文章编号: 1006-2475(2018) 09-0021-06

基于支持向量机的网页正文内容提取方法

梁 东,杨永全,魏志强

(中国海洋大学信息科学与工程学院,山东 青岛 266100)

摘要: 针对网页的正文信息抽取,提出一种基于支持向量机(SVM) 的正文信息抽取方法。该方法采取宽进严出的策略。第 1 步根据网页结构的规律遍历网页 DOM 树,定位到一个同时包含正文和噪音信息的 HTML 标签。第 2 步选择含噪音信息的 HTML 标签的 5 个重要特征,并采用 SVM 训练样本数据。SVM 训练得出的数据模型可以有效去除导航、推广、版权等噪音信息,成功保留正文。将该方法应用于几大常用的网站,实验结果表明该方法具有较好的正文抽取效果和降噪效果,对于传统方法中经常误删的短文本、与正文相关的超链接等信息能够准确保留。

关键词: 支持向量机; 正文抽取; HTML 标签; 降噪; 机器学习
中图分类号: TP391. 1 文献标识码: A doi: 10. 3969 / j. issn. 1006-2475. 2018. 09. 005

Information Extraction of Web Pages Based on Support Vector Machine

LIANG Dong, YANG Yong-quan, WEI Zhi-qiang

(School of Information Science and Engineering, Ocean University of China, Qingdao 266100, China)

Abstract: Aiming at the text information extraction of Web pages, this paper presents a method of extracting text information based on support vector machines. This method adopts "come in easily, out strictly" policy. The first step is to traverse the Web DOM tree according to the rules of the Web page structure, and locate an HTML tag that contains both useful and noise information. The second step is to select five important features of the HTML tag with noise information and use SVM to train the sample data. The model can effectively remove the navigation, promotion, copyright and other noise information, and preserve the useful information of Web pages. The method is applied to several commonly used websites. The experimental results show that this method has good effect of extracting texts and noise reduction, and can preserve short texts, such as hyperlinks related to texts that often mistakenly deleted by traditional methods.

Key words: support vector machine; information extraction; HTML label; noise reduction; machine learning

0 引 言

Web 信息抽取最早在 2001 年提出,目的是将网页中有价值的文本信息提取出来。网页正文是获取 Web 信息的重要来源,但网页中也经常掺杂着与正文无关的导航、推广、版权等噪音信息。正文信息抽取对后续的文本分类、分词、语义分析等工作具有重要意义。本文首先通过分析网页结构的规律,采用一定的规则遍历 DOM 树,定位到一个同时包含正文信息和噪音信息的 HTML 标签; 然后通过提取含噪音信息的 HTML 标签的 5 个重要特征,采用 SVM 训练样本数据,将训练后得出的数据模型用于降噪处理。

1 研究现状

现阶段流行的正文抽取方法主要有 3 类: 基于统计学的正文信息提取方法、基于正文特征和网页结构的信息提取方法、基于机器学习的正文提取方法。

文献[1]提出了一种基于中文标点特征的正文抽取方法。该方法指出正文信息和噪音信息的中文标点具有不同特征,结合网页结构的规律实现正文信息抽取。该方法不针对单一类型的网站,实现相对简单。但由于段落的小标题与噪声信息的标点特征类似,所以该方法易遗漏文章中的短文本。

文献[2-4]通过分析正文信息 DOM 树节点路径

收稿日期: 2018-02-03
基金项目: 海洋科学与技术国家实验室鳌山科技创新计划项目(2016ASKJ07, 2016ASKJ07-08)
作者简介: 梁东(1993-) ,男,黑龙江哈尔滨人,中国海洋大学信息科学与工程学院硕士研究生,研究方向: 大数据,机器学习; 通信作者: 杨永全(1985-) ,男,讲师,博士,研究方向: 云计算,物联网; 魏志强(1969-) ,男,教授,博士生导师,研究方向: 计算机软件与理论,数据挖掘。

的规律提出了相应的正文信息提取方法。在此基础上,研究者发现正文文本密度和噪音信息的文本密度具有不同特征,文献[5]融合了文本密度和标签路径覆盖率2种特征实现了正文信息抽取。该方法融合了多种属性,不易遗漏正文中的短文本。

文献[6]提出一种基于贝叶斯的Web新闻网页正文抽取方法,该方法利用机器学习中的朴素贝叶斯分类方法,将正文信息抽取看作一个二分类问题,即正文信息和噪音信息。该方法首先筛选出正文中出现的高频词汇,根据高频词汇用贝叶斯定理计算文本判断为正文的概率。最后设定一定的阈值,筛选正文。

文献[7-8]是一种基于统计学的信息提取方法,通过分析正文信息在HTML源文件的位置规律,从而实现正文信息提取。文献[9]提出了一种基于标记窗的网页正文信息提取方法,通过分析标题词序列与正文字符串词序列的距离规律实现了正文信息抽取。

本文中抽取正文信息的方法采取“宽进严出”的策略,总体分2个步骤:

1)遍历网页的DOM树,采用一定的规则准确地定位到一个同时包含正文信息和噪音信息的HTML标签。

2)参照文献[10-13]中,SVM在垃圾邮件过滤和其他领域中优秀的分类效果,本文方法选取噪音信息词汇特征、噪音信息标点特征、HTML标签文本长度、HTML标签位置、HTML标签文本长度的比值作为5个重要特征,采用SVM训练样本,使用训练得出的数据模型进行降噪。

本文方法的优势主要如下:

1)不单纯依赖一种特征进行识别,融合多种噪音信息特征共同识别,适用于多种风格的网站,提高了降噪准确率,减少误删正文的情况。

2)SVM训练的模型精度一定程度上依赖于训练数据,随着时代变化,网页数据也会随之发生变化,所以该方法适用于当今数据爆炸式增长的互联网环境。

2 正文信息定位方法

网页中的正文信息大量存在于p标签中,同时还有部分正文信息存在于嵌套在p标签中的a标签。例如: <a>正文信息 ,由于a标签中往往包含噪音信息,所以当p标签嵌套a标签时,需要进行判断。

当出现上述情况时,记p标签的文本长度为L_p,记a标签的文本长度为L_a,采用公式(1)计算比值r。

$$r=\frac{L_a}{L_p},\quad 0\le r\le 1 \tag{1}$$

经过大量统计,发现当r越接近0时,即a标签的文本长度所占比例越小时,a标签中的文本是正文的概率越大。例如:

但<a>Steam国区玩家的人均拥有游戏数量仅为11.78款

此时r等于0.28,a标签中的文本为正文。

反之当r越接近1时,a标签文本是噪音信息的概率较大。

同时人们发现通常所有的正文内容都被一个div标签包含,所以只需找到包含所有正文内容的那个div标签,就能定位到正文位置。定位方法是从网页的HTML源码中找出具有最大文本长度的div标签。下面介绍定位方法的具体步骤,见图1。具体步骤如下:

1)遍历所有的div标签。

2)采用深度优先的方式遍历每个div标签中的全部p标签,目的是计算出每个div标签包含的文本长度。

3)遇到p标签嵌套a标签时,计算比值r,若r≤阈值θ,则保留p标签。

4)遍历标签时,遇到div标签中嵌套的div标签则直接过滤掉,以避免重复计算。

5)计算每个div标签包含的p标签的文本长度。

6)筛选出具有最大文本长度的div标签。

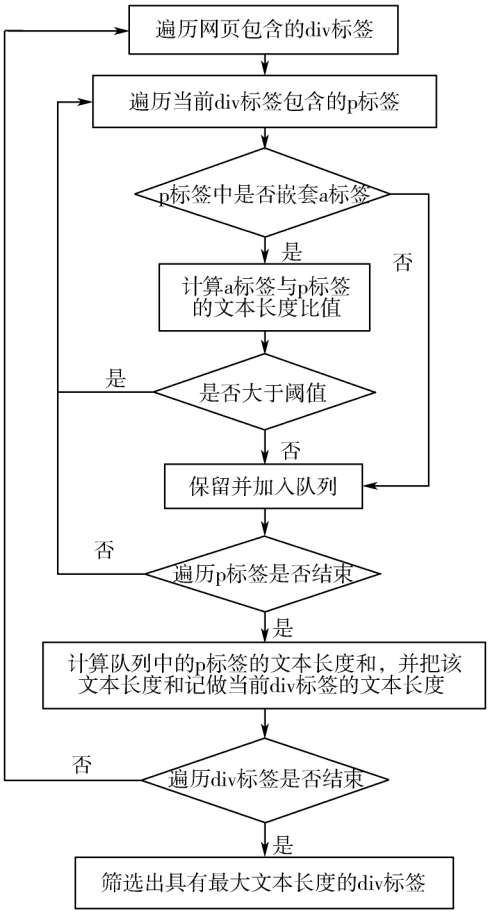

图1 正文定位流程图

定位到全部正文信息的成功率与选取的阈值 θ 有关,实验选择 1000 个网页作为测试网页,测试 θ 取不同值时定位到全部正文的成功率。需要注意的是该 div 标签中还同时混杂着噪音信息,本文将在后续工作中采用 SVM 训练的数据模型进行降噪。测试的结果见图 2。

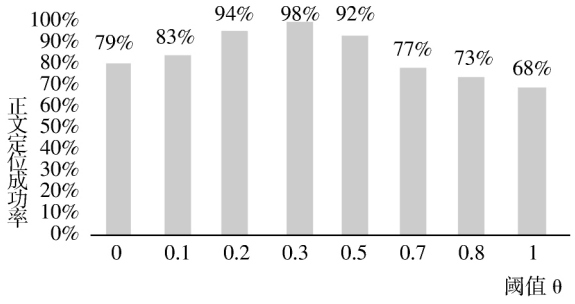

图 2　正文定位成功率与阈值 θ 的关系图

3　支持向量机

本文将降噪看成一个二分类问题,即正文信息和噪音信息。实验中使用 SVM 解决一个二分类问题。假设有一个定义在 n 维空间上的训练集。正类代表正文信息,负类代表噪音信息。通过引入映射,将在低维空间线性不可分的训练集映射到高维空间变成线性可分。SVM 中常用的核函数主要有以下 3 种:

1) 线性核函数:

$$K(x_i, x_j) = x_i \cdot x_j$$

2) 多项式核函数:

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

3) 径向基核函数:

$$K(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{\sigma}\right)$$

其中 d 和 σ 均为核函数的参数,需要人工调整。

在实验中,本文选择能够区分正文信息与噪音信息的 5 个重要特征属性,即最后的训练矩阵为 5 维。第 4 章将逐一介绍这 5 个属性。

4　选择噪音信息标签的特征属性

现在已经找到一个同时包含正文信息与噪音信息的 div 标签。降噪的过程是采用深度优先的方式,遍历当前 div 标签包含的每个 HTML 标签,采用 SVM 生成的模型进行识别。由于绝大多数正文存在于 a 标签和 p 标签中,所以实验设计的 SVM 数据模型只用于识别 p 标签和 a 标签。选择特征属性之前首先要采集训练样本数据。

4.1　采集训练数据

首先从互联网上最流行的 5 大网站: 百度、网易、新浪、腾讯、凤凰网上共随机选取 3572 个网页,使用第 2 章介绍的正文定位方法找到包含正文和噪音信息的 div 标签,将上述 div 标签中的全部 HTML 标签进行二分类,分为只含正文信息的标签和只含噪音信息的标签,依次视为正类样本和负类样本。

其中只含噪音信息的标签数量为 2500 个,相比只含噪音信息的标签数量,只含正文内容的标签数量明显偏多。由于正负类样本数量相差过多会降低 SVM 模型的识别精度,本文从数量偏多的正类样本中随机选出 2500 个标签,使正类样本和负类样本的数量保持 1: 1,从而提高模型识别噪音信息的精确性。下面选取含噪音信息标签的 5 个重要特征:

1) 噪音词汇指数 T。

2) 特殊标点指数 S。

3) 标签文本长度 L。

4) 标签位置 P。

5) a 标签与 p 标签的文本长度的比值 R。

记最后训练样本的 5 维特征向量为(T,S,L,P,R) 。下面对这 5 个特征(噪音词汇指数,特殊标点指数,标签文本长度,标签的位置,a 标签与 p 标签的文本长度比值) 依次作出说明。

4.2　噪音词汇指数

通过观察,发现一些词汇在噪音信息中经常出现,而在正文信息中极少出现。例如 “链接” “扫一扫” “声明”等。本文把这些词汇称为噪音词汇。噪音词汇是判断标签内容是否为噪音信息的重要依据。为了计算噪音词汇指数,首先要判别哪些词汇是噪音词汇,为此需要建立一个噪音词汇库。

4.2.1　建立噪音词汇库

首先对实验中选出的 2500 个只含噪音信息标签的文本内容进行中文分词操作,目的是从这些词汇中挑选出用于识别噪音信息的词汇。本文中,中文分词并不是主要的研究对象,由于篇幅有限,中文分词的内容不在这里展开讨论。

实验中采用开源的中文分词工具 ICTCLAS。ICTCLAS 是一种基于多层隐马尔科夫模型的汉语词法分析系统,主要功能包括中文分词、词性标注等功能,是一种分词精度极高的分词工具。

首先对 2500 条噪音信息进行分词处理,得到 t_1, t_2, t_3, …, t_n。然后依次计算出现词汇 t_i 时,判断为噪音信息的概率 p_i,见公式(2) 。其中 n_i 为词汇 t_i 在噪音信息样本中出现的次数,h_i 为 t_i 在正文信息样本中出现的次数,公式中的比值越接近 1,说明该词汇越频繁地出现在噪音信息中,而极少出现在正文信息中。

$$p_i = \frac{n_i}{n_i + h_i} \tag{2}$$

设定阈值为 α，当 $p_i \geqslant \alpha$ 时，将该词收录进噪音词汇库。此外当标签中含有邮箱地址、网址、电话号码时，标签内容是噪音信息的概率也会增大。把邮箱地址、网址、电话号码作为额外的 3 个特殊噪音词汇。

4.2.2 计算标签的噪音词汇指数

现在已经有了噪音词汇库。假设要识别的 HTML 标签中出现了噪音词汇 t_i，采用公式(3)计算 t_i 的权重值。W_i 为词汇 t_i 的权重值，f_i 为词汇 t_i 在该标签中出现的次数。例如将要识别的标签为:

< p > 更多信息请查看下方链接 < /p >

假设噪音词汇 t_i 为“链接”，f_i 即为“链接”在该 p 标签中出现的次数，在这个标签中“链接”出现的次数为 1，所以此时 f_i 等于 1。

n_i 为词汇 t_i 在噪音信息样本中出现的次数，h_i 为词汇 t_i 在正文信息样本出现的次数。比值 n_i/h_i 越大说明这个词汇在噪音信息中出现的次数越多，在正文中出现的次数越少，说明用该词识别噪音信息的准确率越高。

$$W_i = f_i \times \log_2\left(\frac{n_i}{h_i}\right) \tag{3}$$

假设要识别的 HTML 标签中共出现了 a 个噪音词汇 t_1，t_2，…，t_a。采用公式(4)计算噪音词汇指数 T，即对该标签中出现的每个噪音词汇的权重值求和。其中 T 为指数值，W_i 为噪音词汇 t_i 的权重值。

$$T = \sum_{i=1}^{a} W_i \tag{4}$$

4.3 特殊标点指数

特殊标点指在正文信息中出现频率低，但在噪音信息中出现频率高的标点符号，例如标点 <、>、/、【、[{、|、(、（、\、_、:、：等，记为 s_1，s_2，…，s_n。同时由于噪音文本的结尾常常不带句号，所以当标签的文本结尾不是句号时，本文将该情况视为出现了额外的一个特殊标点。

假设要识别的标签中出现了特殊标点 s_i，仍然采用公式(3)计算标点 s_i 的权重值。其中 W_i 为标点 s_i 的权重值，f_i 为标点 s_i 在标签中出现的次数，n_i 为标点 s_i 在噪音信息样本中出现的次数，h_i 为标点 s_i 在正文信息样本中出现的次数。比值 n_i/h_i 越大说明用该标点识别噪音信息的准确率越高。

假设要识别的 HTML 标签中共出现了 a 个特殊标点 s_1，s_2，…，s_a。采用公式(5)计算该标签的特殊标点指数 S，即对该标签中出现的每个特殊标点权重值求和。其中 S 为指数值，W_i 为 s_i 的权重值。

$$S = \sum_{i=1}^{a} W_i \tag{5}$$

4.4 标签文本长度

观察大量的网页，发现 HTML 标签中文本长度越小，标签内容是噪音信息的概率越大；文本长度越大，标签内容是正文信息的概率越大，因此把文本长度作为第 3 个识别噪音信息的特征。当识别一个 HTML 标签时，记标签的文本长度为 L。

4.5 标签的位置

通过大量观察，发现极短文本也常作为段落的小标题存在于正文中；而短文本作为噪音信息时，通常出现在文章的尾部，由此可见标签的位置也是识别噪音信息的重要特征。采用公式(6)计算标签的相对位置 P，需要注意的是这时已经定位到全部正文在某个确定的 div 标签里，将 div 标签包含的全部标签按顺序排序，n 为该 div 标签包含的 HTML 标签总数，m 为当前要识别 HTML 标签的序号。

$$P = \frac{m}{n} \tag{6}$$

4.6 a 标签与 p 标签的文本长度比值

当 p 标签嵌套 a 标签时，计算 a 标签与 p 标签的文本长度比值。通过大量统计，比值越接近 1 时，p 标签中的内容是噪音信息的概率越大。当 p 标签没有嵌套 a 标签时，默认该比值为 0。当识别一个 a 标签或其他 HTML 标签时，默认该比值为 1。

5 使用 SVM 生成数据模型

1) 数据归一化。对输入 SVM 的训练样本矩阵进行数据归一化处理，映射到 [0,1] 的范围里。数据归一化可以使最优解的寻优过程变得更平缓，更容易正确地收敛到最优解。

2) 确定训练矩阵。将上述 5 种重要的特征作为输入向量，从而形成训练数据的训练矩阵。

3) 确定核函数。实验中分别将线性核函数、多项式核函数、径向基核函数 3 种不同的核函数应用于模型进行对比实验。实验结果表明径向基核函数在实验中的效果较好，因为实验中训练样本的特征数量较少，径向基核函数将样本映射到一个更高维的空间，可以处理当类标签和特征之间的关系是非线性时的样例。

4) 确定参数。对于径向基核函数，其参数的选取直接影响着 SVM 分类器的性能。选择采用 10 折交叉验证法。将数据集分成 10 份，轮流将其中 1 份作为测试数据，另外 9 份作为训练数据，进行实验，以得到最优参数。

6 利用数据模型进行降噪

6.1 噪声过滤器的实现

通过大量观察，当噪音信息出现在一个文本长度

较长的标签时,正文与噪音信息通常掺杂在一个标签中。而当标签的文本长度较短时,通常不会出现上述情况,标签内容或者全部为正文,或者全部为噪音信息。所以降噪时,需要根据标签文本长度分 2 种情况处理。本文设计了一个噪声过滤器,当标签文本长度大于阈值 β 时采用噪声过滤器,剔除标签中的噪音数据,保留标签中的正文内容。具体实现过程见图 3,采用递归算法,整体步骤如下:

1) 由于人们书写文章的习惯,噪声信息往往在标签中文本的尾部,所以本文以句号为分割标识,每次裁剪标签文本尾部的最后一句话,采用 SVM 训练出的数据模型进行识别。

2) 如果最后一句话被识别为噪音信息,则递归,输入剩余文本,并再次调用噪声过滤器,裁剪剩余文本的最后一句话,使用数据模型判断。如果判断为正文,保存正文文本部分,并跳出循环。如果判断为噪音信息,则继续递归,重复上述步骤,直到剩余文本长度为 0 时跳出递归。

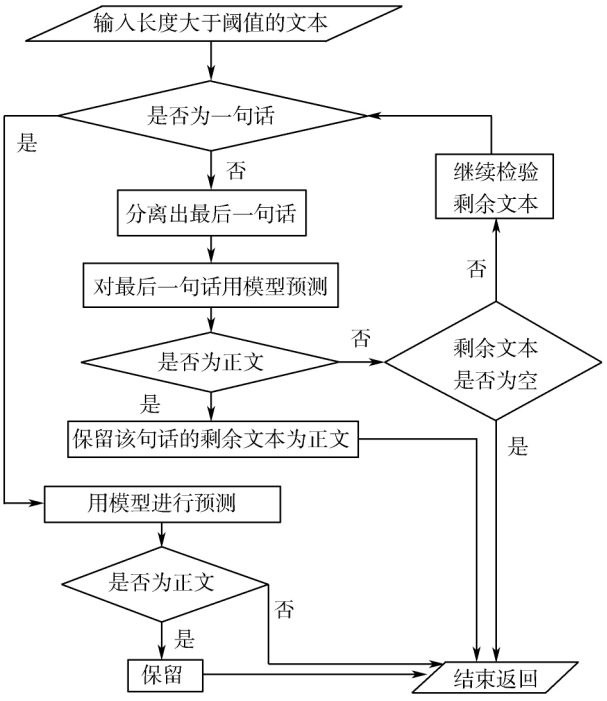

图 3　噪声过滤器的实现流程图

例如下面这个 HTML 标签:

< p > 万豪相关负责人表示已深刻认识问题严重性,将认真配合相关调查,回应社会关切。更多信息请点击下方链接 < /p >

设该 p 标签的文本长度大于阈值,则在噪音信息过滤器中,首先截取 p 标签中的最后一句话,即“更多信息请点击下方链接”,提取 5 个特征属性,然后使用数据模型对该语句进行识别。假设识别结果为噪音信息,则递归,继续识别剩余文本,即“万豪相关负责人表示已深刻认识问题严重性,将认真配合相关调查,回应社会关切。”假设识别结果为正文信息,则跳出递归。

6.2　降噪过程的整体逻辑

降噪的整体逻辑见图 4,需要注意的是这时已经定位到全部正文在某个确定的 div 标签里。具体步骤为:

1) 采用深度优先的方式遍历 div 标签中所有标签。

2) 考虑到表格数据作为正文的一部分,保留 td 标签的文本; 识别 p 标签和 a 标签时,将根据文本长短用 2 种方式分别处理。其他标签直接剔除。

3) 当标签文本长度≤阈值 β 时,直接用 SVM 训练得出的模型识别,当标签文本长度 > β 时采用噪声过滤器,剔除标签中的噪音数据,保留标签中的正文内容。

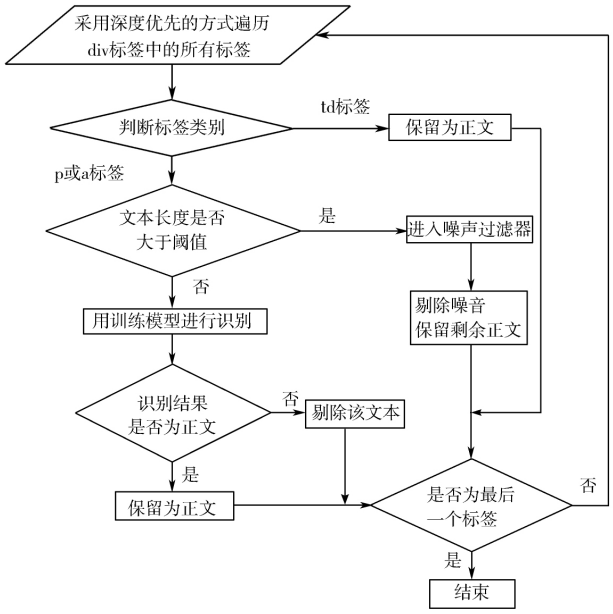

图 4　降噪的整体逻辑流程图

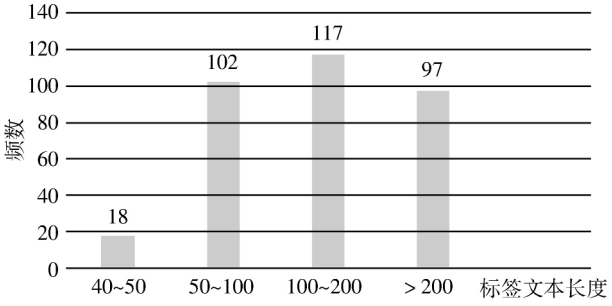

图 5　不同文本长度标签出现的频数图

需要说明的是,若阈值 β 选取过大,SVM 数据模型可能将整个 HTML 标签内容识别为噪音信息,增加误删正文的概率。若阈值 β 选取过小,会减少误删正文的概率,但会影响程序的效率。本文统计了训练

集中正文与噪音信息掺杂在一个标签时，不同文本长度标签出现的频数。统计结果见图 5。根据统计结果将阈值设定为 40。这样可以较好地平衡降噪精确性和程序效率。

7 实验结果与分析

实验中将阈值 θ、α、β 分别设定为 0.3、0.8、40。选取互联网中 5 个最流行的网站：百度、网易、新浪、腾讯、凤凰网。每个网站选取 200 个网页作为测试集。

实验的降噪过程依赖于成功定位到包含全部正文的标签，只有成功定位到包含全部正文的标签，降噪过程才能正常进行。采用公式(7)和公式(8)计算正文定位成功率 P 和降噪准确率 R。

$$P = \frac{C_2}{C_1} \tag{7}$$

$$R = \frac{C_3}{C_1} \tag{8}$$

上述公式中 C_1 表示实验的网页总数，C_2 表示正文定位成功的网页个数，C_3 表示准确降噪的网页个数。定位成功率和降噪准确率均以实验的网页总数 C_1 为前提。实验结果见表 1。

表 1 正文抽取实验结果

网页来源	网页总数	正文定位成功数	成功率/%	降噪准确数	准确率/%
百度	200	195	97.5	190	95
网易	200	196	98	190	95
新浪	200	199	99.5	193	96.5
腾讯	200	198	99	189	94.5
凤凰网	200	196	98	186	93
总计	1000	984	98.4	948	94.8

实验结果表明本文方法对大部分网页的正文提取效果较好，不依赖于网页的风格和样式。由于本文方法中采取“宽进严出”的策略，第 1 步定位到全部正文信息的概率较大，在测试的 5 大网站中，定位到全部正文的成功率均达到 97.5% 以上。

SVM 数据模型可以有效地去除与正文无关的噪音信息，对于引用文献方法中经常误删的与正文相关的链接、短文本等，数据模型可以有效识别为正文并保留。实验结果表明本文方法具有较高的降噪准确率。表 2 为本文方法与文献[1]、文献[2]和文献[5]的实验结果对比。

表 2 本文与其他文献的实验结果对照 单位：%

方法	平均准确率	百度	网易	新浪	腾讯	凤凰网
本文	94.8	95	95	96.5	94.5	93
文献[1]	94.5	96.5	95	94	93	94
文献[2]	93.7	91	90.5	98.5	97	91.5
文献[5]	93.3	92.5	93.5	93.5	94	93

文献[1]中的方法适用于多种类型的网站，但由于段落标题的标点特征与噪音信息相似，该方法易将网页中的段落标题误识别为噪音信息。本文方法对段落标题通常可以准确识别为正文。

文献[2]中的方法主要依赖于网页 DOM 树节点路径相似度，所以对一些网页结构相对规律的网页正文抽取准确率较高。例如新浪网站中网页结构相对规律，该方法抽取正文准确率较高。凤凰网和百度网站中的网页结构相对复杂，该方法抽取正文准确率相对偏低。

文献[5]中的方法主要依赖正文词汇与非正文词汇的不同特征，该方法主要基于文本自身特征，并不依赖于网页的结构，适用于不同风格的网站。由于本文中的方法部分依赖于网页结构，对于网页结构相对复杂的网页，文献[5]的正文抽取效果较好。但本文方法融合了其他重要特征，所以平均的正文抽取准确率更高。

8 结束语

本文中的方法相比传统的正文信息提取方法提高了精度。从实验结果来看，本文所采用的方法适用于大多数网站，但也有个别网页出现正文定位失败的情况，原因在于实验所采用的 HTML 解析器无法正常解析个别网页的网页源码，从而导致定位失败。此外不同类别文章的噪音信息特点也略有差异，比如财经类文章与汽车类文章。如果能够使用文本分类技术将文章准确分类，结合文章类别进行降噪，降噪准确率有继续提高的可能。

参考文献：

[1] 胡露露，刘小勤，孙凯. 基于正文特征和网页结构的网页正文抽取方法[J]. 大气与环境光学学报，2017，12(3)：230-235.

[2] 潘心宇，陈长福，刘蓉，等. 基于网页 DOM 树节点路径相似度的正文抽取[J]. 微型机与应用，2016，35(19)：74-77.

[3] 宋明秋，张瑞雪，吴新涛，等. 网页正文信息抽取新方法[J]. 大连理工大学学报，2009，49(4)：594-597.

[4] Yang Xiudan，Zhu Yuanyuan. Ontology-based information extraction system in e-commerce websites[C]// Proceedings of the 2011 International Conference on Control，Automation and Systems Engineering. 2011，doi：10.1109/ICCASE.2011.5997640.

[5] 刘鹏程，胡骏，吴共庆. 基于文本块密度和标签路径覆盖率的网页正文抽取[J/OL]. http://www.arocmag.com/article/02-2018-06-004.html，2017-06-14.

(下转第 31 页)

参考文献:

[1] Xi Xue-cheng, Poo A N, Chou S K. Support vector regression model predictive control on a HVAC plant [J]. Control Engineering Practice, 2007,15(8):897-908.

[2] Iplikci S. Support vector machines based neuro-fuzzy control of nonlinear systems [J]. Neurocomputing, 2010,73(10-12):2097-2107.

[3] 金晶,王行愚,罗先国,等. PSO-ε-SVM 的回归算法 [J]. 华东理工大学学报(自然科学版), 2006,32(7):872-875.

[4] 任江涛,赵少东,许盛灿,等. 基于二进制 PSO 算法的特征选择及 SVM 参数同步优化 [J]. 计算机科学, 2007,34(6):179-182.

[5] 于化龙,顾国昌,刘海波,等. 改进的离散 PSO 和 SVM 的特征基因选择算法 [J]. 哈尔滨工程大学学报, 2009,30(12):1399-1403.

[6] 姚全珠,蔡婕. 基于 PSO 的 LS-SVM 特征选择与参数优化算法 [J]. 计算机工程与应用, 2010,46(1):134-136.

[7] 陈治明. 改进的粒子群算法及其 SVM 参数优化应用 [J]. 计算机工程与应用, 2011,47(10):38-40.

[8] 单黎黎,张宏军,王杰,等. 一种改进粒子群算法的混合核 ε-SVM 参数优化及应用 [J]. 计算机应用研究, 2013,30(6):1636-1369.

[9] 刘春卫,罗健旭. 基于混合核函数的 PSO-SVM 分类算法 [J]. 华东理工大学学报(自然科学版), 2014,40(1):96-101.

[10] 宁爱平,张雪英,刘俊芳. ABC-PSO 算法优化混合核 SVM 参数及应用 [J]. 数学的实践与认识, 2014,44(18):158-165.

[11] 刘佳,施龙青,韩进,等. 基于 Grid-Search_PSO 优化 SVM 回归预测矿井涌水量 [J]. 煤炭技术, 2015,34(8):184-186.

[12] Kennedy J, Eberhart R. Particle swarm optimization [C] // Proceedings of the 1995 IEEE International Conference on Neural Networks. 1995,4:1942-1948.

[13] 刘俊芳,高岳林. 带自适应变异的量子粒子群优化算法 [J]. 计算机工程与应用, 2011,47(3):41-43.

[14] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York: Springer, 1995.

(上接第 26 页)

[6] 罗永莲,赵昌垣,贾玉芳,等. 基于朴素贝叶斯 Web 新闻内容的抽取方法 [J]. 计算机与现代化, 2016(1):59-63.

[7] 赵欣欣,索红光,刘玉树. 基于标记窗的网页正文信息提取方法 [J]. 计算机应用研究, 2007,24(3):144-145.

[8] Zhu Ningbo, Zheng Bijuan, Zhang Chunfeng. An edge and filter based morphological text extracting method [C] // Proceedings of the 2010 4th International Conference on Intelligent Information Technology Application. 2010.

[9] 李蕾,王劲林,白鹤,等. 基于 FFT 的网页正文提取算法研究与实现 [J]. 计算机工程与应用, 2007,43(30):148-151.

[10] 蒋亚平,梅骁. 基于支持向量机与人工免疫系统的垃圾邮件过滤模型 [J]. 现代计算机, 2016(11):55-57.

[11] 王祖辉,姜维. 基于支持向量机的垃圾邮件过滤方法 [J]. 计算机工程, 2009,35(13):188-189.

[12] 张洁. 改进支持向量机的电子邮件分类 [J]. 现代电子技术, 2017,40(1):77-79.

[13] Bao Jianmin, Pan Lin, Xie Yuanfa. A new BDI forecasting model based on support vector machine [C] // Proceedings of the 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference. 2016:65-69.

[14] 姚潇,余乐安. 模糊近似支持向量机模型及其在信用风险评估中的应用 [J]. 系统工程理论与实践, 2012,32(3):549-554.

[15] 郭晓云. ICTCLAS 中文词法分析的 Delphi 调用研究 [J]. 电脑编程技巧与维护, 2011(24):10-11.

[16] 刘克强. 2009 共享版 ICTCLAS 的分析与使用 [J]. 科教文汇(上旬刊), 2009(8):271.

[17] 罗燕,赵书良,李晓超,等. 基于词频统计的文本关键词提取方法 [J]. 计算机应用, 2016,36(3):718-725.

[18] 赵胜辉,李吉月,徐碧瑢,等. 基于 TFIDF 的社区问答系统问句相似度改进算法 [J]. 北京理工大学学报, 2017,37(9):982-985.

[19] Jiang Hao, Li Wen Qiang. Improved algorithm based on TFIDF in text classification [J]. Advanced Materials Research, 2012,403-408:1791-1794.

[20] Drożdż M, Kryjak T. FPGA implementation of multi-scale face detection using HOG features and SVM classifier [J]. Image Processing and Communications, 2016,21(3):27-44.

[21] Sharma A, Dey S. A boosted SVM based ensemble classifier for sentiment analysis of online reviews [J]. ACM SIGAPP Applied Computing Review, 2013,13(4):43-52.

[22] Cervantes J, García Lamont F, López-Chau A, et al. Data selection based on decision tree for SVM classification on large data sets [J]. Applied Soft Computing, 2015,37:787-798.