

基于森林转化为树的本体模块化方法^①

王忠华

(济宁医学院 实践教学管理处, 济宁 272000)

摘 要: 针对目前本体的构建效率低、重用能力弱的问题, 研究了一种新的本体模块化方法——基于森林转化为树的方法. 将待建模领域逐层分解为若干个不可再分的子领域, 对各个子领域分别进行构建本体模块; 最后根据概念间的继承关系和《数据结构》中森林转化为树的原理, 把若干个本体模块组装成一个模块化本体. 在大学领域本体的构建上, 对该方法进行了验证, 结果证明, 该方法可有效的提高本体的重用能力, 可为构建大型领域本体提供一条有效途径.

关键词: 领域本体; 模块化本体; 本体模块; 模块组装

Modular Method of Ontology Based on Forest-Tree Transformation

WANG Zhong-Hua

(Teaching Practice Department of Jining Medical University, Jining 272000, China)

Abstract: One problem the present ontology faces is low construction efficiency and low reusability. For this issue, this paper proposed a new modular method for ontology based on transforming the forest into trees. In this method, the domain was decomposed into atomic subfields, each of which was modeled individually and later assembled into a modular ontology according to the inheritance relationships among the concepts and the principles for transforming a forest into trees. We examined this modular method by constructing the ontology on the undergraduate studies level, and proved that this modular method can effectively improve the reusability of the ontology, and thus was an effective approach for constructing ontology in large scale domains.

Key words: domain ontology ; modular ontology; ontology module; module assembly

本体^[1-2]是“共享概念模型的形式化规范说明”, 其重要性已在许多方面表现出来并得到广泛认同. 目前本体已被广泛应用于信息检索、人工智能、语义网络和知识管理等诸多领域, 发挥着越来越重要的作用. 模块化可使本体的构建、维护、共享、进化变得更加容易, 而且可以有效的促进知识重用. 国外很多研究人员已提出多种领域本体建模方法, 比如 KACTUS, Methodology 等; 多个本体模块化语言及其构建工具也已被提出, 如 P-DL, OWL, protégé、Swoop 等. 国内对本体模块化的研究人员也日益增多, 他们结合实际研究情况, 借鉴现有方法, 从多个角度提出了自己的思路与方法, 并取得了一定的成果. 张维一

在自己的论文“本体模块化的研究与实现”中介绍了一个用模块化的方法构建的本体实例——汽车驾驶培训领域本体, 包括各模块的构成与模块间的通信方式^[3], 但此方法尽在汽车驾驶领域得到了验证, 重用能力较弱. 北京邮电大学林松涛的博士论文“模块化本体建设研究”主要对模块化本体建设的相关理论性问题进行了探讨^[4], 没有对本体模块的划分、组合、重用等相关方法进行实际应用. 大连海事大学刘洪军的硕士论文“语义 Web 服务中本体的模块化研究”设计了一个基于结构的本体分解方法, 并初步实现了一个本体模块化系统^[5], 此方法通过获取本体的 RDF 三元组, 然后把三元组转化成图, 并为图中不同

① 基金项目: 2013 年山东省软科学项目(2013RKA08007)

收稿时间: 2013-09-16; 收到修改稿时间: 2013-10-16

种类的边设定不同的重要度值,最后使用社会网理论计算结点之间的依赖度值,并对依赖图进行分解.此方法过程繁琐,大大降低了本体分解的效率.郭文丽等人对本体模块化研究的最新进展进行了全面综述,并提出建立基于粒度的本体模块描述方法^[6-7],为本体的构建提供了有意义的建议;冯兰萍提出了一种通过本体模块的裂变、重组、重用实现模块化本体的构建方法^[8],但此方法要求原子本体模块是已经构建良好.

本文在国内外相关研究的基础上,针对目前本体存在的构建效率低、重用能力弱的问题,提出一种基于森林转化为树的本体模块化方法(AMMOBOFIT, A Modular Method of Ontology Based on Forest into Tree),使用该方法可实现分布式的、协同式的开发、测试、维护和进化大型领域本体.

1 基础知识

1.1 本体模块和模块化本体

本体模块,它是本体中的自包含可重用的部分,模块内的概念彼此语义关联,而对模块外部的信息依赖性不高.总起来说,本体模块就是一个独立的小本体,所以每一个本体模块都可表示为一个五元组^[9],即: $O=\{C,R,Hc,Rel,Ao\}$,其中: C 为概念(concept); R 为关系(relation); Hc 表示概念间的分类关系(taxonomy relation); Rel 表示概念间的非分类关系(non-taxonomy relation); Ao 表示本体公理(axop).

模块化本体(Modular Ontology)是由若干个独立的本体模块组成的一个相对大的本体.本体模块之间通过接口连接,模块化本体中单个模块的更新不影响组成本体的其他模块,也不影响模块化本体整体的知识表示.

1.2 森林和树

(1) 定义

树(Tree)^[10]是 $n(n \geq 0)$ 个节点的有限集,在任意一棵非空树中:①有且仅有一个根节点;②当 $n > 1$ 时,其余节点可分为 $m(m > 0)$ 个互不相交的有限集 $T_1, T_2, \dots, T_m$,其中每一个集合本身又是一棵树,称为根的子树.

森林(Forest)^[11]是 $m(m \geq 0)$ 棵互不相交的树的集合.对树中每个节点而言,其子树的集合即为森林.

(2) 树和森林之间的关系

任何一棵树是一个二元组 $Tree=(root,F)$,其中: $root$

是数据元素,称做树的根节点; F 是 $m(m \geq 0)$ 棵树的森林, $F=(T_1, T_2, \dots, T_m)$,其中 $T_i=(ri,Fi)$ 称做根 $root$ 的第 i 棵子树;当 $m < 0$ 时,在树根和其子树森林之间存在下列关系^[10]:

$$RF=\{ \langle root, ri \rangle | i=1,2,\dots,m,m \geq 0 \}$$

本文根据上述本体模块和树的定义及特点,首先把本体模块用树的形式表示出来;然后借助树和森林的关系把本体和其本体模块的关系表示出来;最后再借鉴森林转化为树的方法来分解大型的成熟领域本体,构建本体模块库,降低本体构建的人力、物力,提高本体构建的效率,实现本体重用的目的.

2 基于森林转化为树的本体模块化方法 (AMMOBOFIT)

2.1 AMMOBOFIT 的流程

本文借鉴已有的领域本体构建方法,树与森林的知识,给出 AMMOBOFIT(A Modular Method of Ontology Based on Forest into Tree)的流程:

(1) 确定领域本体的目的和使用范围,并生成自然语言描述的规格说明书.

(2) 划分子领域,这同时也意味着开发任务的划分.这一步骤决定了本体模块的数量和大小,也决定了开发的效率和成本.

(3) 为开发人员分配任务,也就是将子领域本体模块的开发任务分配给开发人员.

(4) 开发人员分布式的、并行的完成各自的开发任务.在开发过程中既可以重用已有的本体模块,也可以从零开始构建各子领域的本体模块.

(5) 对建好的本体模块进行评价审核.根据评价结果,确定对本体模块继续进行优化,还是最终确认为本体模块.当然对本体模块进行评价后,可能会发现先前的子领域划分不尽合理,这时就需要回到步骤 2,重新调整子领域的划分,所以模块化本体的构建是一个螺旋前进、反复迭代的过程.

(6) 本体模块的组装:首先对建好的若干个原子本体模块进行组成,组成一个子领域模块.然后再把若干个子领域模块,组装为更大的模块;依次类推直到组装为描述整个领域的本体模块.这个过程是一个相对复杂的过程,组装时可采取自顶向下和自底向上的混合策略进行;针对如何解决本体模块组装时模块间概念关系的映射问题,例如模块间层次关系、同一概

念在不同模块中有不同表达、模块间组装的规则或方法等等. 国内外相关人员已经做了大量的研究工作, 提出了很多的解决方法^[3-9]. 但不管采用什么方法进行组装本体, 都要在不影响本体模块本身所具有的语义关系的前提下对本体模块进行重组、重用, 而且在组装后要保证语义逻辑表达的一致性.

(7) 形成文档, 这一过程伴随本体模块的开发、评价同时展开, 而且贯穿整个开发过程的始终, 其中部分类型的文档是由本体开发工具自动生成的.

综上所述, AMMOOBOFIT 的流程可用图 1 表示:

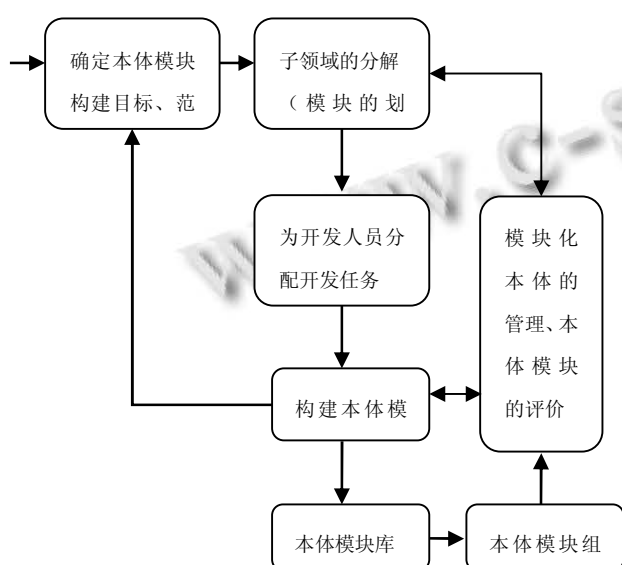


图 1 AMMOOBOFIT 的流程

由于步骤 1、3、7 比较简单, 步骤 4 和步骤 5 的研究目前比较成熟, 所以本文仅对子领域的划分和本体模块的组装做详细的深入的研究.

2.2 子领域的划分

子领域的分解是一个复杂的分析设计过程, 在相当程度上, 它依赖于设计者的领域知识和实践经验, 没有统一标准的理论方法或指导准则, 但是子领域的分解是否合理, 会直接影响到领域系统的功能、质量和实现成本. 所以在划分子领域时, 一般要遵循以下规则^[11-12]:

高内聚, 即在一个模块内部体现最大程度的关联, 只实现单一功能的模块往往具有这种特性;

最低的耦合度, 即不同的模块之间的关系尽可能弱;

模块的层次不能过深, 一般应尽量控制在 7 层以

内;

接口清晰, 信息隐蔽性好; 模块接口是描述其功能并提供与其他模块集成的方法. 所以模块接口应该清晰、简明、可靠, 必要时还要有相应的文档说明, 但是应该隐藏模块的实现细节, 避免不必要的错误发生.

模块大小适度;

尽量采用已有的模块, 提高模块复用率.

本文借鉴以上领域划分原则把待建模的领域 W 自顶向下逐层的分解为若干个相对独立但又相互关联的子领域 W_1 、 W_2 、 $\dots$ 、 W_n , 且子领域之间满足 $W_1 \cap W_2 \cap \dots \cap W_n = \perp$ 和 $W_1 \cup W_2 \cup \dots \cup W_n = W$ 的关系. 对每个子领域 W_i 可进一步划分成更小的子领域 W_{i1} 、 W_{i2} 、 $\dots$ 、 W_{im} 等, 直到满足划分的粒度要求为止, 完成模块的划分过程.

在领域模块划分结束后, 进行构建领域层次树, 具体过程如下:

树的根节点是整个待建模领域 W ;

根的直接后继节点和整个待建模领域 W 的子领域模块 W_1 、 W_2 、 $\dots$ 、 W_n 一一对应, 树枝的方向由直接后继节点指向它的根节点;

以根节点的每个直接后继节点为子节点, 与步骤 (2) 相似, 可以将该直接后继节点对应的模块中的所有子模块组织成一棵有向子树;

采用与步骤 (3) 相似的方法处理每个新产生的叶子节点, 如此类推有限次, 直到所有的叶子节点对应的模块没有子模块为止, 完成领域层次树的构建. 如图 2 所示:

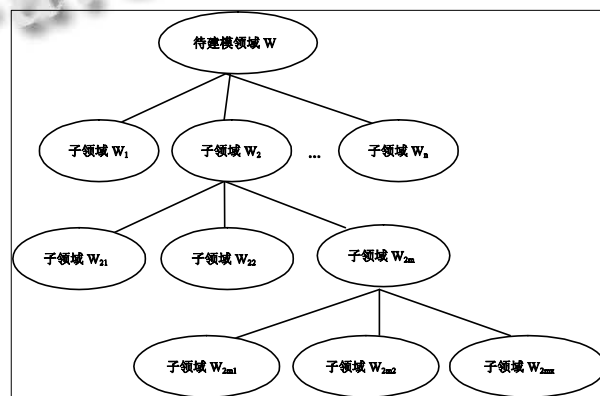


图 2 领域层次树

最后在待建模的层次树中寻找叶子节点, 也就是最后划分的足够小的子领域(在以后的章节中称为“原

子领域”),把它们构建成具有高度内部一致性的模块(本体模块),然后把这些模块按某种方法封装到本体模块库中,当需要构建特定的领域本体时,只需调用满足条件的本体模块组装即可.当然在构建领域本体过程中,应该首选相关子领域已有的本体模块,进行集成组装领域本体——事实上,这也是最理想的情况,也是广大研究模块化本体人员的研究目的,即将来的领域本体建模可以类似于目前的软件开发,从现成的精心建设的可重用本体模块库(对软件开发而言,则是各种形式的可重用类库和构件库)中选择合适的本体模块来组装本体,达到一个省事、省时,节约人力、财力的效果.

2.3 本体模块的组装

本文根据树和森林的关系、模块化本体的特点、诸多文献给出的本体模块组装规则、本体中类之间的类继承关系(sub class of()),提出一种新的本体模块组装方法:一种基于森林转化为树的本体模块组装方法(AAMOOMBOFIT, A Assembly Method of Ontology module Based on Forest into Tree).

(1) AAMOOMBOFIT 的基本思想

① 将构建好的每个本体模块,结合语义学的思想,用有向图的形式表示出来,在有向图中查找具有继承关系的父节点,并把其作为本体模块树的根节点,从而把本体模块表示成为一棵树 T,若干构建好的本体模块树构成了一个森林 F,提取森林中每棵树的根结点,得到一个概念集合 $\sigma(T_0, T_1, T_2, \dots, T_{n-1})$.

② 在组装本体模块时,为保证各模块之间保持紧密的语义关联,达到组装后的本体在语义上具有完整性、一致性,需要根据概念 T_i 到概念 T_j 之间的依赖度计算公式^[5],计算出概念集合 σ 中各概念之间的依赖度值.

$$\omega_{ij} = \frac{\sum_{E_{ij}} \theta_{ij} * e_{ij}}{\sum_k \left(\sum_{E_{ij}} \theta_{ij} * e_{ij} \right)}$$

其中 E_{ij} 表示从结点 i 到结点 j 之间边的集合, e_{ij} 表示 E_{ij} 中某一类边的数目, ω_{ij} 表示结点 i 对结点 j 的依赖度值, $\theta_{ij} = \sum_{e \in E_{ij}} \varphi(e(i, j)), 0 < \varphi \leq 1$;

$$\varphi_{ij} = \varphi(e(i, j)) = \begin{cases} 1, e \in \{SubClassOf\} \\ 0.75, e \in \{SubClassOf\} \\ 0.5, else \end{cases}.$$

③ 把概念集合 σ 中各概念之间的依赖度值 ω_{ij} 表示为一个二维数组,如图 3(a)所示,并把这个二维数组看成是这样一个定长线性表,它的每个数据元素也是一个定长线性表,行向量中的每个数据元素是一个列向量形式的线性表,如图 3(b)所示,

$$\omega_{ij} = \begin{bmatrix} 1 & \omega_{01} & \dots & \omega_{0,n-1} \\ \omega_{10} & 1 & \dots & \omega_{1,n-1} \\ \dots & \dots & 1 & \dots \\ \omega_{n-1,0} & \omega_{n-1,1} & \dots & 1 \end{bmatrix}$$

(a) 各概念间的依赖度值

$$\omega_{ij} = ((1, \omega_{01}, \dots, \omega_{0n}), (\omega_{10}, 1, \dots, \omega_{1,n-1}), \dots, (\omega_{n-1,0}, \omega_{n-1,1}, \dots, 1))$$

(b) 行向量的一维数组

图 3 ω_{ij} 的计算和表示

把行向量中的每个元素以链表的结构存储,以备使用.

④ 首先在第一个行元素的链表中查找与第一个本体模块树根结点 T_0 具有最高依赖度值的结点 T_i , 输出结点 T_j , 并用有箭头的线连接 T_0 和 T_i , 从而把两个模块组装成一个较大的本体模块.

⑤ 然后在其它行元素的链表中查找与结点 T_i 具有最高依赖度值的结点 T_j , 输出结点 T_j , 并用有箭头的线连接 T_i 和 T_j , 从而组装成一棵更大的本体模块.

⑥ 反复调用过程(4)和过程(5),直到把森林 F 中的所有本体模块树组装成一棵模块树为止.模块化本体构建完成.

(2) AAMOOMBOFIT 的算法描述

输入:若干原子领域本体模块.

输出:模块化领域本体.

步骤 1: 构建本体模块的本体图

① 根据待建模领域中的关键词、概念及概念之间的关系,在本体模块库中查找已有的本体模块;

② 把原子领域对应的各个本体模块用 OWL 语言描述,得到 n 个原子本体 $(O_1, O_2, \dots, O_n)$, 其中 $On(Cn, Rn, Fn, An, In)$.

③ 为各个原子本体构建本体图.因为在本体中,继承关系(SubClassOf)是类之间最重要的关系,所以在本体图中将用有箭头的线表示继承关系,无箭头的线表示其它关系.

步骤 2: 把各个本体图看作仅有一个根节点的树 T , 方法如下:

① 在建好的原子领域本体图中寻找具有继承关系的类 C_i ;

② 由于原子领域本体图只有一个继承关系, 所以就把其中的一个父类作为根节点, 其余节点作为子节点. 例如在本体图 4 中, 节点 b 为根节点, 节点 d 、 e 、 f 是和 b 具有继承关系的子节点, 而 a 、 c 则是和 b 具有其它关系的子节点, 而节点 d 、 e 、 f 之间还存在着其它关系;

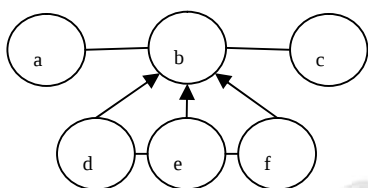


图 4 树形本体图

③ 把为若干个子领域构建的树形本体图看成一个森林 F .

步骤 3: 把由若干棵树组成的森林逐步转化成一棵树.

任意选择一个原子领域树形本体图, 假设是图 4, 找到其根节点 b ;

在概念集合 $\sigma(T_0, T_1, T_2, \dots, T_{n-1})$ 中, 查找和根节点具有继承关系的类 g , 并用有箭头的线把节点 b 和节点 g 进行连接;

在概念集合 $\sigma(T_0, T_1, T_2, \dots, T_{n-1})$ 中, 查找和节点 g 具有继承关系的其它所有类, 并用有箭头的线把这些类节点和节点 g 进行连接; 如图 5 所示.

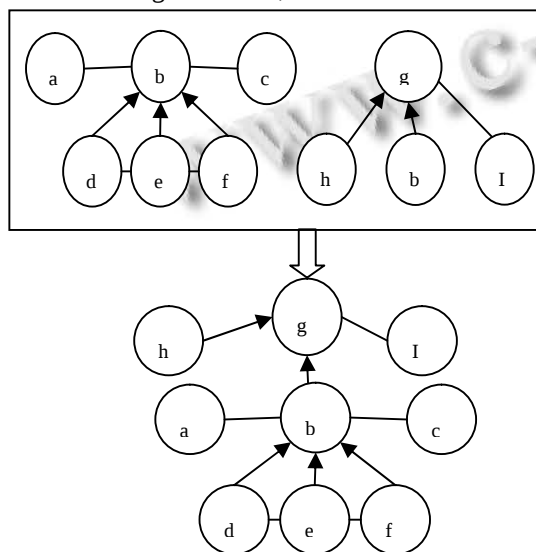


图 5 由两棵树形本体图组成一棵树形本体图

同样的方法, 在概念集合 $\sigma(T_0, T_1, T_2, \dots, T_{n-1})$ 中, 查找节点 g 的父类、和节点 h 具有继承关系的其它类, 并用有箭头的线把这些节点进行连接, 构成一个更大的树形本体图. 重复以上直至整个待建模的本体图构建完成.

步骤 4: 对构建的本体图进行优化处理, 本体模块组装成功, 模块化本体构建完成.

2.4 AMMOBOFIT 的优点

与已有的本体构建方法相比, 本文研究的 AMMOBOFIT 具有以下优点:

(1) 将待建模领域划分成若干个较细粒度的子领域, 使得代码清晰、易于阅读和理解, 同时降低了开发的难度和成本.

(2) 开发人员可以分别独立的开发或测试各个模块, 灵活性和扩展性更强.

(3) 模块化的本体结构可有效控制局部的修改对整体的影响, 使本体产品更加可靠、健壮、易于共享和进化.

(4) 为构建本体模块库提供了可行性.

3 基于AMMOBOFIT的模块化本体构建

为验证本文给出的 AMMOBOFIT 的可行性、有效性和正确性, 本小节使用此方法构建了一个简单的大学本体. 大学本体既涉及教学方面的相关概念, 也包括科研项目的定义, 所以它涉及多学科、多领域, 各种关系错综复杂. 建好的大学本体图如图 6 所示, 图 6 中的数字分别代表的概念如表 1 所示.

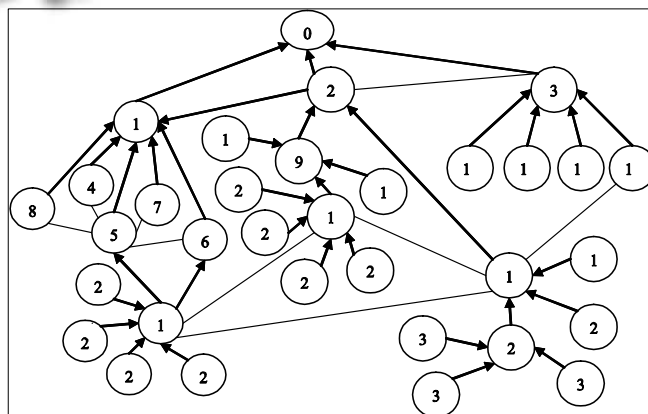


图 6 大学领域的领域层次树

表1 图6中的数字和大学领域中概念对照表

类别 目录	数字 标记	类别 目录	数字 标记
大学	0	教师	17
机构	1	技术人员	18
人员	2	本科生	19
活动	3	研究生	20
党政部门	4	专科生	21
院系	5	博士生	22
实验室	6	硕士	23
编辑部	7	本科	24
后勤处	8	专科	25
教工	9	教授	26
学生	10	副教授	27
教学	11	讲师	28
科研项目	12	助教	29
教改课题	13	博士后	30
社团	14	博士	31
课程	15	硕士	32
行政职工	16		

4 结语

本文的研究工作和创新主要集中在本体模块化的理论性课题上,在进一步的研究中,笔者将针对本体模块库的构建,本体模块的自动组装展开研究,以期大型本体的开发早日走上工业化生产道路。

参考文献

- 1 Studer R, Benjamins VR, Fensel D. Knowledge engineering, principles and methods. Data and Knowledge Engineering, 1998, 25(1-2): 161-197.
- 2 邓志鸿,唐世渭,张铭等.本体论研究综述.计算机研究与发展,2004,41(17):1041-1052.
- 3 张维一,陆汝占.本体模块化的研究与实现.计算机应用研究,2007,24(11):206-209.
- 4 林松涛.模块化本体建设研究.北京:北京邮电大学,2006.
- 5 刘洪军.语义 Web 服务中本体的模块化研究[学位论文].大连:大连海事大学,2008.
- 6 郭文丽,张晓林.本体模块化:特征、描述与方法.图书馆杂志,2008,27(9):5-10.
- 7 郭文丽,张晓林.基于粒度的本体模块描述方法.现代图书情报技术,2010(2):1-6.
- 8 冯兰萍,朱礼军,蒋亚东.一种模块化本体构建方法研究.现代图书情报技术,2010(6):53-59.
- 9 Maedche A. Ontology Learning for the Semantic Web. Boston: Kluwer academic Publishers,2002.
- 10 严蔚敏,吴伟民.数据结构(C 语言版).北京:清华大学出版社,2003,118-121.
- 11 冯兰萍,朱礼军,张继国.基于模块化本体的知识组织与服务体系研究.现代图书情报技术,2008(7):23-27.
- 12 杨倩,冯志勇,饶国政等.物流领域本体的模块化研究与实现.天津大学学报,2009,42(12):1105-1111.