基于单 DOM 树特征预分类的自适应 Web 信息抽取方法

彭艳兵[1,2],谢馨庭[1]
(1.武汉邮电科学研究院 湖北 武汉 430074;2.南京烽火星空通信发展有限公司 江苏 南京 210019)

摘要: 在传统的舆情中多为基于模板采集模式,基于减少人工维护的目的,文中提出一种基于单 DOM 树特征预分类的自适应 Web 信息抽取方法,分为链接预分类与信息抽取两个部分。链接预分类采用 SVM 分类算法,提取信息超链接在页面中的特征进行分类学习,再对分类结果进行同源的 Web 信息提取。实验表明,此方法预分类结果准确率可达 94.48%,召回率为 94.77%。

关键词: DOM 树; 标签路径; 信息抽取; SVM
中图分类号: TN919.6 **文献标识码:** A **文章编号:** 1674-6236(2017)19-0056-04

The adaptive Web information extraction based on single DOM tree characteristics and classification

PENG Yan-bing[1,2], XIE Xin-ting[1]
(1.*Wuhan Research Institute of Posts and Telecommunications*, *Wuhan* 430074, *China*; 2.*Nanjing FiberhomeStarrysky CO.LTD.*, *Nanjing* 210019, *China*)

Abstract: In traditional public opinion, mostly based on the template in acquisition mode, based on the reduction of artificial maintenance purposes, we propose a method based on adaptive Web information extraction single DOM tree features pre-classification, divided into the pre-classification and information extraction link two parts. Links presorting using SVM classification algorithm to extract information about hyperlinks in the pages of features to classify learning, then the results of the classification homologous Web information extraction. Experimental results show that this method of pre-classification accuracy rate of 94.48%, the recall rate was 94.77%.

Key words: DOM tree; tag path feature; information extraction; SVM

网络舆情是指在一定社会空间内,通过网络围绕社会事件的发生,发展和变化,民众对于公共问题和社会管理者产生和持有的社会政治态度,信念和价值观,是国家相关部门了解民意的重要渠道,Web 信息作为舆情系统进行舆情分析的信息输入,采集是否准确,站点是否覆盖全面覆盖,直接影响了舆情系统的性能[1]。在传统的舆情采集中,多采用的是基于模板的方法,对于各个站点进行模板化的定制,随着站点覆盖的越来越多,用于模板维护定制的人力也消耗的越来越大,为了快速准确地获取舆情信息,舆情系统对 Web 信息抽取提出了越来越高的要求,因此如何让计算机程序自动准确地从千变万化的页面中抽取出结构化的目标数据,一直是舆情系统待解决的问题。

目前比较流行的自动化信息抽取工具有 MDR[2]、基于 MDR 的改进方法 Depta[3]等,但这些方法对待抽取网页中信息的结构化程度要求比较严格,但实际网络中存在大量松散结构化信息的网页。文献[4]提出了一种基于文本内容相似度的网页正文提取方法,但未考虑如何区分是否为同源页面。文献[5]提出了一种基于标签路径特征融合的在线 Web 新闻内容抽取方法,引入了页面标签路径特征。

文中深入研究了网页的超链接特征、文本特征和结构特征,构建了面向网络舆情载体类型识别的特征集,在基于 DOM 自动生成模板的方法上引入机器学习中的分类算法对上层页面中的信息超链接进行预分类,提出一种基于单 DOM 树特征预分类的自

收稿日期:2016-08-06 **稿件编号:**201608050
作者简介:彭艳兵(1974—),男,湖北洪湖人,博士,高级工程师。研究方向:海量数据分析,网络行为分析。

适应 Web 信息抽取方法。

1　基于单 DOM 树特征预分类

文中提出的基于单 DOM 树预分类模块如图 1 所示。该模块输入的是各站点 Web 页面,如首页,各版块列表等。将其预处理生成 DOM 树后,提取其中所有超链接及其特征,进入分类器分类。而分类器是选取了 20 个站点页面中所有链接训练生成。

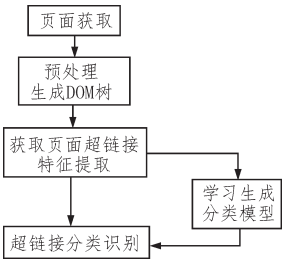

图 1　基于单 DOM 树特征预分类流程图

1.1　DOM 树与树路径

一个 Web 页面可以用 DOM 树表示,DOM 即文档对象模型,定义了 HTML 文档和 XML 文档的逻辑结构,给出了一种访问和处理 HTML 文档和 XML 文档的方法,可以根据 HTML 文档和 XML 文档结构形成一棵对象节点树,称为 DOM 树[6]。在一棵 DOM 树中,各节点的位置可表示为从 DOM 树的根节点到此节点所经过的所有节点标签组成的序列,表示如下:

$$P=(m,t_1,t_2,\cdots,t_n,s_1,s_2,\cdots s_m) \tag{1}$$

其中:m 表示该路径在 DOM 树中出现的次数;$(t_1,t_2,\cdots,t_n)$表示该路径所经历的节点标签组成的序列;$(s_1,s_2,\cdots,s_m)$表示该路径出现的位置,DOM 树所有的路径的叶节点按遍历排序, 用顺序号表示树路径的位置[7]。

树路径是一条从根节点到叶节点经过的所有标签序列, 传统树路径匹配计算采用计算路径序列的相似度,只考虑标签序列,忽略了树路径标签序列在页面中出现的位置, 计算出的相似度结果并不能真实有效地反应实际相似度,因此,本文采用了一种改进的基于树路径匹配的网页结构相似度算法[8]。对于两条树路径

$$P_i=(m,t_{i1},t_{i2},\cdots,t_{in},s_{i1},s_{i2},\cdots,s_{im}),$$
$$P_j=(g,t_{j1},t_{j2},\cdots,t_{jh},s_{j1},s_{j2},\cdots,s_{jg})。 \tag{2}$$

它们之间的树路径相似度定义如下:

$$\mathrm{sim}(P_i,P_j)=w\times\mathrm{st}(P_i,P_j)+(1-w)\times sp(P_i,P_j) \tag{3}$$

其中:

$$\mathrm{st}(P_i,P_j)=\frac{\mathrm{clen}(P_i,P_j)}{\max(\mathrm{len}(P_i),\mathrm{len}(P_j))} \tag{4}$$

表示树路径的标签序列相似度,clen(P_i,P_j)表示两条路径以根节点为开始的最长公共标签序列长度[9],len(P_i)表示路径 P_i 的标签序列长度;

$$\mathrm{sp}(P_i,P_j)=1-\frac{\sum_{k=1}^{m}\frac{[md(si_k)]}{pn}+\sum_{k=1}^{g}\frac{[md(sj_k)]}{pn}+|m-g|}{2\times\max(m,g)} \tag{5}$$

表示两条树路径的位置相似度,md(si_k)表示 P_i 路径在位置 si_k 处与 P_j 的最近距离:

$$\mathrm{md}(si_k)=\min(|si_k-sj_{k1}|,|si_k-sj_{k2}|,|si_k-sj_{k3}|,\cdots,|si_k-sj_{kg}|), \tag{6}$$

$$\mathrm{pn}=\max(pn_i,pn_j)-1, \tag{7}$$

pn_i 和 pn_j 分别表示 P_i 和 P_j 所在 DOM 树的叶节点总数。

路径相似度主要由 st(P_i,P_j)和 sp(P_i,P_j)两部分组成, 分别体现了路径相似性中的标签序列和位置信息,w 为权重,取值 0~1,改变 w 可调节这两部分在路径相似性中的重要性[10]。

例:图 2 是一个简单的网页表示成 DOM 树结构:

$$\{P_1=(2,\mathrm{html},\mathrm{head},\mathrm{title},1,2);P_2=(1,\mathrm{html},\mathrm{body},\mathrm{div},\mathrm{span},\mathrm{a},3)\} \tag{8}$$

由于本文只考虑超链接特征,为简化计算,忽略不包含<a>标签路径, 故该页面共有两个树路径,第一条路径 P1 出现了两次, 叶节点 a 由遍历 DOM 树得到的顺序号为 1 和 2, 因此该路径位置分别为 1,2;第二条路径 P2 出现了 1 次,位置为 3。

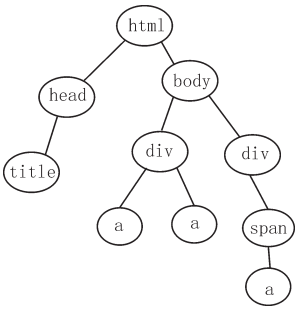

图 2　网页的 DOM 树结构

1.2　特征提取

文中选择支持向量机的分类算法, 对页面中抽取的超链接进行分类。在分类问题中,最重要的是样本的特征选取, 选取特征是否能够反映分类问题的本质,这决定了分类模型的优劣。通过分析大量站点页面, 我们将每个页面中的各超链接视为由超链接特征,文本特征及结构特征构成。

1.2.1　超链接特征

超链接既内容链接,是各网页之间相互连接的有

效路径,我们用同一资源定位符(URL)表示,基本的URL包含协议,域名(或 IP 地址),路径和文件名[11-12]。对于同一站点下的页面,有效的信息帖子 URL 具有以下特征:与站点根域名相同,新闻帖子大多包含日期信息,论坛博客会包含 “bbs”,“thread”,“blog”,“club”等关键词。我们将 URL 是否包含日期及关键词作为特征,进入分类学习。

1.2.2 文本特征

超链接的文本特征,既该链接所对应的<a>标签在 DOM 树中所嵌套包含的文本内容。通过大量观察,其文本内容大多为链接对应网页的标题内容,标签还会携带 title 属性,且由于网页排版限制,大多数帖子标题长度相似,而版块链接文本多限制在 2 到 4 个字符,如新闻,社会,天涯杂谈等。我们将<a>标签所对应文本长度,及是否带有 title 属性作为特征,进入分类学习。

1.2.3 结构特征

根据本文第一节所述,每一个 HTML 页面都可以由一棵 DOM 树结构来描述,它可以将整个页面内容抽象为不同的对象,用结点的方式来表示[13]。通过分析观察,网页中的超链接信息是较均匀的分布在页面主体结构中。简单的版块列表页面超链接都分布在同一区域结构内,而复杂的门户型网页,会存在多个信息超链接区域,单各个信息区域的分界是相似的,且 XPATH 结构路径是相似的。其中每一个链接所在的最小结构体就是该链接所对应的<a>标签在 DOM 树中的 XPATH 路径。基于以上特点,我们可将问题转化为对于<a>标签的路径特征分析。首先,我们根据各链接的 XPATH 路径进行分组,并用公式 1 表示各链接的树路径,相同的 XPATH 路径归为一组。对各组链接进行降序排列,选取链接数最大的组,应用公式计算其他超链接路径与最大链接组路径的相似度,而最大链接数组相似度计为 1。将其作为结构特征。

1.3 分类算法

在本文的分类问题中,我们采用了支持向量机 SVM 算法,这是一个有监督的学习模型,它最终能将训练样本进行划分,求得其最优超平面,它的优势在于是基于系统风险最小化的原则,能够根据有限的样本对给予的任意样本的识别能力和特定样本的学习精度之间寻求到最佳的平衡,以达到最好的学习能力[14]。此算法在解决非线性小样本及高维模式识别等问题中效果良好。应用于本文分类问题时,我们根据第二节所述提取特征,将其抽象表示,选取部分作为训练数据集进行训练学习,获得学习模型。

2 同源页面信息抽取

对于两个同源 Web 信息页面,他们具有相同的结构,页面中不相同的部分即为我们所需抽取的信息内容。我们通过抽取算法比较两个同源页面之间的匹配与不匹配,以获得一个此结构页面的抽取模板。如下图所示,首先我们将网页预处理为 DOM 树结构,选取同组的两个同源页面进行信息抽取计算,生成模板,再通过此模板抽取其他同组页面的内容信息,将其结构化输出。

在信息抽取计算中,其中心思想就是处理两棵 DOM 树之间的不匹配,我们将不匹配分为两种情况,标签不匹配与字符串不匹配,标签不匹配又分为重复项与可选项两种情况[15]。同源页面中的字符串不匹配,很大程度是由于读取数据库内容的不同所造成的,即可认为,字符串不匹配的部分即为我们待抽取的信息内容。对于标签不匹配中的重复项,我们需找到重复标签结点组的最小重复结构,对此重复结构标记并按字符串不匹配处理,通过大量观察可发现,页面中出现的重复结构大多为论坛回贴,新闻评论及博客回复等,同为带抽取的信息内容。而标签不匹配中出现的可选项为信息缺省所导致,将其记录标志,待其它页面验证。将所有不匹配节点标志记录,并依据其结点属性标签内容等特征将其标准化输出为抽取模板。其他同组页面即可通过模板快速抽取信息,而无需进行再次计算。

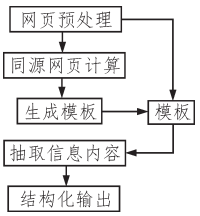

图 3 同源页面信息抽流程图

3 实验与结果

为了验证自动化抽取模型的有效性,编程实现了相关功能算法,并对结果进行评价。在单 DOM 树特征预分类模块中,选取了 21 个站点页面近一万条页面链接,包含门户型综合网站,主流论坛,政府官网等类型,其中 16 个站点页面做为训练集,其余 5 个站点页面为测试集,同时也作为页面信息抽取的

测试集。

表 1　训练数据集站点

训练站点页面	页面超链接数
http://news.163.com/	412
http://news.cqnews.net/	499
http://news.dayoo.com/	137
http://news.eastday.com	277
http://news.gog.com.cn/	139
http://news.sina.com.cn/	1 652
http://news.xhby.net/	292
http://news.youth.cn/	218
http://www.81.cn/	549
http://www.cjdby.net	139
http://www.fecn.net/	1 036
http://www.gmw.cn/	541
http://www.hebei.com.cn	758
http://www.hebnews.cn/	584
http://www.qq.com/	813
http://www.sohu.com/	1 002

实验预分类结果采用准确率与召回率作为评价指标,信息抽取采用准确率与完整性作为评价指标。

将训练数据只提取超链接及文本特征与提取超链接,文本及结构特征两种情况进行分类学习测试,对比加入结构特征后对整体分类的影响效果。

从表的实验结果可以清楚看出，加入了页面结构特征后对于结构单一型的站点页面影响不大,但对于综合门户型网页优化较大。最终总体准确率可达 94.48%,召回率为 94.77%。该特征提取可有效的满足,在单页面中提取其中的有效链接,节省了为各站点定制正则表达来匹配 URL 链接的人力时间。

我们选取了 5 个不同站点进行信息抽取验证,与较流行的基于正文统计算法进行比较，对比结果如下表,可见在传统的新闻站点,二者区别不大,而在含有大量图片新闻的综合门户型网站及论坛等站点,本文提出的抽取方法具有较大优势,且在基于链接预分类的基础上,减少了大量的对比计算。

4　结　论

文中提出了一种基于单 DOM 树特征预分类的自适应 Web 信息抽取方法。针对同一页面中的信息超链接,提取其超链接特征,文本特征及结构特征,

表 2　基于单 DOM 树特征预分类测试结果

站点	页面超链接数	无结构特征准确率	准确率	召回率
http://world.huanqiu.com/	256	70.07%	80.08%	63.36%
http://www.jwb.com.cn/	172	97.67%	98.26%	100.00%
http://www.people.com.cn/	739	93.23%	94.45%	97.41%
http://www.qianlong.com/	549	89.43%	96.91%	99.03%
http://www.xinmin.cn/	421	97.62%	98.57%	100.00%
总计	2137	90.78%	94.48%	94.77%

表 3　页面信息抽取结果

网站	网页数	同源页面抽取算法	基于统计信息抽取
http://world.huanqiu.com/	130	91.53%	84.62%
http://www.jwb.com.cn/	120	100.00%	100.00%
http://www.people.com.cn/	300	96.67%	92.00%
http://www.qianlong.com/	300	95.67%	95.00%
http://www.xinmin.cn/	250	100.00%	96.00%

采用 SVM 分类算法对其进行分类,再对分类结果进行同源的 Web 信息提取。实验结果表明本文提出的方法具有较强的适用性，能有效地对新闻论坛等站点进行信息提取。目前的工作也存在一些不足,需要进一步开展相关的研究,如对于复杂的门户型站点,具有较多的页面样式,在预分类模块中丢失率较高;对于博客新闻类型页面的评论回复无法准确识别。下步计划在预分类中增加页面类型识别，以在信息抽取的过程中针对不同类型页面采取不同抽取方式。

参考文献:

[1] 王元卓,靳小龙,程学旗等.网络大数据:现状与展望[J]. 计算机学报, 2013,36(6):1126-1138.

[2] 王志华,魏斌,李占波,等.基于本体的Web信息抽

(下转第 63 页)

善的系统体系。注重系统的各模块功能互补性,从而共同实现大学生体质测试和数据分析的数字信息化管理。为体质测验后续的成绩评定、报表生成、结果查询以及校方新阶段的体育锻炼计划制定与实施,提供重要的技术支持与数据参考。

参考文献:

[1] 徐叶彤.我国学生体质研究的现状与发展趋势[J].体育文化导刊,2014(2):149-153.

[2] 李国锋. 中国东部与西部汉族中学生体质发育差异动态变化研究[J].兰州文理学院学报:自然科学版,2014,28(6):82-86,96.

[3] 吕俊莉,吴薇.不同专业大学生体质健康状况比较分析[J].西南师范大学学报:自然科学版,2012,37(5):157-160.

[4] 王沛. 一种基于移动手机的大学生体质测试软件设计[J].电子设计工程,2016,24(11) :55-57.

[5] 申良,刘洲洲.一种高校学生体质健康测试管理系统设计与实现[J].电子设计工程,2016,24(1):55-57.

[6] 朱广涛. 基于数据挖掘的学生体质健康测试系统的设计与实现[D].济南:山东大学,2015.

[7] 李文强. 中等职业院校体质评价系统需求分析与设计[D].天津:天津大学,2014.

[8] 李春峰. 高职学生体育成绩管理系统的设计与实现[D].厦门:厦门大学,2014.

[9] 李娜. 基于数据挖掘的高职体育成绩管理系统的设计与实现[D].成都:电子科技大学,2012.

[10]苏锋.基于数据挖掘的中职体育成绩管理系统的设计与实现[D].长沙:湖南大学,2014.

[11]高健.基于数据挖掘的体育成绩管理与体能分析系统[D].成都:电子科技大学,2012.

[12]洪洁.基于数据挖掘的学生管理系统设计与实现[D].昆明:云南大学,2010.

[13]鲍倩.基于Java语言的学生成绩管理系统设计与实现[J].电子科技,2013,26(9):155-156.

[14]张丽娟.基于Web的学生成绩管理系统的设计与实现[D].长春:吉林大学,2009.

[15]杨国林,王飞,贺慧.基于数据挖掘的图书馆数据预处理方法研究[J].电子设计工程,2015(3):23-25.

[16]丁兆云, 贾焰,周斌. 微博数据挖掘综述[J].计算机研究与发展,2014(4):12-117.

(上接第 59 页)

取系统[J].计算机工程与设计,2012,33(7):2634-2639.

[3] 陈钊,张冬梅.Web信息抽取技术综述[J].计算机应用研究, 2010,27(12):4401-4405.

[4] 王利,刘宗田,王燕华,等.基于内容相似度的网页正文提取[J].计算机工程,2010,36(6):102-104.

[5] 吴共庆,胡骏,李莉.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.

[6] 寇月,李冬,申德荣,等.D-EEM:一种基于DOM树的Deep Web实体抽取机制 [J]. 计算机研究与发展,2010,47(5):858-865.

[7] 陈雪,梁永全,赵相彬.改进的基于本体的Web信息抽取[J].计算机应用与软件,2013,30(7):14-16.

[8] 廖浩伟,杨燕,贾真,等.一种改进的基于树路径匹配的网页结构相似度算法[J].吉林大学学报(理学版),2012,50(6):1199-1203.

[9] 高庆宁,吴鹏,张晶晶.基于文档对象模型与行块分布算法的网页信息抽取[J].情报理论与实践,2016,39(4):133-137.

[10]岳国伟,吕楠,申玉三.基于领域本体的Web信息抽取模型研究[J].情报探索,2012(1):105-107.

[11]史西兵,王浩鸣.隐马尔可夫模型解决信息抽取问题的仿真研究 [J]. 计算机仿真, 2010,27(5):132-135.

[12]李伟男,李书琴,景旭,等.基于模拟退火算法和二阶HMM的Web信息抽取 [J]. 计算机工程与设计, 2014,35(4):1264-1268.

[13]李少天,肖基毅,虞乐.基于HMM和小波神经网络混合模型的Web信息抽取[J].微计算机信息,2012(5):136-138.

[14]许世明,武波,马翠,等.一种基于预分类的高效SVM中文网页分类器 [J]. 计算机工程与应用,2010,46(1):125-128.

[15]岳国伟,吕楠,申玉三.基于领域本体的Web信息抽取模型研究[J].情报探索,2012(1):105-107.