《智能信息处理》课程作业

本体基本概念及应用

王煜坤

作业	分数[70]
得分	

2020 年 11 月

本体基本概念及应用

王煜坤

（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，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。本体是人们以自己兴趣领域的知识为素材，运用信息科学的本体论原理而编写出来的作品。本文先简要概述了语义网的概念，进而详细叙述其中本体的基本概念，最后列举了一些本体的应用。

关键词　本体；语义网；智能信息

The basic concept and application of the ontology

WANG Yu-kun

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

Abstract　An ontology is a collection of concepts and a collection of relationships between concepts, which is in fact a formal representation of a set of concepts and their relationships in a particular domain. Ontology is a work written by people using the ontology principle of information science with the knowledge of their field of interest as the material. This paper first briefly summarizes the concept of semantic network, then describes the basic concepts of ontology in detail, and finally lists some applications of some ontology.

Key words　ontology ; semantic network ; smart information

1　引言

本体的概念最初起源于哲学领域，是共享概念模型的明确形式化规范说明[1]。20 世纪 90 年代，本体概念被引入人工智能、图书情报和知识工程等领域，一度成为这些领域的热门研究课题。

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。

本文从先从语义网的概念、目标和设计原则及体系结构三个方面来介绍语义网架构。本体是语义网架构中的一个层次，由此详细展开介绍，具体包括本体的定义、构造准则及构造方法。最后展示了本体在现实生活中的三个应用案例，以便更加深入了解。

2　语义网

2.1　语义网概念

为了给网页中描述的信息带来结构和意义，Tim Berners-Lee 创造了“语义网”这个概念，旨在链接、理解词语及其之间的关系，创建计算机可“理解与处理”的通用信息交换媒介，以支持网络环境下广泛有效地自动推理[2]。

语义网并不是一个独立的网络，而是当前网络的扩展，它赋予信息明确的含义，使得信息共享和重用成为可能，计算机和人们能够更好地协同工作。简单地说，它被称为信息的储存库和表达这些信息所涉及的语言。

2.2　语义网目标及设计原则

语义网的目标是为网上信息提供具有计算机可理解的语义，满足 Agent 对异构、分布信息的有

效检索和访问，实现网上信息资源在语义层上的全方位互联，实现网上信息的更高层、基于知识的智能应用。

语义网设计的原则有：所有资源都能用 URI 来标识；资源与链接可以有类型；部分/片断/不完整信息是容许的；信息不必是绝对真的；能支持、反映信息的变化与演化；以及最小设计原则

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，如下图 1，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。[3]

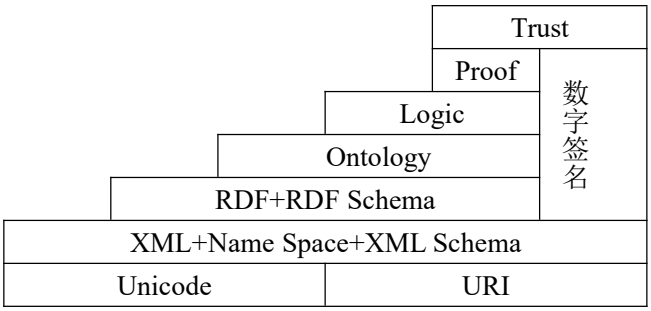

图 1　语义网体系结构

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。

3　本体

3.1 本体定义

本体概念被引入人工智能、知识工程等领域后被赋予了新的含义。然而不同的专家学者对本体的理解不同，所给出的定义也有所差异。

人工智能领域的学者 Neches(1991) 等人对 ontology 进行定义，即:

定义 1 本体是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。[4]

Neches 是最早对本体定义进行研究的学者，从内容的角度给出了本体定义，概括出了本体的基本要素，包括领域术语、关系和规则。这为其后各领域学者对本体的定义研究提供了参考借鉴。

美国斯坦福大学 Gruber (1993) 给出了本体的定义:

定义 2 本体是概念化的规范说明。[5]

Gruber 给出的本体定义最为经典，但是未能全面概括出本体的本质。

随后，Borst 等人(1997) 对 Gruber 给出的定义进行了补充，即

定义 3 本体是共享概念模型的形式化规范说明。[6]

Borst 提出了本体共享的概念，阐明了本体的共享本质，但没有说明概念与概念之间的关系。

德国学者 Studer 等人(1998) 又对 Borst 的定义进行了扩展，提出了概念关系之间的“明确”定义，认为:

定义 4 本体是共享概念模型的明确的形式化规范说明。 [7]

Studer 给出的本体定义被各领域专家学者高度认可，其涵盖了本体的基本特征: 共享、明确、概念化、形式化，被学术界广泛引用，对于后来的本体研究具有重要意义。

中国学者对本体定义也做了很多研究。张晓林教授(2002) 认为“ontology”是概念集，是特定领域公认的关于该领域的对象及其关系的概念化表述[8]。中国标准化研究院的李景(2005) 博士认为，本体是一个关于某些主题的，层次清晰的规范说明。北京大学的汤艳莉、赖茂生教授认为 ontology 作为语义网的重要组成部分，是对世界或者领域知识、概念、实体及其关系的一种明确的、规范的概念化描述[9]。张秀兰教授通过对国内外各领域本体定义的深入研究，总结出了本体定义:

定义 5 本体是通过描述、捕获领域知识，确定领域内共同认可的概念和概念间的关系，以用于领域内的不同主体之间交流与知识共享的形式化规范说明。[10]

3.2 本体构造准则

本体构造准则由 Gruber 在 1995 年提出。包含以下六个方面：清晰性（Clarity），完全性（Completeness），一致性（Coherence），可扩展性（Extendibility）,最小承诺（Minimal ontology commitment）及最小编码偏好（Minimal encoding bias）。其中，最小承诺原则指的是：只定义最必要的术语，只定义约束最弱的关系。最小编码偏好原则指的是不指定术语形式化用何编码。

除了最广为应用的 Gruber 准则之外，也有一些其他的准则，如 Arpirez 规则，这里不多赘述。

3.3 本体构造方法

到目前为止，有五种典型的构造本体的方法，分别是：Uschold 和 King 方法，Gruninger 和 Fox 方法，Berneras 方法，MethOntology 方法，基于 SENSUS 方法。由于篇幅有限，以下仅介绍第一种方法。

Uschold 和 King[11]规定了本体论发展的五个阶段，即：确定目的、构建本体论、评估和文档。本研究的第一阶段（确定目的）分为两个阶段：定义本体论的目的和范围。

然而，没有提供关于如何收集概念以及如何确定概念之间的关系的详细准则,"只给出非常模糊的准则，依靠集思广益技术"。在这项研究中，这些概念被收集起来，并使用定性方法与访谈和文献综述构建领域本体关系。域本体使用描述逻辑以半正式表示形式进行编码。半正式本体在 OWL 中进一步正式编码，以便计算机自动处理。Uschold 和 King 方法的第二阶段通过将域本体与 DOLCE 上层本体学对齐完成。Uschold 和 King 方法（评估）的第三个 TF 阶段通过识别和修复域本体中的语义不一致来执行。本论文中不执行 Uschold 和 King 方法的文档阶段，因为它要么不直接用于本体模型的构建，而本体模型是本研究的主要目的，也不影响本体模型。

4 本体的应用

4.1 家谱知识图谱构建

家谱是指以血缘关系为核心，以家族宗族人员关系的为纽带的家族历史记录资料。家谱对进行国家教育、开展家风宣传、促进祖国统一以及提高中华民族的整体荣誉与自信，具有举足轻重的作用。之所以研究家谱模型系统是为了更好地实现家谱信息的检索与收藏，方便对家谱信息实现共享与保存，以发掘家谱的隐藏价值。这个系统的主要工作是在已有的家谱信息基础上实现对信息的查询，建立家族人物关系网络，并使用图的形式展现家谱的，便于人们了解整个宗族的发展历史与家族文化。

利用本体创建家谱本体的优势在于结构清晰，冗余度较小。该系统采用 protégé 构建家谱本体，先由姓氏作为父类，各种具体姓氏作为子类，然后定义宗族之间的各种关系，为接下来的规则推理提供前提，最后是创建个体属性以及他们的相关属性，个体之间的关系不需要全部定义，只要定义直系亲属的关系。

在 Jena 中没有直接的接口可以实现对 OWL 文件的直接操作，需要借助一个 Model 来辅助进行。首先用模型读取 OWL 文件，把本体相当于复制到 Model 中，然后通过 java 对模型进行内容的增删改查，从而把这个结果在反射给 OWL 文件，实现了对 OWL 文件的修改。[12]

4.2 智慧城市知识图谱模型

智慧城市已经成为一种城市发展理念。据统计，目前 100 %的副省级城市、89 %的地级以上城市、49 %的县级城市已经开展智慧城市建设，累计参与的地市级城市数量有 300 余个。城市管理运营包含民生、交通、教育、医疗、维稳等几十个方面，在智慧城市概念被提出之前，它已经经历了电子化阶段，被称为电子政务、电子警务等。IB M 公司最早在 20 09 年提出智慧城市的概念，中国于 2011 年开始在宁波、上海等城市探索智慧城市建设。

随着物联网、云计算、大数据等技术的发展，智慧城市建设从感知智能到认知智能逐步提升。5 G 技术的应用将加快提升城市的感知能力，数据采集更快、更多、更全。数据包含了文字、图像、音视频等多模态，要把这些数据用好，需要把这些数据组织成大型的知识库，并将其作为智慧城市的基础资源。[13]

4.3 旅游领域本体的语义检索模型

随着旅游产业和互联网相关技术的迅速发展，如何准确、全面地获取旅游信息资源是亟待解决的问题。基于关键字的传统查询方式因为缺乏语义层次上的处理和表示所以无法将用户输入的检索内容与数据资源库中的内容相匹配，造成结果的不完整、不精确，甚至没有符合用户需求的结果。

为了提高旅游信息的查准率和查全率，一些学者建议将本体技术应用于旅游信息的搜索，并获得了一定的进展。但多半的方法只是基于本体技术的基础并没有充分考虑到实例和属性之间的关系。此外，本体的推理功能未充分用于发现本体中的隐式关系。鉴于上述原因，建立在基于旅游领域本体的基础上，提出了用于发现隐含关系的推理规则。通过分析本体中概念间的关系以及结构上的差别，加之推理规则和相似度算法的支持，将现有基于关键字级别的信息组织方式提高到语义级别，设计出一种可以挖掘出隐含信息的搜索方式，通过了解语义的深度，达到优化搜索结果的目的。[1]

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。

[1]参 考 文 献

[1] 张婷,段跃兴,张月琴.基于旅游领域本体的语义检索模型[J].太原理工大学学报,2020,51(02):220-225.

[2] 陈瑞,曾桢.基于语义网技术的网络农业信息资源描述研究[J].信息技术与信息化,2020(05):160-163.

[3] Berners-Lee.Semanticweb-XML2000[EB/OL].[2014-01-20].http://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html

[4] Neches R，Fikes R E，Gruber T，et al． Enabling Technology for Knowledge Sharing［J］． AI Magazine，1991，12(3) : 36－56．

[5] Gruber T． A Translation Approach to Portable Ontology Specifi- cations［J］． Knowledge Acquisition，1993(5) : 199－220．

[6] Borst W N． Construction of Engineering Ontologies for Knowledge Sharing and Reuse［D］． Ph． D thesis，1997．

[7] Studer B，Benjamins V R，Fensel D．Knowledge Engineering: Principles and Methods［J］． Data and Knowledge Engineering，1998，25(1 /2) : 161－197．

[8] 张晓林，李 宇． 描述知识组织体系的元数据［J］． 图书情报工作，2002(2) : 64－69．

[9] 李 景． 本体理论在文献检索系统中的应用研

究［M］．北京: 北京图书馆出版社，2005: 5－6．
[10] 汤艳莉，赖茂生． Ontology 在自然语言检索中的应用研究［J］．现代图书情报技术，2005(2) : 33－36，52．
[11] M.Uschold and M.King, “Towards a Methodology for Building Ontologies,” In Proceedings of IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, Canada, pp. 1-13, 1995.
[12] 孙洪伟,司唯山,纪兆辉.基于本体的家谱知识图谱构建及信息检索系统的设计实现[J].计算机产品与流通,2020(09):156.
[13] 臧根林,王亚强,吴庆蓉,占春丽,李熠.智慧城市知识图谱模型与本体构建方法[J].大数据,2020,6(02):96-106.