

语义 Web 与 Ontology 研究

东野广升 冯丽雅

【摘要】 本文介绍了语义 Web 和 Ontology 的概念特征,探讨了语义 Web 与 Ontology 的关系,结合信息检索研究了 Ontology 在语义 Web 中的应用,并对实现中需解决的问题进行了说明。

【关键词】 语义 Web Ontology XML RDF

Abstract: This article first introduces the concept and construction of Semantic Web and Ontology. It discusses the relationships between Semantic Web and Ontology. After that it researches the application of Ontology in Semantic Web by information searching. It also gives some suggestions of the realization.

Key words: Semantic Web Ontology XML RDF

1 语义 Web 的概念和体系结构

万维网标准化组织 W3C 对语义 Web (Semantic Web) 的定义为:语义 Web 是建立在 RDF 与其它定义的标准基础之上,对 Web 上的数据所进行的一种抽象表示。而根据 W3C 主席 Tim Berners—Lee 的定义,语义 Web 是一个网,它包含了文档或文档的一部分,描述了事物间的明显关系,且包含语义信息,以利于我们的机器自动处理。语义 Web 要求互操作性标准不仅表达文档的构造形式,还要表达其语义内容。Berners—Lee 为未来的 Web 发展提出了基于语义的体系结构,如图 1 所示。

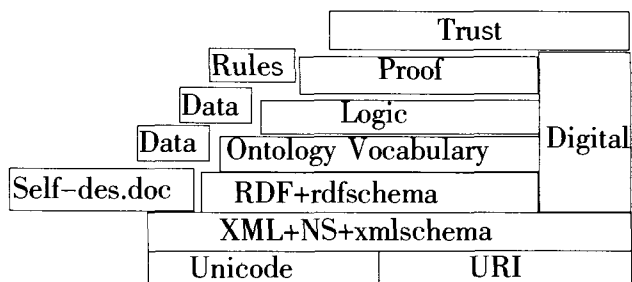


图 1 语义 Web 的体系结构

(1) Unicode 和 URI,是整个语义 Web 的基础,Unicode 处理资源的编码,URI 负责标志资源;

(2) XML + NS + xmlschema,用于表示数据的内容和结构;

(3) RDF + rdfschema,用于描述 Web 上的资源及其类型;

(4) Ontology Vocabulary,用于描述各种资源之间的联系;

(5) ~ (7) 是在下面四层的基础上进行的逻辑

推理操作。

核心层为 XML、RDF、Ontology,这三层用于表示 Web 信息的语义。

从上面的结构和定义可以知道语义 Web 信息组织和描述上一个重要的区别是它注重对信息语义的刻画和在此基础上的联系,而其中 Ontology 是组织、抽象的基本方式。

2 Ontology 的概念和应用特点

Ontology 原本是哲学上的一个概念,称为 Ontology 论、实体论或存在论,是对自然存在及其本质的研究,属于形而上学理论的分支。Ontology 所反映的是事物本质的、科学的内涵。

人工智能等学科将 Ontology 的概念从哲学领域中借用过来,并赋予了一些新的含义,近年来广泛用于知识表示、知识共享、知识集成、知识重用和知识管理等领域中。许多学科的研究都在使用 Ontology 这个词,但却存在不完全相同的定义和理解。

目前在计算机界被广泛接受的是 Gruber 在 1993 年对 Ontology 的定义。Gruber 所定义的:Ontology 是一种共享的概念化的形式化明确说明。这一定义包含了四个方面的含义:概念化、明确、形式化以及共享。概念化是指对世界中一些现象通过标识其相关概念而得到的抽象模型;明确是指所使用的概念的类型以及对这些概念使用上的约束都有了明确的定义;形式化是指 Ontology 是机器可读的(即能被计算机处理),而不是完全用自然语言表达;共享则是指 Ontology 反映的是共同认识的知识,是相关领域中公认的概念集,不为某个人所独有,而为大家所接受。

构建一个 ontology,可以解决以下 5 个问题。

(1) 能够实现信息在人与人之间、软件代理与软件代理之间以及人和软件代理之间的共同理解。假设有若干包含医药信息或提供医药电子商务服务的 Web 站点。如果这些 Web 站点共享相同的底层 Ontology, 那么计算机代理就可以抽提和集成这些来自不同站点的信息。代理软件可以利用这些集成的信息来回答用户的检索式或向用户提供数据。

(2) 实现领域知识的重用。例如许多领域需要应用到对时间的知识表示, 这种时间表示包括时间段、时间点、相对时间等的表示。如果某个研究组织开发出这样一个详细的 Ontology, 其他研究组织就可以轻而易举地将它复用到各自的专业领域。而且, 如果要构建一个大型的 Ontology, 也可以复用一个通用 Ontology 框架, 如 UNSPSContology, 将这个框架 (skeleton) 进行扩展可填充, 来描述人们感兴趣的领域。

(3) 供显性的领域假设。使专业内的假设变得更加明确。显性的领域假设意味着如果人工智能系统的领域改变了, 相应的假设修改也十分容易。对专业领域知识进行明确的规范说明, 对于那些必须理解该领域术语含义的新用户来说很有帮助。与写入到程序中的假设相比, 基于 Ontology 的领域假设更容易被发现、理解和修改。

(4) 从操作知识中分离出领域知识。如将专业领域的知识从运筹学、知识管理的环境中剥离出来。开发 Ontology 的目标在于定义一系列的数据及结构以供其它程序应用。一个 Ontology 可以被问题解答系统应用, 也可以被信息抽取系统应用, 从而实现操作知识和领域知识的分离。我们可以按照一种必须的规范说明和执行程序来完成一项产品的任务。

(5) 对领域知识进行分析。在进行复用现有 Ontology 和扩展这些 Ontology 的尝试中, 对术语进行规范的分析是极有价值的, 一旦对术语及相互关系有了明确的定义, 我们就有可能对领域知识进行分析和推理。为了重用和扩充一个 Ontology, 对它进行形式化的分析是必要的。

3 语义 Web 与 Ontology

为了实现语义 Web 的功能, 需要提供一种计算机能够理解的、结构化的语义描述机制, 以及一系列的推理规则, 以实现自动化推理。在 Tim Berners - Lee 的 Semantic Web 框架中, 几个关键的组成元素就是 XML、RDF 和 Ontology。

XML 允许用户定义自己的文件类型, 允许用户定义任意复杂的信息结构, 但是 XML 只具有语法性, 它不能说明所定义的结构语义。在 Tim Berners - Lee 看来, 语义的描述是通过 RDF 进行的。

RDF 能够表示陈述句, 并且主语、谓语和宾语的

三个组成元素都是通过 URI 所标识的, 所以它具有语义表述的特性。但语义 Web 的要求还远不止于此, 语义 Web 还需要加入逻辑功能: 语义 Web 需要能够利用规则进行推理、选择行动路线和回答相关问题。如果某个医院和某个大学的 web 页面上都有 <Doctor>, 那么 Doctor 代表的是医生还是博士? 用 XML 和 RDF 并不能解决这个问题, 因为 XML 和 RDF 在处理语义上存在两个问题: (1) 同一概念有多种词汇表示; (2) 同一个词有多种含义 (概念)。

为了解决上述问题, 很自然地需要引入 Ontology。Ontology 通过对概念的严格定义和概念与概念之间的关系来确定概念精确含义, 表示共同认可的、可共享的知识。对于 Ontology 来说, Author、Creator 和 Writer 是同一个概念, 而 Doctor 在大学和医院分别表示的是两个概念。因此在语义 Web 中, Ontology 具有非常重要的地位, 是解决语义层次上 Web 信息共享和交换的基础。

语义 Web 研究者也认为, Ontology 是一个形式化定义词语关系的规范化文件。对于语义 Web 而言, 最典型的 Ontology 具有一个分类体系和一系列的推理原则。其中, 分类体系定义对象的类别和类目之间的关系。实体之间的类/子类关系对于 Web 应用具有重要的价值。例如可以在一个地理 Ontology 中加入这样一条规则, “如果一个城市代码与一个省代码相关, 并且一个地址利用了城市代码, 那么这个地址就与相应的省代码相关”。通过这一规则, 程序可以推理出中国科学院文献情报中心在中关村, 应当在北京市。

作为知识表示工具, Ontology 与语义网络非常相似。它们都是表示知识的形式, 并且均可以通过带标记的有向图来表示, 适合用于逻辑推理。但从描述的对象或范围而言, Ontology 与语义有所区别。

Ontology 是对共享概念模型的规范说明, 这里所说的“共享概念模型”指该模型中的概念是公认的, 至少在某个特定的领域是公认的。一般情况下, Ontology 是面向特定领域, 用于描述特定领域的概念模型。语义网络从数学上说, 是一种带有标记的有向图。它最初用于表示命题信息, 现广泛应用于专家系统表示知识。语义网络中节点表示物理实体、概念或状态, 连接节点的边用于表示关系。语义网络中对节点和边没有其他特殊的规定, 因此语义网络描述的对象或范围比 Ontology 广。例如, 语义网络可以表示一句话, 如“我的汽车是红色的”。但是 Ontology 显然不适合于这类的表示, 它侧重于表现整体的内容, 如团体组织 (学校) 的内部构成等。

在表示的深度上, 语义网络不如 Ontology。语义网络对建模没有特殊的要求, 但是 Ontology 却有 5 个要素: 元语、类、关系、函数、公理和实例, 其中公

理可以看作是 Ontology 中的约束。Ontology 通过这 5 个要素来严格、正确地刻画所描述的对象。语义网络的建立可以不要求有相关领域的专业知识,因此比较容易建立。而 Ontology 的建立必须要有专家的参与,相对而言更加严格和困难。需要专家的参与是目前 Ontology 主要缺点之一,如何通过知识挖掘手段自动获取 Ontology 是目前也是今后研究的重点。

4 Ontology 在语义 Web 中的应用

常规的直接基于关键词的信息检索技术已不能满足用户在语义上和知识上的需求,寻找新的方法也就成为目前研究的热点。Ontology 具有的良好概念层次结构和对逻辑推理的支持,因而在信息检索特别是在基于知识的检索中得到了广泛的应用。基于 Ontology 的信息检索的基本设计思想可以总结如下:

在领域专家的帮助下,建立相关领域的 Ontology。

收集信息源中的数据,并参照已建立的 Ontology,把收集来的数据按规定的格式存储在元数据库(关系数据库、知识库等)中。

对用户检索界面获取的查询请求,查询转换器按照 Ontology 把查询请求转换成规定的格式,在 Ontology 的帮助下从元数据库中匹配出符合条件的数据集合。

检索的结果经过定制处理后,返回给用户。

需要说明的是,如果检索系统不需要太强的推理能力,Ontology 可用概念图的形式表示并存储,数据可以保存在一般的关系数据库中,采用图的匹配技术来完成信息检索。如果要求比较强的推理能力,一般需要用一种描述语言(如: Loom, Ontolingua 等)表示 Ontology,数据保存在知识库中,采用描述语言的逻辑推理能力来完成信息检索。由于 Ontology 能通过概念之间的关系来表达概念语义的能力,所以能够提高检索的查全率和查准率。

例如我们要在网上查询“知识工程”领域专家,而已建立起一个关于学科的 Ontology,如图 2 所示。

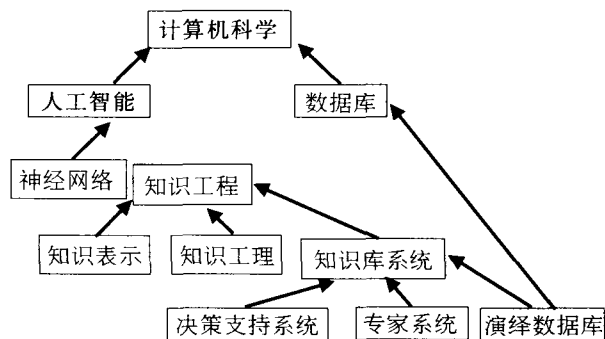


图 2 已建立的一个关于学科的 Ontology

那么上面的简单查询就涉及到概念之间复杂的逻辑、语义和语法关系:

专家是指在某领域具备深厚专业知识的人,如参加或完成相关领域研究和应用项目、发表过相关著作或论文、教授相关课程等人员都可能成为该领域的专家。假设在 Web 上目前有如下人员的资料:

```

<person>
  <name> </name>
  <xlink ref=http://exp.edu.cn/computers/
Mary.html/>
  <books>
    <book>
      <title> Design and Realization of Expert
System </title>
      <topic>Software Design </topic>
      <topic>Expert System, /topic>
      <publisher> Beijing Publish </publisher>
      <publish time>2001 </publish time>
    </book>
  </books>
</person>
  
```

一般认为,是某领域的子领域的专家同样也是这个领域的专家,所以应该是符合条件的结果输出。

可以看出这种基于 Ontology 和语义的信息检索更接近于实际情况,更能获取到我们所需要的信息。

实现过程中需解决的问题 Ontology 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。但目前 Ontology 的开发还没有统一的标准,不同的应用和工程所遵循的创建过程和方法不同。例如现在出现的方法有: Mike Uscholddede & King 的骨架法; TOVE 评价法; KACTUS 工程法等,在具体的应用中,应根据应用的特点选择和借鉴具体的创建和开发方法,为此需解决 3 个问题。

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

第二个问题是为 Ontology 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 Ontology 编码; (2) Ontology 联合和映射、Ontology 集成和 Ontology 翻译工具以及 Ontology 再造工程工具等; (3) 可重用 Ontology 的一致性检查工具等。

第三个问题是 Ontology 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 Ontology 论的版本以及 Ontology 论和标注之间的相对独立性。

语义 Web 是一个新兴的研究方向, Ontology 在其中的应用也仅仅是刚刚开始,还有许多的问题需要研

究和解决。一方面,评价 Ontology 的标准, Ontology 方法学的问题,如何集成 Ontology,研究使远程设计具有一致性的设计工具的研究等;另一方面,不断扩大应用领域,如何更好地引用先验知识,促进知识获取的智能化发展,也是值得注意的问题。因此,支持 Ontology 设计与评价方法和工具的开发,尤其是对系统、全面、完整的方法体系的研究仍是未来的一个研究方向。

参考文献

- 1 邓芳. Ontology 在语义 Web 中的应用研究. 计算机应用研究, 2004 (6)
- 2 刘柏嵩. 基于知识的语义网: 概念、技术及挑战. 中国图书馆学报, 2003 (2)
- 3 T. R. Gruber. A translation approach to portable ontologies.

Knowledge Acquisition, 5 (2): 199 - 220, 1993. ftp://ftp.Ksl.Stanford.Edu/pub/KSL_Reports/KSL-92-71.ps.gz

- 4 Dieter Fensel. Ontologies: Silver bullet for Knowledge Management and Electronic Commerce, Springer-Verlag, Berlin, 2001.
- 5 Natalya F. Noy, Deborah L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. <http://www.Ksl.Stanford.Edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html>
- 6 邓志鸿等. Ontology 研究综述. 北京大学学报, 2002 (5)
- 7 张智雄. Ontology 是什么? 图书情报工作动态, 2004 (3)
- 8 李景, 钱平. 叙词表与 Ontology 的区别与联系. 中国图书馆学报, 2004 (1)

东野广升 冯丽雅 济南大学图书馆。

(上接封三)

题仅靠地方机构的努力是无法实现的。另外,这样长期性的文化工程,不是完全盈利性质的,必然需要财政方面的支持才能很好地得以实现。同时,国家还应该注重人才培养问题,因为古籍善本的保护开发利用工作一直是比较清贫艰苦的工作,工作人员如果得不到重视和尊重是很容易对工作产生倦怠的,这样会不利于古籍人才的接续和技能的提高,而且很难保证“再生”工程的长期性和高质量。

善本“再生”要做好资源共享,减少不必要的浪费和重复的劳动。真正做到“善本得藏,善本得用”并不只是字面上的对善本的开发问题,更包含着在开发过程中再生和合作的一系列相关问题。类似于《四库全书》数字化建设中的重复现象,在很多大大小小的古籍善本“再生”工程中都一定数量地存在着,而且必须得到解决。首先,国家的计划性协调是必须的,每一项工程,即使是地方性质的,也应该制定统一的规划、标准,合理调配资源,增强再生的有序性,这样不仅便于各个领域的共享以及合作,而且有利于以后其他“再生”工程的查找利用,减少不必要的重复和浪费。其次,各个部门的共享协作意识要加强。例如,图书馆在进行一项“再生”项目的时候,不仅仅要熟悉本馆内的馆藏,还要了解其他机构,不能各自为战。如博物馆和档案馆等相关机构是否有可供借鉴的资料,这样才能使“再生”工程更全面和权威。

善本“再生”应该充分运用现代化技术,适应时代的脚步,为现代人所利用,又不失传统文化的精华。传统与现代的结合,历史精髓的吸收与经验的借鉴对现代化建设有着十分重要的作用。因为善本的再生应以发扬传统文化中的精髓,体现中华文明为首要前提,使每项善本“再生”成果都是精品工程,有其各自特色,凝结中华魅力于其中,保证有新意、高质量、高

品位,使读者得到知识积累、艺术享受。但是,在充分运用现代化技术的时候,不能使古籍善本失真,因为许多古籍爱好者正是因为古籍的“原汁原味”才对古籍爱不释手,对他们来说,阅读古人的文章,即使是批注,也是极大的学术和艺术享受。所以在“再生”善本时,必须“因书制宜”,尽量保持原貌不失真。同时,要明确现代技术的运用只是手段,再生古籍,弘扬传统文化才是最终目的,所以不能因为运用了高科技手段而使再生的善本因高价脱离平民,那样则无法达到传承的目的。

中华古籍善本“再生”工程,是关系到国家兴盛、民族振兴、文化传承的大事,需要社会各界坚持不懈地继续下去,在“再生”善本过程中要正确把握方向、坚持原则、注重方法,将踏踏实实的工作作风与勤奋创新的科学理念相结合,为传统文化在新时期发挥应有的作用而努力。

注释

- [1] 彭卫国. 古籍的市场营销. 见: 全国古籍整理出版规划领导小组办公室. 古籍编辑工作漫谈. 济南: 齐鲁出版社, 2003: 65
- [2] 李致忠. 继绝存真传本扬学——《中华再造善本》编纂出版情况简介. 中国出版, 2003: 45 ~ 46
- [3] 冬玲. 记中华再造善本工程. 光明日报, 2003 年 1 月 9 日
- [4] 王伟伟. 古籍整理工作的历史回顾与发展趋势 (上). 图书馆杂志, 2000: 15 ~ 17
- [5] 李修生. 古籍整理与传统文化. 辽宁: 辽宁大学出版社, 1991: 1
- [6] 潘德利. “中华再造善本工程”及其思考. 图书情报工作, 2005, 49 (2): 141 ~ 143

纪晓平 李杨琳 东北师范大学传媒科学学院。