
国家教材委员会专家委员会审核通过
2020
普通高中教科书
信息技术
选择性必修 3 数据管理与分析
华东师范大学出版社

普通高中教科书

信息技术

选择性必修 3

数据管理与分析

华东师范大学出版社

总　主　编：李晓明

副 总 主 编：赵　健

本 册 主 编：张　洁

本册副主编：王　肃

编 写 人 员（按姓氏笔画排序）：

王　肃　毛黎莉　张　洁　高　峰

责 任 编 辑：曹祖红

美 术 设 计：储　平

普通高中教科书　信息技术　选择性必修 3　数据管理与分析

上海市中小学(幼儿园)课程改革委员会组织编写

出版发行　华东师范大学出版社(上海市中山北路 3663 号)

印　　刷　上海四维数字图文有限公司

版　　次　2021 年 3 月第 1 版

印　　次　2021 年 3 月第 1 次

开　　本　890 毫米×1240 毫米　1/16

印　　张　8

字　　数　139 千字

书　　号　ISBN 978-7-5760-0551-6

定　　价　10.10 元

致同学们

亲爱的同学们：

当今，信息技术的发展日新月异，物联网、大数据、人工智能等新技术、新工具扑面而来，显著地改变着人们的生活、学习和工作模式。生存于信息社会中，我们每一个人都不可避免地会接触信息技术、应用信息技术，甚至去创造新信息技术。在具备了基本信息技术应用能力的基础上，高中阶段我们要进一步学习信息技术的知识与技能，能够利用信息技术负责任地解决生活与学习中的问题，全面提升信息素养，迎接信息社会的挑战。

“数据管理与分析”作为高中信息技术学科的选择性必修模块，是高中信息技术学科的重要内容。本教科书采用“项目活动”方式组织学习内容，通过“身边的数据价值以及数据管理与分析”“网上书店数据管理”“在线考试系统的安全维护”“上海市旅游景点数据分析”和“电影数据的数据挖掘”项目，将数据价值、数据管理、数据分析、数据安全、大数据、数据挖掘等知识与技能融入学习活动中。教科书的每章围绕“信息意识”“计算思维”“数字化学习与创新”“信息社会责任”四个学科核心素养提出本章的学习目标，利用“本章知识结构”图示呈现本章知识脉络，帮助同学们从总体上了解本章学习内容。

在学习过程中，同学们可以通过“问题思考”栏目，将现实问题、个人经验与知识技能相关联，带着问题开始学习；通过“项目实践”“探究活动”和“体验思考”栏目，将“做中学”与“学中做”的学习方法相互融合，把知识技能应用于解决实际问题中；根据“作业练习”栏目提供的练习，应用所学的知识技能解决新的实际问题，提高创新能力；按照个人的学习需求，学习“知识延伸”栏目中的内容，拓展个人学习视野。

提升信息素养，要求我们在掌握数据管理与分析的基础技术知识、学会使用数据管理与分析工具的同时，能够用计算思维来分析问题；要求我们在体验数据管理与分析技术给生产生活带来便利的同时，学会运用相关知识创造性地解决实际问题，并且关注数据安全，参与和促进信息社会的伦理与道德建设。同学们可以通过本教科书及其配套资源，学习数据管理与分析技术，负责任地应用数据管理与分析技术，逐步成长为新时代合格的社会主义建设者。

编　者

目　录

第一章

数据管理与分析初步

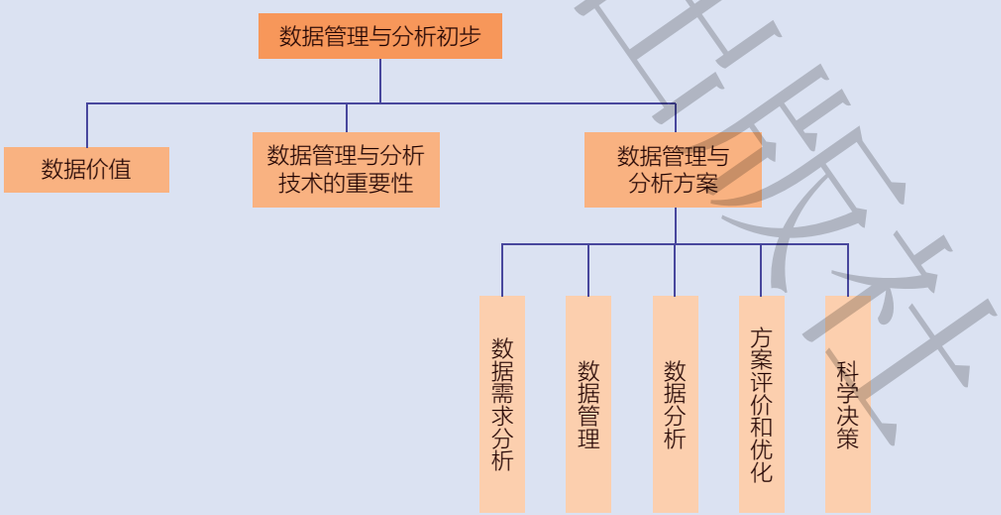
本章学习目标

- 认识到数据是一种重要资源，了解数据的价值，能够发现生活中的数据所蕴含的价值。
- 认识数据管理与分析技术的重要性。
- 结合具体活动了解数据需求分析方法，能结合实际问题进行数据需求分析。
- 了解建立数据管理与分析方案的基本过程，能结合实际问题制定数据管理与分析方案，并对所制定的方案进行评价，针对发现的问题进行方案优化。

数据，古已有之，它是人类改造世界的一种重要资源。古时候的结绳记事记录了数据，货币、度量衡、罗盘的使用都体现了人们对数据的利用。在信息时代，数据更是无处不在，其内涵和价值更为丰富。例如，超市的收银系统会记录顾客购买商品的相关数据；在线学习系统可以记录学习者学习的内容、完成的作业和测试及相应的成绩等数据；人们可以通过在线社交软件发布文字、图片、视频，以及自己所在地理位置等数据。随着科技尤其是人工智能技术的发展，将会有更多的智能设备接入互联网，最终实现万物的互连互通。智能手表、智能家电、无人驾驶汽车、机器人等智能设备利用数据为人们提供服务，同时它们在运行时也记录着大量的数据。例如，智能手表不仅可以为佩戴者提供时间、天气情况等数据，还可以记录佩戴者的实时位置、运动步数、实时心率等数据。

数据的飞速增长给人们带来了更多的机遇和挑战，如何利用好数据，使数据实现其应用价值，是人们越来越关注的问题。“工欲善其事，必先利其器”，实现数据价值是“善其事”，数据管理与分析技术是“善其事”的利器。在信息时代，我们必须利用合理高效的数据管理与分析方法管好数据、用好数据，使数据发挥出更大的价值，为人们的衣食住行提供更好的服务，帮助企业赢得更大的效益和商机，促进国家科技和经济发展。

本章知识结构



项目主题

身边的数据价值以及数据管理与分析

项·目·情·境

年初，学校科创社团开展了一次社会实践活动——参观调研某智能手环研发企业。首先，我们通过地图软件查找学校到该企业的路线，地图软件为我们提供了多种出行方案。到了企业后，经过调研，我们了解了该企业某种畅销智能手环上一年度每个月的销售量数据。我们希望根据这些数据预测明年该智能手环的月销售量，从而帮助企业制定明年的生产计划和合适的营销策略。我们需要对调研到的销售量数据进行管理与分析以便作出预测。

活动结束后，我们把活动报道发布到学生社团网站，让其他同学了解这次有意义的活动。大家对我们的活动非常感兴趣，有许多同学都在活动文章的评论区里发表了评论，还有许多同学转发了这篇文章。同样，我们也可以在网站上查看很多其他社团开展的丰富多彩的活动。学生社团网站上发布了这么多的活动，哪个活动的浏览量最高？哪个活动大家讨论得最热烈？哪个活动的转发量最高？学生社团网站如何向不同的用户推荐他们可能感兴趣的活动呢？我们可以制定一个学生社团网站的数据管理与分析方案来解决这些问题。

项·目·任·务

任务 1

通过“交通路线规划中的数据价值”项目实践活动，了解交通数据的价值。

任务 2

通过“企业商品月销售量数据分析”项目实践活动，了解数据管理与分析技术的重要性。

任务 3

制定学生社团网站的数据管理与分析方案，了解针对具体问题进行数据需求分析、建立数据管理与分析方案的基本过程，以及如何对方案进行评价和优化。

第一节　数据价值

在日常生活中，无论是看新闻、听音乐、购物，还是吃饭、运动，甚至走路、睡觉，人们几乎所有的活动都和数据息息相关。例如：购买火车票时，火车票订票系统会通过对车次运行数据、票务数据、乘客数据等多种数据的有效利用为用户提供方便快捷的订票服务；网络购物时，在线购物网站通过对大量的商品数据、会员数据、订单数据、物流数据等进行管理和分析，为人们提供便利的商品查询和比价、个性化商品推荐等服务。在享受着数据带来便利的同时，我们的各种行为也被智能手机、智能穿戴设备等记录下来，成为数据。数据已经渗透到了日常生活的方方面面以及每一个行业领域。数据蕴含着巨大的价值，合理地使用数据是非常重要的。

问　题　思　考

随着信息技术的发展，数据已经无处不在，并给人们的生产生活带来深远的影响。

请思考：

1. 在生活或学习中，你使用了哪些数据？这些数据对你而言有哪些价值？（请举例说明）
2. 为什么数据管理与分析技术对于实现数据的价值是非常重要的？

在信息社会中，数据价值体现在生产生活以及各行各业中。数据可以为人们的生产生活提供服务和便利，例如，气象数据可以用于预测天气，为人们安排出行和生产生活提供方便。数据可以帮助企业进行创新和决策以提高经济效益，例如，企业利用客户数据和销售数据可以对不同的客户群体进行有针对性的营销。数据可以为政府的科学决策提供支持，例如，公共卫生部门可以利用覆盖区域的居民健康档案数据和电子病历数据，快速检测传染病，进行全面的疫情监测。

项　目　实　践　　交通路线规划中的数据价值

出行时，如果不知道出行路线是一件非常麻烦的事情。可以利用地图软件查找从出发地到目的地的路线，帮助我们快速地做好交通路线规划。规划交通路线需要对出发地、目的地、道路长度、道路状况等多种交通数据进行分析。根据项目情境中的描述，我们要从学校到研发智能手环的企业去开展社会实践活动。请根据图 1.1中的数据，规划从学校到该企业的路线，图中的数字表示道路长度（单位：千米）。

步骤 1　对图 1.1中的所有地点进行编号，如表 1.1所示。

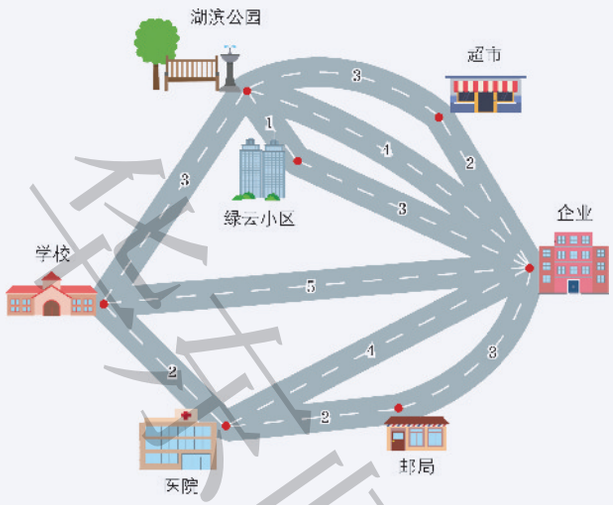


图 1.1　道路图

表 1.1　地点编号表

地点	编号
学校	A
湖滨公园	B
医院	C
绿云小区	D
超市	E
邮局	F
企业	G

步骤 2　请根据图 1.1将不同地点间直接到达(不经过其他地点)的道路长度(单位：千米)填入表 1.2。

表 1.2　地点间直接到达的道路长度表

	A	B	C	D	E	F	G
A	/	3	2	/	/	/	5
B							
C							
D							
E							
F							
G							

步骤 3　计算从学校(A)到企业(G)一共有几条可以到达的路线，并将相关数据填入表 1.3。

表 1.3　学校到企业的路线规划表

路线编号	路线	路线长度(千米)
1	A—G	5

由表 1.3可知，从学校到企业有多条路线。通常，地图软件会推荐最短路线，但是如果最短路线出现堵塞或者路况维护等情况，地图软件很可能会根据路况数据推荐其他路线。即使最短路线路况良好，地图软件也可能会根据出行者的不同需求而推荐其他路线。

步骤 4　根据出行需求及路况，设计推荐路线并填入表 1.4，格式为“路线（长度）”，其中路线长度以千米为单位。

表 1.4　学校到企业的推荐路线表

出行需求及路况	推荐路线（可以有多条）
路线长度最短	A—G（5）
AG 堵塞	
AG、BG、DG 修路，道路不通	

交通数据为人们的生活提供了很多便利，如路线查询、物流配送、实时导航等。人们在使用这些数据时，可以感受到数据的价值。例如，张先生要从上海出发到郑州参加一个重要会议，可是在买火车票的时候发现出发日上海到郑州直达车的车票已经卖完了，他是否必须改乘其他交通工具呢？其实，张先生可以利用火车票订票系统的路线换乘查询功能，查询上海到郑州的中转换乘推荐路线。

生活中还有哪些交通数据为我们提供了便利？这体现出了什么样的数据价值？请思考并填入表 1.5。

表 1.5　交通数据的价值

应用场景	场景中的数据	数据价值
交通路线规划	地点位置、道路长度、道路路况等	

第二节　数据管理与分析技术的重要性

数据本身蕴含着价值，通过数据管理与分析可以发现数据更多的价值，为科学决策提供重要依据。例如，上海的公交车都安装了卫星定位设备，上海城市公交系统的管理中心可以实时获得每辆公交车的当前位置、行驶路线、行驶速度等数据，通过对这些数据进行管理和分析，帮助人们实时查询公交车预计到站时间，为公众出行提供便利。如图 1.2 所示为上海市某公交车站电子站牌实时显示公交车预计到站时间。又如，有的智能手环可以对老人的血压进行实时监测，通过对这些数据进行管理和分析，生成老人的血压曲线图（如图 1.3 所示），并利用手机应用程序推送给其家人，让家人了解老人的血压状况。如果老人身体不适或突发疾病，血压偏离了本人的正常曲线值，手机应用程序可以及时发出警报并通知其家人。

图 1.2　上海市某公交车站电子站牌实时显示公交车预计到站时间

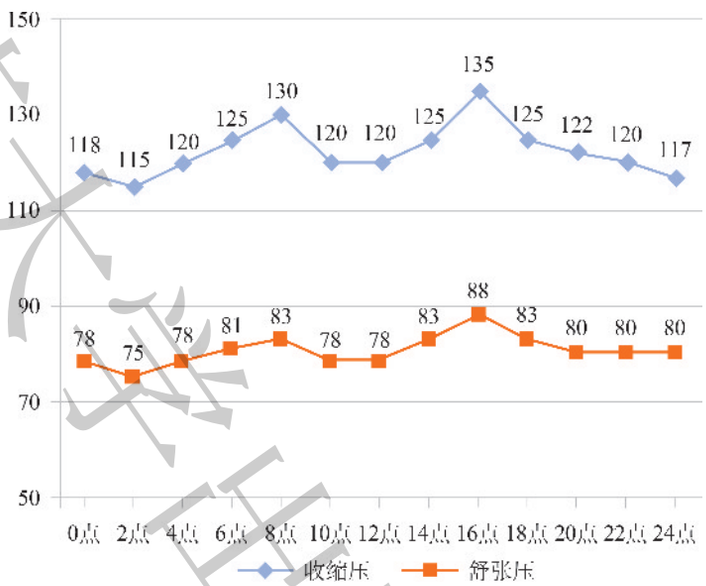


图 1.3　智能手环生成的穿戴者的血压曲线图

项　目　实　践　企业商品月销售量数据分析

在企业进行参观调研时，我们了解了该企业某种畅销智能手环上一年度每个月的销售量。这些商品月销售量数据可以反映该商品的月销售情况，这是数据本身蕴含的价值。而运用数据管理与分析技术对月销售量数据进行分析，可以充分发挥这些数据的价值和作用。例如，我们可以对上一年度每个月的商品销售量进行分析，预测今年每个月的销售量，从而帮助企业制定生产计划或合适的营销策略。

请根据某企业上一年度 1～12月份智能手环的月销售量（如表 1.6所示），预测该商品今年各月的销售量。

表 1.6　智能手环上一年度 1～12月份的月销售量表

月份	1	2	3	4	5	6	7	8	9	10	11	12
月销售量(个)	180	205	220	243	234	257	260	285	290	300	305	288

数据预测是一种常见的数据分析应用。根据预测数据和预测目的的不同，需要选择合适的数据预测方法。简单移动平均法是一种常用的数据预测方法，它主要是利用一组最近的历史数据的平均数来预测未来的数据值，经常被用于预测企业商品的需求量、销售量等。当商品需求量或者销售量既不快速增长也不快速下降，且不受季节性因素影响时，简单移动平均法能有效地消除预测中的随机波动。

简单移动平均法的计算公式如下：

$$P_t = \frac{S_{t-1} + S_{t-2} + \cdots + S_{t-n}}{n}(n \neq 0)。$$

其中，P_t表示未来一期的预测值(即移动平均值)；n表示移动平均时期个数；S_{t-1}表示前一期的实际值，S_{t-2}表示前两期的实际值，以此类推，S_{t-n}表示前 n期的实际值。

由表 1.6可知，智能手环上一年度 1～12月份的销售量变化平稳，没有快速下降或增长，因此可以应用简单移动平均法进行销售量预测。请按照表 1.7设置的移动平均时期个数(n)，计算今年智能手环月销售量的预测值(结果保留到整数)，并进行对比。

表 1.7　今年智能手环月销售量预测表

月份	上一年度的月销售量(个)	n= 2的预测值	n= 3的预测值	n= 4的预测值
1	180	/	/	/
2	205	(205+ 180)/2 = 193	/	/
3	220	(220+ 205)/2 = 213	(220+ 205+ 180)/3 = 202	/
4	243	(243+ 220)/2 = 232	(243+ 220+ 205)/3 = 223	
5	234			
6	257			
7	260			
8	285			
9	290			
10	300			
11	305			
12	288			

由表 1.7可知，移动平均时期个数会影响预测结果，因此选择合适的移动平均时期个数至关重要。通过对历史数据设置不同的移动平均时期个数，并将得到的预测值和实际数据进行对比，可以得到合适的移动平均时期个数。例如，企业可以根据商品 2018年的月销售量，通过设置不同的移动平均时期个数，计算出商品 2019年的月销售量的多个预测值，并将这些预测值和商品 2019年的月销售量的实际值进行对比，从而得到能使预测结果更准确的移动平均时期个数，然后再利用这个移动平均时期个数来预测商品未来的月销售量。这有利于企业更科学地安排生产、制定营销策略。

请同学们根据上述内容填写表 1.8。

表 1.8　商品月销售量数据的价值

数据名称	数据本身的价值	由数据管理与分析技术实现的数据价值
商品月销售量数据		

知 识 延 伸　数据隐私

在信息社会中，人们无时无刻不在和数据打交道。你在社交网络上发布了一条消息或几张图片，社交网络会记录你的信息；物联网中大量的传感器、视频监控摄像头等设备每时每刻都在采集着大量数据。随着无处不在的各类终端不停地收集越来越多的数据，无论你去哪儿，都会留下“脚印”，这可能会存在数据隐私泄露危险。例如，你在某个网站注册时填写了个人资料，包括姓名、手机号、家庭住址等重要信息，经过你的同意，网站有权使用你的资料为你提供服务，但是这并不代表这些数据可以变成网站营销的资源，或者随意流通到其他公司。那么，什么是数据隐私呢？通常，数据隐私就是个人不愿公开的个人信息，包括身份证号、银行账号、手机号、E-mail地址、家庭住址、工作单位、指纹、病史记录等。如何保护数据隐私呢？一方面，每个人都要树立维护数据隐私的意识，既要合法使用数据，也要合理使用数据；另一方面，要有一些可以有效防止数据泄漏的技术手段；当然，还要有健全的隐私保护法律体系。我国先后出台了一系列数据安全相关政策法规。《中华人民共和国网络安全法》自 2017年 6月 1日起施行，其中明确规定个人信息是指以电子或者其他方式记录的能够单独或者与其他信息结合识别自然人个人身份的各种信息，包括但不限于自然人的姓名、出生日期、身份证件号码、个人生物识别信息、住址、电话号码等；任何个人和组织不得窃取或者以其他非法方式获取个人信息，不得非法出售或非法向他人提供个人信息。

（参考资料：维基百科）

第三节　数据管理与分析方案

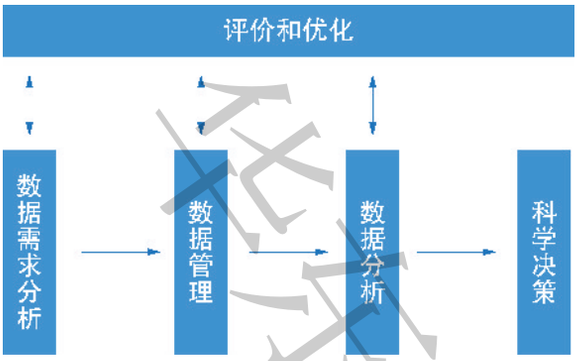


图 1.4　建立数据管理与分析方案的基本过程

数据蕴含着巨大的价值，如果想利用好数据，让其更好地为人们服务，就需要制定合理、有效的数据管理与分析方案。数据管理与分析方案是一个全面系统的综合性解决方案。针对数据需求分析中提出的问题，建立合适的方案对数据进行管理与分析，可以为用户提供服务或决策支持。建立数据管理与分析方案的基本过程包括数据需求分析、数据管理、数据分析、方案评价和优化、科学决策，如图 1.4 所示。

问　题　思　考

在信息社会中，除了电视、报纸、杂志等传统媒体，人们越来越多地通过网络来了解世界各地发生的事情。学生社团网站上发布了各类社团开展活动的通知和相关报道，学生可以在网站上了解丰富多彩的社团活动，并进行评论和交流。

请思考：

1. 学生社团网站需要满足哪些业务需求？
2. 在满足业务需求的基础上，如何对学生社团网站进行数据需求分析？
3. 针对数据需求分析建立数据管理与分析方案，其主要过程是什么？
4. 如何评价数据管理与分析方案并进行优化？

一、数据需求分析

数据需求分析是建立数据管理与分析方案的第一步，是确保数据管理与分析过程正确有效的首要条件。如果数据需求分析不清晰或者出现错误，会导致后面的过程出现问题。

数据需求分析需要对拟解决的问题进行详细分析，弄清楚问题的要求，包括需要输入什么数据、要得到什么结果、最后应以什么方式输出结果。

项　目　实　践　　学生社团网站数据需求分析

根据项目情境的描述，学生社团网站需要解决两个问题。一个问题是了解一周内发布的哪些文章的浏览量最高、评论量最高、转发量最高；另一个问题是网站向用户进行文章个性化推荐，方便学生更快地找到感兴趣的活动。

对于问题一，需要统计出一周内发布的所有文章的浏览量、评论量和转发量，并进行比较。因此需要输入文章数据，如文章的编号、标题、内容、发布时间、发布作者、浏览量、评论量、转发量，通过数据分析，将一周之内每天浏览量最高、评论量最高、转发量最高的文章找到，并用图表可视化方式显示。对于问题二，请思考需要输入的数据、输出的结果、输出方式，并把思考结果填入表 1.9。

表 1.9　学生社团网站数据需求分析表

解决的问题	需要输入的数据	输出的结果	输出方式
找到一周内每天浏览量最高的文章、评论量最高的文章、转发量最高的文章	一周内发布的所有文章的数据，包括文章的编号、标题、内容、发布时间、发布作者、浏览量、评论量、转发量等	一周内每天浏览量最高的文章、评论量最高的文章、转发量最高的文章	图表可视化方式
向学生推荐感兴趣的文章			

二、数据管理

数据管理是利用计算机硬件和软件技术对数据进行有效采集、存储、处理和应用的过程，其目的在于充分有效地发挥数据的作用。数据管理包括对结构化数据、半结构化数据以及非结构化数据的管理（详见第二章第一节）。

数据管理首先要进行数据采集，对数据需求分析中需要输入的数据进行采集，即需要明确数据来源，并利用合理的方式有目的地采集数据，这是保证数据管理与分析过程正确有效的基础。例如，在学生社团网站中，需要采集一周发布的所有文章的数据，这些数据可以从学生社团网站的数据库中导出。但是，如果没有权限，那么也可以编写网络爬虫程序从该网站上采集。采集数据时，应该在保证数据安全可靠的前提下，使采集到的数据尽可能全面、客观、具体、准确。

采集到的数据经过整理后需要进行存储和管理。目前，常用的数据管理方式是应用数据库管理数据。数据库可以对数据进行操作、备

份，并进行数据并发控制、安全性管理。除此以外，数据也可以通过文件系统进行管理。例如，可以利用分布式文件系统管理大数据。分布式文件系统是指文件系统管理的数据不一定在本地计算机上，这些数据可能存储在通过计算机网络连接的其他计算机上。

项 目 实 践　学生社团网站数据管理

将从学生社团网站中采集到的文章数据保存在一张二维表中，请思考该表应该包括哪些列并填写表 1.10。

表 1.10　文章数据表

1	2	3							
文章编号	文章标题	文章内容							

为学生社团网站用户推荐他们可能感兴趣的文章，需要生成用户文章评分数据，用户对文章的评分越高，表示用户对这篇文章越感兴趣。某篇文章的评分数据是从各个用户对该篇文章的浏览、转发、评论、收藏、点赞等行为数据中统计得到的。这些数据也可以保存在二维表中，在数据库中进行管理。表 1.11是用户对文章的访问数据表，包括浏览、转发、评论、收藏和点赞数据。表中的“是”表示用户进行了某种行为，例如表中的第一条数据表示用户 a没有浏览过文章 0001，也就没有转发、评论、收藏、点赞等行为，第二条数据表示用户 a浏览、评论并且收藏了文章 0002。根据表 1.11，可以计算各用户对文章的评分，计算规则如下：用户对某篇文章的初始评分为 0，用户对该文章的浏览、转发、评论、收藏和点赞的每个行为各计 1 分，用户对某篇文章的最高评分为 5分，如果用户没有浏览过该文章，那么评分为 0。请根据表 1.11计算出用户 a～ e对文章 0001～ 0005的评分并填入表 1.12。

表 1.11　用户对文章的访问数据表

文章编号	用户编号	浏览	转发	评论	收藏	点赞
0001	a					
0002	a	是		是	是	
0003	a	是	是	是		是
0004	a	是	是	是	是	是
0005	a					
0001	b	是				是
0002	b					
0003	b	是				

（续　表）

文章编号	用户编号	浏览	转发	评论	收藏	点赞
0004	b	是				是
0005	b					
0001	c	是				
0002	c					
0003	c	是	是	是		
0004	c	是				
0005	c	是	是	是	是	
0001	d	是		是	是	是
0002	d	是		是		是
0003	d	是				
0004	d					
0005	d	是				
0001	e	是				是
0002	e	是	是	是		是
0003	e					
0004	e					
0005	e	是	是	是	是	是

表 1.12　用户文章评分表

	文章 0001	文章 0002	文章 0003	文章 0004	文章 0005
用户 a					
用户 b					
用户 c					
用户 d					
用户 e					

三、数据分析

数据分析需要将采集到的数据进行整理、加工，然后再进行分析并转化为信息，帮助决策者进行科学决策。由于被分析的数据往往

有多个来源，并且数据类型多种多样，因此在分析前需要对数据进行预处理和整理，然后设计合理高效的数据分析方法，再利用数据分析工具对数据进行深入分析，并将分析结果可视化，以图表形式直观、美观、清晰地展示给用户。数据分析具有较强的专业性，目前普遍应用的数据分析工具中，以开源软件为主的有 Python 语言、R 语言等。

数据分析方法多种多样，需要根据数据的特征、数据量大小以及数据需求设计有效的数据分析方法。传统的数据分析主要使用数据统计技术，即从数据中抽取样本，通过统计方法对数据进行排序、筛选、汇总、统计等处理，从而得出一些有意义的结论。但是在面对巨大的数据量和计算量时，许多传统统计方法显得无能为力。这就需要使用新的数据分析方法，例如应用数据挖掘技术。数据挖掘可以利用算法帮助人们从大量的数据中提取隐藏的、人们事先不知道但是又潜在有用的信息。例如：关联规则挖掘算法可以从在线购物网站的大量订单数据中发现商品的潜在规则；协同过滤推荐算法可以从数据中发现购买者的消费行为，从而向购买者进行商品个性化推荐等。

在实际应用中，需要根据解决问题的不同，合理地应用数据分析方法，这样才能得到有效的分析结果，为科学决策提供支持。

项 目 实 践　学生社团网站数据分析

请利用数据分析中常用的统计分析法分别找出浏览量最高、评论量最高、转发量最高的文章。首先，需要将各篇文章的浏览量、评论量、转发量这些数据计算出来。请根据表 1.11统计出文章 0001～0005的浏览量、评论量、转发量，并填入表 1.13。

表 1.13　文章关注度数据表

文章编号	浏览量(次)	评论量(条)	转发量(次)
0001			
0002			
0003			
0004			
0005			

在采集各篇文章的浏览量、评论量和转发量数据时，通常还需要采集文章的发布日期和时间，这样可以分时间段统计出每天、每周、每月浏览量最高的文章、评论量最高的文章、转发量最高的文章。例如，对文章数据表（详见素材库，表 1.14所示为其一部分）中的数据，运用数据分析工具统计出一周内（6月 4日至 6月 10日）每天浏览量最高的文章，并通过图表可视化方式展现，如图 1.5所示。

表 1.14 文章数据表（部分）

发布日期	文章编号	浏览量(次)	评论量(条)	转发量(次)	发布时间
6月 4日	0001	120	10	9	23: 17
6月 4日	0002	360	30	26	22: 32
6月 4日	0003	210	27	26	21: 32
6月 4日	0004	420	49	47	14: 53
6月 4日	0005	130	14	13	13: 38
6月 4日	0006	140	14	13	13: 08
6月 4日	0007	170	21	19	10: 38
6月 4日	0008	190	16	16	8: 54
6月 4日	0009	210	16	15	8: 53
6月 4日	0010	230	27	27	7: 37
6月 5日	0011	100	11	10	22: 34
6月 5日	0012	820	61	53	22: 18
6月 5日	0013	280	33	32	22: 18
6月 5日	0014	300	87	39	22: 17
6月 5日	0015	130	16	14	19: 33
6月 5日	0016	120	9	7	19: 19
6月 5日	0017	300	49	37	17: 32
6月 5日	0018	110	11	9	17: 32
6月 5日	0019	110	7	6	17: 18
6月 5日	0020	170	19	17	13: 04
6月 5日	0021	180	24	23	11: 35
6月 5日	0022	360	49	39	9: 32
6月 5日	0023	639	78	80	7: 31
6月 5日	0024	170	15	13	3: 16
6月 5日	0025	118	325	113	1: 01

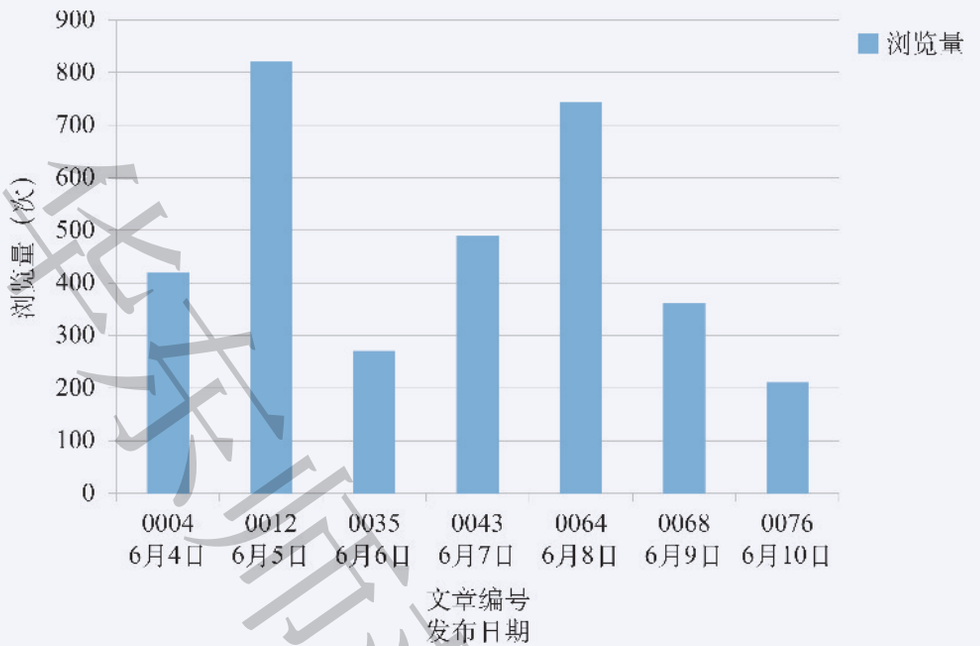


图 1.5　单日最大浏览量对比图

请你对文章数据表中的数据进行数据分析，分别找出 6月 4日至 6月 10日间每天评论量最高、转发量最高的文章，并用图表展现。然后，通过分析结果观察浏览量最高的文章是否也是评论量最高或者转发量最高的文章，思考这三个数据之间有没有什么关系。

探　究　活　动

上述统计方法可以帮助我们直观了解每天最受关注的活动文章，但是没有办法为学生推荐其可能感兴趣的文章。通常可以利用推荐算法来进行文章的个性化推荐。推荐算法通过分析用户对物品的评分数据，推测出用户可能喜欢的物品。例如，利用推荐算法可以进行在线网络购物系统中的商品推荐、云音乐软件中的音乐推荐、新闻网站中的新闻推荐等。推荐算法主要包括协同过滤推荐算法、基于内容的推荐算法、基于知识的推荐算法以及混合推荐算法。

以基于物品的协同过滤推荐算法为例，在在线购物系统中，该算法可以根据用户浏览或购买过的物品的记录，向用户推荐与之相似的物品。该算法首先通过用户对浏览或购买过的物品的评分，计算出物品之间的相似度，再根据物品的相似度和用户行为，预测用户对没有浏览或没有购买过的物品的评分。用户是否喜欢一个物品通过评分表示，评分越高表示用户越喜欢这个物品，因此可以推荐预测评分高的物品给用户。例如，根据表 1.11可知用户 a没有浏览过文章 0001和 0005，因此在表 1.12用户文章评分表中，用户 a对文章 0001和 0005的评分均为 0。应用基于物品的协同过滤推荐算法可以预测用户 a对文章 0001和 0005的评分，预测评分高的文章可能就是用户更关注的文章。因此，可以优先选择预测评分高的文章向用户 a推荐。通过推荐算法这种数据分析方法，不仅可以实现文章的个性化推荐，而且由于网站用户更容易找到自己关注的文章，所以也将提高文章的浏览量和转发量。

请以实现学生社团网站的文章个性化推荐为目标，分组调研一种具体的推荐算法及其实现过程，讨论并学习推荐算法如何向不同的网站用户推荐他们可能感兴趣的文章。

四、数据管理与分析方案的评价和优化

方案评价和优化贯穿于数据管理与分析方案的整个过程中，在每个环节完成后都应该进行该环节的方案评价，如果发现问题，需要立即进行改进和优化。如果整个方案完成后才进行评价和优化，那么一旦中间某个环节有问题，将会导致该环节以及其后各环节的方案都需要进行修改。例如，数据需求分析完成后应该随即进行评价和优化，如果发现问题，可以针对问题进行改进和优化，直至没有问题后再进行数据采集。

方案评价和优化需要根据不同的应用展开。针对不同的过程，评价和优化的方法有多种，主要可以从以下四个方面进行评价：

1. 数据需求目标评价

数据需求分析是否可以解决需要解决的问题，是否可以达到既定目标。

2. 数据真实性和有效性评价

（1）采集数据的目的是否明确。

（2）数据来源以及采集到的数据是否全面、是否真实可信、是否完整、是否合乎法律和伦理要求。

3. 方案合理和有效性评价

（1）数据管理方案是否合理、是否具有扩展性，数据库管理系统选择是否合适。

（2）数据分析方法是否正确高效、是否选择了有效的数据分析工具，分析结果是否可以为用户提供服务和决策支持。

4. 方案安全性和风险性评价

整个数据管理与分析方案是否将风险控制在可接受的范围内，是否符合相关法律法规、标准规范以及伦理要求。

体　验　思　考　　**学生社团网站数据管理与分析方案的评价和优化**

不同应用和解决方案的评价方法不同。一般情况下，首先需要对采集数据的真实性、有效性进行评价。例如，在浏览文章时，有些用户可能因点击错误而打开了文章页面，或者打开页面后发现自己不感兴趣而立即关闭页面。在提取用户浏览数据时，如果根据用户是否打开文章页面来统计，则可能和真实的用户行为有偏差。因此，可以对数据采集方式进行优化，采集用户在页面上的停留时间来判断用户是否浏览了该文章。

在数据管理方案评价和优化中，经常需要对数据组织进行评价。例如，为了提高数据表的规范性和数据完整性，需要对表的结构进行评价和优化。在数据分析方案优化方面，就学生社团网站而言，考虑到只对每天浏览量最高的文章进行统计不够全面，还可以改进方案，对每天浏览量排名前三或者前五的文章进行统计分析。

对于学生社团网站的数据管理与分析方案，除了以上的评价和优化，请分组讨论是否还有其他的评价方法，并尝试对方案进行评价。如果发现了问题，请思考可以用哪些方法对方案进行优化。

五、科学决策

在信息社会中，决策者改变了只依靠知识、经验、思想来决策的传统方式，他们更多地依靠数据分析的结果来进行科学决策，增强了决策的科学性。科学决策并不直接使用数据，而是以数据分析后提取出来的信息为支撑。例如，企业可以通过科学的数据分析方法将产品数据、市场数据、用户数据、项目财务数据等数据转化为可利用的信息，以有利于制定精准的营销方案。又如，城市公交数据分析平台可以对线路站点客流、出行时间段特征、出行次数、出行距离、换乘等数据进行综合分析，判断公交负载效率和营运水平，从而在线路规划、高峰大站车安排、排班调整、运营时间等方面给出优化建议。

作业练习

在信息社会，一切皆可数据化，包括学生的学习过程。请同学们以小组为单位，针对在线学习系统中的某一个问题，设计数据管理与分析方案，并对其进行评价和优化。

知 识 延 伸

数据分析的分类

数据分析通常可以分为四类，即描述性分析、诊断性分析、预测性分析和规范性分析。

1 描述性分析

描述性分析是最常见的一类数据分析，它主要采用数学统计方法对已经发生的事情进行描述和统计。例如，一个在线购物网站每个月完成多少订单、退货多少、利润多少等。找出学生社团网站中每天浏览量最高的文章也是一种描述性分析。描述性分析的结果通常用数据可视化工具来呈现。

2 诊断性分析

通过评估描述型数据，诊断性分析能够深入分析问题的核心原因，即回答“为什么”。诊断性分析主要采用关联分析法和因果分析法。例如，在线购物系统可以对顾客经常一起购买的商品进行关联分析，从而更了解顾客的购买行为，进行商品联合促销。对用户浏览过的文章数据进行分析，推荐他们可能感兴趣的文章，这也是一种诊断性分析。诊断性分析的基础是描述性分析。

3 预测性分析

预测性分析主要采用分类分析、趋势分析等方法，通过建立数据模型对未来进行预测。例如，企业对消费者未来的消费趋势进行预测，企业根据商品的历史销售量预测未来销售量等。预测性分析的基础是描述性分析和诊断性分析。

4 规范性分析

规范性分析主要利用指导性模型对“发生了什么”“为什么发生”以及一系列“可能发生什么”进行分析，帮助用户确定最佳方案。例如，地图软件综合分析路线的长度、公交车的预计到达时间、道路的拥堵情况、实时的交通限制等，帮助人们规划出行路线。规范性分析是数据分析的高级阶段。

（参考资料：《数据科学理论与实践》，朝东门编著）

第二章

数据管理

本章学习目标

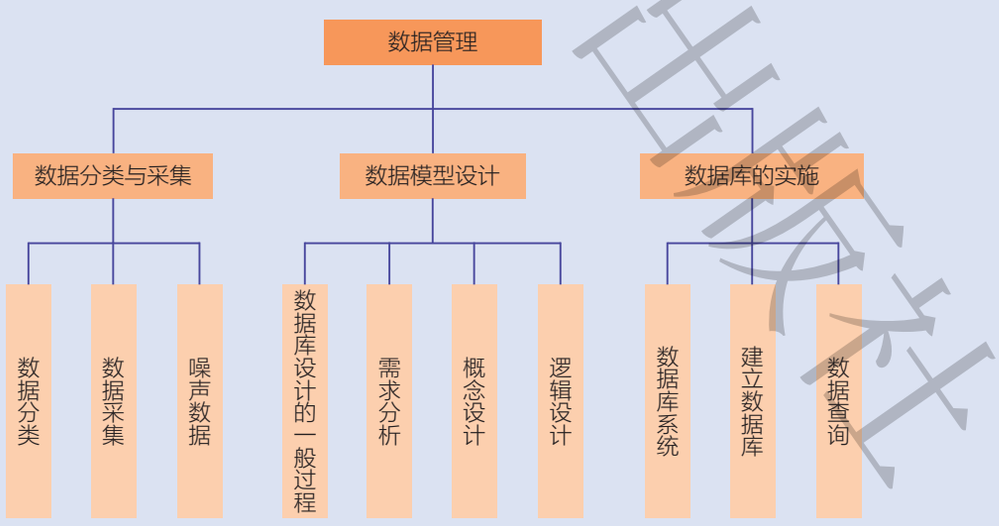
- 了解数据采集途径的多样性，能利用适当的工具对数据进行采集和分类；认识噪声数据的现象和成因；理解不同结构化程度数据的区别，以及在管理与应用上的特点。
- 了解关系数据模型的基本概念，掌握设计简单关系数据库的逻辑结构的方法，设计一个简单的网上书店数据库。
- 使用数据库管理系统建立关系数据库，了解数据库基本的数据查询方法，能使用结构化查询语言进行简单的数据查询，创建和使用一个简单的网上书店数据库。

信息社会中，各个领域产生的数据正以惊人的速度增长，人们往往需要对这些数量巨大、种类繁多的原始数据进行加工处理，从中获取更有价值的信息，作为决策的依据。在数据处理中，需要对数据进行管理，例如采集、分类、组织、编码、存储、维护和查询等。

随着数据管理技术的发展，目前在计算机管理数据过程中普遍使用数据库技术。数据存储在数据库中，利用数据库管理系统可以对数据库进行创建、修改、查询等操作。关系数据库是基于关系模型的数据库，采用二维表的形式存储数据，是目前使用最广泛的数据库。关系数据库管理系统是基于关系模型的数据库管理系统，如 Access、MySQL、SQL Server、Oracle等。

数据库技术已经在各行各业中得到了广泛的应用，图书馆管理、仓储物流管理、网络购物、网上订票、证券交易等系统中一般都使用了数据库技术。随着大数据时代的到来，许多新的数据存储方式应运而生，NoSQL（not only SQL）、分布式数据库、云存储等都可以较好地应用于大数据时代的各种数据管理中。

本章知识结构



项目主题 网上书店数据管理

项·目·情·境

随着互联网的普及，网络购物日益成为一种重要的购物形式，渗透到人们的日常生活中。学校开设了一门软件开发研究型课程，同学们学习软件开发技术时，有一组同学对开发一个购物网站非常感兴趣。在讨论如何开发这个网站的过程中，他们发现需要思考并解决“如何管理数据”等问题。

为了完成开发这个网站的任务，该小组同学首先需要通过探索一个网上书店实例，研究网上书店的数据构成；然后，设计一个简单的网上书店数据库，用以实现网上购书的数据管理；最后，选择一个数据库管理系统来创建和使用这个简单的网上书店数据库。

项·目·任·务

任务 1	任务 2	任务 3
通过多种途径，采集生活中网上书店的图书数据。	对网上书店数据库进行需求分析，确定实体及实体间的联系类型，并建立关系数据模型设计出一个简单的网上书店数据库。	根据任务 2中的设计，使用 MySQL数据库管理系统创建网上书店数据库，并使用 SQL语句在该数据库中查找需要的数据。

第一节　数据分类与采集

信息技术的迅猛发展，使得数据的采集途径越来越多样化，数据的来源和种类也越来越多样化，网络信息系统、网站日志、科学实验、智能设备、大量的传感器等都成为数据的重要来源，数据的种类包括文档、日志、网页、音频、E-mail、视频、图片、地理位置信息等。随着技术的发展，新的数据种类还会不断涌现。

问　题　思　考

日常生活中，每天都有大量的数据产生，这些数据的来源和表现形式多种多样。

请思考：

1. 数据根据结构化程度的不同可以分为哪几类？请举例说明。
2. 数据的采集途径有哪些？请尝试使用不同的途径采集生活中的数据。
3. 什么是噪声数据？请举例说明你在数据采集活动中遇到过哪些噪声数据。

一、数据分类——结构化数据、半结构化数据和非结构化数据

现实生活中所产生的大量数据，根据不同的数据来源、不同的数据使用需求，往往会以各种不同的格式出现。例如：学生考试的成绩数据、企事业单位的财务数据、网络购物的交易数据等，往往以二维表的形式表示并存储在关系数据库中；电话服务部门一般需要对工作人员与客户的通话进行录音并保存，录音产生音频格式的数据；安全监控通常需要用摄像头对监控区域进行录像并保存，录像产生视频格式的数据。

这些不同的数据格式代表了不同的数据组织结构。根据数据结构化程度的不同，可以将数据分为结构化数据、半结构化数据和非结构化数据。二维表数据属于结构化数据；音频、视频等通常以二进制文件格式存储，属于非结构化数据。

结构化数据遵循一个标准的模型，这个模型一般是指关系数据模型，又称关系模型。在关系模型中，数据的结构用二维表来表示，一个关系就是一张二维表。二维表由行和列构成，如表 2.1 图书信息表所示。关系数据库是指采用关系模型来组织数据的数据库，即用二维表

的形式来存储数据的数据库。因此，结构化数据通常存储在关系数据库中。

结构化数据先有结构，再有数据。因此，使用关系数据库存储结构化数据，首先需要定义二维表的结构，即确定表的列数，以及每一列的列名、列宽和数据类型等，然后才能在表中输入数据。如果存储数据的需求发生变化，可能需要修改表的结构。例如，如果要在表 2.1 中增加图书的简介，就必须先修改表的结构，增加一列“图书简介”后才能输入数据。

表 2.1　图书信息表

书号	书名	定价	作者	出版社	出版日期	折扣
1103028726	数据库技术 **	36.00	朱 ** 、张 **	** 教育出版社	2017- 08- 01	0.8
1103134637	大数据 **	80.00	朱 **	** 技术出版社	2018- 01- 01	0.8
1101756339	PHP+ MySQL **	89.80	** 联盟	** 大学出版社	2013- 09- 01	0.8

非结构化数据不遵循统一的模型，即没有固定的数据结构，如图像、音频、视频、各类文档等都属于常见的非结构化数据。非结构化数据可以使用非关系数据库存储。随着互联网的发展，新产生的数据绝大部分都是非结构化数据。

半结构化数据介于结构化数据和非结构化数据之间，它有一定的结构，但又不符合二维表的表示形式，一般使用相关标记（标签）来分隔每项数据。半结构化数据通常存储在文本文件中，常见的有 XML 数据和 JSON 数据等。如下所示为使用 XML 格式存储图书数据：

```
〈books〉
    〈bookID〉1103028726〈/bookID〉
    〈title〉数据库技术 ** 〈/title〉
    〈author〉朱 ** 、张 ** 〈/author〉
    〈publisher〉 ** 教育出版社〈/publisher〉
    〈price〉36.00〈/price〉
    〈pubdate〉2017-08-01〈/pubdate〉
    〈discount〉0.8〈/discount〉
〈/books〉
```

半结构化数据可以简单理解为先有数据再有结构。通过这样的数据格式，可以自由地表达很多有用的信息。所以，半结构化数据的扩展性很好。例如，如果要在如上使用 XML 格式存储的数据中增加图书的简介，可以在标记〈books〉和〈/books〉之间的任意位置增加如下内容：

```
〈abstract〉《数据库技术＊＊》分为三个部分：(一)基础原理；
(二)方法与设计；(三)问题求解……〈/abstract〉
```

体 验 思 考

1 请举出生活中结构化数据、半结构化数据和非结构化数据的例子。

2 查阅 XML数据和 JSON数据的有关资料，深入了解用这两种格式存储数据的方法。

二、数据采集

数据采集的途径有很多，常用的有人工采集、传感器采集、网络爬虫采集和数据库采集等。可以根据采集数据的需要，选择合适的数据采集途径。

1. 人工采集

人工采集是一种传统的数据采集途径，指通过人工观察、调查、访谈等方式进行数据采集。人工采集可以根据需要精准地采集数据，但效率低、成本高、工作量大。

2. 传感器采集

传感器是一种检测装置，它如同人的五官，能感受到被检测目标的数据。通常情况下，传感器以一定的频率采集数据，并将数据发送至相应的数据接收端。传感器广泛应用于工业生产、环境保护、气象观测、资源探测、医疗诊断、交通运输等社会各个领域，对检测目标进行测量、监测、定位、跟踪、导航等。

例如，智慧交通离不开实时监测城市道路交通的状况，这需要安

装在道路上的各种传感器自动、连续地采集不同地点和路段上的实时交通流量数据。

3. 网络爬虫采集

网络爬虫采集主要是指通过网络爬虫或网站公开 API 等方式，从网络上获取公开或授权的数据。它支持图片、音频、视频、附件等数据的采集。

网络爬虫采集数据的一般过程是：

（1）将需要从中抓取数据的网站的 URL 写入 URL 队列；

（2）爬虫从 URL 队列中获取需要抓取数据的网站的 URL；

（3）爬虫从 Internet 上抓取对应网页内容，并抽取出其特定属性的内容值；

（4）爬虫将从网页中抽取出来的数据写入数据库；

（5）数据处理模块对爬虫抓取的数据进行处理；

（6）数据处理模块将处理之后的数据写入数据库。

4. 数据库采集

对于存储在数据库中的科学研究数据、企事业单位的生产经营数据等，可以与单位或个人进行合作，经过授权后，通过从数据库中提取获得需要的数据，此为数据库采集。

探 究 活 动　采集网上书店中的图书数据

新华书店在互联网发展的浪潮下，面对传统书店的激烈竞争，推出了“新华一城书集”网上书店（如图 2.1所示），鼓励读者在朋友圈中传播好书，借此给城市带来更浓书香氛围，助推全民阅读大潮。请在“新华一城书集”网上书店中分别利用人工采集和网络爬虫采集这两种途径采集图书数据。

1 利用人工采集，在“新华一城书集”网上书店中采集你喜爱的图书的相关数据，并填写在表 2.2中，组织成结构化数据。（折扣 = 商城价 定价）

图 2.1 “新华一城书集”网上书店

表 2.2 图书

书号	书名	定价	作者	出版社	出版日期	折扣

2 利用网络爬虫采集，在“新华一城书集”网上书店的商品列表中采集图书类别为“计算机”的图书数据，要求采集前 20页商品列表中图书的书名、商城价、市场价(定价)。采集这些数据的 Python网络爬虫程序如下：

```
import requests      #导入 Python 网络爬虫库 requests
from lxml import html      #lxml 是 Python 的一个解析库，支持 HTML 和 XML 的解析
c=0
for i in range(1,21):
    url = 'https://www.bookmall.com.cn/shop/index.php?act=search&op=index&cate_id=5561&curpage='+str(i)      #需要爬取数据的网址，“5561”表示“计算机”类图书
    page=requests.Session().get(url)      #调用 get()方法获取网页源代码
    tree=html.fromstring(page.text)      #对返回的网页源代码进行处理，方便使用 xpath 定位
    bookinfo=tree.xpath('//div[@class="goods-name"]//a/@title | //div[@class="goods-price"]//em/@title')      #爬取图书的书名、商城价、市场价(定价)数据
    for j in range(0,len(bookinfo),3):
        c=c+1
        print(c,bookinfo[j],"/",bookinfo[j+1],"/",bookinfo[j+2])      #输出爬取的数据
```

数据采集结果示例如图 2.2所示：

```
1 Spring Cloud 微服务:** / 商城价：¥ 63.20 / 市场价：¥ 79.00
2 网络媒体**/ 商城价：¥58.00 / 市场价：¥58.00
3 **趣味编程 / 商城价：¥ 47.20 / 市场价：¥ 59.00
4 **网络爬虫** / 商城价：¥ 39.20 / 市场价：¥ 49.00
5 深度学习**/ 商城价：¥ 47.20 / 市场价：¥ 59.00
……
479 自学 Python** / 商城价：¥ 63.20 / 市场价：¥ 79.00
480 游戏 AI **/ 商城价：¥ 47.20 / 市场价：¥ 59.00
```

图 2.2 数据采集结果示例

请参照采集“计算机”类图书数据的途径，利用网络爬虫采集其他类别图书的数据。

三、噪声数据

在采集到的数据中往往会有一些不符合要求的、无意义的、错误或异常的数据，这类数据通常称为噪声数据。例如，商品的价格数据一般都是数值型，而网络爬虫在“新华一城书集”网站上采集到的每一本图书的“商城价”和“市场价”数据中都分别包含有字符“商城价：￥”和“市场价：￥”（如图 2.2 所示），这些字符不需要出现在图书的价格数据中，在存储商城价和市场价数据时，这些字符会被去除。

产生噪声数据的原因很多，有设备原因、技术原因、人为原因等。如：计算机设备出现硬件故障；数据传输过程中出现错误；数据采集工具出现问题；数据输入时出现错误等。

噪声数据是不能够被接受的，因此需要对采集的数据进行处理，去除噪声数据，以保证数据的质量和可靠性，为后期的数据使用和分析打下良好的基础。

第二节　数据模型设计

计算机不能直接处理现实世界中的事物，必须先把具体事物转换成计算机能够处理的数据。数据模型是对现实世界客观事物及其联系的数据描述，描述的内容包括数据结构、数据操作和数据的约束条件。数据结构描述了数据及数据之间的关系；数据操作定义了对数据对象允许执行的各种操作；数据的约束条件定义了数据模型中的数据及数据之间的关系应具有的制约规则，以确保数据的正确、有效。

问　题　思　考　　**网上书店数据库设计**

数据库设计是数据库应用系统开发的核心问题，只有对数据库进行合理的设计才能保证开发的数据库应用系统能够高效地运行、有效地存储数据，以满足用户的应用需求。

请思考：

1. 如何对网上书店数据库进行需求分析？
2. 如何确定网上书店中实体与实体间的联系类型？
3. 如何建立网上书店关系模型？

一、数据库设计的一般过程

在数据库规范化设计方法中，一般将数据库设计分为四个阶段：需求分析阶段、概念设计阶段、逻辑设计阶段和物理设计阶段。

需求分析阶段，需要认真细致地了解用户的各种需求，在此基础上确定系统的功能。概念设计阶段，需要将现实世界的问题用概念模型来表示。概念模型是按用户的观点对数据建模，是信息世界中数据特征的描述。它概念简单、清晰，易被用户理解，且不依赖于具体的计算机系统。逻辑设计阶段是将概念设计阶段形成的概念模型转换为某个具体的数据库管理系统支持的数据模型。数据模型是按计算机的观点对数据建模，是机器世界中数据之间关系的描述。它有严格的形式化定义，以便于在计算机中实现。在传统数据库领域中较常见的数据模型有层次模型、网状模型和关系模型。物理设计阶段，需要在计算机的物理设备上确定应采取的数据存储结构和存取方法等问题。

本章主要介绍数据库设计中的需求分析阶段、概念设计阶段和逻辑设计阶段。

二、需求分析

简单地说，需求分析就是充分地收集和分析用户的需求，了解用户需要数据库做些什么，实现什么功能。需求分析是数据库设计的第一步，需求分析结果能否完整、准确、全面地表达用户的需求，将直接影响数据库的设计质量。

项 目 实 践　　网上书店需求分析

生活中各种网上书店往往设计了很多功能，为用户提供服务，满足用户的各种需求。例如，“新华一城书集”网上书店为个人用户设计的功能如图 2.3所示。

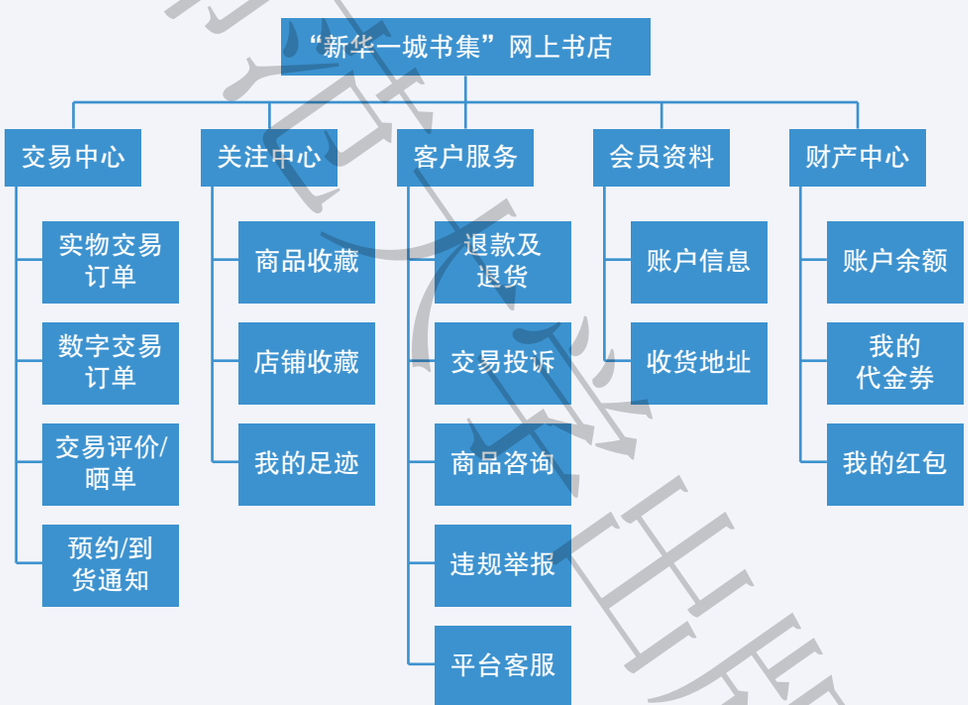


图 2.3 “新华一城书集”网上书店为个人用户设计的功能

请参照“新华一城书集”网上书店中用户订书生成交易订单的功能，构建一个自己的网上书店数据库，对实现网上订书进行需求分析。

用户在“新华一城书集”网上书店中订书，一般需要完成以下四个步骤：

1 注册用户：用户在网站上进行注册，填写个人基本信息并设置密码。

2 登录网站：用户使用注册的用户名和密码登录网站。

3 选购图书：用户可以按图书类别或用关键词搜索等方式查找图书，根据查找到的图书信息判断图

书是否符合自己的需求。

4 生成订单：将符合需求的图书放入购物车，生成订单，如图 2.4所示。

图 2.4 交易订单

网上订书过程中涉及的信息主要有：用户信息、图书信息、图书类别信息、订单信息等。

请分析“新华一城书集”网上书店的用户、图书、图书类别、订单分别登记了哪些信息，填写在表 2.3中。

表 2.3 “新华一城书集”网上书店订书需求分析

	登记的信息
用户	
图书	
图书类别	
订单	

“新华一城书集”网上书店功能设计比较全面，登记的信息也比较详细。参照该网站设计一个自己的网上书店数据库，实现网上订书功能，需求分析如下：

1 用户管理

要对网上订书的用户进行管理，必须要求用户注册，登记用户的详细信息，如用户名称、密码、邮箱、姓名、性别、生日、地址等。注册时，用户名称不允许重复，如果填写的用户名称已经被占用了，系统会给出提示，要求重新填写用户名称。

2 图书管理

对网上销售的图书进行管理，登记图书的详细信息，如书号、书名、定价、作者、出版社、出版日期、折扣等，其中书号不允许重复。用户可以根据登记的图书信息了解图书。

3 图书类别管理

“新华一城书集”网上书店的图书分类方法参考了《中国图书馆分类法》——我国图书馆和情报单位普遍使用的综合性的图书分类法。同时，为了方便用户查找图书，对该分类方法做了一些改进，如图书大类分为计算机、哲学、政治、法律、军事、经济等，计算机大类又分为计算机硬件、计算机类大学教材、网络技术安全等多个小类。

在设计自己的简单网上书店数据库时，可以将图书按大类进行分类，在图书类别中登记类别编号和类别名称。

4 订单管理

订单管理是网上书店的工作重点，包括订单的建立、取消等，因此需要登记的订单信息有订单号、下单时间、交易状态等，其中订单号是每张订单的唯一标识，数据不会出现重复，用户可以根据订单号查询订单信息。每张订单可以用于订购一本或多本图书。

三、概念设计

概念设计需要对用户需求进行综合、归纳和抽象，确定所要研究的事物，找出事物的属性以及事物之间的联系，建立起一个独立于特定数据库管理系统的概念模型。

概念模型常用的设计方法是实体—联系方法，该方法是直接从现实世界中抽象出实体和实体间的联系，并用 E-R 图来表示。E-R 图，即实体—联系模型图，是将实体、实体的属性、实体间的联系用图形的方式描述，使之更为直观。

1. 实体与实体的属性

（1） 实体

实体是客观存在且相互区别的事物，例如用户、图书、订单等。实体可以是具体的人、事、物，也可以是抽象的概念或事件。

（2） 实体的属性

每个实体都可以用一组数据来描述其特性，实体的属性是描述实体特性的数据。例如，用户实体由用户名称、密码、邮箱、姓名、性别、生日、地址等属性组成。

能够唯一地标识某一个实体的属性（或几个属性的组合）称为该实体的关键字。例如，用户注册的用户名称是不允许出现重复的，具有唯一性，因此用户名称可以作为关键字；而用户的姓名、性别等都有可能出现重复，所以不能作为关键字。

2. 实体间的联系

在现实世界中，事物之间是有联系的，这些联系在信息世界中反映为实体间的联系。实体间的联系可以分为三类：一对一联系（1∶1）、一对多联系（1∶N）和多对多联系（M∶N）。

（1）一对一联系（1∶1）

如果实体 A 中的每个实例在实体 B 中至多有一个实例与之联系，反之，实体 B 中的每个实例在实体 A 中也至多有一个实例与之联系，则称实体 A 与实体 B 具有一对一联系。例如，一个班级只有一位班主任，一位班主任只带一个班级，班级与班主任之间存在一对一联系。

（2）一对多联系（1∶N）

如果实体 A 中的每个实例在实体 B 中有 $n(n\geqslant 0)$ 个实例与之联系，反之，实体 B 中的每个实例在实体 A 中只有一个实例与之联系，则称实体 A 与实体 B 具有一对多联系。例如，一个班级有多名学生，一名学生只属于一个班级，班级与学生之间存在一对多联系。班级是“一方”，学生是“多方”。

（3）多对多联系（M∶N）

如果实体 A 中的每个实例在实体 B 中有 $n(n\geqslant 0)$ 个实例与之联系，反之，实体 B 中的每个实例在实体 A 中有 $m(m\geqslant 0)$ 个实例与之联系，则称实体 A 与实体 B 具有多对多联系。例如，一个班级有多位任课教师，每位教师可以任教多个班级，班级与教师之间存在多对多联系。

项目实践　确定网上书店中的实体与实体间的联系类型

根据网上书店的需求分析，对其中的事物进行归纳和抽象，进行数据库的概念设计，确定实体和实体的属性、实体间的联系。

1 分析网上书店中订单、图书、图书类别三个实体的属性和关键字，填写在表 2.4中。

表 2.4　实体及其属性

实体的名称	实体的属性	关键字
用户	用户名称、密码、邮箱、姓名、性别、生日、地址	用户名称
订单		
图书		
图书类别		

2 分析网上书店中的用户、订单、图书、图书类别四个实体之间的联系，填写在表 2.5中。

用户与订单之间的联系：一个用户可以有多张订单，但一张订单只属于一个用户，用户与订单之间存在一对多联系。

订单与图书之间的联系：一张订单可以订购多种图书，一种图书可以由多张订单订购，订单与图书之间存在多对多联系。

图书类别与图书之间的联系：一种图书类别包含多种图书，一种图书只属于一种图书类别，图书类别与图书之间存在一对多联系。

表 2.5 实体间的联系类型

实体 A	实体 B	联系类型
用户	订单	一对多

四、逻辑设计

逻辑设计的任务是要将概念模型转换为某个具体的数据库管理系统支持的数据模型，然后建立用户需要的数据库，把数据组织起来存入计算机。数据模型中目前使用较广泛的是关系模型。

关系模型是用二维表的形式来表示实体与实体间联系的数据模型。二维表又称为关系。如图 2.5 所示，在二维表中，每一列表示实体的一个属性，又称为一个字段。列可以命名，称为列名、属性名或字段名。每一行数据称为一个元组或一条记录。

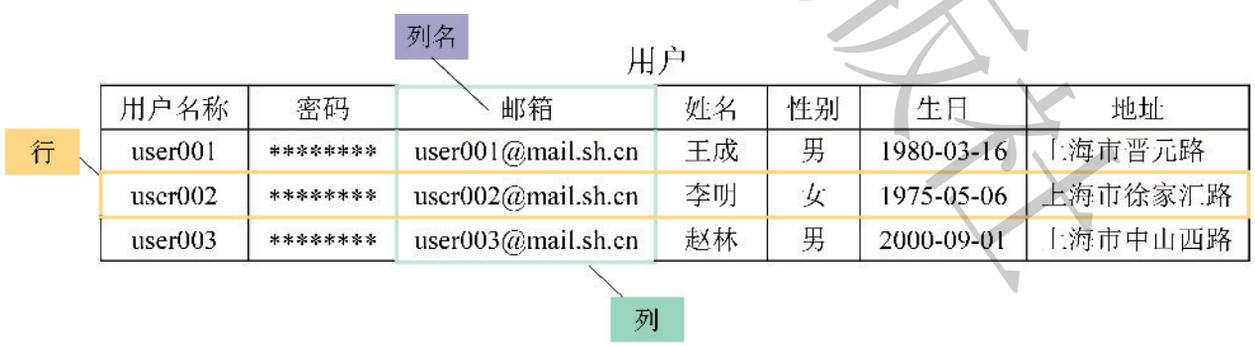


图 2.5 用二维表表示用户实体

在二维表中，可能存在一个或多个关键字，从中选择一个作为主关键字，称为主键。

建立关系模型需要将实体、实体的属性、实体间的联系转换为二维表。

1. 将实体与实体的属性转换为二维表

每个实体转换为一张二维表。实体的名称可以作为二维表的名称，二维表的每一列表示实体的一个属性。用二维表表示用户实体如图 2.5 所示。实体名“用户”转换为表名“用户”，实体的属性名转换为列名——用户名称、密码、邮箱、姓名、性别、生日、地址。

2. 将实体间的联系用二维表来实现

二维表之间可以通过公共属性建立关系。公共属性通常要求作用相同、数据类型一致，属性名可以不同。有两种方法：

（1） 在一张表中加入另外一张表的关键字

如果两个实体间的联系是一对一联系，可以从对应的两张二维表中，任意选择一张，在其中加入另外一张的主键，或其他关键字。

如果两个实体间的联系是一对多联系，需要在对应的两张二维表中，选择“多方”表，在其中加入“一方”表的主键，或其他关键字。

（2） 定义一张新的二维表

如果两个实体间的联系是多对多联系，需要将联系类型转换成一张二维表。这张二维表中的字段包括：两个实体对应的两张二维表中的主键（或关键字）、联系本身的属性。

项 目 实 践　　建立网上书店关系模型

用关系模型表示网上书店中的实体和实体的属性、实体之间的联系。

1 建立“用户”表与“订单”表之间的关系

“用户”表与“订单”表之间具有一对多关系，“用户”表是“一方”，“订单”表是“多方”，因此可以在“订单”表中加入“用户”表的主键“用户名称”，如图 2.6所示。两张表之间通过公共属性“用户名称”建立关系。

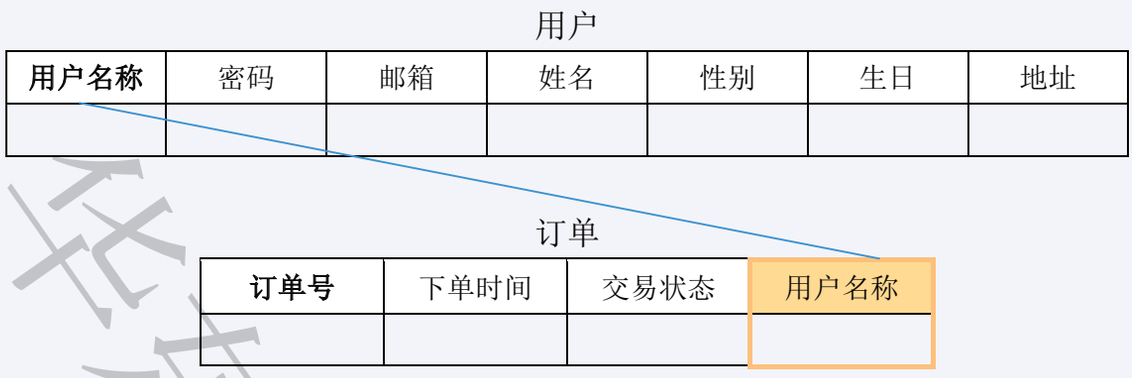


图 2.6　建立“用户”表与“订单”表之间的关系

2　建立“订单”表与“图书”表之间的关系

“订单”表与“图书”表之间具有多对多关系，因此需要定义一个新的二维表来实现两张表之间关系的建立。

如图 2.7所示，建立一张新的二维表“订单明细”表，在“订单明细”表中加入“订单”表的主键“订单号”和“图书”表的主键“书号”，并且还可以增加一些与订单明细相关的信息，如图书的购买数量等。“订单”表与“图书”表之间通过“订单明细”表建立起关系。

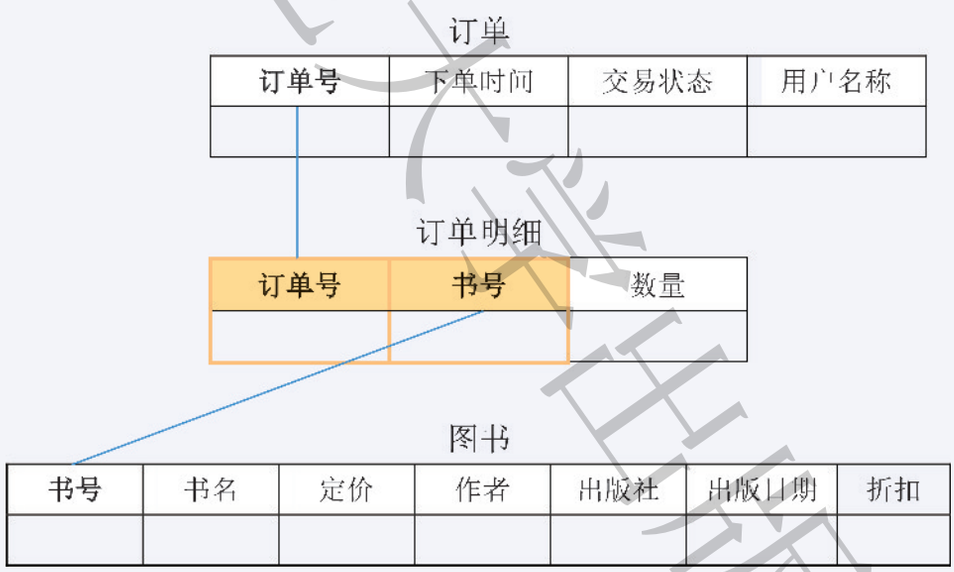


图 2.7　建立“订单”表与“图书”表之间的关系

体　验　思　考

1　网上书店中还有哪些实体？请分析这些实体的属性及实体间的联系类型。

2　如何建立“图书”表与“图书类别”表之间的关系？

第三节　数据库的实施

完成了数据库的结构设计后，就进入数据库的实施阶段。数据库的实施就是根据数据库设计，使用一个具体的数据库管理系统在计算机上创建和使用数据库。

结构化查询语言（structured query language，缩写为 SQL）是用户操作关系数据库的国际标准语言，可以完成数据库的定义、查询、更新、维护、控制等一系列操作。

问　题　思　考　　**网上书店数据库建立与查询**

使用一个具体的数据库管理系统在计算机上创建和使用网上书店数据库。

请思考：

1. 如何使用 MySQL 数据库管理系统创建网上书店数据库？
2. 如何使用 SQL 语句在网上书店数据库中，根据自己的需要查找数据？

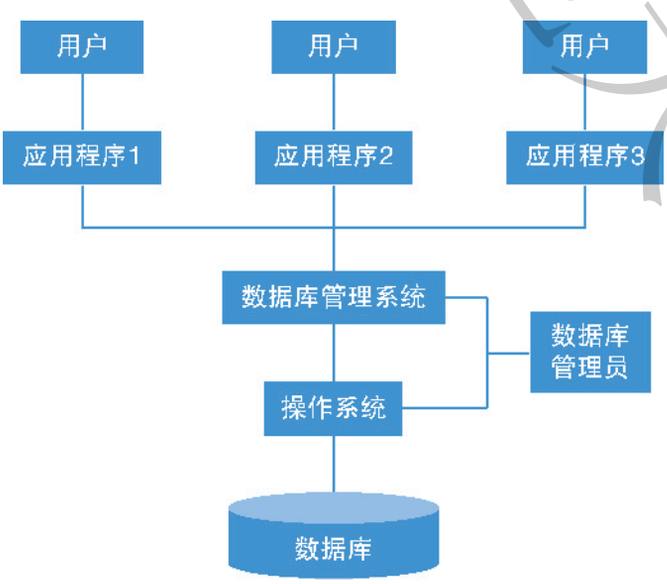


图 2.8　数据库系统

一、数据库系统

数据库系统是基于数据库技术的计算机应用系统，它由计算机系统、数据库、数据库管理系统、应用程序和用户等组成。在数据库系统中，所有的数据都存储在数据库中，应用程序可以通过数据库管理系统访问数据库，如图 2.8 所示。数据库系统的运行需要计算机软硬件环境的支持，同时还要有使用数据库系统的用户和数据库管理员。

数据库（database，缩写为 DB）是用来存储数据的，它是为某一特定主题或某一特定目的而存放于外存储器的相互关联的数据的集合。数据所包含的数据的结构与数据一起存储在数据库中，通过数据库管理系统，用户可以方便地访问数据库中的数据。

数据库管理系统（database management system，缩写为 DBMS）是对数据库中的数据进行存储、处理和管理的系统软件。当用户向数据库发出访问请求后，数据库管理系统接受、分析该用户的请求，并根

据用户请求去查询、存储、更新数据库中的有关数据。常见的数据库管理系统有 Access、MySQL、SQL Server、Oracle 等，这些数据库管理系统都基于关系模型，又称为关系数据库管理系统。

MySQL 是开放源代码的关系数据库管理系统，于 20 世纪 90 年代问世。由于 MySQL 的早期定位主要面向互联网开发，因此其应用实例也大都集中于互联网方向。MySQL 不仅简便易用，支持高性能、可扩展的基于 Web 的数据库应用，而且还提供了一整套数据库驱动程序和可视化工具，可帮助开发人员和数据库管理员自主构建 MySQL 应用，因此深受广大用户的喜爱。同时，由于 MySQL 的开源特性，针对一些对数据库有特别要求的应用，用户可以通过修改代码来实现定向优化。

二、建立数据库

使用数据库管理系统创建数据库，首先要创建一个空白数据库，然后在这个空白数据库中创建数据表。创建数据表首先需要定义数据表的结构，然后才能在数据表中输入数据。

1. 数据库创建、删除、打开

创建数据库：CREATE DATABASE 数据库名
删除数据库：DROP DATABASE 数据库名
打开数据库：USE 数据库名

2. 数据表创建、删除、修改

创建好数据库后，就可以在数据库中创建数据表。数据表是数据库中的一个对象，它以行和列的集合存储数据。创建数据表需要定义表中列的结构，包括列名、数据类型、约束等。

（1） 创建数据表

```
CREATE TABLE 数据表名
(
〈列名 1〉〈数据类型〉〈列约束〉,
〈列名 2〉〈数据类型〉〈列约束〉,
……
[CONSTRAINT 〈约束名〉〈约束条件〉]
)
```

① 数据表名：要创建的数据表的名称。

② 列名：定义的数据表中列(字段)的名称，同一张表中不能有相同的列名。

③ 数据类型：每一列中的数据所属的数据类型。不同的数据库支持的数据类型并不相同，MySQL 支持的常用数据类型如表 2.6 所示。

表 2.6 MySQL常用数据类型

数据类型	说明
int	-2147483648~ 2147483647的整数
decimal	格式：decimal(M, D) M表示存储的总数字位数，不包括小数点，最多为 65位 D表示小数位数，最多为 30位
fbat	绝对值在 1.18× 10^{-38} ~ 3.40× 10^{38} 的实数
char	固定长度的字符串，最多 255个字符。如："user001""男"
date	"1000- 01- 01" ~ "9999- 12- 31"的日期
datetime	"1000- 01- 01 00: 00: 00.000000" ~ "9999- 12- 31 23: 59: 59.999999"的日期和时间

④ 约束：为了防止数据库中存在不符合要求的数据，数据库管理系统需要提供一种机制来检查数据库中的数据，检查其是否满足规定的条件。约束条件作为数据表定义的一部分存储在数据库中，作为数据库管理系统检查数据的依据。常用的约束有：

■ 空值约束：该列是否允许为空值。NULL：允许为空值；NOT NULL：不允许为空值。

■ 唯一约束：该列是否允许出现重复值。UNIQUE：不允许出现重复值。

■ 主键约束：设置该列是表的主键。PRIMARY KEY：主键。

当主键由多个列组成时，"PRIMARY KEY"不能写在每个列后面，应该在所有列描述之后再定义主键，方法是：

PRIMARY KEY(列名 1,列名 2,……)

例如，将 ordernumber(订号单)和 bno(书号)的组合作为主键。

```
PRIMARY KEY(ordernumber, bno)
```

■ 默认值约束：定义列的默认值。例如，购书的"数量"默认为 1：DEFAULT 1。

■ 外键约束：定义该列是表的外键。

外键约束是指限制列的取值要受到其他列的取值范围的约束。例如，“orders(订单)”表中 username 列输入的值，应该来自“users(用户)”表中 username(用户名称)列的值。“orders(订单)”表中的 username(用户名称)是外键，“users(用户)”表中的 username(用户名称)是主键。

外键约束定义了两张表之间的关系。一张表的外键可以是一个或多个列，外键应该是另一张表(主表)的主键或其他关键字。

定义外键约束的方法是：

CONSTRAINT 约束名称 FOREIGN KEY(作为外键的列) REFERENCES 主表(作为主键的列)

例如，创建一张数据表 students。

```
CREATE TABLE students
(
studentID char(8) PRIMARY KEY,          /*学号列设置为主键*/
fullname char(20) NOT NULL,             /*姓名列不允许为空值*/
sex char(2),
birthday date,
mail char(20) UNIQUE)                   /*电子邮件不允许出现重复*/
```

(2) 删除数据表

当删除数据表时，该表的结构和数据以及与该表相关的数据库对象都被删除。如果要删除的表被其他表外键约束，则该表不允许删除。删除数据表的方法：

DROP TABLE 数据表名

例如，删除数据表 students。

```
DROP TABLE students
```

(3) 修改数据表

创建完数据表之后，可以对表的结构进行修改，修改表的结构包括增加或删除列、增加列的约束等：

```
ALTER TABLE 数据表名
[ADD 〈列名〉〈数据类型〉〈列约束〉]          /*增加列*/
|[DROP COLUMN 〈列名〉]                    /*删除列*/
|[ADD CONSTRAINT 〈约束名〉〈约束条件〉]     /*增加约束*/
```

例如，在 students 表中增加一列“联系电话”：phone char(11)。

```
ALTER TABLE students ADD phone char(11)
```

例如，在 students 表中删除新增加的列 phone。

```
ALTER TABLE students DROP COLUMN phone
```

3. 编辑数据表的数据

编辑数据表的数据主要包括插入数据（INSERT）、修改数据（UPDATE）和删除数据（DELETE）。插入、修改、删除数据表中的数据时，必须满足数据的约束条件，否则将会造成操作失败。

（1） 插入数据

在创建完数据表后，就可以使用 INSERT 语句在数据表中添加数据。INSERT 语句的格式是：

INSERT INTO 数据表名［(〈列名表〉)］VALUES（值列表）

〈列名表〉中的列名必须是表中已有的列名。值列表中值的顺序必须与列名表中列的顺序一一对应，并且数据类型也要一致，如果是空值，用 NULL 表示。

如果〈列名表〉省略，则值列表中值的顺序必须与表中列的顺序一致，且每一列必须都有值，可以是 NULL。

例如，在 students 表中插入一条记录。

```
INSERT INTO students VALUES ('20180101', '张信', '男',
'2002-6-10', 'zhangxin@abc.com')
```

如果插入的一条记录中只有部分数据，其中出生日期暂缺，可以使用以下两种方法：

```
INSERT INTO students VALUES ('20180102', '陈息', '女',
NULL, 'chenxi@abc.com')
INSERT INTO students (studentID, fullname, sex, mail)
VALUES ('20180102', '陈息', '女', 'chenxi@abc.com')
```

（2） 修改数据

可以使用 UPDATE 语句对表中已有的数据进行修改。UPDATE 语句的格式是：

UPDATE 数据表名 SET 〈列名〉=〈表达式〉[,〈列名〉=〈表达式〉,……][WHERE 〈条件〉]

例如,在 students 表中将学号为“20180102”的学生的出生日期修改为 2002 年 1 月 2 日。

```
UPDATE students SET birthday = '2002-1-2' WHERE studentID = '20180102'
```

在 MySQL 中,条件表达式常用的运算符有:

算术运算符:+(加)、-(减)、*(乘)、/(除)。

关系运算符:=(等于)、>(大于)、<(小于)、<>(不等于)、<=(小于等于)、>=(大于等于)。

逻辑运算符:NOT(非)、AND(与)、OR(或)。

LIKE:模糊匹配。“_”匹配任何一个字符;“%”匹配 0 或多个字符。例如,表示出版社名称中包含“大学”两个字的出版社:publisher LIKE‘%大学%’。

(3) 删除数据

当确定不需要某些记录时,可以使用 DELETE 语句删除表中的这些记录。DELETE 语句的格式是:

DELETE FROM 数据表名 [WHERE 〈条件〉]

例如,在 students 表中将学号为“20180102”的学生记录删除。

```
DELETE FROM students WHERE studentID = '20180102'
```

如果要删除 students 表中所有的记录,可以使用如下语句:

```
DELETE FROM students
```

项 目 实 践 创建网上书店数据库

在 MySQL数据库管理系统中,使用 SQL语句创建网上书店数据库,并在数据库中创建数据表、编辑数据表的数据。

1 创建和打开数据库

(1) 创建网上书店数据库 bookstore

```
CREATE DATABASE bookstore
```

(2) 打开网上书店数据库 bookstore

```
USE bookstore
```

2 创建数据表

(1) 创建“用户”表和“订单”表

“用户”表和“订单”表的结构如表 2.7和表 2.8所示。

表 2.7 用户(users)

列名	数据类型	长度	说明
username	char	20	用户名称
password	char	20	密码
mail	char	20	邮箱
fulhame	char	20	姓名
sex	char	2	性别
birthday	date		生日
address	char	50	地址

表 2.8 订单(orders)

列名	数据类型	长度	说明
orderID	char	16	订单号
ordertime	datetime		下单时间
state	char	5	交易状态
username	char	20	用户名称

```
CREATE TABLE users
(
    username char(20) PRIMARY KEY,
    password char(20),
    mail char(20),
    fullname char(20),
    sex char(2),
    birthday date,
    address char(50))
```

```
CREATE TABLE orders
(
    orderID char(16) PRIMARY KEY,
    ordertime datetime,
    state char(5),
    username char(20))
```

请你参照前面创建数据表的方法，在 bookstore数据库中再分别创建“图书类别（categories）”“图书（books）”和“订单明细（orderdetails）”三张表。

（2）建立“用户”表和“订单”表之间的关系

按照图 2.9所示的数据库 bookstore中表之间的关系，修改 orders表的结构，定义 orders表的外键约束，建立 users表和 orders表之间的关系。

```
ALTER TABLE orders ADD CONSTRAINT fk_username FOREIGN KEY(username) REFERENCES users(username)
```

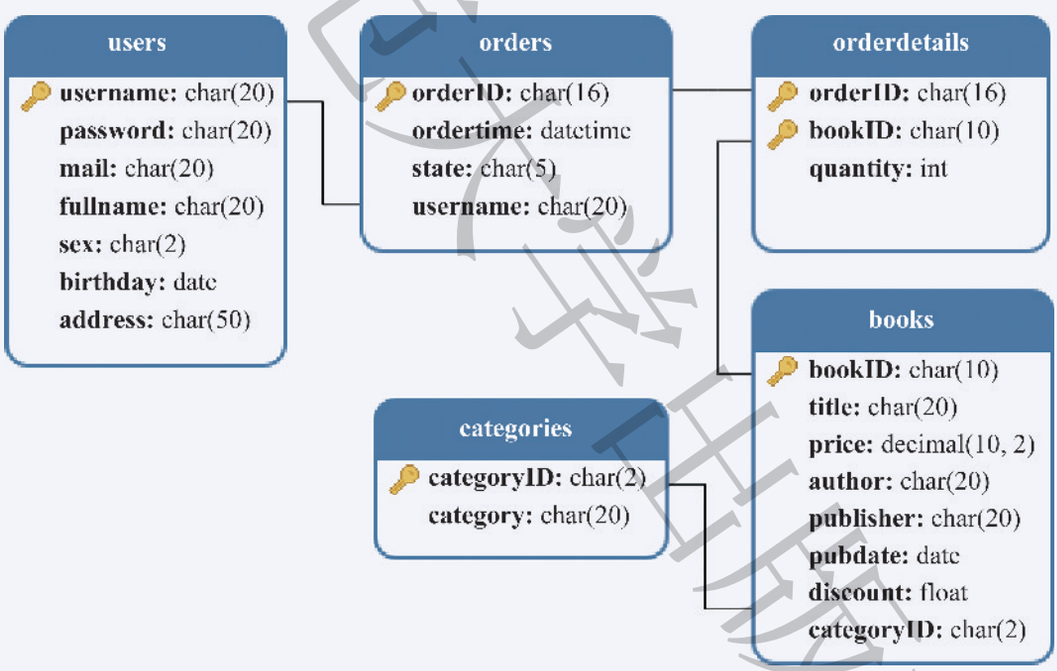


图 2.9 数据库 bookstore中表之间的关系

请你参照前面建立表间关系的方法，在 bookstore数据库中再分别建立 categories表与 books表、books表与 orders表、orders表与 orderdetails表之间的关系。

3 编辑数据表的数据

数据表的结构建立好之后，就可以对表中的数据进行编辑，主要包括插入数据、修改数据和删除数据。

（1）在 users表中插入一行数据

所插数据如下：

username	password	mail	fullname	sex	birthday	address
user001	12345678	user001@mail.sh.cn	王成	男	1980-03-16	上海市晋元路

```
INSERT INTO users VALUES ('user001', '12345678', 'user001@mail.sh.cn','王成','男','1980-03-16','上海市晋元路')
```

（2）将 users表中用户名称为“user001”的用户姓名修改为“王晨”

```
UPDATE users SET fullname='王晨' WHERE username='user001'
```

（3）删除 users表中用户名称为“user001”的用户

```
DELETE FROM users WHERE username='user001'
```

4 导入和导出数据

对于数据库中批量数据的输入和输出，可以使用数据库中数据的导入和导出功能。

操作数据库可以使用结构化查询语言 SQL，也可以使用一些图形化的操作工具。请使用图形化操作工具导入、导出数据库的数据。

（1）导入 users表、categories表、books表、orders表和 orderdetails表中的数据，思考数据导入的顺序。

（2）导出 users表中的数据，分别保存为 XML格式和 JSON格式，并观察这两种半结构化数据的特点。

三、数据查询

数据查询（SELECT）语句是数据库中最基本和最重要的语句之一，其功能是从数据库中查询满足条件的数据（选择运算）。查询的数据源可以是一张表或多张表，查询的结果是由 0 行（没有满足条件的数据）或多行数据组成的数据集，并允许选择一列或多列（投影运算）输出。查询还可以对数据进行分组统计，对查询结果进行排序等。查询语句的基本格式如下：

SELECT〈列名或表达式列表〉
FROM〈数据源〉
[WHERE〈条件〉]
[GROUP BY〈分组列〉]
[ORDER BY〈排序方式〉]

■ SELECT 子句：列出需要输出的列名，也可以是表达式。

■ FROM 子句：用于指定查询的数据，数据可以是一张表或多张表。

查询数据涉及两张或两张以上的表，称为多表连接查询。连接查询是关系数据库中最主要的查询，内连接是最常用的连接类型。使用内连接时，如果两张表的相关字段满足连接条件，则从两张表中提取数据并组合成新的表。FROM 子句中内连接的格式是：

FROM〈表 1〉INNER JOIN〈表 2〉ON〈连接条件〉

■ WHERE 子句：用于从数据表中选择满足条件的数据。

■ GROUP BY 子句：用于按照“分组列”对数据进行分组，然后再对每个组进行计算。例如，按照“性别”列对用户进行分组，分别统计不同性别的用户人数。可以按某一列进行分组，也可以按多列进行分组。

分组统计中一般用到聚合函数，常用的聚合函数如表 2.9 所示。

表 2.9　常用聚合函数

函数名	功能
COUNT()	计数
SUM()	求和
AVG()	求平均数
MAX()	求最大值
MIN()	求最小值

■ ORDER BY 子句：用于对查询的结果进行排序，排序分为升序(ASC)和降序(DESC)，默认为升序排序。ORDER BY 子句的格式为：

ORDER BY〈列名1〉[ASC|DESC],[〈列名 2〉[ASC|DESC],……]

项　目　实　践　在网上书店数据库中查询数据

在网上书店数据库中，根据需要查找满足自己需求的数据。

1　在 books表中查找出版社为“北京大学出版社”的图书记录，结果显示书名、定价、作者、出版社、出版日期。

```
SELECT title, price, author, publisher, pubdate
FROM books
WHERE publisher = '北京大学出版社'
```

查询结果示例：

title	price	author	publisher	pubdate
现代物流 **	38.00	李 **	** 大学出版社	2017- 11- 01
中国现代 **	58.00	程 **	** 大学出版社	2011- 10- 01
自然哲学 **	168.00	牛 **	** 大学出版社	2018- 06- 24

2 查找出版社为“北京大学出版社”的图书信息，结果显示书名、定价、作者、出版社、出版日期、类别名称。

```
SELECT title, price, author, publisher, pubdate, category
FROM books INNER JOIN categories ON books.categoryID = categories.categoryID
WHERE publisher = '北京大学出版社'
```

查询结果示例：

title	price	author	publisher	pubdate	category
现代物流 **	38.00	李 **	** 大学出版社	2017- 11- 01	经济
中国现代 **	58.00	程 **	** 大学出版社	2011- 10- 01	文学
自然哲学 **	168.00	牛 **	** 大学出版社	2018- 06- 24	自然科学

3 查找用户名称为“user001”的订单明细，结果显示订单号、用户名称、书号、数量、书名。

```
SELECT orders.orderID, username, orderdetails.bookID, quantity, title
FROM orders
INNER JOIN orderdetails ON orders.orderID = orderdetails.orderID
INNER JOIN books ON orderdetails.bookID = books.bookID
WHERE orders.username = 'user001'
```

查询结果示例：

orderID	username	bookID	quantity	title
1000020000823301	user001	1101083927	1	** 航母
1000020000823301	user001	1103028726	1	数据库技术 **
1000020000823301	user001	1103188334	1	丝绸之路 **
1000020000827101	user001	1102307520	1	中华文明 **
1000020000827101	user001	1102474857	1	** “一带一路”
1000020000827101	user001	1102710848	1	** 成语大词典
1000020000827101	user001	1103146032	1	Oracle**
1000020000827101	user001	1103149922	2	国际政治 **

4 在 orderdetails表中按订单号统计每张订单的订书总量，结果显示订单号和订书总量。

```
SELECT orderID, Sum(quantity)
FROM orderdetails
GROUP BY orderID
```

查询结果：

orderID	Sum (quantity)	orderID	Sum (quantity)
1000020000823301	3	1000020000829100	5
1000020000823501	5	1000020000829500	6
1000020000826209	7	1000020000830100	7
1000020000827101	6	1000020000830312	6
1000020000828001	3		

5 查找每张订单的订书总额，结果显示订单号和订书总额，并按订单号降序排序。每本书的价格计算方式是：定价 *折扣。

```
SELECT orderdetails.orderID, Sum(quantity * price * discount)
FROM orderdetails INNER JOIN books ON orderdetails.bookID = books.bookID
GROUP BY orderID
ORDER BY orderID DESC
```

查询结果：

orderID	Sum (quantity*price*discount)
1000020000830312	335.56
1000020000830100	337.13
1000020000829500	466.80
1000020000829100	390.00
1000020000828001	168.12
1000020000827101	280.76
1000020000826209	369.00
1000020000823501	218.56
1000020000823301	160.00

作业练习

请选择生活中的一个数据管理项目，如学校运动会管理、学生社团管理、志愿服务管理等，对该项目进行数据需求分析，设计一个关系数据库，并选择一个数据库管理系统创建和使用该数据库。

知识延伸　数据库应用简史

数据库系统的萌芽出现于20世纪60年代。当时计算机开始广泛地应用于数据管理，对数据的共享提出了越来越高的要求，传统的文件系统已经不能满足人们的需要，能够统一管理和共享数据的数据库管理系统应运而生。

早期的数据库系统主要基于层次模型和网状模型。层次型数据库管理系统的典型代表是于1968年推出的IMS数据库管理系统。

1969年，数据系统语言委员会(CODASYL)组织下属的数据库任务组(DBTG)提出了一个系统方案，该方案提出的方法是基于网状结构的，以后开发的许多网状型数据库管理系统都采用了DBTG模型和方法。

1970年，有研究人员发表了奠定关系数据库基础的论文。关系数据库管理系统是目前使用最广泛的数据库管理系统。目前常见的关系数据库管理系统有Access、MySQL、SQL Server、Oracle等。

Access是一个中小型的数据库管理系统。它可以极大地提高数据处理的效率，被广泛应用于财务、金融、统计等众多领域。

SQL Server一经推出，很快就得到了广大用户的积极响应并迅速占领了Windows环境下的数据库领域，成为数据库市场上的一个重要产品，主要面向中小企业。

Oracle数据库管理系统是以分布式数据库为核心的一组软件产品，是目前世界上使用最为广泛的数据库管理系统。作为一个通用的数据库管理系统，Oracle具有完整的数据管理功能和很强的分布式处理功能，并以其良好的兼容性和安全性，高效的可用性和并发性，较强的稳定性和扩展性，以及对复杂计算、统计分析的强大支持，在大型数据库系统中得到了广泛的应用。金融、通信、能源、运输、零售、制造等各个行业的大型公司基本都使用 Oracle。如，银行、金融业对可用性、健壮性、安全性、实时性要求极高，零售、物流业对海量数据存储分析要求很高，Oracle均能满足。

随着大数据时代的到来及互联网 Web2.0网站的兴起，传统的关系数据库在应付海量数据存储和处理方面，已经显得力不从心，非关系型、分布式数据存储得到了快速的发展，NoSQL的概念在 2009年被提出。NoSQL泛指非关系型的数据管理技术，相对于广泛应用的关系数据库来说，这是一种全新的思维方式，对传统的数据管理方式是一次颠覆性的改变。它的产生是为了解决大规模数据集合及多种数据种类带来的挑战，尤其是大数据应用的难题。

NoSQL数据库可以分为四类，即键值存储数据库、列存储数据库、面向文档存储数据库、图形存储数据库。NoSQL数据库可以支持海量的数据存储，数据模型灵活，具有强大的水平可扩展性，可以较好地应用于大数据时代的各种数据管理中。

（参考资料：《数据库系统基础（第 6版）》，Ramez Elmasri, Shamkant B. Navathe著）

第三章

数据安全

本章学习目标

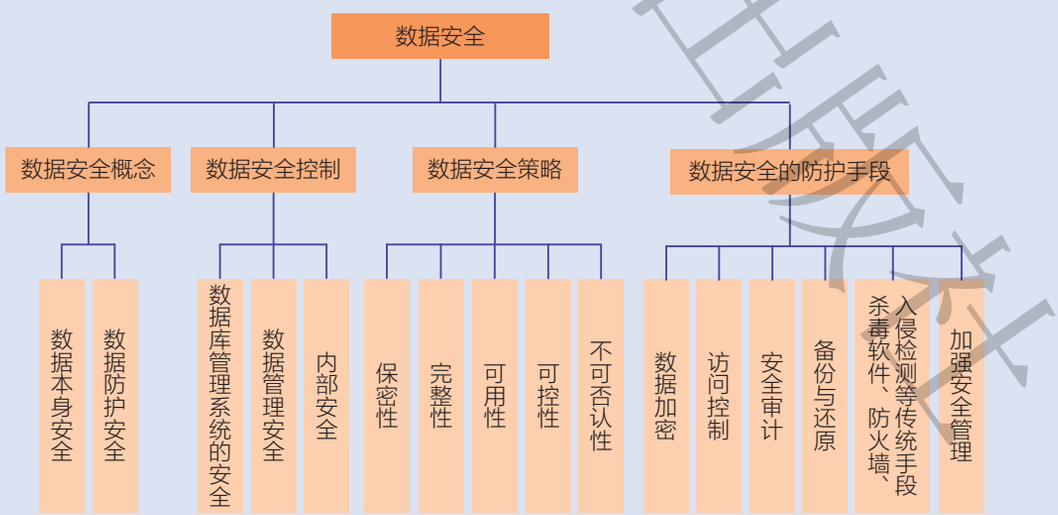
- 结合案例，认识数据丢失和数据泄露的风险，在日常生活与学习中能采取措施降低风险发生的概率。
- 了解什么是数据安全以及威胁数据安全的主要因素，从而能尽量避免数据安全事故的发生。
- 了解数据安全策略。
- 掌握基本的数据安全防范措施，在日常生活中能建立数据安全意识，在不同的应用场景下能采取相应的安全防范措施。
- 掌握数据备份与还原的基本方法。

某电脑病毒全球肆虐，对感染的海量的计算机系统造成了严重的破坏，导致大量数据丢失，无数企业经济损失惨重；某用户的信用卡账号和密码等重要数据被不法分子通过网络窃取，账上的巨额资金被盗刷；某大型网站被黑客攻击，数以亿计的用户真实姓名、身份证号、手机号等重要数据泄露……这些耸人听闻的消息频频出现。互联网浪潮下，我们享受着随时随地能与他人交换数据的便捷，也身处自己的各种数据可能丢失或泄露的危险境地之中。

那么，怎样摆脱这种尴尬的局面呢？我们需要未雨绸缪：了解数据安全的重要性和各种可能威胁数据安全的因素，牢固树立数据安全防范意识，从自我做起养成良好的数据使用习惯，加强软硬件系统防护，阻断病毒传播渠道，关闭可能被黑客攻击的“后门”。同时，做好数据备份工作，通过数据的备份与还原，将因各种主客观原因造成的数据损失降到最低。

数据安全不仅关系着个人和企业，更与国家安全和社会稳定息息相关。为了维护国家安全和社会公共利益，保护公民、法人和其他组织在网络空间中的合法权益，保障个人信息和重要数据安全，很多国家相继出台了相关的法律法规，引导人们文明上网，合理合法地收集和使用数据。因此，我们每一个人都需要增强有关数据安全的法律意识，共同维护国家利益，以及团体和个人的合法权益。

本章知识结构



项目主题

在线考试系统的安全维护

项·目·情·境

在学习过程中，考试是一种常用的评价方式。作为反映教学效果和学习情况的晴雨表，考试既可以帮助学生发现其在学习中存在的问题和不足，也可以帮助教师优化教学内容和改进教学方法。传统考试主要借助于纸和笔，学生用笔在纸质试卷上答题，教师也用笔在纸质试卷上进行批改。为了保证考试的公平公正，重大考试的纸质试卷的命题、印刷、运送、交接和保管等环节都有着严格的保密规定。

现在，无纸化的在线考试逐渐兴起。在线考试是通过网络媒体进行的一种考试形式，通过服务端的在线考试系统，可实现从出题、组卷到发布考试、导入考生信息、自动评卷或人工评卷等完整的考试流程。很多考试、测验、练习都可以通过无纸化考试（练习）系统平台完成。

与传统考试应做好纸质试卷的保密工作一样，在线考试系统对考场行为有很强的纪律约束，对数据安全也有很高的要求。特别是为了保障考试的顺利进行，更要杜绝数据丢失问题。

学校要举行信息技术课程的上机考试，使用在线考试系统作为平台，为了确保考试的顺利完成，考试系统管理员需要考虑很多安全问题。例如：考试试卷的保密；考试时试卷的分发与回收；考试过程中，突然遇到计算机故障等问题必须及时处理；考试结束后考生答卷的保存及成绩等相关数据的查询与分析等等。考试系统管理员需要针对这些安全问题制定预防措施、应急预案。请帮助考试系统管理员一起分析影响在线考试顺利完成的安全问题，并思考如何解决。

项·目·任·务

任务 1	任务 2	任务 3	任务 4
结合案例，认识数据丢失和数据泄露的风险。	结合在线考试系统的安全维护问题，了解可能造成考试数据丢失和相关数据泄露的主要因素有哪些，提升安全意识。	结合在线考试系统在考试前需要做好的防范准备、在考试中可能需要做的应急处理，以及日常的合理维护了解保护数据安全的措施有哪些。	结合在线考试系统中数据的备份与还原，理解数据备份与还原的概念，掌握数据备份与还原的基本方法。

第一节　数据安全威胁与数据安全策略

随着数据价值不断被挖掘，数据泄露事件的发生频率越来越高，规模也呈逐年扩大趋势，互联网、金融、医疗行业尤其如此，成为了数据泄露的重灾区。因此，我们必须不断强化数据安全意识，规范数据操作行为，养成良好的数据使用习惯，发现系统漏洞及时修补，杜绝各种病毒感染的途径。同时，一旦发生数据丢失或泄漏事故，应立即采取措施，以尽量降低其可能造成的各种损失。

问　题　思　考

1. 日常生活中，你或周围的人有过数据丢失或泄露的经历吗？是什么原因造成了数据的丢失或泄露？产生了哪些影响？

2. 你采取过哪些保护数据安全的措施？

3. 应该怎样预防数据丢失或泄露事故的发生？

数据安全不仅关乎个人隐私、企业机密，更涉及国家政治、经济、军事等领域，已上升为一个事关国家政治稳定、社会安定、经济有序运行、国防安全的全局性问题，需要引起足够的重视。感染病毒、黑客入侵、自然灾害、软硬件故障以及人为失误是威胁数据安全的主要因素。

一、数据安全概念

数据安全包含两方面的内容：一是数据本身的安全，主要是指采用复杂的加密算法对数据进行主动保护，如数据加密、数据完整性、身份认证等，以防止数据被泄露、篡改；二是数据防护的安全，主要是指采用先进的信息存储手段对数据进行主动防护，如通过磁盘阵列（磁盘阵列是由很多个独立磁盘构成的冗余阵列，它通过对多个磁盘的管理可以提供比单个磁盘更高的存储性能）、数据备份、异地灾备（在不同的地域构建一套或者多套相同的应用或者数据库，起到灾难发生后立刻接管的作用）等手段保证数据的安全，避免因数据丢失造成损失。

项 目 实 践　在线考试系统的数据安全措施

期末考试就要到了，如果你是在线考试系统管理员，你需要提前对系统做哪些准备工作，对相关的教师和学生做哪些提醒，以保证考试能够在系统上顺利进行？请完成表 3.1。

表 3.1　考前准备措施表

检查项目	采取措施
保证供电正常	提前进行电路检测
保证终端计算机与在线考试系统服务器之间的网络通畅	对每一个网络节点、路由器等进行测试
保障试题在考前不泄露	将考试题目加密保存……

即便考前做了充分的准备工作，在考试过程中，有时也可能因为遇到一些不可预料的突发事件而导致考试中断，或考试数据丢失等。产生这些问题的原因可能有多个，对不同原因导致的突发事件，往往需要实施不同的应急方案，以便保证考试的顺利进行。请完成表 3.2。

表 3.2　考试过程中可能出现的突发事件的原因分析及应急方案表

突发事件	可能的原因	应急方案
停电	电闸跳闸	检查关闭不需要使用的大功率设备，调整电闸，启用 UPS不间断电源
学生无法进入考试系统	登录的用户名或密码错误	重新设置用户名、密码
学生无法上传或提交答卷	网络拥堵	错开时间上传答卷、用优盘拷贝后人工上传

考试结束后，对于客观题甚至部分主观题，在线考试系统可以根据事先保存的标准答案进行自动判卷。当然，它也提供人工阅卷接口。同时，利用保存下来的每位学生的答卷和批改结果，在线考试系统可以自动生成每位学生的成绩、每道题目的正确率等各种数据，这些数据将为教师评价教学效果、学生了解学习情况提供重要依据。因此，妥善保存这些数据、协助相关人员进行数据分析，也是在线考试系统管理员的工作重点之一。此外，考试数据中包含涉及个人隐私的数据，如何妥善保护数据？如何在合理的范围内使用数据，向不同人员提供不同数据？这都是系统管理员需要仔细考虑的问题。请完成表 3.3。

表 3.3　考试数据维护表

问题	原因	措施	预防
学生考试数据丢失	服务器存储设备出现物理故障	更换存储设备，并将之前备份的数据进行还原	对每场考试的数据及时进行备份
试题数据泄露	某位教师使用 U 盘拷贝部分试题时带入了病毒	切断网络，对服务器进行病毒查杀	事先检查要使用的 U 盘是否感染病毒；或禁止外部设备接入，采用其他方式让教师查看试题
学生查询自己成绩时发现也可以查看他人成绩	权限设置出错	重新设置为每位学生只能查看自己的成绩	事先对不同层级的人员设置不同的访问、修改权限

从以上项目实践中可以看出，对数据安全的控制需要针对威胁数据安全的主要因素，从数据库管理系统本身、人对数据的管理以及对可以接触到数据的人的管理三个层面来实现。

1. 数据库管理系统的安全

由于各种数据库管理系统本身就可能存在安全缺陷，因此，提升数据安全首先需要从数据库管理系统入手，利用现代化技术手段不断升级完善数据库管理系统，从而保障数据库管理系统对数据的高效管理和安全保护。同时，由于数据在不断增加和变化，因此，数据库管理系统中的数据库也需要及时地更新，在更新数据库内容时，要特别注意，必须在保证数据库稳定的前提下进行。

2. 数据管理安全——提高数据管理者的安全意识

加强和完善数据库的管理工作还需要提高相关人员的安全意识和业务能力。安全意识的培养主要是让数据管理者认识到数据库安

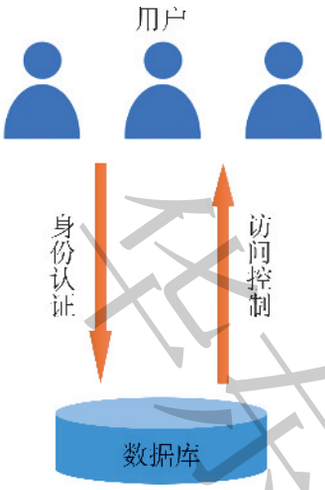


图 3.1　用户身份认证与数据库访问控制示意图

全管理的重要性。同时，随着数据库安全技术的不断增强，数据管理者需要不断地学习和钻研，将新的技术和科研成果应用到数据库安全防护中，将数据库的被动防御变为主动防守，切实加强数据库的安全防护工作。如，可以通过身份认证和访问控制限定不同用户的行为（如图 3.1 所示）；又如，可以通过安全审计主动记录访问者的相关信息，包括 IP 地址、身份信息等，及时发现诸如某一用户无法通过验证却反复对数据库进行访问等异常行为，将这些记录保存并反馈给数据库管理者，以便管理者及时阻止恶意入侵，防患于未然，增强数据库的安全性。当然，数据库管理者必须建立一套科学、完善的数据库管理方案，为数据库的管理提供科学、合理的管理流程，以方便管理工作的顺利进行。

3. 内部安全——提高内部数据使用者的安全意识

由于人为操作的不规范而导致的数据安全问题比比皆是，例如：使用数据时，因操作失误而导致数据泄漏或损坏；存储数据时，数据中心、服务器、数据库的数据被随意下载、共享，或者离职人员通过各种存储设备将机密资料随意拷贝，以及由于个人电脑或存储设备维修、遗失、被盗等造成数据泄漏；传输数据时，通过邮件、即时通讯软件等随意传输机密资料，以及资料传输过程中被窃听、拦截、篡改、伪造等。

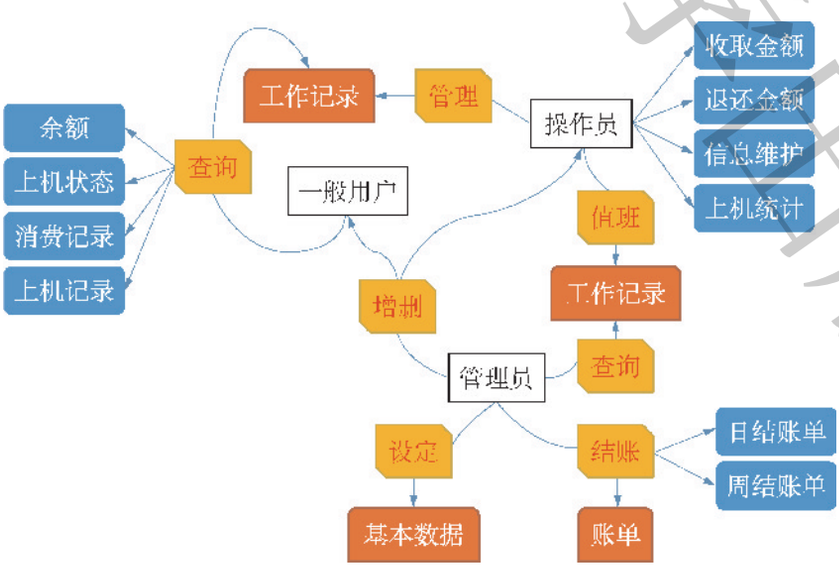


图 3.2　某数据库不同使用者的权限设计

因此，要加强数据的安全管理，首先必须规范数据的各项操作，使每一项操作都有章可循。以数据库安全为例，可对数据库的使用者进行限制，设定不同的访问操作权限，如高权限的使用者可对数据库进行查询、修改、更新等操作，而低权限的使用者只可对数据库进行查询等简单操作。如图 3.2 所示的某数据库不同使用者的权限设计中，管理员可以对一般用户和操作员的账户进行增删，可设定基本

数据，汇总账单，以及查询各种信息等，而对于一般用户而言，只能查询属于自己的余额、上机状态等基本信息。此外，还要规范数据库的操作流程，及时检测数据库运行环境，防止病毒、木马程序等乘虚而入。

为保障计算机中数据文件的安全，可以对存储设备、网络行为、智能设备、操作系统以及端口等进行合理控制。

对于个人而言，不仅要增强数据安全意识，还要遵守信息法律法规，维护信息社会的伦理道德，在面对各种网络数据和信息时，能够理性判断，自觉规范自己的信息行为。

二、数据安全策略

由于现实世界中不存在绝对的安全，因此，要想保证数据安全，首先要准确评估数据的价值，选择合理的数据安全策略。数据安全策略是指为保证提供一定级别的安全保护所必须遵守的规则。数据安全策略通常建立在授权的基础上，未经授权的用户，不得访问、引用或者使用数据，即数据安全策略是对可接受的行为和违规的行为做出相应响应的规定。多数情况下采集到的数据会被保存于数据库系统中，而针对数据库系统的安全策略主要是保证数据的保密性（机密性）、完整性、可用性、可控性和不可否认性。

1. 数据的保密性（机密性）

保护数据秘密，未经授权其内容不会显露。

2. 数据的完整性

保护数据不被非法修改，使数据在传送前后保持完全相同。

3. 数据的可用性

保护数据在任何情况下不会丢失。当需要时，得到授权的用户可随时访问所需数据。

4. 数据的可控性

对数据的内容和传播具有控制能力。

5. 数据的不可否认性

数据交互过程的参与者都不能否认曾经完成的操作和承诺。即数据接收方要发送方承认数据是由其发出的，而不是他人冒名发送的；发送方也要求接收方不否认已经收到信息。

三、数据安全的防护手段

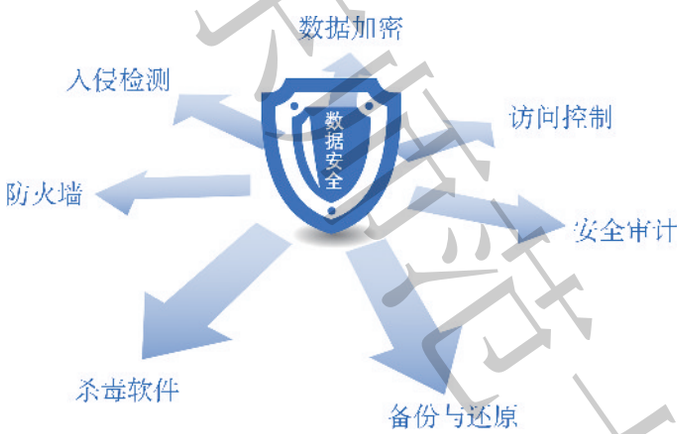


图 3.3　数据安全的防护手段

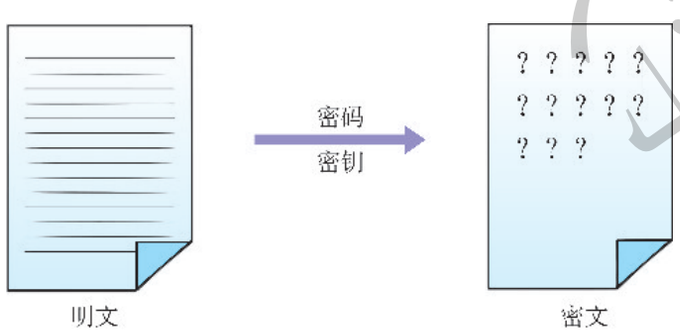


图 3.4　数据加密过程

与针对数据库系统的安全策略对应，数据安全的防护手段主要包括数据加密、访问控制、安全审计和备份与还原等（如图 3.3 所示）。

1. 数据加密

数据加密是对原来为明文的数据按某种算法进行处理，使其成为不可直接读取并理解的乱码（即密文），从而达到保护数据不被非法窃取、阅读的目的。由此可见，明文是加密前的原始数据（消息），密文是加密后的数据，在数据加密中常说的“密码”是指将明文与密文进行相互转换的算法，密钥是在密码中使用且只有收发双方知道的信息。数据加密的过程如图 3.4 所示。

传统密钥加密技术分对称密钥加密系统和公共密钥加密系统两种。对称密钥加密系统中，消息发送方和接收方使用相同的密钥进行加密和解密。对称密钥加密系统速度快、效率高，但密钥的管理、分发困难。公共密钥加密系统则给每个用户分配一对密钥：私有密钥和公共密钥。私有密钥是保密的，只有用户本人知道；公共密钥可以让其他用户知道。公共密钥加密系统的特点是：用公共密钥加密的消息只有使用相应的私有密钥才能解密；同样，用私有密钥加密的消息也只有用对应的公共密钥才能解密。相对于对称密钥加密系统，公共密钥加密系统速度慢、效率低，但密钥管理方便。

目前通过增加密钥的长度，可以提高加密技术的安全强度，从而保证数据在传输过程中的保密性和完整性。但是，由于数据在使用时必须完全解密，对最终用户而言，数据依然是明文，因而无法同时满足数据的保密性和可用性。此外，加密技术作为访问控制、数字签名等其他安全措施的基础，被应用在很多地方，这对密钥的管理和分发也提出了很高的要求。很多用户为方便记忆，使用简单密钥，或多个地方使用相同密钥，都易造成密钥的泄漏。

图 3.5　访问控制的功能

2. 访问控制

指根据预定义的数据模型和用户角色模型，对数据库、数据表的访问行为进行检测和判断，在必要时阻断查询语句以保护数据的安全。访问控制的功能如图 3.5 所示。

3. 安全审计

指对数据请求进行实时严密监控，对数据的访问者、访问时间、访问行为进行详细的审核和记录，通过安全分析检测非法行为，并与其他手段联动对违规事件进行处置。多数系统通过日志进行审计，日志中记录了系统安全事件、用户访问记录、系统运行状态等信息（如图 3.6 所示）。管理员通过日志审计可随时了解整个系统的运行情况，及时发现系统异常，在遇到安全事件或系统故障时进行快速定位，并为故障解除、系统恢复和原因追查提供依据。安全审计的缺点在于，它是一种事后核查机制，在发生数据安全事件后起作用，无法实时对攻击进行拦截和阻断以实现防患于未然。

4. 备份与还原

指通过分布式存储、冗余和恢复来实现数据的容灾安全性，是一种可用性机制。（详见本章第二节“数据备份与还原的实现”）

5. 杀毒软件、防火墙、入侵检测等传统手段

杀毒软件、防火墙和入侵检测等为防止病毒感染和黑客入侵而采用的传统手段仍是保护数据安全的有效手段，不容忽视。

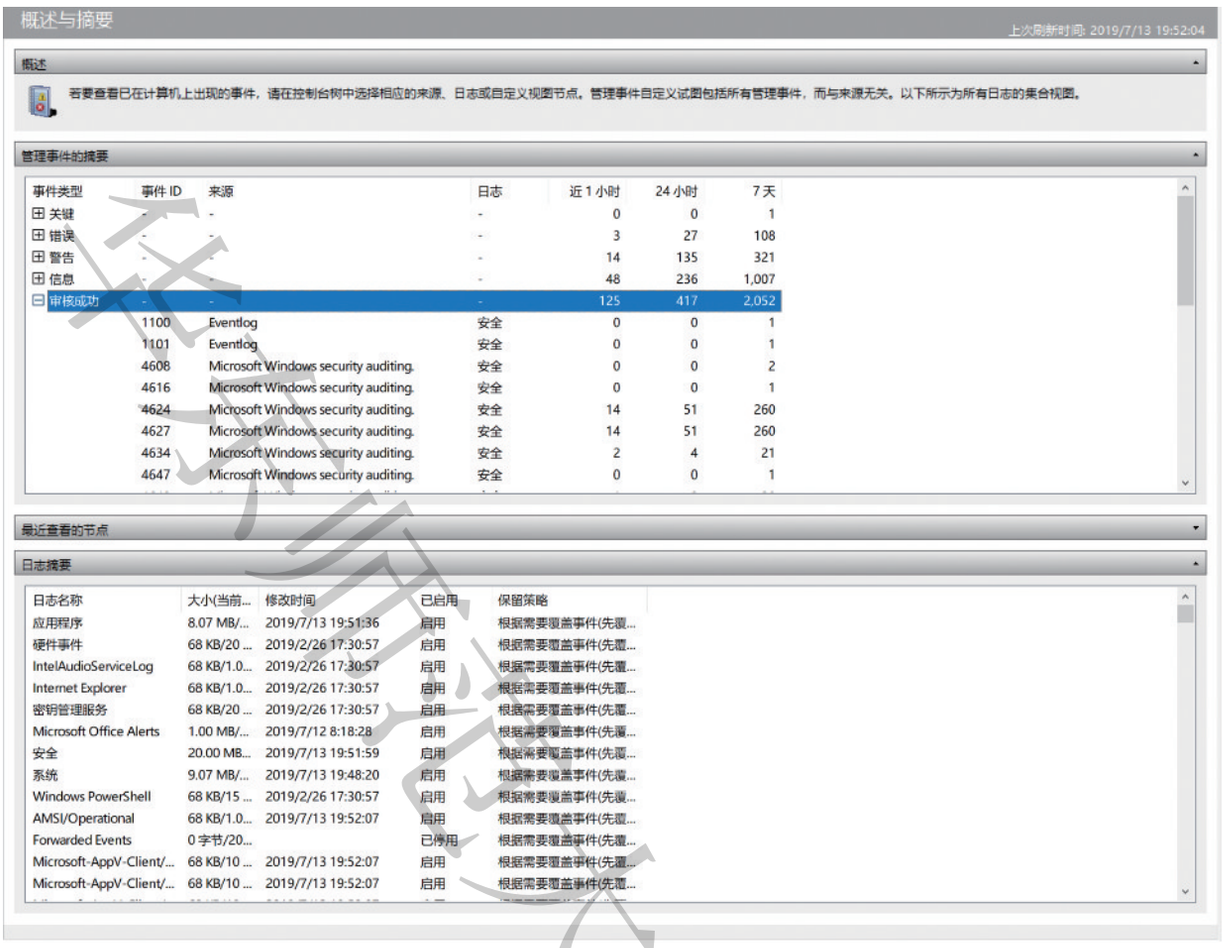


图 3.6　日志示例

杀毒软件：用于侦测、移除计算机病毒、计算机蠕虫和特洛伊木马的程序。

防火墙：一个架设在互联网与内部网之间的信息安全系统，根据预定的策略来监控往来的传输。

入侵检测：一种网上安全设备或应用软件，可以监控网络传输或者系统，检查是否有未经授权的访问或可疑活动。

6. 加强安全管理

针对人为失误，必须加强数据安全的管理，包括不断完善和升级数据库系统，加强对数据的管理控制，以及规范对数据或数据库的操作。

体 验 思 考　手机数据的安全措施

现代社会中，手机已成为我们日常生活中必不可少的设备，其中存储着大量重要的数据。我们在使用手机中的各种应用程序时，该如何保障手机中数据的安全？当我们需要更换手机时，又需要注意哪些问题？

知 识 延 伸　大数据安全

目前，随着数据产业的发展、技术的进步和应用创新的推动，数据挖掘分析在各行各业创造着巨大价值的同时，大数据安全问题也逐渐暴露出来。大数据中蕴含的巨大价值和其集中化的存储管理模式成为网络攻击的重点目标。大数据所涉及的各行业数据资源中往往包含大量的敏感和重要信息，一旦泄露或遭到非法利用，将会给个人甚至是国家带来无法弥补的损失。同时，随着大数据分析技术的成熟和价值挖掘的深入，从看似安全的数据中还原用户的敏感、隐私信息已不再困难。如何在数据交换、共享及使用等过程中实现对敏感数据的定向、精准和彻底脱敏，达到数据安全、可信、受控使用的目标，是数据产生者和管理者亟待解决的技术问题。

大数据安全技术体系分为大数据平台安全、数据安全和个人隐私安全三个层次。大数据平台不仅要保障自身基础组件安全，还要为运行其上的数据和应用提供安全机制保障，包括传输交换安全、存储安全、计算安全、平台管理安全以及基础设施安全；除平台安全保障外，数据安全防护技术为业务应用中的数据流动过程提供安全防护手段，包括数据分类分级、元数据管理、质量管理、数据加密、数据隔离、防泄露、追踪溯源、数据销毁等内容；隐私安全保护是在数据安全基础之上对个人敏感信息的安全防护，具体指利用去标识化、匿名化、密文计算等技术保障个人数据在平台上处理、流转过程中不泄露个人隐私或个人不愿意被外界所知的信息。

第二节　数据备份与还原的实现

任何以预防为目的的保护措施，无论其多么全面周到、细致入微，都只能尽量地减少而不能完全杜绝数据安全事故的发生。从自然灾害、病毒肆虐，到系统故障、操作失误，都会影响信息系统的正常运行，甚至造成整个信息系统完全瘫痪。数据备份与还原就是在这些突如其来的事故发生后，通过备份的数据完整、快速、简捷、可靠地恢复原有数据和系统。在震惊全球的“9·11”事件中，位于美国世贸大厦里的公司就遭遇了数据丢失的悲剧，只有少部分使用了数据异地灾备的公司在灾难发生后迅速恢复业务，而其他很多公司则因数据丢失而遭受毁灭性的打击甚至破产。

生活中，人们经常使用个人计算机或手机存储自己的信息，如果遇到这些电子设备发生故障，又未事先做好数据备份，可能会造成一些重要个人信息的丢失。这样的例子比比皆是。

在学校学习时，每次考试的试卷、成绩等数据都会得到妥善的保管，因为它们牵涉到学校教务工作的正常开展，以及对每位学生学习情况的记录和认证，还能为教师进行教学研究、教学改革提供依据。使用在线考试系统后，这些数据都存储在电子设备中，因此，安全存储并备份数据，在需要的时候还原数据，对学校而言十分重要。

可以说，数据备份与还原是防止数据丢失的最强有力的措施。

问　题　思　考

1. 日常生活中，我们需要经常对哪些数据进行备份？
2. 备份数据有哪些方法？它们各有什么特点？在什么情况下使用？
3. 如何将备份的数据还原？

一、数据备份

备份是为了使信息系统中的数据在损坏或丢失的情况下能够重新恢复而对数据进行的某种保存，这种保存提供恢复过程所需要的信息和数据。数据备份的主要目的是防止数据因为自然灾害、硬件故障、软件错误、人为误操作等因素而损坏或丢失。

数据备份技术源于20世纪70年代，当时主要利用一种海量存储设备——磁带库备份数据，但是由于技术原因磁带设备的利用率较低。20世纪80年代后期个人计算机的发展和20世纪90年代客户机/服务器模式的普及使得网络数据得到了发展，出现了数据分布式存储，即将数据分散存储在多台独立的设备上，这造成数据存储管理的复杂化。之后，随着网络技术的发展和Internet的兴起，信息系统逐渐使用“数据集中”的模式，即业务数据不保存在本地计算机上，而是通过网络集中上传到服务器，由服务器统一进行存储和处理。由此，数据的存储逐渐从传统的本地存储转向集中式存储，也推动了数据备份向大容量、具有先进自动备份管理功能的方向发展。

对于数据备份，我们要警惕一些误区：

1. 认为复制就是备份

早期的备份的确是将相关数据内容复制一份。然而，随着技术的发展和人们对备份要求的提高，今天的备份已不再是简单的拷贝了。因为单纯的数据复制无法保留相关的历史记录和系统状态信息，所以用其进行数据恢复时，无法再现数据的应用环境、属性、历史操作等重要信息。现在，完整的备份应包括自动化的数据管理与系统的全面恢复，因此，从这个意义上说，“备份=复制+管理”。备份管理包括备份的可计划性和自动化操作、历史记录的保存以及日志记录等，这样不仅可以通过自动化的程序设定消除手动备份的麻烦，保证数据的安全性和完整性，而且可以实现对备份的管理和跨平台的备份，满足全面的需求。

2. 以硬盘冗余备份代替备份

很多的服务器都采用硬盘冗余备份的容错设计，如双机热备份、磁盘阵列与磁盘镜像等。

双机热备份，一般也称为双活容灾系统，是指两套系统运行相同的应用，主机失效后，备机提供服务。如果对应到存储系统中，相当于两套存储系统同步存储相同的数据，两套系统中的对应数据同步更新，数据保持实时同步。

磁盘阵列就是把多个独立的物理硬盘按不同的方式组合起来形成一个硬盘组，从而提供比单个硬盘更高的存储性能和数据备份技术。如图3.7所示为磁盘阵列的一个例子。数据以块为单位分布存

储在不同硬盘上。磁盘阵列不对数据块本身进行备份，而是把与数据块对应的奇偶校验信息存储在磁盘上，并且奇偶校验信息和相对应的数据块分别存储于不同的磁盘上。当一个磁盘上的数据损坏后，利用剩下的数据和相应的奇偶校验信息即可恢复被损坏的数据。

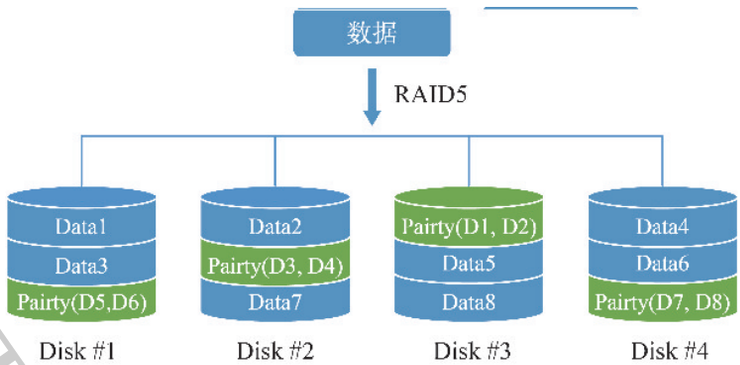


图 3.7　某磁盘阵列

磁盘镜像是指一个存储装置中的数据通过另一个装置或以另一种格式制作的完整复制品。

尽管硬盘冗余备份和恢复的速度很快，但并非理想的备份方案。例如双机热备份中，如果两台服务器同时出现故障，那么整个系统便陷入瘫痪状态，还是有较大的风险。特别是对于逻辑上的错误，如人为误操作、病毒感染、数据错误等，硬盘冗余备份只会将错误复制一遍，无法真正保护数据。

3. 只备份数据文件

有人认为，备份就是对数据文件进行备份，系统文件与应用程序无需备份，因为后者可以通过安装盘重新进行安装。事实上，安装和调试整个系统的时间代价相对较大，而对整个系统进行备份是更加高效、便捷的选择。

4. 不重视备份数据的保管

对于备份数据要妥善保管，否则在出现问题需要恢复数据时，如果备份数据发生损坏、丢失，会导致数据无法还原。因此，常采用异地灾备，将系统和数据备份在相隔较远的系统中，以便在发生如地震等灾害时，也能保证数据的安全和系统持续、稳定地运行。

二、数据备份方式

数据备份的方式是多种多样的，各有优缺点和应用范围。要想最大限度地利用备份介质的容量，合理安排备份的时间，提高备份工作的效率，就应该根据实际情况，选择最合适的备份方式。

从备份的数量角度看，数据备份可以分为全量备份、增量备份和差异备份三种。从备份的时间角度看，数据备份可以分为定时备份和实时备份。

1. 全量备份（full backup）

全量备份指对整个系统（如组成服务器的所有卷）或用户指定的所有文件数据进行一次全面的备份。这是一种最基本也是最简单的备份方式。这种备份方式的好处是：很直观，容易被人理解；如果在备份间隔期间出现数据丢失等问题，只使用一份备份文件就可以快速地恢复所丢失的数据。但是，其不足之处也显而易见：它需要备份所有的数据，因此每次备份的工作量都很大，特别是当数据规模达到 TB 级及以上时，需要大量的备份介质。如果全量备份进行得比较频繁，那么在备份文件中有大量的数据是重复的，这些重复的数据占用了大量的存储空间，这对用户来说就意味着管理成本和硬件成本的增加。与此同时，如果需要备份的数据量相当大，备份数据时进行读写操作所需的时间也会较长，因此，全量备份一般不会进行得太频繁，通常隔较长一段时间才进行一次。但是，这样一旦发生数据丢失，只能恢复到最近一次备份的数据，而在这次备份后较长时间内新产生或更新的数据会丢失。

2. 增量备份（incremental backup）

增量备份指备份上次备份操作（无论是哪种备份）以来新产生或更新的数据。在特定的时间段内只有一定数量的文件发生改变，因此增量备份没有重复备份数据，既节省了备份设备的空间又缩短了备份时间。这种备份方法比较经济，可以频繁地进行。但是在增量备份系统中，一旦发生数据丢失或文件误删除操作时，恢复工作比较麻烦，需要多份备份文件才可以完成。因为恢复操作需要查询一系列的备份文件，从最后一次全量备份开始，将记录在一次或多次的增量备份中

的改变应用到文件上。使用增量备份恢复数据，备份文件之间的关系就像链子一样一环套一环，其中任何一个备份文件出现问题都会导致整条备份链脱节，因此这种备份的可靠性较低。由于恢复过程中需要使用全量备份的数据，因此，所有的增量备份都是在最近一次全量备份以后进行的。

3. 差异备份（differential backup）

差异备份只备份上一次全量备份后新产生和更新的数据。它的主要目的是将数据恢复时所涉及的备份文件的数量限制为两个，以简化恢复的复杂性。差异备份在避免了全量备份和增量备份两种方式的缺陷的同时又具有自身的优点：首先，它无需频繁地做全量备份，工作量小于全量备份，备份所需要的时间短，而且节省存储空间；其次，虽然每次做差异备份的工作量要大于增量备份，但是它的恢复相对简单，只需要两份备份文件，即上次的全量备份文件和最近一次的差异备份文件。

全量备份、增量备份、差异备份这三种备份方式的定义、优缺点和应用范围比较见表 3.4，这三种数据备份方式的示意图如图 3.8 所示。

表 3.4　全量备份、增量备份和差异备份的对比

	全量备份	增量备份	差异备份
定义	对整个系统或用户指定的所有文件数据进行全面的备份	只对上次备份后新产生或更新的数据进行备份	只备份上次全量备份后新产生和更新的数据
优点	备份的数据最全面、最完整。只需利用一份副本，就可以恢复全部数据	没有重复备份数据，可缩短备份时间，快速完成备份，而且能节省备份介质存储空间	恢复数据时，只需要两份文件，一份是上次的全量备份文件，另一份是最新的差异备份文件
缺点	备份工作量大，备份时间长，需要大量备份介质。如果进行得频繁，则备份文件中会有大量重复数据，重复的数据占用大量存储空间，增加了存储成本	可靠性较低，备份数据的份数太多；当发生灾难时，恢复数据比较麻烦，需要按顺序依次恢复每次备份的数据，环环相扣	
应用范围	不适用于业务繁忙、备份时间有限的网络系统。不能进行得太频繁，通常只是在备份的最开始采用		适用于各种备份场合

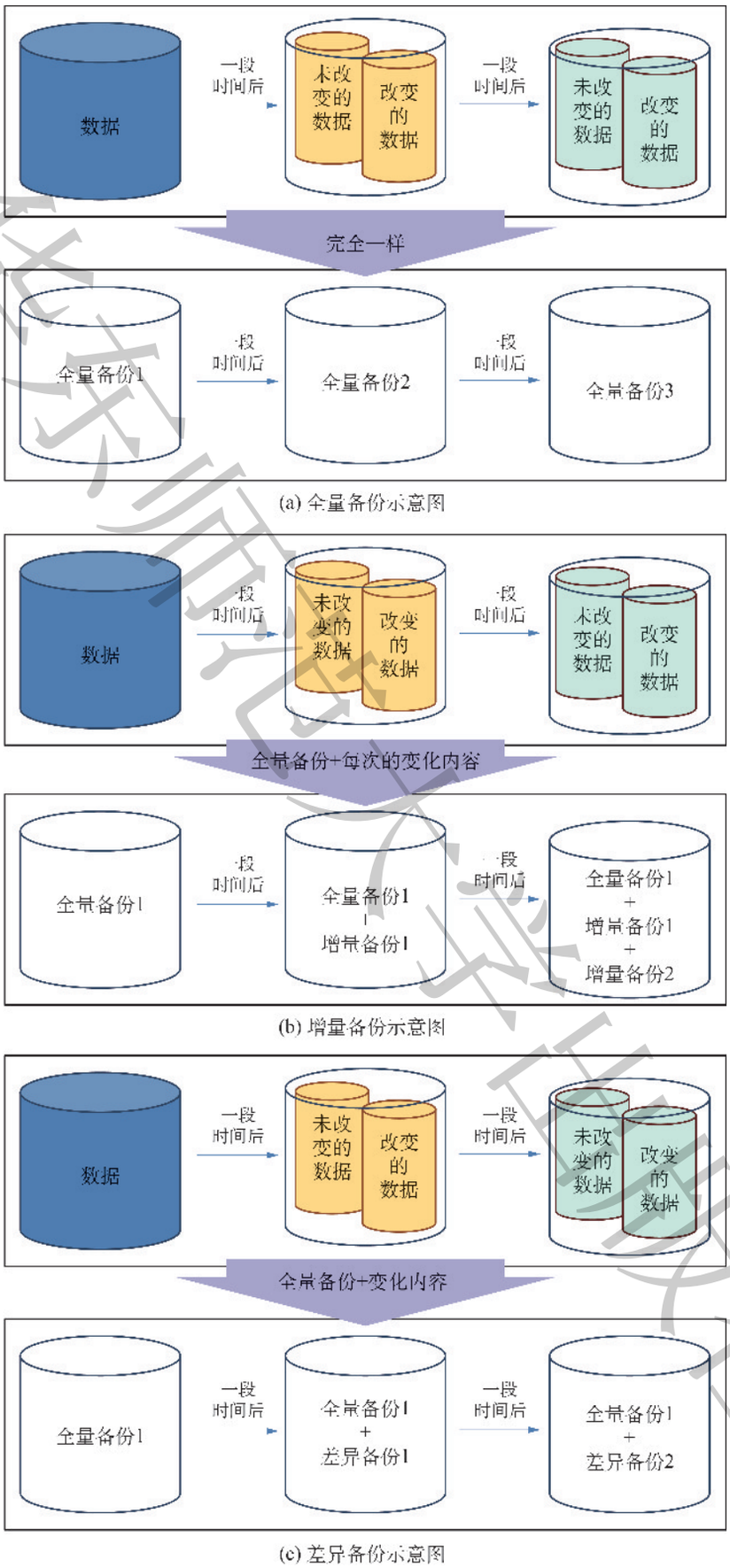


图 3.8　数据备份示意图

4. 定时备份与实时备份

早期的数据备份大都采用了定时备份的方式，即每隔一段时间对数据进行备份。然而，当今社会中数据与人类的紧密度日益增强，如果数据备份存在时间间隔，那么一旦发生数据安全事故，备份间隔之内的数据极易丢失，且数据备份的时间间隔越大，丢失的数据也会越多。因此有了实时备份，即每当数据发生变化，实时执行备份任务，从而实现连续数据保护（continual data protection，缩写为 CDP）。实时备份可将复原点目标（recovery point objective，缩写为 RPO）和复原时间目标（recovery time objective，缩写为 RTO）两个指标值减小，其中，RPO 表示当灾难发生时允许丢失的数据量，RTO 表示系统恢复的时间。现在，容灾备份做得最好的银行系统是将指标设在 RPO = 0，RTO<5 分钟，用户数据一旦发生损坏，利用备份文件可以快速恢复，从而保障用户数据的安全性。

三、建立备份的简单过程

建立备份一般可以采用两种方法：一种是使用现成的备份软件，或有些系统中自带的备份功能；另一种是自己编写程序为系统和数据打造量身定制的备份。以前者为例，一般将备份保存在外部存储设备上，多数备份软件的备份过程如图 3.9 所示。

图 3.9　备份过程

四、数据还原

数据备份对保护数据具有相当重要的作用，但其根本目的是防止数据丢失，或能将丢失、损坏的数据重新加以利用，即数据备份的核心在于数据还原。一个无法还原的备份，对于任何系统而言都是毫无意义的。数据还原，顾名思义就是通过技术手段对保存在存储介质上的数据进行“抢救”和恢复。如果有了前述的数据备份，那么一旦出现问题，通过备份进行数据恢复将十分简单。大部分数据备份系统与数据库管理系统都有数据还原的功能，甚至可自动完成对数据及其运行环境的还原。

项 目 实 践　不同备份方式的数据还原

请根据备份的方式思考如何还原数据，填写表 3.5。

表 3.5　不同备份方式的数据还原

备份方式	还原需要的文件	还原操作步骤	确认结果	注意事项
全量备份				
增量备份				
差异备份				

探 究 活 动

1.如何运用实时备份与定时备份、全量备份和增量备份及差异备份等备份方式，防止在线考试系统中的数据丢失？

2.对于手机中的联系人信息、拍摄的照片和视频、社交平台上的各种往来消息和文件，你都能利用相应的手段备份并在需要的时候还原吗？

知 识 延 伸　冷备份与热备份

数据备份技术在性能和容量方面的优势是首要的，但更为关键的是要确保备份的数据具有完整性。在多用户、高可用服务器环境下，当多个用户正在访问数据时，如果备份系统也在执行拷贝操作，例如备份进程正在拷贝一个文件或数据库，同时发生了文件或数据库记录的更新，那么会发生备份的数据与原数据不一致的现象，即备份数据是不可用的。通常，为了保证数据完整性，可采用冷备份和热备份两种不同的备份方法。

冷备份又叫离线备份，指在执行备份操作时，服务器将不接受来自终端用户或应用系统对数据的更新。冷备份保证了数据的一致性和完整性，但备份进行期间服务器不可用，当备份的时间比较长时，就会对基于服务器的数据可用性造成影响。

热备份又称在线备份，即同步数据备份，当用户或应用正在更新数据时，系统也可以进行备份。它通过采用写前拷贝、软件快照（即软件的一个完全可用的拷贝）等技术解决备份过程中数据的一致性和完整性问题，其基本思想是：对于处于打开状态的数据文件，备份系统给予这些文件单独的写入或修改权限，保证在文件备份期间其他应用不能对其进行更新。

知 识 延 伸　数据备份系统

数据备份系统，也称为容灾系统或灾难恢复系统，就是通过特定的数据备份恢复机制，在各种灾难损害发生后，仍然能够最大限度地保障提供正常应用服务的计算机信息系统。

数据备份系统（如图 3.10所示）是通过在异地（可以是一个城市的两个不同的机房或者是两个不同的城市）建立和维护一个备份存储系统，利用地理上的分离来保证系统和数据对灾难性事件的抵御能力，对企业应用和数据库起到安全性、连续性等方面的支持作用，它是数据保持高可用性的最后一道防线。

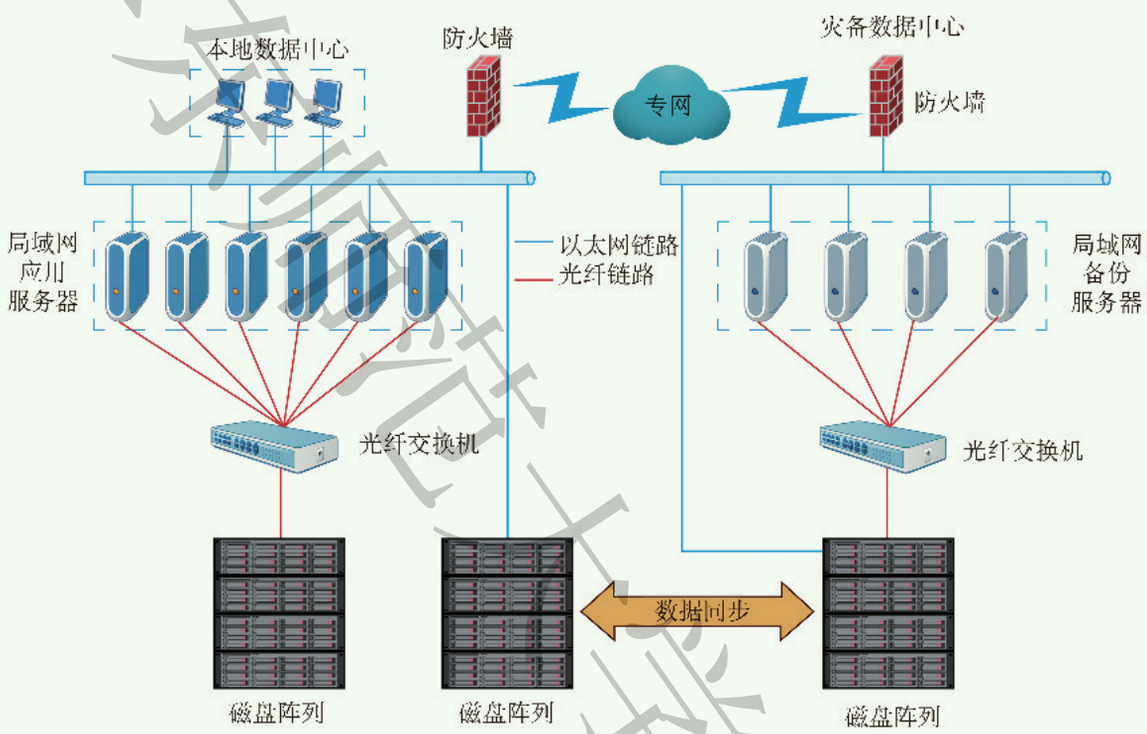


图 3.10　数据备份系统

数据备份的作用，不仅仅像房门的备用钥匙一样，当原来的钥匙丢失或损坏了才能派上用场。有时候，数据备份的作用，更像是我们为了留住美好时光而拍摄的照片，把暂时的状态永久地保存下来供我们分析和研究。虽然我们不能凭借一张儿时的照片回到从前，但却可以利用数据备份使一个存储系统乃至整个网络系统回到过去的某个时间状态，或者克隆出一个指定时间状态的系统，只要在这个时间点上有一个完整的系统数据备份。

数据备份系统按照所保障的内容，可以分为数据级数据备份系统和应用级数据备份系统。数据级数据备份系统是指建立一个异地的数据备份系统，该系统是对本地系统关键数据的复制，当出现灾难时，可将数据从异地系统迅速拷贝至本地系统，从而保证业务数据的完整性与一致性。应用级数据备份系统比数据级数据备份系统层次更高，即在异地建立一套完整的、与本地数据系统相当的备份应用系统（可以同本地应用系统互为备份，也可与本地应用系统共同工作），在灾难出现后，远程应用系统迅速透明地接管或承担本地应用系统的业务运行，保证信息系统提供的服务完整、可靠、安全。

数据备份系统包含数据存储、数据备份和高可用技术，是各种技术的综合应用，并且因用户需求的不同和使用环境的不同而有不同的解决方案。

知 识 延 伸　云备份

云备份主要指通过分布式文件系统集中网络中大量不同类型的存储设备，通过协同工作共同对外提供数据存储备份和业务访问的功能服务。目前，我们在使用手机时，大都可以采用云备份的方式将手机通讯录、短信、照片等数据存储备份在网络上。云备份的特点表现在三个方面：一是备份数据更加安全；二是支持多平台管理；三是数据传输加密更安全。

作业练习

完成本章学习后，请你选择一款日常使用的社交软件，了解人们在使用过程中有可能遇到哪些数据安全威胁、软件自身提供了哪些与数据安全相关的设置、人们在使用期间需注意哪些问题，并据此制订适合自己的数据安全策略。

第四章

数据分析

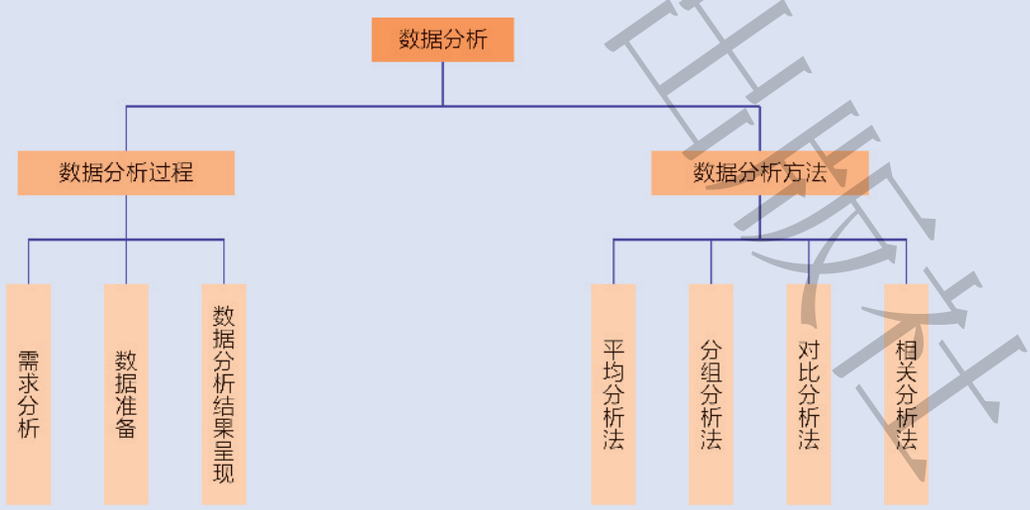
本章学习目标

- 了解数据准备的意义，了解缺失值处理的常用方法。
- 根据需要，选择恰当的方法对采集到的原始数据进行数据提取。
- 了解常用数据分析工具的适用场景。
- 了解常用的数据分析方法，如平均分析法、分组分析法、对比分析法、相关分析法等。
- 在实践中选用适当的数据分析工具，分析、呈现并解释数据。

随着信息技术的不断发展以及信息化应用水平的不断提升，人们在工作、生活的各个方面都生成了多种数据。人们收集、存储这些数据，并对其作适当处理。在实际使用中，为了让数据展示其特征以及发展规律与趋势，通常人们会根据具体需求对它们进行分析。数据分析是指为提取有用信息和形成结论而对收集来的大量数据加以详细研究和概括总结的过程。数据分析的数学基础在 20 世纪早期就已确立，但直到计算机出现，数据分析的实际操作才成为可能，并随着信息技术的发展得以推广。

目前，各行各业都在对收集到的数据作数据分析，例如气象数据分析、金融数据分析、教育数据分析、环境数据分析等。数据分析能为人们认识和改造世界提供新的数据资源，人们可以使用这些数据资源更好地进行科学研究、管理决策、公共服务等。数据分析为人们了解事物发展规律、预测事物发展趋势，甚至影响和改变事物发展进程进而改变生活打开了一扇大门。掌握了数据分析技术，也就在未来的发展和竞争中掌握了主动权。

本章知识结构



项目主题

上海市旅游景点数据分析

项·目·情·境

上海，又称“上海滩”，是一座极具现代化而又不失中国传统特色的国际大都市。她拥有深厚的文化底蕴和众多的历史古迹，老城厢里原汁原味的老弄堂，外滩沿岸拥有百年历史的欧式建筑，陆家嘴地区各式各样的摩天大楼，豫园旅游区里古典的江南园林……多元的城市元素汇聚在这里，让上海独具魅力；众多的景点串联起上海独特的旅行文化，让上海备受世界各地游客青睐。

一所与我校结成友好姐妹学校的国外学校组织了一个学生交流访问团，近期将来我校参观交流。作为学校对外交流社团的一名成员，你将与这些国外同学进行互动交流，通过数据分析向他们介绍上海旅游景点的概况。

首先，从某知名旅游网站上采集旅游景点数据，并对采集到的数据进行预处理；然后，为了了解上海旅游景点的区属分布而提取有效数据；接着，采用常用的数据分析方法分析提取区属信息后的数据，并使用图表呈现分析结果，以了解上海旅游景点的概况；最后，分析景点点评数与门票价格的相关性，以便根据分析结果决定是否推荐游玩某景点。

通过你的介绍，国外的同学可以选择他们喜爱的景点去游玩，了解上海的文化。

项·目·任·务

任务 1

对获取的上海市旅游景点原始数据进行处理，了解数据质量的重要性，对缺失数据进行填充，整合数据并从景点地址信息中提取区属信息。

任务 2

使用平均分析法计算上海 Top50景点的平均评分。

任务 3

使用分组分析法、对比分析法统计上海 Top50景点在各区中的分布并用图展示分布情况，以此了解上海旅游景点的概况。

任务 4

通过计算了解上海地区景点点评数与门票价格之间是否有相关性，从而可以根据某景点的基本情况进行大致推理，以此决定是否推荐游玩该景点。

第一节　数据准备

数据是数据分析要处理的对象，数据分析流程一般从数据准备开始，即根据业务需求，采集相应的数据。由于各种原因，采集的原始数据中可能含有一些错误数据或缺失部分数据，这就需要在数据准备时进行处理。此外，对于一个特定的应用来说，并不是数据越多越好，为了更好地进行数据分析，还需要从原始数据中提取相关的有效数据。

问　题　思　考

每个人心中都有能代表上海的景点。为了客观地展示上海的旅游景点概况，本项目将分析从某知名旅游网站上采集的上海 Top50景点的数据，用数据“说话”。由于是从网站上采集的数据，采集过程中有一些人工操作，因此可能包含一些错误数据或缺失部分数据，需要改正或处理。另外，需要各景点的地理分布，因此需要采集地区数据。

请思考：

1. 可能会存在哪些数据质量问题？
2. 如何知道数据中含有缺失值？它有哪些处理方法？
3. 哪些是常用的数据分析工具？

一、数据预处理

数据质量的优劣对数据分析结果具有很大的影响。高质量的数据是分析结果准确的重要保证。例如，某学校采集学生的身高和体重数据，用于计算 BMI 指数并统计 BMI 指数偏小、正常、偏大人群占总人数的比例，供食堂参考。如果采集的身高、体重数据格式不规范，例如填写体重数据时，有的学生以斤为单位，有的学生以公斤为单位，从数据上看并无异样，但是以斤为单位计算出的 BMI 值是正确值的两倍，数据分析结果中肥胖学生的比例就会增大，食堂依据该数据分析结果可能会提供更多低脂肪饮食餐供学生选择。如果数据单位统一，数据分析结果可能是低体重学生人数较多，食堂原本应该增加更容易被人体吸收的食品。可见，准确的数据有利于得出准确的结果，错误的数据会影响分析结果，进而影响决策，产生不良甚至严重的后果。另外，除了数据准确以外，数据达到一定的数量也是数据分析结果准确的必要条件。

数据质量问题产生的原因有很多，数据的采集、数据的传输、数据

的存储，甚至业务需求分析等方面都有可能导致质量问题。不同领域、不同业务、不同应用的数据质量需求不尽相同。数据质量问题是多方面的，评估数据是否达到预先设定的质量要求，可以通过表 4.1 中的几个方面来进行判断。

表 4.1　数据质量评估标准与数据质量问题

评估标准	数据质量问题
准确性	不正确的数据或含噪声的数据(包含错误或存在偏离期望的值)
完整性	不完整的数据
唯一性	数据重复或者数据属性重复
一致性	数据记录不规范
时效性	数据过时
相关性	应用所需要的数据缺失

由于各种原因，采集的数据有可能不完整，例如，部分调查问卷中一些选项漏填、数据采集的传感器中途损坏等，需要进行缺失值处理。缺失值处理的常用方法如下：

■ 忽略该条记录

这是最简单的方法，但其缺点是，该条记录中除了缺失值以外的剩余属性数据也将不可用，这些剩余属性数据很可能对要进行的数据分析任务是有用的。

■ 人工估算填写缺失值

该方法很费时，并且当数据集很大、缺失值很多时，该方法可能行不通。

■ 使用一个全局常量填充缺失值

即将缺失值用同一个常量填充，该方法很有可能会导致错误的分析结果(例如，将 999999 作为值参与平均数等的计算)。

■ 使用属性的平均数(详见本章第二节中的“一、平均分析法”)填充缺失值

例如，假定在某班学生看课外书的分析中，原始数据的年龄属性有缺失值，可以使用该班学生的年龄值平均数来填充。

■ 使用与给定元组属同一类的所有样本的属性平均数来填充

例如，针对上述提到的原始数据的年龄属性有缺失值，可以分别计算出男生、女生的年龄值平均数，根据缺失值所在记录的性别项选择相应性别的年龄值平均数填充。

项 目 实 践　对上海市旅游景点数据进行预处理

为了解上海市旅游景点的概况，本项目从某知名旅游网站上采集了上海 Top50景点数据（如表 4.2）进行分析。但观察以后发现，其中的景点级别数据有缺失，并且地址数据不规范，提取各景点的区属信息较为困难。因此，需要结合上海市文化和旅游局官方网站 A级景点数据（如表 4.3），进行数据预处理。

表 4.2　从某知名旅游网站上采集的上海 Top50景点数据

排名	景点名称	评分	点评数	地址	市场价	景点级别
1	上海迪士尼度假区	4.6	94808	浦东新区世纪大道 1号	575	
2	外滩	4.7	26041	上海市黄浦区中山东一路		
3	东方明珠	4.6	63859	上海市浦东新区陆家嘴世纪大道 1号	300	AAAAA
4	南京路步行街	4.4	5414	上海市黄浦区河南中路		
5	城隍庙旅游区	4.4	3503	上海市黄浦区方浜中路 249号		
6	上海野生动物园	4.7	42144	上海市浦东新区南六公路 178号	130	AAAAA
7	上海海洋水族馆	4.5	12754	上海市浦东新区陆家嘴环路 1388号	160	
……	……	……	……	……	……	……
24	上海城隍庙	4.5	817	黄浦区方浜中路 249号		
……	……	……	……	……	……	……
50	枫泾古镇	4.3	1129	金山区枫泾镇亭枫公路 8588弄 28号	30	AAAA

表 4.3　上海市文化和旅游局官方网站 A级景点数据

A级	景点	地址
5A	上海东方明珠广播电视塔	浦东新区世纪大道 1号
5A	上海野生动物园	浦东新区南六公路 178号
5A	上海科技馆	浦东新区世纪大道 2000号
4A	上海博物馆	黄浦区人民大道 201号
4A	上海佘山国家森林公园	松江区外青松公路 9258号
4A	上海豫园	黄浦区安仁街 218号
……	……	……
3A	上海南汇大团桃园	浦东新区大团镇赵桥村 888号
3A	上海南汇桃花村	浦东新区惠南镇北门路 289号
3A	上海大宁灵石公园	闸北区广中西路 288号
……	……	……

本项目中使用的某知名旅游网站上海 Top50景点的评分、点评数等数据以及上海市文化和旅游局官方网站 A级景点数据都是有时效性的，为了使分析结果准确，需要在分析的时候根据实时数据进行修正（本项目所用数据为 2018年 8月采集）。表 4.2的数据中，第 5条与第 24条景点重复，需要合并。另外，该表中景点级别数据也有缺失，因此将该表中的数据与表 4.3中的数据进行整合，不是 A级景点的填写缺失值为“未定级”，景点级别数据使用上海市文化和旅游局官方网站的格式，例如 5A。结果如表 4.4所示。

表 4.4 整合后的数据

排名	景点名称	评分	点评数	地址	市场价	景点级别
1	上海迪士尼度假区	4.6	94808	浦东新区世纪大道 1号	575	未定级
2	外滩	4.7	26041	黄浦区中山东一路		未定级
3	上海东方明珠广播电视塔	4.6	63859	浦东新区世纪大道 1号	300	5A
4	南京路步行街	4.4	5414	黄浦区河南中路		未定级
5	城隍庙旅游区	4.4	4320	黄浦区方浜中路 249号		未定级
6	上海野生动物园	4.7	42144	浦东新区南六公路 178号	130	5A
7	上海海洋水族馆	4.5	12754	浦东新区陆家嘴环路 1388号	160	4A
……	……	……	……	……	……	……
24	上海影视乐园	4.4	5469	松江区车墩镇北松公路 4915号	80	4A
……	……	……	……	……	……	……
49	枫泾古镇	4.3	1129	金山区亭枫公路 8588弄 28号	30	4A

二、数据提取

从现有数据源中提取有效数据的过程称为数据提取。数据提取本质上就是对数据源中的数据进行选择加工的过程。

数据提取的常用方法：

第一种，将数据导入数据库中，根据提取的规则，采用合适的查询得到有效数据；

第二种，将数据导入电子表格处理软件或其他数据分析工具中，使用数据筛选（数据筛选是指在数据源的数据表中将满足查询条件的记录选择出来，查询条件一般通过逻辑表达式书写）提取出有效数据；

第三种，使用程序设计语言编写程序，读入原始数据，并根据规则

提取出有效数据。

数据预处理与数据提取以及后续的分析都需要根据具体应用领域，选取合适的工具进行，以提高效率。常用的数据分析工具有 R、Python、Excel、SPSS、Stata、Orange、Weka、SAS、MATLAB 等。它们的优势和缺点、常用应用领域有所不同，如表 4.5 所示。

表 4.5　常用数据分析工具

软件名	优势	缺点	常用应用领域
R	免费，程序库支持、可视化	较难学习	金融、统计
Python	免费，Python是通用编程语言，有很多库支持，如 ScPy、NumPy、Matpbtlb	需编程	工科
Excel	极易使用	运行效率较低、样本量受限、统计功能不完善	小样本数据分析、商务运用
SPSS	易学易用，统计功能全面	运行效率不高	统计学、社会科学
Stata	使用简便，对用户友好，支持程序编写，计算速度快	对同时处理多个文件、数据文件大小有限制	商务运用、自然科学
Orange	免费，具有拖拽式、流程化的可视化编程前端	侧重数据挖掘、统计分析，功能较少	数据挖掘
Weka	免费，基于 Java环境，集合了大量算法	可视化功能相对简单，算法扩展需编程	数据控制
SAS	统计功能强大、大样本分析	需编程	商务运用、政府、科学统计
MATLAB	矩阵支持强大、可视化	需编程，统计功能不完善	工科、自然科学

探　究　活　动　使用 Python编写程序进行数据提取

为了解上海市旅游景点的概况，需要分地区分析数据，因此需要在整合好的数据中提取区属信息。

表 4.4中有景点地址信息，但在具体分析时并不需要详细地址，只需要区属信息，因此可以将其从地址信息中提取出来，Python代码如下：

```
#提取区属信息.py
import pandas as pd
data = pd.read_csv('travel_data0.csv', encoding = 'gb2312')
data['区属'] = data['地址'].str.split('区').str[0] + '区'
data.to_csv('travel_data1.csv', encoding = 'gb2312')
```

也可以将数据导入数据库中，使用如下所示的查询语句得出结果：

```
SELECT *, Concat(Substring_index(地址, '区', 1),'区') as '区属'
FROM Travel_date0
```

部分结果如表 4.6所示。

表 4.6 提取区属信息后的数据表示例

排名	景点名称	评分	点评数	地址	市场价	景点级别	区属
1	上海迪士尼度假区	4.6	94808	浦东新区世纪大道 1号	575	未定级	浦东新区
2	外滩	4.7	26041	黄浦区中山东一路		未定级	黄浦区
3	上海东方明珠广播电视塔	4.6	63859	浦东新区世纪大道 1号	300	5A	浦东新区
4	南京路步行街	4.4	5414	黄浦区河南中路		未定级	黄浦区

体 验 思 考 课程辅导书销售信息的数据预处理和数据提取

收集若干本常用课程辅导书的基本信息及其在网上书城中的销售信息，提出数据分析需求，对数据作预处理并进行数据提取。

第二节　数据分析方法与呈现

大数据时代，人们可以收集到各种各样的数据，通过数据分析来发现数据的特征及其发展规律与趋势，从而帮助决策。数据分析的方法有很多种，平均分析法、分组分析法、对比分析法、相关分析法等都是常用的方法。数据分析所产生的结果数据除了以数据表的形式呈现，还可以用图进行展示。用图来表示数据的方法属于数据可视化范畴。相对于数据表，数据可视化的呈现方式更为直观，更容易让人们发现数据的特征与规律。

问　题　思　考

分析某知名旅游网站上海 Top50景点的评分、点评数、区属、市场价、景点级别等数据，可以计算出上海 Top50景点整体的评分以及上海各区景点的评分、Top50景点和 A级景点在各区的分布，了解上海旅游的概况；还可以通过计算来判断上海地区景点评分与点评数、门票价格、景点级别之间是否有相关性，从而根据某景点的基本情况对其评分进行大致推理，以决定是否推荐该景点。

请思考：

1. 景点评分的平均数如何计算？
2. 如何获得各区 A 级景点的分布？
3. 比较结果如何呈现？
4. 如何判断景点评分与点评数、门票价格、景点级别之间是否有相关性？

一、平均分析法

平均数计算是最基础的分析，几乎所有的数据分析工具都提供了简便的平均数计算方法，例如电子表格软件使用函数计算、数据库管理软件使用查询等。

平均分析法是一种利用平均指标对现象进行分析的方法。平均指标又称为集中趋势量数或平均数，反映总体在一定时间、地点条件下某一数量特征的一般水平，用来说明总体的集中趋势，分为数值平均数与位置平均数两类。

数值平均数是使用总体中每个数据的数值根据给定的计算方法计算得出，常用的有算术平均数、调和平均数、几何平均数、平方平均数等。

简单算术平均数的计算公式如下：

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}。$$

本书第一章第二节的项目实践中，用简单移动平均法预测商品月销量时，就使用了简单算术平均数来进行预测。

除了简单算术平均数以外，算术平均数还有加权算术平均数，计算公式如下：

$$\bar{x}=\frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{f_1+f_2+\cdots+f_n}。$$

其中 f_1，f_2，…，f_n 为 x_1，x_2，…，x_n 的权值。

位置平均数是根据数据的位置或数据出现次数指定的数，常用的有中位数和众数。

中位数是指将总体中的数据从小到大（或从大到小）排列后，居于最中间位置的一个数（或最中间两个数据的平均数）。中位数是样本数据所占频率的等分线，它不受少数几个极端值的影响，有时用它代表全体数据的一般水平更合适。例如，调查网上销售量排名前 10 的钢笔的价格，得到数据集“22.8、22.8、29、78、39、34、22.8、26、18.8、38”，它们的平均数为 33.12。由于其中有一种钢笔的价格为 78 元，比其他钢笔的价格高很多，导致算出的平均数比这些钢笔价格的一般水平要高。事实上，中位数更能代表这些钢笔价格的一般水平。将这些钢笔的价格按照升序排列为“18.8、22.8、22.8、22.8、26、29、34、38、39、78”，则中位数为中间两个数 26、29 的平均数，即 27.5。

众数，顾名思义，就是总体中出现次数最多的数据。例如上述 10 种钢笔价格的数据中，22.8 出现的次数最多，则这 10 种钢笔价格的众数为 22.8。一组数据的众数可能有好几个，即在一组数据中出现次数最多的数据可能不止一个。

一般来说，定量数据的集中趋势使用数值平均数或中位数来描述，定性数据的集中趋势使用众数来描述。

项 目 实 践　计算上海 Top50景点的平均评分

上海有很多景点，人们对上海景点的整体评价如何呢？上海景点的整体水平是好还是差呢？可以根据某知名旅游网站上海 Top50景点评分来计算这些景点的平均评分，以初步了解上海景点的整体情况。

使用本章第一节“二、数据提取”中准备好的数据，计算某知名旅游网站上海 Top50景点评分的算术平均数，用 Python编写程序如下：

```
import pandas as pd
df = pd.read_csv('travel_data1.csv', encoding = 'gb2312')
print('{: .2f}'.format(df.评分.mean()))
print('{: .2f}'.format((df['评分'] * df['点评数']).sum()/df['点评数'].sum()))
```

也可以将数据导入数据库中，使用如下所示的查询语句得出结果：

```
SELECT Avg(评分), Sum(评分 * 点评数)/Sum(点评数)
FROM Travel_date1
```

通过两种方式，均可得出简单算术平均数为 4.53，加权算术平均数为 4.60。由于每个景点的点评数不一样，因此加权算术平均数较之简单算术平均数用于反映整体评分水平更为恰当。评分 4.60在该网站中属于较高水平，说明上海旅游景点整体在网民心目中还是不错的。

二、分组分析法与对比分析法

1. 分组分析法

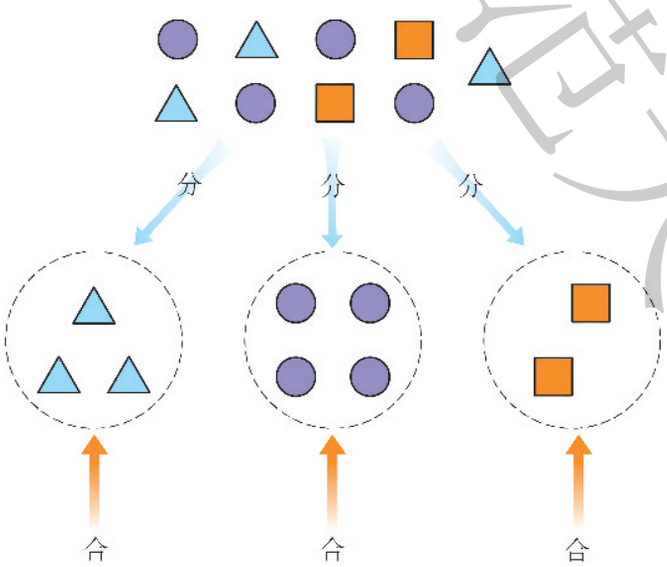


图 4.1 分组分析示意图

分组分析是指根据分析的目的要求，按照一定的标志将所研究的数据分析对象总体划分为若干部分，加以分组整理，进行观察、分析，以揭示其内在联系和规律性，其示意图如图 4.1 所示。分组具有两方面的含义，对总体而言是“分”，即将总体中的对象按照它们的差异区分为若干部分；对个体而言是“合”，即将性质相同的对象合并在一起。

分组分析法是数据分析的基本方法，分组分析法的关键在于正确选择分组标志，以提高分析结果的科学性和真实性。可以使用事物的性质属性进行分组，这种分组标志称为品质标志；也可以使用数量特征进行分组，这种分组标志称为数量标志。品质标志使用文字来表现，数量标志使用数值来表现。在分组之前应考虑研究问题的目的和任务，并对数据对象的特征和发展规律进行理论分析，以便选出能反映问题本质的分组标志。

使用数量标志分组时需要确定组数与组距。组数是根据数据的特点来判断决定的，应该适中。如果组数太少，数据的分布会过于集

中；如果组数太多，数据的分布会过于分散。这两种情况都不便于观察数据分布的特征和规律。组距是一个组中的最大值与最小值之差。根据组距是否相等，可以将分组分为等距分组和异距分组两类。当总体中的数据分布比较均衡、变动比较均匀时适合采用等距分组；相反，当总体中的数据分布不太均衡时采用异距分组更能体现数据关于现象的本质特征。等距分组的组距与组数是反比关系，组距可由全部数据的最大值和最小值及所分的组数来确定，反之，组数也可以由全部数据的最大值和最小值与组距确定。

例如，将某单位在职职工体检数据按年龄分组，由于该数据样本中年龄分布较为均衡，可以使用等距分组。若最小年龄为 18 岁，最大年龄为 59 岁，组数定为 4，则组距为 $(59-18)\div 4\approx 10$，年龄分组为 18 ～ 28 岁、29 ～ 39 岁、40 ～ 50 岁、51 ～ 59 岁。

分组是为了便于对比，数据分组以后，可以将组内对象进行对比，也可以对比组与组之间的差异。

2. 对比分析法

对比分析是指将两个或两个以上的数据进行比较，分析其差异，揭示这些数据所代表的事物的发展变化情况和规律，分辨出事物的性质，从而深刻地认识事物的特点和本质。在数据分析中，对海量复杂的数据单独作分析通常很难发现其特征，而通过对比，数据就有了“好坏”之分。

对比分析法也称为比较分析法，是自然科学、社会科学以及日常生活中常用的分析方法之一。对比分析法的特点是通过准确、量化的数据直观地反映事物某方面的变化或差距。对比分为横比与纵比，横比指在同一时期对不同的事物进行比较，纵比指在不同时期对同一事物进行比较。

在数据分析中，对比的统计指标可以采用绝对数进行分析，也可以采用统计相对数。统计相对数又叫统计相对指标，简称相对数，是两个有联系的统计指标的对比数值。常用的相对数有计划完成程度相对数、结构相对数、利用程度相对数、比较相对数、强度相对数、动态相对数。

分组分析法、对比分析法也会经常结合平均分析法一起使用，即在分组以后，组内取平均数再对比。

探 究 活 动　分析上海 Top50景点在各区的分布

上海是一个国际化大都市，包含了若干个行政区，哪个区好玩的景点最多呢？朋友来上海游玩，在时间有限的情况下，集中在哪个区玩能玩到更多的好玩景点呢？

我们使用本章第一节“二、数据提取”中准备好的数据，使用分组分析统计某知名旅游网站上海 Top50景点的区属分布，并将结果进行对比分析，来了解哪个区好玩的景点最多。

可以使用 Python编写程序来完成统计，程序代码如下：

```
#pivot_data.py
import pandas as pd
df = pd.read_csv('travel_data1.csv', encoding = 'gb2312')
print(pd.pivot_table(df, columns = '景点级别', index = '区属', aggfunc = 'count', values = '景点名称'))
```

结果如图 4.2所示：

景点级别	3A	4A	5A	未定级
区属				
宝山区	NaN	1.0	NaN	NaN
崇明区	NaN	1.0	NaN	1.0
徐汇区	NaN	1.0	NaN	1.0
普陀区	NaN	1.0	NaN	NaN
杨浦区	NaN	NaN	NaN	1.0
松江区	NaN	3.0	NaN	2.0
浦东新区	NaN	5.0	3.0	4.0
虹口区	NaN	NaN	NaN	3.0
金山区	NaN	1.0	NaN	NaN
长宁区	NaN	1.0	NaN	NaN
闵行区	NaN	1.0	NaN	1.0
青浦区	NaN	1.0	NaN	NaN
静安区	NaN	NaN	NaN	3.0
黄浦区	1.0	3.0	NaN	10.0

注：NaN 在 Python 中表示空值

图 4.2　景点区属分布统计结果

也可以将数据导入数据库中，使用如下所示的查询语句得出分组统计结果：

```
SELECT 区属,景点级别,Count(景点名称)AS 计数
FROM Travel_data1
GROUP BY 区属,景点级别
```

还可以使用电子表格软件的数据透视表进行统计，如图 4.3所示。

计数项:景点名称	列标签				
行标签	3A	4A	5A	未定级	总计
宝山区		1			1
崇明区		1		1	2
虹口区				3	3
黄浦区	1	3		10	14
金山区		1			1
静安区				3	3
闵行区		1		1	2
浦东新区		5	3	4	12
普陀区		1			1
青浦区		1			1
松江区		3		2	5
徐汇区		1		1	2
杨浦区				1	1
长宁区		1			1
总计	1	19	3	26	49

图 4.3　使用电子表格统计景点的区属分布

根据统计结果对比可以得知：黄浦区和浦东新区的景点较之其他区是比较多的，且浦东新区的 A级景点更多些。

三、数据可视化

数据分析的结果不仅可以用数据表形式来表示，也可以使用图来表示。例如，人们早就使用地图来表示地域信息。数据可视化是研究如何将数据以图片或图形的方式展现的科学，其主要目的是借助图形化手段，将数据通过可视的、甚至交互的方式进行展示，形象、直观地表达数据蕴含的信息和规律。

数据可视化使人们不再局限于通过数据表来观察和分析数据，还能以更直观的方式看到数据。例如在证券软件中含有每天的行情数据、交易数据等，通过算法可对数据进行分析以发现数据的很多内涵，再使用图表将分析结果简单明了地呈现给股民。

实现数据可视化的常用方法是绘制各种图表，常见图表如下：

1. 柱形图

柱形图是一种最基本的图表，可以包含一组或者多组数据的比较展示，但添加多组数据时比较难以专注其中的一组并得出结论，因此普遍采用一组数据的展示。几乎所有的提供数据可视化功能的数据分析工具都可以实现绘制柱形图。

例如，5 位学生某次考试的成绩数据如表 4.7 所示，可以据此绘制出成绩总分的柱形图，如图 4.4 所示。从图 4.4 中可以很直观地发现王妤总分最高。

表 4.7　5位学生某次考试的成绩表

姓名	语文	数学	英语	总分
陈纯	88	87	85	260
方小磊	93	88	90	271
王妤	82	99	96	277
彭子晖	97	94	84	275
丁海斌	97	94	76	267

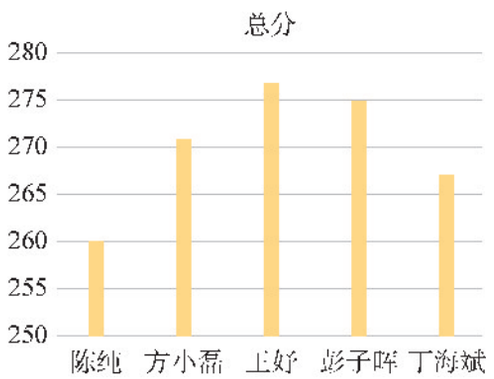


图 4.4　5位学生某次考试成绩总分的柱形图

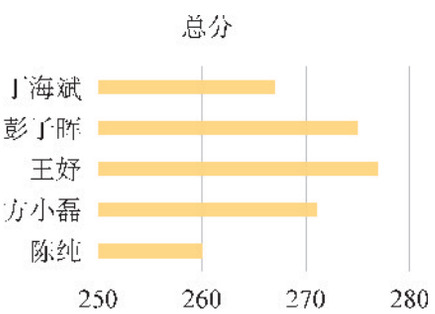


图 4.5　5位学生某次考试成绩总分的条形图

2. 条形图

条形图就像将柱形图旋转 90°，有时也被称为水平柱形图（柱形图对应被称为竖直条形图），它也可以包含一组或多组数据的比较展示。由于大多数人的读写习惯为从左向右，因此与柱形图相比，条形图更容易阅读。例如，根据表 4.7 中的数据绘制 5 位学生某次考试成绩总分的条形图如图 4.5 所示。

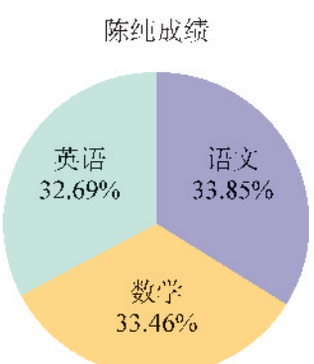


图 4.6　学生陈纯各科成绩饼图

3. 饼图

饼图反映某个部分占整体的比重。例如，从表 4.7 中选取陈纯的成绩绘制各科成绩饼图如图 4.6 所示。从中可以发现，陈纯 3 门课程的成绩占总分的比例差不多。

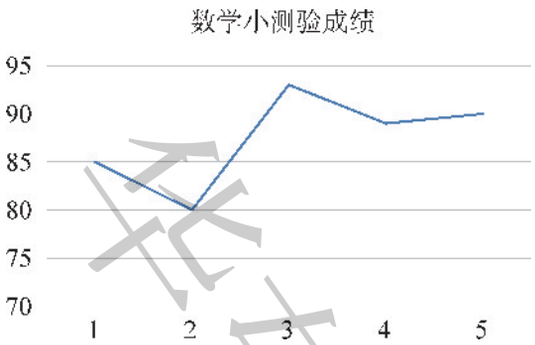


图 4.7　某学生 5次数学小测验成绩的折线图

4. 折线图

折线图也称为线图，是一种非常适合二维数据集展示的图表，尤其在展示数据的时间序列方面很有优势，它还适合用于多个二维数据集的比较。

例如，某学生 5 次数学小测验成绩随时间先后记录为“85、80、93、89、90”，绘制折线图如图 4.7 所示。从中可以看出，该学生的数学成绩总体呈上升趋势，其间略有波动。

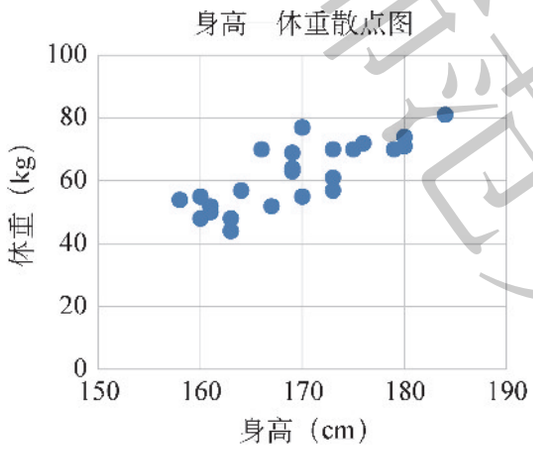


图 4.8　身高—体重散点图

5. 散点图

散点图用于显示和比较数值，表示因变量随自变量而变化的大致趋势，反映不同变量之间的相关关系，适合于描述数据集合的分布状况，常用于科学数据、统计数据和工程数据。

例如，某班学生的身高和体重数据如表 4.8 所示，绘制身高—体重散点图如图 4.8 所示。从中可以看出，该班学生的身高、体重在某个范围内比较集中，并且随着身高的增加，体重大多也增加了，因此身高和体重为正相关（详见本节“四、相关分析法”）。

表 4.8　某班学生的身高和体重表

序号	身高（cm）	体重（kg）	序号	身高（cm）	体重（kg）	序号	身高（cm）	体重（kg）	序号	身高（cm）	体重（kg）
1	176	72	7	160	48	13	161	50	19	160	55
2	161	52	8	163	44	14	173	61	20	170	55
3	173	70	9	175	70	15	180	71	21	180	74
4	167	52	10	163	48	16	169	69	22	170	77
5	166	70	11	164	57	17	169	64	23	158	54
6	169	63	12	173	57	18	184	81	24	179	70

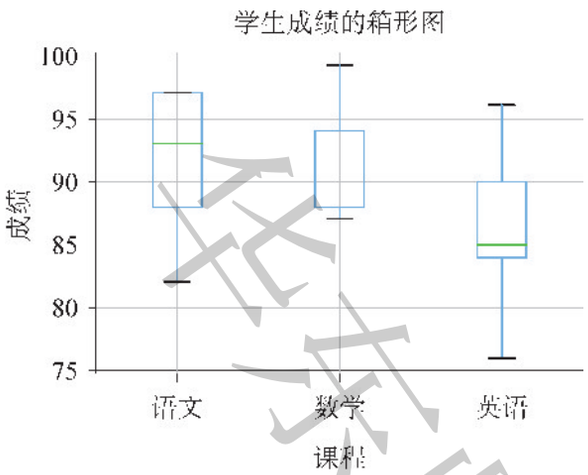


图 4.9　5位学生成绩的箱形图

6. 箱形图

箱形图采用分位数来描述数据集合的分布情况，上、下两条横线表示最大值和最小值，矩形的上、下边表示上四分位数与下四分位数，矩形中间的横线表示中位数。

根据表 4.7 中 5 位学生的成绩数据绘制箱形图如图 4.9 所示(注意：数学成绩中，中位数与上四分位数都是 94，因此矩形中间横线与矩形的上边重合)。从中可以看出：各科成绩的分布都不太均匀；语文成绩除高分段外，其他部分分布较均匀；英语成绩除中下分段外，其他部分分布较均匀。

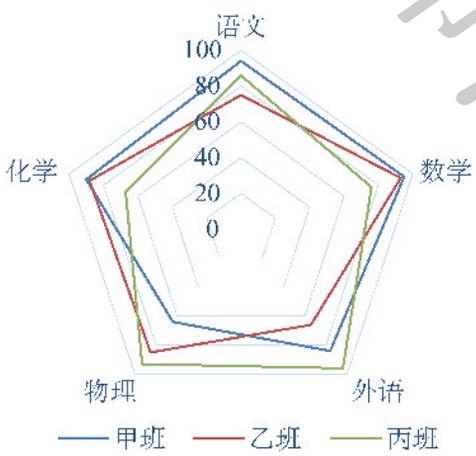


图 4.10　某学校 3个班 5门课程平均成绩雷达图

7. 雷达图

雷达图，也称为网络图、蜘蛛图等，是由一组坐标和多个同心多边形组成的图表，主要用于在同一坐标系内展示多指标的分析比较情况，是综合评价中常用的一种图表。

根据表 4.9 中某学校 3 个班 5 门课程平均成绩绘制雷达图如图 4.10 所示。从中可以看出：甲班语文、数学、化学成绩较好，物理较弱；乙班语文、外语较弱；丙班数学、化学较弱。

表 4.9　某学校 3个班 5门课程平均成绩表

	语文	数学	外语	物理	化学
甲班	94	95	84	64	90
乙班	75	93	66	85	88
丙班	86	76	96	93	67

8. 热力图

热力图，也称为热图，是使用颜色显示不同区域的用户行为数据

的图示。热力图用于各种形式的分析，常用于显示用户在某个特定网页上的行为活动或者是在某些地理区域上的数据展示。

根据表 4.10 中的数据绘制热力图，其示意图如图 4.11 所示，从中可以看出：区 6 内的 A 级景点最多。

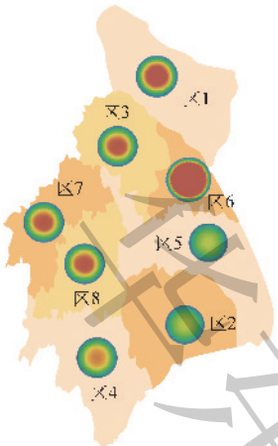


图 4.11　某市某年各区 A 级旅游景点分布热力图（示意图）

表 4.10　某市某年各区 A 级旅游景点分布表

区	数量	区	数量	区	数量
区 1	10	区 4	7	区 7	8
区 2	5	区 5	5	区 8	8
区 3	8	区 6	27		

在使用时，可以将基本图表进行组合与变换，形成新的图，例如甘特图。如图 4.12 所示，通过条形图组合来显示项目、进度随着时间进展的情况。

也可以根据数据分析的主题，使用图表与其他图形的组合形成表现力更强、更美观的新图。如图 4.13 所呈现的某小学所有班级的男女生人数，比普通条形图更一目了然。

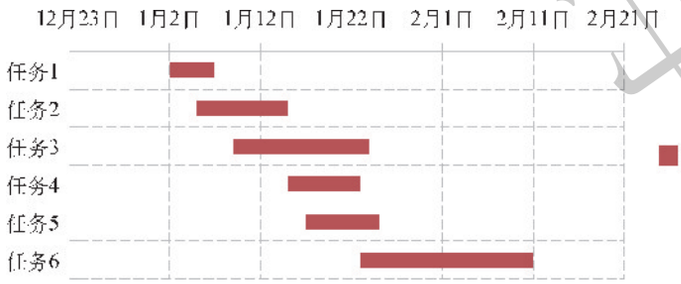


图 4.12　甘特图示例

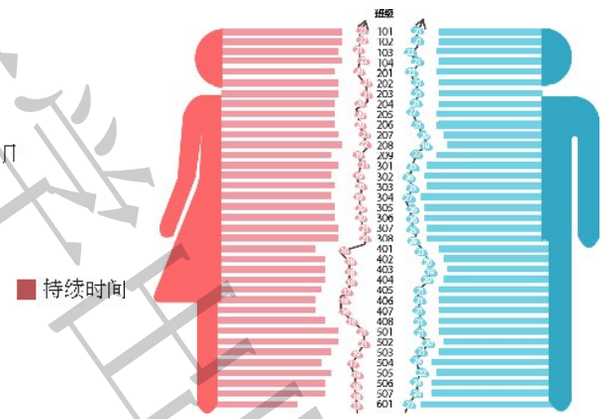


图 4.13　某小学所有班级男女生人数图

图 4.14　《唐诗三百首》中作者名字词云

除了图表以外，也可以使用其他图形呈现数据，用数据讲故事，例如词云。词云是对文本中词语的频率统计的可视化表示，通常会结合分词工具一起使用。图 4.14 为《唐诗三百首》中作者名字词云，从中可以迅速看出杜甫是《唐诗三百首》中诗最多的诗人。

项 目 实 践　用图展现分析结果

上一个活动中，我们分析了某知名旅游网站上海 Top50景点在各区的分布，使用数据表呈现了分析结果。但是基于数据表的观察不太直观，本活动我们将使用图表来一目了然地展示分析结果。

本活动中，使用上一个活动中的分析结果绘制某知名旅游网站上海 Top50景点在各区分布的堆积柱形图。可以使用电子表格软件绘制，方法为选择数据区域，插入“图表”中的“堆积柱形图”；或使用 Python 编写程序来实现，代码如下：

```
import pandas as pd
import matplotlib.pyplot as plt      #导入绘图库
plt.rcParams['font.sans-serif'] = ['SimHei']      #用来正常显示中文标签
df = pd.read_csv('result.csv', encoding = 'gb2312')      #导入数据文件
df.index = df.区属
#绘制柱形图，stacked 参数设置柱形图的类型为堆积
cht = df.plot.bar(title = '上海各区景点分布', stacked = True)
cht.set_ylabel('个数')
cht.set_xlabel('区属')
plt.show()      #显示图表，有些集成开发环境下此句可省略
```

绘制的堆积柱形图如图 4.15所示。

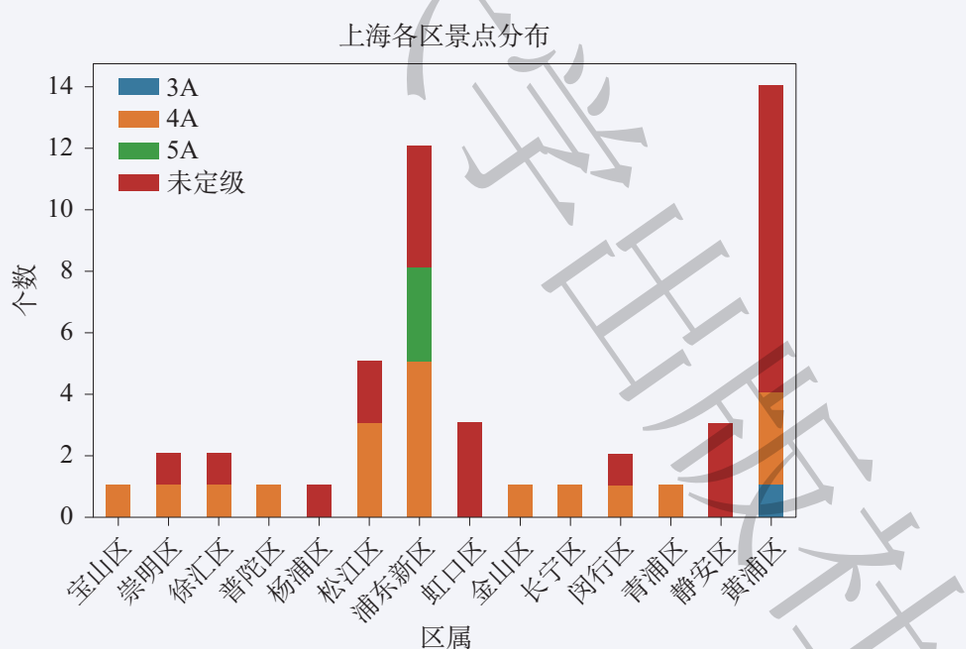


图 4.15　上海各区景点分布的堆积柱形图

从图 4.15中可以很直观地看出浦东新区和黄浦区的景点数量明显多于其他区。

四、相关分析法

世间万物都是存在相关联系的。例如，儿童的年龄和身高显然是有关系的，但是又没有确定的数量上的对应关系，不能通过年龄得出身高，或通过身高得出年龄。这种对应关系称为相关关系。相关关系有两个明显的特点：一是现象之间确实存在依存关系，即某一现象的变化会引起另一现象的变化；二是现象之间的依存关系是不严格的，无法用数学公式表示。例如，商品的单价和销量具有相关关系，单价的变化会引起销量的变化，但是并不能找出一个具体的由单价推出销量的函数来描述这种关系。

相关分析法是研究现象之间是否存在某种依存关系的一种统计方法。它反映现象之间的依存关系在数量上不严格，是一种非确定性的对应关系。相关分析的应用领域非常广泛，在自然科学领域和社会科学领域中，经常需要对两个或两个以上相关的变量作相关分析。如果研究的是一个变量对另一个变量的影响，则所研究的相关关系称为单相关；如果是分析一个变量受若干个变量的影响，则所研究的相关关系称为复相关。在相关关系中，根据两个现象相关的方向，分为正相关和负相关。正相关是指两个现象的标志（也称为相关变量）的数量变动方向一致，即你升我也升，你降我也降；负相关是指两个相关变量的数量变动方向相反，即你升我降。例如，儿童的年龄和身高是正相关，而商品的单价与销量则是负相关。

在相关分析中，根据两个现象相关的程度，分为完全相关、不完全相关和不相关。完全相关指一个现象的数量变化可以由另一个现象的数量变化来确定，可以通过某个函数计算出来，也称为函数关系；不相关指两个现象毫无关系，彼此独立，互不影响；如果两个现象的关系介于完全相关和不相关之间，则称为不完全相关。

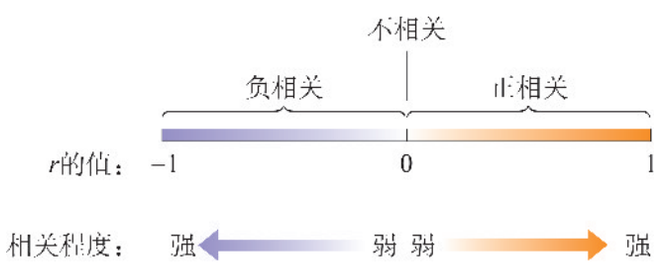


图 4.16　解读相关系数 r

通常使用相关系数 r 来表示两个相关变量之间的相关方向和相关程度。如图 4.16 所示：

相关系数 r 的值在 -1 和 1 之间。

正相关时，r 的值在 0 和 1 之间；负相关时，r 的值在 -1 和 0 之间；r 等于 0 时，表示两个变量不相关。

r 的绝对值越接近 1，两个变量的相关

程度越强；r 的绝对值越接近 0，两个变量的相关程度越弱。

常用的皮尔森(Pearson)相关系数是按积差方法计算的，即以两个相关变量与各自平均数的差为基础，通过两个差相乘来反映两个变量之间的相关程度。

计算公式如下：

$$r=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}。$$

其中 X、Y 是两个变量，有多组数，$\overline{X}$、$\overline{Y}$ 为算术平均数。

除了计算以外，也可以绘制散点图帮助确定相关变量之间的关系，即在直角坐标系中根据两个成对的变量的数据绘制点，查看这些点的分布规律。

探 究 活 动　景点点评数和门票价格有关系吗？

上海有很多景点，有些是收费景点，有些是免费景点。那么，收费景点的门票价格与景点人气有没有关联呢？直觉上可能是门票价格高的景点人会去得少，事实是不是这样呢？由于并未采集景点实际游玩人数，我们可用某知名旅游网站上海 Top50景点的点评数来代替这一指标，认为点评数多的景点游玩人数肯定也多，选取同一级别的景点数据来分析景点点评数和门票价格的相关关系。

我们使用 Python编写程序计算某知名旅游网站上海 Top50景点中收费 4A景点的点评数与门票价格的相关系数，并通过绘制点评数与门票价格的散点图来查看两者之间的关系。

Python的程序代码如下：

```
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
df0 = pd.read_csv('travel_data1.csv', encoding = 'gb2312')
#选择 4A 景点并删除包含缺失值的行
df1 = df0[df0.景点级别 == '4A'].dropna(axis = 0, how = 'any')
x = df1.点评数
y = df1.市场价
print("点评数与市场价的相关系数:", x.corr(y))      #计算相关系数并输出
plt.xlabel('点评数')
plt.ylabel('市场价(元)')
plt.scatter(x,y)      #绘制散点图
plt.show()
```

运行程序，得到相关系数 $r=0.57$，点评数—门票价格散点图如图 4.17所示。

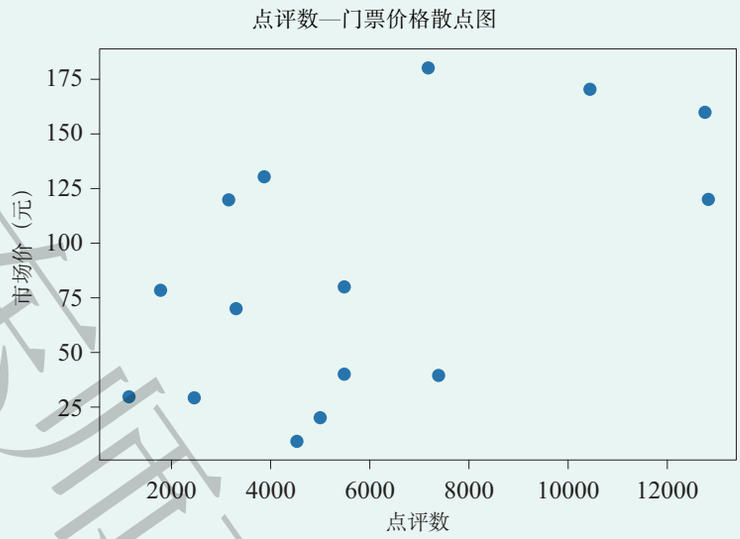


图 4.17　点评数—门票价格散点图

由 $r=0.57$以及图 4.17可以得出点评数与门票价格的关系是正相关。这个结论似乎与我们的直觉不符合，即并不是点评数少的景点的门票价格反而高，而是点评数少（也可以认为是人气低）的景点其门票价格也低，也可以理解为，游客多的景点的门票价格多半比游客少的景点的门票价格贵。在数据不全的情况下，这可以粗略地用来指导我们做旅游攻略。

作业练习

1　对本章第一节中提取区属信息后的数据使用平均分析、分组分析、对比分析、相关分析进行分析，并将分析结果以图表的形式进行展示。

2　在网上书城上查看人气 Top50的书籍，统计这些书的类别，并绘制关于这些图书类别的词云。

知　识　延　伸　　数据描述

数据描述分为数据的集中趋势描述和数据的离散程度描述两种。

数据的集中趋势描述是寻找反映事物特征的数据集合的代表值或中心值，这个代表值或中心值能很好地反映事物的一般水平。例如，平均分析使用的平均数就是数据的集中趋势描述。

数据的集中趋势描述会受到少数几个极端值的影响，不能完全展示数据集合的特征。因此，需要描述数据集合的离散情况，也就是描述数据集合中不同数值之间的差异性，使用离散度（或变异性）来描述。离

散度数值越小，说明数据集合中数值之间的差异越小；反之，离散度数值越大，说明数据集合中数值之间的差异越大。

离散度常用的指标有极差、标准差、方差等。

极差是指数据集合中最大值与最小值的差值，表示数据集合数值之间的距离。例如，对于 10种钢笔的价格数据集合“18.8、22.8、22.8、22.8、26、29、34、38、39、78”，极差为 78－18.8＝59.2。

标准差为数据集合中所有数据与平均数之间的平均距离，计算公式如下：

$$s=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}。$$

其中，s是标准差，X_i是具体的数值，$\overline{X}$是数据集合的平均数，n为数据集合中数据的个数，也称为样本规模。

例如，上述 10种钢笔价格的标准差计算式为：

$$s=\sqrt{\frac{(18.8-33.12)^2+(22.8-33.12)^2+\cdots+(78-33.12)^2}{10-1}}。$$

方差即标准差的平方。

标准差与方差都能描述不同数据分布的数值，方差比标准差夸大了数据集合的离散度。极差仅能描述数据集合中两个极值之间的距离，除这两个极值数据以外的其他数据的数值分布完全不能体现，适合于粗略的离散度的描述。

除了平均分析、分组分析、对比分析、相关分析以外，回归分析（regression analysis）也是一种统计学上常用的分析数据的方法，其目的在于了解两个或多个变量间的相关程度，并建立数学模型，以便通过观察特定变量来预测研究者感兴趣的变量。具体来说，回归分析可以帮助人们了解在只有一个自变量变化时因变量的变化量。

回归分析的目标是建立因变量 Y与自变量 X之间关系的模型。按照涉及的变量的多少，可分为一元回归分析和多元回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。最简单的回归分析只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，称为一元线性回归分析。

第五章

数据挖掘

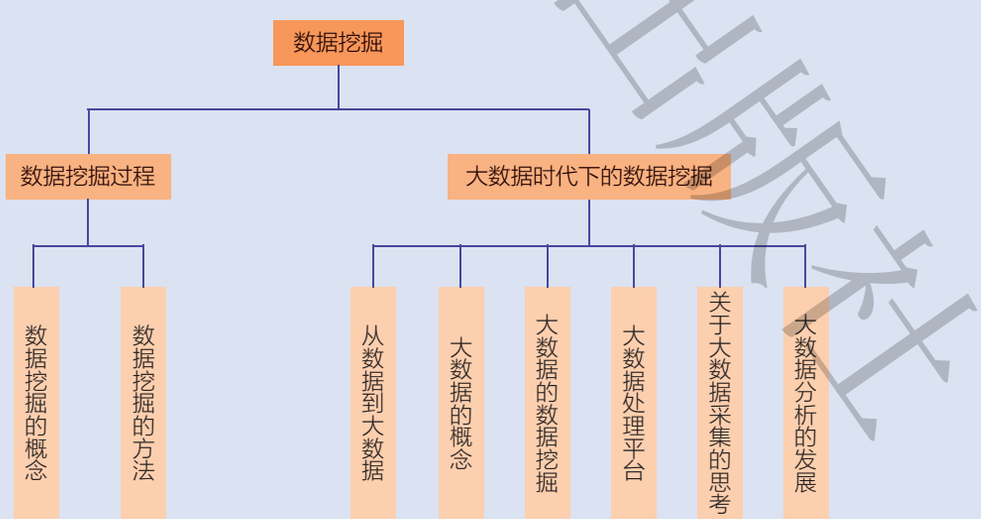
本章学习目标

- 了解什么是数据挖掘，以及数据挖掘对信息社会问题解决和科学决策的重要意义。
- 初步了解数据挖掘的过程和基本算法。
- 了解数据管理与分析技术的新发展，特别是了解什么是大数据，以及在相应环境下的数据挖掘的基本过程。

随着计算机和信息技术的不断普及和应用，各行各业的信息系统规模和数据规模日益增大，这些数据动辄以 PB、EB 等计算。然而，由于数据量太大，从中提取有用的信息已经成为巨大的挑战。尽管从这些数据中常常可以“挖掘”出有用的信息，但传统的数据分析工具和技术显然无法胜任这种“沙里淘金”的工作。因此，在面对大量数据的时候，需要强有力的工具、方法来帮助人们理解这些数据，挖掘出其中隐含的、未知的一些趋势，将数据转变成知识，为人们在实际生活和工作中做决策提供帮助。能满足这一迫切需求的有力工具就是数据挖掘。数据挖掘的目的就是从数据中“淘金”。

现今，数据挖掘已应用到各行各业，在零售业、制造业、金融保险业、通信业以及医疗服务业等领域，其应用案例很多。例如，一家商店根据顾客购物篮的数据集进行数据挖掘，挖掘出了“牛奶和面包两种商品被同时购买的次数非常多”的规则。于是，商店将这两种商品放在相同的区域进行销售，提高了销售收入。又如，一个网站通过搜集多种与机票价格相关的数据来预测机票价格的走势以及升降幅度，为消费者抓住最佳购买时机提供参考。再如，邮箱系统筛选垃圾邮件：首先将邮件内容分成若干词语，然后采用数据挖掘算法计算这些词语组成的邮件属于垃圾邮件的概率，如果概率很大，则该邮件就会被认为是垃圾邮件。

本章知识结构



项目主题

电影数据的数据挖掘

项·目·情·境

观看电影是我们日常文化生活的一个重要组成部分。现今，我国影视行业发展迅速，我国电影市场已成为全球增长最快的市场。随着新影片的不断上映，电影公司、在线订票网站等会积累大量反映电影销售情况的票房数据。而在相关网络论坛与社交媒体上，也源源不断地产生着观影者对影片的评分、评价等数据。

依靠传统的数据分析方法，我们可以回答“今年或本月哪部影片最受欢迎？”“近五年哪类电影最卖座？是历史片，动作片，还是其他？”“今年各月的平均票价是多少”等问题，但是，如果想进一步知道“哪些电影是看过某部电影的人最有可能去看的？”“如何根据社交网站上的电影评分和评论，进行票房预测和电影营销调整？”等诸如此类的问题，则需要利用数据挖掘，从海量电影数据中提取更多有价值的信息。

电影公司为了能推测出观众喜爱的电影，提高票房收入，需要对电影市场进行调查分析。电影公司市场调查员可以从网上搜集电影的相关数据，并通过数据挖掘，从中提取一些有价值的信息，帮助电影公司掌握消费者的喜好、支持其商业决策以赢得市场。

项·目·任·务

任务 1

通过对电影数据的聚类，初步了解数据挖掘的意义。

任务 2

通过在大数据环境中对电影数据的挖掘，了解新发展环境下数据挖掘的基本过程。

第一节　数据挖掘过程

数据挖掘涉及机器学习、统计学、数据库、高性能计算等诸多方面的知识，人们可以应用数据挖掘技术对时序、空间、结构化或半结构化等类型的数据进行分析，从而发现具有一定参考价值的信息。因此，数据挖掘技术已经在电子商务、金融业、保险业等许多领域广泛使用。

问　题　思　考

1. 数据挖掘能做什么？数据挖掘技术会给社会带来什么影响？
2. 数据挖掘的方法有哪些类型？
3. 如何用网络爬虫抓取某个电影网站的数据，对这些电影数据进行聚类分簇，并分析影片的特点？

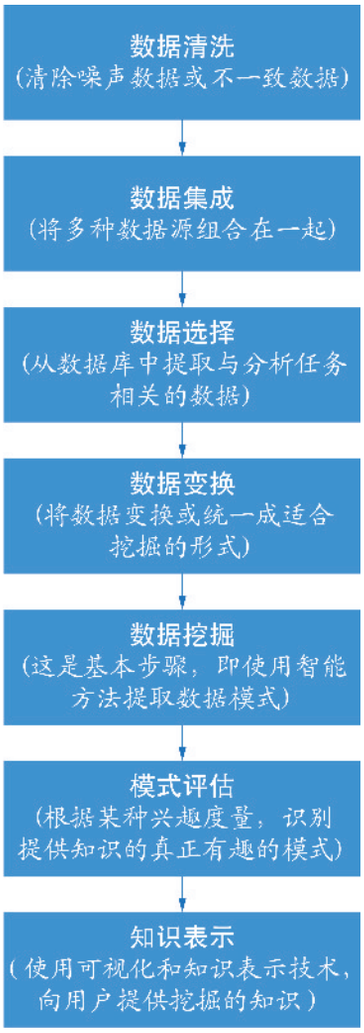


图 5.1　数据挖掘的步骤

一、数据挖掘是什么

数据挖掘是从存放在数据库或从网上获取的数据中挖掘有用信息的过程。数据挖掘是一个跨学科的领域，它是数据分析的一种技术。

具体而言，数据挖掘是将记录下来的海量数据作为挖掘对象，通过一系列的方法、工具或者算法，发现数据中隐含的、未知的信息，以帮助解决现实生活中的问题或者为相关领域的决策提供支持。数据挖掘任务一般可以分为两类，即描述性挖掘任务和预测性挖掘任务。描述性挖掘任务是刻画数据库中数据的一般特性。例如文档聚类，给出新闻文章可以根据它们各自的主题分组。预测性挖掘任务则是在当前数据的基础上进行推断。例如根据花的特征预测花的种类；销售经理希望预测顾客在一次购物期间的消费金额。

数据挖掘的步骤一般如图 5.1 所示。

在对数据进行挖掘之前，经历了好几个准备步骤，这些步骤都是为了使数据挖掘能得到更准确的结果。例如，由于原始数据中存在噪声数据、数据都有各自的数据单位、数据并非都是有用的、有的数据明显是错误的等原因，所以需要清除噪声数据、把数据作标准化处理使之拥有一致的数据单位、把无用的数据暂时剔除等手段，以获得规整的数据集，再根据需求采用相关算法进行

数据挖掘，获得可供科学决策的结论。

任何类型的数据都可以进行数据挖掘，但是需要有明确的目标指引。即数据挖掘的对象可以是结构化数据，如数据库数据、数据仓库、事务数据；也可以是半结构化或者非结构化数据，如时间序列数据、视频监控等数据流、地图、文本图像和视频音频等多媒体数据、社会网络和信息网络等网状数据。对各种各样的数据如何进行挖掘、能挖掘出什么有效的内容等，这些都是专业领域中的问题，有的还待思考和解决。

数据仓库是一个面向主题的、集成的、随着时间变化的、非易失的数据集合，支持管理决策的制定。不同于传统的关系数据库系统，数据仓库需要针对一些明确的主题，关注的是决策者的数据建模和分析，不面向日常操作，不包含对于决策无用的数据。数据仓库不仅可以包含关系数据库，还可以包含跟决策主题相关的文件、事务处理记录等。同时，数据仓库会使用数据清洗等技术，对数据进行一些标准化处理。数据仓库一般提供历史数据，所以数据仓库中的关键结构是包含时间元素的。由于数据仓库并不面向日常操作，所以在存放数据时并不会考虑日常事务处理的便利性。数据仓库只需要承担数据的初始化载入和数据的读取。

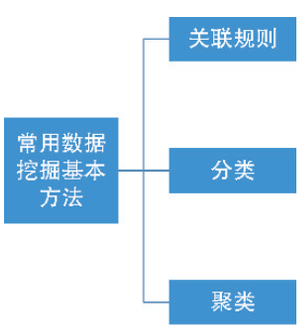


图 5.2　常用数据挖掘基本方法

二、数据挖掘基本方法

数据挖掘利用关联规则、分类、聚类、时序模式等方法，帮助各行各业在生产和研究中挖掘出大量数据中蕴含的价值。常用的数据挖掘方法如图 5.2 所示，有关联规则、分类、聚类。

1. 关联规则

关联规则挖掘是一种简单、实用的分析技术，用于发现存在于大量数据集中的数据的关联性，从而描述一个事物中某些属性同时出现的规律和模式。

关联规则展示了“属性—值”频繁地在给定数据集中一起出现的条件，即描述了一个事物中某些属性同时出现的规律和模式。一个数据集中的项的集合称为项集，包含 k 个项的项集称为 k-项集。关联规则挖掘可以从大量数据中发现项集之间的有趣关联。除了用来分析顾客购物数据中各商品销售之间的关联，为营销决策服务外，关联规则挖掘还常用于生物信息学、医疗诊断、网页挖掘和科学数据分析等。

例如，在地球科学数据分析中，关联规则挖掘可以揭示海洋、陆地和大气之间的有趣联系。关联规则挖掘的经典算法是 Apriori 算法和 FP-Growth 算法。

关联规则挖掘就是希望从数据中找出“买面包的人很可能会买牛奶”这样看起来可能很有意义的模式。如果在 100 位顾客中有 20 位购买了面包，购买面包的 20 位顾客中有 16 位购买了牛奶，那么就可以写成“面包→牛奶[支持度 = 16%，置信度 = 80%]”这样一条关联规则。项集的出现频率是包含项集的事务数，简称为项集的频率、支持度计数或计数。置信度，是指某个关联规则的条件概率。100 位顾客中同时购买面包和牛奶的顾客有 16 人，所以支持度 = 16/100 = 16%。购买面包的 20 位顾客中有 16 位购买了牛奶，所以置信度 = 16/20 = 80%，其表示 80%购买面包的顾客购买了牛奶。挖掘出这样的规则可以有很多用处，例如商店可以考虑把牛奶展柜和面包展柜放到一起以促进销售。实际上，在面对少量数据时关联规则挖掘并不难，可以直接使用统计学中与相关性有关的知识。关联规则挖掘的困难其实完全是由大数据造成的，因为数据量的增加会直接造成挖掘效率的下降，当数据量增加到一定程度，问题的难度就会产生质变。例如，在关联规则挖掘中必须考虑因数据量太大而导致无法承受多次扫描数据库的开销（一般是指系统资源的占用情况，包括中央处理器和存储器等硬件资源）、可能产生在存储和计算上都无法接受的大量中间结果等。关联规则挖掘技术正是围绕着“提高效率”这条主线发展起来的。

2. 分类

分类是找出描述并区分数据类的模型，以便能够使用模型预测一个新的对象的类标号。分类主要用于从大量数据中自动学习，生成分类模型。分类的目的是分析训练数据集，通过这些数据表现出来的特性，为每一个类找到一种准确的描述或者模型，利用这些模型预测新数据所属的类。导出模型可以使用分类规则、决策树、数学公式或神经网络。

分类问题是一个普遍存在的问题，有许多不同的应用。例如，通过电子邮件的标题和内容分拣出垃圾邮件，通过核磁共振的结果图片区分恶性肿瘤和良性肿瘤，银行系统根据贷款用户的年龄、收入等特点预测该用户的贷款申请是安全的还是有风险的。又如，数据挖掘用于 Web 机器人的检测，很多电子商务网站在进行商品的个性化推荐

时需要用到该商品的历史访问记录，但是现实中有一些 Web 机器人会自动访问网站数据。因此在进行商品推荐的时候，必须过滤掉由 Web 机器人产生的访问记录。分类方法建立起分类器模型，从 Web 机器人和正常用户访问网站时的每次会话特性来区分两者的不同，从而使网站的商品推荐更科学更有用。

3. 聚类

与分类不同的是，聚类对已知的数据不提供类标号，也就是不会对已知数据的每个类别进行属性的标记。聚类分析根据在数据对象中发现的描述对象及其关系的信息，将数据对象分组并标记类标号，组内的对象相互之间相似或相关，不同组中的对象是不同或不相关的。组内的相似性越大，组间差别越大，聚类的结果就越好。聚类所形成的分组数据对象的集合称为簇。

聚类可用于商业中，例如：通过收集当前和潜在顾客的大量信息，将顾客划分为若干组，以便进一步分析和开展营销活动；将网络搜索引擎返回的结果分成若干类别，每个类别对应一个特定方面，帮助用户更好使用搜索结果。聚类还可以应用于科学研究中，例如：生物学家用聚类分析大量遗传信息，发现并划分具有类似功能的基因组。

聚类算法有很多，如划分法、层次法、密度算法、图论聚类法、网格算法、模型算法等。比较基础的 k-平均算法属于划分法，其步骤大致如下：首先，随机选择 k 个对象作为簇的中心；然后，对剩余的每个对象，根据其与各个簇中心的相似度，将它们划分为 k 个簇；划分完以后，计算每个簇的平均数作为新的簇中心，再根据相似度重新划分；重复上一步，直到簇中的对象不再发生变化，以使簇尽可能地紧凑和独立。

聚类算法的最终目的是将具有较高相似度的数据点划归到一个簇中。例如，对 $N(N=18)$ 个数据点进行聚类，其 k-平均算法的步骤大致如下：

（1）确定聚类数 K 的值；

（2）从原始数据点中选择 K 个点作为原始簇中心；

（3）计算每个点到 K 个簇中心的距离，并将其划分到最近的那个簇中心所在的簇；

（4）重新计算簇中心；

（5）如果簇中心不再变化，则输出 K 个聚类结果，结束聚类，否则回到步骤 3。

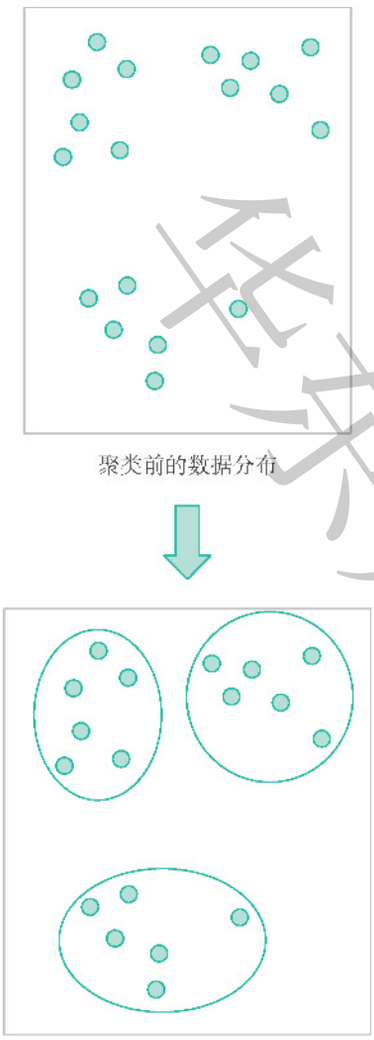


图 5.3　$N = 18$，$K = 3$的聚类数据分布示意图

聚类前后数据的分布示意图如图 5.3 所示。

计算距离的方法有很多，例如欧几里得距离或者其他度量方式。

欧几里得距离是一个最常采用的距离定义，指在 m 维空间中两个点之间的真实距离，或者向量的自然长度（即该点到原点的距离）。在二维和三维空间中，欧几里得距离就是两点之间的实际距离。在二维空间中，点(x_1, y_1)与点(x_2, y_2)的欧几里得距离为$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$。在三维空间中，点$(x_1, y_1, z_1)$与点$(x_2, y_2, z_2)$的欧几里得距离为$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$。

分组之前我们只知道有 N 个数据点，并不知道会得到什么分组结果，每一组的数据会有什么特点事先也无法得知，这种情况下所要进行的分组将根据数据本身的内部特点自动进行，在数据挖掘中称为聚类。

聚类的过程中，首先必须要考虑结果聚成“几”类（即确定 K 的值），这个“几”通常依靠合理的猜测或者一些预定义的要求来确定。聚类是希望获得 K 个分组，每个分组内的数据都具有鲜明的特点，而且跟别的分组内的数据有较高的区分度。

确定 K 值后，需要先确定几个组的“质心”，就是上文讲的“簇中心”，质心是这一组点的均值，然后遍历所有的数据，计算每个数据点和各质心的距离，选择最近距离的质心，并将数据点归入该质心所代表的分组。所有的数据点都分配好分组之后，重新计算一下该组的质心有没有变化，如果有变化，则使用新的质心再去计算和每个数据点之间的距离，重新根据每个数据与新质心的最短距离选择分组。重复上一步骤，直到质心不再发生变化为止，从而得到最终的分组。

探　究　活　动　　电影数据的聚类

看电影是人们比较喜爱的一种娱乐活动，但是市场上的电影的水平参差不齐。无论观影者、投资方还是影院，都希望能有既赚口碑又赚票房的电影，特别是投资方以及影院。投资方在准备投资电影时，要预测影片能否获得市场认可；影院在排片时，要兼顾影片能否受到观影者的认可，从而票房大卖。

我们可以对已有的历史电影数据进行一些简单的数据挖掘，看看能不能发现一些启示信息。首先，通过一些电影论坛或者电影票房网站，采集电影的评分、评论数、票房、类型等数据。然后，根据获得的电影

数据，分析这些电影各有什么特点，看看具有哪些特点的电影是叫座又叫好的。观察数据挖掘结果，看看能否对观影者选择电影、对影院排片给出一些指导意见。

从某电影网站上采集一些电影数据，其中 27部电影的数据如表 5.1所示。在对数据作进一步分析之前需要对数据进行预处理，仅保留数据分析需要的部分，所以表 5.1中用电影编号代替了电影名称。

表 5.1　电影数据

电影编号	评分(分)	评论数(条)	票房(万元)	电影类型
4	6.6	78899	45900	音乐
6	8.5	224792	194200	喜剧
7	8.6	193455	233262	喜剧
12	7.9	383517	221300	喜剧
13	8.2	462345	142300	喜剧
15	6.3	9247	52100	音乐
26	7.1	5565	3040	家庭
30	9.2	602793	17800	爱情
36	7.9	345152	104600	爱情
56	6.5	32628	7545	动画
57	6	1055	5064	动画
62	7.3	2853	142	纪录片
63	8	116523	174900	动作
72	6.1	5577	4207	历史
80	3.9	444	2461	动画
108	6.2	6959	2946	历史
116	4.1	558	1041	动画
136	3	2534	766	悬疑
142	8.6	249610	165200	动作
163	7.8	395310	53000	动作
172	5.7	903	42	动画
184	2.8	120	38	悬疑
193	5.9	2404	137	音乐
204	3.2	266	50	奇幻
205	2.8	863	334	奇幻
218	2.4	646	502	悬疑
231	3.3	5276	1045	悬疑

这些电影中，有的票房很高、有的票房很低，有的评分很高、有的评分很低。能不能对这些电影进行区分，了解哪些电影可以获得高评分和高票房的双丰收呢？

这些电影的特点如果通过人工一个一个地分析并进行归类，会比较麻烦，而且归出的类别会比较主观。而借助数据挖掘中的方法自动对这些电影数据进行分组形成“簇”，可能会获得最佳的分组结果。

表 5.1中的电影数据一共有 27条记录，每条记录中有 5个字段的内容，包括电影的编号、电影的评分、电影的评论数、电影的票房、电影的类型。

虽然可以对多种数据类型的数据进行聚类，包括文本、图形、图像等，但是比较基本的聚类还是针对数字类型的数据。因此表 5.1中可以用来进行聚类的数据为评分、评论数和票房这 3列，如表 5.2所示。

表 5.2 用于聚类的电影数据

评分(分)	评论数(条)	票房(万元)	评分(分)	评论数(条)	票房(万元)
6.6	78899	45900	3.9	444	2461
8.5	224792	194200	6.2	6959	2946
8.6	193455	233262	4.1	558	1041
7.9	383517	221300	3	2534	766
8.2	462345	142300	8.6	249610	165200
6.3	9247	52100	7.8	395310	53000
7.1	5565	3040	5.7	903	42
9.2	602793	17800	2.8	120	38
7.9	345152	104600	5.9	2404	137
6.5	32628	7545	3.2	266	50
6	1055	5064	2.8	863	334
7.3	2853	142	2.4	646	502
8	116523	174900	3.3	5276	1045
6.1	5577	4207			

聚类的过程

1 将表 5.2中的数据标准化，即将大大小小量级相差很大的评分和评论数、票房数据标准化处理到都在 -1～1之间，避免量级过小的数值被边缘化。

2 设置最终的分组的数量为 3，即分成 3个簇。（本活动中设置分组数量为 3，每次实际进行聚类时，可以根据经验设置分组数量）

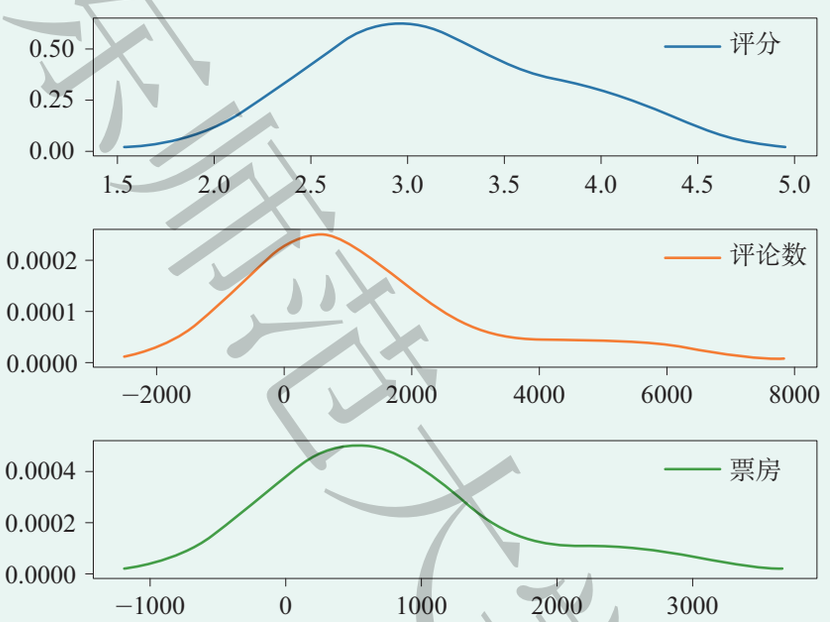
3 使用现成工具实现聚类。本活动中有 3个字段的数据进行聚类，所以是多维聚类。使用 Python实现的代码如下所示：

```
import pandas as pd
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
data = pd.read_csv('film.csv', encoding = 'gbk')
data = data.ix[:,1:]
```

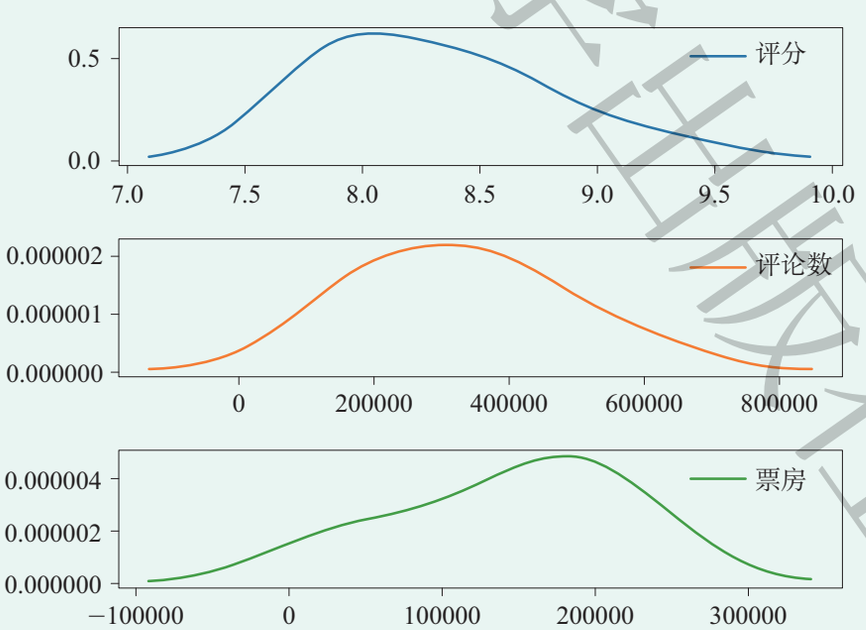
```
#将数据标称化,并输出
data_zs = 1.0 * (data-data.mean())/data.std()
print(data_zs)

k = 3 #分为 3 个类

#使用 K-means 算法实现聚类
model = KMeans(n_clusters = k, max_iter = 100)
kmeans = model.fit(data_zs)

#输出每个样本对应类别,r 中的聚类类别一列用 0、1、2 表示不同类
r = pd.concat([data, pd.Series(model.labels_, index = data.index)], axis = 1)
r.columns = list(data.columns) + [u'聚类类别']

#可视化聚类结果:函数 density_plot 生成密度图
def density_plot(data):
    plt.rcParams['font.sans-serif'] = ['SimHei']
    plt.rcParams['axes.unicode_minus'] = False
    data.plot(kind = 'kde', linewidth = 2, subplots = True, sharex = False)
    plt.legend()
    return plt

#分别产生 3 个类的概率密度图,分别存为 51.png、52.png、53.png
for i in range(k):
    density_plot(data[r[u'聚类类别'] == i]).savefig(u'5%s.png' %i)
    plt.show()

#3 列数据即 3 维数据进行聚类,而散点图的可视化只能是 2 维的,所以需要降维
from sklearn.decomposition import PCA
pca = PCA(n_components = 2)
new_pca = pd.DataFrame(pca.fit_transform(data_zs))

#降成 2 维后的 new_pca 数据生成散点图,存为 scat.png
d = new_pca[r['聚类类别'] == 0]
plt.plot(d[0], d[1], 'ro')
d = new_pca[r['聚类类别'] == 1]
plt.plot(d[0], d[1], 'g.')
d = new_pca[r['聚类类别'] == 2]
```

```
plt.plot(d[0], d[1], 'b*')
plt.savefig('scat.png')
plt.show()
```

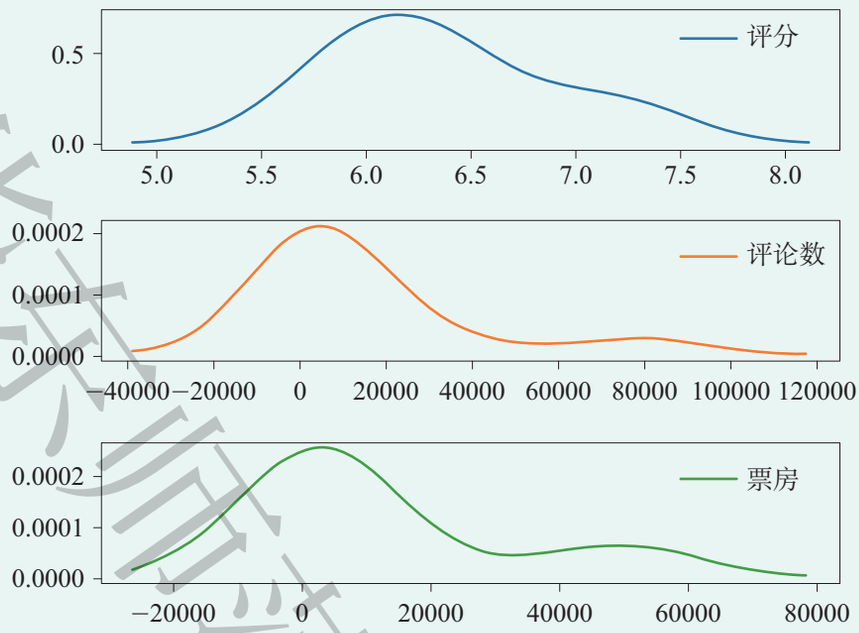
4 输出 3个簇的概率密度函数图(如图 5.4所示),可用于分析每个簇的具体特点。连续型随机变量的概率密度函数是一个描述这个随机变量的输出值在某个确定的取值点附近的可能性的函数。而随机变量的取值落在某个区域之内的概率则为概率密度函数在这个区域上的积分。



(a)以上 3 个图为聚类后第 1 个簇的评分、评论数和票房数据的概率密度函数图



(b)以上 3 个图为聚类后第 2 个簇的评分、评论数和票房数据的概率密度函数图



(c)以上 3 个图为聚类后第 3 个簇的评分、评论数和票房数据的概率密度函数图

图 5.4　概率密度函数图

5 对多维进行降维处理，生成散点图(如图 5.5所示)，用于直观观察聚类的效果。PCA就是主成分分析，也称主分量分析。通过进行主成分分析可以起到降低维度的作用，把多指标合成为少数几个相互无关的综合指标(即主成分)，其中每个主成分都能够反映原始变量的绝大部分信息，而且所含信息互不重复。这种方法在引进多方面变量的同时将复杂因素归结为几个主成分，使问题简单化，同时得到更加科学有效的信息。

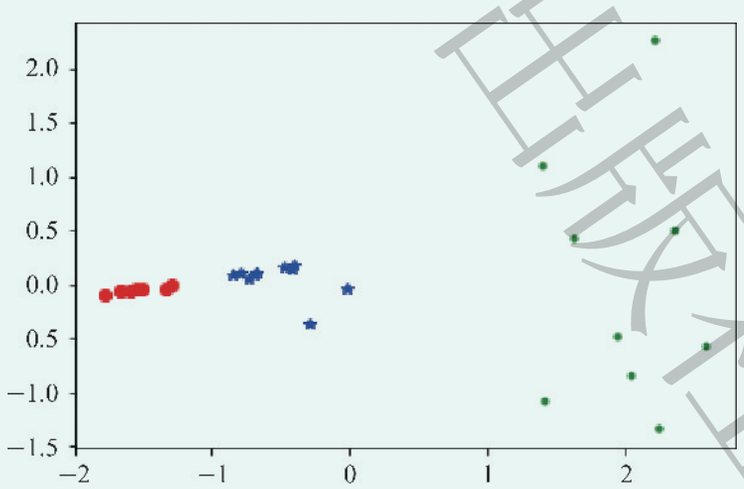


图 5.5　程序运行之后自动生成的分成 3个簇的数据点分组情况

聚类结果分析

表 5.3 聚类结果分析表

评分(分)	评论数(条)	票房(万元)	聚类类别	电影类型
3	2534	766	0	悬疑
3.2	266	50	0	奇幻
2.8	120	38	0	悬疑
2.4	646	502	0	悬疑
3.3	5276	1045	0	悬疑
2.8	863	334	0	奇幻
3.9	444	2461	0	动画
4.1	558	1041	0	动画
8.6	193455	233262	1	喜剧
9.2	602793	17800	1	爱情
8.2	462345	142300	1	喜剧
8.5	224792	194200	1	喜剧
8	116523	174900	1	动作
8.6	249610	165200	1	动作
7.9	383517	221300	1	喜剧
7.8	395310	53000	1	动作
7.9	345152	104600	1	爱情
6.5	32628	7545	2	动画
5.9	2404	137	2	音乐
6.2	6959	2946	2	历史
6.1	5577	4207	2	历史
7.1	5565	3040	2	家庭
6	1055	5064	2	动画
5.7	903	42	2	动画
7.3	2853	142	2	纪录片
6.3	9247	52100	2	音乐
6.6	78899	45900	2	音乐

聚类结果如表 5.3所示：

1 一共聚成 3类，通过字段“聚类类别”的数字 0、1、2进行区分。

2 第 0类，评分低、评论数少、票房低。

3 第 1类，评分高、评论数多、票房高，这类电影无论是对观众还是对制作方而言，都是极受欢迎的，所以影院在排片的时候会尽量多地选择这类电影。

4 第 2类，评分、评论数、票房数据都比较居中。

聚类的结果一般相对比较简单，将原始数据进行自动分类之后，通常会把结果用于进一步的分析和处理。例如，完成聚类后，每个类就可以标出类标号，然后用于预测后来的数据属于哪一个类，这属于数据挖掘的另一个方法——分类。

聚类的过程并非是一蹴而就的，很有可能需要经过多次尝试，例如对于分组的数量、迭代的次数、数据组合的选择等都可能需要进行多次尝试，然后选择最符合预期的、最合理的聚类结果作为最终结果。

体 验 思 考　数据挖掘算法的研究

开展数字化学习，了解数据挖掘的其他算法（至少三个），并举出每个算法的应用实例。

知 识 延 伸　数据挖掘的发展和应用

数据挖掘的概念在 20世纪 80年代就产生了。在我国，国家自然科学基金从 20世纪 90年代开始支持高校、研究院对数据挖掘领域的研究。数据挖掘是一种数据分析技术，一般是指数据库知识发现中的一个步骤，即从大量的数据中通过算法搜索隐藏于其中的信息的过程。

在健康医疗领域，数据挖掘不仅能够辅助完成预防、诊断、治疗、预后等医疗任务，并通过辅助门诊和急诊、住院、医疗费用、医疗保险等的管理为医疗资源的合理配置提供参考，还能帮助理解健康信息需求和行为，优化健康信息获取，进而改善健康信息服务。从预防角度而言，数据挖掘可以有效控制影响疾病发生的危险因素以及疾病的早期筛查；在诊断方面，对病历文本信息和图像数据进行数据挖掘可辅助疾病诊断；在中医药物治疗方面，通过数据挖掘可以发现年龄、性别、症状之间的关系，或者中药方剂中药物组合的规则；在药物不良反应方面，可以通过数据挖掘识别并发出预警。

在教育领域，互联网和信息技术的快速发展使得对学习数据的收集变得更加高效。可以通过智能教学系统来监测学生对于特定知识从无到有的学习过程数据，也可以通过在线学习平台来收集学生对于特定知识的反馈数据，更可以轻松地从学生管理系统中获取学生的个人信息、课程信息和考试信息等数据。使用数据挖掘技术可以发现某一课程中具有相似学习特征的学生并分组；发现具有异常行为或特性的学生，并提前进行干预，根据学习者容易误解的知识点之间的关联，调整教学策略；发现学习者在学习平台上的使用行为和使用过哪些资源与他们的测评成绩之间的关系，来指导学习者提高在学习平台上的学习效果等。

第二节　大数据时代下的数据管理与分析技术的发展

大数据时代下，数据的规模、数据管理涉及的技术都发生了变化。

一、从数据到大数据

数据产生于人们生活、工作、学习的方方面面，人们在处理数据的过程中获得信息。随着计算机与信息技术的迅猛发展和普及应用，各行各业产生的数据都呈爆炸性增长。原来在一台微型计算机上就能存储并处理数据，而现在数据动辄就达到 TB 甚至 PB 数量级。在大数据环境下，无论是传统的存储方式还是传统的算法都成了数据处理中的瓶颈，因此迫切需要新技术来解决问题。与传统数据相比，大数据在数据规模、采集方式、分析方法、价值利用等方面都有了很大的发展。例如，对传统数据进行处理时，通常只考虑集中式存储数据，更多考虑的是抽样数据的分析处理；而大数据环境下需要考虑数据的分布式存储，同时需要考虑的是全面且完整数据的处理分析。今天，数据管理与分析的新技术蓬勃发展，数据处理的方式和效率都发生了变化，大数据分析在帮助人们更好地做出预测和决策方面发挥着越来越重要的作用。

问　题　思　考

1. 大数据对社会产生了什么样的影响？
2. 你了解到的数据管理与分析技术的新发展是怎样的？

二、大数据的概念

人们把大规模数据称为“大数据”（big data），这个概念早在 2008 年就已被提出。2008 年，某权威杂志出版了一期专刊，专门讨论与未来的数据处理相关的一系列技术问题和挑战，其中就提到“Big Data”的概念。大数据不仅指数据规模的大，更是意味着大数据处理需要新技术和新方法。大数据带来的巨大变革，改变着人们的生活、工作和思维方式，影响着经济、政治、科技和社会各个层面。而且，由于数据规模巨大，传统数据规模下的算法很可能无法有效或者高效实现，算

法需要重新设计。

大数据一般可以通过以下四个特征来理解，总结为 4V：

■ 数据规模大（volume）：数量大，存储单位从过去的 GB 到 TB，直至 PB、EB。

■ 数据类型多（variety）：数据种类和来源多样化，多类型的数据对数据的处理能力提出了更高的要求。

■ 处理速度快（velocity）：大数据产生、处理、计算速度较快，能满足实时数据分析需求。

■ 价值密度低（value）：海量数据中有价值的数据所占比例很小，价值密度较低，数据需要经过清洗和挖掘才能发挥效用。

三、大数据的数据挖掘

大数据可以用于开发一个不断分析运动赛事的数据流、基于赛事的激烈程度对比赛排名评分的软件，让用户找到值得收看的比赛。大数据也可以用于将交通的实时数据发送给司机，并且提前告诉司机有空闲停车位的地方。大数据还可以用于通过获得大量病人的临床信息，分析病人的信息，尽早知道可能产生的病症，从而提前采取预防措施，减少疾病的发病率。大数据的数据挖掘，首先要解决的问题是怎么存储这些海量数据。

巨大规模的数据超出了传统的信息系统的处理能力，巨大的数据量导致巨大的计算时间开销，使得原来在小规模数据量时使用的存储方式以及算法实现都遇到瓶颈，必须寻求能适应海量数据的数据处理技术。

随着计算问题规模和数据量的不断增大，一个大问题通常需要被分解成较小的若干个问题。大数据环境下要使用并行计算技术来解决原先单机上的小规模计算问题，这牵涉到并行处理器、不同存储结构下数据的共享、分布式的文件存储和访问、如何将大规模的计算任务分配到并行系统中的各个节点去完成、具体能实现并行计算任务的程序语言等很多方面的技术。

探 究 活 动　大数据下的电影数据的数据挖掘

在探究活动“电影数据的聚类”中，我们完成了对 27部电影的数据的聚类分析。然而，现实世界中的电影数量远远不止 27部，如果有 20万部电影的数据需要进行分析，那么分析过程中遇到的问题会有所不同。

假设聚类数据点远超 27个，有 N 个（$N \geqslant 200000$），需要聚成 K 类。数据不再单独存放在一个数据表中，而是分布在不同的机器上，也就是所谓的分布式存储中。而且，同样的数据通常都会有至少 3份的冗

余备份。

那么，聚类的过程如下：

1 先扫描原始数据集中的所有点，随机选取 K个簇中心；

2 各个分片（大数据集划分成多个数据分片，每个分片对应一个节点）的节点读取存在本地的数据集，用聚类算法生成 K个聚类；

3 把各个节点的结果做归并，重新计算簇中心，如果有变化，则重复步骤 2，直至簇中心不再变化；

4 K个簇中心确定之后，对所有节点上的所有数据点再重新计算一遍其与 K个簇中心的距离，将其划分到最近距离的那个簇中去。

最终，可以将几十万部电影的数据进行聚类，自动将这些电影分成有内在关联的 K个类别。在整个过程中，可能会多次修改 K值，以得到更合理的聚类结果。完成聚类后，还需要有经验的人对聚类结果进行分析，使结果能对观影者和影院决策都有指导意义。

说明：本活动需要在大数据实验平台环境中实现，无法在单机环境中实现。作为教学使用的大数据实验平台，可以在宿主机中运用虚拟化方法构建包含一个主节点和两个从节点的迷你集群。

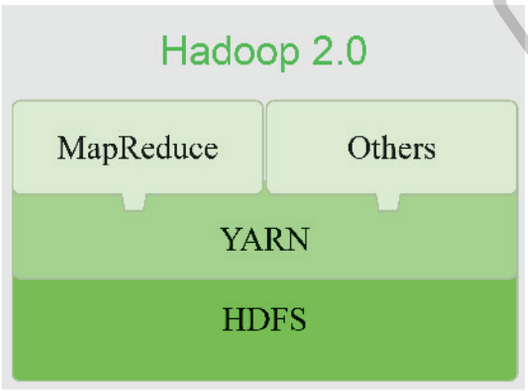


图 5.6 Hadoop 2.0的构成

四、大数据处理平台

2004 年开始，Apache 的开源项目 Hadoop 成为主流的大数据处理平台，Hadoop 2.0 的构成如图 5.6 所示。大数据处理首先面临的是如何解决大数据的存储管理问题，Hadoop 平台上解决该问题的技术是 HDFS——分布式文件系统；其次，大数据处理需要解决如何有效快速完成大规模数据的计算，Hadoop 平台上的 MapReduce 是解决该问题的并行计算框架。

HDFS 是一种分布式的文件存储系统，MapReduce 是一种分布式计算框架。HDFS 文件系统包括若干台机器，每个文件被分成若干块，并且每块都有 3 个节点进行存放，也就是说副本个数为 3。至于选择哪个节点存放数据的副本，也有专门的协议来协调。

分布式计算是一门计算机科学，它研究如何把一个需要非常巨大的计算能力才能解决的问题分成许多小的部分，然后把这些部分分配给许多计算机进行处理，最后把这些计算结果综合起来得到最终的结果。分布式计算通常用于利用互联网上的计算机的中央处理器的闲置处理能力来解决大型计算问题。

MapReduce 采用分而治之的算法来完成大数据处理。其先将大

数据处理的任务分成一个个小的任务，再将处理后的结果归并到一起进行汇总，获得最终结果。

随着数据信息量爆炸性的增长，大规模数据处理任务很难由一般的单机系统在可接受的时间内完成，原有的聚类算法处理海量数据时性能较差，甚至无法完成。在大数据环境下，原始的大数据集将被划分为很多个数据分片，每个分片由一个节点处理。

五、关于大数据采集的思考

数据挖掘可以挖掘出很多有用或者有意思的结果，但大多数时候需要大量的数据作为基础。这些数据的来源往往五花八门，包括社交媒体、传感器、物联网、视频监控等，而且都是通过使用者在使用过程中产生并被采集的，人人都是数据的生产者。数据采集的过程中通常可以获得原始数据，而这些原始数据往往包含或隐含着数据生产者的真实信息。

例如在大数据环境中，可以通过采集使用者的网址搜索记录、手机上网记录、网络购物记录等数据来获取他们的信息，如兴趣爱好、日常生活等。但是，这些数据的采集其实都是在被采集者不知道的情况下进行的，被采集者不清楚自己的这些信息将被用于做什么，亦或是谁用了这些信息，也不清楚这些信息泄露以后由谁来负责。

采集到的数据可能会用在医疗、教育、社会治理等各个方面，数据本身无罪，问题在于数据被使用时是否侵犯隐私，可否找到合适的方式让数据被应用时能有效保护个人隐私。对于用户来说，如果个人信息被泄露，并被用于商业或其他目的，这就是一种隐私泄露的数据安全问题。

对于个人用户的数据安全，现在已有了一些技术保护手段，但还需要国家立法、企业自律和个人防护三位一体来作保护，让人们能正当、合法、必要地使用数据。

我们应从自己做起，维护大数据环境下的数据安全，在学习知识的同时立德，做到明大德、守公德、严私德，从根本上正确使用数据，为人类创造价值。

六、大数据分析的发展

大数据领域每年都会涌现出大量新的技术，为大数据获取、存储、处理分析或可视化提供了有效手段。大数据分析中几个比较核心的技

术包括：大数据的生命周期、大数据技术生态、大数据采集与预处理、大数据存储与管理、大数据计算模式与系统、大数据分析与可视化。

大数据的基本处理流程与传统数据处理流程并无太大差异，主要区别在于：由于大数据要处理大量非结构化的数据，所以在各处理环节中都可以采用并行处理，大数据环境下的数据分析大多在云计算、云存储环境下实现。大数据分析最终通过可视化的交互式视觉表现方式帮助人们探索和理解复杂的数据，成为用户了解复杂数据、开展深入分析不可或缺的手段。

在整个大数据生命周期中，大数据的编程和管理工具的发展以及数据安全问题贯穿始终。目前，基于大规模数据的实时交互可视化分析以及在这个过程中引入自动化的因素是研究的新方向。另外，大数据分析领域中的前沿性研究课题还包括根因分析等。人们一般通过建立数学模型的方式也就是从原因到结果获得知识，而从结果到原因逆向获取知识的过程就是根因分析，可以用于寻找获得结果的基本因素。大数据技术能够将大规模数据中隐藏的信息和知识挖掘出来，为人类社会经济活动的发展提供帮助，提高各个领域的运行效率，提高整个社会经济的集约化程度。

作业练习

请探讨大数据环境下的数据挖掘算法如何实现需要解决的问题。

知 识 延 伸

大数据下的信息安全思考

目前大数据的发展仍然面临着许多问题，其中安全与隐私问题是人们公认的关键问题。当前，人们在互联网上的一言一行都掌握在互联网商家手中，包括购物习惯、好友联络情况、阅读习惯、检索习惯等。多项实际案例说明，即使无害的数据被大量收集后，也会暴露个人隐私。事实上，大数据安全的含义更为广泛，人们面临的威胁并不仅限于个人隐私泄露。大数据在存储、处理、传输等过程中面临诸多安全风险。在大数据背景下，有的商家既是数据的生产者，又是数据的存储者、管理者和使用者，因此，单纯通过技术手段限制商家对用户信息的使用，以实现用户隐私保护是极其困难的事。

我国不仅确立了大数据的战略发展地位，同时对大数据时代的信息安全管理也进行了新的定义。在形成数据开放、安全共享的国家信息安全治理新政策的过程中，应着力提高公民的信息安全素养，重点是提升网络安全素养；在网络强国战略、大数据战略和“互联网 +”行动纲要的实施过程中，应推动把公民信息安全科学素质指标纳入其中，并同时纳入各行业机构文明创建指标考评体系之中，形成公民信息安全素养建设的共建机制，定期开展公民信息素养的评估，形成数据开放、安全共享的民众基础。

后　记

本册教科书依据教育部《普通高中信息技术课程标准（2017 年版 2020 年修订）》编写，并经国家教材委员会专家委员会审核通过。全体编写人员认真领会国家基础教育改革精神，精心研究当代信息社会的人才培养要求，广泛调研上海及各地高中信息技术教育的现状和挑战，深入了解高中学生的学习需求，并汲取了上海市《普通高中信息科技（试用本）》的编写经验。

编写过程中，上海市中小学（幼儿园）课程改革委员会专家工作委员会，上海市教育委员会教学研究室，上海市课程方案教育教学研究基地、上海市心理教育教学研究基地、上海市基础教育教材建设研究基地、上海市信息科技教育教学研究基地（上海高校“立德树人”人文社会科学重点研究基地）及基地所在单位华东师范大学等单位给予了大力支持，郑骏、金莹、陶烨、魏雄鹰等老师作出了重要贡献。在此表示感谢！

本册教科书出版之前，我们已通过多种渠道与教科书选用作品（包括照片、画作）的作者进行了联系，得到了他们的大力支持。对此，我们衷心地表示感谢！恳请尚未联系到的作者与我们联系，以便出版社及时支付相关稿酬。

我们真诚地希望广大教师、学生及家长在使用本册教科书的过程中提出宝贵意见。我们将集思广益，不断修订，使教科书趋于完善。

编　者