贵州省大数据清洗加工规范

Big Data Cleaning Specifications for Guizhou

贵 州 省 大 数 据 发 展 管 理 局

2017 年 6 月

目 次

前　言

本规范按照GB/T 1.1-2009《标准化工作导则 第1部分：标准的结构和编写》给出的规则起草。

本规范由贵州省大数据发展管理局提出并归口。

本规范起草单位：贵州中软云上数据技术服务有限公司、云上贵州大数据产业发展有限公司、上海贝格数据服务有限公司。

本规范主要起草人：郝达治、王文睿、李慧杰、李毅、黄奕超、韦德贵、唐俊、田沥、朱浩。

贵州省大数据清洗加工规范

1　范围

本规范定义了大数据清洗加工的标准方法，为贵州省各级政府部门之间共享交换的数据的清洗加工提供方法指导。

本规范适用于贵州全省范围内政务大数据应用主管单位、设计单位、建设单位、实施单位及评估单位等，用于指导政府数据的清洗加工工作。

贵州省大数据应用建设过程中的数据清洗工作，除了执行本规范外，还需符合国家现行有关技术要求、规范和标准。

2　规范性引用文件

下列文件对于本规范的应用是必不可少的。凡是标注日期的引用文件，仅所注日期的版本适用于本规范。凡是未注日期的引用文件，其最新版本（包括所有的修改单）适用于本规范。

GB/T 1.1-2009 《标准化工作导则 第1部分：标准的结构和编写》

DB 52/T 1123-2016《政府数据 数据分类分级指南》

DB 52/T 1124-2016 《政府数据资源目录 第1部分：元数据描述规范》

DB 52/T 1125-2016《政府数据资源目录 第2部分：编制工作指南》

DB 52/T 1126-2016《政府数据 数据脱敏工作指南》

3 术语和定义

下列术语和定义适用于本规范。

3.1 数据采集 data acquisition

数据采集是指对数据资源进行收集并形成原始记录的过程。

3.2 大数据 big data

一种规模大到在获取、存储、管理、分析方面超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

3.3 脏数据 dirty data

脏数据是指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。

3.4 数据清洗 data cleaning

数据清洗是指利用现有的数据挖掘手段和方法清洗“脏数据”，将“脏数据”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。

3.5 结构化数据 structural data

结构化数据也称作行数据，是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。

3.6 非结构化数据 unstructured data

非结构化数据是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。

3.7 半结构化数据 semi-structured data

非结构化数据处于结构化数据与非结构化数据之间，具有一定的结构，但是缺乏由底层数据模型规定的严格结构。

3.8 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。

3.9 数据完整性准则 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。

3.10 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。

3.11 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。

3.12 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　政务数据来源

政务数据资源是指政务部门在履行职责过程中制作或获取的，以一定形式记录、保存的文件、资料、图表和数据等各类数据资源，包括政务部门直接或通过第三方依法采集的、依法授权管理的和因履行职责需要依托政务信息系统形成的数据资源等。

政务数据不仅有传统的结构化数据，还包含大量的非结构化数据。进行数据清洗前，需将大量非结构化数据转换为结构化数据。在非结构化数据转化为结构化数据的过程中，可以将其利用提取特征值等方法转化为半结构化数据，最后通过技术手段转化为结构化数据。

4.2　数据质量问题与数据清洗的关系

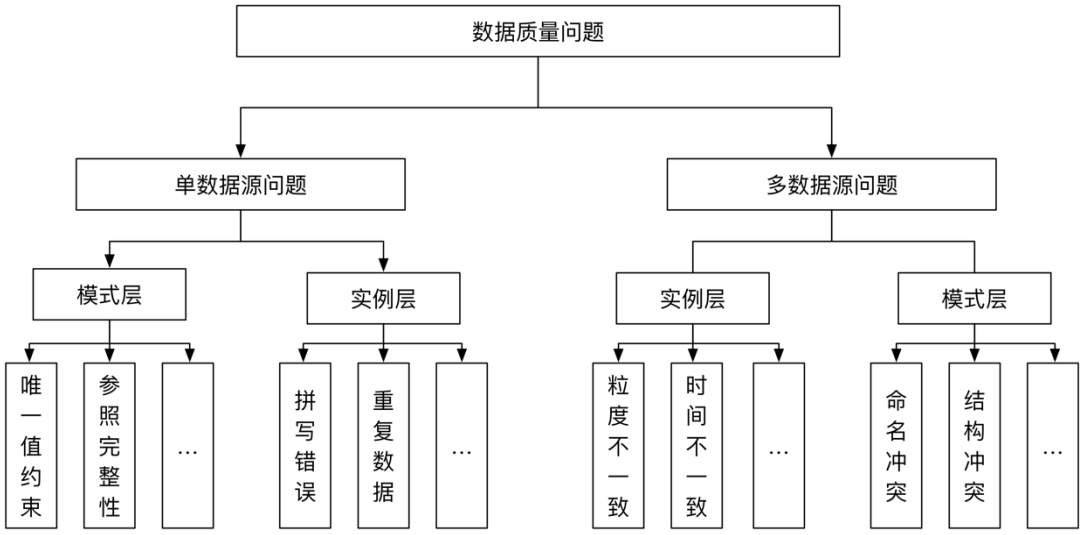

图 1　数据质量问题分类图

数据质量主要针对单数据源数据和多数据源数据两方面，两种类型数据都由实例层数据和模式层数据组成。数据清洗技术是解决数据质量问题的一种有效方法，可以检测和修正实例层的脏数据。但是数据清洗技术无法全面地解决数据质量问题中模式层的脏数据，必须借助数据整合技术。

4.3 数据清洗流程

数据清洗总体流程如下：

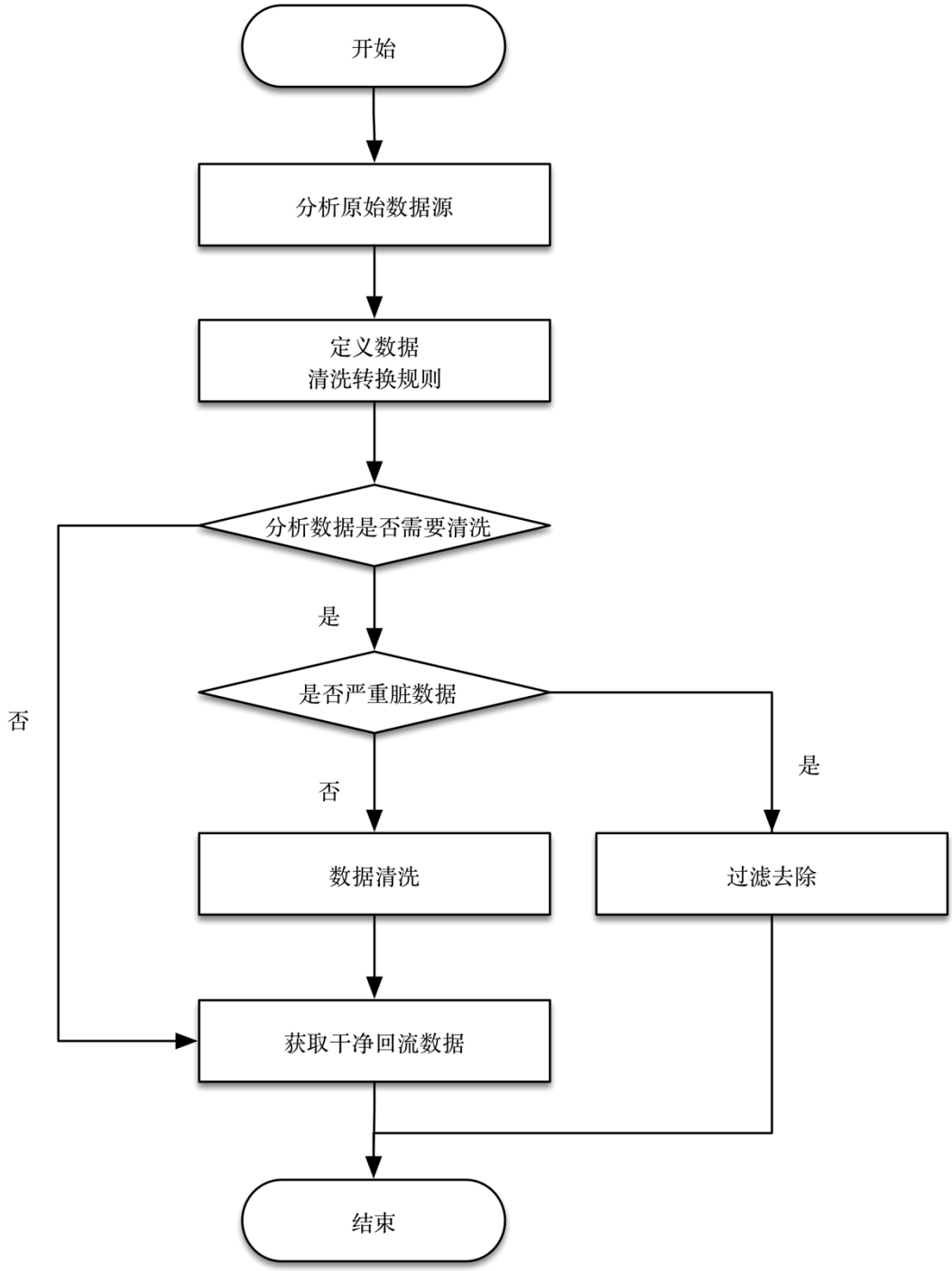

图 2　数据清洗总体流程图

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。

4.4 数据清洗加工原则

4.4.1 方法一致性

数据资源清洗加工工作应统一决策，同一数据库范围内工作方法、技术指标均应当统一，从而达成数据产品的一致性。

4.4.2 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是否一致。

有效性：描述数据是否满足用户定义的条件或在一定的域值范围内。

唯一性：描述数据是否存在重复记录。

4.4.3 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

数据清洗具体流程如下：

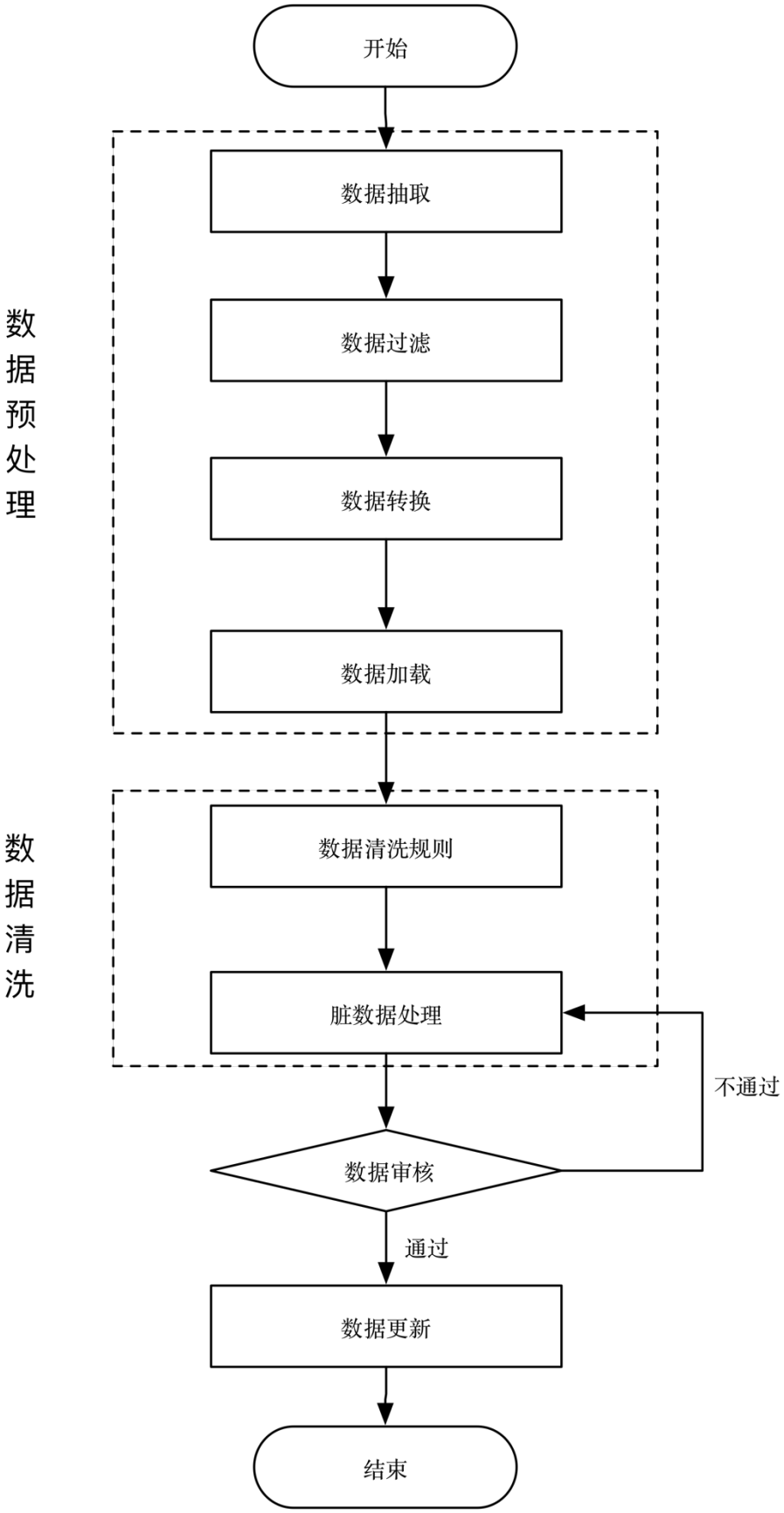

图 3　数据清洗具体流程图

5.1 数据预处理

在汇聚多个维度、多个来源、多种结构的数据之后，需要对数据进行预处理。预处理过程中除了更正、修复系统中的一些错误数据之外，更多的是对数据进行归并整理，并储存到新的存储介质中。

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。

a）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的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。

b）增量抽取（更新）：增量抽取是指在上次抽取完成后，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。常见的业务数据转换规则详见附表“业务数据转换规则表”。

5.1.4 数据加载

数据加载过程进行的主要操作是插入操作和修改操作。将干净数据及脏数据分别插入到不同的数据表中。对于数据加载工作，一般会搭建数据库环境，如果数据量大(千万级以上)，可以使用文本文件存储结合脚本程序处理进行操作。

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、非法代码清洗、非法值清洗、数据格式检核、记录数检核。

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，需进行检核工作。

非法代码、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、乱码等,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，如时间格式、币种格式、多余字符、乱码。

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。

业务约束检核应在实施过程中与业务人员共同确定，业务人员从业务的正确性、一致性、有效性等角度考虑数据的检核规则，如：建档日期、入学日期、民族信息等的有效性检核。

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、基于距离、关联规则等方法来实现数据清洗。

脏数据处理具体步骤如下：

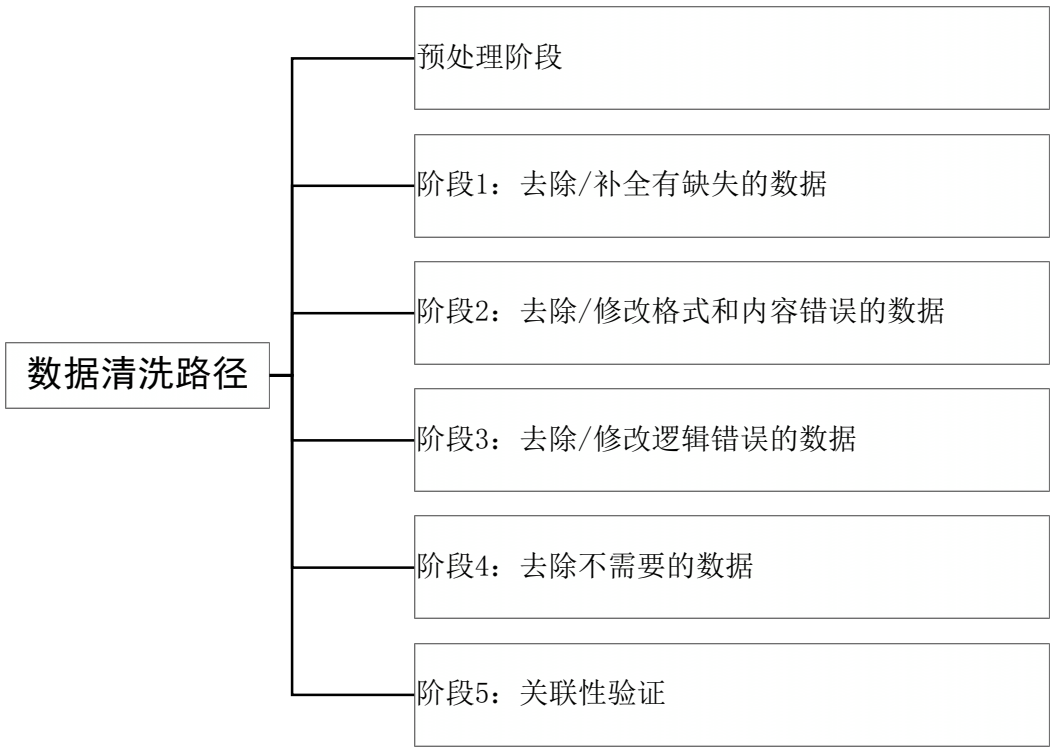

图 4　数据清洗路径图

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“-”等，则认为该记录存在缺失值，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四个步骤进行：

a）确定缺失值范围：对每个字段都计算其缺失值比例，然后按照缺失比例和字段重要性，分别制定策略，策略制定参考下图：

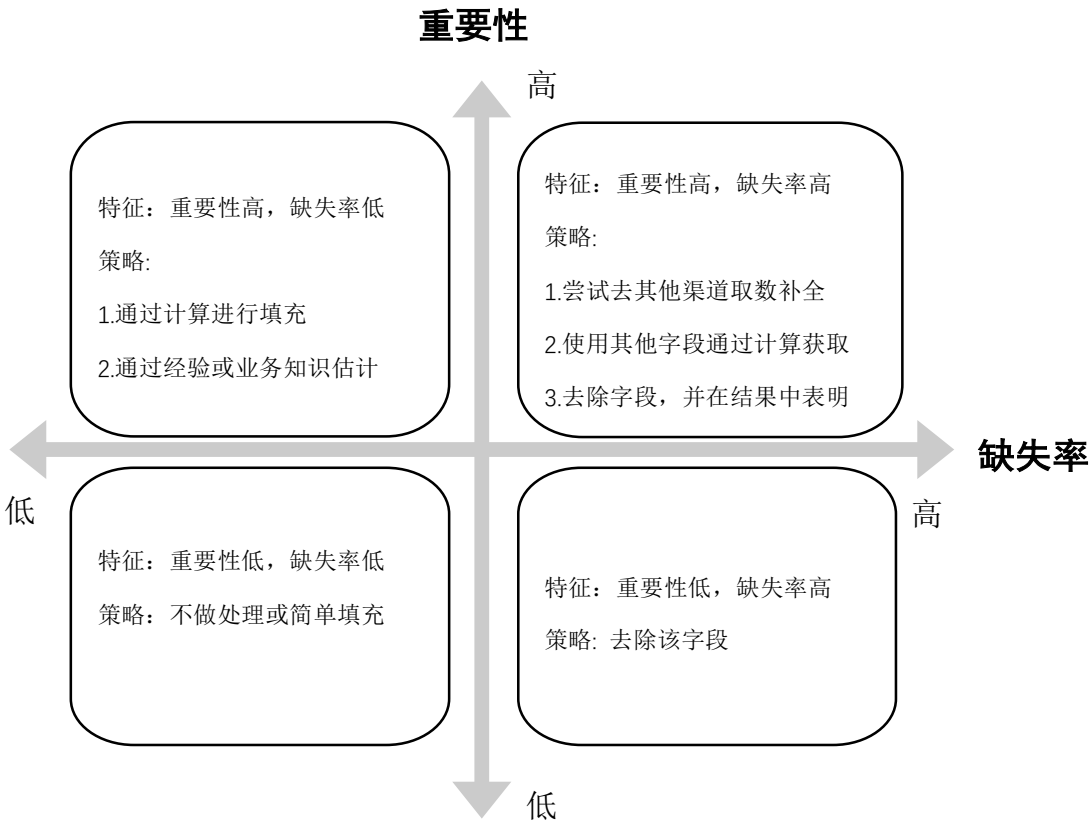

图 5　存在缺失值数据处理策略参考图

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，需要和取数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，直接删掉不需要的字段。

填补空缺值的方法有以下三种：

1）以业务知识或经验推测填充缺失值；

2）以同一指标的计算结果(均值、中位数、众数等)填充缺失值；

3）以不同指标的计算结果填充缺失值。比如：年龄字段缺失，但具有公民身份证号，则可从公民身份证号提取年龄数据。

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：

1）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。这种情况的数据多数由人工收集或用户填写而来，很大可能性在格式和内容上会存在一些问题。另外，在整合多来源数据时也有可能遇到。

2）内容中有不该存在的字符

需要以半自动校验半人工方式来找出可能存在的问题，并去除不需要的字符。典型问题如数据的开始、中间或结尾存在空格，或姓名中存在数字符号、公民身份证号中出现汉字等。

3）数据内容与该字段应有内容不符

该问题不能简单地以删除来处理，因为成因复杂，可能是人工填写错误、前端没有校验、导入数据时部分或全部存在列没有对齐的问题等，因此要详细识别问题类型。

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：

1）去重

去重放在格式内容清洗之后，原因是格式内容清理之后才能总体发现重复的业务数据。

在复杂工作环境中，由于数据多次上报，或是其他人为因素，导致数据重复值的出现是普遍的，主要使用字段相似度来识别判断重复值。

2）离群值（异常值）

采集数据时可能因为技术或物理原因，数据取值超过数据值域范围。为处理离群值，第一步即为识别离群值。识别离群值的方法主要有如下两种：

● 数据分布特征及箱型图方法

一般情况下，对于离散程度并非非常大的数据源来说，数据自身分布将会集中在某一区域之内，所以利用数据自身分布特征来识别离群值，可采用直观的箱型图方法可视化识别离群值及异常值。

● 基于欧几里德距离的聚类方法

一般情况下，利用数据分布特征或业务理解来识别单维数据集中噪声数据是快捷有效，但对于聚合程度高，彼此相关的多维数据而言，通过数据分布特征或业务理解来识别离群值的方法会缺乏有效性。面对这种情况，聚类方法提供识别多维数据集中噪声数据的方法。

识别离群值后，操作人员需要按照经验和业务流程判断其值的合理性：

——若此数值合理，则保留该数值；

——若不合理，则按照其重要性考虑是否需要重新采集。对于重要性较高而又无法重新采集的数值，按照缺失值办法处理。对于重要性较低数值，可直接去除。

3）修正矛盾内容

有些字段可以互相验证。需要根据字段的数据来源，来判定哪个字段提供的信息更可靠，去除或重构不可靠字段。

逻辑错误除以上列举情况，还有很多其他情况，在实际操作中需根据实际情况处理。

5.2.2.3　错误关联数据处理方法

如果数据有多个来源，有必要进行关联性验证。

多个来源的数据整合具有复杂性，要注意数据之间的关联性，尽量在分析过程中避免出现数据之间互相矛盾。

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。不一致数据往往是因为缺乏数据标准或未依照已有标准执行而产生。错误关联数据清洗方法主要有以下方法：

a）统计学方法：将属性当做随机变量，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常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，务必作好数据备份工作。对于明确为非需要字段，可以从数据集中删除。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 数据清洗质量评估要求

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。数据质量评价指标和方法研究的重点在于数据的含义、内容、分类、分级、质量的评价指标等的研究分析。

6.2 数据清洗质量评估指标

数据清洗质量评价可以归纳包含以下 12 个维度的基本评估指标：

a）数据规范(Data specification)：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；

b）数据完整性准则(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；

c）重复(Duplication)：对存在于系统内或系统间的特定字段、记录或数据集意外重复的测量标准；

d）准确性(Accuracy)：对数据内容正确性进行测量的标准；

e）一致性和同步(Consistency and synchronization)：对各种不同的数据仓库、应用和系统中所存储或使用的信息等价程度的测量，以及使数据等价处理流程的测量标准；

f）及时性和可用性(Timeliness and availability)：在预期时段内数据对特定应用的及时程度和可用程度的测量标准；

g）易用性和可维护性(Ease of use and maintainability)：对

数据可被访问和使用的程度，以及数据能被更新、维护和管理程度的测量标准；

h）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；

i）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；

j）可理解性、相关性和可信度(Perception，relevance and trust)：数据质量的可理解性和数据质量中执行度的测量标准，以及对业务所需数据的重要性、实用性及相关性的测量标准；

k）数据衰变(Data decay)：对数据负面变化率的测量标准；

l）效用性(Transactability)：数据产生期望业务交易或结果程度的测量标准。

在评估项目数据质量过程中，需要首先选取几个合适的数据质量维度，再针对每个所选维度，制定评估方案，选择合适的评估手段进行测量，最后合并和分析所有质量评估结果。

7　数据清洗过程管理

7.1　数据清洗角色定义

数据清洗管理涉及的数据管理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系统的基本运行管理、数据清洗规则制定、数据清洗发起等。项目中提供者为数据接入方，管理者为项目建设方。具体针对特殊情况有所变化。

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；提供者应接收数据清洗系统的问题数据，及时修改，并通知管理者。

7.3 管理者管理要求

管理者对数据清洗系统的管理要点应包括:管理者应负责协调并明确数据清洗规则;管理者应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；管理者应负责将问题数据库提交给提供者，并协调提供者修改完善。

7.4 数据审核管理要求

数据审核的目标是确保数据内容与被描述对象相一致，并且质量符合数据产品标准要求。

数据审核可以贯穿于整个数据资源加工过程之中，可以量化评价的内容包括数据来源质量评价、数据加工模型与算法质量评价、数据产品质量评价等。

数据审核可以由数据采集加工人员自检，也可由数据库主要承建单位专门开展。适宜时，数据审核宜采取计算机辅助方法进行。

数据库主要承建单位应明确审核所参照的评估模型和方法以及技术要求等。如果学科领域内已存在相关的数据质量管理国际、国家规范或行业标准，数据审核宜采用这些相关标准。

审核指标的设置应在符合实际的前提下尽可能不应与当前国际

领先水平有太大差距。审核指标可以包括但不限于准确性，真实性误差等技术参数，特色数据和重点数据宜适当提高指标。

数据资源审核通过后方可正式对用户提供服务，未能通过审核的数据一般应返回到必要的流程进行修正或重新加工。

7.5 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。

7.6 数据矫正处理要求

在数据阶段化过程中解决问题，对于那些同意纠正的数据，应当由原始数据提供者和管理者一起制定正确的规则，在数据接入过程中清洗。唯一正确的结果是纠正原始加载的数据并且用当前的数据校正历史数据。原始数据提供者应定期对数据源系统进行检查和清洗。

7.7 数据清洗服务管理要求

对数据清洗服务器的各项操作进行严格管理。提出以下要求：

a）不得随意修改操作系统和数据库系统的用户名及密码，不得随意增加操作系统和数据库用户。

b）不得随意安装和卸载各类软件；不得随意删除任何文件；不得对文件进行更名或移动。

c）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。

d）不得随意修改数据库系统的数据结构，包括增删或修改、字段、存储过程、触发器等。

e）除了数据中心相关系统应用接口程序外，其他任何程序都不得对数据库进行访问操作。

附表

业务数据转换规则示例表

转化规则	规则描述
统一时间日期数据格式	将各类日期统一为八位的字符日期，如 YYYYMMDD
	将各类时间统一为六位的字符时间，如 HHMMSS
	将各类事件日期统一为十四位的字符时间日期，如 YYYYMMDDHHMMSS
统一分类数据取值代码	将人员的性别数据统一转换为国标性别信息代码
	将人员的民族数据统一转换为标准信息代码
	将人员的户籍地址数据统一转换为行政区划代码
	将人员的婚姻登记情况统一转换为标准的婚姻状况代码
	将公民身份证号统一转换为 18 位