第35卷 第12期 计 算 机 工 程 2009年6月
Vol.35 No.12 Computer Engineering June 2009
·开发研究与设计技术· 文章编号：1000—3428(2009)12—0239—03 文献标识码：A 中图分类号：TP301.6

结构化范例库向OWL本体自动映射算法

高俊杰，邓贵仕，张光前
(大连理工大学系统工程研究所，大连 116024)

摘 要： 在分析从关系数据库学习本体方法现状与不足，考虑到结构化范例库蕴含更丰富语义信息且拥有大量可复用的领域知识的基础上，提出一个低时间复杂度的结构化范例库向OWL本体自动映射的算法，并阐述该算法的流程。其优点在于不仅能获取范例库中蕴含的语义信息，而且将范例、规则知识项直接映射为对应的OWL个体，从而实现最大限度的知识复用。应用实例证实了该算法的有效性。

关键词： 结构化范例库；映射算法；OWL本体

Automatic Mapping Algorithm from Structural Case Bases to OWL Ontology

GAO Jun-jie, DENG Gui-shi, ZHANG Guang-qian
(Institute of Systems Engineering, Dalian University of Technology, Dalian 116024)

【Abstract】 On the basis of analyzing the status and shortage of existing methods about ontology learning from relational databases and considering the fact that structural case bases imply more semantic information and include vast domain knowledge, an automatic mapping algorithm from structural case bases to OWL ontology with low time-complexity is proposed and specified in this paper. Semantic information implied in case bases can be discovered and knowledge items about cases and rules stored in case bases can be mapped to individuals of OWL ontology directly to achieve domain knowledge reuse by the greatest extent. Validation of the algorithm is proved by an experiment.

【Key words】 structural case bases; mapping algorithm; OWL ontology

1 本体学习

由于完全手工构建本体十分耗时、费力，易出现倾向性错误、动态更新困难等问题，因此自动或半自动构建本体的方法——本体学习应运而生，并成为国内外研究热点。

当前，本体学习研究的数据源大致可分为文本、词典、知识库、半结构化数据模式和关系数据模式等。其中，已有的从关系数据库学习本体的解决方案大致包括：先将关系数据库模式翻译为一个中间模型，再由中间模型翻译成本体模型[1]；通过对主键、数据、属性的相关性分析，采用一定的映射规则完成向flogic本体的转换[2]；通过相应的一组学习规则从关系数据库模式信息提取本体的方法[3-4]。然而，这些方案存在以下不足：要么需要中间模型的参与，需要完成2次翻译，增加了实现的难度；要么仅在理论层研究相应的学习规则，过分强调表间较为复杂的主、外键关系(实际数据库中很少出现)向本体转换的规则；要么粗糙地分析关系数据库模式和本体间的对应关系，仅生成轻量级本体。更重要的是已有方法基本都是以一般的关系数据库为研究对象，仅从理论层面考虑关系模式中蕴含的领域知识的提取问题。

2 SCBR系统的特点

基于范例的推理作为人工智能的传统研究领域，有着大量研究成果。其中，结构化范例推理依靠事先定义好的属性和值来表述范例；在不同的系统中，属性可以被组织为表、一组关系表或者面向对象的形式。

在不须使用除范例以外的其他知识以获取好的结果的情况下，SCBR方法非常有用，已经成为应用最广、研究最多的主流[5]。由于关系数据库技术的成熟性和应用的广泛性，企业已有SCBR系统中范例被组织为关系表的形式最为常见，有大量的应用。分析发现：已有SCBR系统的开发包括了领域专家和开发者对领域内知识的分析、概括和组织管理过程，系统内蕴含大量的领域知识，并且以各自的形式存储、管理着大量的范例知识。因此，如果能通过分析已有SCBR系统中蕴含的更为丰富的语义信息，自动或半自动地构建本体，将大大简化本体应用系统的开发，并且能充分复用已有系统的范例知识。

基于以上分析，并考虑到在种类众多的本体描述语言中，W3C提出的最新标准OWL无疑是最有前景的。为此，本文提出一个结构化范例库向OWL本体自动映射的算法。

3 结构化范例库向OWL本体自动映射的算法

本研究的数据源设定为结构化范例库中蕴含于关系数据库的领域知识，映射目标是OWL本体；数据源基于关系模型，关系模式是型，元组集是值；而OWL本体是一种具有更多语义、结构更为复杂的模型。以下先分析算法的输入(数据源)、输出(映射目标)，进而确定其映射关系，然后详细阐述自动映射算法的具体流程。

3.1 输入、输出分析

输入信息包括关系数据库模式信息和对应元组集。其中，关系数据库模式信息主要包括基表结构设计和完整性约束申明2个部分。基表结构定义了关系(表)的结构、属性(列)及其

基金项目： 国家自然科学基金资助项目(70671016)
作者简介： 高俊杰(1978－)，男，博士研究生，主研方向：语义Web，知识管理；邓贵仕，教授、博士生导师；张光前，博士
收稿日期： 2008-12-15 **E-mail：** gaojunjie_7821@sina.com.cn

数据类型与长度等；完整性约束定义了语义施加在数据上的约束， 包括关系层的全局约束、元组层的表约束以及属性层的列约束。这 2 部分信息作为元数据存储于数据库的数据字典中。这里设定关系数据库模式信息包括一个名称集合以及一组约束集合。其中，名称集合包括表名集合、数据类型名集合、列名集合；约束集合包括主键约束、外键约束、取值范围(数据类型)约束、唯一性约束、非空约束。关系数据库模式信息的提取技术相对成熟，这里不再赘述。元组集是存储在关系表中的具体范例、规则知识项，可利用 SQL 语句操作、获得。

输出信息是一个OWL本体,其大部分元素是与类(class)、属性(property)、类的实例(instance)以及这些实例间的关系有关的。其中，领域内的基本概念被表示为拥有不同层次关系的各个类；属性(propertyies)可以断言关于类成员的一般事实以及关于个体的具体事实，其又被分为 2 种：数据类型属性(datatype property)和对象属性(object property)；数据类型属性表示类的实例与RDF文字或XML Schema数据类型之间的关系；对象属性表示 2 个类的实例之间的关系。类的一个实例表示相应概念的一个具体成员。个体公理(也被称为事实)是个体及相应属性值的描述，也用于表明类的成员关系。此外，为了更加详尽地说明属性，属性特性(characteristic)提供了一种强有力的机制以增强对于一个属性的推理，属性限制(restriction)可以进一步在一个明确的上下文中限制属性的值域，例如基数限制(cardinality)等。为了便于讨论，这里设定一个用 OWL 表示的本体，包含一个可选的本体标识符以及一组公理(axiom)。其中，本体标识符包括类标识符、对象属性标识符、数据类型属性标识符、个体标识符；公理包括类公理、属性公理以及个体公理。

通过分析并借鉴已有的相关研究成果，将关系表分为实体型表和联系型表 2 种(它们互不相交，实体型表拥有单主键，联系型表拥有复合主键)。进而确定算法输入-输出映射关系如下：一个实体型表映射为一个类标识符以及类公理；一个关系型表映射为 2 个对象属性标识符以及互逆的属性公理；一个非外键列及其取值范围约束映射为一个数据类型属性标识符以及一个属性公理；一个外键列及其外键约束映射为一个对象属性标识符及一个属性公理；数据类型名映射为数据类型标识符，每个数据类型标识符是 OWL 本体中使用的预定义的 XML Schema 数据类型标识符；每个唯一性约束、空值约束、主键约束映射为一个关于基数限制的属性约束；如果一个表的主键同时也是引用另一个表的外键，则映射为一个表述 2 个类是子类-父类关系类的公理;一个元组(范例、规则知识项)对应一个个体标识符及一组个体公理。

3.2 映射算法

自动映射算法的流程如图 1 所示。

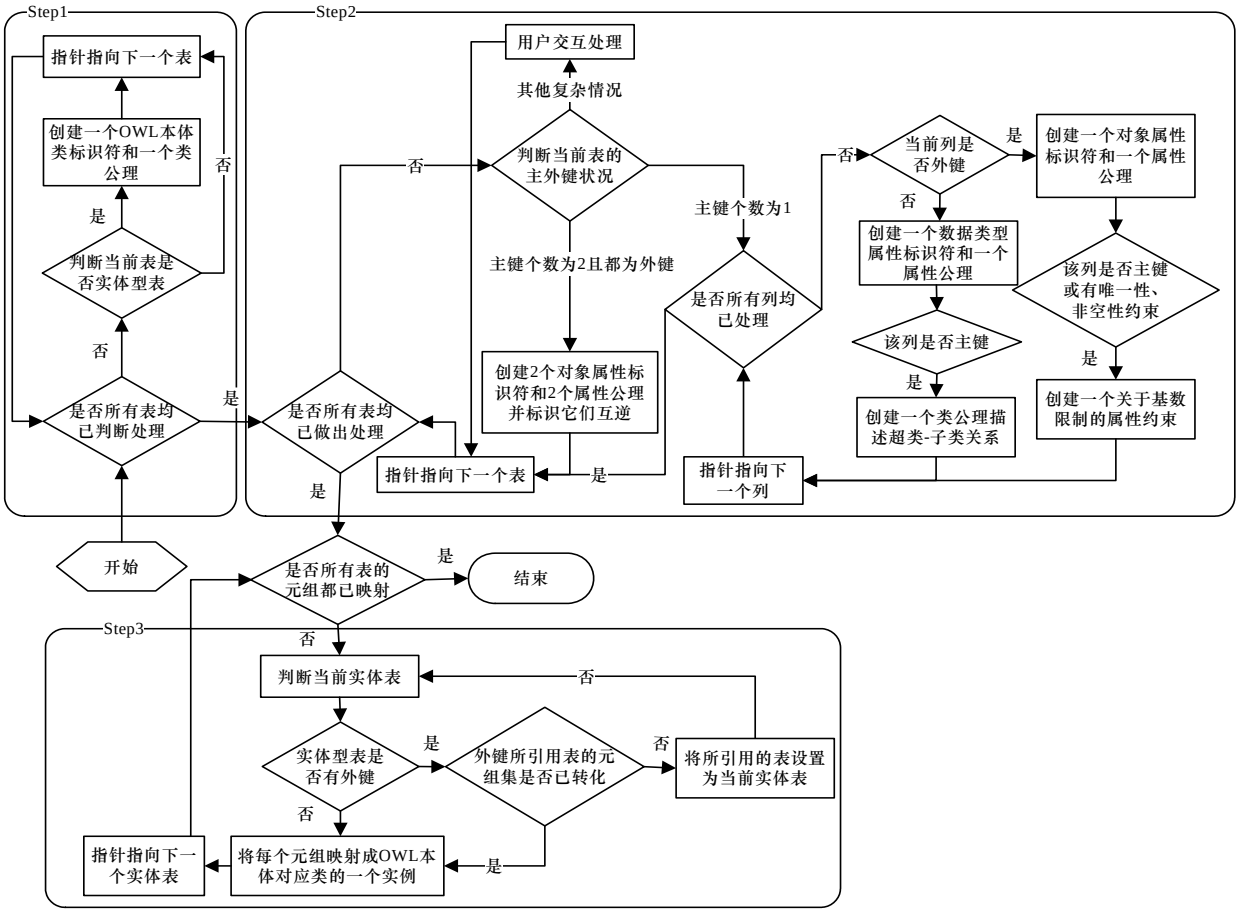

图1 映射算法流程

从功能上看，Step1, Step2 实现关系数据库模式信息向 OWL 本体对应元素的映射，Step3 实现体的范例、规则知识项向 OWL 本体对应元素的映射。以下详细阐述此映射算法：

Step1 逐个判断范例库中每个表是否为实体型表。如果当前表 T 是实体型表，则创建一个 OWL 本体类标识符和类公理。其中，OWL 本体类标识符与表名(T_{name})相同；类公理表述存在一个 OWL 本体类 T_{name}，如<owl: Class rdf: ID= "T_{name}">。

Step2 逐个判断所有表的主外键状况。如果主键个数为1，即为实体型表，则执行(1)操作；如果主键是个数为 2 的复合主键，并且这 2 个主键分别引用其他 2 个实体型表 T_k 和 T_j 的外键，则执行(2)操作；如果非以上经典状况，则为复杂语义结构表，通过交互，由领域专家确定具体处理方案。

(1)逐个判断该表 T_i 中每个列是否外键:1)如果该列(列名为 A)为外键，并且引用 T_j，则创建一个对象属性标识符及属性公理。其中，对象属性标识符为 has_当前列名(如 has_A)，属性公理表明该对象属性的定义域为表 T_i 对应的 OWL 类，值域为表 T_j 对应的 OWL 类。并且，如果该列也是主键，则创建一个类公理，用于描述 T_i 对应 OWL 类是 T_j 对应的 OWL 类的子类。2)如果该列非外键，则创建一个数据类型属性标识符及属性公理。其中，对象属性标识符为 has_当前列名，属性公理表明该对象属性的定义域为表 T_i 对应的 OWL 类，值域为当前列对应的数据类型。并且，如果该列为主键，则创建一个属性约束表明 Cardinality 为 1；如果该列有非空约束为真，则创建一个属性约束，表明 minCardinality 为 1；如果该列有唯一性约束为真，则创建一个属性约束，表明 maxCardinality 为 1。

(2)如果该表 T_i 为关系型表，即由 2 个实体之间 m:n 关系转化而来，则创建 2 个对象属性标识符以及 2 个属性公理，并且这 2 个对象属性互逆。其中，对象属性标识符分别为 has_T_{iname} 和 inv_has_T_{iname}，这 2 个属性公理的定义域和值域分别为 class(T_k)和 class(T_j)。属性公理分别表明 has_T_{iname} 的定义域为表 T_k 对应的 OWL 类，值域为表 T_j 对应的 OWL 类，inv_has_T_{iname} 的定义域为表 T_j 对应的 OWL 类，值域为表 T_k 对应的 OWL 类，has_T_{iname} 和 inv_has_T_{iname} 互逆。

Step3 逐个判断每个实体表是否有外键。如果该实体表无外键，则将每个元组映射成该表对应 OWL 本体类的一个实例(也就是一个 OWL 本体个体)，并将该表标识为已转化。如果该实体表有外键，则进一步判断外键所引用表的元组集是否已经映射为 OWL 个体。如果所引用表已经映射，则将每个元组映射成该表对应OWL本体类的一个实例，并将该表标识为已转化；如果所引用表未映射，则先映射引用表的元组集，再映射该表的元组集。

这里，将每个元组映射成该表对应 OWL 本体类的一个实例的过程就是一个 OWL 本体个体公理生成的过程。具体为：将每个元组的非外键列对应元组值映射为本体个体相应数据类型属性的值；将元组的外键列对应元组值映射为本体个体相应对象属性的值，从而描述 2 个个体间的关系。其中，个体标识符为该个体所属 OWL 类名_该元组的主键对应的元组值，如果该表有 m 个列，则每个元组生成的个体公理为 m 个。

此算法的时间性能可以作以下理论性分析。由于全部标识符创建可以在公理创建中直接进行，因此可认为算法的基本操作为公理的创建。设一个范例库的规模 $N=N_T+N_A+N_i$。其中，N_T 是表的个数；N_A 是所有列的个数；N_i 是所有元组的个数。分析此算法，在极端的情况下，数据库中全为实体表，则第 1 步创建类公理次数最多为 N_T 次；第 2 步分为实体型表和关系型表 2 个部分，对于实体型表部分公理创建次数最多为 $4N_A$，对于关系型表部分公理创建次数不多于 N_T；第 3 步个体公理创建次数不多于 $N_A\times N_i$。所以，最坏情况下，算法的基本操作总执行次数 $T=N_T+4N_A+N_T+N_A\times N_i<N^2$，故算法的时间复杂度低于 $O(N^2)$。

4 算法应用实例

映射算法在J2SE1.4.2 平台上实现，通过调用Jena工具包中的API(主要是com.hp.hpl.jena.ontology包中的接口及相关方法)实现对OWL本体的操作。具体为：先通过ModeFactory接口的createOntologyModel方法建立一个使用OWL语言规则的本体模型OntModel。然后，通过OntModel接口的众多方法创建类、属性、个体等。这些方法的使用方面主要包括：通过 createClass 方法创建相应的 OWL 本体类；通过 createObjectProperty, createDatatypeProperty方法创建相应的OWL本体对象属性和数据类型属性，通过OntProperty接口的addDomain和addRange方法创建属性的定义域、值域；通过 createIndividual 方法创建相应的OWL本体个体；通过 createCardinalityRestriction, createMinCardinalityRestric tion, createMaxCardinalityRestriction方法设置基数约束。

为了验证本算法的有效性，从实际的工装工时定额领域的结构化范例库中自动映射 OWL 本体。该范例库中主要存储 2 种类型的知识——工装范例知识和工装特征规则知识，以关系表形式存储于关系数据库中。其中，工装范例知识是工时定额工程师以往定额完毕，并被事实检验为正确的事实知识。每个工装范例信息包括范例特征的定性描述信息(如加工零件类型)、范例特征的定量描述信息(如模芯尺寸信息)、对应于基准描述的比例系数、定额工时值。工装特征规则知识是工时定额专家总结的规则，主要用于为结构化范例推理提供额外的领域知识支持，包括各种类型工装特征信息的一些典型描述，及其与对应基准特征描述相比影响工时定额结果的比例数值。

为验证算法的有效性，先后对范例库中 5 组范例及其对应规则表进行自动映射，结果显示算法实际运行时间和问题规模 N 的关系与理论分析吻合(优于 N^2 曲线)。

现以略为简化的锻切模范例及其对应的 5 个规则表为例阐述其实体-关系图，如图 2 所示。

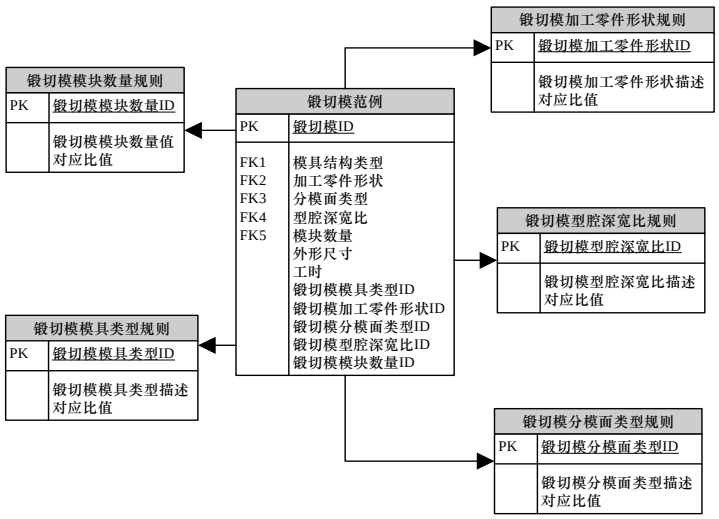

图 2 示例的实体关系图

在映射实例中，锻切模范例表及其对应模具结构类型、加工零件类型、型腔深宽比、分模面类型、模块数量特征规则表中相应的记录为 11, 3, 5, 6, 3, 4。通过映射算法，映射结果如图 3 所示。

公理组成如下：6 条类公理；17 条数据类型属性公理；5 条对象属性公理；22 条基数限制；140 条个体公理。

(下转第 244 页)