文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

大数据环境下软件数据脱敏技术研究与应用

万耀雪
中科万福集团股份有限公司
DOI:10.32629/acair.v4i2.20411

[摘　要] 在大数据 “海量性(Volume)、高速性(Velocity)、异构性(Variety)和价值性(Value)”4V特征驱动下,数据价值与安全风险的矛盾日益凸显,使得软件数据脱敏技术成为平衡数据利用与隐私保护的核心支撑。本文基于数据脱敏技术发展演进脉络,系统梳理k-匿名、差分隐私等核心理论基础,深入剖析大数据环境下静态脱敏、动态脱敏及智能脱敏技术的实现机制与适用场景。结合金融、医疗行业典型应用案例,探讨脱敏技术在实际业务中的落地路径与实施成效,并针对当前技术面临的实时性瓶颈、多模态数据适配不足等问题,提出融合AI智能分析、隐私计算的发展方向。研究成果为提升大数据环境下数据安全防护能力、推动数据要素合规流通提供技术参考与实践借鉴。
[关键词] 大数据；数据脱敏；隐私保护；动态脱敏；行业应用
中图分类号：N37　文献标识码：A

Research and Application of Software Data Desensitization Technology in the Big Data Environment

Yaoxue Wan
Zhongke Wanfu Group Co., Ltd.

[Abstract] Driven by the 4V characteristics of big data—Volume, Velocity, Variety, and Value—the contradiction between data value and security risks has become increasingly prominent, making software data desensitization technology a core enabler for balancing data utilization and privacy protection. Based on the evolutionary trajectory of data desensitization technologies, this paper systematically reviews key theoretical foundations such as k-anonymity and differential privacy, and provides an in-depth analysis of the implementation mechanisms and application scenarios of static desensitization, dynamic desensitization, and intelligent desensitization in the big data environment. Drawing on typical use cases in the financial and healthcare industries, the paper explores the practical implementation paths and effectiveness of desensitization technologies in real-world business scenarios. Furthermore, in response to current challenges including real-time performance bottlenecks and insufficient adaptability to multimodal data, the paper proposes future development directions that integrate AI-driven intelligent analysis and privacy-preserving computing. The research outcomes provide technical reference and practical insights for enhancing data security protection capabilities and promoting the compliant circulation of data elements in the big data environment.
[Key words] big data; data desensitization; privacy protection; dynamic desensitization; industry application

1 引言

1.1研究背景

数字化转型进程中,大数据已成为产业升级核心生产要素,全球数据量呈指数级增长,其中包含大量敏感数据。大数据4V特征重构了数据安全风险格局：PB级规模降低传统脱敏效率,流式数据要求亚秒级时延,80%以上非结构化数据使传统规则失效,数据噪声增加隐私与价值平衡难度。政策层面,《数据安全法》《个人信息保护法》及国际法规明确数据处理安全责任,软件数据脱敏技术作为“数据可用不可见”的关键手段,其研究应用对保障合规流通、释放数据价值具有重要现实意义。

1.2研究现状与意义

数据脱敏技术已历经三代演进：1.0时代以静态脱敏为主,存在链接攻击漏洞；2.0时代聚焦动态脱敏与匿名化,k-匿名等技术提升可用性但存在同质性缺陷;3.0时代呈现智能脱敏与隐私计算融合趋势。国内厂商已推出云原生适配脱敏平台,但多模态处理、跨境合规等方面仍有不足。本文系统研究脱敏技术核

文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

心原理与落地路径,可为企业研发与方案选型提供指导,助力破解“数据安全与价值释放”核心矛盾,兼具学术与实践价值。

2 大数据环境下数据脱敏技术理论基础

2.1核心概念界定

数据脱敏是通过技术手段去除或模糊个人标识信息、破坏“标识-记录”映射关系的泛化技术,包含匿名化、假名化、掩码等具体形式。核心区别在于：匿名化不可逆、隐私保护最强但价值损失最大;假名化通过虚拟标识替代真实信息,可逆且保留价值较好;掩码隐藏敏感字段部分内容,实现简单适用于结构化数据。具体差异如下表所示：

术语	定义	核心区别
数据脱敏	去除或模糊个人标识信息的泛化技术	上位概念,包含多种具体技术
匿名化	彻底去除标识,无法重新识别	不可逆,隐私保护最强
假名化	虚拟标识替代真实标识	可逆,数据价值保留较好
掩码	隐藏敏感字段部分内容	实现简单,适用于结构化数据

2.2核心理论框架

数据脱敏技术的理论基础源于隐私保护模型,核心包括k-匿名、l-多样性、t-接近性及差分隐私,共同构成“从等价类保护到概率性保护”的演进体系。

2.2.1 k-匿名模型：要求数据集中每条记录在准标识符(如年龄、性别、邮编)上,存在至少k-1条相似记录,数学表达式为$\forall r\in D,|\{r'\in D|QI(r')=QI(r)\}|\geq k$。通过构建等价类降低个体识别概率,但无法解决同质性问题——若某等价类敏感属性(如疾病)完全相同,仍可推断个人信息。

2.2.2 l-多样性模型：针对k-匿名缺陷,要求每个等价类的敏感属性至少包含l个不同值,表达式为$\forall E\in \Pi QI(D),|\{s(r)|r\in E\}|\geq l$。通过增加敏感属性多样性,提升隐私保护强度,但未考虑属性分布偏倚问题。

2.2.3 t-接近性模型：要求等价类敏感属性分布与数据集整体分布差异不超过t,常用Earth Mover’s Distance(EMD)衡量分布差异。该模型解决了分布偏倚问题,但计算复杂度较高,不适用于海量数据场景。

2.2.4差分隐私模型：通过向数据添加可控噪声,保证“个体是否存在于数据集不影响输出结果”,核心表达式为$Pr(M(D)=S)\leq e^{\varepsilon}\cdot Pr(M(D')=S)+\delta$,其中 ε 为隐私预算(越小隐私保护越强),δ为松弛参数。该模型无需依赖准标识符,适合大数据高并发场景,是当前主流的隐私保护理论基础。

3 大数据环境下核心软件数据脱敏技术解析

大数据环境下脱敏技术形成“静态+动态+智能”多元体系,各类技术在处理模式、性能与场景上互补适配4V特征。

3.1静态数据脱敏技术

静态脱敏是数据离线导出时的一次性脱敏处理,核心流程为“敏感数据识别→脱敏规则匹配→数据处理→结果存储”。常用技术包括掩码、泛化、替换、哈希等,适用于开发测试、数据分析等非实时场景。

掩码技术通过隐藏敏感字段部分内容实现脱敏,如将手机号“138XXXX1234”处理为“138****1234”,具有实现简单、性能损耗低的特点,但无法抵御链接攻击,仅适用于低安全等级场景。泛化技术通过降低数据精度实现脱敏,如将“25岁”泛化为“20-30岁”、“北京市海淀区”泛化为“北京市”,可保留数据统计特征,但过度泛化会降低数据可用性。

大数据场景下,静态脱敏需适配分布式架构,通过Spark、Flink等大数据工具实现并行处理。例如,某省级农信联社采用静态脱敏系统适配星环TDH大数据平台,可在数小时内完成数百万XML文件的批量脱敏,满足开发测试环境的数据需求。其局限性在于无法应对生产环境的实时数据访问需求,脱敏后数据存在“一次脱敏、终身使用”的安全风险。

3.2动态数据脱敏技术

动态脱敏是数据访问过程中的实时处理技术,核心特征为“数据不落地、按需脱敏、千人千面”,即根据访问主体的权限等级动态调整脱敏策略,确保敏感数据仅对授权用户可见。该技术适用于生产环境的实时业务场景,如在线客服、金融风控等。

动态脱敏的核心实现依赖智能SQL解析引擎,通过拦截数据库访问请求,实时识别敏感字段并执行脱敏策略。例如,腾讯云TCHouse-X基于PostgreSQL内核改造,支持复杂查询的实时字段级脱敏,在TPC-DS 10TB数据集测试中,处理速度达128万行/秒,时延控制在5ms以内,满足百万级QPS的高并发需求。其技术优势在于:一是实时性强,不影响业务系统响应速度;二是安全性高,避免脱敏数据落地存储的泄露风险;三是灵活性高,可根据用户角色、访问场景动态调整策略。

动态脱敏的关键挑战是权限管控与性能平衡。银保机构等高压场景下,需在毫秒级时延约束下完成权限校验与脱敏处理,某股份制银行通过绿盟动态脱敏网关,实现核心交易系统敏感字段毫秒级脱敏,违规访问拦截率达99.9%,性能损耗控制在3%以内。

3.3智能数据脱敏技术

智能脱敏是融合AI技术的新一代脱敏技术,核心能力包括敏感数据智能识别、多模态数据处理、合成数据生成,解决了传统技术对非结构化数据适配不足、脱敏策略依赖人工配置的问题,是2025年数据脱敏技术的主要发展方向。

敏感数据智能识别基于NLP、OCR等技术,可自动识别文本、图像中的敏感信息。例如,腾讯云TCHouse-X通过BERT模型识别文本敏感信息,实现“姓名→虚拟名”“地址→模糊区域”的语义级替换;集成OCR技术提取PDF/扫描件中的身份证、发票信息,识别准确率达99.2%。多模态数据处理针对图像、语音等非结构化数据,采用像素扰动、语义替换等技术,如医疗影像通过像素扰动隐藏患者面部信息,病历文本通过NLP替换敏感词,既保留诊断价值又避免隐私泄露。

合成数据生成基于GAN生成对抗网络,通过学习真实数据的特征分布,生成高保真的虚拟数据,替代真实数据用于AI训练、场景测试。某医院采用GAN生成的医疗数据,特征保留度达92%,既满足科研需求,又规避了患者隐私泄露风险。智能脱敏的局限性在于模型训练需要大量标注数据,对小样本场景适配不足,且存在生成数据与真实数据特征偏差的风险。

4 大数据环境下数据脱敏技术的行业应用实践

金融、医疗行业因数据敏感性高、监管严格,成为脱敏技术核心应用领域,以下结合典型案例分析落地路径与成效。

4.1金融行业应用：省级农信联社数据安全防护

某省级农信联社面临数据类型复杂、场景多样的脱敏需求,原有系统存在大数据平台适配不足、人工干预多、效率低等痛点。解决方案采用静态+动态融合脱敏系统,核心措施包括：完成与星环TDH平台兼容认证,实现分布式并行脱敏；内置多种脱敏算法与规则库,支持多格式批量处理;优化智能操作界面简化策略配置。实施后脱敏交付周期从3天缩短至1天,效率提升60%,保障数据有效性与关联性,完全符合合规要求,审计通过率100%。

4.2医疗行业应用：三甲医院患者隐私保护

某省级三甲医院需满足开发测试、科研交流等多场景脱敏需求,核心诉求是保障数据一致性与业务关联性,实现标准化处理。解决方案采用静动态结合方案：开发测试环境部署静态脱敏系统,自动扫描处理敏感数据,避免中间存储风险；科研交流场景在数据导出时触发动态脱敏,对病历文本、医疗影像执行差异化策略。实施后实现患者隐私全流程保护,隐私字段识别准确率98.7%,科研数据共享效率提升60%,人工成本降低70%,满足医疗数据特殊保护要求。

5 大数据环境下数据脱敏技术现存问题与发展趋势

5.1现存主要问题

5.1.1实时性与高并发适配不足:动态脱敏在PB级流式数据场景下,仍存在时延过高的问题,部分方案在百万级QPS压力下,时延会突破10ms的业务阈值,影响用户体验。

5.1.2多模态数据处理能力薄弱：传统技术对文本、图像、语音等混合数据的脱敏效果不佳,非结构化数据的敏感信息识别准确率平均仅为85%左右,难以满足医疗、工业等领域的复杂需求。

5.1.3脱敏与可用性平衡难度大:过度脱敏会导致数据失去统计与分析价值,而脱敏不足则存在隐私泄露风险,当前缺乏自适应的脱敏强度调整机制,依赖人工经验配置策略。

5.1.4跨境合规适配能力不足:不同国家和地区的法规对脱敏标准要求差异较大,现有系统大多缺乏多法规模板,跨境企业需投入大量成本进行定制化改造。

5.2未来发展趋势

5.2.1 AI全链路赋能：通过深度学习实现敏感数据自动识别、脱敏策略智能生成与动态优化,降低人工依赖。例如,基于强化学习算法,根据数据类型、应用场景自动调整脱敏强度,实现“安全-可用”的最优平衡。

5.2.2隐私计算深度融合：将差分隐私、同态加密、联邦学习与脱敏技术结合,实现“数据可用不可见”的进阶目标。例如,在金融风控场景中,采用差分隐私+动态脱敏的融合方案,既保证交易数据的实时分析需求,又避免敏感信息泄露。

5.2.3云原生与全链路协同:脱敏技术将深度集成于云原生架构,实现与云数据库、数据仓库的无缝对接；同时构建“识别-脱敏-监控-溯源”的全链路安全体系,结合区块链技术实现脱敏日志不可篡改存证,满足等保2.0三级审计要求。

5.2.4行业化定制与合规一体化：针对金融、医疗、政务等不同行业的特点,开发定制化脱敏算法与策略模板;内置多国家法规模板,支持跨境数据流动的合规校验,降低企业合规成本。

6 结论

大数据环境的4V特征对软件数据脱敏技术提出了实时性、多模态、高并发的严苛要求。本文研究表明,静态脱敏、动态脱敏与智能脱敏技术的协同应用,可有效适配不同业务场景的需求；k-匿名、差分隐私等理论为脱敏技术提供了坚实的理论支撑；金融、医疗行业的实践案例验证了脱敏技术在保障数据安全、推动合规流通中的核心价值。

当前数据脱敏技术仍面临实时性瓶颈、多模态处理不足等问题,未来需通过AI赋能、隐私计算融合、云原生集成等方向突破创新。后续研究可聚焦于自适应脱敏强度调整算法、多模态数据统一脱敏框架等关键技术,进一步提升脱敏技术的智能化水平与行业适配能力,为数据要素市场化配置提供更坚实的安全保障。

[参考文献]

[1]张宏科.大数据环境下的数据脱敏技术深度剖析[J].计算机工程与应用,2026,62(3):124-132.

[2]腾讯云开发者社区.2025年数据安全脱敏平台TOP评测:腾讯云TCHouse-X如何领跑行业[EB/OL].https://cloud.tencent.cn/developer/article/2564903,2025-09-04.

[3]李艳红.数据脱敏技术实战:两大行业应用案例解析[J].数据安全与通信保密,2025,(8):78-85.

[4]王建军.动态脱敏技术的难点以及应对[J].金融电子化,2025,(7):67-69.

[5]中国电子技术标准化研究院.数据脱敏技术新趋势与国内优秀数据脱敏产品推荐(2025版)[EB/OL].https://m.sohu.com/a/901371296_122406616,2025-06-04.

作者简介：

万耀雪(1985--),男,汉族,江西人,本科,职称：初级,研究方向：电子信息产业。