GD

中华人民共和国广播电影电视行业暂行技术文件

GD/J 076—2018

电视收视数据清洗规范

Specification of viewership data cleaning

2018－03－01 发布　　2018－03－01 实施

国家新闻出版广电总局科技司　　发 布

目　次

前　言

本技术文件按照GB/T 1.1—2009给出的规则起草。

请注意本技术文件的某些内容可能涉及专利。本技术文件发布机构不承担识别这些专利的责任。

本技术文件由国家新闻出版广电总局科技司归口。

本技术文件起草单位：国家新闻出版广电总局广播电视规划院、中央电视台、央视国际网络有限公司、中国广播电视网络有限公司、国家新闻出版广电总局广播科学研究院、国家新闻出版广电总局广播电视卫星直播管理中心、中国传媒大学、北京歌华有线电视网络股份有限公司、江苏省广电有线信息网络股份有限公司、浙江华数广电网络股份有限公司、成都广播电视台、成都橙视传媒科技股份公司、北京国双科技有限公司、广州市诚毅科技软件开发有限公司、广州欢网科技有限责任公司、北京勾正数据科技有限公司。

本技术文件主要起草人：邓向冬、何剑辉、吴钟乐、聂明杰、孙黎丽、黄卓伟、张文鹏、吉钰丽、邱星华、唐云峰、杨旭、孙于扬、覃毅力、房磊、张福国、尹亚光、施玉海、王鹏、陈杰、柴剑平、万敏、孔令浚、杨利中、黎志。

电视收视数据清洗规范

1 范围

本技术文件规定了各收视数据提供方清洗数据过程中的规范化流程和规则。

本技术文件适用于有线电视、直播卫星电视、地面电视、交互式网络电视（IPTV）、互联网电视等不同收视渠道电视原始收视数据的清洗。

2 规范性引用文件

下列文件对于本技术文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本部分。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本部分。

GD/J 075—2018 电视收视数据交换接口规范

3 术语、定义和缩略语

3.1 术语和定义

下列术语和定义适用于本技术文件。

3.1.1

原始收视数据 original viewership data

在电视收视过程中，从用户、收视终端、服务商等处，直接采集到的、与用户收视过程及内容相关的、未按规范标准进行清洗、转换、整合等处理的数据信息。

3.1.2

心跳数据 heartbeat data

终端设备定时发送的、与具体播出业务和内容无关的联络数据，用于监测终端设备状态。

3.1.3

空数据 null data

终端监测软件根据各类预设逻辑所采集的一类数据，数据值表述为“null”。

3.1.4

路径数据 path data

用户收视电视节目的位置及路径的相关数据信息。

3.1.5

页面数据 page data

页面中所有元素的用户点击数据，包括直播列表页、广告位、推荐位及各类按钮等相关点击行为数据信息。

3.1.6

换台收视数据 channel switch data

用户连续切换频道所产生的收视数据，单次连续收视时长小于3秒。

3.2 缩略语

下列缩略语适用于本技术文件。

ID 标识号（Identification）

QoE 体验质量（Quality of Experience）

QoS 服务质量（Quality of Service）

UTF-8 通用8位编码字符集（8-bit Unicode Transformation Format）

4 概述

数据清洗是按照统一的方法、流程以及输出格式，对来自有线电视、直播卫星电视、地面电视、交互式网络电视（IPTV）、互联网电视等来源的原始收视数据进行处理，过滤不符合要求的数据，清除错误无效数据、处理不一致数据、修补不完整数据，以利于生成计算收视相关指标的有效数据。数据清洗建议流程参见附录A。

5 数据清洗规则

5.1 去除无效数据

5.1.1 基本的无效数据

基本的无效数据包括的类型主要有：

——不完整数据：丢失部分字段的数据；

——错误数据：存在乱码或格式错误或从时间点上存在逻辑错误的数据；

——重复数据：用户行为存在重复的数据。

5.1.2 其他无效数据

其他无效数据是指收视数据提供方采集数据过程中产生的系统辅助数据，其主要用于验证采集环节的完整性与可靠性，对收视数据交换与分析没有实质帮助。无效数据可能存在下述类型中：

——心跳数据；

——空数据；

——路径数据；

——页面数据。

在某些情况下，心跳数据、页面数据可以作为辅助分析数据，用来分析用户的收视行为、业务访问情况，可不进行清洗。

5.2 结合业务特征还原节目数据

收视数据提供方获取的原始数据中，包含节目数据，但不同的业务形式和采集方式使得节目数据不能直接呈现，需要通过清洗还原直观的节目数据。

——直播类节目：终端监测软件采集原始收视数据时，通常只能获取频道号，清洗时应补齐每条行为数据的节目名称；

——非直播类节目：用户对点播、时移、回看等存储在媒资系统的非直播类节目进行收视时，终端监测软件只能监测到相应的内容 ID，清洗时应以节目名称替换内容 ID。

6 数据清洗要求

6.1 数据完整性校验

数据提供方应对原始收视数据进行完整性校验，并确保数据包含GD/J 075中规定的必选字段。

6.2 数据统一编码处理

数据提供方应对原始收视数据进行编码处理，统一采用UTF-8编码。

6.3 无效数据校验与处理

数据提供方应按照无效数据的定义进行数据校验，剔除无效数据信息。

换台收视数据不作为无效数据。

6.4 噪声数据校验与处理

数据提供方应按照收视有效性规则进行数据校验，将直播收视时长大于6小时的数据进行剔除。

6.5 收视数据与节目内容对应

数据提供方应对收视数据与节目内容进行对应。

6.6 时间格式标准化处理

为保障不同数据来源的时间格式统一性，应统一采用纯数字型数据元进行表达，以实现时间的唯一性和全局性。标准格式为：YYYYMMDDHHMMSS。

6.7 数据去重处理

针对完全一致的数据进行去重处理，保证数据信息的唯一性。

附 录 A
（资料性附录）
数据清洗建议流程

数据清洗建议流程如图A.1所示。

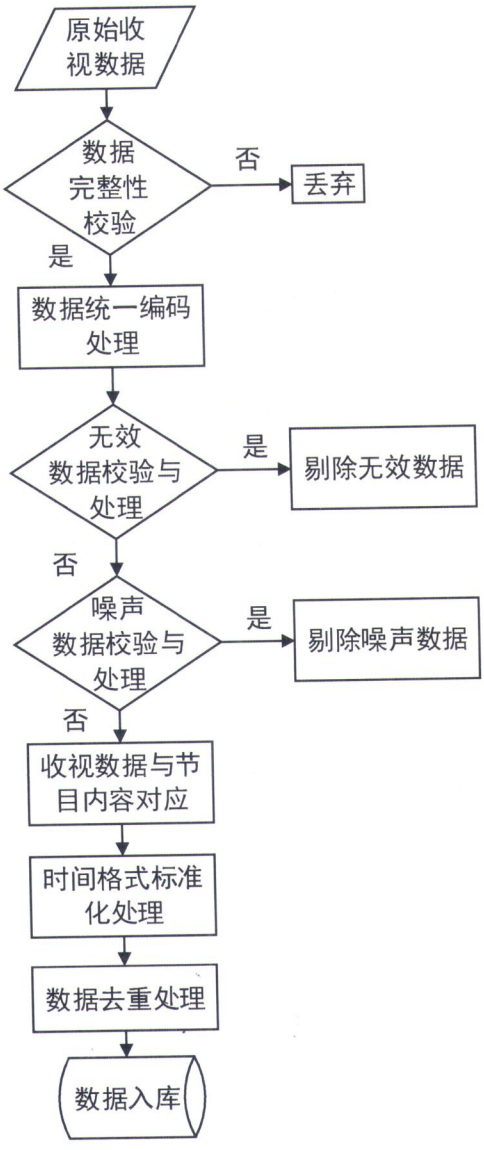

图 A.1 数据清洗建议流程图