doi: 10.12052/gdutxb.170152

基于文本块密度与标签路径等特征的正文提取

杨　贤[1]，唐超兰[1]，李　航[2]

（1. 广东工业大学　艺术与设计学院，广东　广州 510090；2. 广东工业大学　计算机学院，广东　广州 510006）

摘要： 为了解决网页中除正文信息外还包含网页导航、广告和免责声明等噪声信息的问题，本文提出一种基于标签路径等多特征和文本块密度的正文提取方法. 首先根据文本块密度特征确定正文区域，然后在区域内使用标签路径等特征剔去噪音节点，最后抽取该文本块中的正文节点内容. 该方法有效解决了网页正文块中噪声信息难以过滤和标签路径等特征易对正文部分外较长文本误抽取的问题，且无须训练和人工处理. 从知名网站上随机选取新闻网页数据集进行实验，验证了该方法在不同数据源上都具有很好的适用性，抽取精确度优于CETR、CETD等方法.

关键词： 正文抽取；文本块；标签路径；文本密度

中图分类号： TP391　　**文献标志码：** A　　**文章编号：** 1007–7162(2018)02–0051–06

Text Extraction Based on Text Block Density with Tag Path and Other Features

Yang Xian[1], Tang Chao-lan[1], Li Hang[2]

(1. School of Art and Design, Guangdong University of Technology, Guangzhou, 510090, China;
2. School of computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Most of web pages contain content information as well as a lot of noisy information. In order to address this problem and improve the accuracy of web page extraction, a web page extraction method is proposed via text block density with tap path and other features. The proposed method mostly combines the advantages of text block extraction method and label path extraction method. First, the block of the text is determined according to the density feature of the text block, and then the tag path method is used to remove the noisy node in the block, the text node in the text block is extracted from the content finally. This solution effectively solves the problem that the noisy information in the text block is difficult to filter and the tag path method is easy to extract the long text from the noisy block. In the end, experiments show that the solution is better than CETR and CETD in most cases.

Key words: content extraction; text block; tag path; text density

网页信息是互联网文本挖掘以及其他大数据应用的主要来源，但网页信息除了有用的正文内容外，还包含如导航栏、推荐广告、推荐链接、版权信息等噪声信息，这些信息对管理、检索、挖掘和分析等研究造成严重干扰.

目前国内外关于网页正文抽取方面的研究大致分为以下4类：基于包装器的方法、基于模板的方法、基于视觉的方法和基于统计的方法[1]. (1) 基于包装器的网页信息抽取方法最早是通过手工构建包装器来实现的，例如W4F[2]、SCRAP[3]等. 这种方法通常是针对某些特定网站去构建包装器，具有精度高的优点，但耗费代价大，对不同网站实用性差. 随着机器学习技术的发展，人们通过训练学习的方式来自动分析网页结构，生成包装器规则. 如Kim等[4]以文本块的链接密度、标签密度等为特征进行学习训练，然后通过决策树的形式判断文本块是否属于正文块，这种改进需要提供大量标注页面用于学习规则，代价较高. (2) 基于模板的方法是根据网站通常用模板来构建，相同网站的网页结构具有很高的相似度. 这类方法对于结构差别较大的网站效果不佳，此时需

收稿日期： 2017-10-16
基金项目： 广东省部产学研专项资金企业创新平台资助项目(2013B090800042)
作者简介： 杨贤(1982–)，男，助理研究员，主要研究方向为数据挖掘、人机交互.
通信作者： 唐超兰(1969–)，女，教授，主要研究方向为交互设计. E-mail：tangchl@gdut.edu.cn

重新构建新模板，欠缺可维护性[5]. (3) 基于视觉的方法是利用网页内容分布位置和节点的样式等视觉特征来对网页进行分块和信息抽取. 黄文蓓等[6]提出基于TVPS的算法，但计算量很大，难以推广. 总的来说，基于视觉的方法对表现形式单一、正文和噪音内容CSS样式上区别很大的网页有不错的抽取性能，但是欠缺通用性. (4) 随着互联网技术的发展，许多学者开始利用网页内容中的统计特征来进行信息抽取. Weninger等[7-8]提出基于HTML代码的标签比特征值的CETR算法，该方法对新闻网页具有良好通用性. Sun等[9]根据正文部分内容集中且多为长文本的特点，提出了基于文本密度特征的抽取算法(Content Extraction with Text Density, CETD)，该方法能够有效地抽取网页正文块，但是对于部分正文块内有噪音的网页则难以保证抽取精度. 吴共庆等[10]提出了基于标签路径特征的CEPR算法，该算法虽然能够有效地区分网页中的正文内容和噪声内容，但是存在对正文区域和噪音区域边界不敏感的问题，容易将噪音部分的版权申明、长文本评论等误认为正文一并提出.

针对以上问题，本文结合文本密度方法能有效确定网页正文块和标签路径等多特征能有效区分正文节点和噪音节点的优点，设计出基于文本块密度和标签路径等多特征的正文提取方法：首先利用文本块密度特征找到正文所在文本块，再利用标签路径等特征剔除正文块内噪音节点，然后提取正文. 该方法无需训练集，简单有效，且提取效果好.

1 相关技术原理

1.1 DOM树

DOM (Document Object Model, DOM)树是网页的数据结构，通过将网页解析成树状结构，可以很方便地对网页进行分析处理，这是目前表示和处理网页非常有效的技术.

1.2 文本块和文本块密度

观察图1示例网页以及大量主题型网页发现:

(1) 网页的内容都以块状分布，且正文多集中在一块，而噪音则分散为多块.

(2) 网页正文内容通常非常集中，而且有很多长文本，超链接数量少，而噪音内容则一般为短文本，且超链接个数多.

图1 腾讯新闻网页示例

Fig.1 Tencent news web page example

根据以上特征，设T为网页解析树，Tb_v是T上以节点v为根节点的子树，其中v为非文本节点，若Tb_v不为空，则称Tb_v为一个文本块. 同时，设定v的文本块密度(Text block density, TBD)TBD_v为节点v所有子节点为根的文本块中非链接文本字符总数与块内非链接标签总数的比值，公式为

$$\mathrm{TBD}_v = \frac{C_v + 1}{T_v + 1}. \tag{1}$$

其中，C_v为文本块v内的非链接文本字符总数，T_v为文本块v内文本非链接标签个数. 为了避免分母为0，计算时分母加1，保证公式的计算有意义.

图2为图1所示新闻网页对应的TBD特征值的分布图. 纵坐标为文本块对应TBD值，横坐标为文本块对应的序列值. 通过观察发现TBD值最大的对应正文块，此特征能有效地将网页中的正文文本块和噪声文本块区分开.

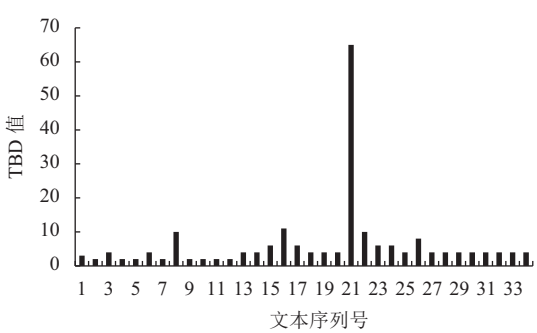

图2 示例网页的TBD直方图

Fig.2 TBD histogram of web page example

图1为单一正文块的网页示例图，对于正文部分分为多块的网页，文献[9]经实验分析表明，以<Body>文本块密度为阈值，取文本密度大于阈值的文本块为正文块有着很好的效果.

根据网页的文本块密度特征很容易找到正文分

布区域，对多数正文区域没有噪音的网页有着良好的抽取效果，但有部分网页（如图3所示）的正文部分包含广告信息（红色框内为噪音），这种情况下，基于文本块的抽取方法则很难保证精度.

图 3　Techweb网页示例

Fig.3　Web page example of Techweb

1.3　基于标签路径等特征的网页正文抽取

1.3.1　标签路径

设T是一棵网页解析树，v是网页解析树的一个节点，从T到v必有一条简单路径，这条路径上经过的所有节点的名称组成的序列即为标签路径. 对于div节点，若其包含class属性，则节点的名称为其属性名，否则与其他元素节点一样都为标签名，对于文本节点，其名称为“Text”. HTML代码如下：

```
<div>
    <div class=“left”>
        <p>文本内容</p>
        <p>文本内容</p>
        <p>文本内容</p>
    </div>
    <div class=“right”>
        <a>链接内容</a>
        <p>噪音内容</p>
        <a>链接内容</a>
    </div>
</div>
```

如图4所示的DOM树中，左侧的p文本节点的标签路径为div/left/p/text，而右侧p节点对应路径为div/right/p/text. 这样定义是因为通常同一个class属性的div块内路径相同的节点多同为正文节点或噪音节点，而在不同块内相同路径的节点则不一定.

1.3.2　文本长度和标点符号的统计规律

随机选取网易、新华网、新浪等网站里的500个主题型网页，统计几种常见标点符号在网页正文中出现个数N_1和在页面中出现的个数N_2，结果如图5所示；统计节点在文本长度取不同值时出现在正文中的个数N_3和在页面中出现的个数N_4，结果如图6所示.

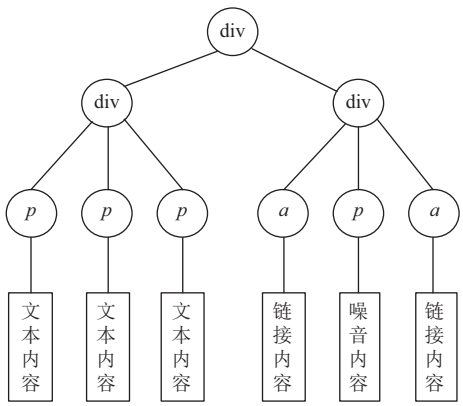

图 4　HTML代码对应的DOM树示例

Fig.4　DOM tree example of HTML code

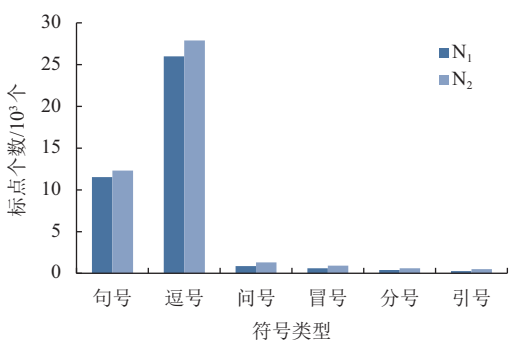

图 5　标点符号统计分布图

Fig.5　Punctuation statistics distribution

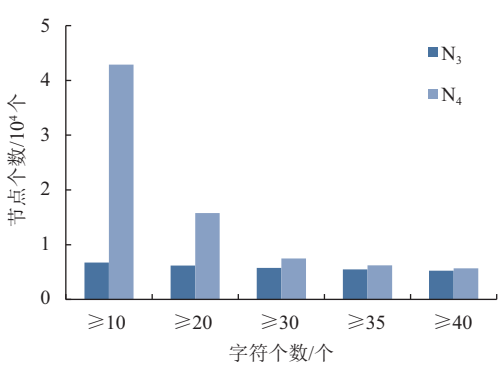

图 6　文本长度统计分布图

Fig.6　Text length statistical distribution

从上述统计结果可以发现以下规律：

(1) 网页中的标点符号主要出现在正文部分，噪音部分出现标点符号较少.

(2) 正文的文本节点有很大概率包含较长文本，噪音节点有很大概率包含短文本.

1.3.3 融合标点符号和文本长度特征

根据标签路径相同的节点多同为正文或同为噪音的特征，设p为网页解析树T的一个标签路径，定义文本标签路径比(Text to tag path ratio, TPR)TPR_p等于路径为p的文本节点的字符之和与节点个数的比值. 同时设定标点标签路径比(Punctuation to tag path ratio, PPR)PPR等于路径为p的文本节点的标点之和与节点个数的比值.

从图5和图6可以看出，文本长度和标点个数均能在一定程度上区分正文节点和噪音节点. 文献[11]的实验表明，融合多个特征可以取得更好的效果. 定义融合特征TPF. 其计算公式为

$$\mathrm{TPF}_p = \mathrm{TPR}_p \times \mathrm{PPR}_p. \tag{2}$$

计算图1示例网页文本节点的TPF值，结果如图7所示. 从图中可以看出，噪音节点和正文节点间的TPF值差距很大，说明融合特征能很好地区分正文节点和噪音节点. 但该特征忽略了正文部分较短正文节点，容易造成部分短文本的丢失，通常需要对其进行平滑处理.

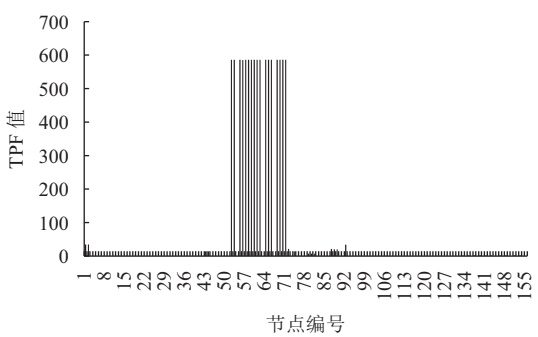

图 7 示例网页TPF值直方图

Fig.7 TPF value histogram of sample web page

1.3.4 特征平滑

借助CETR[7]方法的思想对文本节点TPF值进行高斯平滑，平滑后的效果如图8所示. 从图中可以看出经过平滑后，正文部分之间的许多短文本节点的TPF值有明显增加，设定阈值即可很好地区分开正文内容和噪音内容.但通过实验发现，对于部分有版权申明、较长网络评论的网页，该方法容易将这些较长噪音文本误抽出来.

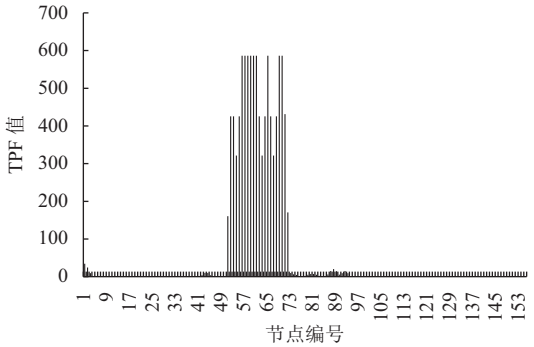

图 8 平滑处理后示例网页TPF值直方图

Fig.8 Smoothed TPF value histogram of sample web page

2 基于文本块密度和标签路径等特征的网页正文抽取

2.1 提取方法

单纯基于标签路径的正文抽取方法对于正文节点和噪音节点有比较好的区分能力，但是容易将版权申明以及网页评论等较长文本内容误抽取. 通过观察发现，这些较长噪音通常和正文内容不在同一个文本块内. 而基于文本块的抽取方法对于正文块和噪音块的划分则较为明确，但是块内可能会有一些噪音节点存在，这些噪音不易剔除. 结合这两种方法，提取正文块内的正文节点，能很好地融合两者的优势，改进不足.

提取方法如图9所示，实线区域内为正文分布块，利用标签路径的方法可以有效地剔除块内一些噪音节点，同时确定正文块的范围可以有效避免基于标签路径方法误提取噪音部分的长文本（实线区域外的几个TPF值较大的节点）.

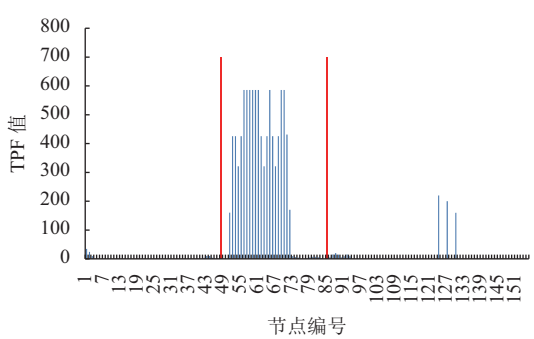

图 9 结合文本块密度与标签路径等特征

Fig.9 Text block density with tag path and other features

2.2 提取步骤

对于给定的网页，正文抽取步骤如下：

(1) 预处理：将网页解析成DOM树，并剔除不可视节点.

(2) 获取待提取文本块：根据网页DOM树计算各个块的文本密度，并将文本密度大于<body>块的文本块的上一级文本块作为待提取块.

(3) 获取标签路径集合：计算每条标签路径的TPR值，设定阈值，获取正文节点候选的路径集合.

(4) 提取正文：将 (3) 的候选路径集合与 (2) 获取的文本块中的路径集合求交集，将交集中路径节点的文本提取，输出为网页正文.

3　实验结果和分析

3.1　实验设置

本文实验数据是从8个知名新闻网站采集得到的. 网站分别为BBC、Freep、Nytimes、Techweb、新浪、网易、搜狐、新华网，每类网站提取200个网页.

实验采用准确率P、召回率R和F均值作为评价网页抽取性能的指标. 其相关计算公式为

$$P=\frac{S_{\mathrm{a}}\cap S_{\mathrm{m}}}{S_{\mathrm{a}}},\tag{3}$$

$$R=\frac{S_{\mathrm{a}}\cap S_{\mathrm{m}}}{S_{\mathrm{m}}},\tag{4}$$

$$F=\frac{2\times P\times R}{P+R}.\tag{5}$$

其中，S_{a}为使用算法自动抽取结果集合，S_{m}为手工标注结果集合. $S_{\mathrm{a}}\cap S_{\mathrm{m}}$为自动抽取内容中抽取正确的内容. 准确率$P$指抽取的正确内容占抽取内容的比率；召回率$R$指自动抽取的正确内容占手工标注正确内容的比例；$F$值是抽取性能的一个综合评价值.

3.2　实验结果

将本文方法和目前主流的Web正文信息抽取方法CETR、CETD进行对比，结果如表1所示.

表 1　实验结果对比

Tab.1　Comparison of experimental results

数据来源	CETD			CETR			本文方法		
	P_{ave}	R_{ave}	F_{ave}	P_{ave}	R_{ave}	F_{ave}	P_{ave}	R_{ave}	F_{ave}
BBC	0.84	0.95	0.89	0.99	0.95	0.96	0.88	0.92	0.90
Freep	0.89	0.89	0.89	0.83	0.92	0.87	0.88	0.99	0.94
Nytimes	0.98	0.95	0.97	0.98	0.94	0.96	0.98	0.90	0.94
Techweb	0.73	0.99	0.86	0.82	0.96	0.89	0.84	0.90	0.87
网易	0.82	0.90	0.86	0.80	0.92	0.86	0.90	0.87	0.89
搜狐	0.80	0.84	0.82	0.75	0.90	0.82	0.86	0.92	0.89
新浪	0.88	0.93	0.91	0.82	0.96	0.89	0.96	0.96	0.96
新华网	0.95	0.93	0.94	0.80	0.94	0.87	0.94	0.89	0.92
平均值	0.86	0.92	0.89	0.84	0.93	0.89	0.90	0.91	0.91

从表1中可以看出CETR和CETD方法在大部分数据集上都能取得不错的实验结果. 但仔细观察可发现CETD方法在Techweb上的准确率最低，通过查看相关网页发现该网站正文块内含有部分噪音信息，而CETD方法正是基于块的提取方法，因此准确率会有所下降. 观察CETR方法可以发现，其在搜狐网站的准确率比较低，查看网页发现，这个网站正文下方大多有较长的评论，CETR方法则非常容易将较长评论误认为正文，因此准确率有所影响. 反观本文算法，其性能非常稳定，其结合了基于块和标签路径两种方法的优势，既能很好地处理正文块内有部分噪音的网页，也能处理长评论干扰的网页，因此其准确率较CETD和CETR方法有很明显的优势. 通过比较整体性能发现，CETD和CETR的召回率较高，这也进一步表明这两种方法更偏向于正文内容的完整抽取，但不一定能很好地保证准确率. 综上所述，本文方法较CETE和CETR方法，其正文抽取的准确率有一定的提高.

4　结论

本文针对基于文本块正文抽取方法块内噪音难以剔除和基于标签路径的方法容易将正文块外较长文本误提取的问题，设计了文本块和标签路径相结合的正文抽取方法. 该方法利用文本块能准确判断正文区域边界和标签路径能有效区分正文节点和噪音节点的优点，在正文块内利用标签路径方法来剔除噪音文本. 经实验表明，该方法较CETR和CETD有着更高的准确率.

然而该方法依然存在缺点，如对于少数正文内容包含连续的短文本的网页，可能会将部分短文本误认为噪音剔除掉，还需进一步改善.

参考文献：

[1] 贺科达, 朱铮涛, 程昱. 基于改进TF-IDF算法的文本分类方法研究[J]. 广东工业大学学报, 2016, 33(5): 49-53.
HE K D, ZHU Z T, CHENG Y. A research on text classification method based on improved TF-IDF algorithm[J]. Journal of Guangdong University of Technology, 2016, 33(5): 49-53.

[2] LIU P F, QIU X P, HUANG X J, Adversarial multi-task learning for text classification[C] //Annual meeting of the association for computational linguistics. Vancouver: Transactions of The ACL Journal, 2017: 1-10.

[3] FAZZINGA B, FLESCA S, TAGARELLI A. Schema-based web wrapping[J]. Knowledge & Information Systems, 2011, 26(1): 127-173.

[4] KIM M, KIM Y, SONG W, *et al*. Main content extraction from web documents using text block context[C] //Database and expert systems applications. Heidelberg: Springer, 2013: 81-93.

[5] 李萍, 朱建波, 周立新, 等. 基于快速构建模板的购物信息抽取方法[J]. 计算机应用, 2014, (3): 733-737.
LI P, ZHU J B, ZHOU L X, *et al*. Shopping information extraction method based on rapid construction of template[J]. Journal of Computer Applications. 2014, 34(3): 733-737.

[6] 杨贤, 何汉武. 基于互联网文本挖掘的用户意图感知[J]. 广东工业大学学报, 2017, 34(3): 54-58.
YANG X, HE H W. Internet text mining for user intent perception[J]. Journal of Guangdong University of Technology, 2017, 34(3): 54-58.

[7] WENINGER T, HSU W H, HAN J. CETR: content extraction via tag ratios[C] //International conference on world wide web. Raleigh: ACM, 2010: 971-980.

[8] WENINGER T, HSU W H. Text extraction from the web via text-to-tag ratio[C] //International workshop on database and expert systems application. Turin: IEEE, 2008: 23-28.

[9] SUN F, SONG D, LIAO L. DOM based content extraction via text density[C] //International ACM SIGIR conference on research and development in information retrieval. Beijing: ACM, 2011: 245-254.

[10] WU G Q, LI L, HU X G, et al. Web news extraction via tag path feature fusion using DS theory[J]. Journal of Computer Science and Technology, 2016, 31(4): 661-672.

[11] 胡骏. 基于标签路径特征的网页正文自适应抽取方法研究[D]. 合肥: 合肥工业大学计算机与信息学院, 2016.

·简讯·

《广东工业大学学报》网站获评“第四届中国高校科技期刊优秀网站”

在媒体融合发展的背景下，为进一步推动高校科技期刊网站建设工作的有效开展，继续提升数字出版能力和水平，中国高校科技期刊研究会组织开展了2017年第四届中国高校科技期刊优秀网站评选活动.《广东工业大学学报》网站在评选活动中荣获“第四届中国高校科技期刊优秀网站”的荣誉称号.

近年来《广东工业大学学报》编辑部在期刊数字化建设方面取得了明显成效，不仅应用了在线投审稿系统，并且实现了XML排版与网刊发布一体化功能，扩展了网站信息发布形式，增设HTML格式发布、预出版功能等，提供了更完善的文献服务，加强了与用户的交流. 同时网站与多媒体平台融合，应用了期刊网站二维码识别，开通了期刊微信公众平台，整合了稿件在线处理系统和网刊发布系统，实现了移动查稿、移动审稿、微信平台优先发布等功能.

《广东工业大学学报》将继续与新技术、新平台接轨，融合使用网站和多媒体平台，提高数字出版水平和学术影响力，努力建设成高水平大学学报.

(广东工业大学学报编辑部)