第 33 卷 第 8 期 计 算 机 工 程 2007 年 4 月
Vol.33 No.8 Computer Engineering April 2007
·软件技术与数据库· 文章编号：1000—3428(2007)08—0083—03 文献标识码：A 中图分类号：TP393

基于页面分解的个性化门户网站构建方法

李军怀[1]，孙 健[1]，张 璟[1]，刘丽娟[2]
（1. 西安理工大学计算机科学与工程学院，西安 710048；2. 大连交通大学软件学院，大连 116028）

摘 要：在分析影响用户感知时间的诸多因素基础上，针对网络带宽一定、网络绝对流量增加的情况，从门户网站页面的静态表现结构多半由一些相对独立的区域（主要是 Table tags 分割的一些区域）构成这一特点出发，研究了通过页面分解技术构建个性化门户网站用以缩短 Web 网站对用户 click 的响应时延，从而提高用户满意度。

关键词：门户网站；Web component；XHTML；树

Method of Personalized Portal Construction Based on Web-page Fragmentation

LI Junhuai [1], SUN Jian [1], ZHANG Jing [1], LIU Lijuan[2]
(1. School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048;
2. Software Institute, Dalian Jiaotong University, Dalian 116028)

【Abstract】 By analyzing the factors which affect user perceived latency, a method is presented that allows the viewers of Web sites to build their own personalized portals, using specific thematic areas of their preferred Web sites. This method is based on an algorithm which fragments a Web page in discrete fragments with pages' internal structure. The method can efficiently reduce the response time and improve user satisfaction in the limited network bandwidth and the rising of network flow over network.

【Key words】 Portal; Web component; XHTML; Tree

随着互联网的普及，人们越来越习惯于将工作、娱乐、通信等行为在网上进行，网络与人们的学习、生活、工作结合得越来越密切。特别是在过去 2 年中，门户网站已经成为当今电子商务的一个主要趋势，它集成了所有与企业相关的数据和服务，是一种信息与服务的Web集成。简而言之，门户是呈现给门户使用者的单一的公共入口点，通过这个入口点，门户网站的使用者可以根据各自的角色和权限进行各种相应操作[1]。然而，近几年由于接入Internet 的用户数量剧增及Web服务和网络固有的延迟，使得网络越来越拥挤，用户的服务质量（QoS）不能得到很好的保证。如果门户网站能够很快的处理浏览器所发出的HTTP连接要求，则使用者就不会因为等待网页下载的时间过长而放弃继续浏览网页，从而增强使用者在网站上消费的意愿。

根据Zona Research[2]的研究指出，如果使用者等待下载网页的时间超过 8s，将有 30%的用户选择停止浏览网页，同样的研究表明，如果下载网页的时间缩短 1s，则这个数字将从 30%降低到 8%。因此，如何缩短终端用户所感到的时间延迟（感知时间，user-perceived latency），提高网站性能已经成为Internet性能研究中的一个重要问题。

1 个性化门户实现框架

1.1 相关技术

目前 Web 网站的页面主要用标记语言来表现。文献[3~6]中首次提出了基于页面分解思想的个性化门户网站构建技术，主要依据 HTML 文件的特点直接构建 Tree，并分析提取 Web component。这为定义 Web component 从而实现页面分解提供了前提和基础。

1.2 实现框架

个性化门户具体实现框架如图 1 所示。

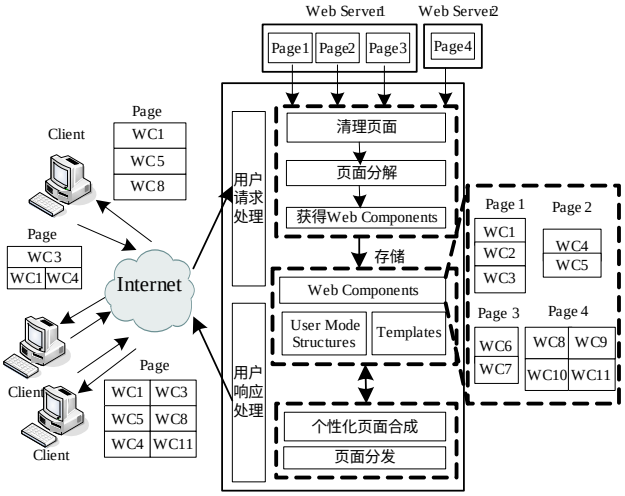

图 1 个性化门户实现框架

在现有的WWW站点中，几乎所有的网站都使用某种静态的结构来为用户提供信息。即使网站的内容时常更换，其静态结构也很少改变。在内容丰富的站点中（如门户网站的

基金项目：国家“863”计划基金资助项目(2002AA414060)；2004年西安市集成电路与软件专项资助项目(ZX04011)
作者简介：李军怀(1969－)，男，副教授，主研方向：分布式计算，CSCW；孙 健，硕士生；张 璟，教授、博导；刘丽娟，硕士生
收稿日期：2006-06-30 **E-mail：**sunjian@xaut.edu.cn

页面），这一结构往往由某些通用的语义内容区域组成（如搜狐主页中的新闻区域、商品销售区域等），这些区域被称为“Web component”或“Web fragment”[3]。针对Web站点中页面的这一特点，可以将页面按照原来的语义区域进行分解，通过分析、提取其所包含的Web component，并给这些WC分配标识，最后将这些component以独立实体的形式提取出来存储在数据库，然后帮助网站浏览者根据其感兴趣的站点主页面的主干区域来构建个性化页面，使用户在一个单个的页面中获取其在多个页面中感兴趣的区域。

2 实现的关键技术

2.1 门户网站页面分析

浏览器所显示的页面是基于 HTML 文件来实现的，而 HTML 文件中的标记(tags)是嵌套的，这意味着页面的 HTML 代码可以用一个 HTML Tree 来表示，因而可以通过抽取 HTML Tree 中的节点来获取页面中不同的 WC。

事实上抽取页面中 WC 的过程就是一个将分解算法应用于 HTML 文件的过滤过程。HTML 文件代码有很多，但是由于表格是最稳定、最常用的结构，因此许多 Web 站点使用 Table 来构建他们的框架。这样，对页面的分解就集中于对 Table 标签的分析和提取了。如果仅保留 HTML Tree 中的 Table 节点，HTML Tree 的复杂性就会大大降低。然后基于每个节点内容（文本长度），选择 component 的节点。

2.2 页面分解算法

(1)Web component 的抽取标准

根据Bouras Christos等人的想法[4]，如果一个节点集p满足下列条件，那么它可以被标记为一个WC，而不必关心这个节点集的内部结构。

$$AverageRatio \le Ratio(p) \le 2AverageRatio \quad (1)$$

式(1)表示了选取 WC 的原始标准，即一个 WC 相对于整个页面必须是中等大小的。其中，*Ratio(p)*表示节点内的纯文本长度与整个页面的文本长度之比，*AverageRatio* 表示假设页面中所有内容节点大小相等时的节点文本平均长度。

$$Ratio(p) = \frac{Pure\ Text\ Length\ in\ node\ p}{Pure\ Text\ Length\ of\ the\ root\ node} \quad (2)$$

$$AverageRatio = \frac{1}{Number\ of\ content\ nodes} \quad (3)$$

确定 Web Component 还有一些其他需要考虑的标准，其中之一是基于 Index Tree 的结构。一个看作 WC 的领域往往含有多个 Table Tag，其中只有一个是 WC 节点集中的根节点。因此，当分解算法找出了 Index Tree 中的一个节点，该节点拥有小于 4 个孩子节点，并且总共有少于 5 个子孙节点的时候，这个节点及其子节点就被确定为一个 WC。如果分解算法到达了叶子节点，那么就将叶子节点选择为一个 WC，以保证内容的完整性。

(2)页面分解算法

本文基于格式良好的 HTML 门户页面，进行 WC 的定义，然后进行页面分解，分析、提取并存储 WC。页面分解过程主要包括以下几个步骤：

1)获取 Web 页面的最新实例，进行页面预处理，使其符合 Well-formed 要求；

2)将 HTML 页面文件转换为 XHTML 文档，并构造 DOM Tree；

3)分析 DOM Tree，并生成 Index Tree；

4)分析 Index Tree，计算并标记 WC 的节点；

5)抽取并存储 WC 及其标识。

下面举一个例子说明页面分解方法。例如，某汽车销售门户页面（图 2）。

图 2 某汽车销售门户页面

对图 2 所示页面的 HTML 文档进行预处理并转换为 XHTML 文档，并生成如图 3 所示的 DOM Tree。由于文档树本身并没有为各个节点建立索引标记，为了实现对文档树操作后与原文档树的映射关系，通过一个广度优先遍历方法为每个节点增加了一个 Index 属性并赋予索引值，标注在图 3 中的节点上。建立索引后的树就是一个 Index Tree，在 Index Tree 中每个节点都有索引值属性。如图 2 中的 Table1、Table2、Table3 分别对应节点 12、13、14。

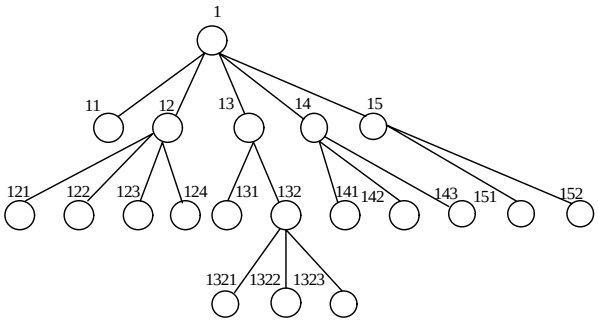

图 3 HTML 的索引树

Index Tree 建立后，重新建一个空的文档树，以<html>节点为根节点，对原文档树进行深度遍历，在遍历的同时，将文档树中的 Table Tag 节点找出来，每找到一个 Table Tag 节点就将其按原文档树中的相应位置插入到新建的文档树中，这样就构建了一个由 Table Tag 组成的树，如图 4 所示。然后广度遍历每个节点并根据式(1)进行计算，抽取 WC，图 4 中实心节点及其子孙节点表示 WC。

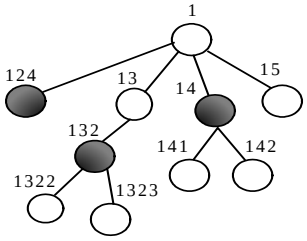

图 4 由 Table Tags 构成的 DOM Tree

最后，如表 1 所示，对于得到的 Web Components 以矩阵的形式存储起来，以满足用户获取个性化门户的需要。

表 1 Web Component 的存储矩阵

Web Component ID	1	2	3
Index	124	132	14
Attributes	Name=“菜单” ……	Name=“价格消息” ……	Name=“资讯” ……

表 1 显示，分析 DOM 树后得到了 3 个 WC，每个 WC 有一个 ID，并对应一个索引值和多个属性，这些索引值和属性就是用来抽取相应 WC 以构建个性化门户的依据。当页面变化时，这些数据还将被提取出来进行比较和更新。

2.3 个性化页面的创建

Web 页面分析与分解的目的是从 Web 页面中获取为用户创建最新的用于个性化页面的 WC。因此，个性化页面创建的目标是面向用户。其目的是创建一个 WC 的列表，用户可以通过增删列表中的 WC 来得到包含自己感兴趣的内容的个性化页面。

实现跟踪用户操作信息（感兴趣的页面区域）的技术之一是在程序中嵌入脚本语言(如JavaScript)，然后使用Xpath表达式来跟踪记录用户在某一页面中点击的WC区域的属性。当用户再次登录时，系统自动根据已经存储的Xpath表达式来提取用户所需要的WC[4]。而个性化页面合成是通过执行Web服务器上的脚本实现的，使用用户所选择的WC的HTML代码来构造的。在个性化页面合成的过程中，来自页面的WC必须遵循一个特殊的过程，使用CSS、JavaScript或者VBScript。

3 测试结果及分析

本文利用集成在 VS.NET 中的 ACT（Application Center Test）工具在一台带有 Intel Pentium 4 CPU 的 Dell 品牌机（主频为 2.4GHz，RAM 为 256MB，HD 为 38GB）上进行测试。ACT 能收集和汇总每个请求的所有响应时间，使用收到第一个字节的时间（TTFB）和收到最后一个字节的时间（TTLB）来准确地表示响应时间。利用 ACT 工具，在相同的测试条件下我们分别对原始页面和采用分解技术后的用户感兴趣的个性化页面进行测试，得到了如图 5 和图 6 所示的 RPS 曲线图以及详细的测试结果。

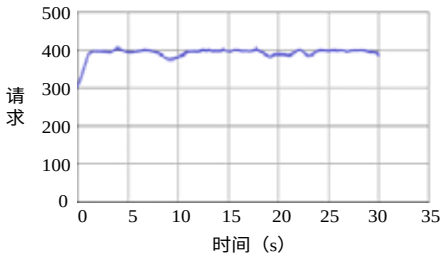

图 5 原始页面的 RPS 曲线图

从图 5、图 6 和表 2 中可以看出，个性化页面无论是在每秒平均请求数还是 TTFB、TTLB 中都比原始页面提高了很多。个性化页面的 RPS 值比原始页面上升了 100 多，同时平均响应时间也缩短了很多，极大地提高了系统的性能，缩短了页面显示的响应时间。

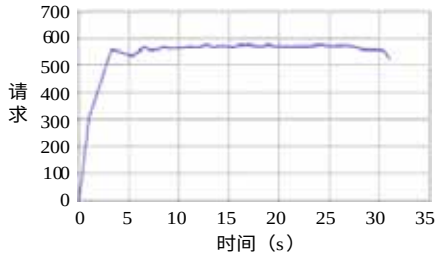

图 6 个性化页面的 RPS 曲线图

表 2 原始页面和个性化页面测试结果比较

	原始页面	个性化页面
请求总数	12 493	17 186
连接总数	12 492	17 185
每秒平均请求数	403.00	572.87
首字节平均响应时间(ms)	1.26	0.60
末字节平均响应时间(ms)	1.31	0.64
每次迭代末字节平均响应时间(ms)	10.51	10.22

4 结束语

为了满足用户需要，提高网络传输效率，减少延迟，本文的页面分解算法使得构建个性化的页面成为可能。但是在实现算法的过程中，由于网页的页面源代码的复杂性，因此在对页面进行分解的时候采用了一些较为理想的页面。这些页面完全符合 XML 规范。相信在处理各种不同页面的时候必然还会遇到一些新的问题，例如一些页面源代码中加入的脚本（如 JScript）是不能被 XMLParser 分析的，在清理页面的时候还要注意这些细节部分。

参考文献

1 Allison C, Bain A, Ling B, et al. MMS: A User-centric Portal for E-learning[C]//Proc. of the 14th International Workshop on Database and Expert Systems Applications. 2003.

2 Zona Research, The Economic Impacts of Unacceptable Web-site Download Speeds[Z]. 1999-04. http://www.webperf.net/info/wp_downloadspeed.pdf.

3 Christos B, Vaggelis K, Ioannis M. A Web Page Fragmentation Technique for Personalized Browsing[C]//Proc. of SAC’04. 2004.

4 Misedakis I, Kapoulas V, Bouras C. Web Page Fragmentation for Personalized Portal Construction[C]//Proc. of the International Conference on Information Technology: Coding and Computing. 2004.

5 Freire J, Kkumar B, Lieuwen D. WebViews: Accessing Personalized Web Content and Services[C]//Proc. of the 10th International Conference on World Wide Web. 2001: 576-586.

6 Bouras C, Konidaris A. Web Components: a Concept for Improving Personalization and Reducing User Perceived Latency on the World Wide Web[C]//Proc. of the 2nd International Conference on Internet Computing, Las Vegas. 2001.