Q/SYY

大 数 据 研 究 院 企 业 标 准

Q/SYY JC105.02—2022

大数据清洗

Big Data Cleaning

2022－07－15 发布　　2022－08－01 实施

中科大数据研究院　　发 布

大数据清洗

1 范围

本文件定义了大数据清洗加工的标准方法，用于指导中科大数据研究院数据的清洗加工工作。

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

数据采集 data acquisition

对数据资源进行收集并形成原始记录的过程。

3.2

脏数据 dirty data

系统中的数据不在给定的范围内或对于实际业务毫无意义，或是数据格式非法，以及在系统中存在不规范的编码和含糊的业务逻辑。

3.3

数据清洗 data clean

利用现有的数据挖掘手段和方法清洗脏数据，将脏数据去除或转化为满足数据质量要求或应用要求的数据的过程。它是发现并纠正数据文件中可识别的错误的一道重要程序。

3.4

结构化数据 structural data

由二维表结构来逻辑表达和实现的数据，严格地遵循数据格式与长度规范，主要通过关系型数据库进行存储和管理。

3.5

非结构化数据 unstructured data

数据结构不规则或不完整，没有预定义的数据模型，不方便用数据库二维逻辑表来表现的数据。

3.6

半结构化数据 semi- structured data

利用现有的数据挖掘手段和方法清洗脏数据，将脏数据去除或转化为满足数据质量要求或应用要求的数据的过程。它是发现并纠正数据文件中可识别的错误的一道重要程序。

3.7

数据规范 data specifications

对数据标准、数据模型、业务规则、元数据和参考数据进行有关存在性、完整性、质量及归档的测量标准。

3.8

数据完整性规则 data integrity fundamentals

对数据进行有关存在性、有效性、结构、内容及其他基本数据特征的测量标准。

3.9

数据覆盖 data coverage

相对于数据总体或全体相关对象数据的可用性和全面性的测量标准。

3.10

表达质量 presentation quality

如何进行有效信息表达以及如何从用户中收集信息的测量标准。

3.11

数据衰变 data decay

对数据负面变化率的测量标准。

4 数据清洗流程与原则

4.1 数据清洗流程

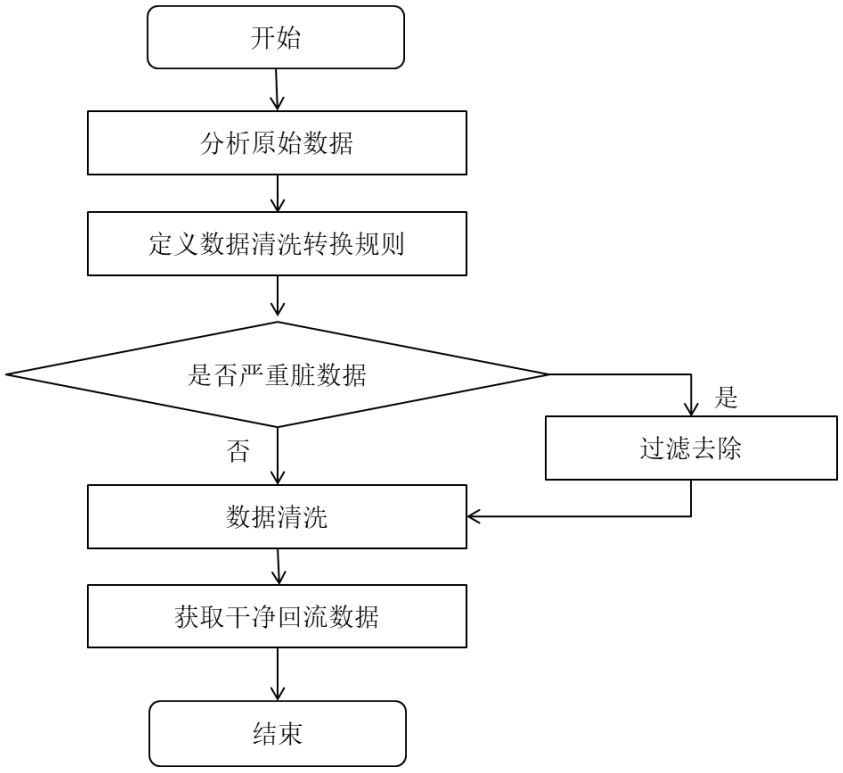

图 1 数据清洗具体流程图

数据清洗的方法包括：缺失数据处理、相似重复对象检监测、异常数据处理、逻辑错误监测、数据不一致性监测等。用不同方法清洗的数据，对后续挖掘应用工作会产生不同的影响。

4.2 数据清洗流程

4.2.1 方法一致性

数据资源清洗加工工作应统一决策，统一数据库范围工作方法、技术指标均应当统一，从而达成数据产品的一致性。

4.2.2 数据可信性

数据可信性包括精确性、完整性、一致性、有效性、唯一性。

精确性：描述数据是否与其对应的客观实体的特征相一致。

完整性：描述数据是否存在缺失记录或缺失字段。

一致性：描述同一实体的同一属性的值在不同的系统石头一致。

有效性：描述数据是否满足用户定义的条件或在一定的阈值范围内。

唯一性：描述数据是否存在重复记录。

4.2.3 数据可用性

数据可用性包括时间性、稳定性等。

时间性：描述数据是当前数据还是历史数据。

稳定性：描述数据是否是稳定的，是否在其有效期内。

5 数据清洗流程控制

5.1 数据清洗流程

数据清洗分为数据预处理与数据清洗，其中数据预处理又包括数据抽取、数据过滤、数据转换、数据加载四个步骤，数据清洗包括加载数据清洗规则、脏数据处理，脏数据处理完毕后进行数据审核，数据审核通过进行数据更新，完成数据清洗，数据审核不通过则再次进行脏数据处理。

5.1.1 数据抽取

数据抽取是从数据源中抽取数据的过程。数据抽取最常用的是ETL技术，具体数据抽取工具种类繁多，可根据实际业务数据的特点进行选择，从数据库中抽取数据一般有以下两种方式。

全量抽取：全量抽取类似于数据镜像或数据复制，它将数据源中的表或视图数据原封不动的从数据库中抽取出来。该方法主要用于在系统数据初始化时使用。

增量抽取：增量抽取是指在上次收取完成后，对数据库中新增或修改的数据的抽取。

5.1.2 数据过滤

数据过滤要初步实现对业务数据中不符合应用规则或者无效的数据进行过滤操作，使得数据标准统一。

5.1.3 数据转换

数据转换要实现对数据的格式、信息代码、值的冲突进行转换。

5.1.4 数据加载

数据加载过程进行的主要操作是操作和修改操作。

5.2 数据清洗

5.2.1 数据清洗规则

数据清洗规则包括：非空检核、主键重复、非法代码及非法值清洗、数据格式检核、记录数据检核。

非空检核：要求字段为非空的情况下，需要对该字段数据进行检核。

主键重复：多个业务系统中同类数据经过清洗后，在统一保存时，为保证主键唯一性，需进行检核和替换工作。

非法代码及非法值清洗：非法代码问题包括非法代码、代码与数据标准不一致等，非法值问题包括取值错误、格式错误、多余字符、编码错误等，需根据具体情况进行校核及修正。

数据格式检核： 通过检查表中属性值的格式是否正确来衡量其准确性。

记录数检核：指各个系统相关数据之间的数据总数检核或者数据表中每日数据量的波动检核。

5.2.2 脏数据处理

数据质量中普遍存在的空缺值、离群值和不一致数据的情况，这些脏数据可以采用人工检测、统计学方法、聚类、分类、关联规则等方法来实现数据清洗。

脏数据处理具体步骤包含六个阶段，分别是数据预处理阶段，去除/补全有缺失的数据，去除/修改格式和内容错误的数据，去除/修改逻辑错误的数据，去除不需要的数据和关联验证。

根据缺陷类型分类，可以将脏数据分为缺失值数据、错误数据和错误关联数据三种核心问题效数据进行数据清洗。

5.2.2.1缺失值处理

不完整的、含噪声的数据是未经清洗的数据集的共同特点。在数据集中，若某记录的属性值被标记为空白或含噪声等，则认为记录存在缺失值，是不完整的数据。缺失值是最常见的数据问题，处理缺失值按照以下步骤进行：

a）确定缺失值范围：对每个字段都计算其缺失值比例，然后按缺失比例和字段重要性，分别制定策珞。

b）对于一些重要性高，缺失率较低的缺失值数据，可根据经验或业务知识估计，也可通过计算进行填补。

c）对于指标重要性高，缺失率也高的缺失值数据，需要和取人员或业务人员了解，是否有其他渠道可以取到相关数据，必要时进行重新采集，若无法取得相关数据，则需要对缺失值进行填补。

d）对于指标重要性低，缺失率也低的缺失值数据，可只进行简单填充或不作处理。

e）对于指标重要性低，缺失率高的缺失值数据，可备份当前数据，直按删掉不需要的字段。

填补空缺值的方法有以下三种：

a）以业务知识或经验推测填充缺失值。

b）以同一指标的计算结果（均值、中位数、众数等）填充失值。

c）不同指标的计算结果填充失值。

5.2.2.2错误数据处理

错误数据包含格式内容问题数据和逻辑问题数据两类。

a）格式内容问题有以下三类：

时间、日期、数值、全半角等显示格式不一致：处理方法是将其处理成一致的某种格式。

内容中有不该存在的字符：需要以半自动校验半人工方式找出可能存在的问题，并去除不需要的字符。

数据内容与该字段应有内容不符。

b）逻辑问题数据处理一般采用逻辑推理的方法，可以去掉一些使用简单逻辑推理即可直接发现问题的数据，防止分析结果错误。主要包含以下三个步骤：

去重；

离群值；

修正矛盾内容。

5.2.2.3错误关联数据处理方法

对于不一致数据的处理，主要体现为数据不满足完整性约束。可以通过分析数据字典、元数据等，还可梳理数据之间的关系，并进行修正。错误关联数据清洗方法主要有以下方法：

a）统计学方法：将属性当作随机变量，通过置信区间来判断值的正误。

b）基于聚类的方法：根据数据相似度将数据分组，发现不能归并到分组的孤立点。

c）基于距离的方法：使用距离度量来量化数据对象之间的相似性。

d）基于分类的方法：训练一个可以区分正好藏数据和异常数据的分类模型。

e）基于关联规则的方法：定义数据之间的关联规则，不符合规则的数据被认为是异常数据。

5.3 非需求数据处理

在数据清理过程中，每一步具体操作前，务必做好数据备份工作，对于明确为非需要字段，可以从数据中删除,对于尚不明确是否需要的字段,原则上数据量在可处理的范围内时,尽可能保留相应字段。

6 数据清洗质量控制

6.1 数据清洗质量评估意义

数据清洗的评估实质上是对清洗后的数据的质量进行评估,而数据质量的评估过程是一种通过测量和改善数据综合特征来优化数据价值的过程。

6.2 数据清洗质量评估指标

数据清洗质量评估指标包含数据规范、数据完整性、重复性、准确性、及时性、可用性、易用性、可维护性、数据覆盖、表达质量、可理解性、相关性和可信度、数据衰变、效用性14个维度的基本评估指标。

7 数据清洗过程管理

7.1 数据清洗角色定义

数据清洗过程管理涉及的角色有提供者和管理者。提供者负责提供清洗的业务数据，管理者负责数据清洗规则制定、数据清洗发起等。

7.2 提供者管理要求

提供者应配合管理者根据接入数据指标规范与接入数据内容、接入数据流程要求，配置与部署接入服务，实现接入数据库的数据交换，提供者应该提供待清洗数据的数据结构。

7.3 管理者管理要求

管理者负责协调并明确数据清洗规则;负责构建清洗后数据及问题数据各自的数据库和数据表的结构。

7.4 数据更新总体原则

数据更新前应订立数据更新计划，计划内容包括更新的频率和周期，数据更新的内容、范围和总量等。

7.5 数据清洗服务管理要求

不得随意修改操作系统和数据库系统的用户名及密码，不得随意修改各用户的权限，不得随意增加操作系统和数据库用户。

不得随意对数据库系统进行表空间增删、数据文件增删等操作。

不得随意修改数据库系统的数据结构，包括但不限于增删或修改字段、存储过程、触发器等。