2018 年 1 月 第 1 期
Vol. 39 No. 1 2018

基于网页聚类的正文信息提取方法

王一洲,陈 星,戴远飞

[1](福州大学 数学与计算机科学学院,福州 350108)
[2](福建省网格计算与智能信息处理重点实验室,福州 350108)
E-mail: chenxing@fzu.edu.cn

摘 要: 精准地抽取 Web 页面中正文内容,在许多 Web 挖掘研究领域有着重要的应用价值. 目前针对该问题主要采用网页分割和密度统计的方法. 但现有的方法在网页中正文内容字符数较少时可能失去作用. 经实例分析发现,网站内部的网页大多都是由一套相同内容模板生成的. 因此本文提出一种基于网页聚类的正文信息提取的方法,该方法主要有 2 个部分组成: 第一,基于网页的结构特征对网页进行聚类; 第二,面向相似网页集合的正文位置特征生成. 采用该方法可以从多种类型的网页中抽取正文信息. 我们针对 5 个网站进行了实验,实验结果表明该方法的可行性和有效性.

关 键 词: 网页聚类; 正文内容块; 节点密度
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2018)01-0111-05

Web Information Extraction Based on Webpage Clustering

WANG Yi-zhou, CHEN Xing, DAI Yuan-fei
[1](College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)
[2](Fujian Provincial Key Laboratory of Network Computing and Inteligent Information Processing, Fuzhou 350108, China)

Abstract: Accurately extracting important content from webpage has important applications for many research fields in Web mining. At present, the method of webpage segmentation and density statistics is used to solve this problem. However, the existing method may lose its function when the number of characters in the webpage is small. In this paper, we propose a method for extracting web information, based on the webpage clustering. This method consists of two components: webpage clustering based on structure feature and text block features generation with similar webpages. The method can extract web information from different types of webpages. We conduct the experiment with webpages from 5 sites, and the experimental results show that the proposed methods are feasibility and effective.

Key words: webpage clustering; text block; node density

1 引 言

Web 技术的迅猛发展,使得 Web 网页成为信息发布的主要载体,因此,Web 网页正文内容的抽取成为了当前学术界的一个研究热点. 然而,在网页内容抽取中存在两个难点: 第一,在一个 Web 页面中除了包含用户感兴趣的正文外,还包含导航条、广告、推荐链接、版权声明等与主题无关的噪音信息. 第二,由于动态脚本和 CSS 技术的广泛应用,使得网页之间的结构差异性不断增大并且网页自身结构的复杂性不断提高. 针对这两个难点,人们提出基于统计的网页正文提取和基于网页分割的网页正文提取,然而,当正文内容字符数较少且噪音信息过长的情况下,网页正文内容的提取可能出错. 例如: 在网页中游客评论的信息过长,就会导致系统将评论误认为正文内容.

实际上,HTML 页面是存储在后台数据库中的数据和 HTML 内容模板的结合体,在网站内部的网页大多都是由一套相同的内容模板生成的,因此可以认为网页的设计是有一定规律的. 通过这个规律,本文提出一种新的基于网页聚类的正文信息提取方法,用于抽取 Web 页面中的正文内容.

本文的主要贡献有两点:

1) 提出一种基于网页的结构特征对网页聚类的方法.

2) 提出一种面向相似网页集合的正文位置特征生成方法.

本文第 2 节简要介绍相关的研究工作. 第 3 节介绍基于网页聚类的正文信息提取方法的框架. 第 4 节提出基于结构特征的网页聚类. 第 5 节提出一种面向相似网页集合的正文位置特征生成方法. 第 6 节是实验和分析部分. 最后,第 7 节对本文工作进行总结和展望

2 相关工作

Web 网页正文信息提取技术随着需求的增加而不断丰富,近年来国内外涌现了多种方法. 根据其原理可分为 3 类: 基于统计的网页正文提取、基于网页分割的网页正文提取和基于机器学习的网页正文提取.

基于统计的网页正文提取方法主要根据网页中文本的分

收稿日期: 2016-11-23 收修改稿日期: 2017-04-07 基金项目: 国家自然基金项目(61402111)资助; 福建省科技平台建设项目(2014H2005)资助. 作者简介: 王一洲,男,1992 年生,硕士研究生,研究方向为 Web 信息抽取、软件工程; 陈 星,男,1985 年生,博士,副教授,CCF 会员,研究方向为系统软件、云计算; 戴远飞,男,1992 年生,硕士研究生,研究方向为智能信息处理、软件工程.

布情况来决定提取的内容. 文献[1]中作者先将 Web 页面转化为 DOM 树,然后提取 DOM 树中每一个节点的标签路径,基于路径标签,提出了 6 个特征,如文本标签路径长度特征,文本标签路径比特征等. 通过将不同特征融合来得到综合特征值,再利用综合特征值来提取网页中的新闻. 文献[2]则是一种完全脱离 HTML 标签的正文内容提取方式,通过统计去除 HTML 标签后网页中各行块的字符数,建立行块分布图,再由行块分布图直接定位网页正文的位置. 基于统计的网页正文提取方法其缺陷在于当网页中正文内容字符数较少时,其统计出来的数据就会导致选取错误的文本.

基于网页分割的正文提取是根据网页中的一些特征对网页进行切割,再从切割后的各个块中选取包含网页正文的块并提取网页正文. 文献[3,4]都是通过网页分割来提取正文内容. 文献[3]则提出 3 个基本的节点密度特征,再利用节点密度特征计算节点密度熵,以节点密度熵为度量将 DOM 树自动分割为若干个块. 最后,利用视觉特征[4]如: 块的大小,位置等信息对分割的块进行分类,最后将页面中的块分为多种类型,方便进行内容提取. 文献[5],则是将 Web 网页根据相应的规则分解成基本元素节点,将这些节点转化为一张图,通过基本元素节点之间的文本相似度比较和在网页中的相对位置给图中每两个节点之间加权值,再利用类似谷歌 PageRank 的排序算法来处理加权图选取包含正文内容的基本元素节点. 基于网页分割的正文提取其缺陷在于当网页正文内容字符数较少,而噪音信息较多时,选取的块可能出错.

文献[6]中提出了基于机器学习的网页正文提取方法,文中将网页内容的提取转化为对 DOM 树中节点的选择,提取 DOM 树中节点的多个特征,利用机器学习将这些特征作为输入训练出相应的模型,再通过模型选出包含有正文的候选节点. 其缺陷在于需要大量的训练样本才能保证准确率.

本文吸收了基于统计的网页正文提取方法和基于网页分割的正文提取方法,在网页聚类的基础上解决了因为正文内容字符数较少,噪音信息较长的特殊情况导致网页正文内容提取出错的问题.

3 方法概览

如图 1 所示,基于网页聚类的正文信息提取由 3 个模块组成,分别是网页解析模块、基于结构特征的网页聚类模块和面向相似网页集合的正文特征生成模块.

基于结构特征的网页聚类模块: 给定网页对应的 DOM 树,遍历 DOM 树提取网页的结构特征. 利用结构特征计算网页之间的相似度. 根据网页之间的相似度对网页集合进行分层聚类. 最后生成一系列的标记类集合,以及每个标记类的网页特征.

面向相似网页集合的正文特征生成模块: 针对同一类网页,对网页进行切割分块,统计各个块的节点密度特征,寻找包含正文内容并且不包含噪声信息的块并提取该块的特征作为该类网页中正文内容块的抽取规则.

图 1 中,用户需要输入一个网站中多个页面,通过网页聚类模块得到聚类结果中每个标记类网页的结构特征. 针对每个标记类提取其网页正文内容块的特征. 当用户输入该网站的网页时,会根据网页的结构特征确定该网页所属的标记类,并利用该标记类网页的正文内容块特征来提取网页中的正文内容.

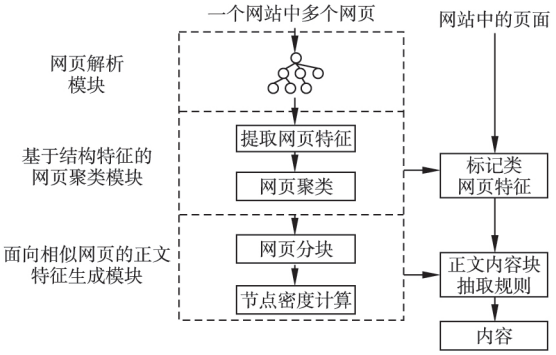

图 1 基于网页聚类的正文信息提取方法

Fig. 1 Webpage clustering based on structure feature

4 基于结构特征的网页聚类

在相关工作中我们分析了基于统计的网页正文提取的缺陷,当网页正文内容的字符数较少时,基于统计的网页正文提取就会失效. 当然,在一个网站中绝大多数网页中的正文内容的字符数是足够用来判断正文内容的位置,而网页正文内容字符数较少导致判断出错的页面是特殊情况. 因此本文利用网页聚类将正文内容所处位置相同(即网页结构相同)的网页放入同一个标记类中,通过对同一标记类中的网页进行统一的正文内容提取操作来消除因为个别网页中正文内容字符数较少引起的错误提取.

4.1 网页结构特征的表示

为了方便表示网页的结构特征,本文引入如下 4 个定义:

定义 1. 每一个 Web 页面均可以表示成一个 DOM 树 T_d, T_d 是一个有向图 $<V,E>$,其中 V 为顶点的集合,$V=\{v \mid v \in$ html 标签集 Tag$\}$. E 为有向边的集合,$E=\{<u,v> \mid u,v \in V$,其中 u 称为 v 的父顶点,而 v 称为 u 的子顶点,且在 html 结构上,v 对应的标签被 u 对应的标签收包含$\}$.

定义 2. 一颗 DOM 树 T_d 可表示为一个页面块的集合 $B=\{b_i \mid b_i \in V$,且 b_i 节点对应的 html 标签为 $<$ div $>$ 或 $<$ table $>\}$,称该节点为块节点.

定义 3. T_d 是一颗以 v_0 为根的 DOM 树,对于任意的节点 $v \in V$,$v_{k0}v_{k1}\cdots v_{kn}$ 是树 T_d 从 v_k 到达 v_{kn} 的节点序列,其中,parent$(v_{kj-1})=v_{kj}(1<=j<=n)$, $v_{kj}=v$,则称 $v_{k0}v_{k1}\cdots v_{kn}$ 为节点 v 的路径,记为 $p(v)$. 如 "body1/div3/div2" 为一个块节点的路径.

定义 4. 给定网页 w 和对应 DOM 树中所有块节点的路径集合 $f=\{p_1,p_2,\cdots,p_n \mid p_i=p(b_i), b_i \in B\}$,可表示为一个网页特征 $F=<w,f>$.

为了能够快速计算网页之间的相似度,本文将 DOM 树中各个块节点的路径作为网页的结构特征[7]. 通过前序遍历的方式遍历网页 DOM 树 T_d,提取块节点的路径集合 $F.f$,构成二元组 $F=<w,f>$ 来表示网页 w 的结构特征. 最后将输入的网页集合 W 转化为网页特征集合 $D=\{F_1,F_2,\cdots,F_k\}$.

4.2 网页聚类

通过上节网页结构特征的表示,我们可以从网页集合中提取出网页的结构特征集合 $D=\{F_1,F_2,\cdots,F_k\}$. 根据网页结构特征 F,可以计算得到网页之间的相似度.

定义 5. 给定有限集合 $A=\{x_1,x_2,\cdots,x_n\}$,将有限集合 A 的元素个数记为 $card(A)=n$.

本文定义了计算两个页面的相似度函数为:

$$sim(F_i,F_j)=\frac{card(F_i.f\cap F_j.f)}{\min(card(F_i.f),card(F_j.f))} \tag{1}$$

其中 F_i 和 F_j 分别表示第 i 个网页和第 j 个网页的结构特征. 为了方便算法表示本文给出了聚类结果后标记类特征的定义.

定义 6. 给定结构相似的网页集合 $W=\{w_1,w_2,\cdots,w_n \mid sim(<w_i,f_i>,<w_j,f_j>)>0.82,0<i,j<=n,n>5\}$ 和网页结构特征 $F_i=\{<w_i,f_i> \mid w_i\in W\}$,则可表示网页聚类结果中每一个标记类特征为一个二元组 $C(c)=<F_i,W>$,其中 c 表示该类的标记,$c\in$ 正整数 N.

本文使用分层聚类算法通过网页相似度计算对一组网页进行聚类,如算法 1.

算法 1. *getClasses*

输入: 网页结构特征集合 $D=\{F_1,F_2,\cdots,F_k\}$,表示 k 个网页的结构特征集合

输出: 标记类集合 $M=\{C_1,C_2,\cdots,C_n\}$,表示聚类结果为 n 个标记类

```
Begin
  For each  F_i ∈ D  then
    C_k ← ∅
    add(C_k, F_i)
    For each F_j ∈ D  then
      if sim(F_i, F_j) > 0.82 then
          add(C_k.W, F_j.w)
          remove(D, F_j)
    if card(C_k.W) > 5 then
      add(M, C_k)
End
```

通过算法 1,可以得到网页标记类网页集合 M,集合中的每一个元素表示该标记类的特征. 为了保证正文特征生成的准确性,在网页聚类的过程中,我们只筛选出标记类中网页数量大于 5 的标记类.

5 面向相似网页的正文特征生成

在获得网页聚类结果后,还需要对同一标记类的网页提取正文内容的抽取规则. 本文采用基于统计的节点密度特征和网页分块相结合的方式来确定网页内容的位置. 在网页中,正文内容的分布一般相对较集中,因此,正文内容所在的节点的文本密度比其他节点的文本密度要高. 从 HTML 文件在浏览器中展现的效果来看,页面是由若干个块构成的,这些块是由 HTML 容器标签(和<table>标签)分割而成的. 所以本文将 HTML 页面切割成块集合 B,再从块集合 B 中选择不包含噪声信息,但包含完整正文内容的正文内容块.

5.1 块节点密度特征

本文采用密度特征来判断块节点是否为正文内容块,下面给出 3 个密度定义:

定义 7. 设 $n\in B$ 为 DOM 树 T_d 中的一个块节点,则 n 的文本密度定义为:

$$p_{text}=\frac{T_n}{1+T} \tag{2}$$

其中 T_n 为块节点 n 包含的纯文本字符数(不含链接文本),T 为 T_d 代表的整个文档中的纯文本字符数(不包含链接文本).

P_{text} 反映了在全局页面中,文本内容在某个块节点中的相对集中程度. 我们发现 P_{text} 越大,往往意味着该节点越有可能包含待发现的正文内容块.

定义 8. 设 $n\in B$ 为 DOM 树 T_d 中的一个块节点,则 n 的链接密度定义为:

$$p_{link}=\frac{lN_n}{1+lN} \tag{3}$$

其中 lN_n 为节点 n 中所包含的链接数,lN 为 T_d 代表的整个文档中所包含的链接数.

P_{link} 反映了在全局页面中,链接在某个块节点的相对集中程度. 我们发现 P_{link} 越大,往往意味着该块节点包含噪声信息的可能性越大.

定义 9. 设 $n\in B$ 为 DOM 树 T_d 中的一个块节点,则 n 的节点文本密度定义为:

$$p_{textl}=\frac{T_n}{1+lT_n} \tag{4}$$

其中 T_n 为块节点 n 的纯文本字符数(不含链接文本),lT_n 为块节点 n 的文本字符数(包含链接文本).

P_{textl} 反映了在某个节点的纯文本集中程度. 我们发现 P_{textl} 越大,往往意味着该节点越有可能包含待发现的正文内容块.

给出了 3 个密度度量后,可以定义块节点的综合密度特征值 $H(b)$:

$$H(b)=\frac{(p1*\ p2*\ p3)}{1+size(b)*\ \alpha} \tag{5}$$

其中 $b\in B$ 表示该块节点,$size(b)$ 表示该块节点中子孙节点个数. p1,p2,p3 分别代表节点的密度特征,取 $p1=p_{text}$,$p2=1-p_{link}$,$p3=p_{textl}$. α 是调节块节点 b 的子孙节点数量对 H 值影响的参数,在实验中取 $\alpha=0.3$. 当 α 设置过低时,选取的块可能带有噪音信息,当 α 设置过高时,利用综合密度特征 H 可能选取错误的块.

5.2 正文特征生成

通过 4.1 节我们可以得到网页标记类集合 M,在同一标记类中的网页,正文内容块的位置是相同的,所以在同一类网页中通过密度特征选择正文内容块,再提取正文内容块的特征作为该类网页正文内容的抽取规则,如算法 2 所示. 在网页中块的特征可以有三种表示方法,块 class 属性对应的值,块 id 属性对应的值和块的路径 path. 为了方便算法表示,本文给出了每个标记类中正文内容块特征的定义,即每个标记类正文内容的抽取规则.

定义 10. 给定聚类结果中标记类的标记 $c = \{c_i \mid c_i = C_i.c\}$，定义该标记类网页正文内容块的特征为一个三元组 $L(c) = <class, id, p>$，其中 $class$ 表示正文内容块 b 的 $class$ 属性对应的值，id 表示正文内容块 b 的 id 属性对应的值，p 表示正文内容块的标签路径，$p = \{p \mid p = p(b) \text{ 且 } b \in B\}$.

将聚类的结果经过算法 2 可以得到每个标记类对应正文内容块的特征 $L(c)$，即为该标记类网页的正文内容提取规则. 在 $L(c)$ 中记录正文内容块的三个特征，根据这三个特征可以从网页中提取出正文内容块. 在一个 Web 网页中并不是每个块都有 id 和 $class$ 属性，所以在 $L(c)$ 中，本文按优先级 $id > class > p$ 依次进行提取，当 id 和 $class$ 属性不存在时用路径 p 来提取正文内容块. 最后，从正文内容块中提取出正文内容.

算法 2. *getBlock*

输入: 标记类集合 $M = \{C_1, C_2, \cdots, C_n\}$，表示聚类结果为 n 个标记类

输出: 抽取规则集合 $N = \{L_1, L_2, \cdots, L_k\}$，表示 n 个标记类对应的 k 个正文内容块特征集合，其中 $k <= n$.

```
Begin
 For each C_i ∈ M then
 BF←∅//BF = { <L_1,H_1>, <L_2,H_2>, ⋯, <L_n,H_n> }
  For each w ∈ C_i.W then
  将网页 w 转化为 DOM 树结构，提取网页中的块集合 B
  计算块 b ∈ B 对应综合密度特征 H(b)
  选择 H(b) 最高的块 b，提取其特征 L(b)，记录该块的特征和其综合密度特征为 bf = <L,H>
  若 ∃ bf_i ∈ BF 且 bf_i.L = bf.L 则 bf_i.H = bf_i.H + bf.H 否则 add(BF,bf)
  End for
  从 BF 集合中选择 H 最高的 L 作为该标记类对应的抽取规则，add(N,L)
 End for
End
```

6 方法评估

为了验证本文提出方法的有效性，我们实现了相应的原型系统. 该原型系统分为两个过程: 基于网页聚类的正文特征生成和网页正文内容提取. 实验环境为 CPU(Inter Pentium CPU,3.10GHz) + RAM(8GB) + Window7 + Eclipse3.10. 在实验中，采用工具 Jsoup 对网页进行解析和块的提取.

实验中所使用的数据集 WebSet 来自包括 5 个网站的 1500 个网页. 该数据集通过半手工方式(种子 URL + 爬虫 + 手工筛选)从互联网网上收集得到的，来源于网易、搜狐、新浪、人民网和新华网，这些网页分布在网站中的不同主题类目. 在具体实验过程中，我们又从 WebSet 中产生 2 个子集: 1) 网页聚类数据集 WebSet-1. 包括 500 个网页从 WebSet 中手工选取，来自 5 个站点并且覆盖每个站点中的主题. 2) 网页正文内容抽取数据集 WebSet-2. 包括 1000 个网页.

我们对数据集 WebSet-1 中的网页进行聚类处理并生成正文特征，其结果如图 2 所示. 在实验中，网易和新浪中出现在同一主题模块的页面中产生多个类别. 图 3 展示了在聚类过程中，不同网站的页面中平均块节点个数，在图 3 结果中网易、搜狐和新浪的页面中平均块节点个数量远远超出新华网和人民网. 而在这些块节点中绝大多数是只包含噪音信息的块节点，因此除了网站中网页本身的设计结构的差异，网页中的噪音信息在一定程度上也影响网页的聚类结果.

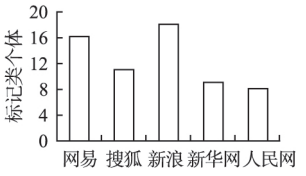

图 2 网页聚类结果

Fig. 2 Experimental result of webpage clustering

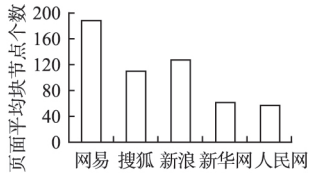

图 3 网页分块结果

Fig. 3 Experimental result of webpage segmentation

在网页正文提取方面，本文对数据集 WebSet-2 中的网页进行内容的提取. 实验分为两种，第一种是不利用网页聚类处理的结果，只通过块节点的综合密度特征来对网页正文内容进行提取，其结果如表 1 所示. 第二种是利用网页聚类和生成的正文内容块特征(抽取规则)来进行网页正文内容的提取，其结果如表 2 所示. 从表 1 和表 2 的对比中我们可以发现网页聚类能够显著提高网页正文内容提取的准确率，基本能够消除因为正文内容字符数较少导致提取错误的块的问题. 在 5 个站点中，网易的提取结果并不理想. 这是因为在网易财经模块中大部分网页并不存在正文内容块，而是将推荐链接等噪音信息与正文内容嵌入在同一个块中，导致实验中提取的正文内容块包含部分噪音信息.

表 1 无聚类处理的正文内容提取结果

Table 1 Experimental result of web information extraction with no clustering processing

DataSet	网页总数	准确率
网易	200	88%
搜狐	200	96.5%
新浪	200	95%
新华网	200	97%
人民网	200	92%

表 2 基于网页聚类的正文内容提取

Table 2 Experimental result of web information extraction with clustering processing

DataSet	网页总数	准确率
网易	200	92%
搜狐	200	100%
新浪	200	98%
新华网	200	100%
人民网	200	99.5%

在时间性能方面，因为网页结构的复杂程度不同，所以不同网站中网页聚类和正文特征生成所耗费的时间也存在差异. 实验中，平均对每 100 个网页进行聚类并生成正文特征的时间为 4571ms. 在网页正文内容提取方面，在无聚类的情况下，平均抽取一个网页的时间为 26ms，在有聚类的情况下平均抽取一个网页的时间 21ms. 从实验结果来看，在有聚类的

情况下平均抽取一个网页的时间比无聚类情况下要快 5ms.

文献 [8] 也是一种基于网页聚类的正文提取方法,该方法采用树编辑距离计算网页之间的相似度,并且利用 DOM 树的结构差异来确定网页的抽取规则,其准确率为 82.5%. 与该方法相比本文采用的方法的准确率高达 97.9%,并且本方法采用路径集合来计算网页之间的相似度降低了网页聚类的时间消耗. 文献 [1] 中的 CEPR 算法在网易,新浪,新华网和人民网的数据集上精确率达到 99.29%,98.57%,94.72% 和 95.11%,基本与本方法相当. 然而,CEPR 算法平均抽取一个网页的时间为 375ms,不适合针对大规模网页的处理.

7　总　结

本文结合现有的 Web 信息提取方法,基于网页分割的正文提取和基于统计的密度特征正文提取,再结合网页聚类,提出了一种基于网页聚类的正文信息提取方法. 该方法利用对结构相同的网页进行统一的提取操作,来提高网页正文内容抽取的准确率. 在实验中,我们将有聚类处理和无聚类处理的网页正文内容提取进行对比,其准确率显著上升. 本方法适用于提取来自同一网站的网页,不需要复杂的计算,简单实用.

未来的工作重点主要包含两个方面: 一方面,将本文提出方法运用到大规模网页处理的环境中. 另一方面,已有的 Web 信息抽取方法主要提取粗粒度的 Web 内容,面向精准的细粒度结构化 Web 信息抽取的精度仍不是很理想,因此,我们的研究重点将转为对网页中细粒度的实体提取.

References:

[1] Wu Gong-qing, Hu Jun, Li Li, et al. Online web news extraction via tag path feature fusion [J]. Journal of Software, 2016, 27(3): 714-735.

[2] Wang J, Wang J. qRead: a fast and accurate article extraction method from web pages using partition features optimizations [C]. Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Lisbon, Portugal, 2015: 364-371.

[3] Zhang Nai-zhou, Cao Wei, Li Shi-jun. A method based on node density segmentation and label propagation for mining web page [J]. Journal of Computer Science and Technology, 2015, 38(2): 349-364.

[4] Cai D, Yu S, Wen J R, et al. Extracting content structure for web pages based on visual representation [C]. Proceedings of the 5th Asian-Pacific Web Conference (APWEB 2003), Xi'an, China, 2003: 406-417.

[5] Yin X, Lee W S. Using link analysis to improve layout on mobile devices [C]. Proceedings of the 13th International Conference on World Wide Web (WWW 2004), New York, USA, 2004: 338-344.

[6] Wu S, Liu J, Fan J. Automatic web content extraction by combination of learning and grouping [C]. Proceedings of the 24th International Conference on World Wide Web (WWW 2015), Florence, Italy, 2015: 1264-1274.

[7] Joshi S, Agrawal N, Krishnapuram R, et al. A bag of paths model for measuring structural similarity in Web documents [C]. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003), Washington, USA, 2003: 577-582.

[8] Yang Tian-qi, Qiu Tao-fen. A method of automatic web information extraction based on page clustering [C]. Proceedings of the 8th World Congress on Intelligent Control and Automation (WCICA 2011), Taipei, 2011: 390-393.

附中文参考文献:

[1] 吴共庆,胡　骏,李　莉,等. 基于标签路径特征融合的在线 Web 新闻内容抽取 [J]. 软件学报,2016,27(3): 714-735.

[3] 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.

[8] 杨天奇,邱韬奋. 一种 Web 信息自动抽取的网页聚类方法 [C]. 第 8 届智能控制与自动化世界大会(WCICA 2011),台北,2011: 390-393.