

语义 Web 技术及其逻辑基础

袁金平, 鲍爱华, 姚莉

(国防科学技术大学信息系统与管理学院, 长沙 410073)

摘要: 随着语义 Web 的产生和发展, 人-机及机-机间的交流与协作将变得更加方便。该文从介绍语义 Web 概念及体系结构入手, 对关键技术 XML, RDF 及 Ontology 进行了对比分析说明, 同时研究了其逻辑基础(描述逻辑), 分析了其语法语义及推理任务。对语义 Web 的未来发展热点问题进行了展望。

关键词: 语义网; 描述逻辑; 本体; 可扩展标识语言; 资源描述框架

Semantic Web Techniques and Its Logic Foundations

YUAN Jin-ping, BAO Ai-hua, YAO Li

(College of Information System and Management, National University of Defense Technology, Changsha 410073)

【Abstract】 With the birth and development of semantic Web, the communication and cooperation between human and computer or computer and computer become more and more convenient. This paper introduces the concept of semantic Web and its architecture, then analyses the key-techniques including XML, Resource Description Framework(RDF) and Ontology contrastively. After that, logic foundations(description logics) of semantic Web is researched, as well as its syntax, semantics and reasoning tasks. And a prospect of the hotspot in the future of semantic Web is given.

【Key words】 semantic Web; description logics; Ontology; XML; Resource Description Framework(RDF)

WWW 的产生改变了人类的交流方式。随着超文本系统的发明, Web 技术有了巨大的飞跃, 使得人与人之间的信息交流和共享更加方便自如。随着互联网上信息的剧增, 人类往往被淹没在信息的海洋里, 急需机器的自动处理和理解能力来帮助人类获得相关的信息。然而, 目前大多数的网页设计都是针对人类本身的, 不便于机器的自动处理和理解, 这就使得人们必须变革当今的 Web, 往其中加入语义——使人类和机器都可以理解所表示的内容。语义 Web 也就有了诞生的需要。如今, 语义 Web 已成为 Web 领域的研究热点, 它是未来实现智能化网络服务的新基础。

1 语义 Web 概述

语义 Web 是使 Web 内容能被机器理解的构想。目标是扩展当今的 Web, 使得网络中的信息具有语义, 实现机器可理解, 从而自动处理信息, 便于人-机间的交互和协作。语义 Web 是 Tim Berners-Lee 在 1998 年首次提出的。在 2 年后的 XML 大会上, 他对这一概念作了进一步完整的阐述, 同时提出了语义 Web 的体系结构框架。

1.1 定义

简单的说, 语义就是资源的含义(资源是一个很广泛的概念, 它可以是网站、网页, 甚至是网页中的部分内容); 语义 Web 是具有明确语义的能够被机器和人理解的网络。也就是说, 语义 Web 是一种能够理解人类语言的智能网络, 它使得人-机间及机-机间的交流变得像人-人间的交流一样便利流畅。

定义 1 语义 Web 是一个网, 它包含了文档和文档的一些部分, 描述了事物间的明显关系, 且包含语义信息, 以便于机器的自动处理^[1]。

这是 Tim Berners-Lee 在 2000 年的 XML 大会上首次给出

的完整定义。他强调语义 Web 是当今 Web 的一个扩展, Web 中的信息被赋予了明确定义的语义, 是机器可以“理解”并处理的。语义 Web 中的各种资源不再只是各种相连的信息, 还包括信息的明确含义, 能够被机器理解并自动处理, 从而提高计算机处理信息的自动化和智能化能力。总的来说, 它不是一个全新的 Web 形式, 而是现有 Web 上的扩展和延伸。

同时, 不同组织从不同角度对语义 Web 也有不同的理解。下面简单介绍一些学术组织和工业上对语义 Web 的几种理解。

定义 2 语义 Web 是对当前 Web 的一种扩展, 其中的信息被赋予明确定义的含义, 使机器和人能更好地协同工作。

定义 3 语义 Web 是一个数据的 Web, 从某种形式上类似于一个全球性的数据库。语义 Web 方法是要开发一种语言, 它能以机器可处理的形式表达信息。

定义 4 语义 Web 是对 WWW 上数据的表示。它是一个由 W3C 带领的, 许多其他研究人员和产业界合作参与的协作研究活动。它的基础是集成基于 XML 语法和 URIs 标识的各种应用的 RDF。

总的来讲, 以上概念描述了语义 Web 的共同特点: 信息是机器可理解并自动处理的; 是当今 Web 的延伸。

1.2 体系结构框架

Tim Berners-Lee 在 2000 年的 XML 大会上给出语义 Web 定义的同时也提出了它的体系结构框架模型, 如图 1 所示, 它

基金项目: 国家自然科学基金资助项目(70371008)

作者简介: 袁金平(1984 -), 男, 硕士研究生, 主研方向: 智能决策技术, 语义 Web; 鲍爱华, 博士研究生; 姚莉, 教授、博士生导师

收稿日期: 2008-04-16 **E-mail:** yuanjp_nudt@sina.com

是个功能逐层增强的层次化结构^[1]。

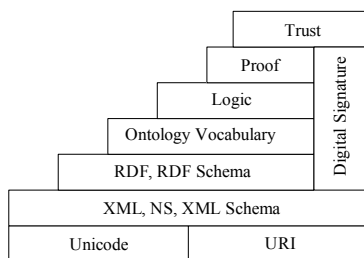


图1 语义 Web 体系结构框架

该体系中从低层到高层分别为：Unicode 和 URI, XML, RDF, Ontology, Logic, Proof, Trust。同时, Digital Signature 贯穿全过程, 建立语义 Web 的安全机制。核心层为 XML, RDF 和 Ontology。

第 1 层 Unicode 和 URI 是整个语义 Web 的基础。

Unicode 是一种字符编码标准, 用于资源的编码; URI 负责无二义地标识 Web 上的任意一个资源。

第 2 层 XML+NS+XML Schema 是语法层, 用于表示数据的内容和结构。

XML 提供了文档结构化的语法, 实现了文档结构与文档表现形式的分离。XML Schema 是约束 XML 文档结构的语言。XML 名字空间是名字的一个集合, 用于文档元素和属性名有效性的验证, 由 URI 引用来标识。

第 3 层 RDF+RDF Schema 是数据层, 用于描述 Web 上的资源及其类型。

RDF 是结构化的元数据编码、交换和重用的一个基础。RDF 数据模型提供了简单的语义, RDF 属性可以看作是资源的属性, 同时又表达了资源之间的关系。RDF Schema 为 RDF 模型提供了一个基本的类型系统。

第 4 层 Ontology Vocabulary 是本体层, 用于描述各种资源之间的联系。

本体层提供一个能明确地形式化地定义术语含义及术语间关系的语言。

第 5 层~第 7 层是 Logic, Proof 和 Trust, 是在下面 4 层的基础上进行的逻辑推理操作, 包括公理和推理规则、认证机制及信任机制。

最后, 贯穿全过程的是数字签名和加密。这是实现 Web 信任机制的关键技术, 为机器提供验证某信息是否可信的证据。

2 关键技术

前面的语义 Web 体系结构框架讲到语义 Web 的核心层为 XML, RDF, Ontology, 主要实现语义 Web 的语义信息表示, 是实现语义 Web 的 3 个关键技术。XML 实现了文档结构化, 但文档信息并不包含任何语义; RDF 数据模型提供了简单的语义; Ontology 使得语义表达能力更加丰富。目前 XML, RDF 技术相对比较成熟, 而本体技术尚待进一步的发展完善。

2.1 XML

XML 作为一种资源描述语言, 因其灵活、通用、丰富的结构化信息表达方式, 而被广泛接受, 成为整个语义 Web 的基石。它可以让信息提供者根据需要, 自行定义标记及属性名, 从而使 XML 文件的结构能复杂到任意程度。它具有良好的数据存储空间和可扩展性、高度结构化以及便于网络传输等优点, 再加上其特有的 NS 机制及 XML Schema 所支持的多种数据类型与校验机制, 使其成为语义 Web 的首要关键

技术, 也是整个体系结构的基础。

2.2 RDF

资源描述框架(Resource Description Framework, RDF)是 W3C 组织推荐使用的用来描述资源及其之间关系的语言规范, 定义了一种用以描述资源及其关系的模型, 是语义 Web 实现的关键技术之一, 也是语义信息描述的有效手段。它不仅是描述数据的框架, 而且也是表示数据的框架。具有简单、易扩展、开放性、易交换和易综合等特点。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问题, 使得所描述的资源成为机器可理解的信息。

RDF 由 3 个部分组成: RDF Data Model, RDF Schema 和 RDF Syntax。

RDF Data Model 提供了一个简单但功能强大的模型, 通过三元组<资源, 属性, 值>来描述特定资源。如: <"Tim", authorOf, "Book">表示 Tim 为 Book 的作者; <"Book", hasPrice, "\$50">表示 Book 的价格是 50 美元。

同时也有其他的表示方式: 如可表示为 authorOf<"Tim", "Book">和 hasPrice<"Book", "\$50">, 也可以表示为有向图, 如图 2 所示。

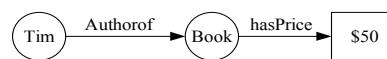


图2 RDF 的有向图

模型定义为: (1)包含一系列的节点 N ; (2)包含一系列属性类 P ; (3)每一个属性都有一定的取值 V ; (4)模型是一个三元组: < N, P, V >或 < N, P, N >; (5)每一个 Data Model 可以看成是由节点和弧构成的有向图。

RDF Schema 使用一种机器可以理解的体系来定义描述资源的词汇, 其功能就像一个字典, 可以将其理解为大纲或规范。其作用是: (1)定义资源以及属性的类别; (2)定义属性所应用的资源类以及属性值的类型; (3)定义上述类别声明的语法; (4)申明一些由其他机构或组织定义的元数据标准的属性类。

RDF Syntax 构造了一个完整的语法体系以利于计算机的自动处理, 它以 XML 为其宿主语言, 通过 XML 语法实现对各种元数据的集成。

2.3 Ontology

语义信息的交流必须以共同的理解为前提, 否则双方信息产生误解或者完全不理解。在语义 Web 中, 这种“共同理解”是由 Ontology 建立和提供的。Ontology 最早是一个哲学上的概念, 从哲学的范畴来说, Ontology 是客观存在的一个系统的解释或说明, 关心的是客观现实的抽象本质。而在人工智能(AI)领域, Ontology 是共享概念模型的明确的规范说明。Gruber 给出的本体定义为: 本体是共享概念模型的形式化规范说明。Fensel 对这个定义进行分析后, 认为本体的本质表现在 4 个关键词上: 概念化(conceptualization)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; 明确(explicit)指概念及它们之间的联系都有明确的定义; 形式化(formal)指精确的数学描述, 计算机可读; 共享(shared)指 ontology 中体现的知识是共同认可的。

语义 Web 可以看作是在本体理论上对当今 Web 所进行的扩展, 使 Web 上的信息具有机器能理解的语义, 同时在本体的支持下实现不同实体间语义层面的互通和互操作。本体从以下 2 个方面使 Web 发挥它的极大潜能:

(1)本体为信息定义形式化语义以便计算机处理;

(2)本体通过双方认可的术语,定义现实世界中的语义,将机器处理的内容与语义联系起来。

2.4 XML, RDF 与 Ontology 比较

XML, RDF 与 Ontology 居于语义 Web 体系结构的不同层次上,解决不同层面的问题。其中,XML 主要是实现信息语法层面的互操作;RDF 实现信息语义层面的互操作;而 Ontology 则是解决不同词汇相同语义及相同词汇不同语义的问题。三者在语义表示过程中相互补充,发挥了各自的优点。

下面,通过表 1 的比较研究来阐述三者的相互联系和区别。

表 1 XML, RDF 及 Ontology 比较

关键技术	比较项			
	地位与作用	语义特性	特点	联系
XML	语法层:用于表示数据的内容和结构,是数据交换格式和数据描述标准	不具备语义描述能力	易扩展、灵活、结构化、易传输	(1)相互结合共同解决语义 Web 不同层面的问题; (2)XML 是 RDF 的一种表示方式,两者互为补充;
RDF	数据层:用于描述 Web 上的资源及其类型	具有简单语义表达能力	简单、易扩展、开放性、易交换、易综合	(3)Ontology 的加入使得 XML 和 RDF 的表示更加具有生命力
Ontology	本体层:解决同义及多义词问题,提供“共同理解”的基础	具备丰富语义表达能力	可靠完备的逻辑基础、对概念的严格描述	

3 语义 Web 的逻辑基础

要让机器能够理解并自动处理语义 Web 上的信息,首先需要解决的就是语义 Web 中知识表示和本体问题。语义 Web 是用本体表达语义的,语义 Web 中本体的使用需要一种合适设计(well-designed)、合适定义(well-defined)及 Web 相容的本体语言。鉴于描述逻辑(Description Logics, DLs)在语义、可判定性以及面向对象的分类表示等方面的优点,一般的本体语言可以建立在描述逻辑的基础上^[2]。因此,一些学者将描述逻辑作为语义 Web 的逻辑基础。目前的标准 OWL(Ontology Web Language)即是以描述逻辑为逻辑基础建立的用于表达语义 Web 的本体语言。

3.1 描述逻辑

描述逻辑是一种基于对象的知识表示的形式化。它是一阶逻辑的一个可判定的子集,具有合适定义的语义,并且具有很强的表达能力^[2]。

描述逻辑是建立在 2 个基本元素——概念(Concept)和角色(Role)之上的,由构造子(Constructor)从简单概念和角色构造出复杂的概念和角色。其中,概念对应经典逻辑中的一元谓词(解释为一个领域的元素集合);角色对应二元谓词(解释为一个领域集合上的元素与元素的一种二元关系),构造子决定着语言的表达能力。如 $Parent(x)$ 表示 x 是家长, $has_child(x, y)$ 表示 x 有孩子 y ; $Person \cap \forall hasChild.Male$ 表示那些所有的孩子都是男孩的人。ALC 是最基本的描述逻辑,包括如下构造子:交(\cap),并(\cup),非(\neg),存在量词(\exists)和全称量词(\forall)。描述逻辑是语法(syntax)+语义(semantics)的;语义是通过解释 $I=(\Delta^I, \bullet^I)$ 来给出。其中, Δ^I 是解释领域(非空集合); \bullet^I 是解释函数(将原子概念 C 映射到 2^{Δ^I} 的子集,将原子角色 R 映射到 $2^{\Delta^I \times \Delta^I}$ 的子集)。ALC 的语法和语义如表 2 所示。

在 ALC 的基础上添加不同的构造子,可以扩展成不同表达能力的描述逻辑^[3]。例如可以添加数量约束(字母 N 表示):

nR (至少的限制), $\leq nR$ (最多的限制),得到 ALCN;添加逆

构造子(字母 I 表示): R^- (角色 R 的逆),可以得到 ALCI;加入模态算子可以扩展为动态描述逻辑(DDL)等,在此不作进一步的介绍。

表 2 ALC 的语法和语义

构造子	语法	语义	举例
原子概念	A	$A^I \subseteq \Delta^I$	<i>Student</i>
原子角色	R	$R^I \subseteq \Delta^I \times \Delta^I$	<i>has_child</i>
顶	T	$T^I = \Delta^I$	<i>Male</i> \neg <i>Male</i>
低	\perp	$\perp^I = \emptyset$	<i>Male</i> \cap \neg <i>Male</i>
交	$C \cap D$	$(C \cap D)^I = C^I \cap D^I$	<i>Teacher</i> \cap <i>Student</i>
并	$C \cup D$	$(C \cup D)^I = C^I \cup D^I$	<i>Teacher</i> \cup <i>Student</i>
非	$\neg C$	$(\neg C)^I = \Delta^I \setminus C^I$	\neg <i>Male</i>
存在量词	$\exists RC$	$\{a \in \Delta^I \mid \exists b.(a, b) \in R^I \wedge b \in C^I\}$	$\exists has_child.Teacher$
全称量词	$\forall RC$	$\{a \in \Delta^I \mid \forall b.(a, b) \in R^I \rightarrow b \in C^I\}$	$\forall has_child.Male$

3.2 推理服务

顾名思义,推理就是在已知知识的基础上发掘蕴含知识,即由显性知识得到隐性知识。要实现机器的自动处理,就必须具备对某些隐含知识的推理,同时推理对保障知识库一致性非常重要。基本推理问题主要有:一致性检测,概念可满足性检测,包含检测,实例检测等。其中概念可满足性检测关键推理,其他推理问题均可转换成概念可满足性问题^[3]。

通常,一个知识库 $KB = \langle TBox, ABox \rangle$ 。其中, $TBox$ 描述概念及其关系的术语公理集; $ABox$ 描述个体实例的断言公理集。对于 $TBox T, ABox A$ 及解释 $I = (\Delta^I, \bullet^I)$, 定义^[3]:

可满足:存在 T 的某个解释 I , 使得 C^I 非空, 则称 C 是可满足的;

包含:对 T 的每一个解释 I , 都有 $C^I \subseteq D^I$, 则 $C \subseteq_T D$;

等价:对 T 的每一个解释 I , 都有 $C^I = D^I$, 则 $C \equiv_T D$;

相离:对 T 的每一个解释 I , 都有 $C^I \cap D^I = \emptyset$, 则 C 与 D 相离。

在描述逻辑中,可利用下面的定理来将其他推理问题转换为概念的可满足性推理。

定理 1 (变形为可满足性检测)对于概念 C 和 D , 有:

- (1) $C \subseteq D \Leftrightarrow C \cap \neg D$ 是不可满足的;
- (2) $C \equiv D \Leftrightarrow C \cap \neg D$ 及 $\neg C \cap D$ 都是不可满足的;
- (3) C 和 D 是相离 $\Leftrightarrow C \cap D$ 不可满足的。

描述逻辑是一簇知识表示语言,以结构化、形式化的方法来表示特定应用领域的知识。它是通过构造子将原子概念和角色来构建复杂的概念和角色,因此构造子对描述能力起着决定作用。研究各种构造子的表达能力及对应的推理复杂性是描述逻辑的研究核心,而如何在描述能力和推理的复杂性进行折中则是描述逻辑的根本问题。描述逻辑具有很好的建模和推理特性,已经成为 Web 本体描述语言的逻辑基础。它为语义 Web 本体语言提供形式化的基础和自动化的推理服务。

4 结束语

语义 Web 在 W3C 及其他研究机构的带领下,取得了巨大的进步,其应用领域不断拓广。虽然语义 Web 的逻辑基础很完备,其关键技术发展相对比较成熟,但整个体系结构的最高层——逻辑表示和处理层,尚待进一步的研究。要实现人们最初对语义 Web 的梦想,其要走的路还很远,面临的研究挑战还很多。如:内容的可获得性;本体的开发、合并及演化;语义 Web 语言的稳定性;证明和信任模型的开发。本

(下转第 199 页)