第35卷 第15期 计 算 机 工 程 2009年8月
Vol.35 No.15 Computer Engineering August 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)15—0190—03 文献标识码：A 中图分类号：TP391

产品配置领域本体学习方法研究

邵伟平，郝永平，魏永合，曾鹏飞
(沈阳理工大学机械工程学院，沈阳 110168)

摘 要：针对大多数本体构建工具只支持手工构建，造成本体构建效率极低、工作量大、容易出错、知识的动态及时更新和维护困难等问题，提出一种领域本体自动构建的框架系统，通过对企业已有数据库及相关领域中大量的知识进行本体学习，实现配置领域本体自动(或半自动)构建，给出不同数据源结构中的本体概念抽取、概念间语义关系抽取等关键技术。

关键词：产品配置；本体学习；概念；关系

Research on Ontology Learning Method of Product Configuration Domain

SHAO Wei-ping, HAO Yong-ping, WEI Yong-he, ZENG Peng-fei
(College of Mechanic Engineering, Shenyang University of Science Technology, Shenyang 110168)

【Abstract】Most ontology building tools only support manual building ontology, which results in low efficiencies, highly workloads and many mistakes etc, And it is very difficult for knowledge updating and maintenance in time. In order to resolve the problem, a framework system of domain ontology automatic construction is proposed. Through ontology learning by using existed enterprise database and vast interrelated domain knowledge in Website, product configuration domain ontology is established. Key technologies of ontology learning are discussed such as domain concepts extraction and semantic relationships between concepts extraction in different data sources structures.

【Key words】product configuration; ontology learning; conception; relation

1 概述

本体是概念模型的明确的规范说明[1]，包含了4层含义：概念模型，明确，形式化和共享。本体作为表达知识的共享概念模型，已被广泛应用于知识工程、知识管理、智能信息集成、信息检索和语义Web、数字图书馆等多个领域。

产品配置技术是适应快速响应创新设计和变型设计的有效策略，基于本体的产品配置知识描述方法正逐渐成为解决领域内概念间语义歧义的热点方法，本体在产品配置领域内的应用越来越广泛。文献[2]用本体描述了虚拟供应网络配置问题；文献[3]将本体论用于产品配置知识表达中，提出了基于本体的产品配置知识共享、配置知识规则表达等方法及相关技术研究。这些文献中对产品配置本体的描述方法研究较多，而对产品配置本体的具体构建工具和构建方法则研究较少。

在近几年涌现的本体构建工具中，较为典型的有OntoEdit, KAON, OntoSaurus, WebOnto, Ontolingua[4]等，这些工具提供了友好的图形化界面和一致性检查机制，但大多数支持完全手工构建本体，而手工构建本体非常费时费力，容易出现倾向性错误，及时动态更新困难，使本体的构建成为一项艰巨而又繁琐的任务。因此，研究人员提出了自动或半自动构建本体的方法，即本体学习。目前本体学习需要解决的关键技术问题有：本体概念的自动抽取，概念间语义关系学习，领域分类体系的自动生成及本体评价等。本文将本体学习方法用于产品配置领域本体的自动构建，以提高本体构建的效率，减少差错，便于信息的及时更新。

2 产品配置领域本体描述

产品配置领域本体是在特定的产品配置领域中提供对该领域特定概念、术语、词汇的定义，并提供对概念之间关系的精确描述。实例是概念的具体化，领域本体还包括对配置领域中主要理论和基本原理的准确描述。随着企业产品的不断更新和发展，企业的知识库信息急剧增加，知识的动态更新频繁，概念及其关系也随之不断变化。如何在产品配置过程中从海量知识库中快速准确地搜索到相关知识，提取相关的本体概念及其关系，是产品配置本体自动构建的关键。本文结合水泵产品说明产品配置领域的知识本体表达方法。

配置领域本体可以用一个概念分类树表示，产品配置领域本体可描述为6元组：

$$PCDO=(C,R,A,H,S,X)$$

C表示产品配置领域中的概念集合，是树中的节点。如部件、功能、端口、资源、关系、各种约束等都是配置领域内的特定概念[5]。对于水泵产品来说，其特定领域本体包含动力机构、减速机构、执行机构、柱塞、液缸体、吸入阀、排出阀、调节丝杠等概念，每个概念都由一些属性变量来描

基金项目：国家自然科学基金资助项目(50875179)；辽宁省自然科学基金资助项目(20082047)
作者简介：邵伟平(1968－)，女，教授、博士，主研方向：产品配置管理；郝永平，教授、博士；魏永合，副教授、硕士；曾鹏飞，讲师、硕士
收稿日期：2009-01-10 **E-mail：**shaoweiping3008@sohu.com

述。$C=C^{\mathrm{I}}\cup C^{\mathrm{II}}$，由 2 个子集组成：$C^{\mathrm{I}}=\{c:\exists instance(c)\}$，表示非基类概念集合，即此类概念中有实例；$C^{\mathrm{II}}=\{c:\neg\exists instance(c)\}$，表示基类概念集合，即此类概念中无实例。

R 是一个关系集合，是树中的边，是概念之间存在的各种关系集合，R 与 C 是 2 个不相交的集合。概念间的关系有多种，如继承关系、部分与整体关系、同义关系、属性关系、概念实例、上下位关系、领域与作用域限制。

H 表示树中概念之间以及概念与实例之间的层次结构关系，一个概念层次 H 是一个偏序集 $(c,\prec)$，其中，c 是一个有限的概念集；$\prec$ 是 c 上的一个偏序。

S 是概念的实例集合，每个实例都有自己的属性集合。

A 是描述概念节点以及实例的数据属性集合，每个属性变量有相应的域。

X 表示公理集合，在产品配置领域中用于求解匹配规则集合、各种约束条件等。

配置领域概念树的结构和概念之间的各种关系十分复杂，为了简化问题的描述和配置本体的构建，假设概念树是一棵有根树，且其任意子树也是有根树。同一个实例 s 不能同时分配到不具备父子关系的 2 个概念之下。

本体学习的任务主要包括概念的获取、概念之间关系(包括分类关系和非分类关系)的获取以及公理的获取。

3 配置领域本体学习框架结构

本文引入知识挖掘和人工智能技术，提出了配置领域的本体学习的基本框架，见图 1。主要包括数据源预处理、候选关键词抽取、概念与概念间语义关系抽取、本体分类体系构建、本体构建及其修剪、本体评估与编辑等。

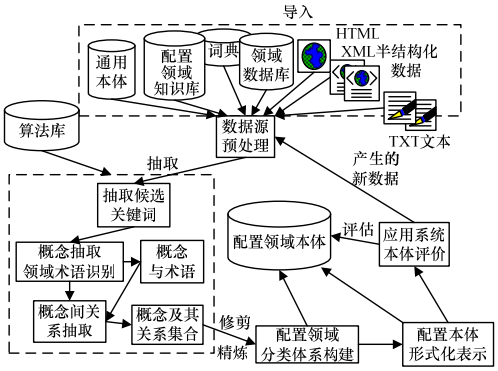

图 1　配置领域本体学习基本框架

3.1 配置域数据源导入及预处理

图 1 表明，本体学习工具的导入数据可以是各种类型的数据源，包括关系数据库或面向对象数据库中的数据等结构化数据、Web 中的大量纯文本文件等非结构化数据、大量的 XML 格式和 HTML 格式的网页等半结构化数据以及它们遵循的文档定义(XML schema 或 DTD)等。预处理模块对数据源进行数据清洗、模式转换、分词、特征化标注、词性标注等预处理，不同的数据源有不同的预处理方法，如词典、知识库或关系模式需要进行模式抽取转换，提取相关主题句，进行分词和词性标注。

3.2 概念及其关系抽取

概念及其关系的抽取阶段利用导入的本体，通过使用算法库中的各种本体学习算法从不同的数据源中抽取新的概念术语及其关系。概念的抽取包括候选关键词抽取、领域术语提取和概念定义。首先从预处理结果集中抽取概念候选词，运用停用词表和过滤规则库过滤非概念词语，选取适当的算法进一步确定领域术语概念，并发现概念定义。概念之间的关系类型很多，可通过相应的算法提取概念间的各种关系，从而得到概念及其关系集合。

3.3 本体修剪与精炼

根据概念之间的语义关系可以构建概念分类层次关系，并将领域概念进行分类组织，得到概念分类体系。在这个过程中，需要对抽取的概念及其关系进行修剪与精炼，利用给定的产品配置领域，移出与该领域不相关的概念和关系，仅保留与该领域密切相关的概念和关系，同时对目标本体进行精炼和调整。

3.4 领域本体形式化描述

通过本体学习所获得的领域本体、概念及其关系集需要用计算机可以理解和处理的语言描述，即形式化表示。常采用 OWL, RDF/S, DAML+OIL, OML 等多种本体描述语言对获取的配置本体进行表达。

3.5 领域本体应用评估

将本体学习得到的配置领域本体应用于产品配置设计应用系统中，通过实际应用评估其有效性。通常采用的评价指标有概念及概念间关系抽取的精确率、召回率和 F 因子等。通过评价和确认后，将最终的结果添加到本体库中。

通过本体学习得到的产品配置领域本体在实际的配置设计应用中又会产生新的知识，即新的概念和概念关系，并将这些新的数据返回到预处理模块中，从而使本体学习形成一个循环过程，使领域本体不断得到动态的及时更新、扩展和丰富。

4 概念及概念间关系抽取

4.1 概念抽取

配置领域中存在多种数据结构形式，根据不同类型的数据源结构，抽取概念时可采取不同的学习方法，这些概念是配置目标领域的核心概念。

4.1.1 基于非结构化数据的概念抽取

对于大量的 Web 文档，利用自然语言处理技术进行预处理，通过统计算法从文本中识别概念。传统本体学习方法一般都是基于“字词(word)”的，首先从文本中识别出这些关键词汇，其中较为典型的“独词词汇”术语较易被识别出来。而许多“多词组合词汇”术语是由这些单字词汇组合而成的，所产生的“多词组合词”术语不能作为正常的概念被提取出来，大多数抽取出来的概念只是“独词词汇”术语，因此，会丢失大量的概念。为此，采用了一种不同的概念抽取策略：首先直接从文本中归纳出“多词组合词”术语，如果这些术语频繁出现在“多词组合词汇”术语中或发现它们通过某种语义关系与“多词组合词”术语相关时就提取出“独词词汇”术语。这种学习策略减小了重要概念被丢失的可能性。这种方法主要根据领域概念与普通词汇所拥有的不同统计特征识别领域概念，如计算概念的领域相关性和领域通用性。

术语 t 在产品配置领域 CD_k 中的领域相关性计算如下：

$$CDR_{t,k}=\frac{P\left(t\left|CD_k\right.\right)}{\max\limits_{1\leqslant j\leqslant n}P\left(t\left|CD_j\right.\right)}$$

其中，条件概率 $P\left(t\left|CD_j\right.\right)$ 计算如下：$E\left(P\left(t\left|CD_k\right.\right)\right)=\dfrac{f_{t,j}}{\sum\limits_{i\in C_k}f_{i,k}}$；$f_{t,k}$ 是术语 t 在配置领域 CD_k 中出现的频率。

术语 t 的领域通用性指该术语在配置领域 CD_k 的分布应用，术语 t 在文档 $d \in CD_k$ 的分布可作为所有 $d \in CD_k$ 的随机变量估计。其通用性程度计算如下:

$$CDC_{t,k} = \sum_{d \in CD_k} \left(P_t(d) \lg \frac{1}{P_t(d)} \right)$$

其中，$E\left(P_t\left(d_j\right)\right) = \dfrac{f_{t,j}}{\sum\limits_{d_j \in CD_k} f_{t,j}}$。

对每一个术语 t 根据相关度排序，并计算其线性组合值:

$$T(t,k) = \alpha CDR_{t,k} + \beta CDC_{t,k}$$

其中，$\alpha, \beta \in [0,1]$ 是权重系数。具有较高 $T(t,k)$ 值的术语被选择添加到初始概念列表 T 中。

4.1.2 基于结构化数据的概念抽取

在产品配置领域中，企业中大量的产品数据都存储在数据库中，结构化数据主要包括关系数据库或面向对象数据库中的数据。因此，利用数据库中丰富的数据构建本体很重要。目前大多数的企业产品数据都采用关系型数据库，因为它采用关系模型，结构简单，二维关系表格形式容易理解。在关系模型中，实体及实体间的联系都用表来表示，因此，在概念及其概念间的关系抽取时，必须区分哪些表是用于描述实体的，哪些是用于描述实体间关系的，将实体映射为配置领域本体中的概念，将联系映射为本体中概念间的关系。本文采用文献[6]的方法进行本体学习，通过分析研究数据库中的表、属性、主外键和包含依赖关系，提出一组从关系模型到配置领域本体的映射规则集合，并基于这些规则得到一个候选本体，然后进一步对候选本体进行评价、修剪和精炼，得到最终的配置领域本体。

4.1.3 基于半结构化数据的概念抽取

半结构化数据是指具有隐含结构但缺乏严格或固定结构的数据，如大量的 XML, RDF 和 HTML 格式的网页。由于这类数据介于结构化数据和非结构化数据之间，因此可以采用基于上述 2 种数据类型的本体学习技术抽取该类的概念。

4.2 概念间语义关系抽取

概念间的各种语义关系是本体学习的重要内容，本文主要考虑概念的分类关系和非分类关系。分类关系是处于不同逻辑层次上的概念之间的关系，如继承关系、实例关系，如图 2 中概念蜗杆 3 是蜗杆的概念实例；而非分类关系反映对象组成结构的关系，如部分/整体关系、同义关系，如图 2 中执行机构是计量泵的组成部分，它们之间是部分/整体关系。

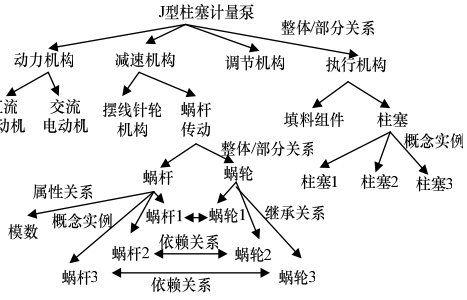

图 2 某 J 型计量泵产品概念及关系

目前，概念间语义关系的获取方法主要有基于模板的方法、基于关联规则的方法、基于概念聚类的方法、基于词典的方法等。本文根据机械产品配置的特点，采用基于概念聚类的方法和关联规则挖掘的方法相结合，进行概念语义关系的学习。概念聚类通过计算概念间的语义距离进行聚类，同一类中的概念具有语义近似的关系，概念层次聚类的结果就是概念间的分类关系。

非分类概念关系学习运用浅层文本处理方法识别概念对，并通过关联规则算法对概念对进行统计分析，挖掘概念间此类关系，然后计算概念对语义关系的支持度(support)和置信度(confidence)。如在水泵产品配置领域中，一个减速器集合 $RS := \{rs_i \mid i = 1,2,\cdots,n\}$，其中每一个减速器 rs_i 都由一个集合组成：$rs_i := \{a_{ij} \mid j = 1,2,\cdots,m_i, a_{ij} \in C\}$，集合中的每一项 a_{ij} 都是配置领域中概念集 C 里的元素。根据用户定义的极限，给定最小支持度和最小置信度，对超过最小支持度的关系集合进行语义关系的置信度计算。只要最小支持度和最小置信度设置合理，就可以有效消除错误关系对。

对于关联规则 $X_k \Rightarrow Y_k$, $X_k, Y_k \subset C, X_k \cap Y_k = \{\phi\}$，其支持度和置信度分别定义为

$$\sup(X_k \Rightarrow Y_k) = \frac{\left|\{rs_i \mid X_k \cup Y_k \subseteq rs_i\}\right|}{n}$$

$$\mathrm{conf}(X_k \Rightarrow Y_k) = \frac{\left|\{rs_i \mid X_k \cup Y_k \subseteq rs_i\}\right|}{\left|\{rs_i \mid X_k \subseteq rs_i\}\right|}$$

规则 $X_k \Rightarrow Y_k$ 的支持度定义为该减速器集合包含 $X_k \cup Y_k$ 作为其子集的百分比，而 $X_k \Rightarrow Y_k$ 的置信度则定义为当 X_k 出现在一个减速器里时，Y_k 也出现在减速器配置里的百分比。

5 结束语

配置领域本体的自动(或半自动)构建是实现产品快速配置的基本条件，在研究本体自动构建的理论和方法的基础上，讨论了企业中不同数据源结构环境下本体学习的方法，分别探讨了配置领域中各种概念的抽取、概念之间各种关系的抽取等关键技术和算法。同时提出了配置领域本体学习的基本框架及其实现的功能模块，克服了手工构建本体的缺点，为产品配置本体构建提供了一种新的方法。当然该方法还处于探索阶段，还有大量的工作要做。比如本体的构建要求对人类语言的深层次理解，而依靠浅层语言处理难以发现其深层次的关系。此外对学习所获得的结果本体需要进行评价，如与目标本体的一致性检查。

参考文献

[1] Gruber T. A Translation Approach to Portable Ontology Specifications[R]. Knowledge System Laboratory, Tech. Rep.: KSL 92-71, 1993.

[2] Alexander V S, Leonid B S, Nikolai C. Ontology-driven Approach to Constraint-based VSN Configuration[C]//Proc. of the 2nd World Conference on POM. Cancun, Mexico: [s. n.], 2004.

[3] 何陈棋, 谭建荣, 张树有. 基于本体论和知识规则的大批量定制配置设计技术研究[J]. 中国机械工程, 2004, 15(9): 783-788.

[4] Maedche A, Staab S. Ontology Learning[M]//Staab S, Studer R. Handbook on Ontologies. Berlin: Springer, 2004.

[5] 邵伟平, 刘永贤, 郝永平. 基于分布式约束满足的产品配置研究[J]. 东北大学学报: 自然科学版, 2007, 28(1): 103-106.

[6] Stojanovic L, Stojanovic N, Volz R. Migrating Data-intensive Web Sites into the Semantic Web[C]//Proc. of the 17th ACM Symp. on Applied Computing. New York, USA: ACM Press, 2002.

编辑 张正兴