



基于 HTML5 的移动互联网应用转换技术^{*}

李慧云¹, 杨新章¹, 胡文胜², 文锦军¹

(1. 中国电信股份有限公司广东研究院 广州 510630;

2. 中国电信股份有限公司广东分公司 广州 510081)

摘 要: HTML5 标准的提出增强了 Web 应用的功能和表现力, 使得许多原本需要插件辅助才能实现的功能可由 HTML 原生支持, 为 Web 网站的移动化提供了可行的解决方案。从 HTML5 适用于移动终端的新特性出发, 研究了将传统 Web 网站转换为适用于移动终端平台的 Web 网站(或应用)的多项关键技术, 并对转换的效果和难度进行了分析和总结。

关键词: HTML5; 应用转换; 网站移动化

doi: 10.3969/j.issn.1000-0801.2013.05.005

Converting Technologies for Mobile Internet Application Based on HTML5

Li Huiyun¹, Yang Xinzhang¹, Hu Wensheng², Wen Jinjun¹

(1. Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China;

2. Guangdong Branch of China Telecom Co., Ltd., Guangzhou 510081, China)

Abstract: The proposed HTML5 standard makes web applications' ability be enhanced, and many functions that originally need many plug-ins to support can be realized directly in HTML. HTML5 becomes a new solution for mobilization. New features of HTML5 for mobile devices were described first, and then the converting technologies were expounded from web portals for PC to web applications for mobilephones. In the end, the converting effects and realization difficulties were summarized.

Key words: HTML5, application converting, application mobilization

1 引言

随着智能终端的普及和服务模式的创新,我国移动互联网市场成长迅速,越来越多的用户通过手机接入互联网。目前智能终端的运行能力几乎可以媲美 PC 机,用户可以如同在 PC 上一样浏览网页,但受限的屏幕大小和不同的操作方式大大降低了用户的体验。因此,许多 Web 应用服务商开始针对移动终端改进用户体验或者推出移动版

网页,以适配终端屏幕;或者开发本地客户端,直接从用户的使用方式上进行适配。

纵观移动终端的产业现状,本地应用仍占据统治地位,原因在于本地应用在表现能力和运行效率上具有较大的优势。但随着 HTML5 标准的出现,Web 应用在表现能力和功能上都得到了很大的提升,而其天然的跨终端平台的特性和更简单的开发语言,可能使 Web 应用在移动终端上获得更大的发展空间和占据更多的市场份额。

^{*} “核心电子器件、高端通用芯片及基础软件产品”国家科技重大专项基金资助项目“移动智能终端操作系统研发”(No.2012ZX01039002-001)

2 HTML5 适用于移动终端的新特性

HTML5 的许多功能就是以能够在移动终端（如智能手机以及平板电脑）上运行为基础而设计的，因此，HTML5 在移动设备上具有更大的优势，将率先在移动设备上应用起来。下面列举了 HTML5 适用于移动终端应用开发的诸多新特性。

(1) 离线存储能力为移动互联网应用的开发提供了基础

相比桌面应用程序，移动设备的网络连线更不稳定，而且有时在移动中并无网络可以使用，此时离线存储能力显得尤为重要。HTML5 提供了 3 种离线存储机制：Web Storage、Indexed DB 和 Application Cache。Web Storage 可以看作增强版的 cookie，不受数据大小的限制，有更好的弹性以及架构，可以将数据写入本机的 ROM 中，还可以在关闭浏览器后再次打开时恢复数据；Indexed DB 是 HTML5 提供的本地数据库机制，用来存储数据量较大、结构化的数据；Application Cache 可称为网页缓存，通过一个 manifest 文件定义需要浏览器下载缓存的网页文件，用于以文件为单位的本地数据存储。

(2) 原生音/视频支持，多媒体嵌入更加灵活

在 HTML5 中，增加了<audio>、<video>标签，用来在 Web 网页中嵌入音/视频播放功能，而无需 flash 和其他嵌入式插件的支持，避免了原生开发方式对于文字和音/视频混排的多媒体内容需要拆分处理的问题，也解决了移动终端浏览器不支持 flash 插件而导致的音/视频内容无法观看的问题。

(3) 增强的用户交互能力

HTML5 增加了拖拽、撤销历史操作、文本选择等交互动作接口，增强了 Web 应用的用户交互能力，适配了移动用户的操作习惯。

(4) 专为移动化平台定制的表单元素

浏览器中出现的 HTML5 表单元素与对应的键盘见表 1。

只需要简单的声明（如<input type="E-mail">），即可完成对不同样式键盘的调用，简捷方便。

(5) 具有鲜明移动特色的地理定位 API

地理定位是移动终端的特色和优势，目前嵌入移动定位的应用越来越多，并不仅局限于导航和地图应用。HTML5 的定位 API 可以综合使用 GPS、Wi-Fi、手机等多种

表 1 HTML5 表单元素与对应的键盘

类型	用途	键盘
Text	正常输入内容	标准键盘
Tel	电话号码	数字键盘
E-mail	电子邮件地址文本框	带有 @ 和.的键盘
URL	网页 URL	带有 .com 和.的键盘
Search	用户搜索引擎	标准键盘
Range	特定值范围内的数值选择器	滑动条或转盘

定位方式，定位更为精准、灵活。

(6) canvas 绘图，提升移动平台的绘图能力

HTML5 canvas 提供了 canvas-2D 的绘图功能支持、canvas-3D 的绘图功能支持、SVG-向量图支持以及图片的移动、旋转、缩放等常规操作。利用 canvas 可直接在终端侧绘制需动态变化的图形、图表，相比在服务器侧生成图片，减少了网络传输，提升了显示效率。

(7) 适用于移动设备显示的标签和样式

移动终端区别于 PC 设备的一个重要特征就是移动终端多样的设备尺寸和屏幕分辨率。如何让 Web 页面能适应各种尺寸的屏幕，让很多 Web 开发人员相当纠结。幸运的是，HTML5 已经为开发者考虑到了这一问题，在网页的头部加上语句<meta name="viewport" content="width=device-width; user-scalable=0;"/>，可以使网页适配设备屏幕。而且，在 CSS3 中，可使用媒体查询(media query)为不同分辨率的终端定制不同的显示样式。

总的来说，HTML5 从设计开始，就考虑了移动终端的特性和需求，将传统的 Web 网站向 HTML5 标准和应用转换，是 Web 网站向移动平台发展的有效手段和趋势。

3 转换技术

比较 PC 时代的 Web 网站和服务，移动版的 Web 网站/应用具有以下特点。

- 简洁：移动设备的屏幕远远小于桌面设备的屏幕，它无法像桌面设备一样同时呈现很多信息，所以需控制好显示的信息数量，尽量用较少的信息清晰地传递出所要传达的内容非常重要。
- 触摸优化：在移动版网页上，要保证用户能用拇指完成所需的所有操作，按钮大小、间隔以及单选、多选的呈现方式都需要根据移动终端的操作方式来设计。



- 强化主题:大部分在移动设备上浏览网页的用户都在寻找一些信息,而不是在网络上“闲逛”。将最重要、用户最需要了解的信息放在关键位置。
- 不依赖插件:移动终端的浏览器受设备能力限制,对插件的支持程度不同,减少对插件的依赖可以保证 Web 应用的普适性。
- 用尽量少的数据传输和重复数据加载:无线网络的网络质量变化非常大,带宽大小不一,尽量缩小网页的体积,去掉网页中的冗余代码,使用本地缓存机制,可提升用户的体验。

基于上述移动网站/应用的特点和原则,可以确定移动互联网 Web 应用转换的目标,即压缩图片,剔除网页中的多余信息,转换需要插件支持的内容,分页展现。因此,转换的关键技术包括:网页结构转换、音/视频转换、图片转换以及文档转换等。总体架构流程如图 1 所示。

3.1 文档转换

将文档转换为 HTML5 的目的是可以直接在 HTML 文件中显示文档(如 Word、Excel、PPT、PDF 等),相比传统使用 Adobe Flash 插件方式显示更为便捷;同时,HTML5 增加了新特性,如 CSS3 可实现翻页特效,canvas 的 2D、3D 模型可实现物理、化学公式和模型,Math ML 可实现数学公式、离线缓存等,增加了在线文档的可阅读性和互操作能力,是 HTML5 的重要应用领域。但由于转换过程需要耗费较长的处理时间,通常在后台完成转换,转换完成后,通过 HTML 页面呈现给用户显示。

3.1.1 Office 文档转换为 HTML5

Office 文档可以说是用户日常操作最多的文档类型,除了用户数最多的 Microsoft Office 办公软件系列生成及

使用的以 doc、xls、docx、ppt 等后缀结束的文件,也包括其他办公软件产品生成的文档,如 OpenOffice 的 odt 和 ott 文件、WPS 的 wps 文件等。Office 文档依赖于应用软件环境,用户在应用软件中打开、查看和编辑文档。

目前 Office 文档的标准主要包括:前国际标准(open document format,ODF)、中国文档标准(unified office document format,UOF)和现有的国际标准(office open XML,OOXML)。OOXML 是微软基于 XML-Extensible Markup Language 的数据存储语言,为 Microsoft Office 开发所制定的技术规范,在 2008 年成为国际标准,主要应用在 Office2007 以后的系列中,文件以 xlsx、docx、pptx 为后缀。在早期的 Microsoft 产品中,采用的是 OLE2 复合文档格式。

Office 文档转换为 HTML,使得用户可以直接通过浏览器查看甚至编辑文档,而不再需要再安装办公软件,实现了办公应用的云服务化,也大大提高了在移动终端上处理办公文档的方便性。总的来说,Office 文档转换分为如下两个步骤。

(1)文档的读取:按照文档格式读取文档内容,包括文本、图片、表格以及各部分内容的位置关系和显示样式。

(2)HTML 文件的生成:将文档内容按照原来的样式和排列位置,通过 HTML 标签和样式进行重新组合,生成 HTML5 文件。

Office 转换为 HTML 主要包括如下两种方式。

- 利用应用软件自带的 API 进行转换:Microsoft Office 和 OpenOffice 都自带了将文档转为 HTML 文档的 API,Microsoft Office API 由于平台限制只能用于 Windows 操作系统。

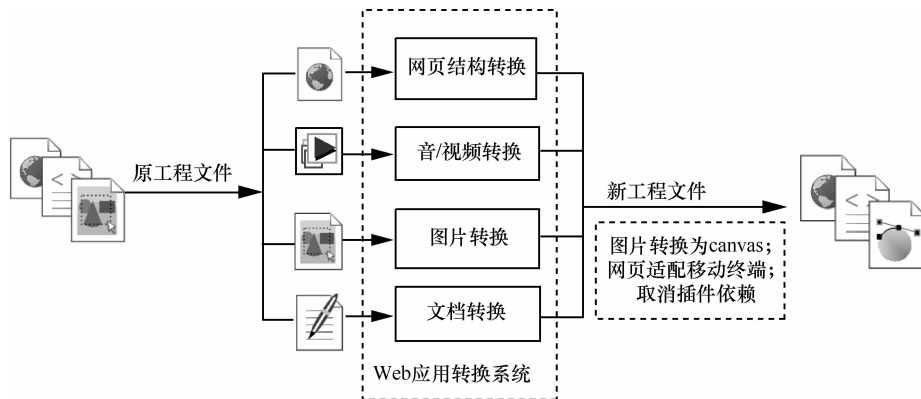


图 1 转换的总体架构流程

- 利用第三方的转换工具:比较有名的是 Apache POI, 是专门用于将 Microsoft Office 文档转换为 HTML 格式的工具, 支持 OLE2 和 OOXML 两个格式的文档转换。

3.1.2 PDF 转换为 HTML5

PDF(portable document format)是一种跨平台、与设备无关、用于文件交换的文件格式。PDF 文件包含一个 PDF 文档和其他支持数据。一个 PDF 文档包含一个或多个页面,每个页面包含与设备和分辨率无关的文字、图形和图像的任意组合,被称为页面描述。PDF 的文档结构是一种树型结构,树的根节点就是 PDF 文件的根对象。根节点下有 4 个子树:页面树(pages tree)、书签树(outline tree)、线索树(article tree)、名字树(named destination)。

PDF 转换为 HTML 的步骤同 Office 文档类似,需要先按照 PDF 的格式和文档结构,将 PDF 文档中的文本、图片、图标等内容单位以及相关的样式信息读取出来,然后按照 HTML5 的标签和样式要求进行重新组合和显示。

目前,PDF 也存在需要开源的转换工具,比较有名的包括 PDF to HTML、PDF2 HTML EX、PDF Masher 等。由于 PDF 内容的设备无关性,一些文字内容也通过图片显示,因此将其转换为 HTML 文档后,在中文和样式的效果上不一定能和原有的保持一致。

3.2 图片转换

为了实现图片到 HTML5 canvas 的转换,需要把 HTML 中的 img 标签的内容替换为 canvas 标签和一段 JavaScript 代码。转换的基本步骤是解析 HTML 文件,将其中的 img 标签的 src 属性的取值读取出来,利用 canvas 的引入图像功能完成图像的呈现,引入的操作分为两步:第 1 步是使用 JavaScript 的 image 对象加载图片;第 2 步是使用 drawImage 方法将图像插入 canvas 中。

其中,图片加载的关键是 img 标签的 src 属性的取值

所代表的图片的 URL。根据 HTML 的规范,img 标签的 src 属性的 URL 可以有两种形式:一种是图片的实际 URL 地址;另一种是图片经 Base64 编码转换生成的 data URI。

- 对于实际 URL 地址的情况,可以用 JavaScript 创建一个新的 image 对象,引入图片的 URL,在脚本执行过程中加载图片,但在图片数量较多时,JavaScript 脚本会因等待图片加载而暂停,需要进行图像预加载处理。
- 对于 data URI 的情况,可以直接通过 data URI 嵌入图像,通过这种方式嵌入的图片内容即时可用,进而减少了页面加载过程中的 HTTP 请求次数,可以优化单次页面加载的时间,但同时会带来编码的图片管理不便、无法被重复利用、无法被独立缓存等问题,同时在客户端需要重新解码和显示,也增加了额外的开销。

3.3 音/视频转换

音/视频转换即将原来由 flash 插件播放的音/视频或音/视频文件转换为 HTML5 标准所支持的音/视频,从而实现浏览器对音/视频的原生支持。

首先,需要了解浏览器对音/视频的支持情况,见表 2。

因此,为了保证音/视频文件能被所有浏览器支持,一般建议音频文件采用 Ogg Vorbis 和 MP3 格式;视频文件采用 WebM 和 MP4 格式。某些媒体文件(特别是 MP4)即使媒体类型为浏览器所支持,也有可能不能正常播放,这是因为文件的编码不被浏览器所支持。

对于音/视频内容提供商来说,只需要提供上述音/视频文件源,同时修改网页音/视频的嵌入代码,即可实现对 HTML5 音/视频的支持,并将服务扩展到移动终端。

3.4 网页结构转换

网页结构转换需要综合分析网页/Web 应用的工程文件,针对移动终端的特点对 HTML、CSS、JavaScript 文件进

表 2 浏览器对音/视频的支持情况

浏览器	媒体类型							
	音频					视频		
	Ogg Vorbis	WAV	MP3	AAC	MP4	TheoraOgg	WebM	MP4
FireFox	√	√				√	√	
Opera	√	√				√	√	
Safari			√	√	√			√
Chrome	√	√	√	√	√	√	√	√
IE9+			√	√	√		√	√
Android 的内置浏览器			√	√			√	√



行调整,以适配移动终端的屏幕显示、操作方式,甚至网络连接状态。

互联网上的页面通常比较复杂、内容元素较多、网页尺寸较大,因此,在适配移动终端时,通常采用的方法是对原网页进行裁剪,裁剪的过程需要考虑到可访问性和易用性两个方面的问题。具体来说,转换出的新页面应可以被大部分手机浏览,实现传统互联网的海量信息和移动互联网的互通,视为可访问性;而移动终端的用户在访问页面的时候需要综合考虑带宽、流量、时间等因素,视为易用性。

因此,页面的转换需要从如下两方面入手:

- 在转换页面的过程中,对页面信息进行有效过滤,屏蔽无关、不重要的信息以及广告信息内容,从而减小页面的大小,减少用户浏览页面所消耗的流量;
- 由于用户移动终端的操作系统以及屏幕大小均存在很大的差异性,页面转换需要对主要网页内容进行合理优化和排版,以适应不同大小的手机屏幕和操作系统所能承载的页面大小,这样便能“因材施教”,根据需要,进行不同的适配合理地完成网页的转换,以满足不同用户的浏览需求,从页面版式、内容、样式、交互和可用性方面提高浏览体验,最终为用户提供低成本的信息获取工具。

为了实现上述功能,网页结构转换工具完成网页结构转换工作的流程可以划分如下。

- 基本页面处理:通过对 HTML 文件(包括 JSP、ASP、PHP 文件)的页面分析,建立 DOM(document object model,文档对象模型)树,完成网页预处理和网页标签的合法性检查等工作。
- 主体内容处理:根据规则完成页面主体内容的识别处理与不重要内容的剔除工作。
- 样式处理:对提取了主体内容的 DOM 树进行样式处理,包括对 CSS 文件的调整与过滤。
- 分页展现:为了实现网页内容在移动终端上的呈现,需要对上述处理结果进行处理,根据终端的信息库,进行标签替换、分页和切页工作,最终实现页面的输出。

在 HTML5 标准中,采用了语义化的标签结构、丰富的表单结构和更具表现力的 CSS3,这些新特性都可以使得网页更加适合移动终端展示。在分页展现时,可充分利用 HTML5 的特性标签和 CSS 样式。

3.5 转换技术比较

分析上述的转换技术,都可以归结为将一种类型的文件或内容中的元数据和样式提取出来,再将元数据进行排列组合,转换为另一种类型的文件或内容,其转换后的样式和转换前大致相同。而影响转换效果和决定转换难度的因素主要在于转换前的文件和内容的编码与 HTML5 的内容编码的相似度以及其样式和 HTML5 能表现的样式之间的差别。

就上述几种转换技术而言,文档转换涉及的转换前后的编码差异最大,尤其对中文的编码和样式、文本样式、表格格式和一些特殊的字符公式等都存在较大的差异性,因此其转换后的效果也千差万别,在某些转换过程中,为了保证转换后的效果,可能会用大块的图片来替代原来的内容,降低了转换的意义和价值。就目前的转换工具而言,Apache POI 是做得比较出色的,其对 xls、doc、ppt 后缀文件的转换效果都不错,对 OOXML 格式的转换仍在开发中,目前对xlsx 后缀文件的支持较好。而以文档转换为基础的产品 Crocodoc 是一个可以将 PDF 和 Office 文档迅速转变成 HTML5 的工具,目前该工具使用范围很广泛,其使得用户可以在线分享和编辑文档,而不受设备和系统的制约。

网页结构转换的主要难点和工作量不是两种编码之间的转换,而是对网页结构的遍历解析和内容重组,难点主要在于以下 3 个方面。

- 文档对象模型是基于浏览器编程的一套 API,是 W3C 组织出台的推荐标准。该标准本身设计相对复杂,在 DOM 树的建立过程中,需要综合考虑标签的位置信息、对应的资源类型信息、语义类型信息,去除页面设计与主体内容无关的元素带来的页面复杂度。
- 为了实现页面内容的最佳呈现效果,需要对页面/Web 应用的 CSS 进行处理,去除移动终端不能显示的 CSS 样式内容数据,修改某些 CSS 的属性值(如调整 CSS 样式中的宽度,适配为屏幕尺寸),并添加某些 CSS 属性值。
- 由于移动终端支持的页面标签并不完善,同时应用开发者提交的页面中可能也存在某些不能转换或语法存在问题的标签,因此还需在预处理的过程中增加合法性检查,去掉不能转换或语法存在问题的标签。

(下转第 39 页)