

一种基于 XML 的中间转码模型设计与仿真

李超建¹, 朱晓姝^{1,2}

(1 玉林师范学院 计算机科学与工程学院, 广西 玉林 537000;

2 桂林电子科技大学 计算机科学与工程学院, 广西 桂林 541004)

摘 要: WAP 的广泛应用, 使原有 HTML 需要转码为 WML, 以便移动无线终端浏览信息. 设计了一种基于 XML 的中间转换模型, 通过 XML 作为中间桥梁, 实现从 HTML 页面转换到 WML 页面. 并实现模型算法设计, 选择纯文本文件实例, 对页面转换进行仿真测试, 记录相关算法运行时间, 对模型性能进行分析, 仿真实验结果表明本模型转换时具有更高的准确性和效率.

关键词: WAP; XML; WML; 转码模型

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2013)10-0161-04

Design and Simulate of an Intermediate Transcoding Model Based on XML

LI Chao-jian¹, ZHU Xiao-shu^{1,2*}

(1 Computer Science Department, Yulin Normal University, Yulin 537000, China;

2 Computer Science Department, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: The development of WAP needs to transcode HTML pages into WML pages. As the meta-language of markup language, XML can be used to solve the transcoding problem as the form of intermediate conversion by using the flexibility of it. Firstly, it completes the transcoding work from HTML pages to XML pages, and then completes the transcoding work from XML pages to WML pages. It achieves the whole transcoding from HTML pages to WML pages via the bridge of XML. Finally, it is proved that the intermediate transcoding model have more accuracy and higher efficiency than other algorithm through experiment.

Key words: WAP; XML; WML; transcoding model

1 引言

无线应用协议(Wireless Application Protocol, WAP)是当前全球性的网络通信协议,在 GSM、CDMA、TDMA、3G 等多种网络中得到广泛应用,使用它可以将 Internet 中丰富的信息和先进的业务引入到移动电话等移动无线终端. 只要无线终端和代理服务器支持,在现有移动通信网络协议的基础上,用户即可使用 WAP 服务. 但是,用户无法在手机、掌上电脑(Personal Digital Assistant, PDA)等

无线终端上直接显示 PC 机使用 IE 浏览器浏览的 HTML 语言编写的页面,因此,必须将原有的 HTML 页面转码到无线标记语言(Wireless Markup Language, WML)页面^[1-2],用户才能在无线终端上正常浏览所需的信息.

目前,许多学者提出一些 WAP 转码方案^[3-7]: 基于服务器端的转码技术、基于客户端的转码技术等. 但是,这些转码方案中存在转码错误或失真、转码效率不高、转换后 WML 格式不规范、难以完整描述原来页面信息等缺陷.

收稿日期: 2013-01-08; **修回日期:** 2013-03-22

基金项目: 广西自然科学基金项目(2013GXNSFAA019337);广西教育厅重点科研项目(2013ZD056);玉林师范学院专项项目(2012YJZX04);广西教育厅研究项目(201010LX427)

HTML 网页页面通常包含导航、音乐、视频、广告、动画等各种多种媒体的内容,如果要设计出一种统一的 WAP 转码技术方案,能解决所有类型多种媒体的格式转换问题,固然很理想,但是难以实现.因此,设计转换方案时,需要针对不同类型页面内容来设计有针对性的转码方案,充分考虑每个 HTML 文件的类型特性,才能使其在兼容性、转码效率、实现细节等方面具有更好地技术优势^[3,8-11]. 本文提出一种优化、高效的基于可扩展标记语言(Extensible Markup Language, XML)的中间转换模型来获得 WAP 页面,使其能针对包含有大量文本类型页面转换时,可以进一步提高转换效率,提高转换准确性,降低转换数据丢失率,并对该模型进行仿真实现.

2 基于 XML 的中间转码模型设计

2.1 HTML、XML、WML 语言

超级文本标记语言 HTML 是标准通用标记语言的应用和标准,它通过标记符号来标记要显示网页的各个部分.可扩展标记语言 XML 是标准通用标记语言下的子集,它可以标记数据、定义数据类型,允许用户定义自己标记语言的源语言,它能统一描述和交换独立于应用程序的结构化数据,适合 Web 传输.与 HTML 相比,XML 更侧重于传输和存储数据,而 HTML 则侧重于显示数据的外观. WML 语言是 XML 语言的子集,它是从 HTML 继承而来的标记语言,它比 HTML 消耗浏览器更少的内存和 CPU 时间,因此对移动终端具有更好的可用性.

2.2 模型结构

基于 XML 的中间转码模型结构如图 1 所示,由三层结构组成.首先,将 HTML 网页页面文件通过 H-X 转换算法转换成中间桥梁 XML 网页页面文件,再将 XML 网页页面文件通过 X-W 算法转换成 WML 网页页面文件.



图 1 基于 XML 的转换模型

2.3 模型算法描述

基于 XML 的中间转码模型采用间接转换的思想,它利用 WML 与 XML 具有格式相似的特点,以 XML 具有强大的数据交换能力为基础,以 XML 中间数据的 WEB 页面可重用性作为技术支撑来实现. HTML 到 WML 的转码过程如图 2 所示.

H-X 算法:实现从 HTML 页面到 XML 页面

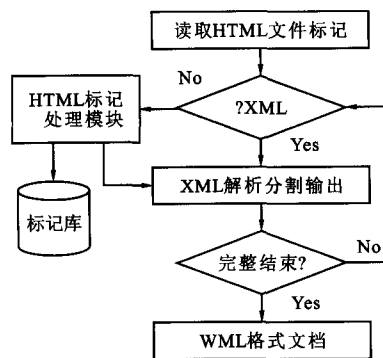


图 2 HTML 到 WML 的转换过程

的转换,内容包括:判断格式,定义 HTML 结构、HTML 分段,构造 HTML 空树,遍历输出 XML 文件等.

对 HTML 格式的文档进行遍历两次的处理.第一次遍历,主要完成与内容转换相关的工作.第二次遍历,主要完成第一次遍历之后的 XML 格式文档结构划分与组合的问题.在第二次遍历中,以第一次遍历生成的中间字符串形式文档作为输入,经过进一步划分处理后生成转换后的 XML 格式文档.

X-W 算法:实现 XML 页面转码到 WML 格式页面,内容包括解析 XML 文件内容,分割 XML 文件,输出 WML 文件等.

XML 标记语言作为标记语言的元语言,在应用中有着很强的灵活性,在分割处理大型 XML 格式文件时,不同的环境对文件的粒度也有不同的要求,在实际应用中提出带有普遍适用性的决策模型来解决这一问题是非常困难的.因此,为了方便,在本文所讨论的内容里,WAP 页面的大小限定在一定范围内.同时,为简化相关操作,在分割处理过程中约定每一块分割的 XML 格式文件块都以 XML 元素的结束符</>结尾.

3 模型实现与仿真

3.1 实现环境

对所构建的基于 XML 的中间转码模型进行模拟仿真实现,测试的环境配置如表 1 所示.

3.2 实现过程

随机抽出一个 HTML 文件 test.html, XML 到 WML 的转换过程如下:

(1)使用 Java 的 SimpleDOMParser 类和 parser()方法实现对 XML 文档结构树的解析;

(2)把 XML 格式文档解析到 XML DOM 对象中,把它解析为一个 DOM 树.选择标准的 XML 文件 test.xml,通过文档的规范处理,过滤掉图片链

接、声音等标签元素。把 test.xml 文档载入到 XML 解析器中,仿真结果如图 3 所示。

表 1 测试的环境

硬件配置	软件配置
戴尔 R210; CPU 主频 2.4 GHz, 内存 2 GB, 硬盘 250 GB; 联想万全 T100; CPU 主频 3.1 GHz 内存 2 GB, 硬盘 250 GB	Windows Server 2003; Lotus 莲花 Domino 8.0.2; WAP 页面开发集成工具包 (UP. SDK); Designer 集成工具; M3Gate 浏览器, 服务器工作时, 直接输入 http://127.0.0.1/文件名的形式对 WAP 页面浏览测试。

基于XML的中间转换模型设计的测试页面

to 接收方
from 发送方

2011年, 中国国内生产总值47.2万亿元, 比上年增长9.2%; 公共财政收入10.37万亿元, 增长24.8%; 粮食产量57 121万吨, 再创历史新高; 城镇新增就业1 221万人, 城镇居民人均可支配收入和农村居民人均纯收入实际增长8.4%和11.4%。我们巩固和扩大了应对国际金融危机冲击成果, 实现了“十二五”时期良好开局。

图 3 测试页面

(3)对 XML 文档进行分割处理,如果有多个 XML 文件,可以假设每个 CARD 显示 X 个字符 (这里取 200 个字符即 100 个汉字),则对应 #textn,分割后 CARD 的总数目为 $m = \text{文字总量} / \text{CARD 总容量}$,由此得到所有 CARD 对应的 WML 格式文件为 text1_card1.wml 一直到 text1_cardm.wml。然后,对 test.xml 分割并生成 4 个 CARD。

3.3 仿真结果

选择仿真工具 M3Gate,设置好 settings 选项等环境变量后,启动 M3Gate 浏览器,并在 url 栏里输入运行地址,即可在 winwap 浏览窗口中查看 test.wml 页面的运行效果。

4 性能测试比较与分析

(1)文本转换率与文本正确率

$$\text{文本转换率}(K) = \frac{\sum_{i=0}^{\infty} M}{\sum_{j=0}^{\infty} N} \quad (i, r = 0, 1, 2,$$

$3 \cdots \infty$),其中 M :转换前的 XML 格式文本数目, N :转换后的 WML 格式文本数目;文本正确率(L) =

$$\frac{\sum_{j=0}^{\infty} D}{\sum_{j=0}^{\infty} Q} \quad (i = 0, 1, 2, 3 \cdots \infty), \text{其中 } D: \text{转换后正确的}$$

WML 格式总文本数目, Q :转换前正确 XML 格式包括总文本数目(注:如果转换时能保证 XML 格式正确,那么 $N = Q$)。

(2)时间复杂度

使用 Rational Test RealTime 工具进行时间复杂度性能比较,记录算法程序的时间耗时并分析。获取算法运行时间的实现代码如下:

```
void GetRunTime(int ms1, ms2)
{
    int subMs=0;
    int UsingHour, UsingMinute, second, ms;
    subMs=ms2-ms1;
    UsingHour =subMs/3600000;
    subMs-=UsingHour * 3600000;
    UsingMinute=subMs/60000;
    subMs-=UsingMinute * 60000;
    second=subMs/1000;
    ms=subMs-second * 1000;
    cout<< UsingMinute <<"分"<<second<<"秒"<<ms<<"毫秒"<<end;
}
```

为对模型进行更深入的测试,从互联网中随机抽取一些纯文本信息页面内容作为测试对象,响应时间分析如表 2 和图 4 所示。

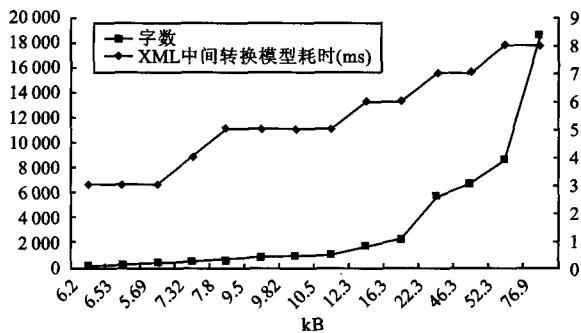


图 4 坐标分析图

可以看到,基于 XML 作为中间转换模型对随机抽取的 HTML 纯文本内容的转换,有较明显的时间响应优势,即使文本字数突然增加,模型耗时并没有显著增加。具体来说,当字数 150 耗时 3 ms,而字数 2309 耗时 6 ms,此时,提高了 100%,这是因为访问页面时都会存在固定的耗时时间。但当字数 5765 时,耗时 7 ms,字数是 18 629 时,耗时只有 8 ms,耗时增加仅 14.29%,远低于之前的 100%增加率,如果不考虑访问页面时存在的固定耗时时间,此算法的耗时更小,效率更高。这种间接转换思想的优势还在于:在 XML 作为 WEB 通用语言逐渐占主流地位的环境下,这也有利于大量 HTML 格式的

WEB 页面资源的充分复制。

表 2 响应时间记录表

序号	文本转换率(K)	文本正确率(L)	文本字数	转换耗时/ms
1	100	100	150	3
2	100	100	225	3
3	100	100	380	3
4	100	100	513	4
5	100	100	680	5
6	100	100	877	5
7	100	100	920	5
8	100	100	1 030	5
9	100	100	1 716	6
10	100	100	2 309	6
11	100	100	5 765	7
12	100	100	6 780	7
13	99.99	100	8 670	8
14	99.99	100	18 629	8

5 语 束 语

本文分析了现有的 WAP 转码方案,设计了一种基于 XML 的中间转换模型,给出相关算法和实现程序,实现从 HTML 页面转换到 WML 页面。最后给出了实例,对页面转换进行了测试,并记录相关的算法运行时间,对模型性能进行分析,证明了本模型转换时的准确性和具有更高的效率。针对其它格式的转码方案是下一步研究的方向。

参考文献:

- [1] 李青山,陈平. 一种基于内容的 HTML 到 XML 转换策略[J]. 计算机工程与应用,2001,37(9):30-32.
- [2] 张仁良. 基于自动生成技术的杉达学院 WAP 网站的设计与实现[D]. 上海:上海交通大学,2009.
- [3] 温泉. WAP 页面转换的实现技术研究[D]. 江苏:河海大学,2004.
- [4] 钱程,阳小兰. HTML 到 XML 转换研究[J]. 计算机与现代化,2011(8):39-41.
- [5] 林碧英,李四达. HTML 页面到 WAP 页面转码器的研究[J]. 中国电力教育(S3),2006,:21-22.
- [6] 李佳,面向移动平台的网络信息获取技术的研究与实现[D]. 哈尔滨:哈尔滨工业大学,2006.
- [7] 董树明. 半结构化 Web 信息抽取技术及其应用研究[D]. 江苏:东南大学,2004.
- [8] 葛冬梅. 基于内容个性化的 HTML 页面转换研究[D]. 黑龙江:哈尔滨工程大学,2008.
- [9] 刘杰,束博. 一种高效的 HTML/XHTML 至 WML 的转换方法[J]. 北京工商大学学报:自然科学版,2006,24(6):45-48.
- [10] 刘洪,贺深,黄河燕. WAP 页面转换代理系统原理及实现[J]. 计算机工程与应用,2002(4):177-179.
- [11] 马黎,基于信息剪裁的 HTML2WML 智能转换算法[J]. 微型电脑应用,2009(25)1:1-4.

作者简介:

李超建 男,(1978—),硕士,讲师.研究方向为计算机网络、嵌入式技术.

朱晓姝 女,(1973—),硕士研究生导师,教授.研究方向为网络计算、数据挖掘.

(上接第 160 页)

- [3] 李伟光,薛龙飞,母晓峰. 智能多媒体教室无线控制研究[J]. 科技通报,2013,29(4):220-222.
- [4] Hu WM, Xie D, Tan TN, et al. A hierarchical self-organizing approach for learning the patterns of motion trajectories[J]. Chinese Journal of Computers, 2003, 26(4):417-426.
- [5] Zelnik—Manor L, Irani M. Event—based analysis of video[C]// Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2001: 123-130.
- [6] 许晓昕,李安贵. 一种基于的网络聊天关键词提取算法[J]. 计算机技术与发展,2006,16(3):122-123.
- [7] Zhou Yanao, Tang Jianbo, Wang Jiaqin. An improved feature selection algorithm based on information entropy[C]//Proceedings of the 26th Chinese Control Con-

ference CCC. [s. l.]:IEEE,2007:312—315.

- [8] 徐燕,李锦涛,王斌,等. 基于区别分类能力的高性能特征选择方法[J]. 软件学报,2008,19(1):82-89.
- [9] 巩知乐,张德贤,胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真,2009,26(7):165-168.
- [10] Zhang D, Gatica-Perez D, Bengio S, et al. Semi-supervised adapted HMMs for unusual event detection [C]// Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2005:611-618.

作者简介:

彭其华 男,(1969—),硕士研究生,讲师.研究方向为计算机软件开发、电子商务.