GY

中华人民共和国广播电视和网络视听行业标准

GY/T XXX—XXXX

广播电视和网络视听收视综合评价数据脱敏规则

Masking rules of radio TV and internet video and audio service big data for comprehensive evaluation

（报批稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家广播电视总局　　发布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由全国广播电影电视标准化技术委员会（SAC/TC 239）归口。

本文件起草单位：国家广播电视总局广播电视规划院、华数传媒网络有限公司、华数数字电视传媒集团有限公司、广东南方新媒体股份有限公司。

本文件主要起草人：李忠炤、郑冠雯、聂明杰、李庆国、曹志、王志豪、胡暐宸、遇琪、诸葛海标、张玮、黄元浩、唐志燕、刘晓敏、殷楚冬、张元迪。

广播电视和网络视听收视综合评价数据脱敏规则

1 范围

本文件规定了广播电视和网络视听收视综合评价数据的脱敏原则、脱敏技术、脱敏流程和脱敏要求。

本文件适用于广播电视和网络视听收视综合评价数据的脱敏。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

个人敏感信息 personal sensitive information

一旦泄露、非法提供或滥用可能危害人身和财产安全，极易导致个人名誉、身心健康受到损害或歧视性待遇等的个人信息。

注1：个人敏感信息包括身份证号码、个人生物识别信息、银行账号、通信记录和内容、财产信息、征信信息、行踪轨迹、住宿信息、健康生理信息、交易信息、14岁以下（含）儿童的个人信息等。

注2：个人信息控制者通过个人信息或其他信息加工处理后形成的信息，如一旦泄露、非法提供或滥用可能危害人身和财产安全，极易导致个人名誉、身心健康受到损害或歧视性待遇等的，属于个人敏感信息。

[来源：GB/T 35273—2020，3.2]

3.2

敏感属性 sensitive attribute

数据集中需要保护的属性，该属性值的泄露、修改、破坏或丢失会对个人产生损害。

注：在潜在的重标识攻击期间需要防止其值与任何一个个人信息主体相关联。

[来源：GB/T 37964-2019，3.10]

3.3

敏感数据 sensitive data

原始数据中具有敏感属性的用户个人信息数据。

3.4

数据脱敏 data masking

按照一定的规则对原始数据进行变形，屏蔽原始数据中的敏感信息，并保留业务环境所需要的数据特征和内容。

3.5

用户 ID user identification

由数据提供方系统生成，用于唯一识别用户的一组不重复的编码。

3.6

用户账号 user account
用户在互联网视听平台中代表自己身份的名称。

4 概述

数据脱敏是按照一定的方法、流程以及输出格式，对敏感数据进行处理，以确保敏感数据不泄露。脱敏后的数据应尽可能体现原始数据的特征和内容，并能在相关业务中继续使用。

广播电视和网络视听收视综合评价数据应为实现收视综合评价目的所必须的最小化数据，不包含用户姓名、出生日期、身份证号码、个人生物识别信息、住址、通信联系方式等个人敏感信息。对广播电视和网络视听收视综合评价数据包含的用户账号、设备信息、IP地址等，应按照本文件进行数据脱敏处理。

5 数据脱敏原则

5.1 有效性

数据经过脱敏处理之后，原始信息中包含的用户个人敏感信息应已被移除，第三方应无法通过处理后的数据得到敏感信息；或需通过巨大的经济代价、时间代价才能得到用户个人敏感信息。

5.2 可用性

脱敏后的数据应保持数据的原有特征，保证数据在非原始环境中的可用性，在脱敏过程中应保留原始数据中的信息，保证收视大数据的开发、测试、培训过程中不会受到脱敏的影响。为保证可用性应满足以下要求：

——保持原数据格式、类型、依存关系；
——保持语义完整性；
——保持引用完整性；
——保持数据统计、聚合特征；
——保持唯一性。

5.3 高效性

应保证数据脱敏的过程可通过程序自动化实现，可重复执行。

5.4 稳定性

为保障数据使用者可正常使用和分析数据，数据脱敏应保证脱敏后的数据与原始数据之间的关联性，脱敏数据之间的关联应是稳定的。

5.5 防御性

应保障数据脱敏算法不被同质属性、概率、知识推断等手段攻击，确保脱敏安全可靠。

5.6 可审计性

在数据脱敏各个阶段应加入安全审计机制，严格、详细记录数据处理过程中的相关信息，形成完整数据整理记录，用于后续问题排查与数据追踪分析。

6 数据脱敏技术

6.1 概述

广播电视和网络视听收视综合评价数据脱敏可采用泛化技术、抑制技术和扰乱技术。

6.2 泛化技术

泛化技术是指一种降低数据集中所选属性粒度的去标识化技术，对数据进行更概括、抽象的描述。泛化技术包括但不限于：

a) 截断：舍弃不需要的信息，仅保留部分关键信息，保证数据的模糊性；
b) 偏移取整：按照一定粒度对时间进行向上或向下偏移取整，保证时间数据满足一定的分布特征，同时隐藏原始时间信息；
c) 规整：将数据按照大小规整到预定义的多个档位进行分类。

6.3 抑制技术

抑制技术即对不满足隐私保护的数据项删除，不进行发布。包括从所有记录中对选定的属性（如屏蔽）、对所选定的属性值（例如，局部抑制），或是从数据集中选定的记录（例如，记录抑制）进行的删除操作；或对敏感数据部分内容使用通用字符进行替换（掩码技术）。

6.4 扰乱技术

扰乱是指通过加入噪声的方式对原始数据进行干扰，以实现对原始数据的扭曲、改变，扰乱后的数据仍保留着原始数据的分布特征，具体的技术方法包括但不限于：

a) 加密：使用加密算法对原始数据进行加密；
b) 重排：将原始数据按照特定规则进行重新排列；
c) 替换：按照特定规则对原始数据进行替换；
d) 均化：针对数值性的敏感数据，在保证脱敏后数据集总值或平均值与原数据相同的情况下，改变数据的原始值；
e) 散列：对原始数据取散列值，使用散列值来代替原始数据。

7 数据脱敏流程

7.1 概述

原始广播电视和网络视听收视综合评价数据经数据预处理后，应按发现敏感数据、标识敏感数据、制定脱敏方案、执行脱敏操作、评估脱敏效果的流程进行数据脱敏处理。

7.2 发现敏感数据

数据提供方应对原始数据进行梳理和分类，将数据分为高度敏感数据、中度敏感数据和非敏感数据；同时，应分析并建立完整的脱敏数据位置和关系库，确保数据脱敏能充分考虑到数据应用的业务范围、脱敏后数据对原始数据业务特性的继承等。

7.3 标识敏感数据

数据提供方应对敏感数据进行标识，并对敏感数据的关系进行调整，以保证数据的关联关系。

7.4 制定脱敏方案

数据提供方应按照广播电视和网络视听收视综合评价的业务需求，根据场景确定脱敏规则和脱敏技术，确定数据脱敏的方案。

7.5 执行脱敏操作

数据提供方应按照脱敏方案，对广播电视和网络视听收视综合评价数据执行脱敏操作。

7.6 评估脱敏效果

数据提供方和数据接收方对数据脱敏效果进行评估，确保脱敏操作有效、脱敏数据可用。

8 数据脱敏要求

8.1 脱敏要求

数据提供方应对用户ID、用户账号、终端设备ID、终端设备网络IP等信息进行脱敏。

数据提供方可根据数据安全要求，在不影响收视综合分析和合规的前提下，对其他敏感数据进行脱敏。

8.2 用户数据

8.2.1 用户 ID

数据提供方可对用户ID进行脱敏处理，脱敏后的用户ID应保证数据关联性，并对数据提供方可逆。

为方便数据分析，数据提供方对用户ID脱敏后，应按图1所示的格式，在用户ID前加上运营商类型和用户地域信息。

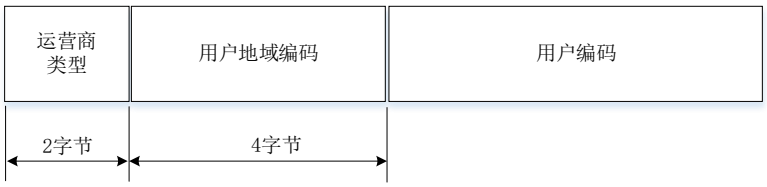

图1 用户 ID 编码规则

用户ID编码规则包含以下信息。

a) 运营商类型，2字节，取值如下：
   - 有线电视：01；
   - IPTV：02；
   - 互联网电视：03；
   - 互联网视听服务：04。

b) 用户地域编码：4字节，取值为地区邮政编码前4位，如北京市西城区为“1000”，上海浦东新区为“2012”，浙江杭州萧山区为“3112”。

8.2.2 用户账号

数据提供方可对用户账号进行脱敏处理。

8.2.3 用户年龄

数据提供方应对用户年龄进行脱敏处理。

用户年龄脱敏采用规整算法，规整档位间隔应为5，如0至5岁规整为5，6至10岁规整为10，11至15岁规整为15。

8.3 设备数据

8.3.1 终端设备 ID

终端设备ID脱敏应保证终端设备ID唯一性。

8.3.2 终端设备网络 IP

数据提供方应对终端设备网络IP进行脱敏处理。

终端设备网络IP脱敏应采用掩码技术，对终端设备网络IP后两段字符使用字符“x”进行替换，替换后的IP地址如“58.100.xxx.xxx”。

参 考 文 献

[1] GB/T 35273—2020 信息安全技术 个人信息安全规范
[2] GB/T 37964—2019 信息安全技术 个人信息去标识化指南
[3] GD/J 075—2018 电视收视数据交换接口规范