

CSDN

博客 学院 下载 GitChat TinyMind 论坛 问答 商城 VIP 活动 招聘 ITeye

正在等待转换...

VIPS基于视觉的页面分割算法[微...

开发一个app多少钱

联系我们

请扫描二维码联系省

webmaster@cs

400-660-0108

QQ客服 客服

关于 招聘 广告服务 网站地

©2018 CSDN版权所有 京ICP证090024

百度提供搜索支持

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心

0

译

写评论

200

月18日 12:26:00

收藏

文出处: <http://www.ews.uiuc.edu/~dengcai2/tr-2003-79.pdf>

微信

微博

VIPS:基于视觉的Web页面分页算法

问题的提出

随着互联网的高速发展, Web已经成为这个世界上最大的信息来源。Web 作为信息技术的载体已成为人们重要的工作、学习、生活、时间和空间界限来共享大量信息。但是如何去获取这些Web信息为我们所用则是大家面临的共同问题。在最基本的层次上, 整个Web相当于获取了Web信息内容。事实上, 目前的很多Web信息获取技术都是基于这种理论。

整个页面作为一个基本的信息获取单位并不是太合理, 尽管用户通常会把一些相关的内容放在同一页面中, 但是大多数情况下, 一个中, 可能包含体育类信息, 可能包含健康类信息, 也可能包含广告, 导航链接等信息。这些信息分布在整个页面的不同位置。因此, 如果要更进一步的语义提取。

Web页面的语义提取在很多方面都有应用。比如, 在Web信息访问中, 为了克服关键字搜索所带来的局限性, 许多研究者开始使用数据库扫描的过程中, 将Web文档分割为一定数目的数据块是首要的工作。目前的工作大多数停留在使用自适应的方法上。如果我们能够获取Web页面的单, 当然语义信息也就很容易提取出来。

语义块的提取另外一个应用场合就是搜索引擎。对于搜索引擎而言, 链接分析是一个极为重要的工作。目前, 对于大部分的搜索引擎而言, 链接关系, 那么这两个页面整体上肯定存在着一定的关系。但是在大部分情况下, 从页面A到页面B的链接仅仅意味着页面A的某部分与页面B的eRank以及HITS都是基于前面的假设。把链接关系定义在两个完整的页面之间与定义在两个页面的某部分之间相比, 则更粗糙一些。因此对于完整页面分割为多个语义块是一个必须的工作。目前已经有一些工作针对这方面进行展开。不过这些工作都是基于DOM树分析页面的结构。但是存在一定的缺陷。

Web页面的语义分块另外一个潜在的用途就是移动终端访问互联网。目前大部分的Web页面都是针对台式机而设计的, 由于移动设备通常屏幕较小, 计算能力有限, 因此这些页面开直接访问。目前通常通过两种手段来解决这种问题: 或者通过服务器进行页面转换或者使用网页缩略图。前者首先将用户访问的页面进行分页和转换, 然后将分页的结果提交给移动设备个Web页面生成缩略页面, 整个页面被分割为数目不等的区域, 用户如果对特定区域感兴趣, 则可以再次访问该区域的内容。通过这两个策略, 基本可以完成移动终端访问互联网的任务还是如何对页面进行语义分割。

如果对Web页面进行有效的分页, 目前已经很多工作展开。[Chakrabarti et al.2002]致力于从HTML DOM树中提取出结构化信息。不过由于HTML语法的灵活性, 目前大部分的网页都C规范, 这样可能会导致DOM树结构的错误。更重要的是, DOM树最早引入是为了在浏览器中进行布局显示而不是进行Web页面的语义结构描述。比如, 即使DOM树中两个结点具有同么这两个结点在语义上也不一定就是有联系得。反之, 两个在语义上有关系的结点却可能分布在DOM树的不同之处。因此仅仅通过分析DOM树并不能完全获取Web页面的语义信息。

从人类的角度来看, 当一个用户观察Web页面的时候, 它总是会自然而然的把一个语义块作为一个单一对象来看待, 而不会管Web页面的内部结构是如何描述的。通常情况下, 在分候, 用户会使用一些视觉因素来进行帮助, 比如背景颜色、字体颜色和大小、边框、逻辑块和逻辑块之间的间距等等。因此如果充分的使用Web页面的视觉提示, 并结合DOM树进行页面以弥补仅使用DOM树所带来的一些缺憾。

在论文中, 我们提出了VIPS(Vision-based page segmentation)算法用以提取给定网页的语义结构。这种语义结构是层次性的结构, 在该结构中, 每一个结点代表一个语义块。每一个个DOC值来描述该语义块内部内容的关联性。DOC的值越大, 则表明语义块内部的内容, 它们之间的联系越紧, 反之越松散。VIPS算法充分利用了Web页面的布局特征: 它首先从DOM的合适的页面块, 然后根据这些页面块检测出它们之间的所有的分割条, 包括水平和垂直方向。最后基于这些分割条, Web页面的语义结构将被重新构建。对于每一个语义块又可以使用分割为更小的语义块。因此整个VIPS算法是自顶向下, 非常高效的。

2.相关工作

忽略不介绍。

3.Web页面的基于视觉的内容结构描述

与[chen et al. 2001]类似, VIPS算法中首先也定义了“基本对象”的概念, 通常DOM树上的叶子结点被定义为基本对象, 因为这些结点已经不能再被继续分割了。在本论文中, 我们首先觉的内容结构, 它里面的每一个结点我们称之为“块”, 这些块或者是一个基本对象或者是一些基本对象的组合。有一点需要注意的是, 基于视觉的内容结构中的块与DOM树中的结点没有系。

与[Tang et al.1999]中文档的描述结构类似, VIPS算法中Web页面的结构定义如下。

对于每一个页面而言, 我们可以将其看作一个三元组 $\Omega=(O, \Phi, \delta)$, 其中

$O=(\Omega_1, \Omega_2, \dots, \Omega_N)$, 表示给定页面上的所有的语义块的集合, 这些语义块之间没有重叠覆盖, 而每一个语义块 Ω_i 又可以被定义为前面所描述的三元组 $\Omega_i=(O_i, \Phi_i, \delta_i)$, 如此迭代

$\Phi=(\phi^1, \phi^2, \dots, \phi^T)$, 表示当前页面上的所有的分隔条的集合。事实上, 一旦确定了一个页面上的两个语义块, 那么这两个语义块之间的分隔条也就被确定了。当然, VIPS中的分隔在的分隔条, 而是虚拟。分隔条包括水平分隔条, 也包括垂直分隔条。每一个分隔条都具有一定的宽度和高度。

$\delta=(\zeta_1, \zeta_2, \dots, \zeta_M)$ 则描述了 Ω 集合中两个语义块之间的关系, 这种关系可以用下面的式子描述: $\delta=O \times O \rightarrow \Phi \cup \{NULL\}$ 。其中的每个 ζ 都是一个形如 (Ω_i, Ω_j) 二元组, 其表示块 Ω_i 和 Ω_j 的割条。

上图演示了Yahoo页面的基于视觉的Web页面内容结构。它同时给出了页面的布局结构和基于视觉的内容结构。在第一层, 整个原有的页面被分割为四个大的可视对象VB1-VB4, 同之间检测出了三个分隔条 $\phi_1-\phi_3$ (原来有五个, 最上面的和最下面的被舍弃)。检测出的四个可视对象并不是这轮分割的最终局部。最终得到的语义块必须根据检测出的四个可视对象和步构建而成, 其中可能需要合并一些语义块, 舍弃一些分隔条等等。

比如, 对于VB2, 从它的内部又可以检测出三个子对象和两个分隔条, 如图1所示。

对于每一个Block, VIPS算法都定义一个DoC(Degree of Coherence)与之对应。该值的大小反映了当前语义块内部内容联系的紧密程度。如果。它具有下面两个重要的特性:

- 1). DoC的值越大, 则语义块内部的内容之间的联系紧密程度就越大, 它们之间就关系就越连续, 反之越小。
- 2). 在层次数上, 语义块的子块的DoC的值肯定要比父块的值大。

浙江大学在职研究生 拓尔思 python null 人脸识别算法 电子工业出版社²⁴

VB2_1块将变的不再允许分割。不同的应用程序可以设置不同的PDoC值来达到自己的要求。

基于视觉的页面分割最主要的目的就是对给定的页面进行语义分割，因此分割后生成的基于视觉的内容结构中的结点通常总是由语义块组成。比如，在图1(a)中，VB2_1_1表示Yahoo宠物商店的目录链接，而VB2_2_1和VB2_2_2则表明了两种不同的comics。

4.VIPS算法描述

这部分我们将详细介绍VIPS算法。整体来说，页面的基于视觉的内容结构是结合DOM树以及一些视觉提示信息而得到的。整个分页过程可以用图2描述。它具有三个步骤：页面块提取以及语义块重构。这三个步骤联合一起作为一次语义块检测的完整步骤。Web页面首先被分割为几次比较大的语义块，同时这几个语义块所组成的层次结构将被记录下来。对于检测出来语义块分页过程又可以继续进行，直到语义块的DoC值达到预先设定的PdoC为止。

在每次迭代循环中，当前逻辑块的DOM树结构以及它的视觉信息都将被获取。然后，从DOM树的根结点开始，逻辑块检测过程将基于视觉信息开始从DOM树中开始检测页面块。每3b中的结点(1,2,3,4,5,6,7)都会被检查它能够构成一个单独的页面块。如果不能，比如图3b中的1, 3, 4结点，那么它的子结点将被执行同样的检查。对于每一个提取出来的页面块，比如6, 7结点，我们都会根据当前页面块的内部可视属性赋予一个DoC值。当本次迭代过程中所有的页面块都被检测出来之后，它们将被保存到页面块池中。基于这些页面块，分隔条检测过这些页面块之间的所有的水平分隔条和垂直分隔条最终将被识别出来并且赋予一定的宽度和高度。基于这些分隔条，页面的布局层次将被重新构建——一些页面块将被合并，形成语义块代过程中的所有语义块都被检测出来。

迭代过程是否需要继续进行取决于本层次的语义块中是否存在DoC值小于PdoC的语义块。对于那些DoC>=PdoC的语义块，分隔过程将停止，否则分隔过程将继续。比如再下图中，值小于PdoC，那么该语义块将被作为新的子Web页面，继续执行分割算法，最终又被分割为两部分：C1和C2，如图4a和4b所示：

当所有的语义块被提取出来后，最终整个Web页面的基于视觉的内容结构也就构建完成。在上面的例子中，我们最终获取的内容层次结构如图5所示。在下面的部分我们将详细的描述分隔条的检测以及内容结构重建过程。

4.1语义块提取

在这步骤中，我们的目标是提取出当前子页面中所包含的所有的可视语义块。通常情况下，DOM树中的每一个结点都可以表示一个可视语义块。不过，在HTML中，一些标签比如<T常用来进行数据组织，因此不适合表示单独的可视语义块。对于这种结点，对他们的提取将被它们的孩子结点替代。而且由于HTML语法的灵活性，很多的Web页面并没有严格遵循W3C这导致DOM树并不能总是能反映不同的DOM结点之间的关系。

对于提取出来的每一个可视语义块，我们将根据它的内部的视觉差异设置它的DoC值。整个迭代提取过程可以用下面的算法描述：

如何判断给定的结点能否被继续分割，我们给出下面的几个方面进行判断：

- 1)、DOM结点本身的属性。比如当前DOM结点的标签，结点的背景色，当前结点所代表的页面块的大小，形状。
- 2)、当前DOM结点的孩子结点。比如孩子结点的标签，孩子结点所代表的区域的背景色，前景色，区域的大小以及不同类型的孩子的数目等等。

基于WWW HTML规范4.0，我们将DOM结点分为两大类：inline结点和line-break结点。

所谓Inline结点是指：如果该结点的标签能够影响文字的外观同时不会引起换行的话，那么这类结点我们称之为Inline结点，比如、<BIG>、、、<I>、类结点通常仅仅影响文字的外观而不会影响文字的布局。

所谓Line-break结点，则就是除了inline结点之外的所有结点。

另外，基于各种结点在浏览器中的显示以及结点的孩子结点属性，我们给出下面的定义：

1)、有效结点(Valid node)：如果一个结点能够在浏览器中表现出来，那么这个结点就是有效结点。通常有效结点的长度和宽度都不为零。另外如果一个结点内部没有任何有用的信息无效结点。比如图7中的第二个和第四个TR结点都是无效结点。

2)、文本结点：这类结点通常是指HTML中的文字，通常它们不被任何标签所包围。

3)、虚拟文本结点(这个定义是递归定义)

φ 如果一个结点是文本结点，那么它自然就是虚拟文本结点

φ 如果一个结点是inline结点，并且它的所有子结点要不是文本结点，要不是虚拟文本结点，那么这个结点也就是虚拟文本结点。

如果一段文字加上了、<BIG>、<I>等标签之后，该文字只是在浏览器中的显示外观发生了变化而已，并不影响这段文字本质上为文字的属性，VIPS中将之称之为虚拟文本结点，可视语义块的提取算法DivideDomtree如图6所示。在该算法中一些很重要的信息可以用于产生推测规则：

φ 标签提示

- 1)、一些标签比如<HR>通常用来从视觉上分隔不同主题的内容。因此如果DOM结点中包含这些标签，那么我们倾向于认为该结点允许被继续分割。
- 2)、如果inline结点的孩子结点存在line-break结点，那么该结点将被倾向于被分割。

φ 色彩提示

如果当前结点的孩子结点中有一个结点的背景色与它的背景色不同，那么我们倾向于分割该DOM结点。同时，具有不同背景色的节点在本次循环中不再被分割。分割由下一次迭代完

φ 文本提示

如果当前结点的大部分孩子结点都是文本结点或者是虚拟文本结点，那么我们倾向于不再继续分割该结点。

φ 尺寸提示

我们通常可以对不同的结点类型预定义一个门槛尺寸(结点的大小与整个页面大小的比较)，如果结点的相对尺寸小于门槛大小，那么分割就停止。

基于上面的这些提示信息，我们给出一些推理规则用以判断当前的结点是否应该被分割。如果一个结点不需要再分割，那么该结点块将被提取出来，同时设置相应得DoC值，并保存推理规则如下表所示：

规则 1	如果当前结点不是文本结点，同时它又没有任何有效的孩子结点，那么该结点将不被分割，并且从结点集合中删除。
规则 2	如果当前结点只有一个有效的孩子结点，同时该孩子结点不是文本结点，那么当前结点将被分割。
规则 3	如果当前的DOM结点是整个子DOM树的根结点(与页面块对应)，同时只有一个子DOM树与当前的页面块关联，那么分割该结点。
规则 4	如果当前结点的所有的孩子结点都是文本结点或者是虚拟文本结点，那么不分割该节点。如果当前所有孩子结点的字体大小和字体重量都是相同的话，那么该结点的DoC设置为10，否则设置为9。
规则 5	如果当前DOM结点的孩子结点中有一个line-break结点，那么该结点将被继续分割
规则 6	如果当前结点的孩子结点中存在<HR>结点，那么该结点将被继续分割

规则 8	如果结点至少具有一个文本或者虚拟文本子结点，同时结点的相对大小小于于门槛大小，那么
规则 9	如果当前结点的所有子结点中最大的尺寸也小于于门槛大小，那么该结点将不再分割，同时DoC值根据HTML标签和结点大小设置。
规则 10	如果前一个兄弟结点没有被分割，那么该结点也不会被继续分割
规则 11	分割该结点
规则12	不要分割该结点，同时基于当前结点的标签和大小设置DoC值

对于不同的DOM结点，我们使用不同的推理规则：

让我们考虑图1的情况。在第一次页面块提取得过程中，最终VB1，VB2_1，VB2_2，VB2_3以及VB3和VB4被提取出来，然后放进了语义块池中。下面我们将详细描述VB2_1，VB2取的过程。

图7(b)显示的是一个表格，该表格是整个Web页面的一部分。它的DOM树结构显示在左边的部分。在页面块的提取过程中，当遇到<TABLE>结点的时候，它只有一个有效的孩子结点则2，我们进入<TR>标签。该<TR>结点具有五个<TD>孩子结点，但是它们中只有三个是有效结点。而且第一个孩子结点的背景颜色与父亲结点的颜色不同。根据规则8，该<TR>结点将个<TD>结点在本次迭代中部进行分割。第一个<TR>结点被保存到页面块池中。第二个和第四个<TR>结点为无效结点，因此它们将被删除。对于第三个和第五个<TD>结点，根据推理规则中不再分割，因此最终我们得到三个页面块VB2_1，VB2_2和VB2_3。

4.2分隔条检测

当所有的页面块被提取出来之后，它们都被保存在页面块池中以便进行分隔条检测。在VIPS算法中，分隔条是Web页面中的垂直的或者水平的行。从视觉的角度而言，separators are used for discriminating different semantics within the page.

在VIPS中，一个可视的分隔条可以用二维向量(Ps，Pe)描述，其中，Ps是分隔条的起始坐标，而Pe则是分隔条的终止坐标。坐标的单位全部为像素pixel。根据Ps和Pe，很容易计算度数和高度。

4.2.1 分隔条检测

分隔条的检测算法如下描述：

- 1)、初始化分隔条列表。最早的分隔条列表中仅仅存在一个分隔条，它的起始和终止坐标为(Pbe,Pee)，分别对应整个Web页面的起始坐标和终止坐标。
- 2)、对于页面块池中的每一个页面块，它与分隔条的关系包括下面三种：
 - 页面块被包含在分隔条中，此时，该分隔条将从页面块的边缘裂变为多个分隔条。
 - 页面块与分隔条发生部分重合，那么根据页面块的边界重新调整分隔条的参数
 - 页面块跨越分隔条，那么此时移除该分隔条。
- 3)、移除页面边缘的四个分隔条

图8演示了分隔条的检测过程。为了简单期间，我们仅仅演示水平分隔条的检测过程。开始的时候我们之后一个大的分隔条，它的起始和终止位置就是整个页面的起始和终止位置。当页面块放入到池中的时候，由于该页面块被包含在分隔条内部，此时原有的分隔条将裂变为S1和S2。同理当第二个和第三个页面块放入到池中的时候，四个分隔条S1，S2，S3和S4被检测到。当第四个页面块放入到池中的时候，它跨越了S3分隔条，同时与S2分隔条有部分重合，此时S3分隔条将被删除，同时S2将被调整，从图中可以看出，调整后，S3明显的变细了。

4.2.2 设置分隔条的权重

分隔条通常用于区别不同语义的页面块，因此基于给定分隔条两边的语义块的在视觉上的差异，我们可以设置分隔条的权重。如果分隔条的权重越重，该分隔条最终成为分隔条的可下面的规则可以用来设置分隔条的权重：

- 1)、分隔条两边的页面块的距离越远，该分隔条的权重就越高。
- 2)、如果某个分隔条是通过检测HTML标签获取的，比如<HR>，那么该分隔条的权重就越高。
- 3)、如果分隔条两侧的页面块的背景色是不相同的，那么该分隔条的权重将相应增高。
- 4)、对于水平分隔条而言，如果分隔条两侧的页面块的字体属性，比如字体大小，字体重量是不同的，那么该分隔条的权重将增加。而且如果分隔条上侧的页面块的字体小于分隔条下侧的字体，那么分隔条的权重将增加。
- 5)、对于水平分隔条而言，当分隔条两侧的页面块的结构非常相似，比如文本，那么该分隔条的权重将递减。

考虑图7中的第三个<TD>。与该结点对应的子页面如图9(b)所示，同时它的DOM树结构如图9(a)所示。我们可以看到根据我们的定义，该DOM树中的很多结点都是无效的，它们无法提取出来，在页面块的提取过程忠，这些结点将被忽略。当这些页面块提取出来之后，六个页面块将保存到池中，同时五个水平分隔条也被检测出来。同时，基于上面的五个分隔条规则，这将被设置。在本例中，页面块2和3之间的分割条要比页面块1和页面块2之间的分割条权重高，这是因为字体不同的原因。同样的原因，4和5之间的分隔条权重也高一些。最终的分隔条以图9(c)所示。

4.2.3 内容结构构建

当分隔条被检测出来，同时权重设置完毕后，相应的内容重建过程就可以开始了。构建过程从最小权重的分隔条开始，该分隔条两侧的页面块将合并在一起组成一个新的页面块。该进行迭代，直到遇到权重最高的分隔条为止。对于每一个新的语义块，相应的DoC也被相应设置。

当页面块最终合并成为语义块之后，本轮的迭代也就结束了。对于这些语义块，每一个语义块的DoC都会与PdoC进行对比，如果DoC的值小于PdoC，那么新的迭代过程将重新开始分隔条检测以及内容结构重构。当所有的语义块的DoC的值都不大于PdoC，迭代过程将停止。同时针对整个Web页面的内容结构将构建出来。

以图9为例，在第一轮迭代中，第一，三以及五个分隔条将被选择出来，同时页面块1和2被合并为新的语义块VB2_2_2_1。同样的合并发生在页面块3和4上，它们被合并为新的语义块5和6最终合并为VB2_2_2_3。新的语义块VB2_2_2_1，VB2_2_2_2以及VB2_2_2_3是语义块VB2_2_2的子结点。对于每一个页面结点，比如VB2_2_2_2_1_1，VB2_2_2_1_1以及它们的DoC的值将被检查，以便确定是否满足PdoC的值。最终的内容结构构建完毕。

转载请注明来源: <http://blog.csdn.net/tingya>

如果你觉得本文不错，请点击文后的“推荐本文”链接！！

上一篇

Apache中的表格实现剖析(2)

下一篇

关于VIPS算法的实现



好用的数据可视化工具

想对作者说点什么？

我来说两句

- Tingya

2009-01-07 20:12:57

#18楼

确实是对原来的VIPS进行了少量的修改。
- xiaoxiaofengqi

2009-01-07 11:04:35

#17楼

这个算法我也写了一下，不过是用C++写的。我不太懂楼上讨论的分页是什么意思，我通过分析dom树然后用文中提到的规则进行dom结点处理以及后续的处理。我日中的内容按网页框架分开，把正文部分合在一起，然后用贝叶斯决策计算正文特征支持率 提取网页内容。现在VIPS基本写完。但是却也发现了些问题，比如说有些结出来会有提取不出分隔条，这是因为有少数坐标有些重叠。这里涉及到一个坐标的确定问题。然后是结点分割规则问题，现在的页面是大部分是通过DIV来组织页面。适TABLE组织的页面，我试过用TABLE组织的页面，分得相当不错。另外，TINYA上面的翻译似乎改了些规则，还有部分翻译不是很准确。比如虚拟文本的定义部分与不知道TINYA有没有注意到。最后，很感谢TINYA 对这个算法的介绍。另外，有对这个算法感兴趣的朋友希望能大家一起讨论下 我的QQ：24888086 msn:trues54120
- xiaoxiaofengqi

2008-11-14 15:41:58

#16楼

英文版原文出处：http://research.microsoft.com/research/pubs/view.aspx?tr_id=690
- xiaoxiaofengqi

2008-11-14 15:41:53

#15楼

英文版原文出处：http://research.microsoft.com/research/pubs/view.aspx?tr_id=690
- xiaoxiaofengqi

2008-10-30 14:05:12

#14楼
- 查看 18 条热评

基于视觉信息的网页分块算法（VIPS）

8249

VIPS: a Vision-based Page Segmentation Algorithm.pdf下载

这篇论文的主要思想：

从人类的角...

VIPS:基于视觉的Web页面分页算法

1477

1.问题的提出 目前，随着互联网的高速发展，Web已经成为这个世界上最大的信息来源。Web 作为信息技术...

VIPS:基于视觉的Web页面分页算法 - CSDN博客

VIPS算法充分利用了Web页面的布局特征:它首先从DOM树中提取出所有的合适的页面块,然后根据这些页面块检...

基于视觉信息的网页分块算法(VIPS) - CSDN博客

VIPS算法的首先从DOM树中提取出所有的合适的页面块,然后根据这些页面块检测出它们之间的所有的分割条,包...



好用的数据可视化工具

百度广告

关于VIPS算法的实现

6806

微软并没有给出VIPS算法的实现,仅仅给出了一个演示程序。而且，目前而言，实现是基于IE，不具有移植性，...

VIPS:基于视觉的页面分割算法[微软下一代搜索引擎核心分页算法](...

VIPS:基于视觉的页面分割算法[微软下一代搜索引擎核心分页算法]<http://www.vipcn.com/chengxukaifa/qitayuya...>

关于VIPS算法的实现 - CSDN博客

微软并没有给出VIPS算法的实现,仅仅给出了一个演示程序。而且,目前而言,实现是...上一篇VIPS:基于视觉的页面...



VIPS算法，按照微软VIPS的思想编程实现，C#实现

2010年09月02日 584KB 下载

VIPS:基于视觉的页面分割算法[微软下一代搜索引擎核心分页算法] (介...

VIPS:基于视觉的页面分割算法[微软下一代搜索引擎核心分页算法]http://www.vipcn.com/chengxukaifa/qitayu...

VIPS基于视觉的页面分割算法

*版权证明: 只允许上传png/jpeg/jpg/gif格式的图片,且小于3M *详细原因: 取消提交 VIPS基于视觉的页面分割...

vips实现-网页segmentation程序

VIPS基于视觉的页面分割算法 立即下载 上传者: wwzhzz1982 时间: 2012-11-23 综合评分: 4 积分/C币:3 VIPS算...

vips实现-网页segmentation程序

下载 2018年07

vips实现代码,实现了著名的vips网页分块程序... vips microsoft 3C币 177下载 VIPS算法实现 3C币 14...

2018筑桥高端会所体验！揭开高级会所神秘面纱！

花韵会所 · 顶新

基于视觉的Web页面分页算法VIPS的实现源代码下载 - CSDN博客

3、采用的合并算法与VIPS算法不相同4、分割条的检测算法与VIPS算法相同具体的...VIPS:基于视觉的Web页面...

基于视觉的Web页面分页算法VIPS的实现源代码下载 - CSDN博客

另外,该源代码实现并未严格遵循VIPS算法,它与VIPS算法存在的差异包括:1、 DOM...VIPS:基于视觉的页面分割算...

基于视觉的Web页面分页算法VIPS的实现源代码下载

4

本来由于尚未优化好，暂时不提供下载的，但是由于众多的user迫切希望获取，因此只能将这个不成熟的版本...



VIPS算法对搜索引擎的意义[转载]

1

&lt;b&gt;VIPS算法对搜索引擎的意义[转载]&lt;/b&gt; 基于VIPS(视觉式版面切...

VIPS算法,按照微软VIPS的思想编程实现,C#实现

VIPS算法源代码 立即下载 上传者: Lexuswj 时间: 2014-03-13 综合评分: 4 积分/C币:3 VIPS基于视觉的页面分割算...

VIPS算法源代码

vips算法的实现,利用C#语言编写,效果还不错,大家可以借鉴一下。



VIPS基于视觉的页面分割算法

2012年11月23日 960KB 下载



关于搜索引擎的几大核心算法浅析

903

关于搜索引擎的几大核心算法浅析



基于单目视觉的车道偏离预警算法研究

3540

基于单目视觉的车道偏离预警算法研究 2016-04-04 15:54 戴秋菊, 陈贤富 (中国科学技术大学 信...

基于视觉的Web页面分页算法VIPS的实现源代码下载

1.2万



VIPS算法源代码

2014年03月13日174KB

下载



惊艳全球数据行业的16个数据可视化例子


百度广告

[转载]VIPS:基于视觉的Web页面分页算法

 753


VIPS:基于视觉的Web页面分页算法目前，随着互联网的高速发展，Web已经成为这个世界上最大的信息来源。...

VIPS:基于视觉的Web页面分页算法（转载）


 543


VIPS:基于视觉的Web页面分页算法1.问题的提出目前，随着互联网的高速发展，Web已经成为世界上最...

PageRank:核心算法|谷歌如何从网络的大海里捞到针

 781

转自：http://mp.weixin.qq.com/s?__biz=MjM5MTQzNzU2NA==&mid=401631721&idx=1&sn=1a...

SEO（搜索引擎优化）浅谈普及一下搜索引擎的核心算法

 1148

外链是搜索引擎算法中，判断网站权重高低的重要指标，当用户在搜索框中输入关键时，搜...

简易垂直搜索引擎的核心算法总结

 2276


1. 倒排索引 倒排索引源于实际应用中需要根据属性值（字段）来查找记录（所在的文件位置）...


数据采集基础知识

百度广告

基于单目视觉的平面模型摄像机定位算法

 1679

基于视觉手势识别系统的方法总结

 6060

一个基于视觉手势识别系统的构成应包括：图像的采集，预处理，特征提取和选择，分类器的设...

Google搜索引擎的奥秘

 3529

1、背景和问题 据统计超过80%的用户靠搜索引擎获取信息网站排名是网络搜索引擎的核心目前Google数据库...



基于机器视觉的停车位检测技术的研究

2014年09月05日420KB

下载


计算机视觉之基于聚类的分割方法


 1188

分割：假设要识别一幅图像中的物体。如果逐个处理每一个像素，那么需要处理的像素就太多了...

舆情监控的市场前景

百度广告

搜索引擎算法

 295

搜索引擎的两大主要任务是：匹配和排名。在实际中，搜索引擎将匹配和排名组合成一个流程以...



搜索引擎算法



视觉的网页分块

2013年03月28日 213KB 下载

基于Matlab的标记分水岭分割算法(imreconstruct)

3722

1 综述 Separating touching objects in an image is one of the more difficult image processing operatio...



【智能驾驶】基于计算机视觉的自动驾驶算法研究综述

889

近年来，随着人工智能技术的迅速发展，传统汽车行业与信息技术结合，在汽车自动驾驶技术方...



一点点加盟

百度广告



浅谈搜索引擎的核心算法

926

外链是搜索引擎算法中，判断网站权重高低的重要指标，当用户在搜索框中输入关键时，搜索引...



基于深度学习的计算机视觉应用之目标检测

672

欢迎大家关注我们的网站和系列教程：<http://www.tensorflownews.com/>，学习更多的机器学习...



搜索引擎与PageRank

962

很早就对Google的PageRank算法很感兴趣，但一直没有深究，只有个轮廓性的概念。前几天趁...



机器视觉算法（数据结构）

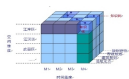
1018

图像（Image）机器视觉中，图像是基本的数据结构，它所包含的数据通常是有图像采集设备传...

计算器核心算法代码实现(Java)

273

在进行一个表达式的计算时，先将表达式分割成数字和字符串然后利用出入栈将分割后的表达式进行中缀转后...



什么是数据仓库

百度广告



立体视觉：算法和应用（五）

1205

本文翻译的外国学者的一份talk，主要内容是关于立体视觉算法和应用的基础知识。限于个人水...

KNN算法源代码分析

3753

KNN的类结构在ml.h头文件中定义，代码如下：KNN类的实现部分在mlknearest.cpp中，代码如下： /*****...



基于区域分割的算法

109

区域分割前面所讲的图像分割方法都是基于像素的灰度来进行阈值分割，本节将讨论以区域为基...



基于图的图像分割（Graph-Based Image Segmentation）

4118

一、介绍 基于图的图像分割（Graph-Based Image Segmentation），论文《Efficient Graph-Base...

基于边缘的分割

206

一 图像边缘检测 基本思路：基于边缘检测的图像分割方法的基本思路是先确定图像中的边缘像素，然后

登录

注册



微软 Surface 手机来了

百度广告

基于机器视觉的车道线检测与追踪



2709

完成的功能： 视频图像采集 图像预处理 车道线检测与识别 实现的效果： 实现思路： 视频采集部分：摄像头...

NGN(下一代网络)的方方面面



1373

NGN （Next Generation Network，下一代网络，缩写为NGN）是一个定义极其松散的术语，泛指一个大量采...

个人资料



tingya

原创
68

粉丝
641

等级： 博客 6

访问

积分： 7753

排名



工作流平台



最新文章

Apache源代码全景分析第
求处理

Linux/Unix下的Curses库于
章 表单开发及应用

Unix/Linux下的Curse库开
菜单开发及应用

Unix/Linux下的Curses库于
章 面板库(panel)开发及应

Unix/Linux下的Curses库于
章 鼠标支持

个人分类

Apache源代码分析

C/C++

Unix/Linux

归档

2010年9月

2009年12月

热门文章

- Apache源代码全景分析第

求处理

阅读量：56629
- Linux下通用线程池的构建

阅读量：23727
- VIPS:基于视觉的页面分割

搜索引擎核心分页算法]

阅读量：21760
- Unix/Linux下的Curses库于

章curses库窗口

阅读量：20940
- Unix/Linux下的Curses库于

章 curses库I/O处理

阅读量：20318

最新评论

- 关于VIPS算法的实现

xzhichen：大神写的咋样了，
- 基于视觉的Web页面分页算

u010925374：请求源码，想学
- Apache数组分析

m0_37595562：请问这个是不

呀，每次pop的都是最后一个元
- 基于视觉的Web页面分页算

u010565037：老师您好，我想

算法，可以发一份源码吗？我

@qq.com谢...
- 公开 《Unix/Linux下的C...

duhengqi：求发邮箱12920719

收藏blog

- 盛普的blog
- 温辉敏的blog
- pandaxcl的专栏
- 卢亮的搜索引擎研究
- 水木尤寒的lucene相关的排
- aimingoo的专栏
- 高性能网络服务器开发
- 我的淘宝店铺[护肤品零售]
- 郭懿心的blog
- 悦色尚品
- Apache中国
- 涛哥的排头兵