本体研究综述

陈世祺

（大连海事大学 信息科学与技术学院 辽宁 大连 116026）

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分析了本体与语义 Web，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

A Survey on Ontology Research

Chen ShiQi

(DaLian Maritime University,Computer Science and Technology,

Liaoning,Dalian,116026，China)

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,analysing Ontology and Semantic Web 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

1. 绪论

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 Grube 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]等。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月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[6] $O:=\{C,R,H^C,\mathrm{Re}l,A^O\}$。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，即概念问的分类关系(taxonomy relation)；Rel 表示概念问的非分类关系（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、概念问关系（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褥针对小同类型的数据源需要采用小同的本体学习技术，所以本文根据数据源的结构化程度，将本体学习技术乡为 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与习技术。

2 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[7]。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[8]。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，实体以及实体问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问关系的获取，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问的联系的，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，即它小能用一张表模型表示出复杂对象的语义，从而小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）。在此期问，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 engineering）[9]。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[10]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一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[11]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问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 和凡都具有属性 A，且 A 小是 R 的主码，满足 R 表中的属性 A 包含依赖于凡表中的属性 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，和凡对应的类。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问的参照关系，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[12]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[13]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，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[14]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[15]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。

3 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[17-19]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[20，21]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 提出了一种基于矢量的聚类方法，Bisson 和 Emde 等人提出了基于 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。

4 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 等人和 Mello 等人使用预定义的规则，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 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。

5 本体与语义 Web

5.1 本体与形式语义学

Walid S 研究指出，在自然语言语义中遇到的问题，大部分只是因为符号处理系统

缺乏相关的背景信息支持。因为在这些系统中，不具备任何类似于人类常识的判断。对于这个问题的解决方法是将反应人类基本常识和自然语言的本体整合到语义中。在这些加工后的逻辑中，有本体的概念，也有逻辑的概念，并且本体概念不仅包括Davidsonian事件，同时也包括抽象对象。Walid S证明了在该架构中，自然语言遇到的语义问题都可以得到正确和统一地解决。

5.2 面向语义服务的本体服务

Khalid、Pasha和Ahmad等人在“基于代理的语义服务与基于OWL的语义服务之间的本体服务”一文中指出，代理和网格服务是两种不同的技术，有着不同的标准和规范。网格服务对于语义的利用推动了两者在软件代理中的发展。Khalid等学者认为，不同的本体语言宜采用不同的语义描述方法，不同的语义描述语言对条目、语法、语义和约束有不同的支持方式。Khalid等学者主要是整合代理技术到网格服务中，并且对代理技术的标准和规范不做修改，提出了一种中间件，作为中间件本体服务，通过对本体的映射，在代理和网格服务中提供语义互操作，如映射OWL本体到FIPA SL本体的代理，映射FIPA SL本体到OWL的网格服务。此外，Khalid等学者还描述了如何注册中间件中的本体，并在网格和代理中进行揭示，以便用于概念、属性和行为的查询。同时，他们还描述了通过一种软件代理来调用和使用OWL发布的网格服务的试验床配置。

5.3Web本体应用

Kim和Su-Kyoung在“面向语义网应用的Web本体应用”一文中分析了语义网应用和Web本体的特点之后，通过描述逻辑和SWRL，建立了基于推理的Web本体，并验证了建立Web本体的推理机制。最重要的是，Kim等学者还通过推理重新生成了新的本体。根据推理规则定义的执行，通过基于本体的知识产生了一个新的本体文件。这种方式的一个重要的目的就是重用和共享本体的功能。该研究建立了一个基于Web本体的试验系统来支持图像检索。试验结果表明，语义Web应用了采用基于推理的Web本体后，在查全率和查准率方面，与一些使用了基于注释的本体目标系统相比，其性能明显要好。

5.4 基于情景族模型的OWL本体概念

语义网的核心是本体，本体支持语义网应用之间的互操作，并且让开发人员能够共享领域知识。建立本体的过程是一个高成本的过程。创建本体更像是一门艺术，而不是一门科学。因此，方法论和支撑工具是帮助开发人员构建合适本体的核心，以便达到既定目标，并确认本体是否与目标相符及其可重复利用性。Dong-Soon Kim、Suk-Hyung Hwang和Hong-Gee Kim等学者提出了一种新颖的方法用来分析基于语境模型的形式概念分析的本体，并建立了一种新的工具用于从OWL的源码中抽取主要的元素（类，属性和个体等），并发现一些结构性的问题。通过这个工具，本体开发人员可以创建良好定义的本体。

5.5 信息技术本体与语义技术

语义技术是相对于语法技术而言的。本体在语法结构以意义为中心的重新配置方法中起到了很重要的作用。Key-Sun Choi等学者研究指出，利用信息本体的目的主要有两个方面：一方面是针对用户需求获取正确的信息和服务；另一方面是为在类和实例之间建立关系提供思路。基于本体的问题—回答模式可以提高其性能。在该模式中，本体提前从相关的信息资源中获取信息，因此每一种问题类型都可以从本体中获取特定的关系。问题是这种关系或者相关的类都是不确定的领域或者依赖于单个特定的领域。本体学习在问题—回答应用的第一步是找到这些不确定性关系的发现机制，并充分考虑针对特定领域资源时，特定的关系—实例的映射。第二步是考虑领域本体获取，针对类似资源（如特定领域词表）的时候采取从上到下的方式，而针对相关资源采取从下到上的方式。但创建者仍需面对的问题是，词表是由类组成的，而不是语料库中的条目实例，它们对于资源的覆盖面比较小，并且在这个层面，类和实例之间的映射并没有充分建立。最后，需要解决两个问题：一是如何评价本体的有效性；二是如何比较每个本体的应用效果。

5.6 面向语义 Web 的模糊本体创建

模糊逻辑在本体中用于表达不确定信息，而模糊本体将模糊概念引入到传统的本体模型中，以解决一定领域的不确定问题。一般而言，模糊本体来自于预定义的概念层级结构，主要由模糊形式概念分析、模糊概念聚类及模糊本体生成等部分组成。Tho 等学者提出了模糊本体生成框架（Fuzzy Ontology Generation Framework，简称 FOGA），用于在不确定信息的条件下，自动生成模糊本体。FOGA 框架由以下方面组成：模型形式概念分析，概念层级结构生成和模糊本体生成。同时，Tho 等学者还对本体新增数据的近似推理方法进行研究，提出了整合模糊技术到属性数据库的建议。

6 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），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。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。

参考文献

[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。

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

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

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

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

[10] 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。

[11] Stojanovic L， Stojanovic N， Volz R。 Migrating data-intensiveWebsites 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

[12] 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。

[13] 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。

[14] 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。

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

[16] 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/
[17] 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。
[18] Missikoff M， Navigli R， Velardi P。 Integrated approach forWebontology learning and engineering。 IEEE Computer， 2002， 35（11）：60-63。
[19] Navigli R， Velardi P， Gangemi A。 Ontology learning and its application to automated terminology translation 。 IEEE Intelligent Systems， 2003，18（1）：22-31。
[20] 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。
[21] 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。