ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.9, September 2006, pp.1837−1847
DOI: 10.1360/jos171837
© 2006 by *Journal of Software*. All rights reserved.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

本体学习研究综述*

杜小勇[1,2+], 李 曼[1], 王 珊[1,2]

[1](中国人民大学 信息学院,北京 100872)
[2](教育部数据工程与知识工程重点实验室,北京 100872)

A Survey on Ontology Learning Research

DU Xiao-Yong[1,2+], LI Man[1], WANG Shan[1,2]

[1](School of Information, Renmin University of China, Beijing 100872, China)
[2](Key Laboratory of Data Engineering and Knowledge Engineering, Ministry of Education, Beijing 100872, China)
+ Corresponding author: E-mail: duyong@ruc.edu.cn

Du XY, Li M, Wang S. A survey on ontology learning research. *Journal of Software*, 2006,17(9):1837−1847.
http://www.jos.org.cn/1000-9825/17/1837.htm

Abstract: Recently, ontology learning is emerging as a new hotspot of research in computer science. In this paper the issue of ontology learning is divided into nine sub-issues according to the structured degree (structured, semi-structured, non-structured) of source data and learning objects (concept, relation, axiom) of ontology. The characteristics, major approaches and the latest research progress of the nine sub-issues are summarized. Based on the analysis framework proposed in the paper, existing ontology learning tools are introduced and compared. The problems of current research are discussed, and finally the future directions are pointed out.
Key words: ontology; ontology learning; concept; relation; axiom

摘 要: 近年来,本体学习技术逐渐成为计算机科学领域的一个研究热点.根据数据源的结构化程度(结构化、半结构化、非结构化)以及本体学习对象的层次(概念、关系、公理),将本体学习问题划分为 9 类子问题.分别阐述了这 9 类问题的基本特征、常用的方法和最新的研究进展,并在此分析框架下进一步介绍和比较了现有的本体学习工具.最后,讨论了存在的问题,指出了未来的研究方向.
关键词: 本体;本体学习;概念;关系;公理
中图法分类号: TP182 文献标识码: A

近年来,在计算机科学中关于本体的研究越来越多.所谓本体,最著名并被广泛引用的定义是由 Gruber 提出的“本体是概念模型的明确的规范说明”[1].通俗地讲,本体是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,这样,人机之间以及机器之间就可以进行交流.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[2].

在过去的 10 年里,已经出现了许多本体构建工具,从最早的 Ontolingua[3],OntoSaurus[4],WebOnto[5],到

* Supported by the National Natural Science Foundation of China under Grant Nos.60496325, 60573092 (国家自然科学基金)
Received 2005-11-22; Accepted 2006-02-23

© 中国科学院软件研究所 http://www.jos.org.cn

Protégé-2000[6],WebODE[7],OilEd[8],OntoEdit[9],以及 KAON[10]等,本体构建工具也日趋成熟.这些工具提供了友好的图形化界面和一致性检查机制.借助这些工具,用户可以把精力集中在本体内容的组织上,而不必了解本体描述语言的细节,而且避免了很多错误的发生,方便了本体的构建.但是,这些工具提供的仅仅是本体编辑功能,支持的仍然是手工构建本体的方式.即使使用这些本体编辑工具,用户依然需要逐个地输入和编辑每个概念的名字、约束、属性等内容.现有的大部分系统,例如 Cyc[11]和 Mikrokosmos[12]等,都是靠手工输入大量的知识,然后才能基于这些知识进行推理或获取新的知识.由于手工方法费时、费力,使得本体的构建成为一项艰巨的任务.因此,如何利用知识获取技术来降低本体构建的开销是一个很有意义的研究方向.目前,国外在该方向的研究很活跃,把相关的技术称为本体学习(ontology learning)技术,其目标是利用机器学习和统计等技术自动或半自动地从已有的数据资源中获取期望的本体.由于实现完全自动的知识获取技术还不现实,所以,整个本体学习过程是在用户指导下进行的一个半自动的过程.

本体的结构(ontology structure)是一个五元组[13]$O:=\{C,R,H^c,Rel,A^o\}$.这里的 C 和 R 是两个不相交的集合,其中:C 中的元素称为概念(concept);R 中的元素称为关系(relation);H^c 表示概念层次,即概念间的分类关系(taxonomy relation);Rel 表示概念间的非分类关系(non-taxonomy relation);A^o 表示本体公理(axiom).从本体的结构可以看出,本体学习的任务包括概念的获取、概念间关系(包括分类关系和非分类关系)的获取和公理的获取.这 3 种本体学习对象构成了从简单到复杂的层次.

现实世界中的数据种类很多,例如纯文本以及 XML,HTML,DTD 等,大部分都可以作为本体学习的数据源.针对不同类型的数据源需要采用不同的本体学习技术,所以本文根据数据源的结构化程度,将本体学习技术分为 3 大类:基于结构化数据的本体学习技术、基于非结构化数据的本体学习技术和基于半结构化数据的本体学习技术.

本文第 1 节~第 3 节分别介绍这 3 大类本体学习技术在本体学习对象的 3 个层次上常用的方法和研究进展.第 4 节介绍并分析比较几个具有代表性的本体学习工具.第 5 节讨论目前本体学习技术中存在的问题以及未来的研究方向.

1 基于结构化数据的本体学习

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.随着数据库在信息管理领域的广泛应用,大量的数据通常存储在数据库中.Lawrence 和 Giles 在 1998 年时估计互联网上有 80%的内容存储在 Hidden Web 中[14].所谓的 Hidden Web 中的数据就是存储在数据库中,而且这些数据一般都是面向主题(领域)的.因此,如何利用数据库中丰富的数据构建本体是一个很有意义的研究课题.

首先看一下关系数据库.众所周知,关系数据库采用的是关系模型,它是对领域信息建模的一种经典模型.这种模型结构简单,二维关系表格形式容易被理解,关系代数理论强有力地支持了关系模型,使得关系数据库得以广泛应用[15].现有的应用大多采用关系数据库来组织和存储数据.在关系模型中,关系(relation)是元组的集合;而关系模式(relation schema)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.所以说,在关系数据库中,关系模式是型,元组集(即关系)是值.与关系模型相比,本体是一种具有更多语义、结构更为复杂的模型.所以,这类本体学习的主要任务就是分析关系模型中蕴涵的语义信息,将其映射到本体中的相应部分.

在关系模型中,实体以及实体间的联系都是用表来表示的.所以,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的联系的,然后才能将实体信息映射为本体中的概念,将联系信息映射为本体中的关系.实际上,早在 20 世纪 90 年代,研究者们就已经开始关注如何自动分析关系模型的语义了.当时的研究动机是他们认为关系模型所能描述的语义信息太少,即它不能用一张表模型表示出复杂对象的语义,从而不适合于对数据类型繁多而语义复杂的领域信息系统的建模.所以,他们提出了将关系模型重新设计成更复杂的结构(例如面向对象模型).在此期间,他们给出一系列技术来获取关系模型的语义结构,并对其重新设计,这些技术被称为关系数据库的逆向工程(relational database reverse

© 中国科学院软件研究所 http://www.jos.org.cn

engineering)[16].这些研究成果中很多都可以用于从关系数据库中获取本体.例如,1994 年,Johannesson[17]提出将关系模型转换为一个概念模型,该概念模型实际上是一个扩展的实体-关系模型的形式化表示,然后由用户对该概念模型进行修订生成最终的本体.由于已有的关系数据库的逆向工程技术都没有考虑到如何将关系模型直接转换成本体,所以 2002 年,Stojanovic 等人[18]通过考察数据库中的表、属性、主外键和包含依赖关系,给出了一组从关系模型到本体的映射规则.基于这些规则能够直接得到一个候选本体,然后可以进一步对该候选本体进行评价和精炼,生成最终的本体.

对于公理的获取,目前还没有查到相关的研究成果.本文认为可以利用数据库中定义良好的结构来获取一些简单的公理.例如,如果数据库中的某个属性具有 Not Null 约束,则可以得到在本体中相应的关系在其对应的类中的 Mincardinality 为 1.除此之外,还可以通过发现属性间的依赖关系来获取公理.例如,假设数据库模式满足 3NF,如果存在两个表 R_i 和 R_j 都具有属性 A,且 A 不是 R_i 的主码,满足 R_i 表中的属性 A 包含依赖于 R_j 表中的属性 A,则可以将 A 映射成一个对象属性 P,且其 domain 和 range 分别是表 R_i 和 R_j 对应的类.该规则表明:如果关系中的某个属性只是用来描述两个关系之间的参照关系,那么可以将其映射成本体中的一个对象属性.

可以看出,现有的研究主要集中在对关系模式进行语义分析,从而获取构建本体所需的概念和关系.由于关系模式中蕴涵的语义十分有限,所以这些方法只能用来构建轻量级的本体(即结构较简单的本体).为此,1999 年 Kashyap[19]提出首先根据关系模式得到一个初步的本体,然后基于用户查询进一步丰富该本体中的概念和关系.由于用户查询具有很大的随机性,所以很难保证结果的质量.实际上,一种更为可行的方法是分析数据库中的元组,得到更多隐含的语义信息.2004 年,Astrova[20]已经通过对元组的分析,得到了概念间的“继承”关系.另外,本文认为还可应用一些基于关系数据库的数据挖掘技术[21],例如概念层次的发现等,来改进这类本体学习技术.

值得强调的是,上述方法的前提都是已知数据库的模式信息,然而在很多情况下,这些信息无法直接获得.此时,如何发现数据库的语义是很有意义的研究课题.2004 年,Astrova 等人[22]提出由于 HTML 表格是 Web 上用户和数据库交互最常用的界面,所以在无法获得数据库模式信息的情况下,可以通过分析这些 HTML 表格的结构和数据来获取关系数据库的语义,从而构建本体.在这方面,最近关于 Hidden Web 的一些研究成果[23]可以借鉴.总之,从关系数据库中学习本体仍然有很多工作可以做.

除了可以从关系模型中获取本体,也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.

2 基于非结构化数据的本体学习

非结构化数据是指没有固定结构的数据.其中,纯文本是 Web 中大量存在的一类非结构化数据,也是最重要的一类,可以用来获取本体的数据源.目前,基于非结构化数据的本体学习技术的研究主要集中在从纯文本中获取本体.纯文本依据一定的造句法表达特殊的语义,使得读者可以基于一些背景知识来理解其中的含义.然而,由于缺乏一定的结构,要使机器能够自动地理解纯文本并从中抽取出所需要的知识,则必须利用自然语言处理(NLP)技术对其预处理,然后利用统计、机器学习等手段从中获取知识.

对于概念的获取,现有的方法可以分为 3 类:基于语言学的方法、基于统计的方法和混合方法.(1) 基于语言学的方法[38]主要根据领域概念的特殊词法结构或模板,寻找和抽取结构符合这些特定模板的字符串.由于这些模板在大多数情况下是与具体语言相关的,因此,这类方法要求针对具体的语言作相应的处理;(2) 基于统计的方法[24−27]主要根据领域概念与普通词汇拥有不同的统计特征(例如,领域相关性和领域通用性),以鉴别出领域概念.大多数基于统计的方法关注于多字词汇(multi word unit,简称 MWU)的抽取,主要方式是计算各组成部分之间的联系程度;(3) 混合方法[28,29]往往是结合语言学和统计学的技术,有的是在统计处理之后采用语法过滤器,以便抽取出经过统计计算有意义的、与给定词法模板匹配的词汇组合;有的则是首先采用语言技术选出候选项,然后再用统计方法对这些候选项进行计算.

与国外相比,国内在领域概念(也称为专业术语)的自动抽取方面,特别是中文领域概念的自动抽取的研究

© 中国科学院软件研究所 http://www.jos.org.cn

工作相对较少.在 2003 年的第 7 届全国计算语言学联合学术会议上,东北大学的陈文亮等人[39]提出利用 Bootstrapping 的机器学习技术,从大规模无标注真实语料中自动获取领域词汇.2005 年,山西大学郑家恒等人[40]提出采用非线性函数与“成对比较法”相结合的方法,综合考虑位置和词频两个因素,给出候选词的权重,实现了关键词的自动抽取.2005 年,上海交通大学的杜波等人[41]提出了一种将统计方法与规则方法相结合的专业领域术语抽取算法.

值得强调的是,无论国内还是国外,统计方法都是主流.我们也曾经尝试着将已有的这些方法应用到经济学领域中,希望能够自动的抽取出中文经济学概念,但结果却不理想.其中的主要困难在于如何识别概念的领域相关性.从理论上讲,可以通过计算概念在领域相关的文本集中出现的频率与其在普通文本集中出现的频率的比值来判断概念的领域相关性,即如果该比值大于指定的阈值,则说明该概念在某个领域中经常出现,而在其他领域中不常用.但是,该方法的结果受普通文本集质量(主要指内容和规模)的影响很大,从而影响了该方法的实际可行性.

对于概念间关系的获取,常用的方法有:基于模板的方法、基于概念聚类的方法、基于关联规则的方法、基于词典的方法,或者这些方法的混和.(1) 基于模板的方法[28,30]是指通过分析领域相关文本,总结出一些频繁出现的语言模式作为规则,然后判断文本中词的序列是否匹配某个模式——如果匹配,则可以识别出相应的关系.例如:可以将一个非常简单的字符串匹配(* is *)作为一个模式,那么,满足该模式的一对概念就可以认为具有“isa”关系.这些模式可以是手工定义的,也可以是从某些样本句子中学习得到的.这类方法的主要缺点是准确度低,因为大量无用的概念对往往也会匹配这些模式,而且模式的获取是否完备对于获取效果影响较大;(2) 基于概念聚类的方法是利用概念之间的语义距离,对概念进行聚类.这样,同一类簇中的概念具有语义近似的关系.同时,也可以进行层次聚类,聚类的结果就是概念间的分类关系.关于概念层次聚类的研究有很多,例如,Fisher[31]提出了一种基于矢量的聚类方法,Bisson[32]和 Emde 等人[33]提出了基于 FOL 的聚类方法.这些方法共同的局限性是只能得到概念间严格的层次关系(即树状的层析结构),然而在本体中一个概念却可以有多个父概念.为此,Faure 等人[34]采用宽度优先的方法对概念进行逐层聚类,较为特殊的是,它在进行每层聚类的时候都要考虑所有的簇而不管这些簇所在的层次.显然,该方法还有一个附加的约束,即一个簇不能和它的父簇进行聚类.这样得到的结果是一个无环图,图中两个结点间的连线表示概念间的层次关系;(3) 关联规则挖掘的方法常用于获取概念间的非分类关系,其基本思想是:如果两个概念经常出现在同一文档(或段落,或句子)中,则这两个概念之间必定存在关系.2000 年,Maedche 等人[35]最先描述并评价了将关联规则应用于本体学习的方法.2001 年,Maedche 等人[36]又提出使用已有的概念层次作为背景知识,然后利用关联规则来发现概念间的非分类关系的方法;(4) 基于词典的方法往往根据一些现有的词汇词典中定义的同义词、近义词和反义词等知识来获取本体中概念间的关系.例如,Nakaya 等人[37]使用 WordNet 来获取概念间的分类关系;(5) 混和方法往往是同时使用上述若干种方法,以期得到更好的结果.其中比较特殊的方法是由 Missikoff 等人[26]和 Navigli 等人[27]提出的,他们提出利用机器学习技术基于已有的通用本体对抽取出来的术语进行语义解释,即为这些术语关联上明确的概念标识符;然后,基于这些语义解释来确定概念之间的分类和相似关系,生成一个领域概念森林.与其他方法相比,该方法的主要特点是对术语进行语义解释,然后使用这些语义解释来获取除分类关系以外的其他概念间的关系,而其他方法都是将术语等同于领域概念.这种做法的好处是可以确定复杂术语的正确含义及其语义关系.对于一个复杂术语,该方法首先确定与该术语的各个组成成分相对应的概念,然后根据这些概念间的语义关系来构造相应的复杂概念.该步骤的结果是得到一个领域概念森林,它表示了这些复杂概念间的分类关系和其他关系.

到目前为止,国际上对概念间关系获取的研究很多,但是,对概念间非分类关系的获取,大部分方法都停留在判断两个概念之间是否存在关系的层次上,无法进一步为获取的关系赋予相应的语义标签,即得到的都是“匿名”关系.为此,2005 年,Kavalec 等人[42]提出使用扩展的关联规则挖掘方法为本体中概念间的非分类关系赋予语义标签.其基本思想是:如果两个概念间存在非分类关系,那么该关系能够用经常出现在这两个词附近的某个动词来表示.所以,可以通过计算某个动词和某两个概念一起出现的条件概率决定这两个概念之间的关系是否可

© 中国科学院软件研究所 http://www.jos.org.cn

以用该动词来表示.Kavalec 等人的方法是对解决该问题的一个初步尝试,但它仅考虑了词频,没有考虑句子结构等其他因素,所以结果并不十分理想.

对于公理的获取,研究成果很少,目前查到的只有 Shamsfard 等人[38]提出的基于模板的抽取方法,即在对句子结构分析的基础上,应用预先定义的模板——如果与模板匹配,则得到相应的本体公理.该方法的局限性很明显,它不仅需要人工预先制定模板,而且无法获取隐含的公理.

3 基于半结构化数据的本体学习

半结构化数据是指具有隐含结构,但缺乏固定或严格结构的数据[13].Web 中的半结构化数据很多,例如大量的 XML 格式和 HTML 格式的网页,以及它们遵循的文档类型定义(XML schema 或 DTD),还有越来越多的用 RDF 标注的网页,都可以作为本体学习的数据源.

由于这类数据是介于结构化和非结构化数据之间的一类数据,所以基于上述两种数据类型的本体学习技术也可以应用到这类数据源.对于 XML,HTML 和 RDF 等格式的网页,可以直接使用那些从纯文本中获取本体的方法.例如,Papatheodorou 等人[47]给出的从 XML 或 RDF 格式的文档中获取概念间分类关系的方法,就是首先抽取出表示每篇文档内容的关键词,然后基于这些关键词使用聚类技术,将文档集分成不同的组,保证同组内的文档内容是相似的;接着,使用统计的方法选出最能表达每组文档内容的关键词;将这些关键词作为本体中的概念,并根据先前聚类的结果给出概念间的分类关系.实际上,由于半结构化数据具有隐含的结构,所以在获取本体的过程中,可以利用这些隐含的结构信息来改善本体学习的结果.例如在进行领域概念抽取时,可以根据文档中的标签区分概念出现的位置,然后通过在传统的统计公式上增加关于位置信息的权重来提高概念抽取的准确度.

对于模式语言(例如 XML schema 或 DTD),因为它们描述了 XML 数据的层次结构,通常认为它们是 XML 的逻辑模型.所以类似于从结构化数据中学习本体,对于这些数据通常采用映射技术,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类:一类是基于学习的方法,即利用一些自学习的手段自动获取,例如 Kavalec 等人[42]重点研究了利用机器学习方法自动地得到映射规则;另外一类是基于预定义规则,即用户预先给出了一些规则,例如,Doan 等人[43]和 Mello 等人[44]使用预定义的规则,从 DTD 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.但是,由于各种模式语言在语法上的差异,需要使用不同的映射规则.为此,Volz 等人[53]提出将这些半结构化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,开始符集和规则集;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,该方法克服了现有模式语言(XML schema 和 DTD)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,例如 key,unique,keyref(key reference)等.

实际上,机器可读的词典(MRD)也是一种特殊的半结构化数据.作为一种通过手工方式认真组织的可靠的领域知识资源,它们也是一种非常好的本体学习数据源.这类数据源的内部结构虽然在很大程度上也是一种纯文本,但对于领域概念及其关系的抽取来说,仍有很多规律可循.所以,对于它们通常使用基于语言学的方法和基于模板的方法.例如:Litkowski[45]通过对词典中每个定义的分析,获取概念之间的分类关系;Rigau 等人[46]使用一组预定义的词典语法模板自动地从词典中发现词与词之间的上下位关系.

另外,随着语义 Web 的发展,Web 中会出现越来越多的用 OWL,RDF(S)等语言描述的本体,它们也是一种半结构化的数据.如何从已有的本体中学习新的本体也是当前国际上比较重视的一个课题,这其中更多地涉及到本体的合并、本体的映射等问题.由于篇幅有限,本文不作讨论.

4 本体学习工具

以上章节分别介绍了基于结构化数据、非结构化数据和半结构化数据的本体学习技术中常用的方法和最新的研究进展.其中一些方法已经应用到本体学习工具之中.由于完全自动的本体学习技术还不现实,所以,现

© 中国科学院软件研究所 http://www.jos.org.cn

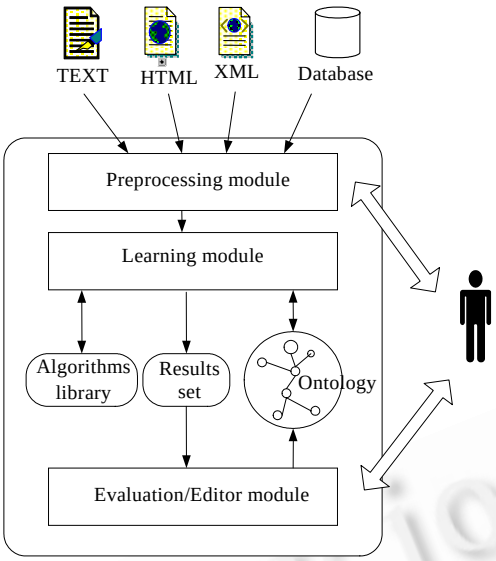

Fig.1 The basic framework of ontology learning tools

图 1 本体学习工具的基本框架

有的本体学习工具都是半自动的.本体学习工具的基本框架如图 1 所示.

图 1 表明,本体学习工具的输入可以是各种类型的数据源.在此框架中,预处理模块首先对数据源进行预处理;接着,学习模块通过使用算法库中的各种本体学习算法从上一步预处理的结果中获取本体;然后,将结果作为候选本体呈现给用户;最后,用户在评价/编辑模块的帮助下对该候选结果进行评价和确认,并将最终的结果添加到本体库中.可以看出,整个过程是在用户参与下的半自动的过程.需要注意的是,学习模块在获取本体的过程中需要参照已有的本体.目前,一些算法已经提出可以利用已有的本体作为背景知识来提高本体学习的精度[13].另外,如果在本体学习之前已经存在了一个初步的本体,那么在将本体学习结果添加到其中时,还要考虑到新添加的知识和已有的知识是否存在不一致性等问题.这就涉及到本体的进化问题,本文不作讨论.

到目前为止,国外已经开发了许多本体学习工具.虽然这些工具的系统框架在细节上各不相同,但都遵循如图 1 所示的基本框架和处理流程.本文重点介绍几个具有代表性的工具:Hasti,OntoLearn,Text-To-Onto,Onto Builder 和 OntoLiFT.

4.1 工具简介

Hasti[38]是 Amir Kabir University of Technology 开发的一个本体学习工具.其主要特点是:使用一个初始的核心本体,然后基于该核心本体自动地从纯文本中获取新的概念、关系和公理,从而不断地扩充这个初始的核心本体.它是为数不多的一个能够获取本体公理的工具.需要注意的是,它使用的这个核心本体是领域独立的,其中包括很少量的手工定义的概念、分类关系、非分类关系和公理.使用该核心本体的主要目的是便于对一些新获取的概念、关系和公理在本体中进行预定位.根据输入的纯文本的领域,Hasti 学习的结果可以是通用本体也可以是领域本体.该工具使用了多种本体学习方法:在获取概念时,它使用了基于语言学的方法;在获取概念间关系时,它使用了基于概念聚类(包括层次聚类和非层次聚类)的方法和基于模板的方法;在获取公理时,它使用基于模板的方法.除此之外,它还使用了启发式的学习方法,即在本体学习过程中,当同时出现多个可能的候选结果时,它利用一些启发式的规则来减少假设空间,消除不确定性.所以说,Hasti 使用的是一种启发式的、混和的本体学习方法.目前,它已经可以做到从波斯文本中获取本体.

OntoLearn[26,27]是 University of Rome 开发的一个基于文本的本体学习工具,它能够获取概念及其关系.其主要特点是:将语义解释的方法应用到本体获取中,即首先使用基于语言学和统计的方法从一组文本集中抽取出领域相关的术语,然后使用通用本体中的概念对这些术语进行语义解释,从而确定术语之间的分类和其他语义关系.OntoLearn 选择 WordNet 作为通用本体,使用 WordNet 中的概念对获取的术语进行语义解释,从而使所构建的领域本体与 WordNet 具有明确的关系,这样的好处是有利于不同领域本体之间的互操作和一致化.

Text-To-Onto[13,48]是 University of Karlsruhe 开发的一个整合的本体学习工具.其主要特点是可以支持从多种数据源中获取本体.目前,它已经可以做到从非结构化数据(纯文本)和半结构化数据(HTML,词典)中获取概念及其关系.对于从非结构化数据中学习本体,它使用加权的词频统计方法来获取概念,使用基于概念层次聚类法来获取分类关系,使用基于关联规则的方法来获取非分类关系;对于 HTML 数据,它将其预处理成纯文本,然后利用基于非结构化数据的本体学习方法从中获取本体;对于词典,它使用基于模板的学习方法.该系统能够处理德文和英文的数据源.

© 中国科学院软件研究所 http://www.jos.org.cn

OntoBuilder[49,50]是 Mississippi State University 开发的一个从 XML 和 HTML 中获取本体(包括概念及其关系)的工具.它看起来像一个 Web 浏览器.当使用它来获取本体之前,需要手工构建一个初始的领域本体;然后,在用户浏览包含相关领域信息的网站的过程中,该工具会为每个网站生成一个候选本体[51];最后,在用户的参与下将这些候选本体与初始本体合并.其中,使用的本体学习方法主要是词频统计和模式匹配(包括子串匹配、内容匹配、词典匹配).OntoBuilder 可以支持英文的网页,但在实际中,它并不能适用于所有的网站,因为有些网站包含了它不支持的技术,例如带有脚本(scripting)的网页.

OntoLiFT[52]是 University of Karlsruhe 开发的一个从半结构化数据(XML schema,DTD)和结构化数据(关系数据库)中获取本体(包括概念及其关系)的工具.对于这两种类型的数据源,它都采用基于映射规则的方法来获取本体.在系统实现中,从 XML Schema 和 DTD 中获取本体的部分是基于一个已有的工具(hMarfra).HMarfra 能够实现从 XML Schema 到本体的映射.然后,OntoLift 开发了一个从 DTD 到 XML Schema 映射的中间工具.这样,将这两个工具合并起来,就实现了从 XML Schema 和 DTD 中获取本体.从关系数据库中获取本体的部分是基于 Java JDBC 标准提供的接口,然后按照一定的命名规范将数据库中的表名和属性名等信息,按照映射规则转换为本体中的元素.

4.2 工具的比较分析

除了本文提到的这 5 个具有代表性的工具外,还有许多各具特色的本体学习工具.由于目前还没有统一的评价标准,很难对它们进行定量的评价.一般来说,本体学习工具之间的主要区别在于:

(1) 数据源:即本体学习工具的输入数据的种类,例如纯文本、Web 页面、机器可读的词典等;

(2) 学习方法:即本体学习工具为了从数据源中获取本体所采用的主要方法,例如:统计方法、机器学习方法和模式匹配等方法;

(3) 本体学习对象:即本体学习工具从数据源中学习到的本体对象,主要包括概念、概念间关系和公理.

根据上述 3 个方面,现将本文介绍的 5 种本体学习工具进行总结(见表 1).表 1 给出了在本文的分析框架下,现有工具对各类本体学习子问题的支持情况及采用的本体学习方法.表中给出的工具名称表示该工具可以支持从相应类型的数据源中获取相应的本体学习对象,工具名称后的括号内给出的是该工具针对该类本体学习子问题所采用的主要方法.

Table 1 Summary of ontology learning tools

表 1 本体学习工具总结

Ontology learning objects	Data resource		
	Structured data	Unstructured data	Semi-Structured data
Concept	OntoLiFT {mapping rules}	Hasti{linguistic analysis / heuristic rules} OntoLearn{linguistic analysis / statistical frequencies } Text-To-Onto{statistical frequencies}	OntoBuilder{statistical frequencies / matching} OntoLiFT{mapping rules} Text-To-Onto{statistical frequencies}
Relation	OntoLiFT {mapping rules}	Hasti{cluster / template / heuristic rules} OntoLearn{semantic interpretation} Text-To-Onto{hierarchical concept clustering / association rules}	OntoBuilder{statistical frequencies / matching} OntoLiFT{mapping rules} Text-To-Onto{hierarchical concept clustering / association rules / template}
Axiom	——	Hasti{template / heuristic rules}	——

表 1 表明:

(1) 支持从非结构化和半结构化数据中获取概念和概念间关系的工具比较多;支持从结构化数据中获取概念和关系的工具只有 OntoLiFT.说明虽然在数据库领域中关于从关系模型中抽取语义或将关系数据模型转换为更复杂模型的研究由来已久,但将这些成果应用于本体学习中的研究还较少;

(2) 支持从结构化数据和半结构化数据中获取公理的工具还没有;支持从非结构化数据中获取公理的工具只有 Hasti.它是 2004 年开发出的一个本体工具,这表明关于本体公理的获取已经逐渐引起人们的注意.虽然 Hasti 中的公理获取方法还有很多不足,但它为今后的相关研究提供了一条思路;

© 中国科学院软件研究所 http://www.jos.org.cn

(3) 这些工具都仅能支持基于某些类型的数据源的本体学习.例如:Hasti 和 OntoLearn 支持的数据源只有纯文本;Text-To-Onto 支持的数据源只有纯文本和某些半结构化数据;OntoBuilder 支持的数据源只有 XML 和 HTML;OntoLiFT 支持的数据源只有 XML Schema,DTD 和关系数据库,说明目前还没有一个整合的本体学习系统;

(4) 大部分工具使用的本体学习方法都比较单一.说明这些工具都只能在某些情况下取得较好的结果.因为任何一种本体学习方法都无法适用于所有的情况,为了提高工具的适用范围,必须利用多种本体学习方法.将获得的结果有效地综合起来,从而保证在大部分情况下都能获得比较理想的结果.资料表明,Text-To-Onto 和 Hasti 都在朝这个方向努力.

另外,本体学习中的很多技术都依赖于对自然语言的处理.所以,本体学习工具具有很强的语言特征.其中:Hasti 支持波斯语;Text-To-Onto 支持英语和德语;OntoLearn,OntoBuilder 和 OntoLiFT 都仅支持英语.目前还没有一个能够支持中文的本体学习工具.

5 存在的问题与未来的研究方向

本文根据数据源的结构化程度(结构化、半结构化、非结构化)以及本体学习对象的层次(概念、关系、公理),将本体学习问题划分为 9 类子问题,分别阐述了这 9 类问题的基本特征、常用的方法和研究进展,并分析比较了现有的本体学习工具.从中可以看出:本体学习虽然是一个新兴的研究领域,但是许多相关领域的研究成果都可以供其借鉴.其中,自然语言处理技术是本体学习的基础.除此之外,领域概念的识别、Web 数据的抽取、数据库的逆向工程、机器学习等技术都极大地促进了本体学习领域的发展.然而,由于本体学习任务自身的特殊性,该领域仍然存在许多有待解决的问题.总结起来有以下几个方面:

- 对本体学习方法的改进

虽然目前已经提出了很多本体学习方法,但大部分方法都不理想.就基于结构化数据的本体学习来说,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.就基于非结构化数据的本体学习来说,它是目前研究较多的一大类问题,但是仍然没有一个成熟的领域概念获取方法,并且无法自动地为非分类关系赋予语义;就基于半结构化数据的本体学习来说,现有的方法往往是将其按照纯文本对待,没有充分地利用其隐含的结构信息;从本体学习对象的层次来看,现有研究主要集中在概念和关系的获取,公理的获取研究很少,然而,公理的定义和维护也是本体构建中一项重要的工作.总之,现有的方法仍然存在许多值得改进的地方(详见第 1 节~第 3 节的讨论).另外,针对同一个学习目标,本体学习技术中的任意一种方法都有自己的适用范围,无法保证在所有情况下都得到好的学习结果.因此,如何将各种方法进行综合从而获得更好的学习结果,是未来的一个研究方向.而且,现有的本体学习方法都需要人的参与,虽然完全自动的方法在短期内是不现实的,但由于 Web 资源的大量性,还需要进一步提高本体学习的自动化程度,尽量减少用户的参与.

- 对本体学习工具的完善

目前的本体学习工具的功能都非常有限,它们都仅能处理某些类型的数据源,获取某些本体学习对象.而且,由于现有的本体学习方法的局限性,这些工具仍然很不成熟,一些最新的研究成果还没有应用其中(详见第 4 节的讨论).虽然由于缺乏客观的评价标准无法准确地对这些现有的本体学习工具进行定量的评价,但由于它们很多都是开放源码或可以免费下载的,所以通过使用可以感觉到它们无论在功能,还是在稳定性、易用性等方面与实际应用还有一段距离.未来需要一个完善的、整合的、能够完成多种学习任务的本体学习工具.

- 对本体学习结果的评价

限于篇幅,本文没有详细讨论对本体学习结果的评价.总的来说,现有方法可以分为 3 类:基于应用的方法、基于“Golden Standard”的方法和基于专家评价的方法.其中:基于应用的方法是通过选择一些相关的应用,根据这些具体应用的结果来评价本体学习的结果;基于“Golden Standard”的方法是使用一些现有的手工构建的本体作为“Golden Standard”,将本体学习的结果与其相比;基于专家评价的方法是邀请一组领域专家对本体学习的

© 中国科学院软件研究所 http://www.jos.org.cn

结果进行人工评价.在这些方法中,相关应用的选择、“Golden Standard”的选择、领域专家的选择都会极大地影响评价的结果,所以说很难使用它们对本体学习结果进行客观的评价.可见,本体学习技术作为一种无监督的学习技术,对其进行评价比对有监督的技术(例如分类技术)的评价更为困难,尤其是标准测试数据集(即标准数据源)的建立和标准结果(即标准本体或标准应用)的制定.目前还没有统一的评价本体学习结果的标准,不利于本体学习方法和工具的进一步发展.所以,如何对本体学习结果进行定量的评价是一个重要的研究方向,也是一个迫切需要解决的问题.

总之,国际上在本体学习方面的研究很活跃,并开发了一些相关的工具.国内在本体方面的研究刚刚起步,并且研究重点主要集中在如何利用本体来解决语义问题,而专门针对本体的快速构建(本体学习)方面的研究成果比较少,还没有一个能够支持中文的本体学习工具.由于中文语法的复杂性,中文本体学习技术确实存在很多困难,单纯依靠统计的手段或现有的与语言无关的算法很难获得令人满意的学习结果,必须结合中文自然语言处理领域的研究成果,使用一些基于规则的方法来改善本体学习的质量.随着本体在计算机科学领域的应用日益广泛,针对中文语言的特点展开相关研究并开发相应的工具是很有必要的.

References:

[1] Gruber TR. A translation approach to portable ontology specifications. Technical Report, KSL 92-71, Knowledge System Laboratory, 1993.

[2] Deng ZH, Tang SW, Zhang M, Yang DQ, Chen J. Overview of ontology. Acta Scientiarum Naturalium Universitatis Pekinensis, 2002,38(5):730−738 (in Chinese with English abstract).

[3] Farquhar A, Fikes R, Rice J. The Ontolingua server: A tool for collaborative ontology construction. Int’l Journal of Human-Computer Studies, 1997,46(6):707−727.

[4] Swartout B, Ramesh P, Knight K, Russ T. Toward distributed use of large-scale ontologies. In: Proc. of the AAAI Symp. on Ontological Engineering. 1996. http://ksi.cpsc.ucalgary.ca/KAW/KAW96/swartout/Banff_96_final_2.html

[5] Duineveld AJ, Stoter R, Weiden MR, Kenepa B, Benjamins VR. Wonder tools? A comparative study of ontological engineering tools. Int’l Journal of Human-Computer Studies, 2000,52(6):1111−1133.

[6] Noy NF, Fergerson RW, Musen MA. The knowledge model of protégé-2000: Combining interoperability and flexibility. In: Dieng R, Corby O, eds. Proc. of the EKAW 2000. Heidelberg: Springer-Verlag, 2000. 17−32.

[7] Arpirez JC, Corcho O, Fernandez-Lopez M, Gomez-Perez A. WebODE: A scalable ontological engineering workbench. In: Gil Y, Musen M, Shavlik J, eds. Proc. of the K-CAP 2001. New York: ACM Press, 2001. 6−13.

[8] Bechhofer S, Horrocks I, Goble C, Stevens R. OilEd: A reason-able ontology editor for the semantic Web. In: Baader F, Brewka G, Eiter T, eds. Proc. of the KI 2001, Joint German/Austrian Conf. on AI. Heidelberg: Springer-Verlag, 2001. 396−408.

[9] Sure Y, Angele J, Erdmann M, Staab S, Studer R, Wenke D. OntoEdit: Collaborative ontology engineering for the semantic Web. In: Horrocks I, Hendler JA, eds. Proc. of the ISWC 2002. Heidelberg: Springer-Verlag, 2002. 221−235.

[10] Bozsak E, Ehrig M, Handschuh S, Hotho A, Maedche A, Motik B, Oberle D, Schmitz C, Staab S, Stojanovic L, Stojanovic N, Studer R, Stumme G, Sure Y, Tane J, Volz R, Zacharias V. KAON—Towards a large scale semantic web. In: Bauknecht K, Mintjoa A, Quirchmayr G, eds. Proc. of the 3rd Int’l Conf. on E-Commerce and Web Technologies. Heidelberg: Springer-Verlag, 2002. 304−313.

[11] Lenat DB. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 1995,38(11):33−38.

[12] Nierenburg S, Beale S, Mahesh K, Onyshkevych B, Raskin V, Viegas E, Wilks Y, Zajac R. Lexicons in the mikrokosmos project. In: Busa F, Johnston M, eds. Proc. of the AISB Workshop on Multilinguality in the Lexicon. Brighton: Ulrich Heid, 1996. 26−33.

[13] Maedche A. Ontology Learning for the Semantic Web. Boston: Kluwer Academic Publishers, 2002.

[14] Lawrence S, Giles CL. Searching the World Wide Web. Science, 1998,280(5360):98−100.

[15] Sa SX, Wang S. Introduction to Database System. 3rd ed. Beijing: Higher Education Press, 2002 (in Chinese).

[16] Ramanathan S, Hodges J. Reverse engineering relational schemas to object-oriented schemas. Technical Report, MSU-960701, Mississippi State University, 1996.

[17] Johannesson P. A method for transforming relational schemas into conceptual schemas. In: Rusinkiewicz M, ed. Proc. of the ICDE’94. Boston: IEEE Computer Society, 1994. 190−201.

[18] Stojanovic L, Stojanovic N, Volz R. Migrating data-intensive web sites into the semantic Web. In: Proc. of the 17th ACM Symp. on Applied Computing. New York: ACM Press, 2002. 1100−1107. http://www.fzi.de/ipe/publikationen.php?id=820

© 中国科学院软件研究所 http://www.jos.org.cn

[19] Kashyap V. Design and creation of ontologies for environmental information retrieval. In: Proc. of the Workshop on Knowledge Acquisition, Modeling and Management.1999. http://sern.ucalgary.ca/KSI/KAW/KAW99/papers/Kashyap1/kashyap.pdf

[20] Astrova I. Reverse engineering of relational database to ontologies. In: Davies J, *et al*, eds. Proc. of the ESWC 2004. Heidelberg: Springer-Verlag, 2004. 327−341.

[21] Han J, Kamber M, write. Fan M, Meng XF, *et al*, translate. Data Mining: Concepts and Techniques. Beijing: China Machine Press, 2001 (in Chinese).

[22] Astrova I, Stantic B. Reverse engineering of relational database to ontologies: an approach based on an analysis of HTML forms. In: Proc. of the Workshop on Knowledge Discovery and Ontologies at ECML/PKDD. 2004. http://olp.dfki.de/pkdd04/astrova-final.pdf

[23] Wang J, Wen J, Lochovsky F, Ma W. Instance-Based schema matching for web databases by domain-specific query probing. In: Mario AN, *et al*, eds. Proc. of the VLDB 2004. San Francisco: Morgan Kaufmann Publishers, 2004. 408−419.

[24] Agirre E, Ansa O, Hovy E, Martinez D. Enriching very large ontologies using the WWW. In: Staab S, Maedche A, eds. Proc. of the ECAI 2004 Workshop on Ontology Learning. 2000. http://ol2000.aifb.uni-karlsruhe.de/

[25] Xu F, Kurz D, Piskorski J, Schmeier S. A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. In: Proc. of the LREC 2002. http://www.dfki.uni-sb.de/~feiyu/LREC_TermExtraction_final.pdf.

[26] Missikoff M, Navigli R, Velardi P. Integrated approach for web ontology learning and engineering. IEEE Computer, 2002,35(11): 60−63.

[27] Navigli R, Velardi P, Gangemi A. Ontology learning and its application to automated terminology translation. IEEE Intelligent Systems, 2003,18(1):22−31.

[28] Daille B. Study and implementation of combined techniques for automatic extraction of terminology. In: Proc. of the ACL'94 Workshop "The Balancing Act: Combining Symbolic and Statistical Approaches to Language". 1994. http://acl.ldc.upenn.edu/W/W94/W94-0104.pdf

[29] Velardi P, Fabriani P, Missikoff M. Using text processing techniques to automatically enrich a domain ontology. In: Proc. of the FOIS. New York: ACM Press, 2001. 270−284.

[30] Hearst MA. Automatic acquisition of hyponyms from large text corpora. In: Bourigault D, ed. Proc. of the COLING. 1999. 539−545. http://www.cs.mu.oz.au/acl/C/C92/C92-2082.pdf

[31] Fisher DH. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 1987,2(2):139−172.

[32] Bisson G. Learning in FOL with a similarity measure. In: Pinkas G, Dechter R, eds. Proc. of the AAAI. San Francisco: Morgan Kaufmann Publishers, 1992. 82−87.

[33] Emde W, Wettschereck D. Relational instance-based learning. In: Saitta L, ed. Proc. of the ICML'96. San Francisco: Morgan Kaufmann Publishers, 1996. 122−130.

[34] Faure D, Nedellec C. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In: Velardi P, ed. Proc. of the LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications. Granada: LREC, 1998. 5−12.

[35] Maedche A, Staab S. Discovering conceptual relations from text. In: Horn W, ed. Proc. of the ECAI 2000. Amsterdam: IOS Press, 2000. 321−325.

[36] Maedche A, Staab S. Ontology learning for the semantic Web. IEEE Intelligent System, Special Issue on the Semantic Web, 2001, 16(2):72−79.

[37] Nakaya N, Kurematsu M, Yamaguchi T. A domain ontology development environment using a MRD and text corpus. In: Proc. of the Joint Conf. on Knowledge Based Software Engineering. 2002. http://panda.cs.inf.shizuoka.ac.jp/mmm/doddle/publication/jckbse2002.pdf

[38] Shamsfard M, Barforoush AA. Learning ontologies from natural language texts. Int'l Journal Human-Computer Studies, 2004,60(1): 17−63.

[39] Chen WL, Zhu JB, Yao TS. Automatic learning field words by bootstrapping. In: Proc. of the JSCL. Beijing: Tsinghua University Press, 2003. 67−72 (in Chinese with English abstract).

[40] Zheng JH, Lu JL. Study of an improved keywords distillation method. Computer Engineering, 2005,31(18):194−196 (in Chinese with English abstract).

[41] Du B, Tian HF, Wang L, Lu RZ. Design of domain-specific term extractor based on multi-strategy. Computer Engineering, 2005,31(14):159−160 (in Chinese with English abstract).

© 中国科学院软件研究所 http://www.jos.org.cn

[42] Kavalec M, Svátek V. A study on automated relation labelling in ontology learning. In: Buitelaar P, Cimiano P, Magnini B, eds. Ontology Learning from Text: Methods, Evaluation and Applications. Amsterdam: IOS Press, 2005. http://nb.vse.cz/~svatek/olp05.pdf

[43] Doan A, Domingos P, Levy A. Learning source descriptions for data integration. In: Suciu D, Vossen G, eds. Proc. of the Workshop on the Web and Database. Heidelberg: Springer-Verlag, 2000. 81−86.

[44] Mello RdS, Heuser CA. A bottom-up approach for integration of XML sources. In: Simon E, Tanaka AK, eds. Proc. of the WIIW. Brazil, 2001. 118−124.

[45] Litkowski K. Models of the semantic structure of dictionaries. Journal of Computational Liguistics, 1978,15(81):25−74.

[46] Rigau G, Rodrigues H, Agirre E. Building accurate semantic taxonomies from monolingual MRDs. In: Proc. of the COLING-ACL. San Francisco: Morgan Kaufmann Publishers, 1998. 1103−1109. http://acl.ldc.upenn.edu/P/P98/P98-2181.pdf

[47] Papatheodorou C, Vassiliou A, Simon B. Discovery of ontologies for learning resources using word-based clustering. In: Kommers P, Richards G, eds. Proc. of the World Conf. on Educational Multimedia, Hypermedia and Telecommunications. Chesapeake: AACE, 2002. 1523−1528.

[48] Maedche A, Staab S. The ontology extraction & maintenance environment Text-to-Onto. In: Proc. of the ICDM 2001 Workshop on the Integration of Data Mining and Knowledge Management. 2001. http://cui.unige.ch/~hilario/icdm-01/DM-KM-Final/Volz.pdf

[49] Modica G, Gal A, Jamil HM. The use of machine-generated ontologies in dynamic information seeking. In: Batini C, Giunchiglia F, Giorgini P, Mecella M, eds. Proc. of the 9th Int'l Conf. on Cooperative Information Systems. Heidelberg: Springer-Verlag, 2001. 433−448.

[50] Avigdor G, Giovanni M, Hasan J. OntoBuilder: Fully automatic extraction and consolidation of ontologies from Web sources. In: Proc. of the ICDE 2004. Boston: IEEE Computer Society, 2004. 853−853. http://csdl.computer.org/comp/proceedings/icde/2004/2065/00/20650853.pdf

[51] Gal A, Anaby-Tavor A, Trombetta A, Montesi D. A framework for modeling and evaluating automatic semantic reconciliation. VLDB Journal, 2005,14(1):50−67.

[52] Volz R, Oberle D, Staab S, Studer R. OntoLiFT prototype. IST Project 2001-33052 WonderWeb Deliverable 11. 2003.

附中文参考文献:

[2] 邓志鸿,唐世渭,张铭,杨冬青,陈捷.Ontology 研究综述.北京大学学报(自然科学版),2002,38(5):730−738.

[15] 萨师煊,王珊.数据库系统概论.第 3 版.北京:高等教育出版社,2002.

[21] Han J,Kamber M,著.范明,孟小峰,等,译.数据挖掘概念与技术.北京:机械工业出版社,2001.

[39] 陈文亮,朱靖波,姚天顺.基于 Bootstrapping 的领域词汇自动获取.见:第 7 届全国计算语言学联合学术会议论文集(JSCL 2003).北京:清华大学出版社,2003.67−72.

[40] 郑家恒,卢娇丽.关键词抽取方法的研究.计算机工程,2005,31(18):194−196.

[41] 杜波,田怀凤,王立,陆汝占.基于多策略的专业领域术语抽取器的设计.计算机工程,2005,31(14):159−160.

杜小勇(1963－),男,浙江开化人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为智能信息检索,高性能数据库,知识工程.

王珊(1944－),女,教授,博士生导师,CCF 高级会员,主要研究领域为高性能数据库,数据仓库,知识工程.

李曼(1977－),女,博士生,主要研究领域为本体,语义 Web,智能信息检索.

© 中国科学院软件研究所 http://www.jos.org.cn