

doi:10.3772/j.issn.1000-0135.2011.01.008

基于 OWL 的 本体半自动进化研究¹⁾

蔡丽宏 马 静 吴一占 谭 胜 薛 浩
(南京航空航天大学经济与管理学院,南京 211100)

摘要 本体随着领域、时间、应用环境等的变化而不断发生着演变,因此,本体的完善以及随环境的变化而进化成为了当前本体研究的重点。本文在引进国外本体进化理论及充分掌握 OWL 语言结构的基础上,提出了一个具有领域特色的本体半自动进化方案,引进了本体原子变化和复合变化,设计了武装直升机领域本体变化获取算法和基于 OWL 语言的本体语义变化算法,提出了本体原子变化集合的实施顺序算法,最后通过实验验证该领域本体半自动进化方案是可行的,且具有很大拓展性的。

关键词 本体半自动进化 一致性 语义变化

Study of Semiautomatic Ontology Evolution Based on OWL

Cai Lihong , Ma Jing , Wu Yizhan , Tan Sheng and Xue Hao
(College of Economics and Management , Nanjing University of Aeronautics and Astronautics , Nanjing 211100)

Abstract Ontology may evolve when the domain, time and application' requirement change. Therefore, the evolution of ontology along with environment change become the key point of current research. This paper proposes a framework for domain ontology evolution semi-automatically based on foreign ontology evolution theory and fully grasping the OWL language structure, introduces ontology atomic changes and complex changes , designs the algorithm of capturing ontology changes semi-automatically and of semantic changes based on OWL language, introduces the operation order of a collection of ontology atomic changes. finally it verifies that this the program of semi-automatic domain ontology evolution is feasible, and scalable through the experiment.

Keywords ontology evolution, consistency, semantic change

1 引 言

随着时间的推进,环境的变化,领域会相应的发生一些变化,这些变化可能是新增加了一些新的词汇、词汇含义的变更或者词汇在领域内寿命的终结等,而这又会导致概念之间关系的变更,可谓“牵一发而动全身”。由于本体本身这种特性,要保证本体

的一致性状态,手动进化本体显然很困难。如果本体很庞大,难度就更大了,可以说基本上不能实现本体的进化,更不要说达到用户需要的进化本体一致性状态。如果想基于本体的系统应用于实际,必须解决本体的进化问题,但是由于本体进化各环节的难度,全自动进化现在几乎是不可能^[1,2]。

目前,本体进化技术尚不成熟,能够实现本体自动进化的工具迄今还没有,走在本体技术研究前沿^[1,3]的德国 University of Karlsruhe 也只是提出一个

收稿日期: 2009 年 11 月 9 日

作者简介: 蔡丽宏,女,1984 年生,硕士生,研究方向:知识管理、信息管理。E-mail: cailihong1340201@126.com。马静,女,1966 年生,教授,研究方向:信息管理,知识管理。吴一占,男,1985 年生,硕士生。谭胜,男,1985 年生,硕士生。薛浩,男,1985 年生,硕士生。

1) 本文系国防技术基础项目研究成果之一。

本体进化的原型系统 KAON, 并没能付诸应用。由该大学 FZI 和 AIFB 研究中心提出的本体进化思路: ①变化捕获; ②变化表示; ③语义变化; ④变化执行; ⑤变化传播; ⑥变化确认, 成为了本体进化研究的指导思想。

本文在借鉴这个思路的基础上, 提出了新的领域本体半自动进化方案, 改造了符合武装直升机领域的本体变化获取思想, 在熟知 OWL 语言的基础上, 设计了一个本体语义变化算法。

2 本体进化的类型

领域本体随目标领域变化而变化, 我们将本体的这些变化分为两个层次: 结构 (structural) 变化^[4]和语义 (semantic) 变化^[5]。结构变化主要包括: ①概念的结构变化; ②概念属性的结构变化; ③属性约束的结构变化; ④公理的结构变化, 例如: 增加/删除概念、增加/删除属性约束等等, 值得注意的是每个结构变化都局限在本体的某个特定部分。语义变化^[6]特别指的是由本体展示的在语义层次上的变化, 主要包括: ①概念泛化变化; ②概念描述的变化; ③属性泛化的变化; ④属性约束的变化; ⑤属性变更的公理变化。

结构变化对本体来说是显式的, 而语义变化不是明显的, 这些语义变化可能是由结构变化引起的显式语义变化, 或由概念和属性间依赖性所带来的隐式语义变化, 可以通过评估可见的结构变化得到的。

3 本体半自动进化方案

3.1 本体半自动进化技术路线

本体进化是一项复杂的工程, 完全自动进化现在几乎是不可能, 因此本文借鉴 FZI 和 AIFB 研究中心提出的本体进化思路, 提出一种本体半自动进化技术路线, 主要包括: ①领域本体变化获取; ②本体变化描述; ③复合变化原子化; ④语义变化; ⑤前项一致性条件检查; ⑥变化实施; ⑦本体确认; ⑧本体的后项一致性条件检查, 如图 1 所示。

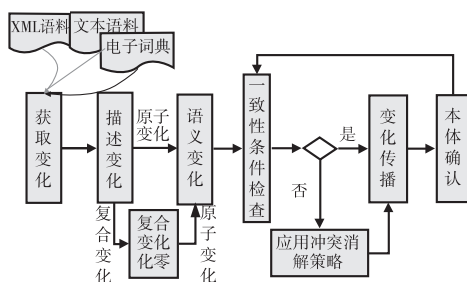


图 1 本体半自动进化技术路线

3.2 本体变化半自动发现技术

3.2.1 思想原型

2005 年, Kavalec 等提出使用扩展的关联规则挖掘方法为本体中概念间的非分类关系赋予语义标签。其基本思想是^[7]: 如果两个概念间存在非分类关系, 那么该关系能够用经常出现在这两个词附近的某个动词来表示。所以, 可以通过计算某个动词和某两个概念一起出现的条件概率决定这两个概念之间的关系是否可以用该动词来表示。这种方法是对解决该问题的一个初步尝试, 但它仅考虑了词频, 没有考虑句子结构等其他因素, 所以结果并不十分理想。

3.2.2 改进思想

笔者受关联规则挖掘方法的启发, 提出一种整体发现本体变化的思路, 通过获取包含了本体变化的语句, 整体发现领域本体变化的新概念、新关系以及变化操作。

由于中文表达思维和习惯, 本体新概念的出现必定会伴随着一些特定的词语, 如果找出包含了这些触发词的语句, 并对这些语句进行分词统计, 只要处理语料范围足够广泛, 根据词频就能够比较准确地找出新概念, 而包含新概念的语句里含有概念间的关系和变化操作信息的几率也会更大, 这样就可以在一句话里找出我们所需要的多项信息, 所以本方案设计的技術路线如图 2 所示。

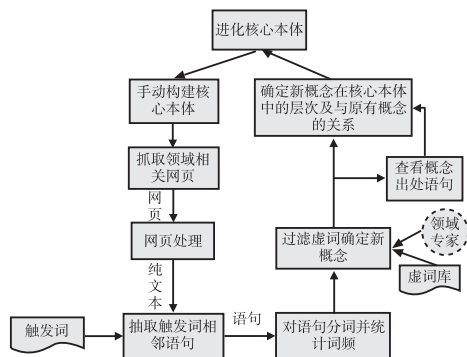


图 2 本体变化半自动发现技术路线

主要有以下步骤:

(1) 手动构建核心领域本体

在本文的进化方案中,我们需要手动构建一个尽可能全的核心本体,在此基础上用来自 Web 网页的文本信息实现本体的半自动进化实验。

(2) 抓取有关 Web 网页

使用信息抽取部分的 Web 信息自动抓取,剔除杂乱 Web 网页的算法,检索出与产品相关的网页,提取出纯文本。因为文本是本体进化的语料库源,网页的纯净度关系到后面操作的正确性和可信度。

(3) 根据触发词抽取相邻语段

从获得的文本中进一步抽取与触发词紧邻的语段,进一步缩小新概念搜索的范围。那么如何定义“紧邻”? 这里,我们假设包含了触发词的那句话,以句号为界限。由于如果一句话的获取范围过窄,会露掉了重要的语段信息;但如果过宽,抽取了没实质用处的信息,就会增加后面基于分词的词频统计工作。因此,笔者认为需要在具体的程序实验中,经过设定不同范围值,选取一个兼顾效率和效果的语段获取范围。

(4) 基于分词的统计

对抽取来的语段进行分词,统计在所有语段中多次出现的连续的几个字,默认为是一个词,作为新概念获取的一个局部词库,并且分析这些词语的来源(即网址)和出现的次数(从高到低排列)。

(5) 根据统计次数确定“新概念”词库

规定一个值,以此为界,出现频率高于这个值的字段即认为是一个词;低的则认为不是。这个值需用试探法,先设定一个数值,通过不断的实验规定一个比较合适的界值。最小值应该不能小于 1,如果是 1 的话,语段中凡是出现的汉字都被认为是词,有的可能只是一个字,比如“的”,这显然是不合适的,而且还影响了后面步骤的效率。

(6) 过滤“虚词”

第 5 步获得的“新概念”词库,由于是计算机判断,机械地把重复出现的连续字段抽取出来,没有人脑的联想和判断能力,抽取到“虚词”的概率非常大,因此,必须把这样的字段过滤掉。可以根据新建的或已有的局部的“虚词”库,从“新概念”词库中减去“虚词”库,就基本上获得了新概念了。

(7) 领域专家判断新概念

第 6 步虽然已经由机器获得了新概念,但是难免会失误,在确定是否为新概念之前,还需领域专家来最终确认,是否符合构词法,是否是实际的概念

信息。

(8) 确定新概念在核心本体中的层次

本方案借鉴图书馆学叙词表和分类表编制技术及机器学习中的自动技术,提出了相似度关联思想。即如果概念之间的相似度达到了一定的阈值,即认为两概念之间有一定的关系,模糊确定新概念在本体中的大致区域。而阈值定哪个数值需要实验来验证。

(9) 确定新概念与核心本体概念的关系

新概念如何与核心本体融合,涉及到多种情况:添加、替换、删除等,每一种操作对本体的改动都不同,难度也不同。本方案根据新概念的源出处,分析新概念和核心本体中概念的关系。

3.2.3 语义变化算法设计

复合变化经分解后成了最基本的原子变化,但是这些分解的原子变化只是显式操作,还有更深层的隐含语义变化。例如,Remove-class C,根据本体进化策略,将 C 的子类都添加到 C 的父类 A 中作为子类。但是与 C 有关的类、属性和实例将如何处理呢,这就涉及了本体的语义变化及一致性状态维护。由于本体是一个由多个概念、关系和实例组成的复杂语义网络,一个概念通常会与其他多个概念相互关联,当这个概念发生变化,就可能导致与其相关的概念与本体定义中的其他部分产生冲突,从而会导致本体的不一致性^[8]。

那么如何确定本体中与原子变化对象相关的所有实体呢? 根据以上本体构建经验,系统需要捕获的相关对象主要是与变化相关的类、属性和实例,因此,笔者主要讨论本体基本元素的语义变化。

3.2.4 语义变化算法

根据 OWL 语言的结构特点即 owl 语言相关实体间标记符,已知原子变化相关参数“A”,找出与 A 相关的所有类、属性和实例,进行语义变化。首先从本体 OWL 文档中搜索出包含 A 的所有语句,以<>为界限,然后根据以下算法确定相关参数种类。

if A 为类(class),则定位相关类、属性和实例算法为:

```
Case1: < rdfs: subClassof .../> or < rdfs:
subClassof>...</rdfs: subClassof>
then 向上寻找第一个<owl: Class...>,
ID=//寻找子类 Case2: <owl: class...>
then 向下寻找第一个< rdfs: subclassof ...>,
```

```

resource = //寻找父类
Case3: <owl: disjointwith.../>
then 寻找与它相连的所有 disjoint 类
resource = //寻找互斥类
Case4: <rdfs: domain...>
then 寻找邻近第一个<owl: objectproperty>...</
owl: objectproperty> or < owl: Dataproperty >...</
owl: Dataproperty>,
ID = or about = //寻找相关属性
Case5: <rdfs: range...>
then 寻找邻近第一个<owl: objectproperty>...</
owl: objectproperty>
ID = //寻找相关属性
Case6: < owl: somevaluesfrom .../> or < owl:
hasvalue .../> or < owl: allvaluesfrom > ... </owl:
allvaluesfrom>
then 寻找邻近第一个<owl: onProperty .../>,
resource = //寻找相关属性
Case7: <owl: ObjectProperty...A...>
then 向下紧邻的第一个<rdfs: subPropertyof
.../>
resource = //寻找相关属性
Case8: <rdfs: subPropertyof...A.../>
then 邻近第一个<owl: ObjectProperty...>...</
owl: ObjectProperty>
ID = //寻找相关属性
Case9: <rdfs: type.../>
then 邻近第一个<owl: Thing...>
Rdf: about = //寻找相关实例
if A 为属性 (property), 则定位相关类、属性算
法为:
Case1: <owl: ObjectProperty...A...>
then 邻近的第一个 domain_range ,
resource = //领域类和范围类
Case2: <owl: onProperty...A...>
then 邻近第一个 class ID, resource//领域类和
范围类
Case3: <owl: ObjectProperty...A...>
then 向下邻近的第一个 subPropertyOf
resource = //父类属性
Case4: <rdfs: subPropertyOf...A...>
then 向上第一个 objectProperty//子类属性
if A 为实例 (individual), 则定位相关类的算
法为:

```

```

Case1: <owl: Thing ...A...>
then 向下第一个<rdfs: type...> , resource = //所
属类
Case2: <A, .../>
then 其中 ID = //所属类
Case3: <rdfs: Description...A...>
then 判断邻近是否有 one of 标记, 如有, 则所属
类为最外围的 Class, ID = .....

```

4 本体半自动进化方案实验验证

4.1 实验设计

这里, 我们开发了航空领域的抓取软件, 以武装直升机为实验领域, 设定抓取任务, 从万维网上找到包含武装直升机本体变化的语句, 进行本体变化描述, 语义变化, 对原子变化进行前向一致性条件检查, 根据需要制定进化策略确定本体原子变化集合, 按序执行, 根据本体一致性模型约束条件进行后向一致性条件检查。

4.2 实验过程

4.2.1 复合变化进化处理

(1) 领域本体变化相关语句搜索

根据本体变化获取思想, 我们利用开发的抓取软件从网上找到一句包含了变化信息的语句: “箭头”系统于去年被陆军选中, 以替换 TADS/PNVS 系统。

(2) 变化描述

ocs = (replace, “箭头”系统、TADS/PNVS 系统, 存在 TADS/PNVS 系统、不存在“箭头”系统)。

(3) 复合变化原子化

replace (“箭头”系统, TADS/PNVS 系统), Remove TADS/PNVS 系统; Add “箭头”系统。

(4) 语义变化

根据语义变化算法, 从 OWL 文件中找出“TADS 与 PNVS 系统”相关的实体, 并删除“TADS/PNVS 系统”的所有相关实体, 包括子类和属性, 得到本体原子进化集合是:

```

Replace (“箭头”系统, TADS 与 PNVS 系统);
Remove TADS 与 PNVS 系统;
Remove-Subclass (一台直视光学装置望远镜,
TADS 与 PNVS 系统);
.....

```


Add “箭头”系统;

Add-Subclass(“箭头”系统, 瞄准系统与导航系统)。

(5) 前向一致性条件检查并按序执行操作

经过本体前向一致性约束条件检查, 发现上述原子操作均满足条件。因此, 定义本体变化操作集合。

(6) 进化结果

将原构建好的武装直升机导入 Protégé 中检验: 将“箭头系统”取代“TADS 与 PNVs 系统”, Run ontology tests 测试之后, 发现进化后的武装直升机本体没有结构和逻辑错误, 符合本体一致性模型约束条件(注: Run ontology tests 是 Protégé 中检查基于 OWL 本体结构和逻辑一致性状态的插件)。

4.2.2 原子变化进化处理

(1) 领域本体变化相关语句搜索

另外, 抓取软件从网上获取信息“AH-64B 是根据……提出的改型, 与 AH-64A 相比主要加大了左前方的电子设备舱, ……”, 加装了卫星全球定位系统(GPS)和自动目标移交系统(ATHs), 并改善了直升机的可靠性、适用性和维护性(RAM)。”

(2) 变化描述

ocs1 = (Add, 卫星全球定位系统(GPS)、卫星导航设备, 不存在卫星全球定位系统(GPS));

ocs2 = (Add, 自动目标移交系统(ATHs)、其他航空电子设备, 不存在自动目标移交系统(ATHs))。

这两项操作都是基本的原子操作, 不需要进行复合变化原子化操作, 接下来就可以进行语义变化。

(3) 语义变化

Add-Subclass (卫星全球定位系统(GPS), 卫星导航设备);

Add-Subclass (自动目标移交系统(ATHs), 其他航空电子设备)。

(4) 前向一致性条件检查并按序执行变化操作

在操作之前, 根据本体一致性模型约束, 检查这两个原子变化操作是否符合条件, 发现满足条件, 系统允许进一步实施操作。这两个操作都是独立的原子变化操作, 不需要考虑执行顺序, 直接在 Protégé 里添加操作即可。

(5) 本体进化结果

进化前的本体, “卫星导航设备”类只有伽利略导航系统, “其他航空电子设备”类没有子类。而根

据请求进化后, “卫星导航设备”类添加了卫星全球定位系统(GPS), “其他航空电子设备”类添加了自动目标移交系统。经一致性检查没有错误。Run ontology test 后发现进化本体符合一致性模型约束条件!

5 总 结

本文在引进国外本体进化理论的基础上, 提出了一个具有领域特色的本体半自动进化方案, 并以实验验证其可行性。当然, 这个思想要取得比较理想的效果, 需要注意: ①处理语料的关联度。语料一定要和领域本体相关联, 语料的不纯洁将会直接或间接影响概念和关系的获取, 加大运算量, 也会加大本体专家的工作量, 因为最后还得由专家来确定概念和关系操作。②触发词集合的完备性。触发词的完善性会影响到相关语句的正确性。而对于每个特定领域的触发词是不一样的, 搜集齐全也很困难。③语句的分词。分词算法的有效性, 直接影响新概念的确立。④本体工程师的监督。以上思想发现的语句是否能反映本体的变化最终还需专家来确定。

参 考 文 献

- [1] 马静, 宋晴晴, 刘思峰. 基于 OWL 的领域本体的综合构建与进化[J]. 情报学报, 2007, 26(6): 827-832.
- [2] Ljiljana S. Methods and Tools for Ontology Evolution [D]. Germany: University of Karlsruhe, 2004.
- [3] FZI, AIFB. KAON is an open-source ontology management infrastructure targeted for business applications [OL]. [2009-06-08]. <http://kaon.semanticweb.org>.
- [4] Noy N F, Klein M. Ontology Evolution: Not the Same as Schema Evolution [J]. KIS, 2004, 6(4): 428-440.
- [5] Li Qin, Vijayalakshmi Atluri. Evaluating the validity of data instances against ontology evolution over the Semantic Web [J]. Information and Software Technology, 2009, 51(1): 83-97.
- [6] 谢强, 张磊. 基于用户自定义变更的本体进化[J]. 南京航空航天大学学报, 2006, 38(6): 796-802.
- [7] 马文峰, 杜小勇. 领域本体进化研究[J]. 图书情报工作, 2006, 50(6): 71-75.
- [8] Plessers P, De Troyer O. Resolving Inconsistencies in Evolving Ontologies [J]. Lecture Notes in Computer Science, 2006, 40(11): 200-214.

(责任编辑 王建平)

