

一种基于 Web 的 HTML 到 XML 数据转换方法

刘江宏 刘金瑄

(西安科技大学计算机学院 西安 710054)

摘 要 随着 Internet 的发展,以 HTML 格式显示的 Web 数据越来越不适应新的发展需求,而用来描述和存储数据的 XML 语言有着许多优于 HTML 的技术,于是将 HTML 格式的数据用 XML 格式表示出来,是现在网络应用中需要解决的问题。这里介绍的基于 Web 的 HTML 到 XML 数据转换方法能够有效地把 HTML 格式的文件转换成 XML(XHTML)格式的文件。

关键词 HTML HTMLDOM 树 XHTML XML

中图分类号 TP311

Study on HTML to XML Conversion Technology Based on Web Data

Liu Jianghong Liu Jinxuan

(Department of Computer Science, Xi'an University of Science and Technology, Xi'an 710054)

Abstract With the development of the Internet, Web data described with HTML can't adapt to the new developed needs increasingly, but XML for describing and storing data is better than HTML. As a result converting the HTML data format to the XML format is a problem that need be solved in the network application now. This paper will introduce a method of HTML to XML conversion technology based on Web Data which can solve the problem effectively.

Key words HTML, HTMLDOM tree, XHTML, XML

Class Number TP311

1 引言

随着互联网的快速发展以及相关产业的应用需求,XML 已经成为互联网上数据描述以及应用系统中数据交换的标准,目前 Web 中大部分的数据是以 HTML 格式显示的,这种非结构化、半结构化的 Web 载体着重数据的显示而缺乏对数据本身的描述,这使得应用程序无法直接解析并利用 Web 上海量的信息,造成资源极大的浪费。于是,研究如何把 HTML 格式的文档转换成 XML(XHTML)格式的文档成为非常有现实意义的工作,这为 Web 数据挖掘、Web 信息抽取等工作带来极大方便,从而有效利用 Web 信息资源。本文将介绍一种 HTML 到 XML(XHTML)数据转化算法的具体实现过程。

2 HTML 和 XML(XHTML)的比较

HTML(Hyper Text Markup Language,超文本标记语言)^[1]是一种描述性的标记语言,起源于标准通用置标语言 SGML(Standard Generalized Markup Language)。用 HTML 编写的超文本文档称为 HTML 文档,它能够独立于各种操作系统平台。HTML 透过标签(Tag)式的指令,将影像、声音、图片、文字等显示出来。

XML(eXtensible Markup Language,可扩展标记语言)^[2],是 WWW 上信息交换的新标准。XML 的语法类似于 HTML,都是用标签来描述数据。HTML 的标签是固定的,只能使用而不能修改,而 XML 允许用户根据需要自行定义标签及属性名,结构化地描述信息内

• 收稿日期:2008 年 9 月 6 日,修回日期:2008 年 10 月 2 日
作者简介:刘江宏,女,硕士研究生,研究方向:数据挖掘、数据库技术等。

容。HTML 用于显示数据,而 XML 用来描述和存储数据。

XHTML(The eXtensible Hyper Text Markup Language,可扩展超文本标记语言)是一种基于 XML,类似于 HTML 的语言,它结合了 XML 的强大功能及 HTML 的简单特性,是一个 HTML 向 XML 过渡的技术,属于 XML 的子集。

HTML 和 XML 都起源于 SGML,与 HTML 相比,XHTML 主要有以下特点:

1)XHTML 解决了制约 HTML 语言发展的问题。HTML 发展到现在主要存在三个缺点:不能满足越来越多的网络设备和应用的需要(如手机、PDA、家电等不能直接显示 HTML);由于 HTML 代码不规范并且臃肿,需要足够智能的浏览器才能正确显示 HTML;数据与表现混合在一起,当页面需要改变显示时,不得不重新制作 HTML。因此,要解决这些问题,HTML 需要进一步的发展。于是,W3C 制定了 XHTML,它是一个 HTML 向 XML 过渡的桥梁。

2)XML 是 Web 发展的趋势,人们希望能够快速地进入潮流,XHTML 是当前替代 HTML 标记语言的标准,它严格地建立在 XML 基础之上,我们可以利用强大的 XML 标准技术来操纵 XHTML 文档^[3],从而简化应用程序的开发和维护。使用 XHTML,只要遵守一些简单规则,就可以设计出既适合 XML,又适合目前大部分 HTML 浏览器的页面,使 Web 平滑的过渡到 XML。同时,XHTML 能够与其它基于 XML 的标记语言、应用程序及协议进行良好的交互工作。

3 基于 Web 的 HTML 到 XML 数据转换算法

3.1 算法思想

数据转换思路为:首先利用 HTML 解析器把 HTML 文档解析成为一棵 HTMLDOM 树^[4],为文档添加 XML 文档声明,然后遍历该 HTMLDOM 树,根据节点的不同类型对节点进行判断,若为文本节点直接打印,若是元素节点则在取出节点之前先打印“<”,然后填充节点(包括节点名称以及节点属性值)内容,最后打印“>”,如果元素节点有子节点,则以同样方式递归打印出所有孩子节点,直到遍历所有节点结束,最后关闭元素节点,文档遍历结束即可得到对应的 XHTML 文档。程序流程图如图 1 所示。

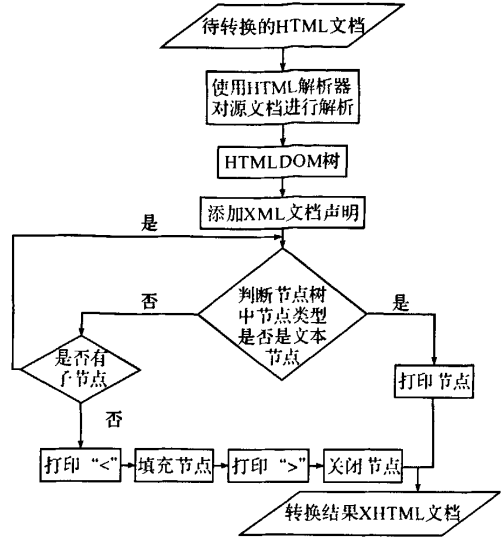


图 1 HTML 到 XHTML 的数据转换流程图

3.2 实现过程

由上面的数据转换思路可知,首先要利用 HTML 解析器把 HTML 文档解析成 HTMLDOM 树。本文使用 NekoHTML 作为解析器,NekoHTML 着重于 Parser 本身,并且 XML 应用一般是围绕 Parser 建立的,因此,这里我们采用 NekoHTML 来解析 HTML 页面,并用标准的 XML 接口来访问其中的信息。这一过程主要是通过 org.w3c.dom 中提供的接口来实现的,这些接口可以通过 nekohtml 和 xerces 来实现。这个包中的接口主要有:Node、Document、Element、Text 等,其中 Node 接口是根接口,它的子接口有 Document、ProcessingInstruction、Element、Comment、Text。通过这些接口可以很方便地访问 HTMLDOM 树中所有的节点。

HTML 到 XHTML 数据转换的具体实现步骤及算法描述如下:

1)创建一个 DOMParser,调用 DOMParser 的 parse 方法把 HTML 文档解析成 HTMLDOM 树。

2)使用 printHTMLDOMTree(Node, PrintWriter)函数递归遍历解析后所得到的 HTMLDOM 树。printHTMLDOMTree(Node, PrintWriter)函数的算法实现具体描述如下:

```

printHTMLDOMTree(Node node, PrintWriter writer)
{
    switch(node.getNodeType())
    {
        //对不同类型的节点进行判断处理
    }
}
  
```

```

case Node . DOCUMENT_NODE: {
    writer.println("<? xml version="1.0" encoding="
gb2312"? >"); //加入 XML 文档声明
    printHTMLDOMTree(((Document)node).getDocument-
tElement(), writer);
    break;
} //打印文档节点
case Node . ELEMENT_NODE: { //节点类型为元素节点
    Element elmt=(Element)node;
    writer.print("<");
    writer.print(node.getNodeName());
    NamedNodeMap attrs=node.getAttributes();
    //获得这个节点的所有属性装入 NamedNodeMap 对
象中
    for (int i=0; i<attrs.getLength(); i++){
        Node attr=attrs.item(i);
        //根据元素的个数打印所有的属性节点
        writer.print(attr.getNodeName()+"="+attr.getNode-
Value()+"");
    }
    writer.println(">");
    //如果元素有子节点则将该元素的子节点装入 NodeList
中
    NodeList children=node.getChildNodes();
    if (children!=null){
        int len=children.getLength();
        for (int i=0; i<len; i++){
            printHTMLDOMTree(children.item(i), writer); //针对
该节点的每一个子节点递归调用 printHTMLDomTree 方法
        }
    }
    break;
}
case Node . TEXT_NODE: { //节点类型为文本节点
    text = URLEncoder.encode (node.getNodeValue(), "
ISO8859_1"); //对文本节点进行编码处理
    writer.println(text);
    break;
}
}
if (type==Node.ELEMENT_NODE){
    writer.print("</");
    writer.print(node.getNodeName());

```

```

writer.println(">");
} //关闭元素节点
writer.flush();
}

```

3)完成遍历,生成 XHTML 文档。


4 结语

随着 Web 上信息组织结构的飞速发展,Web 数据的异构性不断增强,发现 HTML 文档已经不适应新的发展需求,如基于 HTML 的信息检索技术越来越不适应复杂的 Web 数据查询的需要,于是出现了 XML 语言,它有着许多优于 HTML 的技术,那么如何将 HTML 中的信息用 XML 格式描述出来,是现在网络应用中必须解决的问题。本文介绍的基于 Web 的 HTML 到 XML 数据转换方法能够有效地把 HTML 格式的文件转换成 XML(XHTML)格式的文件,从而便于 Web 数据集成和 Web 信息的准确查询等工作的研究。

参考文献

- [1]HyperText Markup Language, W3C Recommendation[EB/OL]. <http://www.w3.org/TR/html401/>. 1999~12
- [2]Bray T, Paoli J, Sperberg-McQueen C M. Extensible Markup Language (XML) 1.0, W3C Recommendation[EB/OL]. <http://www.w3.org/TR/1998/REC-xml-19802108> December, 1997
- [3]The Extensible HyperText Markup Language W3C Recommendation[EB/OL]. <http://www.w3.org/tr/xhtml1>
- [4]Document Object Model, W3C Recommendation[EB/OL]. <http://www.w3.org/DOM/>. 1998~10
- [5]黄晓斌. HTML 向 XML 转换的研究[J]. 现代图书情报技术, 2003, (1): 18~20
- [6]戴怡钧. HTML 转换到 XML 格局以及不同 XML 标准格式之间的转换[D]. 上海: 上海交通大学, 2003
- [7]LaurentSS 云舟工作室译. XML 基础与应用[M]. 北京: 中国水利水电出版社, 2002
- [8]陈艳梅. HTML 到 XML 转换技术的研究与实现[D]. 沈阳: 东北大学, 2002

一种基于Web的HTML到XML数据转换方法

作者: [刘江宏](#), [刘金瑄](#), [Liu Jianghong](#), [Liu Jinxuan](#)
作者单位: [西安科技大学计算机学院, 西安, 710054](#)
刊名: [计算机与数字工程](#) 
英文刊名: [COMPUTER AND DIGITAL ENGINEERING](#)
年, 卷(期): 2009, 37(1)

参考文献(8条)

1. [HyperText Markup Language, W3C Recommendation](#) 1999
2. [Bray T; Paoli J; Sperberg-McQueen C M Extensible Markup Language \(XML\) 1.0, W3C Recommendation](#) 1997
3. [The Extensible HyperText Markup Language W3C Recommendation](#)
4. [Document Object Model, W3C Recommendation](#) 1998
5. [黄晓斌 HTML向XML转换的研究\[期刊论文\]-现代图书情报技术](#) 2003(01)
6. [戴怡钧 HTML转换到XML格局以及不同XML标准格式之间的转换\[学位论文\]](#) 2003
7. [Laurent SS云舟工作室 XML基础与应用](#) 2002
8. [陈艳梅 HTML到XML转换技术的研究与实现\[学位论文\]](#) 2002

本文读者也读过(10条)

1. [于国良, 韩文报, Yu Guoliang, Han Wenbao XML的签名\[期刊论文\]-计算机工程与应用](#) 2006, 42(7)
2. [耿祥义, 刘强, GENG Xiang-yi, LIU Qiang 浅谈XML数据的安全交换技术\[期刊论文\]-电脑知识与技术](#) 2009, 5(22)
3. [魏志华, 黄孝伦, 刘亮, 史林霞, WEI Zhihua, HUANG Xiaolun, LIU Liang, SHI Linxia 基于对称性的HTML到XML的转换方法\[期刊论文\]-武汉理工大学学报\(信息与管理工程版\)](#) 2007, 29(7)
4. [程伟华 权限表达语言在数字内容递送中的应用\[学位论文\]](#) 2003
5. [夏立民, 王华, XIA Li-min, WANG Hua 基于VML的矢量图形动态生成过程的研究\[期刊论文\]-计算机技术与发展](#) 2006, 16(11)
6. [柳翠寅, 刘霞, Liu Cuiyin, Liu Xia XML签名技术的研究与应用\[期刊论文\]-计算机应用与软件](#) 2007, 24(4)
7. [孙雄英, 郭颂, SUN Xiong-ying, Guo Song 基于XML安全技术的算法实现\[期刊论文\]-微机发展](#) 2005, 15(7)
8. [庞闻 XML安全机制在数字化校园中的应用\[学位论文\]](#) 2006
9. [乔航, 冯梦舟, Qiao Huang, Feng Munzhuo 教育资源标准化中的XML数据转换方法\[期刊论文\]-情报杂志](#) 2006, 25(8)
10. [王化宇, Wang Huayu HTML语言在线教学系统的设计\[期刊论文\]-呼伦贝尔学院学报](#) 2011, 19(2)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_jsjyszgc200901010.aspx