文章编号:1671-4598(2024)11-0315-07 DOI:10.16526/j.cnki.11-4762/tp.2024.11.044 中图分类号:TP399 文献标识码:A

基于国密 SM4 及保形加密算法的文件脱敏系统研究

黄 俊[1],刘家甫[1],曹志威[2]
(1.公安部第三研究所 数据安全技术研发中心,上海 201204;
2.海军军医大学 影像医学系,上海 200433)

摘要:针对政务及金融等领域对于内部文件保密要求高,移动介质上存储的文件数据通过传统脱敏方法面临着数据内容量大、数据类型多样导致的脱敏效率低、脱敏内容不彻底等问题,提出了一种基于 SM4 与 FF1 结合的混合数据类型文件脱敏系统,该系统通过内容分割脱敏处理任意类型的数据,提升了文件脱敏的范围、准确性和效率;为了进一步减少脱敏系统代码运行的内存消耗,提出了汉字字典库索引转换算法,该算法通过构建待检测明文与汉字编码库的相对索引关系,优化传统脱敏系统中依赖于构建哈希表的键值映射;通过随机生成 1 000 份测试文件进行脱敏测试,基于混合类型的文本不可识别率达到 99.8%,脱敏以及内容复原的准确率达到 99.9%;通过随机生成 10 份总大小约为 10 MB 的测试文件,纯文本类型的脱敏速率平均可达 2 500 字符/秒。

关键词:国密 SM4;保形加密;数据脱敏;国密算法;文件脱敏系统

Research on File Desensitization System Based on the National Security SM4 and Format Preserving Encryption Algorithm

HUANG Jun[1], LIU Jiafu[1], CAO Zhiwei[2]
(1. Data Security Technology Research and Development Center, The Third Institute of The Ministry of Public Security, Shanghai 201204, China;
2. Faculty of Medical Imaging, Naval Medical University, Shanghai 200433, China)

Abstract: Due to the high confidentiality requirements for internal documents in the fields of government and finance, traditional desensitization methods for data files stored on mobile media face problems such as low efficiency and incomplete desensitization content caused by large data volumes and diverse data types, a hybrid data type file desensitization system based on the combination of the SM4 and FF1 is proposed. This system processes any type of data through content segmentation desensitization, and improves the range, accuracy, and efficiency of file desensitization. In order to further reduce the memory consumption during the execution of desensitized system code, a method for converting Chinese character library indexes is presented. This algorithm constructs the relative index relationship between the plaint text to be detected and the Chinese character encoding library, and optimizes the key value mapping that relies on building a hash table in traditional desensitization systems. By randomly generating 1 000 test files for desensitization test, the text unrecognition rate based on mixed types reaches 99.8%, and the accuracy of desensitization and content recovery reaches 99.9%. The average desensitization rate of pure text type can reach 2 500 characters/second for 10 test files with a total size of about 10 MB.

Keywords: national secret SM4; format preserving encryption; data desensitization; national secret algorithm; file desensitization system

收稿日期:2024-09-29; **修回日期**:2024-10-29。
基金项目:科技部重点研发计划资助(2021YFB3102002)。
作者简介:黄 俊(1981-),男,硕士,副研究员。
通讯作者:刘家甫(1998-),男,硕士。
曹志威(1985-),男,博士,讲师。
引用格式:黄 俊,刘家甫,曹志威.基于国密 SM4 及保形加密算法的文件脱敏系统研究[J].计算机测量与控制,2024,32(11):315-321.

0 引言

在以文件为载体的信息传输过程中，文件中的敏感信息面临着被窃取和泄露的风险，我国的《网络安全法》、《数据安全法》以及欧盟《通用数据保护条例》（GDPR）等都对数据保护提出了明确要求，机关与企业必须采取措施保护个人敏感信息，数据脱敏技术是满足这些法规要求的重要手段之一。

数据脱敏技术可应用在文件系统中，包括函件、内部资料传递、邮件、商务合作等。通过脱敏处理后的文件在传输、存储和使用过程中更加安全，即使文件被非法获取，由于敏感信息已经被处理，攻击者也难以从中获取有价值的信息，从而降低了数据被恶意利用的风险。比如，机关内部秘密发函的文件信息被不法分子盗取后，由于脱敏后的文件完全失去可识别性，不法分子也无法从文件中得到有价值的信息。同时，对文件进行脱敏处理可以使机关单位和企业更加规范地管理数据。通过统一的脱敏规则和流程，可以确保敏感信息得到妥善处理，提高数据管理的质量和安全。以某行业内部信息网使用办法要求为例，涉及连接信息网固定设备的文件传输，严禁连接 U 盘，移动硬盘等移动设备。若如确有需求连接具有存储功能的外界设备，需在申请后通过光盘与光驱进行连接。即使该办法严格控制了设备的接入，一旦信息网的电子文件进入光盘，就仍存在着人工泄露的风险。文件脱敏系统可以在文件传输之前进行文件内容的脱敏，结合保密使用办法的合规流程，为重要文件信息的传输提供安全的“双保险”。

当前，数据脱敏处理主要集中于数字、字母等基于身份识别码的脱敏处理，采用的处理方式有不可逆的替换、截断、隐藏等。可逆替换主要通用对称加密算法针对单一类型的数据通过加密算法进行加解密[1-2]，基于区块链的数据加密研究[3-4]可以实现数据传输的可追溯以及通过类似访问控制的形式实现数据的安全。此外，还有基于保形加密的数据脱敏研究[5-7]。在先前的研究中，基于 SM4 的 FPE 算法主要应用于数据库数据[8-9]同时处理的数据类型格式比较集中于数字和字母。如上述，该研究中针对数据脱敏处理的范围仅限于数据库中特定数据格式的处理，针对文件中混合类型数据格式仍然需要进一步的研究。本研究所设计的文件脱敏系统针对混合类型数据设计汉字字典库索引转换的协议，进而实现了 FPE 算法对于混合数据格式的文件类型进行脱敏处理。

SM4 算法是我国发布的商用密码算法中无线局域网推荐使用的的分组密码算法，也称为国密 SM4 算法。国家鼓励商用密码技术的研究开发、学术交流、成果转化和推广应用，鼓励和促进商用密码产业发展[10]。本研究所应用文件脱敏系统主要应用于机关内部的文件传输，结合 SM4，可进一步实现文件脱敏系统整体知识产权的自主可控。

1 文件脱敏系统基础算法原理

1.1 数据脱敏技术

数据脱敏（Data Desensitization），也称为数据匿名化（Data Anonymization）技术，其核心目标是在保留数据结构的前提下，对关键信息进行处理，使之难以还原，从而降低敏感信息泄露的风险。

脱敏算法类型概述如下。

1）敏感信息替换：

替换是指通过指定规则或者映射进行内容转换，从而破坏其可读性。敏感信息替换处理的数据一般具有不可逆性。替换算法是最为常用的脱敏算法之一，但该算法会导致脱敏后的数据失去价值，不利于数据的后续使用[11]。

2）敏感信息加密：

加密是指通过使用诸如 MD5[12]、Hash[13]、AES 等密码学算法对敏感数据进行加密，进而实现数据脱敏的技术，加密处理后的数据与敏感数据的原始内容在逻辑规则和格式上保持一致，外部未经授权的用户只能访问到无实际意义的密文数据，在特定需求场景下，系统也可以给相关需求方提供解密能力以恢复敏感数据的原始内容。

3）敏感信息混淆：

混淆是指通过对敏感数据内容在指定条件下打乱重排和重新分布，破坏与其他字段数据的关联关系，使混淆后的数据不再具有原始内容的语义，从而实现数据脱敏。

1.2 SM4 分组密码算法

2006 年，SM4 算法为配合 WAPI 无线局域网标准的推广应用公开发布。2012 年 3 月，SM4 算法成为我国密码行业标准。2016 年 8 月，SM4 算法转化为国家标准 GB/T 32907－2016《信息安全技术 SM4 分组密码算法》。

2021 年 6 月 25 日，SM4 分组密码算法作为国际标准[14]由国际标准化组织 ISO/IEC 正式发布。

1.2.1 SM4 算法概述

SM4 算法的分组长度与密钥长度均为 128 比特。SM4 加密算法与密钥拓展算法均采用 32 轮非线性迭代结构，数据加解密过程的算法结构相同，加解密过程的差异取决于轮密钥的使用顺序，其解密过程所采用的轮密钥是加密过程轮密钥的逆序。作为国产密码算法，SM4 算法具有自主知识产权，不受国外密码技术的限制，能够满足国家信息安全的战略需求。

1.2.2 SM4 加解密流程

SM4 加密算法由 32 次非线性迭代运算和一次反序变换 R 组成，SM4 密钥体系包含 128 比特的加密密钥与轮密钥。其中，加密密钥表示为 MK －（MK_0，MK_1，MK_2，MK_3），其中 $i\in$［0，3］。轮密钥表示为（rk_0，rk_1，…，rk_{31}），其中 $i\in$［0，31］，每一位为 32 比特字。轮密钥由加密密钥基于密钥扩展算法生成。轮函数由 128 比特，4 个字组成的输入结合轮密钥和一个可逆的合成置换 T 构成，该函数 T 包含非线性函数 S 盒和线性变换函数。

加密过程中，首先将输入的明文数据 X 分成 4 个 32 比特的字（X_0，X_1，X_2，X_3），然后，使用密钥扩展算法生成 32 个轮密钥 rki 进行轮加密，每一轮加密均涉及非线性

变换、线性变换和轮密钥加等操作。经过 32 轮迭代后，再通过一次反序变换，得到最终的密文输出。解密过程中在相同的密码算法结构基础上，采用轮密钥的逆序（rk_{31}，rk_{30}，…，rk_0）进行数据解密。

1.3　保形加密算法

保形加密（FPE，format preserving encryption）算法主要用于对特定格式的数据进行加密，同时保持加密后数据的格式与原始数据相同。保形加密在大数据[8]等领域拥有广泛的应用场景，在数据库加密方面，可在不影响数据库结构和查询语句的情况下，对身份证号、电话号码等敏感数据加密。在金融领域中[15]，保形加密能对交易数据、账户号码等加密，确保加密后格式与原始数据一致，便于业务处理和数据传输，在个人信息保护方面，医疗记录、社保号码等敏感信息可通过保形加密在保护隐私同时，确保不同系统交互共享时格式一致。

FPE 算法概述如下。

FPE 可应用于非二进制的数据，对于任何有限的符号集，例如十进制数字，FPE 可以通过映射转换格式化为符号序列的数据，即数据的加密形式与原始数据具有相同的格式。其主流实现算法包括 FF1 和 FF3，均基于 Feistel 结构的加密模式。FF1 保形加密算法结构相对独立，没有与其他组件模块之间的嵌套关系。FF3 本质上等同于 BPS[16]的 BPS－BC 组件，是通过与特定的 128 位分组密码进行实例化来实现的。FF1 和 FF3 提供了不同的性能优势。FF1 支持受保护的格式化数据有更大范围的长度，并且在调整值（tweak）的长度方面具有灵活性，这有利于增加数据字典的量级。FF3 通过减少轮数实现了更高的数据吞吐量。

2　文件脱敏系统架构概述及系统应用流程

2.1　文件脱敏系统架构

对于包含敏感信息的文件，文件包含字母、数字、特殊符号等混合类型的数据，且文件内容在格式，大小上具有随机性。因此，通过常规数据脱敏技术时会由于敏感信息识别能力不足导致漏脱敏的情况，基于加密算法的脱敏方式往往会出现加解密信息不一致等问题。本研究所涉及的文件脱敏系统通过设计基于 unicode 汉字库的全量索引转换协议，同时通过使用基于 FPE 的 FF1 算法实现保留格式的加密，以减少敏感信息漏判的情况，其中 FF1 算法中的伪随机函数和 Feistel 循环迭代加密过程的加密函数均由 SM4 算法实现，实现了文件脱敏系统的国产自主化和算法层面的优化。该系统的系统架构图如图 1 所示。

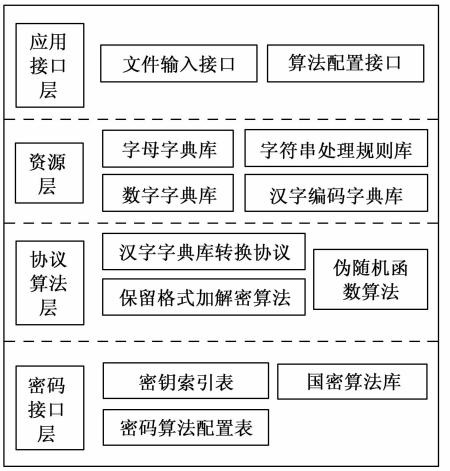

图 1　文件脱敏系统架构图

由图 1 可知，系统自上而下分为应用接口层，资源层，协议算法层，密码接口层。

1）应用接口层：

提供了文件输入接口和算法配置接口，主要用于配置密码接口层的密钥索引信息，对应于密码接口层的密钥索引表；配置 SM4 算法模式包含通过 socket 调用密码接口层 SM4 算法的 IP 和端口，分组密码的类型，包含 ECB（Electronic Codebook，电子密码本模式）和 CBC（Cipher Block Chaining，密码分组链接模式），以及对应的 Padding 模式；配置 FPE 算法信息，包含加解密的输入参数 tweak，该参数可以作为除了提供机密性的固定格式数据之外额外的微调参数，可被视为密钥的可变部分，该参数不需要保证机密性，因此，可以在算法配置层采用默认参数；增加对应的 alphabet 字典库，字典库作为有限集合提供字典内容与索引之间的转换，字典库的容量参数用 radix 表示。应用接口层的配置明确了特定场景下，文件脱敏系统运行的基础信息，应以实际需求进行操作和调整。

2）算法调用资源层：

资源层中的字母字典库、数字字典库，汉字编码字典库为系统处理不同类型的字符提供了索引转换的数据表，通常基于哈希表实现。字符串处理规则库针对不同的数据类型场景提供纯数字、纯字母、混合模式等字符串内容段分割规则库。字符串处理规则库进一步规范了系统对字符串的处理，确保脱敏过程的准确性和一致性。

3）协议算法层：

协议算法层的转换协议为资源调用层提供优化，其中汉字字典库索引转换算法通过算法支撑资源层的汉字字典编码库，通过算法优化，替换哈希表以减少了系统运行的内存消耗，保留格式加解密算法基于 FF1 实现，通过索引映射提供脱敏的加解密算法。伪随机函数基于 SM4 算法实现。

4）密码接口层：

密码接口层通过访问硬件加密机提供 socket 远程加密功能，提供密钥索引表和密码算法配置表，提供稳定安全加密环境。

2.2　文件脱敏系统应用流程

如图 2 所示文件脱敏系统的操作流程。首先，在应用接口层，通过文件输入接口接入待脱敏文件，算法配置用于设置相应的文件脱敏系统运行参数。其次，进行数据内容提取操作，将文件中的数据内容提取。通过调用资源层

的字符串处理规则库针对数据内容进行分割，对于不同内容段的处理包括字母字典、数字字典和汉字字典索引转换算法进行索引转换，通过保留格式加解密算法 FF1 以及基于 SM4 实现的伪随机函数算法等对分割段的数据内容进行遍历脱敏。

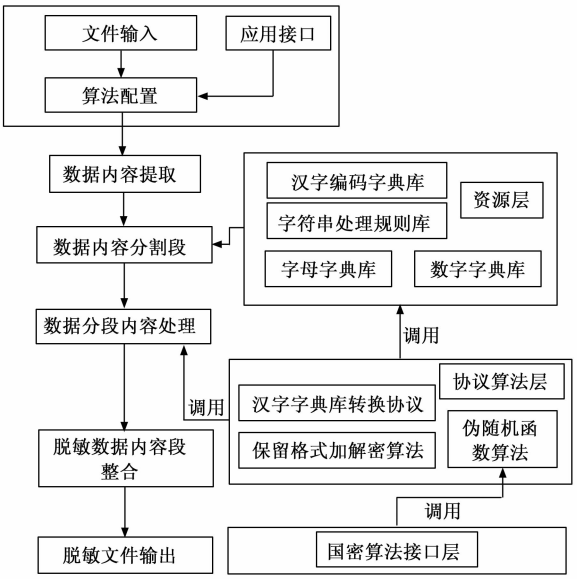

图 2 文件脱敏系统流程

经过对各分段内容的处理后，进行脱敏数据内容段整合操作，将处理后的分段内容重新组合成完整的脱敏后数据。总体而言，该流程通过多个层次的协同工作，实现了对文件中不同类型数据的全面脱敏处理，为保护敏感信息提供了可靠的方法。

在实际应用中，该文件脱敏系统可广泛应用于金融、医疗、政务、电信、电商等多个领域。对于涉及个人隐私信息、商业机密或敏感数据的文件，系统能够有效地进行脱敏处理，降低信息泄露的风险，保护相关方的合法权益。同时，系统的严谨设计和规范操作确保了脱敏过程的可靠性和高效性，为各行业的信息安全保障提供了有力的支持。

3 脱敏系统关键协议设计

3.1 汉字索引转换协议

文件脱敏系统中的字典库存储了不同类型的明文数据有限集，在进行数据加密之前需要针对字典元素转换为对应的序列 index，这为识别元素类型并提供元素（数字、英文、汉字等）与对应索引之间的关系。传统的索引转换协议是通过哈希表来维护字典库与索引的关系，会调用链表、红黑树等数据结构来解决哈希冲突等问题，这种存储键值的方式本身增加了内存消耗的成本，汉字 unicode 编码所容纳的基本汉字数量是 20 902。本研究中，字典库的容量需要考虑字符串处理规则的限定，若待脱敏的字符通过哈希表无法获取对应元素在字典库中的序号，则会出现脱敏失败。因此，对明文字符串处理规则库规则限定域需要大于实际字典库本身的有限集。

如图 3 所示，常规字典库的转换算法基于哈希表实现了字典元素与对应序列的互相转换，本研究创新性的通过相对距离实现对于汉字的索引转换，避免设立一个容量为 20 902 的汉字哈希索引表，优化了系统运行的效率和内存管理。

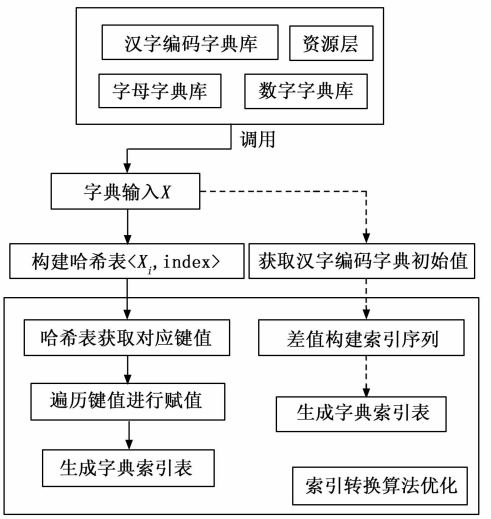

图 3 索引转换算法优化

算法 1：汉字脱敏字典库转换算法 STRTrans（text）

输入：待转换的 String 输入 text

输出：输出 text 的 int[]类型的 index 序列

1：创建 temp = new char[]，转换 text 类型为 char[]。

2：创建 result = new int[] 作为输出。

3：设置汉字表初始索引 initialChinese = Unicode 汉字编码首位索引。

4：遍历 temp 的所有字符：

4.1 计算 temp 的整数转换值与初始索引的差值；

4.2 差值同步到 result 对应索引位中，将 result 数组作为汉字索引表。

5：结束循环。

6：返回 result。

3.2 脱敏算法协议设计

3.2.1 算法函数与参数说明

Tweak：微调值，算法输入缩写为 T，该参数在算法中可选，可以是空字符串。Tweak 存在时可能会对加密和解密过程产生一定的影响，如果 Tweak 不为空，该值会作为一个额外的参数参与到加密或解密的计算中，从而改变加密结果或影响解密的准确性。Tweak 可能会影响加密过程中的数据变换方式，或者在解密时用于确定正确的解密路径。

Radix：字典库容量，表示字典有限集合的元素数量。

Minlen：处理字符串最小长度

Maxlen：处理字符串最大长度，不超过 2^{32}

脱敏函数需要确定实现所支持的最大调整值长度 Maxlen 的前提条件。在加密和解密算法之前会指定对前提条件

值的要求。参数的选择可能会影响互操作性。当给出错误输入时，会影响到脱敏的准确性。

NUM（X）：X 所代表的位字符串按照位权重降序时所代表的字符串大小

NUM_{radix}（X）：X 所代表的位字符串按照以 radix 为基数的位权重降序时所代表的字符串大小。

STR^{m}_{radix}：给定一个小于 radix^m 的非负整数 X，生成数值 X 以 radix 为基数的位权重降序表示。

3.2.2　算法具体步骤

该节所描述的脱敏算法可分为两个阶段，第一个阶段是算法预处理，定义脱敏算法的主要参数，第二阶段是算法的轮加密过程，脱敏算法具体步骤如算法 2 所示。

算法 2：数据脱敏算法 DeSensitive（X，T）

前提条件：128 位分组密码的密码函数 SM4；用于分组密码的密钥 K；基数 radix；支持的消息长度范围 [minlen..maxlen]；调整值的最大字节长度 maxTlen。

输入：基数为 radix、长度为 n 的数字字符串 X，其中 $n \in$ [minlen..maxlen]；tweak 记作 T，一个字节长度为 t 的字节字符串，其中 $t \in$ [0..maxTlen]。

输出：字符串 Y，使得 LEN (Y)$=n$。

1：设置 $u=\lceil n/2\rceil$；$v=n-u$。

2：设置 $A=X[1..u]$；$B=X[u+1..n]$。

3：设置 $b=\lceil\lceil v \cdot \mathrm{LOG}(\mathrm{radix})\rceil/8\rceil$。

4：设置 $d=4\lceil b/4\rceil+4$。

5：设 $P=[1]^1 \| [2]^1 \| [1]^1 \| [\mathrm{radix}]^3 \| [10]^1 \| [u \bmod 256]^1 \| \| [n]^4 \| [t]^4$。

6：i 从 0 到 9 遍历：

6.1　令 $Q=T \| [0]^{(-t-b-1)\bmod 16} \| [i]^1 \| [\mathrm{NUM}_{\mathrm{radix}}(B)]^b$。

6.2　令 $R=\mathrm{PRF\_SM4}(P \| Q)$。

6.3　令 S 取以下字符串的前 d 位：

$$R \| \mathrm{SM4}_K(R \oplus [1]^{16}) \| \mathrm{SM4}_K(R \oplus [2]^{16}) \cdots \mathrm{SM4}_K(R \oplus [\lceil d/16\rceil - 1]^{16})$$

6.4　令 $y=\mathrm{NUM}(S)$。

6.5　如果 i is even，令 $m=u$；否则，令 $m=v$。

6.6　令 $c=(\mathrm{NUM}_{\mathrm{radix}}(A)+y) \bmod \mathrm{radix}^m$。

6.7　令 $A=B$。

6.8　令 $B=\mathrm{STR}^m_{\mathrm{radix}}(c)$。

7：返回 $A \| B$。

1）脱敏算法预处理阶段：

步骤 1～2 中，定义 u，v 为字符串长度 n 的相邻中位索引，基于该索引将输入字符串 X 拆分为 A 和 B 两个子字符串。若 n 是偶数，则 A 与 B 字符串长度相等；否则，字符串 A 将比字符串 B 少一个字符。

在步骤 3～5 中，定义了一个固定数据块 P，将作为初始向量参与 SM4 实现的伪随机函数。

2）Feistel 循环迭代加密：

步骤 6，进行 FF1 的十轮 Feistel 循环迭代，每一轮执行相同的子步骤，对应算法 2 中的 6.1～6.8，循环迭代加密具体算法步骤如下。

步骤 6.1～6.3，首先，将调整值 T、子字符串 B 及其对应的基数 radix 和循环迭代加密的轮数 i 共同编码为一个二进制字符串 Q；接着，将初始向量 P 和 Q 进行字符串接后输入基于 SM4 算法的 PRF 函数，以生成一个数据块 R；最后，R 被截断或扩展为具有适当字节数 d 的字节字符串 S。

步骤 6.4～6.6，计算 NUM（S）与 $\mathrm{NUM}_{\mathrm{radix}}$（$A$），求和之后取模 radix^m，生成结果 c。其中 m 的值为 Feistel 轮中 A 的长度。

步骤 6.7～6.8，首先，交换 A，B 的值，接着，将步骤 6.6 中得到的值 c 通过 转换为一个数字字符串，并赋值给 B。

经过以上步骤完成每一轮的 Feistel 加密，经过 9 轮循环，输出子字符串 A、B 的串接字符串。FF1 算法的 PRF 函数和基于传统 AES 加 CBC[17-18] 模式实现，本研究则通过 SM4 加密算法进行国产化替代。

算法恢复过程所使用的 FF1 解密算法与加密算法的结构相同，如算法 3 所示。解密算法的过程在第 6 步与加密算法存在差异，在 Feistel 循环迭代过程中其序列索引是脱敏过程的逆序，在第 6.7～6.8 步骤中，B 与 A 的值传递方向相反，其中 A 的值由 6.6 步骤中得到的 c 转换为字符串得到，在步骤 6.6 中，转换脱敏算法中的模加为模减来实现密文的恢复，其中解密算法的伪随机函数与步骤 6.3 中的数据块分组加密函数仍由 SM4 算法实现。

算法 3：数据脱敏恢复算法 ReSensitive（X，T）

前提条件：128 位分组密码的密码函数 SM4；用于分组密码的密钥 K；基数 radix；支持的消息长度范围 [minlen..maxlen]；调整值的最大字节长度 maxTlen。

输入：基数为 radix、长度为 n 的数字字符串 X，其中 $n \in$ [minlen..maxlen]；调整值 T，一个字节长度为 t 的字节字符串，其中 $t \in$ [0..maxTlen]。

输出：字符串 Y，使得 LEN (Y)$=n$。

1：设置 $u=\lceil n/2\rceil$；$v=n-u$.

2：设置 $A=X[1..u]$；$B=X[u+1..n]$.

3：设置 $b=\lceil\lceil v \cdot \mathrm{LOG}(\mathrm{radix})\rceil/8\rceil$.

4：设置 $d=4\lceil b/4\rceil+4$

5：设 $P=[1]^1 \| [2]^1 \| [1]^1 \| [\mathrm{radix}]^3 \| [10]^1 \| [u \bmod 256]^1 \| \| [n]^4 \| [t]^4$.

6：i 从 9 到 0 遍历：

6.1　令 $Q=T \| [0]^{(-t-b-1)\bmod 16} \| [i]^1 \| [\mathrm{NUM}_{\mathrm{radix}}(A)]^b$.

6.2　令 $R=\mathrm{PRF\_SM4}(P \| Q)$.

6.3　令 S 取以下字符串的前 d 位：

$$R \| \mathrm{SM4}_K(R \oplus [1]^{16}) \| \mathrm{SM4}_K(R \oplus [2]^{16}) \cdots \mathrm{SM4}_K(R \oplus [\lceil d/16\rceil - 1]^{16})$$

6.4　令 $y=\mathrm{NUM}(S)$.

6.5　若 i 是偶数，令 $m=u$；否则，令 $m=v$.

6.6　令 $c=(\mathrm{NUM}_{\mathrm{radix}}(B)-y) \bmod \mathrm{radix}^m$.

6.7　令 $B=A$.

6.8　令 $A = \mathrm{STR}_{\mathrm{radix}}^{m}(c)$.

7：Return $A \parallel B$.

4　统应用及测试

本节拟对文件脱敏系统的性能和准确率进行测试，测试流程如图 4 所示。

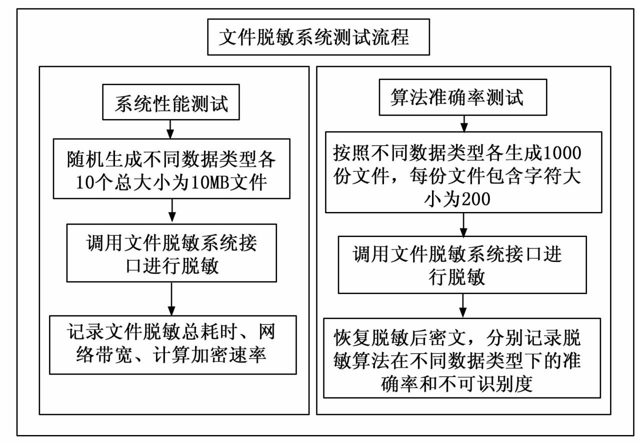

图 4　系统测试流程

通过性能测试与准确率测试，验证文件脱敏系统在实际场景中的可行性，表 1 展示了不同类型数据的脱敏及恢复示例，对于纯数字、数字与字母混合以及任意类型混合的数据，分别给出了示例数据、脱敏结果和恢复结果。

表 1　数据脱敏算法测试结果

	纯数字	数字、字母
示例数据	419123456782090123	41912345678209001p
脱敏结果	141715601562274456	27037342037362634K
恢复结果	419123456782090123	41912345678209001p
	数字、字母、汉字	任意类型混合
示例数据	锁定嫌疑人 ID 为 10034v	邮箱为 1236@163.com
脱敏结果	蚍涗櫟糯俞 op 鼫 82608Q	蓷啰晷 0604@210.QlE
恢复结果	锁定嫌疑人 ID 为 10034v	邮箱为 1236@163.com

4.1　系统测试结果

按照文件内容纯数字、数字结合英文、和混合类型文本，各自随机生成 10 个总大小为 10 MB（约 52 万字符）的文件，在 100 M 带宽，64 G 运存，10 核的 Windows 系统中，将文件导入脱敏系统，待文件脱敏功能完成后对加密的总时长进行统计，测试结果如表 2 所示。

表 2　数据脱敏算法性能测试结果

	纯数字	数字＋字母	混合类型
加密耗时(s)	47.3	82.5	209
加密速率(words/s)	13 000	6 500	2 600

由表 2 可见，系统脱敏算法针对于纯数字类型的内容脱敏速率最快，加密速率达到 13 000 words/s，10 M 的文件加密总消耗时长为 47.3 秒。混合类型的数据加密速率达到 2 600 words/s，10 M 的文件加密总消耗时长为 209 s。本研究中的加密速率满足实际应用中对于敏感文件信息的处理需求。通过调用系统脱敏接口，对不同类型的文件进行脱敏，设置脱敏轮数为 10，记录单次的脱敏时间和平均脱敏的时间，实际性能测试的截图如图 5 所示。

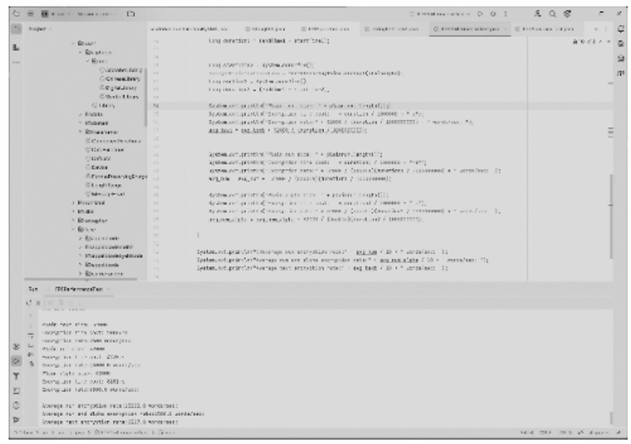

图 5　系统性能测试截图

按照纯数字，数字与字母混合以及任意类型混合（数字、特殊符号、字母与汉字）三种模式各随机生成 1 000 份文件，每份文件中随机生成 200 个字符。将共计 3 000 份文件输入文件脱敏系统，通过数据脱敏与恢复功能来对测试加解密的准确率与不可识别度进行比对，随机截取单次的测试结果如图 6 所示。

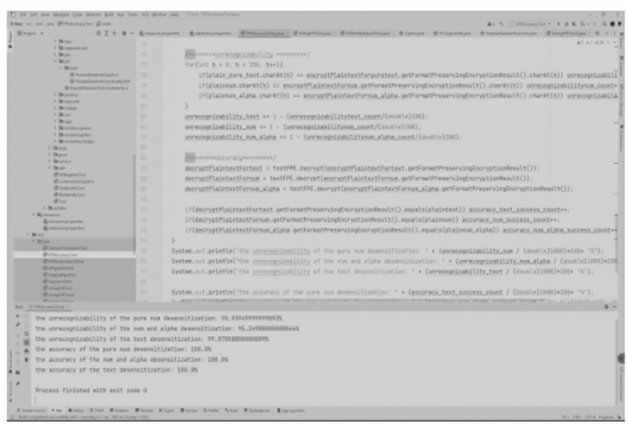

图 6　系统准确率测试截图

其中，算法加解密准确率定义为对于脱敏内容恢复的结果与原文 100%匹配所占全部脱敏条数的比例。不可识别度定义为任意 1 条数据中，信息脱敏后与原文仍重叠的字符占每条数据长度的比例。

通过对于系统准确率测试进行平均统计，如图 7 中的测试结果，可以证明，本研究的文件脱敏系统通过特定的脱敏方法，可有效的将原始数据转换为难以直接识别的形式且同时具备恢复到原始数据的能力，该研究满足在某些场景下既要保护数据隐私又能在需要时还原数据的需求。

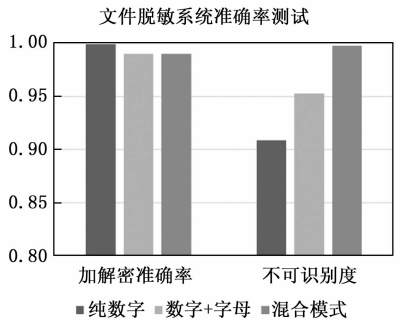

图 7　文件脱敏系统准确率测试结果

4.2　系统应用分析

4.2.1　医疗领域应用

医学研究机构在进行疾病研究、药物研发等工作时，需要大量的医疗数据。文件脱敏系统可以实现对医疗数据进行脱敏[19-20]，基于特定的文件内容过滤规则，可以在保护患者隐私的前提下，为医学研究提供可靠的数据支持。

4.2.2　金融领域应用

银行、证券等金融机构存储着大量客户的个人信息，如姓名、身份证号、银行卡号、账户余额等。在进行数据共享给第三方机构进行市场分析、业务合作，或者内部员工进行测试、培训等场景时，需要对这些敏感信息进行脱敏处理，防止客户信息泄露。

4.2.3　企业内部数据管理

企业内部的文件可能包含员工的个人信息、客户基本信息、竞方商业机密等敏感内容。在进行外部数据合作、内部数据共享、跨部门协作、员工培训等场景时，需要对敏感文件进行脱敏处理，防止敏感信息泄露，以保护客户的商业利益和与个人信息相关的敏感信息。

5　结束语

本研究通过设计汉字索引转换协议，优化了基于哈希表建立字典库索引过程中内存消耗的问题，并通过系统性能测试与加解密准确性测试验证了基于国密算法 SM4 与 FPE 的 FF1 算法构建的文件脱敏系统可以有效的处理任意类型数据的脱敏，提升了数据脱敏的范围、准确性和效率。同时，本研究未来的优化方向是通过引入多线程技术，在文件脱敏系统处理多字符串内容段时进行并行处理，进一步提升文件脱敏系统在处理混合格式文件内容时的脱敏速率[21-22]；研究通过 FPE 的 FF3 算法[16,23]，通过减少加密轮数，进一步优化数据脱敏的效率；通过在硬件 FGPA 下结合 SM4 实现数据脱敏算法的硬件部署[11]，促进文件脱敏系统底层加密接口的轻量化部署。

参考文献：

[1] 张馨方，周江华. 基于轻量型 AES 加密算法的浮空器平台数据传输方案 [J]. 计算机测量与控制，2023，31 (6)：183－190.

[2] 张晓敏. 大数据加密算法在数据安全保护中的应用研究 [J]. 计算机测量与控制，2021，29 (5)：204－208.

[3] 薛中伟. 基于区块链技术的工控数据安全传输系统设计 [J]. 计算机测量与控制，2022 (4)：30.

[4] 孙　煦. 基于区块链技术的健康医疗数据隐私加密控制系统设计 [J]. 计算机测量与控制，2024，32 (3)：188－194.

[5] 顾兆军，蔡　畅，王　明. 基于改进保留格式加密的民航旅客数据脱敏方法 [J]. 信息网络安全，2021，21 (5)：39－47.

[6] 马　龙，薛晨蕾，张　乐. 民航旅客隐私数据动态分级脱敏处理方法 [J]. 无线电通信技术，2023，49 (2)：338－344.

[7] 张玉磊，骆广萍，张永洁，等. 基于格式保留的敏感信息加密方案 [J]. 计算机工程与科学，2020，42 (2)：236－240.

[8] 陈　佳，彭长根，樊玫玫，等. SM4-FPE：基于 SM4 的数字型数据保留格式加密算法 [J]. 小型微型计算机系统，2019，40 (6)：1274－1279.

[9] 白　云. 基于国密 SM4 算法的 FPE 格式保留加密方法及系统：CN202310727051.3[P]. CN116707765A[2024－10－17].

[10] 吕述望，苏波展，王　鹏，等. SM4 分组密码算法综述 [J]. 信息安全研究，2016，2 (11)：995－1007.

[11] 张宏科，袁浩楠，丁文秀，等. 基于 FPGA 的 SM4 算法高效实现方案 [J]. 通信学报，2024，45 (5)：140－150.

[12] 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.

[13] 车力军，杨　蕊，聂　昆. 日志脱敏方法，装置，设备及介质：CN202211679764.9[P]. CN116186758A[2024－09－25].

[14] 李　玮，汪梦林，谷大武，等. SM4 密码算法的唯密文故障分析 [J]. 计算机学报，2022，45(8)：1814－1826.

[15] 沈传年，徐彦婷. 数据脱敏技术研究及展望[J]. 信息安全与通信保密，2023(2)：105－116.

[16] CUI B J，ZHANG B H，WANG K Y. A data masking scheme for sensitive big data based on format-preserving encryption[C]//2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). IEEE，2017，1：518－524.

[17] ABOUSHOSHA B，RAMADAN R A，DWIVEDI A D，et al. SLIM：A lightweight block cipher for internet of health things[J]. IEEE Access，2020，8：203747－203757.

[18] 张育梅. 基于分组密码的网络数据保形加密数学模型[J]. 计算机仿真，2022，39(3)：466－469.

[19] 李　超，张艳玲，张清媛. 面向医疗系统的隐私保护疾病预测研究[J]. 计算机测量与控制，2023，31(4)：219－224.

[20] 贾瑞龙，曹亚州，苗俊青，等. 基于改进 CP-ABE 模型的医疗数据隐私保护管理设计与应用[J]. 计算机测量与控制，2020，28(1)：200－204.

[21] PEREZ-RESA A，GARCIA-BOSQUE M，SANCHEZ-AZQUETA C，et al. A new method for format preserving encryption in high data rate communications[J]. IEEE Access，2020(99)：1.

[22] 张玉磊，骆广萍，张永洁，等. 基于格式保留的敏感信息加密方案[J]. 计算机工程与科学，2020，42(2)：236－240.

[23] ZHANG Y，LUO G，ZHANG Y，et al. A format preserving encryption scheme for sensitive information[J]. Computer Engineering & Science，2020，42(2)：236.