

Semantic Web 研究综述

龚洪泉 张敬周 钱乐秋 任洪敏

(复旦大学计算机信息与技术系 上海 200433)

摘 要 近年来, Semantic Web 逐渐成为 WWW 领域的研究热点以及智能化网络服务和应用开发中的关键技术之一。归纳了 Semantic Web 技术的研究背景和主要发展历史。在分析了典型的 Semantic Web 概念后, 给出了 Semantic Web 的定义。通过讨论 Semantic Web 构想的层次框架模型, 指出了各个层次扮演的角色, 并着重分析了 Semantic Web 的重要研究领域, 指出了它们在 Semantic Web 构架中的核心作用。通过分析 Semantic Web 的应用领域和相关开发工具以及面临的问题和挑战, 指明了 Semantic Web 研究和实践的方向。作为总结, 给出了 Semantic Web 领域下一步的研究趋势。

关键词 Semantic Web 网络服务 RDF RDF schema Topic maps 本体 代理 规则 逻辑 证明 信任

SEMANTIC WEB RESEARCH: A SURVEY

Gong Hongquan Zhang Jingzhou Qian Leqiu Ren Hongmin

(Department of Computing and Information Technology, Fudan University, Shanghai 200433, China)

Abstract Semantic Web is emerging as one of the primary research areas in WWW recently and one of the key technologies to the development of intelligent Web services and applications. The research background and momentous history of Semantic Web are summarized, and the vision of Semantic Web is brought up based on analyzing and comparing several classical definitions about Semantic Web. Based on discussing the hierarchical architecture of Semantic Web vision, the roles of each level are extracted out, and the key research areas are intensely analyzed. After analyzing the application areas and relevant developing tools as well as problems and challenges of Semantic Web, directions of researching and practicing Semantic Web in real world are discussed. Finally, it is concluded with some significantly promising tendency about Semantic Web research.

Keywords Semantic Web Web services RDF RDF schema Topic maps Ontology Agent Rule Logic Proof Trust

0 引 言

WWW 为人们发布和获取信息提供了一个全新的服务平台, 但是目前只能通过人工浏览 Web 信息才能获得相应的服务。为了在 Web 上提供自动的、智能化的服务, 机器必须要能理解它所处理的数据、消息和过程的含义。前几年提出的 Semantic Web 是使 Web 信息能被机器理解的一种构想, 是对当前 WWW 的一种扩展, 其中的信息被赋予明确定义的含义, 使机器和人能够更好地协同工作。有了 Semantic Web 技术和相关的开发工具, 就可以用一种机器可处理的方式来表达 Web 信息的含义, 从而为实现智能化网络服务和应用奠定基础。

本文围绕 Semantic Web 构想的层次框架, 对 Semantic Web 的相关研究进行了讨论。

1 Semantic Web 研究背景

WWW 是 20 世纪 90 年代最引人注目的科技成就, 它彻底改变了人们发布信息 and 获取信息的方式。WWW 的重要性在于它所提供的服务, 从 20 世纪 90 年代中期开始, 互联网及 Web 上提供的服务成了整个 IT 业发展的动力。为了在 Web 上提供智能化服务, 机器应当能够理解它所处理的数据、消息和过程的含义。然而, 目前网络上大部分的内容都不能以算法的形式来描述, 尽管有些网页是根据数据库动态生成的, 但机器并没有一种可靠的方法来处理它们的语义。

在网络服务方面, 尽管提出了 UDDI、WSDL 等技术标准, 但它仍然存在很多尚待解决的问题: 服务描述在语义上存在二义性; 服务发现 (discovery)、匹配、检索的查全率和查准率较低; 服务集成仍然需要人工干预, 不能完全自动化。在 CORBA、COM、EJB 等以跨平台互操作为目标的中间件研究和应用方面, 也同样存在中间件描述、检索和自动化集成等方面的问题。无论是简单的网页信息, 还是网络服务或中间件等复杂的信息和服务的聚集体, 它们的主要问题都在于: 描述信息存在语义二义性, 导致机器无法自动地理解和处理它们; 作为一个全球性的信息网络, WWW 还远没有充分发挥它的潜能。总之, 在目前情况下, 机器还不能自动地、有效地、有目的地发现、集成和复用网络上的各种数据。从技术角度来说, 这是创建智能化网络服务的根本障碍。

Semantic Web 是由 Tim Berners-Lee (WWW、URLs、HTTP 及 HTML 的发明者) 提出的使 Web 内容能被机器理解的构想。它作为有逻辑内涵的 Web 内容提供语义结构及相应的信息访问和集成机制, 从而为构建一个能够按需集成各种信息、完成特定任务的信息管理奠定基础。Semantic Web 已成为 WWW 领域的研究热点, 它是未来实现智能化网络服务的新基础, W3C (World Wide

收稿日期: 2003 - 11 - 19。国家高技术研究发展计划资助项目 (2001AA1100241)。龚洪泉, 博士, 主研领域: Semantic Web, Web Service, 基于构件的系统组装与集成。

Web Consortium)和其它许多机构正在积极地为实现这个构想而努力。

2 Semantic Web 发展史

1989年3月,Time Berners-Lee 在文献[1]中提出的分布式超文本系统图所包含的链接都是蕴含语义的,该图最早提出了Web信息包含语义的思想。1994年5月第一届国际WWW会议召开时,Web完全是一个扁平的、毫无逻辑内涵的世界,Time Berners-Lee 在会议报告中^[2]特别强调了Web信息必须包含语义。这个报告最早明确指出了语义在Web中的重要地位。1997年12月,W3C会议报告中^[3]指出,Web的第一个目标是成为人们通过知识共享进行交流的媒介,第二个目标是成为人们协同工作的媒介。如果所有的数据都有一个公共的、定义良好的接口,Web协作将更加方便。这次关键性的会议为Semantic Web的正式提出做了铺垫。

1998年9月,Time Berners-Lee 在文献[4]中提出了Semantic Web的总体框架,并概括介绍了它的各个组成部分,拉开了Semantic Web研究的序幕。同时文献[5]特别指出了Semantic Web在WWW未来发展中的重要作用。其后,文献[6,7,8]对前两篇文献作了补充说明,对Semantic Web的基础—资源描述框架中的有向标记图(Resource Description Framework [RDF] 's Directed Labeled Graph)模型—与其它数据模型作了详细的对比。这些文章澄清了与人工智能、知识表达、实体联系模型、关系数据库以及推理系统之间的联系和区别,为Semantic Web研究扫清了障碍。2000年12月,Tim Berners-Lee 在XML 2000会议报告中^[9]首次给出了Semantic Web功能逐层增强的层次结构图,指明了Semantic Web的研究框架。

之后,Semantic Web的相关研究工作迅速展开。W3C方面,2001年2月9日启动特别的Semantic Web活动^[10],3月30日特许成立RDF核心工作小组^[11],11月1日成立Web本体(Ontology)工作小组。除了W3C,美国国防部的DAML (DARPA Agent Markup Language)项目也是Semantic Web活动的主要推动力^[10]。欧盟方面,2001年6月13日设立Web本体(OntoWeb)研究网络^[12]。

2001年7月,由国家自然科学基金(NSF)资助的第一届Semantic Web工作研讨会在斯坦福大学成功召开^[13]。会上主要对本体及其维护、语义互操作及其集成和组装、Web Services 及其应用等三个方面进行了讨论。2002年5月,第十一届WWW会议^[14~17]对Semantic Web、RDF 构架、Web 本体语言、Semantic Web 高级开发等领域的发展进行了讨论。2002年6月,第一届国际Semantic Web会议^[18]在意大利撒丁岛召开。这次会议对Semantic Web各个技术发展方向的研究工作进行了讨论,并对Semantic Web在实践面临的各种挑战和困难提出了方向性的建议。



图1 WWW 发展历史

从上面的发展历史可以看出,Semantic Web 属于第三代WWW(图1)^[21]。90年代初期最早的WWW使得通过Internet浏览文档成为可能,即实现了文档表现形式(presentation)与文档位置(location)的分离;90年代后期XML的出现实现了文档结构(structure)与文档表现形式的分离,即文档是存储数据的地方,

但数据可以根据不同的目的有不同的表现形式;现在的Semantic Web就是要实现文档含义与文档结构的分离,也就是使文档具有机器可处理的语义。

3 Semantic Web 的定义和模型

3.1 Semantic Web 的定义

下面介绍并分析Time Berners-Lee 等人和W3C从不同角度对Semantic Web提出的概念和定义。然后,根据我们对Semantic Web的理解,给出能够反映Semantic Web基本特征的定义。

定义1 Semantic Web是一个数据的Web,从某种形式上说类似于一个全球性的数据库。Semantic 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^[4]。

该定义是在Semantic Web整体框架提出之前给出的,主要关注的是信息背后的网络空间关联结构以及如何开发语言来表达Web信息等问题。

定义2 Semantic Web是对WWW上数据的表示。它是一个由W3C带领的,许多其它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(Uniform Resource Identifier)标识的各种应用的RDF^[10]。

这个定义是W3C根据Semantic Web目前的研究内容和技术基础给出的。它强调了Semantic Web研究活动的一个全球性的协作研究活动,强调了研究活动的开放性。

定义3 Semantic 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^[19]。

该定义在较高的抽象层次上对Semantic Web进行了概括,首先指出了“信息被赋予明确定义的含义”是Semantic Web的基础,其次,“协同工作”体现了Semantic Web的最终目标就是要成为协同工作的媒介。

比较上述各种定义可以发现:这三个定义都强调了Semantic Web需要一种数据表示方法。尽管各种定义都从不同角度关注Semantic Web,但都没有完整描述其各个方面的特征。

Semantic Web一方面指代由机器可处理的信息所组成的抽象信息空间,另一方面指代Semantic Web技术所组成的研究对象,“Semantic”含有“机器可处理”的意思,而不是自然语言或人的推理,对信息来说,“Semantic”表达了对信息能作哪些操作。根据我们对Semantic Web的理解,将Semantic Web定义为“Semantic Web”是对当前Web的一种扩展,是一个信息的Web,这些信息被赋予明确定义的含义,是机器可处理的;Semantic Web的技术基础上是XML和RDF;其基本实现方法是开发功能逐层增强的形式化信息规约语言,用以唯一确定信息的含义;其最终目标是成为智能化网络服务和应用开发的基础设施,成为机器与人协同工作的媒介。”

3.2 Semantic Web 模型

在XML 2000会议上,Time Berners-Lee 首次给出了Semantic Web基本构架,它是个功能逐层增强的层次化结构(图2)^[19,20]。

URI和Unicode层是标识Semantic Web对象和使用国际字符集的基本手段。XML层以及名字空间和Schema定义是集成Semantic Web定义与其它基于XML的标准的基础。RDF和RDF Schema[二者合称为RDF(S)]用来描述和定义由URIs引用的对象及词汇,并指定资源和链接的类型。本体层用来定义不同概念之间的关系,以支持词汇的演化。规则层定义通过已有信息导出新信息的规则。逻辑框架层为基于规则的系统提供一个描述公理的框架,证明层执行规则并做出相应的评估。信任层为

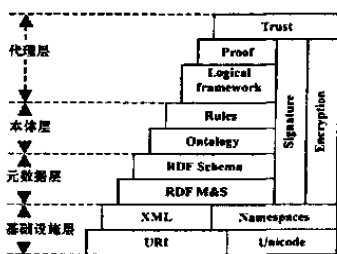


图 2 Semantic Web 层次结构图

应用程序是否信任一个给定的证明提供检测机制。数字签名和加密技术用来检测文档的改动情况,是增强 Web 信任的手段。

3.2.1 URI URI 是 WWW 的核心概念之一,它能够无二义地标识 Web 上的任意一个资源,其思想是在需要的时候通过链接引用资源,因此不需要对资源进行拷贝或集中管理。

3.2.2 Unicode Unicode 是一种新的字符编码标准,它支持世界上所有的语言。不论在什么平台上,不论在什么程序中,不论使用什么语言,每个字符都对应于一个唯一的编码值。它是 Semantic Web 多语种支持的基础。

3.2.3 XML 及名字空间 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,根据不同的目的同一个文档可以有不同的表现形式。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。

3.2.4 RDF(S) 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,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,因此 RDF 数据模型对应于传统的属性-值对,又类似于实体联系图。RDF Schema 为 RDF 模型提供了一个基本的类型系统,其目的就是定义资源的属性,定义被描述的资源的类,并对类和关系的可能组合进行约束,同时提供约束违例的检测机制。

3.2.5 本体层 虽然 RDF(S) 能够定义对象的属性和类,并且还提供了泛化 (generalization) 等简单语义,但它不能明确表达描述属性或类的术语的含义及术语间的关系。本体层就是要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。

3.2.6 逻辑、证明和信任 除了本体层定义的术语关系和推理规则外,还需要有一个功能强大的逻辑语言来实现推理。证明语言允许服务代理在向客户代理发送断言的同时将推理路径也发送给客户代理。这样应用程序只需要包含一个普通的验证引擎就可以确定断言的真假。但是,证明语言只能根据 Web 上已有的信息对断言给出逻辑证明,它并不能保证 Web 上所有的信息都为“真”,因此软件代理还需要使用数字签名和加密技术用来确保 Web 信息的可信性。

3.2.7 数字签名和加密 数字签名简单地就说就是一小段数据加密块,机器和软件代理可以用它来无二义地验证某个信息是否由特定的可信的来源提供,它是实现 Web 信任的关键技术。公共密钥加密算法是数字签名的基础。

4 Semantic Web 研究重点

4.1 语义元数据描述

解决如何让机器理解 Web 数据的一种方法是使用元数据来描述 Web 中的数据。元数据就是“关于数据的数据”,在 Web 环境中特指“描述 Web 资源的数据”。在描述 Web 资源的元数据方面,目前主要有 W3C 的 RDF(S) 标准和 ISO 的 Topic Maps 标

准^[3,4,7,21]。

4.1.1 RDF(S) RDF 是处理元数据的基础,RDF 模型和语法包括表达 Web 资源及其关系的元数据模型以及用 XML 来编码和传输元数据的语法。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,因此 RDF 数据模型对应于传统的属性-值对,又类似于实体联系图。

基本 RDF 数据模型包含如下三种对象类型^[22]:

资源 (Resources) 所有用 RDF 表达式描述的事物都叫做资源。

属性 (Properties) 属性是资源所具有的特定方面、特征或关系。

语句 (Statements) RDF 语句由资源、命名属性以及属性的取值组成。属性值可以是文字或其它的资源。

每个 RDF 语句都可以用三元组,RDF/XML 或有向标记图表示,它们所表达的含义是等价的。从逻辑语言学角度说,三元组就是简单语句的“主-谓-宾”。例如,“人民日报由人民日报社主办”和“人民日报是一种综合性日报”可表示成表 1 所示:

表 1 RDF 语句的不同表示

三元组	RDF/XML	有向标记图
主语 谓语 宾语 <人民日报,主办,人民日报社> <人民日报,报刊类型,综合日报>	<pre> <rdf:RDF> <rdf:Description about="人民日报"> <s:Sponsor>人民日报社</s:Sponsor> <s:JournalType>综合日报</s:JournalType> </rdf:Description> </rdf:RDF> </pre>	

除了上面的简单对象类型,为了便于引用一组资源,RDF 还定义了如下三种容器对象:

包 (Bag) 资源或文字的无序列表。用来表示属性有多个取值,但取值顺序并不重要。

序列 (Sequence) 资源或文字的有序列表。用来表示属性有多个取值,而且取值顺序是重要的。

替换集 (Alternative) 代表某个属性 (单一) 取值选项的资源或文字的列表。可以取列表中的任意一个值。

在 RDF 中,资源是通过语句来描述的,但语句不是资源,要复用语句所包含的信息,就必须用一个新的资源 X 对语句建模,这样 X 就可以被其它语句作为宾语引用,其它语句也可以对 X 进行更详细的描述。该建模过程在知识表达领域形式化地叫作具体化 (reification),语句模型 X 叫作具体化语句,它有如下四个属性:

主语 (subject) 主语属性标识原始语句所描述的资源,即主语属性的取值为原始语句中的资源。

谓语 (predicate) 谓语属性标识原始语句中的属性,即谓语属性的取值为原始语句中的属性。

宾语 (object) 宾语属性标识原始语句中的属性值,即宾语属性的取值为原始语句中的属性取值。

类型 (type) 所有具体化语句 X 都是 RDF:Statement 的实例,即类型属性的取值为 RDF:Statement。

RDF 基本对象类型只能表达简单事实,使用具体化语句,可以方便地复用已有的信息,实现 Web 信息的扩展,为表达任意复杂的事物奠定基础。例如,要表达“王先生说人民日报是一种综合日报”,先用 X 对语句{人民日报,报刊类型,综合日报}进行具体化,然后对 X 的属性“谁说的”进行赋值: {[X], subject, [人民日报]} {[X], predicate, [报刊类型]} {[X], object, [综合日报]} {[X], type, [RDF:Statement]} {[X], whoSays, [Mr Wang]}。

RDF 提供了用语句来表达资源和资源属性的形式化模型,但是没有对属性自身进行刻画,也没有提供描述属性间关系和资源间关系的机制。RDF Schema^[23]为 RDF 模型提供了一个基本的类型系统,其目的就是定义资源的属性,定义被描述的资源类,并对类和关系的可能组合进行约束,同时提供约束违例的检测机制。其中, `rdf:type` 表达实例-类关系。`rdfs:Class`, `rdfs:Property`, `rdfs:subClassOf` 和 `rdfs:subPropertyOf` 为类和属性提供简单的泛化层次结构,实现继承扩展。`rdfs:domain`, `rdfs:range` 用来限制属性的定义域和值域,保证数据的一致性。`rdfs:comment` 和 `rdf:label` 为资源提供简单的注解能力。

RDF(S) 是一个用精确的形式化词汇表达命题的简单断言逻辑语言,其中每个三元组表达一个简单的命题。为了定义和分析 RDF(S) 的语义属性, W3C 用精确的模型理论规约为它提供了一个严格的语义基础^[24]。

综上所述, RDF(S) 是描述元数据的基本框架,它借鉴了实体联系模型、图数据模型、知识表达、结构化文档、数据库约束、面向对象建模等各方面的研究成果。RDF 的核心是关系模型,其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,因此它对应于传统的属性-值对,又类似于实体联系图。属性的定义域和值域来自于数据库中的数据完整性观念。三元组就是逻辑语言学中简单语句的“主-谓-宾”。具体化语句是实现信息复用和信息扩展的基本手段,为表达任意复杂的事物奠定了基础。最后 RDF Schema 还引入了面向对象建模范型 (paradigm) 中的类层次结构和属性继承机制。

4.1.2 Topic Maps Topic Maps(TM)^[25]是用来描述 Web 资源元数据的 ISO 标准,它最初的目标是开发一种合并图书索引的标准,其思想是在 90 年代初 WWW 被广泛应用之前就出现了。TM 有如下三个核心元素:

主题 (Topics) 任意的可以附加信息的话题、概念、人或事物都是主题。

关联 (Associations) 关联用来连接相关的主题。本联本身也可以是主题。

具体取值 (Occurrences) 具体取值是主题的不同具体化形式。

与 RDF 一样, TM 的底层数据模型也是有向图结构。从这个角度来说, TM 可以看作是人工智能中知识表达方面的“语义网络”。一个主题由它的名字、与其它主题的关联,以及具体取值来刻画。可以沿着从一个主题到另一个主题的关联实现对主题的信息搜索。

RDF(S) 和 TM 都是描述 Web 资源元数据的方法。它们的底层数据模型都是图,因此易于实现文档的合并。TM 更面向于应用,因为它一开始是用于图书索引的。RDF 则是一个更一般的底层框架和 Web 数据模型,在其上可以创建特定的语言和应用。

4.2 语义本体描述

虽然 RDF(S) 能够定义对象的属性和类,并且还提供了泛化等简单语义,但它不能明确表达描述属性或类的术语的含义及术语间的关系。本体层就是要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。典型的 Web 本体通常包含一个分类法和一组推理规则。分类法定义类及类间的关系,对于 Web 来说,类、子类以及实体之间的关系是十分强大的工具。通过指定类的属性和允许子类继承属性就可以表达实体间的各种复杂关系。推理规则提供更进一步的能力,它定义通过已有信息导出新信息的规则。根据推理规则和术语间关系,机器就可

以执行一定的推导,或判定两个不同的术语实际上指代同一事物,或从术语间已有的关系推导出新的关系,或完成更高级的推理任务^[19,21]。

本体的概念最早来源于亚里士多德 (公元前 384-322 年) 在形而上学方面对事物存在本质的研究,后来成为哲学的研究分支^[26]。在人工智能领域,最早定义本体的是 Neches 等人^[27]: “本体通过定义主题领域词汇中的基本术语和关系以及术语和关系的组合规则来定义词汇的外延”。即本体不仅包括明确定义的术语,还包括根据规则导出的术语,是术语的一个闭包。Gruber 在 1993 年给出的定义最为著名^[28]: “本体是对概念化的明确规约”。1997 年, Borst 对 Gruber 的定义作了一点改动^[29]: “本体是对共享的概念化的形式化规约”。后来 Studer 等人对这两个定义作了如下的解释^[30]: “概念化”是指在识别出与世界上某些现象相关的概念后得出的关于这些现象的抽象模型; “明确的”意味着概念类型及概念的使用约束都有明确的定义; “形式化”是指本体应当是机器可读的; “共享”的指本体捕获的是已达成共识的知识,即本体并不是个人私有的,而是一个组织所公认的。从 Guarino 等人在文献[31]中对七种主要的本体定义所作的比较可以看出,对不同的研究者,本体可以是一个逻辑理论、一个形式化语义记录、逻辑理论词汇、或概念化规约。

在 Semantic Web 中,本体主要用来对描述属性或类的术语的含义及术语间的关系进行规约。形式化的本体需要良好的设计,良好的设计需要用形式化的语言为本体中所使用的概念和术语指定机器可处理的含义。过去十几年中,出现了许多传统的本体规约语言^[30]: Ontolingua, LOOM, OCML, Hologic, CARIN 等,它们的语法大部分都包含无格式的文本 (类似于 Lisp 的语法),目前这些语言的开发都已经比较稳定。另外,还有很多其它的本体规约语言。有些是面向特定本体的,如 CycL 用于 Cyc 知识库, GRAIL 用于医学领域;有些并不是特别为表达本体而开发的,如 NIKL。最近,出现了许多基于 Web 的本体语言^[30,32~36]: RDF(S), SHOE, XOL, OML, OIL, DAML+OIL 以及最新的 OWL。除了 SHOE 是基于 HTML 语法外,其它的都是基于 XML 语法的,它们之间的关系如图 3 所示:

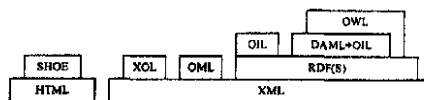


图3 基于 Web 的本体语言金字塔

RDF(S) 提供了简单的本体定义机制 (类层次结构, 属性及属性约束), 但对完整地规约本体来说还存在一定的差距^[37]: (1) 不能定义属性的特征 (刻画); (2) 不能区分成员关系中的必要和充分条件; (3) 不能表达类间的等价关系和不相交关系; (4) 只有属性的定义域和值域约束可以使用; (5) 语言的语义规约过少。

OWL 是 W3C 在 RDF(S) 和 DAML+OIL 的基础上提出的 Web 本体描述语言, 其中心思想是向下兼容 RDF(S) 的语义, 同时又朝形式化描述逻辑 (Description Logic) 方向扩展, 它是基于一阶语义的框架 (面向对象) 描述逻辑系统, 比 RDF(S) 提供了更丰富的属性和类的描述机制^[32~36]: (1) 公理描述。表达本体中属性和类的事实。比如等价类、等价属性、相同个体、相异于、全不同、不相交等。(2) 属性特征约束。比如互逆属性、传递属性、对称属性、函数属性、反函数属性等。(3) 属性类型限制。比如全部取值限制 (所有取值都在某个类型范围内)、部分取值限制 (至

少有一个取值在某个类型范围内)、确切取值限制(在一定条件下属性取值是确定的)等。(4)基数限制。最小基数、最大基数等。(5)类的逻辑组合。类的交、并、补等。

综上所述,本体层为 Semantic Web 提供了一个能够明确地、形式化地定义术语含义及术语间关系的语言,Web 本体语言 OWL 的基础是描述逻辑。作为 Semantic Web 的核心概念,本体用形式化的逻辑语言描述后,应用程序或软件代理就可以根据术语间的关系和推理规则执行自动化的推理,从而为智能应用提供高级的服务。

4.3 软件代理

Semantic Web 是一个层次化的系统,元数据层和本体层对 Web 资源进行了形式化的规约,赋予了 Web 信息机器可理解的含义。机器真正理解和利用 Web 信息来提供智能化的服务是在代理层完成的。随着 OWL 的标准化,基于描述逻辑的本体层研究已经比较成熟。下一步的研究将是软件代理如何使用形式化的 Web 信息完成真正的 Web 任务。软件代理的概念最早在 20 世纪 50 年代中期由麻省理工大学提出,其研究可分为两个阶段:第一阶段从 20 世纪 70 年代后期开始,主要是分布式人工智能的研究;第二阶段从 1990 年左右开始,主要是各种代理和代理系统的研究^[38]。

目前 Semantic Web 中软件代理的研究核心是逻辑层和证明层。逻辑层在本体层定义的术语关系和推理规则之上提供功能更强大的逻辑语言来实现推理,它不仅能够表达逻辑引用,还能够表达谓词逻辑(与、或、非)和量化(存在量词、全称量词)。当逻辑层在本体规约语言之上提供图灵完全的逻辑语言后, Semantic Web 就有足够的表达能力给出关于事物及事物间关系的断言。证明层的证明语言允许服务代理在向客户代理发送断言的同时将推理路径(从客户代理能够接受的假设开始推导出这个断言的推理路径)也发送给客户代理。这样客户代理只需要包含一个普通的验证引擎就可以确定断言的真假。但是,证明语言只能根据 Web 上已有的信息对断言给出逻辑证明,它并不能保证 Web 上所有的信息都为“真”,因此软件代理还需要使用数字签名和加密技术用来确保 Web 信息的可信性。

有了逻辑层和证明层提供的功能更强大的表达能力和推理能力,就可以用本体和规则来描述网络服务,客户代理就可以使用规则推理来发现所需的服务,然后根据服务的使用规则与服务代理自动集成,从而完成客户代理中的任务。通过交换本体,互相通信的代理之间可以对所使用的术语达成共识,代理甚至可以通过发现新的本体来增强自己的推理能力,代理的一个重要特点是可以交换证明,服务代理向客户代理提供信息服务的同时可以向客户提供关于这些信息的证明。利用数字签名技术,客户代理中的推理引擎就可以验证所提供的证明是否可靠,然后决定是否使用这些信息,从而实现自动化的、智能的网络服务^[19,21]。

5 Semantic Web 应用领域及开发工具

Semantic Web 是一个层次化的构想,目前研究比较成熟的最高一层是本体层。为了更好地促进各种相关研究工作的开展,必须开发实际应用和相关的支持工具。Semantic Web 的主要应用领域有^[19,21,25,30,39,40]:(1)网络服务。Semantic Web 中最重要的 Web 资源之一就是网络服务,使用 Semantic Web 技术,软件代理可以自动地发现、调用和集成网络服务,并对网络服务的执行

进行监控。(2)基于代理的分布式计算。Semantic Web 使用形式化的逻辑语言来表达 Web 知识的语义,从而赋予软件代理更多的智能和移动性,传统的客户/服务器计算范型将被基于代理的分布式计算范型取代。(3)基于语义的搜索引擎。目前典型的搜索引擎都是基于关键字的,它们的查准率和查全率往往受到多义词和同义词的影响。Semantic Web 使用本体来表达 Web 资源的语义,基于语义的搜索都是在一定的本体周境中执行的。本体能够消除多义词或同义词的影响,从而提高搜索结果的查准率和查全率。(4)基于语义的数字图书馆。随着 Web 上多媒体数据的日益增加,对它们的管理和检索也变得越来越重要。传统的多媒体检索技术使用颜色、纹理和形状等特征来描述图像或视频。基于语义的数字图书馆将使用本体来描述各种多媒体信息和图书信息,从而支持基于语义的检索和导航。Semantic Web 在这些领域中的应用将促进各种相关研究工作的开展,从而为知识管理、电子商务、自然语言处理、智能化信息集成、智能化教育等应用领域提供技术和实践经验。

为了促进 Semantic Web 的各种相关研究工作和实际应用的开展,目前已经出现很多具体的研究和开发工具。这些工具主要分成如下几类: RDF 编辑器、RDF 解析器、RDF 数据库接口工具、Topic Maps 创建及管理工具、本体构造工具、本体合并工具、本体注解工具、本体学习工具等^[21,35,30,41]。

6 Semantic Web 研究面临的问题和挑战

尽管 Semantic Web 在元数据描述和本体领域的研究已经基本成熟,要充分发挥 Web 的潜能,完全实现 Semantic Web 的构想,还面临许多问题和挑战。WWW 成功的重要因素是参与方便和工作结果立即可见,而 Semantic Web 中对信息的形式化逻辑规约是普通开发者和小型内容提供商难以参与的最大障碍。其中主要的问题有^[42,43]:(1)向 HTML 文档添加 RDF 信息或本体信息后会产生信息冗余,难以维护信息的一致性;(2)缺乏完整的易于理解的工具和实例,参与较难;(3)Semantic Web 的信息是面向机器的,需要通过“网络效应”才能看出它的优势,因此对开发者来说,工作结果不是立即可见的。

要完全解决上述的问题和充分利用 Semantic Web 所带来的机遇, Semantic Web 研究还面临如下几个方面的挑战^[21,38,39,43~46]:(1)内容的可获得性。目前 Web 上只有极少量的 Semantic Web 内容,要为 Semantic Web 奠定资源基础,现有的 HTML 内容、XML 内容、动态内容以及多媒体和网络服务等信息都必须转化成 Semantic Web 内容。(2)本体的开发、合并及演化。本体是 Semantic Web 的核心,在公共本体的创建、本体的变更管理、演化控制和注解等方面都需要大量的工作。(3)可视化技术。直观的、可视化的 Semantic Web 内容管理工具有助于用户识别和使用相关的内容。(4)Semantic Web 语言的稳定性。Semantic Web 层次结构中的所有语言都应标准化以确保相关支持技术的稳定性。(5)证明和信任模型的开发。由于 Semantic Web 并不能保证所有的信息都为“真”,所以必须开发相应的模型来验证断言的真实性以确保信息来源的可信任性。

7 总 结

WWW 的最终目标是提供服务。从目前 Web 上信息检索性能、网络服务发现与集成和各种电子商务模型中出现的问题可

以看出,实践发展与理想模型之间还是存在很大的差距。Semantic Web 是开发基于知识的、自动化的、智能的和易于使用的网络服务的新起点。是当前 WWW 研究的核心。Semantic Web 是个功能逐层增强的层次化结构,目前的研究主要集中在本体层。随着本体规约语言 OWL 的标准化,本体层的研究逐渐趋向成熟。在真正提供服务的代理层,基于 Semantic Web 框架的逻辑、证明和信任方面的研究相对较少,目前只有一些简单的演示系统。随着元数据描述和本体应用及相关工具和环境日趋成熟,本体之上的代理层将是 Semantic Web 下一步的研究重点。

参 考 文 献

- [1] Berners-Lee T. Information management: A proposal. W3C Publication. 1989. <http://www.w3.org/History/1989/proposal.html>.
- [2] Berners-Lee T. Web future directions. Technical Report, First International W3 Conference in Geneva, Switzerland, 1994.
- [3] Berners-Lee T. Realising the full potential of the Web. W3C publication. 1997. <http://www.w3.org/1998/02/Potential.html>.
- [4] Berners-Lee T. Semantic Web road map. W3C Design Issues. 1998. <http://www.w3.org/DesignIssues/Semantic>.
- [5] Berners-Lee T. Web architecture from 50,000 feet. W3C Design Issues. 1999. <http://www.w3.org/DesignIssues/Architecture.html>.
- [6] Berners-Lee T. Relational databases on the Semantic Web. W3C Design Issues. 1998.
- [7] Berners-Lee T. Why RDF model is different from the XML model. W3C Design Issues. 1998.
- [8] Berners-Lee T. What the Semantic Web can represent. W3C Design Issues. 1998. <http://www.w3.org/DesignIssues/RDFnot.html>.
- [9] Berners-Lee T. RDF and the Semantic Web. Technical Report, XML 2000 Conference in Washington, D. C., USA, 2000.
- [10] W3C. Semantic Web Activity Statement. W3C Publication. <http://www.w3.org/2001/sw/Activity>.
- [11] W3C. RDFCore Working Group. W3C Publication. <http://www.w3.org/2001/sw/RDFCore/>.
- [12] W3C. Web Ontology (WebOnt) Working Group. W3C Publication. <http://www.w3.org/2001/sw/WebOnt/>.
- [13] Cruz IF, Decker S, Euzenat J, McGuinness D. Forward. In: Cruz IF, Decker S, Euzenat J, McGuinness D, eds. Proceedings of SWWS 01. The First Semantic Web Working Symposium. 2001. 1 ~ 1.
- [14] Miller E. The Semantic Web. Technical Report, Eleventh International World Wide Web Conference (WWW2002), Honolulu, Hawaii, USA, 2002.
- [15] McBride B. RDF architecture. Technical Report, Eleventh International World Wide Web Conference (WWW2002) in Honolulu, Hawaii, USA, 2002.
- [16] Hendler J. OWL: A Web Ontology Language. Technical Report, Eleventh International World Wide Web Conference (WWW2002) in Honolulu, Hawaii, USA, 2002.
- [17] Swick RR, Brickley D. Semantic Web Advanced Development. Technical Report, Eleventh International World Wide Web Conference (WWW2002) in Honolulu, Hawaii, USA, 2002.
- [18] Horrocks I, Hendler J. Preface. In: Horrocks I, Hendler J, eds. The Semantic Web- ISWC 2002, Proceedings of First International Semantic Web Conference. Heidelberg; Springer-Verlag, 2002.
- [19] Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Scientific American, 2001, 284(5): 34 ~ 43.
- [20] Berners-Lee T. The Semantic Web - LCS Seminar. Technical Report, MIT: MIT LCS. 2002.
- [21] Hyvönen E. The Semantic Web - The new Internet of meanings. In: Hyvönen E, ed. Semantic Web Kick-Off in Finland: Vision, Technologies, Research, and Application. Helsinki: HIIT Publications, 2002. 3 ~ 26.
- [22] Lassila O, Swick RR. Resource Description Framework (RDF) model and syntax specification. W3C Recommendation. 1999.
- [23] Brickley D, Guha RV. Resource Description Framework Schema specification 1.0, W3C Candidate Recommendation. 2000.
- [24] Hayes P. RDF semantics. W3C Working Draft. 2003.
- [25] Hyvönen E, Harjula P, Viljanen K. Representing metadata about Web resources. In: Hyvönen E, ed. Semantic Web Kick-Off in Finland: Vision, Technologies, Research, and Application. Helsinki: HIIT Publications, 2002. 47 ~ 76.
- [26] Kivela A, Hyvönen E. Ontological theories for the Semantic Web. In: Hyvönen E, ed. Semantic Web Kick-Off in Finland: Vision, Technologies, Research, and Application. Helsinki: HIIT Publications, 2002. 111 ~ 136.
- [27] Neches R, Fikes RE, Finin T, Gruber TR, Senator T, Swartout WR. Enabling technology for knowledge sharing. AI Magazine, 1991, 12(3): 36 ~ 56.
- [28] Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition, 1993, 5(2): 199 ~ 220.
- [29] Borst WN. Construction of Engineering Ontologies. PhD. thesis, University of Twente, Enschede, 1997.
- [30] Corcho O, Fernández-López F, Pérez AG. OntoWeb D1. 1. Technical Roadmap IST- 2001 - 29243. 2001.
- [31] Guarino N, Garetta P. Ontologies and knowledge bases: Towards a terminological clarification. In N. J. I. Mars Towards very large knowledge bases: Knowledge building & knowledge sharing. IOS Press, 1995. 25 ~ 32.
- [32] Deborah L, McGuinness, Harmelen FV. OWL Web Ontology Language overview. W3C Working Draft. 2003.
- [33] Smith MK, Welty C, McGuinness D. OWL Web Ontology Language guide. W3C Working Draft. 2003.
- [34] Harmelen FV, Hendler J, Horrocks I, McGuinness D, Patel-Schneider PF, Stein LA. OWL Web Ontology Language reference. W3C Working Draft. 2003.
- [35] Patel-Schneider PF, Hayes P, Horrocks I. OWL Web Ontology Language semantics and abstract syntax. W3C Working Draft. 2003.
- [36] Heflin J. Web Ontology Language use cases and requirements. W3C Working Draft. 2003.
- [37] McIlraith S, Son T, Zeng H. Semantic Web services. IEEE Intelligent Systems, March/April 2001. 46 ~ 53.
- [38] Laamanen H, Helin H, Laukkanen M. Semantic Web and software agents meet wireless world. In: Hyvönen E, ed. Semantic Web Kick-Off in Finland: Vision, Technologies, Research, and Application. Helsinki: HIIT Publications, 2002. 207 ~ 242.
- [39] Lu SY, Dong M, Fotouhi F. The Semantic Web: Opportunities and challenges for next - generation web applications. Information Research, 7(4), 2002.
- [40] Narayanan S, McIlraith SA. Simulation, verification and automated composition of Web services. WWW 2002, Honolulu, Hawaii, USA. ACM 1-58113-449-5/02/0005. 2002.
- [41] Silvenen P, Hyvönen E. Semantic Web tools. In: Hyvönen E, ed. Semantic Web Kick-Off in Finland: Vision, Technologies, Research, and Application. Helsinki: HIIT Publications, 2002. 137 ~ 152.
- [42] Haustein S, Pleumann J. Is participation in the semantic web too difficult? In: Horrocks I, Hendler J, eds. The Semantic Web- ISWC 2002, Proceedings of First International Semantic Web Conference. LNCS 2342, Heidelberg; Springer-Verlag, 2002. 448 ~ 453.
- [43] Brickley D, Buswell S, Matthews BM, Miller L, Reynolds D, Wilson MD. SWAD- Europe: Semantic Web advanced development in Europe- A position

(下转第 119 页)

```

    < IPAddressRangeTo > any < /IPAddressRangeTo >
  < /SourceIPAddress >
  < SourceIPPort >
    < IPPortRangeFrom > any < /IPPortRangeFrom >
    < IPPortRangeTo > any < /IPPortRangeTo >
  < /SourceIPPort >
< /Source >
< Target >
  < TargetIPAddress >
    < IPAddressRangeFrom > 192. 168. 1. 3 < / IPAddressRangeFrom >
    < IPAddressRangeTo > 192. 168. 1. 3 < / IPAddressRangeTo >
  < /TargetIPAddress >
  < TargetIPPort >
    < IPPortRangeFrom > any < /IPPortRangeFrom >
    < IPPortRangeTo > any < /IPPortRangeTo >
  < /TargetIPPort >
< /Target >
< /RuleHead >
< RuleOption >
  < flags > < FIN/ > < SYN/ > < /flags >
< /RuleOption >
< /Rule >

```

然后, A-BOX 针对上述 Event GIDO, 依据检测规则进行分析。经分析, 得出结果是: 该 Event GIDO 与 SYN-FIN 扫描规则相匹配, 故 A-BOX 生成 Alert GIDO, 其内容包括 Alert GIDO 生成的时间, 对 Event GIDO 进行分析的 Analyzer 的一些相关信息, 攻击来源及目标, 经分析确认后的攻击类型, 以及攻击对系统的影响程度等。

```

< GIDO-Information >
  < Alert ident = " 1 " >
    < CreateTime >
      < date > 2003/4/3 < /date >
      < time > 10:01:25. 156832 < /time >
    < /CreateTime >
    < Analyzer analyzerid = " 1 " >
      < Node ident = " zd213 " >
        < ProcessID > 1356 < /ProcessID >
      < /Analyzer >
    < Source >
      < Address port = " 4867 " > 202. 120. 100. 35 < /Address >
    < /Source >
    < Target >
      < Node ident = " abc123 " >
        < Address port = " any " > 192. 168. 1. 3 < /Address >
      < /Node >
    < /Target >
    < Classification >
      < name > SYN-FIN scan < /name >
    < /Classification >
    < Assessment >
      < Impactseverity = " low " completion = " succeeded " impacttype = " reconnais-
sance " >
    < /Assessment >
  < /Alert >
< /GIDO-Information >

```

最后, R-BOX 依据 Alert GIDO 进行判断, 作出相应的响应方式并生成 Response GIDO, 其内容包括 Response GIDO 生成的时

间, 对 Alert GIDO 进行分析的 Responder 的一些相关信息, 针对攻击类型为 SYN-FIN scan 的报警作出的响应是通知防火墙阻隔 SYN-FIN 数据包。

```

< GIDO-Information >
  < Response ident = " 1 " >
    < CreateTime >
      < date > 2003/4/3 < /date >
      < time > 10:01:25. 235638 < /time >
    < /CreateTime >
    < Responder ident = " 1 " >
      < Node ident = " zd213 " >
        < ProcessID > 2563 < /ProcessID >
      < /Responder >
    < Classification >
      < name > SYN-FIN scan < /name >
    < /Classification >
    < Action type = " block data " >
      < signature >
        < Flags > < FIN/ > < SYN/ > < /Flags >
      < /Signature >
    < /Action >
    < Receiver > Firewall < /Receiver >
  < /Response >
< /GIDO-Information >

```

5 总 结

本文提出了采用具有简单、灵活、可扩展、易理解、跨平台等特性的 XML 语言来描述通用入侵检测对象 GIDO 和入侵检测规则, 从而弥补了采用 CISC 语言来描述 GIDO 而造成的难以理解和不易推广的缺点; 同时, 在修改和完善了 Snort 系统的检测规则的定义方法的基础上, 使用 XML 来描述检测规则也使得规则的描述更为灵活与完整。

参 考 文 献

- [1] Curry D, Debar H. Intrusion Detection Message Exchange Format, Internet Engineering Task Force, 2003 - 01 - 30.
- [2] <http://www.w3c.org/>.
- [3] 罗妍.《通用入侵检测框架研究》,《华南理工大学学报》,2001 年.
- [4] 傅华, 温巧燕, 杨义先. "XML 语言在入侵检测系统中的应用",《网络信息安全》,2002 年第 9 期.
- [5] 唐屹. "CIDF 框架上的入侵检测规则扩充",《计算机应用》,2002 年第 11 期.
- [6] 唐正军. 黑客入侵防护系统源代码分析, 机械工业出版社, 2002. 3.

(上接第 6 页)

- paper. In: Horrocks I, Hendler J. eds. The Semantic Web- ISWC 2002, Proceedings of First International Semantic Web Conference. LNCS 2342, Heidelberg: Springer-Verlag, 2002. 409 ~ 413.
- [44] Benjamins VR, Contreras J, Corcho O, Gómez-Pérez A. Six challenges for the Semantic Web. ISOCO White Paper. 2002.
- [45] Berners-Lee T. The Semantic Web and challenges. Technical Report, W3C Talks. 2003.
- [46] McBride B. Four steps towards the widespread adoption of Semantic Web. In: Horrocks I, Hendler J. eds. The Semantic Web- ISWC 2002, Proceedings of First International Semantic Web Conference. LNCS 2342, Heidelberg: Springer-Verlag, 2002. 419 ~ 422.