智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

非结构化文档敏感数据识别与异常行为分析

喻波， 王志海， 孙亚东， 谢福进， 安鹏

引用本文:
喻波, 王志海, 孙亚东, 等. 非结构化文档敏感数据识别与异常行为分析[J]. 智能系统学报, 2021, 16(5): 932–939.
YU Bo, WANG Zhihai, SUN Yadong, et al. Unstructured document sensitive data identification and abnormal behavior analysis[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(5): 932–939.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202104028

吴文俊人工智能技术发明奖一等奖

成果名称：基于数据流的一体化数据安全管理关键技术及平台应用
获 奖 人：喻波、王景璟、王志海、严寒冰、何能强、王志华、何晋昊、孙加光、彭洪涛、安鹏、孙亚东
完成单位：北京明朝万达科技股份有限公司、清华大学、国家计算机网络与信息安全管理中心

喻波

毕业于清华大学，北京明朝万达科技股份有限公司首席科学家、高级副总裁，正高级职称（教授）、兼任公安部通信标准化技术委员会委员，20年信息安全领域科研经验，擅长大型安全系统架构设计、操作系统内核安全分析、人工智能技术与密码技术。2005年，其与王志海先生一起创办明朝万达，负责产品研发与管理工作。在喻波带领下，研发中心由成立初的5人扩展到目前150余人，形成一支以清华大学博士和硕士为骨干力量的核心团队，中心先后取得ISO9001、ISO27001、CMMI3等资质。截止目前，申请发明专利300余项，授权发明110项，其本人获发明专利近100项。作为项目负责人或骨干先后承担国家级、市级科研课题近20项。其带领团队研制安元数据安全系列产品，先后获得密码局、保密局、公安部、国家重点新产品认证与资质，并成功应用于中国银行、光大银行、审计署、质检局、海关总署、浙江移动等多个重点行业。

喻波在加强公司技术自主创新和推进成果转化等方面做出了卓越贡献，取得显著经济效益，助力公司快速发展。历经十余年发展和积累，明朝万达现已成为金融、政府、公安、电信运营商等国内高端客户的首推品牌，签约用户超过3000家，领跑国内数据安全市场。

团队简介

北京明朝万达科技股份有限公司、清华大学、国家计算机网络与信息安全管理中心联合研制的《基于数据流的一体化数据安全管理关键技术及平台应用》项目，围绕数据生命周期，通过源头创新，实现云网端体系化、全方位数据安全保护。北京明朝万达科技股份有限公司是中国新一代信息安全技术企业的代表厂商，专注于人工智能、数据安全、云安全、大数据安全及加密技术的研究与应用，客户覆盖金融、政府、公安等众多行业。在此项目中提出了基于双缓冲的跨平台文档透明加解密、基于复杂语境下 N 元中文语言模型的文档分类分级等关键技术，实现文档用户无感知加解密、数据隔离存储，显著提升了文档分类分级准确率；清华大学是位列“世界一流大学和一流学科”、“211 工程”、“985 工程”的中国著名高等学府、中国高层次人才培养和科学技术研究的重要基地，被誉为“红色工程师的摇篮”。在本项目中提出了时变网络用户异常行为分析技术，大幅提升用户异常行为识别准确率；国家计算机网络与信息安全管理中心，为中央网络安全和信息化委员会办公室直属事业单位，是我国网络安全应急体系的核心协调机构。在本项目中提出了通过动态检测与静态检测相结合的技术路线，改进了动态内容分析与静态内容融合分析方法，显著提升了未知恶意代码检测率。完成单位基于多年行业经验，将本项目的研制成果，广泛推广到金融、公安等众多行业，取得了显著的经济效益与社会效益。

DOI: 10.11992/tis.202104028
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20210722.1124.002.html

非结构化文档敏感数据识别与异常行为分析

喻波，王志海，孙亚东，谢福进，安鹏
（北京明朝万达科技股份有限公司，北京 100876）

摘　要：在海量数据中快速、准确地对数据进行分类分级，快速识别用户异常行为是目前数据安全领域的重要研究内容。在数据分类分级研究领域，自然语言处理技术提升了分类分级的准确率，但是中文语体混杂、无监督学习准确率低、有监督学习样本标注工作量大等问题亟待取得关键突破。本文提出多元中文语言模型和基于无监督算法构建样本，突破数据分类分级领域面临的关键问题。在用户异常行为分析研究领域，由于样本依赖度过高，导致异常行为识别准确率较低，本文提出利用离群点检测方法构建异常行为样本库，解决样本依赖过高问题。为验证方法可行性，进一步构建实验系统开展实验分析，通过实验验证所提出方法可以显著提高数据分类分级和异常行为分析的准确率。

关键词：数据安全；人工智能；分类分级；语言模型；用户异常行为分析；样本；自然语言处理；监督学习
中图分类号：TP18; TP319; TP309　　**文献标志码**：A　　**文章编号**：1673−4785(2021)05−0932−08

中文引用格式：喻波，王志海，孙亚东，等. 非结构化文档敏感数据识别与异常行为分析 [J]. 智能系统学报，2021，16(5)：932–939.
英文引用格式：YU Bo, WANG Zhihai, SUN Yadong, et al. Unstructured document sensitive data identification and abnormal behavior analysis[J]. CAAI transactions on intelligent systems, 2021, 16(5): 932–939.

Unstructured document sensitive data identification and abnormal behavior analysis

YU Bo, WANG Zhihai, SUN Yadong, XIE Fujin, AN Peng
(Beijing Wondersoft Technology Co., Ltd, Beijing 100876, China)

Abstract: It is an important research content in the field of data security to classify data quickly and accurately in mass data, and to quickly identify user abnormal behavior. In the field of data classification research, natural language processing technology improves the accuracy of classification, but the problems of mixed Chinese language, low accuracy of unsupervised learning, and large workload of supervised learning sample labeling need to be Chinese made urgently. In the field of user anomaly analysis, due to high sample dependence, which leads to low accuracy of abnormal behavior recognition, this paper proposes to use outlier detection to build an abnormal behavior sample library to solve the problem of excessive sample dependence. In order to verify feasibility of the method, the experimental system is further constructed to carry out experimental analysis, and the proposed method can significantly improve the accuracy of data classification and abnormal behavior analysis.

Keywords: data security; artificial intelligence; classification; language model; user's behavior analysis; sample; nlp; supervised learning

伴随大数据时代来临，数据同土地、劳动力一样列入生产要素之一，为加快培育数据要素市场，数据安全防护是基础保障[1]。由于业务场景复杂多变、数据量持续增长，基于数据流的信息化架构演进对数据安全防护提出了更高要求。传统的数据安全防护措施难以适应新时代架构复杂、场景多样、数据海量、交互频繁的“大物移云智”环境。在数据安全领域，可以把管控场景划分为强管控、中管控、弱管控 3 种场景。针对强管控场景，在数据传输、数据存储等环节采取全量数据加密手段，可以解决强管控场景的安全诉求。但是在更为普遍的中管控场景，敏感数据与非敏感数据混合使用，如果采用强管控，必然对

收稿日期：2021−04−16.　**网络出版日期**：2021−07−23.
基金项目：国家电子发展基金项目 (工信部财 [2014]425 号).
通信作者：喻波. E-mail: yubo@wondersoft.cn.

操作效率造成影响，但如果弱管控，又会造成敏感数据泄露。因此中管控场景，识别敏感数据与异常行为是目前数据安全领域研究的重要问题。

1　研究背景

近年来企业内部人员窃取数据事件频发，如银行内部人员窃取、贩卖上万条用户信息，根据 Gartner 的调查结果[2]，内部人员窃取敏感数据、盗用账号等行为已经成为企业数据泄露的最主要原因。由于内部人员具备企业数据资产的合法访问权限，且了解企业敏感数据的存放位置，因此内部人员在合法外衣的保护下，自由地游走在企业内网，进行长期而隐蔽的数据窃取行为。

导致上述现象的根本原因是企业经营活动复杂、业务系统多、数据规模大，无法穷举人员、终端、应用、数据之间多对多的使用与访问规则，从而为内部恶意人员留下了巨大的非法活动空间。这种情况下，安全部门只能把管控重点放在已知的、规则明确的安全威胁上，导致内部人员长期、隐蔽数据窃取行为发现难，数据泄露事件依然层出不穷[3-5]。

因此，研究文档分类分级技术，识别需要重点保护的敏感数据；研究用户异常行为分析技术，识别用户异常操作，是解决内网数据泄露的重要课题。

2　基于 N 元语言模型的文档智能分类分级技术

根据数据安全治理理论，数据安全防护的基础是识别数据资产、分级分类数据资产，只有明确了数据保护对象及安全等级，才能对数据实施按需防护，避免一刀切式的静态防御，促进数据的安全流通与共享。

在数据资产的识别过程中，按载体形态差异其识别方式也存在不同，数据按载体形态主要划分为结构化数据与非结构化数据两类。针对结构化数据，数据集中承载在数据库中，可以通过元数据的精确定义，较为直接地进行数据资产识别并实施分级分类管控，但针对广泛分布在终端、网络、云存储中的非结构化敏感数据 (如商务合同、会议纪要、监管报告、技术资料等敏感文档)，鉴于文档是由字、词及上下文语义构成的，单纯的依赖关键词、正则表达式等传统规则手段对文档分类进行识别，缺乏词与词之间的上下文语义级分析，导致文档分类分级的高误报率与高漏报率。如商务合同与涉诉文档，虽然都会存在甲方、乙方、联系电话、联系方式等关键词，但在不同类型文档语境下，其文档表达含义与分类截然不同。而且，由于文档分类定级误判，会产生敏感文档的非授权明文外发风险，进而导致数据泄露。由于分类误判对非敏感级文档进行加密、阻断处理，影响正常业务开展[6-8]。

2.1　面临的技术问题

通过自然语言处理技术检测识别文档敏感内容，已经演化为文档智能分类分级的技术趋势，在落地实施过程中，也面临如下突出的技术问题[9]。

在文档语义特征表达层面，各领域文档形式多样、内容丰富、中文语体混杂的情况普遍存在，例如公安电子笔录形式多样，官方语体和方言语体交相混杂，且上下文语义高度相关。而目前的语言模型建立在朴素贝叶斯独立性假设的基础上，现实情况明显无法满足独立性假设要求，从而导致文档分类分级的准确率较低[9-10]。

在文档内容识别层面，敏感文档的分类检测技术分为有监督和无监督两种模式，无监督文档分类技术不需要样本但准确率很低，有监督文档准确率相对较高，但存在样本标注工作量大、人工标注质量无法保证的问题，直接导致产品研发周期长、成本高、效果差的问题[11]。

针对文档分类分级在实际应用中面临的技术问题，本文提出如下的解决思路：构建 N 元中文语言模型，解决当前语言模型面临的独立性假设不成立问题。构建自动无监督样本库，解决当前样本标注工作量大，质量难保证问题。

2.2　技术方案

2.2.1　语言模型基本原理

语言模型的发展经历了专家语法规则模型、统计语言模型和神经网络语言模型。专家语法规则模型在初始阶段利用模式匹配技术，以自然语言的语法规则为切入点，模式匹配归纳总结，提供自然语言建模能力，但随着语法规则规模急剧扩大，专家语法规则已不可持续[12]。

统计语言模型认为文档由单词序列构成，通过建模，分析文档中单词的概率分布。统计语言模型基于朴素贝叶斯的独立性假设，即将句子 S 看成一组相互独立的单词序列 $(w_1w_2\cdots w_n)$，那么对于任意一个单词 $w_i(i\leqslant n)$在句子 S 中出现的条件概率为 $P(w_i|w_1w_2\cdots w_{i-1})$，那么整个句子的概率模型 $P(S)=P(w_1w_2\cdots w_n)$，根据贝叶斯公式和全概率公式可推导出概率模型：

$$P(S)=P(w_1w_2\cdots w_n)=\prod_{i=1}^{n}P(w_1)P(w_2|w_1)\cdots P(w_i|w_1w_2\cdots w_{i-1}) \tag{1}$$

式 (1) 必须在基于朴素贝叶斯独立性假设条

件的情况下才成立，但是在实际应用中朴素贝叶斯独立性假设往往不成立。

神经网络语言模型为了应对独立性假设不成立的问题，采用因果思路，在语料库庞大到能覆盖所有可能语义表述的情况下，通过训练的方法获得条件概率 $P(w_i|w_1w_2\cdots w_{i-1})$。但是收集这样一个庞大的语料库本身就是一个问题，在特别的应用领域，例如各地方言与官方语体混杂语料的收集明显不可行[13]。

2.2.2　*N* 元中文语言模型构建过程

因此在统计语言模型的基础上提出 *N* 元中文语言模型，其核心思想是将语言模型划分为 *N* 个相互独立的子模型，分段训练，利用线性插值公式分步整合。其实施过程：

1) 将语料库按领域和语体划分为 *N* 个语料库子集；

2) 针对各个语料子集训练特定语言模型；

3) 使用线性插值公式，获得整个语言模型：

$$\widehat{P}\left(w_i|w_1^{i-1}\right)=\sum_{j=1}^{n}\lambda_j\widehat{P}M_j\left(w_i|w_1^{i-1}\right)$$

式中 $0\leqslant\lambda\leqslant1$。$\lambda$ 的期望最大值 (expectation maximum, EM) 的迭代计算方法为

1) 对于 *N* 个语言模型，使用随机数初始化 λ；

2) 根据整个语言模型公式计算新的语言模型概率；

3) 第 *r* 次迭代，第 *j* 个语言模型在第 $i(i\leqslant n)$ 类上 λ 的计算公式为

$$\lambda_{ij}^{r}=\frac{\lambda_{ij}^{r-1}P_{ij}(w|h)}{\sum_{i=1}^{n}\lambda_{ij}^{r-1}P_{ij}(w|h)}$$

式中 *h* 为历史。

4) 不断迭代，重复步骤 2)、3)，直至收敛。

通过上述方法，将中文语言模型的困惑度值从 320 降低到 150 以下，为自然语言处理后续任务奠定基础。

2.2.3　无监督算法构建样本库基本原理

常用的无监督聚类包括 *K* 均值聚类、均值漂移聚类、基于密度的聚类方法、高斯混合模型的最大期望聚类和层次聚类等。经实验验证效果最好的方法是 *K* 均值聚类。*K* 均值聚类首先随机选择 *K* 个中心点，其次计算每个样本到 *K* 个中心点的欧氏距离，然后将每个样本划分到离它最近的中心点所属类族，最后更新每个类族的中心点，重复迭代直到所有的样本不再被重新分类为止[14]。此过程不需要人工标注样本，无须人工干预就可以自动区分样本及类别，但是 *K* 均值聚类面临两个问题：

1) *K* 的取值问题。在业务层面，企业往往也无法提供文档类别数；在技术层面，*K* 均值聚类算法的 *K* 取值本身也是一个关键问题。

2) 准确率低问题。*K* 均值聚类算法，虽然是效果最好的无监督算法，但是准确率也只能到 70%，与构建样本库的要求还相去甚远。

2.2.4　基于无监督算法的样本库构建过程

针对上述问题，基于无监督算法的样本库构建过程如下：

1) 从生产环境网络出口收集大量文档，作为样本集 *Q*；

2) 采用 *N* 元中文语言模型对样本 *Q* 中的文档进行特征提取；

3) 采用数据分析工具 Pandas 对样本集 *Q* 中的文档特征进行特征对齐；

4) 采用非线性降维的算法 (uniform manifold approximation and projection for dimension reduction, UMAP) 降维文档特征，降低文档特征复杂性，提高聚类准确率；

5) 确定 *K* 均值聚类算法的 *K* 值，具体原理和方法为：假设真实类别数为 *N*，所有样本到其所属类族中心的距离的平方和为 *D*，随着 *K* 值增加，样本划分的类族越来越精细，每个类族的内聚程度会越来越高，那么平方和 *D* 会越来越小；当 *K* 值小于 *N* 时，增加 *K* 值时会大幅增加每个族的聚合程度，故平方和 *D* 下降梯度会很大，当 *K* 值大于等于 *N* 时，继续增加 *K* 值，类族内部的聚合效果不再明显，所以平方和 *D* 下降梯度会急剧变小，平方和 *D* 下降梯度拐点即为真实聚类数 *K*：

$$D=\sum_{i=1}^{K}\sum_{P\in C_i}|P-\boldsymbol{M}_i|^2$$

式中：*i* 为类族编号；C_i 为第 *i* 个类族；*P* 为 C_i 中的某一个文档数据；$\boldsymbol{M}_i$ 为 C_i 的类族中心点向量；*D* 为所有样本到其所属类族中心的距离的平方和；

6) 按照 5) 确定的 *K* 值，对样本集 *Q* 进行聚类，得到聚类结果，由于 *K* 均值聚类算法准确率离样本库准确率要求太远，优化 *K* 均值聚类算法提升准确率的投入大而回报小，所以不可取。因此采用将无监督转换为有监督，分批迭代，投票筛选样本的办法来解决此问题；

7) 从 *K* 均值聚类结果中挑出一部分离类族中心距离小于预设阀值 *M* 的文档作为样本集 *Y*，可以调节 *M* 的取值，确保挑中样本集 *Y* 的准确性；

8) 将样本集 *Y* 分为训练集和验证集；

9) 分别采用支持向量机、TextCNN、邻近算法建模，并使用样本集 *Y* 训练模型；

10) 从 Q 中取出一批样本，分别使用已训练的支持向量机、TextCNN、邻近算法模型进行预测；

11) 使用少数服从多数的投票法，对预测结果进行合并，挑出至少有 2 个模型预测结果均一样的样本，将这些样本合并到样本集 Y 中；

12) 重复步骤 10)~11)，直至样本数量达到要求。

3　基于无标记样本的实时用户异常行为分析技术

随着企业 IT 架构日益复杂，业务系统逐步增多，大量内部隐藏、持久、缓慢的数据泄露行为成为企业数据安全关注的重点。传统的单点式安全防护措施，对于具备一定反侦查能力的数据泄露行为，毫无办法。为了解决隐蔽数据泄露问题，采用基于人工智能的用户异常行为分析技术，从传统的特征规则和人工分析转向大数据和机器学习驱动的新型安全模式[14]。

基于从正常行为中发现规律，从规律中挖掘异常的思路：首先，基于历史数据构建用户个体行为、群体行为、场景行为三大行为基线，从个体、群体、场景 3 个角度诠释正常用户行为；然后，采用神经网络，建立用户异常行为模型，对用户异常行为进行跟踪、预测与判定。模型具备实时流处理、离线批量处理两项能力，实时流用于实时分析日志，动态发现异常，并进行跟踪。离线批量处理，用于对历史数据进行分析判定，在历史数据中发现已发生的异常行为[15] 如图 1 所示。

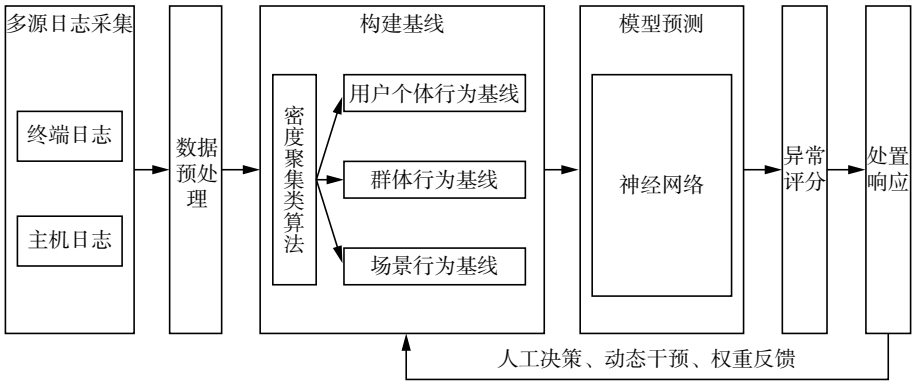

图 1　模型构建过程

Fig. 1　Build model process

在用户异常行为分析中，采用大数据和机器学习构建异常检测模型，显著提高异常行为识别能力，缩短策略升级严重滞后问题，异常行为检测时间控制在 24 h 以内，自动优化模型，实现自适应动态风险识别[15]。

3.1　面临的技术问题

目前采用机器学习技术帮助企业明确用户与实体的行为基线，检测偏离基线的异常行为，防止数据泄露。但在实际生产环境中，由于样本依赖度过高而导致异常行为识别准确率较低。

现有的用户异常行为分析系统需要通过大量已标注样本数据，训练生成用户异常行为分析模型，但是在实际应用中，用户往往不具备足够的已标注样本数据，从而导致系统识别准确率偏低[16-17]。

3.2　解决思路

为解决目前用户异常行为分析过程中样本依赖过高的问题，构建无样本依赖、实时识别异常行为的分析系统。本文采用无监督异常点检测算法，以海量行为日志数据为基础，计算单一操作行为异常，自动构建异常行为样本库，解决样本依赖问题。

3.3　技术方案

3.3.1　离群点检测构建异常行为样本库基本原理

由于异常行为与正常行为日志数据混杂存储，异常数据呈现出以下特点：

1) 异常数据的特征与大多数数据的特征是不一致的，例如同一个员工，大多数情况下都会使用自己的账号登录业务系统，员工每天的工作内容都基本相同，其行为也应基本相同。如果某天在自己办公电脑上使用其他账号登录，就可能是非正常行为。

2) 异常数据在企业数据中的占比较小，例如同一岗位的不同员工，大多数情况下工作内容基本相同，其工作产生的行为日志特征也基本相同，如若某员工与其他同岗位员工行为不太一致，就可能是非正常行为。

针对上述日志数据的特点，可以采用基于统计的、距离的、密度的量化指标，刻画异常数据与全集数据的疏离程度，即离群点检测方法[18]。

3.3.2　离群点检测构建异常行为样本库构建过程

依据企业日志数据的特点，企业用户异常行为样本库构建过程如下：

1) 在用户终端部署终端日志采集器，收集用户操作行为数据，如系统登录、文档操作、软件使用、邮件外发、上网行为等日志数据；

2) 在数据库、应用服务器、交换机等位置部署设备日志采集器，采集防火墙、业务系统、数据库等系统与设备的用户操作日志；

3) 定义用户异常行为场景；

4) 依据异常行为场景，确定场景所需的用户行为数据；

5) 依据场景所需的用户行为数据，从用户及操作类型的角度切分数据；

6) 对切分好的数据按照场景要求进行特征提取；

7) 使用邻近算法、离群点检测算法和 Feature Bagging 等算法，分析建模特征数据，分离出正常数据和异常数据；

8) 按场景聚合正常数据，形成样本库。

4 实验系统

4.1 系统架构

构建数据分类分级与异常行为分析实验系统，分析验证技术可行性与应用效果。

实验系统采用 Spring Cloud 微服务架构，使用应用容器 Docker 进行部署，采用关系型与文件型混合存储模式，通过 Nginx Web 服务器实现负载均衡。系统架构包括采集层、分析层、存储层、可视化层。

4.2 数据采集层

实验系统采用探针主动采集和接口被动采集两种模型，管理员可以通过图形化界面选择采集模型，定义探针主动采集频率、被动采集服务地址。主动采集探针部署在操作终端、交换机、应用服务器、数据库审计系统，主动采集日志。被动采集服务部署在服务端，通过标准接口，对接操作终端、交换机、应用服务器、数据库审计等系统上报的日志数据。日志数据上报后，数据流转引擎将日志数据推送至分布式消息系统 Kafka 数据队列组件中准备进行数据清洗处理与持久化。

采集探针利用日志处理引擎 Logstash 的 filebeat 组件，采用 TCP、UDP 协议将日志数据上报到采集层，通过网络和应用监测系统 Zabbix Agent 将日志数据上报至 Zabbix 服务器。

采集层汇集日志数据后，经过数据匹配、解析以及组装操作，将数据推送到采集层中的 Kafka 队列中，通过数据分流操作以及动态模板匹配操作将数据推送到数据搜索引擎 Elasticsearch 中完成日志数据的存储流程。

采集层通过 Logstash 的 input、filter、output 组件对数据进行匹配、解析、组装，通过 HTTP Client 读取 Zabbix 服务端的数据，进行处理操作后将数据上报到数据存储层。

4.3 数据存储层

存储层存储训练用户异常行为分析模型需要的日志数据、训练分类分级模型需要的文档样本数据、用户异常行为告警数据、安全事件审计数据。其中日志数据包括终端、网络设备、主机设备、数据库、应用系统日志数据。文档样本数据包括覆盖业务范围内各业务分类与各安全等级的文档。

数据存储层使用 Spark Streaming 从 Kafka 中读取主动探针采集、被动接口采集汇总的日志数据，使用 Spring Boot 同步数据、生成文件，通过 Common-pool2 创建数据库连接池，通过 AbstractRoutingDataSource 访问关系型数据库，使用 Elasticsearch 和 Mysql 存储用户异常行为告警数据、安全事件审计数据，使用 Redis 存储日志数据与文档样本数据，采用 Mysql 存储规则、用户权限等系统管理数据。

4.4 数据分析层

分析层包括基本规则与高级模型两种分析方式，其中基本规则是采用关键字、正则表达式等简单匹配模型识别文档类别、安全等级与异常行为。高级模型包括文档分类分级模型、用户异常行为分析模型。

数据分析层包括工作流框架、特征提取、算法模型、格式化输出等主要功能。工作流框架负责构建业务流、调度数据处理任务，特征提取负责选择各种数据的业务特征、特征提取和特征对齐，根据业务场景和数据分布选择算法模型，对接业务系统输出分析结果。

业务工作流采用 SpiffWorkflow 工作流框架，实现动态配置业务流程、动态选择数据特征，动态配置算法模型、灵活调度业务场景。

特征选择和特征提取采用 *N*-Gram 语言模型、词向量方法、线性差别分析、主成分分析、奇异值分解等技术对文档数据和行为数据进行特征选择和特征提取。

算法模型采用插件化封装，根据数据分布特点选择算法模型。算法模型包括 *K* 均值聚类、密度聚类方法、高斯混合模型，以及支持向量机、决策树、邻近算法、长短时记忆网络、神经网络。

采用 JSON、XML 协议封装输出数据，满足上层业务系统对接要求。

4.5 数据可视化层

可视化层用于审计分析用户异常行为安全事

件、查看文档分类分级结果。

可视化层使用前后端完全分离的机制, 采用 VUE.JS 结合 Webpack 搭建前端架构, 使用 HTTP 协议进行前后端的数据通信。

可视化层的接口服务模块, 外部接口使用 Netty 进行 TCP 协议数据交互、HttpClient 进行 HTTP 协议数据交互。内部接口使用 Aviator、Elasticsearch、Mysql、Common-pool2 以及 Redis 实现。

5 实验分析

实验环境包括硬件和软件配置两部分。

硬件配置: 两台测试机器, CPU Intel Core i3-4130 3.40 GHz 4 核, 内存 8 GB, 硬盘 5 TB, 网卡 1000 MB。

软件配置: 操作系统 CentOS7.6, 数据库 MySQL5.7.29。

5.1 基于 *N* 元语言模型的文档智能分类分级技术实验过程

基于 *N* 元语言模型的文档智能分类分级技术, 验证数据分类分级准确率, 采用业界与学术界公认的测试基准数据集。首先使用训练数据集构建模型, 然后使用测试数据集评价模型, 其中标准准确率 $ACC = \frac{TP}{ALL}$, TP 为预测正确的样本数量, ALL 为测试集样本总数。

1) 实验准备

①服务器、客户端系统部署完成、网络通信正常;

②准备 24 000 份样本文件, 其中财经、体育、娱乐、时政各 6 000 份;

③测试发件箱 test@wondersoft.cn, 收件箱 test@shou.com。

2) 实验输入

①登录控制台, 创建财经、体育、娱乐、时政分类, 创建 4 个安全等级、定义每个安全等级涉及的敏感数据形式, 导入样本文件, 构建分类分级模型;

②使用 foxmail 发件箱 test@wondersoft.cn 发送各种类型、各种安全等级文档至 test@shou.com, 检查分类分级结果。

5.2 基于无标记样本的实时用户异常行为分析技术实验过程

基于无标记样本的实时用户异常行为分析技术, 验证用户异常行为识别准确率, 采用业界与学术界公认的测试基准数据集。首先使用训练数据集构建模型, 然后使用测试数据集评价模型, 其中标准准确率 $ACC = \frac{TP}{ALL}$, TP 为预测正确的样本数量, ALL 为测试集样本总数。

1) 实验准备

①服务器系统搭建, GPU 驱动安装、tensorflow 环境搭建, 客户端系统搭建和客户端软件安装;

②准备 3 000 名用户、30 万条日志数据。日志数据包括邮件外发日志数据、U 盘拷贝文件日志数据、账号准入登录日志数据 3 种类型。其中邮件外发日志数据包括操作终端信息、发件人信息、收件人信息、抄送人信息、正文信息、附件信息、时间信息; U 盘拷贝文件日志数据包括 U 盘设备信息、操作终端信息、拷贝操作信息、操作文件信息、操作时间信息; 账号准入登录日志数据包括人员账号信息、接入设备信息、接入网络信息、管控策略信息、接入时间信息。

2) 实验输入

①首先将日志数据导入 mongodb, 3 种行为日志数据单独存放。其次登录实验系统控制台, 配置数据源和检测模型。最后使用准备数据训练得到用户行为模型;

②使用客户端执行邮件外发、U 盘拷贝文件和账号登录系统操作;

③服务端接收客户端日志信息, 对用户操作进行异常行为检测, 并将结果展示在控制台。

5.3 实验结论

基于 *N* 元语言模型的文档智能分类分级技术, 数据分类分级准确率达到 93%。

4 种分类数据、4 个安全等级数据组成了 16 种数据集合, 每种数据集合各执行 500 次数据发送。进入实验系统数据分类分级界面, 可查看到 8 000 次数据发送邮件, 按 16 组统计分析分类分级模型准确率分别为 96.2%、93.6%、94.8%、87.4%、90.1%、92.1%、97.2%、86.2%、89.2%、90.1%、96.4%、92.7%、91.6%、87.4%、90.1%、88.3%。

基于无标记样本的实时用户异常行为分析技术, 用户异常行为分析准确率达 86% 以上。

邮件外发、U 盘文件拷贝和账号上下线操作各执行 500 次, 进入实验系统用户行为分析界面, 可查看到邮件外发记录 500 次、U 盘文件拷贝 500 次、账号登录系统 500 次, 统计分析异常行为识别准确率分别为 86.7%、88.6%、86.2%。

6 结束语

在数字经济时代, 数据安全是保障数字经济发展的基础条件, 数据安全也正在从传统的边界防护逐步转向以数据为基础的纵深全链条防护, 在这个转变过程中, 准确地识别数据, 找到管控重点, 成为关键问题。人工智能赋能数据安全创新地解决了海量数据中敏感数据识别、数据分

类分级和用户异常行分析识别问题。本文为数据安全领域使用人工智能技术解决行业普遍性问题进行了前期的尝试与探索。同时，我们也应看到，人工智能技术在数据安全领域的应用才刚刚开始，如何通过人工智能技术低成本、高效率地解决数据安全问题，还需要广大从业者不断努力探索。

参考文献：

[1] 新华社. 中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见 [EB/OL]. 新华社,2020(2020-03-20). http://www.gov.cn/zhengce/2020-04/09/content_5500622.htm.

[2] ERIC OUELLET, JEFFREY WHEATMAN. Typical elements of an enterprise data security program[EB/OL]. Gartner,2015(2020-10-16).https://www.gartner.com/en/documents/1153112.

[3] 罗军舟, 韩志耕, 王良民. 一种可信可控的网络体系及协议结构 [J]. 计算机学报, 2009, 32(3): 391−404.
LUO Junzhou, HAN Zhigeng, WANG Liangmin. Trustworthy and controllable network architecture and protocol framework[J]. Chinese journal of computers, 2009, 32(3): 391−404.

[4] 王琨, 陆艳军. 数据文件安全管控技术的研究与实现 [J]. 信息安全研究, 2018, 4(1): 84−90.
WANG Kun, LU Yanjun. Research and implementation of security management of data files[J]. Journal of information security research, 2018, 4(1): 84−90.

[5] MENEZES A J, VAN OORSCHOT P C, VANSTONE S A. Handbook of applied cryptography[M]. Boca Raton: CRC Press, 1997.

[6] NOVAK R. SPA-based adaptive chosen-ciphertext attack on RSA implementation[C]//5th International Workshop on Practice and Theory in Public Key Cryptosystems. Paris, France, 2002: 252−262.

[7] FOUQUE P A, MARTINET G, POUPARD G. Attacking unbalanced RSA-CRT using SPA[C]//5th International Workshop on Cryptographic Hardware and Embedded Systems. Cologne, Germany, 2003: 254−268.

[8] 李增局, 彭乾, 史汝辉, 等. CRT-RSA 算法的选择明文攻击 [J]. 密码学报, 2016, 3(5): 447−461.
LI Zengju, PENG Qian, SHI Ruhui, et al. Chosen plaintext attacks on CRT-RSA[J]. Journal of cryptologic research, 2016, 3(5): 447−461.

[9] 赵冬斌, 邵坤, 朱圆恒, 等. 深度强化学习综述: 兼论计算机围棋的发展 [J]. 控制理论与应用, 2016, 33(6): 701–717.
ZHAO Dongbin, SHAO Kun, ZHU Yuanheng, et al. Review of deep reinforcement learning and discussions on the development of computer[J]. Control theory and applications, 2016, 33(6): 701–717.

[10] 刘朝阳, 穆朝絮, 孙长银. 深度强化学习算法与应用研究现状综述 [J]. 智能科学与技术学报, 2020, 2(4): 314–326.
LIU Zhaoyang, MU Chaoxu, SUN Changyin. An overview on algorithms and applications of deep reinforcement learning[J]. Chinese journal of intelligent science and technology, 2020, 2(4): 314–326.

[11] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述 [J]. 计算机学报, 2018, 41(1): 1–27.
LIU Quan, ZHAI Jianwei, ZHANG Zongchang, et al. A survey on deep reinforcement learning[J]. Chinese journal of computers, 2018, 41(1): 1–27.

[12] 万里鹏, 兰旭光, 张翰博, 等. 深度强化学习理论及其应用综述 [J]. 模式识别与人工智能, 2019, 32(1): 67–81.
WAN Lipeng, LAN Xuguang, ZHANG Hanbo, et al. A review of deep reinforcement learning theory and application[J]. Pattern recognition and artificial intelligence, 2019, 32(1): 67–81.

[13] 窦宇宸, 胡勇. 基于 BERT 的安全事件命名实体识别研究 [J]. 信息安全研究, 2021, 7(3): 242–249.
DOU Yuchen, HUYong. Research on name entity recognition of security events based on BERT[J]. Journal of information security research, 2021, 7(3): 242–249.

[14] 刘思琴, 冯胥睿瑞. 基于 BERT 的文本情感分析 [J]. 信息安全研究, 2020, 6(3): 220–227.
Liu Siqin, FengXuruirui. Text Sentiment Analysis Based on BERT[J]. Journal of information security reserach, 2020, 6(3): 220–227.

[15] 陈红松, 王钢, 宋建林. 基于云计算入侵检测数据集的内网用户异常行为分类算法研究 [J]. 信息网络安全, 2018, 18(3): 1–7.
CHEN Hongsong, WANG Gang, SONG Jianlin. Research on anomaly behavior classification algorithm of internal network user based on cloud computing intrusion detection data set[J]. Netinfo Security, 2018, 18(3): 1–7.

[16] 张建平, 李洪敏, 贾军, 等. 一种基于流量与日志的专网用户行为分析方法 [J]. 信息安全研究, 2020, 6(9): 783−790.
ZHANG Jianping, LI Hongming, JIA Jun, et al. A method of user behavior analysis based on network flow and log in private network[J]. Journal of information security research, 2020, 6(9): 783−790.

[17] 蹇诗婕 卢志刚 姜波, 等. 基于层次聚类方法的流量异常检测 [J]. 信息安全研究, 2020, 6(6): 474–481.
SAI Shiyi, LU Zhigang, JING Bo. Flow anomaly detection based on hierarchical clustering method[J]. Journal of information security reserach, 2020, 6(6): 474–481.

[18] SUTTON R S. Learning to predict by the methods of temporal differences[J]. Machine learning, 1988, 3(1): 9–44.

作者简介：

喻波，北京明朝万达科技股份有限公司首席科学家、高级副总裁，兼任公安部通信标准化技术委员会委员，主要研究方向为数据安全。主持国家重点研发计划、国家自然科学基金重点项目 6 项。获吴文俊人工智能科技进步奖一等奖、授权发明专利 120 余项。

王志海，北京明朝万达科技有限公司董事长、总裁，计算机安全专业委员会常务委员、《信息安全技术》编委会委员，中国大数据生态产业联盟专家委员，国家移动信息产业技术创新战略联盟理事长，主要研究方向为数据安全。获吴文俊人工智能科技进步奖一等奖、授权发明专利 120 余项，出版专著 1 部。

孙亚东，数据安全解决方案专家，主要研究方向为数据安全。获吴文俊人工智能科技进步奖一等奖、申请国家发明专利 13 项，授权 3 项，参与编写数据安全国家标准 2 项。

第三届国际高性能大数据暨智能系统会议

The 3rd International Conference on High Performance Big Data and Intelligent Systems

第三届国际高性能大数据暨智能系统会议（The 3rd International Conference on High Performance Big Data and Intelligent Systems, HPBD&IS 2021）拟于 2021 年 12 月 5-7 日在中国澳门举办。会议旨在搭建高性能计算、大数据及人工智能领域高端前沿交流平台，促进海内外专家学者的交流与合作，推动智能技术进步和智能产业发展。

会议主旨：

本次会议将汇聚全球顶级专家、学者和产业界优秀人才，共同围绕国际研究热点、核心关键技术、产业发展及挑战等进行开放式研讨。会议由中国计算机学会 (CCF)、中国人工智能学会 (CAAI) 联合主办，IEEE Computer Society 技术支持，澳门大学、中国科学院深圳先进技术研究院、中国科学院半导体研究所、CCF 高性能计算专委会、CAAI 神经网络与计算智能专委会、中国自动化学会 (CAA) 模式识别与机器智能专委会、北京联合大学、天津理工大学、深圳市龙岗区机器人协会、深圳国际机器人城产业园共同承办。会议论文集将由 IEEE Xplore®出版，EI 收录，优秀论文将推荐至 Computation Practice and Experience (CCPE)、Optoelectronics Letters 等 SCI、EI 期刊发表。

征文内容 (包括但不限于)：

1）高性能计算技术及应用 (高性能计算机体系结构，高性能计算机系统软件，高性能计算环境，高性能微处理器，高性能存储技术，多核多线程体系结构方法，并行分布式系统的体系结构、软件及算法，高性能普适计算，高性能自适应进化计算，高性能区块链技术，高性能应用，大数据并行处理，大数据硬件/操作系统加速)；2）大数据技术及应用 (大数据模型、处理算法及编程技术，多媒体大数据的表示，大数据学习与分析，大数据持久性和保存，大数据的质量和来源控制，大数据保护、完整性和隐私，大数据存储与计算融合技术，大数据搜索与挖掘，大数据管理与可视化分析，大数据业务模式创新，大数据应用)；3）智能系统 (神经网络与学习系统，计算机视觉，机器人科学与控制工程，智能传感器与传感器融合，智能存储设备与系统，实时系统，区块链系统，自适应系统，AR/VR/MR，复杂系统和网络，智能制造，模式识别，SLAM)。

重要时间：

全文投稿截止日期：2021 年 9 月 30 日

论文录用通知日期：2021 年 10 月 31 日

全文提交截止日期：2021 年 11 月 15 日

早鸟注册截止日期：2021 年 11 月 15 日

投稿链接：https://easychair.org/conferences/?conf=hpbdis2021

会议官网：http://www.hpbdis.org/