软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(3):714−735 [doi: 10.13328/j.cnki.jos.004868]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 基于标签路径特征融合的在线 Web 新闻内容抽取*

吴共庆[1], 胡 骏[1], 李 莉[1], 徐喆昊[1], 刘鹏程[1], 胡学钢[1], 吴信东[1,2]

[1](合肥工业大学 计算机与信息学院,安徽 合肥 230009)

[2](Department of Computer Science, University of Vermont, Burlington, VT 05405, USA)

通讯作者: 吴信东, E-mail: xwu@hfut.edu.cn

**摘 要**: 精准地抽取新闻网页的内容,是提高 Web 新闻分析等应用系统工作质量的关键技术之一.由于缺少 Web 新闻出版的标准,存在大量不同的出版格式,并且 Web 本身是一种具有高度异构性的大数据载体,导致 Web 新闻内容抽取成为一个开放性问题.经大量实例分析发现,新闻网页内容与其上的标签路径存在潜在的关联性.因此,设计了标签路径特征系,以从不同视角区分网页内容和噪音.在特征相似性分析的基础上,提出了一种基于组合特征选择的特征融合策略,并设计了基于融合特征的 Web 新闻内容抽取方法 CEPF.CEPF 是一种快速的通用、无需训练的在线 Web 新闻内容抽取算法,可抽取多种来源、多种风格、多种语言的 Web 新闻网页.在 CleanEval 等测试数据集上的实验结果表明,CEPF 方法优于 CETR 等抽取方法.

**关键词**: 内容抽取;Web 新闻;标签路径特征;组合特征选择;特征融合

**中图法分类号**: TP311

中文引用格式: 吴共庆,胡骏,李莉,徐喆昊,刘鹏程,胡学钢,吴信东.基于标签路径特征融合的在线 Web 新闻内容抽取.软件学报,2016,27(3):714−735. http://www.jos.org.cn/1000-9825/4868.htm

英文引用格式: Wu GQ, Hu J, Li L, Xu ZH, Liu PC, Hu XG, Wu XD. Online Web news extraction via tag path feature fusion. Ruan Jian Xue Bao/Journal of Software, 2016,27(3):714−735 (in Chinese). http://www.jos.org.cn/1000-9825/4868.htm

## Online Web News Extraction via Tag Path Feature Fusion

WU Gong-Qing[1], HU Jun[1], LI Li[1], XU Zhe-Hao[1], LIU Peng-Cheng[1], HU Xue-Gang[1], WU Xin-Dong[1,2]

[1](School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

[2](Department of Computer Science, University of Vermont, Burlington, VT 05405, USA)

**Abstract**: Accurately extracting content from Web news is a key technology for quality improvement in Web news analysis and applications. Due to the lack of publication standards, differences in publishing formats, and a highly heterogeneous big data carrier of the Web itself, Web news extraction has become an open research problem. Extensive case studies by this research indicate that there is potential relevance between Web content layouts and their tag paths. Inspired by this observation, this paper designs a series of tag path extraction features to distinguish the Web content and noise from different perspectives. Based on the similarity analysis of these features, the paper proposes a features fusion strategy with group feature selection, and provides a Web news extraction method via feature fusion, CEPF. CEPF is a fast, universal, no-training and online Web news extraction algorithm. It can extract Web news pages across multi-resources, multi-styles, and multi-languages. Experimental results with public data sets such as CleanEval show that the CEPF method achieves better performance than the state-of-the-art CETR method.

* 基金项目: 国家自然科学基金(61273297, 61229301, 61273292); 教育部创新团队发展计划(IRT13059); 国家重点基础研究发展计划(973)(2013CB329604); 国家高技术研究发展计划(863)(2012AA011005)

Foundation item: National Natural Science Foundation of China (61273297, 61229301, 61273292); The Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education (IRT13059); National Program on Key Basic ResearchProject of China (973 Program) (2013CB329604); National High-Tech R&D Program of China (863 Program) (2012AA 011005)

收稿时间: 2015-01-31; 修改时间: 2015-05-08; 定稿时间: 2015-06-09

**Key words**: content extraction; Web news; tag path feature; group feature selection; feature fusion

互联网是大数据的一个重要载体.《2013 互联网趋势报告》(2013 Internet trends)(http://www.kpcb.com/insights/2013-internet-trends)指出,互联网用户数量激增,2012 年,全球互联网用户达 24 亿,同比增长 8%.《2014 互联网趋势报告》(2014 Internet trends)(http://www.kpcb.com/insights/2014-internet-trends)指出,互联网用户数量仍保持着继续增长的趋势.根据中国互联网信息中心(CNNIC)的调查显示,阅读网络新闻是互联网用户的主要行为之一.据 CNNIC 测算,在中国,阅读互联网新闻的互联网用户普及率为 78.7%(http://research.cnnic.cn/img/h000/h11/attach200911091321380.pdf)、移动互联网用户普及率为 59.5%(http://www.cnnic.net.cn/hlwfzyj/hlwfzzx/qwfb/201306/W020130614603405171795.pdf),位居第一.各种各样的 Web 新闻网页成为人们获取信息的快捷途径,成为相继报纸之后新一代的信息载体.

然而,Web 新闻网页除了包含用户感兴趣的正文内容外,还包含导航条、广告、推荐链接、版权声明、免费声明等与网页主题无关的噪音信息.2005 年初,Gibson 等人研究表明:网页中的噪音数据占整个网页数据的 40%~50%,且预测该比例可能会持续增加[1].这些庞大的噪音数据降低了 Web 新闻网页的价值密度,同时给 Web 新闻网页的存储、Web 内容的管理与分析、Web 新闻聚合等研究和应用带来了巨大的挑战.如何过滤这些噪音信息,得到干净的 Web 新闻网页,促使 Web 新闻数据从量到质的转变,正是 Web 新闻内容抽取的研究目的.

从多源、海量、异构、价值密度低的 Web 新闻网页中精准地抽取网页正文内容,是极具挑战性的问题.由于缺少 Web 新闻出版的标准,存在大量不同的出版格式,并且 Web 本身是具有高度异构性的载体[2],导致 Web 新闻内容抽取成为开放性问题[3].近年来,随着新闻网页中噪音数据的增加、CSS 技术的广泛应用、不同新闻网站网页结构差异性的增大以及网页自身结构复杂性的提高,Web 新闻网页内容抽取研究面临更多的挑战.Web 新闻网页具有海量异构的特点,对手工构造包装器技术以及基于规则学习的包装器技术提出了挑战.这两类技术通常适合针对特定的网站构建包装器,面向海量异构的 Web 新闻网页抽取内容时通常会失效. Web 新闻网站的显示风格、页面结构会不断变化,导致基于视觉特征、模板推导类的包装器技术面临挑战.

为了从多源、海量、异构的 Web 新闻网页中精准地抽取网页内容,解决实际应用的需求,通过大量的实例分析研究发现,Web 新闻网页内容与其对应的 DOM 树中的标签路径间存在潜在关联,具体表现在:

(1) Web 新闻网页的内容部分有相似的标签路径,且包含较长的文本内容和较多的标点符号;

(2) Web 新闻网页的噪音部分有相似的标签路径,且包含较短的文本内容和较少的标点符号.

此外我们还发现,根据网页噪音的表现形式,一般可以分为两类:一类包含较短的文本和较少的标点符号,文本长度或标点个数与正文内容差别较大;一类分布于网页布局的边缘,有着背景、颜色、超链接等修饰,虽然文本长度与正文接近,但是这类噪音的标签路径的层次数多于正文的标签路径层次数,即,噪音的标签路径中标签的个数多于正文的标签路径中标签的个数.

根据上述观察,本文设计了基于标签路径的抽取特征系,以从不同的视角区分新闻网页内容和噪音.这些抽取特征有各自独特的优势,且相互之间也有一定的关联性.在抽取不同的 Web 新闻网页时,每个特征的优势的强度、特征之间的相关性也有所不同.因此,在抽取特定的 Web 新闻网页时,如何去除冗余的抽取特征、提高抽取的精度和效率是一个关键问题.本文采用高斯相似函数度量特征相互间的冗余度,基于图划分的思想提出了一种组合特征选择策略,并融合所选择的特征为一个综合特征.使用该综合特征,判断某个标签路径所达到的内容是否为正文内容.该方法是一种在线抽取方法,无需用户参与、无需对样本网页进行标记和训练、不依赖于网页的模板且适合抽取不同语言的新闻网页,具有较好的通用性、抽取精度和抽取效率.

本文的贡献包括 3 个方面:

(1) 根据 Web 新闻内容与其对应的解析树中的标签路径间存在着潜在关联的现象,设计了用于抽取新闻网页内容的标签路径特征系,从不同的视角区分 Web 新闻内容和噪音;

(2) 提出了一种组合特征选择策略,在此基础上,设计了标签路径特征系的特征融合方法,以获得具有更强区分能力的综合特征;

(3) 基于标签路径特征系的融合特征,设计了在线 Web 新闻内容抽取算法 CEPF,并验证了该算法适于海量、异构、多语言的 Web 新闻内容抽取工作.

本文第 1 节介绍基于标签路径特征融合的 Web 新闻内容抽取框架.第 2 节详细介绍用于 Web 新闻内容抽取的标签路径特征系.第 3 节提出一种组合特征选择策略.第 4 节提出基于特征融合的 Web 新闻内容抽取方法.以现实的数据源为实验数据,第 5 节给出抽取算法的性能测试结果和对比实验结果.第 6 节介绍相关工作.最后,第 7 节对本文工作进行总结和展望.

## 1 基于标签路径特征融合的 Web 新闻内容抽取框架

图 1 给出了基于标签路径特征融合的 Web 新闻内容抽取框架,用户或应用程序提交一个 URL 作为初始输入,返回被抽取的网页内容.

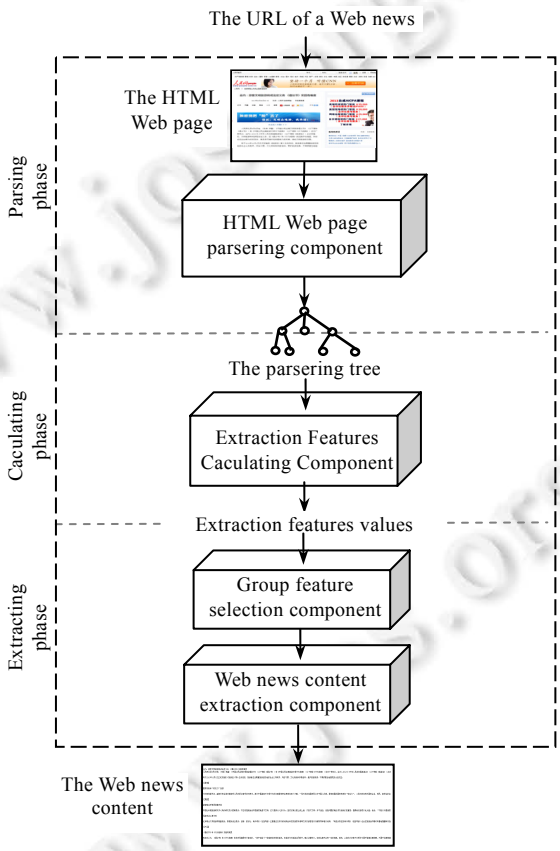



Fig.1 Architecture of the Web news extraction system via tag path features fusion

图 1 基于标签路径特征融合的 Web 新闻内容抽取

基于标签路径特征融合的 Web 新闻内容抽取框架由 3 个模块组成,分别是 HTML 网页解析模块、抽取特征计算模块和基于标签路径特征融合的 Web 新闻内容抽取模块.

- HTML 解析模块:给定一个 HTML 网页,从网页中移去脚本、注释、样式标签,因为这些信息在页面上是不可见信息,不必纳入计算范畴.并将网页解析成一棵 DOM 树,可使用已有的任意 HTML 网页解析器,如 HTMLParser 等;
- 标签路径特征系计算模块:给定 HTML 网页的解析树,遍历 DOM 树,获取所有的标签路径,并统计标签路径的相关信息.在此基础上,计算所有标签路径的特征值.如何定义标签路径特征?什么样的标签路

径适合抽取 Web 新闻网页的内容?这是标签路径系计算模块需要探索研究的两个核心问题;

- 基于标签路径特征融合的 Web 新闻内容抽取模块:该模块包含两部分,分别是组合特征选择模块和抽取模块:组合特征选择模块用于过滤冗余特征,选择一组相对独立的特征;抽取模块首先将选择后的特征集合融合为一个综合特征,然后根据抽取控制策略对该解析树上的 Web 新闻内容节点进行抽取,并返回抽取后的 Web 新闻内容.如何进行组合特征选择?如何进行特征融合?以及使用何种抽取控制策略?这些是抽取模块需要探索研究的问题.

## 2 标签路径特征系

以图 2 的新闻网页为例(http://www.freep.com/article/20140213/NEWS07/302130081/Brazil-woman-fetus-stone-baby),该图展示了一个典型的 Web 新闻网页.实线框内的部分为 Web 新闻内容,虚线框内的部分为噪音内容.导航栏、广告、推荐链接等噪音部分占据了一半以上的页面.


Fig.2 Example of an article from Freep website

图 2 底特律自由报新闻网页样例

为了准确抽取新闻网页中的正文内容,我们观察分析了大量不同网站、不同设计风格的新闻网页并发现以下现象:

(1) 新闻网页的主题部分是一个整体,每个段落具有相似的显示格式;

(2) 噪音内容有很多块,同一块中的噪音内容有相似的显示格式;

(3) 噪音内容大部分都分布在整个新闻网页页面的边缘;

(4) 新闻网页的主题部分和噪音部分在文本内容的表现形式上有显著的区别;

(5) 进一步,在分析这些内容块和噪音块的网页结构后发现,同一块的信息片段对应的标签路径有着类似的结构.

图 2 中的样例网页也有上述特征.基于以上发现,本节将首先描述 HTML 网页 DOM 树及其上的标签路径,然后根据正文内容与噪音内容在标签路径和文本内容上的一些显著区别,定义了一系列基于标签路径的抽取特征,这些特征具有区分内容和噪音的能力,并称为标签路径特征系.

### 2.1 标签路径

DOM 树模型,又称为文档对象模型,是 W3C 组织推荐用于处理 HTML 和 XML 文件的标准模型.每个 Web

新闻网页都对应着一棵以标签为内部节点、文本或图像为叶子节点的 DOM 树,对 HTML 文档的处理可以通过对 DOM 树的操作实现.

为简洁、方便地描述问题,本文采用标记有序树作为 HTML 网页的表示模型;根据信息抽取的需求,扩展标记有序树模型.在扩展标记有序树模型的同时,确保该表示模型上的操作能在 DOM 模型上方便地实现.在不引起上下文混淆的情况下,本文将一个网页的 DOM 解析树和扩展标记有序树视为同一概念.

**定义 1(扩展标记有序树)**. $L=\{l_0,l_1,l_2,\ldots\}$是有限字母表的集合,$l_i$为 HTML 的网页标记元素.$L$ 上的扩展标记有序树定义为七元式 $T=(V,E,v_0,\prec,L,l(\cdot),c(\cdot))$,其中:$V$ 是节点的集合;$E$ 是边的集合;$v_0\in V$;$\prec$是节点的兄弟关系集合,$G=(V,E,v_0,\prec)$是一棵以 $v_0$ 为根的有序树;映射 $l:V\to L$ 是标记函数,$\forall v\in V,l(v)$为节点 $v$ 上的标记;映射 $c:V\to String$ 是内容函数,$\forall v\in V,c(v)$为节点 $V$ 上的内容.

扩展标记有序树 $T$ 上的映射 $c(\cdot)$可根据抽取任务的需要具体定义,本文定义 $c(\cdot)$为节点上的文本内容.

**定义 2(标签路径)**. $T$ 是 $L$ 上的以 $v_0$ 为根的扩展标记有序树,$\forall v\in V,v_0v_1\ldots v_k$ 是树 $T$ 从 $v_0$ 到达 $v_k$ 的节点序列,其中,$parent(v_i)=v_{i-1}(1\leqslant i\leqslant k),v_k=v$,则称 $l(v_0)l(v_1)\ldots l(v_k)$为节点 $v$ 的路径,记为 $path(v)$.

### 2.2 标签路径特征系的设计

大量的实例研究发现:Web 新闻网页的正文内容和噪音内容在文本特征方面存在显著的区别,而且 Web 新闻网页中的内容分布和标签路径之间存在着潜在的联系.这些区别和联系具体可以描述为:

(1) 标签路径信息:Web 新闻网页的正文内容有着相同或相似的标签路径,噪音内容有着相似或相同的标签路径;

(2) 节点信息:所有的文本节点都是叶子节点;

(3) 节点的文本信息:内容节点的文本长度通常较长,噪音节点的文本长度通常较短;

(4) 节点的标点符号信息:内容节点上的文本包含较多的标点符号,噪音节点上的文本包含较少的标点符号;

(5) 网页的文本信息:新闻网页的正文内容的总文本长度一般大于网页中噪音的总文本长度;

(6) 网页的标点符号信息:新闻网页中,正文内容中所包含的文本节点的个数一般多于噪音中标点符号的个数;

(7) 修饰信息:Web 新闻网页的正文内容有较少的修饰信息,即,正文内容的标签路径层次较低;网页的噪音部分有着较多的修饰信息,如超链接、背景、字体颜色等多方面的修饰,涉及较多的标签,即,噪音的标签路径层次较高.

根据网页正文内容和噪音在基本信息方面的差异,本文定义了一个基于标签路径的抽取特征系,简称标签路径特征系.标签路径特征系中的特征分别是文本标签路径长度特征(text to tag path length,简称 TPL)、文本标签路径比特征(text to tag path ratio,简称 TPR)、文本标签路径层次比特征(text to tag path level ratio,简称 TPLR)、标点标签路径长度特征(punctuation to tag path length,简称 PPL)、标点标签路径比特征(punctuation to tag path ratio,简称 PPR)和标点标签路径层次比特征(punctuation to tag path level ratio,简称 PPLR),在给出相关概念之后,将分别介绍每个抽取特征.

给定解析树 $T$ 上的节点 $v$,记 $c(v)$为节点 $v$ 上的文本内容,$length(c(v))$为节点 $v$ 上的文本长度,$puncNum(c(v))$为节点 $v$ 上的标点符号的个数.

设 $p$ 是解析树 $T$ 上的一条标签路径,记 $accNodes(p)=\{v_p^1,v_p^2,\ldots,v_p^m\}$为标签路径 $p$ 在树 $T$ 上的可达文本节点的集合,可以看出:$p$ 在树 $T$ 中的标签路径数为 $p$ 在树 $T$ 上的可达文本节点的个数,即$|accNodes(p)|$;$level(p)$表示标签路径 $p$ 的层次数,即,标签路径 $p$ 中包含的标签个数.

**定义 3(文本标签路径长度特征)**. 设 $p$ 是树 $T$ 上的标签路径,定义 $p$ 的文本标签路径长度 $TPL(p)$为 $p$ 上可达文本节点的文本长度之和,即:

$$TPL(p)=\sum_{v\in accNodes(p)} length(c(v)) \tag{1}$$

**定义 4(文本标签路径比特征)**. 设 $p$ 是树 $T$ 上的标签路径,定义 $p$ 的文本标签路径比 $TPR(p)$为 $p$ 上可达文本节点的文本长度之和与标签路径数的比值,即:

$$TPR(p)=\frac{\sum_{v\in accNodes(p)} length(c(v))}{|accNodes(p)|} \tag{1}$$

**定义 5(文本标签路径层次比特征)**. 设 $p$ 是树 $T$ 上的标签路径,定义 $p$ 的文本标签路径层次比特征 $TPLR(p)$为 $p$ 上可达文本节点的文本长度之和与标签路径层次数的比值,即:

$$TPLR(p)=\frac{\sum_{v\in accNodes(p)} length(c(v))}{|level(p)|} \tag{3}$$

**定义 6(标点标签路径长度特征)**. 设 $p$ 是树 $T$ 上的标签路径,定义 $p$ 的标点标签路径长度 $PPL(p)$为 $p$ 上可达文本节点的标点符号个数之和,即:

$$PPL(p)=\sum_{v\in accNodes(p)} puncNum(c(v)) \tag{4}$$

**定义 7(标点标签路径比特征)**. 设 $p$ 是树 $T$ 上的标签路径,定义 $p$ 的标点标签路径比 $PPR(p)$为 $p$ 上可达文本节点的标点符号个数之和与标签路径数的比值,即:

$$PPR(p)=\frac{\sum_{v\in accNodes(p)} puncNum(c(v))}{|accNodes(p)|} \tag{5}$$

**定义 8(标点标签路径层次比特征)**. 设 $p$ 是树 $T$ 上的标签路径,定义 $p$ 的标点标签路径层次比 $PPLR(p)$为 $p$ 上可达文本节点的标点符号个数之和与标签路径层次数的比值,即:

$$PPLR(p)=\frac{\sum_{v\in accNodes(p)} puncNum(c(v))}{|level(p)|} \tag{6}$$

一旦 HTML 文本被解析成 DOM 树后,通过对 DOM 树的遍历,可以获得 DOM 树中所有的标签路径,统计每个文本节点的文本长度、包含的标点符号的个数.因此,每个标签路径的所有标签路径特征值都可以方便地计算出来.

以 TPL 特征为例,图 3 给出了图 2 所示 Freep 样例网页中的每个标签路径特征值的柱状图.图中的横坐标表示先序遍历解析树过程中文本节点的出现序号,纵坐标表示该文本节点的标签路径特征值.其中,红色部分为网页中的内容节点特征值,蓝色部分则为噪声节点特征值.通过观察该直方图可以发现,内容节点的特征值比大部分噪声节点的要大:左边噪声节点特征值比较高的原因是由该新闻网页中的导航文本节点含有大量文本,且这些节点处于同一标签路径下;直方图右边的噪声节点特征值比较高的原因则为该网页中存在大量的相关新闻超链接和网站的免责声明,导致这部分噪声节点的特征值比较高.

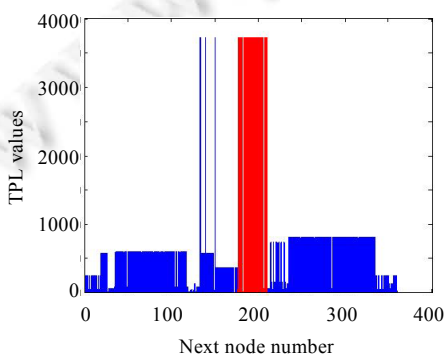



Fig.3 TPL of each text node from the Web news of Freep

图 3 Freep 样例网页的 TPL 柱状图

### 2.3 标签路径特征系分析

Web 新闻网页具有多源异构的特点,不同的 Web 新闻网页中噪音与内容的区别点也有所不同.因此,在抽取不同的 Web 新闻网页时,标签路径特征系中每个特征准确识别网页内容的能力也是不同的.

根据文本标签路径长度特征(TPL)的定义可知:TPL 特征将所到达的文本节点的文本长度之和的值较大的标签路径识别为内容路径,并抽取节点中的内容.但事实上,并不是所有的新闻网页所包含的正文内容都多于噪音内容.因此,文本标签路径长度特征在抽取正文内容多于噪音内容的 Web 网页时会达到的较好的抽取效果,但对于噪音内容较多的 Web 网页,其抽取能力较差.

文本标签路径比特征(TPR)是根据 Web 新闻网页中的内容与噪音在节点的文本信息上的差异而设计的.文本标签路径比特征是在文本标签路径长度特征的基础上,平均了标签路径所到达的所有文本节点的文本长度,不会受到内容或噪音的文本总长度的影响.但对于一些仅到达一个文本节点且包含较长文本的噪音节点,如网页的版权声明信息,文本标签路径比特征无法准确过滤,而文本标签路径长度可以较好地过滤掉这类噪音信息.

文本标签路径层次比特征(TPLR)不同于文本标签路径长度特征(TPL)和文本标签路径比特征(TPR),其利用了网页的修饰信息,主要解决分布于网页边缘、有较多修饰的噪音信息的识别问题.一般网页中的噪音都位于网页的边缘,有着特殊的修饰,如超链接、背景等修饰,即,到达网页噪音的标签路径包含的标签的个数一般多于到达内容的标签路径所包含的标签个数.文本标签路径层次比抽取特征的缺点是,其抽取性能易受到网页设计风格的影响.

标点标签路径长度特征(PPL)、标点标签路径比特征(PPR)和标点标签路径层次比特征(PPLR)是在 TPL 特征和 TPR 特征的基础上,将文本长度信息替换为标点符号数量信息.一般网页的噪音部分包含较少的标点符号或不包含标点符号,而正文部分包含较多的标点符号.相对于文本信息,标点符号信息区分网页正文和噪音的能力更强.但当网页的内容是诗体或类似的形式时,PPL 特征、PPR 特征和 PPLR 特征在区分内容和噪音时会失效.

根据标签路径特征系的定义,在遍历 HTML 网页解析树的过程中,可以方便地计算各个标签路径特征.因此,计算标签路径特征系的时间复杂度为 $O(n)$,其中,$n$ 为解析树上节点的个数.

## 3 组合特征选择

组合特征选择是从一组数量为 $D$ 的特征中选择出数量为 $d(D>d)$、彼此之间相关联程度较小的一组特征子集的过程.组合特征选择的目的是去除冗余的特征,以便于特征融合过程在减少计算量的同时提高融合特征的质量.

根据标签路径特征系中每个抽取特征的物理意义,我们知道每个标签路径特征的抽取能力都是唯一的.在抽取不同类型的 Web 新闻网页时,每个标签路径特征会表现出不同的优势;同时,标签路径特征之间也有不同的相似度.为提高抽取的效率和精度,需要去除一些冗余特征,然后综合各特征的优势,将被选择的特征集合融合为一个综合特征.针对一个特定的 Web 新闻网页,如何度量特征间的相似性并过滤掉冗余的抽取特征,是本节所要研究的问题.

本文解决该问题的思路是:设 Web 新闻网页 $wp$ 的解析树 $T$ 上有 $n$ 个文本节点,其先序遍历的顺序为 $\langle v_1,v_2,...,v_n\rangle$.标签路径特征系中的每个特征在这 $n$ 个文本节点上的取值为一个 $n$ 维向量,例如,$TPR$ 特征的取值为 $n$ 维向量 $\langle TPR(path(v_1),TPR(path(v_2),\ldots,TPR(path(v_n)\rangle$.因此,标签路径特征系中的每个特征均可看做是 $n$ 维向量空间中的一个点.标签路径特征系中的每个特征对应当点之间均有一条边连接,每条边都有一个权重,其值为两点(特征)之间的相似度.由此,我们得到一个无向加权图 $G(V,E)$,其中,$V$ 为顶点的集合,$E$ 为加权边的集合.于是,组合特征选择的问题转化为图的划分问题,即:基于图的最优划分准则,使得划分出的子图内部相似度最大、子图间的相似度最小.根据划分的结果,我们可以认为:子图内的点(特征)之间存在冗余,子图间的点(特征)彼此不冗余.从而解决了组合特征选择问题.

对于特征(无向图中的点)的相似度度量问题,我们选取了高斯相似函数来度量特征相似性.对于图的划分问题,我们选取了谱聚类方法.

### 3.1 标签路径特征相似性度量

在进行图的划分(谱聚类)之前,选择一种合适的方法度量特征之间是相似性尤为重要的.我们采用的是高斯函数度量两特征之间的相似性,见公式(7):

$$similarity(X,Y)=\mathrm{e}^{-\frac{(dist(X,Y))^2}{2\sigma^2}} \tag{7}$$

其中,$X$ 和 $Y$ 为标签特征系中任意两个特征在 $n$ 维向量空间中特征值形成的两个 $n$ 维向量,$X=\langle x_1,x_2,\ldots,x_n\rangle$,$Y=\langle y_1,y_2,\ldots,y_n\rangle$;$\sigma$为用户指定的尺度参数;$dist(X,Y)$为 $X$ 和 $Y$ 的距离,计算方法见公式(8).

$$dist(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_n-y_n)^2} \tag{8}$$

高斯相似函数是经典谱聚类算法中计算两点间相似度的常用方法,当谱聚类算法使用高斯相似函数度量两点的相似性时,其收敛性已被证明[4].

### 3.2 基于谱聚类的组合特征选择

根据公式(7),我们可以计算出标签路径特征系中任意两个标签路径特征的相似度.将标签路径特征视为 $n$ 维向量空间中的一个点,任意两点(特征)间的边的权重为两点(特征)的相似度,我们可以得到一个无向加权图 $G(V,E)$.其中:$V$ 是图 $G$ 的顶点集合,即,标签路径特征系;$E$ 是图 $G$ 的边集合.于是,组合特征选择问题转化为了图的划分问题.

鉴于谱聚类算法完整的理论基础和优越的实验性能,我们选取谱聚类算法解决图的划分问题,也就是本文的组合特征选择问题.谱聚类算法建立在谱图理论基础上,与传统的聚类算法相比,它具有能在任意形状的样本空间上聚类且收敛于全局最优解的优点[5].谱聚类算法有多种划分准则,划分准则的好坏直接影响到聚类结果的优劣.常见的划分准则有最小割集准则(minimum cut)[6]、平均割集准则(average cut)[7]、规范割集准则(normalized cut)[8]等.本文采用规范割集准则(normalized cut),该划分准则不仅能够衡量类内样本间的相似程度,也能衡量类间样本间的相异程度.经谱聚类划分后,可得到聚类结果 $C=\{C_1,C_2,\ldots,C_k\}(1\leqslant k\leqslant m)$,其中:$m$ 为标签路径特征系中特征的数量,在本文中 $m=6$;$k$ 值由谱聚类算法自动确定.同一类中,标签路径特征间存在冗余,不同类别的标签路径特征是不冗余的.为保证被选特征子集中不存在冗余特征,我们从每个类别中任意选择一标签路径特征构成我们的特征子集.

图 2 所示样例网页的标签路径特征系的划分结果为$\{\{TPL\},\{TPR\},\{TPLR,PPL,PPR,PPLR\}\}$.因此在样例网页中,TPLR,PPL,PPR,PPLR 特征间存在冗余.在特征融合时,只能从这 4 个特征中选择一个.因此,样例网页的组合特征选择结果为$\{TPL,TPR,f\}$,其中,$f\in\{TPLR,PPL,PPR,PPLR\}$.

## 4 基于组合特征选择的 Web 新闻内容抽取方法

### 4.1 特征融合

特征融合技术就是利用网页的各个特征之间的互补信息,充分发挥各自的优势,弥补各自的局限,将多个特征融合为一个综合能力更好的特征.

在进行组合特征选择之后,被选择的特征子集中的各个特征之间是相互独立的,而且每个特征都有自己独特的抽取优势,不存在某种特征能够在所有的新闻网页抽取中都优于其他特征,也不存在某种可以完全被其他特征所替代的特征.如何综合考虑每个特征的优势/劣势,将多个特征融合成一个综合决策特征,是决定抽取效果的一个关键性问题.

信息融合(information fusion)起初被称为数据融合(data fusion),起源于 1973 年美国国防部资助开发的声纳信号处理系统.信息融合实现的基本原理是:模拟人类大脑对接收到的各种信息进行加工处理,然后根据经验或相关理论知识对数据进行综合分析做出最终判断的过程.依据信息融合系统中数据抽象层次,融合可以划分为 3 个级别:数据级融合、特征级融合和决策级融合.特征融合是中间层次的信息融合.

信息融合从提出到现在已有几十年的历史,其研究应用已经非常广泛.然而直到现在,信息融合尚未形成统

一的理论框架、通用的融合模型和算法[9].目前,融合研究大多是针对某一特定领域的具体问题进行的,因此需要在解决具体问题时充分分析问题的特点,针对融合的目标及数据特性选择融合算法.从运行环境、处理的信息类型、信息表示、信息的不确定性、融合技术和适用范围等角度,表 1 给出了常用融合方法性能比较[10,11].

**Table 1** Performance comparison of common methods

**表 1** 常用融合方法性能比较

| 融合方法 | 运行环境 | 信息类型 | 信息表示 | 不确定性 | 融合技术 | 适用范围 |
|---|---|---|---|---|---|---|
| 加权平均 | 动态 | 冗余 | 原始值读取 | | 加权平均 | 低层融合 |
| 乘法 | 动态 | 冗余互补 | 原始值读取 | | 乘法 | 低层融合 |
| 卡尔曼滤波 | 动态 | 冗余 | 概率分布 | 高斯噪声 | 系统模型滤波 | 低层融合 |
| 贝叶斯估计 | 静态 | 冗余 | 概率分布 | 高斯噪声 | 贝叶斯估计 | 低层融合 |
| 统计决策 | 静态 | 冗余 | 概率分布 | 高斯噪声 | 极值决策 | 高层融合 |
| 证据理论 | 静态 | 冗余互补 | 命题 | | 逻辑推理 | 高层融合 |
| 模糊理论 | 静态 | 冗余互补 | 命题 | 隶属度 | 逻辑推理 | 高层融合 |
| 神经元网络 | 动静 | 冗余互补 | 神经元输入 | 学习误差 | 神经网络 | 低或高 |
| 产生式规则 | 动静 | 冗余互补 | 命题 | 置信因子 | 逻辑推理 | 高层融合 |

从 Web 新闻网页中动态提取的标签路径特征系,通过组合特征选择算法,我们得到一组冗余互补的原始值表示的特征.从表 1 可以发现:乘法在运行环境、信息类型、信息表示、不确定性处理、适用范围等方面,非常适合作为融合多个标签路径特征的方法.另外,乘法是一种非常简单直观的融合方法,即:将多个选出的标签路径特征值相乘,所得的积作为融合结果.此方法的最大特点是计算量小,能够实时对海量 Web 新闻网页数据抽取过程的特征进行融合.根据选择出的特征子集融合算法,每个标签路径的综合特征值 $TPF(p)$计算如下:

$$TPF(p)=\prod_{f\in M}F(f,p) \tag{9}$$

其中,$M$ 是选择出的特征子集,$M\subseteq\{TPL,TPR,PPL,PPR,TPLR,PPLR\}$.$F(f,p)$的计算见公式(10):

$$F(f,p)=\begin{cases}TPL(p), & \text{if } f=TPL\\ TPR(p), & \text{if } f=TPR\\ TPLR(p), & \text{if } f=TPLR\\ PPL(p), & \text{if } f=PPL\\ PPR(p), & \text{if } f=PPR\\ PPLR(p), & \text{if } f=PPLR\end{cases} \tag{10}$$

与采用单个标签路径特征区分内容路径和噪音路径的方法类似,我们按以下方式进行判定:$TPF(p)$值相对较高的标签路径聚合了内容节点;反之,$TPF(p)$值较低的标签路径通常聚合了噪音节点.图 4 为 TPF(标签路径的综合特征值)的直方图.

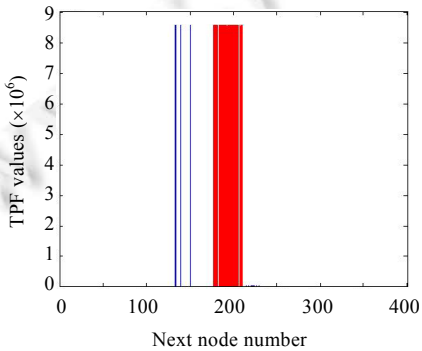



Fig.4 TPF of each text node from the Web news of Freep

图 4 Freep 样例网页的 TPF 柱状图

在该实例中,基于谱聚类的组合特征选择方法,选择了 TPL,TPL 和 PPLR 特征,根据乘法融合策略计算得到 TPF 值.相对之前提出的 6 个单特征,TPF 有更好的内容噪音区分能力.通过观察该直方图易发现:内容节点比绝大部分噪声节点的特征值要大,但是部分特征值之间差距不大.这是因为部分噪声节点文本内容数和标点符号数量都比较多,导致噪声节点的 TPF 值依然比较大,内容节点的特征值无法有效地与其区分出来.与图 3 比较可以看出:在综合特征值柱状图中,噪音节点和内容节点间的区分度增大.

### 4.2 融合特征的扩展

大量的实例研究统计发现:Web 新闻网页中的内容分布和标签路径之间存在着潜在的联系,除了第 2.2 节所描述的基本信息外,还存在以下区别:

(1) 文本节点内容长度标准差:网页内容的文本节点内容长度通常差异较大,具有相对较高的标准差值;反之,网页噪音的文本节点内容长度通常差异较小,具有相对较低的标准差值;

(2) 文本节点内容标点符号数标准差:网页内容的文本节点内容标点符号数量通常差异较大,具有相对较高的标准差值;反之,网页噪音的文本节点内容标点符号数量通常差异较小,具有相对较低的标准差值.

基于以上所描述的统计信息,可按公式(11)对 TPF 特征进行扩展,得到扩展的综合特征 TPFE:

$$TPFE(p)=TPF(p)\cdot\sigma_{cs}\cdot\sigma_{ps} \tag{11}$$

其中,$TPF(p)$为标签路径 $p$ 的综合特征值,$\sigma_{cs}$ 为标签路径聚合的文本节点内容长度标准差,$\sigma_{ps}$ 为标签路径聚合的文本节点内容标点符号数量标准差.图 5 为 TPFE(扩展的综合特征)的直方图,在 TPF 的基础上引入$\sigma_{cs}$和$\sigma_{ps}$,观察图 5 发现:TPFE 明显地将内容节点和噪声节点的特征值区分开来,但是部分内容节点的特征值也受到一定的影响,导致其特征值偏小.与图 4 比较可以看出:在综合特征值柱状图中,内容节点的特征值远高于噪音节点的综合特征值.

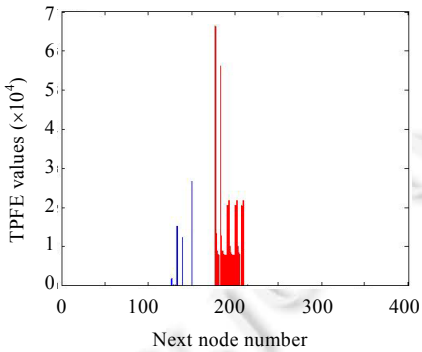



Fig.5 TPFE of each text node from the Web news of Freep

图 5 Freep 样例网页的 TPFE 柱状图

### 4.3 基于标签路径编辑距离的加权高斯平滑

根据标签路径特征值的定义和根据高特征值判定节点为内容节点的策略可知:基于标签路径特征值的 Web 新闻内容抽取方法适合于抽取 Web 新闻内容块中的长文本,易丢失一些重要的短文本内容.虽然我们的计算方法本身已提高了同一路径上的短文本的标签路径特征值,然而对于内容块中特别格式化的短文本,如内链接等,我们的方法还不能准确地识别.为了解决这个问题,我们提出了一种基于标签路径相似性的高斯平滑方法,以提高短文本的标签路径特征值.

图像处理中,标准高斯平滑方法的思想是相邻区域间像素的相互平均.与普通平滑方法的区别是,图像进行邻域平均时,邻域中不同位置的元素具有不同的权值.由于标准的高斯平滑方法是对连续属性的一个处理过程,不适用于我们的特征值直方图,因此,我们借鉴高斯平滑的优点及标签路径特征值直方图的特性,实现了一种应

用于一维离散特征上的高斯平滑操作.

当一个网页的标签路径特征值直方图 $H$ 被计算出来后,横坐标中文本节点的顺序是先序遍历的顺序,即,考虑到相邻节点物理属性的相似性,即,同为内容节点或者同为非内容节点.因同一内容块或噪音块中的文本节点的标签路径具有相似性,因此,我们还考虑到了相邻文本节点的标签路径的相似距离对数据的平滑作用.设 $r$ 为设置直方图 $H$ 的平滑窗口的参数,给定大小为 $2r+1$ 的平滑窗口,高斯内核的构建公式见公式(12):

$$k_i = e^{\frac{-i^2}{2r^2}}, -r \leqslant i \leqslant r \tag{12}$$

按照公式(13),对公式(12)中的高斯内核 $k$ 进行归一化处理:

$$k_i' = \frac{k_i}{\sum_{j=-r}^{r} k_j}, -r \leqslant i \leqslant r \tag{13}$$

最后,如公式(14)所述,根据高斯内核 $k'$对直方图 $H$ 进行加权卷积计算,得到直方图 $H'$:

$$H_i' = \sum_{j=-r}^{r} w_{ij} k_j' H_{i-j}, r \leqslant i \leqslant len(H) - r \tag{14}$$

其中,

$$w_{ij} = \begin{cases} 1, & p_i = p_j \\ \dfrac{1}{(dist(p_i, p_j))^{\alpha}}, & p_i \neq p_j \end{cases} \tag{15}$$

公式(15)中的 $p_i$,$p_j$ 分别表示解析树 $T$ 上的第 $i$ 个和第 $j$ 个文本节点的标签路径;$dist(p_i,p_j)$是标签路径 $p_i$,$p_j$ 的字符串编辑距离,反映了标签路径 $p_i$ 和 $p_j$ 的相似性(此处将标签路径中的一个标签视为一个字符).例如,标签路径“div.div.div.p”和标签路径“div.div.div.p.a”的字符串编辑距离是 1.$\alpha$为平滑调节参数,用于调节路径间的相似性对平滑的影响力.$w_{ij}$ 表示路径间的相似性对平滑的贡献,路径间相似性越大贡献越大,否则贡献越小.通常,内容块中内容节点的标签路径的相似性较大,而内容节点的标签路径与掺杂在内容块中的噪音节点的标签路径的相似性较小,所以 $w_{ij}$ 可用于提高内容块中特殊格式化的内容节点(如内链文本)的标签路径特征值,降低掺杂在内容块中的噪音节点(如推荐链接)的标签路径特征值.

图 6 为 STPFE(通过高斯平滑的扩展的综合特征值)的直方图.为了避免某些内容节点特征值过低问题,对 TPFE 进行高斯平滑.对比图 5 可以发现:图 5 中所示直方图的一些被特征格式化的短文本谷底在图 6 中被平滑了,一些内容节点的文本标签比特征值被拉高;同时,一些噪声节点的特征值被拉低.所以,STPFE 更能够有效地区分内容节点和噪声节点.

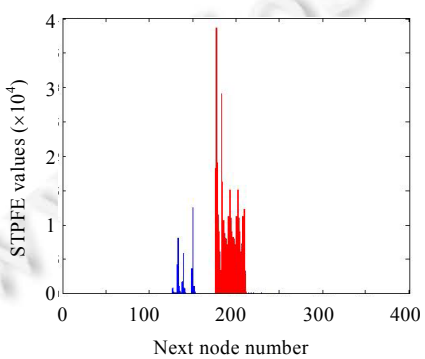



Fig.6 STPFE of each text node from the Web news of Freep

图 6 Freep 样例网页平滑的 TPFE 直方图

### 4.4 自适应阈值设置与内容抽取算法

基于标签路径特征值抽取内容的思路很直接,选定{*TPL*,*TPR*,*PPL*,*PPR*,*TPLR*,*PPLR*,*TPF*,*TPFE*}中的任一标

签路径抽取特征 $F$,设置一个阈值 $\tau$,设 $v$ 为解析树上的一个文本节点,当 $F(path(v))\geqslant\tau$ 时,将节点 $v$ 判定为内容节点;反之,将其判定为非内容节点.简而言之,具有高过阈值的特征值的节点是内容节点,具有低于阈值的特征值的节点是噪音节点.由此,抽取问题核心就成为确定最佳阈值以抽取完整的内容.

常见的阈值选取方法有中间值、平均值和标准差.标签路径特征直方图中,标签路径特征值较高的文本节点很少,标签路径特征值较低的文本节点较多,而且标签路径特征值直方图中的高低值相差较大,因此,标签路径特征值直方图的中间值偏向于低值,而平均值易被高值左右.因此,我们选择标准差作为设置阈值的基准,设置阈值 $\tau=\lambda\sigma$,其中,$\lambda$ 为阈值参数,$\sigma$ 为标准差.因此,阈值的设置问题转化为阈值参数 $\lambda$ 的设置问题.

参数 $\lambda$ 的设置,最简单的方法是设置为经验值,即,通过大量实验选择出使得平均抽取效果最佳的 $\lambda$ 值作为经验值.显然,这种方法缺乏通用性.随着数据集的变化,经验值 $\lambda$ 也要变化,且不能保证使得每个网页都达到很好的抽取效果.为了解决这个问题,我们研究了一种自适应阈值选择方法.

Web 新闻内容抽取的过程,实质上是将网页解析树中的文本节点划分为内容节点和噪音节点的过程.设置合适的阈值 $\tau$,将文本节点划分为内容节点和噪音节点两类,且使得类间距离最大.衡量类间距离的方法有很多种,包括类间最大距离法、均匀性度量法和最大类间方差等.根据标签路径特征直方图的特点,我们选择最大类间方差作为度量类间距离的标准.在标签路径特征值直方图的基础上,依据类间距离极大准则来确定分割阈值.

设网页解析树 $T$ 中有 $N$ 条互不重复度标签路径,具有 $L$ 个不同的标签路径特征值,这些特征值自小到大的排列顺序为 $\langle a_0,a_1,\ldots,a_{L-1}\rangle$.若特征值为 $a_i$ 的不重复标签路径个数为 $n_i$,标签路径特征值为 $a_i$ 的概率为 $p_i$,则有 $p_i=n_i/N$.显然,$\sum_0^{L-1}p_i=1$.

设置阈值 $\tau=\lambda\sigma$,把所有标签路径分为两类,分别是内容标签路径类 $A$ 和噪音标签路径类 $B$.$A$ 类中的标签路径对应的标签路径特征值 $\{a_0,a_1,\ldots,a_t\}$,$B$ 类中的标签路径对应的标签路径特征值为 $\{a_{t+1},a_{t+2},\ldots,a_{L-1}\}$.两类出现的概率分别为

$$p_A=\sum_{i=0}^{t}p_i,p_B=\sum_{i=t+1}^{L-1}p_i=1-p_A \tag{16}$$

$A$,$B$ 两类的标签路径特征值均值分别为

$$\overline{u}_A=\sum_{i=0}^{t}v_ip_i\Big/p_A,\overline{u}_B=\sum_{i=t+1}^{L-1}v_ip_i\Big/p_B \tag{17}$$

整个网页的标签路径特征值均值为

$$\overline{u}=\sum_{i=0}^{L-1}a_ip_i=p_A\overline{u}_A+p_B\overline{u}_B \tag{18}$$

由此,可以得到 $A$,$B$ 两类的类间方差为

$$g(\tau)=p_A(\overline{u}_A-\overline{u})^2+p_B(\overline{u}_B-\overline{u})^2 \tag{19}$$

使类间方差最大的分割,意味着错分概率最小,则使得类间方差最大的阈值 $\sigma^*$,即为所求的最佳阈值:

$$\tau^*=\arg\max_{\lambda\sigma}g(\tau),\tau=\lambda\sigma \tag{20}$$

在迭代搜索最佳阈值时,阈值参数 $\lambda$ 的搜索范围为[0~2.5],迭代步长为 0.01.搜索到最佳阈值后,遍历新闻网页的文本节点集合,若文本节点的标签路径特征值大于阈值时,判定节点为内容节点,并抽取节点内容;否则,判定为噪音节点.

需要说明的是:我们在不重复的标签路径节点基础上计算阈值,为的是避免重复的高特征值路径计算出过高阈值的倾向;另外,计算经过高斯平滑后特征值的阈值时,采用的是平滑前的特征值计算阈值,以避免局部平滑操作导致计算出过高阈值的倾向.

结合标签路径特征融合和阈值设置过程,图 7 给出了基于标签路径特征融合的内容抽取算法 CEPF.该算法第 1 步先将一个网页 *wp* 解析为树 $T$;第 2 步调用 *CalculateFeatures* 函数,遍历解析树 $T$,为每条标签路径计算它所有的标签路径特征值;第 3 步调用 *GetTextNodeSet* 函数获得所有的文本节点集合;第 4 步使用组合特征选择方法选择互不相关的特征构成特征子集 $F$;第 5 步调用特征融合方法获取每条标签路径的综合特征值 *TPFE*,并生成所有标签路径综合特征值直方图 $H$;第 6 步根据公式(20)计算最佳阈值 $\tau$;第 7 步~第 10 步,逐个判断每个文本

节点是否为内容节点,若是,则抽取该文本节点的内容;最后,第 11 步输出新闻网页的内容.计算树 $T$ 中所有标签路径特征值的过程主要操作是遍历树,时间复杂度为 $O(n)$,其中,$n$ 为树 $T$ 上节点的个数.设 $m$ 为标签路径特征的个数,用于自适应阈值求解的谱聚类方法需要求解相似度矩阵的前 $k$ 个特征值和特征向量,其时间复杂度为 $O(m^3)$;另外,谱聚类方法调用 $k$-means 算法得到 $k$ 个聚类,其时间复杂度为 $O(m\times k\times t)$,其中,$t$ 为迭代次数.因此,算法 CEPF 的时间复杂度为 $O(n+m^3+m\times k\times t)$.

在算法第 5 步也可生成 TPF 综合特征值直方图.记采用 TPF 特征值并未进行加权高斯平滑的 CEPF 算法版本为 CEPF-S,采用 TPFE 特征值并未进行加权高斯平滑的 CEPF 算法版本为 CEPF-E,采用 TPFE 特征值并进行加权高斯平滑的 CEPF 算法版本为 CEPF-ES.在不引起混淆的情况下,统称它们为 CEPF.

**Algorithm CEPF**.
输入:HTML 网页 *wp*,参数$\lambda$;
输出:网页内容 *content*.
1: 将网页 *wp* 解析为树 *T*;
2: *FS*←*CalculateFeatures*(*T*);
3: *NodeSet*←*GetTextNodeSet*(*T*);
4: *F*←*SelectGroupFeatures*(*FS*)
5: *H*←*GetHistogramByFusion*(*F*);
6: calculate the best $\tau$ with Eq.(20);
7: *content*←"";
8: **for** *i*=0 to *NodeSet*.size **do**
9: if (*H*(*path*(*NodeSet*[*i*]))≥$\tau$)
10: *content*←*content*+*c*(*NodeSet*[*i*]);
11: **output** *content*.

Fig.7 CEPF algorithm for Web content extraction
图 7 网页内容抽取算法 CEPF

## 5 抽取性能评估

本文评估的算法均为在线算法,无需训练,给定一组标注好的新闻网页语料,可全部做为测试集.以下实验的硬件配置为 Intel(R) Core(TM) i5-3470S CPU @ 2.90GHZ,2.90GHZ,8GB RAM,操作系统为 Windows 7 Professional x64,开发平台为 JDK 1.6.

### 5.1 实验数据集和抽取性能评估指标

实验中所使用的数据集主要分为以下 3 类:

(1) CleanEval DataSet:CleanEval 比赛[12]使用的数据,在 CleanEval 数据集中包含 932 个英文网页(EN)和 971 个中文网页(ZH);同时,CleanEval 数据集也是第 5.4 节对比实验对象之一——CETR 算法[13]使用的数据集;

(2) News DataSet:该数据集包括 13 个新闻网站的 3 450 个新闻网页.其中,数据集 NY Post,Suntimes,Techweb,Tribune,Nytimes,Freep,BBC,Reuters 也是对比实验 CETR 所使用的数据集,但原始数据集中每个网站仅包含 50 个网页,为了更好地验证算法的有效性,本文对原有数据集做了不同程度的扩展.此外,本文还增加了一个英文数据集(Yahoo!)和 3 个中文数据集(人民网、网易、新华网);

(3) Microblog DataSet:微博数据集.文献[14]指出,面向一个应用领域的信息抽取技术很难复用到另外一个不同的应用领域.虽然 CEPF 算法是针对新闻网页内容抽取而设计,但为了验证 CEPF 算法的可复用性以及标签路径特征在处理新型 Web 应用的优劣性,我们选取 900 个微博网页进行了实验与分析,该 900 个微博网页分别来自腾讯微博、新浪微博和搜狐微博,各 300 个测试网页.

表 2 给出了新闻和微博网站数据集的详细信息.

与 CETR 算法一样,采用精确度 Precision、召回率 Recall 和 $F$ 值作为 Web 新闻内容抽取的性能评估指标.将网页的抽取结果、网页手工标记的结果都看作是字符(英文是空格分隔得到的符号串,中文是汉字和字母)的

集合.记 $S_e$ 为抽取结果,$S_l$ 为手工标记结果集合,则 $S_e \cap S_l$ 表示抽取结果中应该被抽取的内容集合.精确度 Precision、召回率 Recall 和 $F$ 值这 3 个指标的定义见公式(21),其中,精确度 Precision 定义为抽取结果中应该被抽取的内容集合的字符个数和抽取结果集合的字符个数的比例,即,抽取结果中被正确抽取部分的比例;召回率 Recall 定义为抽取结果中应该被抽取的内容集合的字符个数和手工标记结果集合的字符个数的比例,即,应该抽取的内容被正确抽取的比例;$F$ 值是一个评价的综合指标.

$$P=\frac{|S_e \cap S_l|}{|S_e|}, R=\frac{|S_e \cap S_l|}{|S_l|}, F=\frac{2\times P\times R}{P+R} \tag{21}$$

**Table 2** Details on news sites and microblog sites

**表 2** 新闻和微博网站数据集的详细信息

| Name | News sites | Number of pages |
|---|---|---|
| NY Post | http://www.nypost.com/ | 300 |
| Suntimes | http://www.suntimes.com/ | 300 |
| Techweb | http://www.techweb.com/ | 250 |
| Tribune | http://www. chicagotribune.com/ | 300 |
| Nytimes | http://www.nytimes.com/ | 150 |
| Freep | http://www.freep.com/ | 50 |
| BBC | http://www.bbc.com/ | 300 |
| Reuters | http://www.reuters.com/ | 300 |
| Yahoo! | http://www.yahoo.com/ | 300 |
| sina | http://www.sina.com.cn/ | 300 |
| 人民网 | http://www.people.com.cn/ | 300 |
| 网易 | http://news.163.com/ | 300 |
| 新华网 | http://www.xinhuanet.com/ | 300 |
| 腾讯微博 | http://t.qq.com/ | 300 |
| 新浪微博 | http://weibo.com/ | 300 |
| 搜狐微博 | http://t.sohu.com/ | 300 |

### 5.2 平滑调节参数设置

平滑调节参数$\alpha$用于调节标签路径之间的相似性(即,标签路径的编辑距离)对标签路径特征值直方图平滑的贡献权重.由公式(14)和公式(15)知:$\alpha$值越小,标签路径的编辑距离对平滑的贡献越大;反之越小.图 8 给出了未平滑及$\alpha$=0,1,2,3,4,5 时,融合特征 TPFE 在抽取 Tribune 数据集上的性能比较,其中,窗口大小设为 3.

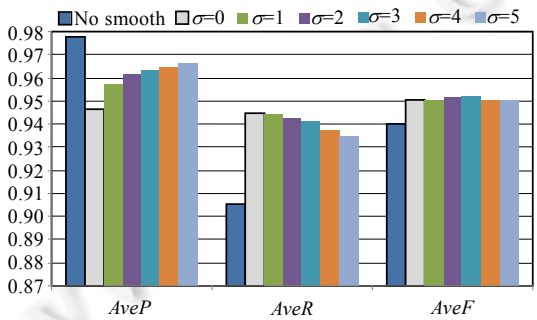



Fig.8 Extraction performance of CEPR-E for the Tribune corpora with different $\alpha$ settings

图 8 $\alpha$变化时,CEPR-E 在 Tribune 数据集上抽取性能比较

观察图 8 发现:在$\alpha$=3 时,融合特征 TPFE 在 Tribune 数据集上的抽取性能最好;在未平滑的情况下,TPFE 的抽取性能最差.相对于未平滑的情况,$\alpha$=0 的平滑虽然降低了抽取的精度,但同时也极大地提高了抽取的召回率,且综合抽取指标 *AveF* 值也得到了提高.

比较$\alpha$=0,1,2,3,4,5 时的融合特征 TPFE 的抽取性能发现:随着$\alpha$的增大,平均抽取精度增大,平均抽取召回率降低.在$\alpha$=0 的平滑时,由公式(14)和公式(15)知:不论标签路径间的相似度为多少,平滑权重 $w_{ij}$ 都为 1.例如,一对

相邻文本节点的标签路径分别为“div.div.div.p”和“div.div.div.p.a”,另一对相邻节点的标签路径分别为“div.div.div.p”和“div.div.div.ul.li.p”.第 1 对相邻节点的标签路径编辑距离为 1,平滑过程中的影响权重 $w_{ij}$ 为 1;而第 2 对相邻节点的标签路径编辑距离为 3,平滑过程中的影响权重同样为 1.这显然是不合理的.$\alpha$=1,2,3 的平滑使得融合特征 TPFE 的平均抽取性能得到了提高,$\alpha$=3 时的抽取性能最好.当$\alpha$>3 时,相对于未平滑时的抽取效果,其平滑后的抽取性能也有所提高,但并不明显.因为仅当两标签路径相同或编辑距离为 1 时,平滑权重为 1,有一定的平滑效果;否则,当 $dist(p_i,p_j)\geqslant 2$ 时,影响权重 $w_{ij}\leqslant 1/16$,平滑能力较弱.

综上所述,在本文实验中设置平滑调节参数$\alpha$为 3.

### 5.3 实验结果与分析

记基于标签特征系中单个特征抽取的方法为其使用的特征名.基于融合特征值的算法为 CEPF,如第 4.4 节所述,CEPF 有 CEPF-S,CEPF-E,CEPF-ES 这 3 个版本.表 3~表 5 给出各种方法在不同数据集上的抽取性能.表 3~表 5 同时也给出了和 CETR 方法的性能比较实验结果,详细的对比实验分析见第 5.4 节.

**Table 3** $F$_score for each algorithm in each source (the best sources are marked in bold) (%)

**表 3** 每种算法在各个数据集上的抽取 $F$ 值(优胜者标记为粗体) (%)

| DataSet | TPL | TPR | TPLR | PPL | PPR | PPLR | CEPF-S | CEPF-E | CEPF-ES | CETR | CEPR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CleanEval-En | 88.41 | **89.56** | 88.12 | 87.84 | 86.65 | 87.27 | 87.40 | 87.02 | 88.39 | 88.30 | 75.33 |
| CleanEval-Zh | **87.38** | 86.98 | 87.14 | 86.62 | 86.41 | 86.91 | 87.29 | 86.81 | 86.86 | 83.36 | 75.65 |
| NY Post | 70.76 | 54.87 | 72.49 | 65.88 | 71.38 | 66.49 | 78.58 | 89.28 | **90.04** | 58.19 | 81.02 |
| Freep | 81.06 | 49.45 | 83.68 | 84.19 | 78.43 | 86.19 | 88.11 | **89.87** | 87.79 | 70.36 | 86.00 |
| Suntimes | 71.70 | 70.73 | 76.04 | 85.20 | 93.51 | 88.41 | 92.91 | 93.70 | **94.08** | 82.20 | 85.90 |
| Techweb | 79.07 | 33.53 | 76.75 | 79.96 | 34.04 | 77.54 | 49.36 | 88.67 | **90.70** | 74.56 | 88.86 |
| Tribune | 90.47 | 94.76 | 92.75 | 93.76 | 93.79 | 94.12 | 94.12 | 94.00 | **95.21** | 89.83 | 90.32 |
| Nytimes | 90.83 | 89.66 | 91.30 | 91.00 | 91.06 | 91.36 | 91.69 | 91.33 | **92.31** | 91.14 | 86.91 |
| BBC | 84.72 | 80.86 | 86.09 | 82.59 | 81.69 | 83.86 | 85.06 | 88.89 | **89.53** | 72.76 | 80.13 |
| Reuters | 85.98 | 87.09 | 87.70 | 89.38 | 82.32 | 88.86 | 91.30 | 91.88 | **94.40** | 71.73 | 84.26 |
| Yahoo! | 88.55 | 85.52 | 89.65 | 89.05 | 85.76 | 90.54 | **91.31** | 89.09 | 89.33 | 82.06 | 84.96 |
| Sina | 77.22 | 93.91 | 78.24 | 77.55 | 96.01 | 79.77 | 95.16 | **97.41** | 96.92 | 73.99 | 90.63 |
| 人民网 | 72.85 | 74.15 | 73.79 | 81.53 | **92.20** | 78.49 | 89.62 | 91.41 | 89.27 | 86.23 | 85.32 |
| 网易 | 45.24 | 50.44 | 48.86 | 47.95 | 65.08 | 54.82 | 59.18 | 82.32 | 79.84 | 38.28 | **88.56** |
| 新华网 | 81.05 | 93.86 | 87.47 | 80.99 | 93.19 | 88.02 | 94.55 | 94.62 | **95.08** | 83.32 | 81.24 |
| 腾讯微博 | 80.30 | 82.67 | 81.50 | **84.61** | 68.27 | 83.17 | 74.91 | 73.52 | 83.72 | 79.36 | 17.75 |
| 新浪微博 | 82.03 | 65.54 | 81.23 | 72.57 | 74.70 | 70.94 | **82.32** | 81.70 | 79.38 | 57.99 | 18.88 |
| 搜狐微博 | 86.59 | 84.07 | 86.88 | 87.20 | 83.97 | 86.97 | 86.61 | 87.13 | **93.51** | 87.16 | 92.32 |
| Average | 80.23 | 75.98 | 81.65 | 81.55 | 81.02 | 82.43 | 84.42 | 88.81 | **89.80** | 76.16 | 77.45 |

**Table 4** Precision for each algorithm in each source (the best sources are marked in bold) (%)

**表 4** 每种算法在各个数据集上的抽取精度(优胜者标记为粗体) (%)

| DataSet | TPL | TPR | TPLR | PPL | PPR | PPLR | CEPF-S | CEPF-E | CEPF-ES | CETR | CEPR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CleanEval-En | 91.70 | 89.49 | 91.80 | 91.36 | 90.02 | 91.23 | 93.19 | 93.92 | 92.93 | 89.42 | **95.96** |
| CleanEval-Zh | 87.16 | 84.99 | 87.09 | 87.05 | 87.21 | 87.12 | 89.25 | **89.42** | 88.37 | 79.60 | 85.23 |
| NY Post | 60.46 | 38.38 | 62.93 | 57.43 | 59.08 | 57.89 | 74.15 | 92.01 | 92.10 | 42.68 | **98.28** |
| Freep | 70.61 | 33.32 | 74.72 | 75.54 | 65.33 | 78.85 | 82.13 | 85.24 | 80.66 | 57.38 | **86.75** |
| Suntimes | 60.79 | 55.39 | 65.93 | 81.80 | 90.85 | 85.32 | 94.60 | 96.09 | 94.73 | 70.82 | **98.10** |
| Techweb | 69.61 | 36.18 | 65.75 | 70.72 | 39.18 | 67.13 | 49.28 | 88.52 | 86.96 | 60.29 | **94.04** |
| Tribune | 89.58 | 95.05 | 94.68 | 96.82 | 92.73 | 97.74 | 97.92 | 97.77 | 96.32 | 84.99 | **99.43** |
| Nytimes | 94.69 | 89.12 | 95.61 | 95.10 | 92.92 | 96.06 | 97.31 | 96.67 | 96.13 | 88.98 | **99.73** |
| BBC | 83.93 | 74.88 | 85.39 | 83.80 | 79.31 | 84.76 | 88.11 | 93.51 | 91.75 | 60.11 | **97.83** |
| Reuters | 82.43 | 79.92 | 85.61 | 88.29 | 77.31 | 87.59 | 93.43 | 95.83 | 95.41 | 58.48 | **98.13** |
| Yahoo! | 86.30 | 79.01 | 89.15 | 86.60 | 79.29 | 90.21 | 95.26 | 95.04 | 93.04 | 72.67 | **97.83** |
| Sina | 63.90 | 89.92 | 65.28 | 64.92 | 94.83 | 67.71 | 93.60 | 98.31 | 96.29 | 59.09 | **98.57** |
| 人民网 | 58.58 | 59.54 | 59.82 | 71.31 | 88.30 | 66.54 | 86.55 | 90.36 | 84.47 | 77.04 | **95.11** |
| 网易 | 32.64 | 35.39 | 36.09 | 39.56 | 53.26 | 48.56 | 55.42 | 80.48 | 73.59 | 24.40 | **99.29** |
| 新华网 | 70.03 | 93.52 | 80.72 | 70.02 | 92.50 | 81.63 | 95.44 | **96.66** | 96.26 | 72.10 | 94.72 |
| 腾讯微博 | 99.28 | 99.22 | 99.28 | 97.77 | 98.57 | 98.83 | 99.27 | 99.33 | **99.36** | 66.37 | 94.25 |
| 新浪微博 | 78.77 | 48.83 | 77.26 | 61.50 | 66.41 | 59.36 | 79.86 | 79.19 | 66.90 | 41.81 | **86.84** |
| 搜狐微博 | 97.12 | 89.69 | 98.09 | 97.28 | 89.74 | 98.09 | 97.54 | **99.07** | 98.59 | 81.00 | 96.00 |
| Average | 76.53 | 70.66 | 78.62 | 78.72 | 79.82 | 80.26 | 86.79 | 92.64 | 90.21 | 65.96 | **95.34** |

**Table 5** Recall for each algorithm in each source (the best sources are marked in bold)(%)

**表 5** 每种算法在各个数据集上的抽取召回率(优胜者标记为粗体) (%)

| DataSet | TPL | TPR | TPLR | PPL | PPR | PPLR | CEPF-S | CEPF-E | CEPF-ES | CETR | CEPR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CleanEval-En | 85.35 | **89.64** | 84.71 | 84.58 | 83.52 | 83.65 | 82.29 | 81.07 | 84.28 | 87.20 | 68.00 |
| CleanEval-Zh | 87.60 | **89.07** | 87.19 | 86.20 | 85.62 | 86.70 | 85.41 | 84.35 | 85.40 | 87.48 | 72.67 |
| NY Post | 85.29 | **96.22** | 85.49 | 77.25 | 90.15 | 78.09 | 83.57 | 86.70 | 88.07 | 91.40 | 71.43 |
| Freep | 95.12 | 95.86 | 95.08 | 95.07 | **98.12** | 95.05 | 95.04 | 95.04 | 96.29 | 90.93 | 90.46 |
| Suntimes | 87.38 | 97.83 | 89.82 | 88.89 | 96.34 | 91.74 | 91.28 | 91.43 | 93.44 | **97.95** | 78.25 |
| Techweb | 91.52 | 31.25 | 92.17 | 91.98 | 30.09 | 91.77 | 49.43 | 88.81 | 94.77 | **97.68** | 87.31 |
| Tribune | 91.38 | 94.47 | 90.90 | 90.88 | 94.86 | 90.77 | 90.60 | 90.51 | 94.12 | **95.24** | 83.62 |
| Nytimes | 87.27 | 90.19 | 87.36 | 87.24 | 89.28 | 87.10 | 86.68 | 86.55 | 88.78 | **93.40** | 78.29 |
| BBC | 85.54 | 87.89 | 86.80 | 81.41 | 84.21 | 82.97 | 82.22 | 84.71 | 87.41 | **92.17** | 70.88 |
| Reuters | 89.85 | **95.67** | 89.89 | 90.49 | 88.02 | 90.17 | 89.26 | 88.24 | 93.41 | 92.75 | 77.01 |
| Yahoo! | 90.91 | 93.19 | 90.14 | 91.64 | 93.37 | 90.87 | 87.68 | 83.85 | 85.91 | **94.22** | 78.86 |
| Sina | 97.53 | 98.26 | 97.62 | 96.26 | 97.22 | 97.04 | 96.76 | 96.52 | 97.55 | **98.94** | 85.54 |
| 人民网 | 96.31 | **98.27** | 96.26 | 95.15 | 96.47 | 95.69 | 92.92 | 92.49 | 94.64 | 97.89 | 79.12 |
| 网易 | 73.68 | 87.80 | 75.65 | 60.86 | 83.65 | 62.94 | 63.49 | 84.24 | 87.24 | **88.76** | 81.46 |
| 新华网 | 96.18 | 94.21 | 95.44 | 96.03 | 93.89 | 95.49 | 93.69 | 92.66 | 93.93 | **98.68** | 73.94 |
| 腾讯微博 | 67.42 | 70.86 | 69.12 | 74.58 | 52.22 | 71.80 | 60.15 | 58.35 | 72.34 | **98.66** | 10.25 |
| 新浪微博 | 85.58 | **99.63** | 85.63 | 88.49 | 85.34 | 88.14 | 84.94 | 84.38 | 97.61 | 94.56 | 11.10 |
| 搜狐微博 | 78.12 | 79.12 | 77.97 | 79.02 | 78.89 | 78.11 | 77.87 | 77.75 | 88.92 | **94.33** | 89.53 |
| Average | 87.33 | 88.30 | 87.62 | 86.45 | 84.51 | 86.56 | 82.96 | 85.98 | 90.23 | **94.01** | 71.54 |

观察表 3~表 5 发现:在大部分数据集上,TPL 特征能取得较高的召回率,但抽取精度很低,以至于综合指标 *F* 值也不高,特别是在 NY Post、网易、Suntimes 等数据集上表现较差.这是因为 NY Post、网易、Suntimes 的新闻网页包含大量的导航、推荐链接和用户评论信息,以总文本长度为抽取标准的 TPL 特征误将它们识别为正文内容并抽取.但在数据集 CleanEval-Zh、BBC 以及新浪微博上取得了不错的抽取效果.实验结果表明:当网页的新闻内容的文本长度与噪音内容的文本长度接近时,标签路径特征 TPL 的抽取性能会受到影响.

以平均文本长度为抽取标准的标签路径特征 TPR 在 CleanEval-En、Sina、新华网等数据集上都取得了不错的抽取结果.特别是在 Sina 和新华网上,其抽取性能比 TPL 特征高 10 多个百分点.但是在数据集 Freep 和 Techweb 上,TPR 特征的抽取效果不是很好.这是因为 Freep 和 Techweb 数据集中网页的版权声明较长,且版权声明所在文本节点的 TPR 值远大于内容节点的 TPR 值,以至于部分新闻内容未被抽取而版权声明却被误抽.实验结果表明:当网页中包含较长的版权声明或类似的噪音信息时,标签路径特征 TPR 的抽取性能会受到影响.

标签路径抽取特征 TPLR 与 TPL 和 TPR 的不同之处在于:其利用了网页布局信息,能够识别和过滤文本长度较长的噪音信息.相比于 TPL 和 TPR 特征,TPLR 在 Freep,NY Post 上的抽取性能有所提高.TPLR 特征在所有数据集上的平均抽取性能也略优于 TPL 和 TPR.实验结果表明,TPLR 抽取特征能够有效地过滤文本长度较长的噪音信息.

与标签路径特征 TPL,TPR 和 TPLR 不同,标签路径特征 PPL,PPR 和 PPLR 用标点符号信息代替文本信息.虽然平均抽取性能优于基于文本信息的抽取特征,但在数据集 CleanEval、BBC 及新浪微博上的抽取性能并不是很理想.实验结果表明,文本信息不可完全被标点符号信息取代.

观察比较 CEPF 算法和单个标签路径特征的的抽取性能发现:CEPF 算法的平均抽取性能要优于任意一个标签路径特征,特别是在数据集 NY Post,Freep 和 Reuters 上.实验结果表明:基于标签路径特征融合的 Web 新闻抽取方法 CEPF 能够有效地融合标签路径特征系,且提高了抽取 Web 新闻的性能.

观察比较 CEPF-S 算法和 CEPF-E 算法的抽取性能发现:CEPF-E 算法的平均抽取性能优于 CEPF-S 算法,特别是在 Techweb 数据集和网易数据集上.实验结果表明,统计信息(文本节点内容长度标准差、文本节点内容标点符号数标准差)有利于提高标签路径特征区分 Web 新闻内容和噪音的能力.

观察比较 CEPF-ES 算法和 CEPF-E 算法的抽取性能发现:CEPF-ES 算法在大部分数据集上的抽取性能优于 CEPF-E 算法,特别是在数据集 Techweb、Yahoo!及搜狐微博上的抽取性能尤为突出.实验结果验证了基于标签路径编辑距离高斯平滑方法的有效性,平滑后的标签路径特征具有更强的抽取能力.

### 5.4 对比实验

面向开放环境下的 Web 新闻在线抽取问题,基于机器学习的抽取方法因需人工标注网页内容,在实际应用中不具有可操作性.然而还有一些方法能实时在线工作,这些算法包括 FE,KFE,BTE,DSC,ADSC,LQ,LP,CCB,CETR,CEPR.

本文选取的实验比较对象之一 CETR 是一种基于标签比对网页内容在线进行抽取的算法[13],该算法的基本思路是:首先,计算 Web 网页中每一行的标签比;然后,在此基础上进行聚类,再根据聚类结果提取网页的内容.由于基于一维标签比的中心内容提取算法没有考虑标签比数组的特征,其进行改进并构造了二维标签比,实现了基于二维标签比的网页内容提取算法 CETR.其实验结果表明:在大多数情况下,CETR 的抽取性能优于 FE,KFE,BTE,DSC,ADSC,LQ,LP,CCB 等当前大多数 Web 信息抽取算法.由于 CETR 使用标签比作为其抽取特征,使其具备抽取多种语言网页的能力.另外,CETR 是一种无监督的 Web 信息抽取方法.

另一实验比较对象是CEPR算法[15],也是本文部分作者合作研究的一项成果.CEPR算法设计了一种文本标签路径比特征及其扩展特征,研究了基于文本标签路径直方图区分内容和非内容的阈值方法,提出了基于路径编辑距离的加权高斯平滑方法,有效地解决了短文本抽取问题和新闻内容中非新闻内容过滤问题.CEPR 是一种快速、通用、无需训练的网页内容抽取算法,可抽取多种来源、多种风格、多种语言的 Web 信息网页.在 CleanEval 测试数据集上的实验结果表明:大多数情况下,CEPR 方法优于 CETR 等抽取方法.本文设计了标签路径特征系,以从不同视角区分网页内容和噪音.在特征相似性分析的基础上,提出了一种基于组合特征选择的特征融合策略,并设计了基于融合特征的 Web 新闻内容抽取方法 CEPF.因此,CEPF 是一项在 CEPR 算法基础上的进一步深入和拓展研究的工作.原始的 CEPR 算法采用的是经验阈值的方法,而 CEPF 是一种自动阈值方法,为了比较方便起见,我们使用的是 CEPR 算法自动设置阈值的版本,并且 CEPR 的自动设置阈值方法和 CEPF 的自动设置阈值方法相同.

CleanEval 竞赛数据集包括了测试和评估两个部分,因在线抽取问题不需训练,Weninger 等人在实验中将这两部分数据集合二为一,形成了一份数据集[13].CleanEval 竞赛中取得优胜的方法[16]是一种基于模板检测的方法,难以解决面向开放环境下的 Web 新闻在线抽取问题.因此,文献[13,15]和本文均未与该方法做对比实验.

表 3~表 5 也给出了 CETR 方法的抽取效果,可以发现:CETR 方法在多个数据集上的召回率 Recall 优于 CEPF 方法,抽取精度 Precision 和 $F$ 值低于 CEPF 方法.实验结果说明:CETR 方法能尽量将应该被抽取的内容抽取,但抽取结果中被正确抽取部分的比例较低,说明 CETR 方法抽取的结果中噪音内容比较多,其误抽率较高,导致其综合评价指标 $F$ 值低于本文所提 CEPF 方法.除 TPR 特征外,基于其他标签路径特征的平均抽取性能都要优于 CETR 方法,尤其是 CEPF-ES 方法,在 NY Post、Reuters、Sina 和网易数据集上抽取性能 $F$ 值分别比 CETR 方法高 31.85%,22.67%,22.92%和 41.56%,平均抽取性能 $F$ 值高 13.64%.

表 5 给出了各算法在各个数据集上的抽取召回率测试,CEPR 方法在大多数数据集上的精确度 Precision 优于 CEPF 方法,召回率 Recall 和 $F$ 值低于 CEPF 方法.实验结果说明:CEPR 方法的能尽量做到抽取的内容是应该被抽取的内容,但抽取结果中应该被抽取的内容的比例较低,说明 CEPR 方法漏抽取了网页内容,其漏抽率较高,导致其综合评价指标 $F$ 值低于 CEPF 方法.

CEPR 方法的抽取特征 ETPR[15]是在 TPR 特征的基础上设计的扩展特征.从表 3 可以发现:CEPR 方法的综合抽取性能 $F$ 值仅比 TPR 和 CETR 方法好,比 TPL,TPLR,PPL,PPR,PPLR,CEPF,CEPF-E,CEPF-ES 方法都差.其中,CEPF-ES 是综合抽取性能最好的方法.除了在网易数据集上,CEPR 比 CEPF 方法的抽取性能 $F$ 值好之外,在其他所有数据上,CEPF 方法的抽取性能均超过 CEPR,平均抽取性能 $F$ 值高 12.35%.实验结果表明:我们设计的标签路径特征系具有较好的抽取性能,并且我们设计的特征融合算法能有效提升标签路径特征系的抽取性能.

在时间性能方面,在 CleanEval DataSet,News DataSet 和 Microblog DataSet 上的实验结果表明:CEPF 算法平均抽取一个网页的时间为 408.9ms,CETR 算法平均抽取一个网页的时间为 395.03ms,CEPR 算法平均抽取一个网页的时间为 375ms,三者抽取时间性能相当.另外,CEPF,CEPR 方法采用的是网页解析树中标签路径特征,能够反映网页的结构特征,比 CETR 方法采用的行标签比特征具有更好的解释性.与 CETR,CEPR 方法一样,CEPF 方

法也具有抽取多种语言网页的能力,且都是无监督的抽取方法.

虽然 CEPF,CEPR,CETR 算法是针对新闻网页内容抽取而设计的,为了检测不同 Web 新闻抽取算法复用到其他应用领域的能力,我们选取了与 Web 新闻类似的 3 个微博数据集分别应用不同的算法进行内容抽取.从表 3 可以发现:CEPF 在微博数据集上仍能表现出较好的内容抽取能力,其次是 CETR,最差的是 CEPR.这说明 CEPF 算法有一定的被复用的能力.另外,我们也尝试了其他性质的数据集,包括论坛、电子商务数据,发现 CETR,CEPR,CEPF 算法都处于失效的状态.对于论坛数据,我们需要采用面向社交网络应用的信息抽取方法;对于电子商务数据,我们需要采用面向 Web 表格抽取的方法.

综上所述,CEPF 抽取方法优于 CETR,CEPR 抽取方法.

## 6 相关工作

Web 信息抽取(Web information extraction,简称 WIE)工作可以追溯到结构化和半结构化等异构数据源整合研究.利用包装器(wrapper)封装单个信息源上的信息存储操作,构建异构信息整合系统,使得用户可以通过统一的查询接口实现多个异构信息源上的操作.其中,Web 页面信息抽取技术是包装器的关键[17].除了强烈的应用需求外,美国国家标准技术研究所(NIST)组织的自动内容抽取(automatic content extraction,简称 ACE)评测会议推动了该领域研究的进展,这项评测从 1999 年 7 月开始酝酿,2000 年 12 月正式开始启动,从 2000 年~2007 年已经举办过多次评测[18].

面向传统的 HTML 页面开展信息抽取已有十几年的相关研究,可按以下应用领域进行分类[14]:上下文敏感广告、客户服务、数据库逆向工程、互联网应用逆向工程、商业智能和竞争智能、Web 应用流程集成和频道管理、Web 应用功能测试、比较购物、Mashup 信息聚合、观点挖掘、引用数据库的构建与维护、提高 Web 网页的易读性、内容抽取、Web 网页博物馆等.文献[14]也指出,面向一个应用领域的信息抽取技术很难复用到另外一个不同的应用领域.近年来,随着 Web 新闻聚合、舆情分析、Web 新闻事件追踪[19]等研究和应用工作的开展,针对网页新闻标题、正文等特定属性的抽取研究被广泛研究[20].本文的研究重点是 Web 内容抽取的 Web 新闻抽取领域,以下将重点阐述 Web 新闻抽取、相关抽取技术和抽取特征这 3 个方面的工作.

从抽取技术的特征来看,手工构建一个包装器是最简单和直接的方法.使用 Java,Perl 这样通用的程序设计语言或用户自行设计的特定语言,用户面向特定的网站构建包装器.手工构造的 WIE 系统包括 TSIMMIS[21],W4F[22]和 XWRAP[23]等,这类系统需要用户具有编程的能力、熟悉数据源和输出结果的格式、理解各个数据项的语义以及抽取规则的内涵.其优点是能较好地解决特定领域的问题,缺点是自动化程度不高,构造代价高,难以扩展和推广,当数据源和输出结果的格式和数据项的语义发生变化时,需手工修改系统.鉴于手工抽取系统存在以上问题,WIE 系统研究和发展的趋势之一是在保证抽取精度的同时提高系统的自动化程度[14].将监督学习[24,25]、半监督学习[26]和无监督学习[2]的理论与方法引入 Web 信息抽取领域,通过数据挖掘和机器学习的方法自动配置抽取系统规则库,致力于提高抽取的精度和减少标注语料的规模和代价等工作.

大多数方法假定被抽取的网页对象有特定的结构或共享某种结构,Bar-Yossef 和 Rajagopalan 针对这种现象,首次提出了从网页中提取模板的问题[27].然而,基于模板归纳的方法通常会存在以下问题:首先,已有的包装器难以有效抽取未知模板生成网页的内容,如果需要抽取这些网页的内容,需要针对该模板构建新的包装器;其次,抽取如新闻这样特定领域的内容,需要对成百上千的数据源构建包装器,任何模板的变化将导致包装器失效,产生和在线维护这些模板的代价极高.

在 Web 新闻抽取领域,Web 信息抽取算法面临以下问题:首先,相似的内容可能在不同的网站被不同的标签表示,例如不同的网站分别用〈B〉,〈H1〉,〈FONT〉,〈SPAN〉等表示标题;其次,从不同 DOM 树中产生的模板可能具有完全不同的拓扑结构.由于不同网站的结构缺少一致性,很难推导一个基于 DOM 树结构的传统包装器.基于视觉特征的包装器可在一定程度上解决这种不一致性.Cai 等人于 2003 年引入了基于视觉的页面分块技术[28],采用启发式的方法将页面分解为视觉块树.由于在计算视觉特征之前仍需构建 DOM 树,一般而言,基于视觉的页面解析方法较其他方法要占用更多的计算资源.Zheng 等人[29]将每个 DOM 树节点视为一个矩形视觉块,从一系

列视觉特征派生出一个复合特征,通过机器学习的方法为新闻站点生成包装器.然而,它使用手工标注的数据作为训练集,手工标注的过程是一个费时费力的工作,另外,如果训练集不够大,将导致抽取结果的不准确.Wang 等人[30]提出了一些新颖的视觉特征,通过机器学习的方法从一个网站的训练数据中建立训练模型,面向 12 个网站进行测试,取得了较好的抽取效果,表明视觉特征具有较好的泛化性能.然而,由于使用一些新闻网页布局的假定以及监督学习方法限制了该工作的使用范围.Sun 等人于 2011 年提出一种基于文本密度和特征从 DOM 树中抽取 Web 内容的方法[31],这种方法是一种内容区域抽取[32]方法,没有过滤内容中的噪音.

随着 Web 的发展,一些被称为 Web 2.0 的技术元素被大量应用[14],如 RSS,Mashup,Ajax,Deep Web,CSS 技术,使得传统的 Web 信息抽取技术面临一些新的机会与挑战.其中,RSS 和 Mashup 技术能够获取更有结构化的数据,使得信息抽取问题变得更加容易;AJAX 可以使网页实现异步更新,因为通过 AJAX 技术获得的网页和传统的 HTML 网页没有区别,所以,AJAX 技术对 Web 信息抽取过程而言没有什么影响;Deep Web 须通过动态请求产生网页信息,这种变化使得使用网页先验知识设计的信息抽取算法失效;现代 Web 标准中典型的应用模式是 CSS+XHTML,摒弃了 HTML 4.0 中的表格定位方式.其优点是网站设计代码规范,加快了网站访问速度,并使符合 Web 标准的网站对于用户和搜索引擎更加友好.但是,这种变化使得使用 HTML 网页中表格特征的信息抽取算法失效.

更进一步地,面向 Web 大数据这样一个开放的环境和领域,目前的 Web 信息抽取技术受到更大的挑战[33].例如在 CleanEval[12]任务中,每个网站仅有少量甚至一个网页.目前,大多数方法面向开放环境下的信息抽取问题时通常会失效,这是因为这些方法通常假设有以下假设:网页具有特定的结构(例如含有〈table〉标签)、主要内容是相邻的、网页具有预知的格式、甚至整个网站是根据模板生成的.基于这些假设,抽取系统可以对几个网页进行分析,试图归纳出模板以丢弃菜单和其他重复块.这是因为现代网页大量采用 CSS 技术,网页中主要包括〈div〉标签,无需被层次化地组织(就像基于〈table〉标签的设计);其次,现在很多网页是基于 Deep Web 技术动态自动生成的,它往往是难以根据先验知识进行分析;另外,当我们没有或者只有少量未标注的网页,已有方法难以有效地推导模板以进行信息抽取.因此,需要更通用的方法解决这类抽取问题.

然而,还有一些方法能实时在线工作.实时意味着无需对网页进行预处理或预先知道它们的结构,在线意味着能适应任何网页.其中之一是 CCB[34],该技术使用 CCV(content code vector)描述文档中的元素(标签、字符/单词符号),每个元素按照其属性初始定义为内容元素或代码元素.在 CCV 描述的基础上,根据环绕元素的内容元素和代码元素的数量迭代计算其内容代码比(content code ratio).最后,根据 CCV 中 CCR 信息判断文档中的内容块.另一个可实现在线抽取的算法是 CETR[13],该方法基于标签比对网页内容进行抽取.其实验结果表明:在大多数情况下,CETR 的抽取性能优于当前多数 Web 信息抽取算法.

上述在线方法的主要问题是基于网页中字符或行进行计算,以至于完全忽略网页的结构.大量的实例研究表明,网页中的内容、噪音和其 DOM 树路径有潜在的联系.W4F[22]是一个生成 Web 包装器的 Java 工具包,生成的包装器包括 3 个独立组件:检索组件、抽取组件和映射组件,其中,抽取组件的生成依赖于用户根据 HEL 语言书写的抽取规则.将一个 HTML 页面解析成一棵 DOM 树,抽取组件根据抽取规则的路径表达式定位抽取对象,得到粗粒度的抽取结果,再根据抽取规则中的 match 和 split 操作获得细粒度的抽取结果,其中:match 操作获得匹配的字符串内容,过滤掉不匹配的部分;split 操作将一个字符串按照指定的分隔符分解为一个字符串列表.XWrap[23]通过和用户的交互,生成一个输出为 XML 形式的包装器.其中,XWrap 的信息抽取组件负责推导抽取规则细分为 3 个步骤:① 识别网页中用户感兴趣的数据区域;② 识别重要的语义符号串及其在网页解析树中的路径和节点位置;③ 识别数据区域中相关数据的层次结构.上述 3 个步骤均产生相应的抽取规则.XWrap 无需用户编写抽取规则,然而其需要用户理解 HTML 解析树,识别表格中分隔行和列的 HTML 标签等专业知识.另外,XWrap 没有使用特定的学习算法,主要依靠网页 DOM 解析树路径进行对象定位.Vertex[35]是基于 XPath 的抽取规则学习的大规模 Web 信息抽取系统,是 Yahoo!开发的用于从模板生成网页中抽取结构化数据表的包装器软件系统,已部署到实际应用的产品,抽取了 200 多个网站共 2.5 亿条以上的数据记录.文献[36]基于 XPath 进行扩展,设计了一种自动爬取和抽取语言 OXPath,以解决面向 Deep Web 的大规模数据处理问题.PPWIE[25]是

Wu 等人提出的一种基于路径模式挖掘的信息抽取方法,以解决新闻网页过滤与总结系统[20]内容抽取模块研发过程中的路径模式自动获取问题.将 HTML 网页解析成 DOM 树,设计了可视化用户标注工具,通过在树节点中添加(属性,值)对实现训练数据的标注工作.提取训练数据集中已标注的正例节点和反例节点的路径,设计区分路径模式挖掘方法,获取泛化路径模式.以 CNN 英文网站和 China News 中文网站上随机选取的网页为实验数据集,实验结果表明:通过合理设置容噪阈值,基于路径模式挖掘的新闻网页内容抽取方法可达到 98%以上的宏平均 $F$ 值,同时也验证了路径模式具有语言不敏感性的特点.CEPR 算法[15]设计了一种文本标签路径比特征及其扩展特征,研究了基于文本标签路径直方图区分内容和非内容的阈值方法,在 CleanEval 测试数据集上的实验结果表明:大多数情况下,CEPR 方法优于 CETR 等抽取方法.以上工作表明,标签路径特征可有效地应用于 Web 信息抽取任务.

考虑到从不同视角设计的特征有助于提升抽取性能[37],本文设计的 CEPF 算法充分利用 DOM 树节点的标签路径信息,设计了标签路径特征系;在此基础上,探索研究基于 DOM 树路径的 Web 新闻网页在线内容抽取方法.该方法不依赖于任何特定的 HTML 标记,不假定网页有任何特定的结构,不假定网页之间共享特定的结构,能适应开放环境的 Web 新闻内容抽取问题.

## 7 结论和展望

从多源、海量、异构、价值密度低的 Web 新闻网页中抽取 Web 新闻内容,是 Web 大数据处理与应用中的一项重要任务.本文根据 Web 新闻网页中的内容分布和标签路径之间存在潜在联系的现象,从网页内容中的文本信息、标点符号信息、修饰信息等视角,设计了用于 Web 新闻内容抽取的标签路径特征系.针对在抽取不同 Web 新闻网页时标签路径特征系中的特征存在不同的差异和联系的现象,基于高斯相似度函数度量特征间的相似性,提出了一种基于谱聚类的组合特征选择方法,并设计了一种基于标签路径特征融合的 Web 新闻内容抽取算法 CEPF.在 CleanEval 等数据集上的实验表明:CEPF 是一种通用的、无需训练的 Web 新闻内容抽取算法,可抽取多种来源、多种风格、多种语言的 Web 新闻网页,在抽取精度、召回率、$F$ 值等方面均优于现有的 CETR 等抽取方法;并且算法的时间开销较小,简单实用,易于实现,适用于开放环境下的 Web 新闻内容的在线抽取.

但是,基于标签路径特征融合的 Web 新闻内容抽取算法 CEPF 还有很多改进空间:首先,在设计标签路径特征系时,我们仅考虑了文本、标点符号等方面的信息,设计的特征存在多样性不足的问题,有可能存在其他标签路径抽取特征,有更好的抽取性能和可解释性;其次,本文仅考虑了抽取新闻内容,如果将网页分块技术、噪音过滤技术和内容抽取技术有机结合,有可能会进一步提高抽取的精度.另外,本文所研究的 Web 新闻内容抽取方法都是针对网页中的文本信息的抽取,对视频、图片等形式的网页内容的抽取没有提出解决方案.希望以后可以通过标签路径特征或树结构特征对文本信息、视频和图片信息进行综合处理,提高抽取方法对 Web 信息多样性的适应能力.

Web 内容抽取问题已有十多年的研究,但面向开放环境下的 Web 新闻在线抽取问题也是近年才逐渐被研究者关注.Web 新闻聚合、舆情分析、Web 新闻事件追踪等实际应用系统需要精准的在线 Web 新闻抽取工作的支持.在实际的工程应用中,已有的 Web 新闻在线抽取方法主要抽取粗粒度的新闻内容,面向精准的细粒度结构化新闻的在线抽取任务表现仍不是很理想.精准地在线抽取 Web 新闻的标题、发布时间、来源、内容等结构化信息,仍需做进一步的深入研究.

**References**:

[1] Gibson D, Punera K, Tomkins A. The volume and evolution of web page templates. In: Proc. of the Special Interest Tracks and Posters of the 14th Int'l Conf. on World Wide Web (WWW 2005). New York: ACM Press, 2005. 830−839. [doi: 10.1145/1062745.1062763]

[2] Reis DC, Golgher PB, Silva AS, Laender AHF. Automatic Web news extraction using tree edit distance. In: Proc. of the 13th Int'l Conf. on World Wide Web (WWW 2004). New York: ACM Press, 2004. 502−511. [doi: 10.1145/988672.988740]

[3] Parapar J, Barreiro Á. An effective and efficient Web news extraction technique for an operational NewsIR system. In: Proc. of the 12th Conf. of the Spanish Association for Artificial Intelligence, Vol.2. 2007. 319−328.

[4] Luxburg U, Belkin M, Bousquet O. Consistency of spectral clustering. The Annals of Statistics, 2008,36(2):555−586. [doi: 10.1214/009053607000000640]

[5] Fiedler M. Algebraic connectivity of graphs. Czechoslovak Mathematical Journal, 1973,23(2):298−305.

[6] Wu Z, Leahy R. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1993,15(11):1101−1113. [doi: 10.1109/34.244673]

[7] Sarkar S, Soundararajan P. Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(5):504−525. [doi: 10.1109/34.857006]

[8] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(8): 888−905. [doi: 10.1109/34.868688]

[9] Wang XT. The research of multi-source information fusion method [MS. Thesis]. Harbin: Harbin Engineering University, 2012 (in Chinese with English abstract).

[10] Hall DL, Llinas J. An introduction to multisensor data fusion. Proc. of the IEEE, 1997,85(1):6−23. [doi: 10.1109/5.554205]

[11] Khaleghi B, Khamis A, Karray FO, Razavi SN. Multisensor data fusion: A review of the state-of-the-art. Information Fusion, 2013, 14(1):28−44. [doi: 10.1016/j.inffus.2011.08.001]

[12] Baroni M, Chantree F, Kilgarriff A, Sharoff S. Cleaneval: A competition for cleaning Web pages. In: Proc. of the 6th Int'l Conf. on Language Resources and Evaluation (LREC 2008). 2008. 638−643.

[13] Weninger T, Hsu WH, Han J. CETR: Content extraction via tag ratios. In: Proc. of the 19th Int'l Conf. on World Wide Web (WWW 2010). 2010. 971−980. [doi: 10.1145/1772690.1772789]

[14] Ferraraa E, Meob PD, Fiumarac G, Baumgartnerd R. Web data extraction, application and techniques: A survey. Knowledge-Based Systems, 2014,70(C):301−323. [doi: 10.1016/j.knosys.2014.07.007]

[15] Wu G, Li L, Hu X, Wu X. Web news extraction via path ratios. In: Proc. of the 22nd ACM Int'l Conf. on Information and Knowledge Management (CIKM 2013). Burlingame: San Francisco Airport Marriott Waterfront, 2013. 2059−2068. [doi: 10.1145/2505515.2505558]

[16] Marek M, Pecina P, Spousta M. Web page cleaning with conditional random fields. In: Proc. of the 3rd Web as Corpus (WAC3). 2007.

[17] Chang CH, Kayed M, Girgis MR, Shaalan KF. A survey of Web information extraction systems. IEEE Trans. on Knowledge and Data Engineering, 2006,18(10):1411−1428. [doi: 10.1109/tkde.2006.152]

[18] Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassel S, Weischedel R. The automatic content extraction (ACE) program —Tasks, data, and evaluation. In: Proc. of the 4th Int'l Conf. on Language Resources and Evaluation (LREC 2004). 2004. 837−840.

[19] Zhang C, Soderland S, Weld DS. Exploiting parallel news streams for unsupervised event extraction. Trans. of the Association for Computational Linguistics, 2015,3:117−129.

[20] Wu X, Wu GQ, Xie F, Zhu Z, Hu XG, Lu H, Li H. News filtering and summarization on the Web. IEEE Intelligent Systems, 2010, 25(5):68−76. [doi: 10.1109/mis.2010.11]

[21] Hammer J, McHugh J, Garcia-Molin H. Semistructured data: The TSIMMIS experience. In: Manthey R, Wolfengagen V, eds. Proc. of the 1st East-European Conf. on Advances in Databases and Information systems (ADBIS'97). Swinton: British Computer Society, 1997. 1−8.

[22] Sahuguet A, Azavant F. Building intelligent web applications using lightweight wrappers. Data and Knowledge Engineering, 2001, 36(3):283−316. [doi: 10.1016/s0169-023x(00)00051-3]

[23] Liu L, Pu C, Han W. XWRAP: An XML-enabled wrapper construction system for Web information sources. In: Proc. of the 16th IEEE Int'l Conf. on Data Engineering (ICDE 2000). 2000. 611−621. [doi: 10.1109/icde.2000.839475]

[24] Kushmerick N, Weld DS, Doorenbos R. Wrapper induction for information extraction. In: Proc. of the 15th Int'l Conf. on Artificial Intelligence (IJCAI'97). 1997. 729−735.

[25] Wu G, Wu X. Extracting Web news using tag path patterns. In: Proc. of the 2012 IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI 2012). 2012. 588−595. [doi: 10.1109/WI-IAT.2012.107]

[26] Hogue A, Karger D. Thresher: Automating the unwrapping of semantic content from the World Wide Web. In: Proc. of the 14th Int'l Conf. on World Wide Web (WWW 2005). New York: ACM Press, 2005. 86−95. [doi: 10.1145/1060745.1060762]

[27] Bar-Yossef Z, Rajagopalan S. Template detection via data mining and its applications. In: Proc. of the 11th Int'l Conf. on World Wide Web (WWW 2002). 2002. 580−591. [doi: 10.1145/511446.511522]

[28] Cai D, He X, Wen JR, Ma WY. Block-Level link analysis. In: Proc. of the 27th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2004). 2004. 440−447. [doi: 10.1145/1008992.1009068]

[29] Zheng S, Song R, Wen JR. Template-Independent news extraction based on visual consistency. In: Columbia B, Cohn A, eds. Proc. of the 22nd National Conf. on Artificial Intelligence (AAAI 2007), Vol.2. Vancouver: AAAI Press, 2007. 1507−1512.

[30] Wang J, Chen C, Wang C, Pei J, Bu J, Guan Z, Zhang WV. Can we learn a template-independent wrapper for news article extraction from a single training site? In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2009). 2009. 1345−1354. [doi: 10.1145/1557019.1557163]

[31] Sun F, Song D, Liao L. DOM based content extraction via text density. In: Proc. of the 34th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2011). New York: ACM Press, 2011. 245−254. [doi: 10.1145/2009916.2009952]

[32] Sleiman HA, Corchuelo R. A survey on region extractors from web documents. IEEE Trans. on Knowledge and Data Engineering, 2013,25(9):1960−1981. [doi: 10.1109/tkde.2012.135]

[33] Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. IEEE Trans. on Knowledge and Data Engineering, 2014,26(1):97−107. [doi: 10.1109/TKDE.2013.109]

[34] Gottron T. Content code blurring: A new approach to content extraction. In: Proc. of the 19th Int'l Conf. on Database and Expert Systems Application (DEXA 2008). Washington: IEEE Computer Society, 2008. 29−33. [doi: 10.1109/dexa.2008.43]

[35] Gulhane P, Madaan A, Mehta R, Ramamirtham J, Rastogi R, Satpal S, Sengamedu SH, Tengli A, Tiwari C. Web-Scale information extraction with vertex. In: Proc. of the 27th IEEE Int'l Conf. on Data Engineering (ICDE 2011). 2011. 1209−1220. [doi: 10.1109/icde.2011.5767842]

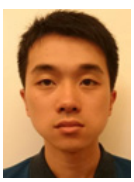
[36] Furche T, Gottlob G, Grasso G, Schallhart C, Sellers A. OXPath: A language for scalable data extraction, automation, and crawling on the deep Web. The VLDB Journal, 2013,22(1):47−72. [doi: 10.1007/s00778-012-0286-6]

[37] Peters ME, Lecocq D. Content extraction using diverse feature sets. In: Proc. of the 22nd Int'l Conf. on World Wide Web Companion (WWW 2013). 2013. 89−90.

**附中文参考文献**:

[9] 王新涛.多源信息融合方法研究[硕士学位论文].哈尔滨:哈尔滨工程大学,2012.

**吴共庆**(1975－),男,安徽岳西人,博士,副教授,CCF 会员,主要研究领域为 Web 智能,数据挖掘.



**胡骏**(1990－),男,硕士生,主要研究领域为 Web 数据集成,数据挖掘.

**李莉**(1989－),女,硕士生,主要研究领域为 Web 数据集成,数据挖掘.

**徐喆昊**(1989－),男,硕士生,主要研究领域为 Web 数据集成,数据挖掘.

**刘鹏程**(1991－),男,硕士生,主要研究领域为 Web 数据集成,数据挖掘.

**胡学钢**(1961－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据挖掘,人工智能,算法设计与分析.

**吴信东**(1963－),男,博士,教授,博士生导师,主要研究领域为数据挖掘,专家系统,万维网信息处理.