基于结构化信息源的本体构建方法综述*

车成逸，马宗民†，焦晓龙
（东北大学 信息科学与工程学院，沈阳 110819）

摘　要：作为一种能够在语义层和知识层上描述信息系统的概念建模工具，本体在许多领域得到了广泛应用。由于本体的构建和维护工作费时费力，本体的构建方法研究成为了实现语义 Web 应用的最重要技术。综述了从不同的结构化信息源（数据库、XML 文档以及 Web 表格）构建本体的方法，进行了详细分析与对比，并给出其存在的不足之处以及未来可能的研究方向。

关键词：本体构建；结构化信息源；数据库；XML 文档；Web 表格
中图分类号：TP391.13　　**文献标志码**：A　　**文章编号**：1001-3695(2012)07-2406-05
doi:10.3969/j.issn.1001-3695.2012.07.02

Survey on methodology for constructing ontology based on structured information source

CHA Song-il, MA Zong-min, JIAO Xiao-long
(*College of Information Science & Engineering, Northeastern University, Shenyang* 110819, *China*)

Abstract: Ontology is a tool for modeling concept of information systems on the semantic and knowledge layer, and is widely employed in many areas. Since building and maintenance of ontologies must cost much time and energy, research on ontology building method becomes the most important technology for realizing semantic Web applications. This paper surveyed the methodology for constructing ontologies from different structured information sources (database, XML document, Web table), analyzed and compared with each other. And, it pointed out some remaining defects and future research directions.

Key words: ontology construction; structured information; database; XML document; Web table

0　引言

本体(ontology)是用来描述广泛范围内或某个领域内的概念以及概念之间的关系，使得这些概念和联系在共享的范围内有着明确唯一的定义，达成一种共识，这样人机就可以进行交流[1]。随着本体在人工智能、信息检索以及知识管理等研究领域中的应用不断扩展，人们对本体的要求也越来越多。近年来，研究者们利用本体思想从不同角度对信息集合进行标引，表示信息内容与知识组织体系之间的链接关系，可以将本体与信息系统进行链接，从而使用户在使用信息的过程中更加便捷地浏览和理解相关概念和资源，还可以利用本体中的语义关系及推理规则集合进行推理，从而实现基于本体的智能分析和知识组织，并通过智能分析来预测知识增长点。目前已经有很多建成并在实践中应用的本体，这些本体在不同的领域取得了不同程度的成功。

本体的应用建立在本体的构建基础上，优质的本体构建正成为本体开发与应用的一个瓶颈。本体构建方法直接影响所构建本体的质量及效率，规定本体在特定应用或特定环境中的性能和适用性，影响本体能否在语义网中大规模应用。本体构建是一项花费巨大的工程，因此，越来越多的研究者们开始重视和研究最适合的本体构建方法。由于本体工程到目前为止仍处于相对不成熟的阶段，每一个工程拥有自己独立的方法。结构化数据作为巨大的数据源，从结构化数据中提取知识是当前本体研究的热点之一。

本文从基于结构化数据源本体自动构建方法论的角度出发，总结分析了现有领域本体的构建技术与方法，并在此基础上展望了本体构建未来可能的研究方向。

1　结构化信息源的本体创建方法综述

本体的质量贯穿从本体构建到应用的整个生命周期，如何选用合适的数据源及挖掘手段快速获取高质量本体是当前和未来一段时间内的重要研究领域。从文本（包括 Web 文本）数据资源中抽取信息，需要进行词法分析和句法分析，目前该领域仍然需要更加深入的研究，同时由于存在着自然语言固有的模糊性，所以抽取本体的准确性不高。当前 Web 上最大的数据资源是以结构化的形式展示的，基于结构化数据来源获取本体的方法是一种能很好地解决低准确性弊端的方法。Web 上的数据库、XML 文档和表格数据源结构化程度有所不同，但都是能以简便、快捷的方式构建高质量本体的数据源。以下分别针对从三个数据源构建方法进行分类和归纳。

1.1　基于数据库的领域本体学习方法

目前万维网上绝大多数数据仍然以关系数据库的方式存

收稿日期：2012-01-12；**修回日期**：2012-03-21　　**基金项目**：国家自然科学基金资助项目(60873010,61073139)；中央高校基本科研业务费资助项目(N090504005,N100604017,N090604012)；国家教育部新世纪优秀人才支持计划资助项目(NCET-05-0288)

作者简介：车成逸(1969-)，男，朝鲜民主主义人民共和国人，博士研究生，主要研究方向为数据库、信息抽取、本体；马宗民(1965-)，男(通信作者)，教授，博士，主要研究方向为智能数据与知识工程(mazongmin@ise.neu.edu.cn)；焦晓龙(1987-)，男，硕士研究生，主要研究方向为本体工程.

储(约占 77.3%)[2],实现关系数据库和本体之间数据的互操作性可以通过构建关系数据库模式和本体间映射的途径来解决。从广义上将基于关系数据库的本体学习问题分为两大类:从关系数据库抽取本体;关系数据库模式和本体间映射。

第一类方式又可以分为四种主要类型:基于逻辑数据模型的抽取方法;基于概念模型的抽取方法;基于模式描述本体的方法;基于辅助知识的方法。

基于逻辑数据模型的抽取方法将关系模型作为关系数据库的逻辑模型,从中抽取本体。在关系模型中,用二维表表示实体,用列表示实体属性,实体之间联系无法显式表示出来,而是通过主键和外键的方式隐式表达的。为了抽取本体,需要从关系模型中挖掘出这些隐含的语义关联,并发现表之间联系的基数比约束和参与度约束,这是该方案的难点。Stojanovic 等人[3]利用所定义的规则将关系模式转换为一个本体,通过数据迁移过程将关系数据库中的数据转换为本体的实例。该方法为了保持关系模型中的数据类型信息,将每个数据类型也转换为本体中的一个概念。另外,Astrova 等人[4]将数据库定义用 SQL 表示出来(DDL),将 DDL 作为系统的输入,构建表映射规则、列映射规则、数据类型映射规则、约束映射规则和行映射规则。该方法同时指出需要对转换的质量进行评价,即考察是否能够根据所生成的本体还原原来的数据库。

基于概念模型的抽取方法是将 ER 模型转换为本体的方法。由于关系模式一般是从 ER 模型转换而来的,而在 ER 模型转换为关系模式时,一些语义信息会丢失,如类层次关系等。所以通过将 ER 模型直接转换为 OWL 本体能够获取更多的语义信息。这种方式的输入是设计关系数据库时生成的 ER 模型或扩展的 ER 模型,然而 ER 模型往往是一种图形化表示,所以很难被计算机所理解和操作。Calvanese 等人[5]讨论了利用描述逻辑来表示 ER 模型的方法,通常采用的方法是将 ER 模型转换为一种模型表达语言(如转换为 XML 文档),然后再用程序进行解析。

基于模式描述本体的方法引入一个用于描述数据库模式信息的本体(以下称为模式描述本体)来描述关系模式[6]。该方法中数据源的相关元数据可以利用这个本体中所提供的词汇进行描述。这种方式只是利用模式描述本体中所定义的概念术语对数据模式进行描述,即将数据库的模式信息(如表名、列名等)作为模式描述本体的实例信息。该方式实质上是自动化地将数据库模式信息转换为计算机可以操纵和识别的本体表示语言。该方式转换算法简单,自动化程度高,但只注重数据模式结构信息的转换,而对于模式中的约束信息一般不能完全地转换到本体中。

基于辅助知识的方法为了获取更为丰富的语义信息,利用一些辅助知识生成本体。除了基于模式的方法外,还可以利用一些辅助知识来对要生成的本体进行优化。例如,针对数据库中的实例数据信息进行学习,根据学习到的知识来辅助本体的抽取;根据用户对数据库的查询日志[7]获取辅助知识等。通过这些方法可以为进一步修改本体提供依据,使所生成的本体更加符合用户的要求。实例数据的处理所需时间较长,但是能够从数据库中挖掘出更加丰富的语义信息。前三种类型基本上都是基于数据库模式的抽取方法,通过转换规则将其转换为本体,没有考虑语义保持性。

第二类方式从数据库模式和本体间建立映射的目的出发,又可以进行分类:基于模型转换途径的分类;基于映射所针对的数量的分类;基于映射结果表达形式的分类。

目前关系数据库模式和本体间模型转换的途径主要有两种:把关系数据库模式转换为类似本体形式表达;把关系数据库模式和本体分别转换到某种中间模型。有些研究工作[8,9]采用了把关系数据库模式用本体的形式表达的转换途径。通常这类工作首先通过一些转换规则,比如采用关系数据库逆向工程(relational database reverse engineering)的思想,自动或半自动地把关系数据库模式表达为本体的形式(以 RDFS 或 OWL 最为常见),然后再寻找转换本体和输入本体之间的映射。现有的研究主要集中在第二种模型转换途径上,即把关系数据库模式和本体分别转换到某种统一的中间模型[10,11]。通常这类工作首先定义一个表达能力适中的中间模型,比如有根的有向无环图(rooted directed acyclic graph)和 Web-PDDL 中间模型等,然后分别把关系数据库模式和本体转换到中间模型。对于关系数据库模式到中间模型的转换,可以增加某些语义信息,比如通过机器学习和数据挖掘的方法获取更多更复杂的关系;而对于本体到中间模型的转换,则需要裁剪丢弃不兼容的语义信息,比如把本体图模型转换为树型的连接公式。

从映射方法针对的关系数据库模式和本体的数量上将映射方法分为三种:两者皆为任意数目[12];两者数目都固定[10];两者之一数目固定。其中最为常用的方法是面向任意多个关系数据库模式和一个已知本体之间的映射[13]。通常此类方法主要面向某些特殊领域的数据集成问题,在这些领域中存在被普遍认同的通用本体,且这些通用本体覆盖了该领域中绝大多数的概念知识,这时只需要考虑多个关系数据库模式到该通用本体的映射问题。

目前关系数据库模式和本体间映射结果的表达形式有简单对应关系的表示方式[9,10,12]和较复杂的包含语义信息的表示方式[11]。

除了可以从关系模型中获取本体外,还可以从面向对象模型中获取本体。面向对象模型与本体有许多相似之处,所以从面向对象模型中获取本体的方法比较简单。另外,由于目前面向对象数据库应用范围有限,所以这方面不是研究的重点。由于篇幅有限,本文不作讨论。

1.2　基于 XML 的领域本体学习方法

XML 的文档结构可以嵌套任意复杂的句子,因此它适于构成大型和复杂的文档。这不仅使用户可以指定一个定义了文档中元素的词汇表,而且还可以指定元素之间的关系。DTD(document type definition)和 XML Schema 是 XML 文档的模式,用来对 XML 文档的逻辑结构进行定义。XML 文档的模式规定了 XML 文档中的元素、属性、元素间以及元素和属性之间的关系。可以将基于 XML 的本体学习方法分为两大类:从 XML 文档抽取本体;XML 文档和已存在的本体间映射。

第一类方式又可以分为两种主要类型:利用 XML 文档模式的转换方法;关注 XML 文档原始内容的转换方法。

利用 XML 文档模式的转换方法只关注 XML 文档的模式(DTD 或 XML Schema),一般忽略 XML 文档中的许多原始信息,即利用一些映射规则将其中的一些元素映射到本体。其中的研究重点是映射规则的发现,现有的方法可以分为两种类

型:a)基于学习的方法,即利用一些自学习的手段自动获取映射规则,如 Kavalec 等人[14]重点研究了利用机器学习方法自动地得到映射规则;b)基于预定义规则,即用户使用预定义的规则,从 XML 文档的模式中提取语义信息生成相应的概念模式,然后对这些概念模式进行语义集成得到本体。这类方法以 Ferdinand 等人[15]提出的为代表,即从 XML Schema 生成 OWL 本体和将 XML 文档转换到 RDF 图的方法。从 XML Schema 到 OWL 映射是基于预定义规则:从 XML Schema 的复杂类型、模型组定义和属性组定义生成 OWL 类;从复杂类型的元素生成 OWL 对象属性;从简单类型元素和属性生成 OWL 数据类型属性;从 XML Schema 的约束性继承及扩展性继承生成类继承。Thuy 等人[16]的 DTD2OWL 系统也是将 XML 文档模式(XML DTD)自动映射到 OWL 领域知识和将 XML 实例转换到 OWL 个体的框架。该方法可以将 XML 文档的所有元素转换成 OWL,而且在 XML 模式的要素定义的基础上自动描述类和属性。上述方法由于各种 XML 文档的模式在语法上的差异,需要使用不同的映射规则。为此,Volz 等人[17]提出将这些 XML 数据映射成一棵语法树,该语法树是一个四元组:非终结符集、终结符集、开始符集、规则集;然后使用一些规则将这些非终结符集和终结符集中的元素映射为本体中的概念和关系。通过使用语法树,该方法克服了现有 XML 文档的模式在语法上的差别。但当把 XML Schema 映射成语法树时,该方法没有考虑 XML Schema 的完整性约束。

关注 XML 文档原始内容的转换方法的重点是解决提取 XML 语义信息的问题。Melnik[18]只考虑了 XML 文档本身不考虑 DTD 或 XML schema 的存在,它认为每个 XML 文档都会包含一个 RDF 模型。该方法通过使用一个简化的语法形式来检测 XML 实例中的语义信息,可以将 XML 实例映射为 RDF 文档。Xu 等人[19]提出了利用实体联系模型从 XML 文档中生成 OWL 本体的方法。该方法将 XML 文档映射到实体联系模型的 XTR(XML-to-relational)映射方法和将实体联系模型映射到 OWL 本体的 RTO(relational-to-Ontology)映射方法。该方法中利用表示关系数据库的特殊词汇描述 OWL 本体。

第二类方式是进行 XML/XML Schema 和本体间映射的方法。目前通过将 XML 文档与本体建立映射,为 XML 增加语义或集成异构 XML 文档。Cruz 等人[20]提出了利用本体作为媒介集成异构 XML 数据源的方法。本体的集成处理包括模式转换和本体集成两个阶段。该方法为将每个 XML 数据源做成局部 RDF 本体利用了 RDFS,从 XML 转换到 RDF 的方法为:将复杂类型元素转换到 rdfs:class;将属性和简单类型元素转换到 rdfs:property;元素—部分元素关系利用 rdfx:contain 编码为类—类关系。另外,An 等人[21]在模式和本体之间的映射中,利用了某种中间模型。

1.3 基于 Web 表格的领域本体学习方法

表格是一种表达结构化信息强有力的手段,当前许多领域信息都采用表格形式展现。因此能够简明清楚地表达出数据之间的关系的、且不需自然语言分析的表格是能够构建高质量本体的数据源。属于关系数据库的数据表格是其单元格按照关系模式的严格规定布置的表格,它的结构一目了然。但网页上用于向用户展现结构化信息的表格(以下称为 Web 表格)是以非结构化语言 HTML 编码的,因此对 Web 表格信息获取比较困难。Web 表格是按照设计者的想法任意地布局,表格中的句法和语义概念是相互混合的,以表格逻辑单元格的相对位置信息来获得该内容,因而此类句法结构比自然语言更为复杂。

从 Web 表格构建本体的过程大致分为表格信息抽取和本体学习两个阶段。Web 表格信息抽取包括:Web 表格定位、Web 表格结构识别、Web 表格内容识别等。

Web 表格定位是指从 Web 页内找到表格区域,并判断真假表格。目前有基于机器学习分类、基于人工构造规则分类及基于本体辅助分类三种方式的真假表格分类方法。Wang 等人[22]提出了 Web 表格定位时需考虑的三类特征,即布局特征、内容类型特征和词组特征,并且给出了基于决策树和 SVM 学习方法的真假表格分类算法。Jung 等人[23]提出了基于决策树学习方法的独立于具体领域的表格定位方法。

Web 表格结构识别是指生成表格的逻辑结构模型,其包括标题行和内容行识别、表格展开方式识别以及表头和表体识别,如 Pivk 等人[24]比较表格行之间相似度和列之间相似度的数值大小来确定表格的展开方式,依据匹配表格区域的内容模式来识别属性—值区域并规整表格逻辑单元;Yoshida 等人[25]使用特定领域的词汇信息,通过期望最大化算法匹配预定义的九种表格结构模型。

Web 表格内容识别是指对于其存放的内容进行识别并存储到数据库中。在表格内容抽取中,目前可以使用的方法有:基于规则的抽取、基于决策树的抽取以及基于本体的抽取。例如,Cui[26]通过表格中关系源属性到本体中目标属性的推理映射实现内容整合;Chen 等人[27]利用 Web 表格单元格的一些标记规则对属性单元格进行内容规整,如 Rowspan、Colspan 属性等。

Web 表格本体学习过程包括:利用已抽取的表格信息本体元素生成、本体内部的映射发现和本体合并。这方面的代表性工作是 BYU 研究小组的 TANGO 系统[28],其基本过程包括:理解 Web 表格的结构和概念内容、发现概念内容间的相互约束关系并生成小型本体、利用已构建的应用本体对小型本体进行概念匹配来发现本体内部的映射、将小型本体合并到应用本体。

2 结构化信息源本体创建方法的分析与对比

本体的构建是对概念本身以及概念与概念之间的关系进行形式化描述,多是面向特定领域。由于应用领域的不同,对本体研究的侧重点也有所不同。出于对不同学科领域和具体工程的不同考虑,领域本体构建的过程各不相同。

2.1 结构化信息源本体创建方法的分析

现行的结构化信息源本体创建方法都不是经权威标准化机构认证的方法,而是从具体的本体构建项目中总结获得的。由于目前还没有统一的评价标准,很难对以上提到的三种结构化数据源本体创建方法进行定量的评价。上述的每个构建方法对其他结构化信息源本体构建有着参考意义,但不能直接套用,因此需要对结构化信息源的特点进行分析。

下面从结构化信息源本体创建的角度对结构化数据源的特点进行分析。

a)从数据源获取的难易程度来看　出于安全和保护商业秘密等因素的考虑,网站后台的数据库中数据往往对普通用户

是不可见的,Web 上很少提供现成的关系模型的模式和数据;Web 上大多数信息以 HTML 文档的格式呈现,但 HTML 不能描述结构化数据,目前 Web 上大量使用 XML 文档描述结构化数据,XML 已成为当前因特网上信息交换的一种标准语言,因此 XML 文档较为容易获取;大量的 HTML 文档中包含动态和静态表格,其中动态表格是基于后台数据库生成并以 HTML 形式展现的,包含了结构化数据,因此,Web 表格较容易获取。

b)从是否包含模式信息来看　关系数据库一般都包含模式信息,它采用的是关系模型,在该模型中,关系是元组的集合,而关系模式是用来描述关系的结构的,即它由哪些属性构成、这些属性来自哪些域以及属性和域之间的映像关系;XML 文档也大多包含模式信息,而 Web 表格通常不含模式信息以及表格标题(表格名)。

c)从数据间联系程度来看　关系数据库一般面向特定的领域应用设计搭建,因此其中的关系表格之间联系程度高。表格之间通过外键约束保持联系,实体及实体间的联系都是用表来表示的。无论是概念的获取还是概念间关系的获取,首先必须区分出哪些表是用来描述实体的,哪些表是用来描述实体间联系的,然后才能将实体信息转换为本体中的概念,将联系信息转换为本体中的关系;XML 文档之间的联系程度较低,而 Web 表格通常不含表格间的联系信息(表格单元格中的超链接元素将其他网页关联起来,但很难实现表格间的联系)。

2.2　结构化信息源本体创建方法的对比

与以上三种结构化信息源相比,本体是一种具有更多语义、结构更为复杂的模型,所以本体创建的主要任务是分析结构化信息源中蕴涵的语义信息,将其转换到本体中的相应部分。而由于输入信息源类型的不同,结构化信息源的本体创建中采用的方法及创建结果也有所不同。

下面从基于不同结构化数据源构建本体方法的成熟度来进行对比。

由于关系数据库的模式明确,结构清晰,基于关系数据库构建本体的研究工作提出的方法较多,也相对成熟。本体描述语言 OWL 虽然基于 XML 构造,但不能在 XML 文档上简单添加字段来构建本体,而必须根据 XML 文档本身的模式信息来构建本体。基于 XML 文件构建本体的研究工作进展居于其他两者之间,成熟度次于基于关系数据库构建本体;由于 HTML 语言用于在浏览器中展现 Web 页面,缺乏对数据的描述,因此 Web 表格不含模式信息。基于 Web 表格构建本体的研究工作最为困难,提出的方法较少,相对不成熟,尚待研究的问题仍然很多(表 1)。

表 1　结构化信息源本体创建方法的比较

比较项	数据库	XML 文档	Web 表格
获取的难易程度	难	容易	容易
模式信息的包含程度	包含	包含	不含
数据间联系程度	高	较低	低
方法的成熟程度	成熟	较成熟	不成熟

3　总结与展望

3.1　总结

本文从广义的角度考察影响本体构建的整个过程的因子(依据的数据源、构建的步骤、构建结果),据此,详细分析了从现有的三种结构化信息源(数据库、XML 文档、Web 表格)构建领域本体的方法。从中可以看出,针对不同类型的数据源需要采用不同的本体学习技术,虽然结构化信息源是一个能以简便、快捷的方式构建本体的新兴的研究领域,许多相关领域的研究成果可以供其借鉴,但是由于结构化数据源具有自身的特殊性,在利用上述方法构建实用的领域本体时,该领域仍然存在不足之处。总结起来有以下几个方面:

a)在当前的本体构建研究成果中还没有某种标准的方法能够支持所有形式的领域本体构建。而本体学习方法大部分依赖于本体的构建者所采用的方法,因而对于任何领域都适合的途径或典型模式不存在。

b)在构建不同领域的本体时,没有能够明确表达本体的构建要求的手段和能够评价构建的本体是否符合其要求的方法。

c)领域本体构建方法还未能像软件工程那样成为一种成熟的工程方法,更没有构建过程的规范管理。

d)领域本体构建的目标是获取相关的领域知识,提供对领域知识的共同理解,确定领域内共同认可的概念,并从不同层次的形式化模型上给出这些概念和概念之间相互关联的明确定义,提供其领域中发生的活动以及主要理论和基本原理等,即领域本体的主要目的是为不同系统提供语义交流的手段,因而共享和重用是领域本体构建中很重要的问题。但当前的构建方法忽视构建后的本体共享和重用。

3.2　未来的研究和发展方向

从结构化 Web 数据源构建本体的研究尚属较新的研究领域,虽然已经出现了一些研究成果,但仍然存在许多值得研究和解决的问题,可以总结为以下几个方面:

a)从关系模式生成语义映射

在当前的关系数据库与本体之间的映射中,现有方法一般只考虑关系模式的语义,而没有进一步去挖掘大量元组中包含的语义信息,所以获取的概念数量和关系种类都非常有限。一般来说,数据模式总伴随着数据实例,因此如何利用数据实例来辅助生成语义映射是一个重要的研究方向。因为很多具体映射问题都是针对具体领域而言的,所以在关系数据库与本体映射的研究中,采用领域知识来提高映射的准确度是一个可行的思路。

b)关系模式与本体之间语义映射结果的评价

对于关系模式与本体之间的映射,需要建立合理的评价方法。现在的一些映射方案中,有些采用正确映射和自动生成的候选映射之间的比较来评价其算法。如何对所生成的模式与本体之间的映射进行定量的评价是一个重要的研究方向,它的研究有利于推动相关软件方法和工具的进一步发展。

c)从 Web 表格语义信息抽取

为了准确地抽取表格信息,需要利用自然语言词语特征抽取表格中隐含的语义特点。例如从 Web 表格抽取 OWL 属性时,利用自然语言词汇特征可以生成 Is-a 关系及类—实例关系等[29]。表格中的句法和语义概念是相互混合的,只是以表格逻辑单元格的相对位置信息来获得语义非常有限。在目前的本体构建研究中,只着重利用表格的结构信息转换成本体元素,很难获取准确的语义信息。因此,采用基于自然语言文本

的本体抽取方法构造符合表格本体建设任务自身的构建方法具有重要的意义。

d)较为完整的属性元素抽取

目前从 Web 表格抽取本体的研究大部分集中在类、实例和三元组的获取,构建的本体属性元素还较少且有限。为了生成更加符合语义 Web 特点的本体,需要进行更为丰富的、能够应用于本体推理等实证研究的属性获取。

总之,基于结构化 Web 数据源本体构建是语义 Web 建设中的一个重要研究分支,其研究成果将极大地促进本体管理和应用技术的发展与应用。目前有些领域还不具有领域内通用的本体,这就给该领域中的语义交互带来了更大的困难,所以从现有的结构化数据源中快速、自动地生成本体具有十分重要的意义。今后的研究工作会沿着结构化数据源的语义特点抽取和本体属性元素获取两个方面深入。

参考文献:

[1] STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineering: principles and methods[J]. Data and Knowledge Engineering, 1998, 25(1-2): 161-197.

[2] CHANG K C, HE Bin, LI Cheng-kai, *et al*. Structured databases on the Web: observations and implications[J]. ACM SIGMOD Record, 2004, 33(3): 61-70.

[3] STOJANOVIC L, STOJANOVIC N, VOLZ R. Migrating data-intensive Web sites into the semantic Web[C]//Proc of the 17th ACM Symposium on Applied Computing. New York: ACM Press, 2002: 1100-1107.

[4] ASTROVA I, KORDA N, KALJA A. Rule-based transformation of SQL relational databases to OWL ontologies[C]//Proc of the 2nd International Conference on Metadata and Semantics Research. Corfu Island: Ionian Academy, 2007: 415-424.

[5] CALVANESE D, LENZERINI M, NARDI D. Unifying class-based representation formalisms[J]. Journal of Artificial Intelligence Research, 1999(11): 199-240.

[6] DeLABORDA C P, CONRAD S, PÉREZ C, *et al*. Relational. OWL: a data and schema representation format based on OWL[C]//Proc of the 2nd Asia Pacific Conference on Conceptual Modelling. Darlinghurst, Australia: Australian Computer Society, 2005: 89-96.

[7] ZHANG Jie, XIONG Miao, YU Yong. Mining query log to assist ontology learning from relational database[C]//Lecture Notes in Compu-ter Science, vol 3841. Berlin: Springer-Verlag, 2006: 437-448.

[8] CHEN Hua-jun, WU Zhao-hui, WANG Heng, *et al*. RDF/RDFS-based relational database integration[C]//Proc of the 22nd International Conference on Data Engineering. Washington DC: IEEE Computer Society, 2006: 94-94.

[9] KOROTKIV M, TOP J L. From relational data to RDFS models [C]//Lecture Notes Computer Science, vol 3140. Berlin: Springer-Verlag, 2004: 430-434.

[10] DRAGUT E, LAWRENCE R. Composing mappings between schemas using a reference ontology[C]//Lecture Notes Computer Science, vol 3290. Berlin: Springer-Verlag, 2004: 783-800.

[11] DOU D, LEPENDU P, KIM S. Integrating databases into the semantic web through an ontology-based framework[C]//Proc of the 3rd International Workshop on Semantic Web and Databases. Washington DC: IEEE Computer Society, 2006: 54-54.

[12] AN Yuan, BORGIDA A, MYLOPOULOS J. Inferring complex semantic mappings between relational tables and ontologies from simple correspondences[C]//Lecture Notes Computer Science, vol 3761. Berlin: Springer-Verlag, 2005: 1152-1169.

[13] LUO Xi-xi, CHEN Xiao-wu, ZHAO Q. OOML-based ontologies and its services for information retrieval in UDMGrid[C]//Proc of the 6th International Confercence on Workshop on Advanced Parallel Processing Technologies. Berlin: Springer-Verlag, 2005: 342-352.

[14] KAVALEC M, SVATEK V. A study on automated relation labelling in ontology learning[C]//Ontology Learning from Text: Methods, Evaluation and Applications. Amsterdam: IOS Press, 2005: 44-58.

[15] FERDINAND M, ZIRPINS C, TRASTOUR D. Lifting XML schema to OWL[C]//Proc of the 4th International Conference on Web Engineering. Berlin: Springer-Verlag, 2004: 354-358.

[16] THUY P T T, LEE Y K, LEE S Y. DTD2OWL: automatic transforming XML documents into OWL ontology[C]//Proc of the 8th IEEE/ACIS International Conference on Computer and Information Science. Washington DC: IEEE Computer Society, 2009: 125-131.

[17] VOLZ R, OBERLE D, STAAB S, *et al*. OntoLiFT prototype, IST Project 2001-33052 WonderWeb[R]. 2003.

[18] MELNIK S. Bridging the gap between RDF and XML[EB/OL]. (1999)[2010-12-11]. http://www-db. stanford. edu/melnik/rdf/syntax. html.

[19] XU Jiu-yun, LI Wei-chong. Using relational database to build OWL ontology from XML data sources[C]//Proc of International Conference on Computational Intelligence and Security Workshops. Washington DC: IEEE Computer Society, 2007: 24-127.

[20] CRUZ I F, XIAO H, HSU F. An ontology-based framework for XML semantic integration[C]//Proc of the International Database Engineering and Applications Symposium. Washington DC: IEEE Computer Society, 2004: 217-226.

[21] AN Yuan, BORGIDA A, MYLOPOULOS J. Constructing complex semantic mappings between XML data and ontologies[C]//Proc of the 4th International Semantic Web Conference. Berlin: Springer-Verlag, 2005: 6-20.

[22] WANG Y, HU J. A machine learning based approach for table detection on the web[C]//Proc of the 11st International World Wide Web Conference. New York: ACM Press, 2002: 242-250.

[23] JUNG S W, KANG M Y, KWON H C. Hybrid approach to extracting information from Web-tables[C]//Proc of the 21st International Conference on Computer Processing of Oriental Languages. Berlin: Springer-Verlag, 2006: 109-119.

[24] PIVK A. Automatic ontology generation from Web tabular structures [J]. AI Communications, 2006, 19(1): 83-85.

[25] YOSHIDA M, TORISAWA K, TSUJII J. A method to integrate tables of the World Wide Web[C]//Proc of the 1st International Workshop on Web Document Analysis. 2001: 31-34.

[26] TAO Cui. Schema matching and data extraction over HTML tables [D]. Bringham: Brigham Young University, 2003.

[27] CHEN H H, TSAI S C, TSAI J H. Mining tables from large scale HTML texts[C]//Proc of the 18th International Conference on Computational Linguistics. San Francisco: Morgan Kaufmann, 2000: 166-172.

[28] TIJERINO Y A, EMBLEY D W, LONSDALE Y, *et al*. Toward ontology generation from tables[J]. World Wide Web: Internet and Web Information Systems, 2005, 8(3): 261-285.

[29] CHA S I, MA Zong-min, CHENG Jian-wei, *et al*. Learning of ontology from the web-table[C]//Proc of the 8th International Conference on Fuzzy Systems and Knowledge Discovery. Washington DC: IEEE Computer Society, 2011: 1454-1458.