第38卷 第21期
Vol.38 No.21

计 算 机 工 程
Computer Engineering

2012年11月
November 2012

·人工智能及识别技术· 文章编号：1000—3428(2012)21—0153—04 文献标识码：A 中图分类号：TP391

基于正文相关度的维吾尔网页正文提取

王 瑞[1,2]，周 喜[1]，李 晓[1]

(1 中国科学院新疆理化技术研究所，乌鲁木齐 830011；2. 中国科学院研究生院，北京 100049)

摘 要：网页表达的主要信息通常隐藏在大量无关的结构与文字中，使正文信息不能被迅速获取，影响文本检测的效率。为此，根据维吾尔网页的非规范化编码、论坛型网页较多等特点，提出一种基于正文相关度的正文提取算法，并建立上下文正文密度和节点间正文比例等数学模型对算法进行改进。对大量维吾尔网页的实验结果表明，该算法具有较好的正文提取正确率和召回率，能够有效地从维吾尔网页中提取到所需的正文信息。

关键词：正文提取；正文相关度；信息安全；自然语言处理；正文密度

Content Extraction of Uighur Web Based on Content Correlativity

WANG Rui[1,2], ZHOU Xi[1], LI Xiao[1]

(1. Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China;
2. Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】 In addition to the main content, most Uighur Web contain noises such as navigation panels, advertisements which are not related to the main content. To improve the efficiency of security detection, this paper presents a content extraction algorithm of Uighur Web based on Web text correlativity, and designs the model of text density and content scale to improve the algorithm. Experimental result shows that this algorithm can extract the main content from the Uighur Web efficiently.

【Key words】 content extraction; content correlativity; information security; natural language processing; content density

DOI: 10.3969/j.issn.1000-3428.2012.21.041

1 概述

截止2011年新疆维吾尔自治区的统计，散布在互联网的维吾尔文网页已经超过2 000 000张。与此同时，各种不良信息也同样开始大量充斥在其中，所以，对维吾尔网页进行内容检测变得十分必要。但是Web页面通常含有大量与网页主题无关的信息，如广告链接和图像等，它们分布于网页四周，或附着在正文旁边，使正文信息不能被迅速定位，降低了网络内容检测的效率。所以，设计并开发维吾尔网页正文提取系统变得十分必要。

本文针对维吾尔网页编码不规范现象较多的特点，提出一种基于正文相关度的正文提取算法，该算法对维吾尔网页进行预处理，使其编码规范化，利用DOM树结构的方法解析出网页的基本结构，计算每个节点的正文相关度，根据一个阈值对节点进行过滤。并且针对维吾尔语论坛型网页较多的特点进行数据平滑，提出节点间正文比例的概念，利用它和上下文正文密度对每个节点是否为正文节点再次判断，可提高正文提取的效果。

2 相关研究

网页“正文提取”这个词是由Rahman A R、Alam H、Hartono R[1]等人在2001年提出的。其后，大量不同的算法被研究并开发出来。这些算法针对不同的网页发挥其作用，目前比较主流的网页正文提取方法主要分为3种：早期的一种网页正文提取方法是基于对一些网页独特的特征的了解，以及一些通用规则，比如正文中含有标点符号、噪音区含有较多链接等。并针对这些特征，对不同网页编写特定的算法，有些还建立特定的包装器(wrapper)，用于从特定的信息源中抽取需要内容[2]。近期，一些学者改进了根据网页特征总结的规则，将不同的判定特征给出不同的权值，并进行优化[3]。基于网页特征规则的方法的优点是准确率高，但同时缺点也显而易见，针对不同网页的特点，要编写不同的规则；同时，相同的网站在不同时间依然可能改变其自身的结构特点，所以这种算法可移植性较低，不能广泛地应用到互联网的海量网页中。

基于视觉的网页正文抽取方法利用网页的布局结构：首先分析网页整体结构，然后寻找块间的分割线，并根据分割线确定每个分块，每一个分块都会有一个表示其与正文相关程度的阈值，最后根据计算分块阈值来确定哪些分块是正文[4-5]。这种自顶向下的结构效果很高，但是这种

基金项目：新疆维吾尔自治区高技术研究发展基金资助项目(201012112)；新疆维吾尔自治区电子发展专项基金资助项目(XJDZZXZJ20109)
作者简介：王 瑞(1985－)，男，硕士研究生，主研方向：自然语言理解；周 喜，副研究员；李 晓，研究员、博士生导师
收稿日期：2012-02-07 **修回日期**：2012-03-13 **E-mail**：xiaoli@ms.xjb.ac.cn

算法的前提是能够获取网站的样式表，比如层叠样式表(CSS)，在一定程度上增大了网页正文提取的条件和难度。所以，对比其他算法，这种算法是具有取材局限性的。

基于密度的网页正文抽取方法。该类方法一般利用网页分析器，如(htmlparser)解析出网页的 DOM 树结构，然后根据不同的算法判断各树节点是否为网页正文。文献[6]利用正文密度对 DOM 节点进行分类，并提出了“DensitySum”方法对数据进行平滑。文献[7]提出了一种基于 STU-DOM 树的模型，并对其进行基于结构的过滤和基于语义的剪枝。但是这 2 种算法所提出的密度或者相关度算法在处理维吾尔网页，尤其是论坛型网页时局限性明显，因为维吾尔语论坛型网页中含有大量的较少字符的回复，而这种回复的正文密度与噪声节点的差异并不大，如果采取这样的方法，就会造成剪掉了较多的有用的正文信息，而同时有保留了较多的噪声节点。

目前主流的网页正文提取方法要么自身存在缺陷，要么其算法在处理维吾尔网页时存在局限性，因此，本文提出了针对维吾尔网页的正文相关度算法，建立了上下文正文密度和节点间正文比例模型来对这一算法进行改进。

3 正文提取算法

本节介绍维吾尔网页预处理的方法，将处理好的网页按照 DOM 树的结构进行解析和遍历，提出正文相关度的概念，并利用其对 DOM 树中的每个节点进行正文相关度计算。同时，针对维吾尔网页的特点，提出基于节点间正文比例的数据模型，并且利用这一模型和上下文正文密度对之前的计算结果做数据平滑。

3.1 维吾尔网页预处理

维吾尔语与维吾尔文是中国新疆维吾尔自治区自治民族维吾尔族使用的语言和文字，是新疆维吾尔自治区的官方文字，使用人口 860 余万。现代维吾尔文是在晚期察合台文基础上形成的以阿拉伯文字母为基础的拼音文字。现行维吾尔文有 8 个元音字母和 24 个辅音字母，共 32 个字母，自右向左横写。图 1 为在维吾尔地区较流行的论坛 Bilik Uyghur(Uighur) Universal BBS Website 的一个网页的缩略图。

图 1 网页缩略图

目前，维吾尔网页的编码方式主要存在以下几种：Windows-1252/1256，utf-8，ISO-8859，unicode 等。由于编码方式的不同，导致网页源代码显示方式的不同。部分网页源代码中并非是维吾尔字符本身，而是其对应的编码方式的代码，如果直接用这种编码格式的网页进行正文相关度计算，就会导致提取的网页正文与实际偏差很大。编码转化前后的网页源代码缩略图如下，其中，上面部分为转换前，下面部分为转换后。

<meta http-equiv="Content-Language" content="ar-sa">

<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">

<title>شىۋېتسىيە … ش قىلدى</title>

↓

<meta http-equiv="Content-Language" content="ar-sa">

<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">

<title>شىۋېتسىيەدىكى ئۇيغۇرلار 1- نۆۋەتتە</title>

所以，对网页源代码进行正文相关度计算前，要将其转化为与模式统一的 utf-8 编码格式。针对维吾尔网页的特点，本文设计并实现了一种基于正则表达式的编码转化方法，在对 HTML 网页源代码进行扫描时，如果连续匹配到“&”和“#”，即出现“&#”时，而且后面是数字字符，即判断其为一个转义字符串。将后面的数字串，查找维吾尔字符代码转换表(如表 1 所示，以 Windows1252/1256 与 utf-8 编码转换为例)，如果能够找到匹配项，则将其转化为对应的 utf-8 格式编码并替换为对应的维吾尔字符。由于维吾尔语的特殊性，一个同样的维吾尔字母出现在单词不同位置(词首、词尾、词中等)其编码和字形也可能不同，本文算法将不同的编码和字形统一为基本型，通过这种算法，将维吾尔网页统一转化为统一的 utf-8 的编码格式。

表 1 维吾尔字符代码转换表略图(Windows1252/6-utf-8)

维吾尔字符(基本型)	utf-8 编码(基本型)	Windows 1252/56 编码
ر	0x0631	&#1585,&#65198/&#65197
ز	0x0632	&#1586,&#65200/&#65199
س	0x0633	&#1587,&#65201/&#65202/&#65203/&#65204
…	…	…

3.2 DOM 树解析及遍历

W3C DOM(Document Object Model)[8]被设计成平台无关、可使用任意编程语言实现的规范。每个 HTML 网页都可以被解析成一个 DOM 树，其中顶层节点为树的根节点，而子节点包括文本节点和链接节点，网页中每个结构标签被解析成一个 DOM 树节点，现在大部分的维吾尔网页是采取 作为网页构架，但是本文的算法也要兼容传统的以<TABLE>等标签作为构架的网页。与此同时，在进行 DOM 树解析时，<script>、<remark>和<style>等不可能含有正文的节点应该被过滤掉。利用 HTMLPARSER

得到解析好的 DOM 树，并按照节点在源代码出现的先后顺序编号(0,1,…,*n*)后，按照后根遍历的顺序遍历计算出每个节点；即从子节点开始计算其含有的链接节点个数和非链接文本字符数，如果一个节点含有子节点，则要计算该节点及其子节点所含的链接节点个数与非链接文本字符数。下面以计算每个节点的非链接文本字符数(TA)为例，给出遍历算法的伪代码：

```
Com(Node CurNode)
1:  If (CurNode 没有子节点) do
2:      TA  ← 节点 CurNode 本身的非链接文本字符数
3:      登记(CurNode, TA)
4:      Return TA
5:  Else if  (CurNode 有子节点)  do
6:      temp  ← 节点 CurNode 本身的非链接文本字符数
7:      Childs  ← 得到节点 CurNode 的所有子节点
8:      While( Node j  in  Childs)  do
9:          temp ← temp + Com(Node j)
10:     End While
11:     TA  ←  temp
12:     登记(CurNode, TA)
13:     Return temp
14: End if
```

因为每个节点需要被登记一次 TA，所以其算法时间复杂度为 $O(n)$。如图 2 所示，为对 Bilik BBS 中某个网页的 DOM 树遍历后的缩略结构，其中，TA 和 LA 分别表示非链接文本字符数与链接节点个数。

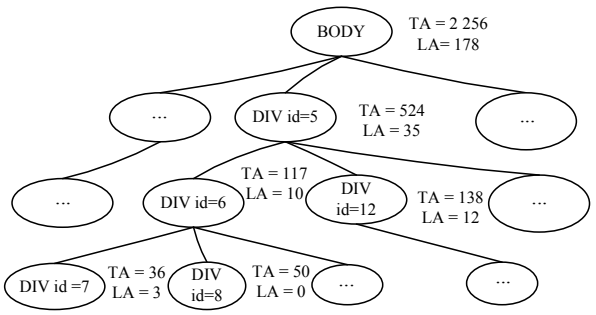

图 2　Bilik BBS 中某个网页 DOM 树遍历后的缩略结构

3.3　正文相关度

通常来说，在一个含正文的节点内，其所含的链接节点个数较少，非链接文本较多；反之，在一个噪声节点中，其所含的链接节点个数较多，非链接文本较少[9]。根据针对大量人工标注的维吾尔网页的源代码的分析，这个原则也是成立的。可以计算出每个节点的相关的非链接文本字符数与链接节点个数的比例，定义正文相关度如下：

定义 1　对于网页的一个节点 i，它的正文相关度表示为该节点及其子节点所含的非链接文本字符数与该节点及其子节点所含的链接节点个数的比例：

$$CC_i = \frac{TA_i}{LA_i} \tag{1}$$

其中，LA_i 表示链接节点个数该节点及其子节点所含的链接节点个数；TA_i 表示该节点及其子节点所含的非链接文本字符数，如果 LA_i 为 0，则将其置 1。

利用这个算法遍历整个 DOM 树，得到每个节点的正文相关度。对图 1 中的网页整体所解析后的 DOM 树进行正文相关度(CC 值)的计算，所得的结果如图 3 所示，其中，节点编号为其在 HTML 源代码中出现的顺序。

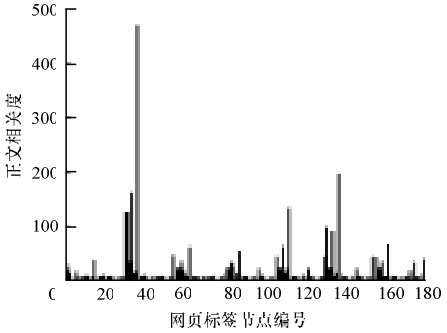

图 3　Bilik BBS Website 一个网页每个节点的正文相关度

3.4　阈值计算

在不考虑其他因素时，要设计一个阈值 C，对于前面计算出每个节点 i 的正文相关度值 CC_i，与阈值进行比较，如果 $C \geqslant CC_i$，则认定该节点为噪声节点；反之，如果 $C < CC_i$，则认定该节点为正文节点。一般来说，一个 HTML 网页的平均正文相关度，即其<body>节点的文本正文相关度，因为<body>作为根节点内包含了 DOM 树内的所有其他节点。但是根据 HTML 网页中平均无关链接比大于 90%[7]这一特性，如果采取文本平均绝对密度作为阈值，就会存在很多无关链接节点没有过滤掉的情况。当阈值 C 较大时，可以删除更多的无关链接，减少冗余；当 C 值较小时，可以保留更多与主题有关的链接。经过对维吾尔网页大量的实验，最终确定当阈值 C 为 29 时，能够过滤掉较多无关链接，同时保留适当的主题链接，效果较为理想。

3.5　数据平滑

但是如果仅按照一个单纯的阈值来判断一个节点是否是正文节点是远远不够的，按照图 3 所示的情况能够发现，有一部分节点，它们的正文相关度值是明显不同于周围节点的。这样的话，就可能出现以下 2 种情况：(1)在有些维吾尔语论坛网页中，部分用户的回复很短，其中还可能包括链接或图片，这种节点一般处于网页的正文附近，其正文相关度值明显小于周围节点，如果仅根据阈值判断，就会将其错判为噪声节点。(2)有些节点，例如网页的 copyright 节点一般不含有链接节点，而且含有较多的文字，它的正文相关度值一般来说明显大于周围的链接节点，如果仅根据阈值判断，则会将其错判为文本节点。所以，必须设计一种方法来平滑正文相关度值，避免遗漏和错判。

定义 2　对于网页的一个节点 i，它的上下文正文密度(text density)表示为该节点的父节点及其子节点所含的非链接文本字符数与该节点及其子节点所含的链接节点数的比例：

$$TD_i = \frac{TA_{(p_i)}}{LA_i} \tag{2}$$

其中，LA_i 表示该节点及其子节点所含的链接节点数；$TA_{(Pi)}$表示该节点的父节点及其子节点所含的非链接文本字符数，如果 $TA_{(Pi)}$为 0 的话，则将其置 1。

定义 3 对于网页的一个节点 i，它的正文比例(content scale)是指，它在 DOM 树中的同层节点中正文相关度大于(或等于)阈值 C 的节点数占其总兄弟数的比例，而且这些同层节点的标签需要与节点 i 的标签相同：

$$CS_i = \frac{TNbrother_i}{Nbrother_i} \times 100\% \tag{3}$$

其中，$TNbrother_i$ 表示它的兄弟节点中正文相关度大于(或等于)阈值 C 的节点数；$Nbrother_i$ 表示它的兄弟(包括其本身)节点总数。

对于 DOM 树中的每个节点，利用式(2)、式(3)进行数据平滑：对于正文相关度值小于阈值 C 的节点，如其正文比例值 CS 大于设定的阈值 S，且上下文正文密度(TD)同时大于设定的阈值 D，则将其视为正文节点，该节点正文相关度值更改为阈值 C。针对维吾尔语的特点，采取对大量维吾尔网页进行统计学习的方法，最终确定阈值 S 为 57%，当 D 为 63 时，数据平滑效果最好。同时，上述公式可以判断出正文节点在 DOM 树中所处的层次结构，对于明显偏离这一结构但是正文相关度值大于设定的阈值的节点，如 copyright 节点，将其正文相关度值更改为 0。

数据平滑后的 DOM 树的每个节点的正文相关度值如图 4 所示。对比图 3，可以看出一些原本被过滤掉的正文节点被重新判定为正文节点，而一些原本很突出的单个噪声节点被或滤掉了，其中，节点编号为其在 HTML 源代码中出现的顺序。

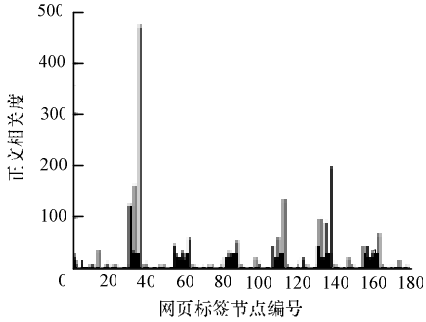

图 4 平滑后的该网页每个节点的正文相关度值

得到数据平滑后的正文相关度大于阈值的节点后，需要做去重处理，即如果得到的 2 个节点是父子节点关系的话，需要将父节点中与子节点重复的信息去除。最后将这些节点的文本内容提取出来，得到的内容就是最终的正文信息。

4 实验与结果分析

本文首先提出一种评价维吾尔网页正文提取的标准，并选取不同类型的维吾尔网页对本算法进行测试，然后给出实验结果，并对其进行分析。

4.1 评价标准

本文采取利用评价网页正文提取是否能提取出全部的正文信息的召回率 R、评价提取出的正文信息的准确率 P[10]。同时利用 F 值，对召回率 R 和准确率 P 做出综合的评价：

$$P = \frac{COM(c,s).length()}{c.length()}$$

$$R = \frac{COM(c,s).length()}{s.length()}$$

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

其中，$COM(c,s)$表示本文提取出的字符串 c 和实际上的正文字符串 s 中相符的字符数；$a.length$ ()表示相应的字符串 a 的长度。

4.2 数据选取

针对维吾尔语网页的不同特点，本文选取 6 个网页源，其中，3 个为政府新闻类网站的维文版(表 2 的前 3 个网站)，3 个为维吾尔常见的论坛网站(表 2 的后 3 个网站)。每个网页源，选取平均大概 200 个左右的网页作为样本集。将这近 1 200 个网页作为输入，对本文算法进行效果测试，并且对比相关研究中的传统的基于正文密度的算法。从表 2 和表 3 的对比可以看出，传统的算法在处理新闻型网页(表 2 中前 3 个网站)时效果较好，而处理论坛型网页(表 3 中后 3 个网站)时相比而言有较大的差距，正确率和召回率都有很大的下降。而本文算法由于采用了针对维吾尔网页的算法改进，对论坛型网页(后 3 个网站)的召回率和正确率有明显提升，而对传统的新闻型网页的正确率也有提升。

表 2 传统的基于正文密度的算法实验结果

网站名	正确率/(%)	召回率/(%)	F 值
http://uyghur.people.com.cn/	92.5	96.3	0.944
http://www.xjtsnews.com/normal/content/bak/index.htm	90.3	92.3	0.913
http://uyghur.xjkunlun.gov.cn/	89.8	92.4	0.911
http://www.qutyar.com/ud/portal.php	83.2	84.3	0.837
http://www.halastan.com/	85.5	87.2	0.863
http://bbs.bilik.biz/	86.7	92.2	0.894

表 3 本文算法的实验结果

网站名	正确率/(%)	召回率/(%)	F 值
http://uyghur.people.com.cn/	94.6	96.3	0.955
http://www.xjtsnews.com/normal/content/bak/index.htm	92.3	93.3	0.928
http://uyghur.xjkunlun.gov.cn/	91.3	94.2	0.927
http://www.qutyar.com/ud/portal.php	87.0	92.5	0.896
http://www.halastan.com/	88.6	93.8	0.911
http://bbs.bilik.biz/	90.3	94.3	0.922

5 结束语

传统的维吾尔网页的正文提取方法有明显的局限性，基于对维吾尔语网页编码格式不规范、论坛型网页较多等特性，本文提出了改进的基于正文相关度的正文提取算法。经过实验和项目验证，该方法在处理维吾尔语网页正文提取中取得了良好的应用效果。

(下转第 160 页)