 DOI: https://doi.org/10.12349/iser.v6i12.8469

Research on Data Desensitization Methods in Big Data Applications

Zhenzhen Si Min Wang

China Electronics Technology Group Corporation 22nd Research Institute, Xinxiang, Henan, 453000, China

Abstract

With the popularization of big data technology and increasingly strict data privacy regulations, data anonymization has become an indispensable key technology in big data applications. This article provides a systematic overview of data anonymization technology and proposes a layered technology architecture that dynamically adapts to multiple scenarios, covering business processes, anonymization algorithms, rule systems, and application design. Experimental verification shows that this architecture balances practicality and security, providing important references for privacy protection scheme design.

Keywords

Big data security; Privacy protection; Data desensitization; Dynamic desensitization; Anonymization

大数据应用中的数据脱敏方法研究

司祯祯 王珉

中国电子科技集团公司第二十二研究所，中国 · 河南 新乡 453000

摘 要

随着大数据技术普及与数据隐私法规日益严格，数据脱敏已成为大数据应用中不可或缺的关键技术。本文系统综述数据脱敏技术，提出多场景动态适配的分层技术架构，涵盖业务流程、脱敏算法、规则体系与应用方式设计，经实验验证表明该架构兼顾了实用性与安全性，为隐私保护方案设计提供重要参考。

关键词

大数据安全；隐私保护；数据脱敏；动态脱敏；数据匿名化

1 引言

在大数据技术迅猛发展的时代背景下，数据资源已成为驱动社会进步与经济转型的核心要素。数据脱敏技术作为平衡数据利用与隐私保护的核心手段，通过系统化处理使敏感信息在保留必要特征的同时消除可识别性，已成为保障数据安全应用的核心技术之一 [1]。

国内外当前研究已初步构建了基于变形、替换、加密及合成等多样化脱敏技术路径，但面对大数据环境的高并发、多模态特征，以及数据规模从 GB 级向 TB、PB 级跃升，传统脱敏方法存在明显不足 [2]。

针对上述挑战，本文提出了一种支持多场景动态适配的数据脱敏分层技术架构，突破了大规模数据高效处理、敏感信息智能识别等关键技术，成为兼顾安全性和实用性的解决方案。

【作者简介】司祯祯（1982–），女，中国河南新乡人，硕士，高级工程师，从事计算机软件信息化研究。

2 数据脱敏技术概述

2.1 数据脱敏概念

敏感数据即隐私信息，其泄露可引发多重风险与损失。常见类型如姓名、电话号码、身份证号码、银行账号、邮箱地址、密码凭证、组织机构名称、营业执照号码等数据。

数据脱敏通过系统性技术操作降低敏感数据的敏感性与可辨识度，在满足业务需求的同时有效规避隐私泄露风险。数据脱敏通常涉及数据变形、替换、加密、屏蔽等多种技术手段的综合运用，使脱敏后的数据保留原有特征与统计价值，但失去直接关联到具体个体或实体的能力 [2]。

数据脱敏技术的实施须遵循最小化、可逆性、安全性原则以确保其合法性和有效性 [2]。数据脱敏实践可结合具体应用场景与数据类型进行差异化设计。针对结构化数据，可通过替换、泛化、随机化等方法实现字段级脱敏；对于非结构化数据，可借助自然语言处理技术识别并替换文本中的敏感信息 [3]。

2.2 数据脱敏技术分类

根据应用场景与需求差异，数据脱敏技术可分为静态脱敏与动态脱敏两类。

（1）静态脱敏通过预处理手段对原始数据进行操作，生成独立的脱敏数据副本，适用于存储 / 传输场景，核心采用敏感信息替换、字段乱序排列、加密处理等技术手段 [4]，处理效率高、结果稳定性强，但灵活性差。

（2）动态脱敏技术聚焦于数据使用时的实时脱敏，核心是根据用户权限、访问场景和业务需求，动态生成脱敏结果。该技术通过构建多维度的脱敏规则库与权限控制模型，实现对敏感数据的按需访问控制。其优势是灵活性高，支持复杂权限管理和差异化策略，但技术实现较复杂，需集成实时解析、规则匹配和性能优化模块 [4]。

这两类技术适用场景不同，实际应用时，需结合数据类型、安全需求、使用频率和系统架构，选择或组合使用两种技术，以达到最佳隐私保护效果。

3 大数据应用中数据脱敏架构设计

根据大数据应用中数据敏感等级和业务场景需求自动调整脱敏强度，可采用动态识别应用场景、集中化规则管理、自动配置脱敏策略、智能选择脱敏算法、适应批流一体处理的分层技术架构，不仅提高了脱敏数据的自动化程度，而且提高了脱敏效果。下面围绕脱敏流程、脱敏算法、脱敏规则和脱敏方式等分别进行架构设计。

3.1 脱敏流程

数据脱敏的业务流程分为敏感数据识别、敏感数据整理、脱敏方案制定、脱敏任务执行四个步骤 [2]，在各个步骤中结合数据脱敏算法、数据脱敏规则来实现最佳数据脱敏效果。

（1）敏感数据识别：敏感数据发现一般采用自动识别为主，结合人工发现和审核，来完成敏感数据的发现和定义，最终形成完善的敏感数据字典集。

（2）敏感数据整理：在识别敏感数据后，进行敏感数据项和敏感数据关系的整理，确定数据项之间的关联关系，为脱敏算法选择提供支撑依据。

（3）脱敏方案制定：根据不同的数据类型的脱敏需求，分别配置脱敏方案，并为脱敏操作配置适宜的脱敏算法。

（4）脱敏任务执行：根据脱敏方案对脱敏数据进行脱敏操作，包括脱敏任务的停止、启动、暂停等操作，在大数据应用中支持任务并行处理和支持脱敏任务的中断续延等。

3.2 脱敏算法

通常根据不同敏感数据类型和保护特征选择不同的脱敏算法，脱敏算法包括替换、扰乱、泛化、生成等四类。

3.2.1 替换类算法

随机值替换：用随机生成的数值或字符串来替换原始数据中的敏感信息。可以确保数据的唯一性和不可恢复性，但可能会影响数据分析的准确性。

固定值替换：用一个固定的数值（如“未知”、“*”、“#”等）来替换原始数据中的敏感信息，方法简单易行，但可能降低数据的多样性。

哈希替换：使用哈希函数将敏感信息转换为哈希值进行存储。保证数据的唯一性，但无法从哈希值中恢复出原始数据。

加密替换：通过加密算法将敏感数据加密后存储。这种方法可以提供较高的安全性，但需要额外的解密步骤才能使用数据。

3.2.2 扰乱类算法

数值扰乱：在原始数据的基础上添加噪声（如随机数、偏移量等），以改变数据的真实值，可以在一定程度上保护数据隐私，但可能会引入误差并影响数据分析结果。

时间扰乱：对日期和时间信息进行修改，可以隐藏数据的实际发生时间，但需要注意保持数据的逻辑一致性和合理性。

空间扰乱：对地理位置信息进行模糊化处理，如将精确位置替换为一定范围内的区域，可以保护用户的地理隐私，但同样需要注意保持数据的逻辑一致性和合理性。

3.2.3 泛化类算法

数值泛化：将具体数值替换为其所属的区间范围或类别标签。例如，将年龄从具体的年份替换为年龄段（如 30-39 岁），减少数据的粒度，从而保护个人隐私。

文本泛化：将具体的文本信息替换为其所属的类别或关键词。例如，将姓名替换为姓氏加占位符（如张 XX），可以保留部分信息以供分析使用，同时减少泄露个人身份信息的风险。

3.2.4 生成类算法

合成数据生成：基于原始数据的统计特征和分布规律，生成与原始数据相似但不包含敏感信息的合成数据，可以提供丰富的数据集用于分析和训练模型，但需要确保合成数据与原始数据在关键特征上保持一致。

差分隐私技术：通过在查询结果中添加噪声来保护用户隐私的技术。差分隐私技术可以确保即使单个用户的记录被添加到数据集中也不会显著影响查询结果的准确性，尤其适用于大数据环境下的隐私保护需求。

3.3 脱敏规则

在大数据应用中数据脱敏规则可以进行配置管理，脱敏规则包括可恢复与不可恢复两类 [2]。

（1）可恢复类：指脱敏后的数据可以通过一定的方式，可以恢复成原来的敏感数据，此类脱敏规则主要指各类加解密算法规则。

（2）不可恢复类：指脱敏后的数据使用任何方式都不能被恢复，一般可分为替换算法和生成算法两大类。替换算法即将需要脱敏的部分使用定义好的字符或字符串替换，生成类算法要求脱敏后的数据符合逻辑规则，即是“看起来很真实的假数据”。

3.4 脱敏方式

通过对大数据平台中数据脱敏处理，常见有流式数据脱敏方式和批量数据脱敏方式。

（1）流式数据脱密

流式数据是指连续产生、动态增加且有更新时效要求的数据。对流式数据的脱敏处理技术的优势是从数据传递的

同时进行了数据处理，劣势是无法利用全量数据进行复杂关联处理。

（2）批量数据脱敏

批量数据是指通过数据扫描稳定数据源的方式批量录入到大数据平台，数据经常以历史数据为主。批量数据脱敏可以在数据导入的过程中进行批量脱敏，或在大数据平台应用过程中进行批量脱敏 [2]。

4 实验与分析

4.1 实验条件设计

围绕实验环境准备、实验数据集、脱敏效果评估指标三方面分别进行实验条件设计。

（1）实验环境准备：本研究实验环境基于服务器硬件环境进行搭建，硬件配置采用 Intel Xeon E5-2686 v4 处理器（2.3GHz 主频，64 核）、256GB DDR4 内存及 4TB SSD 存储单元，采用 Windows 操作系统，核心实验框架基于 Python 3.8 开发环境，集成 Pandas、NumPy 等数据处理库。数据存储采用 MySQL 关系型数据库和 MongoDB 非关系型数据库，兼顾结构化数据查询与非结构化数据扩展的实验需求。

（2）实验数据集：采用 Apache Faker 工具生成符合隐私保护场景的结构化数据集，包括姓名、身份证号、电话号码等敏感字段的用户信息表共 1000 万条，通过正态分布参数控制数据特征分布 [5]，确保数据逼真性，最终形成包含数值型、类别型、文本型数据的多模态实验数据集。

（3）效果评估指标：脱敏效果评估采用信息损失率、分类准确率下降幅度及隐私保护强度三个子指标，其中信息损失率通过原始数据与脱敏数据的字段相似度计算得出，分类准确率下降幅度通过对比脱敏前后机器学习模型的预测结果量化数据可用性损失，隐私保护强度则基于 K- 匿名、l- 多样性等标准进行量化评估。

4.2 实验实施过程

本研究实验实施过程可分为数据预处理、脱敏处理及效果评估的三阶段分别设计如下。

（1）数据预处理阶段：实验首先对原始数据集进行多维度清洗操作，包括缺失值处理、异常值检测与修正以及重复记录的过滤。针对缺失值问题，采用基于特征相关性的插值算法和统计学均值填补策略，结合领域知识判断缺失机制后选择最优填充方案。对于异常值，通过箱线图法与孤立森林算法识别离群点，并根据数据特性采取截断修正或数据重构等处理手段。

（2）脱敏处理阶段：采用分层策略实现隐私保护与数据可用性的平衡。针对结构化数据，基于敏感信息分类建立多级脱敏规则库，采用分级加密、字符替换及泛化等基础脱敏技术处理个人身份信息字段。对于包含复杂关联关系的非结构化数据，引入差分隐私机制与数据扰动技术，通过拉普拉斯噪声注入控制隐私预算，结合局部敏感哈希（LSH）算法维护数据分布特征。针对批量数据处理场景，设计基于并行化脱敏架构，利用分布式脱敏任务调度，通过动态负载均衡策略优化计算资源利用率。

（3）效果评估阶段：采用定量指标与定性分析相结合的方法。通过计算脱敏前后数据集的统计特征相似度及分类模型预测准确率变化率，量化脱敏对数据可用性的影响。隐私保护效果通过 k- 匿名、l- 多样性及 t- 接近度等指标衡量，基于关联规则挖掘的攻击模型测试隐私泄露风险。在性能评估方面，记录不同脱敏算法的计算时延、内存占用及 I/O 吞吐量进行评估分析。

4.3 实验结果分析

对实验数据通过单因素方差分析对比不同方法的性能差异，对数据脱敏方法的脱敏效果与性能进行了多维度验证，包括信息损失率（ILR）、可逆性指标（RMI）及业务规则符合度（BRC）等核心指标对不同算法进行了量化评估。

通过实验对比验证了所有方法在保证隐私安全的前提下，采用本分层技术架构，在医疗数据场景中成功将个人身份信息的再识别风险降低至 0.3% 以下，较传统方法平均提升 20% 以上的信息保留率，且处理速度在百 GB 级数据集上达到分钟级响应 [5][6]，满足大数据平台的实时脱敏要求，在隐私保护与数据可用性之间实现了有效平衡。

5 结论与展望

数据脱敏是大数据安全和治理的基石，本研究探讨了大数据应用中数据脱敏相关技术，创新性提出了一种面向高维异构数据动态场景适配的脱敏分层技术架构。通过实验分析验证了该分层架构兼顾了实用性和安全性，提出了多维度的脱敏效果评估体系，支撑对脱敏方法的系统化评价，有效识别不同脱敏方法的优劣势，为技术选型与参数优化提供了支撑依据。

当前数据脱敏技术的研究正朝着智能化与自动化方向发展，通过机器学习算法识别敏感数据模式、自动推荐脱敏策略已成为热点趋势。随着隐私计算技术的发展，联邦学习、多方安全计算等新兴技术正在与数据脱敏技术深度融合 [2][6]，为数据共享场景提供了兼顾隐私保护与协作计算能力的创新解决方案。这些技术进步不仅拓展了数据脱敏的应用边界，也为构建更安全高效的数据安全生态系统奠定了基础。

参考文献

[1] 徐双,刘文斌,李佳龙等.大数据背景下的数据安全治理研究进展[J].太原理工大学学报.2024,55(01):127-141

[2] 佟玲玲,李鹏霄,段东圣等.面向异构大数据环境的数据脱敏模型[J].北京航空航天大学学报.2022,48(02):249-257

[3] 彭婧,尹立夫,王洲等.电力数据脱敏安全防护体系[J]. 计算机应用 2022,42(S1):191-194

[4] 周在彪等.云计算环境下的数据隐私保护策略[J].软件学报 2025,7(7):92-94

[5] 金红军.医疗数据脱敏技术在医院信息共享与安全中的应用研究[J].信息系统工程.2025,10(10):45-48

[6] 张志立,杨红,庞娟等.大语言模型在医疗数据脱敏中的实践与表现[J].北京大学学报.2025-10-11网络首发