XML 相关语言解析

An Analysis on XML and Related Languages

冒东奎　（银川 西北第二民族学院计算机系　750021）

摘要：本文主要介绍 XML 相关语言，包括定义语言 DTD 和 XML Schema、样式表语言 XSL、样式表转换语言 XSLT、文档格式化对象 XML – FO、路径语言 XPath、链接语言 XLink、指针语言 XPointer、查询语言 XQuery 和可扩展的超文本标记语言 XHTML 的技术内涵，论述这些语言的实现模式、功能、以及相互之间的联系，同时还介绍这些语言的标准化状况。

关键词：XML　XML Schema　XSLT　XPath　XML – FO　XLink　XPointer　XQuery　XHTML

1　引言

万维网联盟 W3C 创建的可扩展标记语言 XML 是一种可以让用户自己定义标记的语言。XML 和 HTML 一样，都源于标准通用标记语言 SGML（Standard Generalized Markup Language）。W3C 于 1996 年 11 月提出最初的 XML 工作草案，1998 年 2 月推出 XML1.0 标准，2000 年 10 月推出 XML1.0 标准的第二版，2004 年 2 月推出 XML1.0 标准的第三版。

W3C 创建 XML 的目的是为了克服 HTML 可扩展性差的局限。HTML 文件里的标记告诉浏览器应该如何显示信息，但却没有告诉浏览器这些信息是什么。人的智慧能理解这些标记的含义，而机器却不能。与 HTML 的设计目标不同，XML 语言是为机器设计的。用 XML 可以给文档中的标记赋予某种含意，从而使机器也容易检索和处理其中的信息。例如机器在以下一段 XML 文档中只要找到 <postal – code> 和 </postal – code> 标记之间的内容，就很容易从该 XML 的地址文档中取出邮政编码。

```
<address>
  <name>王海峰</name>
  <street>东土城路 132 号</street>
  <city>北京</city>
  <state>中国</state>
  <postal – code>100013</postal – code>
</address>
```

XML 的这种自描述性及可扩展性可大大简化 Web 应用之间的数据交换，并支持智能搜索。XML 还可以用结构化形态标识每个信息片段以及这些片段之间的关系。所以能编写用机器直接处理而无需人工干预的 XML 文档。

由于以上的突出优点，XML 近年来在 Web 应用领域展现出强大生命力和重大实用价值，并派生出一系列相关语言，如定义语言 DTD（Document Type Definition）和 XML Schema、样式表语言 XSL（XML Stylesheet Language）、样式表转换语言 XMLT（XSL Transformations）、文档格式化对象 XML – FO（XSL Formatting Objects）、路径语言 XPath、链接语言 XPink、指针语言 XPointer、查询语言 XQuery 以及可扩展超文本标记语言 XHTML（Extensible HTML）等，形成一个语言家族，有力地推动 XML 向实用化方向发展。但是，这些语言之间的关系错综复杂，令人难以掌握它们的来龙去脉，更难跟踪其最新发展。为此，作者拟通过本文分析和透视 XML 及相关语言的技术内涵，包括实现模式、主要功能以及相互之间的联系，同时还介绍这些语言的标准化状况。

2　XML 的定义语言

XML 文档有两种定义语言：一种是文档类型定义 DTD。另一种是 XML Schema，即 XML 模式。模式可以定义能在 DTD 中使用的所有文档结构，还可以定义数据类型和比 DTD 更复杂的规则。

（1）文档类型定义 DTD。DTD 的目的是定义合法的 XML 文件的构造块。它用一个合法元素的列表定义文件结构。DTD 定义可以在 XML 文档中出现的元素、这些元素出现的次序、嵌套关系以及 XML 文档结构的详细信息。DTD 可以在用户的 XML 文件中声明，也可以单独作为一个文件，由 XML 文档从外部引用。DTD 是最初的 XML 规范的一部分，与 标准通用标记语言 SGML 中的 DTD 非常相似。

上述 DTD 代码片段的含义：第 1 行注释这是用 DTD 定义的代码。第 2 行定义 address 元素又包含 5 个元素：name、street、city、state 和 post – code。第 3 行定义 name 元素属"#PCDATA" 类型，即已解析字符数据类型。以下各行定义的元素类型与第 4 行相同。

（2）XML Schema。XML Schema（模式）基于 XML，是 DTD 的替代语言。其功能比 DTD 优越，具体表现在以下几个方面：

· 对数据类型的支持更强：易描述许可的文件内容；易

有效地校正数据;易操作数据库的数据;易定义对数据的限制条件;易定义数据格式;易转换不同类型的数据。

· 采用 XML 的语法:使用户不必学习其他语言,直接可以用 XML 编辑器编辑模式文件,用 XML 解析器解析模式文件,用 XML 的文档对象模型 DOM 操纵模式文件,还可以用 XML 样式表转换 XSLT 来转换模式文件。

· 允许可靠的数据通信:借助 XML Schemas,数据发送方可以按照接收方可以理解的格式发送数据,因此不会因使用的格式不同而发生误解。

· 可以扩展:因为 XML Schema 使用 XML 书写,所以像 XML 一样能扩展。一个模式定义在其他模式文件中能重用。用户可以从标准数据类型创建自己的数据类型。从同一文件里可以引用多个模式定义。

· 语法更严格,并且支持验证:如 XML 模式定义必须以 XML 声明开头、根元素必须惟一、起止标签必须匹配、标签对大小写敏感、所有元素必须封闭、嵌套必须得当、属性值必须提供、XML 实体必须使用专用字符。

W3C 于 2000 年 2 月发布了 XML Schema 基本部分工作草案,2001 年 5 月发布了 XML Schema 基本部分、结构部分和数据部分的推荐标准。

该 XML Schema 代码片段的含义:第 1 行指明按 XML 1.0 标准定义。第 2 行声明 XML Schema 的名字空间。其余各行定义了一个名为 name 的元素,是复杂类型,包含 5 个元素,都是字符串类型,顺序为 name、street、city、state、post – code 。

3 XML 的样式表语言

网页设计者都知道,HTML 文档需要用层叠样式表语言 CSS 给 HTML 页面元素添加显示风格,告诉浏览器以何种字体和颜色显示 HTML 页面元素。同样,XML 文档也需要有一种样式表语言描述文档怎样显示,这种语言就是可扩展样式表语言 XSL。XSL 包含三个部分:XML 文档的转换语言 XSLT、引用 XML 文件的某部分内容的路径语言 XPath、以及? XML 的文档格式化对象 XSL – FO。

W3C 于 1998 年 8 月发布了 XSL 1.0 工作草案,后来多次发布修订版。2001 年 10 月 W3C 发布了 XSL 1.0 推荐标准。

(1) XML 文档的转换语言 XSLT。XSLT 是 W3C 的 XSL 标准的最重要的部分,用于将一个 XML 文档转换为另一个 XML 文档,或者可以被一种浏览器识别的另一类文档,例如超文本标记语言 HTML 文档和可扩展的超文本标记语言 XHTML 文档。通常 XSLT 通过将每个 XML 元素转换为以上两种文件的元素的方式来实现文档转换。XSLT 可以在输出的目标文件中添加新元素,或者去掉某些元素。还可以将元素重新排列、分类、测试,并决定显示哪一些元素等。描述 XSLT 转换过程的通用说法是将 XML 源树转换为目标树。XSLT 使用 XPath 定义转换的匹配模式。在转换过程中,XSLT 使用 XPath 确定源文件中匹配一个或多个预定模板的部分。当找到匹配的部分时,XSLT 将源文件的匹配部分转换到目标文件。源文件与模板不匹配的部分在目标文件中不做修改。

W3C 于 1999 年 11 月发布了 XML 文档转换语言推荐标准 XSLT 1.0 。从 2001 年 12 月到 2003 年 12 月 W3C 又连续发布了五个 XSLT 2.0 版的工作草案。

(2) XML 路径语言 XPath。XPath 是对 XML 文档中某些部分进行寻址的语言。XPath 并不用 XML 书写。XPath 使用路径表达式识别 XML 文档里的节点,这些路径表达式看起来很像计算机文件系统使用的传统的文件路径。XPath 定义了一个标准函数库,用于操作字符串、数字和逻辑表达式。XPath 是 XSLT 的主要成员。Xpath 可以使用在 XSLT、XQuery(查询语言)、XPointer(指针语言)和其他 XML 解析软件中。

W3C 于 1999 年 11 月发布了 XPath 的推荐标准 XPath 1.0,2001 年 12 月 W3C 又发布了 XPath 2.0 的工作草案。XPath 2.0 是由 XPath 1.0 和 Xquery1.0 派生的语言。XPath 2.0 和 XQuery 1.0 工作草案共享同样的语法,大部分文本也相同。

(3) XML 文档格式化语言 XSL – FO。XSL – FO 是格式化 XML 数据的语言。XSL – FO 的全名是可扩展样式表语言格式化对象。XSL – FO 是一个说明 XML 文档格式化语义的词汇表。它基于 XML 的标记语言,并描述输出到屏幕、纸张或其他介质上的数据的格式。格式化实际上是一个将 XSL 转换结果再转换成适合阅读者或收听者接收的输出的过程。

XSL – FO 后来被正式命名为 XSL。实际上 XSL – FO 和 XSL 是同一个东西。因为样式的含义既包括转换信息,也包括格式化信息。W3C 当初在制定第一个 XSL 工作草案时,在草案里既包含转换 XML 文档的语法,也包含格式化 XML 文档的语法。后来 W3C 的 XSL 工作组将原工作草案分成三个单独的推荐标准,即转换信息的语言 XSLT、格式化信息的语言 XSL 或 XSL – FO 以及对 XML 文档某些部分进行寻址的路径语言 XPath。

W3C 对可扩展样式表语言的格式化对象没有单独的标准文件,但它的规范可以在 W3C 的 XSL 的推荐标准 XSL 1.0 中找到。

以下是 XSL – FO 文件示例,第 2 行的 <fo:root> 元素意为包含 XSL – FO 文件,也声明该文件的名字空间。其余部分的含义见代码中间插入的注释。

```
<? xml version = "1.0" encoding = "ISO – 8859 – 1" ? >
<fo:root xmlns:fo = "http://www.w3.org/1999/XSL/For-
```

```
mat" >
<fo:layout-master-set>
  <fo:simple-page-master master-name="A4">
    <!-- Page template goes here -->
  </fo:simple-page-master>
</fo:layout-master-set>
<fo:page-sequence master-reference="A4">
  <!-- Page content goes here -->
</fo:page-sequence>
</fo:root>
```

4 XML 的链接语言和指针语言

(1) XML 链接语言 XLink。XML 链接语言 XLink 允许将一些元素插入 XML 文档,以创建和描述 Web 资源之间的链接。XLink 使用 XML 的语法创建一些结构,可以描述类似于现今的 HTML 的简单、单向的超级链接,也可以描述更复杂的多终点的和定型的链接。

W3C 于 1998 年 3 月提出了 XLink 规范的工作草案,后来发布了一系列工作草案修订版,2001 年 6 月发布了 XLink 1.0 推荐标准。

以下是用属性元素 ABC 做 XLink 简单链接的一段代码,其中定义了 URI 资源及其作用和标题,并定义该资源在被请求时激活,覆盖当前窗口。

```
<ABC
    xlink:type = "simple"
    xlink:href = "http://www.w3.org/"
    xlink:role = "w3chome"
    xlink:title = "W3C Home Page"
    xlink:show = "replace"
    xlink:actuate = "onRequest" >
    The W3C
</ABC>
```

(2) XML 指针语言 XPointer。XML 指针语言 XPoint 是用来对互联网上采用统一资源标识(URI)的信息做片段识别的语言,这些资源的介质类型可以是 text/xml、application/xml、text/xml 的外部解析实体,或者 application/xml 的外部解析实体之中的任何一种。

XPointer 基于 XML 路径语言 XPath,支持在 XML 文档的内部结构里寻址。允许检查层次型的文档结构,并根据各种属性,如元素类型、属性值、字符内容和相对位置等选择文档的部分内容。XPointer 可以指向字符数据的子串和整个树型结构的若干片段。

W3C 于 1999 年 7 月发布 XPointer 的工作草案。后来虽然发布了多个工作草案修订版,但 2002 年 8 月最新发布的还是工作草案。

以下是 XPointer 的一个示例,整个字符串是用统一资源标识符(URI)表示的资源:

http://www.foo.org/bar.xml#xpointer(article/section[position()<=5])

其中"#"号之后的部分,即 xpointer(article/section[position()<=5])是 XPointer 指向的资源片段的标识符,article/section[position()<=5]是 XPointer 表达式。含义为该 XML 指针指向 article 根元素的前 5 个段落的元素。

5 XML 的查询语言

XML 的查询语言 XQuery 是用于从 XML 数据源查询数据的一种语言,它还有能从多种关系型数据库和许多数据文档中查询数据的特点。XQuery 的设计目标是提供一种从 Web 和多种数据库的真实的或者虚拟的 文件里提取数据的灵活手段,以最终为 Web 世界和数据库世界之间提供一个必要的接口,使访问 XML 文档像访问数据库一样方便。

XQuery 构建在 XPath 规范之上。XQuery 1.0 和 XPath 2.0 共享同样的数据模型、同样的函数和同样的语法。

XQuery 是一种将查询表示成表达式的功能语言,它能够用 XPath 表达式从文档里选择特殊的节点序列,也能实现各种不同形式的查询。XQuery 的表达式可以互相嵌套,也支持子查询。目前,数据库业界的三大主流厂商 Oracle、IBM、Microsoft 都已经在各自的产品中提供了对 XQuery 规范的支持。

W3C 于 2004 年 7 月将 XQuery 1.0 和 XPath 2.0 全文作为工作草案一起发布。

下面是一个简单的 XQuery 查询样例:

```
//Catalog[@name="disk"]/@price
```

相当的 SQL 查询语句为:

```
select Catalog.price from Catalog
where Catalog.name="disk"
```

6 可扩展超文本标记语言 XHTML

可扩展的超文本标记语言 XHTML 由 HTML 4.01 的所有元素和 XML1.0 的语法结合而成。XML 的设计目标是描述数据,而 HTML 的设计目标是显示数据。XML 的语法很严谨,所以用它设计出的文档也很规范。而目前互联网上的 HTML 页面就未必规范。如今的市场上有不同的浏览器技术,有些浏览器在计算机上运行,而另一些在移动电话或者手持设备上运行。后者就没有资源和能力解释不规范的标记语言。

(下转第 52 页)

(上接第 49 页)

因此需要将 HTML 和 XML 的优势结合到一起,得到一种更好用的标记语言。XHTML 使程序员现在就可以编写格式规范的网页,在所有浏览器上可以工作,落后的浏览器也兼容。

XHTML 是新一代的 HTML,它与 HTML 最重要的差别是:

- XHTML 的元素必须适当嵌套。
- XHTML 文档格式必须规范。
- 标记名称必须小写。
- 所有 XHTML 元素必须封闭。

XML 的设计目标不是为了替代 HTML,而 XHTML 的设计目标却是为了替代 HTML。XHTML 使程序员现在就可以编写规范的、既可以在当前的浏览器上运行,又可以与落后的浏览器兼容的网页。

W3C 于 2000 年 1 月发布 XHTML1.0 推荐标准,2002 年 8 月推出 XHTML1.0 推荐标准的第二版。

7 结束语

XML 及其相关语言为不同应用程序之间的消息传输提供了所需的灵活性,目前已经成为 Web 应用领域事实上的数据表示和数据交换的标准。随着近年来 Web 服务的蓬勃发展,XML 语言家族越来越多地活跃在数据交换和存储领域,用 XML 表示的结构化数据的应用越来越普遍,在 B2B 电子商务行业中尤其明显。面对 XML 应用的指数级增长,Web 应用开发者必须尽快学会这些语言的使用方法,并掌握这一技术领域的最新发展。

参考文献

1 http://www.w3.org/TR/REC-xml/.
2 http://www.w3.org/XML/Schema#dev.
3 http://www.w3.org/Style/XSL/.
4 http://www.w3.org/TR/xslt.
5 http://www.w3.org/TR/xpath.
6 http://www.w3.org/TR/WD-xptr.
7 http://www.w3.org/TR/xlink/.
8 http://www.w3.org/XML/Query.
9 http://www.w3.org/TR/xhtml1/.
10 http://www.vbxml.com/xsl/tutorials/intro/default.asp.