《智能信息处理》课程考试

本体研究综述

学　　院：信息科学技术学院

专　　业：软件工程

姓　　名：郭爱彬

学　　号：1120201307

授课老师：李冠宇

考核	课程成绩
得分	

本体研究综述

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

Review of ontology research

Abstract In recent years, ontology learning technology has gradually become a research hotspot in the field of computer science. According to the degree of data source structure (structured, semi-structured and unstructured) and the level of ontology learning object (concept, relation and axiom), ontology learning problem is divided into 9 sub-problems. The basic characteristics, common methods and the latest research progress of these 9 types of problems are described respectively, and ontology and semantic Web are further analyzed under the analysis framework, the existing problems are discussed, and the future research direction is pointed out.
Keywords ontology,ontologylearning,concept,relation

1 引言

“本体论”最早是哲学中的基本概念，它是研究“是”之所以为“是”的理论，可以说是哲学中的哲学,甚至可以认为西方哲学自身的发展就是一个“本体论”的产生、发展、怀疑和批判的过程。近年来，本体论的方法在知识工程领域得到了越来越广泛的应用，在很多有名的知识系统中，如美国 D.Lenat 教授领导研制的大型常识知识库系统 Cyc,Pricnteno 大学 Berkeley 分校研制的语言知识库 WrodNet 等,本体论都有一定的应用。一方面,本体论研究深层次上的指示，把知识工程研究中的知识向更深更本质的方向上推进,另一方面,本体论的研究独立于任何语言,因此本体论将会为不同系统之间知识的共享和互操作提供手段。早在 1998 年,Gruber 就已经给出了本体的一个流行定义，即“本体是领域概念化对象的明确表示和描述”。Guarino 把概念化对象 C 定义为:C<D,W,R>,其中 D 是一个领域，W 是该领域中相关的事务状态集合，R 是领域空间<D,W>概念关系的集合。因此，从概念化对象的定义来看，本体把现实世界中的某个领域抽象成组概念(如实体、属性、进程等)及概念间的关系。某个领域的本体不仅提供了关于该领域的一个公认的概念集，同时也表达了各概念间所具有的各种语义联系。

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 念以及概念之问的关系，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，人机之问以及机 器之问就可以进行交流。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。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 本体定义

在哲学里，ontology 是历古而恒新的基本研究领域。西方在笛卡尔之前哲学探讨的重心都在 ontology。按一般理解，它主要研究与“存在、本体”有关的道理。中国学界通用的汉译名是“本体论”。近来有学者认为该译名不准确，建议改译为“存在论”，但也有反对意见。在人工智能界， 最早给出 Ontology 定义的是 Neches 等人，在文献[1]中， 他们将 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。

1993 年，Gruber 给出了 Ontology 的一个最为流行的定义[2] ， 即 Ontology 是概念模型的明确的规范说明。此后，Borst 在此基础上， 给出了 Ontology 的另外一种定义[3]：Ontology 是共享概念模型的形式化规范说明.Studer 等对上述两个定义进行了深入的研究， 认为 Ontology 是共享概念模型的明确的形式化规范说明。这包含 4 层含义[4]：概念模型、明确、形式化和共享。概念模型指通过抽象出客观世界中一些现象的相关概念而得到的模型。概念模型所表现的含义独立于具体的环境状态。明确指所使用的概念及使用这些概念的约束都有明确的定义。形式化指 Ontology 是计算机可读的(即能被计算机处理)。共享指 Ontology 中体现的是共同认可的知识，反映的是相关领域中公认的概念集， 即 Ontology 针对的是团体而非个体的共识。Ontology 的目标是捕获相关领域的知识， 提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇和词汇间相互关系的明确定义。

2.1 本体的语义性质

若以 Gruber 定义为标志，Ontology 作为计算机领域的一个研究与应用的分支，形成于 20 世纪 90 年代。作为该分支名称的“Ontology”，却是借用了在哲学领域已有近 400 年历史的最活跃也是最有争议的基础概念（Ontology 一词由德国哲学家 R.Goclenius 在 1613 年首创使用）。现在要思考的是：哲学术语 Ontology 的语义为何？当它被引入计算机工程领域其语义是否被改变？

Ontology 在哲学领域有两种含义：一是指始于巴门尼德的“真理之路”（公元前 500 年）而终于黑格尔的《逻

辑学》(1820 年）的西方哲学的历史形态。二是指“深入到事实后面去”认识世界万物的思维方式[3]。也就是说，哲学术语 Ontology 既是指 19 世纪之前的西方哲学，又是指延续至今的一种认识思维方式。二者同归类于抽象的“理论”范畴，而非具体的客观存在。

Ontology 的语义性质是理论。Goclenius 在 17 世纪初新创该词汇时正是这样认定的。Ontology 由 onta 和后缀- logy 组合而成。前者究竟为何意哲学界迄今尚在争议之中；后者的作用学者们的看法却是一致的。构词后缀- logy 的主要含义是“….学”或“….论”，另一个含义为“言词”。使用 WordNet（其为当今计算机领域最著名的 Ontology 之一）进行检索，可知现代英语中含有后缀- logy 的名词总数约为 160 个。其中绝大多数是指称“….学”或“….论”的，只有很少几个是“言词”的意思。

有人会以常用名词 Technology 举例反驳，既然 Technology 可以是“技术”，Ontology 的汉译为何必须带一个“论”的限定？殊不知，汉语中的“术”恰与“学”或“论”具有相似的含义。Technology 由 Technic 和- logy 组合而成。因为 Technic 对应于汉语的“工艺”或“技巧”，所以 Technology 的汉译理解就是：“工艺之学”或“技巧之学”。“工艺”侧重于理论，“技巧”侧重于操作，因而就有“工艺之学”和“技巧之术”（简称技术）的区分。“学，识也；术，道也”（《广雅》），两者都是人类关于某类事物的系统性知识，仅是层次上有所差别而已。在汉语中与“工艺学”与“技术”相类似的情况还有“医学”与“医术”、“算学”与“算术”等。

既然英文后缀- logy 的汉译规律如此，那么 Ontology 也只能被看作是“××学”或“××论”。至于具体是什么，就取决于对其词干“onta”做何种理解了。哲学领域的 Ontology 是一种理论，被引入计算机工程领域其含义可以有所改变，但其所指还应该属于理论范畴。“本体”的说法改变了 Ontology 的理论属性，使其成为一种具体的人造实在。

2.2 XML

XML (Extensible Markup Language) 是针对包 含结构化、半结构化信息的文档而设计的一种标记 语言。XML 已经成为了 Web 上数据表示和交换的事实标准， 是应用之间或者机器之间共享数据的一 种有效方式。XML 的可扩展性是 XML 区别其他标 记语言的最基本特征， 其核心在于以一种标准化的 方式来建立数据表示的结构，而将具体标记的定义 留给了用户。但是， XML 并不能解释它标记的含义，并且 XML 模式只能对 XML 的语法合法性进行验 证， 而不能区分 XML 属性和元素在含义上的不同。 因此对于同样的信息内容， 我们可以将其映射为多 种不同的 XML 结构， 同时对同一种 XML 结构， 也 可以存在多种不同的解释， 而相同的应用也可能对不同的 XML 做出同样的解释。 2.2.2 RDF 与 RDFS RDF 定义了一种用以描述资源及其相互关系 的简单模型，是语义互联网实现的关键技术之一，也是语义信息描述的有效手段。其基本数据模型包
含三类对象：资源、属性和陈述， 资源之间的关系 通过属性和值来描述，描述特定资源特定属性的 值， 就构成 RDF 中的一个陈述， 通常可以用三 元 组 <subject， predicate， object> 描述。其中， 被描述的资 源称为 subject， 描述资源的属性称 predicate， object 则是属性对应的值。 RDF 建立在 XML 和 URI 的 基础上。 RDF 通过属性和值描述了资源以及资源之间 的关系， 但并没有提供描述这些属性及属性间关系 的机制。RDF. Schema 提供了这种表达机

制，它描 述了RDF中 properties 的使用规则， 为RDF 定义了 领域字典，并用类型层次结构来组织该字典， 从而 构成完备的语义空间。RDF 的数据模型实质上是一 种二元关系的表达，由于任何复杂的关系都可以分 解为多个简单的二元关系， 因此 RDF 的数据模型 可以作为其他任何复杂关系模型的基础模型。 RDF. Schema 在提供了简单的机器可理解语义模型的同时， 为领域化的本体语言提供了建模基础， 并 使得基于 RDF 的应用可以方便地与这些本体语言 所生成的本体进行合并。RDF 的这一特性使得基 于 RDF 的语义描述结果具备了可以和更多的领域 知识进行交互的能力，也使基于 XML 和 RDF 的 Web 数据描述具备了良好的生命力。

2.3 OWL

OWL 目前是本体的标准描述语言。它是结合 了 DAML + OIL 应用经验而改进的修订版， 建立在 RDF 基础上，以 XML 为书写工具。主要用来表达需要计算机应用程序来处理的文件中的知识信息， 而不是呈递给人的知识。OWL 能清晰的表达词表 中各词条的含义及其之间的关系， 这种表达被称为 本体。OWL 相对 XML、RDF 和 RDF. Schema 拥有 更多的机制来表达语义，从而 OWL 超越了 XML、 RDF 和 RDF. Schema 仅仅能够表达网上机器可读的文档内容的能力。[4] 2.3 本体的构建方法

本体的构建主要有手工构建、复用已有本体和 自动构建本体等多种方式，由于构建本体的领域范 围、设计标准与原则等不相同的特点，本体的构建工作没有统一的实现标准。构建本体的目的有多种 多样，构建本体的步骤和过程也各不相同。一般的 来说，本体的构建应该遵循明确性和客观性、完全 性、一致性等原则[1]。一般说来，建构一个知识领 域的本体， 包括以下 6 个步骤: (1)确定本体的领域和范围

首先， 要明确建构的本体将覆盖的专业领域、 应用本体的目的、作用及其系统开发、维护和应用 的对象。 (2)列举知识领域中重要的术语、概念 在创建本体的初始阶段， 尽量列举出系统想要 陈述的或准备向用户解释的所有概念， 不必考虑概 念之间语义的重叠及表达方式(类、属性、实例) 。 (3)建立本体框架 上一步生成的知识领域中的大量概念是一个 没有组织结构的词汇表， 需要按一定的逻辑规则， 将其分组， 构成不同的工作领域， 并对同一工作领 域内的概念相关性和重要性进行评估， 选出关键性术语， 尽可能准确而精炼地表达出该领域的相关知 识， 形成该领域知识的框架系统。 (4)设计元本体， 重用已有的本体， 定义领域中概念 之间的关系 元本体是元概念的本体， 其术语用于定义本体 中的高层次的抽象概念，如实体、关系、角色等。 设计元本体时， 一要尽量作到领域无关性; 二要包 含的元概念尽可能少。一个概念可采用元本体中定 义的元概念进行定义， 或采用本体中已被定义的概 念进行定义或重用已有的本体。除了定义概念之外， 还要定义概念之间的关 系。这些关系不仅涉及同一工作领域中的概念， 而 且还与其它工作领域的概念相关联。 (5)对领域本体进行编码、形式化 选用合适的本体描述语言， 对上述建立的本体 进行编码， 形式化。现有本体描述语言约 28 种， 大 都基于一阶逻辑，也有基于描述逻辑。常用的本体 语言有: Ontolingua、CycL、LooM 等。本体模型实 现形式化可提供比自然语言更严格的格式， 能增强 机器的可读性，便于交换及本体模型自动逻辑推理及检验。 (6)对本体进行检验和评价 本体在形式化以后， 是否满足用户需求， 是否 符合本体的建构准则， 是

否术语、概念定义清晰， 是否关系定义完整等， 都要在本体建构后进行检验 和评估。

3 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
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]DengZH, TangSW, ZhangM, Yang DQ, ChenJ。 Overview ofontology。 Acta Scientiarum Naturalium Universitatis Pekinensis, 2002, 38（ 5 ）： 730 - 738 （ in ChinesewithEnglishabstract)。

[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]SwartoutB, RameshP, Knight K, Russ T 。 Toward distributed use of large-scale ontologies。 In: Proc 。 oftheAAAI Symp 。 OnOntologicalEngineering。 1996。 http: //ksi 。cpsc 。ucalgary 。ca/KAW/KAW96/swartout/Banff_96_final_2。

[5]DuineveldAJ ， StoterR ， WeidenMR， KenepaB, BenjaminsVR。 Wondertools?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 SemanticWeb。 Boston: KluwerAcademic Publishers, 2002。

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