运营商数据安全合规检查技术研究与实践

安 鹏[1] 李宏飞[1] 高 铭[2] 王世彪[1] 喻 波[1]
[1](北京明朝万达科技股份有限公司 北京 100142)
[2](中国移动通信集团宁夏有限公司 银川 750002)
(anpeng@wondersoft.cn)

Research and Practice on Data Security Compliance Check Technology for Operators

An Peng[1], Li Hongfei[1], Gao Ming[2], Wang Shibiao[1], and Yu Bo[1]
[1](*Beijing Wondersoft Technology Co., Ltd., Beijing* 100142)
[2](*China Mobile Ningxia Co., Ltd., Yinchuan* 750002)

Abstract In the context of the development of the global digital economy, data has become an important asset for enterprises. China positions data as one of the national basic strategic resources and innovative elements of social production. In recent years, the proliferation of ransomware attacks from hackers has posed a significant risk of data leakage to enterprise data security management. Secondly, unconscious data-sharing operations by employees during the production process are also an important way for enterprise data asset leakage. With the promulgation of the Data Security Law, regulatory agencies have made data security reviews a part of the industry security inspections for operators. Therefore, based on regulatory compliance, research and practice related inspection technologies to help operators enhance their security inspection capabilities, ensure data security, and meet the needs of compliance regulation and business development.

Key words compliance check; content identification; clustering algorithm; operator; data security

摘 要 在全球数字经济发展的大背景下,数据已成为企业的重要资产.我国将数据定位为国家基础战略性资源和社会生产创新要素之一.近年来黑客勒索攻击的泛滥对企业数据安全管理造成很大的数据泄露风险,员工在生产过程中无意识的数据分享操作也是当前企业数据资产泄露的重要途径之一.随着《中华人民共和国数据安全法》的颁布,监管机构把数据安全审查作为运营商行业安全检查的内容.因此基于监管合规性,研究相关检查技术并进行实践,从而帮助运营商增强安全检查能力,保障数据安全,满足合规监管与业务发展需要.

关键词 合规检查;内容识别;聚类算法;运营商;数据安全

中图法分类号 TP309.2

收稿日期:2023-04-27
通信作者:李宏飞(lihongfei@wondersoft.cn)
引用格式:安鹏, 李宏飞, 高铭, 等. 运营商数据安全合规检查技术研究与实践[J]. 信息安全研究, 2023, 9(7): 643-647

1 研究背景

近年来，随着国家对数据安全的重视程度空前，监管也愈加严格，极大地促进了数据安全行业的发展，推动企业数据安全建设进入“快进模式”.但随着安全建设的不断深入，各种规划、制度、要求的增多，在进入安全运营阶段可能会出现制度未落地、要求未执行到位的情况，合规检查的存在就尤为重要.

对于运营商来说，在数字化转型的进程中，数据作为生产要素的重要性越发凸显.而随着生产组织模式的逐步转变，数据在个人终端存储、内部业务系统流转、数据离网及外发等环境下都存在泄露的风险[1].因此，通过实施合规检查工作，帮助运营商克服数据质量和数据安全问题，保障企业数据的安全合规，对企业数据资产的安全稳定和今后业务的良性发展有着重要的意义.

因此，本文通过对数据检查技术研究，提出适用运营商行业数据安全合规的关键技术，提供对数据安全合规检查工作的多样手段，提高数据安全合规检查准确性，帮助运营商增强安全检查能力，杜绝信息泄露，满足监管要求，保障客户隐私权益.为数据资源可视、可管、可控，促进数据有序流动保驾护航.让数据赋能千行百业，充分发挥数据生产要素的价值，推动数字化转型，提升国家的综合竞争力.

2 现状分析

2.1 合规检查现状

1）新形势下常规手段力有未逮.

目前，数据类型多且杂，散落在各个应用系统、办公人员电脑中，数据结构、数据类型、存储形式、重要程度各不相同.据调研机构统计，从泄露数据的数量来看，个人信息数据约有868.8亿条，占比为91.4%，其次是运营数据，约有79.5亿条，占比为8.4%，二者之和占泄露数据总量的99.8%.在商业数据泄露方面，主要涉及文档数据泄露、代码数据泄露和文化版权数据盗版3大类数据泄露风险类型.数据的复杂性使得常规依靠问询和调研性质的检查工作已不能达到预期效果，无法保证检查测评结果的全面和准确.

2）需保护的目标数据无从发现.

无论是数据安全保护还是数据安全合规，其目标都是重要数据和敏感数据.其界定标准来自于国家法规和行业、企业标准，由于缺乏精准高效的内容识别手段，要找到法规标准所描述的需保护的目标数据如同盲人摸象，事倍功半.

3）如何平衡数据使用的便利性、安全性和合规性.

运营商面临的最大挑战是既要保护敏感数据资产的安全，保持符合监管合规性，又要在不改变现有业务流程的前提下快速部署实施，降低自身员工生产过程中使用数据资产的操作复杂度，提升生产效率.

2.2 监管要求

下面以落实《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律关于数据安全管理规定的《网络数据安全管理条例（征求意见稿）》（以下简称《征求意见稿》）中关于合规监管要求为例进行介绍.

1）分类分级，重点保护.

《征求意见稿》第5条规定：“国家建立数据分类分级保护制度.按照数据对国家安全、公共利益或者个人、组织合法权益的影响和重要程度，将数据分为一般数据、重要数据、核心数据，不同级别的数据采取不同的保护措施”.同时要求“各地区、各部门应当按照国家数据分类分级要求，对本地区、本部门以及相关行业、领域的数据进行分类分级管理”，可见对于数据分类分级管理，要根据不同地区、不同行业制定出针对性和个性化的数据分类分级标准，使分级管理合乎规范.

2）技术措施，红线要求.

《征求意见稿》对于数据处理者的数据安全管理和数据安全事件的应对，提出了更高的要求，从被动应对的消极保安全到自主自发的积极保安全，提出了基于风险管理的红线要求.第9条规定：“数据处理者应当采取备份、加密、访问控制等必要措施，保障数据免遭泄露、窃取、篡改、毁损、丢失、非法使用，应对数据安全事件，防范针对和利用数据的违法犯罪活动，维护数据的完整性、保密性、可用性”“数据处理者应当按照网络安全等级

保护的要求，加强数据处理系统、数据传输网络、数据存储环境等安全防护，处理重要数据的系统原则上应当满足3级以上网络安全等级保护和关键信息基础设施安全保护要求，处理核心数据的系统依照有关规定从严保护”.

3）违法必究，执法必严.

《征求意见稿》首先明确了监管部门和其对应的职能，第55条规定：“国家网信部门负责统筹协调数据安全和相关监督管理工作”“公安机关、国家安全机关等在各自职责范围内承担数据安全监管职责”“工业、电信、交通、金融、自然资源、卫生健康、教育、科技等主管部门承担本行业、本领域数据安全监管职责”.对于违法追究方面，除了常规的警告、罚款、责令暂停相关业务，停业整顿，吊销相关业务许可证或者吊销营业执照，没收违法所得，特别值得注意的是已经上升至刑法高度，第70条规定：“数据处理者违反本条例规定，给他人造成损害的，依法承担民事责任；构成违反治安管理行为的，依法给予治安管理处罚；构成犯罪的，依法追究刑事责任”.

3 数据安全合规检查思路

3.1 现状调研

运营商单位在进行数据安全合规检查之前，首先要进行充分的现状调研，梳理运营商单位的基本信息、数据情况、整体制度和安全防护情况.如果现状不明确可能会导致数据安全合规检查工作没有头绪，无法开展.因此，要在数据安全合规检查工作开始前进行现状调研，明确合规检查的范围、对象，并将整理结果汇总至相关责任人.

3.2 确定内容

数据安全合规检查可以按内容分为管理检查项和技术检查项.也可以按照检查方式分为人工检查和工具检查.还可以根据数据级别分为数据安全通用防护和数据安全分级防护，为不同的数据对象确定不同的检查内容.

3.3 合规检查

在明确检查内容后迅速成立检查小组，按既定的检查计划，针对不同的检查项进行数据安全合规检查.

对于管理制度和台账文档清单可以人工进行检查，但针对数据生命周期中的数据资产清单、敏感数据识别、异常行为监测、敏感数据泄露、非法越权访问、数据跨境传输、数据非法外联等人工难以检查的内容则应该采用专用工具，通过数据安全合规检查技术进行检查，减少人工检查工作量的同时也可以量化检查.对于大部分数据生命周期技术检查项，可以使用工具进行扫描、收集和分析数据安全情况.

3.4 总结提升

数据安全合规检查不是结束，而是一个开始.在数据安全合规检查后，应基于对国家法律法规、行业标准、技术的研究，基于运营商行业安全现状，梳理出数据安全的管理需要遵循安全政策，建立科学的数据安全保护制度，解决实际存在的安全问题，增强抗击数据安全威胁的能力.结合业务发展规划、数据安全现状和数据安全合规检查情况，编制数据安全能力提升规划，为数字化转型和业务发展、数据安全建设提供决策依据.

4 关键技术

本文基于内容匹配的基础检测技术（正则表达式检测（标示符）、关键字和关键字对检测、文档属性检测）[2]，结合运营商数据安全合规要求，解决数据检查准确率低的问题，研究运营商终端数据合规检查关键技术，包括文档指纹提取对比、机器学习、K-means聚类算法等.

4.1 基于文档指纹提取对比的合规检查技术

该技术可提升合规检查数据的识别能力，使数据识别具有普遍性，进一步缩小误报率，并且在高识别率的情况下指纹文件的压缩比例能达到100∶1；设计并实现一种前后端分离的结合文档指纹库获取、快速对比的技术，可以做到实时比较是否与指纹库文档存在重合；可以识别多篇小段摘抄情况，并且可以识别具体的摘抄部分与涉密文章的重合比例.

文档指纹提取流程为：

1）启动服务端接收请求；接收到训练请求时读取请求训练路径，获取全部文档文件；对文档文件进行分词去停用词，取字数大于1的词作为基础的元素列表.

2）将元素列表进行K-grams模型划分，生成

$(N-k)/s$ 个元素列表，N 为大于 0 的整数，表示词的个数，k 为大于 2 的整数，表示窗口大小，s 为大于 0 的整数，表示滑动窗口的大小[3].

3）将元素列表合成 1 个文本，然后使用 UTF-8 编码格式把文本转换为 1 个“单词”.将这些“单词”按照设定阈值分到 n 组，$n=N/t$ 向上取整，N 为大于 0 的整数，表示“单词”的个数，t 为大于 0 的整数，表示每块最多“单词”的个数.每块“单词”个数小于等于阈值大小[3].

4）初始化一个 num 位的最小哈希签名列表，num 为大于 0 的整数，表示使用随机哈希函数的个数.根据随机系数随机排列这些单词块里的单词，将其哈希值按照随机函数内容更新进入最小哈希表中，取更新后的 num 位的最小值，此过程重复 N 遍，N 为“单词”块中“单词”的个数.最后得到 1 个更新完毕的 num 位哈希签名.

5）对全部文档进行以上操作，将文件名和对应的最小哈希签名一起存储.

文档指纹快速对比流程为：

1）读取保存的指纹库加载，得到文档名和其对应的哈希签名.给哈希签名设置合适易读的索引值.

2）初始化局部敏感哈希表，设定阈值 t（$0<t<1$）.此哈希表默认搜索相似度高于此阈值的数（要达到最大的搜索范围，一般将此值设定为无限接近于 0 的小数）.

3）将读取的哈希签名更新，进入局部敏感哈希表中，将此哈希表存入内存中备用，完成指纹库初始化过程.接收客户端发送的请求，读取文本内容.

4）将这些内容施以与训练步骤相同的文本处理和哈希表生成操作.将得到的哈希签名与内存中初始化完成的指纹库哈希表进行快速搜索，便可快速得到与其有重合元素的哈希值和其索引值，通过索引值便可定位到此重合部分与哪（几）篇文档有重合（对比 1 次可得到多篇重合结果）.

5）得到文档后计算其哈希签名与重合文档哈希签名的相似度，再通过转换并且求和便可得到具体的重合度.

4.2 基于机器学习的合规检查技术

自然语言识别是语言信息处理的一个重要组成部分，设计人工智能的算法将设定的自然语言识别机制用于数据安全合规检查的扫描引擎中，构造出能够理解和识别自然语言的智能数据合规检查能力，利用人工智能技术，提升敏感数据检测的准确性.

在机器学习基础上与数据分类相结合，分为样本训练阶段、模型预测阶段、人工纠错阶段.样本训练阶段主要利用基于深度学习的神经网络相关算法，分析大量人工标注的原始样本集，根据文本内容的语义特征和格式自动按照内容进行主题梳理，并可通过人工干预灵活调整语义相似度，获得满意的分类效果并生成预测模型；模型预测阶段主要利用预测模型实时感知合规检查数据的使用状况，实现对合规检查数据的有效预测与精准分类；人工纠错阶段主要对预测模型的误报行为，手动标注误报功能，对预测模型的准确率造成积极正向干预，为数据动态分类分级奠定基础.

4.3 基于 K-means 聚类算法的合规检查技术

在不需要人为设定的情况下，实现由机器自动计算数据最佳聚类，并且支持对同一批样本的持续性的增量训练，从而不断地优化聚类模型，提升文档识别效果.

K-means 聚类算法逻辑流程：

1）在第 1 次聚类时，通过多次聚类，在每次聚类之后计算聚类的 davies bouldin 指数，指数最小时的 K_1 值为最佳聚类的个数，对数据进行 K-means 聚类.最后保存模型和模型的聚类中心，为下一次聚类作准备.

2）第 2 次作增量聚类时，数据模块提供增量数据和原始数据，将 2 个数据合并进行同样的多次聚类取指数操作，得到最佳聚类个数 K_2.

3）将 K_1 和 K_2 作对比，分以下 3 种情况讨论：当 $K_2<K_1$ 时，会提醒使用新增数据的分布与源模型部分，退出聚类以免损毁原始模型；当 $K_2=K_1$ 时，说明新增的数据没有产生新的类别，就以原来的聚类中心作为新一次聚类的初始聚类中心，然后再按照 K-means 的聚类规则进行聚类直至聚类中心不再发生变化；当 $K_2>K_1$ 时，说明新增的数据中加入了新的分类，需要产生新的聚类中心，那么将新增数据中离各个已有聚类中心最远的点选作新的聚类中心，将新选定的聚类中心加入聚类中心.这种方法重复 K_2-K_1 次，新增 K_2-K_1 个新聚类中心，将这些中心作为 K-means 的初始化聚类中心，进行聚类中心的迭代，如此获得最佳的聚类效果.

5 融合实践

5.1 实践概述

数据资产是当今企业和社会发展的核心生产要素，也是运营商在生产过程中不可或缺的生产对象.而针对数据采取合规检查是为清晰数据安全现状.采用上述数据安全合规检查技术，基于内容识别构建数据合规安全体系，切实防范化解数据风险，确保符合相关法律法规和标准要求.

运营商数据安全合规检查在对数据合规现状环境调查的基础上，确定需要检查的内容项，输出检查表，之后以“人工＋工具”的方式开始进行合规检查，采用内容分析引擎，利用本文提到的合规检查技术，对照行业数据分类分级规章制度和数据安全管理法律法规进行合规性检查，满足运营商数据合规检测的管理要求.

5.2 数据安全合规检查技术实践案例

某运营商监管部门业务是对二级单位的数据进行合规监管.首先监管部门检查人员对现场具体被检查人员及被检查环境进行调研，梳理运营商单位的基本信息、数据情况、整体制度和安全防护情况.下一步以运营商内部数据分类分级规章制度和数据安全管理法律法规为基础，输出检查项并与被检查单位进行确认，通过现场抽查的检查方式，收集其终端环境下的各类数据文件，对终端的数据进行解析与内容扫描，因终端数据复杂，使用文档指纹提取对比、机器学习、K-means 聚类算法技术，提高终端数据安全合规检查的准确性与高效性.减少常规识别手段效率低、识别差、人工投入多等问题.根据预置的运营商数据分类分级标准和数据安全法律法规进行合规性检查，以是否被命中判断是否存在敏感数据，一经发现数据风险，责令其整改，以满足被监管单位数据的合规要求.

6 结　　语

本文通过上述研究，形成数据安全合规检查技术，相较于传统检查技术，使敏感数据识别的漏报率降低 60%，误报率降低 50%，识别种类提高 140%，同时使得数据合规检查的精准度提升至 93%以上.可满足合规检查或常态化自查处理及为后续的防泄露建设提供基础输入.希望可以为运营商企业开展数据安全合规检查提供帮助.

参　考　文　献

[1] 严敏，何庆. 基于大数据平台敏感数据流转全生命周期监控的研究与应用[J]. 信息安全研究，2018，4(2)：145-149
[2] 马晓芳. 政府建设政务数据共享交换平台存在的风险及安全措施探讨[J]. 网络安全技术与应用，2017，12(26)：120-121
[3] 北京明朝万达科技股份有限公司. 文档检测处理方法、装置、存储介质及电子设备：中国，CN202210044868.6 [P]. 2022-05-06

安　鹏
硕士，工程师.主要研究方向为数据安全.
anpeng@wondersoft.cn

李宏飞
工程师.主要研究方向为运营商数据安全.
lihongfei@wondersoft.cn

高　铭
硕士，工程师.主要研究方向为大数据建模分析、骚扰诈骗电话数据分析、信息安全、数据安全.
gaoming@nx.chinamobile.com

王世彪
硕士，工程师.主要研究方向为数据安全、可信计算.
wangshibiao@wondersoft.cn

喻　波
正高级工程师.主要研究方向为数据安全、可信计算.
yubo@wondersoft.cn