信息抽取中领域本体的设计和实现

于江德[1]，李学钰[1]，樊孝忠[2]

(1. 安阳师范学院计算机与信息工程学院 河南 安阳 455002; 2. 北京理工大学计算机科学技术学院 北京 海淀区 100081)

【摘要】领域本体在信息抽取系统有着重要作用。该文介绍了本体、领域本体的基本概念，并探讨了领域本体与信息抽取的关系。提出了信息抽取系统中领域本体的设计准则和实施步骤，实施步骤包括领域本体需求分析、收集本体信息、构建领域本体框架、形式化编码、确认和评价等。在信息抽取原型系统中实现了显示器领域本体，并将该领域本体应用到信息抽取中的命名实体识别、抽取模式获取和主题概念提取等任务中，应用结果表明该方法、步骤是可行的。

关 键 词 概念; 本体构建; 信息抽取; 领域本体

中图分类号 TP391 **文献标识码** A

Design and Implementation of Domain Ontology for Information Extraction

YU Jiang-de[1], LI Xue-yu[1], and FAN Xiao-zhong[2]

(1. School of Computer and Information Engineering, Anyang Normal University Anyang Henan 455002;
2. School of Computer Science and Technology, Beijing Institute of Technology Haidian Beijing 100081)

Abstract Domain ontology plays an important role in the system of information extraction. In this paper, after a discuss of the relation of domain ontology and information extraction, the design principles and steps of domain ontology in the system of information extraction are proposed. The domain ontology about monitor is implemeneted in a prototype system of information extraction, and the domain ontology is applied to some tasks of information extraction, including named entity recognition, extraction pattern acquisition, and thematic concept extraction. The application results show that the principles and steps are feasible.

Key words concept; domain ontology; information extraction; ontology constructing

近十多年来，本体(ontology)被广泛应用于信息科学和计算机领域，并已成为当今信息科学研究的一个热点。在知识发现和管理[1]、知识库设计和集成[2]、信息检索和抽取[3]等领域，本体扮演着越来越重要的角色。信息抽取(information extraction)是从自然语言形式的文本中抽取用户感兴趣的事实、事件以及卷入其中的特定类型的实体等信息，并将这些信息转换为结构化的数据并存储的过程。作为一种自然语言处理系统，信息抽取系统需要强大知识库的支撑[4]。在不同的信息抽取系统中，知识库的结构和内容是不同的，但一般来说，都要有一个领域本体(domain ontology)，该领域本体通常是面向特定领域或场景的，是通用概念层次模型在特定领域或场的细化或泛化。基于领域本体的信息抽取系统能提供用户感兴趣的特定信息，并通过领域本体为信息源提供必要的语义标注信息，从而使系统对领域内的概念、概念之间的联系有统一的认识，可进一步提高系统的查准率和召回率，为用户提供更有价值的信息。所以，领域本体对信息抽取系统有至关重要的作用。近几年来，构建领域本体的方法已经成为一个新的研究热点。本文旨在探索信息抽取系统中领域本体的构建方法。

1 领域本体与信息抽取

1.1 本体的基本概念

本体是从西方语言的“ontology”转译而来，原本是一个哲学概念，指哲学中研究世界本原或本性的部分。哲学上把本体定义为“对世界上客观事物所进行的系统描述”[5]。文献[6]最早将本体定义为“给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延的规则的定义”，使得知识工程中的本体研究有了一个基

收稿日期：2007－06－09; 修回日期：2008－01－12
基金项目：教育部博士点基金(20050007023)
作者简介：于江德(1971－)，男，博士，讲师，主要从事自然语言处理、信息抽取、信息检索等方面的研究.

本方向。文献[7]给出了在信息科学领域被广泛接受的定义“An ontology is an explicit specification of a conceptualization”。文献[8]试图明确说明本体和概念化两者之间的差别，把概念化对象C定义为$C=(D,W,P)$，其中D是一个领域；W是该领域中相关事务状态的集合；P是领域空间$\langle D,W\rangle$上概念关系的集合。

1.2　领域本体的概念

将本体引入信息科学，就是从语义层次上考察事物的运动状态及状态的变化方式，把本体意义上的信息赋予更具体的内涵。一般而言，一个本体由以下几个方面构成：该领域概念类的层次体系、概念类的属性及属性的取值范围、概念之间其他的语义关系、一定的推理规则等。

领域本体是用于描述特定领域知识的一种专门本体。它给出了领域实体概念、领域属性概念、领域属性值及相互关系，以及该领域所具有的特性和规律的一种形式化描述。文献[2]认为，领域本体由属性、对象、关系和子领域本体组成。

1.3　领域本体在信息抽取中的应用

领域本体通过对特定领域内概念及概念间关系的精确描述，成为人机之间、机器与机器之间互相理解的语义基础。在信息抽取系统中，特定关系的抽取、事件的抽取都需要浅层的句法分析，并需要进行一定的篇章分析与推理。这些分析都需要领域内的语义信息提供坚实的支持。信息抽取按抽取对象的不同分为命名实体识别、实体关系抽取和事件抽取3个不同的任务。在这3个不同层次的任务中，领域本体都有大量的应用，能有效地提高抽取的查准率和召回率，为用户提供更有价值的信息。

2　信息抽取中领域本体的设计

出于对各自问题领域和具体工程的考虑，在构建领域本体时所遵循的标准和过程各不相同。下面结合信息抽取系统的具体情况，给出领域本体的设计准则和实施步骤。

2.1　领域本体的设计准则

就本体的设计准则而言，文献[9]总结的本体设计准则有：清楚、一致、可扩展、最小本体承诺、编码偏好程度最小等。结合信息抽取系统的具体情况，信息抽取系统中领域本体的构建准则如下。

(1) 清晰明了。领域本体必须能有效地说明所定义的类、概念、属性、属性值的含义。领域本体中所有的类名、概念、属性等术语应该能清楚地表达所要传递的意义，不能有二义性。

(2) 一致性。领域本体应该是一致的，也就是说，领域本体所定义的公理、某些推理以及领域本体的描述文档都应该具有一致性。

(3) 可扩展性。领域本体应该为可预料到的任务提供概念基础，并可支持在已有概念基础上定义新的术语，以满足特殊需求，而无需修改已有的概念定义。

(4) 简洁高效编码。概念的编码应该简洁高效，以便计算机容易处理。

2.2　领域本体的构建步骤

在应用实践中也形成了几种构建领域本体的方法，如骨架法[9]、企业建模法[10]、METHONTOLOGY法[11]等。参照这些构建领域本体的方法，结合信息抽取中的具体情况，并参考软件工程中的某些思想，领域本体的构建步骤如图1所示。

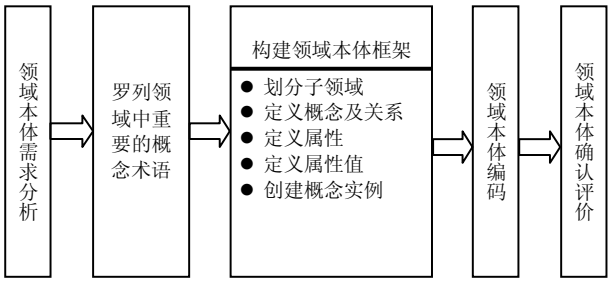

图1　领域本体的构建步骤

1) 需求分析。确定领域本体的应用目的、领域与范围、表示方法与用途等。

2) 罗列领域中重要的概念、术语。在领域本体创建的初始阶段，尽可能地列举出该领域内所有能够看到、想到的概念和术语。

3) 构建领域本体框架。在步骤2)中罗列出的领域中的大量的概念、术语，是一张无组织结构的词汇表，需要按照一定的逻辑规则对它们进行分类，形成不同的子领域，在同一子领域下的概念、术语，相关性较强。另外，对其中的每一个概念、术语的重要性要进行评估，选出关键性的概念、术语，摒弃那些不必要或者超出领域范围的概念、术语，并确立概念及概念间的等级关系，尽可能准确而简洁地表达出领域的知识，从而形成一个领域知识的框架体系，得到领域本体的框架结构。构建领域本体框架包括：

(1) 定义领域本体中的类，即划分子领域本体；

(2) 定义领域本体中的概念及概念间的关系，即采用自上向下的方法定义领域本体中的概念(先定义领域中综合的、概括性的概念，然后逐步细化、说明)；

(3) 定义属性值(属性值既可以是一个具体数值也可以是一个描述)，即通过属性值来说明属性的取值类型、值个数及有关值的其他特征；

(4) 创建实例，即创建概念的特征词。

4) 对领域本体编码、形式化。选用合适的本体描述语言对上述所建立的领域本体进行编码、形式化，以便对领域本体进行计算机处理。

5) 领域本体的确认和评价。评价包括本体的清晰性、一致性、可扩展性等方面。

3 信息抽取中领域本体的实现

依据上面关于领域本体的设计准则和实施步骤，本文在开发受限领域信息抽取原型系统时，采用手工和半手工的方式实现了显示器领域本体，在系统的多项应用中效果显著。

3.1 显示器领域本体的框架结构

本文构建的信息抽取原型系统主要用于对Web页面中的显示器产品信息及相关评价等信息进行抽取。为了更好地识别该领域的命名实体，判定网页或某些段落所评述的主题概念，在关系抽取模式、事件抽取模式获取时进行抽取模式的自扩展。本文利用自己开发的辅助工具构建了显示器领域本体。该领域本体由显示器领域的实体概念、实体概念特征词、属性概念、属性概念特征词、属性值概念、属性值概念特征词以及概念关系、概念属性关系等十多个数据表组成。下面对较重要的数据表及其作用分别叙述如下。

(1) 实体概念数据表用于保存显示器领域中可能的实体概念。最初建立时通过人工查找、辨别受限领域的实体概念，并输入到数据库中；以后逐渐通过计算机识别领域实体概念，并增加到数据库中人工进行查验。

(2) 与实体概念数据表相对应的是实体概念特征词数据表，该数据表主要用于存储每个实体概念在真实文本中可能出现的变形术语或形式。例如对于实体概念数据表中的实体概念“显示器”而言，在真实文档中有可能变形为“监视器”、“监示器”、“显示屏”、“显示屏幕”，所以在实体概念特征词数据表中与之相对应的特征词有监视器、监示器、显示屏、显示屏幕、显示器。

(3) 属性概念数据表用于保存领域中可能的属性概念。与属性概念数据表相对应的是属性概念特征词表，该数据表主要用于存储属性概念在真实文本中可能出现的变形术语或形式。

(4) 属性值概念数据表用于保存领域中可能的属性值概念。与属性值概念数据表相对应的是属性值概念特征词数据表，该数据表主要用于存储每个属性值概念在真实文本中出现的术语或形式。

(5) 概念关系表主要描述领域的实体概念之间存在的各种关系。实体属性和属性值关系表主要描述受限领域的实体、属性、属性值之间的关系。

3.2 领域本体中概念间关系的实现

显示器领域本体中概念间的关系的表示通过几个概念关系表和各数据表之间的关系来实现。受限领域本体由十多个数据表组成，它们不仅描述受限领域的实体概念、属性概念、属性值和相对应的特征词，而且对实体概念之间的关系、实体概念、属性、属性值之间的关系进行描述。各数据表之间存在着多种关系，实体概念表和实体概念特征词表、属性概念表和属性概念特征词表、属性值表和属性值特征词这3对数据表中的每一对都是概念和概念对应特征词的关系。属性值是某个属性的具体值；属性又是某个实体概念的属性。

而实体概念关系表描述的是实体概念表中的两个实体概念之间的关系。实体概念、属性、属性值关系表描述的是实体概念、属性、属性值之间的相互关系，如图2所示。

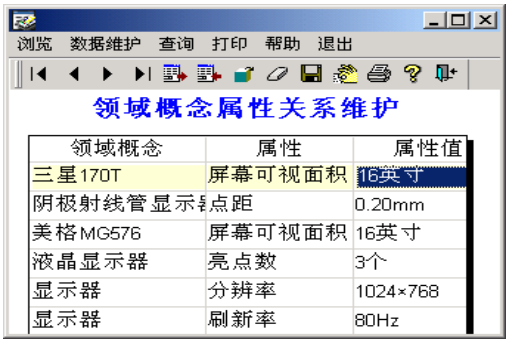

图2 实体概念、属性、属性值关系示意图

3.3 显示器领域本体的具体应用

本文的信息抽取原型系统的命名实体识别、实体关系抽取和事件抽取3个抽取任务3，对显示器领域本体都有大量的应用，主要集中在如下几方面。

(1) 在命名实体识别阶段，领域本体能够提供大量语义信息，这些语义信息对识别命名实体的类别有非常重要的意义。而且领域本体能够用于术语整合和相同命名实体的有效判定。

(2) 信息抽取系统基本上都是基于模式匹配的，即首先从文本中学习出感兴趣的关系或事件抽取模式，然后再用抽取模式去发现新的关系和事件。而领域本体的语义信息能够用于抽取模式的获取过程

和关系抽取模式、事件抽取模式的自扩展过程，能够对已有的抽取模式进行语义扩展。

(3) 另外在进行文档或某些段落的主题概念提取时，领域本体也有重要作用。

4 总　结

近几年来，本体的研究和应用在知识工程、自然语言理解等领域日益受到重视。本体的设计是一个创造性的过程，而领域本体的构建更是一个极具挑战性的工作。本文给出了信息抽取系统中领域本体的设计准则和实施步骤，并采用手工和半手工的方式实现了一个显示器领域本体。在信息抽取原型系统中的应用表明该准则、步骤符合人们的思维认知，逻辑性强，并具有良好的可操作性以及可拓展性。今后将致力于信息抽取系统中领域本体的自动构建研究。

参　考　文　献

[1] 王海涛, 曹存根, 高　颖. 基于领域本体的半结构化文本知识自动获取方法的设计和实现[J]. 计算机学报, 2005, 28(12): 2010-2018.
WANG Hai-tao, CAO Cun-gen, GAO Ying. Design and implementation of a system for ontology-mediated knowledge acquisition from semi-structured text[J]. Chinese Journal of Computers, 2005, 28(12): 2010-2018.

[2] 陈　刚, 陆汝钤, 金　芝. 基于领域知识重用的虚拟领域本体构造[J]. 软件学报, 2003, 14(3): 350-355.
CHEN Gang, LU Ru-qian, JIN Zhi. Constructing virtual domain ontologies based on domain knowledge reuse[J]. Journal of Software, 2003, 14(3): 350-355.

[3] 万　捷, 滕至阳. 本体在基于内容信息检索中的应用[J]. 计算机工程, 2003, 29(4): 122-123.
WAN Jie, TENG Zhi-yang. Application of ontology in content-based information retrieval[J]. Computer Engineering, 2003, 29(4): 122-123.

[4] 俞士汶, 段慧明, 朱学锋, 等. 综合型语言知识库的建设和利用[J]. 中文信息学报, 2004, 18(5): 1-10.
YU Shi-wen, DUAN Hui-ming, ZHU Xue-feng, et al. The construction and utilization of a comprehensive language knowledge-base[J]. Journal of Chinese Information Processing, 2004, 18(5): 1-10.

[5] 李善平, 尹奇韡, 胡玉杰, 等. 本体研究综述[J]. 计算机研究与发展, 2004, 41(7): 1041-1052.
LI San-ping, YIN Qi-wei, HU Yu-jie, et al. Overview of researches on ontology[J]. Journal of Computer Research and Development, 2004, 41(7): 1041-1052.

[6] NECHES R, FIKES RE, FININ T, et al. Enabling technology for knowledge sharing[J]. AI Magazine, 1991, 12(3): 36-56.

[7] GRUBER T R. A translation approach to portable ontology specification[J]. Knowledge Acquisition, 1993, 5(2): 199-220.

[8] GUARINO N. Formal ontology and information systems[C]//In: Proc of the 1st Int'l Conf on Formal Ontology in Information Systems. Trento, Italy: IOS Press, 1998: 93-155.

[9] USCHOLD M. Ontologies principles, methods and applications[J]. Knowledge Engineering Review, 1996, 11(2): 93-155.

[10] GRUNINGER M, FOX M S. Methodology for the design and evaluation of ontologies[C]//In: Proc of IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing. Montreal: AAAI Press, 1995.

[11] FERNANDEZ M, GOMEZ P A, JURISTO N. METHONTOLOGY: From ontological art towards ontological engineering[C]//In: Proc of AAAI-97 Spring Symposium on Ontological Engineering. Stanford: AAAI Press, 1997.

编　辑　熊思亮