《智能信息处理》课程考试

本体的基本概念及应用

刘鑫

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 10 日

本体的基本概念及应用

刘鑫

（大连海事大学 信息科学技术学院， 大连 116026）

摘 要 本体是概念的集合及概念间关系的集合的集合，事实上，它是一组特定领域的概念及其相互关系的正式表达。本体论是人们在信息科学中应用的基本理论之一。本文首先简要介绍了语义网络的概念，详细介绍了语义网络的基本概念，然后列举了语义网络的应用。

关键词 本体；语义网；智能信息

The basic concept and application of the ontology

Liu Xin

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026)

Abstract Ontology is a collection of concepts and relationships between concepts. In fact, it is the formal expression of a group of domain specific concepts and their relationships. Ontology is one of the basic theories applied in information science. Firstly, this paper briefly introduces the concept of semantic network, introduces the basic concept of semantic network in detail, and then lists the application of semantic network.

Key words ontology ; semantic network ; smart information

1 引言

本体英文术语“ontology”一词源于哲学领域，且一直以来存在着许多不同的用法。在计算机科学领域，其核心意思是指一种模型，用于描述由一套对象类型（概念或者说类）、属性以及关系类型所构成的世界。

本体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之间相互关系的明确定义。本体的概念有四层含义：概念化(conceptualization)、形式化(formal)、明确(explicit)、共享(share)。它是一种能在语义和知识层次上描述信息系统的概念模型。本体应用的基础是本体的构建[1]。

本文首先从语义网的概念、目标、设计原则和体系结构三个方面介绍了语义网的体系结构。本体是语义 web 体系结构中的一个层次，详细介绍了本体的定义、构建标准和构建方法。最后给出了本体在现实生活中的三个应用实例，加深对本体的理解。

2 语义网

2.1 语义网概念

由万维网联盟的蒂姆·伯纳斯-李（Tim Berners-Lee）在 1998 年提出的一个概念是能够根据语义进行判断的智能网络，实现人与电脑之间的无障碍沟通。它好比一个巨型的大脑，智能化程度极高，协调能力非常强大。在语义网上连接的每一部电脑不但能够理解词语和概念，而且还能够理解它们之间的逻辑关系，可以干人所从事的工作。它将使人类从搜索相关网页的繁重劳动中解放出来，把用户变成全能的上帝。语义网中的计算机能利用自己的智能软件，在万维网上的海量资源中找到你所需要的信息，从而将一个个现存的信息孤岛发展成一个巨大的数据库[2]。

2.2 语义网目标及设计原则

语义网的目标是为在线信息提供计算机可理解的语义，满足 agent 对异构分布式信息的有效检索和访问，实现在线信息资源在语义层的全方位互联，实现在线信息的更高层次、基于知识的智能化应用。

语义 web 设计的原则是：所有资源都可以通过 URI 进行标识；资源和链接可以有类型；允许部分

/部分/不完整的信息；信息不一定绝对真实；它能够支持和反映信息的变化和演变；和最低设计原则。

2.3 语义网体系结构

Berners-Lee 于 2000 年提出了语义网的体系结构，并对此做了简单的介绍。该体系结构共有七层，自下而上其各层功能逐渐增强。

2.3.1 第一层：“字符集”层。

Unicode 和 URI。Unicode 是一个字符集，这个字符集中所有字符都用两个字节表示，可以表示 65536 个字符，基本上包括了世界上所有语言的字符。数据格式采用 Unicode 的好处就是它支持世界上所有主要语言的混合，并且可以同时进行检索。URI(Uniform Resource Identifier)，即统一资源定位符，用于唯一标识网络上的一个概念或资源。在语义网体系结构中，该层是整个语义网的基础，其中 Unicode 负责处理资源的编码，URI 负责资源的标识。

2.3.2 第二层：根标记语言层。

XML+NS+xmlschema。XML 是一个精简的标准通用标记语言，它综合了标准通用标记语言的丰富功能与 HTML 的易用性，它允许用户在文档中加入任意的结构，而无需说明这些结构的含意。NS(Name Space）即命名空间，由 URI 索引确定，目的是为了避免不同的应用使用同样的字符描述不同的事物。XML Schema 是文档类型定义（DTD）的替代品，它本身采用 XML 语法，但比 DTD 更加灵活，提供更多的数据类型，能更好地为有效的 XML 文档服务并提供数据校验机制。正是由于 XML 灵活的结构性、由 URI 索引的 NS 而带来的数据可确定性以及 XML Schema 所提供的多种数据类型及检验机制，使其成为语义网体系结构的重要组成部分。该层负责从语法上表示数据的内容和结构，通过使用标准的语言将网络信息的表现形式、数据结构和内容分离。

2.3.3 第三层：“资源描述框架”层。

RDF+rdfschema。RDF 是一种描述 WWW 上的信息资源的一种语言，其目标是建立一种供多种元数据标准共存的框架。该框架能充分利用各种元数据的优势，进行基于 Web 的数据交换和再利用。RDF 解决的是如何采用 XML 标准语法无二义性地描述资源对象的问题，使得所描述的资源的元数据信息成为机器可理解的信息。如果把 XML 看作为一种标准化的元数据语法规范的话，那么 RDF 就可以看作为一种标准化的元数据语义描述规范。Rdfschema 使用一种机器可以理解的体系来定义描述资源的词汇，其目的是提供词汇嵌入的机制或框架，在该框架下多种词汇可以集成在一起实现对 Web 资源的描述。

2.3.4 第四层：“本体词汇”层。

“本体词汇”，（外语：Ontology vocabulary）。该层是在 RDF(S）基础上定义的概念及其关系的抽象描述，用于描述应用领域的知识，描述各类资源及资源之间的关系，实现对词汇表的扩展。在这一层，用户不仅可以定义概念而且可以定义概念之间丰富的关系。

2.3.5 第五至七层：Logic、Proof、Trust。Logic 负责提供公理和推理规则，而 Logic 一旦建立，便可以通过逻辑推理对资源、资源之间的关系以及推理结果进行验证，证明其有效性。通过 Proof 交换以及数字签名，建立一定的信任关系，从而证明语义网输出的可靠性以及其是否符合用户的要求。

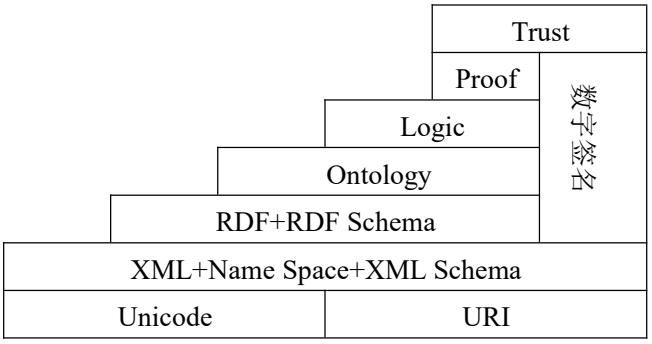

图1 语义网体系结构

3 本体

3.1 本体定义

在计算机科学与信息科学领域，理论上，本体是指一种“形式化的，对于共享概念体系的明确而又详细的说明”。本体提供的是一种共享词表，也就是特定领域之中那些存在着的对象类型或概念及其属性和相互关系；或者说，本体就是一种特殊类型的术语集，具有结构化的特点，且更加适合于在计算机系统之中使用；或者说，本体实际上就是对特定领域之中某套概念及其相互之间关系的形式化表达（formal representation）。

3.2 本体分类

关于本体的研究非常广泛，最为常用的分类方法是根据本体应用主题，将这些为数众多的本体划分为五种类型:领域本体、通用或常识本体、知识本体、语言学本体和任务本体。而依据本体的层次和领域依赖度，Guarino 等人将其分为四类：顶层本体、领域本体、任务本体和应用本体。

（1）顶层本体：研究通用的概念以及概念之间的关系，如空间、时间、事件、行为等，与具体的应用无关，完全独立于限定的领域，因此可以在较大范围内进行共享。

（2）领域本体：研究的是特定领域内概念及概念之间的关系。

（3）任务本体：定义一些通用任务或者相关的推理活动，用来表达具体任务内的概念及概念之间关系。

（4）应用本体：用来描述一些特定的应用，既可以引用领域本体中特定的概念，又可以引用任务本体中出现的概念。

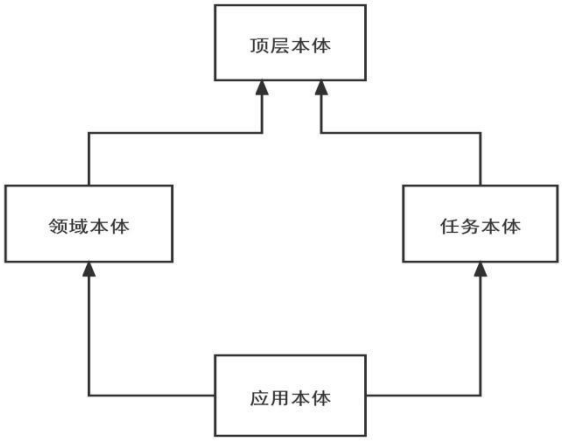

图2 本体分类图

3. 3 本体构建方法

本体的构建方法多种多样，常见的有如下几种。

1、TOVE 法：TOVE 是指多伦多虚拟企业（Toronto Virtual Enterprise）专门用于构建 TOVE 本体（关于企业建模过程的本体）。

2、METHONTOLOGY 法：专用于构建化学本体（有关化学元素周期表的本体）。

3、骨架法：专门用于构建企业本体，建立在企业本体基础之上，是相关商业企业间术语和定义的集合。

4、KACTUS 工程法：是基于 KACTUS 项目而产生的，KACTUS 是指关于多用途复杂技术系统的知识建模工程。

5、七步法：斯坦福大学医学院开发的七步法，用于领域本体构建。七步骤为：确定领域本体的范畴；复用现有的本体；列出领域内的术语；定义类和类的等级关系；定义类的属性；定义属性的分面；填充实例。

6、基于叙词表的领域本体构建：叙词表又称为主题词表，是一种语义词典，由术语及术语之间的关系组成，能够反映某学科领域的语义相关概念。由于叙词表包含丰富的领域概念和一定的语义关系，在表达和知识结构上与本体有着天然联系，包含了本学科比较完整的术语，因此，国内外很多学者都在尝试将叙词表转换为本体。国内目前基于叙词表已经转化成型的本体原型有《国防科学技术叙词表》和《中国农业科学叙词表》。

3.4 本体描述语言

本体的描述语言众多，而 W3C 推荐的本体描述语言主要有 RDF、RDFS 和 OWL。

1、RDF（Resource Description Framework，资源描述框架）

客观世界中任何一种关系都可以用一个三元组（主体/主语、谓语、客体/宾语）来进行表达。RDF 用于描述 web 上的资源，是使用 XML 语言编写、计算机可读的，不是为了向用户展示。RDF 使用 web 标识符（主体/主语）来标记资源，使用属性（谓语）和属性值（客体/宾语）来描述资源。这里的资源、属性和属性值就构成了一个陈述（或者被称为陈述中的主体、谓语和客体）。

本体中的类（概念）就是 RDF 三元组中的主体/客体，类的属性就是 RDF 三元组中的谓语。RDF 数据也可以被表示为一个带有标记的有向图，图上的节点对应三元组中的主体和客体，边对应谓语。

2、RDFS（RDF Schem，RDF 词汇描述语言）

RDFS 是在 RDF 基础上对其进行扩展而形成的本体语言，解决了 RDF 模型原有的缺点，定义了类、属性、属性值来描述客观世界，并且通过定义域和值域来约束资源，更加形象化表达了知识。

3、OWL（Web Ontology Language，Web 本体语言）

OWL 是由 W3C 开发的网络本体语言，用来对本体进行语义描述。OWL 保持了原有 RDF、RDFS 的兼容性，有保证率较好的语义表达能力，根据表达能力的增强顺序 OWL 分为三种子语言：OWL-Lite、OWL-DL 和 OWL-Full。OWL 本体中有 3 中基本元素：类、属性和实例。

4 本体的应用

4. 1 生物医学本体及应用

GO 计划是在 1998 年由三种模式生物(果蝇、酵母和小 鼠)数据库系统联合推出的,现在其成员单位已经发展为 16 家。GO 计划的目的在于:开发一个结构化的受控词汇表，用 来描述分子生物学领域内的概念，例如基因产物的属性和生 物序列;用 GO 词表对生物学数据库中的记录进行注释;为公 众提供集成式的本体数据库和相关软件[3]。

GO 计划提供描述基因及其产物的三个独立的本体: 分子 功能(Molecular Function, MF),生物过程(Biological Process, BP) 和细胞组分(Cellular Component, CC),最近又开发了描述生物 序列属性的序列本体(Sequence Ontology,SO)，本体之间相互独 立，本体内部的词汇形成有向无环图结构，每个词为该结构的 一个节点，并由唯一的 GOID 标识节，节点之间由关系:is - a 或 part-of 连接。目前整个本体库共有约 16 500 个词。

现在 GO 在数据整合、数据库注释和文本挖掘领域都有 广泛的应用。例如 GOA(Gene Ontology Annotation)计划将 GO 用于对 SWISS - PROT、TrEMBL 和 InterPro 数据库的注释;Jensen U,通过 GO 分类体系来预测人类蛋白质的功能；Jung- Hsien Chiang 建立的用于挖掘基因功能的文本挖掘系统中采 用 GO 作为领域专家词典。

虽然 GO 已经在生物学研究中得到广泛的应用,但是其 自身的结构还需要不断的完善。Leipzig 大学的研究人员通过 对 GO 本体结构的分析，发现 GO 缺乏描述逻辑的特性，基于 GO 的层次结构还无法进行深入的知识推理,GO 三个子集内 部以及子集之间的逻辑关系并不明确,缺乏必要的规则用来判断一个概念是否应该被 GO 收录。因为 GO 的开发目的在于生成生物学领域内的受控词表，因此缺乏本体论的理论规范，这些缺点在 Anand Kumar 的文献中也有详细的论述。最 近的 GONG(The Gene Ontology Next Generation)计划正尝试用 DAML + OIL 语言对其进行扩充和重写,使其组织性和表示能力更强、语义和关系的表达更丰富。

4.2 叙词表的领域本体构建

叙词表又称为主题词表，它是一种语义词典，由术语及术语之间的各种关系组成，能反映某学科领域的语义相关概念[4]。叙词表收录了某一领域的所有叙词和非叙词，按照一定顺序排列。叙词表的语义关系包括“用、代、分、属、参”，分别用来表示叙词款目之间的等同、等级、相关等语义关系。由于叙词表包含丰富的领域概念和一定的语义关系，在表达知识结构上与本体有着天然联系，包含了本学科领域中相对比较完整的术语，因此，国内外很多学术团体都在尝试着基于叙词表进行本体的构建，研究重点在于叙词表向本体转换的方法。目前由叙词表进行转换的思路主要有两种: ① 直接用某种本体表示语言表示叙词表中的词汇和关系；② 仅将叙词表作为本体中概念的来源。这两种方式都需要对转换得到的本体进行属性、关系的添加和修正, 并添加公理和函数[5]。

国外已经有 10 多种叙词表用各种方法转换为本体，如由联合国粮农组织转换为农业本体的 Agrovoc 叙词表，教育资料网关(GEM) 中的受控词表，艺术和建筑叙词表(AAT) 等。国外在这方面研究得比较成熟的是通过何种本体表示语言对叙词表的词语和关系进行转换，总结起来有以下几种:① 用 XML Schema 构建叙词标记语言。如澳大利亚 CSIRO 的 M. Lee 等所开发的叙词标记语言(TML) ，构建了叙词描述本体的框架。② 用 RDF Schema 关系表示叙词内容。典型的如 AAT 一类的分面形式的叙词表，可以将叙词表某个子集作为本体某一类属性的值直接引入。③ 用 RDF Schema 表示叙词关系。大多数叙词表采用的是这种方式转换，如 L IMER 和 ELSST 社会科学叙词表等。④用 DAML + OIL 关系表示叙词关系。DRC 提出了一个用 DAML +OIL 表示叙词关系的建议。

5 总结

随着智能信息的爆炸式增长，传统的 HTML 语言已经在很多情况不能满足人们检索信息等功能的需求。XML 语言、语义网及本体的出现，使信息具有了针对各领域的语义系统，使得信息的检索和操作更加智能便捷。从本文最后一部分的本体应用中，我们也可以看出，本体的研究具有光明的前景，值得去深入学习，不断优化。

参 考 文 献

[1] Tom Gruber (2008). "Ontology". To appear in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2008.

[2]刘清堂,黄景修,吴林静,郭志强.基于语义网的教育应用研究现状分析[J].现代远距离教育,2015(01):60-65.DOI:10.13927/j.cnki.yuan.2015.0010.

[3]Gene Ontology Consortium, The Gene Ontology (GO) Database and Infbnnatics Resource[j] .Nucleic - Acids - Res, 2004 Jan 1; 32 Database issue ： D258 - D261.

[4]孙倩，万建成. 基于叙词表的领域本体构建方法研究[J]. 计算机工程与设计，2007，28(20)：5054-5056.

[5]丁晟春，李岳盟，甘利人. 基于顶层本体的领域本体综合构建方法研究[J].情报理论与实践，2007，30(2)：236-240.