Computer Science and Application 计算机科学与应用, 2020, 10(6), 1234-1242
Published Online June 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.106127

Hans汉斯

Modular Style Template Matching Method for Mobile Webpages

Shengnan Zhang, Jiawei Wu, Lianqiang Niu, Kun Yang

School of Software, Shenyang University of Technology, Shenyang Liaoning
Email: zsncjr@sina.com, 598583879@qq.com, niulq@sut.edu.cn, 133977475@qq.com

Received: May 28th, 2020; accepted: Jun. 5th, 2020; published: Jun. 12th, 2020

Abstract

Aiming at the problem that the traditional template-based methods for webpage mobile adaptation cannot deal with too complex pages, this paper proposes a new hybrid matching method based on modular style templates. It first categorizes the page blocks that make up webpages, designs the styles suitable for browsing on the mobile terminal for each type of page blocks, then identifies the type of the page block after webpage segmentation, and finally associates the determined page blocks with the corresponding modular styles to complete the page conversion. Experimental results show that the presented method only needs a small number of templates to process different pages and solves the problem of page style loss caused by segmentation.

Keywords

Mobile Webpage Adaptation, Style Template, Page Block Classification, Webpage Segmentation

移动网页模块化样式模板匹配方法

张胜男，吴嘉惟，牛连强，杨　坤

沈阳工业大学软件学院，辽宁 沈阳
Email: zsncjr@sina.com, 598583879@qq.com, niulq@sut.edu.cn, 133977475@qq.com

收稿日期：2020年5月28日；录用日期：2020年6月5日；发布日期：2020年6月12日

摘　要

针对传统的基于模板的Web页面移动端适配方法中无法处理过于复杂的页面问题，本文提出了一种基于模块化样式模板匹配的混合式适配方法。该方法首先对构成网页的页面块进行归类，并对每一类设计适合在移动端浏览的样式，然后对网页分割后的页面块进行类型识别，最后将识别后的页面块与对应的模块化样式相关联，进而完成页面转换。实验结果表明，本文方法仅需少量模板即可应对多种风格的页面，并解决了因分割造成的页面样式的丢失问题。

文章引用: 张胜男, 吴嘉惟, 牛连强, 杨坤. 移动网页模块化样式模板匹配方法[J]. 计算机科学与应用, 2020, 10(6): 1234-1242. DOI: 10.12677/csa.2020.106127

关键词

移动端网页适配，样式模板，页面块分类，网页分割

1. 引言

移动互联网的快速发展使得手机、iPad 等成为人们从互联网获取信息的主要终端设备[1]。不过，现存的网站大多是为大屏幕 PC 机设计的，并没有对移动端如手机进行适配，直接用手机浏览时用户体验较差。通过页面转换技术重新优化布局页面，能使其在移动终端具有良好的展示，从而实现网页对移动端的自动适配[2]。

现存的网页自动适配方法主要分为两类。样式模板适配法[3] [4]将整个网页看作一个不可分割的单元，先对大量网页进行聚类，生成每一类网页对应的模板，构成模板库。在适配网页样式时，计算网页的结构与库中模板的相似度，并选取相似度最高的模板作为该网页的样式模板。由于网页数量庞大，结构复杂多变，这种针对完整网页设计的模板与待适配网页特性的吻合度低，且相似度计算开销大，难以满足网页适配的实时性要求。

网页分割适配法[5] [6] [7]将网页分块后再进行移动端适配，在一定程度上避免了样式模板适配法的缺陷。由于目前的多数网页均采取标签与样式分离的原则设计，其网页样式信息主要保存在外部 CSS 样式表中，网页内部含有的样式信息很少，很容易造成分割后的页面样式信息丢失[8]。要使分割后的页面块符合移动端设计要求，必须重新设计块的样式，如何有效处理这些样式是算法成功的关键。

本文提出了一种基于模块化样式模板匹配的混合式适配方法。该方法先将组成网页的页面块归结为有限类型并建立样式模板库。在适配样式时，利用分割技术将网页分块，并映射到有限类型，再依据页面块所属类型与对应类别的样式模板关联，进而完成各页面块乃至整个网页的适配。该方法能够简化样式模板库的构造，提高效率，对结构复杂的页面也有更好的自适应性。

2. 页面块分类

2.1. 页面块分类依据和流程

页面块分类是模块化样式模板匹配的基础。按照页面块功能或风格特征，本文共设置了图片组/轮播图、导航、页面主体/正文、侧边栏、图片链接列表、非导航链接列表、页眉和页脚 8 个类别，基本覆盖了常见的 Web 页面组成元素。页面块分类算法主要依据 HTML5 中的如下标签和属性进行：

1) HTML5 中的语义化标签；

2) 常见的 class 和 id；

3) 页面块的属性和标签。

具体分类时，算法将按照 1)、2)、3)顺序依次对页面块进行判别。这里假设 n_i 为某页面经分割后形成的一个页面块，具体分类步骤如下：

Step1：判断页面块 n_i 是否包含语义化标签，如果包含，使用语义化标签分类；否则，进入下一步；

Step2：判断页面块 n_i 是否包含常见的 class 和 id，如果包含，使用常见的 class 和 id 进行分类；否则，

进入下一步；

Step3：判断页面块 n_i 是否至少满足一个属性值和标签条件，如果包含，依据页面块的属性值和包含的标签判断类别；否则，将页面块归为其他类别。

2.2. 基于 HTML5 中的语义化标签的类别判断

HTML5 引入了一批新的标签和属性，这些标签使得 HTML 在定义了网页结构的基础上还可以定义网页的内容类型，使开发者构建网页更加便利且易于搜索引擎识别[9]。表 1 列出了 HTML5 引入的新的语义化标签及其定义的内容区域描述，页面分类可以根据这些语义化标签直接判断页面块类型。例如，<nav>标签表示该区域是导航模块，包含<article>、<details>和<section>标签的页面块均可划分为正文类页面块。若该页面块包含多个语义化标签，则以更靠近根节点的标签为准。

Table 1. New semantic tags introduced by HTML5
表 1. HTML5 引入的新的语义化标签

标签	描述
<article>	定义页面独立的内容区域。
<aside>	定义页面的侧边栏内容。
<details>	用于描述文档或文档某部分的细节。
<footer>	定义文档的页脚。
<header>	定义文档的页眉。
<nav>	定义导航链接的部分。
<section>	定义文档中的节和区段。

2.3. 基于 class 和 id 的类别判断

现有很多的网页样式表中用于 CSS 选择器的 class 和 id 的命名已经形成了行业内的默认规范，例如 header、footer、nav、article、left、right 等。表 2 为部分常见的 class 和 id 以及它们对应的含义。依据 class 和 id 判别页面类别时，其中的版权信息可划分为正文类，其他相关链接可划分为链接列表类。若页面块包含多个常用 class 和 id，则以更靠近根节点的标签的 class 和 id 为准。

2.4. 基于页面块属性值和包含关系的类别判断

组成网页的各个不同内容块的标签和内容往往差别很大。例如导航块就是由一个或多个短文本链接列表组成的，网页的正文通常包含大量的文字，版权信息一般在页面底部等。因此本文利用这些不同页面块的特征来判断页面块类别，将页面块的图片数量、链接数量等 9 个属性作为判断依据[9]，其对应的特征量和判定条件如表 3 和表 4 所示。其中，表 3 中中心相对横坐标代表的含义是页面块的中心坐标点与页面垂直中线的距离。

Table 2. Some common classes and ids
表 2. 部分常见的 class 和 id

class 或 id	代表的含义
#sidebar, #aside	侧边栏
#nav, #subnav, #menu, #submenu, #tool, #toolbar, #drop, #dorpmenu	工具条，导航，菜单

Continued

#list, #news	链接列表
#friendlink, #joinus, #partner	其他相关链接
#copyright, #siteinfo, #siteinfoLegal, #siteinfoCredits	版权信息，法律声明等
#foot, #footer	页脚部分
#main, #article, #details	页面主体，正文
#head, #header, #banner	页眉部分

Table 3. Feature attributes of page blocks as the judgment basis
表 3. 作为判断依据的页面块特征属性

属性名	特征量	单位
图片平均大小	x_1	像素
图片数量	x_2	个
链接数量	x_3	个
链接文本数量	x_4	字节
非连接文本数量	x_5	字节
中心相对横坐标	x_6	百分比
相对纵坐标	x_7	百分比
宽度	x_8	像素
高度	x_9	像素

以导航块的判定条件为例，表 4 的判定条件意味着当页面块中链接数量大于等于 3 个且平均链接文本长度小于 12 字节，而且页面块的 HTML 代码包含
<a>的结构时，就判定该页块为导航类。

当页面块的属性符合多个条件时，规定当前页面块的类别优先级由高到低依次为：图片链接列表→导航、非导航链接列表→页面主体→图片组→页脚、页头、侧边栏。

Table 4. Determination conditions of the page block type
表 4. 页面块类型判定条件

属性值条件	标签条件	类型
$x_7 < 20\% \ \& \ x_8/x_9 > 2$		页眉
$x_2 \geq 2$		图片组/轮播图
$x_3 \geq 3 \& \ x_4/x_3 \leq 12$	<a>	导航
$x_5 > 80$		页面主体/正文
$x_8 < 30\% \& \ x_6 > 70\%$		侧边栏
$x_3 \geq 3 \ \& \ x_2 = x_3 \ \& \ x_1 < 10000$	<a><img>	图片链接列表
$x_3 \geq 3 \ \& \ x_4/x_3 > 12$	<a>	非导航链接列表
$x_7 > 80\% \ \& \ x_8/x_9 > 2$		页脚

3. 页面块样式模板库的构造

网页分割后，需要对分割后形成的页面块进行模板的匹配，因此需要针对每一类页面块的特点设计

相应的模板。

3.1. 样式模板设计要求

针对用户操作移动设备的习惯和移动端屏幕的特点，保证适配后的页面拥有良好的用户体验，设计的样式模板应遵循以下要求：

1) 将垂直滚动设为浏览页面的主要方式，避免水平滑动。
2) 样式模板的元素使用相对宽度与位置，确保能够适配不同的移动终端分辨率。
3) 图片等比例缩放至与屏幕同宽，但缩放过程不能超过图片的原始尺寸。
4) 每个页面块的宽度与屏幕宽度相同，并能够根据屏幕宽度自动微调布局。
5) 响应元素一律设置为块级元素，适当扩展可点击区域的边缘。

3.2. 样式模板设计

以文字导航模块为例说明模板设计方法。本文设计的导航块样式模板分为标签式文字导航、隐藏式文字导航两种，是否隐藏以及标签列数可以根据目标页面的导航链接的数量自动调整。图 1 和图 2 分别为 3 列和 4 列文字导航的样式模板示例。

由于移动端屏幕空间有限，不宜展示 PC 端复杂繁多的导航列表，所以当网页中导航栏的链接数量过多时，应考虑将导航隐藏，用户需要导航时则可以点击按钮来弹出导航。这种隐藏式导航在移动端的显示效果如图 3 和图 4 所示。

Nav1	Nav2	Nav3
Nav4	Nav5	Nav6
Nav7	Nav8	Nav9

Figure 1. Three-column text link navigation
图 1. 三列文字链接导航

Nav1	Nav2	Nav3	Nav4
Nav5	Nav6	Nav7	Nav8
Nav9	Nav10	Nav11	Nav12

Figure 2. Four-column text link navigation
图 2. 四列文字链接导航

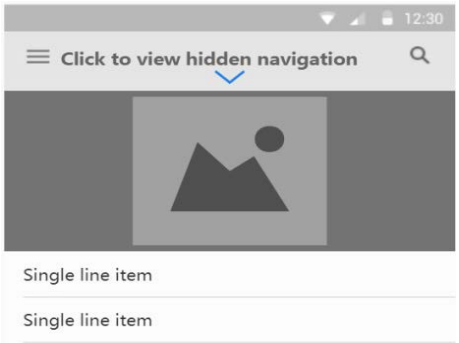

Figure 3. Hidden navigation
图 3. 隐藏式导航

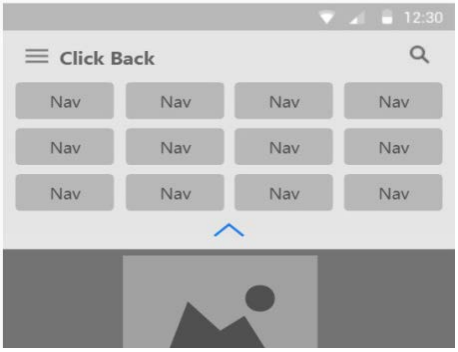

Figure 4. Hidden navigation expansion
图 4. 隐藏式导航展开

当一个页面中所有的页面块均匹配了相应的模板后，也就完成了页面的适配过程，如果页面块分类过程中出现了其它类，则此页面块会适配通用样式模板。通用样式模板只对页面中图片、列表和链接等组成网页最基本的元素进行处理，不添加影响网页结构的样式。图 5 为完整的页面模板示例。

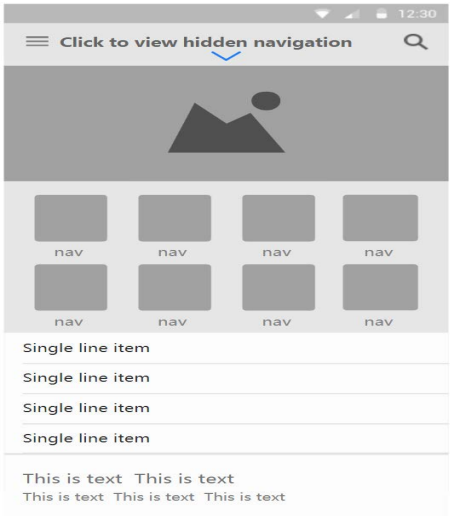

Figure 5. Full page template example
图 5. 完整页面模板示例

4. 实验与结果分析

实验分为两个部分：页面块分类方法的准确性测试、经适配后的网页在移动终端的显示效果测试。其中，页面块分类测试之前，首先要进行网页的分割。

4.1. 网页分割算法的选取

网页分割算法主要分为基于视觉特征、基于文本信息、基于标签以及基于 DOM 树的方法。其中作为当前广受关注的分割方法，DOM 树法是利用了 HTML 文档经过 DOM 解析后可以形成准确描述网页中各个元素之间层次关系，且便于计算机处理的树型结构这一特点而提出的。本文采用的网页分割算法为一种改进的基于 DOM 树的网页分割算法[10]，综合考虑了 DOM 树节点的内容特征和结构特征，利用节点信息熵和最大子树文本密度来衡量一个节点是否为独立的页面块，最后通过节点融合生成网页的分割结果。该算法相对准确且能快速地对网页进行分割。

4.2. 页面块分类方法测试

本文随机选取了新闻类网站、企业机构的门户网站、论坛社区类网站的主页及其子页等 100 个结构不同的网页作为分类方法的测试对象。设网页 P_i 为网页集中第 i 个网页($1 \le i \le 100$)，A_i 为 P_i 的页面块分类准确度，其定义如公式(1)所示：

$$A_i = \left(C_i \times 100\%\right) / T_i \tag{1}$$

其中，C_i 为 P_i 中正确分类的页面块数，T_i 为 P_i 经分块算法分割形成的页面总块数。则评价所有页面块分类准确度 A 的定义如公式(2)所示：

$$A = \left(\sum_{i=1}^{100} A_i\right) \Big/ 100 \tag{2}$$

测试方法为页面块自动分类后的类别与人工标注的类别进行对比，类别相同即为分类正确，反之分类错误。经实验测得的准确度 A 为 78.15%。通过对错误分类的页面块进行分析，发现造成错误分类的主要原因是网页分割错误，将原本应属于同一区块的内容进行了过度分割，或没有将本应属于多个页面块的内容进行分割。图 6 为错误分割的页面示例。

Figure 6. The webpage segmentation result of the original algorithm
图 6. 原始算法的 Web 页面分割结果

通过对网页分割算法的阈值进行手动调整(阈值在算法中决定分割的粒度)，使其能够正确地对网页进行分割。对分割结果进行修正后，经实验再次测得的分类准确度为 89.52%，此时造成错误分类的主要原因包括页面块结构的特殊性、页面块 HTML 代码的不规范，以及存在无法自动修复的语法错误等。

4.3. 样式模板匹配实验

样式模板匹配以国内某高校门户网站的主页为例进行测试，该网页的 PC 版页面经调整阈值后的页面分割算法分割后的结果如图 7 所示。其中，每个矩形框内的内容即代表经页面分割后此区域的内容为一个独立页面块，该页面经适配后的结果如图 8 和图 9 所示。

图中可见，该页面分割后的页面块被归类为页眉、导航、轮播图、图文链接导航、链接列表等类型，并分别匹配了对应的样式模板。其中，原网页中文字导航的数目为 13 个，此时导航块采用隐藏式导航的样式模板；图文链接导航数量为 9 个，采用九宫格式的样式模板；原网页中的图片进行了缩放处理，链接列表也进行了自动换行处理。样式模板匹配完成后，将所有页面块在垂直方向重新组合即可形成适合在移动端浏览的完整的页面，此时该页面的页面元素布局和字体大小比较合理，用户浏览页面时只需上下滑动，无需缩放或左右滑动页面，部分不适合在移动端显示的内容也做了精简处理。经模块化样式模板匹配后的页面非常适合在移动端进行显示且更符合用户的操作习惯。

Figure 7. The webpage segmentation results after adjusting the threshold
图 7. 调整阈值后算法的 Web 页面分割结果

Figure 8. Adapted page 1
图 8. 适配后的页面 1

Figure 9. Adapted page 2
图 9. 适配后的页面 2

5. 结语

本文提出了一种模块化样式模板匹配方法。利用网页分割结果，将结构简单、类型有限的页面块作为模板匹配的基本单元，相对于传统的基于网页的整体匹配，模块化样式匹配更加灵活也更易实现，且通过组合简单的样式模板就能应对较为复杂的页面，使其更符合移动端页面的显示要求，用户体验更好。在未来的工作中，可以将页面块分类规则与现有的视觉信息提取技术和文本的语义信息分析技术相结合，以提高分类的准确率。

基金项目

本文受辽宁省教育厅科学技术研究项目：产业信息服务平台的微信平台迁移技术研究(LFGD2017014)资助。

参考文献

[1] 中国互联网络信息中心. 第 44 次中国互联网络发展状况统计报告[EB/OL]. http://www.cnnic.net.cn, 2019-08.

[2] 文星. 基于移动终端适配技术的网站页面信息显示方法[J]. 自动化与仪器仪表, 2019(12): 126-129.

[3] 马倩. 基于下一代 Web 技术的移动网页模板匹配的设计与实现[D]: [硕士学位论文]. 北京: 北京邮电大学, 2015.

[4] 任胜兵, 王志健, 王宇. Web 页面自动化设计中布局挖掘和样式匹配算法[J]. 计算机工程与应用, 2018, 54(3): 227-232.

[5] 王宪发, 郭岩, 刘悦, 等. 基于视觉特征的网页信息抽取方法研究[J]. 中文信息学报, 2019, 33(5): 103-112.

[6] Zeleny, J., Burget, R. and Zendulka, J. (2017) Box Clustering Segmentation: A New Method for Vision-Based Web Page Preprocessing. *Information Processing & Management*, **53**, 735-750. https://doi.org/10.1016/j.ipm.2017.02.002

[7] 李进生, 乐惠骁, 童名文. 基于标题机器学习的网页分割方法[J]. 计算机科学, 2018, 45(S1): 583-587.

[8] 彭红超, 童名文. 基于规则的网页分割预处理算法研究[J]. 计算机科学, 2013, 40(11A): 379-388.

[9] 韦佳佳. 基于 HTML5 语义化标签的 Web 文本提取技术[J]. 贵阳学院学报: 自然科学版, 2017, 12(3): 25-28.

[10] 杨坤. Web 信息系统的微信平台迁移技术研究[D]: [硕士学位论文]. 沈阳: 沈阳工业大学, 2017.