一种 DOM 树标签路径和行块密度结合的 Web 信息抽取方法

马晓慧，李泓莹

（山西大学 商务学院，太原 030031）

摘 要：本文提出了一种标签路径和行块分布函数相结合的信息抽取方法来实现 Web 页面的信息抽取。该方法将 Web 页面解析成 DOM 树，使用视觉特征和标签过滤的规则将树进行剪枝，引入标签路径特征的方法粗略划分出网页的正文内容和噪音内容，最终使用行块分布函数的方法进行抽取，获得正文文本。实验结果表明，这种抽取方法有效地防止了正文内容误删及噪音内容漏删的现象，使得提取的正文信息更加准确，准确度达到 91%，召回率达到 95%，*F* 值达到 93%。本算法对于包含过多短文本的网页抽取的准确度还有待提高。

关键词：DOM 树；视觉特征；标签路径特征；行块分布函数

中图分类号：TP391 文献标志码：A 文章编号：2095-2163(2017)04-0013-05

Web information extraction based on label path of DOM tree and block density

MA Xiaohui，LI Hongying

（*Business College*，*Shanxi University*，*Taiyuan* 030031，*China*）

Abstract：In this paper，an information extraction method combining tag path and block distribution function is proposed to extract information from Web pages. The Web page is parsed into a DOM tree in first step. Secondly，the DOM tree is pruned by using visual features and label filtering rules. And then introducing label path characteristics，Web information is roughly divided into two parts：text content and noise content. Finally，using row block distribution function to extract text，the text is utterly obtained. The experimental results show that this method can prevent that the text is mistaken to delete and the noise content is missed to delete effectively，making the extraction of text information more accurately. The results shows that the precision reaches 91%，the recall rate 95%，*F* score 93%. The accuracy of the algorithm for Web pages which are containing too many short texts still has to be improved.

Keywords：DOM tree；visual features；label path characteristics；block distribution function

0 引 言

Web 页面是目前人们获取信息的主要方式之一，也是舆情监测、数据分析和处理的一个重要来源。常见的 Web 页面除了包含有用的正文信息外，还包含了大量的与正文主题无关的链接、图片、脚本等内容。因此，从纷杂的信息中快速准确地提取所需信息就显得尤为重要，Web 页面的信息抽取也成为了研究的一个热点问题。

刘秉权[1]等提出了基于 DOM 树的方法，根据 HTML 标签把网页解析为一颗树，在树上通过 DES 算法、MDR 算法等应用算法抽取网页中有效信息。聂卉[2]等深入研究了一种基于 GATE 语义标注的 Web 信息自动抽取技术，这种技术通过领域本体对网页进行语义标注准确定位目标项，再通过从构建好的 DOM 树中抽取语义项的特征描述构建样本实例，最后运用归纳算法实现抽取。陈鑫[3]则重点探讨了在经过初步过滤后的 HTML 中，结合正文区的密度，以行为自变量，行块长度为因变量建立线性行块分布函数，通过分布函数图找出阈值，从而得到有效的正文内容。朱泽德[4]等建立了一个融合结构和语言特征的统计模型，利用高斯平滑运算对密度序列进行计算以获取平滑文本密度，再由最大子序列分割平滑文本密度抽取正文内容。张乃洲等[5]用节点密度熵为度量分割 DOM 树，再采用 K 最近邻标签传播的半监督法和 SVM 分类器对页面进行分类，抽取有用类。微软亚洲研究院[6]最早开展了基于视觉特征的信息抽取技术研究，可将人对一个网页的视觉感受作为依据，区分出不同主题的主题块，对所需主题块进行提取。孙璐等人[7]还在此方法基础上做出了实用升级改进，利用 VIPS 算法将去除无关信息的 DOM 树来拓展执行语义分块，而后根据位置特征找到基准块，以此作为中心，遍历 DOM 树找到所有相似块并引入抽取处理，提高了抽取效率。此外，还有基于模板的技术。顾韵华等人[8]在领域本体的引导下建立了双模板——DIV 块模板和表格模板，可用其分别实现粗粒度和细粒度的信息抽取。郭少华等[9]基于模板提出正交过滤算法，过滤掉模板中的噪音信息，改善了生成的模板。随着研究的深入，后期出现了基于机器学习等多种抽取技术，在上述抽取方式中，以 DOM 树方法应用最为广泛。本文即在 DOM 树的基础上提出了一种根据标签特征、行块分布函数以及链接密度精确抽取正文的方法。研究设计内容可做如下论述。

基金项目：山西大学商务学院 2016 年科研基金(2016008)。

作者简介：马晓慧(1982-)，女，硕士，副教授，主要研究方向：计算机应用技术、信息检索、并行算法；李泓莹(1994-)，女，本科生，主要研究方向：Web 信息抽取、信息检索。

收稿日期：2017-07-20

1 抽取系统实现框架

本文致力于探讨的这种研究方法大致可分为3步。首先将经过规范化的网页解析成DOM树，由标签过滤和链接密度过滤的方式去除不必要的分枝，使一颗结构复杂的树简洁化。其次，遍历DOM树，对树中的所有标签路径、文本标签及标点个数提供数理运算统计，分别计算所有可到达文本长度之和与标签路径的比值和所有可到达标点路径之和与标签路径的比值，大致区分正文和噪音部分。最后，使用行块分布函数法对已划分出的正文和噪音完善推演、并设计进一步的过滤、抽取，最终能够高精度地从网页中抽取得到有效信息。

1.1 构建DOM树

DOM树具有结构性强，将无序网页有序化的特点，能够清晰地展示一个网页的结构。因此，为了使网页结构更加直观，方便正文抽取工作，首先可将网页转换成DOM树。设计过程可详述如下。

1.1.1 规范化HTML语法

在将网页解析成DOM树前，需使用W3的HTML Validator工具检验HTML代码是否合法，对不合法的代码进行修正，获取规范的HTML文档。本文所采用的部分语法规范准则如表1所示。

表1 语法规范准则表

Tab. 1 Grammatical norms

语法规则	规范写法	错误写法
标签匹配。每个开始标签都对应一个结束标签		或
属性放于对应的标签中，如：button按钮定义在form之间，tr、td定义在table之间	<form action="……"><input type="button" ></form>	<table><input type="button" ></table>
正确嵌套标签	… …</br>…	… … …</br>
标签的属性值放置在引号中	< img src = "/img/style.gif" >	< img src =/img/style.gif>
所有字符必须为英文字符	<form action="……"></from>	<form action="……"></from>
不能省略结束标签	<link rel=icon type="image/png" sizes="16×16" href = "/logo - 16.png" />	<link rel=icon type="image/png" sizes="16×16" href = "/logo - 16.png" >

1.1.2 解析DOM树

通过标签属性对，将获取的HTML文档解析为一颗以html为根节点的DOM树，现以图1所示网页为例，解析后生成的DOM树结构如图2所示。

图1 网页样例图

Fig. 1 An example of Web page

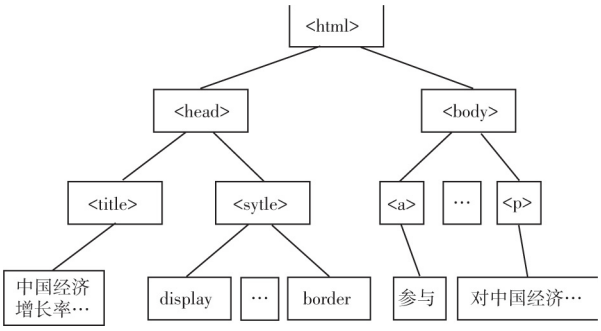

图2 DOM树结构图

Fig. 2 A structure chart of DOM tree

1.2 降噪处理

初步构建好的DOM树分支多，其中大量分支放置着无意义内容，如脚本信息、链接广告等。这样的树若是不拓展设置降噪环节，不但会将之后正文抽取的工作复杂化，还会在一定程度上降低抽取的效率和精确度。本文用视觉特征和标签过滤的方法对网页做降噪处理，对通常不含正文文本内容的标签做剪枝处理，得到一个简洁的DOM树。这里将给出研究分述如下。

1.2.1 视觉特征降噪

经比对多个网页发现，大量的网页布局基本类似，都由head、foot、right、left、center这5个区域中的任意几个构成，其中97%的网页均含有head、foot区域，right、left区域选择性拥有。以图1为例，对应的区域结构则如图3所示。

参阅文献[10]所示，依据其中提出的可视布局去除网页噪音的算法，去除网页中的head，foot区域。设计得到步骤如下：

1) 显示已解析的DOM树，由此获取网页实际大小。

2) 由网页的实际大小分别得出上、下边界的阈值，相应记为*top*、*lower*。

3) 将网页内除body标签外的所有元素取得的绝对坐标及其实际大小同由阈值划分的区域展开比较。以任一元素*Element*为例，其所属区域左上角的原点绝对坐标为(X,Y)，

且设定所占区域大小为（*Height*，*Width*）。若 *Element.Y* + *Element.Height* < = *top*，则元素 *Element* 属于 *head*。若 *Element.Y* > = *lower*，则元素 *Element* 属于 *foot*。据此规则对 DOM 树进行遍历，对区域进行划分，去除 head、foot 区域，实现初步噪音处理。未去除的元素均暂时作为正文。

图 3　网页区域结构图

Fig. 3　A structure chart of Web page

1.2.2　标签过滤

在初步获取的正文文本中，可能包含 right、left 区域，这些区域中的元素都作为正文包含其中，需要通过标签过滤的方式再次去除网页噪音。这里的标签过滤分为 2 种，一种是过滤不含正文的标签，另一种是通过链接密度过滤正文中难以识别的超链接。

网页中的标签大体可分为 2 类，一类是构建网页框架，显示正文文本的有用标签，一类是用来修饰网页使其美观的无用标签。本文的目的是抽取正文信息，因此需删去 DOM 树上的无用标签，如：script 和 hidden 的标签及其内容，文本样式修饰标签<style>，文本格式修饰标签<b>、<i>、、<u>、等。只保留可能包含正文的有用标签<p>、 、<table>、<li>、等。

在过滤无用标签时能够过滤掉一部分具有明显标签的链接，但如果在正文文本的一行中超链接长度所占比率较大，以上方法则无法准确识别，还需通过链接密度进行判断。这里的链接不仅包括广告链接和正文内容中的某些链接，还包括图片链接等多种广义上的链接。

使用链接密度除噪的方法需统计出树中的纯文本数量，记作 *NumText*，每个节点下的纯文本数量，记作 *note.NumText*，每个节点下的链接数量，记作 *link.NumText*。计算链接密度 *D*，*D* 为节点下链接数量与纯文本数量的比值。数学公式可表述如下：

$$D = \frac{link.NumText}{node.NumText} \tag{1}$$

由公式（1）得到，密度值越大，表明正文中此节点包含链接文本的可能性越大，找出该节点并删去链接，即获得了正文文本。综合上述对网页中噪音的处理，即可得到一个较为简洁的 DOM 树。下面将以此为基础，进一步剖析论述正文内容的抽取的研究过程。

1.3　抽取正文内容

抽取正文内容目的是将有效信息同噪音区分开，传统的抽取方法大都只从单一的角度出发，或选用文本密度作为区分，或采取链接密度作为依据。如今网页风格不一，若采取单一的方法只能针对具有明显特征的网页，文献 [4] 已经验证文本密度的方法就在传统正文内容形式单一且为连续密集结构的文本中抽取效果显著。但还易发生将噪音作为正文抽取出来（下文称之为误抽现象）或将正文当做噪音删除（下文称之为漏抽现象）的现象。因此，本文将标签特征与行块分布函数联系起来，综合考虑，进而研究提出一种能精确抽取多种类型网页的方法。在此，可得其研究分述如下。

1.3.1　基于标签特征进行抽取

经过观察大量网站可以看出，正文多以段落的形式表达，特点是文本内容长，标点符号多，而噪音则相反。为此可以根据这一特点，由式（2），式（3）计算可到达文本路径长度之和与所有标签路径的比值（*LTR*），与可到达标点路径长度之和与所有标签路径的比值（*PTR*），由这 2 步计算可以粗略地划分出噪音内容和正文内容，设 *n* 为树上的节点，*t* 为节点上的文本内容。具体步骤如下。

1）遍历 *DOM* 树，得到树中的所有标签路径（*tagpath*）。

2）统计出每个文本标签的文本长度（*length*）。

3）统计出树中标点个数（*number*）。

4）计算 *LTR* 和 *PTR* 值，值大的路径上的文本内容暂视为正文。研究推得计算公式如下：

$$LTR = \frac{\sum length(t(n))}{tagpath(n)} \tag{2}$$

$$PTR = \frac{\sum number(t(n))}{tagpath(n)} \tag{3}$$

通过以上步骤，虽可抽取出正文内容，但对有些种类的网页准确率不高。若正文中包含文本短且标点少的文本，比如诗歌则会出现误抽现象；若噪音中包含诸如评论信息较长类似于正文的文本，会出现漏抽现象。因此还需对已经区分开的正文和噪音文本做进一步检查和抽取。

1.3.2　基于行块分布函数进行抽取

文献 [3] 提出的基于行块分布函数算法可以有效解决上述问题，下面给出行块分布函数算法中的 3 个定义。对此，可详细阐释如下：

定义 1　将上一步划分出的噪音内容称为 *Noisetext*，以 *Noisetext* 中的第 *i* 行为轴，取其上文或下文 *k*（*k* < = 5）行，组成一个行块 *Block*，叫做行块 *i*。

定义 2　将一个行块 *Block* 中的所有换行、空格等占位符去除后，剩余文本长度称为行块长度。

定义 3　以 *Noisetext* 中的每行为轴，设 *Noisetext* 中共有 *n* 行，共有 $\frac{n}{k}$ 个 *Block*。做出以 *X* 轴为行号（*i*），以 *Y* 轴为行块 *i* 的行块长度（*length*（*i*））的行块分布函数。

将正文内容称为 *ContentText*，正文的行块分布函数定义同上。

通过行块分布函数绘制得到行块分布函数图，以图中骤升和骤降点作为边界，分别在噪音文本和正文文本中，确定起始

行块 X_{start} 与终止行块 X_{end}。其中，X 为行号，$Y(X_i)$ 为行块 i 的行块长度，需满足以下要求:

1) $Y(X_{start}) > Y(X_i)$，行块 i 为起始行块。

2) $Y(X_n) \neq 0(n \in [start+1, start+k])$ 紧随起始行块的下一个起始行块长度不为0。

3) $Y(X_m) = 0(m \in [end, end+1])$ 终止行块及其后的 K 个行块长度为0。

4) 当取 $Y(X_{\max})$ 时，$\exists X \in [X_{start}, X_{end}]$，保证此区域为取到行块最大值的区域。

将起始行块与终止行块中的内容作为正文提取，得到一个较为纯净的正文文本。

2 实验数据比对与分析

为验证本抽取方法的有效性，随机爬取了新闻、军事、体育和财经等4种类型的800个网页，使用单一的基于视觉特征网页信息抽取[11]方法、行块分布函数分别进行抽取，并同本抽取方法展开了研究对比。

2.1 实验数据集

本文的数据集来源于8个网站，分别是: 今日头条、环球网、网易军事、新浪军事、搜狐体育、体坛周报、东方财富、凤凰财经，从中随机抽取800个网页，因而得到抽取网页数据的结果信息则如表2所示。

2.2 评价标准

在从Web页面中抽取有效信息的实验中，采用准确率(*Precision*)、召回率(*Recall*)和 F 值作为实验结果的性能评估指标。准确率、召回率、F 值的计算公式可分见公式(4)~(6)。

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F = \frac{2PR}{P + R} \tag{6}$$

其中，TP 为抽取出的所有信息中的有效信息量，FP 为抽取出的所有信息中包含的无效信息量，FN 为未被抽取出的信息中的有效信息量。

表2 抽取网页数据集

Tab. 2 Web data sets

网站	网址	页面数量
今日头条	https://www.toutiao.com/	50
环球网	http://www.huanqiu.com/	120
网易军事	http://war.163.com/	100
新浪军事	http://mil.news.sina.com.cn/	80
搜狐体育	http://sports.sohu.com/	150
体坛周报	http://www.titan24.com/	100
东方财富	http://www.eastmoney.com/	110
凤凰财经	http://finance.ifeng.com/	90

2.3 比对结果及分析

在抽取的这8个网站中，今日头条网站中含有噪音内容较少，正文以段落长文本为主体，无图片及链接的干扰。环球网、体坛周报与新浪军事网结构类似，噪音内容集中在头部、尾部及右侧区域，正文中有图片插入，文字为长文本。网易军事网站中的网页正文区域以图片为主，文字为辅，文本内容较少，多为1~2句话。搜狐体育除以长文本构成的正文区域外，右侧区域也含有比赛时间这类短文本正文内容。东方财富与凤凰财经这2个财经类网站的正文中有大量短句格式，二者区别在于，东方财富中网页内容均由短句、数字构成，而凤凰财经中网页内容中除短句外也含有长文本，广告链接插在正文中。

将研究选用的800个网页分别用视觉特征网页信息[11]的方法、行块分布函数法与本文方法进行信息抽取，抽取结果如表3所示。由表3中的数据可以看出，本文方法的抽取效果较为理想，但对于包含过多短文本的网页抽取的准确度还有待提高。

表3 Web信息抽取结果比对表

Tab. 3 Comparison of Web information extraction results

网站名称	页面数量	文献[3]方法			文献[11]方法			本文方法		
		P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
今日头条	50	0.90	0.92	0.91	0.92	0.95	0.93	0.94	0.95	0.94
环球网	120	0.87	0.92	0.89	0.83	0.94	0.88	0.94	0.96	0.95
网易军事	100	0.75	0.95	0.84	0.75	0.96	0.84	0.84	0.92	0.88
新浪军事	80	0.84	0.96	0.90	0.89	0.93	0.91	0.92	0.95	0.93
搜狐体育	150	0.72	0.95	0.82	0.87	0.95	0.92	0.95	0.98	0.96
体坛周报	100	0.85	0.94	0.89	0.82	0.97	0.88	0.93	0.94	0.93
东方财富	110	0.74	0.96	0.84	0.85	0.94	0.89	0.82	0.95	0.88
凤凰财经	90	0.77	0.93	0.85	0.82	0.96	0.90	0.92	0.96	0.94

(下转第20页)

ciphers mCrypton and LED against biclique cryptanalysis [J]. Peer to-Peer Networking and Applications, 2013,8(4):716-732.
[4] Kang J, Jeong K, Sung J, et al. Collision attacks on AES-192, AES-256, Crypton-192/256, mCrypton-96/128 and Anubis [EB/OL]. [2013-11-10]. http://downloads.hindawi.com/journals/jam/aip/713673.pdf.
[5] MALA H, DAKHILALIAN M, SHAKIBA M. Cryptanalysis of mCrypton-A lightweight block cipher for security of RFID tags and sensors [J]. International Journal of Communication Systems, 2012, 25(4): 415-426.
[6] SHAKIBA M, DAKHILALIAN M, MALA H. Non-isomorphic biclique cryptanalysis and its application to full-round mCrypton [EB/OL]. [2013-03-08]. http://eprint.iacr.org/2013/141.pdf.
[7] PARK J H. Security analysis of mCrypton proper to low-cost ubiquitous computing devices and applications [J]. International Journal of Communication Systems, 2009, 22(8): 959-969.
[8] BIRYUKOV A, WAGNER D. Slide attack [C]//LNCS 1636: Proc of FSE 1999. Berlin: Springer, 1999: 245-259.
[9] Knudsen L R. DEAL-a 128-bit block cipher [R]. Bergen: University of Bergen, 1998.
[10] Biryukov A, Wagner D. Advanced slide Attack [C]//LNCS 1807: Proc of UROCRYPT 2000. Berlin: Springer, 2000: 589-606.
[11] WANG Q, LIU Z, VARICI K, et al. Cryptanalysis of reduced-round SIMON32 and SIMON48 [M]//MEIER W, MUKHOPADHYAY D. Progress in Cryptology-INDOCRYPT 2014. INDOCRYPT 2014. Lecture Notes in Computer Science. Cham: Springer, 2014, 8885: 143-160.
[12] TSUNOO Y, TSUJIHARA E, SHIGERI M, et al. Impossible differential cryptanalysis of CLEFIA [C]//LNCS 5086: Proc of FSE 2008. Berlin: Springer, 2008: 398-411.
[13] BOURA C, NAYA-PLASENCIA M, SUDER V. Scrutinizing and improving impossible differential attacks: Applications to CLEFIA, Camellia, LBlock and Simon [C]//LNCS 8873: Proc of ASIACRYPT 2014. Berlin: Springer, 2014: 179-199.
[14] BAHRAK B, AREF M R. Impossible differential attack on seven-round AES-128 [J]. IET Information Security, 2008, 2(2): 28-32.
[15] BIHAM E, BIRYUKOV A, SHAMIR A. Cryptanalysis of Skipjack reduced to 31 Rounds [C]//Advances in Cryptology—EUROCRYPT 1999. Berlin: Springer-Verlag, 1999: 12-23.
[16] LU J, DUNKELMAN O, KELLER N, et al. New impossible differential attacks on AES [C]//LNCS 5365: Proc of INDOCRYPT 2008. Berlin: Springer, 2008: 279-293.

(上接第16页)

3 结束语

本文提出了一种Dom树标签路径剪枝和行块密度结合的Web信息抽取方法。将Web页面解析成DOM文档,在此基础上根据路径标签比和行块分布函数对信息进行抽取,获取精确度颇高的抽取结果。经实验表明,本文研发方法的准确率达到91%,由此可知方法设计效果高效可行,但对于类似财经类这种包含过多短文本的网页,对其抽取的准确度还有待后续的改进提高。在今后的研究中,将对本文方法设计引入进一步优化处理,扩大抽取方法的适用范围,提高抽取系统的性能。

参考文献:

[1] 刘秉权,王喻红,葛冬梅,等. 基于结构树解析的网页正文抽取方法[C]//黑龙江省计算机学会2007年学术交流年会论文集. 大庆:黑龙江计算机学会,2007:14-17.
[2] 聂卉,黄贵鹏. 基于GATE语义标注的Web信息的自动抽取[J]. 图书情报工作,2010,54(5):110-114.
[3] 陈鑫. 基于行块分布函数的通用网页正文抽取[EB/OL]. [2016-02-23]. https://www.doc88.com/p-912707793066.html.
[4] 朱泽德,李淼,张健,等. 基于文本密度模型的Web正文抽取[J]. 模式识别与人工智能,2013,26(7):667-672.
[5] 张乃洲,曹薇,李石君. 一种基于节点密度分割和标签传播的Web页面挖掘方法[J]. 计算机学报,2015,38(2):349-364.
[6] Cai Deng, Yu Shipeng, Wen Jirong, et al. VIPS: A vision-based page segmentation [R]. Redmond, WA: Microsoft corporation, 2003.
[7] 孙璐,陈军华,廉德胜. 一种基于视觉特征的Deep Web信息抽取方法[J]. 计算机与数字工程,2016,44(6):1107-1111,1126.
[8] 顾韵华,高原,高宝,等. 基于模板和领域本体的Deep Web信息抽取研究[J]. 计算机工程与设计,2014,35(1):327-332.
[9] 郭少华,郭岩,李海燕,等. 可扩展的网页关键信息抽取研究[J]. 中文信息学报,2015,29(1):97-103.
[10] 荆涛,左万利. 基于可视布局信息的网页噪音去除算法[J]. 华南理工大学学报(自然科学版),2004,32(S1):84-87,98.
[11] 安增文,徐杰锋. 基于视觉特征的网页正文提取方法研究[J]. 微型机与应用,2010(3):38-41.