基于视觉特征的网页正文提取方法研究

安增文,徐杰锋

(中国石油大学(华东) 计算机与通信工程学院,山东 东营 257000)

摘 要:利用网页的视觉特征和 DOM 树的结构特性对网页进行分块,并采用逐层分块逐层删减的方法将与正文无关的噪音块删除,从而得到正文块。对得到的正文块运用 VIPS 算法得到完整的语义块,最后在语义块的基础上提取正文内容。试验表明,这种方法是切实可行的。

关键词:页面分块;信息提取;视觉特征

中图分类号:TP391 **文献标识码**:A

The research on vision-based Web page information extraction algorithm

AN Zeng-Wen, XU Jie-Feng

(College of Computer & Communication Engineering, China University of Petroleum, Dongying 257000, China)

Abstract: To get the useful information blocks, this paper first segmented the Web page into blocks with its visual features and its DOM tree's characteristics, and then deleted the noise blocks. This is a recursive process until no block can be deleted. Then handled the reserved blocks with the VIPS algorithm to get the semantic blocks. At last, got the text content by handling the semantic blocks. Experiment shows that this method is feasible.

Key words: page segmentation; information extraction; visual features

随着互联网的迅速发展,互联网上的信息量以几何级数倍增。人们需要在海量的信息库中查找自己需要的信息。虽然搜索引擎能帮助人们快速地搜索到想要的信息,但每个网页除了正文内容外还掺杂了很多用户不需要的信息。例如,为了方便用户浏览而加入的导航链接、出于商业利益而加入的广告链接、版权信息以及相关主题阅读推荐链接等。这些信息掺杂在网页中,影响了用户对主题内容的浏览。因此,如何从包含大量噪音内容的网页中将正文信息准确、完整地提取出来成为众多研究者研究的课题。

1 相关工作

在 Web 信息抽取领域,已经有大量的研究工作,包括 HTML 结构分析方法(如 XWRAP 和 Lixto)、基于自然语言处理的方法(如 SRV 和 WHISK)、机器学习方法等。但是这些方法都是针对特定网站或特定格式的,不具有通用性,并且不能完成自动抽取。众多的 Web 网页正文信息提取方法都有各自的优缺点。

参考文献[1]采用机器学习的方法提取网页正文信息。此方法通过对网页集的学习,不断生成新的模板,从而建立模板库。提取信息时,查找对应的模板,利用模板中主题结点信息,直接定位主题信息块,快速提取主题信息。虽然此方法采用自动抽取的方式,其智能化程度也在一定程度上方便了用户的使用,但对于一个新的网页,若找不到匹配的模板,此方法就不适用了。而且随着模板数量的增加,模板库的维护工作也变得越来越复杂。

从页面视觉特征的角度对网页结构进行挖掘也是很有效的途径。典型的代表就是微软亚洲研究院提出的 VIPS(Vision-based Page Segmentation)算法[2]。它利用背景颜色、字体颜色和大小、边框、逻辑块和逻辑块之间的间距等视觉特征,通过制定相应的规则把页面分成了各个视觉信息块。这能在一定程度上满足复杂页面对算法的要求,但由于视觉特征的复杂性,运用的启发知识往往较为模糊,需要人工不断地总结调整规则,因此如何保证规则集的一致性是一大难点。

有许多研究者考虑使用 HTML 标签信息来划分页面。其中,中科院计算所软件研究室提出利用 TABLE 标

记和视觉特征对页面进行语义块划分，并识别各语义块属性的算法 TVPS(Table and Vision based Page Segmentation)[3]。TVPS 算法中的分块方法只考虑了各个最底层的 TABLE 标记，但是实际情况中网页样式结构和 TABLE 标记的嵌套关系都非常复杂，网页正文信息不一定全在最底层的 TABLE 标记中。如果只考虑最底层的 TABLE 标记，会遗漏部分正文信息。

参考文献[4]根据正文字数多、标点符号多 2 个特征，提出一种基于正文特征的网页正文信息提取方法。该方法利用 HTML 标签对网页内容进行分块，把具有正文特征的块保留，不具有正文特征的块舍弃，从而进行网页正文信息的提取。这种方法对于新闻、财经、科技等类型网页提取效果较好，但对于图片多文字少或对于用户回帖字数较少的论坛型网页提取效果较差。

以往的基于分块的网页信息提取算法都是对整个网页进行处理，并分完块后再对页面块进行取舍，确定正文块。这类方法对与页面主题无关的噪音信息也进行了处理，增加了算法的复杂度。本文在前人工作的基础上结合参考文献[3]、[4]、[7]，提出采用逐层分块逐层删减的方法对 Web 网页进行信息提取，以降低算法的复杂度，提高抽取的准确度，并用试验验证其可行性。

2 正文提取算法

Web 网页通常分为 3 种类型：主题型网页、图片型网页、目录型网页[5]。主题型网页通常通过成段的文字描述 1 个或多个主题(如新闻网页)；图片型网页中内容是通过图片体现的，只用少量文字对图片进行说明；目录型网页通常不会用成段的文字描述，而是提供指向相关网页的超链接，也可称为索引页。本文所研究的网页正文提取是针对主题型网页展开的。

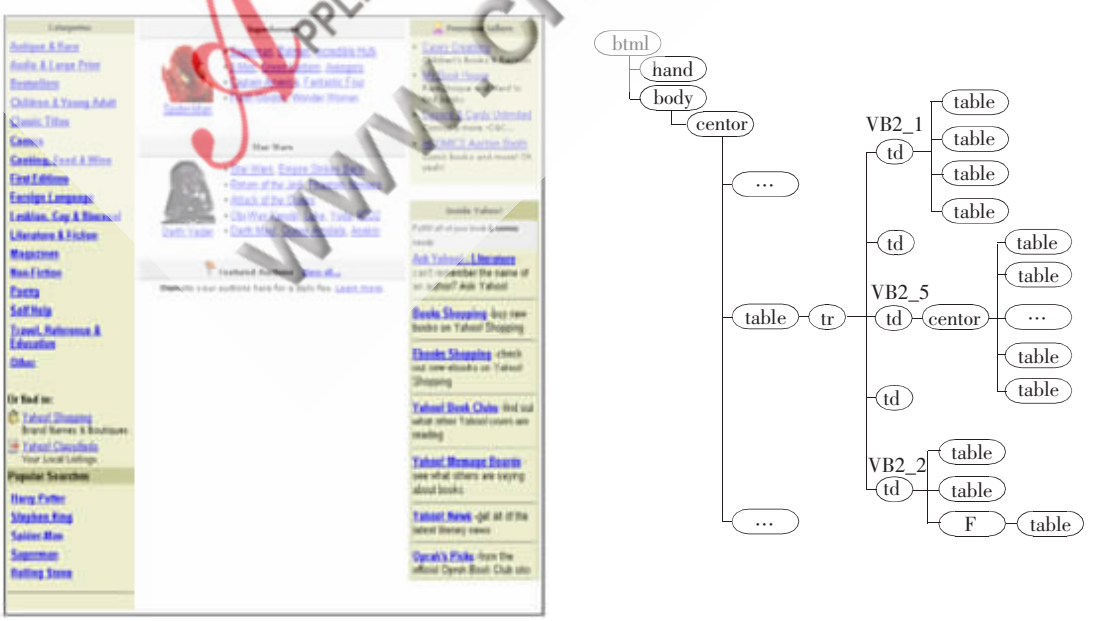
2.1 VIPS 算法

VIPS 算法充分利用了 Web 页面的布局特征。它首先从 DOM 树中提取出所有合适的页面块，然后根据这些页面块检测出它们之间所有的分割条，包括水平和垂直方向；最后基于这些分割条，重新构建 Web 页面的语义结构。对于每一个语义块又可以使用 VIPS 算法继续分割为更小的语义块。该算法分为页面块提取、分隔条提取和语义块重构 3 部分，并且是递归调用的过程，直到条件不满足为止。在此仅对页面块提取方法做简单介绍。

整个 VIPS 算法自顶向下，图 1(a)显示的是一个表格，该表格是整个 Web 页面的一部分，它的 DOM 树结构如图 1(b)所示。在页面块的提取过程中，当遇到<TABLE>结点时，它只有 1 个有效的孩子结点<TR>。根据参考文献[2]中规则，进入<TR>标签。该<TR>结点具有 5 个<TD>孩子结点，但是它们中只有 3 个是有效结点，而且第 1 个孩子结点的背景颜色与父亲结点的颜色不同。根据参考文献[2]中规则，该<TR>结点将被分割，而第 1 个<TD>结点在本次迭代中不进行分割，将其保存到页面块池中。第 2 个和第 4 个<TD>结点为无效结点，因此将被删除。对于第 3 个和第 5 个<TD>结点，根据参考文献[2]中规则，在本次迭代中不再分割，被保存到页面块池中。因此最终得到 3 个页面块 VB2_1、VB2_2 和 VB2_3。

2.2 页面块提取和过滤

由于 VIPS 算法的重点是对网页进行分块，所以其要对网页上的所有内容进行处理。而对于网页信息提取，只有与主题相关的正文信息才有意义，其他内容(如导航栏、相关阅读、广告链接、用户评论等)都属于噪音信息，只需要识别出来，并不需要对其进行处理。如果直



(a)一个 Web 页面　　(b)Web 页面对应的 DOM 树结构

图 1　一个 Web 页面及对应 DOM 树结构

接利用 VIPS 算法对网页进行页面块的划分，则会将那些与正文内容无关的噪音内容也进行处理，需要大量的内存来保存结点信息，增加了算法的时间和空间开销，也不利于提高提取的准确度。所以本文采用逐层分块逐层删减的方法，将网页中的噪音信息尽早地删除，以节省开销、提高准确度。本文在页面块提取和过滤阶段只利用了 VIPS 算法中的页面块提取方法，并未进行分隔条的提取和语义块的重构，而是在过滤完成后对保留的页面块进行相应的处理。

由于 VIPS 算法中的页面块提取过程是结合 DOM 树和视觉特征，利用人工制定的规则来判断该结点是否需要再分，并用页面块池来存储要被继续提取的页面块，所以可以对每一层完全提取后页面块池中保存的页面块进行相应的判断，将与标题内容无关的噪音块删除。这样将每 1 次提取出来的页面块中的噪音块都删除，当页面块提取完后，剩下的页面块即为要进行信息提取的页面块。

对网页信息的抽取包括：页面主题、发表时间、正文内容，抽取流程如图 2 所示。

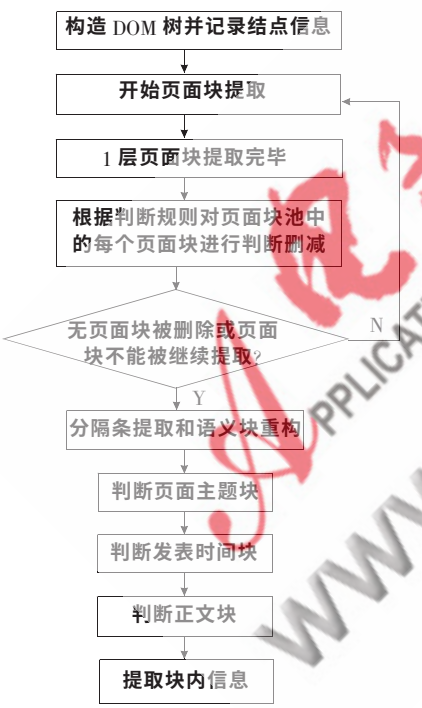


图 2 正文信息抽取算法流程

(1)构建 DOM 树。由于 HTML 书写的随意性，首先对 HTML 代码进行预处理，例如对书写不规范的标签进行补全处理，以免在后面的程序处理中造成错误，并去除一些无用标签，如<script>、注释信息等。网络上有很多网页预处理工具可以将不规范的 html 代码规范化，例如 HTML Tidy 等。构造 DOM 树的过程中要保存每个结点的字体大小、颜色、粗细、背景色等视觉信息，方便后续处理。

(2)对网页进行逐层分块逐层删减。根据对大量网页的统计，处于网页最上方和最下方的页面块基本全部是网站的导航链接和版权声明，这 2 个页面块可以直接删除。而页面块的中心位置与网页的左边框或右边框的距离小于一定阈值的基本全为广告信息。在实验数据中有的广告可以占到网页宽度的 40%。可以利用这一特征将网页四周的噪音块删除。定义网页的左上角顶点为坐标原点，网页的右下角顶点坐标为(WIDTH，HEIGHT)，每个页面块的中心点坐标为(Center_X，Center_Y)，定义 4 个阈值：上临界值(TOP)、下临界值(BOTTOM)、左临界值(LEFT)、右临界值(RIGHT)，据此可以得出对页面块进行删减的 2 个判断规则：

规则 1：IF Center_X<LEFT ‖ Center_X>RIGHT，则删除该块。

规则 2：IF Center_Y<TOP ‖ Center_Y>BOTTOM，则删除该块。

进行完第 1 次页面块提取后可以利用这 2 个判断规则将位于页面四周的导航栏、广告内容、版权声明等页面块删除。

网页中不仅包括正文标题、发表时间、主题相关图片、正文内容等要抽取的信息，还包括相关阅读等与页面主题相关但不需要抽取的信息以及图片广告、搜索栏等噪音内容。在页面块提取过程中记录该页面块的中心位置的坐标、文字长度(TextLength)、链接文字长度(LinkTextLength)、图片数量(ImageNum)。记正文字数和链接个数的比值为 F。设置阈值 T(试验中 T=2)。据此可以得出对页面块进行删减的 3 个判断规则：

规则 3：IF $F<T$ && ImageNum=0，则说明该块为相关阅读或文字广告链接，删除该块

规则 4：IF $F<T$ && ImageNum>0 && CENTER_Y>HEIGHT/2，即链接较多，文字较少且位于网页的下方，则为图片广告，删除该块。

规则 5：IF TextLength<100，可能为搜索栏或用户评论等噪音，删除该块。

对逐层提取出来的页面块按照以上 5 个判断规则逐层将噪音块删除。

2.3 正文信息提取

逐层分块逐层删减后仍保留在内存块池中的页面块被认为是正文页面块，下面的工作是对这些页面块进行信息抽取。在正式提取内容前要对这些页面块进行分隔条提取和语义块的重构，以保证提取内容的语义完整性。

(1)提取页面主题。包含主题的页面块一般具有以下视觉特征：字号比其他页面块都大；字体颜色与其他块不同；周围有较多的空白；位置在网页的上方。在此假定满足以上条件中的 3 个或 3 个以上即被认为是页面主题块。

(2)提取发表时间。本文利用视觉信息识别包含发表时间的页面块。在视觉上,发表时间一般位于页面主题下方,且字号相对其他内容块较小。利用参考文献[3]中提到的位置和词性双重约束的方式对发表时间进行识别:考虑页面块标题和正文之间的文字,判断它们的词性,对词性为“数词(m)”或“时间词(t)”的文字串,把它挖掘出来作为发表时间。本文采用中国科学院计算技术研究所软件室研发的词法分析器 ICTCLAS[6]进行词性的判断。其对时间信息的分析结果如下所示:

“2004-06-15/m 08:/m 57:/m 45/m”、“2004 年/t06 月/t 5 日/t05:/m 37/m”。

(3)提取正文内容。需要注意的是:有的正文中有小标题,其视觉信息与其他正文内容不同,这上面的分块中已有体现。

3 试验与分析

为验证该方法的可行性,从新浪、搜狐、网易、新华网、人民网 5 大热门网站中各抽取 100 篇网页,共 500 篇网页进行试验,网页内容涉及新闻、财经、军事等多个领域。从页面标题、发表时间、提取的完整率和准确率等方面进行评价。为验证其性能,从中抽取 200 篇进行人工抽取,并进行比对。实验结果如表 1 所示。

表 1 实验结果

网页来源	网页主题提取/%	发表时间提取/%	完整率/%	准确率/%
新浪	93	95	95.2	99
搜狐	94	94	96.1	97
网易	93	95	95.5	96
新华网	90	93	94.9	93
人民网	92	95	95.8	96

准确率和完整率的计算公式如下:

准确率=(正确提取正文信息网页个数/网页总数)×100%

完整率=(完整提取正文信息的网页个数/正确提取正文信息网页个数)×100%

通过对实验结果的分析发现,有些网页的发表时间前后都带有网站的链接,导致该页面块被当作噪音删除。实验数据表明,正文抽取完整率和准确率都达到 90%以上,证明了该方法的可行性。

本文在 VIPS 算法的基础上结合网页正文抽取的特点,实现了一种根据页面视觉特征对 Web 页面进行逐层分块逐层删减的正文信息抽取方法。下一步将对判断规则进行完善,以达到更好的抽取效果。

参考文献

[1] 欧健文,董守斌,蔡斌.模板化网页主题信息的提取方法[J].清华大学学报,2005,45(S1):1743-1747.

[2] CAI D, YU S, WEN J R, et al. VIPS: A vision-based page segmentation algorithm. Microsoft Technical Report, MSR-TR-2003-79. 2003:10.

[3] 于满泉,陈铁睿,许洪波.基于分块的网页信息解析器的研究与设计[J].计算机应用,2005,25(4):974-976.

[4] 孙桂煌,刘发升.基于正文特征的网页正文信息提取方法[J].现代计算机,2008(9):34-37.

[5] JOHNSON R, HOELLOR J, ARENDSEN A, et al. Spring 框架高级编程[M].蒋培,译.北京:机械工业出版社,2006.

[6] 张华平.ICTCLAS[EB/OL]. [2009-08-15].http://mtgroup.ict.ac.cn/~zhp/ICTCLAS.htm. 2002.

[7] 黄文蓓,杨静,顾君忠.基于分块的网页正文信息提取算法研究[J].计算机应用,2007,27(B06):24-26.

(收稿日期:2009-0-0)

作者简介:

安增文,男,1983 年生,硕士研究生,主要研究方向:Web 信息挖掘。

徐杰锋,男,1964 年生,教授,博士,主要研究方向:Web 信息挖掘、数据库应用。