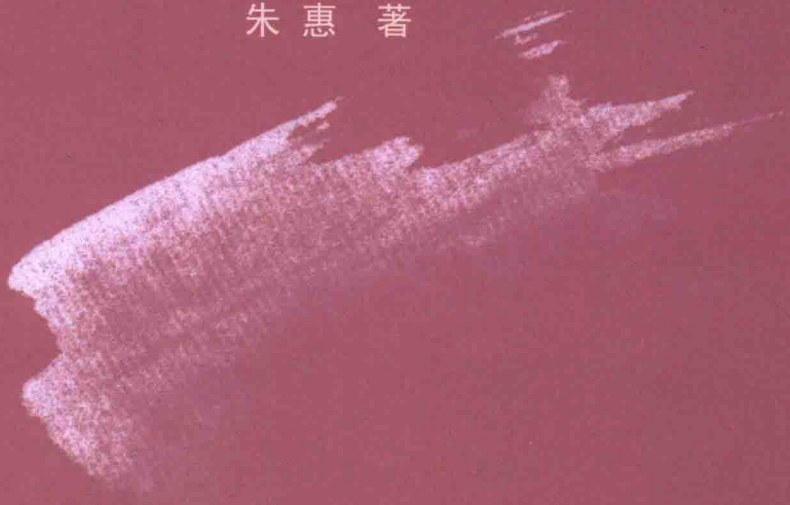
# 学科术语本体构建

Subject Term Ontology Construction

朱惠 著



南京大学出版社

# 学科术语本体构建

Subject Term Ontology Construction

朱惠 著

南京大学出版社

**图书在版编目(CIP)数据**

学科术语本体构建 / 朱惠著. —南京：南京大学出版社，2018.9

ISBN 978-7-305-20975-8

Ⅰ. ①学… Ⅱ. ①朱… Ⅲ. ①专业—术语—研究 Ⅳ. ①H083

中国版本图书馆 CIP 数据核字(2018)第 219670 号

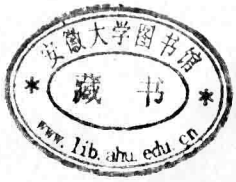
出版发行 南京大学出版社
社　　址 南京市汉口路 22 号　　邮　编 210093
出 版 人 金鑫荣

**书　　名 学科术语本体构建**
著　　者 朱　惠
责任编辑 卢文婷

照　　排 南京紫藤制版印务中心
印　　刷 常州市武进第三印刷有限公司
开　　本 787×1092　1/16　印张 10.5　字数 200 千
版　　次 2018 年 9 月第 1 版　2018 年 9 月第 1 次印刷
ISBN 978-7-305-20975-8
定　　价 36.00 元

网　　址 http://www.njupco.com
官方微博 http://weibo.com/njupco
官方微信 njupress
销售咨询 (025)83594756





# 序

本书作者朱惠博士在读博期间一直跟随我学习，于2015年获得南京大学管理学博士学位。她学习努力刻苦、悉心钻研，为专业研究打下了较为坚实的理论基础，同时也提高了自己的实践应用能力。在她的博士论文答辩过程中，评审专家对论文和答辩都给出了较好的评价。当她向我征询是否可以将博士论文修改后出版专著时，我积极鼓励并支持她完成这项工作。经过一年多的努力，她的书稿终于完成，我应邀写上数语，是为序，以鼓励她在未来的学术研究中取得更多更好的成绩。

本书的主题是利用本体学习方法和技术构建学科术语本体，这是一项颇具挑战性的工作，它突破了传统的手工构建知识本体的模式，引入知识自动抽取技术，通过计算机自动构建知识本体，大大提高了构建效率和质量。而早期构建知识本体是依靠本体工程师和领域专家通过手工方式来完成的，这样的构建方式不但耗费了大量的时间、人力和物力，还会受到领域专家知识结构、领域背景等主观因素的影响，缺乏客观性，从而导致结果出现偏差。现在借助本体学习方法和技术自动构建知识本体，能够避免手工构建模式的缺陷。本书正是从这一点出发开展研究，构建了领域本体学习系统模型，用于领域本体的自动生成，并对模型的合理性、可行性和有效性进行了实验验证，其研究成果在相关领域有很高的参考借鉴价值。

本书重点研究了本体学习系统的理论模型构建和关键技术解决，并将其有效地运用于“数字图书馆”这一学科领域，实现了学科知识的语义描述和合理组织，构建了学科术语本体。因此，该研究完善了知识发现方法，为有效的知识描述和知识组织提供了新的思路和途径，也为进一步的知识服务提供了结构基础。

对于呈现在读者面前的这本专著，我认为其价值主要有这样几个方面：

第一，建立了面向汉语非结构化文本的基于技术集成的领域本体学习系统模型。作者在总结分析国内外本体学习系统功能组成和学习流程的基础上，结合汉语非结构化文本的特点，集成多种数据挖掘技术和统计分析方法，建立了一个面向知识服务的领域本体学习系统模型，提出并论证了模型中关键组件的具体实现方案。

第二，实现了面向汉语非结构化文本的术语识别。目前中文自然语言处理处于瓶颈期，还不成熟，作者以关键词作为候选术语，经过统一规范、设置筛选指标和筛选原则获得了领域术语集。

第三，实现并验证了术语分类关系（层次关系）抽取的方法和策略。首先基于非结构化领域文档构建术语的向量空间模型，在此基础上，利用 BIRCH 预聚类和层次聚类相融合的两步聚类法挖掘术语的分类关系，并利用术语综合语义相似度指标确定各聚类类别的标签。

第四，实现并验证了术语非分类关系抽取的方法和策略。首先基于非结构化领域文档构建句子×术语向量空间模型，运用关联规则分析方法以及评价关联规则有效性和实用性的指标获取关联术语对，然后基于句子×〈术语，动词〉向量空间模型，再次利用关联规则分析方法和评价指标获取术语对与动词的关联规则，对两次关联规则分析的结果运用过滤规则进行过滤，获得最终的术语非分类关系及其标签。

第五，实现了学科术语本体的逻辑描述和关系数据库存储。运用网络本体描述语言 OWL 对构建的“数字图书馆”学科术语本体进行逻辑描述。OWL 把本体中的概念（术语）描述为类（Class），概念（术语）的分类关系通过 subClassOf 这一属性来描述，概念（术语）的非分类关系通过用户自定义属性来描述。同时，也为学科术语本体设计了关系数据库存储模式，通过 4 张表来存储术语、术语分类关系、术语非分类关系、属性等本体元素。

第六，实现了学科术语本体的可视化。运用本体编辑工具 Protégé 中的可视化组件 OntoGraf 对学科术语本体进行可视化展示，借助工具中的一些功能，可以对术语以及术语间的关系有更直观形象

的了解，并且可以从中发现新的领域知识。

面向汉语非结构化文本的本体学习是当前计算机科学和信息科学领域的前沿研究，是知识组织和知识管理得以实现和推广的技术基础。作者利用数据挖掘、统计分析等方法和技术，以“数字图书馆”学科领域的汉语非结构化文本为数据源，自动构建了术语本体，并实现了本体的描述、存储和可视化，力图在促进我国本体自动构建研究发展方面做出一定贡献。

作为一名青年学者，朱惠同志谦虚好学，勤于思考。这本书是她的第一本学术专著，作为她的导师，衷心期望她能以这本著作的出版为契机，不断进步和发展，为我国信息管理事业的发展做出更大的贡献。

苏新宁

2018 年 7 月 1 日

# 目　录

# 第1章 绪 论

## 1.1 语义网

万维网(World Wild Web,WWW)上信息资源的组织方式以超文本链接为基础,这样的组织方式使得人们能很好地理解和识别这些信息资源。随着越来越多信息资源的产生,人们发现这样的资源组织方式对于资源的检索和浏览越来越困难,因此,有必要对资源组织方式进行优化,以便让计算机能理解和识别这些信息资源[①]。针对这些问题,万维网的发明人即万维网联盟(W3C)的主任蒂姆·伯纳斯-李(Tim Berners-Lee)于1998年提出了语义网(Semantic Web,SWeb)这一新的概念[②],他希望能对信息资源进行语义描述,从而使得计算机能更好地理解信息资源的含义。相较于万维网,语义网是一种智能网络,它不但能够理解信息资源中的概念(术语),还能理解它们之间的逻辑关系,这使得人与计算机之间的交流变得更有效率和价值,也更轻松。图1-1展示了语义网的分层体系结构。

语义网的分层体系结构共有7层[③④]:

(1)"Unicode + URI"层(编码定位层)。该层是语义网的最底层,它是语义网体系的基础,主要功能是利用Unicode对资源进行编码以及利用URI(Uniform Resource Identifier,统一资源定位符)对资源进行标识。Unicode是一个字符集,共有65536个字符,基本上涵盖了世界上所有语言的字符,全部字符均由两个字节表示,这也是使用Unicode对资源进行编码的优势。URI主要用于对网络上的资源进行唯一标识。在语义网中,URI不仅能标识资源,还可以描

① 谷俊. 本体构建技术研究及其应用——以专利预警为例[D]. 南京:南京大学, 2012.

② Semantic Web Road Map [EB/OL]. http://www.w3.org/DesignIssues/Semantic.html, 2014-05-05.

③ Semantic Web [EB/OL]. http://en.wikipedia.org/wiki/Semantic_Web. 2014-05-05.

④ 语义网 [EB/OL]. http://baike.baidu.com/link?url=NBEP7QD36iYdTv-bFDgXbn_hjBuuVddcx_4X9kQ9z1QkxNmfxy1uy8j64Kynoh2Gbka4FzsFRhSINGN8zZr3Pq, 2014-05-05.

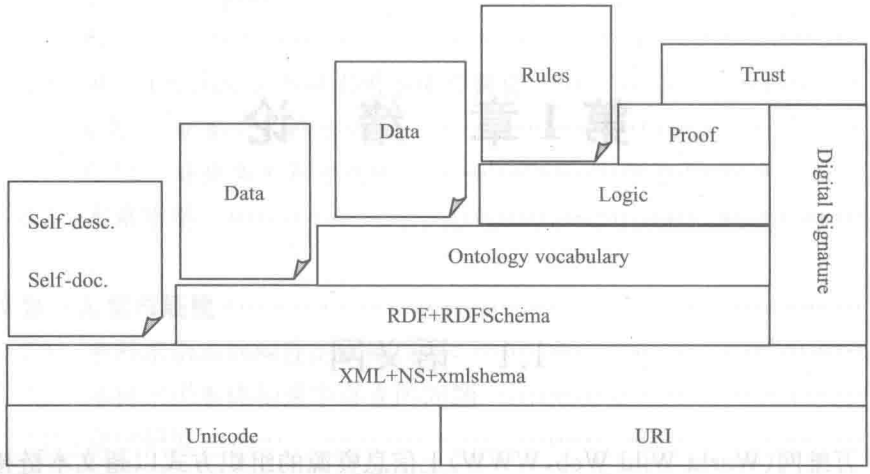



**图 1-1　语义网分层体系结构**[②]

述资源间的关系。

(2) "XML + NS + xmlschema"层(XML 结构层)。该层的主要功能是利用 XML(eXtensible Markup Language,可扩展标记语言)从语法上描述数据的内容和结构,通过 XML 将网络资源的数据结构和内容进行分离,这有利于对资源进行维护和管理。XML 不仅具备 SGML(Standard Generalized Markup Language,标准通用标记语言)的丰富功能,而且还具备 HTML(Hyper Text Markup Language,超文本标记语言)的易用性。XML 允许用户在不说明结构含义的情况下在文档中加入任意的结构。NS(Name Space,命名空间)的作用是为了避免不同的应用使用相同的字符描述不同的事物,它由 URI 索引确定。xmlschema 用 XML 语法来描述多种数据类型,是 DTD(Document Type Definition,文档类型定义)的替代品,但它比 DTD 更灵活,所以,能更好地为有效的 XML 文档服务并提供数据校验机制。XML 灵活的结构、NS 的数据确定性以及 xmlschema 所提供的多种数据类型及检验机制,使得它们成了语义网体系结构的重要组成部分。

(3) "RDF + RDF Schema"层(资源描述层)。该层的主要功能是对网络信息资源以及资源间的关系进行描述。RDF(Resource Description Framework,资源描述框架)是由万维网联盟(W3C)在 1997 年开发的一种描述网络信息资源的语言,它的目标是构建一个供多种元数据标准共存的框架。该框架能充分利用各种元数据的优势对网络数据进行交换、共享和再利用。RDF 采用 XML 无二义性地描述资源,且可以将资源间复杂的关系分解为若干个"主语—谓语—宾语"的三元组形式。采用 RDF 描述的资源以及资源间的关系能被计算机理解。

RDF Schema 丰富了 RDF 的框架，使用一种计算机可理解的体系定义了一组概念(术语)，这些概念(术语)可以用来描述资源以及资源间的关系。例如，rdfs:class 可以用来定义一个概念(术语)，rdfs:subClassOf 可以用来定义某个类是另一个类的子类。

(4) "Ontology vocabulary"层(本体层)。该层是语义网分层体系中的核心层，主要功能是提供网上互操作体之间关于信息资源的共同理解，即"语义"，故担当着语义互操作的重要角色。本体层在 RDF(S)基础上定义概念以及概念间的关系，用来描述特定领域的知识。在具体应用过程中，利用 RDF(S)定义网上信息资源，再利用本体定义互操作的语义空间，从而构成一个基本的语义网应用环境。本体层的存在，增强了信息资源的语义表达能力和逻辑推理能力。目前，一般采用 OWL(Web Ontology Language，网络本体语言)对本体进行描述，包括对概念以及概念间关系的描述，除此之外，OWL 还可以用来对不同系统的本体库进行链接。

(5) "Logic"层(逻辑层)。该层的主要功能是提供公理和推理规则，为智能推理提供基础，通过逻辑推理对资源、资源之间的关系和推理结果进行验证，证明其有效性。在利用 RDF(S)和本体对资源进行描述后，语义网需要通过具有强大逻辑推理能力的智能代理(Agent)搜集和处理信息，而仅依靠本体层的推理能力并不能满足这一需求，因此，必须要有一套与语义网开放、分布式体系结构相适应的规则系统来做这项工作，这便是逻辑层的任务。

(6) "Proof"层(验证层)。该层的主要功能是提供一种验证机制，执行逻辑层产生的规则。为保证逻辑层工作的可靠性，验证层使用逻辑层的规则、本体层的数据描述和本体层的逻辑推理进行"验证"，从而为数据或结论的可靠性提供保证。

(7) "Trust"层(信任层)。该层的主要功能是提供一种信任机制，保证资源交互的安全可靠。信任层位于语义网分层体系结构的最顶层，通过 Proof 交换和数字签名(Digital Signature)建立信任关系，保证语义网的可靠性。

从上述分析可以知道，本体层位于语义网分层体系结构的第 4 层，而且是整个体系结构的核心层。因此，本体是语义网的重要组成部分，是进行信息资源交互与共享的基础，对其进行研究能促进语义网的发展，研究人员也越来越关注对本体构建的研究。

## 1.2 本体构建

最初对本体的构建都是通过人工方式进行，由本体构建工程师和领域专家对领域知识进行建模，抽取领域概念以及概念间的关系。由于领域概念众多，且

概念间的关系复杂，并且还有不容易被研究人员察觉的潜在的概念间关系，因此，本体构建工作是一项庞大的系统工程。手工构建方式需要耗费大量的人力、时间和精力，且由于领域专家知识背景和结构的局限性，获得的本体往往带有偏见，具有误差倾向。因此，研究人员正在考虑是否可以利用知识自动抽取技术来自动构建本体，以降低本体构建的成本，这也形成了一个很有意义的研究方向——本体学习(Ontology Learning)。本体学习也可称为本体抽取(Ontology Extraction)、本体生成(Ontology Generation)或本体获得(Ontology Acquisition)。本体学习是利用一些方法和技术自动或半自动地从自然语言文本生成本体，包括抽取领域概念、概念间的关系、公理以及其他的本体元素，并利用本体描述语言对它们进行编码，从而对领域知识进行检索、更新、共享和再利用，同时也为基于本体的相关应用提供基础①。

本体学习需要利用数理统计、数据挖掘、机器学习等方法和技术，通过计算机自动或半自动地从已有的数据资源中发现概念和概念间的关系，再根据获得的本体元素的准确性和可靠性，由领域专家辅以修正和评价，从而获得期望的知识本体②。

## 1.3 本体学习的研究历程和现状

本体学习的研究起源于20世纪90年代，目前国外学者已经取得了比较丰富的成果，提出了一系列用于实现本体学习的方法和技术，并在此基础上成功开发了若干个可实用的字符语系本体学习系统，在一定程度上满足了自动或半自动构建领域本体的需要。这些本体学习系统存在如下特点：(1) 多支持非结构化和半结构化数据源；(2) 均集成了多种本体学习方法技术，以适用于不同本体元素的自动识别；(3) 自动生成公理的系统很少，目前只有 Hasti③ 支持从文本中获得公理。这些本体学习系统已经陆续应用于生物基因、化学化工等领域，开发出了一些重要的领域本体。

进入21世纪以后，国内学者开始借鉴国外的开发模式，在现有资源基础上以手工方式构建领域知识本体，并逐渐关注本体的自动构建研究，陆续出现了对外文资料的综述和局部技术的实现，部分学者开始探讨面向特殊结构数据的本体学习。

① Ontology Learning [EB/OL]. http://en.wikipedia.org/wiki/Ontology_learning, 2014-05-07.

② 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.

③ Amir Kabir University of Technology 开发的一个本体学习工具。

### 1.3.1 国外研究概述

针对各领域中存在的结构化程度不同的信息资源，国外学者提出并验证了其所适用的多种本体学习方法和技术，包括语言分析、机器学习、模板驱动、聚类挖掘、关联规则挖掘、形式概念分析、基于潜在语义索引的奇异矩阵分解、基于规则的方法、图形映射、数据库逆向工程、模型映射等。在采用上述方法和技术处理各类数据时，能够获得本体中的元素：概念、概念间关系、公理等。所采用的方法不同，得到的结果可能也不尽相同。研究人员目前主要聚焦于概念和概念间关系这两种本体元素的自动或半自动抽取研究，公理的抽取相对困难。将上述经过实验论证的有效方法和技术结合在一起，即可形成适合特定语种、面向特定数据结构的本体学习系统，典型的如 Text-To-Onto、OntoLearn、Hasti、OntoBuilder 和 OntoLiFT。

◆ 本体学习系统

德国 Karlsruhe 大学的 Alexander Maedche 等人提出了一个本体学习框架，如图 1-2 所示[①②]。该框架包括 4 个主要部件：(1) 文本处理组件，其包含文本处理管理（Text & Processing Management）和文本处理服务器（Text

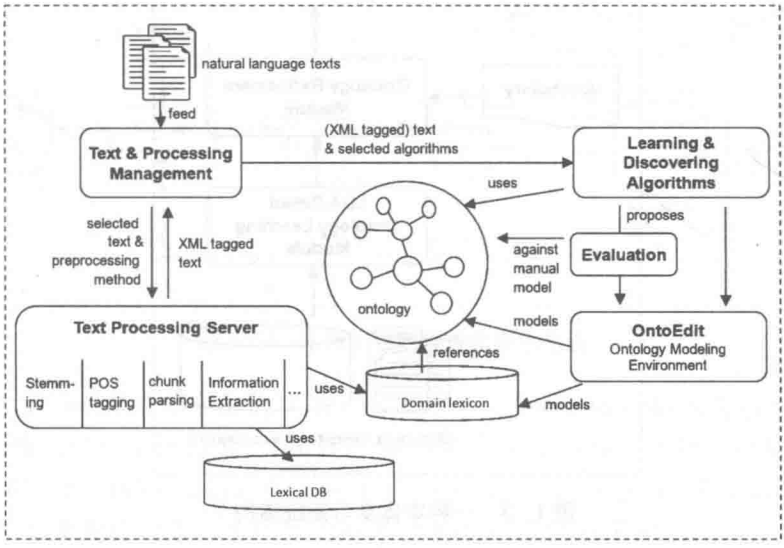



图 1-2 一种本体学习框架[③]

① Maedche A, Staab S. The Text-To-Onto ontology learning environment[C]. Citeseer, 2000.

② Maedche A, Staab S. Ontology learning for the semantic web[J]. Intelligent Systems IEEE, 2001, 16(2): 72-79.

③ Maedche A, Staab S. The Text-To-Onto ontology learning environment[C]. Citeseer, 2000.

Processing Server)，主要功能是选取合适的处理方法并且基于词典数据库对自然语言文本进行相关处理，返回 XML 标签文本；(2) 本体学习算法库(Learning & Discovering Algorithms)，主要功能是针对 XML 标签文本，选取合适的算法；(3) 本体编辑组件(OntoEdit)，主要功能是为本体建模提供适当的环境，创建本体；(4) 本体评估组件(Evaluation)，主要功能是对本体学习获得的本体和人工构建的本体进行对比评价。Alexander Maedche 等将该学习框架应用在了 KAON 和 Text-To-Onto 系统中。

Francesco Colace 等认为语义网的成功很大程度上依赖于有效的本体，它使得机器能理解相关数据。然而，目前仍没有一般的方法来自动获得本体，特别是基于各种资源运用自然语言处理和机器学习技术获得领域本体。他们提供了一个结合了统计和语义方法的本体学习系统(参见图 1-3)，并进行了实验验证，说明了相关方法的有效性。

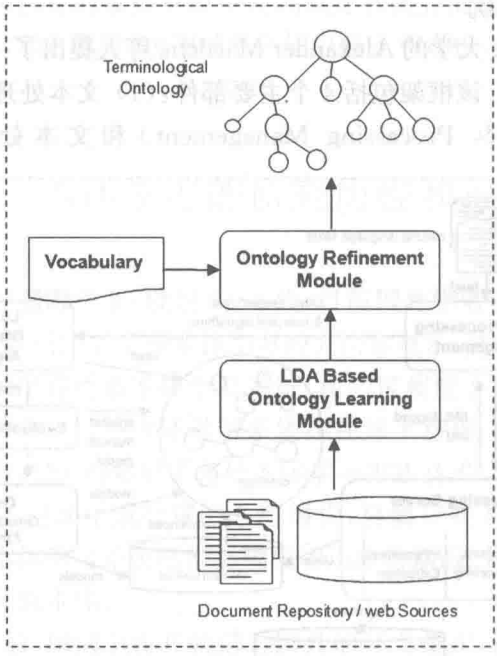



**图 1-3　一种本体学习系统结构**[①]

Brewster 等提出了一种基于用户的本体学习框架。该框架的流程如下：(1) 用户选择一批相关文档和种子本体，种子本体必须有词库予以支撑(每个节

① Colace F, Santo M D, Greco L, et al. Terminological ontology learning and population using Latent Dirichlet Allocation[J]. Journal of Visual Languages and Computing, 2014, 25(6): 818-826.

点至少一个词);(2) 系统自动从来源文档中抽取包含上述词库内容的句子作为样本,交由用户检验;(3) 验证通过后,系统进行自动学习获得正确与错误样本的学习规则;(4) 系统运用以上学习规则从来源文档中继续抽取相应样本,交由用户验证;(5) 系统得到用户的验证反馈后自动修正抽取规则,进一步抽取;(6) 直到结果满足用户的需求后,停止抽取过程,最后交给用户检验并进行人工修正[①]。

M. Missikoff 等人借助 OntoLearn 工具,研究开发了一套用于本体构建和本体评估的软件,可以为虚拟社区提供智能信息服务。该软件的流程如下:(1) 利用 OntoLearn 工具从相关的网络文档中抽取概念和概念间的分类关系,生成领域概念集;(2) 使用 Consys 进行机器辅助的概念有效性评估;(3) 最后在 SymOntoX 环境中,由本体工程师最终确定本体中的概念和概念间的关系。此外,SymOntoX 还可以自动将已学习的概念树附加到顶层本体相应节点中,丰富和完善现有本体,并对其进行一致性检测,从而最终完成本体半自动构建。整个流程如图 1-4 所示。

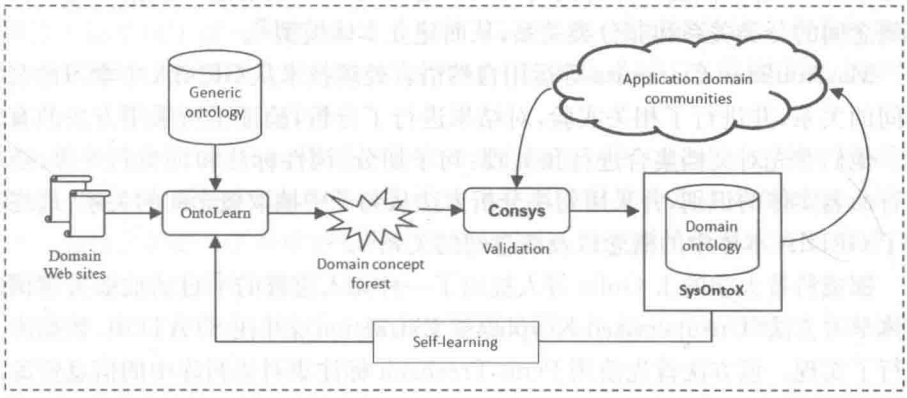



**图 1-4 本体半自动构建和评估系统[②]**

N. Weber 等介绍了如何从 Wikipedia、Wiktionary 和 DWDS 等在线词典中抽取领域词汇和实体标签,并描述了如何构建相应的领域本体。他们开发了 ISOLDE 系统(Information System for Ontology Learning and Domain Exploration),该系统使用命名实体识别技术(NER)抽取已有本体中的概念,随后抽取概念间的逻辑关系,最后利用在线词典等网络资源对现有概念的语义进行扩

① Brewster C, Ciravegna F, Wilks Y. User-centred ontology learning for knowledge management [J]. Natural Language Processing and Information Systems, 2002: 203-207.

② Missikoff M, Navigli R, Velardi P. Integrated approach to web ontology learning and engineering [J]. Computer, 2002, 35(11): 60-63.