基于XML的Web数据抽取研究

吕 锋,余 丽
(武汉理工大学,湖北 武汉 430070)

摘 要:文中介绍了三种常用的Web数据抽取的方法:直接解析HTML文档的方法,基于XML的方法(也称作为分析HTML层次结构的方法)以及基于概念建模的方法。重点研究其中的基于XML的数据抽取方法,基本做法是将原始的HTML文档通过一个过滤器检查并修改HTML文档的语法结构,从而形成一篇基于XML的XHTML,然后利用XML工具来处理这些HTML文档。实现了从非结构化的HTML文档向结构化的XML文档转化的预处理过程,给在Web挖掘中使用传统的数据抽取方法进行数据抽取创造了有利条件。

关键词:XML;Web;数据抽取
中图分类号:TP274^{+}.2 **文献标识码**:A **文章编号**:1673-629X(2007)06-0053-03

Study on Web Data Extraction Based on XML

LÜ Feng, YU Li
(Wuhan University of Technology, Wuhan 430070, China)

Abstract: Introduces three common methods for Web data extraction: method that directly analyses HTML document, method that bases on XML(it is also called method that analyses the structure of HTML document) and conceptual-model-based approach. especially, Web data extraction based on XML is studied. The original HTML document gets through a filter which checks and corrects the syntax structure of HTML document, then forms an well-formed XHTML, XML stools can be used to dispose these HTML documents. Implemented a data preprocessing which transformed the semi-structured HTML document to the structured XML document. Also it created a good condition of using the traditional data extraction methods to deeply data extraction.

Key words: XML; Web; data extraction

0 引 言

随着Internet及其相关技术的飞速发展,信息的发布与传播变得非常简便和迅速。Web也因此成为人们最大的信息和知识来源,堪称有史以来最为成功的网上"百科全书"。然而,网上信息浩如烟海,获取有用的信息好比大海捞针,以至于人们越来越依赖于搜索服务。搜索服务只是对用户获取和集Web上的信息方面提供了有限的能力,但人们却有多方面的关于Web信息获取的需求,例如:希望得到相关信息的一个集成视图,同时消除信息中的异构(heterogeneity)数据。希望得到相关信息的一个即时视图。希望得到相关信息的结构化拷贝,以便存储和进行结构化查询。希望能够利用自动化工具——例如智能代理程序来帮助处理信息。

XML的出现为解决Web数据挖掘中异构数据的难题带来了机会。由于XML能够使不同来源的结构化的数据很容易地结合在一起,因而使搜索多样的不兼容的数据库能够成为可能,从而为解决Web数据挖掘难题带来了希望。XML的扩展性和灵活性允许XML描述不同种类应用软件中的数据,从而能描述搜集的Web页中的数据记录。同时,由于基于XML的数据是自我描述的,数据不需要有内部描述就能被交换和处理。作为表示结构化数据的一个工业标准,XML为组织、软件开发者、Web站点和终端使用者提供了许多有利条件。

1 Web数据抽取的方法

Web数据抽取方法取决于Wrapper的构造机制。根据Wrapper中数据抽取器的实际运行机制,可以将目前主要的Web数据抽取方法归为三类[1]:

(1)直接解析HTML文档的方法。

该方法利用Perl, Java, YACC, Phyon等程序语言

收稿日期:2006-09-16
基金项目:教育部重点实验室开放研究基金(TKLJ0203)
作者简介:吕 锋(1957-),男,山东滨州人,教授,研究方向为计算机网络通信、信息系统与信息安全技术、计算机控制与仿真、灰色系统理论与应用。

或其他自行设计的程序语言,编写可执行程序直接对 HTML 网页进行分析和处理。这种方法主要利用规则表达式对内容进行模式匹配,不涉及 HTML 文档的层次结构。这种方法的缺点主要的是程序的健壮性和可维护性较差。因为抽取规则固化在程序中,一旦网页内容和结构发生变化,就必须对 Wrapper 进行重新设计。后来出现的一些 Web 数据抽取方法中引入了规则文件的概念。抽取逻辑从程序中被分离出来放入规则文件中,一旦结构发生变动,或者需要抽取同类网页数据,只需改写规则文件。这在很大程度上弥补了上面谈到的缺陷。规则文件以各种形式存在,例如描述文件(specification files)、XSLT 文件、DEL 脚本文件等。

(2)分析 HTML 层次结构的方法。

这种方法主要是利用 XML 技术,因此也称作基于 XML 的方法。随着 XML 技术的出现,XML 已成为 Web 上重要的数据表示和交换标准。因此,Web 数据抽取就不能不考虑到 Web 上将会出现大量的 XML 文档和利用强大且日益成熟的 XML 技术。基于 XML 的 Web 数据抽取也已经成为一种趋势。该方法首先将 HTML 文档根据 DOM 转换为一棵具有层次结构的 HTML 树。基本做法是将原始的 HTML 文档通过一个过滤器(filter),该过滤器检查并修改 HTML 文档的语法结构,从而形成一篇良构(well - formed)的 HTML 文档,即 XHTML。由于 XHTML 是基于 XML 的,因此下一步就可以利用 XML 工具来处理这些 HTML 文档。

(3)基于概念建模的方法(Conceptual - Model - Based Approach)。

该方法主要基于 Ontology 概念。BYU Data Extraction Group 对此进行了大量研究。该方法先用 Ontology 建立数据模型,再把可能抽取的数据项映射到 Ontology 中的元素上,用户选择 ontology 中的元素以决定抽取的对象。Ontology 的引入既保证了结构的一致性,又保证了数据的一致性,使不同来源的数据都能以统一的视图呈现,方便了信息的继承和交换。

2 DOM 概述

2.1 DOM 的定义

基于 XML 的数据抽取方法首先将 HTML 文档根据 DOM 转换为一棵具有层次结构的 HTML 树。所谓的 DOM 就是文档对象模型(Document Object Model)[2],是一种供 XHTML 和 XML 文档使用的应用程序编程接口(API)[3],它定义文档的逻辑结构以及访问和操作文档中各个部分的标准方法。对于符合语法要求的文档,DOM 是以树的结构形式进行存储和处理的,它是 XML 文档的逻辑表示,DOM 树中的基本构成单一位为节点(Node),节点对象通过逐层间的继承关系形成一棵 DOM 树,它包含了 XML 文档的全部信息。DOM 的核心是一组基本的程序接口,利用这些接口,程序可以访问和维护已经解析过的 XML 或 XHTML 文档。

简单地说,DOM 就是一组对象的集合,通过操纵这些对象,就能操纵 XML 和 HTML 数据。它是一个与语言无关的接口,应用通过这个接口来和 XML 或 HTML 内的数据打交道。无论是在浏览器里,还是在浏览器外,在服务器上,还是客户端,只要用 XML 就会碰到 DOM。

2.2 DOM 的整体结构图

DOM 的整体结构图如图 1 所示。

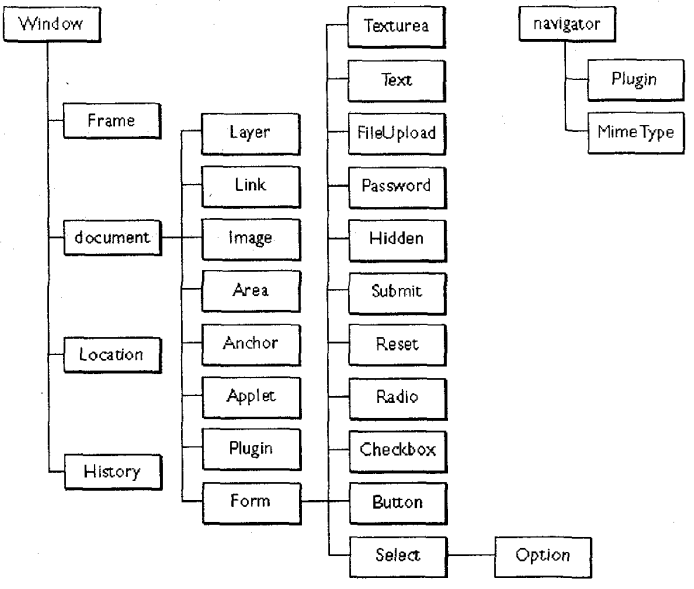

图 1 DOM 的一个整体结构图

2.3 DOM 的三部分

目前的 DOM 分为 核心(core)、HTML、XML 三部分。核心是结构化文档比较底层的对象的集合,但它们就已经可以表达出任何 HTML 和 XML 文件了。HTML 和 XML 两部分是专为 XML 和 HTML 提供的另外的高级接口,使操纵 HTML 和 XML 更方便。

3 非结构化的 HTML 文档向结构化的 XML 文档转化的过程

选取一个 HTML 语言编写的网页,网址 http://

ie. zzu. edu. cn/jiaoxue/asp/2/2 - 25. htm,网页快照如图 2 所示。

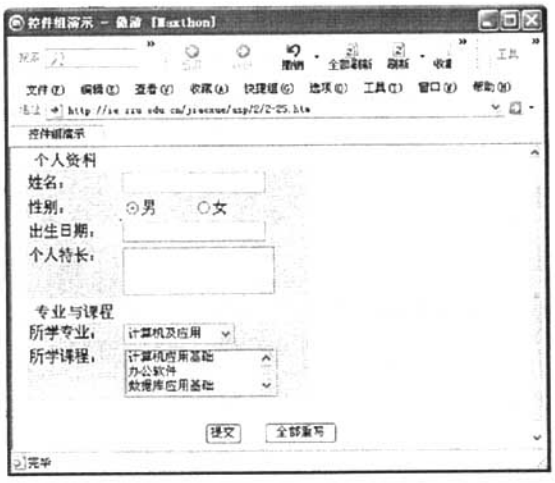

图 2　网页快照

此网页的 HTML 源代码如图 3 所示。

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<HTML xmlns="http://www.w3.org/1999/xhtml" lang="zh-CN">
<HEAD>
<TITLE>控件组演示</TITLE>
</HEAD>
<BODY>
<FORM>
<FIELDSET>
<LEGEND><B>个人资料</B></LEGEND>
<TABLE BGCOLOR = "#D6D3CE" WIDTH = "300">
<TR>
<TD>姓名:</TD>
<TD><INPUT TYPE = "text" NAME = "T1" SIZE = "20"></TD>
</TR>
<TR>
<TD>性别:</TD>
<TD><INPUT TYPE = "radio" NAME = "R1" CHECKED VALUE = "男">男
   
<INPUT TYPE = "radio" NAME = "R1" VALUE = "女">女</TD>
</TR>
<TR><TD>出生日期:</TD>
<TD><INPUT TYPE = "text" NAME = "T2" SIZE = "20"></TD>
</TR>
<TR>
<TD VALIGN = "top">个人特长:</TD>
<TD><TEXTAREA ROWS = "3" NAME = "S1" COLS = "20"></TEXTAREA></TD>
</TR>
</TABLE>
</FIELDSET>
```

图 3　HTML 源代码

利用 HTML Tidy 工具处理以上 HTML 源代码。输入的转换命令为:

Tidy - asxhtml index. html - big5 index. html

其中 - asxhtml 参数的意思是将 HTML 转换成符合标准的 XHTML。- big5 是指以 big 编码输入和输出文档,- gb2312 是指以 gb2312 编码输入和输出文档。还有更多的参数可以使用,可以输入 tidy - help(或者 - h)查看帮助信息。

转换后得到的 XHTML 代码如图 4 所示。

由于 XHTML 是基于 XML 的,因此下一步就可以利用 XML 数据抽取工具进行处理了[4]。

4　结束语

由于 Web 的存在,数据量在爆炸式地不断增长。将传统的数据抽取方法应用到 Web 上成为近年来讨论的焦点。Web 数据抽取就是利用数据挖掘技术从网络文档和服务中发现和提取信息。Web 上各种形式的文档和用户访问信息就构成了 Web 数据抽取的对象。Web 上的数据具有一定的结构性,但不同于传统数据库的结构化数据,它们没有特定的数据模型,数据本身具有自我描述、动态可变和半结构化的性质。XML(eXtensible Markup Language)对半结构化数据提供了良好的支持,它以标记文本格式存放数据,成为 Web 数据交换的标准[5]。以 XML 家族为基础的新一代的 WWW 环境是直接面对 Web 数据的。不仅可以很好地兼容原有的 Web 应用,而且可以很好地实现 WWW 这一分布计算环境下的信息共享与交换。因此,它已经成为 Web 信息发展的可喜的趋势。

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN">
<html lang="zh-CN">
<head>
<meta name="generator" content=
"HTML Tidy for Windows (vers 14 February 2006),see www.w3.org">
<title>控件组演示</title>
</head>
<body>
<form>
<fieldset><legend><b>个人资料</b></legend>
<table bgcolor="#D6D3CE" width="300">
<tr>
<td> 姓名:</td>
<td><input type="text" name="T1" size="20"></td>
</tr>
<tr>
<td>性别:</td>
<td><input type="radio" name="R1" checked value="男">男
   <input type="radio" name="R1" checked value="女">女
</td>
</tr>
<tr>
<td>出生日期:</td>
<td><input type="text" name="T2" size="20"></td>
</tr>
<tr>
<td valign="top">个人特长:</td>
<td>
<textarea rows="3" name="S1" cols="20">
```

图 4　转换后的 XHTML 文档

参考文献:

[1] 杨　鲲,孟　波. 一种基于 XML 的 Web 数据挖掘方法[J]. 计算机应用,2003,23(6):160 - 164.

[2] 欧建雄,张礼平. HTML 数据内容的抽取与集成[J]. 华东理工大学学报,2003,29(6):613 - 616.

[3] 刘晓鹏,邢长征. 基于 WEB 文本数据挖掘的研究[J]. 计算机与数字工程,2005,33(9):75 - 79.

[4] 王建丽,丁振国. 一种基于 XML 的 Web 数据挖掘技术[J]. 西安科技学院学报,2002,22(3):337 - 340.

[5] 沈　洁,薛贵荣. 一种基于 XML 的 Web 数据挖掘模型[J]. 系统工程理论与实践,2002(9):74 - 77.