HTML到XML转换技术的研究与实现

陈艳梅 张 斌

（东北大学图书馆 沈阳 110004）（东北大学信息与工程学院 沈阳 110004）

【摘要】 网络上大多数的信息都是用HTML写的，这种语言不能处理网络上的很多需求，因为它只是一种用于浏览信息的语言，不能表达数据本身，网络还没有形成一个良好的结构化文档的存贮，而只是一个可变的HTML页的聚集，我们迫切希望来自网络资源的信息以一种结构化的方式来存贮。XML和它的各种扩展功能如数据模型、查询语言等是实现结构化方式的一种，是一种元语言，可以弥补很多HTML的不足。未来的网页会使用具有很好结构化的XML语言，但是现在这一阶段是过渡阶段，必须思考一种方法来实现HTML到XML的转换，以更好地利用网络资源。本文提出了一种实现HTML到XML转换的方法。

【关键词】 包装器 信息抽取 HTML解析 HTML－XML转换技术 【分类号】 TP39

The Research and Realization of Technology Converting HTML to XML

Chen Yanmei

(*Northeastern University Library, Shenyang 110004, China*)

Zhang Bin

(*Information Engineering Institute of Northeastern University, Shenyang 110004, China*)

【Abstract】 Nowadays, the whole world can possibly communicate with all different people by using web. Internet usually uses HTML, it cannot handle the various requirement of Internet and also express the data itself. To do so, information from web sources needs to be accessible in a structured way. XML and its various extensions are a step in this direction. Unfortunately, the web is not yet a well organized repository of nicely structured XML documents but rather a conglomerate of volatile HTML pages, for which structure has to be extracted. This thesis shows the design and implementation of a conversion system of HTML to XML

【Keywords】 Web wrapper Information extraction HTML parsing HTML to XML conversion

1 前 言

现在网络上大容量的、有用的、有价值的基于HTML的信息均可以被设计给人们浏览，但是对于这些面向用户的HTML页，很多的程序很难解析和捕获。能否研究一种技术，通过对Web上这些分布、异构的信息源实施有效抽取，集成，将它们的分布性和异构性屏蔽起来，向用户提供一致的数据界面和高效、简便的查询服务呢？基于Internet的信息集成技术是在这一目标推动下发展起来的。本文中提出了一个自适应性的方法来构建一个交互式的系统，即半自动化包装器（HTML－XML包装器），其目标是实现HTML到XML的半自动转换。

2 HTML－XML转换包装器系统设计

2.1 HTML－XML转换包装器的系统设计思想

随着Web上信息资源的不断增加，人们已不再仅仅满足于对信息的简单利用，需要对一些信息进行再加工处理，以满足日益增长的进一步需要。通过对Web上信息的组织结构的分析和研究，发现HTML文档已经不适应新的发展需求。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。我们设计的HTML－XML包装器以说明性方式对网页执行透明性的检索处理，从HTML源抽取信息，向目标格式作映射以备将来应用。我们的工具包可以提供集成和转换的方法，实现HTML文档到XML文档的转换。描述了怎样访问资源，抽取哪部分信息，使用何种结构，生成一些Java代码，执行HTTP请求，进行抽取，返回结构化的结果。本包装器完全是用Java程序编写的，也可以用来构建WebAPI，也可以从其它的Java程序中进行调用。

2.2 HTML－XML包装器的软件结构

一个包装器是一个软件组件库，可以将一种模式的数据转换成另一种模式。在网络环境下，包装器的角色是将以HTML文档方式存贮的信息，主要由纯文本加上一些标记组

收稿日期：2003－03－19

成，转换成以某种数据结构存贮的信息。我们设计的包装器主要思路是首先用HTML句法规范器对网页上的HTML文档进行句法规范，形成规范化的HTML文档，再对此HTML规范化的文档进行解析，形成基于DOM的解析树，形成节点信息，对这些节点信息再进行抽取，抽取出来的信息被存为嵌入结构列表NSL形式，形成一系列抽取规则。映射模块主要负责将被提取的信息转换成满足用户需要的数据结构，主要说明将对应的NSL信息映射到XML形式的数据。HTML－XML包装器主要包括六个软件结构模块，即HTML句法规范器，HTML解析器，HEL信息抽取器，嵌入结构列表生成器，XML模板映射器及XML信息生成器等。如图1所示：

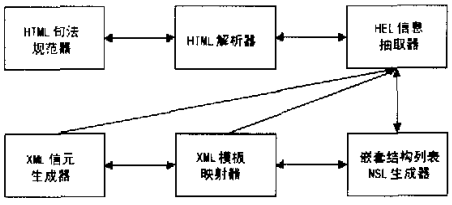

图1 HTML－XML包装器软件结构图

3 HTML－XML转换包装器实现

我们的HTML－XML包装器整个过程是这样的：

(1) 从网络数据资源取一个HTML文档的内容，对该HTML文档进行规范化检验，我们主要是通过HTML Tidy工具的使用来进行验证，对不正确的HTML句法进行修复，若是还不符合用户的需求或是还不规范，则可以请用户参与进行修复，以便得到一规范化的HTML文档。我们的HTML－XML包装器的实现就是基于规范化的HTML文档的。

(2) 我们使用HTML Parser即HTML解析器对这规范化HTML文档进行解析。形成一棵基于文档对象模型即DOM的解析树，树中形象地表示出了HTML各个节点的层次关系及各自的属性及相应的文本值。

(3) 对HTML解析树进行遍历，因为HTML解析树揭示了HTML文档的层次结构，我们可以对HTML文档进行深度优先遍历，即沿着文档流的方向进行遍历。从而可以到达HTML的任一个节点。

(4) 我们对各个节点信息进行抽取，采用HEL，即HTML抽取语言，来生成一些相应的抽取规则，来实现对各个节点信息的抽取。若抽取不完全，即没有将所有的节点信息抽取出来，则重新进行抽取。

(5) 抽取出来的节点的信息，以嵌入结构列表的形式即NSL的形式来进行存贮。

(6) 如果NSL对应着简单的数据类型，即string，int，float等，我们事先设计了相应的XML映射模板，可以直接通过这些映射模板映射成XML信息。而对于其它的复杂的数据类型，可以由用户生成相应的XML模板，来实现到XML信息的转换。在此NSL到XML信息的映射过程中生成相应的映射规则，存入映射规则库中

(7) 在生成XML信息的同时，可相应地生成文档类型定义DTD。通过以上七个步骤，我们的HTML－XML包装器就可以实现HTML到XML的转换。

具体的HTML－XML包装器的整个的系统流程图如图2所示：

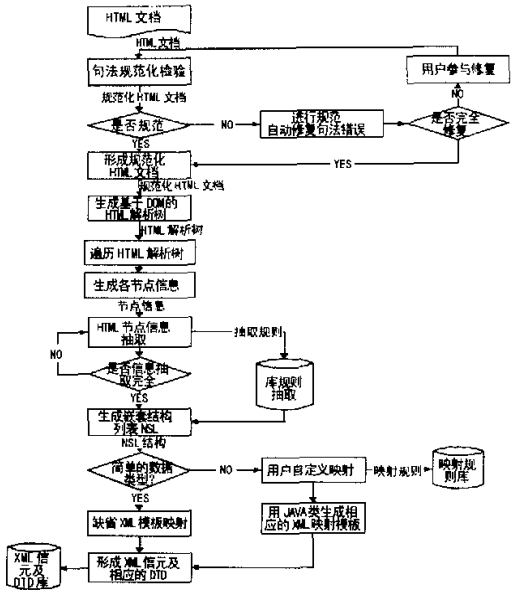

图2 HTML－XML包装器的系统流程图

4 结束语

任何一项新技术的产生都是有其需求背景的，XML是在HTML遇到不可克服的困难之后诞生的。近年来HTML在许多复杂的Web应用中遇到了难题，要彻底解决这些难题，必须用功能强大的XML来替代HTML作为Web页面的书写出工具。XML有利于信息的表达和结构化组织，从而使数据搜索更有效；XML可以使用URL别名，使Web的维护更方便，也使Web的应用更稳定；XML可以使用数字签名，使Web的应用更广阔地拓展到安全保密领域。可以认为：未来绝大部分的Web书写工具必定是XML。而XML广泛使用必然能推动Web不断发展，从而开创Web应用的新时代。目前关于HTML－XML转换的文章较少，且这部分的研究还有待进一步深入，希望更多的同行加入到这个研究队伍中，来更好地解决HTML－XML的转换问题，使我们的网页更好地提供信息，给我们使用网页带来更大的便利。

参考文献：

[1] Ling Liu, Calton Pu, Wei Han, XWRAP: an XML－enabled wrapper construction system for web information sources [J]. 2000 IEEE on data engineering

[2] S. Abitebonl, D. Quass, J. McHugh, J. Widom, and J. L. Wiener. The Lorel Query Language for Semistructured Data[J]. Journal on Digital Libraries, 1997

[3] Brad Adelberg. XoDoSe－A Tool for SemiAutomatically Extracting Semi－Structured Data from Text [J]. In Proc. Of the SIGMOD Conference, Seattle, June 1998

(下转第90页)

购置了80台IBM PII200、64M SDRAM的二手计算机，由于没有光驱、软驱、硬盘，节省了大量的费用，通过市场考察，购置了80块CENTERM－2100 Windows终端仿真卡，该终端防真卡支持ICA(独立计算结构)协议，从卡中内嵌系统直接启动可以做Windows终端使用。配置两台服务器，每台服务器配置双PⅢ733E、512ECC SDRAM内存、18G SCSI硬盘、双10Mbit/S/100Mbit/S服务器网卡，操作系统采用Windows 2000 SERVER，加装CITRIX公司提供的ICA协议及META FRAME FOR Windows 2000 V1.8软件，以便终端能够访问终端服务器，终端服务器通过专线连接Internet。通过访问终端服务器，运行Windows Terminal以及Windows 2000 Server，可以访问终端服务器的信息资源，也可以进入Internet，最大限度节约构建、升级，更新操作系统和软件投资，有效延长了旧PC机的生命周期。由于终端没有本地的存储设备，可有效的防范好动的孩子或对计算机不太熟悉的人操作错误损坏机器，防范病毒和游戏等不允许使用的软件从客户机侵入系统。使网络的安全性进一步加强。

4 采用Thin－Client/Server方案的优越性

(1) 可有效的降低总拥有成本(TCO)：TCO(Total of Ownership)为构建信息系统总拥有成本。以Windows 2000的终端服务和Windows终端组成的Thin－Client/Server网络计算体系，是Windows 2000系统降低TCO的最新途径。据调查结果显示，构建一个Thin－Client/Server计算体系，比以PC/PC－Server方式构建一个同样功能的系统节约20%左右的系统投资。

(2) 该系统对客户端设备要求低，可以利用老的386/486机器配置客户端。终端要求的网络带宽低(每台终端机通常只需要零点几K，而PC机则需要几M)。有效的防止网络"堵车"现象。

(3) 系统和软件的安装方便，系统和应用软件只安装在服务器端，管理员可以在服务器上为每个用户配置好系统和应用软件，供给所有的终端用户使用，终端用户不用进行任何安装调试。

(4) 系统软件、硬件的升级只考虑服务器，所有的软件升级在服务器端一次完成。作为客户端设备面临的升级压力极小，甚至根本谈不上硬件升级，客户端设备不会因为服务器软硬件的升级而频繁遭到淘汰。

(5) 客户端桌面与服务器界面一样。软件100%在终端服务器上运行，客户端用户完全可以像使用本地系统和软件一样使用服务器、网络上和终端本地的软硬件资源，与使用普通PC完全一样。

(6) 系统使用RDP协议或ICA协议，系统管理员可以通过管理员权限在任意一台终端客户服务端来管理服务器。

(7) 系统安全可靠，客户端没有软驱、硬盘等存储设备，可有效的防范数据的丢失。有效防范病毒从客户机进入网络。服务器端的NTFS文件系统，可有效防止非法访问系统和其他用户的的数据。管理员可以有效的监控并记录用户的使用状态。Windows 2000中的磁盘配额管理可防止用户占用过多的公共资源。

(8) Windows 2000支持多客户端同时登录服务，由于服务器为每一个终端客户分配独立的内存空间，不同终端拥护在服务器上所使用的内存空间，以及中断和系统用户之间的通信都是相互独立的，每个用户就像单独使用一台服务器，其它用户并不存在一样。一个终端客户机由于误操作引起的死机，不会影响其它终端客户机。

(9) 由于终端客户机本身很大程度上不易因环境和人为原因损坏，从一定程度上减轻管理员对客户端设备维修、管理上的压力，大大降低了系统的维护管理成本，提高了设备的利用率。

5 结束语

Thin－Client/Server计算模式在社区图书馆建立电子阅览室中，节省了大量的经费，取得了明显的社会效益和经济效益。这是一种集中式管理、只在服务器端进行软硬件升级、进行安全应用配置的全新的管理模式。

中文Windows 2000 Server的发布加速了国内的终端市场的进一步发展。Thin－Client/Server这种终端应用模式得到了众多厂家的支持，如Wyse、NCD、Boundless、Tektronix、IBM等。随着对终端服务技术的进一步了解，终端产品的更加成熟，相信这种应用模式在不久的将来会被越来越多的社区图书馆采用。

参考文献

[1] Windows 2000 Server，终端服务技术概览，Microsoft公司技术资料，1999，9

(上接第67页)

[4] Gustavo Arocena and Alberto Mendelzon. WebOQL：Restructuring Documents，Databases，and Webs [J]. In Proc. ICDE'98，Orlando，February 1998

[5] Jean－Robert Gruser，Louiqa Raschid，M. E. Vidal and L. Bright. Wrapper Generation for Web Accessible Data Sources [J]. In COOPIS，1998

[6] J. Hammer，H. Garcia－Molina，J. Cho，R. Aranba，and A. Crespo. Extracting Semistructured Information from the Web [J]. In Proceedings of the Workshop on Management of Semistructured Data. Tueson，Arizona，May 1997

[7] Gerald Huck，Peter Fankhauser，Karl Aberer，and Erich J. Neuhold. JEDI：Extracting and Synthesizing Information from the Web [J]. In COOPOIS，New－York，1998

[8] Mary Tork Roth and Peter Schwartz. A Wrapper Architecture for Legacy Data Sources [J]. Technical Report RJ10077，IBM Almaden Research Center，1997

[9] World Wide Web Consortium(W3C). The Document Object Model，1998. http://www.w3.org/DOM

[10] Jon Bosak. XML，Java and the Future of the Web[J] http://sunsite.unc.edu/pub/sun－info/standards/xml/why/xmlapps.html

HTML到XML转换技术的研究与实现

作者： 陈艳梅， 张斌
作者单位： 陈艳梅(东北大学图书馆,沈阳,110004)， 张斌(东北大学信息与工程学院,沈阳,110004)
刊名： 现代图书情报技术 PKU CSSCI
英文刊名： NEW TECHNOLOGY OF LIBRARY AND INFORMATION SERVICE
年，卷(期)： 2003(5)
被引用次数： 2次

参考文献(10条)

1. Ling Liu;Calton Pu;Wei Han XWRAP: an XML - enabled wrapper construction system for web information sources 2000

2. S Abitebonl;D Quass;J McHugh;J Widom, and J L Wiener The Lorel Query Language for Semistructured Data 1997

3. Brad Adelberg;XoDoSe-A Tool for SemiAutomatically Extracting Semi- Structured Data from Text 1998

4. GUSTAVO AROCENA;Alberto Mendelzon WebOQL:Restructuring Documents Databases and Webs 1998

5. Jean-Robert Gruser;Louiqa Raschid;M. E.Vidal Bright.Wrapper Generation for Web Accessible Data Sources 1998

6. J.Hammer;H Garcia- Molina;J.Cho;R Aranba,and A Crespo.Extracting Semistructured Information from the Web 1997

7. Gerald Huck;Peter Fankhauser;Karl Aberer;Erich J Neuhold JEDI : Extracting and Synthesizing Information from the Web 1998

8. Mary Tork Roth;Peter Schwartz A Wrapper Architecture for Legacy Data Sources 1997

9. World Wide Web Consortium The Document Object Model 1998

10. Jon Bosak;XML Java and the Future of the Web

本文读者也读过(10条)

1. 詹志飞 XML技术及应用解析[期刊论文]-科学咨询2009(15)

2. 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)

3. 黄晓斌 HTML向XML转换的研究[期刊论文]-现代图书情报技术2003(1)

4. 王伟 标记语言及HTML和XML的比较分析[期刊论文]-现代图书情报技术2000(5)

5. 黄冠能 HTML文档信息抽取及语音再表达的研究与实现[学位论文]2007

6. 贾长云.程永上.JIA Chang-yun.CHENG Yong-shang HTML表格向XML的智能转换[期刊论文]-计算机工程2009,35(14)

7. 黄伟.刘娟.HUANG Wei.LIU Juan 一种基于DOM树的HTML转换为XML的方法[期刊论文]-电脑知识与技术（学术交流）2006(7)

8. 戴怡钧 HTML转换到XML格式以及不同XML标准格式之间的转换[学位论文]2003

9. 刘芳.卢正鼎.LIU Fang.LU Zheng-ding 有效地检索HTML文档[期刊论文]-小型微型计算机系统2000,21(9)

10. 李晓溪.王异 基于HTMLParser的HTML解析研究[期刊论文]-网络财富2009(8)

引证文献(2条)

1. 陆伟.寇广增.魏泉 Web环境下的内容抽取及RSS发布[期刊论文]-情报杂志 2005(9)

2.吴治宗 基于XML的文档处理技术的研究与实现[学位论文]硕士 2006