基于 SVM 和 Word2Vec 的 Web 应用入侵检测系统*

凌仕勇[1]，龚锦红[2]
（1.华东交通大学 网络信息中心，江西 南昌 330013；
2.华东交通大学 电气与自动化工程学院，江西 南昌 330013）

摘 要： 高校应用系统中的 Web 日志数据是系统运维、安全分析的重要来源。针对数据中心产生的 Web 日志进行研究，同时考虑 GET 和 POST 请求的所有数据，采用 Word2Vec 构造特征向量，利用支持向量机进行模型构建。并基于MapReduce 并行计算模型，给出了一种异常入侵检测算法，构建了一套基于 Web 日志的安全分析平台。系统运行结果表明，该平台可以有效地发现校园网中的异常入侵，检索效率高，能有效提高运维效率和异常排查速度。

关键词： 支持向量机；Word2Vec；MapReduce；入侵检测
中图分类号： TP391　　**文献标识码：** A　　**DOI：** 10.20044/j.csdg.2097-1788.2022.02.003
引用格式： 凌仕勇，龚锦红. 基于 SVM 和 Word2Vec 的 Web 应用入侵检测系统[J]. 网络安全与数据治理，2022，41(2)：13-19.

Intrusion detection system of Web application based on SVM and Word2Vec

Ling Shiyong[1], Gong Jinhong[2]
(1.Network Information Center, East China Jiaotong University, Nanchang 330013, China;
2.School of Electrical and Automation Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: The Web log data in the university application system is an important source of system operation and security analysis. This paper mainly studies the Web log generated by the data center, with considering all data for both GET and POST requests, constructs the feature vector with Word2Vec, and builds the model with support vector machine. Based on MapReduce parallel computing model, an anomaly intrusion detection algorithm is proposed, and a security analysis platform based on Web log is constructed. The system operation results show that the platform can effectively find abnormal intrusion in the campus network, with high retrieval efficiency, and can effectively improve the operation and maintenance efficiency and abnormal troubleshooting speed.

Key words: Support Vector Machine(SVM); Word2Vec; MapReduce; intrusion detection

0 引言

随着高校信息化的发展，高校应用系统中积累了大量的师生、教学、科研、管理方面的业务数据。而随着各业务系统的对外访问，网络安全问题日趋严重。目前，校园网安全运维主要是通过网络安全产品如防火墙、IDS、IPS 等设备来实现，总体效果不佳，一个重要的原因是忽视了日志在校园网管理中的作用。校园网中的网络产品、服务器、应用系统等软硬件运行过程中产生大量的日志，记录了系统运行，使用者、攻击者的访问行为，可以通过对这些日志的综合分析和处理，有效解决校园网运行中遇到的安全问题。

Web 入侵检测是针对 Web 应用的一种入侵检测技术，通过对 Web 应用的请求分析，检测和识别 Web 攻击行为。在已有研究中，周勇禄[1]使用 Web 日志中动态页面的参数值长度、字符分布等数据，建立了基于统一异常的检测模型。Estevez-Tapiador 等[2]对日志 URL 进行了划分，对应到马尔科夫模型的不同状态，使用状态转移矩阵，根据模型达到终态的概率判断日志的合法性。Le[3]将 Web 入侵的

* 基金项目：江西省教育厅科技项目(GJJ190317)

URL 根据不同部分进行切割,包括域名、路径、参数等,并对每个部分进行选定特征的提取。Ma[4]等人提取入侵 URL 中的 host 等特征,以此进行 Web 应用入侵威胁检测。Kolar[5]等人则采用词袋模型解决 Web 威胁入侵的检测问题。

高校信息系统一般分散部署在各个服务器中,导致所产生的日志也比较分散。高凯[6]研究了大数据环境下,采用分布式数据流的四个子系统:数据采集子系统、消息处理子系统、流式计算子系统和数据存储子系统,进行用户大规模日志安全分析。陈付梅等[7-9]介绍了大规模系统的日志模式提炼算法的优化方法。上述研究从不同角度构建了针对 Web 应用的入侵检测模型或系统,但主要是通过对 URL 的分析,提取基于文本的统计特征,从而构建分析模型,而没有考虑到 POST 请求体的数据,且在对文本数据的特征向量构建上,主要以统计特征为主,较少考虑到文本本身的词汇特性。本文主要针对数据中心产生的 Web 日志进行研究,采用 Word2Vec 构造特征向量,利用支持向量机进行模型训练,并基于 MapReduce 并行计算模型,给出了一种海量数据异常入侵检测算法。通过此系统对日志事件进行并行挖掘分析,可以很好地发现安全攻击事件,得出平台整体的安全态势,为数据中心正常运转提供安全保障。

1 相关理论

1.1 支持向量机

支持向量机(Support Vector Machine,SVM)是结合向量机和统计学习理论的机器学习方法,其基于结构风险最小化原则,通过寻找一个满足分类要求的超平面,使得在确保分类精度的同时,在该超平面两侧的边界超平面达到最大化间隔。除了进行线性分类之外,SVM 还可以使用多项式、径向基、Sigmoid 等核函数将原空间的数据映射到高纬特征空间,然后在高纬特征空间进行线性分类模型训练,从而有效实现非线性分类。

1.2 词向量

目前,采用机器学习算法识别恶意请求,对文本数据进行向量化的方法,一般多采用词频、词袋或者 TF-IDF(Term Frequency-Inverse Document Frequency,词频-逆文档频率)[10],且以 TF-IDF 居多,即在构建词向量时,根据词语在文档中出现的次数与在词库中的权重系数来衡量该词语在文档中的重要性,这种方法在进行安全识别时,需要对词库的质量以及与实际问题的符合度均要求较高。特别重要的是,在 Web 日志中,很多恶意请求包含的敏感词汇出现频率很低却十分重要,因此采用 TF-IDF 构建恶意请求 URL 的特性向量时,对真实环境下的线上攻击很难作出比较好的响应。

Word2Vec 是 Mikolov 在 Google 带领的研究团队于 2013 年开发的一款将词表征为实数值向量的高效词嵌入工具,它基于 CBow(Continuous Bag-of-words)算法和 Skip-Gram 算法[11-12],将字符串作为输入进行训练得到一个模型。该模型将每个词映射到一个向量,也可以用来表示词与词之间的关系。Word2Vec 系列模型是无监督的,其将大量文本语料作为输入并生成每个单词对应的向量空间,而生成的向量维数往往小于该单词对应的独热编码的维数,即词汇量的总数。这也意味着与独热编码相比,Word2Vec 生成的向量空间更加稠密。

2 入侵检测模型

2.1 数据集

在本次研究中,采用多种渠道获取训练和测试所需要的数据。

(1)分别从 GitHub[13]和 CISC-2010[14]中提取正常和异常的数据,并去除请求主机头信息和冗余信息。

(2)通过网络信息中心的态势感知安全系统,获取被标记为异常的数据。考虑到异常标注的准确性,本文对此数据进行了人工筛查,并对重复数据进行了处理。

(3)通过校园门户前置 Nginx 代理获取 Web 日志数据。

总体训练数据包含 8 种类型,分别为 SQL 注入、XSS、命令执行、路径遍历、远程文件包含、缓冲区溢出、参数篡改、正常数据。总体搜集的训练集数据量为 50 929 条,测试数据集数量为 30 000 条。

2.2 特征工程

2.2.1 数据预处理

数据预处理包括异常值的处理,如删除空数据、去除中文字符、转换字符为小写、urlDecode 解码、对 HTML 转义字符进行处理、数字统一替换为 0,以及每种样本的标签属性定义等。

2.2.2 分词

采用 Java 开源分词算法 ansj_seg[15]对样本进行词粒度分词,分词后进行Word2Vec 向量化处理。如

样本：><script>alert(1)</script><，分词后为：>，<，script，>，alert，(，1，)，<，/，script，>，<。相对于很多文献采用的字符级别的分词，本文以词粒度进行分词既可以保留特殊字符“>”，“<”的特征，又可以保留样本敏感词攻击特征 alert、script，在对样本进行 Word2Vec 向量化时，Word2Vec 模型可以保留 SQL 注入、XSS、命令执行等重要的样本攻击特征词汇。

2.2.3 特征提取

很多文献采用样本的统计特征，如 URL 长度、URL 中数字占比、特殊字符个数、攻击敏感词数，以及参数特征，如参数对个数、参数值最大长度、参数值字母占比、参数中特殊字符个数等，这些特征很难完全描述样本的攻击特性，导致在对模型进行构建时无法达到很好的效果。本文采用 Word2Vec[16]对模型进行训练，在对每个样本进行分词的基础上，通过 Word2Vec 对样本进行学习，生成 Word2Vec 模型。因 Skip-Gram 在大量的数据中效果更好，所以模型采用 Skip-Gram 模式，上下文窗口大小设置为 5，每个样本训练 50 个特征。

图 1 为本文根据训练样本数据提取出的 Word2Vec 特征向量分布，为了在二维空间上直观描述 50 维的特征向量分布，本文对 Word2Vec 特征向量进行了主成分分析降维，横纵坐标值为降维后的二维空间向量特征值。可以看出，经 Word2Vec 对样本进行特征提取后，保留了大量的攻击敏感词汇且相似程度越高的攻击敏感词汇在使用 Word2Vec 构造特征向量后在距离上更近。

2.3 模型训练

采用开源 LibSVM[17]算法训练 SVM 模型，图 2 给出了具体的 SVM 模型训练流程：首先读取所有样本数据；然后通过遍历所有样本数据，在对单个样本进行分词后，基于上述训练出来的 Word2Vec 特征向量模型，分别计算单个样本词汇的特征向量，并对每个样本计算其特征词汇的特征向量均值，即单个样本的特征向量；接着将样本特征向量和样本对应的标签值添加到待训练的数据集中；最后通过 SVM 进行训练生成模型，并保存模型供系统在线检测使用。

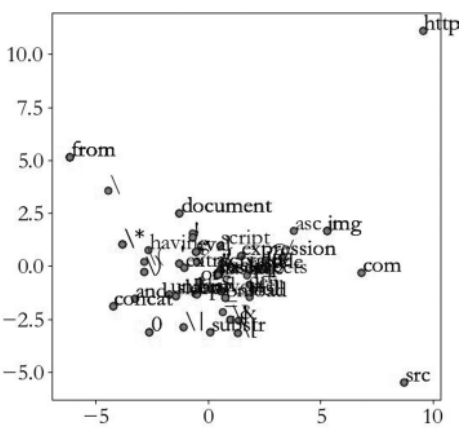


图 1 Word2Vec 特征向量分布

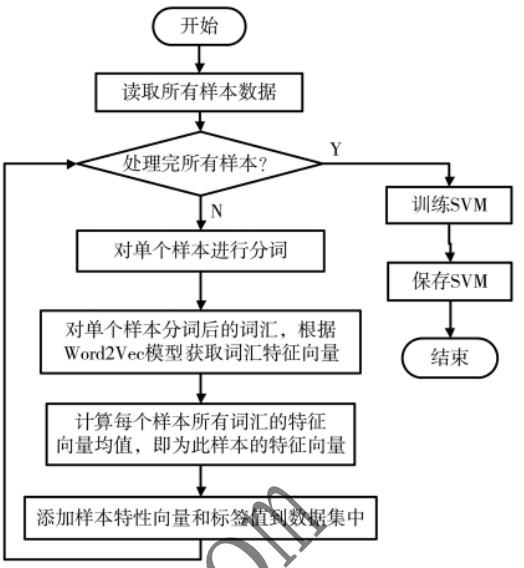


图 2 SVM 模型训练流程

2.4 模型比较

本文采用了两种特征提取方式，一种是前文所述 8 个统计特征向量，另一种是本文的 Word2Vec 特征向量。基于同样径向基核的 SVM 分类器，在检出率、误报率等方面对模型进行了比较，结果如图 3

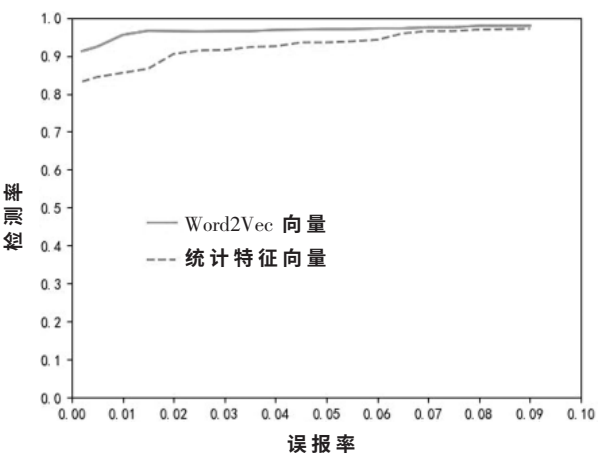


图 3 基于 Word2Vec 向量和统计特征向量的 ROC 曲线

所示。结果显示,通过 Word2Vec 对样本提取特征向量,整体效果好于样本统计特征向量。

2.5 模型评估

模型评估在样本处理方式上与模型训练类似:首先读取所有测试样本数据,对单个样本进行分词、词汇特征向量计算、词汇特征向量均值计算,得到单个样本的特征向量,通过读取训练好的 SVM 模型预测此样本所对应的分类值,并与真实值进行比较,以得到正确和错误个数。

模型评估采用批量测试和单个测试相结合的方式,批量测试采用总计 30 000 条的数据集进行测试,预测值与真实值不一致的有 501 条,正确率为 98.33%,基本可以满足真实环境入侵检测需求。另外,通过提取 8 个样本综合特征(URL 长度、URL 中数字占比、特殊字符个数、攻击敏感词数、参数对个数、参数值最大长度、参数值字母占比、参数中特殊字符个数)构建 SVM 模型,对 30 000 条数据集进行测试后的错误数据,再用本文基于 Word2Vec 向量特征构建的 SVM 模型再次验证,错误率大大降低。

表 1 显示了不同攻击的检测率。因缓冲区溢出的字符输入长度一般明显异于正常请求,SQL 注入、路径遍历攻击的字符分布明显呈现异常,这几种情况的样本特征向量均值与正常样本特征值有着明显的差异,在本次有限样本模型测试中,检出率很高。对于参数篡改,此模型的识别效果还不是很好,需要借助于其他模型如基于用户行为的马尔科夫模型作进一步的分析。

表 1 不同攻击的检测率

攻击类型	检测率
SQL 注入	1.0
XSS	0.982 0
命令执行	0.976 8
路径遍历	1.0
远程文件包含	0.991 5
缓冲区溢出	1.0
参数篡改	0.968 7
总计	0.983 3

3 系统实现

3.1 日志采集

图 4 给出了日志采集的具体流程,采用 Logstash 从多个数据源获取数据,并对数据进行转换和过滤,Logstash shipper 支持各种数据源的 Web 日志,多个不同的 shipper 抽取不同数据源的各种数据,缓存到统一的 Redis 模块。每个 shipper 根据不同的数据来源,自动添加相应的系统标签 systemLabel,如对于门户,添加 systemLabel="portal",以便于后续异常检测中基于不同系统、不同访问请求的识别。同时,在每个 shipper 的配置中,自动过滤掉无关 URL,如对图片、html 文件、样式文件的访问。Logstash indexer 从 Redis 模块获取数据,结合 WebHDFS 插件,将数据存储到 HDFS,供入侵检测系统进行异常检测;同时推送到 Elasticsearch 提供快速检索,供大屏实时展示。

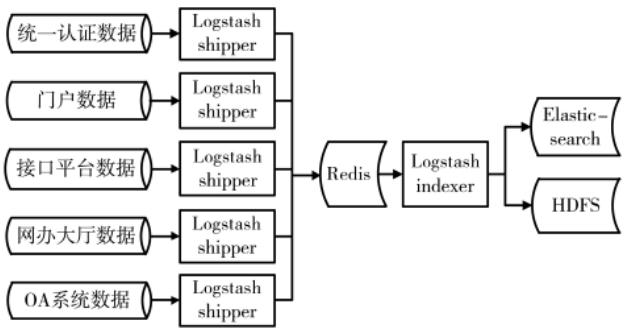


图 4 日志采集流程

3.2 异常检测

异常入侵检测算法采用分布式的 Hadoop MapReduce 架构[18-19],含一组 map()和一组 reduce(),流程如图 5 所示。与一般仅对 GET 请求的 URL 数据进行处理不同,本文的异常检测模型既考虑到 GET 请求的 URL 参数值,也考虑到 POST 请求参数值,系统对用户请求的所有数据进行在线检测,以确保恶意攻击检测的完备性。

map()阶段对每一条 HTTP 请求日志进行数据解析,解析出系统标签 systemLabel、请求所在的日期 day 以及客户 IP,组合构成输出的 key,请求串为输出的 value,以供 reduce()进行详细数据分析使用。

reduce()阶段对 map()输出的数据进行处理。首先,对 map()输出的数据解析出请求 URL,如果请求 URL 串被加密,需要对 URL 串进行解密。其次对于所有的请求,若存在请求参数,解析 URL 中的所有请求参数串对,并对所有的参数值进行 SVM 检测。对于 POST 请求,封装 jsonToMap()和 keyValueToMap()函数,通过这两个通用函数,解析 JSON 格式的请求体或者键值对形式的请求体,然后对所有参数值进行 SVM 检测。若系统检测到异常数据,持久化提取异常

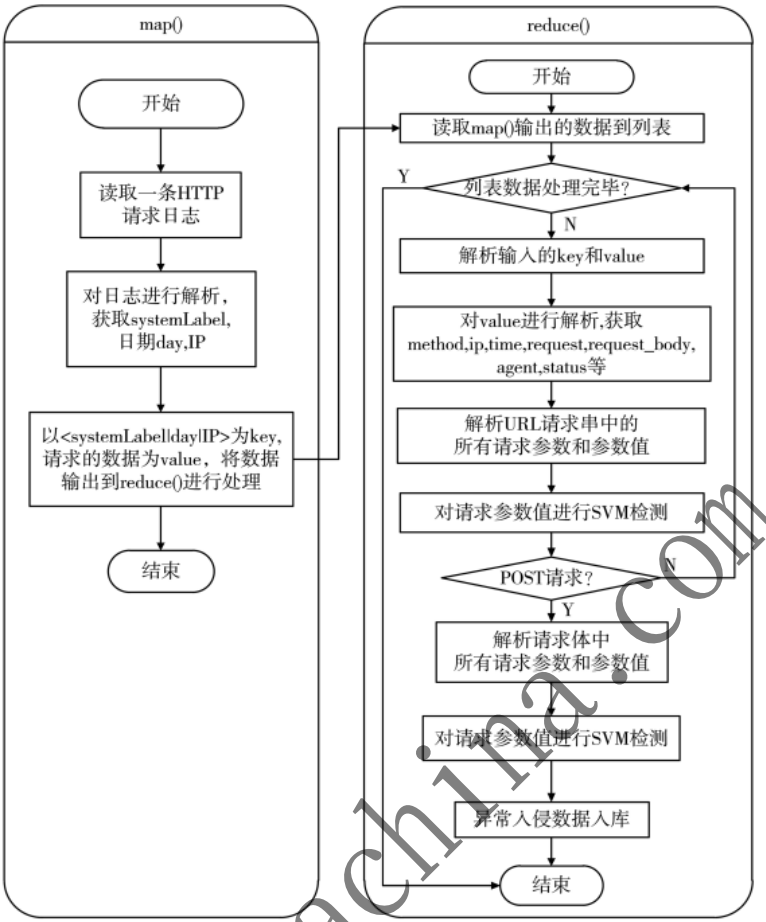


图 5 异常检测算法

数据到数据库中，供 Web 前端进行统计分析与查询。

例如，对于 URL 请求“/?m=offer&s=offer_list&id=1004 and(select 1 from(select count(*),concat(md5(333),floor(rand(0)*2))x from information_schema.tables group by x)a)#”，系统首先将 URL 参数以 & 分割为 m，s，id 3 个参数对，然后对此 3 个对应参数值分布采用 SVM 模型进行检测，识别参数 id 的参数值为 SQL 注入攻击，以 JSON 描述方式记录到数据库中：

```
[{"modelType":"31",
"modelTypeName":"svm_sql 注入",
"reqFunction":"/",
"reqParams_i_param":"id",
"reqParams_i_value":"1004 and(select 1 from(select count(*),concat(md5(333),floor(rand(0)*2))x from information_schema.tables group by x)a)#"}]
```

4 系统效果

4.1 大屏展示

在系统首页，如图 6 所示，实现了对 Web 应用进行实时监测的大屏展示，系统可以在线实时展示当前最高 IP 请求、热门 URL、非 200 状态码请求、搜索引擎占比、最近 30 min 的访问量、最近每天的访问量以及各省份请求地图分布等，有利于管理员对校园应用系统访问状态进行实时在线监测。

4.2 异常入侵访问

图 7 展示了近一年来系统检测到的五个校园应用系统(认证、门户、接口服务平台、网上办事大厅、OA 系统)异常入侵的访问趋势。

4.3 不同类型攻击

图 8 显示了基于异常检测系统构建的不同类型攻击展示，系统目前实现了除静态规则外的三种模型：即 fpm 参数统计分类模型(含参数枚举值、参数集、参数值长度、参数值分布等 4 种类型)，马尔科夫模型(HMM 参数值分布)，SVM 分类模型(含 SQL 注入、XSS、命令执行、路径遍历、远程文件包含、缓冲区溢出、参数篡改 7 种类型)。

图 6 大屏展示

图 7 异常入侵访问

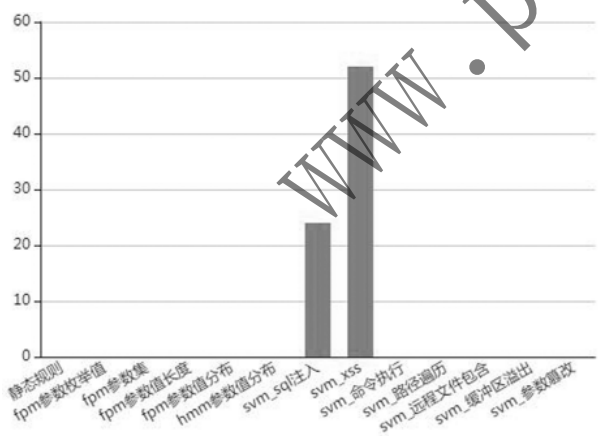


图 8 不同类型攻击

5 结论

校园网 Web 日志数据对业务系统安全运维至关重要，本文针对数据中心产生的 Web 日志，在比较多种样本特征提取基础上，采用 Word2Vec 构造样本特征向量，并利用支持向量机训练出模型。在此基础上，基于 MapReduce 并行计算模型，提出了一种同时兼顾 GET 请求参数值和 POST 请求参数值的异常入侵检测算法，并构建了一整套包括数据采集、数据分发与传递、安全分析、数据存储与可视化呈现的校园网 Web 安全分析平台，能较好地检测到系统访问状态，发现异常访问行为，并通过对访问数据的分析，找出系统可能存在的漏洞，为系统安全正常运维提供良好的基础。

本系统目前仅采集部分 Web 日志数据，下一步将着力集成更多的业务系统数据，并将服务器系统日志、交换机和安全设备日志进行数据集成，结合多个数据源的数据进行联动分析，通过一系列攻击模式挖掘算法，做好校园网安全实时监控和风险预警。

参考文献

[1] 周勇禄，吴海燕，蒋东兴.基于统计异常的 Web 应用入侵检测模型研究[J].计算机安全，2012，12(5)：8-12.

[2] ESTEVEZ-TAPIADOR J M, GARCIA-TEODORO P, DIAZ-VERDEJO J E. Detection of Web-based attacks through Markovian protocol parsing[C]//Proceedings of 10th IEEE Symposium on Computers and Communications, 2005 (ISCC 2005). Los Alamitos: IEEE, 2005: 457-462.

[3] LE A, MARKOPOULOU A, FALOUTSOS M. Phish Def: URL names say it all[C]//IEEE INFOCOM, 2011.

[4] MA J, SAUL L K, SAVAGE S, et al. Identifying suspicious URLs: an application of large-scale online learning[C]//International Conference on Machine Learning (ICML 2009). ACM, 2009: 681-688.

[5] KOLARI P, FININ T, JOSHI A. SVMs for the blogosphere: blog identification and splog detection[C]// 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006: 92-99.

[6] 高凯. 大数据搜索与日志挖掘及可视化方案-ELK Stack: Elasticsearch、Logstash、Kibana(第2版)[M]. 北京: 清华大学出版社, 2016.

[7] 陈付梅, 韩德志, 毕坤, 等. 大数据环境下的分布式数据量处理关键技术探讨[J]. 计算机应用, 2017, 37(3): 620-627.

[8] 赵一宁, 肖海力. 对于大规模系统日志的日志模式提炼算法的优化[J]. 计算机工程与科学, 2017, 39(5): 821-828.

[9] 姚攀, 马玉鹏, 徐春香. 基于ELK的日志分析系统研究及应用[J]. 计算机技术与发展, 2018, 39(7): 2090-2095.

[10] KIM S W, GIL J M. Research paper classification systems based on TF-IDF and LDA schemes[J]. Human-centric Computing and Information Sciences, 2019, 9(1): 30.

[11] JOHNSON R, ZHANG T, Effective use of word order for text categorization with convolutional neural networks[J]. arXiv: 1412.1058, 2014.

[12] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems, 2013: 3111-3119.

[13] Foospidy. Payloads[EB/OL]. (2021-04-22). https://github.com/foospidy/payloads.

[14] CSIC. 安全数据集[EB/OL]. (2012-01-20). https://www.isi.csic.es/dataset/.

[15] NLPchian. Ansj_seg 分词[EB/OL]. (2021-08-15). https://github.com/NLPchina/ansj_seg.

[16] YuyuZha0. Word2vec[EB/OL]. (2021-02-19). https://github.com/YuyuZha0/word2vec.

[17] CHANG C C, LIN C J. LibSVM-A library for support vector machines[EB/OL]. (2021-04-14). http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[18] 丁宣宣, 黄伟. 基于MapReduce和Bagging的并行组合支持向量机[J]. 信息工程大学学报, 2018, 19(2): 196-204.

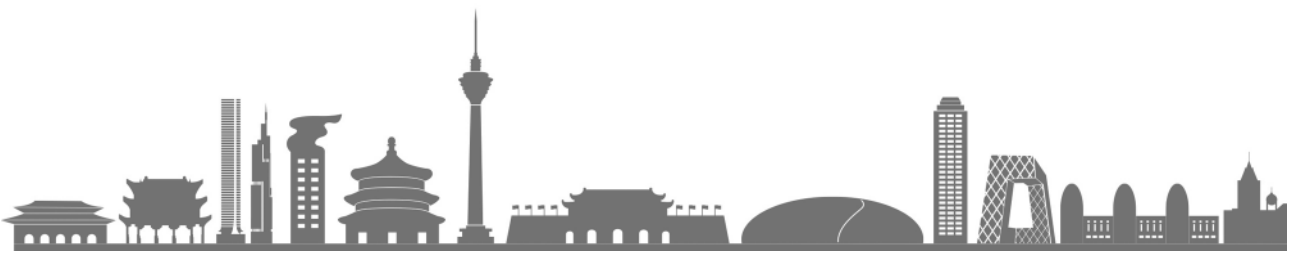
[19] 甘冬连, 张永, 刘博. 基于MapReduce并行SVM的垃圾邮件分类[J]. 软件导刊, 2016, 15(6): 197-206.

(收稿日期: 2022-06-01)

作者简介:

凌仕勇(1974-), 男, 硕士, 副教授, 主要研究方向: 大数据、信息安全。

龚锦红(1985-), 女, 硕士, 讲师, 主要研究方向: 自动化、智能技术。



版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部