基于树结构的 Web 页面适配方法的研究

高集荣[1],田 艳[2],江晓妍[1]
(1.中山大学 计算机科学系,广东 广州 510006;
2.西安财经学院 信息学院,陕西 西安 710061)

摘 要:设计和实现了从互联网页面到手机页面的适配转换机制,提出了基于树结构分析的 Web 页面适配方法,该适配方法首先对互联网页面建立对应的文档模型树结构,依据用户硬件数据信息,对这棵树进行网页去噪声、对 Frameset/Iframe 适配、分页重排、智能缓存以及多国语言字符集支持的操作,最终得到 XHTML MP 页面,完成了 Web 页面到手机页面的转换。通过实验,验证了整个页面适配过程和方法的可行性。

关键词:互联网页面适配网关;树结构;页面适配;文档对象模型
中图分类号:TP393.092 **文献标识码**:A **文章编号**:1674-7720(2014)01-0077-04

The research of the Web page adaptation method based on tree structure

Gao Jirong[1], Tian Yan[2], Jiang Xiaoyan[1]
(1.Dept.of Computer Science, SUN YAT-SEN University, Guangzhou 510006, China;
2.School of Information, Xi'an University of Finance and Economics, Xi'an 710061, china)

Abstract: In this thesis, a conversion mechanism which adapted the Internet page to mobile page is designed and implemented, a webpage method based on tree structure analysis is proposed, and the design of system algorithm with its C++ implementation is introduced. The Internet page adaptation method in this paper creates the corresponding document model tree structure of Internet page firstly, and removes the page noise, adapts Frameset/Iframe, paginates, restricts, deals with intelligent caching and multi-language character sets supported operating, finally gets the XHTML MP page. The feasibility of the process and methods for internet page adaptation is verified by a series of experiments.

Key words: internet page adaptation proxy; tree structure; page adaptation; document object module pattern

随着信息化时代的到来,手机越来越智能化,但手机用户访问 WWW 站点的需求仍得不到很好的解决。目前主流的应对措施是研发手机浏览器,但通过这种方式,只能使较高端的手机得到较好的用户体验,不能解决大量中低端的手机硬件限制的问题,中低端用户群依然得不到很好的业务使用保障,也无法有效地访问 WWW 站点[1]。

网页适配网关是随着移动互联网的发展需要新产生的一种设计架构,近几年内不断有国内外的专家学者提出了类似网页适配网关这种设计概念。2005 年 LAAKKO T 和 HILTUNEN T 提出内容适配网关解决针对 Web 页面手机终端开发的问题,使得手机用户有效地使用互联网信息[2]。2010 年甘玉珏等人也做了相关的研究[3]。

本文在 WAP 2.0 的基础上,设计和实现了从互联网页面到手机页面的适配转换机制,提出了基于树结构分析的 Web 页面适配方法,并对系统进行了算法的设计和 C++实现。算法实现了 HTML 标签语言到 XHTML MP 标签语言的转换,依据用户硬件数据信息,对互联网上的内容进行了适配,突破了手机屏幕尺寸,内存等硬件限制,满足了各种类型的手机终端访问 WWW 站点的需求。

1 算法描述

本文设计和实现了一个基于树结构的的网页面适配方法,实现了以下 7 个主要功能:

(1)DOM 树结构解析 HTML;

(2)去掉网页内复杂的节点,实现网页去噪声;

(3)按照一定规则处理页面上的框架,输出经过处理的 HTML 或序列化后的 DOM 树;

(4)对不符合手机终端屏幕大小的大页面进行网页切片;

(5)页面重排;

(6)分页缓存的设计与实现;

(7)多国语言字符集支持。

该方法包括接收和处理用户 HTTP 请求,将网页内容的标签语言 HTML 转换为符合手机标准的标签语言 XHTML MP,同时要将大页面的内容根据手机屏幕大小进行分页,并且将分页的数据结果存入到模块的分片缓存中。本文的 Web 页面处理方法是页面适配网关的核心环节和数据处理速度的瓶颈,基于 DOM Tree 的 HTML 解析过程和分片缓存的设计大大提高了网关的效率。

1.1 DOM 树结构化 Web 页面

Web 网页大多由 HTML 标签语言构成,HTML 是非结构化的,很难直接应用于下一步的研究和开发中。因此本文在 Web 网页适配方法中首先将 HTML 页面解析为 DOM Tree 的逻辑结构,提交给下一步的操作[4]。

本文算法采用的是基于标签结构的网页内容分析方法,所有对于 Web 页面内容的分析都是通过 HTML 标签来加以识别和分析的,例如通过 frameset、frame 和 iframe 来识别网页中的框架元素,并加以提取和分析处理。但由于 HTML 文档是无结构的,因此需要基于 DOM 树结构分析对 HTML 文档标签进一步的处理[5]。

HTML 原本是专为方便网页浏览器访问电子文档而设计的,但随着 HTML 的发展而进行改进设计的浏览器可以容许错误的编码、忽略错误的语法和允许“邋遢”的 HTML 代码。网页浏览器包含了许多可以忽略错误的程序指令,例如丢失结束标记的错误。即使 HTML 文档在语法上错误百出,PC 浏览器一样能读它。由此造成了 Web 上的 HTML 文档内容大部分格式不友好,手机浏览器不能支持,显示这些文档的数据变得比较困难。为了解决这个问题,本文采用了支持 DOM 的开放源代码的 Tidy 库。

DOM 可以将整个 HTML 页面文档规划成由多个相互连接的节点集合构成的文档,文档中的每个部分都可以被看作是一个节点的衍生物。这样一个节点的集合在逻辑形式上被看作为一棵 DOM 树。DOM 树中的文本内容和层次结构可以有效地指导数据区域定位和实体区域定位,这令 HTML 结构解析的难度大大减低。HTML 文档中的所有节点组成了一棵文档节点树,树中的节点代表了 HTML 文档中的元素、属性和文本等。HTML DOM 节点之间存在等级关系。树的根节点为 HTML 文档节点,并由此派生出它的子树,直到这棵树的所有最低级别的文本节点为止。

(1)对 Tidy 进行封装

在系统的实际开发中,本文对 Tidy 进行了客户化的操作,使其可以将 HTML 页面转换为符合 W3C 标准的 XHTML-MP。Tidy 最初设计的目的是用来自动修正 HTML 中的错误和松散的标签,检查 HTML 代码,并指出其中没有完全符合 W3C 发布标准的地方,它可以用来分析一个 HTML 文件或者一个包含 HTML 语句的字符串,还可以自动进行必需的修改以使代码符合相关标准的要求。

由于 Tidy 支持 HTML 与 XHTML 的互相转换以及 HTML 转换为 XML,因此本文在 Tidy 基础上,对它进行了一系列的封装和客户化,进行了应用级别的开发。这些新的函数包括 checkXhtmlMP ( ), parseDomTree ( ), paginate ( ), FrameSetProcessor ( ), restruct ( ) 等,它们是在 Tidy 基础上,进行二次开发的。

(2)建立 HTML DOM Tree

HTML 文档结构化为一棵 DOM Tree 的过程实际上就是新建一个 DOM Tree 对象并且初始化的过程。本文的实现方法是新建一个 DOM Tree 列表:TidyDoc*docs=new TidyDoc[out_num]。

1.2 网页去噪声的设计与实现

网页中的与主题内容无关的元素,如版权、广告、导航栏等,称为网页噪声。据统计,每张网页中有多达 40%~50%的内容是模板噪音。噪音会对网页信息处理造成很大干扰。

1.2.1 去噪声策略

HTML 原本是一种简单的标签语言,但 Web 网页发展迅速,Javascript 使得动态网页的可交互性变得更为强大,CSS、Image、Flash 以及各种插件使得 Web 家族变得日益庞大,也给手机用户浏览 Web 页面带来了困难[6]。因此本论文在 HTML DOM Tree 建立的基础上,保留了 Frame 相关的元素,裁剪了其他的所有标签,并交付给网关框架中的其他模块做适配和处理。

另外,网页中有些次要元素包括广告信息、不可见的外部链接、用于网页修饰的图片元素对手机用户获取信息并不是必要条件的内容或者不可见的内容,需要网页适配网关采取必要的裁剪,将更简单清晰的 Web 页面内容返回到手机用户的终端屏幕上。

1.2.2 本文网页去噪声的具体实现

(1)从文档根节点开始,深度遍历已经建立的 HTML DOM Tree;按照表 1 所示的标签删减子节点,将 script 块的内容交付给 JavaScript 模块进行处理。

(2)将去噪声后的 HTML 文本内容保存为一棵新的 DOM Tree。

1.3 Frameset/Iframe 适配算法

1.3.1 ICAP 模式

本文是基于 ICAP 协议的应用网关开发。ICAP 协议在结构和用法上是和 HTTP 协议有一样具有请求和应答模式，它包括两种模式：REQMOD 和 RESPMOD。

到 origin server 取回 Frame 页面时，需要在 ICAP REQMOD 下进行，因此本文研究的 Frame 处理方法是基于 REQMOD 模式的。

1.3.2 Web 页面中 Frameset 的处理

HTML 页面中的 Frameset 代表一个框架集，用于组织多个框架和嵌套框架集。Frame 在 Frameset 元素内表示单个框架。通过使用框架集，在同一个浏览器窗口中可以显示不止一个页面，每个页面被称为一个框架，每个框架独立于其他框架显示。

目前大多手机终端不能处理网页中的 Frameset，因此对网页中的 Frame 进行重新适配，变得非常必要。

(1)对 Frameset 的处理原则

多数手机终端不能处理 Frameset，所有 Frame 页面的内容应该被整合到同一个页面中，才能使页面内容被有效使用。基于这个原则，本文将 Frameset 转换成标准页，HTML 文档中与 Frame 相关的标签将被删除，如表 1 所示，Frameset 标签被替换为 body，而 Frame 则被替换为 div，适配后将得到新的 XHTML MP 文档。

表 1 Framset 标签处理

Html 文档	操作	Xhtml MP 文档
Frameset	替换	body
Frame	替换	div

当 http request 的 header 经过 squid 代理服务器传递给 Icap server，再传达给 Web 页面适配网关，网关提取 Http header 中的 User-Agent 信息，并去用户数据信息缓存中取手机终端信息，根据 Frame 处理过程，相对的 URL 地址会被改写成绝对 URL，并按照这个绝对的 URL 地址到 original server 取回每个 Frame 的内容，对标签 Frame 用 div 进行替换；并对网页重排后，将三个 Frame 的内容重新排列到同一个页面里，得到如下适配结果，经过 checkMP 函数，得到审核后的适配布局。

(2)Web 页面中 Iframe 的处理

Iframe 是 HTML 页面中的内联框架。HTML 文档在浏览器端的显示结果是一张页面内嵌的图片。通过例子中的 HTML 文档，可以观察到 Iframe 最主要的特点是内置于 HTML 标签 body 内，所以 Iframe 也被称为内嵌框架。

表 2 显示的是在 HTML DOM Tree 中（即 HTML 文档中），iframe 节点要被替换为 div 节点，符合 XHTML MP 移动概要，才能在手机终端上显示 Iframe 的内容。

表 2 Iframe 标签处理

Html 文档	操作	Xhtml MP 文档
Iframe	替换	div

1.4 页面重排设计与实现

本文研究的网页重排方法是一种基于通用规则的网页重布局方法，需要将 HTML 语言到 W3C 定义的手机标准格式语言 XHTML MP。网页重排后输出的 DOM Tree 将由 HTML DOM Tree 转换为 XHTML DOM Tree。

网页重排是指让 Web 页面上的内容在手机上的显示效果主次分明，适合在手机上阅读，方便用户迅速定位到自己感兴趣的内容。

现在的网页排版优化方法主要有基于通用规则的网页内容局部变换，以及基于网页封装器的页面定制这两种手段。基于通用规则进行网页内容局部变换是网页重排的一种方法，其基本原理是根据页面内容本身的结构特点，研究出可以让网页在手机上显示效果更好的局部变换规则，这些规则具有明确的触发条件，当条件成立时，在输出数据中添加特定结构或内容，或者选择忽略或输出原页面特定的结构和内容。对页面内容的局部变换通常会使用一些比较保守的策略。其优点是可以应用于各种页面，并能在一定程度上降低页面数据流量和页面排版的复杂度，缺点是精确性比较低。

本文 Web 页面到 XHTML MP 页面的网页重排过程如图 1 所示。

图 1 网页重排示例

网页重排时先从用户信息缓存中取回重排需要的一个参数 maxResponseSize（最大响应大小）。这个参数决定一页内容的大小，在网页分片中使用过。另外还需要得到手机的品牌型号。

1.5 分页的设计方案

为解决 Web 页面适配后整个页面内容冗长的问题，本文提出要对适配后的网页进行分页。网页分页就是在 HTML 文档树中确定分页的节点，将 Web 页面进行切分的过程。

为了提高系统的效率，本文在系统的实际开发中设置了分片缓存这个子模块。分页缓存指的是在页面适配网关中设置一个只能缓存模块存储用户访问过的 Web 网页，这些网页是以往同一型号手机用户通过页面适配

网关访问过的页面，已经经过适配但并没被更新。因此手机用户不必在每次访问 WWW 站点时都要重新进行页面适配，而只在该型号手机第一次访问时再进行页面适配。

当缓存中有相应的内容，分片缓存直接将缓存的相应页面发给用户，可以避免重复访问 Web 网站上的静态内容，重复再做网页适配，从而减少中间件到网站获取网页的连接时延和传输耗时。设立分片缓存同时可以使用户直接跳到目标页，提高了访问速度。

2 算法验证

2.1 不含 Frame 的页面适配实验

在本模拟实验中采用大小为 50 KB 的 W3C 测试页面。这个测试页面非常典型，是不含有 Frame 的普通网页，这使实验的测试点足够小，去掉了 Frame 的干扰情况。在这个模拟实验中，只需要验证这个网页是否经过了本文所实现的适配流程：建立 DOM Tree->去噪声->网页重排->网页切片。W3C 测试页面经适配前的显示效果如图 2 所示。

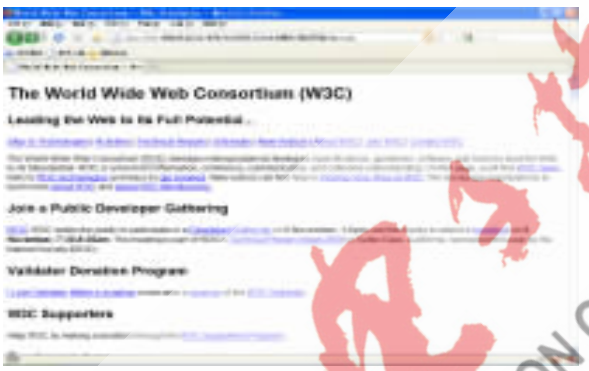

图 2 W3C 页面原始页面

HTTP Request 的 header 经过 squid 代理服务器传递给 Icap server 后，再传达给 Web 页面适配网关，网关提取 Http header 中的 User-Agent 信息，并去用户数据信息缓存中取手机终端信息。根据用户信息，Web 页面适配网关进行了网页解析、去噪声、网页分页、网页重排，W3C 页面经页面适配网关的适配结果如图 3 所示。从结果中，可以看到经过网页分页处理，页面被切成了两页。经过网页重排，网页的布局有所改变，符合本模拟实验中测试手机的屏幕大小。网页中的噪声点被取出，得到的结果更简洁，适合手机显示。这个实验很好地验证了 Web 适配系统的功能。

用户信息数据库用在终端适配中，这个数据库存储了目前世界上主流手机的品牌型号以及操作系统。页面适配网关在收到 HTTP 请求时，将查看 HTTP 头部的字段，得到手机品牌型号后，在用户信息数据库中查找出该手机品牌型号的操作系统以及屏幕尺寸大小、分辨率、手机内存等性能参数。这些信息同时将存入用户信息缓存中，向页面适配网关的其他处理模块传递相应的数据参数。

2.2 Frame 页面适配实验

为验证本文工作的 Frame 页面适配子模块，本文模拟了 Web 页面中含有 Frameset 的情况。

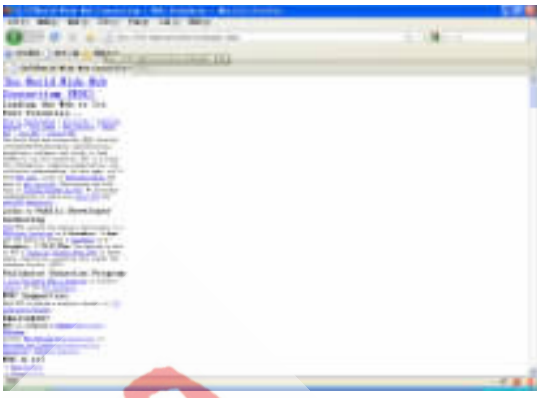

图 3 W3C 页面适配后结果

2.3 分页和重排黑盒实验

分页和重排黑盒实验测试了 Web 网关将大页面进行分页和重排的过程。

实验结果显示，其中分页的结果不完全与测试页面完全对应，原因主要有两点：(1) 在整个互联网页面适配网关中，广告投放模块会将格外地在适配后的结果中投放广告等内容。(2)当测试页面经过适配方法的去噪声模块后，不能被网关适配且手机终端不能支持的模块内容从页面中被删除了。

本文提出的基于树结构的互联网页面适配方法解决了互联网页面适配网关中的 HTML 适配，并分别通过黑盒实验验证了互联网页面适配方法的可行性，白盒实验验证适配算法 C++程序代码路径的正确性。该项目成果已经在某一品牌手机中得到了具体验证。

参考文献

[1] 小雨.爱立信 Web 网关提升移动互联网应用体验[J].世界电信(企业报道版)，2009(4)：76.

[2] LAAKKO T, HILTUNEN T. Adapting Web content to mobile useragents[J]. IEEE Internet Computing, 2005, 9(2): 46-53.

[3] 甘玉珏，杨杰，苏军根，等.移动互联网 Web 网关的设计[J].移动通信，2010(22)：71-74.

[4] 李猛. 基于 DOM 的 Web 信息抽出技术的研究与实现[D].大连：大连理工大学，2008.

[5] 寇月，李冬，申德荣，等.D2EEM：一种基于 DOM 树的 Deep Web 实体抽取机制[J].计算机研究与发展，2010，47(5)：858-865.

[6] 时达明，林鸿飞，杨志豪.基于网页框架和规则的网页噪音去除方法[J].计算机工程，2008，33(19)：276-278.

(收稿日期：2013-10-25)

作者简介：

高集荣，男，1960 年生，博士，副教授，主要研究方向：数据库技术、网络技术及其应用。

田艳，女，1962 年生，硕士，教授，主要研究方向：信息管理技术、网络技术及其应用。

江晓妍，女，1987 年生，硕士，主要研究方向：数据库技术、数据挖掘、网络技术及其应用。