Journal of Computer Applications
计算机应用,2021,41(1):133 - 138
ISSN 1001-9081
CODEN JYIIDU
2021-01-10
http://www.joca.cn

文章编号:1001-9081(2021)01-0133-06
DOI:10.11772/j.issn.1001-9081.2020081379

基于注意力机制和集成学习的网页黑名单判别方法

周超然,赵建平*,马　太,周　欣
(长春理工大学 计算机科学技术学院,长春 130022)
(*通信作者电子邮箱jpzhao@yeah.ne)

摘　要:搜索引擎作为互联网主要应用之一,能够根据用户需求从互联网资源中检索并返回有效信息。然而,得到的返回列表往往包含广告和失效网页等噪声信息,而这些信息会干扰用户的检索与查询。针对复杂的网页结构特征和丰富的语义信息,提出了一种基于注意力机制和集成学习的网页黑名单判别方法,并采用本方法构建了一种基于集成学习和注意力机制的卷积神经网络(EACNN)模型来过滤无用的网页。首先,根据网页上不同种类的HTML标签数据,构建多个基于注意力机制的卷积神经网络(CNN)基学习器;然后,采用基于网页结构特征的集成学习方法对不同基学习器的输出结果执行不同的权重计算,从而实现EACNN的构建;最后,将EACNN的输出结果作为网页内容分析结果,从而实现网页黑名单的判别。所提方法通过注意力机制来关注网页语义信息,并通过集成学习的方式引入网页结构特征。实验结果表明,与支持向量机(SVM)、K近邻(KNN)、CNN、长短期记忆(LSTM)网络、GRU、结合注意力机制的卷积神经网络(ACNN)等基线模型相比,所提模型在所构建的面向地理信息领域的判别数据集上具有最高的准确率(0.97)、召回率(0.95)和F_1分值(0.96),验证了EACNN在网页黑名单判别工作中的优势。

关键词:网页黑名单;判别模型;网页结构特征;语义信息;注意力机制;集成学习;深度学习
中图分类号:TP391.1　　**文献标志码:**A

Web page blacklist discrimination method based on attention mechanism and ensemble learning

ZHOU Chaoran, ZHAO Jianping*, MA Tai, ZHOU Xin
(*School of Computer Science and Technology*, *Changchun University of Science and Technology*, *Changchun Jilin* 130022, *China*)

Abstract: As one of the main Internet applications, search engine can retrieve and return effective information from Internet resources according to user needs. However, the obtained returned list often contains noisy information such as advertisements and invalid Web pages, which interfere the user's search and query. Aiming at the complex structural features and rich semantic information of Web pages, a Web page blacklist discrimination method based on attention mechanism and ensemble learning was proposed. And, by using this method, an Ensemble learning and Attention mechanism-based Convolutional Neural Network (EACNN) model was built to filter useless Web pages. First, according to different categories of HTML tag data on Web pages, multiple Convolutional Neural Network (CNN) base learners based on attention mechanism were established. Second, an ensemble learning method based on Web page structural features was used to perform different weight computation to the output results of different base learners to realize the construction of EACNN. Finally, the output result of EACNN was used as the analysis result of Web page content to realize the discrimination of Web page blacklist. The proposed method focuses on the semantic information of Web pages through attention mechanism, and introduces the structural features of Web pages through ensemble learning. Experimental results show that, compared with baseline models such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), CNN, Long Short-Term Memory (LSTM) network, Gate Recurrent Unit (GRU) and Attention-based CNN (ACNN), EACNN has the highest accuracy (0.97), recall (0.95) and F_1 score (0.96) on the geographic information field-oriented discrimination dataset constructed. It verifies the advantages of EACNN in the task of discriminating Web page blacklist.

Key words: Web page blacklist; discrimination model; Web structural feature; semantic information; attention mechanism; ensemble learning; deep learning

0　引言

在互联网的实际应用中,当用户使用搜索引擎进行信息检索时,出于商业、用户点击量、规范性差及网络安全系统不完善等原因会导致返回数据列表中包含部分不符合用户搜索目标的结果网页。比如,房地产及招聘广告网页、新闻资讯网

收稿日期:2020-07-31;**修回日期:**2020-10-15;**录用日期:**2020-10-29。
基金项目:吉林省科技发展计划项目(20190303133SF);吉林省教育厅"十三五"科学技术项目(JJKH20200796KJ)。
作者简介:周超然(1994—),男,吉林吉林人,博士研究生,主要研究方向:数据挖掘、城市计算;　赵建平(1964—),男,吉林榆树人,教授,博士,主要研究方向:数据挖掘、计算机网络;　马太(1996—),男,河北秦皇岛人,硕士研究生,主要研究方向:城市计算、深度学习;　周欣(1997—),女,吉林磐石人,硕士研究生,主要研究方向:自然语言处理、行人建模与仿真。

页、失效网页、非法网页等不包含用户需求信息的网页。现有网页黑名单识别技术应对新威胁的速度太慢，使得超过90%的访问者可以在页面被列入黑名单之前对其进行查看[1]。过分或不适当的网络资源访问不仅给网络运维厂商带来生产力的负担和传输资源的浪费，网络上的违法内容甚至会给用户带来经济、精神等方面的损失，严重影响网络环境[2]。本研究的目的是在使用搜索引擎检索信息时，对搜索引擎返回的网页内容进行分析，保留有效信息，提升搜索引擎对用户的适用性。

本文在网页黑名单判别任务中，需解决如下问题：1）网页具有复杂的结构信息和语义信息，实现合理的网页特征构建来表达网页信息是一个关键问题；2）由于网页数据复杂性很高，选择哪种技术来构建判别模型，并保证判别模型的鲁棒性和准确性也是十分关键的。针对上述问题，本文通过网页动态内容分析提出了一种基于注意力机制和集成学习的网页黑名单判别方法，并训练了基于集成学习和注意力机制的卷积神经网络（Ensemble learning and Attention mechanism-based Convolutional Neural Network，EACNN）模型。EACNN将网页HTML标签中的文本数据转化为词向量，采用基于注意力机制的卷积神经网络（Attention-based Convolutional Neural Network，ACNN）提取文本特征。通过不同标签的样本集来训练基学习器的构建。再采用Bagging集成学习方法将各标签的基学习器的结果选择优化输出权重，完成集成学习器的构建。将集成学习器的输出作为该网页是否为黑名单的判别结果。EACNN相较于其他相关工作优势如下：

1）基于神经网络的判别模型。相较于统计学习方法（*K*近邻（*K*-Nearest Neighbor，*K*NN）[3]、支持向量机（Support Vector Machine，SVM）[4]等，神经网络模型[5-9]的准确度更高，学习能力更强，能实现良好的非线性映射关系并具有很强的鲁棒性。

2）引入注意力机制。对文本转化后的嵌入向量执行注意力计算，提升对判别结果影响力大的词汇的关注度，降低对判别结果影响力小的词汇关注度，使文本数据的特征表达更加合理、充分。

3）基于网页结构特征的集成学习。考虑网页中不同标签数据对网页的信息涵盖程度不同，为引入网页结构特征，采用集成学习将不同网页标签基学习器的判别结果进行整合，构建基于集成学习和注意力机制的卷积神经网络（EACNN）模型。实验结果表明EACNN模型性能优于其他基线模型，证明了采用集成学习方法引入网页结构特征的有效性。

1 相关工作

网页分析从20世纪90年代初的专业管理人员甄别，逐步向后期的关键词检索分析，发展到现在已经开始采取启发式内容分析方法。网页分析的主要目标是获取包含有效信息的网页并剔除无关的网页，即实现黑名单网页的判别。早期的网页分析方法以关键词检索、特征设计和引入知识库、统计学习等技术为主要手段。文献[10]基于关键词匹配构建混合模型来过滤网页中的色情文本。Sheu等[11]通过设计网页特征结合决策树算法从网页集合中区分医疗类网页和情色类网页。徐雅斌等[12]基于*K*近邻算法来过滤互联网资源中的不良网页。顾敏等[13]基于Naive Bayesian Model结合网页结构特征，采用多特征融合的方法进行网页分类。Kan等[14]将URL作为特征实现快速的网页分类。由于网页包含很多非结构化信息，完全依赖于人工设计特定的方式来分析网页是不合理的；而单纯采用关键词检索和统计学习方法在非线性特征方法的表达能力较弱，难以生成准确性高、鲁棒性好的模型。

随着深度学习的不断发展，基于神经网络的网页分析模型开始成为了主流。邓玺[5]基于深度卷积神经网络搭建了网页特征提取模型并实现了网页类别划分工作。Buber等[6]基于循环神经网络（Recurrent Neural Network，RNN）的深度学习架构并将标题、描述和关键字的元标签信息作为特征来构建网页分类模型。另外有许多基于深度神经网络的方法[7-9]被用来解决文本分类判别问题，尽管它们没有考虑网页特征，但为基于深度学习的网页黑名单判别模型的研究开展起到了铺垫和推进作用。

本文在深度神经网络分类模型的基础上引入注意力机制，在提升网页语义信息表达能力的同时完成基础学习器的构建。注意力机制在本文的应用方式为，以人的关注要点行为依据设计网页文本嵌入向量的注意力计算。注意力机制在深度神经网络上的应用，最早应用在图像处理领域[15]，如今已广泛应用于自然语言处理[16]、语音识别[17]、模型设计[18]等领域。ACNN[19]将深度神经网络中的CNN引入注意力机制，是注意力机制与神经网络结合来解决文本分类任务的探索性工作。

Zhou等[20]发现，对多个网络进行随机赋权，并利用遗传算法对权值进行进化，在一定程度上能够提升神经网络的表现力。基于上述贡献，采用类似Zhou等[20]工作的集成学习方式来引入网页结构特征，将不同标签数据训练的基学习器进行集成，实现网页黑名单的判别模型构建。集成学习通过构建并结合多个学习器来完成学习任务。Bagging[21]是并行式集成学习方法最著名的代表，基于自助采样法，利用不同的采样集训练出多个学习器，再将这些基学习器进行结合。Bagging通常对分类任务使用简单投票法，对回归任务使用简单平均法。蒋芸等[22]提出了基于Bagging的概率神经网络集成分类算法，该算法在分类误差、准确率、泛化性以及执行速度方面优于传统BP（Back Propagation）神经网络。鉴于引入注意力机制并结合网页特征的集成深度学习模型在特征提取方面的优势，本文设计基于注意力机制和集成学习的网页黑名单判别方法。

2 网页黑名单判别方法设计

对网页信息表达而言，不同标签中的文本数据对用户理解信息的影响程度是不同的，相较于子标签，标题和摘要标签中的数据对网页信息的概述效果更好，显然更便于用户理解网页信息。基于上述思想，本文提出一种基于注意力机制和集成学习的网页黑名单判别方法，此方法被用于构建基于集成学习和注意力机制的卷积神经网络（EACNN）模型。EACNN结构如图1。

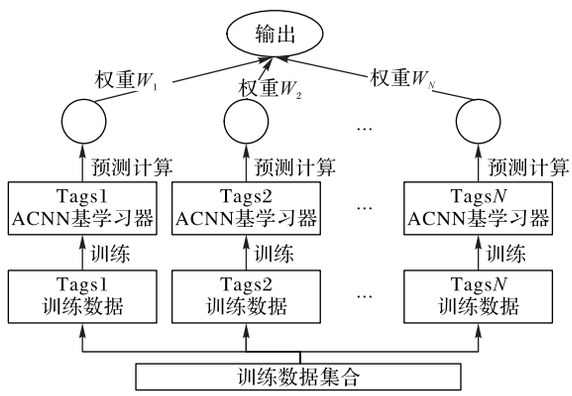


图1 EACNN结构

Fig. 1 EACNN structure

主要内容如下：首先，将网页HTML数据根据标签类型抽

样为若干个子训练集;然后,对不同标签(Tags)的训练子集采用基于注意力机制的CNN(ACNN)来构建基学习器;最后,通过集成学习的方法对不同的基学习器赋予一个不同的权重W_k,实现网页黑名单的判别输出。

2.1　基于注意力机制的CNN基学习器

基于注意力机制的CNN(ACNN)网页判别模型结构如图2所示。

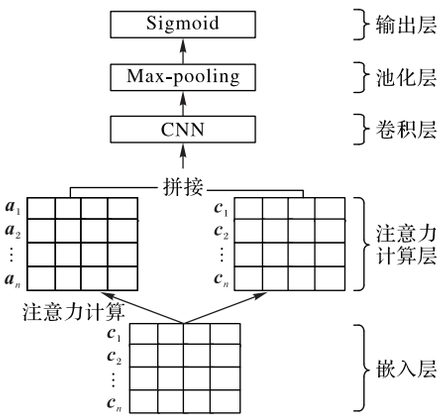


图2　ACNN网页判别模型

Fig. 2　Web page discrimination model of ACNN

嵌入层　将文本映射到多维实数空间,实现语义表达。基于Word2vec的Skip-gram策略预训练词嵌入模型[23],用于将原始输入文本序列转化成嵌入矩阵,并将嵌入矩阵作为注意力计算层的输入。

注意力计算层　对嵌入矩阵进行局部注意力的计算,获取文本序列的注意力信息来生成注意力矩阵。卷积层的输入矩阵为嵌入矩阵和注意力矩阵的拼接结果。

卷积层　对输入的嵌入矩阵执行卷积操作,提取特征信息。

最大池化层　对卷积结果矩阵进行特征的再提取,减少模型参数量的同时提高模型的鲁棒性。

输出层　数据归一化处理。对最大池化层输出的特征矩阵采用Sigmoid函数计算输出结果,完成网页是否为黑名单的判别。

2.1.1　词嵌入

嵌入层采用Jieba(https://github.com/fxsjy/jieba)开源工具进行删除停用词和中文分词的预处理工作,并基于Word2vec方法的Skip-gram模型实现词向量的构建。嵌入层通过Word2vec嵌入模型将输入文本映射到多维实数空间上,构建嵌入向量来表达文本特征和语义内容。表达方式为:将一个词汇长度为n的句子通过预训练的词嵌入模型,生成输入矩阵$\boldsymbol{X}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n\}$。其中输入矩阵$\boldsymbol{X}$为一个$n*d$的矩阵,$n$代表输入文本的词长度,$d$代表词向量长度,$\boldsymbol{c}_i$表示句子中第$i$个词汇的嵌入向量。嵌入表达采用基于Wikipedia_zh中文维基百科语料贡献的词向量模型[24]。词向量模型的基本设置:动态窗口大小为5;消极采样为5;迭代次数为5;低频词汇为10;二次采样概率为1E-5。

2.1.2　注意力计算

注意力计算层通过引入注意力机制来提升嵌入矩阵的特征表达能力。注意力计算层为更好地关注与判别操作紧密相关的关键词,设计了自注意矩阵。采用滑动窗口来计算局部嵌入矩阵的权值,滑动窗口大小为j,每个窗口的权值是不共享的。为保证窗口的中心词都是原始向量中矩阵的词,实现覆盖全部嵌入向量,在输入矩阵首尾各加入$(j-1)/2$个随机初始向量。随后执行局部嵌入重要程度的计算。

评价局部嵌入重要程度的计算公式如下:

$$\boldsymbol{s}_i = f(\boldsymbol{W}_{\text{att}} \cdot \boldsymbol{X}_{i:i+j-1} + b_{\text{att}}) \tag{1}$$

其中:$\boldsymbol{s}_i$为窗口中心词的重要程度,$\boldsymbol{X}_{i:i+j-1}$为输入的第i个到第$i+j-1$个窗口内的嵌入矩阵,$\boldsymbol{W}_{\text{att}}$为输入词的注意力权值矩阵,$b_{\text{att}}$为注意力偏置值,$f()$表示Sigmoid激活函数。

关键词阈值设定公式:

$$\lambda = \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{s}_i|$$

通过将词的重要程度$\boldsymbol{s}_i$的大小与阈值λ的对比,得到词的关键向量$\boldsymbol{a}_i$,定义如下:

$$\boldsymbol{a}_i = \begin{cases} \boldsymbol{0}, & |s_i| \leqslant \lambda \\ \boldsymbol{c}_i, & |s_i| > \lambda \end{cases} \tag{2}$$

$\boldsymbol{a}_i$保留了重要程度更高的词汇,并将低于平均影响力的词汇的权重设置为零向量($\boldsymbol{0}$)。其中$\boldsymbol{c}_i$为词的原始输入向量,将$\{\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n\}$拼接得到经过自注意矩阵的$\boldsymbol{X}_{\text{att}}$。矩阵$\boldsymbol{X}$和矩阵$\boldsymbol{X}_{\text{att}}$拼接构成卷积层的输入矩阵$\boldsymbol{X}_{\text{con}}$。

2.1.3　卷积与预测

卷积层通过对输入矩阵的卷积操作来提取文本的局部特征,基学习器的卷积运算如下:

$$\boldsymbol{co}_i = f\left(\sum \boldsymbol{W}_{\text{con}} \cdot \boldsymbol{X}_{\text{con},i:i+h-1} + b\right) \tag{3}$$

其中:$\boldsymbol{co}_i$即卷积层的特征提取计算结果,$\boldsymbol{co}_i$(i=1,2,$\cdots$,$n-h+1$)表示卷积运算后的结果矩阵,h为窗口大小,$\boldsymbol{X}_{\text{con},i:i+h-1}$表示卷积层输入的第$i$个到第$i+h-1$个窗口内的矩阵,$\boldsymbol{W}_{\text{con}}$为权值矩阵,$b$为偏置值,$f()$表示Sigmoid激活函数。

然后采用最大池化操作,对卷积操作提取的特征进行压缩并提取主要特征,将池化得到的最大值进行拼接,得到一条一维特征向量$\hat{\boldsymbol{c}}$的池化操作如下。

$$\hat{\boldsymbol{c}} = \max\{\boldsymbol{co}_1, \boldsymbol{co}_2, \cdots, \boldsymbol{co}_{n-h+1}\} \tag{4}$$

最后将池化后的输入矩$\hat{\boldsymbol{c}}$与判别神经元进行全连接操作,得到判别结果y,定义如下:

$$y = \text{Sigmoid}(\boldsymbol{W}_{\text{f}} \cdot \hat{\boldsymbol{c}} + b_{\text{f}}) \tag{5}$$

Sigmoid函数适用于二分类问题中将内部函数转化为概率函数。其中$\boldsymbol{W}_{\text{f}} \in \mathbf{R}^{1\times M}$为用于网页黑名单判别的1维矩阵,$b_{\text{f}}$为偏置值。

2.1.4　模型训练

ACNN是一个二分类模型,因此激活函数选择为Sigmoid,通过将空间矩阵映射到[0,1]区间,实现网页是否属于黑名单的预测。

Sigmoid计算公式为:

$$f(x) = \frac{1}{1+\mathrm{e}^{-x}} \tag{6}$$

代价函数选择交叉熵损失函数,定义如下:

$$L(\hat{y}, y) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}) \tag{7}$$

其中:y为实际类别标签值,$\hat{y}$是Sigmoid激活函数计算得到的判别结果,区间为[0,1]。在模型训练阶段采用梯度下降法来加快收敛并减少计算量,引入dropout策略[25]和k折交叉验证来防止过拟合。本文dropout和k折交叉验证的参数分别为0.5和3。

2.2　基于网页结构特征的集成学习器构建

考虑到不同的网页标签对网页信息的表现力不同,需要分析不同网页结构文本对网页判别的权重系数。将不同标签数据集设置为不同基学习器的训练集,每个训练集对应训练一个基学习器。与传统Bagging采用投票方式不同的是,本文为每个基学习器的输出赋予一个优化权重,即经过每个基学习器输出结果为原始输出与优化权重的乘积,总体结果为基

学习器输出结果累加和的平均数。具体计算过程如下。

将网页HTML数据集根据标签类别抽样成N个子数据集，表示为$T=\{T_1,T_2,\cdots,T_N\}$。标签类别数目与基学习器个数相同，均为N。第k类子数据集表示为$T_k=\{t_{k1},t_{k2},\cdots,t_{km}\}$，$t_{km}$表示第$k$类标签子数据集的第$m$条文本数据。

第k类基学习器对t_{km}的预测输出结果表示为$O_k(t_{km})$，经过集成学习概念对第m个网页是否为黑名单的判别结果计算公式为：

$$O(t_m)=\frac{1}{N}\sum_{k=1}^{N}W_kO_k(t_{km}) \tag{8}$$

其中$O(t_m)$表示对第m条网页进行集成计算的输出结果，即引入权重系数W_k表示第k个基学习器的输出权重。为保证集成学习器效果的优秀，需要对不同学习器的文本权重进行最优解计算。粒子群优化算法[26]具有对连续参数进行目标函数优化的能力，且具有搜索速度快、效率高，适合于实值型处理等方面的优点。因此本文采用粒子群优化算法（算法1）来实现集成学习器权重W的参数求解。

$$score(W)=\sum_{k=1}^{N}\sum_{i=1}^{m}rank(t_{ki}) \tag{9}$$

$$rank(t_{ki})=\begin{cases}1, & \hat{y}(t_{ki})=y(t_{ki})\\ 0, & \hat{y}(t_{ki})\neq y(t_{ki})\end{cases} \tag{10}$$

其中：$score(W)$表示当前输入文本设置权重对模型判别的影响效果，$W=\{W_1,W_2,\cdots,W_N\}$表示各个基学习器的输出权重集合。$rank(t_{ki})$表示基学习器k对第i个网页的判别效果，$\hat{y}(t_{ki})$表示基学习器k对第i个网页是否为黑名单的预测标记，$y(t_{ki})$表示第i个网页是否为黑名单的实际标记。

算法1 基于粒子群优化算法的EACNN权重设置方法。

输入 粒子数$Part_n$，迭代次数$Iter_n$，速度$Speed$，EACNN模型，网页的标记集合$Y=\{y_1,y_2,\cdots,y_m\}$。

过程：

1） 随机初始化$Part_n$个粒子集合，表示为$W_p=\{w_{p1},w_{p2},\cdots,w_{pn}\}$

2） 循环迭代次数$Iter_n$，操作

3） 循环每个粒子W_{pi}，操作

4） 利用粒子W_{pi}训练EACNN模型

5） 利用式（9）评价当前粒子的参数效果

6） 结束循环

7） 更新最佳粒子参数：$particle_{best}$为具有最高评价结果的粒子

8） 循环每个粒子W_{pi}，操作

9） 更新粒子参数：$W_{pi}\leftarrow W_{pi}+(particle_{best}-W_{pi})\times speed$

10） 结束循环

11） 结束循环

12） 返回最佳粒子参数$particle_{best}$

输出 最佳粒子参数$particle_{best}$。

算法1中每个粒子的训练参数表示基学习器的输出权重系数集合。算法1相关参数设置如下：粒子数$Part_n$=100，迭代次数$Iter_n$=10，更新权重$Speed$=0.2。

3 实验与结果

3.1 实验准备

基于爬虫技术从百度搜索引擎（https://www.baidu.com）中收集了4 390条地理信息类的网页数据作为实验数据集，并采用人工标注的方法对实验数据进行标记，其中正例2 790条、反例1 600条。将黄页、商业购物、投放广告、社区问答类网页标记为黑名单，蕴含地理信息数据的网页标记为白名单。被标记为正例的数据属于白名单数据，标记为反例的数据为黑名单数据。实验数据的训练集和测试集的数据数量比例为7∶3。每条数据包含如下属性：1）网页标题数据；2）网页元标签数据；3）网页URL数据；4）网页各级小标题数据；5）网页特殊字体数据；6）网页正文数据；7）网页正反类型的布尔数据，数据样本示例如图3。

序号	标题	描述	url	tags1	tags2	tags3	是否有效
1	求吉林长春净	最佳答案: 东北	http://www.bai	求吉林长春净	知道求吉林长	我来答经验	0
2	长春市实验中	长春市实验中	http://www.bai	长春市实验中	长春	长春	1
3	2018东北师范	3天前 - 学校B	http://www.bai	2018东北师范	东北师范大学	24 长春今：	晴 1
4	长春南关区宽	2017年10月30	http://www.bai	ZAKER 聚合了	墨西哥客机坠	环球网2小时前	0
5	吉林省长春市	吉林省长春市	http://www.bai	吉林省长春市	全部城市意见	【切换城市】	1
6	长春市第二实		http://www.bai	长春市第二实	百科长春市第	清除历史记录	1
7	南关区宽城区	2017年10月30	http://www.bai	本地宝长春教	南关区宽城区	加入收藏长春	1
8	...培训学校部分	2018年1月26日	http://www.bai	学校以音乐围		净月	2018年01月26	0
9	长春南关区小	新湖镇西湖小	http://www.bai	长春南关区小	国内城市国外	你好，请登录	1

图3 数据样本示例

Fig. 3 Data samples

实验是在macOS系统下进行的，CPU为Intel Core i5，内存为8 GB 1 600 MHz DDR3。词向量维度d=300。其他实验条件如表1。

表1 实验环境及配置

Tab. 1 Experimental environment and configuration

集成开发环境	PyCharm CE
编程语言	Python 3.6
词向量模型	Chinese Word Vectors[24]
数据预处理工具	Jieba
深度学习框架	Tensorflow
实验用例（正例/反例）	4 390（2 790/1 600）

3.2 基学习器权重优化

为引入网页结构特征并验证集成学习方法的效果，在实验环节将网页HTML文本数据根据其元标签类型分为如下子数据集：1）标题；2）元数据标签页；3）各级小标题；4）特殊字体类标签；5）网页正文和table标签。根据观察得出不同标签类别对判别网页结果的影响不同，本研究中将网页文本数据分为如下类别，如表2。采用上述5类数据进行ACNN基学习器的构建。

表2 网页标签的类别划分

Tab. 2 Categories of Web page tags

类别	HTML标签
1	<title>标题
2	<meta>元标签数据页
3	<h1>,<h6>…各级小标题
4	,,<b>,<i>…特殊字体
5	<p>网页正文<table>table标签

采用粒子群算法（算法1）实现各基学习器的最优权重计算结果如表3。结果发现，基于标题文本数据和元标签数据构建的基学习器的权重系数高于其他基学习器的权重系数。考虑原因，标题和元标签更好地对网页的元信息进行了概述，而网页正文数据蕴含的信息相较其他标签内容更多样且噪音信息更多，所以基于网页正文构建的基学习器的集成权重系数为最低的0.169 0。在后续实验环节沿用表3中的参数完成EACNN构建。

表3 基学习器的输出权重

Tab. 3 Output weights of base learners

基学习器名称	权重类别	权重系数
Tags1	W_1	0.213 4
Tags2	W_2	0.241 5
Tags3	W_3	0.196 7
Tags4	W_4	0.179 4
Tags5	W_5	0.169 0

3.3　数据样本长度设置

网页数据的样本长度为样本所包含的词汇数目。各Tags数据集样本长度统计曲线均为近对数正态分布。由于不同Tags集合的样本长度分布不同,单纯取长度最长的样本,并把其他样本填充成同样的长度,会浪费计算资源。为避免样本间信息量偏差过大,所以设计长度阈值L来优化输入内容。通过计算样本长度的平均值与二倍此类样本长度标准差的和作为数据的样本长度阈值L的设置公式。公式为:

$$L = \mathrm{mean}(words) + 2 \times \mathrm{std}(words) \tag{11}$$

当样本长度超过阈值L时截取末尾多余长度的样本,对样本长度低于阈值L的样本执行前补零操作。将采用L和原始样本长度(Original Length, OL)的不同数据集基学习器进行了F_1分值的对比实验,来观察设置样本阈值长度对模型的影响。如图4所示。

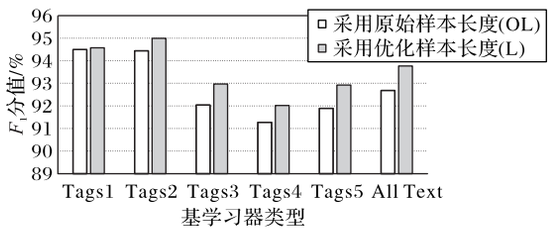


图4　样本长度阈值L对不同学习器的F_1分值影响

Fig. 4　Influence of sample length threshold L to F_1 score of different learners

实验结果表明对各学习器设置优化的样本长度阈值L能够不同程度地提升模型的整体性能。同时发现由于Tags1和Tags2的样本长度范围较小,样本长度阈值的设置对模型的提升效果较低。在样本长度普遍更长的Tags3、Tags4、Tags5和All Text数据集,样本长度阈值的设置使基学习器的F_1分值提升更显著。证明了样本长度阈值设置的合理性。

3.4　模型对比实验

为了验证EACNN模型的先进性,选择了SVM、*K*NN、LSTM、GRU和ACNN作为基线模型,实验结果如表4。

表4　不同模型的性能对比　　单位:%

Tab. 4　Performance comparison of different models　unit: %

模型	Accuracy	Precision	Recall	F_1
SVM[3]	76.77	80.00	38.89	52.34
*K*NN[4]	74.87	71.58	45.73	55.81
CNN[5]	88.15	84.33	78.47	81.29
LSTM[7]	87.93	86.40	75.00	80.30
GRU[27]	89.29	87.60	78.47	82.78
ACNN[17]	91.80	97.95	90.19	93.77
EACNN	96.90	96.23	94.96	95.58

从表4可以看出,使用深度学习的模型判别效果优于基于SVM和*K*NN的传统机器学习模型,证明了深度学习模型在特征提取和学习方面性能优于传统机器学习模型。此外,基于深度神经网络的网页黑名单判别模型在使用本数据集时能力相近,CNN、LSTM和GRU的F_1分值分别为81.29%、80.30%和82.78%。值得关注的是,引入注意力机制后的CNN模型的判别效果得到了一定的提升,达到了91.80%的Accuracy和93.77%的F_1分值,这说明注意力机制的引入增强了网页语义特征的表达能力。EACNN模型的Accuracy和F_1达到了最高的96.90%和95.58%。这证明了本文提出的方法的有效性,将注意力机制和基于网页结构特征的集成学习方法结合能为网页黑名单判别模型的构建起到积极作用。

3.5　集成学习器性能实验

为进一步分析基于网页结构特征的集成学习方法的合理性,将EACNN与基学习器的网页判别能力进行了对比。实验结果如表5。

表5　EACNN与基学习器的性能对比　　单位:%

Tab. 5　Performance comparison between EACNN and base learners　unit: %

学习器类型	Accuracy	Precision	Recall	F_1
Tags1	92.94	94.75	94.63	94.57
Tags2	93.74	93.88	96.39	94.99
Tags3	90.89	88.78	97.87	92.97
Tags4	89.64	88.04	96.73	92.02
Tags5	90.66	88.89	97.61	92.92
All Text	91.80	97.95	90.19	93.77
EACNN	96.90	96.23	94.96	95.58

由表5可以看出,基于网页结构特征的集成学习器相较基学习器而言,具有更高的Accuracy和Precision和F_1值,Recall小幅下降,这说明采用本研究提出的方式,网页黑名单判别的效果更好。可以证明,随着网页结构特征概念的引入,模型的各方面效果都得到了提升。All Text学习器是网页HTML全文本训练的ACNN学习器。

实验分析　与Tags3、Tags4、Tags5基学习器相比,EACNN的Recall略有下降。这是因为Accuracy和Recall互相影响,集成学习的策略为理想状态下追求两个值均高为优,但实际情况是两者相互"制约":追求准确率高,则召回率就低;追求召回率高,则会影响准确率。在集成学习方式选择上采用粒子群智能算法逼近真实答案,这是一种追求准确率的方式,因此会损失一部分召回率作为代价,由于F_1值的提升,证明了集成学习能够提升模型整体性能。此外,本文发现基于Tags1和Tags2数据的基学习器的整体性能优于其他基学习器,这说明网页的标题和元描述数据更好地包含了关于Web页面特征的信息,且此类数据的样本长度较短,易于提取特征且噪声信息少。实验结果发现引入不同网页标签特征的EACNN模型的判别效果优于单一基学习器,说明采用集成学习方法在引入网页结构特征的同时对网页黑名单判别有积极作用。

4　结语

本文提出一种基于注意机制和集成学习的网页黑名单判别方法。本方法依据网页HTML标签类型构建多个网页黑名单判别的基学习器。基学习器采用词向量表达和注意力的卷积神经网络进行特征提取,经过Sigmoid函数进行二分类输出。然后采用基于网页结构特征的集成学习方法对基学习器的输出进行权重赋值,形成引入网页结构特征的最终输出结果,实现网页黑名单的判别。通过设计样本长度实验、基线模型对比实验和集成效果性能实验多角度的观察本方法的性能。实验结果表明,本方法构建的EACNN模型性能在整体性能上优于其他基线模型,证明了引入基于网页结构特征的集成学习和注意力机制能够更好地提取网页结构和文本特征,使网页黑名单的构建效果更好,说明本方法能够为净化互联网环境做出贡献。未来我们计划开展网页多分类工作的研究,结合知识图谱技术,实现面向互联网资源的知识抽取,达到提升互联网信息检索的效率和质量的期许。

参考文献(References)

[1] JAVED M A, YOUNIS M S, LATIF S, et al. Community detection in networks: a multidisciplinary review[J]. Journal of Network and Computer Applications, 2018, 108: 87-111.

[2] 沈昌祥,张焕国,冯登国,等. 信息安全综述[J]. 中国科学(E辑:信息科学),2007,37(2):129-150. (SHEN C X, ZHANG H G, FENG D G, et al. Information security overview[J]. Science in China (Series E: Information Sciences), 2007, 37(2):129-150.)

[3] 胡燕,吴虎子,钟珞. 基于改进的*k*NN算法的中文网页自动分类方法研究[J]. 武汉大学学报(工学版),2007,40(4):141-144. (HU Y, WU H Z, ZHONG L. Research on Chinese Web page automatic classification method based on improved *k*NN algorithm [J]. Engineering Journal of Wuhan University, 2007, 40(4): 141-144.)

[4] 贺海军,王建芬,周青,等. 基于决策支持向量机的中文网页分类器[J]. 计算机工程,2003,29(2):47-48. (HE H J, WANG J F, ZHOU Q, et al. A Chinese Web page classifier based on SVM-decision tree[J]. Computer Engineering, 2003, 29(2): 47-48.)

[5] 邓玺. 基于深度学习的网页分类技术研究[D]. 北京:中国地质大学(北京),2019: 27-41. (DENG X. Research on Web classification technology based on deep learning [D]. Beijing: China University of Geosciences (Beijing), 2019: 27-41.)

[6] BUBER E, DIRI B. Web page classification using RNN [J]. Procedia Computer Science, 2019, 154: 62-72.

[7] RAY A, RAJESWAR S, CHAUDHURY S. Text recognition using deep BLSTM networks [C]// Proceedings of the 8th International Conference on Advances in Pattern Recognition. Piscataway: IEEE, 2015: 1-6.

[8] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016: 2873-2879.

[9] 赵富,杨洋,蒋瑞,等. 融合词性的双注意力Bi-LSTM情感分析[J]. 计算机应用,2018,38(S2):103-106. (ZHAO F, YANG Y, JIANG R, et al. Sentiment analysis based on double-attention Bi-LSTM using part-of-speech[J]. Journal of Computer Applications, 2018, 38(S2): 103-106.)

[10] 苏贵洋,李建华,马颖华,等. 用于中文色情文本过滤的近邻法构造算法[J]. 上海交通大学学报,2004,38(S1):76-79. (SU G Y, LI J H, MA Y H, et al. A *K*NN algorithm on Chinese erotic text filtering[J]. Journal of Shanghai Jiaotong University, 2004, 38(S1):76-79.)

[11] SHEU J J. Distinguishing medical Web pages from pornographic ones: an efficient pornography websites filtering method [J]. International Journal of Network Security, 2017, 19(5): 839-850.

[12] 徐雅斌,李卓,陈俊伊. 基于改进*K*最近邻分类算法的不良网页并行识别[J]. 计算机应用,2013,33(12):3368-3371,3379. (XU Y B, LI Z, CHEN J Y. Parallel recognition of illegal Web pages based on improved *K*NN classification algorithm[J]. Journal of Computer Applications, 2013, 33(12): 3368-3371, 3379.)

[13] 顾敏,郭庆,曹野,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 M, GUO Q, CAO Y, 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.)

[14] KAN M Y, THI H O N. Fast webpage classification using URL features [C]// Proceedings of the 14th ACM International Conference on Information and Knowledge Management. New York: ACM, 2005: 325-326.

[15] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2204-2212.

[16] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2020-03-22]. https://arxiv.org/pdf/1409.0473.pdf.

[17] CHOROWSKI J, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015:577-585.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2017: 6000-6010.

[19] YIN W, SCHÜTZE H, XIANG B, et al. ABCNN: attention-based convolutional neural network for modeling sentence pairs [J]. Transactions of the Association for Computational Linguistics, 2016, 4: 259-272.

[20] ZHOU Z, WU J, TANG W. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence, 2002, 137(1/2):239-263.

[21] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24: 123-140.

[22] 蒋芸,陈娜,明利特,等. 基于Bagging的概率神经网络集成分类算法[J]. 计算机科学,2013,40(5):242-246. (JIANG Y, CHEN N, MING L T, et al. Bagging-based probabilistic neural network ensemble classification algorithm[J]. Computer Science, 2013, 40(5):242-246.)

[23] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2013:3111-3119.

[24] LI S, ZHAO Z, HU R, et al. Analogical reasoning on Chinese morphological and semantic relations[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 138-143.

[25] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[26] KENNEDY J, EBERHART R. Particle swarm optimization[C]// Proceedings of 1995 International Conference on Neural Networks. Piscataway: IEEE, 1995: 1942-1948.

[27] 黄磊,杜昌顺. 基于递归神经网络的文本分类研究[J]. 北京化工大学学报(自然科学版),2017,44(1):98-104. (HUANG L, DU C S. Application of recurrent neural networks in text classification [J]. Journal of Beijing University of Chemical Technology (Natural Science), 2017, 44(1): 98-104.)

This work is partially supported by the Science and Technology Development Program of Jilin Province (20190303133SF), the "13th Five-Year Plan" Science and Technology Project of Jilin Provincial Education Department (JJKH20200796KJ).

ZHOU Chaoran, born in 1994, Ph. D. candidate. His research interests include data mining, urban computing.

ZHAO Jianping, born in 1964, Ph. D., professor. His research interests include data mining, computer network.

MA Tai, born in 1996, M. S. candidate. His research interests include urban computing, deep learning.

ZHOU Xin, born in 1997, M. S. candidate. Her research interests include natural language processing, pedestrian modeling and simulation.