一种本体构造方法及其在Web页面建模中的应用

杨德仁，顾君忠
（华东师范大学 计算机应用研究所，上海 200062）

摘 要：为了便于机器理解和有效挖掘Web内容，为Web页面结构及查询接口进行了建模．采用基于元本体的分层体系结构，分离了本体与语境，分层了概念和实例，是一个逐层实例化的过程．与数据库模式一样，Web页面结构及查询接口也由一些基本组件构成，这些组件及其之间的关系能灵活设置．模型的层次分明，具有良好的共享性和扩展性．利用基于元本体的分层体系结构，不但能为页面结构和查询接口建模，而且利用这种本体与语境相分离的机制能为查询结果页面之间的语义关系建模，提高搜索引擎的检索精度．

关键词：本体方法学；HTML页面；模型
中图分类号：TP 393 **文献标识码**：A **文章编号**：0254－0037(2006)09－0853－06

本体作为语义Web[1]的关键技术日益流行，旨在实现知识在软件代理之间共享和复用．本体还是一种有效的知识表示方法，已广泛应用到电子商务、知识管理(如知识集成)等领域．多个应用程序互操作的基础是数据要具有语义，以便进行数据映射．形式化的、规范的本体可为数据提供语义．Gruber[2]把本体定义为概念模型的明确的规范说明．Borst[3]则为共享概念模型定义明确的、形式化的规范说明．实际上，本体是对有关实体的概念级的内在本质及其规律的反映．其中，概念化指领域中某些现象的一种抽象模型，它以识别现象的概念为手段；明确化指概念类型及其使用限制被明确定义；形式化指本体应该是机器可读的；共享指本体中体现的是共同认可的知识，反映了相关领域中公认的概念集．传统本体设计方法的主要缺陷有：①没有分离本体及其模式，即把本体元素及其元数据放在一个逻辑文件(即本体)中；②设计过于笼统，过分强调了本体的领域性，粒度大；③本体中概念系统的层次关系不清．这些都不利于本体的设计、实现、共享和复用．作者描述的本体建模方法的主要特点是：把语境从本体中分离出去，形成纯本体，实现本体的复用，则概念上下位关系是纯粹的属种关系；把实例层从概念层面中分离出去，形成语义Web数据层，便于知识表示．

1 一种基于元本体的本体建模方法

1.1 一种修正的概念系统分类方法

国家标准GB/T15237.1—2000[4]的相关定义：概念是通过对特征的独特组合而形成的知识单元．概念域是同某一主题相关的非结构化的概念的集合，是建立概念体系的基础．概念体系是根据概念间相互关系建立的结构化的概念的集合．概念的层级关系是属种关系或整体-部分关系中的上位概念和下位概念间的关系．概念间的横向关联关系(即语用关系)是主题相关的概念间的关系，包括序列关系(在空间或时间上邻近)和因果关系．

这种标准存在的问题是：整体-部分关系并没有按概念的内涵特征分层，属于不同范畴的概念之间的关系；把语用关系定义为序列关系和因果关系也不全面．因此，有必要对该标准进行修正，把概念之间的关系分为相同范畴内的层次关系和属于不同范畴的概念之间的横向实用关系．层次关系严格按照概念的

收稿日期：2005-05-08．
基金项目：上海市科学技术发展基金资助项目(055115001)．
作者简介：杨德仁(1964－)，男，宁夏同心人，副教授，博士生．

特征划分,是一种纯属种层次,属性(内涵)定义良好. 概念具有特定的属性(内涵)和实例(外延). 而横向关系是基于实用的,即根据应用而异,包括整体-部分关系. 把整体-部分关系归属为横向关联关系,有利于提高设计的灵活性,表示的知识也有良好的扩展性.

1.2 元本体与元元本体

元本体数据是构建本体模式的组件库,其具体数据用于表示本体模式. 一个具体本体用到的模式数据都包含在元本体的数据中,元本体是为表达本体而设计的. 其特点是:"元"性体现在为本体模式提供组件能力上,它全面描述了本体模式中可能用到的所有概念及其关系. 同样,元本体的数据用元元本体中的术语描述. 元元本体(即术语层)是最抽象的层次.

定义 0(元元本体) 术语 T:=Terminology.

术语(Terminology)是描述元本体数据的概念,术语的取值将被实例化为元本体的数据. 元元本体的基础是集合理论[5]. 它描述了元本体的数据:元概念(Meta_Concept)及其之间的元关系(Meta_Relation).

定义 1(元本体) 元本体 MO:=MC|RC.

其中,MC:=OC|CC,OC 指本体概念,可取多值,分别对应概念纵向各层,CC 指语境概念,取值与 OC 相似;RC 指关系概念,取值为纵向概念层次关系、属性关系、实例关系、本体和语境的概念之间的关系及其实例之间的关系.

元本体模式和数据为:

元概念(Meta_Concept)

本体概念(Ontology_Concept)

语境概念(Context_Concept1)

……

元关系(Meta_Relation)

本体概念与语境概念之间的关系(Relation_OC1_CC1)

其中,OC1 代表 Ontology_Concept1,依次类推. 具体的本体和语境要用到或至少部分用到这些数据.

1.3 本体

本体中概念的纵向层次数目与具体概念的特征属性的划分和归类有关,多层模型有利于语义伸缩:向上则宽泛,向下则具体. 其模式是元本体中的全部或部分概念和关系.

定义 2(概念层本体) 本体 O:={OC;OP;$OH^{OC;OP}$;oprop}.

其中,OC 为类,OP 为性质,类的层次 OH^{OC}:$OH^{OC}\subseteq OC\times OC$($OH^{OC}$(OC1,OC2)指 OC1 是 OC2 的子类),关系函数 oprop:OP→OC×OC(域 dom:OP→OC,且 dom(OP):=Π_1(oprop(OP)),值 rang:OP→OC 且 range(OP):=Π_2(oprop(OP));oprop(OP)=(OC1,OC2),可以写成 P(OC1,OC2);性质层次 OH^{OP} 与 OH^{OC} 的定义相似.

以微机在高校中的应用为例构造本体. 图 1 所示是微机的概念层次. 实际中还应该考虑微机的等当概念如 PC、notebook PC 等.

计算机(Computer)

微机(Micro_Computer)

486

586

1.4 语境及其与本体的横向关系

语境(context)在许多学科中都起着关键性作用. 语境概念有 3 种表示方法[6]:Kamp 的表示理论、Barwise 和 Perry 的情形语义和 Sowa 的概念图. 概念图把本体当作 1 个图,节点代表概念,弧线代表 1 个关系或属性,并把概念的语境定义为在本体图中概念的邻居,便于计算概念的相似性. 本文讨论的语境与

此相似.

传统的本体涉及多个概念系统,粒度大,很难重用. 本文的方法把传统的本体分为基于单纯概念的层次关系的本体和语境,语境是对本体关联领域中的概念的描述. 语境与本体一样,在层次上属于概念级别,在逻辑上属于本体的关联域.

本体与语境的关系是一种域值关系和/或限制关系,这种语用关系随实际应用不同而不同,即本体和语境之间的关系是动态的、可伸缩的和可变的. 只用更改语境,本体便可跨域复用. 与把语境归类在本体核心概念中的传统的垂直化本体结构相比,这种扁平化结构更有利于本体的映射、重用、共享等. 当然,1 个本体也可以同时与多个语境关联(如整体-部分关系),可用于改进搜索引擎技术,实现基于语境的概念查询,为用户提供精确信息.

定义 3(本体层语境)　语境 $C:=\{CC;CP;CH^{CC;CP};cprop\}$. 与定义 2 相似.

例如,在为微机在工程中的应用建模时,只用为语境"工程"建模,实现了微机本体的完全重用(见图 1).

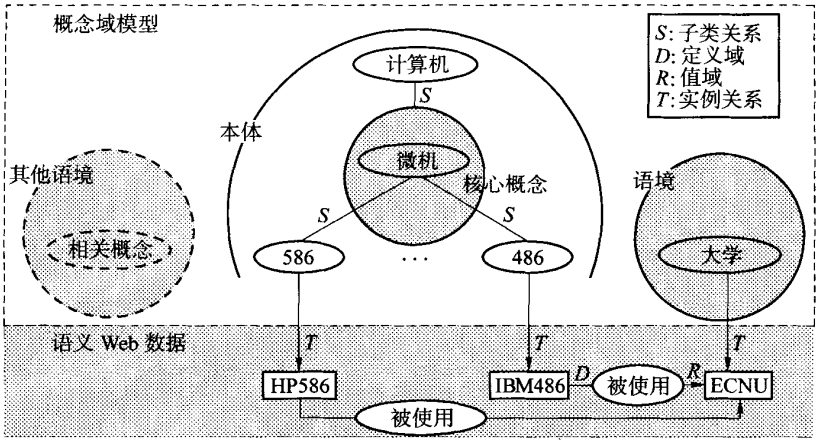

图 1　微机本体和高校语境的关系示意图

Fig. 1　An illustration for the relation of ontology and context

1.5　语义 Web 数据

本体和语境都属于概念模型,应用要涉及到与之相关的实例数据及其之间的关联. 考虑到大多应用是面向 Web 的,故称语义 Web 数据.

定义 4(语义 Web 数据层)　语义 Web 数据的结构为 $SWMD:=\{O;C;IO;IE;inst;instr\}$.

其中,O 为本体,C 为语境,IO 为本体实例,IC 为语境实例;本体类实例函数 $inst:OC\rightarrow 2^{IO}$($inst(OC)=IO$ 即 $OC(IO)$),性质实例函数 $instr:OP\rightarrow 2^{IO\times IO}$与 inst 相似;语境类实例函数 $inst:CC\rightarrow 2^{IE}$($inst(CC)=IE$ 即 $CC(IE)$),性质实例函数 $instr:CP\rightarrow 2^{IE\times IE}$与 inst 相似.

在实例层面的语义 Web 数据与具体应用密切相关,除概念对应的对象外,还包含一些重要关系,如实例之间的语用关系. 如要查询"HP 586 微机在华东师范大学的应用情况",把本体中"586"实例化为"HP586",语境中的大学实例化为"华东师范大学"(图 1 中 ECNU),2 者之间的关系是"Used by",其域为"HP586",值为"华东师范大学".

实例之间的关系因应用(语境)不同而不同,这种关系一般有逆关系,如 Used by 的逆关系是 Use. 还可以对关系实施进一步的限制,如基数限制等.

1.6　元本体体系结构和建模步骤

如图 2 所示,基于这种分层模型的本体建模工具可以在 4 个层面之间实现实例化映射. 此模型的通用性体现在本体的重用性和语境的可更换性上,可利用 OWL[7]语言实现. 这种本体建模的关键是把握建

模的层次性,实行分而治之. 在元本体层面要详尽列出在本体层面要用到的实体及其之间的关系. 而本体和语境层面是具体应用的模式,语义 Web 数据层是这种具体模式的数据.

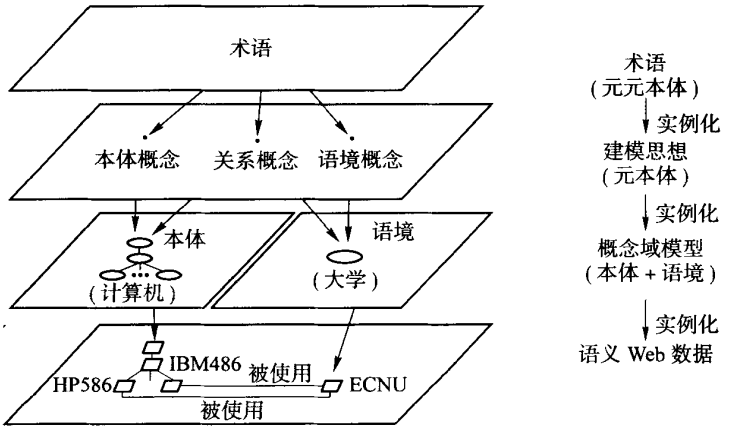

图 2 本体建模流程

Fig. 2 Ontology modeling flow

元元本体层是通用的,与领域无关. 作为该方法的通用步骤,术语的取值为"元概念"和"元关系",在模型设计中不用列出. 具体建模步骤如下.

1) 在元本体层面,要详尽列出领域的所有概念及其之间的所有关系. 这个层面与应用域有关,而与具体应用无关. 这个过程需要领域专家的参与.

2) 在本体和语境层面,根据具体应用中的被建模的概念,利用元本体层的概念及其关系来建模. 这个层面与领域中具体的应用有关.

3) 在语义 Web 数据层,为对应于模式中的概念和关系赋值. 其结果便是对应模型的知识表示. 这是实例层面,对应于本体和语境层面的本体模型.

根据上述步骤,用斯坦福大学的本体编辑器 Protégé 编辑本体和语境.

2 Web 页面的知识表示

虽然 Web 上的半结构化数据没有确定的和明显的模式,但研究发现,静态和动态页面的内容及其结构、查询接口等都具有语义性. 用上述方法为其建模,有利于复用和机器理解.

2.1 Web 查询接口的语义建模

互联网信息的搜索和集成都要用 Web 查询接口(即表单[8]),从 Web 数据库提取结构化数据. 表单是面向人设计的,对机器处理的最大挑战是如何识别表单、查询元素及其语义、正确准备数据提交查询,因此有必要为表单建立知识模型.

表单是用户与系统交互的窗口,具有明显的语义,体现在:①它有背景知识,目的是查询后台数据库中的数据;②它有明确的应用领域,这与用户的域本体相关;③查询接口中的查询元素的标签有语义,也有自己的模式(数据库模式的部分视图);且查询元素越多,越有利于数据库视图的建立;④查询接口中的选择性输入框的可选值有语义,是数据库中的数据或数据分类;⑤查询结果有语义,体现在它是数据库的一种视图,与查询接口中的元素有一定的联系.

作者用基于元本体的本体方法为表单建立语义模型[9]. 表单元本体模式为概念类及其关系. 概念类分 3 级:一级类是表单;二级类有表单说明(即上下文)、标签、输入文本框、输入选择框和按钮等;三级类有动作、方法、类型、名称、值、ID 等. 概念类之间的关系为:一级类拥有二级类;一级、二级类拥有或部分拥有三级类;标签则说明了输入框和选择框等.

在本体层面，利用上述组件对具体表单建模．如图 3 所示，Google 的本地服务查询表单[10]含有 2 个文本输入框和对应的说明标签、1 个可选框和 1 个提交按纽．建模后，准备查询数据，提交查询并处理结果页面[9]，并用 OWL 语言实现这种模型的表示[9]．

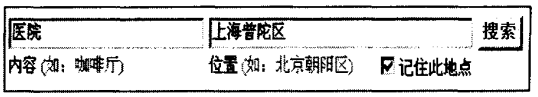

图 3　Google 本地搜索接口

Fig. 3　Google local search interface

2.2　Web 页面结构的知识建模

Web 页面是半结构化 HTML 文件，作为一个知识载体，其语义并不是单一的．利用页面的元数据和逻辑结构先分类网页信息，然后提取，并用本体方法把页面的主题性内容及其逻辑结构表示出来，以便深入处理．页面的层次结构为：

```
〈html〉
    〈head〉
        〈title〉 * * * * * * * 〈title〉
        〈meta name="* * * * *" content="* * * * * *"〉
        ……
    〈/head〉
    〈body〉
      〈block_element1〉
        〈block_element12〉
          〈line_element121〉 * * * * * * 〈/line_element121〉
          〈block_element122〉 # # # # # # # 〈/block_element122〉
        〈/block_element12〉
    〈/block_element1〉
    〈block_element2〉
        * * * * * * * * *
    〈/block_element2〉
  〈/body〉
〈/html〉
```

其中，〈block_element?〉是块级元素标签，? 代表层次编号；〈line_element?〉是行内元素标签；〈Body〉元素只直接容纳块级元素．

块级元素作为其他元素的容器，是构成整个文件结构的主要部分．块级元素能产生新行，包含块级元素和行内元素．〈form〉、〈hr〉、〈div〉、〈table〉、〈p〉等 30 多个元素都属于块级元素．文本内容处于块级元素〈p〉和〈/p〉之间，标题位于〈hn〉和〈/hn〉之间．〈hn〉可为〈h1〉，〈h2〉，…，〈h6〉．块级元素分 2 种：一种如〈div〉，能容纳其他块级元素或行内元素；另一种如〈p〉，只能容纳行内元素和文字．行内元素不产生新行，只能包括行内元素和文字．

根据〈title〉和〈meta〉内容，对〈body〉进行语义分割，找出主题内容，假定〈block_element1〉是主题信息块，并且详细结构信息具有 B1(B12(B122(H1P)))型结构．其中，B1 为〈div〉；B12 为〈div〉；B12 为〈hr〉；最内层的"# # # # # # # #"是〈h1〉、〈p〉等标签及其所表示的主题性内容．这是一种嵌套的整体-部分关系．

先罗列 HTML 标签(tag)，再根据 HTML 规则建立标签之间的关系，然后针对此文档建模，标签的命名为其绝对路径．此模型可用于处理批量的相似页面，如新浪中的新闻页面或 Deep Web 中的动态页面．

2.3　在 Web 文档分类中的应用简介

除基本内容外，Web 文档(简称文档)也包含了一套指向其他相关文档的超链接，文档中的超链接提供了与其他文档的关系的信息．通过分析文档中的超链接，与宿主文档的标题、关键词等元信息进行比

较,就能识别文档之间的关系(层次关系或其他关系). 用本文的知识建模步骤为文档关系建模,以便搜索引擎能分类显示检索结果.

3 结束语

用元本体构建本体模型是本体工程的研究热点[5]. 与文献[5]中元本体体系结构相比较,作者提出的基于元本体的本体分层模型建模方法的特点是:区分了核心本体与语境,分层了概念和实例,是一个逐层实例化的过程,因此模型具有良好的可扩展性. 该方法适于为那些组件固定的领域建模,如针对Web页面包括表单、表格、列表等元素的建模,有利于Web知识提取的智能化. 用OWL语言实现的这种模型是一种通用的表示形式,可广泛应用于信息共享和系统集成等领域.

参考文献:

[1] BERNERS-LEE T, HENDLER J, LASSILA O. The semantic web[J]. Scientific Am, 2001, 284(5): 28-37.
[2] GRUBER T. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.
[3] BORST W N. Construction of engineering ontologies[D]. Enschede: University of Twente, 1997.
[4] 国家质量技术监督局. GB/T15237.1—2000 术语工作(词汇) [S]. 北京: 中国标准出版社, 2000.
[5] HEINRICH H, FRANK L. A meta-ontological architecture for foundational ontologies[C] // MEERSMAN R, TARI Z. On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE. Berlin: Heidelberg Springer, 2005: 1398-1415.
[6] VAROL A, MEHMET S. Steps toward formalizing context[J]. AI Magazine, 1996, 17(3): 55-72.
[7] World Wide Web Consortium. OWL web ontology language overview[EB/OL]. [2005-10-11]. http: //www.w3.org/TR/owl-features.
[8] World Wide Web Consortium. Forms[EB/OL]. [2005-10-11]. http: //www.w3.org/TR/html4/interact/forms.html.
[9] 杨德仁, 顾君忠. 一种Web查询接口的语义模型研究[J]. 华东师范大学学报(自然科学版), 2006(5): 93-98.
YANG De-ren, GU Jun-zhong. A semantic model for web search interface[J]. Journal of East China Normal University (Natural Science), 2006(5): 93-98. (in Chinese)
[10] Google. Google local search interface[EB/OL]. [2005-10-11]. http: // bendi.google.com/clochp.

A Ontology Methodology and Its Application to Web Page Modeling

YANG De-ren, GU Jun-zhong
(Institute of Computer Application, East China Normal University, Shanghai 200062, China)

Abstract: In order to make machine understand and mine web content effectively, it is necessary to model web page structure and querying interface included. Using a layered architecture based on meta-ontology, the modeling method separates contexts from ontology, and instances from concept. It is a process of instantiating hierarchically. Like database schema, web page structure and querying interface included have some basic components. These components and the relations among them can be established flexibly. The resulting models are well defined, have clear layers, and can be shared and extended. Using the architecture based on meta-ontology, modeling web page structure and interface is feasible; its separation mechanism is also useful to class resulting pages of search engine, and hence improves search precision.

Key words: ontology methodology; HTML pages; models