称为腰，而另一条边被称为底边。”可以得到等腰三角形和腰的关系：<等腰三角形，等腰三角形中的两条相等的边被称为腰，腰>等。其广度扩展本体片段如图 10 所示。

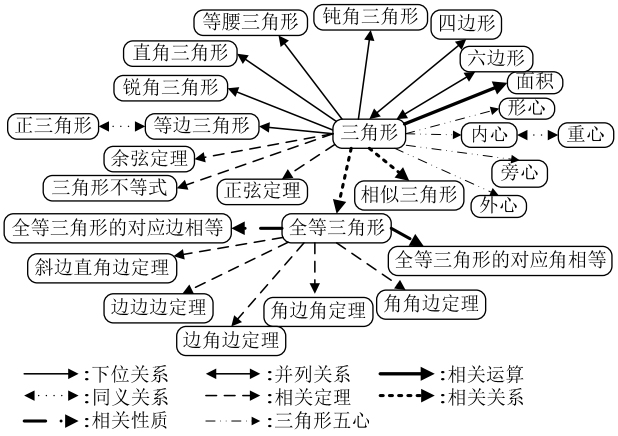

图10　广度扩展本体片段

通过 WordNet 和 Wikipedia 对本体中术语和术语关系的扩展，共获得新的术语 270 个，层次关系 288 个，一般关系 625 个，其数量对比如图 11 所示，从图 11 的实验数据可以看出，平面几何领域本体中的术语数目和关系数目大大增加。通过 WordNet 的深度扩展，建立起了术语的层次结构，使其形成了一个树状结构；通过 Wikipedia 对领域本体中的术语和关系从广度上进行扩展，形成了一个更完善的网状结构。几何本体构建完成后，就可以实现以此本体为基础的相关应用，例如资源标准、主题导学、资源推荐等。

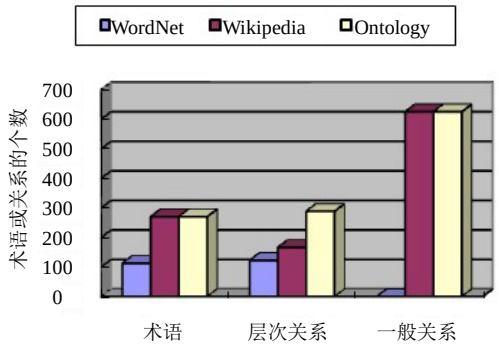

图11 本体与WordNet、Wikipedia的对比图

5 结 论

通过 WordNet 和 Wikipedia 与自然语言处理技术相结合的方法半自动构建本体，能够在领域顶层本体的基础上，进行本体的扩展，这是一种高效的领域本体构建方法。使用该方法向本体中添加更多的术语是一个迭代的过程，能够根据本体构建者的需求，很好地进行过程控制。基于规则的术语间关系的确定，对于领域本体中术语之间关系的扩展具有重要的意义。由于规则的构建是通过对 Wikipedia 中的页面信息进行挖掘，其中的语句能够建立更丰富的关系。

对语句中的语义关联进行挖掘，形成更强大的规则系统，这将是今后的研究重点，对本体的完善将有着重要意义。

参 考 文 献

[1] 高志强, 潘越, 马力, 等. 语义Web原理及其应用[M]. 北京: 机械工业出版社, 2009.
GAO Zhi-qiang, PAN Yue, MA Li, et al. Principle and application of the semantic Web[M]. Beijing: China Machine Press, 2009.

[2] FORTUNA B, GROBELNIK M, MLADENIC D. Semi-automatic construction of topic ontology[C]//The Second International WorkShop on Knowledge Discovery and Ontologies. [S.l.]: [s.n.], 2005.

[3] NAVIGLI R, VELARDI P. Learning domain on to ontologies from document warehouses and dedicated web sites[J]. Computational Linguistics, 2004, 30(2): 151-179.

[4] LIU Bai-song, GAO Ji. General ontology learning framework[J]. Journal of Southeast University, 2006, 22(3): 381-384.

[5] LEE Chang-shing, KAO Yuan-fang, KUO Yau-hwang, et al. Automated ontoIogy construction for unstructured textdocuments[J]. Data & KnowIedge Engineering，2006, 60(3): 547-566.

[6] SUCHANEK F M, KASNECI G, WEIKUM G. YAGO: a core of semantic knowledge unifying WordNet and Wikipedia[EB/OL]. [2012-08-12]. http://www2007.org/papers/paper391.pdf.

[7] Princeton university. Wordnet[EB/OL]. [2012-08-12]. http://wordnet.priceton.edu.

[8] Wikipedia.Wiki[EB/OL]. [2012-08-12]. http://zh.wikipedia.org/wiki.

[9] NOY N F, MCGUINNESS D L. A guide to creating your first ontology[EB/OL]. [2012-08-12]. http://protege.stanford.edu/publications/ontology_development/ontology101.pdf.

[10] 钟秀琴, 符红光, 丁盘苹. 基于本体和prolog的平面几何定理证明[J]. 电子科技大学学报, 2011, 40(3): 429-434.
ZHONG Xiu-Qin, FU Hong-guang, DING Pan-ping. Geometry theorem proving on ontology and prolog[J]. Journal of University of Electronic Science and Technology of China, 2011, 40(3): 429-434.

[11] 钟秀琴, 刘忠, 丁盘苹. 基于混合推理的知识库的构建及其应用研究[J]. 计算机学报, 2012, 4(35): 761-766.
ZHONG Xiu-qin, LIU Zhong, DING Pan-ping. Construction of knowledege base on hybrid reasoning and its application[J]. Chinese Journal Of Computers. 2012, 35(4): 761-766.

编 辑 税 红