第 3 章

大数据采集与预处理

本章学习目标

- 了解大数据的来源和大数据采集的基本方法。
- 了解常用的大数据采集工具。
- 熟悉大数据预处理的基本方法。
- 培养兼收并蓄、去芜存菁的理念与价值观及精益求精的工匠精神。

互联网和物联网的快速发展与广泛应用产生了海量的数据，如物联网传感数据、电子商务交易数据、股票交易数据等。要对这些海量的数据进行处理，首先需要进行大数据采集与预处理，这是大数据分析至关重要的一个环节，也是大数据分析的入口。

面对海量的数据，大数据采集面临着巨大的挑战。一方面数据源的种类多，数据的类型繁杂，数据量大，并且数据产生的速度快；另一方面需要保证大数据采集的可靠性和高效性，同时还要避免采集重复的数据。

大数据采集与预处理是获取有效数据的重要途径，也是大数据应用的重要支撑。本章向读者介绍大数据采集与预处理的方法，主要包括大数据的来源和采集途径，常用的大数据采集工具，以及数据预处理的基本方法，包括数据清洗、数据集成、数据变换等方法。

3.1 认识数据

认识数据是处理和分析数据的前提，不仅要了解数据的属性（维）、类型和量纲，还要了解数据的分布特性、洞察数据的特征、检验数据的质量，以便进行后续的分析和处理工作。

3.1.1 数据的属性和类型

所谓“属性”是指数据对象的特征，也称为数据字段、维。iris 数据集的属性如图 3-1 所示，iris 数据集（鸢尾花数据集）有 4 个属性（特征），分别为 sepallength、sepalwidth、petallength 和 petalwidth，以及一个类别属性 class。其中 class 属性为标称数据（表示类别的标称数据），其他 4 个属性为定量数据（数值型数据）。在数据分析领域，数据的属性也称为特征、自变量、解释变量或观测值，数据的标签也称为因变量（如 iris 数据集中的 class 属性）。

属性、特征、自变量、解释变量或观测值　　标签或因变量

sepallength	sepalwidth	petallength	petalwidth	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa

图 3-1　iris 数据集的属性

数据的基本属性可以分为两大类：定性属性和定量属性。

1．定性属性

具有定性属性的数据是表示事物性质、规定事物类别的文字表述型数据。

（1）标称数据：是一些符号或事物的名称，每个值代表某种类别、编码或状态，如性别分类中的“男”“女”，表示颜色的“红”“黄”“蓝”等。

（2）二元数据：只有两个值或状态，0 或 1，又被称为布尔属性（值为 true 或 false）。

（3）序值型数据：其可能的值之间具备有意义的序或秩评定，如成绩（“差”“中”“良”“优”）。

以上 3 种类型都是对数据对象特征的定性描述，而不给出实际大小或数量。

2．定量属性

定量属性是描述数据对象的数值大小，如长度、速度、半径等的数量值，定量属性有整数（离散）和浮点数（连续）两种形式。

实际上，在许多真实的数据集中，数据对象通常是混合类型的，一个数据对象的属性可能包含上面列举的多种类型。

3.1.2　数据的量纲

数据属性的值，有时是有单位的，称作量纲数据。所谓量纲，是指物理量的基本属性。例如，物理学中的 7 个基本量——长度、质量、时间、电流、热力学温度、物质的量和发光强度，其量纲分别用 L、M、T、I、Θ、N 和 J 表示。

有些数据是有量纲的，如身高、长度、时间、质量、速度；而有些数据是没有量纲的，如男女比例、两个长度之比。无量纲化，是指去除数据的单位限制，将其转换为无量纲的纯数值，便于不同单位或不同量级的指标进行比较和加权。

不同的评价指标往往具有不同的量纲，数据之间的差别可能很大，不进行处理会影响数据分析的结果。为了消除指标之间的量纲和取值范围差异对数据分析结果的影响，需要对数

据进行标准化处理，也就是说，把数据按照比例进行缩放，使之落入一个特定的区间，以便进行综合分析。例如，归一化处理是把数变为[0,1]之间的小数，把有量纲的数据转换为无量纲的纯数值。

3.2 大数据的来源和采集途径

3.2.1 大数据的来源

大数据的来源非常广泛，如互联网、物联网、信息管理系统、科学实验和计算机系统的日志等。按照产生数据的主体来划分，大数据主要有 3 个来源，它们分别如下。

1．对现实世界的测量

通过感知设备获得的数据，这类数据包括传感器采集的数据（如环境监测、工业物联网和智能交通的传感数据）、科学仪器产生的数据、摄像头的监控影像等。例如，在新冠疫情期间，许多商场都在门口安装了红外线体温检测仪，路过的行人在其视野范围内都会接受测量，测得的体温信息会被马上记录下来，这便是一种大数据的来源。

此类数据的特点是：数据的模式比较固定、数据的规模极大、数据的质量参差不齐、数据的价值密度低。

2．人类的记录

由人类录入计算机形成的数据，如信息管理系统、社交软件、电子商务系统、企业财务系统等产生的数据。例如，个人的电子邮件、Word 文件、照片、视频、音频、QQ 空间、微信朋友圈、社交软件的聊天记录等，以及电子商务系统记录的交易数据、信用卡的刷卡数据等。

此类数据的特点是：数据的模式多样、数据的规模较大、数据的质量参差不齐、数据的语义不明确、数据的价值密度低。

3．计算机产生的数据

由计算机程序生成的数据，如服务器的日志、计算机的运算结果、软件生成的图像和视频等。

此类数据的特点是：由程序自动生成，数据的模式固定、数据的质量高、数据的语义明确，数据的价值密度与产生数据的程序相关。

3.2.2 大数据的采集途径

数据采集是指从真实世界中获得原始数据的过程，它是大数据应用的基础环节，具有非常重要的意义。数据采集的手段通常有以下 5 种。

1．传感数据的采集

通过传感器采集物理世界的数据，如通过环境监测传感器、工业传感器采集热、光、气、

电、磁、声和力等数据，通过多媒体数据采集设备获取图像、音频和视频数据等。

2．系统日志的采集

主要采集数字设备和计算机系统运行状态的日志。许多数字设备和计算机系统每天都会产生大量的日志，一般为流式数据，如信息管理系统的操作日志、服务器的运行状态和搜索引擎的查询结果等。收集和处理这些日志通常需要专门的日志采集系统，如 Apache Flume、Apache Chukwa、Meta（原名为 Facebook）的 Scribe 和 LinkedIn 的 Kafka 等。

3．数据库数据的采集

许多企业和单位在信息化过程中结合自身业务建立了各种各样的数据库系统，积累了大量的数据。从传统的关系型数据库中采集数据，需要利用对接关系型数据库和大数据平台的数据采集引擎（如数据采集工具 Sqoop），从数据库中进行数据的采集和汇聚，并将数据传输到大数据平台上。

4．网络数据的采集

网络数据的采集通常可以分为网页内容的采集和网络流量的采集两种类型。

网页内容的采集是指通过网络爬虫或网站公开 API 等方式从网站上获取数据，该方法可以将非结构化的数据从网页中抽取出来，将其存储为统一的本地数据文件，并以结构化或非结构化的方式存储，图 3-2 所示为网络爬虫的工作原理示意图。

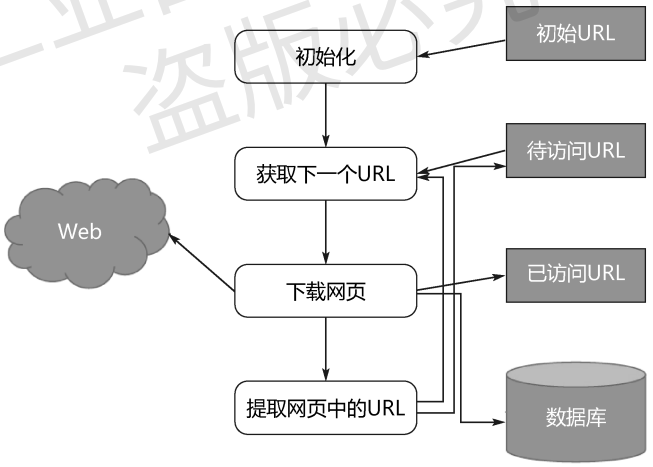


图 3-2　网络爬虫的工作原理示意图

网络流量的采集可以使用 DPI（Deep Packet Inspection，深度包检测）和 DFI（Deep/Dynamic Flow Inspection，深度/动态流检测）等带宽管理技术进行处理。DPI 技术在分析包头的基础上，增加了对应用层的分析，这是一种基于应用层的流量检测和控制技术。DFI 仅对网络流量的行为进行分析，因此只能对应用的类型进行笼统分类，但速度很快。

5．外包和众包

外包（Outsourcing）是指将所要进行的数据采集任务交付给已知的雇员去完成。

众包（Crowdsourcing）是指数据采集任务由一群不固定、通常数量很大的参与者共同协作完成，如 Wikipedia（维基百科）就是一个成功应用“众包”方法构建的庞大知识库。

3.3 常用的大数据采集工具

在数据规模不断扩大的情况下，运用数据采集自动化工具，从外部系统和互联网等自动获取、传输和记录数据已经成为必要的技术手段。

目前常用的大数据采集工具有爬山虎采集器、八爪鱼采集器、Scrapy、Flume、Chukwa、Kafka 和 Sqoop 等，如表 3-1 所示。

表 3-1 常用的大数据采集工具

名称	类型	功能和特点	研制者或项目来源
爬山虎采集器	网页爬虫	简单易用、无须编程、智能识别	合肥简数信息技术有限公司
八爪鱼采集器	网页爬虫	简单易用、无须编程、自定义模板	深圳视界信息技术有限公司
Scrapy	基于 Python 的网页抓取框架	具有高度可定制性的网页采集工具	开源 GitHub 项目
Flume	日志采集中间件	分布式海量日志采集、聚合和传输系统；数据源可定制、可扩展	Cloudera 公司开源的 Apache 项目
Chukwa	基于 Hadoop 集群的日志处理/分析	开源的用于监控大型分布式系统的数据收集系统	Apache Hadoop 项目的系列产品
Kafka	基于消息发布-订阅的流处理平台	分布式的消息发布-订阅系统，用于将数据流从一个应用程序传输到另一个应用程序	LinkedIn 公司开源的 Apache 项目
Sqoop	数据传输工具	在 Hadoop 和关系型数据库之间传递数据	Apache 开源软件

3.3.1 爬山虎采集器

爬山虎采集器是一款通用的网页采集软件，它能够采集互联网上的大部分网站数据，包括网页表格数据、文档、图片及其他各种形式的文件，自动批量下载到本地计算机。它不仅可以将采集到的数据导出成多种格式文件（如 TXT 文本、Excel 表格、CSV 文件）、数据库（如 MySQL、SQLite、Access 和 SQL Server）、网站 API，还可以定时运行、自动发布、增量更新采集数据，完全实现自动化运行，无须人工干预，极大地提高了人们从互联网上获取数据的效率。爬山虎采集器目前只提供 Windows 版本客户端软件，其主界面如图 3-3 所示。

爬山虎采集器内置了上百种简易的采集规则，能够对网页源码自动进行智能分析，用户只需要通过一些简单的参数（如关键词、网址）就可以开始采集。图 3-4 所示为作者使用自定义采集功能对京东网上的手机商品列表进行数据采集的截图，从中可以看出爬山虎采集器可以对该网页进行智能分析，获取商品名称、商品价格和详情链接等信息。

图 3-3　爬山虎采集器客户端软件的主界面

图 3-4　爬山虎采集器采集京东网上手机商品数据的示意图

选择需要采集的数据字段，并自定义字段名称（如商品名称、价格、商家），最后设置简单的浏览器参数就可以开始采集，图 3-5 所示为爬山虎采集器采集京东网上手机商品数据的结果截图。

运行日志 数据(300)

商品名称	价格
HUAWEI nova 10 【内置66W华为超级快充】 前置6000万超广角镜头 6.88mm轻薄机身 128GB 10号色 华为手机	2539.00
荣耀X40 120Hz OLED硬核曲屏 5100mAh 快充大电池 7.9mm轻薄设计 5G手机 8GB+128GB 彩云追月	1699.00
华为智选 Hi nova 9z 5G全网通手机 6.67英寸120Hz原彩屏 6400万像素超清摄影 66W快充8GB+128GB亮黑色	1299.00
OPPO K9x 8GB+128GB 银紫超梦 天玑810 5000mAh长续航 33W快充 90Hz电竞屏 6400万三摄 拍照5G手机oppok9x	1199.00
vivo iQOO Z5 8GB+256GB 蓝色起源 骁龙778G 5000mAh长续航 120Hz高刷原色屏 双模5G全网通手机iqooz5	1799.00
京品手机OPPO 一加 Ace 竞速版 12GB+256GB竞技灰享OPPO官方售后 天玑8100-MAX 120Hz变速电竞直屏游戏稳帧引擎5G手机	1899.00
小米12S Pro 骁龙8+处理器 徕卡光学镜头 2K超视感屏 120Hz高刷 120W秒充 12GB+256GB 黑色 5G手机	5099.00
华为mate40 5G手机 可选TD Tech M40 亮黑色8GB+256GB【TDTech M40】 官方标配【全网通】	2999.00
Redmi 10A 5000mAh大电量 1300万AI相机 八核处理器 指纹解锁 4GB+64GB 暗影黑 智能手机 小米 红米	699.00
vivo iQOO Neo7 12GB+256GB 几何黑 天玑9000+ 独显芯片Pro+ E5柔性直屏 120W超快闪充 5G全网通手机iqooneo7	2999.00

图 3-5　爬山虎采集器采集京东网上手机商品数据的结果截图

3.3.2　八爪鱼采集器

八爪鱼采集器是一个智能网页爬虫工具，其原理是模拟人的网页浏览行为，通过内置谷歌浏览器浏览网页数据。使用者根据网页特性和采集需求，设计采集流程，八爪鱼采集器会根据流程全自动采集数据。

八爪鱼采集器的客户端软件目前提供 Windows 和 Mac 两种版本。本书中下载八爪鱼 Windows 客户端软件。安装完成后，在“开始”菜单或桌面上找到八爪鱼采集器的快捷方式，启动八爪鱼采集器，进行注册和登录。八爪鱼客户端软件的界面示意图如图 3-6 所示。

图 3-6　八爪鱼客户端软件的界面示意图

1．通过模板采集数据

所谓的“采集模板”是由八爪鱼官方提供的、做好的采集模板，目前已有 200 多种采集模板，涵盖主流网站的采集场景。使用模板采集数据时，只需要输入几个参数（网址、关键词、页数等），就能在几分钟内快速获取目标网站的数据。（类似 PPT 模板，只需要修改关键

信息就能直接使用，无须自己从头配置。）

2．自定义配置采集数据

自定义配置采集数据比通过模板采集数据更为复杂，自定义配置采集数据有两种方法：智能识别方法和手动配置采集流程。

使用智能识别方法时，只需要输入网址，八爪鱼采集器就能自动地智能识别网页数据，并且支持自动识别列表型网页数据、滚动和翻页。

手动配置采集流程是指，自定义从特定网页上抓取数据的指令。由于每个网站的页面布局是不同的，因此采集流程不能通用。一般情况下，每个网站都需要配置一个采集流程。自己动手配置采集流程需要用户了解网页的结构和确定需要采集的网页字段，自定义采集的流程。该方法可灵活应对各类采集场景，包括翻页、滚动、登录等。

3.3.3　基于 Python 的网页抓取框架 Scrapy

Scrapy 是一个用 Python 语言开发的基于异步模型的网页抓取框架，用于抓取网站并从页面中提取结构化数据。它的用途广泛，可以用于数据挖掘、监控和自动化测试等。

Scrapy 的吸引人之处在于它是一个框架，任何人都可以根据需求修改，它也提供了多种类型的爬虫基类，如 BaseSpider、SiteMap 爬虫等，其最新版本还提供了对 Web 2.0 爬虫的支持。接下来介绍一个 Scrapy 的应用案例。

【案例 3-1】用 Scrapy 框架编写基于 Python 语言的爬虫程序，用于采集豆瓣读书评分 9 分以上的图书数据，要求采集到的每本图书的数据包括：书名、评分、作者、出版社和出版年份，并保存为本地 CSV 文件。

豆瓣读书评分 9 分以上榜单的网页示意图如图 3-7 所示。

图 3-7　豆瓣读书评分 9 分以上榜单的网页示意图

在编写 Scrapy 爬虫程序时，需要对图 3-8 所示的网页源码进行解析，并模拟单击翻页链接的动作，把所有评分 9 分以上的图书数据（共 34 个页面）都采集下来。

图 3-8　豆瓣读书评分 9 分以上榜单的网页源码

采集完成后保存到本地文件，Scrapy 爬虫程序采集的网页数据示意图如图 3-9 所示。

	A	B	C	D	E
1	title	author	press	year	rate
2	万物生光辉	[英] 吉米·哈利	中国城市出版社	2012-3	9
3	我亲爱的甜橙树	(巴西)若泽•毛罗•德瓦斯康塞洛斯	天天出版社	2010-6	9
4	教父	(美)马里奥·普佐	江苏文艺出版社	2014-4	9
5	故事	[美] 罗伯特·麦基	天津人民出版社	2014-9	9
6	树上的男爵	[意] 伊塔洛·卡尔维诺	译林出版社	2012-4-1	9
7	罗马人的故事2	盐野七生	中信出版社	2012-1	9
8	神秘岛（全三册）	(法)儒勒·凡尔纳	中国青年出版社	1979	9
9	罗杰·艾克罗伊德谋杀案	阿加莎·克里斯蒂	贵州人民出版社	1998-10-1	9
10	呼兰河传	萧红	中国青年出版社	2003-01-01	9
11	孽子	白先勇	广西师范大学出版社	2010.10	9
12	木心作品八种	木心	广西师范大学出版社	2009-1	9
13	灵魂的事	史铁生	百花文艺出版社	2005-4	9
14	万物既聪慧又奇妙	[英] 吉米·哈利	中国城市出版社	2010-7	9
15	局外人	[法] 阿尔贝·加缪	上海译文出版社	2010-8	9
16	荒原狼	[德]赫尔曼·黑塞	上海译文出版社	2010-9	9
17	所罗门王的指环	(奥)劳伦兹	和平出版社	1998-07-01	9
18	故道白云	一行禅师	线装书局	2007-6	9

图 3-9　Scrapy 爬虫程序采集的网页数据示意图

3.3.4　日志采集工具 Flume

Flume 是 Apache 的顶级项目，用于日志数据的采集。Flume 提供一种分布式、可靠的服务，用于高效地收集、聚合和移动大量的日志数据，它具备可调节的可靠性机制、故障转移和恢复机制，具有强大的容错能力。

Flume 支持在日志系统中定制各类数据源，可以对数据进行简单的处理，并将数据输出到各种数据接收方，其设计的原理是将数据流（如日志数据）从各种网站服务器上汇集起来，存储到大数据平台中。接下来，介绍两个 Flume 的应用场景。

场景 1：某个在线购物 App 需要建立用户推荐系统，它可以根据用户访问的节点区域、浏览的商品信息等来分析用户的行为或购买意图，以便更加快速地将用户可能想要购买的商品推送到界面上。为了实现这一功能，就需要收集用户在 App 上点击的产品数据、访问的页面和访问时间等日志信息，并保存到后台的大数据平台上去进行分析和挖掘，这样的需求就可以用 Flume 来实现。

场景 2：目前许多新闻类 App（如今日头条、腾讯新闻等）大都具有内容推送、广告定时投放和新闻私人订制等功能，这需要收集用户操作的日志信息（如用户曾经看过的新闻、视频、观看时间和 IP 地址等），以便使用智能推荐系统进行分析，更精准地向用户推荐可能感兴趣的内容和广告。

Flume 的核心流程是把数据从数据源收集过来，经过传送通道将收集到的数据送到指定的目的地，图 3-10 所示为 Flume 架构的示意图。

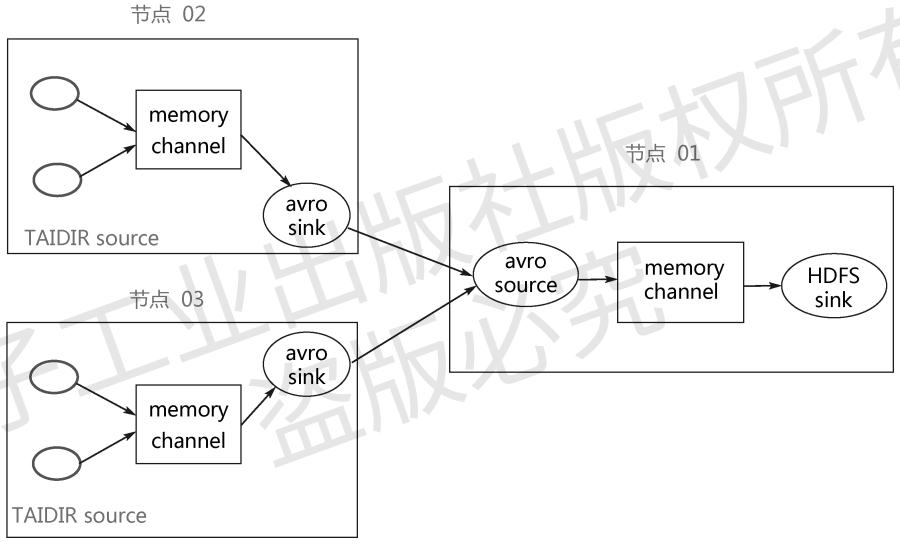


图 3-10　Flume 架构的示意图

【案例 3-2】利用 Flume 采集日志文件。

本案例利用 Flume，从多台 Linux 主机的特定目录中采集日志文件（数据源），将其汇总传输到指定的主机节点中，并存到 Hadoop 集群的 HDFS 文件系统（目的地）中。

1）任务描述

利用 Flume 实时监控并采集 Linux 主机节点 02 和节点 03 中/work/logs 文件夹下的所有日志文件，将采集到的日志文件传输到主机节点 01 中汇总，保存到 Hadoop 集群的 HDFS 文件系统中。图 3-11 所示为利用 Flume 采集日志文件的示意图。

2）解决方案

本案例中 Flume 数据流模型的示意图如图 3-12 所示，我们分别在主机节点 02 和节点 03 中运行 Flume 代理程序（Flume Agent），监视/work/logs 文件夹中日志文件的变动情况，实时地将新产生的日志数据以内存作为通道，通过 Avro 协议传输到主机节点 01 的指定网络端口。在主机节点 01 上运行 Flume 代理程序，监视指定端口的数据，将该端口接收到的数据保存到 Hadoop 的 HDFS 分布式文件系统中。

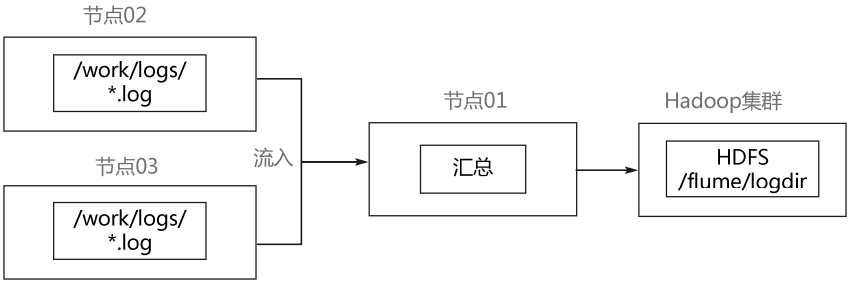


图 3-11　利用 Flume 采集日志文件的示意图

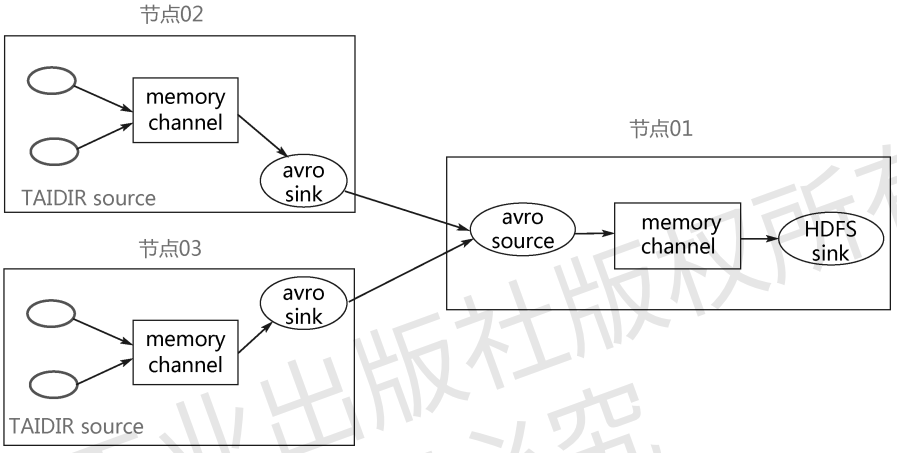


图 3-12　Flume 数据流模型的示意图

3.3.5　分布式消息服务工具 Kafka

Kafka 项目起源于 LinkedIn，2011 年成为开源 Apache 项目，2012 年成为 Apache 的一流项目。目前 Kafka 已发展为功能完善的基于分布式的消息发布-订阅系统。

Kafka 架构模型的示意图如图 3-13 所示。其中，消息发布者能够发布消息。消息接收者可以订阅一个或多个话题，并从 Kafka 集群（Kafak Cluster）上的消息服务节点中拉数据，从而消费这些已发布的消息。

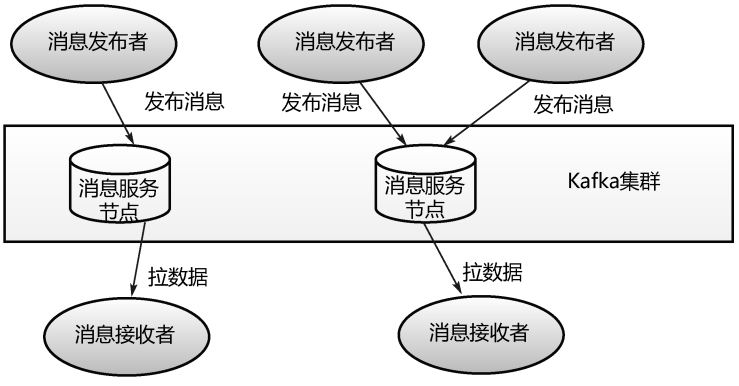


图 3-13　Kafka 架构模型的示意图

我们可以将 Kafka 与 Flume 配合使用，将 Kafka 作为 Flume 的源端，也可以将 Flume 采集的数据按照不同的类型输入 Kafka 不同的话题中。

3.4　数据预处理

数据预处理是指在数据进行分析挖掘之前，对原始数据进行变换、清洗与集成等一系列操作。通过数据预处理工作，可以使残缺的数据完整，对错误的数据予以纠正，将多余的数据去除，有效地提高数据的质量。

没有高质量的数据，就没有高质量的数据挖掘结果，低质量的数据会对许多数据挖掘算法有很大的影响，甚至“挖掘”出错误的知识。数据预处理的目的，是为后续的数据分析与挖掘提供可靠的高质量数据，提高数据挖掘的效率。数据预处理的流程如图 3-14 所示。

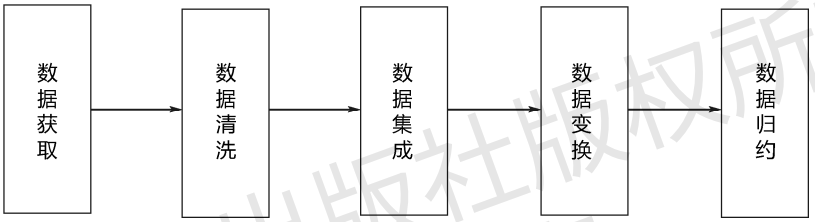


图 3-14　数据预处理的流程

3.4.1　数据清洗

数据质量的衡量标准涉及多种因素，包括数据的一致性、精确性、完整性、时效性和实体同一性。数据清洗就是为了提升数据的质量，即将“脏”数据清洗“干净”，数据清洗的示意图如图 3-15 所示。

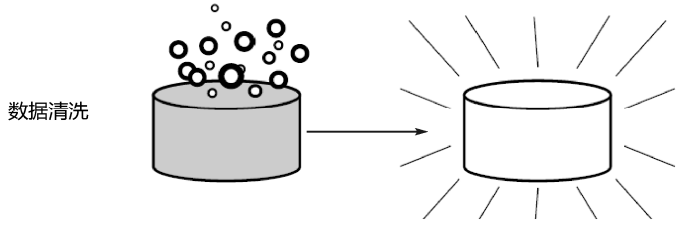


图 3-15　数据清洗的示意图

数据清洗的手段主要包括缺失值填充、数据平滑、识别和去除离群点、检测与修复不一致性、识别实体与发现真值等。

1．缺失值填充

数据集中的缺失值可能是由于错误或没有记录观察结果导致的。当数据集中存在缺失值时，某些数据分析或挖掘的算法可能不起作用，也可能无法得到所需的结果。数据集中的缺失值示例（其中，NA 表示缺失值）如图 3-16 所示。

年龄	教育	工龄	收入	负债率	信用卡负债	其他负债	违约
41	3	17	176	9.3	11.36	5.01	1
27	1	10	31	17.3	1.36	4	0
40	1	15	55	5.5	0.86	2.17	0
41	1	15	NA	2.9	NA	0.82	0
24	NA	2	28	17.3	1.79	3.06	1
41	2	5	25	10.2	0.39	2.16	0
39	1	20	67	30.6	3.83	NA	0
43	1	12	38	3.6	0.13	1.24	0
24	1	3	NA	24.4	1.36	3.28	1
36	1	0	25	19.7	2.78	2.15	0
27	1	0	16	1.7	0.18	0.09	0
25	1	4	23	5.2	0.25	0.94	0
52	1	24	64	10	3.93	2.47	0
37	1	6	29	16.3	1.72	3.01	0

图 3-16　数据集中的缺失值示例（NA 表示缺失值）

1）均值填充

如果缺失值是数值型的，就根据其他所有对象取值的平均值来填充该处缺失的变量值。

例如有以下一组数据：99、100、NaN、91、95（注：其中 NaN 表示缺失值），用平均值填充后得到：99、100、96.25、91、95。

这是因为：96.25 = (99 + 100 + 91 + 95) / 4。

2）众数填充

如果缺失值是非数值型的，通常使用众数来补齐该处缺失的变量值。

例如有以下一组数据：Apple、Orange、Banana、NaN、Apple、Apple、Orange

使用众数填充时，缺失值 NaN 将会被 Apple 替代，这是因为 Apple 出现的次数最多（3 次），这组数据中的众数为 Apple。

3）其他填充方法

其他的数据填充方法有拉格朗日插值法、回归填充法、热卡填充法、就近补齐法、极大似然估计法、期望最大化法、k 最近邻域法等。

2．数据平滑——去除噪声/异常值

噪声数据是指存在错误或异常的数据，这些数据对数据的分析和挖掘造成了干扰。在实际应用中，经常会遇到初始结果中噪声数据太多的问题，如音频中的背景杂音、光谱信号抖动得太厉害等。数据中的异常样本示意图如图 3-17 所示。

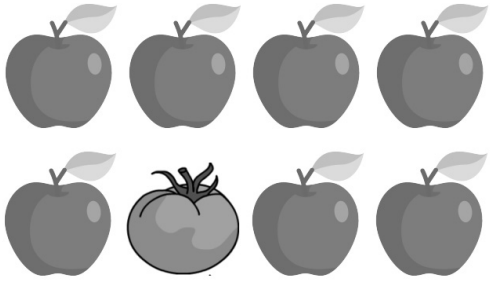


图 3-17　数据中的异常样本示意图

数据平滑就是去除数据中的噪声波动，使数据平滑，常用方法有分箱法、移动（滑动）平均法、3σ 探测法和 k 最近邻域法等，以下是常用的预处理方法。

1）分箱法

分箱法是通过考察“邻居”（周围的值）来平滑存储数据的方法，存储数据被分布到一些桶或箱中，平滑各个分箱中的数据，常用的分箱法如下。

（1）等深（频）分箱法：各个箱里有相同个数的数据。

（2）等宽分箱法：各个箱的取值区间相同。

常用的数据平滑方法有 3 种。

（1）平均值平滑：箱中每一个数据被箱的平均值替换。

（2）中位数平滑：箱中每一个数据被箱的中位数替换。

（3）箱边界平滑：箱中每一个数据被离它最近的箱边界值替换。

【案例 3-3】假设有 8、24、15、41、6、10、18、50、25 共 9 个数，先对数进行从小到大的排序，写为 6、8、10、15、18、24、25、41、50。

按等深（频）分箱法分为如下 3 箱。

箱 1：6、8、10。

箱 2：15、18、24。

箱 3：25、41、50。

分别用 3 种不同的平滑技术平滑噪声数据。

（1）按箱平均值求得平滑数据：如箱 1 的平均值为 8，这样该箱中的每一个数据被箱平均值替换，结果为(8,8,8)。

（2）按箱中位数求得平滑数据：如箱 2 的中位数为 18，这样该箱中的每一个数据被箱中位数替换，结果为 18。

（3）按箱边界值求得平滑数据：如箱 3 为(25, 41, 50)，箱中的最大值和最小值被视为箱边界值，这样该箱中的每一个数据被最近的箱边界值替换，结果为(25, 50, 50)。

按等宽分箱法（箱宽度为 10），可分为如下 3 箱。

箱 1：6、8、10、15。

箱 2：18、24、25。

箱 3：41、50。

分别用 3 种不同的平滑技术平滑噪声数据。

（1）按箱平均值求得平滑数据：如箱 1 的平均值为 9.75，这样该箱中的每一个数据被箱平均值替换为(9.75, 9.75, 9.75,9.75)。

（2）按箱中位数求得平滑数据：如箱 2 的中位数为 24，这样该箱中的每一个数据被箱中位数替换，结果为(24,24,24)。

（3）按箱边界值求得平滑数据：如箱 2 的箱边界值为 18 和 25（箱中的最大值和最小值被视为箱边界值），这样该箱中的每一个数据被最近的箱边界值替换，结果为(18, 25,25)。

2）移动（滑动）平均法

移动（滑动）平均法是一种用于滤除噪声的简单数据处理方法，该方法将观测数据替换为前后若干次观测数据的平均值，以便平滑噪声。

移动（滑动）平均法的计算公式为

$$p_t = \frac{\sum_{i=1}^{n}\left(x_{t-i} + x_{t+i}\right) + x_t}{2n+1}$$

式中， p_t 表示对 t 时刻观测数据的修正；n 表示滑动窗口半径；x_t 表示观测数据。

【案例 3-4】采用移动（滑动）平均法处理表 3-2 所示的观测数据，滑动窗口的半径取 2。

表 3-2　某段日期的观测数据

日期	2022 年 12 月 1 日	2022 年 12 月 2 日	2022 年 12 月 3 日	2022 年 12 月 4 日	2022 年 12 月 5 日	2022 年 12 月 6 日
观测 数据	18	15	16 （用 15 替换）	14 （用 14.4 替换）	12	15

使用移动（滑动）平均法，滑动窗口的半径取 2，则有如下结果。

（1）2022 年 12 月 3 日的测量数据 16 用(18+15+16+14+12)/5=15 替换。

（2）2022 年 12 月 4 日的测量数据 14 用(15+16+14+12+15)/5=14.4 替换。

3）3σ 探测法

3σ 探测法的思想来源于切比雪夫不等式，假设一组数据的平均值为 μ，标准差为 σ，一般情况下数据分布有以下特点。

（1）在所有数据中，数据分布在 $(\mu-\sigma,\mu+\sigma)$ 中的概率为 0.6827。

（2）在所有数据中，数据分布在 $(\mu-2\sigma,\mu+2\sigma)$ 中的概率为 0.9545。

（3）在所有数据中，数据分布在 $(\mu-3\sigma,\mu+3\sigma)$ 中的概率为 0.9973。

可以认为，数据的取值几乎全部集中在 $(\mu-3\sigma,\mu+3\sigma)$ 中，超出这个范围的可能性仅占不到 0.3%，凡超过这个区间的噪声数据应予以剔除。3σ 适用于有较多数据的情况。

4）k 最近邻域法

根据某种距离度量方法来确定噪声数据最近的 k 个近邻，然后将这 k 个数据加权（权重可取距离的比值），然后根据自定义的阈值，将距离超过阈值的数据当作噪声数据。

5）聚类算法

采用某种聚类算法把相似的数据聚成一个“簇”，落在各个簇之外的数据可以看成噪声数据，采用聚类算法去除噪声数据的示意图如图 3-18 所示。

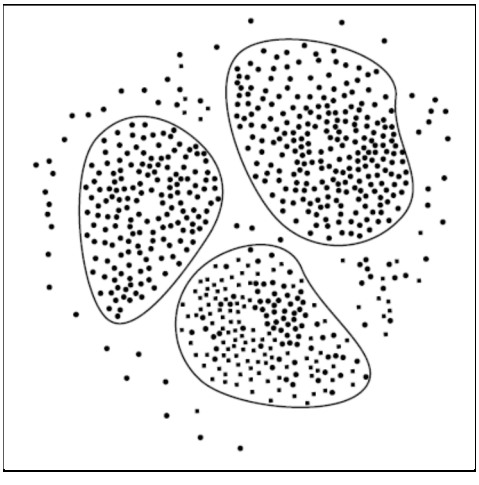


图 3-18　采用聚类算法去除噪声数据的示意图

3.4.2　数据集成

数据集成是把不同来源、格式的数据有机地集中起来，通过一致的、精确的表示法，对同一种实体对象的不同数据进行整合的过程。

在数据采集过程中，数据可能来自不同的系统，难以确保数据的模式、模态和语言的一致性，在很多应用中要将不同来源的数据集成汇总，这样才能正常使用。

根据方式的不同，数据集成可以分为传统数据集成和跨域数据集成。

1．传统数据集成

传统数据集成将来自多个数据集的数据以统一的模式进行集成汇总，以达到数据共享的目的（见图 3-19），例如，将多个来源的数据存储到一个关系数据库中。

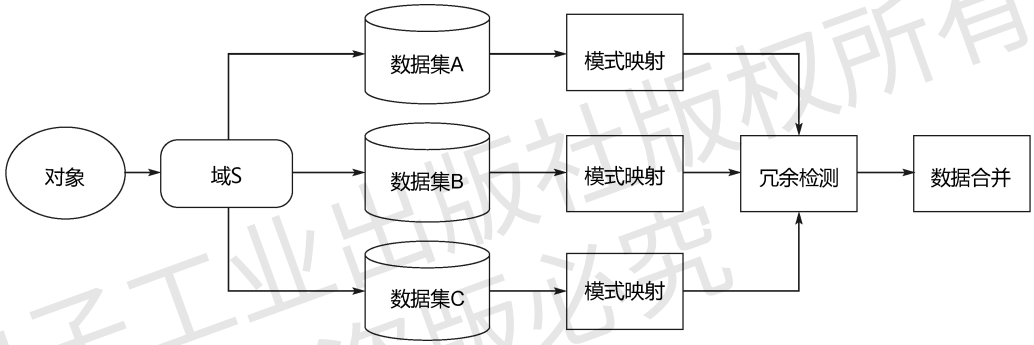


图 3-19　传统数据集成的示意图

2．跨域数据集成

跨域数据集成将来自不同领域的数据进行集成汇总（见图 3-20），例如，将社交网络关系、电商和金融三个领域的数据有机地整合在一起。

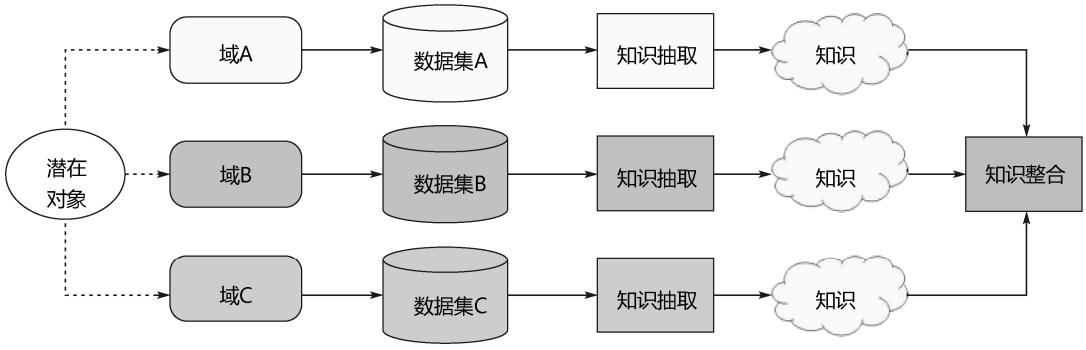


图 3-20　跨域数据集成的示意图

数据集成解决的主要问题包括实体识别、数据冗余问题和数据值冲突的检测与处理等。

1）实体识别

来自多个数据源的实体有时并不一定是匹配的，表 3-3 所示为来自不同数据源的同一个

人的姓名数据，要解决这样的问题：Wei Wang ?= Wang Wei。

表 3-3　来自不同数据源的同一个人的姓名数据

姓名	归属
Wei Wang	新加坡国立大学
Wang Wei	新加坡国立大学

2）数据冗余问题

在一个数据集合中重复的数据被称为冗余数据，产生冗余的因素有很多，如每天备份公司的数据会产生冗余，在多个系统中存储相同的信息时，最后也可能得到冗余数据。

3）数据值冲突的检测与处理

由于编码、数据类型、单位等不同，对于同一个实体，不同数据源的属性值可能不同。例如，某实体的重量属性可能在一个系统中以公制单位存放，而在另一个系统中以英制单位存放。

3.4.3　数据变换

所谓数据变换，是将数据转换成适当的形式以便更好地理解和处理，常用的数据变换方法如图 3-21 所示。

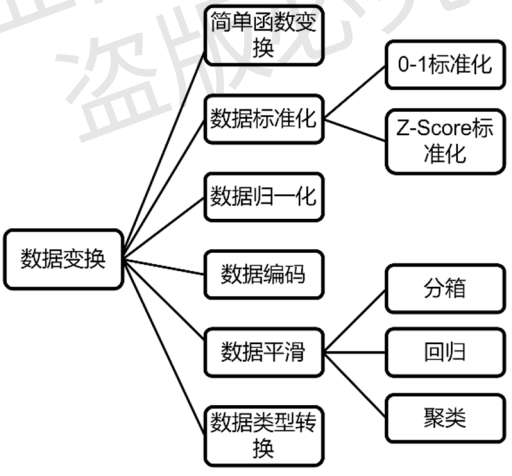


图 3-21　常用的数据变换方法

1．数据归一化

一个数据集中的数据由于各个属性的数据单位不同，各个属性的数据的值域范围可能相差很大，在进行数据分析和挖掘时可能会影响结果的准确性，不同尺度数据的示例图如图 3-22 所示，年龄和月薪这两个属性的数据尺度相差很大。

数据归一化是将数据变换为[0,1]之间的小数，这样可以把有量纲的数据转变为无量纲的数据，避免值域或量纲对数据的影响，便于对数据进行分析和挖掘。

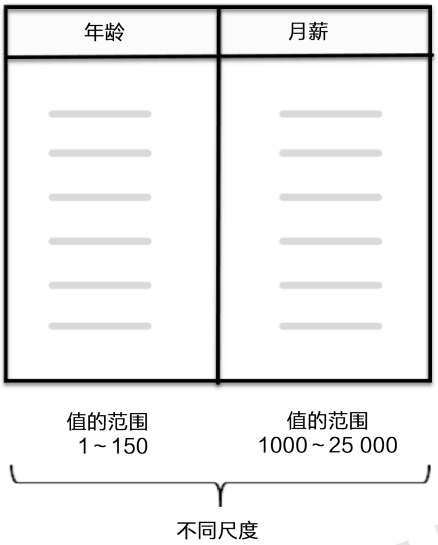


图 3-22　不同尺度数据的示例图

1）Max-Min 归一化方法

Max-Min 归一化也被称为离差归一化，该方法对原始数据进行线性变换，变换后的数据区间为[0,1]，公式为

$$f(x)=(x-\min)/(\max-\min)$$

这种方法的适应性非常广泛，它对数据归一化的同时还能较好地保持原有数据的分布结构。

【案例 3-5】Max-Min 归一化实例。

在一次考试中，某学生的语文成绩是 90 分，英语成绩也是 90 分。单从这次考试分数来评价，似乎该学生的语文和英语学得一样好。但是，语文总分是 150 分，而英语总分是 120 分，你还认为该学生的语文和英语成绩是一样的吗？

由于各科的考题难度不尽相同，假设班级中语文的最低分是 60 分，最高分是 140 分；英语的最低分是 80 分，最高分是 110 分。

根据 Max-Min 归一化公式，计算出该学生归一化的语文成绩：

$$0.375 = (90-60)/(140-60)$$

归一化的英语成绩：

$$0.33 \approx (90-80)/(110-80)$$

因此，该学生的语文成绩优于英语成绩。

2）用于稀疏数据的 MaxAbs

MaxAbs 根据最大的绝对值进行标准化，变换后的数据区间为[−1, 1]。

$$f(x)=x/\text{最大的绝对值}$$

MaxAbs 具有不破坏原有数据分布结构的特点，可以用于稀疏数据，或者稀疏的 CSR 或 CSC 矩阵。

2．数据标准化

数据标准化的目的是将不同规模和量纲的数据经过处理，消除数据间的量纲关系，从而使数据具有可比性，以减少规模、特征、分布差异等对模型的影响。

Z-Score 标准化可将数据转换成标准的正态分布，其转换公式如下：

$$f(x)=(x-\text{均值})/\text{标准差}$$

Z-Score 标准化之后的数据符合以 0 为均值、以 1 为方差的正态分布。Z-Score 标准化方法是一种中心化方法，会改变原有数据的分布结构，不适用于对稀疏数据进行处理。

【案例 3-6】Z-Score 标准化实例。

Z-Score 标准化的主要目的就是将不同量级的数据转化为同一个量级，统一用计算出的 Z-Score 值衡量，以保证数据之间的可比性。

假设两个班级考试，所采用的试卷不同。A 班级的平均分是 80 分，标准差是 10，刘同学考了 90 分；B 班的平均分是 400 分，标准差是 100，马同学考了 600 分。利用 Z-Score 标准化公式计算他们的标准分数，看看谁更优秀。

刘同学：(90−80)/10 = 1

马同学：(600−400)/100 = 2

因此马同学更优秀。

3．数据类型转换

在对数据进行分析和挖掘时，为了适应不同算法或不同应用场景的要求，可能需要在定性数据和定量数据之间进行转换，即有时需要把定性数据转换为定量数据，而有时需要把定量数据转换为定性数据。以下介绍一些常用的数据类型的转换方法。

1）标称数据的数值化编码

最常用的是独热编码（One-Hot Encoding）和标签编码（Label Encoding）两种编码方法。

独热编码又称一位有效编码，主要采用 *N* 位状态寄存器（有 0 和 1 两个状态）来对数据的 *N* 个状态进行编码，独热编码的示例如表 3-4 所示，“公司名”属性的每个数据被编码为 3 个属性值（因为该属性分为 3 个不同的属性：丰田汽车、本田汽车和大众汽车），每个数据转换后只有一个属性为 1，其他为 0。

表 3-4　独热编码的示例

公司名	丰田汽车	本田汽车	大众汽车
丰田汽车公司	1	0	0
本田汽车公司	0	1	0
大众汽车集团	0	0	1

标签编码将标称数据转换为数值型数据，即对不连续的数字或文本进行顺序编号，对表 3-4 中的数据运用标签编码的结果如表 3-5 所示。

表 3-5　标签编码的示例

公司名	公司编码
丰田汽车公司	0
本田汽车公司	1
大众汽车集团	2

可以看出，标签编码将原本无序的数据变成有序的数值序列，这是标签编码的明显缺陷。

2）连续变量的离散化处理

对连续变量进行离散化处理是把连续变量转换为离散值，如某门课程的成绩可以根据分数段转换为差、中、良、优（或 D、C、B、A）；又如客户的年龄数据，可以转换为<30、30～40、>40，如图 3-23 所示，纵轴代表人数。

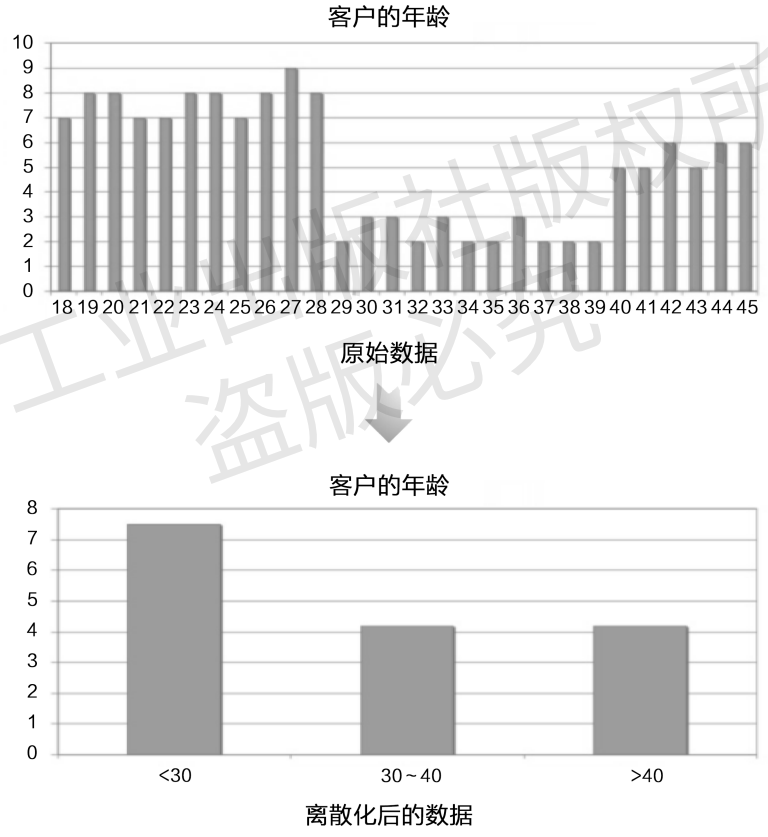


图 3-23　数据离散化的示例图

常用的数据离散化方法有等宽法、等频法和基于聚类的分析方法等。例如，有一组实数（浮点数）的数值在 0～100 之间，采用等宽法将数据划分为 10 个区间，将处于[0, 10)区间的数转换为 1，将处于[10, 20)区间的数转换为 2，以此类推。

4．数据编码

用统一的编码标准对信息记录进行编码，用一个编码符号代表一条信息或一串数据。运用数据编码方法可以对采集的大数据进行规范化管理，提高处理效率和精度。

例如，我国的身份证号码就是一种带有特定含义的编码方案，它可以表示某人所在省市区、出生年月、性别等，因此可以用特定的算法进行分类、校核、检索、统计和分析操作。

3.4.4 数据归约

数据归约是指在尽可能保持数据原貌的前提下，最大限度地精简数据量，得到数据的简化表示。

1．特征归约（降维）

特征归约是指从原始特征中排除不重要或不相关的特征（特征选择），或者通过重新组合特征来减少特征的数量（特征投影）。其原理是降低特征向量的维度，同时保持甚至提高原有的分辨能力。

2．数量归约（数量削减）

数量归约是指用较简单的数据表示形式替换原数据，或者采用较小的数据单位，或者用数据模型代替数据以减少数据量。常用的方法有直方图、用簇中心表示实际数据、抽样和参数回归法，以及特征值离散化技术等。特征值离散化技术是将连续的特征值离散成少量的区间，每个区间被映射成一个离散的符号，该技术的优点是简化了数据的描述，使人们更容易理解数据和最终的挖掘结果。

3．样本归约（数据抽样）

样本的数量通常很大，质量或高或低。样本归约就是从数据集中选出一个有代表性的样本的子集。在确定子集大小时，应考虑计算成本、存储要求、估计的准确性，以及与算法和数据的特点有关的其他因素。

3.5 使用 OpenRefine 对数据进行预处理

OpenRefine 是一种开源的交互式数据转换工具，可以对结构型数据进行数据清洗、修正、分类、排序、筛选与整理，它的功能强大，使用较为简便。

OpenRefine 软件的前身是 MetaWeb 公司于 2009 年发布的一个开源软件，谷歌在 2010 年收购了 MetaWeb 公司，把项目的名称从 Freebase Gridworks 改成了 Google Refine。2012 年，谷歌放弃了对 Refine 的支持，让它重新成为开源软件，名字改成了 OpenRefine。

目前的 OpenRefine 有 Windows、Mac 和 Linux 版本，以下用 OpenRefine 3.6.2 for Windows 版本和示例数据集 movie-metadata 为例介绍该软件的基本使用方法。

1．下载安装 OpenRefine 软件

访问 OpenRefine 的官网，下载所需的版本，如图 3-24 所示。运行 OpenRefine 需要 Java 运行环境的支持，因此首先需要安装 OpenRefine 要求的 JDK（Java Development Kit）版本（注：OpenRefine 3.7.4 要求 JDK 版本大于 11）。

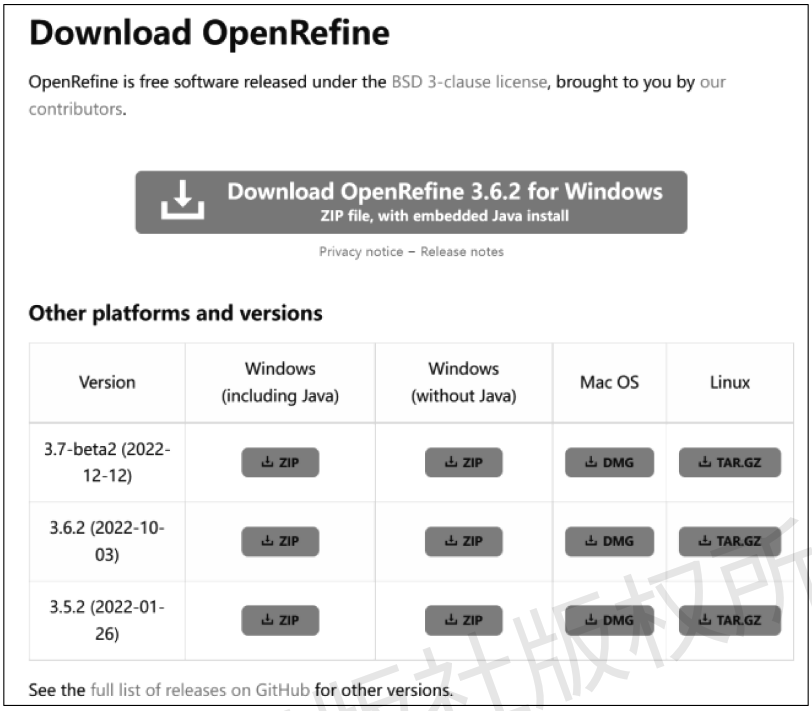


图 3-24　下载 OpenRefine

将下载的 OpenRefine 压缩包解压到某个文件夹中，然后运行 openrefine.exe 程序，如果当前计算机环境中没有配置 Java 运行环境，会提示下载并安装 JDK。

2．建立 OpenRefine 项目

OpenRefine 启动后是以本地 Web 服务（默认端口号为 3333）运行的，可以用浏览器访问 127.0.0.1:3333 打开初始界面，如图 3-25 所示。

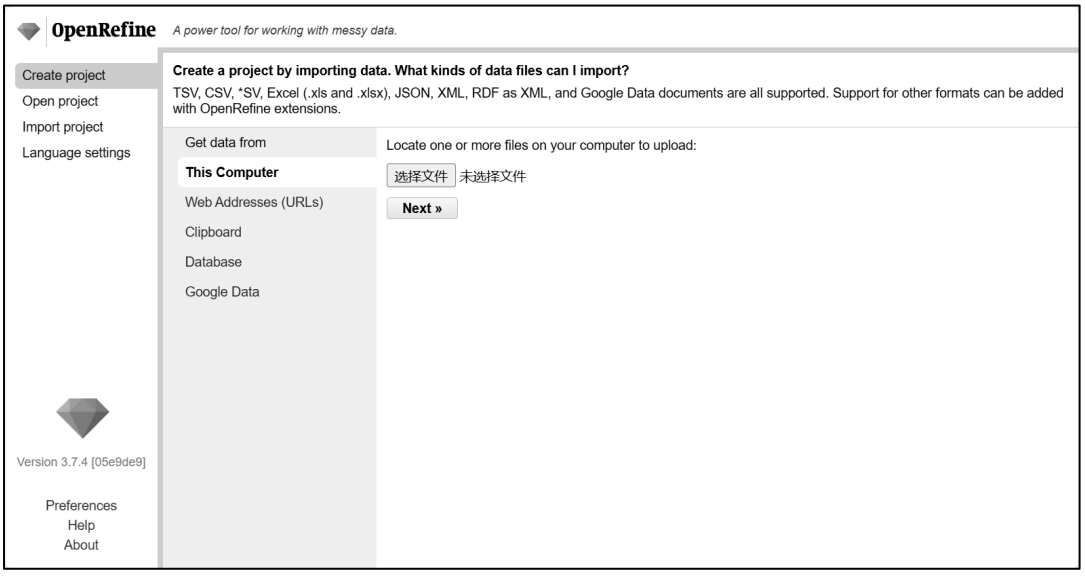


图 3-25　OpenRefine 初始页面

可以把 OpenRefine 页面的语言更改为简体中文，如图 3-26 所示。

在 OpenRefine 初始页面中单击“选择文件”按钮打开本地数据文件，接着单击“下一步>>”按钮进入文件格式选项设置对话框。

注意：如果是 CSV 或 Excel 文件，需要在弹出的对话框下方的选项中勾选“将单元格中的文本解析为数字”复选框，如图 3-27 所示，否则 OpenRefine 会将实数列作为文本处理。

图 3-26　更改 OpenRefine 页面的语言

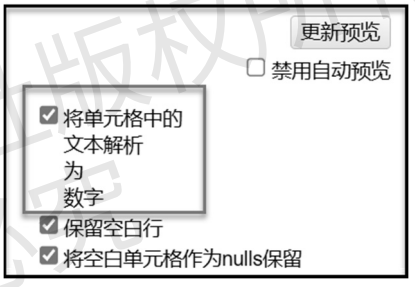


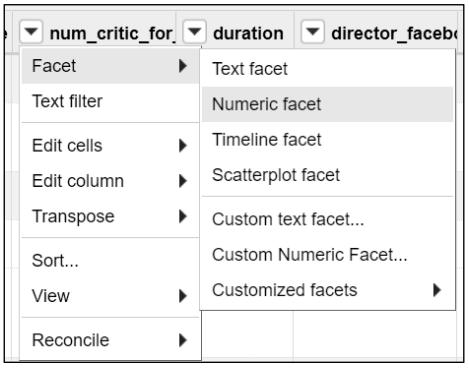
图 3-27　CSV 和 Excel 文件需要勾选的格式复选框

输入项目名称，单击“新建项目”按钮建立 OpenRefine 项目。

3．OpenRefine 的数据预处理功能

选择想要处理的列（属性/特征），单击列标题左边的箭头出现下拉菜单，选择“归类”选项（如果是英文界面，则为“Facet”选项），如果是文本数据就选择“文本归类”选项，如果是数字数据就选择“数值归类”选项，如图 3-28 所示。

（a）中文界面



（b）英文界面

图 3-28　选择想要处理的列

数据归类是 OpenRefine 中常用到的功能之一，它主要用于从数据中获得子集，方便用户从多个角度查看数据，并且不会改变数据本身。OpenRefine 支持多种归类，包括文本归类、数值归类、时间线归类、散点图归类及自定义归类。

选择 movie-metadata 数据集的 num_critic_for_reviews 列，单击“数值归类”选项后，页面的左边会出现该列的数据透视结果，该结果是以频率图的形式显现的，从左到右按升序排列，柱形的高度表示其密集程度，下面 4 个复选框表示数据类型，列数据透视结果的示意图如图 3-29 所示。

图 3-29　列数据透视结果的示意图

1）缺失值填充

可以从图 3-29 中看出数据集的 num_critic_for_reviews 列有 50 个缺失值（Blank）。取消勾选“数值型”复选框，则只会显示包含缺失的行，在缺失值单元格中单击“编辑”按钮，填入想要的值（如该列数据的平均值），缺失值填充的示意图如图 3-30 所示。

图 3-30　缺失值填充的示意图

2）数据转换

可以对数据进行一些必要的转换，从而解决由于英文单词大小写不统一、额外的空格、首行有空格等造成的偏差。

在打开的数据文件中选择某一文本列，在下拉菜单的“编辑单元格”选项中选择“常用转换”选项，则出现常用的数据转换操作选项。例如：选择“移除首尾空白”命令则会去除该列所有的行首和行尾空格；选择“全大写”命令则会将该列所有字母从小写转换到大写。数据转换操作的示意图如图 3-31 所示。

图 3-31　数据转换操作的示意图

3）数据排序

选择某列，在下拉菜单中选择“排序”选项，则会出现图 3-32 所示的排序对话框，OpenRefine 支持 4 种排序依据，文本、数字、日期和布尔，并为每种排序依据提供了相应的两种排序方式。

4．OpenRefine 的数据分析功能

在选择数据时，筛选条件并不是那么严格时可能会选择好几类相似的数据，此时可以用聚类方法对相似的数据进行聚类。下面对 genres 列数据进行聚类，打开“genres”下拉菜单，在下拉菜单中选择“编辑单元格”中的“簇集并编辑”选项，如图 3-33 所示。

图 3-32　排序对话框

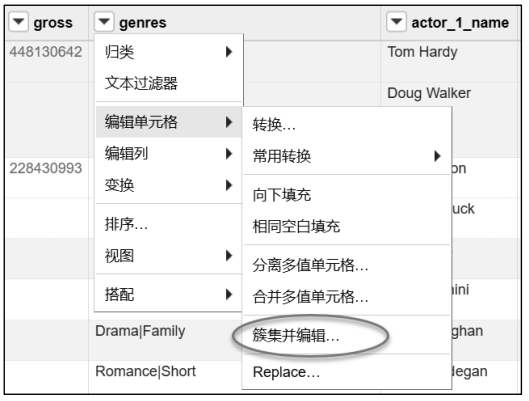


图 3-33　对 genres 列数据进行聚类

在聚类分析界面中选择“方法”下拉列表中的“就近原则”选项，采用 ppm 计算字符串之间的距离并进行聚类，聚类分析结果的示意图如图 3-34 所示。

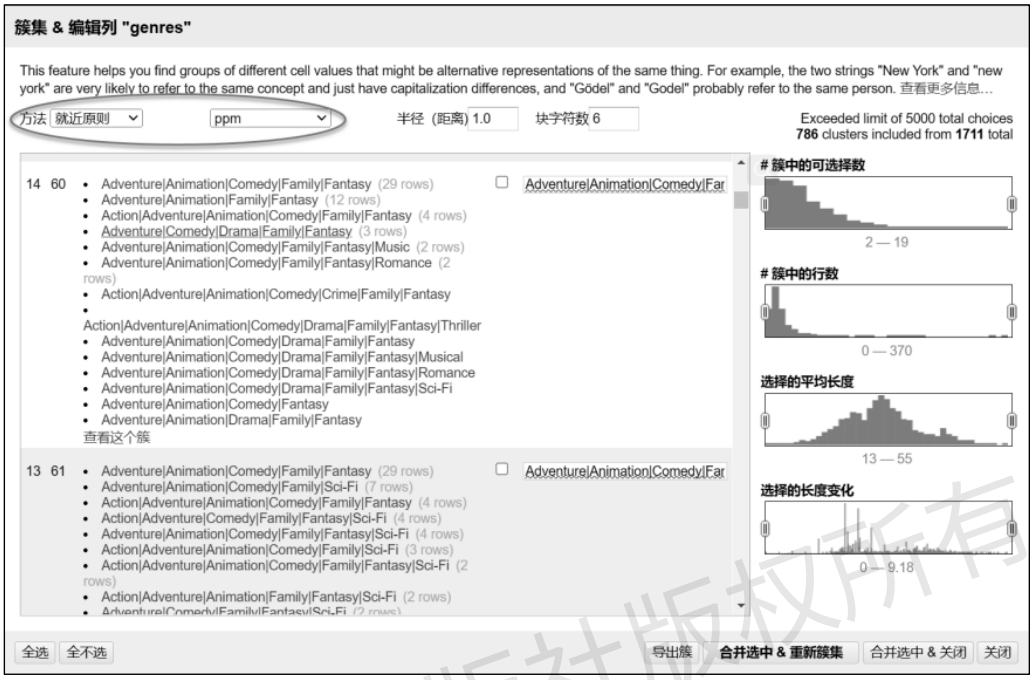


图 3-34　聚类分析结果的示意图

OpenRefine 的特点是对列的操作十分便利，可以进行列的隐藏、展开、转换、移动、重命名和删除等操作，可以直观方便地观察、分析和操纵数据。

OpenRefine 会在项目创建后保存所有的操作步骤，所以可以随便尝试各种变换数据，也可以将操作完成后的数据保存为 CSV、TSV、Excel 和 Open Document 格式，还可以导出 OpenRefine 压缩包和进行自定义导出设置等。

3.6　本章小结

本章的主要内容包括大数据采集工具的使用及数据预处理的常用方法，首先介绍了大数据的来源和采集方法，接着介绍了常用的大数据采集工具，最后对常用的数据预处理的方法进行了简要介绍。

3.7　习题

一、选择题

1．按产生数据的主体来划分，大数据主要有三个来源，它们分别是（　　）。

A．信息管理系统的记录、计算机产生的数据和对现实世界的测量

B．对现实世界的测量、人类的记录和计算机产生的数据

C．对现实世界的测量、人类的记录和物联网监测的数据

D．对现实世界的测量、传感器采集的数据和计算机产生的数据

2．以下哪项处理属于数据集成要完成的工作（　　）。

A．数据平滑　　B．标准化处理

C．去除噪声　　D．实体识别

3．以下哪项不属于数据预处理中去除噪声的方法（　　）。

A．分箱法　　B．移动（滑动）平均法

C．归一化　　D．3σ 探测法

4．假设有 12 个从小到大的有序数据：5、10、11、13、15、35、50、55、72、92、203、215，使用等宽分箱法（假设箱宽为 30）将它们划分成 4 个箱，92 在哪个箱中（　　）。

A．第 1 个　　B．第 2 个

C．第 3 个　　D．第 4 个

二、填空题

1．网络数据的采集通常可以分为_____和_____两种类型。

2．数据容量单位 1 GB=_____KB。

3．数据预处理是指在数据进行分析挖掘之前，对原始数据进行_____、_____与_____等一系列操作。

三、简答题

1．简述大数据采集主要采用的方法。

2．简述归一化处理和标准化处理的差别。