计算机科学2001Vol 28№. 5

XML 及语义 Web 技术

On the XML and Semantic Web Technologies

聂培尧[1,2]
(西北工业大学计算机科学与工程系　西安710072)[1]
安世虎[2]
(山东财政学院计算机信息工程系　济南250014)[2]

Abstract The semantic Web would be built on Markup languages and document descriptions that would let machines understand the nature of page's content. This would take the Web beyond the era of HTML, which lets machines understand the nature of a page's appearance. The occurrence of XML makes the semantic Web possible Some issues of XML-based semantic Web technologies are discussed in this paper

Keywords XML, HTML, Semantic Web

1 引言

XML 的目标就是要改变 Web 的基本结构，超越 HTML 并代之以更强大、更具有可扩展的体系结构。XML 旨在使 Web 返回到基于内容的结构，而不再是开发人员强加给它的基于格式的结构。但是到目前为止，几乎所有的 Web 页面都是用 HTML 编写的。虽然 HTML 具有通用、简单易学、句法简单紧凑等许多优点，使得它得以在 Web 网页上大显身手，但是随着 Web 应用的越来越深入，HTML 过于简单的弱点也越来越突出了。其中一个明显的弱点即是由 HTML 编写的 Web 页面缺乏其语义信息，因为 HTML 只是一种表达的技术，它并不一定能揭示 HTML 标记说明中所表示的含义。举一个简单的例子，〈H2〉Apple〈/H2〉虽然在网络浏览器中有其特定的表现，但是 HTML 却并没有告诉我们它到底是什么，其中的 Apple 只不过是一个英文单词罢了。它在不同的环境下可能会有不同的意思，或许是一个计算机公司、一个水果、也可能是一个姓氏?HTML 并没有明确地告诉我们 Apple 的具体含义。

给 HTML 页面增加语义信息并不容易。很多程序曾经试图用一些非标准的方法来解决这一问题，比如在 HTML 注释中隐藏信息。但是，这样的注释是很难使用的，因为对象模式并不能理解它。正是由于 HTML 无法理解 Web 页面内容，导致了最终不能使用它来开发基于内容的语义 Web。

考虑到这种情况，由 Tim Berners-Lee 作为主任的 WWW 协会(W3C)正在开发一种用来创建语义 Web 的标准。

语义 Web 将由标记语言和文档描述来建造，这将使得 Web 超越 HTML 时代，因为在 HTML 时代机器只能理解页面外观的性质，而并不能理解页面内容的性质。这种进步对几种基于 Web 的关键技术来说是具有广泛含义的。首先，语义 Web 能更准确和有效地对 Web 进行查找，而这恰是基于 Web 活动中的最主要的一环。

Web 查找传统上来说是基于在文档中寻找关键字，因为关键字常常具有多层意思或以多种方法使用之，所以这种方法常常会产生太多的与关键字无关的内容，究其原因，这主要是因为 HTML 无法区分信息与元信息而造成的。而且，HTML 不支持信息嵌套体系结构，限制了全文检索功能。如果 Web 文件中含有标记或其它可对文件的内容提供准确描述的元素的话，这种查找将是更精确的。同时，智能软件代理对语义 Web 站点的理解将会比传统的 Web 站点做得更好，这是十分重要的，某些业界人士估计到2002年，在 Internet 文档查看中有75%之多将由代理来完成。

语义 Web 技术对其它活动也是十分有用的，如内容监探(例如，一个 Web 站点是否适合于儿童)以及对页面集进行描述以表示一个单一的文档。

聂培尧　博士生，教授，主要研究方向为数据库系统理论与实现、Web 技术与数据库系统。**安世虎**　副教授，主要从事智能 DSS 和 MIS 的研究。

为了推进语义 Web 的创建，W3C 已经完成了或者正在制定几种标准，包括 XML（可扩充标记语言）RDF（资源描述框架）以及 XHTML（可扩充 HTML）。但是，要想真正获得语义 Web 还有很多困难，例如，数据库中的大量文档将不得不重新加以标记以使其符合语义 Web 的标准。然而，应当看到，语义 Web 是 Web 发展的必然趋势，因此有必要在这个方向上下大气力来克服这些困难。

2 开发语义 Web

W3C 已经把开发 XML 作为实现语义 Web 的开始，并将 XML 作为 RDF 和 XHTML 标准的基础来使用。

2.1 XML

W3C 自1996年开始开发 XML，并于1998年采纳 XML1.0作为标准。作为 SGML 的一个子集，XML 是一种元语言，可允许用户创建自己的标记语言来对 Web 文档的内容进行描述。XML 提供了用于定义文档句法和组织的一系列规则。XML 使用标记、模式（标记的集合及数据结构的规则），以及文档类型定义（DTD）来描述内容。例如，一个〈PHONENUM〉标记可标明一个表示电话号码的信息。

XML 的关键结构是 DTD 和文档。DTD 定义了许多描述文档内容的标记，大多数的 XML 文档都是通过专门的 DTD 设计的。XML 允许开发者创建自定义的 DTD 来精确地描述特定种类的信息，接收应用使用这一信息来对文档的标记进行解释。大多数的用户，如相同业界的公司，为了满足其自身的需要可以开发他们自己的基于 XML 的标记和应用程序。有些用户已经这样做了，例如金融机构已开发了 FinXML（http://www.finxml.org）。计算机制造商也可使用 XML 来商定一种标准的描述产品（处理器速度、内存大小等）信息的方法，然后可以在他们的 Web 站点中使用这些描述格式以便用户对在这些站点上找到的内容进行理解。

同样，由于 XML 提供了一种公共的数据格式，文档可通过应用程序、平台和界面进行处理和显示。正是由于这一点，很多企业将 XML 看做一种非常重要的应用集成方法。这一方法的思想是，可以使用可扩充的元语言来对构成文档的各个不同部分的意义进行描述，我们可对这种思想进行拓广以支持机器对文件内容的理解，反过来也允许对 Internet 上内容的更丰富的索引和分类。

W3C 现在正在继续精化和扩充 XML。另外，各种不同的群体正在以一种新的使用方法来对 XML 进行扩展。例如 AT&T、Lucent Technologies 及 Motorola 形成了 VoiceXML 论坛（http://www.vxmlforum.org）来创建一种 VXML 标准，从而可创建一种通过电话存取 Web 的内容及服务的方法。

2.2 RDF

1999年早些时候，W3C 通过了 RDF，目前尚待其他的相关组织审核。RDF 是资源描述框架，它提供了一种超信息（元数据）的功能，用来提供一种通用项以描述 Internet 资源（正文页面、图形、音频文件、视频剪辑等）的方法。RDF 为开发者使用 XML 来提供对诸如 Web 站点的页面组织或文档的主题、作者，或目标听众的机器可理解信息的一种方法。

如图1所示，RDF 使用了括号、左斜杠、标记名称、属性及其他一些 XML 中的句法元素。与 XML 情形相同，用户为满足自身的需要，可创建他们自己的句法元素集合。XML 自身并不能作为元数据框架，因为 XML 需要仔细安排元素顺序这样一种复杂的文档构造。这种复杂的文档构造型式在很多一般的 Web 文件中是找不到的。RDF 为开发者提供了一种一致的方法来使用他们的文档以提供元数据，这将对开发者是一种鼓励以促使他们这样做，并因此促使语义 Web 的开发。

```
〈RDF xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:DC="http://purl.org/dc/elements/1.0"〉
〈Description about="http://www.w3.org/folio.html"〉
〈DC:title〉The W3C Folio 1999〈/DC:title〉
〈DC:creator〉W3C Communications Team〈/DC:creator〉
〈DC:date〉1999-03-10〈/DC:date〉
〈DC:subject〉Web development, World Wide Web Consortium, Interoperability of the Web〈/DC:subject〉
〈/Description〉
〈/RDF〉
```

图1 一个由标记和语句构成的 RDF 文档

另外，通过对描述和查询的数据提供一种标准的格式，这种标准将允许在 RDF 应用程序间交换其机器可理解的 Web 数据的互操作性。

在将来，个别的用户及业界群体应当在一起共同工作来建造他们自己共同的 RDF 字典，这将形成一个在概念上相等价的 Web，而对一个机器来说，该 Web 看起来将像一个大的数据库。

2.3 XHTML

W3C 根据在 XML 中对 HTML4.0进行重构而开发了 XHTML。与 XLM 相同，XHTML 用户集可根据情况定义新的标记及属性。XHTML 使用 HTML4.0中的相同的标记类型和属性，这就使得 Web 开发者很容易地转到基于 XML 的文档上。

根据 W3C，XHTML 的主要优点是可扩充性及可移植性。由于用户可定义新的标记和属性，他们可以对 XHTML 进行扩展以满足新的及改进的 Web 方法和技术需求。另一方面，使用 HTML，用户不得不等待，

直到新版本出来以获得新的功能，同样，XHTML 是很灵活的，它为文档定义了标记复杂性的可变的层次，W3C 正在研究 XHTML1.0建议并且最近将返回给协会的 HTML 工作组以做进一步的修改。

3 语义的需求

当前的语义 Web 标准仍在日常的电子商务使用中进行测试，现在还仍面临着很多困难和障碍。例如，销售方和用户必须开发使用该技术的标准方法。因此，应用程序应当是可互操作的。人们预期语义 Web 技术的发展将经历与浏览器最初发展所遇到的有关兼容性类似的问题，为了解决这一问题，商家应当坚持开放性，众所周知，正是由于浏览器制造商，特别是 Microsoft 和 Netscape，各自开发他们自己的 HTML 版本从而导致了浏览器的不兼容问题。这意味着如果开发者不将他们的文档根据不同的浏览器设计多种不同的版本的话，他们的某些文档则不可能在所有的浏览器中进行查看。同样，语义 Web 的处理对用户也必须是透明的。这样可更加方便，而不是更加复杂地进行各种在线活动。

另一个需要确认的问题是数据库中的很多文档应能通过语义 Web 技术进行存取，但是，对这些文档重新进行标记其工作量将是很大的。

Berners-Lee 等人认为目前语义 Web 仅是开发的开始。他估计全部的电子商务按这种逻辑分类方法出现还需十五到二十年的时间。这意味着应当采取审慎的步骤，以一种与创建语义知识库相一致的方法来组织和表示 Internet 中的信息。正文必须标记其语义，书名亦是以这样的方法表示出的。随着越来越多的机器可理解的 WWW 内容的产生，我们会发现我们用来阅读 WWW 内容的工具应当不断地对所含的知识进行收集和过滤。Web 将变成知识的储存地，而不仅仅是事实的一种简编。

结束语 Web 中语义的概念可以允许在 Web 页面的内容中使用更加丰富的字典项，这样就可使机器能够理解文档中的短语的含义，如文档中标记的短语实际上是一本书的名字，一个地址的邮政编码等。这种语义 Web 技术对开发用于市场的 Web 站点是非常有意义的，特别是其中的基于语义的信息检索机制对于用户存取 Web 中的信息是至关重要的，因为这种机制中的语义分析技术将极大地改进用户信息检索的关联性。

语义 Web 技术在电子商务的开发中具有极大的潜力。由于越来越多的 B-to-C（Business-to-Consumer）和 B-to-B（Business-to-Business）事务的出现，有效的电子商务活动也将急需机器能对其内容进行更好、更准确的理解。可以预计，在未来五年内，XML 将对语义 Web 的发展起到关键的推动作用。

参 考 文 献

1 叶文川. 构造未来的 Web 页面的工具语言 XML. Available at: http://www.xml.org.cn/resource/

2 褚建. XML 结构 Available at: http://www xml.org.cn/resource/article/XMLarch.html

3 郁桦. XML 的产生. Available at: http.//www.xml.org.cn/resource/article/xmlintro html

4 Nie Peiyao, Hu Zhengguo. Developing Enterprise-based Web Applications. In: Proc. of the 7th Joint Intl Computer Conf. Santou University, China, 2000

5 Nie Peiyao, Hu Zhengguo. On the Web Publishing Using Markup Language. In. Proc. of the 7th Joint Intl Computer Conf. Santou University, China, 2000

6 Hellman R. A Semantic Approach Adds Meaning to the Web. Computer, 1999, 32(12)

7 XML: A Primer, 2nd Edition, by Simon St. Laurent, IDC Books World Wide, Inc., Foster City, California, USA, 2000

8 XML: In Record Time, by Natanya Pitts, Sybex Inc, USA, 1999

（上接第10页）

参 考 文 献

1 范玉顺，吴澄. 工作流管理技术研究与产品现状及发展趋势 计算机集成制造系统-CIMS, 2000, 6(1): 1～7

2 Luo Haibin, Fan Yushun. CIMFlow: A Workflow Management System Based on Integration Platform Environment. In: 7th IEEE Intl. Conf. on Emerging Technologies and Factory Automation, 1999. 233～241

3 范玉顺，吴澄，王刚，高展. 集成化企业建模方法与工具系统研究. 计算机集成制造系统-CIMS, 2000, 6(3): 1～5

4 Alonso G, Agrawal D, Abbadi El A, et al. Functionality and Limitations of Current Workflow Management Systems. IEEE Expert, 1997, 12(5)

5 WfMC. Workflow Management Coalition Terminology and Glossary (WfMC-TC-1011): [Technical Report] Workflow Management Coalition, Brussels, 1996

6 Edelweiss N, Nicolao M. Workflow Modeling: Exception and Failure Handling Representation Computer Science, SCCC'98. 1998

7 Han Dong-Soo, Shim Jae-Yong. Design and Implementation of a Distributed Transactional Workflow System. IEEE TENCON, 1999