ICS 35.240.70
L 70

中华人民共和国国家标准

GB/T 38667—2020

信息技术　大数据　数据分类指南

Information technology—Big data—Guide for data classification

2020-04-28 发布　　2020-11-01 实施

国家市场监督管理总局
国家标准化管理委员会　发布

库七七 www.kqqw.com 提供下载

目　　次

库七七 www.kqqw.com 提供下载

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

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这些专利的责任。

本标准由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本标准起草单位:中国科学院信息工程研究所(信息安全国家重点实验室)、国家信息中心、浪潮软件集团有限公司、智慧神州(北京)科技有限公司、方正国际软件(北京)有限公司、国网安徽省电力有限公司(电力科学研究院)、中国铁道科学研究院集团有限公司、中国电子技术标准化研究院、上海三零卫士信息安全有限公司、联通大数据有限公司、中国保险信息技术管理有限责任公司、九次方大数据信息集团有限公司、中电长城网际系统应用有限公司、广东电网有限责任公司信息中心、中电科大数据研究院有限公司、北京大学、山东省计算中心(国家超级计算济南中心)。

本标准主要起草人:陈驰、马红霞、马书南、田雪、高亚楠、黄先芝、单震、张慧敏、张煜、顾广宇、吴艳华、郑金子、尹卓、叶林、干露、关泰璐、李燕超、郎佩佩、闵京华、魏理豪、禄凯、张吉才、冯念慈、赵俊峰、史丛丛、孙嘉阳。

库七七 www.kqqw.com 提供下载

信息技术　大数据　数据分类指南

1　范围

本标准提供了大数据分类过程及其分类视角、分类维度和分类方法等方面的建议和指导。

本标准适用于指导大数据分类。

2　规范性引用文件

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

GB/T 4754—2017　国民经济行业分类

GB/T 35295—2017　信息技术　大数据　术语

3　术语和定义

GB/T 35295—2017 界定的以及下列术语和定义适用于本文件。为了便于使用，以下重复列出了GB/T 35295—2017 中的某些术语和定义。

3.1

大数据　big data

具有体量巨大、来源多样、生成极快、且多变等特征，并且难以用传统数据体系结构有效处理的包含大量数据集的数据。

注：国际上，大数据的 4 个特征普遍不加修饰地直接用 volume、variety、velocity 和 variability 予以表述，并分别赋予了它们在大数据语境下的定义：

a)　体量　volume：构成大数据的数据集的规模。

b)　多样性　variety：数据可能来自多个数据仓库、数据领域或多种数据类型。

c)　速度　velocity：单位时间的数据流量。

d)　多变性　variability：大数据其他特征，即体量、速度和多样性等特征都处于多变状态。

[GB/T 35295—2017，定义 2.1.1]

3.2

数据集　data set

数据记录汇聚的数据形式。

注：它可以具有大数据的体量、速度、多样性和易变性特征。数据集的特征表征的是数据本身或静态数据，而数据的特征，当其在网络上传输时或暂时驻留于计算机存储器中以备读出或更新时，表征的是动态数据。

[GB/T 35295—2017，定义 2.1.46]

3.3

大数据分类　big data classification

根据大数据的属性或特征，将其按一定的原则和方法进行区分和归类，并建立起一定的分类体系和排列顺序的过程。

库七七 www.kqqw.com 提供下载

3.4

分类主体 classification subject

大数据收集、存储、使用、分发、删除等过程中对大数据进行梳理归类的组织或个人。

3.5

分类视角 classification angle

分类主体观察和开展大数据分类活动的角度。

3.6

分类维度 classification dimension

用于实现分类的数据所具有的某个或某些共同特征。

注：常见数据分类维度包括产生来源、结构化特征、业务归属、处理时效性要求等。

3.7

分类方法 classification method

根据选定的分类维度，将数据类别以某种形式进行排列组织的逻辑方法。

3.8

数据分发 data distribute

将原始数据、处理数据、分析结果等形式的数据传递给内部或外部实体的过程。

注：数据分发包括线上或线下等多种方式，如数据交换、数据交易、数据共享、数据公开等。

3.9

类别 category

具有共同属性(或特征)的数据的集合。

4 缩略语

下列缩略语适用本文件。

ETL：提取、转换和加载(Extract-Transform-Load)

FTP：文件传输协议(File Transfer Protocol)

SQL：结构化查询语言(Structured Query Language)

5 分类过程

5.1 概述

大数据分类过程划分为分类规划、分类准备、分类实施、结果评估、维护改进5个阶段，如图1所示。

库七七 www.kqqw.com 提供下载

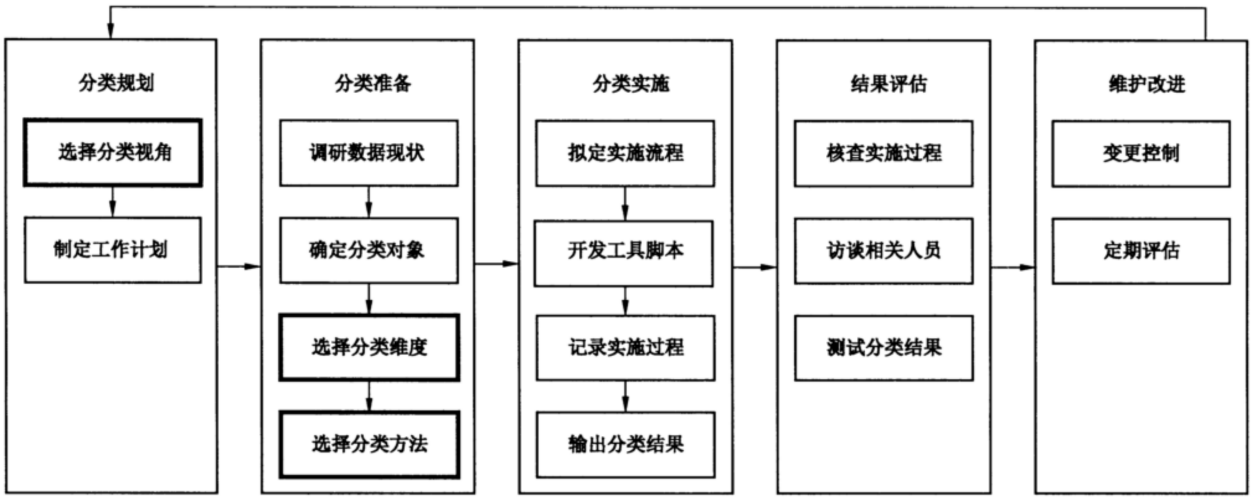

图 1　大数据分类过程

本章规范了大数据的分类过程，并根据大数据实际应用场景，在第 6 章、第 7 章、第 8 章分别对分类视角、分类维度、分类方法 3 个关键步骤进行规范，具体分类示例参见附录 A。

5.2　分类规划

5.2.1　选择分类视角

选择分类视角过程包括：

a）　明确分类业务场景；

b）　根据业务场景选取分类视角。

注：分类视角见第 6 章。

5.2.2　制定工作计划

制定工作计划过程包括：

a）　明确规划拟开展分类的数据范围；

b）　明确拟采用的分类维度和方法；

c）　明确预期分类结果；

d）　明确分类工作实施方案及进度安排；

e）　明确对分类结果的评估方法；

f）　明确对分类结果体系的维护方案。

5.3　分类准备

5.3.1　调研数据现状

调研数据现状过程包括：

a）　调研数据产生情况，包括但不限于数据产生的场景、主体、方式、频率、稀疏稠密、合法合规性等；

b）　调研数据存储现状，包括但不限于数据内容的格式、存储方式、存储位置、存储量等；

c）　调研数据质量情况，包括但不限于数据的规范性、完整性、准确性、一致性、时效性、可访问性等；

d）　调研数据业务类型，如组织人事管理数据、经营数据、财务数据等；

库七七 www.kqqw.com 提供下载

e) 调研数据敏感程度，包括但不限于数据的涉密程度、安全性、保护需求等；
f) 调研数据应用情况，包括但不限于数据的使用目的、应用领域、使用方式等；
g) 调研数据时效性情况，包括但不限于数据处理的时效性要求、数据价值时效性等；
h) 调研数据权属情况，包括但不限于数据的所有权、管理权、使用权等。

5.3.2 确定分类对象

确定分类对象过程包括：
a) 确定数据分类的业务场景；
b) 确定数据产生的起止时间；
c) 确定数据量大小；
d) 确定数据产生频率；
e) 确定数据结构化特征；
f) 确定数据存储方式；
g) 确定数据处理时效性；
h) 确定数据交换方式；
i) 确定数据产生来源；
j) 确定数据流通类型；
k) 确定数据质量；
l) 确定数据敏感程度。

5.3.3 选择分类维度

选择分类维度过程包括：
a) 梳理分类视角的数据特征；
b) 根据数据特征选取分类维度。

注：分类维度见第7章。

5.3.4 选择分类方法

选择分类方法过程宜明确分类维度的排列顺序和组合方式。

注1：分类方法见第8章。

注2：若选择混合分类法，还需考虑以哪种分类维度为主，哪种分类维度作为补充。

5.4 分类实施

5.4.1 拟定实施流程

拟定实施流程宜结合大数据的生命周期，拟定具体的分类实施流程，包括但不限于明确实施步骤、启动实施工作、开展实施工作、总结实施过程等。

5.4.2 开发工具脚本

开发工具/脚本宜根据实施流程、分类维度和分类方法编写分类算法，遵循软件开发或者脚本编制的规范开发分类工具/脚本。

5.4.3 记录实施过程

记录实施过程宜记录分类实施过程的各个步骤及其分类结果，输出文档。

库七七 www.kqqw.com 提供下载

5.4.4 输出分类结果

输出分类结果宜梳理各个步骤的分类结果，形成数据分类表。

5.5 结果评估

5.5.1 核查实施过程

核查实施过程包括：

a) 核查数据分类表，明确类别划分是否合理；
b) 核查分类过程记录，明确分类结果与预期目标的偏离程度；
c) 核查分类维度，确保分类维度符合业务需求、分类目标；
d) 核查分类方法的合理性；
e) 根据核查结果调整大数据分类过程。

5.5.2 访谈相关人员

访谈相关人员包括：

a) 访谈数据分类执行者，询问分类视角、范围、维度、方法与业务场景的关联性等；
b) 访谈数据所有者，询问数据分类结果中的数据权属类别划分、产生频率类别划分等是否符合实际情况；
c) 访谈数据管理者，询问数据分类结果中的数据结构化类别划分、数据存储方式类别划分、稀疏程度划分、敏感程度划分等是否符合实际情况；
d) 访谈数据使用者，询问数据分类结果中的数据处理实时性划分、交换方式类别划分、业务归属类别划分、流通类型类别划分等是否符合实际应用情况；
e) 核查意见和问题，调整大数据分类过程。

5.5.3 测试分类结果

测试分类结果包括：

a) 对分类后的数据执行分类脚本或程序，查看是否有不符合分类策略的分类结果；
b) 核查意见和问题，调整大数据分类过程。

5.6 维护改进

5.6.1 变更控制

变更控制包括：

a) 分析变更的必要性和合理性，确定是否实施变更；
b) 制定变更计划，评估变更对大数据分类工作的影响，包括分类维度、分类方法的改变等；
c) 执行变更，对分类结果进行更改，记录变更过程；
d) 对新的大数据分类结果进行评估；
e) 发布新的大数据分类结果。

5.6.2 定期评估

定期评估包括：

a) 定期评估大数据分类维度和方法的合理性，检查其是否符合业务场景变化和分类视角变化；

库七七 www.kqqw.com 提供下载

b) 定期评估大数据分类结果的有效性和应用情况，检查其是否满足业务应用需求的更新；
c) 核查意见和问题，调整大数据分类过程。

6 分类视角

6.1 概述

大数据分类视角分为技术选型视角、业务应用视角和安全隐私保护视角。

6.2 技术选型视角

技术选型视角包括但不限于：

a) 理清数据产生频率，明确数据产生规律，确定数据更新周期和存储策略，确定数据存储平台配型等存储资源分配方案；
b) 理清数据产生方式，分析数据的来源和质量，确定在整个数据处理流程中数据所处的位置，及数据处理及存储技术；
c) 分析数据的结构化特征，确定数据存储与处理方案；
d) 明确数据的存储方式，确定数据建模模型与数据的访问方式，支撑各类数据应用场景；
e) 理清数据稀疏稠密程度，明确数据稀疏稠密规律，确定数据存储策略和分析方法，选择数据存储方案和分析方案；
f) 明确数据处理时效性要求，明确数据处理时机，确定数据处理策略，选择包括计算平台和资源匹配等的数据处理方案；
g) 理清数据交换方式，确定数据共享方式及策略，支撑构建信息交换体系。

6.3 业务应用视角

业务应用视角包括但不限于：

a) 理清数据产生来源，明确数据权属和访问权限，便于数据追踪溯源；
b) 明确数据应用场景，确定数据业务主题，判断数据应用价值，选择数据分析方案；
c) 明确数据分发场景，确定数据应用行业，明确可用数据的种类和范围；
d) 理清数据质量情况，明确数据应用需求，确定数据质量管理方案。

6.4 安全隐私保护视角

安全隐私保护视角包括但不限于：

a) 明确不同敏感程度的大数据在存储、传输、访问、分发时的安全要求；
b) 明确不同敏感程度的大数据的隐私保护要求；
c) 指导分类主体制定隐私保护方案；
d) 指导分类主体制定安全管理方案。

7 分类维度

7.1 概述

本章从技术选型、业务应用和安全隐私保护三种视角给出不同的分类维度，以及用于描述每种分类维度的分类要素、数据类别和适用场景。

库七七 www.kqqw.com 提供下载

7.2 技术选型维度

7.2.1 按产生频率分类

7.2.1.1 概述

按产生频率分类是指根据数据产生的频率(单位时间内产生的数据量或达到指定数据量的频率)对数据进行分类。

7.2.1.2 分类要素

按产生频率分类的要素包括:

a) 数据产生周期,如秒、分、时、天、周、月、季度、半年、年等;

b) 单位周期中数据的产生量,可以以记录条数表示或者以数据占用空间表示,如百万条记录、千万条记录、GB级数据、TB级数据等。

7.2.1.3 类别

按产生频率可分为:每年更新数据、每月更新数据、每周更新数据、每日更新数据、每小时更新数据、每分钟更新数据、每秒更新数据、无更新数据等。

7.2.1.4 适用场景

按产生频率分类的适用场景,如根据数据产生频率判断资源分配合理性和数据分析价值等。

7.2.2 按产生方式分类

7.2.2.1 概述

按产生方式分类是指按照数据的产生方式对数据进行分类。

7.2.2.2 分类要素

按产生方式分类的要素包括:

a) 数据被获取或被采集的方式,如人工采集、通过信息系统采集等;

b) 数据被加工的程度,如原始数据、二次加工数据等。

7.2.2.3 类别

按产生方式分类可包括:人工采集数据、信息系统产生数据、感知设备产生数据、原始数据、二次加工数据等。

7.2.2.4 适用场景

按产生方式分类的适用场景,如确定数据采集方案、数据保护方案和数据处理方案等。

7.2.3 按结构化特征分类

7.2.3.1 概述

按结构化特征分类是指根据数据的结构化程度对数据进行分类。

库七七 www.kqqw.com 提供下载

7.2.3.2 分类要素

按结构化特征分类的要素包括：

a) 是否有预定义的数据模型；

b) 数据结构是否规则；

c) 数据长度是否规范；

d) 数据类型是否固定。

7.2.3.3 类别

按结构化特征分类可划分为：结构化数据，如零售、财务、生物信息学、地理数据等；非结构化数据，如图像、视频、传感器数据、网页等；半结构化数据，如应用系统日志、电子邮件等。

7.2.3.4 适用场景

按结构化特征分类的适用场景，如根据数据结构规划数据处理和存储架构。

7.2.4 按存储方式分类

7.2.4.1 概述

按存储方式分类是指根据数据适合采用的数据存储方式对数据进行分类等。

7.2.4.2 分类要素

按存储方式分类的要素包括：

a) 数据建模适合采用的数据模型，如关系模型、文档模型、图模型等；

b) 数据访问使用的查询语言，如 SQL、类 SQL、图查询语言等。

7.2.4.3 类别

按存储方式可划分为：关系数据库存储数据、键值数据库存储数据、列式数据库存储数据、图数据库存储数据、文档数据库存储数据等。

7.2.4.4 适用场景

按存储方式分类的适用场景，如选择数据存储采用的数据库系统、确定应用系统与数据存储系统之间的数据访问方式等。

7.2.5 按稀疏程度分类

7.2.5.1 概述

按稀疏程度分类是指根据数据的稀疏稠密程度对数据进行分类。

7.2.5.2 分类要素

按稀疏程度分类的要素主要包括数据稀疏程度评价标准，即数据集中数值缺失或者为零的数据所占比例。如空值或零值小于 50%的数据为稠密数据，空值或零值大于或等于 50%的数据为稀疏数据。

7.2.5.3 类别

按稀疏程度可划分为：稠密数据和稀疏数据。

库七七 www.kqqw.com 提供下载

7.2.5.4 适用场景

按稀疏程度分类的适用场景，如根据单位时间内数据的量级进行数据价值密度分析判断等。

7.2.6 按处理时效性分类

7.2.6.1 概述

按处理时效性分类是指根据数据处理的时间延迟要求对数据进行分类。

7.2.6.2 分类要素

按处理时效性分类的要素包括：

a) 数据处理延迟时间要求，即应用场景是否对处理延迟时间有明确的上限要求；
b) 数据价值时效性，即数据应用价值随时间推移的有效性；
c) 数据处理量，即延迟上限时间内需处理多少量级的数据。

7.2.6.3 类别

按处理时效性可划分为：实时处理数据、准实时处理数据和批量处理数据。

7.2.6.4 适用场景

按处理时效性分类的适用场景，如根据数据时效要求安排业务顺序和资源投入等。

7.2.7 按交换方式分类

7.2.7.1 概述

按交换方式分类是指根据数据在提供方和接收方之间交换的方式对数据进行分类。

7.2.7.2 分类要素

按交换方式分类的要素包括：

a) 数据交换双方之间的网络状况，即交换双方之间的网络是否互通；
b) 数据在交换双方之间的同步实时性要求；
c) 单次交换的数据量；
d) 数据交换的频次，如固定频率交换、固定时间交换或按需交换等。

7.2.7.3 类别

按交换方式可划分为：ETL 方式、系统接口方式、FTP 方式、移动介质复制方式等。

7.2.7.4 适用场景

按交换方式分类的适用场景，如根据不同交换方式对大数据共享便利程度的影响，规划信息交换系统架构等。

7.3 业务应用维度

7.3.1 按产生来源分类

7.3.1.1 概述

按产生来源分类是指根据数据产生的实际情景对数据进行分类。

库七七 www.kqqw.com 提供下载

7.3.1.2 分类要素

按产生来源分类的要素包括：

a) 数据产生主体，如人工、机器、传感器、应用软件、信息系统等；

b) 数据权属，即数据所有权的归属。

7.3.1.3 类别

按产生来源可划分为：人为社交数据、电子商务平台交易数据、移动通信数据、物联网感知数据、系统运行日志数据等。

7.3.1.4 适用场景

按产生来源分类的适用场景，如根据数据来源确定数据归集策略、预测服务提供和数据交易定价等。

7.3.2 按业务归属分类

7.3.2.1 概述

按业务归属分类是指根据数据所属的业务类型对数据进行分类。

7.3.2.2 分类要素

按业务归属分类的要素包括：

a) 分类主体的业务类型划分，如生产类业务、管理类业务、经营分析类业务；

b) 生成数据的业务所属的职能，如产品研发、市场营销、财务管理、人力管理等；

c) 生产数据的具体业务，如商品交易、会员注册、人才招聘等。

7.3.2.3 类别

按业务归属可划分为：生产类业务数据、管理类业务数据、经营分析类业务数据等。

7.3.2.4 适用场景

按业务归属分类的适用场景，如按业务属性评价数据应用价值等。

7.3.3 按流通类型分类

7.3.3.1 概述

按流通类型分类是指根据数据在流通交易过程中的交易类型进行分类。

7.3.3.2 分类要素

按流通类型分类的要素包括：

a) 数据权责，即数据需求方可获取的数据权益，如所有权、经销权、使用权、可复制权等；

b) 计费方式，即数据供应方和数据需求方之间计算数据交易费用的方式，如按使用量计费、按使用时长计费等；

c) 交付内容，即数据供应方向数据需求方提供的数据内容，如原始数据集、数据分析报告等；

d) 行业主题，即流通数据所属的行业领域，如农业、林业、医疗、交通、科研等；

e) 敏感程度，即流通数据是否涉及国家秘密、行业秘密、企业秘密或个人隐私等，如公开数据、脱

库七七 www.kqqw.com 提供下载

敏数据、涉密数据等。

7.3.3.3 类别

按流通类型可划分为：可直接交易数据、间接交易数据、不可交易数据等。

7.3.3.4 适用场景

按流通类型分类的适用场景，如以大数据分析和大数据交易为经营内容的企业进行产品规划等。

7.3.4 按行业领域分类

7.3.4.1 概述

按行业领域分类是指根据数据内容所属的行业领域范畴对数据进行分类。

7.3.4.2 分类要素

按行业领域分类的要素包括：

a) 数据产生行业，即产生数据的活动所属的国民经济行业；
b) 数据应用行业，即分析和使用数据的活动所属的国民经济行业。

7.3.4.3 类别

按行业领域分类可划分的类别见 GB/T 4754—2017。

7.3.4.4 适用场景

按行业领域分类的适用场景，如公安、气象、水文等行业大数据分析等。

7.3.5 按数据质量分类

7.3.5.1 概述

按数据质量分类是指根据数据的质量差异对数据进行分类。

7.3.5.2 分类要素

按数据质量分类的要素包括：

a) 数据的准确性，即数据是否存在异常、错误或过时；
b) 数据的完整性，即数据是否存在缺失及缺失程度；
c) 数据的一致性，即数据内容是否遵循统一规范；
d) 数据的及时性，即所需数据是否及时到达目标应用；
e) 数据的重复性，即是否存在大量重复数据。

7.3.5.3 类别

按数据质量可划分为：高质量数据、普通质量数据、低质量数据等。

7.3.5.4 适用场景

按数据质量分类的适用场景，如根据不同数据质量的比例确定数据利用的价值和数据质量管理工作难易程度等。

库七七 www.kqqw.com 提供下载

7.4 安全隐私保护维度

7.4.1 概述

按数据安全隐私保护维度分类是根据数据内容敏感程度对数据进行分类。

7.4.2 分类要素

按安全隐私保护维度分类的要素包括：

a) 数据的敏感性，即数据本身或其衍生数据是否涉及国家秘密、企业秘密或个人隐私；

b) 数据的保密性，即数据可被知悉的范围；

c) 数据的重要性，即数据未经授权披露、丢失、滥用、篡改或销毁后对国家安全、企业利益或公民权益的危害程度。

7.4.3 类别

按数据安全隐私保护维度可划分为：高敏感数据、低敏感数据、不敏感数据等。

7.4.4 适用场景

按安全隐私保护维度分类的适用场景，如根据数据内容敏感程度确定大数据应用边界、数据保护策略、数据脱敏方案等。

8 分类方法

8.1 线分类法

8.1.1 概述

线分类法旨在将分类对象(即本标准界定的数据)按选定的若干个属性或特征，逐次分为若干层级，每个层级又分为若干类别。同一分支的同层级类别之间构成并列关系，不同层级类别之间构成隶属关系。同层级类别互不重复，互不交叉。

线分类法适用于针对一个类别只选取单一分类维度进行分类的场景。

8.1.2 确定分类类别之间关系

采用线分类法确定分类类别之间关系的过程包括：

a) 确定一个分类维度；

b) 确定该分类维度的分类类别；

c) 针对每一个分类类别：如果该分类类别不需要再进一步划分子类，则转 d)步，否则确定该分类类别进行子类划分的分类维度，转 b)步；

d) 所有分类类别均不需进一步划分，则分类类别之间关系确定。

注：上述过程完成后，将形成一棵分类类别关系树。树的叶节点为最终的分类项，通常称为基本类别；其余节点为中间类别。

8.1.3 特点

线分类法的特点包括：

a) 层次性好，能较好地反映类别之间的逻辑关系；

b) 实用方便，便于机器处理信息；

库七七 www.kqqw.com 提供下载

c) 结构弹性较差，分类结构一经确定，不易改动；
d) 效率较低，当分类层次较多时，影响数据处理速度。

8.2 面分类法

8.2.1 概述

面分类法是将所选定的分类对象（即本标准界定的数据），依据其本身的固有的各种属性或特征，分成相互之间没有隶属关系即彼此独立的面，每个面中都包含了一组类别。将某个面中的一种类别和另外的一个或多个面的一种类别组合在一起，可以组成一个复合类别。

面分类法是并行化分类方式，同一层级可有多个分类维度。面分类法适用于对一个类别同时选取多个分类维度进行分类的场景。

8.2.2 确定分类类别之间关系

采用面分类法确定分类类别之间关系的过程包括：

a) 确定分类对象的若干个特征面，即分类维度，每一个分类维度构成一个分类面。
b) 确定分类面的排列顺序，应当按照分类维度的重要性或使用频率的高低由左向右进行排列。
c) 划分每一个分类维度的分类类别。为每一个分类维度确定一个分类规则，并按此规则划分各个分类维度的分类类别。
d) 通过上述步骤所得到的各个面的类别将分类对象划分成了若干个对象类。

8.2.3 特点

面分类法的特点包括：

a) 弹性较大，一个“面”内类别的改变，不会影响其他的“面”；
b) 适应性强，可根据需要组成任何类别；
c) 易于添加和修改类别；
d) 可组配的类别很多，但实际应用的类别不多。

8.3 混合分类法

8.3.1 概述

混合分类法是将线分类法和面分类法组合使用，克服这两种基本方法的不足，得到更为合理的分类。混合分类法的特点是以其中一种分类方法为主，另一种做补充。混合分类法适用于以一个分类维度划分大类、另一个分类维度划分小类的场景。

8.3.2 特点

混合分类法的优点包括：

a) 可以根据实际需要，对两种分类方法进行灵活的配置，吸取两种分类方法的优点；
b) 适应一些综合性较强、属性或者特征不是十分明确的数据分类。

库七七 www.kqqw.com 提供下载

附 录 A
（资料性附录）
大数据分类示例

A.1 业务场景和分类视角

根据本标准中提出的分类过程、分类视角、分类维度和分类方法，以铁路大数据为例，进行大数据分类实践和验证。

铁路大数据涵盖铁路勘测设计、建设和运营等各阶段，在铁路数据目录梳理、铁路数据交换共享、铁路数据建模分析、铁路数据安全保护等铁路大数据管理场景下均需对铁路大数据进行分类。

对铁路大数据进行分类的视角是规范国铁集团、铁路局两级数据管理的相关标准，同时规范铁路数据与外部数据交换共享的类型等。

A.2 分类范围、分类维度和分类方法

铁路大数据分类范围包括由铁路客运、物流、基础设施、移动设备、工程建设、资产经营、企业管理等各铁路业务领域的结构化、非结构化数据所汇集而成的数据集合。

分类维度选择按结构化特征分类、按产生来源分类、按产生频率分类、按业务归属分类。

分类方法采用以线分类法为主、面分类法为辅的混合分类法。

A.3 分类实施和分类结果

在进行铁路大数据分类实施时，考虑到铁路大数据的多源性和异构性等特点，首先，采用线分类法，选择按结构化特征、按业务归属、按产生来源和按产生频率等维度对铁路大数据进行大类划分；其次，针对具体的某一大类数据，采用面分类法，选择按产生来源、使用标记等维度进行小类划分。

具体分类过程如下：

a) 第一级分类：

1) 按结构化特征分类，将铁路大数据分为结构化数据和非结构化数据两大类。

b) 第二级分类：

1) 针对结构化数据，按业务归属分类，分为主数据、事务数据和分析数据；

2) 针对非结构化数据，按产生来源分类，分为文本数据和多媒体数据。

c) 第三级分类：

1) 针对事务数据，按产生频率分类，分为实时数据和非实时数据；

2) 针对文本数据，按业务归属分类，分为法律数据、制度数据、办公数据、事务数据。

d) 第四级分类：

1) 针对第三级分类结果和部分第二级分类结果，进一步按业务归属分类，形成第四级分类。

分类结果如图 A.1 所示。

库七七 www.kqqw.com 提供下载

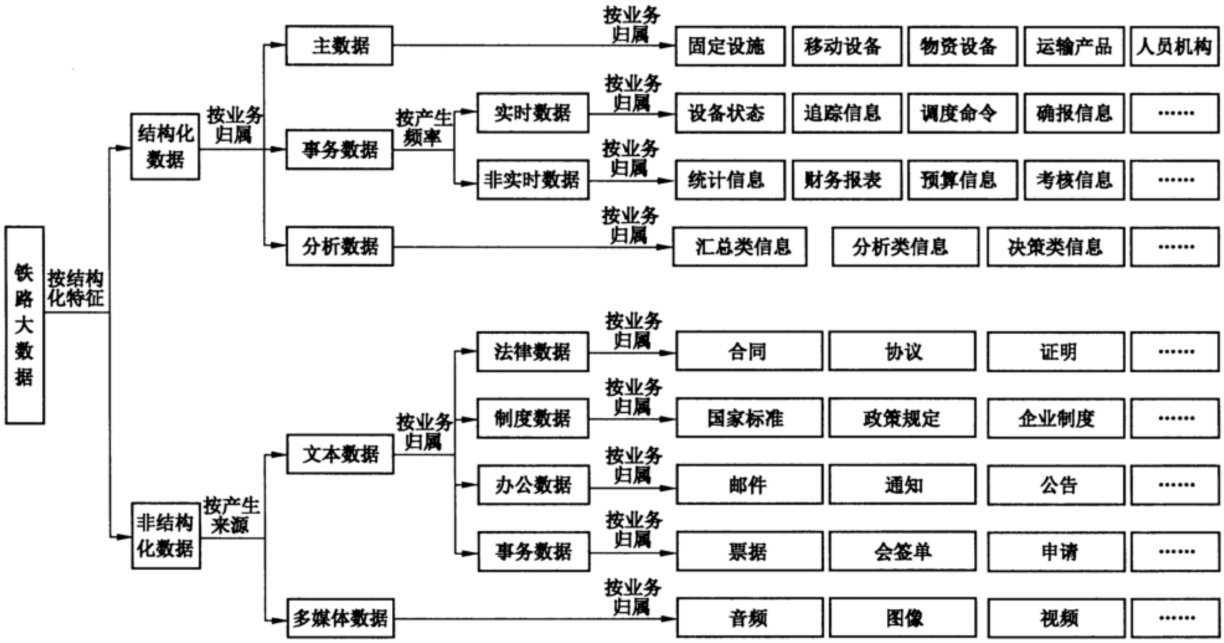

图 A.1　铁路大数据线分类方法示例

经过四级线分类法已将铁路大数据划分到具体业务层面，而根据实际应用需求，需采用面分类法将数据进行进一步地细分。主数据中的固定设施类数据按业务归属分类（见图 A.2 实线箭头），可分为车站主数据和专用线主数据，以专用线主数据为例，介绍面分类过程。

针对专用线主数据，可分别按产生来源和使用标记这两个“面”进行分类，如图 A.2 虚线箭头所示：

a）　按产生来源分类，即根据产生数据的专用线对数据进行分类，分类实例如客运专用线主数据、货运专用线主数据等；

b）　按使用标记分类，即根据数据使用标记对数据进行分类，分类实例如 A 类主数据、B 类主数据、C 类主数据等。

固定设施类主数据的面分类结果如图 A.2 所示。

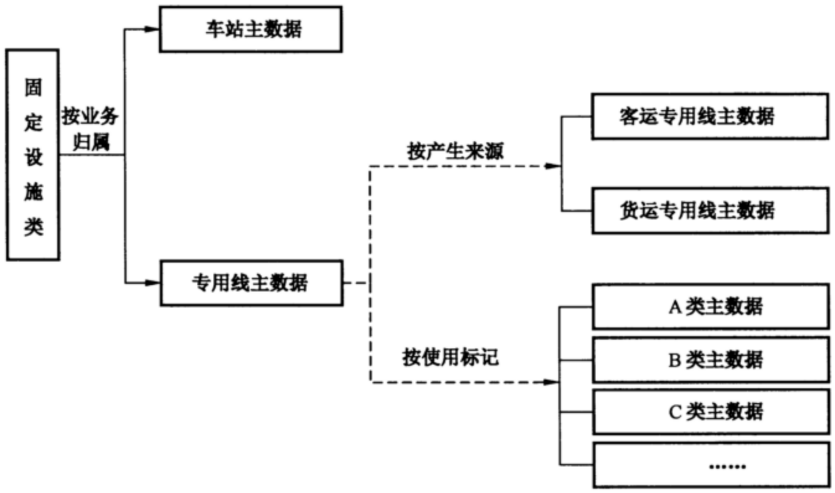

图 A.2　铁路大数据面分类示例

库七七 www.kqqw.com 提供下载