第47卷第4期 中国科学技术大学学报 Vol.47,No.4
2017年4月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Apr. 2017

文章编号:0253-2778(2017)4-0290-07

基于结构和文本特征的网页分类技术研究

顾 敏[1,2],郭 庆[1],曹 野[1],朱 峰[1],顾彦慧[1],周俊生[1],曲维光[1,2]

(1.南京师范大学计算机科学与技术学院,江苏南京 210023;2.福建省信息处理与智能控制重点实验室,闽江学院,福建福州 350121)

摘要:Web网页中含有丰富的信息资源,通过网页分类可以更好地对其内容进行抽取和管理,方便用户阅读.针对网页复杂的结构信息和丰富的文本内容,提出了一种基于网页文本和结构的网页分类方法,利用众创相关网页的结构特点和文本信息,选择联合特征和原子特征相结合的方法进行分类.实验表明,这种方法有一定的可行性,且比单一使用文本信息进行分类的方法具有更高的正确率和召回率.

关键词:网页分类;朴素贝叶斯;原子特征;联合特征

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.04.002

引用格式:顾敏,郭庆,曹野,朱峰,等.基于结构和文本特征的网页分类技术研究[J].中国科学技术大学学报,2017,47(4):290-296.

GU Min, GUO Qing, CAO Ye, et al. Research on web page automatic categorization based on structural and text information[J]. Journal of University of Science and Technology of China, 2017,47(4):290-296.

Research on web page automatic categorization based on structural and text information

GU Min[1,2], GUO Qing[1], CAO Ye[1], ZHU Feng[1], GU Yanhui[1], ZHOU Junsheng[1], QU Weiguang[1,2]

(1. *School of Computer Science and Technology, Nanjing Normal University, Nanjing* 210023, *China*;
2. *Fujian Province Key Laboratory of Information Processing and Intelligence Control, Minjiang University, Fuzhou* 350121, *China*)

Abstract: Since web pages contain abundant information resources, a better extraction and management of the information can be achieved through web page categorization. Considering the complex structure and abundant text information, a method was proposed for web page categorization based on the structure and text. The method of combining joint features and atomic features was employed to classify the web pages. The experiment result shows that the proposed method is feasible to some extent and has a higher precision and recall rate than using text information only.

Key words: web page classification; naïve Bayes; atomic feature; joint feature

0 引言

随着“大众创业,万众创新”理念的发展,众创相关的网页数量也呈指数增长.如何从众创相关的网页中获取有价值的信息成为一个值得考虑的问题.网页分类就是组织和利用这些网页信息的一种有效

收稿日期:2016-03-01;**修回日期**:2016-09-17

基金项目:国家自然科学基金(61472191),江苏省高等学校自然科学基金(15KJA420001),留学回国人员科研启动基金(教外司留[2015]1098号),福建省信息处理与智能控制重点实验室(闽江学院)开放基金(MJUKF201705),山东省语言资源开发与应用重点实验室开放课题(211180A41601),江苏省普通高校研究生科研创新计划(KYLX16_1293)资助.

作者简介:顾敏,女,1993年生,硕士生.研究方向:自然语言处理. E-mail:15205150477@163.com

通讯作者:顾彦慧,博士/副教授. E-mail: gu@njnu.edu.cn

途径.网页分类是指按照预先定义的主题类别,根据海量网页文本的内容,确定相应网页的类别.目前,网页分类采用的技术基础是基于内容的纯文本分类.在抓取到的网页集合中,先对每篇网页文本进行纯文本的内容抽取,得到相应的纯文本文档;再将抽取出的纯文本组成新的文档集合,在新的文档集合上运用纯文本分类算法进行分类.基于纯文本的方法具有局限性,因为网页是一个半结构化的文档,含有丰富的结构信息,不能仅仅考虑文本信息而忽略结构信息进行分类.网页分类的难点在于结合网页的结构信息选择合理的表达方式和分类算法[1].网页分类的一般过程包括:文本表示,特征选择,特征提取和分类器.将网页文本内容抽取后进行分词,然后进行文本表示,从而将网页内容形式化表示出来,常用的方法是向量空间模型(vector space model, VSM).特征选择和提取是指通过互信息、卡方检验等方法量化特征权值,构建特征向量.分类器是指构建分类器,用测试样本对训练得到的特征向量进行测试、评价.

在传统文本分类技术的基础上,许多研究者针对 Web 网页特性开展了研究.一些研究使用邻居网页的文本信息训练分类模型.Fürnkranz[2] 使用 Ripper 算法进行分类,用指向网页的所有链接周围的文本和所在段落的标题及上级标题来表示网页,其正确率比直接使用网页局部文本提高了 20%.Shen 等[3] 对网页的隐含链接和显示链接的作用进行了比较,发现使用显示链接可以取得更好的分类效果.

在特征选择和提取方面,Sriurai 等[4] 使用基于潜在语义分析的主题模型对特征进行降维处理.由于使用词袋模型的特征过于简单,忽视了特征项的语义信息,所以表示相似含义的不同特征项在特征向量中的位置不同,这一定程度上会影响分类效果,所以他使用主题模型进行分类,分类结果的 F 值较词袋模型相比提升 23.31%,达到了 84.51%.由于网页是多视图的数据,通常包含两种以上的数据类型,如文本、超连接、图片等.每一种数据类型都可以看作一个视图.Jing 等[5] 由此首次提出了特征提取方案:半监督的视图内和视图间的多方面判别特征提取方法(SI2MD);当标记的训练样本占 60%时,正确率达到了 80%以上.

目前,网页分类中使用的分类算法有决策树[6]、朴素贝叶斯[7]、支持向量机[8]、KNN[9]、神经网络[10] 和 LLSF 等.Markov 等[6] 将网页表示成图的结构,提出了一种利用网页的标记信息进行分类,使用决策树和朴素贝叶斯的混合模型进行分类,实验结果表明,混合模型具有更高的分类精度.郑诚[7] 提出了基于 Naïve Bayes 的协调分类方法,利用超文本页面中的文本信息和结构信息进行分类,将<title></title>和<hn></hn>之间的单词提取出来,用基于 Naive Bayes 和文档相似性的分类器进行分类,结果显示综合文本信息和结构信息的分类器具有更好的分类效果.SVM 具有处理巨大特征空间的潜力,能够很好地解决具有高维数、特征相关、向量稀疏、线性可分特点的文本分类问题,但是 SVM 适合二分分类问题,而网页分类一般是多元分类.所以,张亮等[8] 提出了一种基于最小损失的支持向量机模型(SLMBSVMs),在网页分类中,采用 l-vs.rest 策略来将多类问题转化成 SVM 可以解决的两类问题.训练正例是该类所包含的全部网页;而反例是在训练集中不属于该类的所有其他类的网页.这就解决了 SVM 只适用于二分分类的问题.针对两类问题的错误损失的不同,构建整体损失率最小的 SVM.在该数学模型下,系统将以较高的概率正确输出某一类样本,而牺牲另一类的正确率.文献[8]借助模型的特性构造了两个 SVM,并将其与 KNN 结合提出了新的 SLMSVM-KNN 分类算法.实验结果显示,分类效果有显著提高.总体上看,SVM、KNN、神经网络等算法的分类正确率占优势,但是朴素贝叶斯在保持一定分类正确率的前提下有着较高的工作效率,所以本文选择了朴素贝叶斯分类算法.在网页分类的研究中,使用基于结构和文本特征进行分类的实验虽然实验精度较高,但是也存在着一些不足. Quek 等[11] 使用组合的分类方式,他将网页的<title>,<meta>,<a>和文本内容抽取出来,用于判断网页的分类.Yang 等[12] 总结了网页之间超链接存在的一些规则,利用这些规则设计了网页的表示方法和分类算法.正确率达到了 83%.刘欣[13] 使用基于结构的方法进行网页分类.网页中的文本内容按其所在的结构进行划分,每一个结构单独训练分类器,最后将组合的结果作为整个网页的分类结果输出.实验的宏观 F 值达到了 87%,但是她只使用了网页的标题和主题文本进行组合,没有利用其他可用的结构,如元数据标签等.此外,层次化的分类与单一分类器相比可能产生阻滞问题,上一层分类错误会影响下层的结果.

在总结前人工作及对网页进行分析的基础上，本文使用了原子特征，即文本信息进行分类.此种方法称为“基于原子特征的分类方法(single feature categorization,SFC)”.因为原子特征的分类区分度有限，所以本文使用了基于众创网页结构和文本联合特征的网页分类方法(joint feature categorization,JFC).利用网页中具有明显类别区分度的标签，而不仅仅使用标题和主题文本.将网页内容按其结构进行划分抽取，通过加权词频的计算使提取出的特征更具有类别区分度.此外，本文将文本特征与结构特征相结合，通过转化为DOM(document object models)树进行结构特征的提取.在训练获得特征向量后使用朴素贝叶斯分类器进行分类.本文的研究方法实际已应用于众创相关网页分类和信息抽取.

本文的贡献简要列举如下：

(Ⅰ)分类特征提取.本文分析与众创相关的网页的结构和文本特征，总结出判断规则，提取结构特征和文本特征；

(Ⅱ)基于结构特征和文本信息分类.根据前期提取的特征，采用贝叶斯分类算法进行网页分类；

(Ⅲ)测试与优化.分析测试结果，修改通过训练提取的特征，提高分类正确率及召回率.

1 网页特征

本节主要介绍众创相关网页的特征，对其特征进行分析可以为特征选择提供更多的依据.众创相关的网站一般包含孵化器的介绍，新闻动态，活动信息，项目信息及相关政策等网页.通过分析大量的众创网站，本文总结出众创网站的特点如下：

(Ⅰ)各种类别网页均会出现一些明显的特征词，这些特征词会以强调的形式出现，如加粗显示等.例如活动类网页中会出现“活动时间”，“活动地点”等词语加粗显示，这些特征词在选取时需要考虑适当增加权重.

(Ⅱ)项目的介绍多以<table>的形式呈现，内容简短，而新闻往往以段落形式出现.在网页预处理抽取网页信息时需要按实际情况处理.

(Ⅲ)网页的导航栏部分一般以超链接<a>或列表<li>形式显示内容，超链接会显示出当前指向的页面，如标签<a>+“孵化器介绍”表明当前网页为孵化器类的网页.

(Ⅳ)新闻类网页会有一个div用于显示发布时间和文章来源等新闻的基本信息，可以抽取出该部分内容作为结构特征.如：“ 发布时间 ”，在网页中用于显示网页的发布时间.

(Ⅴ)活动类网页有标签用于显示活动时间，活动地点等信息，如：“ 活动地点 ”，在网页中显示出活动的地点.

(Ⅵ)孵化器类网页的底部会有一个div供用户填写孵化器的申请信息，如：“ +申请”，这是孵化器网站单独具有的结构特征.

(Ⅶ)项目类网页一般会有标签<p>+“项目”的结构特征，如：“<p class=”path”>+项目”，这种联合特征可用于区分项目类的网页.

(Ⅷ)项目类网页中会出现大量的复合词，如“项目展示”，“项目负责人”，“项目开发”等复合词，复合词可作为联合特征用于分类.

网页分类一般会选择原子特征作为分类的特征，但是原子特征具有一定的局限性，所以本文选择将这些结构和内容的特征作为联合特征用在分类器中，从而有效改善分类效果.

2 基于结构和文本的网页分类方法

基于结构和文本联合特征的分类方法中，对于网页含有的丰富结构信息加以利用对网页分类方法的改进有重要作用.本文分类方案的示意图如图1所示.首先提取网页文本信息并分词，对结构标签进行分类别提取.然后进行特征的选择和特征提取，在训练得到特征向量后，使用朴素贝叶斯分类器进行分类.

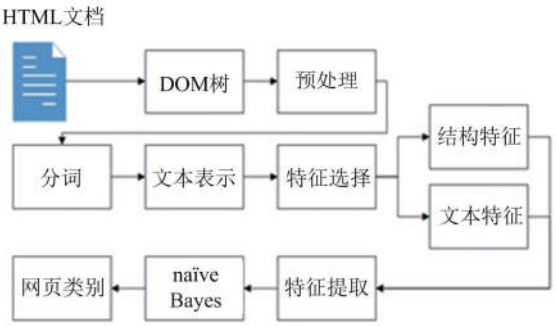

图1 分类方案示意图

Fig.1 Schematic of classification scheme

2.1 网页预处理

本文采用基于DOM树的网页内容提取方法[14].它的主要思想是按照HTML的内部标签结

构将其表示为一棵语法树，每一个 tag 标签对应一个结点，将信息提取转化为对语法树的操作，遍历语法树获取需要的信息.在 DOM 树中，HTML 文档中的每一个成分都是一个节点，主要分为 5 类节点：①文档节点(root)：整个文档的根节点；②元素节点(element)：每一个 HTML 标签是一个元素节点；③属性节点(attribute)：HTML 标签属性；④文本节点(text)：HTML 元素中的文本；⑤注释节点(annotation)：HTML 文档中的注释.将一个 HTML 文档表示成一棵文档树，树的叶子节点是文本节点，包含对应的文本信息.一个孵化器类的网页的部分 HTML 代码和对应 DOM 树如图 2 所示.

```
<html>
<head>
<title>i.Center清华大学创客空间</title>
<meta name="keywords" content="i.Center清华大学创客空间">
<meta name="description" content="i.Center清华大学创客空间">
</head>
<body>
<div class="topbar"><p>创客孵化公共服务平台</p></div>
</body>
</html>
```

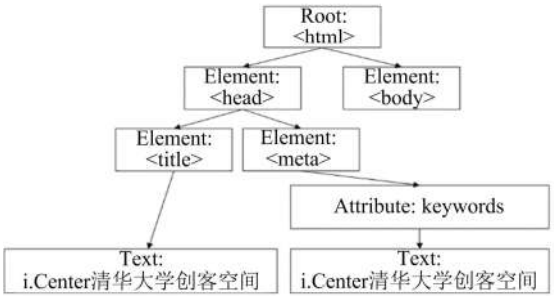

图 2　DOM 树举例

Fig.2　Examples of DOM tree

遍历 DOM 树可以获取需要的 tag 的属性信息和文本信息.本文使用 Jsoup 解析器[①]进行网页清洗，去除网页的广告和干扰判断的项目，然后进行内容的提取.网页中对分类有帮助的信息可以分为：

①标题.标题是＜title＞＜/title＞标签中的文字，是对网页内容的概括.

②元数据标签.＜meta＞＜/meta＞标签主要用于说明一个网页的页面信息，其中 name 为 description 和 keywords 的内容表示网页的关键词和描述.

③表示强调的标签.各级小标题＜hl＞-＜h6＞，强调＜strong＞，强调＜em＞，黑体＜b＞，斜体＜i＞表示网页中的一些重要信息.

④网页正文中的文本.网页正文的文本是网页分类中最常使用的信息，一般是＜p＞＜/p＞标签或者＜table＞标签中的信息.

预处理的主要流程包括：①使用 Jsoup 解析器将 HTML 文档转为 DOM 树；②网页清洗，删除无用标签；③遍历 DOM 树，如果是提取的元素节点，则提取出该标签的文本信息，否则，继续读下一个节点，直到遍历完所有节点.

在取出文本信息后，需要对文本信息进行分词和去除停用词处理，使用 Lucene 分词开源工具包(IKAnalyzer[②])进行处理，在分词后删除相应的停用词和符号，保留文本中的有用信息.

2.2　特征选择与提取

特征选择是从原始训练集合中选择出具有较强区分类别能力的特征.通常，标题和元数据标记中信息是对网页主题的直接描述，而强调标记虽然不能直接描述网页的主题，但是可以体现出网页的内容结构，具有很强的类别归属信息.此外，网页中的普通文本的类别指示性虽然不及以上三类，但是由于文本内容长度较长，其中也包含一些可以指示网页所属类别的信息.本文将需要抽取文本内容信息的 tag 类型分为三类，如表 1 所示.

表 1　HTML tags 分类

Tab.1　Categories of HTML tags

类别	Tags
1	＜title＞,＜meta＞
2	＜hl＞-＜h6＞,＜strong＞,＜em＞,＜b＞,＜i＞
3	＜p＞,＜table＞,＜span＞

特征词是从以上标签中分别抽取出的文本.网页中特征词的权重主要取决于三个方面：①特征词在网页中出现的词频；②特征词在网页中出现的位置，即所属 tag 的类别[15]；③含有特征词的训练样本数目.

定义特征词 c 的加权词频为：

$$\mathrm{tf}_c = \sum_{p=1}^{3} (\lambda(p) \times f_{c,p}) \tag{1}$$

$f_{c,p}$ 表示特征词 c 所在 tag 在属于类别 p 时出现的词频，$\lambda(p)$ 为加权函数，在实验中，不使用加权特征权值的分类方法达到的平均正确率为 61%.对比实验结果，参考兰均等[16]的研究结果，确定加权函数为

① Jsoup 是 java 语言编写的 HTML 解析器，下载地址为 http://jsoup.org.

② IKAnalyzer 是基于 java 语言开发的轻量级的中文分词工具包.

$$\lambda(p)=\begin{cases}4, p=1\\2, p=2\\1, p=3\end{cases} \tag{2}$$

式中，p 为特征词 c 所在 tag 属于的类别.特征词 c 的复合权重 w_c 计算公式为

$$w_c=\mathrm{tf}_c\times\mathrm{idf}_c=\mathrm{tf}_c\times\log\left(\frac{N}{N_c}+0.01\right) \tag{3}$$

tf_c 为特征词 c 的加权词频，idf_c 为反文档频率，N 为训练样本中网页总数，N_c 为含有特征词 c 的网页数.复合权重作为特征权值用于贝叶斯分类器中计算类条件概率.按照复合权重排序，选取 top 40-45 的特征词构成文本特征向量.

在特征选择时不但需要考虑网页的文本内容，而且需要将网页的结构特征加入其中.本文根据总结出的规则，选取了 37 个结构特征作为联合特征加入分类中.如表 2 所示.

表 2 联合特征举例

Tab.2 Examples of joint feature

类别	tag	词语
项目类	div[class=place]	项目
新闻类	div[class=location]	新闻
活动类	div[class~=position*]	活动报名
	a[class=active]	街区活动
孵化器	a[class=on]	孵化

结构特征的权重 w_s 计算只考虑出现的频次，不考虑特征出现的位置.将文本特征和结构特征整合构成联合特征向量.

2.3 分类器

在通过训练样本得到特征后，本文使用朴素贝叶斯算法进行分类.朴素贝叶斯分类器是一种有监督学习，常见有两种模型：多项式模型(multinomial model)和伯努利模型(Bernoulli model)[17].二者的计算粒度不一样，多项式模型以单词为粒度，又称词频型；伯努利模型以文件为粒度，又称文档型.本文选用了多项式模型，根据先验概率来计算网页属于某一类别的后验概率.设 $D=\{d_1,d_2,\cdots,d_n\}$ 是训练样本集合，$C=\{C_1,C_2,\cdots,C_n\}$ 是类别集合，$V=\{v_1,v_2,\cdots,v_n\}$ 是特征集合，包括结构特征和文本特征.将待分类网页 d 描述为 k 个特征项的序列 $<t_1,t_2,\cdots,t_k>$，求出后验概率最大值对应的 C 即为网页所属类别.假设特征之间相互独立.

后验概率定义为

$$P(C_i\mid d)=\frac{P(C_i)P(t_k\mid C_i)}{P(d)} \tag{4}$$

式中，$P(d)$ 为常数，所以后验概率的计算可以转化为求类条件概率 $P(t_k\mid C_i)$ 的最大值.$P(C_i)$ 是类别 C_i 的先验概率，类条件概率 $P(t_k\mid C_i)$ 是特征 t_k 在类别 C_i 出现的概率.

先验概率为

$$P(C_i)=\frac{\text{类}C_i\text{下单词总数}}{\text{训练样本单词总数}} \tag{5}$$

类条件概率估计公式为

$$P(t_k\mid C_i)=\frac{1+w(t_k,C_i)}{|V|+\sum_{j=1}^{|V|}w(t_k,C_i)} \tag{6}$$

式中，$w(t_k,C_i)$ 是特征项在类别 C 中的特征权值，取值可以是文本特征的复合权值 w_c，也可以是结构特征的权值 w_s.V 是训练样本的特征词表，$|V|$ 表示训练样本中包含的特征总数.网页所属类别 C_x 定义为

$$C_x=\underset{C_i\in C}{\operatorname{argmax}}(P(C_i)P(t_k\mid C_i)) \tag{7}$$

3 实验及结果分析

实验所采用的数据集是从网络中抓取的 3 763 个网页，进行人工标注，选取 3 228 个网页作为训练样本，535 个网页作为测试样本.首先使用训练样本训练得到模型文件，模型文件中存储特征向量.将模型文件作为贝叶斯分类器的输入，使用测试样本进行测试，结果与人工标注结果进行对比分析.实验中使用的特征数目、训练样本和测试样本的规模如表 3 所示.

表 3 实验中特征数目及样本数目

Tab.3 Number of features and samples

类别	特征数目	训练样本数	测试样本数
创业相关类	100	1 097	430
创业不相关类	200	2 131	105
孵化器类	51	248	57
活动类	44	221	63
项目类	47	218	98
新闻类	60	318	212

特征数目包括了结构和文本特征的总数目.本文在实验中对比了单独选用文本特征作为原子特征进行分类(SFC)和将文本特征和结构特征作为联合特征进行分类(JFC)的结果.单独文本特征和采用结构和文本联合特征的实验结果对比如图 3 所示.

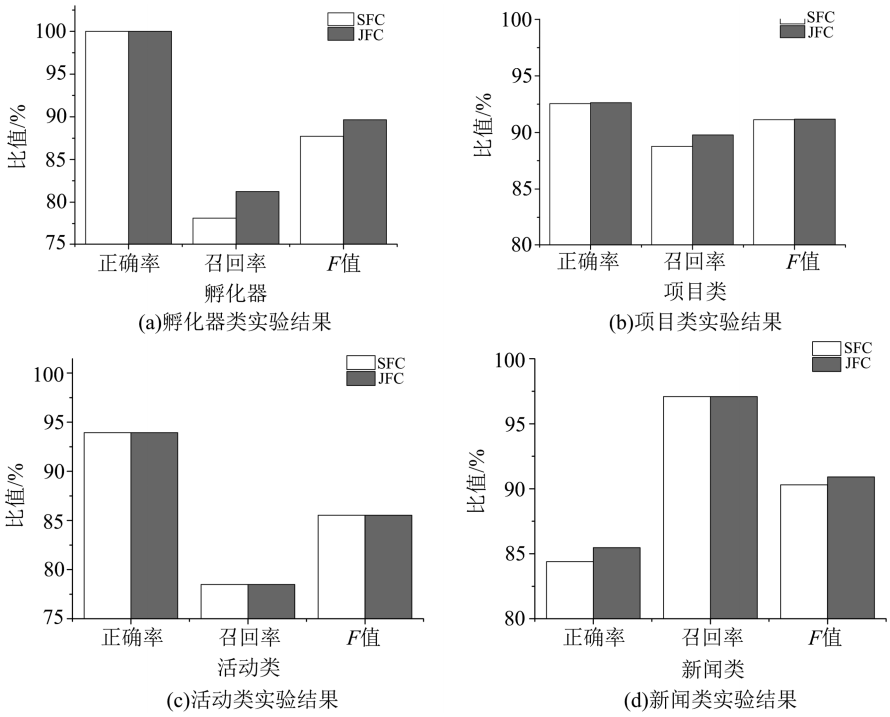

图 3　原子特征与联合特征结果对比图

Fig.3　Comparison of SFC and JFC

联合特征分类中孵化器类和活动类的召回率低于其他类别的召回率，仅为 81%.因为从训练样本中获取的结构和文本特征是有限的，使选取的特征的类别区分度有限，造成了某些类别的召回率比较低.从实验结果可以看出，使用联合特征时，新闻类分类的正确率高于使用原子特征 1.08%，孵化器类召回率提高 3.12%，项目类召回率提高 1.01%.结构特征和文本特征联合进行分类的正确率和召回率都高于未使用联合特征的情况.这主要是因为网页是一个半结构化数据文件，结构在网页的分类中会产生一定的影响.本文实验的四种分类在结构上有着明显的区分，所以选择结构和文本作联合特征进行分类的效果会优于只考虑文本特征的情况.活动类分类效果没有显著提升是因为活动类的结构特征不适用于所有网页，类别区分度有限，具有一定的局限性.这也是本文的一个局限.

4　结论

因为网页的结构和内容比较复杂，所以网页分类的问题还有许多值得研究的方面.本文使用了结构和文本作为联合特征，用贝叶斯分类算法进行分类.本文选择作为特征的词语都是原子词语，在接下来的工作中可以考虑将复合词加入到特征中.此外，在网页抓取时可以利用网页所处的层次结构信息[18].训练和测试语料库的大小对分类也有一定的影响，本文的结论是在一个小规模语料的基础上得出的，今后将扩大语料的规模进行测试和研究.

参考文献(References)

[1] 孙建涛，沈抖，陆玉昌，等. 网页分类技术[J]. 清华大学学报(自然科学版)，2004，44(1)：65-68.
SUN Jiantao，SHEN Dou，LU Yuchang，et al. Web classification technology [J]. Chinese Journal of Tsinghua University(Natural Science)，2004，44(1)：65-68.

[2] Fürnkranz J. Exploiting structural information for text classification on the WWW[C]// Proceedings of the 3rd International Symposium on Advances in Intelligent Data Analysis. Springer，1999：487-498.

[3] SHEN D，SUN J T，YANG Q，et al. A comparison of implicit and explicit links for web page classification [C]// Proceedings of the 15th International Conference on World Wide Web. Edinburgh，UK：ACM Press，2006：643-650.

[4] SRIURAI W，MEESAD P，HARUECHAIYASAK C. Improving web page classification by integrating neighboring pages via a topic model [C]// 10th International Conference on Innovative Internet Community Systems. Bonn，Germany：Gesellschaft

Für Informatik, 2010: 238-246.

[5] JING X Y, LIU Q, WU F, et al. Web page classification based on uncorrelated semi-supervised intra-view and inter-view manifold discriminant feature extraction[C]// Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Agentina: AAAI Press, 2015: 2255-2261.

[6] MARKOV A, LAST M, KANDEL A. Model-based classification of web documents represented by graphs [C]//Proceedings of the Conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA: ACM Press, 2006. oai:CiteSeerX.psu:10.1.1.86.3971.

[7] 范焱, 郑诚, 王清毅, 等. 用 Naive Bayes 方法协调分类 Web 网页[J]. 软件学报, 2001, 12(9):1386-1392.
FAN Yan, ZHEN Cheng, WANG Qingyi, et al. Using naïve Bayes to Coordinate the classification Web pages [J]. Chinese Journal of Software, 2001, 12(9): 1386-1392.

[8] 张亮, 叶允明, 于水, 等. SLMBSVMs-KNN: 一种新的网页分类算法[C]// 全国搜索引擎和网上信息挖掘学术讨论会. 北京, 2003: 80-85.

[9] 李蓉, 孙媛. SVM-KNN 分类器在网页分类中的应用[J]. 科学技术与工程, 2009, 9(16): 4653-4656.
LI Rong, SUN Yuan. Application of SVM-KNN classifier into Web page classification [J]. Chinese Journal of Science Technology and Engineering, 2009, 9(16): 4653-4656.

[10] SUN S, LIU F, LIU J, et al. Web Classification Using Deep Belief Networks[C]// Proceedings of the 17th International Conference on Computational Science and Engineering. Chengdu, China: IEEE Computer Society, 2014: 768-773.

[11] QUEK C Y, MITCHELL T. Classification of world wide web documents[J]. Senior Honors Thesis, 1997: 1-12.

[12] YANG Y, SLATTERY S, GHANI R. A Study of Approaches to HypertextCategorization[J]. Journal of Intelligent Information Systems, 2002, 18(2): 219-241.

[13] 刘欣. 基于结构信息的中文网页自动分类技术研究[D]. 南京: 南京航空航天大学, 2010.

[14] 侯小静, 王黎明. 利用 HTML 标签筛选网页分类样本[J]. 微机发展, 2005, 15(3): 142-144.
HOU Xiaojing, WANG Liming. Using HTML tag to screen Web page classification [J]. Chinese Microcomputer Development, 2005, 15(3): 142-144.

[15] 郭晓, 蒋宗礼. 基于网页结构与链接关系的中文文本分类方法[J]. 现代电子技术, 2010, 33(22): 54-56.
GUO Xiao, JIANG Zongli. A novel Chinese text classification method using webpage tags and hyperlinks[J]. Chinese Journal of Modern Electronic Technology, 2010, 33(22): 54-56.

[16] 兰均, 施化吉, 李星毅, 等. 基于特征词复合权重的关联网页分类[J]. 计算机科学, 2011, 38(3):187-190.
LAN jun, SHI Huaji, LI Xingyi, et al. Associated web page classification based on the weight of composite features [J]. Chinese journal of computer science, 2011, 38(3):187-190.

[17] 张海雷, 王会珍, 王安慧, 等. 基于朴素贝叶斯模型的垃圾邮件过滤技术比较分析[A]// 全国网络与信息安全技术研讨会论文集(下册). 2007: 551-557.

[18] 王振宇, 唐远华, 郭力. 面向分层结构的网页分类与抓取[J]. 计算机工程与科学, 2012, 34(11): 1-6.
WANG Zhenyu, TANG Yuanhua, GUO Li. Categorization and extraction of Web pages based on Hierarchy [J]. Chinese Journal of Computer Engineering and Science, 2012, 34(11): 1-6.