文章编号:1671 - 6833(2002)04 - 0017 - 04

XML 文档解析技术及其应用

逯 鹏

(郑州大学电气工程学院,河南 郑州 450002)

摘 要:介绍了 XML 文档应用的主要技术环节,分析了基于事件驱动和基于树结构的两种不同的 XML 文档解析技术,比较了基于事件驱动和基于树结构的解析技术的特点和适用环境,同时给出了利用树结构的 XML 解析技术实现流式文本数据转化为 XML 数据的方法,这一方法解决了专利申请文件遗留数据转换成 XML 文件的问题.

关键词:可扩展标记语言;文档对象模型;解析技术

中图分类号:TP 311.12 文献标识码:A

1996 年,W3C 组织开始设计一种 Web 数据存储和交换的标准,这就是 XML(eXtensible Markup Language,可扩展标记语言),它使得现有的因特网协议和软件更为协调,从而简化了对数据的处理和传输.XML 所拥有的可扩展性、自描述性、自相容性以及跨文种等优点,使得它非常适于 Web 上的数据交换与信息发布,被广泛应用到电子商务、电子政务、Web 服务等许多领域.目前很多国际著名的公司已完全加入到 XML 支持者的行列,如微软 IE 6.0 已广泛使用了 XML, Netscape 新版本也将会支持 XML,其它公司,包括 IBM,Adobe,Sun 和 Xerox 等也宣布支持 XML,并都在着手相关产品的研制.相对于 HTML 的"所见即所得",XML 将数据和显示信息分离,被称为"文档数据库",这就使 XML 文档很适合于描述数据库中的数据.而其它非标准化、非结构化的数据转换为 XML 文档后,就可以将大量遗留数据实现信息共享和交换.XML 解析技术是操作 XML 文档的重要环节,它在实现把非结构化数据转换为 XML 数据的过程中有着重要的作用.本文分析了不同的 XML 文档解析技术,并利用基于树结构的 XML 文档解析技术实现了将流式文本数据转化为 XML 数据.这一方法已经得到实际应用.

1 应用 XML 文档的主要技术环节

应用 XML 文档主要有三个技术环节,包括对 XML 文档的编辑、XML 文档的解析和显示或使用 XML 文档的数据.

XML 文档的编辑目前有两种方式:一种是使用通用的编辑器生成 XML 文档,如 Notepad;另一种是使用 XML IDE(集成开发环境),如 XML,SPY 等工具.也可以针对具体应用开发某种专用的可视化 XML 编辑器.

XML 文档解析的主要任务是检查编辑的 XML 文档是否是结构完整(Well - formed)和合法的(Validate).如果 XML 解析器发现 XML 文档中的数据或者结构不完整,就会向应用程序报告一个"致命"错误,而且不再会以正常的方式向应用程序传递数据或 XML 结构.

目前的 IE 浏览器已经支持 XML 文档的显示,可以用 IE 直接打开 XML 文档显示嵌套的结构.如例 1 所示的 XML 文档的片断.也可以将数据翻译成数据库数据,存入数据库,或者传递给运行中的 java 程序.

例 1 一个 XML 文档的片断:

```
<?xml version="1.0" encoding="UTF-8"?><发明专利请求书>
  <发明名称>多功能鼠标</发明名称>
  <申请人>
    <姓名> 张三</姓名>
    <省自治区直辖市名称>河南省</省自治区直辖市名称>
    <市县名称>郑州市</市县名称>
    <城区乡街道门牌号>文化路 97 号</
```

收稿日期:2002 - 08 - 10;修订日期:2002 - 09 - 28

作者简介:逯 鹏(1974 -),男,河南省滑县人,郑州大学助教,北京航空航天大学博士研究生,主要从事电子商务与信息安全方面的研究.

城区乡街道门牌号 >
< /申请人 >
< /发明专利请求书 >

2 XML 文档的解析技术

XML 的解析技术在 XML 文档的应用过程中有着重要的作用,它的行为减少了应用程序处理 XML 数据的负担,为应用程序和数据库提供了可操作的数据. 关于 XML 解析技术,目前的争论非常之多,与许多其它技术问题一样,XML 文档的处理需求有着很大的区别,不同的技术实现方案会适合不同的问题域.

2.1 基于事件驱动的 XML 解析技术

基于事件驱动的解析技术主要是围绕着事件源以及事件处理器来工作的. 一个可以产生事件的对象被称为事件源,而可以针对事件产生响应的对象就被称作事件处理器. 事件源和事件处理器是通过在事件源中的事件处理器注册方法连接的. 这样当事件源产生事件后,驱动事件处理器相应的处理方法,一个事件就获得了处理. 当然在事件源调用事件处理器中特定方法的时候,会传递给事件处理器相应事件的状态信息,这样事件处理器才能够根据事件信息来决定自己的行为. 这种方式需要的内存小,运行速度快[1].

基于事件驱动的处理模式是一种通用的程序设计模式,被广泛应用于 GUI 设计. 在 JAVA 的 AWT,SWING 以及 JAVA BEANS 中就有它的身影.

访问 XML 文档的 SAX(Simple API for XML, SAX)接口是采用基于事件驱动 XML 解析技术的代表. 使用 SAX 接口可以将 XML 文档转化成一系列的事件,由单独的事件处理器来决定如何处理. SAX 并不是一个实际可以使用的 XML 文档解析器,而是其它兼容 SAX 的解析器要实现的接口和帮助类的集合. 实现了 SAX 的解析器有很多,比如 Apache 的 Xerces,Oracle 的 XML Parser 等等. 在 SAX API 中有两个包,org. xml. sax 和 org. xml. sax. helper. 其中 org. xml. sax 中主要定义了 SAX 的一些基础接口,如 XMLReader, ContentHandler, ErrorHandler,DTDHandler,EntityResolver 等. 而在 org. xml. sax. helper 中则是一些方便开发人员使用的帮助类,如缺省实现所有处理器接口的帮助类 DefaultHandler、方便开发人员创建 XMLReader 的 XMLReaderFactory 类等等.

如例 1,SAX 从第一个“ < ”开始分析 ,读到“ 发明专利请求书 ”后的“ > ”即反映出它是一个开始标注,之后 SAX 将这个小信息包传送给应用程序.

2.2 基于树结构的 XML 解析技术

基于树结构的 XML 解析技术是将结构完整的 XML 文档定义为一棵树,如图 1 所示. XML 文档的组织结构是层层嵌套而完成的,每个 XML 文档都有一个根结点,后跟一个或者一些元素,根结点代表文档本身,元素可以看作根结点的孩子或者是树的分支,孩子元素还可以有孩子元素,从而构成了整个 XML 文档树. 在解析 XML 文档树时,处理器从文档内的第一个元素——根元素开始解析,然后解析它面对的每个分枝形成的路径,最终解析整个 XML 文件 ,将该 XML 文档各元素组成相对应的树形结构,然后向应用程序发送. 树是广泛应用的一种数据结构,将 XML 文档解析成树结构以后,许多成熟的算法都可以用来遍历、搜索、编辑 XML 文档树.

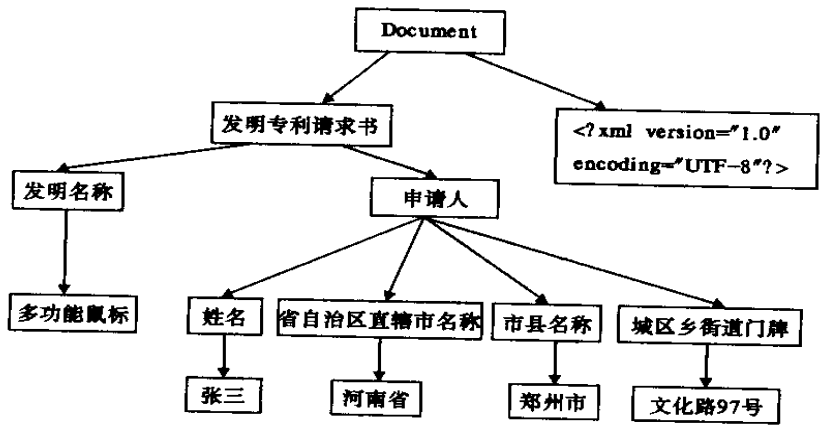

图 1 XML 文档树

Fig. 1 Document tree of XML

使用基于树结构的 XML 解析技术的解析器通常遵守 W3C 的文档对象模型 DOM(Document Object Model). DOM 是一种独立于平台、语言的接口,它能够对树结构的文档进行操作.操作文档之前必须在内存中生成 DOM 树[2],树结构生成之后,应用程序就可以通过相关的 API 访问 DOM 树.根据 W3C 的 DOM1.0 规范,DOM 的一些核心接口包括[3]:① Document,整个文档;② Document Fragment,文档的一部分;③ Dom Implementation,DOM 的实现方式;④ NodeList,节点顺序接口;⑤ NamedNodeMap,使用节点名属性获得节点句柄;⑥Text,文本数据;⑦Node,文档树中的单个节点;⑧CharacterData,Node 的扩展;⑨ Attr,属性;⑩ Element,文档的元素;⑪Comment,注释内容.

2.3　两种解析技术的比较

基于树结构的解析技术的树结构思想与 XML 文档的结构相吻合,它把整个 XML 文档以一棵树的形式存放在内存中,通过树结构应用程序可以很容易实现随机访问.这种访问方式给应用程序的开发带来了很大的灵活性,它可以任意地控制整个 XML 文档中的内容.然而,当 XML 文档比较大或者文档结构比较复杂时,对内存的需求就比较高.而且,对于结构复杂的树的遍历也是一项比较耗时的操作,会影响应用系统的执行效率.

基于事件驱动的解析技术提供的是一种对 XML 文档的顺序访问机制,对内存的要求比较低,程序执行效率高.由于基于事件驱动本身是有时序性的,对于已经分析过的部分,不能回溯重新处理.此外,与基于树结构的解析技术相比,目前支持基于事件驱动的解析器只能读 XML 文档,不能写 XML 文档,对 XML 文档的处理也缺乏一定的灵活性.

所以,两种解析技术根据不同的需求各有优势和劣势,可以在开发中互补.

3　基于树结构解析技术的一种应用

遗留数据格式转换为 XML 格式是实现企业和政府机构开展电子商务和电子政务的关键一步.但多数遗留的数据结构不符合 XML 的格式要求.例如,国家知识产权局专利局的专利申请文件的遗留格式是流式文本数据,这些数据保存在.rtf 格式的文件中,如例 2 所示,本身不存在层次结构,是以文本流的形式存在的.

例 2　一条描述发明专利请求书的片断:

【发明名称】多功能鼠标【申请人】【姓名】张三【省自治区直辖市名称】河南省【市县名称】郑州市【城区乡街道门牌号】文化路 97 号

在实现专利申请文件转换为 XML 文档的过程中,借鉴了基于树结构的解析技术,采用了建立基于 TOM(Text Object Model)模型的 RTF 索引树(如图 2 所示)及与 XML DOM 树转换的方法,以实现流式文本数据到 XML 数据的转换.TOM 模型是微软定义的访问 RTF 文档的一组接口[4].TOM 模型将 RTF 文档解析为树型结构,通过 TOM 接口,程序员可以操作由 RTF 文档组成的内容.

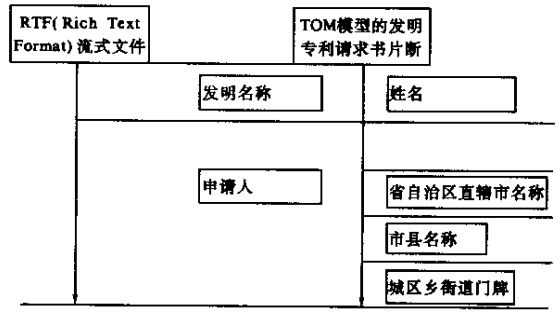

图 2　发明专利请求书 RTF 文件的 TOM 模型示意图

Fig.2　TOM model of patent application RTF files

实现两种树结构的转换的思路(如图 3 所示)如下:①顺序读取 RTF 流式文件;②以文件中的特殊字符【';'】为标志建立 RTF 索引树 TOM 模型;③通过 DOM 接口将相应文件的 XML 空文档导入;④按照转换规则将 TOM 树的内容通过 DOM 接口函数写入 XML 文档.

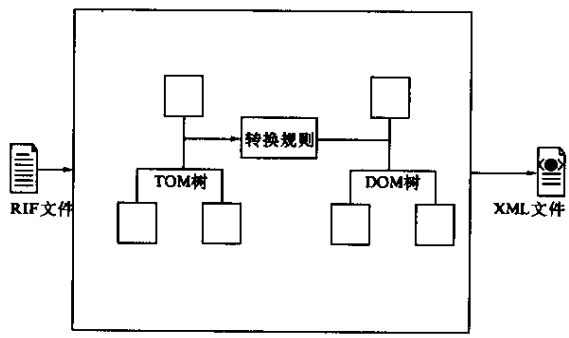

图 3　RTF 流式文件到 XML 文件的转化示意图

Fig.3　Transforming RTF stream file to XML file

4　结束语

XML 的两种解析技术具有各自的优点,在应用过程当中要针对不同的需求,运用相应的技术.采用基于树结构的解析技术解决了 RTF 流式文本到 XML 文本的转换,这一实现方法已成功地应

用到了国家专利局电子申请系统当中,取得了良好的效果.这一转换思路也可以作为其它形式转换的借鉴.需要指出的是,这一应用是有针对性的,并不适用于所有的流式文本数据,进一步的工作是要结合两种解析技术给出通用的转换模式.

参考文献:

[1] DIDIER Martin. XML 高级编程[M]. 李 喆,译. 北京 机械工业出版社,2001.

[2] DOM Working Group. Document Object Model (DOM) [DB/OL]. http://www.w3c.org/dom,2002-01-10.

[3] 郭洪锋. DOM 文档操作和 XML 文件互相转换的 java 实现[DB/OL]. http://www-900.ibm.com/developerWorks/cn/xml/x-dmj/index.shtml,2001-12-01.

[4] STEVE M. An overview of TOM model [DB/OL]. http://www.vbaccelerator.com/codelib/richedit/tomdemo.htm, 2002-02-03.

The Technique of Parsing XML Document and Its Application

LU Peng

(College of Electric Engineering, Zhengzhou University, Zhengzhou 450002, China)

Abstract: This paper introduces the main part of XML technique and puts the emphasis on two parse techniques of XML. It also compares two parse techniques, one of which is based on event-driven and the other is based on tree structure, including their characters and the environments of application. The method of transforming the text data based on stream format into the text data based on XML structure, and it solves the problem of transforming the left data of patent application files into XML files.

Key words: XML; DOM; parse technique