文章编号：2095-2163(2020)07-0001-05 中图分类号：TP311 文献标志码：A

面向 Web 新闻与博客的内容提取方法

王金麟，方滨兴，于海宁，马雪阳
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：Web 深刻地改变了社会生活，新闻和博客网站作为其中代表性的消息来源，为人们提供了方便的信息获取方式。在 Web 分析的实际业务中，广告、文章推荐等无关信息的存在，给新闻和博客网页中主要内容的提取带来了负面影响。本文提出了一种区别于抽取模板的新闻和博客内容提取方法 CEVC，通过定义有效字符，对网页内容文件的 DOM 树进行递归计算，确定最具代表性的子节点作为主要内容节点。实验选取了中文与英文网页作为数据集，定义了提取新闻和博客内容的性能指标。对比实验的结果表明，CEVC 在 Web 内容提取方面的性能优于现有方法。

关键词：Web 分析；内容提取；DOM 树

A Content Extraction Method for Web News and Blogs

WANG Jinlin, FANG Binxing, YU Haining, MA Xueyang
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Web has profoundly changed the social life. News and blog sites, as a representative source of information, provide a convenient way for people to obtain. In the actual business of web analysis, the existence of irrelevant information such as advertisements and article recommendations has negatively affected the extraction of the main content in news and blog pages. This paper proposes CEVC, a method for extracting news and blog content, which is different from extracting templates. By defining valid characters, the DOM tree of web content files is recursively calculated to determine the most representative child node as the node of main content. Chinese and English web pages were selected as the data set, and the performance indicators were defined for extracting content of news and blog. The results of the comparative experiments showed that CEVC outperforms the existing method in Web content extraction.

【Key words】 Web analytics; Content extraction; DOM tree

0 引 言

随着互联网产业的蓬勃发展，Web 深刻地改变了社会生活，为信息共享和流动提供了最便捷的方式。在线新闻和博客网站作为两种代表性媒体，在 Web 中起着重要作用。在线新闻具有即时传播事件信息的优势，而博客网站允许用户表达自己的观点并与他人分享。各个国家数十亿的用户被媒体的便利性和即时性所吸引，使得 Web 成为了世界上最大的数据存储库。为了有效地利用 Web 信息，需要提取网页中有用的信息，即主要内容。网页中几乎 40%至 50%的内容都是与主要内容无关的噪音，包含导航面板、广告、相似或推荐的文章、免责声明或版权[1]。因而在许多实际应用中，从网页中提取主要内容至关重要。在文本分析过程中，存储网页的主要内容比存储页面文件消耗的空间要小得多，可以通过自动内容提取进行相对较低的成本构建[2]。同时，随着移动设备的普及，提取适合在手机、平板电脑上显示网页的主要内容成为热点需求[3]。从新闻和博客网站中提取具有不同结构和布局的主要内容，由于网页的庞大性和异构性，这些内容有时也会发生变化，这给提取过程增加了难度。尽管定制化爬虫技术可为特定网站提供非常高的精度，但它们不适合进行大规模内容提取。同时，大多数基于模板的爬取应用都无法跟上网站结构的变化，而有监督技术因人工标记的培训数据集的高昂成本而遭受损失。

本文提出了一种高度自动化的内容提取方法 CEVC，用于从 Web 新闻和博客页面中提取主要内容。与现有方法不同，本文方法具有良好特性：

(1)模板独立方法；

(2)不需要训练数据；

(3)不对网页的结构做任何假设。

导航面板或广告中通常包含许多超链接，以吸

基金项目：国家重点研发计划(2016QY03D0501, 2017YFB0803300)；国家自然科学基金(61601146, 61732022)；四川省科技计划项目(2019YFSY0049)。

作者简介：王金麟(1991-)，男，博士研究生，主要研究方向：网络空间安全；方滨兴(1960-)，男，博士，教授，博士生导师，中国工程院院士，主要研究方向：网络空间安全；于海宁(1983-)，男，博士，助理研究员，硕士生导师，主要研究方向：网络空间安全；马雪阳(1991-)，男，硕士，主要研究方向：网络空间安全。

收稿日期：2020-03-30

引用户点击。分散各处的冗杂文本通常包含的停用词并不多,无法形成一个结构合理的句子。在常见的新闻和博客网页中,与任何超链接无关但包含停用词的文本片段更有可能是主要内容。主要内容通常指新闻和博客页面中的专有内容。因此,可以利用有效字符的分布来定位主要内容块。将 Web 网页文件解析为 DOM 树后,首先使用自上而下的递归方法计算与每个 DOM 节点关联的有效字符数。有效字符数越多,意味着该节点越有可能代表内容段落,有效字符的密集程度成为获取主要内容块的重要特征。最终,将主要内容块中除噪声部分之外的有效字符作为主要内容的输出结果。

1 相关工作

Web 内容提取方面,采取最直接的方法是运用手工模板[4],由业务人员使用 XPath,正则表达式,甚至编程语言直接从 HTML 模板中提取嵌入式文本而构建的。尽管这样的做法可以产生高精度的提取结果,但是在大规模提取各类型网页的应用场景下,这种做法在模板的构造和维护上的巨大成本是不可接受的。为了解决自动化程度低的问题,现有研究倾向于开发自动模板检测功能[5]。该功能使用由同一模板生成的一组网页来学习通用结构。与此同时,需要为每个网站更精确地构建一个模板。在实际应用中,这些模板需要随目标网站的变化而保持日常维护,模板故障的检测和重构成为不可避免的棘手问题。

业界研究中存在一些基于统计信息和网页启发式的内容提取方法,这些方法是完全自动的并且独立于模板。Bar Yossef 等将网络文档(例如文本,图像,脚本等)分割成多个块,并分析它们的特定特征的普遍性[6]。通过选择对应的块来提取内容文本最好的所需功能。链接配额过滤器(LQF)使用文档块中比例很高的链接内容来检测导航菜单或类似结构[7];Gottron 在检测过程中使用了诸如停用词比率和其他启发式等方法[8];Davison 通过使用机器学习的方法识别广告、冗余和不相关的链接,从而过滤掉网页中的噪音[9]。这些系统的缺点在于,它们需要大量的手动标记训练数据,并且提取结果可能对注释的质量产生敏感影响。

页面 HTML 源文件中的标签和文本之间存在相对稳定的映射关系,这给解决 Web 内容提取问题提供了可靠的解决思路。正文文本提取方法(BTE)将 HTML 文档解释为单词和标签标记的序列,并将包含大部分单词但不包含大多数标签的连续区域识别为有效提取文本块[10]。内容代码模糊方法(CCB)在源代码字符序列中找到代表均匀格式文本的区域[11]。内容标签比率方法(CETR)在逐行计算文本与标签的比率后,将结果值聚类为内容和噪声区域[12]。虽然计算方法简单有效,但是该方法易受页面源代码样式更改的影响。以上提到的研究方法依赖于非结构化的文本特征,并未过多地关注文本结构。Sun 等提出了基于 DOM 树和统计信息的方法,通过计算 DOM 节点的文本密度进行内容提取[13]。Wu 等通过计算文本内容的路径比率来区分新闻内容与非新闻内容[14]。结合文本结构特征,对于 Web 网页进行内容提取,成为值得研究的解决思路。

2 内容提取

在从新闻和博客网页中提取内容的过程中,通过研究总结发现与任何超链接无关但包含停用词的文本片段更有可能是真正的网页主要内容。本文将详细描述基于有效字符的内容提取方法。

2.1 有效字符

文档对象模型(Document Object Model, DOM)是一种标准化的、与平台和语言无关的接口,用于访问和更新文档的内容、结构和样式。每个 HTML 页面对应一个 DOM 树,其中标记是内部节点,具体的文本和图像是叶节点。新闻与博客网页中的非内容文本通常位于吸引用户点击的超链接中。DOM 树中的超链接由 < *a* > 标签组成,但是一些详细的文本可能并不直接属于标签 < *a* >。例如: < *a* > < *span* > TEXT < */span* > < */a* > 同样也是锚定文本,因此需要检查文本在 DOM 树中所有的祖先来确定文本的有效性。停止词通常指的是一门语言中最常见的单词,比如英语中的 on、the、in。虽然这些词在 NLP 处理中毫无用处,但是可以使用它们来过滤不相关的文本。结合超链接和停止词启发法,接下来对有效字符进行定义。

定义 1 如果 DOM 树中的字符不是任何标记的后代,且至少包含一个停止词,那么它们就是有效字符。

有效字符的数量是衡量 DOM 节点重要性的指标。与同一级别的其他兄弟节点相比,具有更多有效字符的 DOM 节点更有可能携带重要的文本信息。这有助于确定 Web 页面的某一部分是否有意义。显然,将内容对应具有更多有效字符的 DOM 节点是合理的。

将 HTML 文档解析并用 DOM 树表示之后,可以使用 *countVC* 算法自顶向下地递归计算与每个节点关联的有效字符数。在将 DOM 树中的 < *style* >

等无关标签去掉后，计算每个节点的有效字符数，并将其累加到父节点上。如算法 1 所示，算法 *countVC* 遍历整棵 DOM 树的运行时间为 $O(n)$，与树中节点的数量成线性关系，即 n 为 DOM 树中节点的数量。通过算法计算后，有效字符的数量作为一个属性附加到 DOM 树的每个节点上。

算法 1　算法 *countVC*

Input：DOM node *N*

Output：DOM node *N*

```
1：if N has child nodes then
2：    for NC ∈ N.childnodes do
3：        countVC(NC)
4：    end for
5：    N.parent.amountVC += N.amountVC
6：else
7：    if N has valid characters then
8：        length = getNonSpaceLength(N)
9：        N.parent.amountVC += length
10：end if
11：end if
```

2.2　定位内容区块

准确地找到含有内容的网页区块，可以通过寻找最多有效字符的方式进行判断。DOM 树中存在有效字符的子节点有很多，需要对判断内容区块的标准进行定义。

定义 2　最大有效字符比率（*MVCR*）表示 DOM 树节点结构内包含有效字符的程度，式(1)。

$$MVCR_n = \max nc \in childnode(n) \frac{VC_{nc}}{VC_n}. \quad (1)$$

其中，n 是一个 DOM 树的节点；nc 是 n 的子节点；VC_n 表示 n 节点中的有效字符数量。

在对于 DOM 树中所有节点的 *MVCR* 值进行计算后可以发现，比较含有内容区块的节点的父节点，其 *MVCR* 值较小。这是由于内容区块节点的子节点往往携带较少的内容信息。可以通过设立阈值，建立算法运用 *MVCR* 定位有效区块。算法 *locateCB* 显示了定位主要内容块的过程。如果当前节点的 MCR 小于阈值 α，该算法将收集此节点内的有效字符并终止。当递归进入子节点，阈值比较条件未被达成时，计算结果返回父节点。如算法 2 所示，算法 *locateCB* 的时间复杂度为 $O(n)$，其中 n 表示 DOM 树的高度。

算法 2　算法 locateCB

Algorithm 2　Algorithm locateCB

Input：DOM node *N*

Output：Content block node

```
1：maxNode = null
2：maxVC, totalVC = 0
3：for NC ∈ N.childnodes do
4：    totalVC += NC.amountVC
5：    if NC.amountVC>maxVC then
6：        maxVC = NC.amountVC
7：        maxNode = NC
8：    end if
9：end for
10：if maxNode is not null then
11：    if maxVC/totalVC<α then return N
12：    else
13：        locateCB(maxNode)
14：    end if
15：else return N.parent
16：end if
```

3　实　验

将使用来自各种新闻和博客网站的真实数据集，对本文方法 CEVC 的有效性进行实验性验证。

3.1　数据集

实验使用数据集的数据来源分为二类：

(1)中文网页数据。通过搜索引擎，挑选出知名的新闻与博客网站各 8 个；

(2)英文网页数据。这里使用了 CETR 所使用的数据集，该数据集中包含 8 个流行的英语新闻网站。

将以上数据来源中的网站形成列表，依次在这些网站中随机收集各 50 个网页，并通过专家知识对网页主要内容区块进行标记，形成中文与英文数据集，见表 1。

3.2　性能指标

如公式(2)所示，本文使用三种指标来评估和比较不同方法的性能水平，分别是精密度 P、召回率 R 和综合评价指标 $F1$。

$$P = \frac{LCS(e,g).length}{e.length},$$
$$R = \frac{LCS(e,g).length}{g.length}, \quad (2)$$
$$F1 = \frac{2 * P * R}{P + R}.$$

在性能指标计算过程中，e 表示抽取结果文本，g 表示经过专家知识标记的正确文本，$LCS(e,g)$ 表示 e 和 g 之间的最长公共子序列。文档中的每个词汇都被认为是不同的，即使其中两个词汇内容一致，

这两个词汇在位置上的不同也会影响性能指标的计算结果。值得注意的是,中文较之于英文的分词难度更大,使得指标的计算更为复杂。

3.3 参数设置

算法 locateCB 中存在一个阈值 α,其设定了收集有效字符过程的停止条件。在 0 到 1 的取值范围内,如果 α 较低,那么最终取得的有效字符内容可能较少,召回率 R 较低;反之,准确率 P 将随着提取内容的增多而降低。

本文随机使用 20% 的中文数据集分析指标与阈值 α 之间的关系,实验效果如图 1 所示。根据观察可知,准确率P和召回率R的变化符合预期。当 α 的取值范围在 0.3 到 0.6 之间时,三类指标的变化并不敏感。在后续实验中,α 值被设定为 0.5。

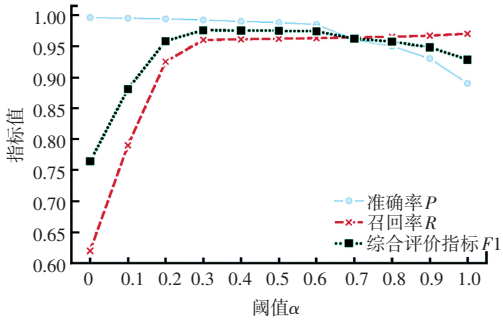

图 1 对应不同 α 值的指标表现

Fig. 1 Indicator performance corresponding to different values of α

表 1 实验结果

Tab. 1 Experimental result

网站	类型	语言	URL	CETR			CEVC		
				P	*R*	*F*1	*P*	*R*	*F*1
新浪新闻	新闻	中文	news.sina.com.cn	0.829	0.977	0.897	0.998	0.964	0.981
网易新闻	新闻	中文	news.163.com	0.537	0.998	0.698	0.999	0.966	0.982
搜狐新闻	新闻	中文	news.sohu.com	0.39	0.993	0.560	0.96	0.935	0.947
凤凰新闻	新闻	中文	news.ifeng.com	0.365	0.976	0.531	0.999	0.948	0.973
腾讯新闻	新闻	中文	news.qq.com	0.556	0.999	0.714	0.984	0.957	0.970
新华网	新闻	中文	www.xinhuanet.com	0.949	0.993	0.971	0.999	0.977	0.988
CCTV 新闻	新闻	中文	news.cctv.com	0.502	0.981	0.664	0.972	0.983	0.977
环球网	新闻	中文	www.huanqiu.com	0.607	0.965	0.745	0.995	0.967	0.981
新浪博客	博客	中文	blog.sina.com.cn	0.839	0.94	0.887	0.991	0.914	0.951
网易博客	博客	中文	blog.163.com	0.504	0.801	0.619	0.994	0.978	0.986
搜狐博客	博客	中文	blog.sohu.com	0.772	0.976	0.862	0.993	0.96	0.976
凤凰博客	博客	中文	blog.ifeng.com	0.657	0.666	0.661	0.991	0.935	0.962
博客中国	博客	中文	www.blogchina.com	0.85	0.932	0.889	0.993	0.959	0.976
天涯博客	博客	中文	blog.tianya.cn	0.758	0.919	0.831	0.988	0.979	0.983
博客网	博客	中文	www.bokee.com	0.763	0.984	0.860	0.996	0.984	0.990
博客园	博客	中文	www.cnblogs.com	0.863	0.985	0.920	0.971	0.942	0.956
NY Post	新闻	英文	www.nypost.com	0.614	0.984	0.756	0.919	0.935	0.927
Freep	新闻	英文	www.freep.com	0.581	0.897	0.705	0.759	0.86	0.806
Suntimes	新闻	英文	www.suntimes.com	0.87	0.987	0.925	0.971	0.975	0.973
Techweb	新闻	英文	www.techweb.com	0.567	0.986	0.720	0.862	0.935	0.897
Tribune	新闻	英文	www.tribune.com	0.949	0.92	0.934	0.973	0.912	0.942
NYTimes	新闻	英文	www.nytimes.com	0.953	0.959	0.956	0.954	0.975	0.964
BBC	新闻	英文	www.bbc.com	0.711	0.977	0.823	0.919	0.987	0.952
Reuters	新闻	英文	www.reuters.com	0.402	0.91	0.558	0.929	0.901	0.915
中文数据集				0.671	0.943	0.769	0.989	0.959	0.974
英文数据集				0.706	0.953	0.797	0.911	0.935	0.922
总体				0.683	0.946	0.779	0.963	0.951	0.957

(下转第 10 页)