基于OWL DL的关系数据知识提取研究

张国强 贾素玲 王 强
(北京航空航天大学经济管理学院 北京 100191)

摘 要 关系型数据是企业的重要资源,针对当前关系数据库中海量数据的知识提取转换问题,对如何将关系型数据提取表示为OWL DL本体进行了研究。在形式化关系数据记录和本体模型的基础上,设计了转换算法来提取关系数据并将其转换为本体形式,对其中的标识和关联等关键问题进行了重点研究,最后以实例验证展示算法。该方法的提出可以使本体开发者更容易地将关系数据提取为知识,为快速构建企业知识本体提供了新的途径。

关键词 本体,OWL DL,关系数据,知识提取
中图法分类号 TP301 **文献标识码** A

Ontology-based Knowledge Extraction of Relational Data

ZHANG Guo-qiang JIA Su-ling WANG Qiang
(School of Economics and Management, Beihang University, Beijing 100191, China)

Abstract For converse the massive relational data to knowledge, the paper studied how to extract relational data to construct a OWL DL ontology. Based on the formalization of relational data records and knowledge ontology, an algorithm was designed to realize the extraction and conversion. Some key issues such as the identity and relationship table were studied. Finally a case was designed to test and verify the algorithm. The proposed method can help ontology developers extract the relational data into knowledge more easily, and provide a rapid way to build ontology for the realization of knowledge management.

Keywords Ontology, OWL DL, Relational data, Knowledge extraction

1 引言

自语义Web提出以来,本体(ontology)已成为人工智能和知识工程中一种重要的工具。通过本体支持,用户和系统可以在共同的领域知识下互相交互。尽管学术界和企业界均把本体作为知识展示的一种重要工具,但到目前为止,本体的创建还远没有达到系统化和自动化的程度。一些通用的本体如WordNet,Cyc和知网等虽然可以直接使用,但是在更多的实际应用中还需要建立某一特定领域的本体知识来描述领域中的概念和关系。由于缺乏结构化的领域知识,传统的本体开发更多依赖于领域专家。这种创建方式虽然知识准确,但耗时长,工作量大而且极为繁琐和枯燥,容易出错。虽然有很多工具如Protégé,OntoEdit等支持本体开发,但它们均需要大量的人工参与[1]。

随着信息化建设的不断进步,很多企业和组织先后投入巨资建立了各种生产自动化控制系统、经营管理信息系统和信息发布系统。这些系统在企业的运营中积累了海量的数据,且绝大部分数据均以关系数据的方式保存。如果能够直接将这些关系数据转换为本体知识,则可以避免大量的人工工作,更为方便切实地建立相关应用本体,将业务数据转换为企业的显性知识。

虽然企业的信息系统中保存大量的数据,隐含丰富的信息,但这并不等同于知识。首先二者理论基础不同,关系数据库中的数据均以关系的形式进行表示,是基于封闭世界的假设,而知识本体基于描述逻辑相关理论,是基于开放世界的假设,即对事实的每一个方面都需要进行概念化明确的说明。其次,目前学术界就如何进行知识产生和知识推理进行了大量的研究,这些研究的基础便是基于描述逻辑的各种语义模型,而关系模型并不是基于语义的,无法直接对其进行逻辑描述和推理,成为知识产生和发现的一个障碍。因此如何将结构化的关系数据转换成本体知识已成为领域专家关注和研究的热点。

本文第2节对本体建模语言及其建模语言进行了简要介绍,并对要建立的知识本体进行形式化定义;第3节在现有ER模式转换的基础上,研究如何将关系数据提取转换为基于OWL的知识本体,并引入相应的解决算法;第4节设计了一个简单的数据模型实例对算法进行演示验证;最后进行总结。

2 知识本体建模与OWL DL

本体(ontology)最著名并被广泛引用的定义是由Gruber

到稿日期:2009-04-28 返修日期:2009-07-02 本文受国家科技支撑计划项目(2006BAG01A05),教育部人文社会科学研究基金(06JD6300001),航空基金(2007ZG51078)资助。

张国强(1982--),男,博士生,主要研究方向为管理信息系统,E-mail:zhanggq@163.com;贾素玲(1954-),女,教授,博士生导师,主要研究方向为管理信息系统等;王 强(1966-),男,副教授,主要研究方向为管理信息系统等。

提出的"本体是概念模型的明确的规范说明"[2]。通俗地讲，本体用来描述某个领域甚至更广范围内的概念以及概念之间的关系，使得这些概念和关系在共享的范围内具有大家共同认可的、明确的、唯一的定义。通过本体的建立可提高异构系统和不同参与者之间的互操作性，以促进知识共享。本体描述现实世界的能力十分强大，它既可以用来描述简单的事实，又可以用来描述信念、假设、预测等抽象的概念；既可以描述静态的实体，又可以描述与时间推移相关的概念，如事件、活动、过程等。本体的表示有多种：有仅表示概念的简单表示，有表示概念、属性的框架和语义网络表示，还有能表达丰富语义的逻辑表示。目前本体已经被广泛应用于语义 Web、智能信息检索、信息集成、数字图书馆等领域。

本体的表达需要一定的描述语言支持，其中 OWL(Web Ontology Language)是 W3C 最新推荐的本体描述语言标准，它是 W3C 在已有的 DAML+OIL 基础上改进得来的，目前已经成为国际通用的标准语义 Web 语言[3]。OWL 提供了 3 种表达能力递增的子语言：OWL Lite，OWL DL 和 OWL FULL。OWL Lite 表达能力最弱，用于提供给那些只需要一个分类层次和简单约束的用户。OWL DL (DL 表示描述逻辑)支持那些需要最强表达能力的推理系统的用户，且这个推理系统能够保证计算的完全性和可判定性，它包括了 OWL 语言的所有成份，但有一定的限制。OWL FULL 包括了 OWL 语言的所有成份，并且取消了 OWL DL 中的限制，支持那些尽管没有可计算性保证，但有最强的表达能力和完全自由的 RDF 语法的用户[4,5]。OWL DL 语义是基于描述逻辑的，描述逻辑是一种基于对象的知识表示的形式化，也叫概念表示语言或术语逻辑。它是一阶逻辑的一个可判定的子集，具有合适定义的语义，并且具有很强的表达能力。一个描述逻辑系统包含 4 个基本组成部分：表示概念和关系的构造集；TBox(Terminology Box)术语断言；ABox(Assertional Box)实例断言；TBox 和 ABox 上的推理机制。一个描述逻辑系统的表示能力和推理能力取决于对以上几个要素的选择以及不同的假设[6]。描述逻辑中有两个基本元素，即概念和关系(Role)。概念解释为一个领域的子集；关系则表示在领域中个体之间所具有的相互关系，是在领域集合上的一种二元关系。为了更清晰地描述关系数据的本体表示，首先需要对建立的本体模型进行形式化定义。

定义 1 本体定义为一个三元组 $O=\langle C,A,I\rangle$。其中：C 表示本体概念构造集，其中的元素称为概念(concept)，包括类概念 C_c，属性概念 C_p，因此概念可以用一个二元组表示为 $C=\langle C_c,C_p\rangle$；A 是一个包含断言的有限集合，也称为术语公理的集合，它是一个描述领域结构的公理集，对应于描述逻辑的 TBox。A 中的元素称为公理(Axiom)，其一般形式为 $C\sqsubseteq D$，C 和 D 都是概念。公理包括类公理 A_c，类约束 A_r，属性公理 A_p，公理可以用一个三元组表示：$A=\langle A_c,A_r,A_p\rangle$；$I$ 代表实例(Instance)，其中又包含实例标识(Instance Identify，ID)和属性值(property value，IV)，$I=(ID,IV)$。

OWL DL 具有确保计算完备性和可判定性的最大表达力，它有两种语法：①交换语法(exchange syntax)，即 RDF/XML 语法，以一组 RDF 三元组(triples)的 XML 序列化格式来表示一个本体，以便在 Web 上发布和共享本体；②抽象语法(abstract syntax)，即框架风格的语法，其中，关于一个类或属性的一组信息用一个尽可能大的、类似于自然语言语法的构造子来表示，便于用户理解和评价一个本体。同一本体使用不同语法时，具有相同的形式语义[5]。为了便于理解，文中采用抽象语法进行问题说明。

3 关系数据知识提取算法研究

3.1 模式转换

目前对关系数据的知识表示已经有一定的研究，主要成果为对 ER 模式的描述逻辑的表示和转换映射，如 Calvanese 等人提出的 ER 模式的一阶形式化定义[7]和语义转换，许卓明教授提出的 ER 模式本体保值翻译算法[8]。这些研究在将数据库概念模型转换为本体模型方面取得了重要成果，为具体物理实施的数据库与数据的本体表示起到了重要的基础作用。本文以笔者提出的 ER 模式本体转换算法为基础(具体如表 1 所列)，对关系数据的本体表示进行研究[9]。

表 1 ER 模式转换算法

ER 模式元素	转换后
E-R 模式 $S=(ls,isas,atts,rels,cards)$	本体类公理模型 $O=\langle C,A\rangle$
实体 $E_i\in E_s$	Class($\phi(E_i)$)
实体对 $E_1,E_2\in E_s$，且 $E_1\ isa_s E_2$	SubClassOf($\phi(E_1)\phi(E_2)$)
域 $D\in D_s$	XML Shema 类型映射或 EnumeratedClass($\phi(D)$)；
属性 $A\in A_s$	Class($\phi(E)$ partial restriction($\phi(A_i)$ allValuesFrom ($\phi(D_i)$) cardinality (1)))； DatatypeProperty($\phi(A_i)$ domain ($\phi(E_i)$) range ($\phi(D_i)$) [Functional]
联系 $R\in R_s$	Class($\phi(R)$ partial restriction($\phi(U_1)$ allValuesFrom ($\phi(E_1)$) cardinality (1)) … restriction ($\phi(U_k)$ allValuesFrom($\phi(E_k)$) cardinality (1))) ObjectProperty($\phi(U_i)$ domain ($\phi(E_i)$) range ($\phi(R)$))；
角色 $U\in U_s$	Class($\phi(U_i)$ domain($\phi(R_i)$) range($\phi(E_i)$))； Class($\phi(E_i)$ partial restriction ($\phi(U_i)$ minCardinality (m)))； Class($\phi(E_i)$ partial restriction ($\phi(U_i)$ maxCardinality(n)))。
不相关概念 X_1 X_2	类公理 DisjointClasses(X_1 X_2)

通过表 1 所展示的算法，可以将 ER 模型各元素映射为知识本体概念公理，以解决模式转换问题，为下一步实现关系数据的提取和知识表达提供基础。

3.2 关系数据定义

从上面的映射关系，可以看到概念模型 ER 模式与本体元素和 OWL 表示间的对应关系，但 ER 模式并不等同于物理数据库，二者之间还存在着一定的差别。实际的物理数据涉及各方面的内容，包括服务器、操作系统、数据库管理系统、模式、表空间、关系、索引、触发器、存储过程等，此外出于对物理存储和检索效率的考虑，实际关系模式也会与原有概念模型有一定差别，如将一张表按照时间或者属性性质的不同横向或者纵向进行拆分，主键的采用考虑检索或者其他原因不采用业务主键等。因此对概念模型的转换并不等同于关系数据的转换。由于要实际处理的对象为业务数据记录，因此主要考虑与业务数据相关的内容，在此，先对数据库记录进行如下定义。

定义 2 关系数据库记录可表示为 $R=(RV,RR)$。其中 RV 由一组具体数值构成，代表该记录中非关联字段的属性值，$RR=[U_1:R_1,\cdots,U_k:R_k]$ 代表该记录与其他表的关联关

系，其中 U_i，R_i 分别表示该关联的角色与实体记录，这种关联关系在数据库中体现为主外键关系。

3.3 知识提取

建立如上的数据记录形式化定义后，接下来讨论如何提取记录并进行知识表示。

3.3.1 实例标识

对记录 R 进行转换，首先需要对记录在知识本体中的对应实例进行标识，转换为知识本体标识 $I=\varphi(R)$。数据库中记录以表的形式保存，主键也按照表进行唯一定义，这样不同表中的记录可能存在值相同的主键，因此直接采用主键作为实例标识违反了 OWL DL 标识唯一的约束，在提取记录时需要对标识进行重新定义。标识定义最基本的原则就是在本体的语义表达范围内取值唯一，因此采用何种方式进行编码可以根据企业现有的数据编码标准和实际需要来确定，这方面高复先教授曾有很好的表述[10]。本文中采用“模式名＋表名＋表内主键值”的方法进行定义，主键为组合键时把各键值字符串相连。

3.3.2 无关联表及主键

无关联的独立表数据处理较为简单，表模式已经通过表 1 算法进行了转换，因此这里直接将数据作为实例进行处理即可。表转换为类公理 Class($\phi(E)$)，则每一个实例可映射为：

Individual ($\varphi(R)$ type ($\phi(E)$) value ($\phi(A_1)RV_1$)…value ($\phi(A_n)RV_n$))

其中，$\phi(A_i)$表示属性 A_i 转换后的类公理，RV_i 代表该属性的值。

3.3.3 关联表

关系数据库最大的特点便是其中各类关联，因此将关联关系提取并转换是整个知识提取的重点。实际的数据库关联包含 1∶1，1∶N，M∶N 3 种类型，其中 M∶N 关系采用关联表来实现，因此其映射最为直接。

对关联表记录 R 创建实例（与算法中联系映射部分对应）：

Individual($\varphi(R)$ type($\phi(E)$) values($\phi(A_i)RV_i$)…
value($\phi(U_i)\varphi(R_i)$)…)

其中，$\phi(E)$代表该关联表对应类公理，$\varphi(R_i)$代表该关联字段也就是外键字段所对应的类公理实例。

1∶1 和 1∶N 的关系类型为 M∶N 关系的特殊情况，它们的关联不是以关联表的形式体现，而是将关联建立在子表一方，如果要对这类记录进行知识提取，则首先要对其关联进行标准化。

为了讨论方便，假定记录 R 只包含一个外键，则可以表示为 $R=(RV,U_1:I_1)$，对 R 进行标准化：

$R'=(IV)$

$RR'=(U_1:I_1,U_2:\varphi(R'))$

其中，R'代表原记录中的一般属性，RR'代表将原父子关系提出后并进行完整描述的关联。如此转换之后，就可以得到该记录的 OWL 本体描述：

Individual($\varphi(R')$ type($\phi(R')$) value($\phi(A_i)IV_i$)…)

Individual($\varphi(RR')$ type($\phi(RR')$) value($\phi(R_i)\varphi(R')$)…)

通过如上算法的转换，可以对关系数据库中的业务数据记录进行提取，并将其转换为知识本体。这样可以直接方便地将企业已有数据迅速构建为知识本体原型，为企业管理者构建出一个相对完整的业务信息框架体系和信息实体内容，同时也为下一步本体的丰富和推理提供知识基础。

4 算法演示

为了对算法进行清晰的展示，设计了一个简单的关系数据库，对其进行知识提取。数据库结构如图 1 所示，此模型为学校教师教学的一个数据模型，共有 5 个实体，学校人员（SchoolStaff）、教师（Teacher）、课程（Course）、学生（Student）、论文（paper）；其中教师和学生属于学校人员的一种，教师教授课程，学生选修课程，教师撰写论文（为了演示考虑，这里只有教师可以撰写论文且每篇论文只有一个作者）。每一个实体和联系均有自由的属性，具体实体关系及实体属性如图 1 所示。

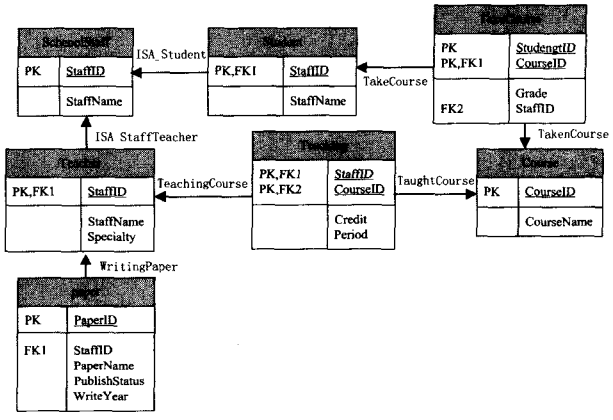

图 1 教学数据库物理模型

对教师教学和撰写论文进行分析，通过模式转换，获得教师实体的概念。

SchoolStaff 类公理：

Classs(“SchoolStaff” partial restriction(“StaffID”
allValuesFrom(“&xsd;nonNegativeInteger”)
cardinality(1) “StaffName”
allValuesFrom(“&xsd;nonNegativeInteger”)
cardinality(1)))

Teacher 为 SchoolStaff 的子类：

SubClassof(“Teacher” “SchoolStaff”)

Teaching 类公理：

Class (“Teaching” partial restriction(“TeachingCourse”
allValuesFrom(“Teacher”) cardinality(1)) restriction
(“TaughtCourse”) allValuesFrom(“Course”) cardinality
(1)))……

Course 类公理：

Classs(“Course” partial restriction(“CourseID”
allValuesFrom(“&xsd;nonNegativeInteger”)
cardinality(1) “CourseName”
allValuesFrom(“&xsd;nonNegativeInteger”)
cardinality(1)))

Paper 类公理：

Classs(“Paper” partial restriction(“PaperID”
allValuesFrom(“&xsd;nonNegativeInteger”)
cardinality(1) “PaperName”
allValuesFrom(“&xsd;nonNegativeInteger”)

（下转第 164 页）

[R]. Dept. of Computer Science Technical Report 8805. University of NSW, Kensington, Australasian, 1988

[12] Liu Chao, Chen Chen, Han Jia-wei, et al. GPLAG: Detection of Software Plagiarism by Program Dependence Graph Analysis [C]// ACM SIGKDD'06. 2006: 872-881

[13] West A. Coping with plagiarism in Computer Science teaching laboratories[C]// Computers in Teaching Conference. Dublin, July 1995

[14] Gitchell D, Tran N. Sim: A Utility for Detecting Similarity in Computer Programs[C]// Proceedings 30th SIGCSE Technical Symposium. New Orleans, LA, USA, March 1999

[15] Ji Jeong-Hoon, Woo Gyun, Cho Hwan-Gue. A Source Code Linearization Technique for Detecting Plagiarized Programs[C]// ITiCSE'07. United Kingdom, June 2007

[16] Arwin C, Tahaghoghi S M M. Plagiarism Detection across Programming Languages[C]// Twenty-Ninth Australasian Computer Science Conference (ACSC2006). Australia, January 2006

[17] 赵长海，晏海华，金茂忠. 一个基于编译优化和反汇编的程序相似性检测方法[J]. 北京航空航天大学学报，2008，34(6)：711-715

[18] 边肇祺，张学工. 模式识别(第二版)[M]. 北京：清华大学出版社，2000

[19] Thonmas，等. 算法导论(第二版)[M]. 潘金贵，等译. 北京：机械工业出版社，2006

[20] Theodoridis S, Koutroumbas K. Pattern Recognition (Third Edition)[M]. Beijing: China Machine Press, 2006: 214-215

[21] 孟庆磊，姚春莲，宋建斌. 一种面向 H.264/AVC 的快速帧内预测选择算法[J]. 北京航空航天大学学报，2007，33(2)：119-223

(上接第 151 页)

cardinality(1))…)

其中 Paper 中包含 1：*N* 的关联，该关联的类公理为：

Class ("WritePaper" partial restriction ("WritingPaper" allValuesFrom("Teacher") cardinality(1)) restriction ("WritenPaper") allValuesFrom("Paper") cardinality (1)))……

完成以上工作后，我们建立了模型的本体类公理模型，接下来将对该数据库中的数据记录进行抽取，并将其转换为本体实例。

对教师、课程实体分别抽取实例：

Individual(dboTeacher1 type(Teacher) value(StaffName "Barclay") value(StaffID 1)…)

Individual(dboCourse0 type(Course) value(CourseName "Philosophy") value(CourseID 0)…)

Teaching 表记录的实例为：

Individual (dboTeaching01 type(Teaching) value(TeachingCourse dboTeacher1) value(TaughtCourse dboCourse0) value (Credit "20") value(Period "0"))

对 Paper 表中的记录，按照 1：N 关联表记录的转换算法进行知识提取：

Individual (dboPaper0 type(Paper) value(PaperID "0") value (PaperName "ontologie review"))

Individual (dboWritePaper01 type(WritePaper) value(WritingPaper dboTeacher1) value(WritenPaper dboPaper0))

结束语 关系数据的知识化是实现企业知识管理的重要部分。本文在形式化数据库记录模型和本体模型的基础上，研究了如何将关系数据转换为知识本体，提出了相应的信息提取算法，对其中的关键问题如标识和关联进行了重点研究，最后以实例对转换方法进行了展示。该算法生成的本体模型遵守 W3C 定义的 OWL DL 语法规则，因此具有良好的机器可读性和人机互操作性。本算法的提出可以使本体开发者更容易地将关系数据提取成为知识，为实现企业的知识管理提供了一种迅速构建本体的途径。同时，基于业务数据知识本体的创建也为企业下一步的知识进化及推理建立了基础。

参 考 文 献

[1] Lee Chang-shing, Kao Yuan-fang, Kuo Yau-hwang, et al. Automated ontology construction for unstructured text documents [J]. Data & Knowledge Engineering, 2007, 60(3): 547-566

[2] Neches R, Fikes R E, Gruber T R, et al. Enabling Technology for Knowledge Sharing[J]. AI Magazing, 1991, 12(3): 36-56

[3] McGuinness D L, van Harmelen F. OWL Web Ontology Language Overview. W3C Recommendation 20040210[OL]. http://www.w3.org/TR/owl-features/

[4] Dean M, Schreiber G. OWL Web ontology language reference. W3C Recommendation. 20040210[OL]. http://www.w3.org/TR /owl2ref/

[5] Patel-Schneider P F, Hayes P, Horrocks I, et al. OWL Web ontology language semantics and abstract syntax. W3C Recommendation. 2004202210[OL]. http://www.w3.org/TR/2004/REC-owl-semantics-20040210/

[6] Borst W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[D]. University of Twente, Enschede, 1997

[7] Calvanese D, Lenzerini M, Nardi D. Unifying class-based representation formalisms[J]. Journal of Artificial Intelligence Research (JAIR), 1999, 11: 199-240

[8] 许卓明，董逸生，陆阳. 从 ER 模式到 OWL DL 本体的语义保持的翻译[J]. 计算机学报，2006，29(10)：1786-1795

[9] Zhang Guoqiang, Jia Suling. Ontology-based knowledge extraction for relational database schema[C]// 2009 Second International Symposium on Electronic Commerce and Security. 2009: 585-589

[10] 高复先. 信息资源规划——信息化建设基础工程[M]. 北京：清华大学出版社，2002

[11] Kingston J. Multi-perspective ontologies: Resolving common ontology development problems[J]. Expert Systems with Applications, 2008, 34(1): 541-550

[12] Chandra C A T. Organization and problem ontology for supply chain information support system [J]. Data & Knowledge Engineering, 2007, 61(2): 263-280

[13] Weng S-S, Tsai Hsine-Jen, Liu S-C, et al. Ontology construction for information classification[J]. Expert Systems with Applications, 2006, 31(1): 1-12