信息抽取中领域本体建模方法研究

易利涛[1],周肆清[1],丁长松[2]
(1. 中南大学 信息科学与工程学院,湖南 长沙 410083;
2. 吉首大学 数学与计算机科学学院,湖南 吉首 416000)

摘 要:近几年来,本体作为一种知识重用、知识共享和建模的重要工具,尤其是领域本体,在信息抽取系统中扮演着越来越重要的角色。但是,目前领域本体的创建还缺乏系统的、工程化的方法。首先介绍了本体的概念及本体的建模准则,然后分析了现有的几种常见的本体建模方法,并通过对比分析各种方法的优缺点,再结合信息抽取的原理以及软件工程的思想,提出了一种新的领域本体的建模方法。该方法具有很强的逻辑性和可操作性,可被一些领域本体在建立时采用。

关键词:信息抽取;本体;领域本体;建模准则;建模方法
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2011)10-0023-05

Research on Modeling Method of Domain Ontology in Information Extraction

YI Li-tao[1],ZHOU Si-qing[1],DING Chang-song[2]
(1. School of Information Science and Engineering,Central South University,Changsha 410083,China;
2. Mathematics and Computer Science College, Jishou University, Jishou 416000,China)

Abstract:As an important tool for knowledge reuse, knowledge sharing and modeling, ontology, especially domain ontology,plays an more and more important role in the information extraction system in recent years. The conception of ontology and the modeling criterion of ontology is introduced firstly. And then the ordinary existing modeling methods of ontology are analyzed. After comparatively analysing those kinds of modeling methods of ontology, a new modeling method of domain ontology is advanced according to the principle of information extraction and the concept of software engineering. It performs well in logic and operation and can be adopted in modeling certain domain ontology.

Key words:information extraction;ontology;domain ontology;modeling criterion;modeling method

0 引 言

近十几年来,本体(Ontology)被广泛地应用于计算机科学的诸多领域,如知识工程、数字图书馆、信息系统以及软件复用等,今已成为普遍研究的一个热点。随着对本体的理论和应用的深入研究,取得了丰富的研究成果,本体理论与技术也日趋成熟。基于领域本体的信息抽取系统能够提供给用户特定的感兴趣的信息,并可以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以及概念之间的联系能够有统一的认识,进一步提高系统的查准率和召回率,最终为用户提供更有价值的信息。可见,领域本体在信息抽取系统中起着十分重要的作用。但目前领域本体的构建尚缺乏系统的、针对所有领域的、工程化的方法。

文中旨在通过研究和分析已有的领域本体的各种构建方法以及其存在的问题,探索一种信息抽取系统中新的领域本体建模方法。

收稿日期:2011-03-07;**修回日期**:2011-06-14
基金项目:湖南省科技厅软件学课题(2009ZK3046)
作者简介:易利涛(1985-),男,硕士研究生,研究方向为信息检索、数据库应用技术等;周肆清,副教授,硕士生导师,研究方向为计算机应用、数据库应用技术。

1 领域本体与信息抽取

1.1 本体与领域本体

本体(Ontology)的概念源于哲学。在计算机科学中,本体是共享概念模型的明确的形式化规范说明[1]。这一概念具体包含以下四层含义:

(1)概念模型(conceptualization)。概念模型不依赖于具体的环境(计算机系统),它是纯粹反映信息需求的概念结构。

(2)明确(explicit)。概念和概念的约束都有明确的、无歧义的定义。

(3)形式化(formal)。通过对本体的形式化,使得

本体可以被计算机识别处理。

(4)共享(share)。本体所包含和体现的领域知识可以被共同认可,本体所反映的相关领域内概念集可以被公认。

本体所强调的是特定领域中有着公认语义本质的概念和概念之间的关联,同时借助概念和概念之间的关联来实现语义的表达。

领域本体(Domian ontology)是专业性的本体,它专门用于描述特定学科领域的知识。领域本体定义了有关该领域内概念的词表及概念间的关系。领域本体的组成元素包括属性、对象、关系和子领域本体[2]。

1.2 信息抽取中的领域本体

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[3]。

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,同时也需要一定的篇章分析与推理。由于领域本体能够对特定领域中的概念及概念之间的关系给予比较精确的描述,从而为人机之间、机器与机器之间的相互理解提供了语义基础。这时,领域内的语义信息给这些抽取分析提供了依据。按抽取对象的不同,信息抽取的主要任务:命名实体识别、实体关系抽取和事件抽取。领域本体能够大量应用于这些不同层次的任务中,从而有效地提高了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构建

2.1 领域本体的建模准则

Perez 等用分类法对本体进行了组织,并归纳出本体的五个基本构成元素(即建模元语):类或概念、关系、函数、公理、实例[4]。当然,在实际的应用中,构建本体时不一定要严格按照上述五类建模元语,应该视不同的情况选择所需要的建模元语。

1995 年 Gruber[5] 提出了构造本体的 5 条准则:明确性和客观性、一致性、可扩展性、最小编码偏差以及最小本体承诺。

1998 年 Arpirez 又提出了 3 个补充规则:概念名称命名标准化、概念层次多样化以及语义距离最小化。

根据本体的建模准则,构建本体概念模型的具体方法总的来说可归为两种构建模式,一是利用现有文献和领域专家使用手工方式创建概念关联,二是将已有的叙词表改造成本体,或采用学习机制进行自动或半自动化的本体构建。

2.2 领域本体的建模方法

目前,在相关研究及实践中产生了一些面向不同应用需求的本体建模方法,其中比较知名的有 IDEF5 法、骨架法、TOVE 法、KACTUS 法、METHONTOLOGY 法、SENSUS 法及七步法等。

2.2.1 IDEF5 法

美国 KBSI 公司(Knowledge Based Systems, Inc.)的 IDEF5[6] 法是开发出来用于描述和获取企业本体的。IDEF(integration definition for function modeling)是 KBSI 开发的一系列"面向功能建模的集成定义"项目。IDEF5 方法提供了一种在理论上和实践上均有充分根据的方法,该方法专门用于帮助创建、修改和维护本体。标准化的程序,以一种直观和自然的形式代表本体信息的能力、更高质量的结果,使得 IDEF5 的应用可以减少项目的活动经费。

IDEF5 法主要采用 IDEF5 原理图语言和 IDEF5 细节说明语言。原理图提供的示意图以图表的方式非常直观地描述了本体,加上详尽的细节说明,使得描述的信息一目了然。其中,语义规则必须提供一切可能的解释原理。为了解释最基本的语言结构,必须概括这些规则,然后将它们应用于递归构造更复杂的构图。

IDEF5 本体开发过程包括以下五项活动:

(1)组织和范围(organizing and scoping);

(2)数据收集(data collection);

(3)数据分析(data analysis);

(4)初始化本体开发(initial ontology development);

(5)本体的细化和验证(ontology refinement and validation)。

IDEF5 方法提供了一种结构化的方法,领域专家可以利用此方法有效地开发出可用的、精确的领域本体,并很好地维护之。

2.2.2 骨架法

Mike Ushold & Micheal Gruninger 的 Skeletal Methodology 即骨架法[7],专门用于为企业本体的开发提供指导方针,故又称 Enterprise 企业建模法。骨架法的流程图如图 1 所示。

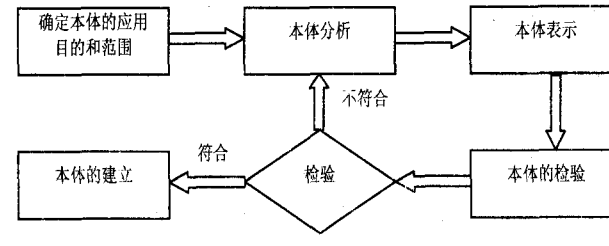

图 1 骨架法基本流程

其中,本体分析要求对本体内所有的概念术语及概念术语间的关系给出明确的定义。本体的评价标准基本上符合 Gruber 构造本体的准则,如清晰性、一致性、完整性、可扩展性等。

2.2.3　TOVE 法

TOVE 法,又称 Gruninger & Fox 评价法[8],产生于 Toronto 大学企业集成实验室(Enterprise Integration Lab.)的 TOVE 项目,主要使用一阶逻辑(first-order logic)进行集成。TOVE 项目的主要目标是为了给商业和公共企业建模而开发一套集成本体(integreted ontology),即 TOVE 本体。

TOVE 法主要包括以下步骤:

(1)定义一套激发场景(motivating scenario)。

(2)定义一系列非正式的能力问题,为了支持激发场景本体必须回答这些能力问题。

(3)使用一阶逻辑定义本体的术语。

(4)使用一阶逻辑和术语重新定义那些能力问题。

(5)使用一阶逻辑定义术语的语义以及术语的约束。

2.2.4　KACTUS 法

KACTUS 工程法是在欧洲 ESPRIT 工程的 KACTUS(modelling Knowledge About Complex Technical systems for multiple USe)项目中产生的。该项目旨在开发一种本体的构建方法,以使得 technical systems 生命周期(life-cycle)过程中可以知识重用。这意味着使用同一知识库就可以进行设计、诊断、操作、维护、重新设计和指导等操作。该方法包括的基本步骤如下:

(1)基于应用的说明。

(2)初步设计相关的本体范畴。

(3)构造本体。

2.2.5　METHONTOLOGY 法

METHONTOLOGY 法[9]是由西班牙马德里大学 AI 实验室提出来的一种非常接近软件工程的本体开发方法。该方法将整个本体开发过程划分为三个阶段:管理阶段、开发阶段和维护阶段。其本体开发的具体流程如下:

(1)规格说明(specification)。此部分的结果是一份以自然语言形式的本体规格说明文档。该文档中必须明确本体开发的目的与用途、本体的描述范围(如描述的词汇、描述的特点及描述的粒度等)、本体的形式化程度等。

(2)知识获取(knowledge acquisition)。知识来源的途径繁多,如专家、书籍、网络等,另外还有其他的可共享或可重用的本体。因此,知识获取的方法也多种多样。

(3)概念化(conceptualization)。用领域术语描述领域知识,再把领域术语识别为概念(类)、关系、属性、实例,之后用便于应用的非形式化方式表示它们。

(4)集成(integration)。通过集成已有的本体,以实现本体的共享。当重用其他本体中已建好的定义时,可以查看元本体,选择自己概念模型中的语义和实现相一致的术语定义。

(5)实现。用某种形式语言编码实现本体。

(6)评价(Verification)。本体的评价方法可以参照知识系统(KBSs)中的知识验证和评价技术,一些关于 METHONTOLOGY 的文献中也给出了检查不完备性、不一致性和冗余的指导性方法。

(7)文档化:在本体建设的每个阶段都应该有对应的文档。

2.2.6　SENSUS 法

SENSUS 法是美国 USC 信息科学研究基于自然语言处理的 SENSUS 语言本体而开发出来的。目前大概有7万多个电子科学领域的概念包含在 SENSUS 语言本体中,要从中构造出特定的领域本体,还必须对这些术语或概念进行筛选操作(剪除)。SENSUS 法构造本体的基本流程如下:

(1)定义"叶子"术语。

(2)通过手工方式连接叶子术语和 SENSUS 术语。

(3)从叶子节点出发,找出一条到达 SENSUS 根节点的"路径"。

(4)当遇到 SENSUS 本体中不存在的术语,而与其领域相关的术语时,将此术语增加到其中。

(5)按照启发式思维找出所有特定域的术语。如果节点处在两条或两条以上的路径上,则此节点一定是一棵子树的父节点,这说明这棵子树上的所有节点均与该特定域相关,故它们均是要增加的术语。

2.2.7　七步法

七步法[10]是由斯坦福大学医学院开发的。该方法主要应用于领域本体的构建。七步法的7个基本步骤如下:

(1)确定本体的专业领域和范畴。确定构建本体的目的,本体所覆盖的领域等。

(2)考查复用现有本体的可能性。复用本体不仅使得本体可以共享,还可以实现系统与其它应用平台之间进行交互。

(3)列出本体中的重要术语。制订一份本体中全面的术语清单。

(4)定义类和类的等级体系。类的等级体系的完善主要有自顶向下法、自底向上法以及综合法。

(5)定义类的属性。类的体系所能提供的信息并不足以回答系统能力问题。在类定义好之后,还必须描绘术语或概念间的内在结构。

(6)定义属性的"分面"。一个属性可能由多个"分面"所组成。所谓属性的"分面",指属性取值的类

型、取值的范围、取值的个数等。

(7)创建实例。

有学者利用IEEE 1074-1995标准对以上7种方法进行比较(如表1所示),得出结论:与IEEE标准相比较,没有一种方法体系是完全成熟的。7种方法体系的成熟度依次为:七步法、METHONTOLOGY法、IDEF5法、TOVE法、骨架法、SENSUS法、KACTUS法[11]。

表1 本体建模方法比较结果表

项目 / 名称	生命周期	相关技术	本体应用	方法细节
KACTUS法	没有	不确定	仅一个域	很少
SENSUS法	没有	不确定	多个域	一般
骨架法	没有	不确定	仅一个域	很少
TOVE法	非真正的	不确定	仅一个域	较少
IDEF5法	没有	不确定	多个域	详细
METHONTOLOGY法	有	有,不全	多个域	详细
七步法	非真正的	有	多个域	详细

目前大多数本体构建方法是针对具体的项目开发的,它们都有各自的构建原则和设计标准,因此难以实现本体的共享、重用和互操作。而且本体的构建大多是手工构建,还缺乏系统的、工程化的方法,应用领域有限,方法细节比较粗,相关技术比较少,自动化程度不高,因此存在着一定的局限性。

通过借鉴以上几种本体构建方法,汪方胜等提出了一套构建领域本体的知识工程方法[12]。该方法较符合人类的认知思维,具有较强的逻辑性、可操作性及拓展性。知识工程方法的流程如图2所示。

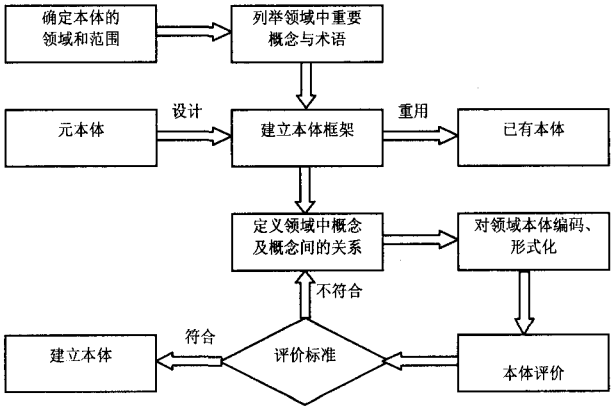

图2 构建领域本体的知识工程方法流程

2.3 一种新的领域本体的建模方法

领域本体的构建是一个增量迭代式开发的过程,是一个不断添加新概念并精准已有概念的过程。因此,在领域本体开发的初期及开发的各阶段之间,如何保持各个概念的定义及描述的完整一致性是至关重要的。

借鉴软件工程中的思想及以上领域本体的建模方法,再结合一种原型法(基于螺旋模型的领域本体构建方法),提出了一种可维护易于扩展的领域本体的建模方法。该领域本体的构建过程中包括以下几个阶段:领域本体的需求分析、领域本体的规格说明、领域本体的框架构建、领域本体的元本体设计、领域本体的表示与评估、领域本体的维护。

(1)领域本体的需求分析。

在此需求分析阶段,首先应该明确构建领域本体的目的与用途,确立其使用对象。然后,确定该领域本体的领域和范围,制订"需求说明书"文档。之后,通过统筹资源,由领域本体建设人员制订出构建领域本体的"项目计划书"文档。

(2)领域本体的规格说明。

该阶段要尽可能地通过专家、网络、书籍等各种媒体渠道搜集领域本体相关信息,充分了解领域知识。在做好资料准备工作后,尽可能罗列出系统所要描述的术语和概念(包括一些向用户解释的概念),并确立这些概念之间的关系。最后产生一份以自然语言编写的领域本体的"规格说明书"文档。

(3)领域本体的框架构建。

这个阶段的主要任务是按计划逐步搭建领域本体的框架。由于具体领域知识的复杂性和领域边界的模糊性,加上领域专家参与程度的不同,领域知识中描述的概念及关系也不精确不全面。因此,领域本体的框架构建时采用自顶向下的方法,即从领域中最大的最顶层的概念(类)开始,逐步添加子类将其细化,这样将得到领域本体中类的层次体系,经过进一步地精化细化,最终成为领域知识的框架体系。

(4)领域本体的元本体设计。

元本体即本体的本体,它是领域中概念的最高层次的抽象。元本体设计的原则是领域无关性,尽可能减少包含的元概念数目。设计元本体的基本步骤如下:

a. 定义类及类的层次关系;

b. 定义类的属性;

c. 定义属性值;

d. 创建实例。

另外,在设计元本体时,应该考虑重用已经存在的本体。因为本体的主要作用就是解决知识的共享和重用问题。

(5)领域本体的表示与评估。

领域本体的表示就是通过使用本体描述语言对领域本体进行编码和形式化。目前,本体描述语言主要有可扩展标识语言(extensible markup language,XML)、资源描述框架(resource decription framework,RDF)、网络本体语言(web ontology language,OWL)等。

形式化的领域本体需要检验和评估,但是目前还

没有本体评估的标准方法。通常,在实际应用系统中,考查构建的领域本体是否满足了需求、达到了目的,是否符合本体的构建准则,是否具有可扩展性、可重用性等。

(6)领域本体的维护。

随着具体领域的不断发展,本体的内容也会发生变化,需要通过知识的进一步获取、概念的进一步扩充,不断改进和扩展领域本体。另外,对领域本体的相关维护记录形成文档化。

3　结束语

本体的开发是一个复杂的工程,在本体的实际开发与维护过程中应该借鉴软件工程的思想。本体的构建方法指导本体构建的整个过程,是实现本体构建的理论基础。目前,领域本体构建没有一个统一、完善的标准。在现有本体构建方法的基础上,提出了一种强调需求分析,有效减少领域专家参与度的领域本体构建方法。这种构建方法具有很强的逻辑性及可扩展性,有利于本体的共享及重用。

参考文献:

[1] Studer R, Benjamins V R, Fensel D. Knowledge engineering, principles and methods[J]. Data and Knowledge Engineering, 1998,25(1-2):161-197.

[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.

[3] 苗夺谦,卫志华. 中文文本信息处理的原理与应用[M]. 北京:清华大学出版社,2007:279-303.

[4] 刘燕玲,华庆一,郭晓娟. 基于领域本体面向问题的需求分析与领域建模[J]. 计算机技术与发展,2007,17(8):99-100.

[5] 刘　琼,李宝敏. 一种果品领域本体库的构建方法[J]. 计算机技术与发展,2009,19(1):197-199.

[6] KBSI. IDEF5 Ontology Description Capture Method[EB/OL]. 2001. http://www.idef.com/IDEF5.htm.

[7] Siau K. Informational and computational equivalence in comparing information modeling methods[J]. Journal of Database Management,2004(15):73-86.

[8] Ushold M, Gruninger M. Ontologies Principles, Methods and Applications[J]. Knowledge Engineering Review, 1996, 11(2):36-40.

[9] 杨秋芬,陈跃新. Ontology 方法学综述[J]. 计算机应用与研究,2002,19(4):5-7.

[10] 杜文华. 本体构建方法比较研究[J]. 情报杂志,2005,24(10):24-25.

[11] 李　景,苏晓鹭,钱　平. 构建领域本体的方法[J]. 计算机与农业,2003(7):7-10.

[12] 汪方胜,侯立文,蒋　馥. 领域本体建立的方法研究[J]. 情报科学,2005(2):241-244.

(上接第 22 页)

5　结束语

提出了基于会话层的行为特征垃圾邮件识别方法,运用了支持向量机算法,该算法具有较好的分类能力,并且只是针对邮件头部信息的分类,具有很强的通用性。行为识别技术在处理速度上有相当大的优势,对邮件头部信息进行处理,可以在其接收整个邮件内容之前进行过滤,这样可以节省由于垃圾邮件的泛滥而浪费的网络带宽。该技术与基于内容识别技术相结合同时提高准确率和召回率会是接下来的研究方向。目前现有的反垃圾邮件技术各有优缺点,单一的依赖某一种技术不可以完美地解决所有垃圾邮件问题,多种技术融合的过滤技术将会是反垃圾邮件领域未来发展趋势。

参考文献:

[1] 陈　勇,李卓桓. 反垃圾邮件完全手册[M]. 北京:清华大学出版社, 2006:8-9.

[2] 赵治国,谭敏生,丁　琳. 垃圾邮件行为识别技术的研究与实现[J]. 计算机应用研究,2007,24(11):128-131.

[3] Terri O, Tony W. Developing and Immunity to Spam[C]//In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2003). Chicago:[s. n.],2003.

[4] 白秋颖,章　睿,张耀龙. 基于网络会话层的垃圾邮件行为识别[J]. 计算机工程与应用,2007,43(1):167-169.

[5] 石　义,钱步仁. 基于内容与行为特征的反垃圾邮件系统[J]. 网络安全技术与应用,2009(4):121-124.

[6] Vapnik V N. 统计学习理论的本质[M]. 张学工,译. 北京: 清华大学出版社,2000.

[7] 段凤娟,朱吉胜,王华建. 支持向量机快速算法的实现技术[J]. 现代计算机,2008(9):73-76.

[8] 周彩兰,虞　珊,张亚芳. 基于 SMTP 协议解析的垃圾邮件防治技术[J]. 计算机技术与发展,2008,18(1):188-191.

[9] Cristianini N, Shawe-Taylor. An introduction to Support Vector Machines and other kernel-based learning methods[M]. Cambridge:Cambridge University Press, 2000.

[10] CCERT Data Sets of Chinese Emails (CDSCE) [EB/OL]. [2011-04-11]. http://www.ccert.edu.cn/spam/sa/datasets.htm.

[11] ICTCLAS[EB/OL]. [2011-04-11]. http://ictclas.org/.

[12] 潘洁珠,周　晓,吴共庆,等. 基于小样本学习的垃圾邮件过滤方法[J]. 计算机工程,2010,36(21):245-247.