基于深度学习的网页篡改远程检测研究

印 杰[1],蒋宇翔[1,2],牛博威[2],严梓宸[1,2],郭延文[3,4]

(1.江苏警官学院 计算机信息与网络安全系,江苏 南京 210031;2.江苏省公安厅 网络安全保卫总队,江苏 南京 210024;
3.南京大学 计算机科学与技术系,江苏 南京 210023;
4.南京大学 计算机软件新技术国家重点实验室,江苏 南京 210023)

摘 要:为了更为精确和全面地对网页篡改攻击进行远程检测,基于语料库建设和深度学习方法改进了检测过程。大规模获取了可能被篡改攻击的网页,并基于语料库建设方法,人工建立了网页篡改数据库。基于深度神经网络,提出了融合文本特征、结构特征和网络特征的自动检测算法。该文提出的方法可以判断网页是否被篡改和被篡改的类型。经过实验,该文提出的方法在测试数据集上的精确率、召回率和 F 值分别为为 95.6%、96.7%和 96.1%,显著超过基准方法。

关键词:网页篡改;暗链检测;深度学习;神经网络;网络表示学习
中图分类号:TP393.08 **文章编号**:1005-9830(2020)01-0049-06
DOI:10.14177/j.cnki.32-1397n.2020.44.01.008

Remote detection of web page tampering based on deep learning

Yin Jie[1], Jiang Yuxiang[1,2], Niu Bowei[2], Yan Zichen[1,2], Guo Yanwen[3,4]

(1.Department of Network Security Corps, Jiangsu Police Institute, Nanjing 210031, China;
2.Jiangsu Public Security Bureau, Nanjing 210024, China;
3.Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China;
4.State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China)

Abstract: The paper proposes a method that can detect attacks of web page tampering based on corpus construction and deep learning, which can obtain results with high precision and recalling. This paper obtains a large amount of web pages which are potentially tampered, and manually builds the web page tampering database based on the method of corpus construction. Secondly, this paper proposes an automatic detection algorithm based on deep neural network, which integrates text features, structure features and network features. The proposed method can predict whether a webpage has been tampered or not, as well as the attack type. Extensive experiments are conducted to show the effectiveness of the proposed method, with accuracy of 95.6%, recall of 96.7%, and F value of 96.1%, which significantly outperforms the baseline method.

收稿日期:2019-09-17 **修回日期**:2019-12-09
基金项目:国家自然科学基金(61802155);江苏省高等学校大学生创新创业训练计划项目(201710329049X)
作者简介:印杰(1977-),男,高级工程师,主要研究方向:互联网安全和数字取证,E-mail:yinjie@jspi.edu.cn。
引文格式:印杰,蒋宇翔,牛博威,等. 基于深度学习的网页篡改远程检测研究[J]. 南京理工大学学报,2020,44(1):49-54.
投稿网址:http://zrxuebao.njust.edu.cn

Key words: web page tampering; hidden hyperlink detection; neural network; deep learning; network representation learning

据 2019 年国家互联网应急中心(National Internet Emergency Center, CNCERT)发布的第 26 期网络安全信息与动态周报报告[1],我国境内被植入后门的网站数量约为 3 259 个。而根据本文的研究发现,在 2019 年第一季度,仅江苏省内网站就有 2 416 个网站被非法入侵篡改,我国的网站安全形势之严峻远超预期。

大部分网站缺乏专业的维护人员,管理方难以发现和应对种类多样、手段隐蔽、进化速度快的网页篡改攻击。因此,由政府部门提供的、基于远程的自动检测服务是治理网站篡改攻击问题的重要途径。文献[2]使用静态文本识别技术检测已知类型的网站挂马和暗链植入等恶意行为。文献[3]使用机器学习方法识别网页中的暗链,他们结合了暗链的域名、文本和隐藏结构特征,使用了分类与回归树、梯度提升决策树和随机森林 3 种方法来构建检测模型。这些研究在网页篡改方面进行了有益的探索,但仍存在两个主要问题:(1)使用的训练集过小,考虑的网页篡改类型简单,难以面对复杂多变的篡改行为。(2)使用的方法简单,不能充分挖掘不同特征中的复杂非线性关系,限制了模型的表示能力。

也有一些研究致力于检测相近领域的攻击行为。文献[13]对脚本攻击行为进行了智能检测,他们首先使用类图像处理方法对数据进行预处理,再通过词向量方法获取特征,最后使用深度卷积神经网络进行智能识别。文献[14]提出了基于网页主视觉区域的结构化文档对钓鱼网站进行检测,他们主要思路是识别网站中的结构化文档(Document based on main visual area, DMVA),再从中提取特征进行分类。这些研究为识别网页篡改行为提供了有益的思路,然而,由于研究对象并不直接匹配,这些方法和所用的数据集并不能直接移植到网页篡改识别工作中。

近年来,基于神经网络的方法在大数据的知识挖掘上展现出了强大的学习能力[4]。文献[5]提出的 Doc2Vec(Document to vector)是经典词向量表示学习算法 Word2vec[6](Word to vector)的一个改进,可以对短语、句子、段落和文章提取出高质量的分布式特征,以供下游分类任务使用。文献[7]提出的网络表示学习算法 DEEPWALK,可以从拓扑网络中学习出有效的连续特征表示,在分类和链接表示任务上显示出了优异的结果。这些研究成果表明,深度神经网络在提取和处理复杂特征上有着传统算法所不能比拟的优势。

为了更为精准、全面地进行网页篡改的远程识别,本文基于语料库的建设和深度学习方法来解决这一问题。

1 网页篡改攻击数据集

科学、全面、定义完整的数据集不仅可以加深人们对我国网页篡改攻击现状的认识,更重要的是,为比较不同的自动识别算法提供了统一的平台和评测标准,有助于推动网页篡改研究这一领域的发展。

1.1 数据集概述

为了使数据集中数据尽可能丰富,本研究收集了容易遭受网页篡改攻击的政府、教育、小型企业网站网址共 22 350 个,编写爬虫进行页面信息采集,将失效的、爬取错误的链接、页面信息删除,仅留下访问正常、可供研究的页面信息建立数据集,并对搜集到的 22 247 个样本进行了人工标注。数据集中共包含 22 247 个样本,其中被篡改的为 1 112 个。整个数据集按照 8∶1∶1 比例随机切分为训练集、开发集和测试集,每个样本均包含 1 个超文本标记(HyperText markup language, HTML)源代码文件,对应从网络上抓取的 1 个网页。其中,对恶意篡改的部分进行人工标记,并标注出攻击的类型。经观察和统计,主要的攻击类型包括“SEO(Search engine optimization)攻击”、“重定向攻击”、“诱导攻击”和“其它攻击”。

SEO 攻击:通过在网页中插入恶意链接,以提升该链接在搜索引擎中的自然排名。

重定向攻击:通过让页面引用恶意 Javascript(JS)文件,使得正常网站的访问被劫持到恶意网站中。

诱导攻击:通过向页面中插入具有诱导性的文字、qq、微信等方式,诱导用户主动访问恶意网站。

其它攻击:由于大部分攻击行为都可以通过前三类概括,现仅将利用篡改网页分发软件,直接破坏网页内容等目的难以划分的极少量篡改行为

归入其他。统计数据见表 1。

表 1　网页篡改攻击数据集统计

类型	样本数	特征
SEO	263	只有链接
重定向	479	有链接和脚本
诱导	366	有文字、可能有链接,无脚本
其它	4	不定
负样本	21135	未被攻击的网页

1.2　数据集建立

数据搜集阶段,本研究使用自动扫描的策略获取可能被篡改的网页样本。首先,使用网页采集技术正常访问目标网页。对于已经爬取到的网页,定期重新爬取 5 次。每次爬取结束后,对网页中的标签字段建立文档指纹。正常网页的标签字段更新并不频繁,而被篡改网页的标签字段一般更新十分频繁。因此,如果同一网页的 5 次指纹都不一致,则认为该网页可能被篡改。

人工标注阶段,本研究雇佣了 30 名标注者,对搜集到的 22 247 个样本进行了人工标注,标注内容为该网页的被篡改类型。对于每个标注者的标注结果,使用随机抽样的方法进行了可靠性检验,最终,被篡改的页面数为 1 112 个。

2　网页篡改识别算法

2.1　定义和算法框架

在本文中,向量和矩阵将用粗体字母表示。本文将网页篡改识别问题定义为单分类问题:给定训练集$(x_1,\boldsymbol{y}_1),(x_2,\boldsymbol{y}_2),\cdots,(x_n,\boldsymbol{y}_n)\in D^{\text{train}}$,其中 x_i 为网页源代码,$\boldsymbol{y}_i$ 为篡改攻击的类型,为 5 维的单热点表示,分别对应被攻击的 5 种情况,0 表示未被篡改。网页篡改识别的目标则是在训练数据中训练出一个模型 $\boldsymbol{M}$,对于测试数据集$(x_k,y_k)\in D^{\text{test}}$,使得测试数据集中的 $M(x_k)\approx\boldsymbol{y}_k$。

本文的框架见图 1。首先,从源文本中分别提取出文本特征、结构特征和网络特征,并拼接得到文档特征向量,之后,使用深度神经网络分类器进行计算,得到最终的篡改类型。

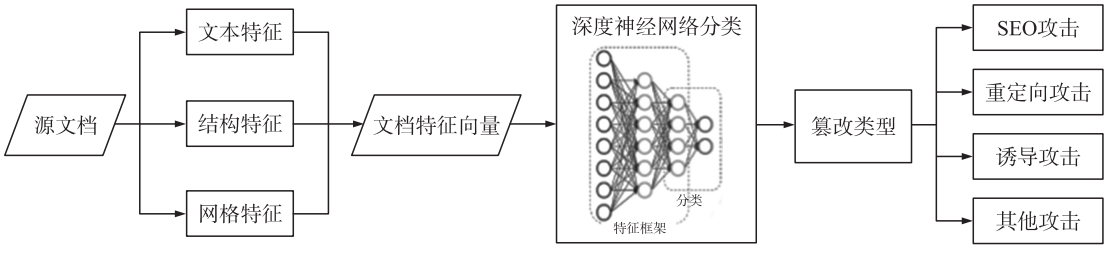

图 1　算法框架图

2.2　文本特征提取

被篡改的网页中常常包含较强的文本语义信息,例如赌博、彩票、色情的相关词,有助于判定篡改的类型。但仅使用人工搜集特征词往往覆盖率较低,并常常造成误判,例如在正常的新闻网站的报导中也可能出现赌博、色情相关的特征词。

首先,使用正则表达式以规则方式提取源文档中的文本,再使用斯坦福中文分词工具包进行自动分词。因此每一个网页会抽取到一串 $w_1,w_2,\cdots,w_n$ 词串。本文使用 Doc2vec[4] 从词串中无监督地抽取特征。Doc2vec 可以自动地学习出高质量的复杂文本特征。文档 $\boldsymbol{d}$ 的向量 $\boldsymbol{d}$ 的具体优化目标函数见式(1)。其中,w_i 为文档 $\boldsymbol{d}$ 中的每一个词,$context(w_i)$ 为该词的上下文词。向量 $\boldsymbol{d}$ 为定长的连续特征向量。对于所有的文档,算法将按顺序遍历文档中的每一个词 w_i,将文档的向量 $\boldsymbol{d}$ 和上下文词 $context(w_i)$ 的向量进行拼接,再最大化目标词 w_i 的似然率。假设所有的单项是条件独立,则总体概率是所有单个概率项的乘积。

$$\boldsymbol{d}=\text{argmax}_d\prod_i p(w_i\mid\boldsymbol{d},context(w_i)) \tag{1}$$

2.3　结构特征提取

网页篡改者常使用 Javascript 脚本或者层叠样式表(Cascading style sheets,CSS)样式来控制篡改,影响内容的呈现、网页的跳转和 SEO 的排名。例如,通过在 CSS 中写入“display:none”、“visibility:hidden”等方式在界面隐藏插入的文本或链接,但可以被搜索引擎索引到。文献[3]对于高频暗链隐藏结构特征进行了很好的总结,本文直接使用了他们的提取方法。

此外,为达到隐藏和劫持等目的,网页篡改的脚本和 CSS 有其本身的统计特点,本文建立了 CSS 和 JAVASCRIPT 词典,包含了常用的函数名 127 个,在计算与网页篡改攻击的互信息后,保留了与其最为相关的 25 个,表 2 给出了部分示例。

表 2 常见的网页篡改相关函数

类型	函数名	效果
JAVASCRIPT	window.navigate	重定向
JAVASCRIPT	window.location.href	重定向
CSS	display:none	隐藏样式
CSS	visibility:hidden	隐藏样式

结构特征为单热点特征,每一个维度对应于一个函数或者规则。本研究一共设定了 35 个结构特征,10 个来自于文献[3],25 个来自于互信息筛选。

2.4 网络特征提取

网页及其链接关系可以形式化为社会网络,每一个网页是网络上的一个节点,而相互之间的链接引用则是网络上的边。通过挖掘网页社会网络的结构特征,常常可以揭示网页本身的性质。例如,Pagerank[8]基于对网页社会网络的挖掘,可以评价网页间的相关性和重要性。

本文使用 Deepwalk[6]算法以无监督的方式获取网页的结构特征。Deepwalk 首先使用随机游走策略从网络 G 中采样出节点序列 $\boldsymbol{v}^1,\boldsymbol{v}^2,\cdots,\boldsymbol{v}^n$,其中,$\boldsymbol{v}^j$ 可以看做节点 j 的特征向量。之后,使用随机梯度下降法优化式(2)中的损失函数

$$O(G)=\sum_{j=1}^{n}\sum_{k=-w}^{w}p(\boldsymbol{v}_k\mid\boldsymbol{v}_j) \tag{2}$$

式中:w 为上下文窗口的大小。概率 $p(\boldsymbol{v}_k\mid\boldsymbol{v}_j)$ 可以直接使用点乘获得,见式(3)

$$p(\boldsymbol{v}^k\mid\boldsymbol{v}^j)=\frac{e^{\boldsymbol{v}^k\cdot\boldsymbol{v}^j}}{\sum_m e^{\boldsymbol{v}^m\cdot\boldsymbol{v}^j}} \tag{3}$$

2.5 分类器

本文使用深度神经网络分类器,可以充分挖掘异构的文本特征、结构特征和网络特征之间的复杂非线性关系。

首先,对于某一个网页,将其文本特征向量 $\boldsymbol{d}$,结构特征向量 $\boldsymbol{f}$ 和网络特征 $\boldsymbol{v}$ 拼接为最终的特征向量 $\boldsymbol{a}$。

深度神经网络一共有 m 层,每一层的输出是下一层的输入,第一层的输入则是 $\boldsymbol{a}$。假设第 g 层的输入向量为 $\boldsymbol{x}^g\in\mathbf{R}^m$,输出向量为 $\boldsymbol{y}^g\in\mathbf{R}^n$,公式 4 中的 h^g 为第 g 层的神经网络函数变换

$$\boldsymbol{y}^g=h^g(x^g)=\mathrm{sigmoid}(\boldsymbol{W}^g\boldsymbol{x}^g+\boldsymbol{b}^g) \tag{4}$$

式中:$\boldsymbol{W}^g\in\mathbf{R}^{m\times n}$,$\boldsymbol{b}^g\in\mathbf{R}^n$,Sigmoid 为逐元素的非线性变换

$$\mathrm{sigmoid}(z)=\frac{1}{1+e^{-z}} \tag{5}$$

因此,如果某个网页的特征向量 $\boldsymbol{a}$ 输入到 l 层的神经网络中,最后得到的输出向量为

$$\boldsymbol{y}^l=h^1h^2\cdots h^l(\boldsymbol{a}) \tag{6}$$

分类器的目标是预测目标网页的被篡改类型,由于网页的篡改类型共有 5 种,因此,分类的标签数为 5,也就是 $\boldsymbol{y}^l\in\mathbf{R}^5$。最终的标签概率输出向量为 $\boldsymbol{y}^o\in\mathbf{R}^5$,本文使用 softmax 函数进行计算。

$$\boldsymbol{y}_i^o=\frac{e^{y_i^l}}{\sum_{i=1}^{5}y_i^l} \tag{7}$$

令 $\boldsymbol{y}^t\in\mathbf{R}^5$ 为网页的真实被篡改标签,使用单热点向量表示,分类器的损失函数为式(4)中的交叉熵损失函数。

$$L=-\sum_{i=1}^{5}y_1^t\log_2 y_i^o \tag{8}$$

在整个数据集上的损失函数则是所有单个损失函数的和。

3 实验

3.1 数据集和评测标准

实验所用的数据集是本研究建立的网页篡改攻击数据集,具体统计数据见第二节。

本研究的目标是判断目标网页的被篡改类型,是单分类问题,使用精确率、召回率和 F 值进行评测。

3.2 基准方法

目前,专门研究网页篡改的研究还不多。与本文较为相关的为文献[3],该研究使用了分类和回归树算法(Classification and regression trees, CART)、梯度推进决策树(Gradient boosting decision tree, GBDT)和随机森林(Random forest, RF)三种基于决策树的算法进行了暗链识别,发现 RF 算法最为强大。然而,暗链只是网页篡改中的一个子集,诱导攻击中就不一定含有链接。

因此,本研究的基准分类方法使用了 RF 算法和支持向量机(Support vector machine, SVM)算法,前者已经在文献[3]中被验证为性能最高,而后者是常见的强大的分类算法,使用的特征即 2.2-2.4 中介绍的特征提取结果:

随机森林:RF 是一种以决策树为基学习器的集成学习算法[9],在多样性增强上不但引入了数据样本扰动,还使用了输入属性扰动,所以一般具

有极好的准确率。

支持向量机:SVM 的目标函数为最大化边界距离函数,并使用核函数将低维数据隐射到高维空间中,从而可以进行分类。本研究使用的是 Libsvm[10] 版本。

3.3　训练细节

使用 Doc2vec 训练文本特征时,窗口大小为 10 个词,特征向量由 PV-DBOW 和 PV-DM 拼接而成,其维度各为 32,总维度为 64。

使用 Deepwalk 训练网络特征时,窗口大小为 10 个节点,最大采样长度为 100,迭代轮数为 10,向量长度为 64 维。

综上,总的特征维度数为 64+64+35(结构特征)= 163 维。这些无监督获取到的特征被固定,不在之后的分类器训练中更新。

在训练分类器时,本研究使用深度为 2 的前馈神经网络,第一层维度为 128,第二层维度为 128。批处理随机梯度下降算法被用来加速训练,每个批次的大小为 20。本研究使用 Dropout[11] 技术来解决过拟合问题,$p=0.2$。在优化时,使用自适应矩估计(Adaptive moment estimation, Adam)策略[12] 来更新参数,可以动态调整学习速率。

3.4　结果

分类结果见表 3。

表 3　分类结果

%

算法	精度	召回	*F*-估值
RF	0.921	0.913	0.917
SVM	0.942	0.903	0.932
本文(文本)	0.912	0.921	0.916
本文(网络)	0.382	0.351	0.366
本文(结构)	0.678	0.725	0.701
本文(全部)	0.956	0.967	0.961

由表可知,SVM 算法超过了 RF 算法,在 *F* 值上取得了 0.5%的提高,而本文使用了全部特征的方法有着显著的提高,比 SVM 分类器提升了将近 2.9%。这表明在结合利用异构数据方面,深度神经网络有着更为强大的性能。

对于不同类型的特征,本文方法在只使用文本特征时,效果最为突出,达到了 0.916 的 *F* 值,这表明文本特征是识别网页篡改的最佳方法,毕竟网页篡改的主要目标主要吸引用户直接或间接的访问赌博、色情、政治等站点,语义特点较为明显。网络特征效果最差,只有 0.366,但结合所有的特征后,分类效果有显著提升,这表明不同类型的特征都可以为最后的识别提供有效信息。

综上,使用深度神经网络来利用多种异构特征识别网页篡改行为,是一种有效的技术方案。

3.5　参数敏感性分析

图 2 给出了深度神经网络的参数敏感性分析。由图 2(a)可见,当中间层维度在 128 至 256 时,分类效果最佳,这表明过多的中间层维度并没有必要,过拟合现象会损害效果,可能未来有更大的训练集才能支撑更大的维度。图 2(b)表明,当层数为 2 时,分类效果最好,更多的层数叠加会造成反向传播时候的梯度难以传播,可能未来引入更先进的网络结构,例如跳层和批标准化能够支撑更多的层数。

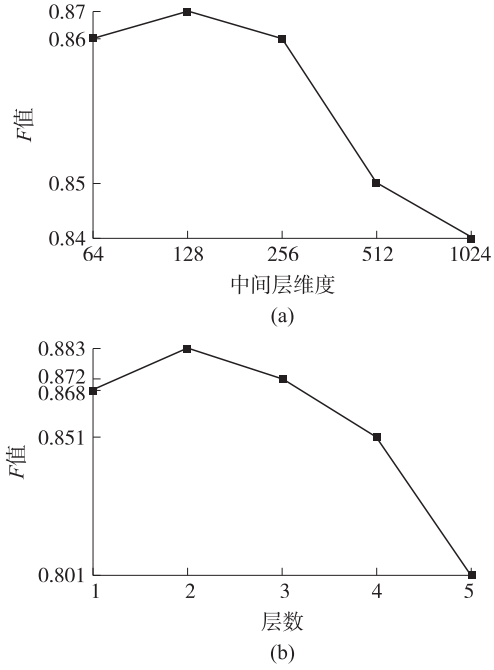

图 2　参数敏感性分析

3.6　案例分析

本文介绍的网页篡改远程检测系统已经部署运行,快速准确地捕获到多次攻击事件,其中最为严重的是某省级重要网站连续遭受名为“365bet”赌博集团的攻击,网页被多次植入关于该赌博集团的广告。该网站分别于 2018 年 9 月 5 日、9 月 6 日、9 月 8 日遭受黑客篡改。该网站主页被黑客植入恶意代码,用户正常访问时不会发现异常,但是通过搜索引擎爬虫访问,会被重定向到赌博网站。

此外,将本次攻击行为特征提取做二次大范围检测,发现名为“365bet”赌博集团做违法广告植入的攻击事件 220 个,其中包括某些重要政府机关网站。

然而,本系统仍然有一些漏检和误检案例,本文对容易发生混淆的主题词进行了可视化,结果见图 3。由图可知,例如阿里巴巴商业圈传销赌博集团等词,通过向网页中插入语义十分模糊的语句比如“阿里巴巴商业人”等语句来诱导用户,和正常商业用语极为接近。提高识别这类网页篡改的能力是未来进一步值得探索和研究的方向。

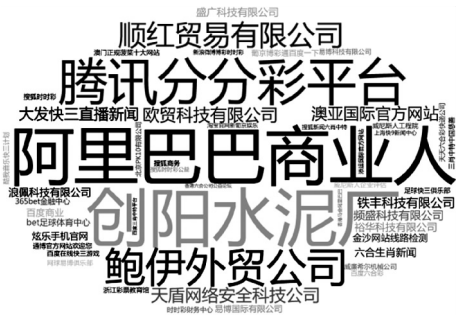

图 3 易混淆的主题词

4 结束语

本研究建立了网页篡改攻击数据集,并使用深度神经网络建立了基于异构特征的自动识别模型。本文的主要贡献有 3 点:首先,建立了网页篡改攻击数据集,和之前研究所用的数据集相比,收录类型更多,数据量更大,为训练自动识别模型提供了数据支撑。其次,提出了基于深度学习的异构特征分类算法,可以有效融合文本特征、结构特征和网络特征,更为精准地识别不同种类的网页篡改行为。最后,实验表明,本文提出的方法在识别的查全率和查准率上显著超过已有技术,证明了方法的有效性,可实现网页篡改远程检测的无人值守。

参考文献:

[1] 网络安全信息与动态周报-2019 年第 26 期[EB/OL]. 2019-7-5,https://www. cert. org. cn/publish/main/upload/File/2019week26. pdf National Internet Emergency Center. Network security.

[2] 邢容.基于文本识别技术的网页恶意代码检测方法研究[D]. 北京:中国科学院大学,2012.

[3] 周文怡,顾徐波,施勇,薛质. 基于机器学习的网页暗链检测方法[J]. 计算机工程,2018,44(10):22-27.
Zhou Wenyi, Gu Xubo, Shi Yong, Xue Zhi. Detection method for hidden hyperlink based on machine learning [J]. Computer Engineering,2018,44(10):22-27.

[4] 张捷,薄煜明,吕明. 基于神经网络预测的网络控制系统故障检测[J]. 南京理工大学学报,2010,34(1):19-23.
Zhang Jie, Bo Yuming, Lv Ming. Fault detection of networked control systems based on neural network prediction [J]. Journal oF Nanjing University of Science and Technology,2010,34(1):19-23.

[5] Le Q, Mikolov T. Distributed representations of sentences and documents[C]//Proceedings of the 31st International Conference on Machine Learning. Beijing,China:PMLR,2014:1188-1196.

[6] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]//Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, USA: Curran Associates, Inc, 2013:3111-3119.

[7] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA:ACM,2014:701-710.

[8] Svetnik V. Random Forest:A classification and regression tool for compound classification and QSAR modeling [J]. Journal of Chemical Information & Computer Sciences,2003,43(6):1947.

[9] Chih Chung, Chang Chihjen. LIBSVM: A library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3):27.

[10] 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.

[11] Kingma D P. Generating sequences with recurrent neural networks [J]. arXiv preprint, https://arxiv.org/abs/1412. 6980.

[12] 张海军,陈映辉.类图像处理及向量化:大数据脚本攻击智能检测[J/OL]. 计算机工程. https://doi.org/10. 19678/j. issn. 1000-3428. 0053360

[13] 刘博文,王雨琪,林果园. 基于结构化文档的钓鱼网站检测算法[J]. 计算机工程与设计,2019,40(10):2791-2798.
Liu Bowen, Wang Yuqi, Lin Guoyuan. Phishing detection algorithm based on structured document[J]. Computer Engineering and Design, 2019, 40(10): 2791-2798.