人工智能环境下网页用户行为数据与安全风险预警机制

何亚玲
云南省曲靖应用技术学校，云南 曲靖 655000

摘　要：在人工智能技术快速发展的背景下，网页用户行为数据成为安全风险识别的重要依据。本文构建了一套集行为数据采集、深度学习识别与风险预警于一体的机制，利用 CNN-LSTM 模型提升异常行为识别准确率，并通过多维指标与动态阈值策略实现高效预警响应。系统具有良好的实时性与适应性，能够应对高并发下的复杂攻击行为，为网页安全防护提供有效支持。

关键词：人工智能；网页行为数据；深度学习；风险预警；CNN-LSTM

在人工智能技术迅速发展的背景下，网页用户行为数据的采集与分析已成为网络运营与安全防护的重要基础。用户在网页中的点击、滑动、停留时间等行为不仅反映其兴趣与意图，同时也被越来越多地用于个性化推荐和广告投放。然而，这些行为数据的广泛使用也带来了新的安全隐患，例如恶意流量伪装成正常用户操作、自动化脚本模拟人类行为进行信息爬取、优惠漏洞利用等现象频发，使得传统的规则式安全防御机制逐渐难以应对。当数据本身变得复杂且高频，人工识别与人工干预变得不再现实，迫切需要依托人工智能技术对用户行为模式进行动态学习和异常检测，从而实现更加及时有效的风险预警机制。目前国内外研究虽提出了多种检测算法，但在中文环境下的网页业务中尚缺乏完整的、面向实际部署的智能预警框架。因此，本文聚焦于人工智能环境下的网页用户行为数据分析与安全风险预警机制的构建，尝试通过实证与系统设计提出一种可落地、可扩展的技术路径，弥补现有研究在实时性与准确性方面的不足。

1 网页用户行为数据的采集与特征

随着网页交互的不断复杂化，用户在网页端的每一次点击、滑动、页面跳转甚至停留时间，都构成了行为数据的重要组成部分。这些数据不仅是理解用户行为的重要依据，也为网页安全风险的识别提供了基础支撑。在实际运营中，网页行为数据的采集一般通过前端嵌入脚本进行埋点，实现对用户行为全过程的捕捉。以常见的电商网站为例，前端埋点通常覆盖点击事件、页面进入与退出、滚动条变化、鼠标悬停区域及表单输入等操作。这些信息会与时间戳、设备信息、浏览器类型、Referer 等技术参数一并上传至服务器，用于后续分析。此类行为数据的丰富程度极高，粒度甚至可以精确到每一次点击坐标的位置。但也因为采集频次高、字段类型杂，数据质量往往存在明显问题，如埋点失败、字段缺失、值不规范等情况，严重时还会影响前端页面的正常加载。例如，某次大型活动期间，多地反馈前端加载缓慢，排查后发现是因嵌入的行为采集 SDK 与浮窗广告组件发生脚本冲突，造成加载阻塞，最终通过调整数据上报策略才得以缓解。

网页用户行为数据的来源主要分为三类：第一类是由页面前端直接上传的结构化事件数据；第二类是 Web 服务器或 CDN 节点生成的访问日志；第三类则是由第三方数据统计工具（如神策、GrowingIO）提供的用户行为报告。当前较为常用的采集架构大多采用“前端 SDK + 日志系统 + 流处理引擎”的形式，结合 Kafka、Flume 等中间件实现高并发下的日志接入和统一处理。这种设计虽然可以满足基本的日志传输与处理要求，但仍面临字段冗余、事件时序错乱、数据漂移等一系列挑战。为了实现后续的风险行为识别，必须对原始日志进行系统性的清洗与转换，确保数据的统一性与可用性。

在数据预处理阶段，首先需要对原始日志中相同用户的多个行为进行归一化处理，建立完整的用户访问 Session。这一过程通常需要结合用户的设备标识（如 Cookie 或 Token）、IP 地址及时间戳信息进行合并重建。其次，针对字段缺失或异常的情况，常用的数据修复方法包括前向填充、均值替代、字段回推等方式。对于恶意刷流、脚本

操作或自动化程序伪装行为常伴随的大量重复请求，还需进行过滤去噪。经过处理后，原始日志数据可以被转换为标准化的行为序列输入模型进行分析。在实际部署中，经过上述流程处理后的结构化行为数据往往压缩比高、结构清晰，便于快速调用与建模。

在特征工程方面，为提升后续人工智能模型识别风险行为的准确性，需从用户行为中提取高质量、可区分度强的特征变量。传统的统计类特征，如访问频率、页面跳转路径、平均停留时长、浏览深度等，仍然具备一定的识别能力。但面对更加复杂的伪装行为，仅靠这些低阶特征难以实现有效判断。因此，近年来逐渐引入了行为序列类与交互节奏类特征。例如，滑动节奏密度、事件间隔波动系数、页面交互区域集中度、行为 n-gram 序列模式、点击 - 滚动配比等特征被广泛用于识别模拟脚本与真实用户之间的差异。特别是在多个 IP 地址共用相同 Cookie、或一个终端在极短时间内多次完成完全一致操作的场景中，这类特征可以有效暴露出潜在的恶意行为。

此外，在数据标签稀缺的条件下，一些无监督方式如基于聚类的异常检测或密度估计也被引入到预处理流程中，用于识别出明显偏离正常行为分布的样本。行为数据本身具有较强的时序性与上下文关联性，因此在构建最终模型特征时，还需要考虑序列中的动作依赖关系，这也为后续采用序列类模型（如 LSTM 或 Transformer）奠定了基础。

2 人工智能在用户行为模式识别中的实践

网页用户行为数据呈现出高频次、强时序性与复杂上下文特征，仅依靠传统的规则库或静态分析方法，难以准确识别潜在的风险行为。在实际运维中，不少企业仍使用基于 IP 黑名单、访问频率设限等方式进行初级风险管控，但这些方式无法识别经过伪装的自动化程序或模拟点击行为，尤其在面对复杂的人机混合流量时，误判与漏判现象尤为严重。因此，采用人工智能技术，尤其是基于深度学习的行为模式识别方法，成为提升安全风险预警能力的现实选择。

在众多模型中，传统的聚类算法如 K-means 和密度聚类（DBSCAN）虽然可用于无监督的用户行为聚类，但在处理高维、序列化行为数据时，效果并不理想。这类模型对于行为的时间顺序缺乏敏感性，无法捕捉用户连续行为中的潜在逻辑。相比之下，基于神经网络的模型，如卷积神经网络（CNN）和长短期记忆网络（LSTM），在处理网页交互行为时展现出更强的特征抽象与时间序列理解能力。其中，CNN 适合从局部行为片段中提取高频模式，如快速点击、连续滑动等特征；而 LSTM 擅长捕捉行为之间的长期依赖关系，适用于识别用户在多个页面间的跳转路径、操作顺序等逻辑行为。

在本研究中，综合考虑网页用户行为的局部操作密度与整体行为节奏，构建了一个 CNN-LSTM 混合模型。模型的输入是经过预处理后的用户行为序列矩阵，每一行代表一个用户的完整 Session，每一列代表一次具体的操作事件（如点击、滑动、输入等）。通过一维卷积层提取操作片段的局部模式后，将特征序列送入双层 LSTM 网络，捕捉其时序变化趋势。模型输出为一个风险评分，范围在 0 到 1 之间，分值越高表明该行为序列越可能为异常。训练过程中，采用交叉熵损失函数进行反向传播优化，使用 Adam 优化器控制学习率动态调整，最大迭代轮次设为 10 轮，早停条件为验证集精度不提升超过 3 轮。在实际训练样本中，正负样本比例控制在 1:3，主要来源于历史日志中被确认的恶意请求与真实用户行为标注数据。

在企业内部的试验部署中，选取某大型电商网站近一月的网页交互数据进行训练与测试。该网站日均活跃用户在 180 万左右，页面请求总量超过 2 亿条。系统通过 Kafka 实时抽取行为流，经 Flink 处理生成行为序列后送入模型进行推理。模型部署在一台 32 核 CPU、128G 内存的服务器上，使用 TensorRT 对模型推理过程进行了加速，整体延迟控制在 2.3 秒以内。在评估指标上，模型在测试集中的精确率达到 0.86，召回率为 0.91，F1 值为 0.88，显著优于基准的基于规则的风险识别机制。在典型的“优惠薅羊毛”场景中，该模型成功识别出通过代理 IP 与多账号切换操作频繁访问促销接口的异常用户群体，避免了大额无效补贴支出。

值得一提的是，该模型并非完全依赖行为序列本身，还融合了部分上下文信息，如访问设备类型、地理位置、Referer 路径等作为辅助特征，用于提高模型对跨区域异常行为的感知能力。在实际运行中发现，加入这些上下文信息后，对于代理服务器行为、VPN 穿越行为的识别率提升了约 15%。这表明，在行为识别过程中，仅依赖序列建模仍然存在一定盲区，合理融合多维数据来源，可以显著提高人工智能模型的实际判别能力。

3 安全风险预警机制设计与实现

在完成用户行为数据的采集与模型识别之后，如何将这些识别结果及时转化为可以执行的预警信号，是风险控制中非常关键的一环。传统网页安全体系多依赖静态规则库，如设置固定访问频率上限、识别重复请求行为等，这类方式实现成本低，但在面对复杂的自动化攻击和行为伪装时明显力不从心。为此，需要构建一套依托人工智能判断、具备动态调整能力的网页风险预警机制，形成从识别到响应的闭环系统。

本研究设计的预警机制，核心由三个部分组成：风险指标体系、风险评分与分级策略、预警流程的实时响应。首先在指标设计方面，我们基于前文提取的行为特征，构建了五大类风险行为：包括访问频次异常、页面跳转路径不规律、滑动与点击节奏不协调、身份信息冲突（如 IP 与设备不一致）以及资源请求过密等，共设定 17 项指标。这些指标均可自动从行为数据中提取，通过标准化与加权处理，进一步形成统一的风险评分体系。

风险评分采用模型输出值与指标分布权重结合的方式进行计算。系统使用熵权法动态确定每个指标的参考权重，避免人为主观设置；再结合深度模型的输出结果进行加权求和，得出最终的风险得分。该得分划分为三个等级：0–40 为绿色，表示用户行为正常；41–70 为黄色，说明存在可疑特征；71 以上为红色，表示高风险行为，应立即引发警报并触发处理动作。分级策略有利于系统根据不同场景制定响应方案，避免“一刀切”式的处理方式。

在预警流程方面，系统采用“行为识别—评分判断—告警联动”三步结构。用户行为数据从 Kafka 管道输入后，先经由流处理框架 Flink 进行行为合并与序列整理，然后送入部署好的模型进行快速识别。识别后的风险得分会与预设的阈值进行比对，如超过警戒值，则将结果推送到 Redis 短时缓存，并通过 Webhook 接口向指定平台发送告警信息，如企业微信、钉钉群机器人等，同时也可触发封禁 IP、限制接口调用频率等联动操作。整个流程从数据进入到预警生成，一般控制在 3 秒以内，满足中等流量网站的实时响应要求。

在实际运行中，为确保系统高效稳定，各处理模块均采用容器化部署，便于分布式扩展。模型推理服务使用 TensorRT 加速，提升并发处理能力，测试结果显示：在单机 8 核 CPU、32GB 内存配置下，每秒可处理约 5000 条行为记录，预警准确率达 92% 以上。该系统曾成功识别出一次通过模拟点击参与大促活动的自动化操作行为，并在短时间内完成用户限流与接口保护，避免了运营资源的非正常消耗。此外，系统还设置了阈值动态调整机制。考虑到实际网站在促销、高峰访问等阶段行为模式会波动，静态阈值容易引起误判，因此系统每天会根据过去 24 小时的风险均值与波动幅度自动调整触发阈值，提高灵敏度的同时降低误报率。

4 结论

本文围绕人工智能环境下网页用户行为数据的分析与安全风险预警展开研究，构建了从数据采集、行为识别到实时预警的完整机制。实践表明，基于深度学习的识别模型能够有效发现异常行为，结合动态阈值与风险分级策略，提高了预警的准确性与响应效率。该机制适用于高并发场景，具备一定的实用价值。尽管仍存在算力依赖与误报控制等问题，但整体框架具备良好的扩展性，为网页安全防护提供了新的技术路径。

参考文献：

[1] 李忠霖 . 大数据背景下用户行为数据分析课程教学改革研究 [J]. 电脑与电信 ,2022,(10):27–30.

[2] 兰坤 , 吴琼 , 耿艳兵 . 基于 Python 的社交网站用户行为数据采集方法 [J]. 智能计算机与应用 ,2024,14(06):219–223.

[3] 林玲 . 基于主题搜索的校园用户行为挖掘系统的设计与实现 [D]. 北京邮电大学 ,2020.

[4] 叶力铭 . 基于 Spark 电商用户行为数据的分析与研究 [D]. 沈阳师范大学 ,2020.

[5] 周爱娟 . 基于计费系统的校园用户行为分析与建模 [D]. 北京交通大学 ,2019.

[6] 向大为 , 吴燕波 . 互联网用户行为数据收集与分析的研究 [J]. 现代信息科技 ,2019,3(06):14–16.

[7] 华圩 . 基于大数据的用户行为数据分析系统设计和实现 [D]. 华中科技大学 ,2018.

[8] 孙宇 . 基于海量数据的用户行为数据分析系统研究与实现 [D]. 山东大学 ,2017.