面向文本的本体学习研究概述

贾秀玲[1] 文敦伟[1,2]

(中南大学信息科学与工程学院 长沙 410083)[1]

(阿萨巴斯卡大学计算机与信息系统学院 阿萨巴斯卡加拿大 T9S3A3)[2]

摘 要 对本体(ontology)的研究在计算机领域变得越来越广泛,但手工构造本体是一项繁琐而辛苦的任务,还会导致知识获取瓶颈。本体学习技术是利用本体工程技术和机器学习技术等众多学科技术来实现本体的(半)自动构建。本体的学习可以面向文本、知识库、结构化数据、半结构化数据和无结构数据。本文主要介绍了面向文本的本体学习,并对其中的学习内容、学习方法、学习工具、学习过程和系统评价等关键技术进行了说明,特别介绍了学习方法中的基于统计的方法、词汇句法模式法和形式概念分析法并对其优缺点做了简单的分析。

关键词 本体,本体学习,知识获取,学习方法,评价方法

A Survey of Ontology Learning from Text

JIA Xiu-Ling[1] WEN Dun-Wei[1,2]

(School of Information Science and Engineering, Central South University, Changsha 410083)[1]

(School of Computing and Information Systems, Athabasca University, Athabasca Canada T9S3A3)[2]

Abstract Research on ontology is increasingly becoming widespread in the computer science community. But the manual construction of ontology is a time-consuming task and easily leads to the bottleneck of knowledge acquisition. Ontology learning aims at the integration of a multitude of disciplines such as ontology engineering techniques and machine learning techniques to construct the ontology (semi)automatically. There are different ontology learning approaches according to the type of input: ontology learning from text, from knowledge base, from structured-data, from semi-structured data and from unstructured data. Ontology learning from text is mainly introduced. The key technologies of ontology learning from text are presented, including learning content, learning approach, learning tool, learning process and system evaluation. Authors especially introduce the statistical method, lexico-syntactic pattern method and formal concept analysis method and simply analyze the advantage and disadvantage of these methods.

Keywords Ontology, Ontology learning, Knowledge acquisition, Learning method, Evaluation method

1 引言

本体(ontology)的概念最初起源于哲学领域,是客观存在的一个系统的解释或说明,关心的是客观现实的抽象本质。在人工智能界,Neches[1]将 Ontology 定义为"给出构成相关领域词汇的基本术语和关系,以及利用这些术语和关系构成的规定这些词汇外延的规则的定义"。而本体最为流行的定义[2]是"ontology 是概念模型的明确的规范说明"。

目前对于本体的研究在计算机科学领域变的越来越广泛,本体被广泛地用于许多领域如语义网、搜索引擎、电子商务、自然语言处理、知识工程、信息提取、多 agent 系统、数据库设计和数字图书馆等。为何本体有如此大的魅力?我们知道[28]全面地理解自然语言需要整合大量的知识源,而以本体形式表示的领域知识是深入理解文本的基础;知识管理主要是处理一个组织中知识的获取、维护和访问,而本体可以用于对无结构信息进行语义标注,从而使得信息的整合和访问更容易;在电子商务中,交易的自动化要求对商品进行形式化描述,因此需要一个标注化的词汇表——本体。本体有助于对内容意义的精确高效通信,同时促使系统的交互式操作、重用和共享等一系列性能得以提高。从上述几个应用方面可以看出,这些领域的一个共同需求是共享某个领域内的知识,而这正是本体的主要目标。

本体在众多领域的应用都是在构建本体的基础之上实现的,但本体的构建却是一项繁琐而辛苦的任务。手工方式构建的本体需要耗费大量的人力和时间,像 Cyc[31]和 Wordnet[32]等系统需要使用人工为本体输入大量的知识,然后系统才能使用其庞大的知识库进行推理或是获取新的知识。这就容易导致知识获取瓶颈,无法保持本体的更新。因为本体中的知识是变化的,它总是在不断地发展和更新。这就决定了本体不能以手工方式构造,我们需要自动或半自动方式来构建本体。因此,本体学习(ontology learning)技术应运而生,它可以实现本体的自动或半自动构建。

2 面向文本的本体学习

本体学习技术[3]旨在综合众多的学科技术来促进本体的自动或半自动构建,特别是本体工程技术[28]和机器学习技术[27]。本体学习涉及到从输入数据中提取本体学习内容(概念知识)并用这些内容构建本体。在最近几年一些本体学习

贾秀玲 硕士研究生,主要研究方向为自然语言处理、语义网与本体工程;**文敦伟** 教授,主要研究方向为分布式人工智能与知识工程、自然语言处理与机器学习、智能系统与计算智能。

方法和系统已经建立，这些系统中大多数是依赖于语言分析和机器学习算法以发现潜在的感兴趣的概念和这些概念之间的关系。本体学习有面向文本、面向字典、面向知识库和面向关系数据库等等[3]。本文中提到的本体学习系统指的都是面向文本的，通过分析比较这些系统，本文对面向文本的本体学习中的一些关键技术做了简单的总结，希望能起到抛砖引玉的作用。

2.1 学习内容

Alexander Maedche[5]对 ontology 结构定义为一个 5 元组 $O:=\{C,R,H^C,rel,A^O\}$。其中，C 为概念集合；R 为关系集合；H^C 为概念层次或分类层次，$H^C\subseteq C\times C$ 是一种有向关系，$H^C(C_1,C_2)$表示 C_1 是 C_2 的子概念；$rel:R\rightarrow C\times C$ 是一个函数，表示概念之间的非分类关系。$rel(R)=(C_1,C_2)$亦可表示为 $R(C_1,C_2)$；A^O 为使用某种逻辑语言表达的 ontology 的公理集。多数本体学习系统[6,10,12,39]学习的主要是本体的知识(ontological knowledge)，即所学内容主要是概念、关系和公理。也有一些系统[37]是学习如何从输入数据中提取本体知识的元知识(meta knowledge)。它们学习的元知识是诸如从 Web 中提取实例和关系的规则或者是从文本中提取知识的模式等等。本文介绍的本体学习所学习的主要是本体的知识，下面就本体知识中的概念、概念间关系和公理做更详细的说明。

2.1.1 概念

本体是共享的概念模型的形式化的规范说明，Fense[4]对这个定义进行分析后认为本体的概念包括 4 个主要方面：概念化(conceptualization)，本体是通过抽象出客观世界中的一些现象的相关概念而得到的模型；明确(explicit)，概念及它们之间的联系都被精确定义；形式化(formal)，精确的数学描述；共享(share)，本体中反映的知识是其使用者共同认可的，反映的是相关领域中公认的概念集。

概念的含义很广泛，可以指任何具体的或抽象的事物，如工作描述、功能、行为、策略和推理过程等。概念可以在本体模型图中用节点表示也可以在本体学习系统中通过学习得到，概念可以从输入数据中提取也可以在本体精炼(refine)[5]的过程中从其它的概念中产生。换句话说在本体学习系统的输入数据中，可能有也可能没有元素和概念对应。在基于术语的概念获得中，一个概念节点可以从相应的提取的术语中产生，这些术语是自然语言的词或短语。然而在基于语义的概念获得中，概念常常在本体精炼的过程中产生。

2.1.2 概念间关系

许多本体学习系统把概念间关系分为分类关系(taxonomic relation)和非分类关系(non-taxonomic relation)两种[35]。分类关系被广泛地用于组织本体的知识，许多系统都把上下位关系(hyponymy relation)作为分类关系来处理。下位/上位关系也称为从属/上属关系，子集/超集关系，或 ISA 关系。像{枫树}是{树}的下位词，{树}是{植物}的下位词，如："An x is a (kind of) y."为框架构造的句子，则同义词集合{X, X…}表示的概念称为同义词集合{Y, Y…}表示的概念的下位关系。非分类关系是指除了 ISA 关系以外的概念间的任何关系，如：同义词关系(synonymy)、属性关系(attribute-of)和实例关系(instance-of)等等。一类关系是两个或多个概念之间的联系，因此，关系应该作为 n 个概念($n>1$)的一个产物的子集来学习。

2.1.3 公理

公理就是无需证明的正确假设，也是逻辑的前提[2]，公理包含在本体中有以下几个目的：约束包含在本体中的信息，校验它的正确性或推导出新的信息。对于自动半自动的学习公理存在一个"公开"的问题。Hasti[12]系统在限定的条件下学习公理，在条件化定量化的自然语言语句中 Hasti 系统把外在的公理转换为以 KIF(Knowledge Interchange Format)编码形式的逻辑化、形式化公理。

2.2 学习方法

知识提取方法的范围是[4]从一般的知识提取方法(统计的方法)到比较精深的方法(逻辑的方法)。在这个范围之内有不同的方法和技术来提取本体知识和学习本体。这些方法可以是监督的也可以是无监督的，可以是在线的，也可以是离线的。本体学习包含了逻辑学、语言学、模板驱动和语义分析等学科技术，也可把启发式方法应用到其中。在面向文本的本体学习中我们可以采用不同的方法和技术来提取本体知识和学习本体，本体学习算法可以采用基于统计的方法[14~16,36]、词汇句法模式(lexico-syntactic patterns)[17]、关联规则(association rules)[18]、形式概念分析(formal concept analysis)[19,20]和聚类[13,30]等方法。下面重点说明学习算法中的基于统计的方法、词汇句法模式法和形式概念分析法。

2.2.1 基于统计的方法

本体学习系统中概念的提取多采用基于统计的方法，简单的提取概念的方法是计算术语的频率，通常这种方法是基于假设的，即在一个特定领域的文本集合中术语的频率可以表明这个概念在相关领域出现的频率，文[36]介绍了一种比简单计算频率更有效的方法：设文本集合 D，文本 $d\in D$，$f_{l,d}$：术语 l 在文本 d 中的出现的频率；$f_{d,D}$：在文本集合 D 中包含术语 l 的文本 d 的数量；$f_{l,D}$：在所有的文本集合 D 中术语 l 出现的频率；可以看出：$f_{d,D}\leqslant f_{l,D}$，而且 $\sum d\times f_{l,d}=f_{l,D}$。

定义 1 (文本 d 中术语的频率)：

$$F_{l,d}=f_{l,d}*\log\left(\frac{/D/}{f_{d,D}}\right)$$

可见在概念的提取中，低频术语和高频术语的区分度较低，而中等频率术语往往与文本的主题有关。一个术语在一个文本中的等级可以反映它在整个文本集中的等级。

定义 2 (文本集 D 中术语的频率)：

$$F_l=\sum_{d\in D}F_{l,d}\quad F_l\in R$$

用户可以定义和修改一个阈值 $k\in R^+$，F_l 必须大于这个值，这个阈值用来选择文本集中的术语形成概念集合。

基于统计的方法也可以用于概念间关系的学习，在 Text-To-Onto[38]系统中分类关系的提取就用到了基于统计的聚类提取和基于统计的分类提取的方法。基于统计的方法最大的缺点就是容易产生数据稀疏(data sparse)现象，特别是对于领域文本中的一般性概念。基于统计的方法的另一个特点是多用于非增加(non-incremental)的本体学习系统，所谓非增加是指用学习的方法立即从整个输入数据中学习，输入数据不再改变的那些系统。

2.2.2 词汇句法模式法

在本体学习中常采用词汇句法模式方法提取概念间的语义关系特别是分类关系，一般基于模式的匹配方法都是启发式的，文[17]详细说明了词汇句法模式，如有以下英文模式：

例 1 Such NP as {NP,} * {(or and)} NP

… Works by such authors as Goldsmith, and Shakespeare.

⇒Hyponym("author","Goldsmith"),Hyponym("author","Shakespeare")。当在某个语句中应用到这个模式时,or | and右边的概念(NP)是左边概念(NP)的子概念。

例2 NP {,} especially {NP,} *{or | and} NP

… most European countries, especially France, England, and Spain.

⇒Hyponym("European country","France")

Hyponym("European country","England")

Hyponym("European country","Spain")

这种词汇句法模式可以是一个循环过程,开始可以人工总结出一些模式如上例,利用这些模式学习得到一些上下位关系,然后从中取出一对上下位关系例如 hyponym("European country","France")对语料库进行搜索又可以发现此上下位关系的新的模式,再利用这些新的模式又可以从语料库中提取出新的上下位关系。这种词汇句法模式法对新模式的发现和定义是一个循环过程,这种方法对于人工发现和定义模式有了一定的改进。这种模式匹配法的特点是结果更准确,和基于统计的方法比深度更大,但广度不如基于统计的方法,虽然这种模式匹配法对于模式的定义和选择有所改进,但在模式的定义广度上还不够。

2.2.3 形式概念分析法学习概念层次

形式概念分析是应用数学的一个分支,它是建立在概念和概念层次的数学化基础之上的。目前,形式概念分析的方法已经大量运用在概念聚类、数据分析、信息检索、知识发现和本体工程的应用之中。在文[20]中用形式概念分析的方法发现概念之间的内在关系,这些概念是通过一些属性来描述的。为了从文本中获得属性,解析文本并从中提取出动词/介词短语、动词/主语短语、动词/宾语短语。对于出现的每个名词我们用相对应的动词作为属性来构造形式上下文,然后以这个为基础来计算形式概念格(concept lattice)[9]。

定义3(形式上下文 formal context) 一个形式上下文是一个三元组(G,M,I)。其中,G和M是集合,$I\subseteq G\times M$是G和M之间的一种二元关系。G中的元素称为对象,M中的元素称为属性。对于$A\subseteq G,B\subseteq M$:

$A':=\{m\in M\mid \forall g\in A:(g,m)\in I\}$

$B':=\{g\in G\mid \forall m\in B:(g,m)\in I\}$

对于对象集合A,A'是A中所有对象的共性的集合;对于性质集合B,B'是具有B中所有性质的对象的集合。

定义4(形式概念) 对集(A,B)是形式上下文(G,M,I)的形式概念,当且仅当$A\subseteq G,B\subseteq M,A'=B,A=B'$。一个给定上下文的形式概念也可以用子概念—超概念(sub-super concept)之间的关系来定义:$(A_1,B_1)\leqslant(A_2,B_2)\Leftrightarrow A_1\subseteq A_2$($\Leftrightarrow B_2\subseteq B_1$)"$\leqslant$"反映了概念间的层次关系,$(A_1,B_1)$被称作$(A_2,B_2)$的子概念,$(A_2,B_2)$则是$(A_1,B_1)$的超概念。由层次关系搭构的所有$(G,M,I)$的概念叫做概念格。

例如:The museum houses an impressive collection of medieval and modern art. The building combines geometric abstraction with classical references that allude to the Roman influence on the region.

通过分析器分析这些句子,我们可以提取出:

houses_subj(museum)

houses_obj(collection)

combines_subj(building)

combines_obj(abstraction)

combine_with(references)

allude_to(influence)

提取出这些对集以后会产生两个问题[20]:首先,分析器的输出结果可能是错误的,也就是说并不是所有得到的动词属性对都正确。其次,并不是所有获得的动词属性对在语义上都是有意义的,即可以帮助区分不同的对象。为了解决以上问题,可以用某些信息计算公式来测量对象/属性对,只有这些测量结果超过某一阈值时,这些对象/属性对才是我们需要的。形式概念分析法的特点是更一般更容易实现,覆盖度比较大。但不足之处是准确度不高而且容易产生数据稀疏现象。

2.3 学习工具

目前面向文本的本体学习工具有很多,如 OntoLearn(Velardi et al.[10,11], 2002), Text-To-Onto(Maedche and Volz[6,38],2000;Maedche and Staab[8],2003),SOAT(Wu and Hsu[39], 2002), ASIUM (Faure and Ne-dellec[13], 1999), SVETLAN' (Chaelandar a-nd Grau[40], 2000) 和 Hasti (Shamsfard[12], 2003)等等,下面就其中的几种做一简单介绍。

(1)Text-To-Onto[5,8]

由卡尔斯鲁厄大学(University of Kar- lsruhe)的 AIFB (Institute of Applied Inform-atics and Formal Description Methods)开发,它是一个一体化的本体学习环境。用这个环境可以发现概念间的关系从而构建本体。Text-To-Onto 有一个学习算法库可以满足不同的需要。它的学习方法采用的是一种多策略综合的方法,对于不同的输入数据和任务可以结合不同的方法去学习。Text-To-Onto 方法支持半自动化的本体开发,这种方法的好处是本体创建和使用与文档生成和使用之间有可选的反馈。一个组织内部的知识结构的演进可用通过进入和产出文档的语言分析来进行监测,其结果将用于更新对应的本体。

(2)OntoLearn[10]

OntoLearn 工具是 2002 年 Velardi 和 Missikoff 等人提出的,目的是从一个文本语料库中提取相关的领域术语,然后使用自然语言处理和统计的技术来过滤这些术语。他们的工作是基于 WordNet 的。虽然 WordNet 并不提供语义 Web 的本体标准,但是它是使用最广泛的通用在线词汇数据库和英语词汇资源领域的标准。OntoLearn 工具的一种新方法是术语的语义解释,可以使用这个语义解释来探测不同于分类的其他类型的关系,因一般的方法总是将术语识别成为领域概念。以 OntoLearn 工具为中心,能够建立和评估领域本体以支持虚拟用户社区的智能信息集成。还在两个欧洲项目中测试了 OntoLearn,在项目中它作为语义交互平台的基础用于小一中规模的旅游企业。

(3)ASIUM[13]

ASIUM 是"Acquisition of Semantic K-nowledge Using Machine Learning Method-s"的缩写,是由 Paris-Sud 大学计算机科学研究室开发的。ASIUM 的主要目的是用分析的方法帮助专家获得技术文本的语义知识。学习方法是基于概念上和层次上的聚类。使用一个矩阵计算各个类之间的语义相似度。

(4)Hasti[12]

Hasti 是 2003 年 Shamsfard 等人提出的一种自动化本体构造方法,是从一个小的本体内核出发,通过文本理解来自动

化建造本体。本体内核包含建立本体所需的基本概念、关系和操作符，还包含了添加、移动、删除和更新本体元素的基本元知识。他们提出模型的特性之一是它是与领域和应用独立的，在一个小的基本内核上建造本体，学习单词、概念、分类关系、非分类关系和公理。Hasti 是一个用于实现和测试自动本体构造方法的项目，它从波斯语文本中抽取词汇和本体知识。

2.4 学习过程

此小节主要讨论面向文本的本体学习系统的学习过程的自动化程度。知识获取可以从手工、半自动到全自动来实现，本文所谈的学习系统都是全自动或半自动的。像 Hasti 系统是全自动的，Text-To-Onto 是半自动的，ASIUM 是协作式的系统(cooperative system)[13]。全自动的本体学习系统，它们的适用性不强，很多方面受到限制而且和半自动或协作式系统相比性能比较差。换句话说协作式系统提供更多可接受的结果，因为该系统在学习的过程中使用者可做出一些解释决定，使用者的作用可以在很大的范围内变化，使用者可以选择最初的本体，在类关系中选择所需模式，处理噪声和标注新的概念。对于基于文本的信息提取系统更适合采用协作式，此系统的难点是受语言工具(标注器，分析器)的限制。

2.5 系统评价

目前对本体学习系统的评价还没有形成一个统一的评价标准。常用的评价方法[21~33]有以下两种：

(1)在同一个领域使用交叉评价的方法比较两个或多个本体。

(2)基于应用的评价，即通过应用来评价领域本体本身。

因为不同的本体学习系统学习的本体内容不同，对于不同的输入数据采用的方法不同，所以通过一种方法来比较它们的结果是很困难的。因此，许多本体学习系统都有自己的测试和评价方法，这些方法是基于本体所应用的环境和所选择的领域的。像很多本体学习系统[5]通过计算学习模型的查全率(recall)和查准率(precision)来评价学习系统。查全率是指正确概念的数量除以测试集中概念的总数，查准率是指正确概念的数量除以所提取概念的总数。显然在一个系统中查准率的值要大于查全率的值。系统的评价也可以采用概念级之间的比较，如对于分类关系的比较，综合考虑被比较概念的父子概念的相似度。对于非分类关系的比较，可以考虑被比较关系的 domain 和 range 的相似度。但这些结果仍然是不可以在系统间相互比较的，因这些系统用于不同的领域，用于不同的环境而且有不同的输入数据。

结束语 本体学习技术是当前研究的一个热点[3,5,10]，是利用本体工程技术和机器学习技术等众多学科技术实现本体的自动或半自动构建。这就为本体在电子商务、自然语言处理、知识工程和信息提取等众多领域的应用提供了一个平台，而且本体学习的研究对计算机网格和语义网的向前发展并最终普及应用起很大的推动作用。但是虽然目前对本体学习各个方面的研究做了一些工作，可仍有一些问题需要我们注意：一方面，许多对本体学习中概念间关系的学习都是针对分类关系而言的，对非分类关系研究却很少。另一方面，用本体学习来促进本体构建的方法多用于领域本体，而对于一般本体的自动构建仍然有大量的工作要做。而且许多本体学习系统都是在小的有限的领域中进行测试，还需要增加在实际应用中的工作。本体学习有面向文本、面向字典、面向知识库和面向关系数据库等等，本文只是对面向文本的本体学习中的一些关键技术做了介绍和总结，希望能对相关领域的研究人员在各自的研究中有所启迪和帮助。

参 考 文 献

1 Neches R, Fikes R E, Gruber T R. Enabling technology for knowledge sh-aring. AI Magazine, 1991, 12(3): 36~56
2 Gruber T R. A translation approach to portable ontology specifications. Kn-owledge Acquisition, 1993, 5(2): 199~220
3 Perez G, Macho M. A survey of ontology learning methods and techniq-ues. OntoWeb Deliverable D1. , 2003, 5: 1~86
4 Studer R, Benjamins V R, Fens-e. lD. Knowledge engineering, princ-eples and method. Data and Knowledge Engineering, 1998. 161~197
5 Maedche A, Staab S. Ontology l-earning for the semantic web. IEEE Intelligent Systems, 2001, 16(2): 72~79
6 Maedche A, Volz R. Discovering conceptual relations from text. In: Pr-oceedings of 14th European Conferenc-eon Artificial Intelligence, Berlin, 2000. 321~325
7 Maedche A, Staab S. Semi-autom-atic engineering of ontologies from te-xt. In: Proceedings of the 12th Intern-ational Conference on Software Engin-eering and Knowledge Engineering, Chicago, 2000
8 Gabel T, Sure Y, Voelker J. A KAON-Ontology management infrastru-cture. Institute AIFB, 2004
9 Navigli R, Velardi P. Learning d-omain ontologies from document ware-houses and dedicated web sites. Com-putational Linguistics, 2004, 30(2): 151~179
10 Navigli R, Velardi P, Gangemi A. Ontology learning and its application toautomated terminology translation. I- EEE Intelligent Systems, 2003, 18(1): 22~31
11 Missikoff M, Navigli R, Velardi P. Integrated approach to web ontology l-earning and engineering. IEEE Compu-ter, 2002, 35 (11), 60~63
12 Shamsfard M, Barforoush A A. Learning ontologies from natural lang-uage texts. International Journal of H-uman Computer Studies, 2004, 60(1): 17~63
13 Faure D, Nedellec C. A corpus-bas-ed conceptual clustering method for verb frames and ontology acquisition. In: Proc. LREC-98 Workshop on Adap-ting Lexical and Corpus Resources to Sublanguages and Applications, Europ-ean Language Resources Distribution Agency, Paris, 1998
14 Pantel P, De Kang- Lin. A statistic-al corpus-based term extractor. In: Pr-oceedings of the 14th Biennial Confer-ence of the Canadian Society on Com-putational Studies of Intelligence: Adv-ances in Artificial Intelligence, 2001. 36~46
15 Manning C D, Schutze H. Foundations of statistical natural langu-age processing. Computational Linguis-tics, 2000, 26: 1~3
16 De Kang-Lin, Pantel P. Concept disc-overy from text. In: Proceedings of the 19th International Conference on Computational Linguistics, 2002. 1~7
17 Hearst M A. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Internatio-nal Conference on Computational Ling-uistics, Nantes, France, 1992. 539~545
18 Buitelaar P, et al. A Protégé Plug-in f-or ontology extraction from text basedon linguistic analysis. In: Proceedings of the 1st European Semantic Web Sy-mposium, 2004
19 黄伟，金远平. 形式概念分析在本体构建中的应用. 微机发展, 2005, 15(2): 28~31
20 Cimiano P, Hotho A, Staab S. Learning concept hierarchies from text corpora using formal concept analysi-s. Journal of Artificial Intelligence Re-search, 2005, 24: 305~339
21 Corcho O, Gomez P A. Evaluatin-g knowledge representation and reason-ing capabilities of ontology specificati-on languages. In: Proceedings of the ECAI2000 Workshop on Applications of Ontologies and Problem-Solving M-ethods, 2000
22 Sabou M. Extracting ontologies fro-m software documentation: a semi-aut-omatic method and its evalution. In: Proceedings of the ECAI-2004 Works-hop on Ontology Learning and Popula-tion, Valencia, Spain, 2004
23 Kietz J U, Maedche A, Volz R. Amethod for semi-automatic ontology a-cquisition from a corporate intranet. In: Proceedings of the EKAW 2000 workshop on Ontologies and Texts, Fr-ance, 2000
24 刘柏嵩，高济. 面向知识网格的本体学习研究. 计算机工程与应用, 2005, 20: 1~5

25 裴柄镇,陈晓明,胡褶,等.一种建立中文概念分类关系的新算法.计算机工程与应用,2004,36:18~21
26 Wang B B, Mckay R I(Bob), Abbass H A, et al. Learning text classifier using the dom-ain concept hierarchy. IEEE, 2002. 1230~1234
27 Zheng De-Quan, Zhao Tie-Jun, Yu Fe-ng, et al. Machine learning for automatic acquisition of Chinese lingu-istic ontology knowledge. IEEE, 2005. 3728~3733
28 Shauntrelle D D, Tia B W. Engi-neering knowledge. In: Proceedings of the 42nd Annual Southeast Regional Co-nference, Huntsville, Alabama, 2004. 406~407
29 Suryanto H, Compton P. Learning classification taxonomies from a classi-fication knowledge based system. In: Proceedings of the ECAI 2000 Works-hop on Ontology Learning (OL'2000), 2000
30 Bisson G, Nedellec C, Canamero D. Designing clustering methods for ontology building: The Mo'K workbe-nch. In: Proceedings of the ECAI 2000 Workshop on Ontology Learning (OL'2000), 2000
31 Lenat G. Building large knowledgebased system: representation and infer-ence in the CYC project. 1st edition. Boston: Addison Wesley Press, 1989
32 Snasel V, Moravec P, Pokorny J. WordNet ontology based model forweb retrieval. IEEE, 2005. 220~225
33 Gruninger M, Fox M. Methodologyfor the design and evaluation of ontol-ogies. In: Proceedings of the IJCAI 95 Workshop on Basic Ontological Iss-ues in Knowledge Sharing, 1995
34 Emde W, Wettschereck D. Relatio-nal instance based learning. In: Proc-eedings of 13th International Confere-nceon Machine Learning (ICML'96), 1996. 122~130
35 Yamaguchi T. Acquiring conceptual relations from domain-specific texts. In: Proceedings of the IJCAI 2001 W-orkshop on Ontology Learning, 2001
36 Maedche A, Staab S. Ontology le-arning. In: Proceedings of 14th Euro-pean Conference on Artificial Intellig-ence, 2000
37 Craven M, DiPasquo D, Freitag D, et al. Learning to constru-ct knowledge bases from the World Wide Web. Artificial Intelligence, 2000. 69~113
38 Maedche A, Volz R. The Text-To-Onto ontology extraction and mainten-ance environment. In: Proceedings of the ICDM Workshop on Integrating Data Mining and Knowledge Manage-ment, California, 2001
39 Wu S H, Hsu W L. SOAT: A semi-automatic domain ontology acqui-sition tool from Chinese corpus. In: the 19th International Conference on Computational Linguistics, Howard In-ternational House and Academia Sinic-a, Taipei, Taiwan, 2002
40 Chaelandar G, Grau B. SVETLAN'a system to classify words in context. In: Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany, 2000

(上接第 151 页)

(4) 算法的第四步是删除冗余的元素。这是因为在运用变换规则的时候,有可能使得某些元素既没有属性也没有任何元素,而且也不可以取字符串值。此时,可以根据实际应用的需求删除这些冗余的元素。

例 2 为了避免 DTD *D*1 中存在多值依赖的数据冗余,可以把 *D*1 通过变换规则和 MVD 无损联接分解算法,成为如下 DTD *D*2:

〈! ELEMENTdepartment (teacher * ,student *)〉
〈! ATTLIST department dname CDATA # REQUIRED〉
〈! ELEMENT student EMPTY〉
〈! ATTLIST student sname CDATA # REQUIRED〉
〈! ELEMENT teacher EMPTY 〉
〈! ATTLIST teacher tname CDATA # REQUIRED〉

例 2 中的 DTD *D*2 对应的一个 XML 文档树,如图 3 所示。

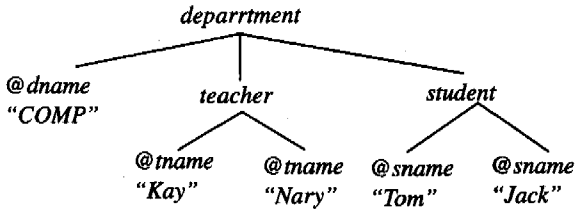

图 3 符合 D2 的一个 XML DTD 文档

通过这种变换,把直接联系的实体放在一起形成嵌套关系,在此处是把教师和学生都作为系的子元素类型,从而消除了 *D*1 中存在的数据冗余和操作异常。

通过此算法,我们可以获得满足 MXNF 的 DTD 文档。这样的文档很好地消除了当存在 MVD 时的数据冗余,而且文档有较好的结构。但是,通过这一算法得到的 DTD 并不一定保持依赖,这与关系数据库中的 4NF 的分解很相似,这里不再对此详细实证[9]。

结论与进一步的工作 本文研究了 XML 文档中的多值依赖问题,讨论了当 XML DTD 中存在 MVD 时的规范化问题,分析了 XML 文档中由多值依赖引起的数据冗余和各种操作异常现象。提出了 DTD 的多值依赖的概念,定义了 XML DTD 的一种范式 MXNF,给出了把 DTD 无损联接地分解成符合 MXNF 的算法,它不仅消除了 XML DTD 文档中的数据冗余和各种异常,而且更好表达了现实世界中实体的语义关系。此外,对于 XML DTD 相关的多值依赖的变换规则以及如何充分发挥 XML DTD 的特性来更好地描述数据,提出 MXNF 规范化算法。这将对未来的 XML 函数依赖保持、XML 完整性约束、推理规则、XML 多值依赖以及 XML 模式的进一步规范化研究奠定理论基础[10]。

参 考 文 献

1 Arenas M, Libkin L. A Normal Form for XML Documents. In: Symposium on Principles of Database Systems (PODS'02), Madison, Wisconsin, U. S. A. ACM Press, 2002. 85~96
2 Provost W. Normalizing XML. http:// www. xml. com/pub/a/2002/11/13/normalizing. html
3 Lee Mong Li, Ling Tok Wang, Low Wai Lup. Designing Functional Dependencies for XML. In: VIII Conference on Extending Database Technology (EDBT), Prague, March 2002
4 谈子敬,施伯乐. DTD 的规范化. 计算机研究与发展,2004,41(4):594~601
5 Vianu V. A Web Odyssey: from Codd to XML. In: Proceedings of ACM PODS, Santa Barbara CA USA, 2001. 148~160
6 张忠平,王超,朱扬勇. 基于约束的 XML 文档规范化算法. 计算机研究与发展,2005,42(5):755~764
7 Tan Zi-Jing, Shi Bo-Le. Propagating Functional Dependency and Normalization Between Relations and XML. Journal of Software, 2005,16(4):533~539
8 吕腾,顾宁,闫萍等. XML 文档的范式. 小型微型计算机系统[J], 2004,10(25):1836~1840
9 Fan W, Libkin L. On XML Integerity constraints in the presence of DTDs. In: Proceedings of ACM Symposium on Principles of Database Systems (PODS), Santa Barbara, California, May 2001. 114~125
10 Fan W, Simen J. Integerity constraints for XML. In: Proceedings of ACM Symposium on Principles of Database Systems(PODS), Dallas, Texas, May 2000. 23~34