基于本体描述的构件库中本体演化研究

杨明华 钱乐秋 赵文耘 唐姗
(复旦大学计算机科学与工程系软件工程实验室,上海市智能信息处理重点实验室上海 200433)

摘要：在构件的描述与检索中引入本体，可以更好地表示构件间的语义信息。本体是对知识的表达，可以对知识进行有效的组织和管理，实现知识共享和重利用，从而充分有效地利用知识资源。但是知识不是一成不变的，基于本体描述的构件库系统中的本体库需要更新和维护。本文提出了一种本体演化理论框架，并探讨了其研究方向和重点。

关键词： 本体 知识 演化
中图法分类号： TP301 文献标识码： A 文章编号：

A Study on Ontology Evolution for the component library system

Yang Minghua Qian Leqiu Zhao Wenyun Tangshan
(*Department of Computer Science and Engineering, Shanghai Key Laboratory of Intelligent Information Processing,Fudan University, Shanghai* 200433)

Abstract： By introducing ontology into the field of software-component description and retrieval ,the semantic information among the components can be presented more precisely. Ontology is the representation of knowledge. It can make knowledge sharable and reusable by managing it efficiently. Thus we can exploit the utility of knowledge resource more effectively. However, we should update and maintain ontology because intelligence is not invariable. This paper propose a framework of ontology evolution from owl files, the research direction and key points are also discussed.

Key words：Component Description, Ontology, evolution model

引言

软件复用的必由之路是软件构件化，构件描述是构件入库、组装的前提条件，基于刻面的构件描述对于反映构件静态信息已经有成功的应用，但是缺乏语义信息和动态性。在构件描述中引入本体，可以表示丰富的语义信息，通过推理机制，更加可以得到隐含的语义信息，在构件中引入本体，可以实现面向服务的构件检索。而这些受益都需要对构件进行良好的本体描述，本体库的建立和演化就成了对构件进行本体描述的先决条件。本体是用来哦描述和表示知识的，信息世界知识是海量的，无法依赖手工处理来管理本体的演化。现有的一些研究工作表明[1,2]，如何让本体具有演化能力，成为亟待解决的问题。

1 本体简介

1.1 本体的定义

本体的发展覆盖的领域很多，如哲学、人工智能领域、知识工程领域等，在不同的领域有不同的定义，关注的焦点也不同。本体源自哲学，在哲学中指的是对客观存在的系统的解释和说明。在 AI 领域，Neehes 等人将本体解释为 定义了包含相关领域词汇的基本术语和关系，以及组合这些术语和关系定义词汇外延的规则 [3]。

基金项目：国家自然科学基金项目（项目号：60473062），国家 863 项目 (2004AA1Z2330)，上海市科委攻关项目 (04D215022)
作者简介：杨明华（1979-），男，硕士研究生，主要研究方向：软件工程。钱乐秋 (1942-)，男，教授、博士生导师，主要研究方向：软件工程，软件复用，构件技术。赵文耘，男，教授、博士生导师，主要研究方向：软件工程；唐姗，女，博士研究生，主要研究方向为软件工程

1.2 本体的功能

总的来看，本体可作为知识表达的基础，避免重复的领域分析，并通过统一的术语和概念达成知识共享的目的。目前，虽然在计算机领域对本体的研究越来越热，但是由于技术固有的难度以及研究的时间还不长，距离实际应用尚有一段差距。为实现构件语义表示需构建大量的本体，尤其是领域本体,来满足其需求。但本体和知识库从何而来?相对于因特网海量信息而言，目前只有很少手工构件的本体，例如WORDNet[4]。另一方面用手工构建本体需要耗费大量的人力和时间，与此同时，诸如 WordNet[4]等这些通用本体只包含非常少的领域知识。而且，如何维护现有本体，尤其是如何保持同步更新也是很重要的问题，因为新的概念、新的知识和属性需要不断引入。为了解决本体工程中知识瓶颈问题，我们需要自动化或半自动化工具来构建、维护和更新本体。

2 本体的演化过程

将本体引入计算机软件领域就是因为本体含有丰富得语义信息，基于这些信息可以进行语义推理，得出隐含得语义信息，然而，知识不是一成不变的，尤其是当前信息技术发展迅速，知识更新的更快。本体库中的知识需要修改、添加、删除，如何让非结构化、半结构化，结构化本体库支持半自动、自动的协作式的本体工程，不断的演化，从而将原有的本体不断扩大与完善。在本体演化过程中，不能引入语义冲突和不一致性，已经进行语法和规则检查的本体库，不能因为加入新的知识而变得不一致。首先我们给出知识本体一致性的形式化定义：

定义 1 本体是一个三元组<V，A，E>，其中词表 V 是谓词符号的子集，公理 A 是合适公式的子集，而 E 是本体扩展集。

定义 2 令 K 为：$R \to 2^W$，将每一资源映射为一组合适公式的函数，称 K 为 知识函数 ,因其提取包含在资源中的知识并为其提供公理。R 为因特网资源集，包括任何通过因特网提供信息的事物,如 Web 页面、新闻组或 E-mail 信息等。

O_u =<{Faculty}，§¶ §¶>

K(r1)={Faculty(Dr Li)}

K(r2)= {Faculty(Dr Zhang)}

变化后的本体和资源：

O_u' = ({Faculty,AssistProf,AssocProf,Professor},

{AssistProf(x)→Facuhy(x);AssocProf(x)→Facuhy(x);Professor(x)→Facuhy(x)},§¶>

K(ri')=(Faculty(Dr Li)}

K(r_2')=(Faculty(Dr Zhang)}

K(r_3')={AssocProf(Dr Wang)}

上述定义表明了当一个新的术语项添加到本体中时所带来的影响。该例中 O_u,r_1 和 r_2 表示一个简单的 University 本体及两个遵从该本体的资源。O_u 包含一个术语 Faculty，而 r_1 和 r_2 应用 Faculty 谓词的资源。经过演化后，O_u'，r_1'、r_2' 和 r_3'代表相关 Web 对象的状态。本体 O_u'表示一个新版本的 O_u,它包括表示 Faculty 子类的术语。当本体设计者以这种方式增加术语项时，就有可能增加公理。如 Professor(x)→Facuhy(x)以帮助定义术语。其中 r_1' 和 r_2'是变化后的 r_1 和 r_2，因为 $K(r_1')=K(r_1)$且 $K(r_2')=K(r_2)$，这些资源并未发生变化。由于词表 O_u'的词汇表 V 是 V 的一个超集，对于 O_u'而言 r_1 和 r_2 仍是定义明确的。一旦 O_u 改变为 O_u'，我们就可应用 O_u'中的新术语项创建资源，r_3'就是包含有关 Dr Wang 断言的一种资源。

3 本体演化框架模型

本文旨在实现从 OWL 文件自动抽取本体，从 xml 数据中找出本体语义概念的模式及其关系。它通过分析同一应用领域 OWL 文件集来半自动化地抽取本体。其基本步骤包括：(1)OWL 文档集的收集、选择和预处理；(2)生成候选关键词集；(3)抽取领域术语；(4)领域概念过滤、筛选 (5)确定语义关系分类层次体系和语义相关度；(6)得到新的领域本体

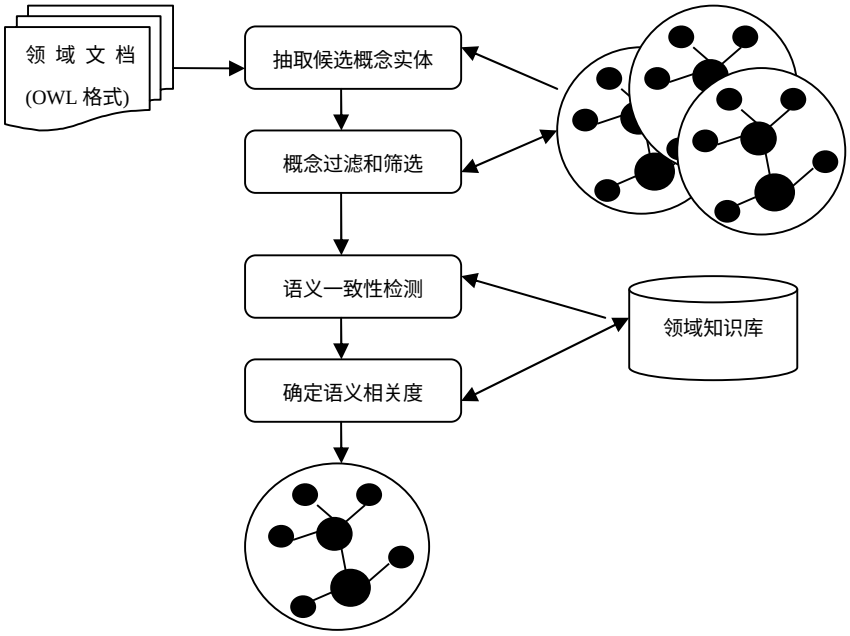

图 1 本体演化框架

3.1 概念术语抽取

术语可以看成是相关领域概念的外部表现。为了简单起见，系统只处理正文文本信息。输入一个OWL[5]文件，输入接口扫描该 OWL[5]文件，通过 XML TAG 定位，进行文档的特征抽取，并找到 OWL 文本中可能分类为实体的词组或短语。

步骤 1文档的特征抽取：一个OWL文件中包含了丰富的信息，但最主要的还是正文的文字信息。我们并没有应用自然语言处理技术，过滤掉那些不包含子主题或关键概念的文档。

步骤 2 句子锚定：选择那些可能包含本体关系的句子 S，如果 S中包含了是 C的成员的提示词，则某一句子被锚定。更准确地说，某一锚定句子可以重写为字符串其中和为名词短语或术语序列。

步骤 3每一候选规范词的查询结果进行汇总。重复每一候选规范词，将所有候选规范词和候选本体概念导入语言模式以提取假设短语。

3.2 概念过滤和筛选

构成本体的词汇，数量是相当大的，表示本体的向量空间的维数也相当大，可以达到几万维，因此我们需要进行维数过滤和筛选的工作．对于每一类，我们应去除那些表现力不强的词汇，筛选出针对该类的特征项集合。目前，存在多种筛选特征项的算法，本系统中采用了词和类别的互信息量进行特征项抽取的判断标准．我们使用如下算法进行概念的过滤和筛选：

输入：概念与属性权重值和邻近层次概念权重值，综合权衡因子 α

输出：概念 c与 c'的结构相似值

Step1: 初始化 Psim = DCsim = 0;

Step2: 初始化 Ssim = 0;

Step3: 获得概念 c和 c'的直接概念环境 PS(c)和 PS(c')

Step4: 根据概念的属性计算相似度

for 概念 c的每个属性 pi

for概念 c'的每个属性 pj

//m和 n分别为概念 c和 c'的最大属性值

Step5: 根据概念的邻近层次概念来计算相似度

for 概念 c的每个邻近概念 ci

for概念 c'的每个邻近概念 cj

//p和 q分别为概念 c和的 c'最大邻近概念值

Step6: 对属性相似度和邻近层次概念相似度进行加权计算

Ssim = α *Psim + (1-α)*DCsim;

Step7：Return Ssim;

3.3 语义一致性检测

经过解析得到的概念结果集，是需要引入现有本体库的，但还需要进一步分类。语义解析和分类得主要任务是进行语法分析，检查是否存在语义冲突。知识本体的一致性检测是通过将知识本体转换为一阶谓词表示形式，我们可以利用一阶谓词逻辑的一致性检测方法来检测知识本体的一致性，具体方法如下：

1．对于一个知识本体 O,如果它不引用其他知识本体，则将知识本体转换为一阶谓词形式,构成其公理集合,检测一致性。

2．对于一个引用其他知识本体的知识本体 O,我们首先检测其所引用到的知识本体的一致性，对于每个被引用到的知识本体 O∈I，重复使用本转换检测方法，检查 O 的一致性。然后，在每个知识本体都一致的基础上,检测对应一阶谓词公式集合的一致性，如果这些一阶谓词公式集合都是一致的，可知本知识本体 O 是一致的。

3.4 确定语义相关度

可以根据概念兼容程度赋予相似值，如果概念*c*和是完全兼容的，即概念等价，则其相似度为1，如果不兼容的，则相似度为0。我们描述推理级相似性算法如下：

输入：概念*c*与c’

输出：概念*c*与c’的推理相似值

Step1：初始化Rsim = 0;关系评价值*RE*;

Step2：计算概念*c*与*c’*的间接概念环境*IContext(c)*和*IContext(c’);*

Step3：将间接概念环境根据HowNet[6]中的语义表示为语义原的并集，作为概念的逻辑表达;

Step4：Ssum = 0;

for si ∈ *IContext(c)*

for *sj* ∈ *IContext(c’)*

Switch 义原和的关系

Case Hyponym Ssum +=RE[0]; break;

Case Meronym Ssum +=RE[1]; break;

Case Holonym Ssum +=RE[2]; break;

Case Other Ssum +=RE[3]; break;

Step5：Rsim = Ssum / (|*IContext(c)*|*|*IContext(c’)*|);

Step6：return Rsim;

4 相关研究

在本体建立方面，目前存在的绝大多数本体都是手工生成的，该方法费时费力还容易出错，更难以维护和更新。由于网格上的信息量大，研究如何自动化、半自动化生成本体具有重大的意义。为此，研究者提出了本体演化这一涉及人工智能中信息获取、自然语言处理等多领域交叉的研究课题。Maedche等首先正式提出了本体进化的概念，并给出一个半自动化的需人工干预的本体演化框架，采用平衡的协作建模方式来构造语义Web中的本体，这个框架用半自动化的本体构造工具对典型的本体工程环境进行扩展，在这个框架中本体的建模周期由 5个步骤组成：本体引用、抽取、剪枝、精炼和评估，这个框架将能够为本体工程师提供丰富的本体协作建模工具。在本体建造方面，仍然有不少问题需要解决，如从纯文本和异类数据源中演化本体还停留在实验室阶段，实际应用仍然存在很多困难；关系抽取也是一个非常复杂和难以解决的问题，已经成为本体演化和应用的主要障碍。因此，在本体体现其在信息组织、管理和理解方面的优越性之前，还有大量的工作需要做。

5 结语

本体是对世界或者领域的概念化描述。我们通过研究，发现还有以下几个问题需要解决：首先，如何

建立本体还没有一个很好的理论支持,并且本体中的概念一般都是通过人工提取的,这使得基于本体的应用不能大规模开展,因此需要开发出能够自动或半自动提取概念的工具;其次,利用领域本体进行实体构建时,需要从文本中抽取出实体的属性,而本文中采用的基于规则和模板的信息抽取的方法并不能够很好的胜任这一工作,因此需要开发出基于语义的信息抽取工具。对这些问题都有待进一步的研究。本文的研究仅仅是一个起点，关于本体的自动构建，后续的工作还有很多，主要包括本体集成、本体映射和评价等。

参考文献

[1] **HELIOS: a general framework for ontology-based knowledge sharing and evolution in P2P systems**

Castano, S.; Ferrara, A.; Montanelli, S.; Zucchelli, D.;

Database and Expert Systems Applications, 2003. Proceedings. 14th International Workshop on

1-5 Sept. 2003 Page(s):597 - 603

[2] **Requirements for the visualisation of ontological evolution**

Blundell, B.; Pettifer, S.;Theory and Practice of Computer Graphics, 2004. Proceedings,2004 Page(s):18 - 23

[3]. G. Webb, J. Wells, and Z. Zheng, “An ExperimentalEvaluation of Integrating MachineLearning with Knowledge Acquisition,” *Machine Learning*, vol. 35, no. 1, 1999, pp.5–23.

[4]. www.cogsci.princeton.edu/~wn/

[5]. http://www.w3.org/TR/2004/REC-owl-ref-20040210/

[6] 知网. http://www.keenage.com. 2005.2.