基于视觉特征的网页最优分割算法

李文昊[1] 彭红超[1] 童名文[1] 石俊杰[2]
(华中师范大学教育信息技术学院 武汉 430079)[1] (解放军63981部队 武汉 430311)[2]

摘 要 网页分割技术是实现网页自适应呈现的关键。针对经典的基于视觉的网页分割算法VIPS(Vision-based Page Segmentation Algorithm)分割过碎和半自动的问题,基于图最优划分思想提出了一种新颖的基于视觉的网页最优分割算法VWOS(Vision-based Web Optimal Segmentation)。考虑到视觉特征和网页结构,将网页构造为加权无向连通图,网页分割转化为图的最优划分,基于Kruskal算法并结合网页分割的过程,设计网页分割算法VWOS。实验证明,与VIPS相比,采用VWOS算法分割网页的语义完整性更好,且不需要人工参与。

关键词 网页最优分割,网页视觉特征,网页自适应呈现,最优划分
中图法分类号 TP311.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.058

Web Page Optimal Segmentation Algorithm Based on Visual Features

LI Wen-hao[1] PENG Hong-chao[1] TONG Ming-wen[1] SHI Jun-jie[2]
(School of Education Information Technology, Central China Normal University, Wuhan 430079, China)[1]
(63981 Ttroops, The Chinese People's Liberation Army, Wuhan 430311, China)[2]

Abstract The Web page segmentation technique is a key point to realize Web page adaptive presentation. To overcome the shortcomings of the classical Web page segmentation algorithm VIPS(Vision-based Page Segmentation Algorithm) including fragmented content and semi-automatic, a novel Web page segmentation VWOS(Vision-based Web Optimal Segmentation) was proposed based on the optimal division of graph. The Web page is constructed as the weighted undirected connected graph from the perspective of visual features and structure of the Web page. Therefore, the problem of Web page segmentation is transformed into the optimal division of graph. VWOS was designed by combining Kruskal algorithm and the process of the Web page segmentation. It was proved by the experimentation that the effect of Web page segmentation produced by VWOS is better than that by VIPS.

Keywords Web page optimal segmentation, Web page vision features, Web page adaptive presentation, Optimal division

1 引言

随着移动通信技术的迅猛发展,人们通过移动终端访问网页的活动日渐频繁。然而,移动终端屏幕尺寸的限制往往造成Web页面无法正常显示,给用户带来了很大的困扰。为了解决这个问题,早在20世纪90年代,研究人员便开始研究网页自适应呈现技术,提出了若干算法。这些算法可归纳为3类,即网页重构、网页转码、网页分割。其中,网页分割是实现网页自适应呈现的主流技术之一。它首先将网页分割成若干个语义相关的内容段(也称为内容块);然后在内容服务过程中,服务器根据移动终端特征,选择合适的内容段并推送给用户,以确保网页内容在移动终端上得以正常显示。网页分割技术具有两个优点:一方面,它不需要占用大量的计算资源;另一方面,用户也不需要反复拖动滚动栏查看网页内容,使网页内容的服务质量得以保证。

近年来,关于网页分割技术的研究受到了广泛关注,并且取得了丰富的研究成果。其中经典算法是Cai等研究人员提出的基于视觉的网页分割技术(Vision-based Page Segmentation Algorithm, VIPS)[13]。VIPS根据人的视觉特点,总结出一些网页分割的规则,然后基于这些规则实现网页分割。此后,许多研究者在该方法的基础上提出了许多改进的网页分割技术,但基于规则的思想没有本质变化。目前,基于视觉的网页分割技术主要存在两方面问题:其一,网页分割结果过碎,不利于网页重构;其二,分割规则的总结需要人工参与,规则的好坏也直接影响网页分割效果。因此,如何划分网页分割的粒度,如何能减少分割过程中人工参与,从而降低主观因素影响,均是需要进一步研究的问题。

本文将网页分割转化为图的最优划分问题,提出一种新

到稿日期:2014-10-08 返修日期:2015-03-23 本文受教育部科技发展中心网络时代的科技论文快速共享专项研究资助课题:基于学术社交网络的多粒度科技论文共享技术研究(2013123),中央高校基本科研业务费项目:内容适配系统中最优适配决策器模型及分布式寻优算法研究(CCNU14A02012)资助。

李文昊(1977—),男,博士,副教授,主要研究方向为多媒体学习理论与技术;彭红超(1987—),男,博士生,主要研究方向为智慧学习理论与技术;童名文(1975—),男,博士,教授,主要研究方向为多媒体内容适配技术、教育资源管理,E-mail:tmw@mail.ccnu.edu.cn(通信作者);石俊杰(1975—),男,高级工程师,主要研究方向为机械故障诊断。

颖的网页最优分割算法(Vision-based Web Optimal Segmentation,VWOS)。VWOS算法首先基于人的视觉特点设计内容相似度计算模型,然后利用网页结构特征和内容相似度模型,将网页构造为加权无向连通图,并将网页分割转化为图的最优划分问题,最后基于Kruskal算法求解图的最优划分问题,实现网页最优分割。VWOS算法是一种自动算法,不需要人工参与。实验分析表明,该算法能够有效地对网页进行分割,分割效果和算法性能优于VIPS算法。

2 相关研究

网页是一类特殊的文本文件,它具有内容特征、结构特征、布局特征和视觉特征。针对上述4种特征,网页分割技术可以分为4种类型:基于内容特征的分割技术、基于结构特征的分割技术、基于布局特征的分割技术和基于视觉特征的分割技术。基于内容特征的网页分割技术主要是基于网页标签。20世纪90年代末的手机浏览器不支持CSS层叠样式,也不支持JavaScript,只能访问简单的静态网页。因此,当时的学者只需基于标签的类型进行分割,即可达到很好的效果。Yanlei Diao等人提出具有自学习功能的Web查询处理系统[1],利用有效标签类型(如〈p〉、〈table〉、〈ul〉、〈h1〉~〈h6〉)进行网页分割;Wai-ching Wong提出标签检测算法[2]来检测具有同类型信息的相似标签,并定义标签类型进行网页分割;Eija Kaasinen[3]与Orkut Buyukkokten[4]仅仅利用像〈p〉〈table〉〈ul〉这样的简单标签进行Web网页分割。基于结构特征的网页分割技术采用了DOM(Document Object Model,DOM)技术,将网页表示成DOM树结构,然后根据各内容块在DOM树中的位置对网页进行分割。文献[5-7]均采用的是基于DOM树的分割技术,Richard Romero[8]在DOM树的基础上进行聚类分析,实现网页分割。基于布局特征的网页分割技术主要包括基于位置的网页分割技术与基于模板的网页分割技术两种。Gen Hattori提出的基于距离的网页分割技术[9],利用标签的相对位置与层级关系计算内容块的距离,以此对网页进行分割。然而HTML中某些特殊标签具有布局作用,降低了分割的准确率。通过对HTML标签的研究与分析,Gen Hattori于2007年提出改进技术:混合分割技术[7]。混合分割技术将〈div〉与〈table〉作为布局信息,进行初步分割,之后将标签间的距离作为内容块的距离做二次分割。基于模板的网页分割技术的主要思想是分割前定义好各类模板,通过将欲分割的网页或内容块与模板匹配来进行分割。Yu Chen[10]将网页分成上、下、左、右和中间5个部分,之后根据这5个部分的特征将网页的内容提取后纳入到定义好的特征模版中,从而实现网页分割。这种技术适合于结构标准的网页,对于其他结构的网页则无法正确分割。文献[11,12]将网页归类于八大布局模板,之后依据网页布局(此处考虑的是标签形成的布局,而非样式信息形成的布局)与标题块进行网页分割。基于视觉特征的网页分割技术的原理是标签本身携带内容显示信息,根据人眼的视觉特征,利用这些显示信息实现网页的内容分割。Deng Cai提出了一种基于视觉特征的网页分割技术(Vision-based Page Segmentation Algorithm,VIPS)[13],该算法具有良好的网页分割效果。VIPS存在的问题在于需要人工不断地去总结和调整分割规则,而且当新规则产生后,将影响以前的分割效果。基于VIPS算法,国内外学者提出了一系列的改进技术[14-18],这些技术在一定程度上优化了VIPS,但上述的本质问题却没有解决。此外,这些技术均没有考虑CSS样式信息对视觉特征的影响。

3 VWOS算法设计

根据网页的标签,可以将网页划分为许多语义完整的原子内容块,这些内容块是网页内容的最小组成单元。基于网页视觉特征定义两个原子内容块的相似度计算公式,并利用该公式构造原子内容块相似度矩阵。因此,网页可以视为由原子内容块为顶点、相似关系为边、相似度为权的加权无向连通图,网页分割就转化为图的最优划分问题。

3.1 网页最优分割模型

为便于表述网页最优分割模型,对其中包含的重要概念做如下定义:

定义1 网页 $P=\{AC_1,AC_2,\cdots,AC_n\}$,其中 AC_i 是网页中不可再分的原子内容块。

定义2 内容块 $CS=\{AC_1,AC_2,\cdots,AC_m\}$,$m<n$,是 P 的子集。

定义3 原子内容相似度 ACS_{ij} 是原子内容块 AC_i 与 AC_j 内容语义相似性测度。

定义4 原子内容块相似度矩阵 $S[n][n]$是网页原子内容块两两相似度构造的数值矩阵。

定义5 块相似度 $CSSD$ 是内容块中所有元素相似度之和。

定义6 内容块面积 CSS 是内容块所占据的像素面积。

通过解析网页得到内容块,并利用内容块相似度公式计算内容块两两之间相似度,得到相似度矩阵。在此前提下,网页可以构造为加权无向连通图。因此,网页分割转化为图的最优划分问题,其最优化模型如式(1)所示:

$$\begin{aligned}&\text{求网页 } P \text{ 的一个内容块划分}\{CS_1,CS_2,\cdots,CS_l\},\text{满足:}\\&\forall CS_k\subset P\ Max.CSSD_k\\&\text{s.t. } CSS_k\leqslant S_t^k\end{aligned}\tag{1}$$

其中,CS_k 是网页的内容块 k,S_t^k 是内容块 k 像素面积阈值。

式(1)的最优化模型具有3个典型性质:最优子结构、重叠子问题与贪心选择性质。最优子结构指问题的最优解包含子问题的最优解。如果上述问题的最优解包含了原子内容块 n,那么其余原子内容块一定构成子问题 $n-1$ 个原子内容块在组阈值为 S_t-S_n 时的最优解。如果最优解不包含原子内容块 n,那么其余原子内容块一定构成子问题 $n-1$ 个原子内容块在组阈值为 S_t 时的最优解。重叠子问题指用递归算法自顶而下解决上述问题时,每次产生的子问题并不总是新问题,有些子问题被反复计算了多次。贪心选择性质指所求问题的整体最优解可以通过一系列局部最优解的选择来达到。若所有原子内容块构成的集合为 V,组内已确定的原子内容块构成的集合为 U,$s[u][v]$表示原子内容块 u、v 间的相似度。$u\in U$,$v\in V-U$,由于 U 中所有原子内容块构成一棵相似度最大的生成树,根据MST性质,V 的所有相似度最大的生成树中一定存在一棵包含边(u,v)。

3.2 VWOS算法

网页最优分割算法VWOS分为4个步骤:第一,根据手机分辨率信息,确定网页分割阈值 S_t;第二,建立原子内容块池 P_c 与相似度池 P_s,相似度按值从大到小排序;第三,构建网页加权无向连通图 $G(V,E)$;第四,求解图 $G(V,E)$ 最优划分问题。

(1)网页分割阈值确定

不同手机的分辨率不同，所能呈现的信息量亦不同。因此网页分割时，需要设置不同的阈值，以达到在不影响正常显示与用户体验的情况下，子页所呈现的信息量最大化。网页分割算法采用像素面积确定分割阈值 S_t。使用诺基亚5800W手机，随机浏览100个手机版网页，统计分析知，平均每个网页需要3.51屏显示；随机抽取100位大学生手机网民，对"手机版网页出现几屏显示时，会心生埋怨心理"问题进行调查，分析发现其平均值为2.87。考虑到用户体验的重要性，对上述两个结果以比例1:2进行加权求均值得3.08。因此，可确定最适合手机显示的网页大小为手机屏幕的3倍。由于VWOS算法所用的网页分割算法的特殊性，有时分割形成的子页结果为所设阈值的2倍。对此，将网页分割阈值 S_t 设为1.5屏，即 $S_t=1.5\times$水平分辨率$\times$垂直分辨率。

(2)原子内容块池 P_c 与相似度池 P_s 的建立

为了更有效地实现网页分割，需要建立原子内容块池 P_c 与相似度池 P_s。P_c 中存放原子内容块算法获得的所有内容块。原子内容块相似度计算及后期建立连通分支时均从 P_c 中获取所需的内容块。P_s 中存放采用基于网页视觉特征的相似度公式得到的相似度值，并按值从大到小的顺序排列。内容相似度基于网页视觉特征，在此将网页视觉特征定义为6维向量，根据向量中维度的不同度量属性，采用不同的计算公式计算各维度相似度值，最后用加权求和的方式计算出最终的内容块相似度值[19]。P_c 存放的内容块类型为PerfectNode，而 P_s 存放的相似度以本文自定义的Similarity类型标识，其类图如图1所示。对 P_s 而言，由于构建连通分支时，以相似度值从大到小的顺序连通各分支，因此，P_s 中的Similarity数据是按值递减的有序队列。

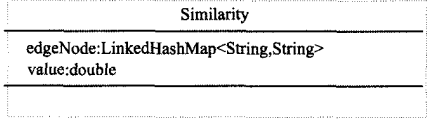

图1 Similarity类

(3)连通分支构建

P_c 中含有网页所有的原子内容块，P_s 中含有两原子内容块间的相似度值，并按值从大到小排列。对 P_c 与 P_s 采用如下算法构建连通分支，以确保每个分支的相似度权值最大，且每个分支中所有顶点的像素面积和 S_k 均小于分割阈值 S_t。这样便实现了网页分割所需的每个连通分支均可转换为各个子页，这些子页不仅可在手机浏览器中正常显示，而且具有较好的用户体验。连通分支构建算法如图2所示。

1)将 P_c 中 n 个内容块看成 n 个孤立的连通分支，并建立关联池cr。

2)计算各连通分支像素面积和 S_k，并与 S_t 比较：

如果 $S_k\geqslant S_t$ 或所有边都被查看过，则将连通分支中的顶点从 P_c 中取出存入关联池cr中；

如果 $S_k<S_t$，按下述方法连接两个不同的连通分支：设查看到第s条边，若该边两端点分别是当前两个不同的连通分支T1和T2中的顶点，则用该边将T1和T2连成一个连通分支；若该边两端点在当前的同一个连通分支中，直接查看第s+1条边。

3)如果所有边都被查看过或 P_c 中已经没有原子内容块，则连通分支构建算法结束，否则转步骤2)。

图2 连通分支构建算法

需要特别指出的是，若以是否含有孤立的连通分支或查看边数是否达到 $n+1$ 作为结束算法的条件，虽然可以大幅度减少算法循环次数，但是，却不能保证最后生成的连通分支不能再合并。而仅仅以判断所有边是否被查看过作为结束算法的条件，虽然可以保证 S_k 在小于 S_t 的情况下最大化，然而因需要过多的循环次数导致时间复杂度过大，从而影响VWOS算法的性能。通过分析与测试发现，多数情况下，各连通分支均不能再合并时，有很多边没有被查看，对此，连通分支构建算法采用"所有边都被查看过或 P_c 中已经没有原子内容块"作为算法结束条件，以达到在满足 S_k 最大化的同时性能最大化。

(4)求解网页最优分割问题

网页最优分割模型如式(1)所示，基于模型的最优子结构和贪心选择性质，可采用贪心策略求解该模型。因为，加权无向连通图 $G(V,E)$ 可构造为一棵生成树，使生成树上各边权值最大，于是网页分割可变为在特定阈值 S_t 条件下构造子生成树的过程，每个子生成树均满足特定阈值 S_t。采用最优化理论中的Prim算法与Kruskal算法所需的时间复杂度分别为 $O(n^2)$ 与 $O(e\log e)$，其中 e 为图的边数。当 $e=\Omega(n^2)$ 时，$O(n^2)<O(e\log e)$，当 $e=O(n^2)$ 时，$O(n^2)<O(e\log e)$。因为网页对应的加权无向连通图 $G(V,E)$ 是一张完全图，即 $e=n(n-1)/2=O(n^2)$，所以用Kruskal算法比用上述其它算法时间复杂度低。因此，本文采用Kruskal算法实现网页最优分割问题求解。

4 实验与分析

为检验VWOS算法的执行效果和性能，设计了一组网页分割对比实验。实验基于Web服务设计，通过移动终端访问网页。将VWOS算法和VIPS算法部署在代理服务器，以国家精品课程站点随机选取的100个网页为对象，移动终端采用分辨率为360×640像素的Android2.3手机模拟器，分割阈值 $S_t=1.5\times360\times640=3.456\times e5$。采用3个评价指标：平均分割块数、语义完整度和平均执行时间。其中语义完整度定义见式(2)。通过网页在移动终端的呈现结果，比较VWOS算法与VIPS算法在3个指标上的表现，评价算法效果和性能。

$$sid=\frac{\text{语义属于内容块且选中的原子块数}}{\text{内容块总原子块数}+\text{语义属于内容块未选中原子块数}}\times100\% \quad (2)$$

4.1 结果与分析

本节以具体的2个网页呈现的结果为例，比较两种算法分割效果和性能，并分析其中的原因。结合专业背景，选取的网页定为北京师范大学的国家精品教育技术学导论与南京师范大学国家精品课程教育社会学。

图3(a)为北京师范大学的国家精品教育技术学导论经VWOS算法分割的效果。VWOS算法将该网页分割为两个子页，如图3(b)与图3(c)所示，图3(b)为主页，图3(c)为子页。从图中可以发现，VWOS算法分割该页面后得到的两个子页语义完整且适合手机浏览器显示，具有较好的用户体验。

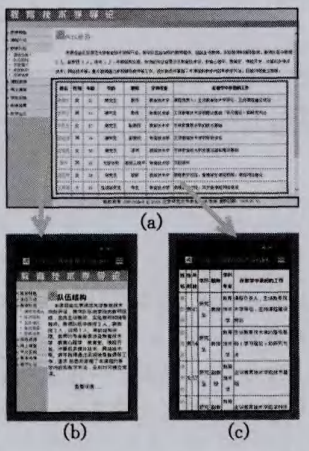

图 3　教育技术学导论 VWOS 分割效果

图 4 为教育技术导论网页采用 VIPS 算法的分割结果。VIPS 将该网页分割为 6 个子页。其中只有表格子页的像素面积与分割阈值接近，而其他 5 个子页尺寸均远小于阈值。VIPS 算法之所以将网页分割得过碎，主要因为其以 DOM 树为基础，对每个内容块用 DoC(Degree of Coherence)表示紧密程度。按照 VIPS 的规则，DoC 在 DOM 树中呈现自顶而下逐渐增大的规律。而 VIPS 采用自顶而下的方式分割，因此当 DOM 树底层的内容块符合分割阈值时，上层内容块因 DoC 小于 PDoC(Permitted Degree of Coherence)而被过度分割。VWOS 算法基于最优化理论，将网页分割看作分组最优化问题，并设计网页分割算法以自底而上的方式对网页进行分割，有效地避免了分割过碎问题。由此可以看出在分割后形成子页数方面，VWOS 算法较 VIPS 算法内容块语义更完整，也更适合移动设备显示。

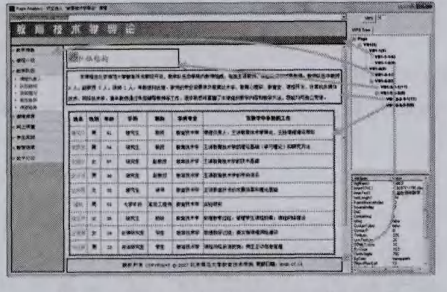

图 4　教育技术学导论 VIPS 分割效果

图 5 为南京师范大学国家精品课程教育社会学分别经 VWOS 与 VIPS 分割的效果。采用 VWOS 算法进行分割后形成两个子页，其中图 5(b)所示子页为原网页右下角部分。该部分像素面积略大于分割阈值 S_t，按照 VWOS 算法设计的网页分割算法，该部分会作为一个完整子页存在。采用 VIPS 算法，将页面分割为 3 个子页，如图 5(c)所示，其中黑色方框内的部分为分割后保留下的内容块。观察图 5(c)很容易发现丢失了部分内容块。分析该网页的代码可发现，丢失的部分均为样式信息，该部分的样式信息存储在 CSS 文件中，而非 HTML 标签的 style 属性中。由此，再一次证明了由于“数据内容-样式信息”的分离，致使 VIPS 分割效果无法满足手机用户需求的假设，也再一次说明了网页分割预处理算法的必要性。VWOS 算法充分考虑了〈link〉、〈style〉与 HTML 标签 style 属性中的样式信息，并将样式信息与数据内容融合，以此保证内容块视觉特征的全面性与精确度。因此可以看出，在分割后形成的内容块方面，VWOS 算法得到的内容块具有语义完整的特点，而 VIPS 算法分割过程中，会造成内容块视觉特征的丢失，甚至会造成部分内容块的丢失。此外观察图 5(b)可以发现，分割形成的子页的像素面积比手机尺寸大，这主要因为该部分采用〈table〉标签进行布局，而 VWOS 算法并未对〈table〉标签的宽高信息进行处理。

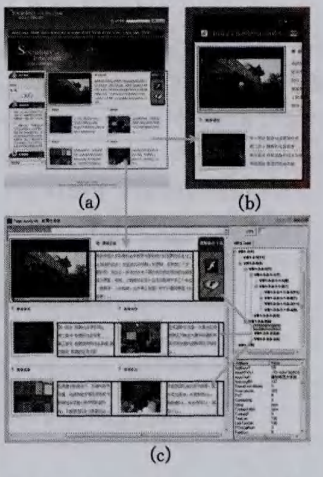

图 5　教育社会学网页经 VWOS 与 VIPS 的分割效果

实验采用 VWOS 算法和 VIPS 算法共对 100 个精品课程站点网页进行网页分割，并在 3 个性能指标上进行统计对比，结果如表 1 所列。

表 1　网页分割性能统计

	平均内容块数(个)	语义完整度 sid	平均执行时间(s)
VWOS	2.3	83.2%	2.47
VIPS	4.8	76.6%	2.43

通过上述实验，初步证明 VWOS 算法在内容块语义完整性和网页适应性方面，比 VIPS 算法具有更好的性能。具体而言，VWOS 算法比 VIPS 算法具有以下 4 点优势：第一，VWOS 算法不需要人工参与，是一种网页分割自动处理方法；第二，在同样的分割阈值条件下，VWOS 算法生成的子页数少，因此用户在各子页中遍历浏览时，不易迷航；第三，VWOS 算法生成的每个子页的像素面积 SP 均在$[S_t, 2S_t)$区域中，没有过度分割的子页；第四，VWOS 算法充分利用视觉特征表示内容块的特征，分割得到的每个内容块均具有高度的语义完整性。

结束语　网页分割技术被广泛应用于网页信息获取和网页自适应呈现等领域。目前，经典的网页分割算法仍存在需要人工参与和分割过碎的问题。针对这些问题，综合视觉特征和网页结构，将网页构造为加权无向连通图，并将网页分割转化为图的最优划分问题，最后基于经典的最优化算法，结合网页分割的过程，提出了一种基于视觉特征的网页最优分割算法 VWOS。实验证明，VWOS 算法在语义完整性和网页适应性方面，性能优于经典分割算法 VIPS。与 VIPS 算法相比，VWOS 算法有两个优点。其一是网页分割结果没有过多的内容碎片，较好地保留了内容块的语义完整性；其二是它属于自动算法，不需要人工参与。当然，VWOS 算法仍存在一些不足之处，集中表现在由于网页样式采用技术不同对构造

(下转第 309 页)

[J]. Journal of Yanshan University, 2012, 36(5): 428-432, 464

[13] Lean G, Moshe B, Eli S, et al. Actions as Space-Time Shapes [OL]. http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

[14] 杨杨，张田文. 一种基于特征光流的运动目标跟踪方法[J]. 宇航学报，2000，21(2)：8-15
Yang Yang, Zhang Tian-wen. Moving target tracking based on feature-optical-flow[J]. Journal of Astronautics, 2000, 21(2): 8-15

[15] 王晋疆，刘阳，吴明云. 基于快速鲁棒特征的 Camshift 跟踪算法[J]. 计算机应用，2013，33(2)：499-502
Wang Jin-jiang, Liu Yang, Wu Ming-yun. CamShift tracking algorithm based on speed-up robust features[J]. Journal of Computer Applications, 2013, 33(2): 499-502

[16] 汪鑫，刘嘉敏，李敏，等. 基于 SIFT 特征匹配算法的目标跟踪及视频采集与传输研究[J]. 重庆理工大学学报(自然科学版)，2014，28(11)：89-93
Wang Xin, Liu Jia-min, Li Min, et al. Target Tracking Video Information Acquisition and Transmission Based on SIFT Feature Matching Algorithm[J]. Journal of Chongqing University of Technology(Natural Science), 2014, 28(11): 89-93

(上接第 287 页)

网页无向连通图 G 影响较大，因此该算法的鲁棒性存在不足。

下一步研究将从 3 方面展开。第一，将采用更多的客观评价指标(如信息检索领域评价指标)，全面对比 VWOS 和 VIPS 两种算法的性能，并以此为依据对 VWOS 算法做改进。第二，在算法中增加对网页样式技术的识别，并做相应的处理，提高算法的鲁棒性。第三，将以 VWOS 算法为核心，研究网页自适应呈现技术，以期达到 Web 学习资源移动访问的目标，提高 Web 学习资源的利用率，为移动学习服务打下技术基础。

参 考 文 献

[1] Diao Y, Lu H, Chen S, et al. Toward Learning Based Web Query Processing[C]// VLDB. 2000: 317-328

[2] Wong W, Fu A W C. Finding Structure and Characteristics of Web Documents for Classification[C]// ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 2000(s1): 96-105

[3] Kaasinen E, Aaltonen M, Kolari J, et al. Two approaches to bringing Internet services to WAP devices[J]. Computer Networks, 2000, 33(1): 231-246

[4] Buyukkokten O, Garcia-Molina H, Paepcke A. Accordion summarization for end-game browsing on PDAs and cellular phones [C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2001: 213-220

[5] 吴鹏飞，孟祥增，刘俊晓，等. 网页区域分割与识别技术[J]. 现代计算机(专业版)，2006(6)：48-50
Wu Peng-fei, Meng Xiang-zeng, Liu Jun-xiao, et al. Segmentation and Identification of Web Page's Areas[J]. Modern Computer, 2006(6): 48-50

[6] 王畴，唐世渭，杨冬青，等. 基于 DOM 的网页主题信息自动提取[J]. 计算机研究与发展，2004，41(10)：1786-1792
Wang Qi, Tang Shi-wei, Yang Dong-qing, et al. DOM-based automatic extraction of topical information from Web pages[J]. Journal of Computer Research and Development, 2004, 41(10): 1786-1792

[7] Hattori G, Hoashi K, Matsumoto K, et al. Robust web page segmentation for mobile terminal using content-distances and page layout information[C]// Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 361-370

[8] Romero R, Berger A. Automatic partitioning of web pages using clustering[M]// Mobile Human-Computer Interaction-MobileHCI 2004. Springer Berlin Heidelberg, 2004: 388-393

[9] Hattori G, Matsumoto K, Sugaya F. Auto Web Page Distilling Scheme Using Content Distance Based on Depth of Tag Hierarchy[J]. DBSJ Letters, 2005, 4(1): 1-8

[10] Chen Y, Xie X, Ma W Y, et al. Adapting Web pages for small-screen devices[J]. Internet Computing, IEEE, 2005, 9(1): 50-56

[11] Sano H, Shiramatsu S, Ozono T, et al. A Web Page Segmentation Method based on Page Layouts and Title Blocks[J]. International Journal of Computer Science and Network Security, 2011, 11(10): 84-90

[12] Sano H, Swezey R M E, Shiramatsu S, et al. A Web Page Segmentation Method by using Headlines to Web Contents as Separators and its Evaluations[J]. International Journal of Computer Science & Network Security, 2013, 13(1): 1-6

[13] Cai D, Yu S, Wen J R, et al. VIPS: a vision-based page segmentation algorithm: MSR-TR-2003-79[R]. Microsoft, 2003

[14] 蒙韧，邵延振，袁鼎荣. 一种基于页面 Block 的 Web 信息提取方法[J]. 计算机技术与发展，2010，20(1)：197-200
Meng Ren, Shao Yan-zhen, Yuan Ding-rong. A Web Information Extraction Algorithm Based on Web Page[J]. Computer Technology and Development, 2010, 20(1): 197-200

[15] Li L, Liu Y, Obregon A. Visual segmentation-based data record extraction from web documents[C]// IEEE International Conference on Information Reuse and Integration, 2007(IRI 2007). IEEE, 2007: 502-507

[16] 王静，姚勇，刘志镜. 基于广义隐马尔可夫模型的网页信息抽取方法[J]. 山东大学学报(理学版)，2007，42(11)：49-52
Wang Jing, Yao Yong, Liu Zhi-jing. Web information extraction based on a generalized hidden Markov model[J]. Journal of Shandong University(Natural Science), 2007, 42(11): 49-52

[17] 史晶，吴庆波，杨沙洲. 移动终端个性化页面显示优化技术研究[J]. 计算机工程，2012，38(18)：277-281
Shi Jing, Wu Qing-bo, Yang Sha-zhou. Research on Personalized Page-display Optimization Technology in Mobile Terminal[J]. Computer Engineering, 2012, 38(18): 277-281

[18] Song R, Liu H, Wen J R, et al. Learning block importance models for Web pages[C]// Proceedings of the 13th international conference on World Wide Web. ACM, 2004: 203-211

[19] 彭红超. 一种基于视觉的网页分割技术及应用研究[D]. 武汉：华中师范大学，2014：21-26
Peng Hong-chao. Research on Technique and Application of a Vision-based Webpage Segmentation[D]. Wuhan: Central China Normal University, 2014: 21-26