

面向现代汉语动态流通语料库的 html To xml 转换工具的设计和实现

唐长宁, 张志平, 赵小兵

(内蒙古师范大学 计算机与信息工程学院, 内蒙古 呼和浩特 010022)

摘 要: 分析了目前 HTML 与 XML 格式的特点及现有的 html to xml 转换软件的不足, 给出面向现代汉语动态流通语料库的 html To xml 软件的设计和实现过程. 编码过程利用面向对象思想, 使用 Java 编程语言, 可以实现跨平台运行. 用测试用例对软件做了相应测试, 达到了预期效果.

关键词: html to xml 转换软件; 动态语料库; XML; HTML

中图分类号: TP 391.2 **文献标识码:** A **文章编号:** 1001-8735(2008)01-0063-04

随着互联网应用需求及其相关支撑技术的发展, XML(eXtensible Markup Language)已经成为互联网环境中数据描述和网上应用系统间数据交换事实上的标准^[1]. XML 是一种元标记语言, 用户可以定义自己需要的标记. 它提供了描述结构化数据的格式, 可以通过独立运行的方法来共享数据. 与 HTML 相比, XML 具有内容与形式相分离的特性, 以及良好的可扩展性、跨平台移植性和自描述性等特性. 当前 Web 信息大多数都是 HTML(Hyper Text Markup Language)格式, 由于具有简单、易用等特点, 所以目前被广为接受. 尽管作为信息的主要载体, HTML 提供了一种能方便地向读者呈现信息的方法, 但它可能并不是一个很好的自动提取与数据驱动服务或相关应用程序的信息机构.

动态语料库(Dynamic Circulating Corpus, 简称 DCC)是历时语料库, 它与静态语料库和共时语料库是相对而言的. 这种语料库可以对语言的变化进行检测和监测^[2], 通过对语料库的分析, 可以观察到语言现象的发生、发展和消亡. 从 2001 年开始, 北京语言大学的现代汉语动态流通语料库已收集整理了 15 家主流报纸的语料, 约 15 亿的语料, 获得的初始语料是 Html 格式的, 每年都要追加新的语料. 对于我们的研究而言, 在 Html 格式的文件中存在大量的垃圾信息(Html 脚本语言、广告内容等), 而垃圾信息过滤的质量直接影响对话料计算的精确度和研究的可信度. 语料库数据量庞大, 采用人工去除垃圾信息是不可取的, 所以需要自动转换软件实现 Html 格式文件到 xml 格式文件的转换, 而且能够把原始语料中的垃圾信息去掉.

1 现有 html to xml 转换软件的不足

目前直接把 html 格式文件转换为 xml 格式的免费软件很少, 这样的软件都是科研部门内部开发的, 主要用于内部研究, 而且存在的缺点也很明显, 只能够去除 html 格式文件中的部分垃圾信息. 图 1 和图 2 是一个免费转换软件转换前后的结果. 从文本格式转换后得到的结果可以看到以下一些特点: ① 能够保留网页的部分重要信息; ② 基本能保留原来网页中的段落格式; ③ 特殊的转义字符(如 >)不能去掉; (4) 转换后的文中存在无关的文本内容(如“图片桌面”).

2 转换软件的设计和实现

2.1 转换软件的要求 为以后研究的需要, 在文件格式转换时, 我们将 HTML\HML 网页格式的文件中所带的文件相关信息, 如“媒体名称(mediaName)”、“文件栏目(nodeName)”、“文件标题(title)”、“作者(authorName)”

收稿日期: 2007-06-22

基金项目: 国家自然科学基金资助项目(60663008)

作者简介: 唐长宁(1980-), 男, 辽宁省丹东市人, 内蒙古师范大学硕士研究生; 赵小兵(1967-), 女, 内蒙古呼和浩特市人, 内蒙古师范大学教授, 主要从事自然语言信息处理技术研究.

thor)、“发布时间(publishTime)”“正文内容(content)”等与该文档相关的信息尽可能同时提出,并按照 XML 格式写入纯文本文件中。根据网页的内容和实验的要求,转换后的文件格式如图 3 所示。

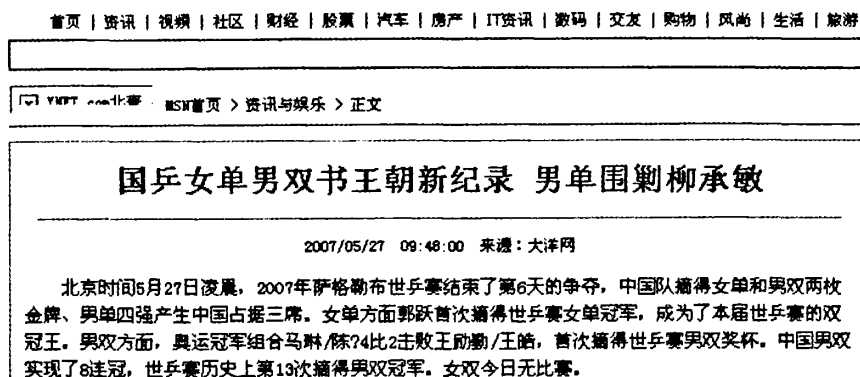


图 1 文本格式转换前的内容

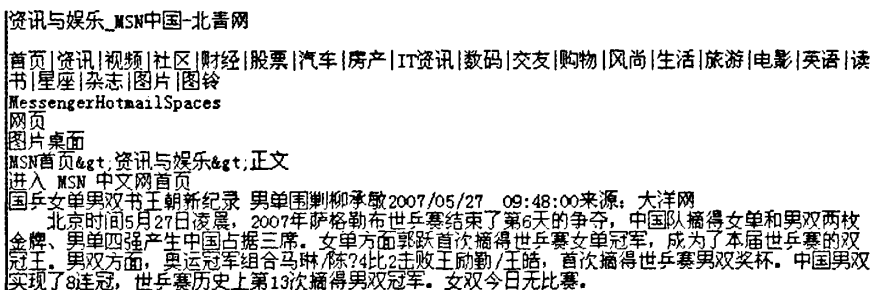


图 2 文本格式转换后得到的结果

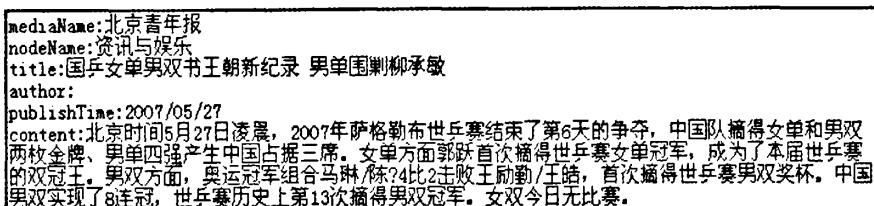


图 3 符合要求的文本格式转换后得到的结果

2.2 转换软件的设计 格式转换中需要考虑的问题有:① 不同类报纸的信息提取标志一般是不同的;② 同一类报纸在不同时间段,信息提取标志会有所改变;③ 语料扩充,增加新的报纸种类。

鉴于以上考虑,为了保证程序在后期使用中不修改,我们采用了 Hibernate 的配置文件方法^[3],把信息提取的标志写到一个固定的文件中,不论是将来语料的扩充,还是不同时间段信息提取标志的改变,只需修改配置文件就可以了,配置文件的格式如图 4 所示。

在图 4 的信息中,第 1 行中的数字表明语料库中报纸的类数,数字要和后面给出的媒体的个数一致。图中给出的是配置文件的一部分,实际配置文件中还有北京青年报、人民日报、环球时报、法制日报、羊城晚报、北京晚报等 6 类报纸,所以我们看到图中的数据是 6。第 2 行到第 7 行给出媒体的名称,以及 nodeName、title、author、publishTime、content 的开始和结束标志。有用信息都以 # 作为开始和结束的标志,例如数字 6 在图中是“#6#”。一类信息要求占用一行。

2.3 转换软件的实现 (1) 面向内容的 HTML 分析。HTML 由内容及这些内容的呈现格式数据组成。内容是有特定逻辑语义的数据,是到 XML 机构化文档的转换数据,而各式数据主要描述内容如何表达,通过

```

#6# /*数字6表示要处理的报纸有6类*/
#北京青年报# /*媒体的名称*/
#>nodeName:<#<node>#</node>#/*<node>和</node>是提取nodeName信息的开始和结束标志*/
#>title:<#<title>#</title>#/*title的开始和结束标志*/
#>author:<#<!--author>#</author>#/*author的开始和结束标志*/
#>publishTime:<#<!--timer>#</timer>#/*publishTime的开始和结束标志*/
#>content:<#<enpcontent>#</enpcontent>#/*content的开始和结束标志*/
#人民日报# /*媒体的名称*/

```

图4 配置文件的格式

一些有特定含义的标记及其属性值来格式化内容数据。HTML 中可以放置内容数据的有页面头(head)、段落(p)、图形(img)、表单(form)、表格(table)以及多窗口页面(frame)等^[4],但是从数据的原子性角度看,HTML 中的内容数据主要表现为文本、图形链接及文本链接。表单、表格及多窗口页面给出这些原子内容数据的组织形式。例如,表格的语法形式是<table>...</table>, <tr>用于定义表行, <th>用于定义表头, <td>用于定义表元。表头中的原子数据信息都在<th>和</th>之间,表格的具体数据都在<td>和</td>之间,根据这些规则可以界定表格中的内容数据。要实现 HTML 内容到 XML 数据的转换,关键是给出 HTML 的这些内容数据及其关系的一种组织方式,以及这种方式在 XML 模式中的相应表达规划。

(2) 标记规则。基于对 HTML 机构和语法特点的分析,首先定义一套标记规则用来提取 HTML 文档中含有特定含义的内容数据。利用该标记规则,可以在 HTML 页面数据对内容和呈现内容的格式数据进行分离,按照内容数据间本来的关系组织这些数据。在此基础上,软件设计了几个类,类中还设计了 FindContent 等方法来实现。标记规则也是转换后的 XML 结构数据所依赖的 XML 模式的直接构建依据。在标记规则作用下,HTML 源文件的标记分两个层次展开,一是根据页面不同部分本身

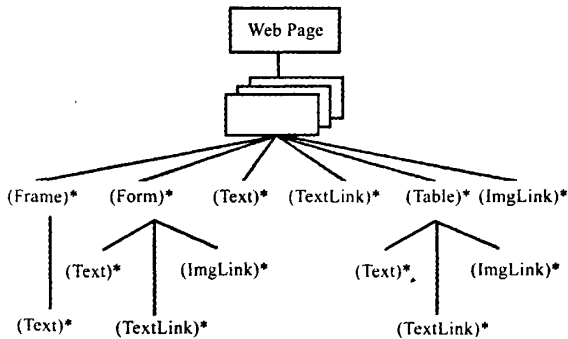


图5 标记规则层次图

语义的组织关系进行区域逻辑上的划分,二是在区域内对各种不同类型内容数据进行内容块的划分。从规则上讲,标记分为区域标记和内容标记,定义规则也适用于内容标记。内容标记是标记特定区域内某一类型内容块的标记,分为文本标记、文本链接标记、图形链接标记、表单标记、表格标记和多窗口标记。每种标记也有起始标记和结束标记。文本起始标记用<!--# \$ %TEXT表示区域标记序号,内容块标记序号用# \$ %-->格式表示。其中“TEXT”是文本标记的关键字。同样,文本链接标记、图形链接标记、表单标记、格式标记、多窗口标记所对应得关键字分别是“TEXT-LINK”、“IMG-LINK”、“FORM”、“TABLE”、“FRAME”。所有内容标记的内容块标记序号按出现的先后顺序统一编号,编号的范围是自然数。“区域标记序号内容块标记序号”唯一界定了特定区域内某一特定内容块,标记规则层次图如图5所示。

另外,软件还给出读取标志信息的程序代码。在程序中用到 Java 语言中 String 类的一个重要方法 String[] split(String regex)^[5],该方法能够把信息标志的前后标志分解出来。

2.4 软件测试 测试用例(Test Case)是为某个特殊目标而编制的一组测试输入、执行条件以及预期结果,以便测试某个程序路径或核实是否满足某个特定需求。测试用例可以分为基本事件、备选事件和异常事件。通过对表1中的测试计划进行多次软件测试,完全符合期望结果。期望结果是否有效的衡量标准主要依据图3的格式。转换后的格式如图3所示为有效,否则为无效。

表1

测试 数据	测试范围	期望结果
北京晚报的 html 格式文件	有效等价类	有效
人民日报的 html 格式文件	有效等价类	有效
环球日报的 html 格式文件	有效等价类	有效
人民日报的 doc 格式文件	无效等价类	无效
环球日报的 bmp 格式文件	无效等价类	无效

综上所述,使用面向现代汉语动态流通语料库的 html To xml 软件,可以有效地将 HTML 文件转换为 XML 格式,形成 XML 数据源。该软件克服了现有转换软件的不足,简化了信息提取工作,为处理 XML 数据,并检索出适当的数据做了有效的铺垫,为基于动态语料库的基本词汇的提取和通用词汇的提取提供了有效的帮助。本文的研究工作中还存在一定的问题,需要进一步完善和改进,例如在搜索、查找算法的选择上软件使用的是蛮力法,软件运行效率不高,另外在算法的选择上还有待于进一步提高。

参考文献:

- [1] Khare R, Rifkin A. XML: A Door to Automated Web Applications [J]. IEEE Internet Computing, 1997, 1(4): 78-87.
- [2] 张普. 关于大规模真实文本语料库的基点理论思考 [J]. 语言文字应用, 1999(1): 34-43.
- [3] 计磊, 李里, 周伟. 精通 J2EE Eclipse, Struts, Hibernate, Spring 整合应用案例 [M]. 北京: 人民邮电出版社, 2006: 47-48.
- [4] Dave Raggett. HTML 4.01 Specification W3C Recommendation [EB/OL]. [1999-09-24]. <http://www.w3.org/TR/1999/REC-html401-19991224>.
- [5] 郑阿奇. Java 实用教程 [M]. 北京: 电子工业出版社, 2005: 42-43.

The Exchange Software of the html to xml Base on Dynamic Circulating Corpus

TANG Chang-ning, ZHANG Zhi-ping, ZHAO Xiao-bing

(College of Computer and Information Engineering, Inner Mongolia Normal University, Huhhot 010022, China)

Abstract: The characteristics of HTML and XML were analyzed. Some disadvantages of the current html-xml exchange software were pointed out. Based on Dynamic Circulating Corpus, a new html-xml exchange software were designed and achieved with object oriented method, such as Java. It was shown that this software can run without single platform. Test Cases in the software do the corresponding tests to achieve the desired effect.

Key words: the exchange software of the html to xml; dynamic circulating corpus; XML; HTML

【责任编辑 陈汉忠】

(上接第 62 页)

Effects of Pulse Electric and Magnetic Field on GSH2px Activity and MDA Content

SHI Ji-fei


(Department of Medical Physics, Baotou Medical College, Baotou 014010, Inner Mongolia, China)

Abstract: The effects of pulse electric and magnetic field on the malonaldehyde (MDA) and glutathione peroxidase of hepatic tissue were studied. The magnetic induction intensity inside the coils and its rising time from zero Tesla to the maximum were calculated. Different pulse electric and magnetic field with averaged change rates of rising edge were applied to mice for 4 hours per day. After 15 days, the MDA and glutathione peroxidase of hepatic tissue were detected. Compared with control group, MDA content decreased and GSH2px activity increased obviously, when mice exposed in pulse electric and magnetic field ($P < 0.01$). Pulse electric and magnetic field can affect the lipid peroxidation of mouse hepatic tissue, improve the activity of antioxidant and decrease production of oxide.

Key words: Glutathione Peroxidase; Malonaldehyde(MDA); pulse electric and magnetic field

【责任编辑 陈汉忠】

面向现代汉语动态流通语料库的html To xml转换工具的设计和实现

作者：[唐长宁](#)，[张志平](#)，[赵小兵](#)，[TANG Chang-ning](#)，[ZHANG Zhi-ping](#)，[ZHAO Xiao-bing](#)
作者单位：[内蒙古师范大学, 计算机与信息工程学院, 内蒙古, 呼和浩特, 010022](#)
刊名：[内蒙古师范大学学报（自然科学汉文版）](#) 
英文刊名：[JOURNAL OF INNER MONGOLIA NORMAL UNIVERSITY \(NATURAL SCIENCE EDITION\)](#)
年，卷(期)：2008, 37 (1)

参考文献(5条)

1. [Khare R;Rifkin A](#) XML:A Door to Automated Web Applications[外文期刊] 1997 (04)
2. [张普](#) 关于大规模真实文本语料库的基点理论思考 1999 (01)
3. [计磊;李里;周伟](#) 精通J2EE Eclipse、Struts、Hibernate、Spring整合应用案例 2006
4. [Dave Raggett](#) HTML 4.01Specification W3C Recommendation 1999
5. [郑阿奇](#) Java实用教程 2005

本文链接：http://d.wanfangdata.com.cn/Periodical_nmgsdxb200801016.aspx