DOI:10.20033/j.1003-7241.(2026)02-0123-05

基于 Web 中文自适应分词算法的网络安全风险识别

蔡　翔, 李卫国, 刘立亮, 程兰芳, 张科健, 文　涛

(国网安徽省电力有限公司宣城供电公司, 安徽 宣城 242074)

摘　要:由于缺少对资产文本数据二元切分路径的提取,使得挖掘出的风险特征不完善。为此,提出基于 Web 中文自适应分词算法的网络安全风险识别。采集网络安全的运行数据,进行预处理和分类,构造风险影响因素集,确定网络潜在安全系数,采用 Web 中文自适应分词算法对网络安全文本数据进行分词处理,并利用二级 Hash 表加载词频字典,以获取词长的二元切分路径,对风险特征维度进行等级赋值,确定网络安全的风险等级,实现对网络安全的风险识别。实验结果可知,所提技术得到的约登指数始终控制在 0.85 以上,识别结果与实测值更为接近,识别精度较高。

关键词:Web 中文自适应分词;网络安全;风险识别;攻击检测;风险因素

中图分类号: TP393.08　　**文献标志码**: A　　**文章编号**: 1003-7241(2026)02-0123-05

Network security risk identification based on Web Chinese adaptive word segmentation algorithm

CAI Xiang, LI Weiguo, LIU Liliang, CHENG Lanfang, ZHANG Kejian, WEN Tao

(Xuancheng Power Supply Company, State Grid Anhui Electric Power Co., Ltd., Xuancheng 242074, Anhui, China)

Abstract: Due to the lack of extraction of binary segmentation path of asset text data, the risk characteristics mined are not perfect. Therefore, a network security risk identification method based on web Chinese adaptive word segmentation algorithm is proposed. It collects the operating data of network security, preprocesses and classifies them, constructs a set of risk influencing factors, and determines the potential safety coefficient of the network. It uses the web Chinese adaptive word segmentation algorithm to segment the network security text data, and uses the two-level Hash table to load the word frequency dictionary, so as to obtain the binary segmentation path of word length, grade the risk feature dimension, determine the risk level of network security, and realize the risk identification of network security. The experimental results show that the Jordan index obtained by this technology is always controlled above 0.85, and the recognition result is closer to the measured value, with high recognition accuracy.

Keywords: Web Chinese adaptive segmentation; network assets; risk identification; attack detection; risk factors

互联网技术飞速发展,网络安全问题日益凸显,成为制约网络空间健康发展的重要因素。在网络安全防护体系中,风险识别作为首要环节,其准确性和效率直接关系到后续防护措施的针对性和有效性。中文文本作为网络空间信息传播的主要载体,其处理与分析能力对于网络安全风险识别至关重要。然而,中文与英文等西方语言在书写结构上的显著差异,使得中文文本处理,尤其是中文分词,成为网络安全风险识别中的一项关键技术挑战。中文分词,即将连续的汉字序列切分成有意义的词语单元,是中文文本处理的基础。与英文等以空格分隔单词的语言不同,中文词与词之间没有明显的界限,这增加了中文分词的复杂性。特别是在网络环境下,随着网络用语、新词、缩写词等不断涌现,传统的基于词典的分词方法已难以满足实际需求。因此,研究基于 Web 的自适应中文分词算法,对于提高网络安全风险识别的准确性和效率具有重要意义。

文献[1]通过收集和分析 APT 攻击数据,提取可以表征攻击行为的特征向量,采用 LSTM 模型训练这些特征向量,对网络安全进行风险检测与识别。该方法对输入数据中的噪声和异常值具有一定的容忍度,可以保持较为稳定的识别性能。然而,该方法并行处理能力相对较弱,无法保证最终识别精度。文献[2]收集了与网络安全相关的多源异构数据,并使用图数据结构表示网络安全及其关系,对网络安全进行风险评估,并生成相应的风险报告。该方法可以实时反映网络状态的变化,动态调整风险评估和应对策略。但该方法具有强数据依赖性的缺陷,会影响风险识别的准确性和可靠性。文献[3]基于数据预处理,

收稿日期:2024-09-15;**录用日期**:2024-11-29

基金项目:国网安徽省电力有限公司科技项目资助(B312G0240004)

作者简介:蔡　翔(1983—),男,硕士研究生,高级工程师,研究方向:网络安全。

引用本文:蔡翔, 李卫国, 刘立亮, 等. 基于 Web 中文自适应分词算法的网络安全风险识别[J]. 自动化技术与应用, 2026,45(2):123-127. (CAI Xiang, LI Weiguo, LIU Liliang, et al. Network security risk identification based on Web Chinese adaptive word segmentation algorithm[J]. Techniques of Automation and Applications, 2026,45(2):123-127.)

利用超图理论构建网络安全及其安全风险的超图模型,对提取的风险特征进行分析和处理,实现网络安全风险的自动识别。该方法可以根据实际需要调整模型结构和参数,以适应不同规模和类型的网络安全风险识别任务。然而,该模型的可解释性相对较差,使得最终识别结果的准确度还有待提高。文献[4]借鉴人类免疫系统的结构和功能,构建基于人工免疫算法的网络安全风险识别模型,将提取的风险特征输入到基于人工免疫算法的模型中,识别潜在的网络安全风险。该方法可扩展性较强。但模型中的参数设置对识别性能有重大影响,若设置不当,则会影响识别结果的准确度。

针对以上分析,为切实提高网络安全风险识别的准确率,本文以Web中文自适应分词算法为技术依托,对此展开研究。旨在通过自然语言处理技术的创新应用,提高网络安全风险辨识的智能化水平。该方法的创新点如下:

1) 采集网络安全的运行数据,进行预处理和分类,构造风险影响因素集,确定网络的潜在安全系数;

2) 采用Web中文自适应分词算法对网络安全文本数据进行分词处理;

3) 利用二级Hash表加载词频字典,以获取词长的二元切分路径,对风险特征维度进行等级赋值,确定网络安全的风险等级,实现对网络安全的风险识别。

1 识别关键技术设计

1.1 网络潜在攻击行为检测

网络安全风险识别是指对网络环境中的服务器、数据库、网络设备、应用程序等各种资产进行全面梳理和评估,以识别出可能存在的安全风险或漏洞,旨在发现潜在的安全威胁和薄弱环节,为制定针对性的安全防护措施提供依据[5]。而通过实时监控和动态发现潜在的网络攻击行为,能够为网络安全风险识别提供依据。

在检测网络潜在攻击行为时,主要从网络安全状态变化情况进行分析,通过网络流量捕获和日志收集等方式获取网络安全数据,并根据网络安全属性,对原始数据进行分类处理,以此建立风险影响因素集合[6],表达式为

$$\begin{cases} v(t)=\dfrac{g_s(a_s+h_0)}{\boldsymbol{A}_{\mathrm{g}}} \\ v'(t)=\dfrac{v(t)}{f_{\mathrm{k}}} \\ C_x=\dfrac{g_{\mathrm{e}}/m_h}{f_{\mathrm{k}}\times(j_0/n_{\mathrm{f}})} \end{cases} \tag{1}$$

式中,g_s表示网络安全风险状态信息的二元变量;a_s表示s时刻采集到的网络流量数据;h_0表示网络安全风险的可信度;$\boldsymbol{A}_{\mathrm{g}}$表示叠加标度矩阵;$v(t)$表示原始网络安全数据;$f_{\mathrm{k}}$表示信用奖励函数;$v'(t)$表示预处理后的网络安全数据集合;$g_{\mathrm{e}}$表示主机受到攻击的危险性;$m_h$表示网络中第$h$个节点分配的资源量;$j_0$表示每个节点的平均度数;$n_{\mathrm{f}}$表示复杂网络中的特征路径长度;$C_x$表示风险影响因素集合。

假设网络中的节点数量为N,存储的资源量为M_i,则构建的具有小世界特性的网络数学模型可表示为

$$A_t=\sum_{i=1}^{N}M_i\times C_x\times\exp\left(-\frac{\beta_{\mathrm{e}}\times d_s}{\boldsymbol{h}_s/k_l}\right) \tag{2}$$

式中,β_{e}表示网络安全访问复杂度;d_s表示网络安全风险源数量;$\boldsymbol{h}_s$表示资产访问向量;k_l表示分辨系数。

基于攻击行为的攻击强度与网络漏洞的脆弱性,对攻击行为的决策向量进行编码处理,获取网络处于风险威胁中的资产数据量[7],结合网络安全的风险预测概率,确定网络潜在安全系数,计算公式为

$$\begin{cases} q_s=\dfrac{A_t\times\lambda_{\mathrm{g}}}{r_{\mathrm{u}}} \\ s_0=\dfrac{q_{\mathrm{s}}}{d_t}\times\dfrac{\alpha_{\mathrm{g}}}{p_x} \end{cases} \tag{3}$$

式中,λ_{g}表示网络处于风险威胁中的资产数据量;r_{u}表示网络安全价值评估结果;q_s表示网络环境的脆弱性程度量化数值;d_t表示攻击强度;α_{g}表示攻击行为的路径集合;p_x表示网络安全的风险概率;s_0表示网络安全的潜在安全系数。

通过采集网络安全的运行数据,并对其进行预处理和分类,构造风险影响因素集,结合网络环境的脆弱性程度量化数值,确定网络安全的潜在安全系数。当网络安全潜在安全系数超过预设阈值时,则可判定网络存在攻击行为[8],以此实现对网络攻击行为的检测,为接下来挖掘网络安全风险特征作准备。

1.2 网络安全风险特征挖掘

由于网络安全数据具有非结构化特性,直接从中提取出与资产风险相关的特征信息具有较大的难度。因此,采用Web中文自适应分词算法将网络安全文本数据切分为有意义的词汇单元,即分词处理,从而提取出更有效的风险特征,用于后续风险识别。

首先将网络安全相关的文本数据作为待分词文本,利用有限状态机将文本划分为子句,并结合语料库中的词频字典对数据进行平滑处理[9],公式为

$$G_t=\frac{s_0}{\tau_{\mathrm{d}}}\times\frac{\theta_{\mathrm{g}}}{B_{\mathrm{f}}} \tag{4}$$

式中,s_0表示网络潜在安全系数;τ_{d}表示文本数据的观测序列;θ_{g}表示文本给定状态序列;B_{f}表示分词概率;G_{t}表示平滑后的数据。

采用二级Hash表加载词频字典,以获取词长的优先级和二元切分路径,即

$$\begin{cases} r_t=G_t\times\delta_x\times v_{\mathrm{g}} \\ y_i=r_t\times\boldsymbol{\eta}_0/z_{s'} \end{cases} \tag{5}$$

式中,δ_x表示字符串x在整个语料库中出现的累计次数;v_g表示语料库中文本数量;r_t表示词长优先级;$\boldsymbol{\eta}_0$表示词频向量;$z_{s'}$表示文本s'的隶属度;y_i表示二元切分路径。

在二元切分路径内设定词性搭配阈值,以此为依据对

网络安全的文本数据进行分词处理[10],公式为

$$E(t) = y_i \times \xi_g \times c_h / j_h \tag{6}$$

式中,ξ_g 表示词性搭配阈值;c_h 表示中权系数;j_h 表示经验常数。

在网络安全文本数据分词基础上,利用特定模式匹配方法计算攻击模式中关键词的频率[11],即

$$F_k = \frac{E(t)}{\gamma_k} \tag{7}$$

式中,γ_k 表示匹配系数。

结合风险因素的贡献度[12],提取与网络安全风险相关的特征 D,表达式为

$$D = F_k \times \frac{\mu_f \times w_f}{H_p} \tag{8}$$

式中,μ_f 表示网络被攻击的次数;w_f 表示时间窗函数;H_p 表示风险因素的贡献度。

在网络攻击行为检测基础上,采用 Web 中文自适应分词算法对网络安全文本数据进行分词处理,得到文本的二元切分路径,结合风险因素的贡献度提取网络安全的风险特征,为最终实现网络安全风险识别奠定基础。

1.3 网络安全风险识别

以提取的网络安全风险特征为基础,结合网络安全的攻击路径,采用攻击图的方法来描述前端对网络安全的攻击意图[13]。该过程可表示为

$$S_j = \frac{D}{\exp\left(-\frac{\xi_j}{u_l}\right)} \tag{9}$$

式中,D 表示网络安全风险特征;ξ_j 表示攻击想法量化数值;u_l 表示攻击态度的量化数值;S_j 表示网络安全的攻击图。

假设网络安全在受到入侵后,资产面临的安全风险影响为 χ,则此时风险特征的空间维度为

$$Z_l = S_j \times \psi_b \times Q_d \times \chi \tag{10}$$

式中,ψ_b 表示分类常数;Q_d 表示属性节点指标;Z_l 表示第 l 个风险特征的维度。

利用加权平均方法对风险特征维度进行等级赋值[14],公式为

$$h_y = \frac{Z_l}{\nu_c / \boldsymbol{n}'} \tag{11}$$

式中,ν_c 表示比例系数;$\boldsymbol{n}'$ 表示特征平移向量。

进一步可采用下式计算网络安全风险检测的后验概率,即

$$P = h_y \times \zeta_c \times \varepsilon_g \tag{12}$$

式中,ζ_c 表示模糊合成算子;ε_g 表示权系数。

根据式(12)求取网络安全风险识别的后验概率,并依据其取值将网络安全风险等级进行划分[15],即:$0 \leqslant P \leqslant 0.5$ 为极低风险;$0.5 < P \leqslant 1.0$ 为低风险;$1.0 < P \leqslant 1.5$ 为中风险;$1.5 < P \leqslant 2.0$ 为高风险;$P > 2.0$ 为高风险。参照网络安全风险等级,即可实现对网络安全风险的识别。

2 实例论证分析

为验证本文设计的基于 Web 中文自适应分词算法的网络安全风险识别关键技术在实际应用中的效果,将本文方法应用在某企业网络中,并对其进行风险识别,根据实验结果分析该方法的识别效果。企业网络安全风险识别如图 1 所示。

图 1 网络安全风险识别

Fig. 1 Network security risk identification

2.1 实验准备

本次实验选用某企业级内部网络为主要研究对象。该网络包含约 500 台服务器、2 000 台工作站及移动设备,分布在多个地理位置的数据中心和办公室中。网络采用星型与树型混合的拓扑结构,包括核心层、汇聚层和接入层。其中,核心层由高性能交换机和路由器组成,负责在整个网络中实现高速的数据传输和流量交换;汇聚层位于接入层和核心层之间,用于连接各部门子网,负责将从接入层聚合来的流量进行分发,并将其发送到网络的其他部分;接入层为网络的最外层,是用户设备连接网络的入口,负责连接电脑、手机和其他网络设备等终端用户设备。该网络的物理架构如图 2 所示。

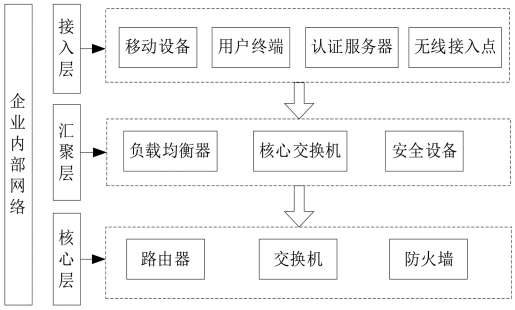

图 2 网络的物理架构

Fig. 2 Physical architecture of the network

利用网络爬虫技术采集该网络安全的公开数据,并将其作为实验数据,将其汇总得到实验数据集,具体如表 1 所示。

以表 1 中的网络安全数据为基础研究数据,对本文方法的应用性能展开测试分析。

2.2 实验过程

实验环境的硬件平台包括服务器和客户端配置;软件平台包括 Ubuntu Server 20.04 LTS 操作系统、Windows 10

Professional 客户端开发系统 Python 3.8 编程语言组件、PyCharm 开发工具和 MySQL 5.7 数据库。实验中,设置自定义词典大小为 5 000 字,包括网络安全领域的专业术语和缩写;最大单词长度为 10;训练集大小为 5 000 条数据,包括不同风险等级的网络安全记录;交叉验证折扣系数为 0.3。

表 1 实验数据集描述

Tab. 1 Experimental dataset description

字段名称	字段表示	字段描述
AssetID	INT(主键)	网络安全的 ID
AssetName	VARCHAR(255)	Web 服务器 1
AssetType	ENUM	网络安全类型
AssetLocation	VARCHAR(255)	逻辑位置
Owner	VARCHAR(255)	资产的所有者
IPAddress	VARCHAR(45)	资产的 IP 地址
OperatingSystem	VARCHAR(255)	操作系统
SoftwareVersion	VARCHAR(50)	服务的版本号
LastUpdated	DATETIME	资产信息最后更新的时间
RiskRating	INT	资产的风险评级
Vulnerabilities	TEXT	安全漏洞列表
ProtectionMeasures	TEXT	防护措施列表

实验过程如下:从多个网络安全数据库收集 IP 地址、域名、端口号、服务类型等原始数据;使用 Pandas 进行数据分析和特征提取,并为训练准备数据集;加载 jieba 字典并导入自定义字典,对网络安全描述等文本进行分割处理;从分割结果中提取关键字和短语作为特征,并结合网络安全的服务类型、端口号等其他属性识别网络安全的风险因素,进而计算网络安全风险的后验概率。

基于以上实验准备与相关参数的设置,将本文设计的方法应用于该网络安全风险识别中,并通过对比实验的形式验证其正确性。

2.3 网络安全风险识别结果与分析

实验中引入基于 LSTM 的方法(方法 1)、基于超图的方法(方法 2)作为本文方法的对比方法。分别采用 3 种方法对该企业内部网络中的网络安全进行风险识别,实验结果如图 3 所示。

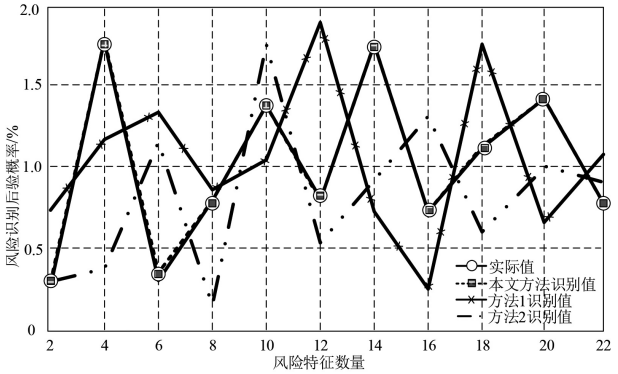

图 3 网络安全风险识别结果

Fig. 3 Results of network security risk identification

通过分析图 3 可以看出,基于相同风险因素数量,应用本文方法对该网络安全进行风险识别分析,得到的风险检测后验概率与实测值基本一致,而方法 1 和方法 2 得到的风险识别值与真实值存在显著偏差,表明这两种方法的识别准确率还有待提高。由此可以证明本文方法在网络安全风险识别中的可行性。

2.4 对比实验与分析

为进一步体现本文方法的优越性能,引入约登指数作为评估指标,对 3 种风险识别方法的性能进行定性评估。在风险识别领域,约登指数越高,表明计算值与实际值之间的特异程度越低,识别精度越高。对比结果如表 2 所示。

表 2 基于不同方法的网络安全风险识别结果

Tab. 2 Results of network security risk identification based on different methods

实验次数	约登指数		
	方法 1	方法 2	本文方法
1	0.36	0.52	0.96
2	0.62	0.41	0.88
3	0.47	0.60	0.94
4	0.39	0.51	0.93
5	0.51	0.44	0.89
6	0.37	0.63	0.95
7	0.49	0.47	0.97
8	0.50	0.55	0.87

从表 2 中的数据可以直观看出,在所研究方法的应用下,得到的识别结果的约登指数明显高于对照组方法,表明本文方法的风险识别值与实际值更加接近,具有较高的识别准确度。

3 结论

信息技术飞速发展,网络安全问题日益凸显。在数字

化时代,保护网络系统和数据的安全至关重要。本研究提出基于 Web 中文自适应分词算法的网络安全风险识别关键技术,实现了对网络安全特征的深度挖掘和海量网络数据的智能分词,为网络安全风险识别提供了新的思路和方法。本研究不仅具有重要的理论价值,可以丰富和完善网络安全领域的知识体系,而且具有广阔的应用前景,可以为政府和企业等用户提供高效可靠的网络安全风险识别解决方案。

参考文献

[1]冯志伟. 基于 LSTM 的 APT 攻击识别与网络风险分析[J]. 电脑编程技巧与维护, 2024(5):160-162.

[2] LI Z, SUN Y, BI X, et al. Multi-temporal heterogeneous graph learning with pattern-aware attention for industrial chain risk detection [J]. World Wide Web, 2023, 27(4):128-132.

[3]王大军, 王震华, 武文元. 基于超图的铁路网络安全风险检测技术研究[J]. 网络安全技术与应用, 2024(5):104-107.

[4]童炜华. 基于人工免疫算法的网络安全风险检测系统研究[J]. 信息记录材料, 2024, 25(4):214-216, 219.

[5]胥明凯, 朱坤双, 李元良, 等. 电力作业多源要素风险的自适应识别模型[J]. 清华大学学报(自然科学版), 2024, 64(6):1047-1059.

[6]刘淑芝. 养殖 RFID 测温定位电子耳标研究--基于智慧农业区块链[J]. 农机化研究, 2025, 47(2):181-185.

[7]MOUTI S, SHUKLA S K, ALTHUBITI S A, et al. Cyber Security Risk management with attack detection frameworks using multi connect variational auto-encoder with probabilistic Bayesian networks [J]. Computers and Electrical Engineering, 2022(103):1-11.

[8]吴昊, 张萍, 周本伟, 等. 基于区块链技术的地震行业网络风险检测系统设计[J]. 现代信息科技, 2023, 7(18):110-113.

[9]周浩. 基于区块链的复杂网络近邻入侵风险检测算法[J]. 成都工业学院学报, 2023, 26(5):60-64.

[10]马秋微, 赵书良, 赵妍. 基于异质信息网络的文本相似性度量方法[J]. 中文信息学报, 2023, 37(9):108-120.

[11]韩冬松, 沙乐天, 赵创业. 基于蠕虫和代理的工控系统攻击建模[J]. 计算机与现代化, 2023(10):107-114.

[12]王孔耀, 安希胜, 陶修冬, 等. 基于人工智能技术的电网基建安全状态自动估计研究[J]. 自动化技术与应用, 2025, 44(3): 164-167,175.

[13]王家鑫, 王瑞琪, 孟海波, 等. 基于深度神经网络 AdaMod 优化模型的来袭目标攻击意图识别[J]. 计算机测量与控制, 2023, 31(6):274-279.

[14]刘欣. 基于危险理论的数字电网网络安全风险预警研究[J]. 自动化技术与应用, 2024, 43(2):102-106.

[15]宋明珍, 徐润婕, 王鹏涛. AIGC 学术知识服务的信息质量风险识别与治理[J]. 现代情报, 2024, 44(8):89-98.

(上接第 86 页)

参考文献

[1]崔岩松, 王方, 陈科良. 基于 VR 直播的远程教育系统设计[J]. 实验技术与管理, 2020, 37(6):132-136, 140.

[2]樊泽明, 余孝军, 王鹏博, 等. 面向工业机器人专业的混合式教学系统[J]. 高等工程教育研究, 2021(6):183-189.

[3]聂永涛, 赵书锐. 基于 VR 技术的柴油机保养维修教学仿真系统设计[J]. 佳木斯大学学报(自然科学版), 2022, 40(3):85-88.

[4]刘泽宇, 张思维, 陆靓, 等. 基于“互联网+”的虚拟现实(VR)计算机辅助教学系统设计[J]. 科教导刊, 2019(18):129-130.

[5]杨横, 张肖如. 基于 VR 技术的算法沉浸式教学系统的设计[J]. 九江职业技术学院学报, 2021(2):51-53.

[6]何丽琴. VR 技术支持下高校英语教学生态系统重构——评《基于虚拟现实的计算机辅助语言教学研究》[J]. 科技管理研究, 2020, 40(21):269.

[7]赵峰, 余江斌, 浦正国, 等. 基于虚拟现实的电力信息交互式系统设计[J]. 自动化技术与应用, 2024, 43(6):161-165.

[8]潘长学, 王兴宇, 张蔚茹. 基于虚拟现实技术的医学解剖教学沉浸式交互设计研究[J]. 装饰, 2020(3):66-69.

[9]樊艺蕾, 丁伟. 沉浸式虚拟现实技术在科学教学中的应用述评[J]. 化学教育(中英文), 2020, 41(5):84-90.

[10]王萍, 颜文贞, 王芳, 等. 沉浸式虚拟现实技术在静脉注射法实验教学中的应用[J]. 中国护理管理, 2020, 20(2):176-180.

[11]邵知宇, 刘添元, 李精伟, 等. 基于虚拟现实的泵站虚拟巡检系统[J]. 自动化技术与应用, 2025, 44(7):43-46, 133.

[12]刘晓曦, 丛晓丹, 宋昌江. 基于虚拟现实的“七星砬子东北抗日联军密营遗址”再现研究[J]. 自动化技术与应用, 2024, 43(7):181-183.

[13]汪海, 李传洋. 基于三维建模和激光点云的建筑物目标监测系统[J]. 自动化技术与应用, 2024, 43(10):148-152.