

●刘柏嵩

基于知识的语义网:概念、技术及挑战

摘要 目前万维网由松散结构数据组成。为改变这种状况,提出了语义网的思想。大量的数据需要结构化,更为重要的是赋予含义。这种语义网将提供对异构和分布式信息的智能访问,使得软件产品可作为用户需求和可获取信息源的中介。图1。参考文献13。

关键词 语义网 知识表示 本体论 基于知识的系统

分类号 G350

ABSTRACT At present, the WWW is composed of loosely-structured data. To change the situation, the concept of semantic Web is proposed. Large amount of data need to be structured and given meanings. Such a semantic Web will provide intelligent access to heterogeneous and distributed information, and enable software products to become intermediaries of user's needs and available information. 1 fig. 13 refs.

KEY WORDS Semantic web. Knowledge representation. Ontology. Knowledge-based system.

CLASS NUMBER G350

1 语义网的概念

我们正处在新一代万维网的边缘, W3C 主席 Tim Berners-Lee 的《编织万维网》一书将其称作语义网(Semantic Web)^[1]。不同于现存的万维网,其数据主要供人类使用,新一代 WWW 中将提供也能为计算机所处理的数据。这将使得大量的智能服务成为可能,如信息中介、搜索代理、信息过滤等,这一过程可以说是“充分发挥 Web 的潜力”^[2]。该领域研究的重要性,从美国的 DAML 启动计划可略见一斑,美国国防部高级研究计划局(简称 DARPA)为旨在开发语义网的项目一共提供 7000 万美元的资助。

语义网的基本思想是提供基于机器可处理的数据语义,并应用这些元数据的启发式进行自动化的信息访问^[3]。数据语义的显性表示和领域理论(本体)将使得 Web 提供一种全新质量的服务。其最终目标是将人类知识编织成一个巨大的网络,并以机器处理的方式来实现它。各种自动化服务将帮助用户以机器可理解格式访问和提供信息,并使得计算机自动化处理过程和 Web 信息集成更为方便。

将 Web 变成一个“知识万维网”,一下子就把知识获取和知识表示推到中心位置并成为热门课题。

2 知识表示和处理

2.1 知识表示

知识表示是任何基于知识系统中的一个重要元素。人工智能领域已设计了大量的知识表示模式,

以便知识能应用于推理过程和进行问题求解。这些技术可大致分为两类:(1)陈述式表示,知识表示为静态事实集和方法,包括语义网络、框架和剧本等。它所强调的是事物所涉及的对象是什么,是对事物有关知识的表态描述,是知识的一种明确表达形式。(2)过程化表示,它将有关某一问题领域的知识,连同如何使用这些知识的方法,均隐性地表达为一个求解问题的过程,其描述形式就是程序。

在 Web 这个分布式、动态的环境下进行知识表示时,采用单一的知识表示方法往往不够。目前,在语义网中采用最广泛的是基于框架、谓词逻辑及描述逻辑的方法,而且有把三者融合起来使用的趋势。

发展语义 Web 的两个关键技术已经形成:可扩展标记语言(XML)和资源描述框架(RDF)。XML 允许每个用户创建自己的标记并任意增加结构,且无须说明其结构的含义。含义是由 RDF 和 RDFS(RDF Schema)来描述的,它以成组的三元组编码,这些三元组可序列化 XML 形式。但是由于缺乏对规则、变量和推理的支持,作为知识表示语言 XML 和 RDF 都非常受限制。而由 DARPA 提出的 DAML + OIL 的目标就是将当前的 Web 进化为语义 Web,它建立在 RDF 和 XML 的基础上,并提供描述资源的规则和语义^[4]。

2.2 本体

知识表示能将已知世界进行程序化建模,而应用于 Web 的本体(Ontology)可解决信息异质性,并实现信息的结构化。本体定义了用于描述和表示领域

知识的术语,它用于人、数据库和应用之间共享信息。它通常表达为一组对象(概念)、关系、函数、定理和示例。目前最常用的本体定义是:“本体是一个共享概念化的形式化、明确的说明。”“概念化”(conceptualization)是指某一概念系统所蕴涵的语义结构,它是对某一事实结构的一组非正式的约束规则^[5]。它可以理解、表达为一组概念(如实体、属性、过程)及其定义和相互关系^[6]。“明确”意味着所采用概念的类型和它们应用的约束进行明确的定义:“形式化”是指本体中的事实应是机器可读的;“共享”反映本体应捕捉该社区中一致公认的知识。

本体通常以基于逻辑的语言来表示,因此可在类、属性和关系之间做出详细、准确、一致且完备的区别。一些本体工具能采用本体进行自动推理并因此为智能应用提供高级服务,如概念或语义检索,软件代理,决策支持,自然语言理解,知识管理和电子商务等。对 Web 而言,最典型的本体类型是分类(如 Yahoo! 的分类目录)和一组推理规则。本体可从多个方面提高 Web 功能,它们既可以提高 Web 搜索的精度,也可方便实现查询应答(Question Answering, QA)。

3 面向语义网的本体语言

3.1 基于 XML 的本体交换语言

XOL (Xml-based Ontology exchange language)是一种基于 XML 语法和 OKBC 语义的本体交换语言。它由美国生物信息学术团体设计用于其领域的一组异构软件系统之间进行本体定义的交换,它是以 Ontolingua 和 OML (本体标记语言)作为基础的,融合了 OKBC 的高层表达方式和 OML 的语法。目前没有支持 XOL 本体开发的工具,但由于它采用 XML 语法,可以采用 XML 编辑器来创建 XOL 文件。

3.2 RDF 和 RDF(S)

由 W3C 为描述 Web 资源而开发的 RDF 允许在 XML 的基础上以一种标准化、互操作的方式对数据语义进行定义^[7]。它的数据模型与语义网络形式等价,其目的主要是为元数据在网络上的各种应用提供一个基础结构,使应用程序之间能够在网络上交换元数据,以实现网络资源的自动化处理。在 RDF 的基础上,RDF(S)用来建立类模型、属性层次以及其他可从 RDF 模型中引申出的基本模型限制。也就是说 RDF(S)定义了一套简单的本体,可以控制 RDF 文档的一致性。

RDF(S)仅为本体对象提供一些常见的模型限制。不论选择哪种正式的语言,还必须在语言的易处理和表达性之间采取折中原则。因为 RDF(S)要

应用于大量的网络资源中,它已经十分接近最基本的表示。与常用的知识表示语言相比,它并不是所有知识表示问题的最终解决方案,而仅是一种可扩展的核心语言。RDF(S)的命名空间和具体化机制使得不同领域的用户,通过扩展核心的定义和语义,以 RDF(S)的形式定义自己的标准,包括不同命名空间之间的映射、版本升级或者本体公理的定义等。

3.3 简单 HTML 的本体扩展(SHOE)

由马里兰大学开发,并用于开发 OML,它作为 HTML 的一个扩展,将机器可读语义知识结合进 HTML 文档或其他 Web 文档中,它允许直接在万维网的基础上设计和应用本体。近来,SHOE 的语法已转向 XML。SHOE 使得 Agents 能够收集有意义的 Web 页面和文档的信息,改善搜索机制和知识收集。该过程包括 3 个阶段:定义本体;用本体信息注解 HTML 页面以描述自身和其它页面;通过搜索所有现存页面和保持信息更新,有一个 Agent 语义级检索信息。

3.4 本体标记语言 OML

由 Washington 大学开发的 OML 部分基于 SHOE。OML 有 4 种层次:OML 核心层与语言的逻辑层相关;简单 OML 直接映射 RDF(S);简化 OML 包括概念图特征;标准 OML 是 OML 最详尽的版本。除了 XML 编辑器外,没有其它 OML 创作工具。

3.5 本体交换语言 OIL

由 Onto Knowledge 项目开发的 OIL,允许 Web 资源之间的语义交互性。它的语法和语义是基于已有的语言(OKBC、XOL 和 RDF(S)),提供本体工程中基于框架方法中广泛采用的建模原语,以及描述逻辑(一阶逻辑的子集)中的规范语义和推理支持^[7]。

OIL 是建立在 RDF 的基础上,具有如下层:核心 OIL 包括直接映射到 RDF(S)原语的 OIL 原语;标准 OIL 是完整的 OIL 模型,采用 RDF(S)以外的原语;范例 OIL 将概念和任务的范例增加到已有模型;重 OIL 是作为 OIL 的将来的扩展。

OIL ED, Protege2000 和 WebODE 可用于编辑 OIL 本体。OIL 的语法不但是以 XML 表示的,而且以 ASCII 码形式提供。

3.6 DAML + OIL

DAML + OIL 是由 US 和欧盟在 DAML 背景下共同开发的,它与 OIL 有着相同的目标,是目前应用最广泛的本体语言^[8]。DAML + OIL 是建立在 RDF(S)基础之上的,意在取代初始的定义 DAMLONT^[9]。OIL Ed、OntoEdit、Protege2000 和 WebODE 都可作为设计 DAML + OIL 本体的工具。

3.7 面向语义网的语言层次

W3C 提出的语义网的语言层次如图 1 所示。

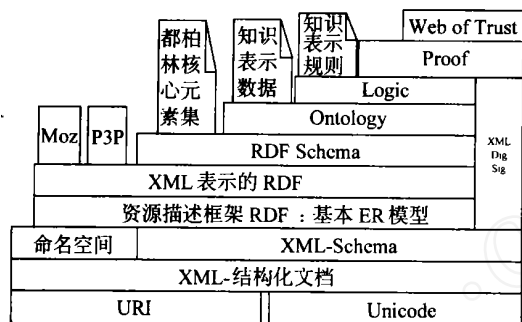


图 1 语义网的语言塔形结构图

4 语义网面临的挑战

4.1 内容的可获取性

语义网的内容是根据某些特定的本体进行 Web 页面标注的,这些本体定义了页面中词或概念的含义。由于语义网的基础架构正在建设中,目前语义网的内容很少^[10]。除了基础架构外,研究人员正在开发支持 Web 内容语义标注的工具,这些工具对于语义网的成功非常关键。但是它们存在两点不足:绝大多数只标注静态页面;还有很多只强调创建新的内容。由此导致了如下不利局面:

忽略了动态产生的内容,动态内容是指由数据库产生的内容,据研究,其容量是静态内容的 400 ~ 550 倍^[11]。虽然 XML 内容是作为内容共享的基础,仍有部分内容有可能被语义网所排除。即使人们期望最终整个 Web 将语义化,但如果目前高质量的 Web 内容不能在语义网中访问也非常可惜。

该问题的解决方案如下:根据静态和动态 Web 文档来创建一组标注服务(中间件),这些 Web 文档包括多媒体和 Web 服务等。标注服务应根据语义网语言层次来生成,以便理解语义网的不同语言的各种 Agents 总能理解和处理不同语言所表示的内容。标注资源的方法虽然有创建性,但目前大多数提出对 Web 资源进行标注的方案只是采用这些语言中的某一种。同时,这种分层的思想也被语言设计所采纳。

静态内容的自动标注基本上是采用已有的封装技术,而动态内容的标注则更为复杂,主要方法如下:

(1) 从信息源抽取动态内容,对其标注并存储。该方法的缺点是从动态站点(数据库驱动)可能产生的静态页面数量非常庞大,而且不断更新,需要在数据库变化时创建或移除静态页面。

(2) 将内容保持在数据库中,只标注检索相关内容的查询式。这种方法节省空间,保证了标注与底层信息源的一致性,因其内容是在执行检索时动态产生和标注的。

还需考虑 Web 服务的标注,由此语义 Web 的服务可采用已有的服务。这项工作需要在现有的标准如 SOAP、WSDL、UDDI 和 WSDL 的基础上进行。对于更为复杂的服务,DAML-S 和 UPML 是很好的选择。但该语言的定义还需进一步研究。DAML-S 是一个基于 DAML 的 Web 服务本体论,它以无二义性和机器可理解方式,为 Web 服务供应商提供了描述 Web 服务的属性和能力的标记语言的一个核心集。UPML 是实现标记任务的高层服务语言框架。在 I-BROW 项目中采用。

至于多媒体内容的标注,目前至少有两种形式:图像和视频处理及采用 NLP 技术的信息抽取,包括图像的文本化解释、自动语音识别等,其标注的思想与上述的静态和动态方法类似。

4.2 本体的开发和演化

本体是实现语义网的关键,它们携带语义网中内容的含义,即提供标记的词汇表和语义。为应对该挑战需解决 3 个问题。

第一是能应用于所有领域核心本体的构建。已应用于不同领域的核心本体包括:IEEE 的标准顶层本体 SUO,电子商务领域的 UNSPSC、ROSETTANET 等。

第二个问题是为本体开发过程的大部分活动提供方法论或技术上的支持,包括:(1) 知识获取、概念建模和语义 Web 语言的本体编码;(2) 本体联合和映射、本体集成和本体翻译工具以及本体再造工程工具等;(3) 可重用本体的一致性检查工具等。

第三个问题是本体的演化(Evolution)及其与已标注数据的关系,配置管理工具必须控制每一本体论的版本以及本体论和标注之间的相对独立性。

4.3 内容的可扩展性

有了语义网的内容后,接下来需要考虑的是如何以可扩展方式来管理它,包括如何组织、存储和查找等,主要问题有:

第一个问题是与语义网页面的存储和组织相关。“基本”语义网由基于本体的标注页面组成,这些页面的链接结构反映了万维网的结构,即页面之间通过超链接的方式互联。这种超链接方式并不能完全利用语义 Web 页面的底层语义,必须采用语义索引基于特定主题对语义 Web 内容进行分类。这是让应用程序能够聚合内容以提供增值服务必须的一步。语义索引将采用本体信息和标注文档动态生成。

第二个问题是在语义网中方便地查找信息有关,主要是语义索引的协调。考虑到 Web 资源的语义,必须有方便查找语义网内容的机制。在此,可以采用 P2P 架构,它与目前 WWW 的路由器配置类似,可称之为“语义 TCP/IP 协议”,有一项欧洲语义 Web 项目正是关于该主题的。索引可看作是活动的 A-

gents,它了解自己的能力和。在某一索引中没有出现的主题可以语义级地路由到相邻的索引。在此需采用 Agents 协商技术以获取主题的语义路由,同时也应用了 WordNet 中的“语义距离”的概念。因此,就不需要进行语义内容的中心注册,这样使得整个架构的可扩展性更强,语义网的内容可不断添加。同时,这种组织方式与已有的 WWW 思想保持一致。

4.4 多语种支持

WWW 上内容的语言分布的研究表明,虽然英语占绝对的主导地位,但其他语言也不容忽视,其比重为:英语 68.4%,日文 5.9%,德文 5.8%,中文 3.9%,法文 3.0%,西班牙文 2.4%,俄文 1.9%,意大利文 1.6%,波兰文 1.4%,韩文 1.3%,其他 4.6%(摘自:www.vilaweb.com)。对欧洲的 WWW 资源而言,语言的多样性更为明显。多语种性在以下方面起着越来越重要的作用:本体级、标注级和用户界面级。

(1)本体级:本体创建者一般采用其母语来开发本体,标注将在此基础上进行。既然不是所有的人都是本体创建者,该级别的优先级最低。现有的多语种和语言资源,如 WordNet、EuroWordNet 和 HowNet 可用于支持该级别的多语种性。

(2)标注级:可以采用多种语言对内容进行标注,由于更多的用户(尤其是内容提供者)是标注内容而不是开发本体论,为内容提供者采用母语标注内容提供支持很有必要。为了使语义网内容尽可能简便,不能要求一个意大利人以西班牙文来标注内容,反之亦然。

(3)用户界面级:不管源标注采用何种语言,人们喜欢以母语的方式访问相关内容。虽然目前绝大多数内容是英语的形式,但其他语种的内容也会越来越多。任何语义网方法应方便地访问不同语言的信息,采用国际化和本体化技术实现基于用户母语的个性化信息访问。

4.5 本体语言的标准化

语义网是一个新兴的研究领域,W3C 将就应用于该领域的语言和技术作出推荐。经过一年左右的工作和讨论,目前该工作组(www.w3.org/2001/sw/we-bont)已提出了 Web 本体语言的需求分析^[12,13],并于 2002 年 8 月初推出标准本体语言草案(OWL)1.0 版。

为构造语义网,相关的工具支持也必不可少,而这些工具也或多或少与语义网语言有关联。目前得到一致公认的本体语言是 DAML + OIL (www.daml.org),相关的项目有 Onto Knowledge (www.ontoknowledge.org)、IBROW 和 Esperanto (www.esperanto.net)等。

5 结论

语义网无疑是迷人的。按照 Berners-Lee 的预

想,它不仅能理解词语和概念,而且还能理解其中的逻辑关系,具有不可估量的潜力。目前在数字处理方面,计算机已经远胜于人类。能够同样熟练地处理语言和推理的计算机将不仅能够帮助人们获得新的深邃见解,而且还能自己开辟新的道路。

在未来语义网中,对网上知识库进行知识开采的工作也不再会要求人们费力地一屏一屏地浏览无关的数据。相反,计算机将派遣智能代理——也就是软件 Agents 去探索大量网站,按照逻辑关系筛选出真正切题的内容,大量网页上的相关思想便可以得到提炼和总结。这将提高评估信息并把它们整合至新高度的能力。以实现信息资源和知识资源的智能共享为目标,基于知识的语义网是因特网的下一代,也是信息服务的未来平台,可以预见,它的技术经济和社会应用前景十分广泛。

因特网技术改变了人类通信的方式,而万维网技术改变了人类工作的方式。人们正在从 EDI 过渡到 XML,从 CORBA 发展到 SOAP,所有这些技术变化对相关领域提供的服务已产生深刻的影响。同样,语义网将有助于从 HTML 到 XML 再发展到 OWL (Ontology Web Languages)。但是语义网在解决其面临的挑战前,距真正的大规模应用尚有一段时间。

参考文献

- 1 T. Berners-Lee, Weaving the Web, Harper, San Francisco, CA, 1999.
- 2 Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. Scientific American, 284(5):34-43, May 2001
- 3 T. Berners-Lee, and James Hendler, Publishing on the semantic web, Nature 410, 1023-1024 (26 April 2001)
- 4 F. van Harmelen, P. F. Patel-Schneider and I. Horrocks (eds), DAML + OIL, March, 2001.
- 5 V. Richard Benjamins et al., Six Challenges for the Semantic Web. KR 2002 (19 April 2002)
- 6 Steffen Staab. Ontologies' KISSES in standardization. IEEE Intelligent Systems, Mar/Apr 2002: 70-79.
- 7,12 Jeff Heflin et al., Requirements for a web ontology language. W3C working draft 08 July 2002.
- 8 <http://www.daml.org>
- 9 Ian Horrocks et al., Reviewing the design of DAML + OIL: an ontology language for the semantic web. May 2002
- 10 Peter F. Patel-Schneider and Dieter Fensel. Layering the Semantic web problems and directions. Feb 2002
- 11 Lawrence & Giles. Accessibility of Information on the Web. Nature 400: 107-109, July 8, 1999.
- 13 Dieter Fensel. Language standardization for the semantic web: the long way from OIL to OWL. Jan 2002

刘柏嵩 博士生。通讯地址:浙江宁波大学 41[#]信箱。邮编 315211。(来稿时间:2002-08-09)