基于 WordNet 和自然语言处理技术的半自动领域本体构建*)

徐力斌　刘宗田　周　文　宋二伟

(上海大学计算机科学与工程学院　上海 200072)

摘　要　现有的大多数本体都是通过手工构建的,然而,本体的构建是一项非常费时费力的过程,近年来有关如何半自动地构建领域本体的研究越来越多。本文提出了一种基于 WordNet 和自然语言处理技术的领域本体半自动构建方法,该方法能够大大提高本体的构建效率,并且一定程度上能够保证结果本体的质量。实验表明,本文的方法在一定程度上令本体的生成过程实现自动化。

关键词　领域本体,语义单元,WordNet,自然语言处理

A Semi-automatic Domain Ontology Construction Method Based on WordNet and Natural Language Processing Technologies

XU Li-Bin　LIU Zong-Tian　ZHOU Wen　SONG Er-Wei

(School of Computer Science and Engineering, Shanghai University, Shanghai 200072)

Abstract　Ontology is more and more popular in all kinds of knowledge management system, but the traditional ontology constructing is of high costs and time-consuming. This paper introduces a semi-automatic ontology constructing method based on the general ontology WordNet and natural language processing technologies.

Keywords　Domain ontology, WordNet, Natural language processing

1　引言

本体作为能在语义和知识层次上描述信息系统的概念模型,自被提出来以后就得到了广泛关注,它能够被广泛应用在计算机科学领域的通信、互操作、系统工程等方面。近些年来,基于本体的信息检索也成为本体应用的热点。本体应用价值凸现的同时,一个现实的问题摆在我们面前:如何快速有效地构建本体?特别在各种特定的领域,如何快速构建一个该领域特定的本体,成为研究的热点。

本文对已经被采用的一些本体构建方法进行介绍,在此基础上,提出一种新的基于通用本体 WordNet 和自然语言处理技术的半自动本体构建方法,本文的结构安排如下:第 1 部分给出本体的定义以及相关知识;第 2 部分对当前已经被采用的一些本体构建方法进行介绍;第 3 部分介绍本文所采用的构建方法;之后,介绍本文所采用的构建方法,最后给出本文的实验结果。

2　本体

2.1　本体定义

本体最早是一个哲学上的概念,指的是对客观存在的一个系统解释或说明,因此,它描述的是客观现实的抽象本质。在计算机科学领域中最早引入本体概念的是人工智能界,1993 年,Gruber 给出了本体最为流行的定义,即"本体是概念模型的明确的规范说明",后来 Studer 经过对本体定义的进一步研究,认为本体是"共享概念模型的明确的形式化规范说明"。总而言之,本体包括四层含义:概念模型,共享,明确,形式化[1]。总的来说,本体可以被认为是用来描述概念以及概念之间关系的模型。

2.2　本体构建的一般原则与构建方法

本体的应用价值已经被广泛认可,但在我们能够达到应用本体的目的之前,我们首先要考虑的是如何得到一个本体,具体地讲,如何从各种领域信息中有效地得到各种概念以及概念与概念之间的关系,最后用本体的建模元语正确表达出来。本体的应用领域和应用方法使得各种目标本体的构建方法有着很大的区别,目前还不存在标准的本体构建工程方法,一般认为,针对特定的领域构建本体需要该领域的专家参与,专家参与的程度影响着所构建本体的有效性。在这种情况下,也有一些比较有价值的本体构建原则,其中比较有影响的是 Gruber 在 1995 年提出的 5 条规则[1]:

(1)明确性和客观性:知识本体应该用自然语言对术语给出明确、客观的语义定义。

(2)完整性:所给出的定义应该是完整的,能表达特定术语的含义。

(3)一致性:知识推理产生的结论与术语本身的含义不会产生矛盾。

(4)最少约束:对建模对象应该尽可能少列出限定约束条件。

(5)最大单向可扩展性:向知识本体中添加通用或专用术

*)基金项目:本文受国家自然科学基金(60575035)资助。徐力斌　硕士研究生,从事基于语义的知识管理系统的研究;刘宗田　博士生导师,教授,主要研究领域为人工智能和软件工程等;周　文　博士研究生,研究方向为人工智能、形式概念分析与本体;宋二伟　硕士研究生,研究方向为信息检查、自然语言处理。

语时，通常不需要修改已有的内容。

本体构建的一般过程可以确定为7个步骤[3]。1)首先确定本体的领域和范围，本体的应用范围确定了本体的内容。2)考虑重用现有本体的可能性，重用的方式可以根据具体情况而定。3)列举领域内的重要术语，也就是确定本体中的概念。4)定义类和类的层次关系。5)定义类的属性，也就是槽。6)定义槽的侧面，也就是对槽的约束条件。7)创建实例，需要在本体中包含的实例根据实际情况而定。

从目前本体构建的现状来看，领域本体的构建很大程度上依赖于手工，半自动与自动构建领域本体的方法远远没有达到成熟的应用。一般的构建方法往往在确定领域范围后先考虑重用其他本体的可能性，这一过程本质上来讲就是在现有本体的基础上半自动或者自动地生成应用领域内的本体。不管采用何种技术或者何种方法，这一过程总的来说可以定义为本体学习。本体学习的知识源根据其结构化程度，分为三种：非结构化信息，半结构化信息，结构化信息。在本体学习的过程中，可以结合使用上述三种知识源。从现状来看，大多数领域中并不存在可以被本体学习使用的半结构化和结构化信息，而非结构化信息却大量存在，例如领域专业文献、Internet上的网页信息，甚至可以是与领域专家的交谈笔记。以非结构化信息为基础进行本体构建，从本质上都可以被理解为从文本中进行本体学习，这一方法又可细分为以下几种，以下进行详细说明。

(1)本体提炼：该方法的核心思想是先获得一个领域相关的通用本体，该通用本体可以是已经存在的本体，也可以在领域专家的帮助下建立(可能是不完备的)，在此基础上，通过领域内的各种文本语料库提炼目标本体，提炼过程往往是对现有概念和关系的增加或减少。根据目标本体的应用效果，这个过程可以反复，最后的目标本体往往是某一时期针对某一领域的有效本体。该方法在领域相关通用本体能够获得的情况下，能够快速地构建一个领域目标本体，但是由于领域相关通用本体的广泛缺乏，并且在已有的案例中，缺乏有效的方法对本体进行提炼，因此，现有应用的效果并不好。

(2)概念聚类：该方法的核心思想是通过计算概念之间的语义距离，对概念进行分类，由此得到类的层次结构，语义距离的计算公式可以由本体构建者自行决定。该方法的核心是进行正确的语义距离计算，由于在计算过程中缺乏有效的领域背景知识指导，概念聚类的过程不能进行有效控制，此外，通过概念聚类得到的本体中只包含了分类关系。

(3)形式概念分析：该方法的核心思想是这样的，首先通过NLP中的一些方法获得领域概念和属性，由此可以获得形式背景，再由形式背景获得和形式背景同构的概念格，将概念格中的概念和本体中的相关概念联系起来，从而得到本体。运用形式概念分析方法构建领域本体只能得到本体中的分类关系，此外，由于通过NLP技术分析得到的形式背景相对比较大，不容易形成概念格。

(4)关联规则：该方法的核心思想是使用关联规则发现概念之间分类关系之外的关系。关联规则主要被运用在关系型数据库中，而领域本体构建中能够得到的领域知识往往只是文本数据，因此，它只能作为本体学习的一种辅助技术。

(5)模板规则：该方法通过对领域文献进行一定的分析，分析过程结合统计方法学和专家指导，最后形成从某一类领域文献中提取本体中概念和关系的模板规则，然后应用这些模板规则进一步生成领域本体，根据生成本体的质量，可以对模板规则进行调整，直到达到满意为止。该方法中模板规则的形成需要领域专家的大量指导，并且对模板规则的调整过程中容易产生振荡效应，不能保证最终得到最好的模板规则。

(6)概念学习：该方法的核心思想是从现实的文本中抽取领域相关的新概念并更新已有的概念分类。概念学习只能作为本体学习的一部分内容。

3 通用本体WordNet与自然语言处理技术相结合的半自动本体构建方法

本文采用的本体构建方法是一种通用本体WordNet[4]与自然语言处理技术相结合的本体构建方法，该方法结合了上文所描述的本体提炼与概念学习的思想，同时，又增加了关系学习的部分，关系学习的核心思想与概念学习相同，但是，关系学习建立在概念学习的基础之上，即先从领域文献中抽取相关概念，然后抽取概念之间的关系。该方法的总体流程如图1，下面对上述本体构建流程进行分析。

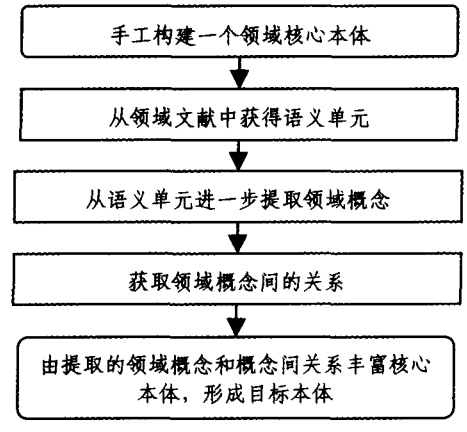

图1 本体构建流程

3.1 领域核心本体的构建

领域核心本体的构建基于手工并辅以领域专家的帮助，技术上相对比较简单，在领域核心本体中，我们只包括该领域的核心概念。一般情况，我们将领域的名称作为该核心本体的顶级概念，并且在核心本体中的概念以概念层次的方式组织，这样，核心本体中包括了概念之间的一般分类关系，得到的本体模型如图2。

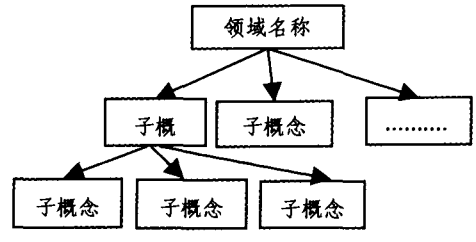

图2 领域核心本体模型

3.2 领域文献处理

本文所描述的本体构建方法采用领域文献作为知识源，将领域文献转化成自然语言处理工具所能分析的文本格式后，进行自然语言处理。这里所采用的自然语言处理工具主要包括两部分：一部分为谢菲尔德大学开发的GATE框

架[5]，该框架是一个开源的应用程序，用当前流行的 Java 语言开发，它既能作为一个图形化工具对简单文本格式的文献进行分析，也提供了能被其他应用软件所使用的 Java 开发包；另一部分为斯坦福大学开发的 StanfordParser 工具[6]，该工具提供了句法分析的功能，能够对一般的英文句子进行正确的句法分析，该工具也是一个开源的应用程序，并提供了 Java 开发包，能被其他应用程序使用。使用上述工具进行领域文献处理的过程如图 3，经过上述处理，最后得到语义单元，并将其保存在关系数据库中，这一过程同时保存领域文献的其他信息，以备后期处理使用。这里的语义单元指的是一个句子中最小的可独立成句的句子片断，如句子“Some experts believe your PC will make you blind.”中可得到语义单元“your PC will make you blind.”。

3.3 利用通用本体 WordNet 扩展领域核心本体

WordNet 是普林斯顿大学认知科学实验室开发的一个通用词典，它通过词语的语义属性来组织词典，其中基本的构建单位是同义词集合。WordNet 不仅仅用同义词集合的方式来罗列概念，同义词集合之间还以一定数量的关系相关联，典型的关系包括上下位关系、整体部分关系、继承关系等等。领域核心本体的扩展过程如下：

1)取得一个领域概念 C，从 WordNet 中取得与该概念相关的直接下层概念及通过其他直接语义关系与其关联的概念，将其放入一个集合 R。

2)验证 R 中的概念是否在已经分析得到的领域文献语义单元中出现，如果出现，将其添加到领域核心本体中，并将其与概念 C 的关系也添加到领域核心本体中。这里，如果 R 中的概念已经在领域核心本体中存在，只需将其与概念 C 的关系添加到领域核心本体中。

3)先序遍历图 2 所示的领域核心本体树，重复上述 1-2 步骤。

4)最后得到一次扩展后的领域本体。

5)查看扩展后的领域本体，根据需要，可对领域本体进行多次扩展，直到得到满意的结果。

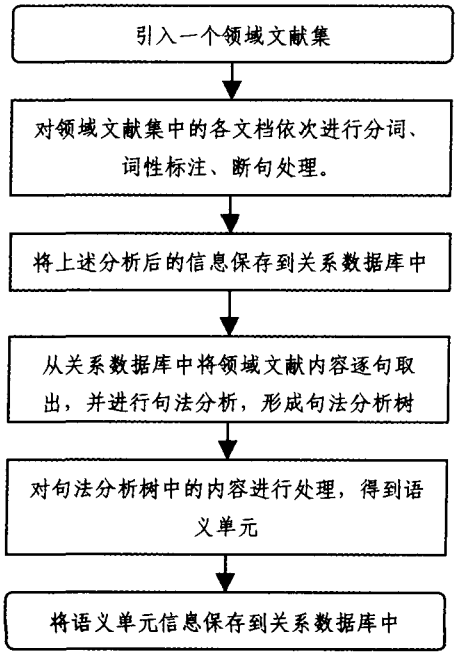

图 3 领域文献处理

3.4 向扩展后的领域本体中添加更多概念和关系

扩展后的领域本体中已经包括了由 WordNet 继承而来的概念和概念之间的关系，但是，通过使用语义单元中的语义属性，向领域本体中添加更多概念和概念之间的关系是十分有意义的，添加过程如下：

1)取得所有包含扩展后领域本体中领域概念的语义单元。

2)如果语义单元中包含两个或两个以上领域概念，查看领域概念的表达词汇在语义单元中的句法成分(通过 StanfordParser 工具进行句法分析得到)，将满足下列条件的概念或概念之间的关系加入到扩展后的本体中：

如果两个表达词汇之间存在动词，将该动词作为领域概念之间的关系加入到扩展后的领域本体中。

如果语义单元中存在其他非修饰名词且该名词与表达词汇之间存在直接动词(即在该名词与表达词汇之间不存在其他概念的表达词汇)，查看该名词在 WordNet 中的一般性指标 F_1(在 WordNet 中，每个词汇都有一个一般性指标，用来指示该词汇在所有文献中被使用的频率指数，用来表明该词汇是否被人们经常使用，使用频度越高，一般性指标也就越大)，再根据领域文献处理步骤的处理结果，得到该名词在领域文献中出现的频率，该频率的计算公式如下：F_2 = 该名词在领域文献中出现的次数 / 领域文献的数量，最后，得到该词汇的领域相关度指数 $RTD = F_2/F_1$，将领域相关度指数高于某一阈值的名词作为概念加入扩展后的领域本体，并且将上述动词也作为概念之间的关系加入扩展后的领域本体。

3)如果语义单元中包含一个领域概念，按照步骤 2 中的第二种情况进行同样处理。

上述领域本体构建的过程，自动化程度非常高，由于使用通用本体 WordNet 作为构建的一个基础，因此对领域专家的需求相对较小，但是，如果在领域核心本体的构建中能够加入专家帮助，此构建过程的效果将更好。此外，通过在构建过程中使用自然语言处理技术，能够很好地发现领域中概念之间的关系，这些关系用动词来进行表示，很好地补充了领域本体从 WordNet 中继承的关系。

4 系统实现

根据上文所述的本体构建过程，我们设计了一个本体构建系统，系统总体框架如图 4，利用该系统，我们尝试构建了一个医学领域的本体。实验中，我们将 medicine 作为医学领域的顶级概念，初始的领域核心本体中只包含该顶级概念，领域文献集来自于医学杂志的刊物摘要，该文献集包含将近 8000 个文献，每篇文献的长度都不超过 500 个词汇，RTD 阈值的大小为 0.05。经过一次扩展后，新的领域本体中概念和关系的变化如下表 1。增加了四类从 WordNet 中继承而来的关系以及 457 个由语义单元中动词产生的关系，新增加的概念分为两部分，其中由 WordNet 继承而来的概念有 639 个，从语义单元中挖掘产生的概念有 78 个，扩展后本体的一个片断如图 5。

由实验结果可以看出，经过上述系统进行本体扩展后，本体中的概念数目大大增加，其中大多数概念是从 WordNet 继承而来的，从语义单元中挖掘产生的概念相对偏少，但是，从语义单元中挖掘了很多 WordNet 中所不包括的概念之间的关系。

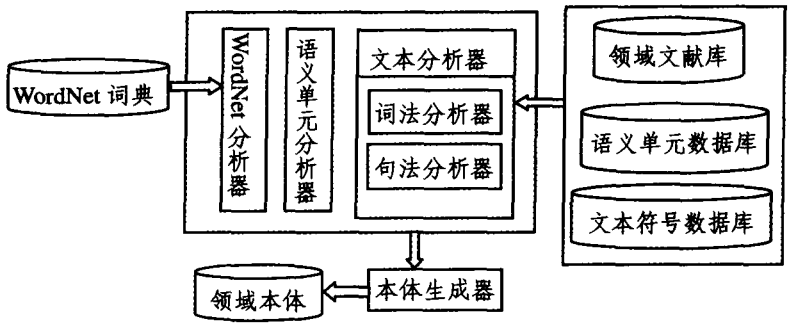

图 4　本体构建系统总体框架

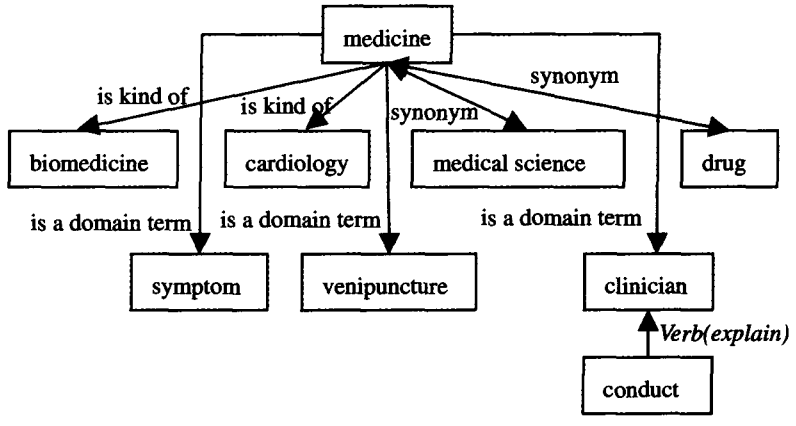

图 5　目标本体片断

表 1　目标领域本体概要

概念类型	数量	所涉及到的关系	关系数量
同义词概念	2	通过同义关系添加到本体中	2
上下位概念	117	通过概念之间的上下位关系添加到本体中	117
整体部分概念	47	通过政体部分关系添加到本体中	47
领域术语概念	373	通过隶属领域关系添加到本体中	373
文本挖掘概念	78	通过动词关系添加到本体中	457

结束语　使用通用本体 WordNet 与自然语言处理技术相结合的方法构建本体，能够在领域核心本体的原型基础上，进行本体的提炼，这是一种快速的本体构建方法。使用这种方法向领域本体中添加更多的概念是一个迭代的过程，能够根据本体构建者的要求，很好地进行过程控制。基于语义单元中动词的概念间关系的确定，对于领域本体中概念之间关系的扩展具有重要的意义。

由于语义单元反映的是一个简单句子片段的语义属性，在以后的研究中，如果能够通过统计方法，对语义单元之间的语义关联性进行挖掘，使更多的语义单元进行合理组合，形成更强的语义属性，将使本体的构建无论从质量上和效率上都会大大提高，这也是本文所描述的本体自动构建方法的重点改进方向。

参 考 文 献

1　邓志鸿，唐世渭，张铭，杨冬青，陈捷．ontology 研究总述．北京大学学报，2002，38(5)

2　Chandrasekaran B, Josephson J R, Benjamins V R. What Are Ontologies, and Why Do We Need Them? IEEE Intelligent Systems, 1999, 14(1): 20～26

3　Noy N F, McGuiness D L. Ontology Develo- pment 101: A Guide to Creating Your First Ontology. SMI technical report SMI-2001-0880 (2001), Stanford University

4　WordNet. a lexical database for the English langu- age. http://wordnet. princeton. edu/

5　GATE. A General Architecture for Text Engineer- ing. http://gate. ac. uk/

6　The Stanford Parser. http://nlp. stanford. edu/softwa re/lex-parser. shtml

(上接第 215 页)

效率如何呢？我们用 Java 构建了一个电子商务平台，在 20 个人中进行模拟应用，收集了他们的浏览喜好和购物记录形成了兴趣文件，并将数据进行分析，用上述方法进行了用户分类，当用户上网站时进行货品推荐，均达到了比较高的相关度和满意度。

结束语　正确识别客户是客户关系管理成功的基础，也是电子商务网站对客户进行有效管理的前提条件。应该对客户进行正确的分类，明确各类客户之间的关系以及各类客户管理工作之间的相互联系，以便更好地留住客户，产生更大的效益。

参 考 文 献

1　Michie D, Spiegelhalter D J, Taylor C C. Machine learning, neural and statistical classification[M]. New York: Ellis Horwood, 1994

2　曹渝昆，李云峰，汪成亮，等. 改进型模糊神经网络在顾客分类中的应用研究[M]. 计算机工程与应用，2006，19：218～221

3　Mitchell T M. 机器学习[M]. 曾华军，等译. 北京：机械工业出版社，2003