基于节点权重的网页去噪方法的研究

王 健,张 金
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:随着网络信息的不断增多,网页信息不仅成为用户的重要信息来源,同时也是数据挖掘、信息检索等研究的重要数据来源。为提供高质量的文本信息源,页面去噪已经成为网页处理中不可忽视的步骤。随着网页制作技术的不断提升,页面中的视觉元素日益增多,网页节点信息愈加丰富。视觉信息已经成为页面去噪中不可忽视的重要部分。从用户的角度,在浏览网页时,视觉的信息网页能够第一时间反映页面中模块的重要程度。传统的页面去噪技术过多地忽略了页面的视觉特性,面对现今复杂的页面结构,去噪效果大大下降。文中在综合视觉信息和节点信息的基础上,提出了一种基于节点权重的去噪方法,该方法充分考虑了节点的视觉特性和内容特性。实验结果表明,该方法在网页去噪的准确率和召回率上有所提高。

关键词:视觉特性;节点权重;准确率;召回率
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2017)10-0083-04
doi:10.3969/j.issn.1673-629X.2017.10.018

Research on Web Page Denoising Method Based on Node Weight

WANG Jian, ZHANG Jin
(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: As the network information is increasing continuously, website information is not only an important information resource of users, but also important data source for data mining, information retrieval and other studies. To provide the text information with high quality, website denoising has become a nonnegligible step for webpage processing. With the continuous improvement of webpage making technology, visual elements in webpage are raised increasingly, and the information of webpage node becomes richer and richer. Visual information has been a nonnegligible and important part in webpage denoising. From a user's point of view, the visual information can immediately reflect the importance of module in the page when browsing the web page. Traditional webpage denoising technology is neglected in the visual characteristics of webpage too much. Facing to the current complex webpage, the denoising effects are decreased greatly. Based on the comprehensive visual information and node information, a noise weight-based denoising method is proposed which fully considers the visual and content characteristics of nodes. The experimental results indicate that its accuracy rate and recall rate is improved to certain content.

Key words: vision characteristics; node weight; accuracy rate; recall rate

0 引 言

互联网的飞速发展使得网络上的信息急剧增加,网络已经与人们的生活紧紧相联。网络就像一个巨大的信息库,提供了各种各样的信息。人们可以从中查询自己需要的知识,在丰富了生活的同时,给工作和学习也带来了巨大的益处。但是,在网页规模不断扩大的同时,网络上的信息并不像图书馆的书那样编排得分类整齐,使得搜索和获取信息变得非常困难。不仅如此,当浏览网页获取信息时,往往发现网页中充斥着大量的无关信息,如导航栏、广告信息、浮动提示窗等,称之为“噪声”。这些“噪声”往往与页面的正文内容无相关性,而且影响了网页的可观性,因此网页的噪声给信息的获取增加了一道门槛。

面对海量信息,用户能够快速定位所需信息就成了当务之急。为帮助用户获取更精确的信息,信息检索[1]、文本挖掘[2]等技术应运而生。而数据源的质量将直接与这些技术息息相关。网页中的噪声内容对于检索技术来说,不仅有可能导致搜索的主题漂移[3],而

收稿日期:2016-11-15 **修回日期**:2017-03-07 **网络出版时间**:2017-07-19
基金项目:教育部专项研究项目(2013116)
作者简介:王 健(1991-),男,硕士,研究方向为大数据。

且会影响索引建立,使得上述技术的处理效率和效果不够完美。因此,研究网页去噪技术,即去除页面中无关的导航栏、无用标签等信息,准确抽取网页的正文内容,是非常有必要的。

目前的网页去噪领域已存在许多研究成果,噪声的处理也能提高页面正文的提取效果。随着硬件水平的提高,能够展现的页面内容也越来越丰富,网页的制作技术也百花齐放。网页的制作者为了使网页美观,为页面增加了更多的样式修饰,提升了用户体验,但同时随之而来的是越来越复杂的页面结构,以及越来越多的噪声信息。传统的页面从 HTML4 发展到 HTML5[4],原有的 table 布局也被取代。传统的页面去噪声技术已经不能适应复杂的新型网页结构。因此研究出相对简单的去噪算法,能够应对复杂的新型网页的处理需求,让算法具有更好的健壮性,也是一个很有挑战性的课题。

1 相关工作

目前国内外关于页面噪声去除的研究中,已经取得了不少成果,采用的方法多种多样,涉及各个领域。但从去噪方法的成果来看,可以分为如下几种:

(1)基于 DOM 结构的方法。

DOM(Document Object Model)即文档对象模型[5]的缩写,根据网页的标签结构信息能够将网页松散的 HTML 代码表示成结构清晰的 DOM 树,因此较多的页面去噪工作都是在 DOM 结构的基础上进行的。文献[6]提出了基于 DOM 的页面分块方法,利用这种方法去除页面的噪音信息,抽取正文内容。文献[7]在建立 DOM 树的基础上,以页面中文本节点的视觉特性为特征,分别使用聚类等方法对 DOM 节点进行分类,得到页面的正文信息。由于 DOM 结构的节点中包含的语义等信息较少,文献[8]在 DOM 树的基础上移入 STU(Semantic Textual Units,语义文本单元),添加语义特征,构造相应的 STU-DOM 树,并对其进行基于树结构的过滤和基于特征的剪枝,完成对网页的去噪工作。

由于在制作网页时,为了可维护性,采用嵌套的方式书写 HTML 的标签,将网页信息转换成 DOM 结构的方法有较好的可适应性。但是直接将网页转换成对应的 DOM,不考虑网页原有的布局位置等信息,在 HTML 代码不具有嵌套规范的情况下,生成的 DOM 树将严重影响叶子节点的分析以及后续的剪枝操作。

(2)基于生成模板的方法。

该方法通常是根据一类网站共同的结构特征。在此假设的基础上,通过训练模式匹配和归纳学习生成包装器得到该类网站的模板。包装器的目的就是把网页中有用的结构信息以结构化的形式存储起来,在抽取页面时,利用生成的包装器去除页面中的噪音。李文立等利用标签对形式生成的树结构抽取网页模板[9],效果较好,但没有对前期的页面进行去噪操作,抽取页面正文的算法时间复杂度较高且没有分类,模板健壮性较低。文献[10]采用一些网页作为训练集找到相应的 Xpath,将其用来抽取相似网页的正文。基于生成模板的方法,往往由于训练集的数量问题,不仅效率较慢,且训练集的不准确会严重影响网页的清洗率。

(3)基于统计的方法。

该方法克服了传统网页内容抽取方法需要针对不同网页结构的问题,具有一定的普遍性,不需要生成模板,大大提高了正文抽取速度。文献[11]针对制定规则,提出对标签建立标签库,并根据中文字符数最多的决定网页正文块,但不利于短正文页面的抽取。孙承杰等[12]提出正文信息只能位于<table>节点,将页面 HTML 表示成 DOM 树后对每个<table>节点进行处理,比较节点中的中文字符数量。该方法虽利用了中文网页的特性,实现简单,健壮性强,但未考虑英文网页,且对短正文网页效果不理想。

(4)基于视觉分块的方法。

通常在浏览网页时,人们往往将不同的功能区域看成不同的语义块。较早的分块方式是按照 HTML 的树形结构进行[13],但随着 HTML 的发展,仅仅依赖树形结构,不足以满足通用性。2003 年,微软亚洲研究院提出基于页面视觉分块的算法(VIsion-based Page Segmentation,VIPS),利用页面的可视化信息在树形结构的基础上进行网页分块。然而它仅仅是一种分块算法,利用已有的视觉信息,并未对页面进行净化操作,可以在算法的基础上加入规则进行页面净化操作。文献[14]通过修改 VIPS 算法迭代过程,在块划分后进行一系列的分隔条提取和语义块重构,采用制定规则对页面进行去噪操作。VIPS 算法充分考虑了用户的视觉习惯,但由于分隔条提取和语义块重构需要过多的人工参与,复杂度较高,且缺乏对网页中和信息的利用。

文中在 VIPS 算法分块的基础上,提出样式树,再根据链接比及树路径距离生成相应的权重树,自动调整权重,根据权重进行剪枝操作,生成去噪页面。

2 样式树定义

样式树由 DOM 树演化而来[15],主要包含两类虚拟节点:样式节点(Style nodes)和元素节点(Element nodes)。样式节点描述了节点布局或者展现风格,样式节点 A 的表现样式 S_A 是一个序列 $< l_1, l_2, \cdots, l_n >$ 。

其中 l_i 是一个二元组(Tag,Styles)元素,通常 Styles 表示为{width:300,height:200,bg-Color:red}, n 表示样式长度。节点 E 描述节点的属性信息,表示为 E(Tag,Attrs,Content),其中 Tag 表示节点标识,Attrs 表示属性信息,Content 表示节点的文本信息。基本样式树如图 1 所示。

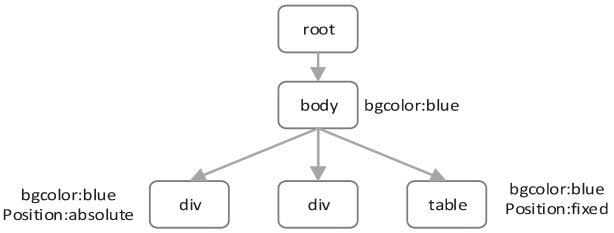

图 1　基本样式树

3　基于节点权重的网页去噪算法

3.1　算法基本思想

基于节点权重的去噪算法在 VIPS 基础上,将 VIPS 生成的基本视觉块树进行样式树的转化,利用样式树节点中的样式特性,将叶子节点划分成细粒度的样式树,再对样式树进行权重标注,根据权重标注进行剪枝,生成去噪页面。基本流程如图 2 所示。

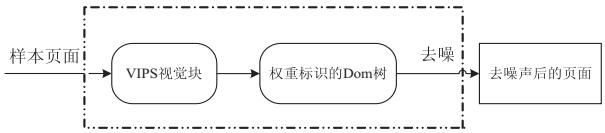

图 2　从样本页面到净化页面的总体流程

通常生成的样式树,无权重表示,在属性节点的基础上,引入权重节点的概念。权重节点 T 表示为 Q_T,记为 $Q(k,d,t,m)$。其中, k 表示链接比,即当前节点中链接数占总链接数的比值; d 表示树路径距离,即当前节点与容器节点在树形结构上的距离; t 表示文本比,即当前节点文本占总文本的比例; m 表示节点私有属性的权重系数。为了使 $H(Q_i)$ 的值落在[0,1]之间,使用节点的标签个数 n 将 $H(Q_i)$ 归一化。

$$H(Q_i) = -\sum_{j=1}^{n}(1-k_j)t_j m_j \log_n(d_j/D) \tag{1}$$

其中, k_i 表示第 i 个标签的链接比; t_i 表示第 i 个标签的文本系数; d_i 表示第 i 个标签的树路径距离; D 表示权重树中的节点路径和。

3.2　视觉块树细粒度化

通常,VIPS 生成的视觉树,只是初步提取了页面的基本布局信息,粗粒度的视觉块树将噪声和正文融合到了相同的块中,必须进行细粒度化。此时对生成的样式树进行样式节点和属性节点的标注。对已经标注完的块节点,进行子元素的相似度分析。子元素的样式节点用二元组 $<l_1,l_2,\cdots,l_n>$ 表示,属性节点标识为 E(Tag,Attrs,Content),由于 l_i 的 Styles 是以键值对的形式存在,在此将键值对转化为样式系数 C_i,将块标签 Tag 表示为 HTML 中对应的 NODE 值,此时 l_i 表示为 (T_i,C_i)。节点相似度判断如下:

$$P(l_i,l_j) = \frac{\sqrt{(T_i-T_j)^2+(C_i-C_j)^2}}{\sqrt{T_i^{\,2}+C_i^{\,2}}\sqrt{T_j^{\,2}+C_j^{\,2}}} \tag{2}$$

当相关系数较小时,将子节点进行分裂。采用自顶向下的层次遍历方式,完成对视觉树的初步分裂。

3.3　细节树剪枝

此时得到的是一棵基于样式的视觉树,在样式和基本属性上已经不可细分,在此基础上进行噪声的判断。根据大量线上页面的统计,噪声区域往往有比正文区域更多的链接比,更少的文本比,以及更浅的树距离。故此处引入权重节点的概念,对细粒度化的视觉块树进行自顶向下的标注,对权重低的节点进行剪枝操作。在初次遍历的过程中,可进行一次简单的预处理,对含有样式树节点中含有键值对 display:none 和 position:fixed 的节点进行删除操作,前者是网页中不做显示的元素,后者是悬浮窗,据大量网页的观察,两者都是判断噪声节点的重要依据。

剪枝算法描述如下:

(1)获取样式树,设样式树为 T_i;

(2)For(样式树的每个节点 Q_i)

(3)if(该节点的 css 属性中含有 position:fixed,display:none 等键值对时) then

(4)删除该节点;

(5)Else if

(6)计算出文本比,节点的距离深度,计算权重值 $H(Q_i)$;

(7)For(样式树的每个节点 Q_T);

(8)删除平级节点中权重小的节点。

4　实　验

4.1　数据集

为了验证文中算法的去噪效果,使用该算法对含有噪音的网页进行处理。考虑到页面抽取时信息获取的客观性,选取网易、新浪等页面各 200 个,考研论坛等论坛型网页 200 个,从网页处理的整体效果出发,进行网页去噪的实验。

4.2　评价指标

在实验中,常见的评测指标有准确率和召回率。由于准确率和召回率介于[0,1]之间,而且不相互独立。所以文中引入同时兼顾准确率和召回率的 F_1,即 F-measure,作为综合评价指标。

准确率为:

$$P = t_0/t_1 \tag{3}$$

召回率为：

$$R = t_0/t_2 \tag{4}$$

其中，t_0 表示当前页面被抽取出的正文块；t_1 表示当前页面中全部的正文块；t_2 表示被当做正文中抽取出来的信息块。

由于在 F-measure 公式中 β 通常用来调节准确率和召回率的权重，而此处重点考虑的是网页抽取的准确率和召回率，所以取 β 为 1，最终用来判断实验效果的公式如下：

$$F_1 = \frac{2P * R}{P + R} \tag{5}$$

4.3 实验结果与分析

为了验证文中算法，分别进行了两组实验，结果如表 1 和表 2 所示[16]。

表 1 文中算法

网页来源	网页数量	P	R	F_1
新浪	200	92.4	95.1	93.7
搜狐	200	95.3	96.2	95.7
网易	200	93.4	97.6	95.6
考研论坛	200	93.9	94.2	94.0
天涯	200	91.3	93.2	92.2

表 2 基于行块分布函数算法

网页来源	网页数量	P	R	F_1
新浪	200	84.2	86.1	85.2
搜狐	200	85.7	85.8	85.7
网易	200	83.2	86.6	84.9
考研论坛	200	86.4	87.4	86.9
天涯	200	82.5	86.4	84.4

从上述实验可以看出，文中算法在准确率和召回率方面要优于基于行块分布函数算法的页面处理效果。基于行块分布函数的方法虽然实现简单，但是对去除标签后的文本分块的数量选取将直接影响网页正文提取的准确率，而且去除标签同时也去除了页面中大量可用的视觉信息，当噪音文本与正文文本混杂时，将会被提取。文中充分考虑了页面的视觉特征，在当前视觉元素丰富的网页中，从网页制作者的方向出发，利用大量的视觉特性，提取视觉系数，再利用正文内容特征，合理去除页面中的噪音块，使正文块更易被识别。

5 结束语

文中在 VIPS 分块的基础上，引入了样式树的概念，取消了原有的基于视觉繁杂的启发式的规则，只使用了 VIPS 粗粒度的视觉分块，对粗粒度的视觉块树进行细粒度的划分，进一步考虑了视觉块之间的相关性，再对标注完权重的样式树进行去噪操作。实验结果表明，该算法可以更好地去除页面中导航栏等局部噪声以及隐藏中正文块的全局噪声。该算法主要针对主题型页面、论坛型页面，但当正文内容和噪音内容相似度较高时，去噪效果不够理想，这是该算法的局限性。在以后的研究中，将进一步分析这些网页的特征，寻求改进方法，增强算法的健壮性。

参考文献：

[1] 欧石燕，唐振贵，苏翡斐. 面向信息检索的术语服务构建与应用研究[J]. 中国图书馆学报，2016，42(2)：32-51.

[2] Witten I H，Frank E. Data mining：practical machine learning tools and techniques[M]. [s. l.]：Morgan Kaufmann Publishers Inc. ，2011：206-207.

[3] 高 琪，张永平. 超链接导向搜索算法中主题漂移的研究[J]. 计算机应用，2009，29(11)：3100-3102.

[4] 刘华星，杨 庚. HTML5-下一代 Web 开发标准研究[J]. 计算机技术与发展，2011，21(8)：54-58.

[5] 李效东，顾毓清. 基于 DOM 的 Web 信息提取[J]. 计算机学报，2002，25(5)：526-533.

[6] 胡金栋. 网页正文提取及去重技术研究[D]. 杭州：浙江大学，2011.

[7] 汪建伟，杨冬青，高 军，等. 一种基于分类算法的网页信息提取方法[J]. 计算机科学，2008，35(3)：91-93.

[8] 王 琦，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.

[9] 李文立，王乐超，宋春雷. 基于 HTML 树和模板的文献信息提取方法研究[J]. 计算机应用研究，2010，27(12)：4615-4617.

[10] Fu Y，Yang D，Tang S，et al. Using XPath to discover informative content blocks of web pages[C]//Proceedings of third international conference on semantics，knowledge and grid. [s. l.]：[s. n.]，2007.

[11] 赵 文，唐建雄，高庆锋. 基于统计的中文网页正文抽取的研究[J]. 电脑知识与技术，2008(1)：120-123.

[12] 孙承杰，关 毅. 基于统计的网页正文信息抽取方法的研究[J]. 中文信息学报，2004，18(5)：17-22.

[13] 刘晨曦，吴扬扬. 一种基于块分析的网页去噪音方法[J]. 广西师范大学：自然科学版，2007，25(2)：149-152.

[14] 穆 琼. 基于视觉特征的网页清洗研究与实现[D]. 北京：北京邮电大学，2013.

[15] Yi L，Liu B，Li X. Eliminating noisy information in Webpages for data mining[C]//Proceedings of the 9th ACMSIGKDD international conference on knowledge discovery and data mining. New York：ACM，2003：296-305.

[16] 高庆宁，吴 鹏，张晶晶. 基于文档对象模型与行块分布算法的网页信息抽取[J]. 情报理论与实践，2016，39(4)：133-137.