普通高中教科书

信息技术

选择性必修 3

数据管理与分析

上海科技教育出版社

普通高中教科书

信息技术

选择性必修 3

数据管理与分析

上海科技教育出版社

编写人员名单

主　　编：郑　骏

分册主编：金　莹

分册副主编：钱卫宁

主要编写人员（以姓氏笔画为序）：

毛嘉莉　张　召　罗轶凤

金澈清　周　烜　高　明

陶　烨　黄定江　董启文

蔡福民

欢迎广大师生来电来函指出教材的差错和不足，提出宝贵意见。

上海科技教育出版社地址：上海市闵行区号景路 159 弄 A 座 8 楼

邮政编码：201101

联系电话：021-64702058

邮件地址：office@sste.com

写给学生的话

亲爱的同学：

如今信息技术快速发展，各种各样的数据不断充斥、影响着我们的生活。对交通数据进行分析，可以为制定交通方案提供科学依据；多维度地了解用户购买需求，则为网上商店进行精准营销提供了可能……大数据时代，人们正以从前无法想象的方式从海量数据中挖掘有价值的信息，作为合理决策的有力武器。

在《数据管理与分析》的学习中，我们将带领你通过具体的生活事例，了解各种数据采集途径，掌握设计简单关系数据库的方法，利用适当的数据分析方法从给定的数据中提取出有用信息，根据需求形成最终解决方案，从而感受数据管理与分析的重要性，以及数据安全的重要性。

为了让你在学习《数据管理与分析》的过程中获得更大的成功，请浏览本书的栏目介绍。

单元引言、学习目标和单元挑战

从生活经验出发引入本单元将要学习的内容，提出本单元学习要达成的学习目标，预告学习完本单元后要接受的单元挑战。

项目引言和学习目标

描述项目产生的背景和意义，介绍项目学习的主要内容，并提出一些具体问题，引导你带着问题探究。

项目学习指引

通过剖析真实的项目实施过程，帮助你了解学科思想方法，理解相关概念，掌握具体技能。

核心概念和小贴士

解释一些重要概念和术语，或提示相关知识和技术，帮助你抓住重点，扫除认知障碍。

思考与讨论??

提出若干问题引导你对技术背后的原理以及人、信息技术与社会的关系等进行思考和讨论。

数字化学习

引导你利用网络、数字化工具和数字资源进行学习。

活　动

提出活动任务，并引导你运用所学知识，使用信息技术工具进行探究、总结和展示。

知识链接

系统整理和归纳本项目的知识要点，方便你学习。

拓展阅读

补充更丰富的阅读材料，开阔你的视野。

单元挑战

布置面向真实情境的项目任务，希望你综合运用本单元所学的知识与技能去解决问题。

单元小结

用思维导图可视化呈现本单元的知识脉络，提供基于学科核心素养的评价表，为你的学习表现进行自我评价。

在学习过程中，希望你勤实践体验、多思考讨论，借助各种数字化工具、资源进行学习与创新，不仅要理解和掌握具体的信息技术知识与技能，还要把握用信息技术解决问题的思想方法，并思考将信息技术应用于社会时所引发的各种挑战，以开放、包容的心态与信息技术、信息社会一起进步。

编　　者

目 录

第一单元　初识数据管理与分析 1

项目一　探究交通数据的管理与分析——认识数据资源与价值 2

1. 采集路口交通数据 3

2. 管理交通数据 5

3. 分析交通数据 6

4. 了解交通数据资源及其价值 8

知识链接 9

项目二　了解网络购物数据的管理与分析——经历数据管理与分析的流程 13

1. 分析业务需求 14

2. 管理网上商店订单数据 14

3. 分析订单数据 17

4. 完成科学决策 18

5. 评价、优化整体方案 18

知识链接 19

单元挑战　调查校园数据管理现状 21

单元小结 22

第二单元　数据管理 23

项目三　了解健身数据的采集与分类——认识数据的结构化 24

1. 采集会员健身数据 25

2. 分类存储会员健身数据 28

3. 认识噪声数据 29

知识链接 30

项目四　建立简易网上书店数据库——了解关系数据库的建立 33

1. 分析数据库设计需求 34

2. 建立实体集和联系 35

3. 建立数据模型 37

4. 创建数据库 39

知识链接 40

项目五　管理网上书店数据库——使用结构化查询语言 43

1. 添加数据 44

2. 查询数据 45

3. 更新数据 46

4. 删除数据 47
知识链接 47
单元挑战　建立年级作业评价数据库 50
单元小结 51

第三单元　数据分析 53
项目六　分析城市交通拥堵状况——了解常用的数据分析方法 54
1. 了解城市道路交通拥堵状况 55
2. 分析造成城市道路交通拥堵的相关因素 59
知识链接 62
项目七　揭示网上书店图书销售情况——分析、呈现并解释数据 65
1. 分析并呈现网上书店图书销售情况 66
2. 发现用户数据的相关性 72
知识链接 75
项目八　探索网上书店图书推荐——认识数据挖掘的重要意义 78
1. 了解数据管理与分析技术的新发展 79
2. 挖掘用户阅读兴趣 80
3. 用协同过滤推荐方法推荐图书 82
知识链接 86
单元挑战　分析在线社交平台用户情况 90
单元小结 91

第四单元　数据备份与数据安全 93
项目九　探秘网上书店数据库系统容灾方案——应对数据丢失风险 94
1. 了解数据丢失风险 95
2. 备份网上书店数据 96
3. 优化数据丢失防范方案 99
知识链接 100
单元挑战　探索 MySQL 数据库的实时备份 103
单元小结 104

附录　部分名词术语中英文对照 106

第一单元

初识数据管理与分析

信息技术与经济社会的交汇融合引发了数据量和数据处理速度的迅猛增长。数量巨大、来源分散、格式多样的数据就像一个个宝藏，被不同的组织或者个人获取、管理、分析和使用着，最终实现其价值。

利用技术工具有效管理和分析数据，提取和发现有价值的信息，已经成为人们解决问题的一种重要方式。商家多渠道采集消费者的购物数据，分析其消费习惯和规律，为营销决策提供支持；企业利用产品设计、制造、营销、售后等各环节的数据，为新产品研发和企业创新发展提供支持；医院充分挖掘临床医疗数据中的价值，用于远程诊疗、医疗研发等；社会保障部门建设公共服务数据平台，为公众提供个性化和精准化的服务……数据管理与分析在生产与生活中占据着越来越重要的地位，可以帮助人们更好地应对未来的挑战。

在本单元中，我们将结合生活实际，认识数据管理与分析的价值和意义，并初步了解数据管理与分析的一般流程。

单元挑战

调查校园数据管理现状

学习目标

- 认识到数据是一种重要的资源。
- 感受数据管理与分析技术的重要性。
- 初步了解分析业务需求、建立数据管理与分析问题整体解决方案的基本过程。
- 尝试对既定方案进行分析、评价，发现问题并优化方案。

项目一

探究交通数据的管理与分析
——认识数据资源与价值

为了解决交通管理问题，不少城市的交通管理部门都在主干道、路口安装了视频监控、地感线圈等设备，实时采集交通数据，如图 1-1 所示。此外，随着移动网络、全球定位系统等技术的发展，还产生了大量通过手机、车载设备甚至遥感卫星等采集的交通数据。

这些海量的、形式多样且来源丰富的交通数据，可以帮助交通管理部门了解实时路况，及时处理交通事故；可以为公交公司、出行者以及相关企业提供信息服务；还可为政府各部门规划道路建设、开发公交线路等提供决策支持。交通数据作为一种重要资源正被不同组织共享利用，并发挥着价值。

图 1-1　采集交通数据的视频监控

项目学习目标

在本项目中，我们将以解决某十字路口的拥堵问题为例，了解数据的价值，认识到数据是一种重要资源，感受数据管理与分析技术在其中扮演的重要角色。

完成本项目学习，须回答以下问题：

1. 数据管理与分析及数据价值之间有怎样的关系？
2. 数据为什么是一种重要的资源？
3. 数据的价值体现在哪些方面？

项目学习指引

1. 采集路口交通数据

生活中经常会出现如下的问题：某十字路口在特定时段异常拥堵，虽然有交警帮助疏导车辆，但是效果不佳。

要解决以上问题，首先要知道路口拥堵的具体状况，以便分析出原因，这就需要获取该路口的交通数据。路口的交通数据很多，其中，车流量数据能够反映在一天中某时段通过的车辆数量、车辆流向及车辆分类情况，是反映路口车辆行驶情况的重要数据。因此，从采集路口的车流量数据入手，可能会发现有价值的信息。

交通管理部门一般会在路口布设自动化的交通数据采集设备，如地感线圈、视频监控等，进行全天候的实时数据采集，如图 1–2、图 1–3 所示。

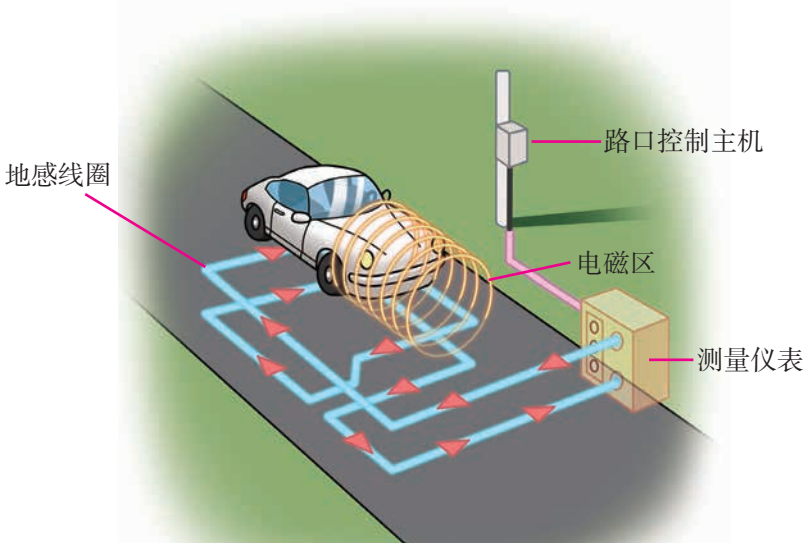


图 1–2　利用地感线圈采集数据

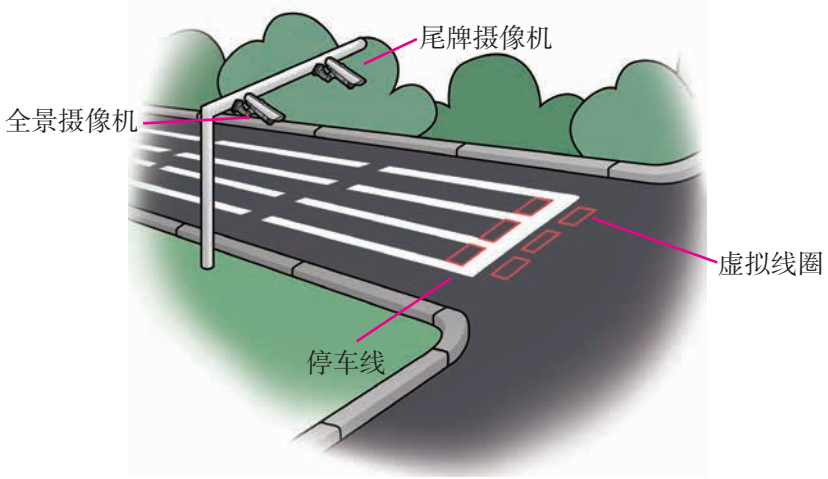


图 1–3　利用视频监控采集数据

小贴士

除了利用设备自动采集车流量数据以外，还可以利用人工的方式采集。人工采集是由人通过手工揿按计数器来统计某个时间段内经过的车辆数，从而得出车流量数据。虽然利用设备自动采集交通数据方便省力，但是对于特定路段，在缺少采集设备的情况下，人工采集仍然不失为一种有效的数据采集方式。

例如，地感线圈依靠埋在路面下的一个或一组感应线圈产生的电磁感应变化，来检测通过车辆的状况，包括车辆数量、车辆速度等。又如，视频监控采用虚拟线圈的方式触发摄像机，对经过道路卡口的每辆车进行抓拍，并对所拍摄的图像进行分析，从而自动获取车辆的通过时间、车牌号码、车型、号牌颜色、车身颜色等数据。

然而，实际情况非常复杂，会影响交通数据的采集，可能产生错误的、异常的或不完整的噪声数据。例如，大雾天气会导致视频监控设备无法获取清晰的图像数据；在夜间或光照较差的情况下，可能获取错误的车牌数据；地感线圈故障也会导致相关车辆数据丢失；等等。因此，一般在做数据分析之前，需要对数据进行预处理，从而保证数据分析结果的可靠性。

这些设备采集到的数据会传输到专门的数据库系统中进行存储，供交通管理部门分析和使用。为了使数据蕴含的价值被深入挖掘、充分利用，有些城市的交通数据经过数据脱敏后会开放给科研机构或企业，甚至免费向社会公众开放。

小贴士

数据脱敏的目的是在数据交换、共享、使用等过程中实现对敏感数据的定向、准确和彻底的变换，使数据安全、可信、受控使用。要达到上述目的，需要依据相应的脱敏原则，针对敏感级别制定脱敏策略。

思考与讨论

1. 除了以上方式，你还知道哪些交通数据采集的设备和途径？

2. 某商店店主为了防盗，在店铺里安装了视频监控，并定期将偷窃视频公布在网上。某饭店为了提高知名度，在进餐区安装了视频监控并在直播平台进行直播。你是否赞同这两种行为？为什么？

活　动

1.1 走访学校或者家附近的路口，观察有无交通数据采集设备，再通过上网学习，了解这些设备可以采集哪些数据，交通管理部门利用这些数据可以解决哪些问题。

2. 管理交通数据

无论是利用设备自动采集到的交通数据，还是用人工的方式采集到的交通数据，都需要进行存储与管理，以方便后续的数据分析。人工采集到的交通数据可以通过录入的方式存储到相应的数据库中，而利用设备采集到的交通数据会被自动存储到数据库中。数据存储后，还需要对数据进行查询、添加、删除、更改等操作。

参见 P10 知识链接“数据管理与分析技术”

为使采集到的数据保持连贯性、持续性和有效性，以便在数据库系统之间实现共享，还需要对数据进行标准化处理。比如，对数据的名称、代码、分类编码、数据类型、精度、单位、格式等，要规定其标准形式。

例如，路口的视频监控系统所采集的数据，自动存储到交通管理部门的数据库后，经过处理，可得到某年 5 月 8 点到 9 点时间段内某路口平均车流量数据，如表 1-1 所示。表中用统一代码 NS、SN、WE、EW 分别代表由北向南、由南向北、由西向东、由东向西四个车辆行驶方向。

表 1-1　某路口平均车流量数据表

月份	起始时间	终止时间	方向	直行车辆数	左转车辆数	右转车辆数
5 月	8:00	9:00	NS	925	607	327
5 月	8:00	9:00	SN	1248	127	466
5 月	8:00	9:00	WE	660	223	151
5 月	8:00	9:00	EW	548	316	796

思考与讨论

在管理和分析交通数据时，需要规避或转换哪些数据，避免车主隐私信息的泄露？

活　动

1.2　公交一卡通能够准确地反映乘坐公交车出行者的位置分布情况，其采集的公交车客流量数据是公交客流预测、公交线路优化、公交合理调度等应用的重要数据基础。尝试选择恰当的工具将公交一卡通数据表（表 1-2）存储到计算机中，注意按需要设置数据类型、精度等。

表 1-2　公交一卡通数据表

卡号	交易日期	交易时间	公交 / 地铁站点	行业名称	交易金额	交易性质
6021411280	2015-04-01	07:51:08	703 路闵行医院	公交	2.00	非优惠
6021411280	2015-04-01	09:07:57	11 号线昌吉东路	地铁	6.00	优惠
2201252167	2015-04-01	19:20:33	7 号线场中路	地铁	4.00	非优惠
2201252167	2015-04-01	08:55:44	1 号线陕西南路	地铁	4.00	非优惠

3. 分析交通数据

造成路口拥堵的原因有多种，可以选用适当的数据分析工具对路口不同方向的车流量作分析，如图 1-4 所示。常用的数据分析工具有电子表格软件、专业的数据分析软件以及可完成复杂数据分析任务的 Python 等编程语言。这些分析工具各有优缺点，应根据实际需求选用。

例如，利用电子表格软件对 5 月 8 点到 9 点时间段内某路口各方向的平均车流量数据进行分析，可以得到各方向车辆驶出数据表（表 1-3）和驶入数据表（表 1-4）。

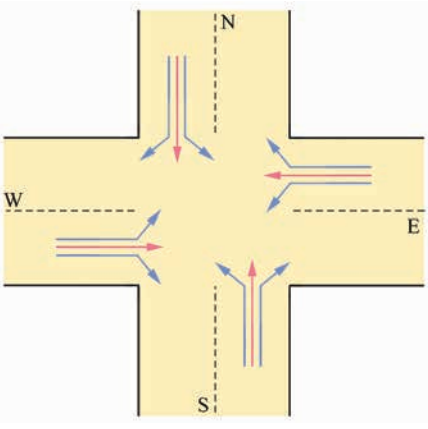


图 1-4　十字路口

表 1-3　某路口各方向车辆驶出数据表

方向	直行车辆数	左转车辆数	右转车辆数
N	925	607	327
S	1248	127	466
W	660	223	151
E	548	316	796

表 1-4　某路口各方向车辆驶入数据表

方向	直行车辆数	左转车辆数	右转车辆数
N	1248	223	796
S	925	316	151
W	548	127	327
E	660	607	466

对比各方向车辆的驶入驶出数据，可以发现该路口北方进出的车辆数均超过其余方向进出的车辆数（表 1-5）。

表 1-5　某路口各方向车辆进出数据表

方向	进	出
N	2267	1859
S	1392	1841
W	1002	1034
E	1733	1660

为了对各个方向的车流量数据有比较直观的感受，可以通过可视化图表展示数据。例如，利用电子表格软件将某路口各方向车辆进出数据表进行可视化，得到如图 1-5 所示的直方图。从图中可以看到该月南北方向上行驶车辆的数量偏多，这可能是造成路口拥堵的原因之一。

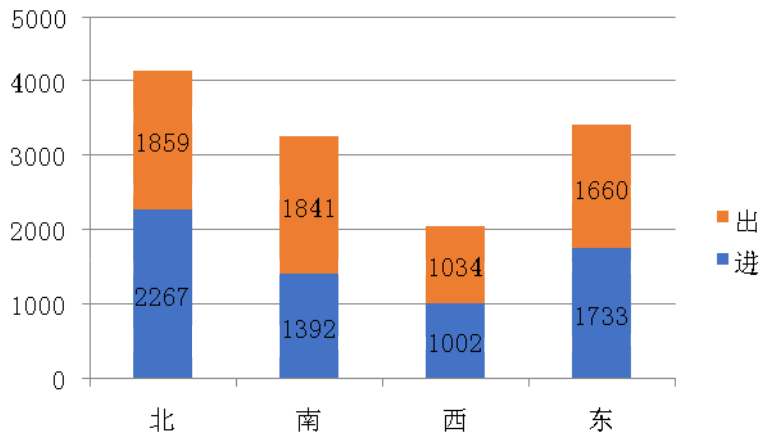


图 1-5　某路口各方向车辆进出情况图

分析该路口拥堵的原因，仅仅用一个月的数据是远远不够的，还需要对该年其他月份的路口数据进行分析，或对历年来每个月的路口数据进行分析，同时综合考虑该路口及周边路口的交通数据。这样得到的数据分析结果可以为交通管理部门缓解早高峰路口交通压力的决策提供支持：例如，在早高峰期间延长该路口车流量较多那一方向的绿灯时长。

思考与讨论

为什么一个月的路口车流量数据尚不能为决策提供支持？

活　动

1.3　以小组为单位，各组分别尝试利用一种数据分析工具，对本项目中的路口平均车流量数据表进行分析，交流分析结果并对工具进行比较。

4. 了解交通数据资源及其价值

如今，人们管理和分析的交通数据来源广泛、形式多样，并不仅仅只有车流量数据。在智能交通系统中，通过地感线圈、视频监控、手机、公交卡等传感设备和移动终端采集的人、车、路等交通要素的数据是一种重要的资源，对交通行业及其他各行业组织的运营和管理都十分重要。

以城市公交数据为例，对于一个大中型城市来说，每天从公交车辆、公交站点、公交司机或乘客等数据源处采集的公交数据类型多、数据量巨大，如图 1-6 所示。经过一段时间以后，数据的规模更是超出了传统意义上的尺度，传统的软件和工具难以胜任数据的管理和分析工作，需要采用新的、合适的数据管理工具和分析方法，才能有效地挖掘数据资源中潜在的巨大价值。

小贴士

智能交通系统是将先进的信息技术、数据通信传输技术、电子传感技术、控制技术及计算机技术等有效地集成运用于整个地面交通管理系统，而建立的一种在大范围内全方位发挥作用的，实时、准确、高效的综合交通运输管理系统。

参见 P9 知识链接“数据资源与数据价值”

图 1-6　城市公交数据资源

管理和分析公交数据，对公交公司来说，可以很好地帮助其进行车辆营运时间调整、车辆调度等，从而提高公交车辆的利用率；对交通管理部门来说，可以为道路疏导、道路规划等提供决策依据；对地图导航企业来说，可以在导航平台上为人们的公共出行提供服务；对计划选址开业的商家来说，可以获得不同地区人流量的信息，帮助商家选择合适的经营位置。数据作为资源被不同行业或企业有效管理和分析后，会创造出各种价值。

小贴士

根据公交卡的刷卡记录和居民使用公交卡的比例，可以很容易地获取不同站点公交车辆的人流量分布情况。此外，根据移动手机信令数据（手机用户与发射基站之间的通信数据）也可以获取区域内的各种人流量数据。

活　动

1.4　以小组为单位，选择某一类数据（如教育数据、医疗数据、环境数据、人口数据），查阅与数据资源、数据价值相关的案例，并在班级里开展以“数据资源与数据价值”为主题的交流会。

知识链接

数据资源与数据价值

随着时代的发展，数据已俨然成为人类社会赖以生存和发展的一项重要资源，它对国家和民族的发展、对人们的工作和生活至关重要，广泛存在于经济、社会的各个领域和部门。公司管理、商业决策、科学研究、政府政策制定，都离不开对数据资源的利用。例如，与医疗卫生和生命健康活动相关的健康医疗数据就是宝贵的数据资源。对健康医疗数据的分析与挖掘在医学临床、分子生物学、预防医学、医院管理等领域都发挥着重要作用。对各卫生医疗机构采集的患者就诊数据进行深入挖掘后，医生能优化治疗过程，精准用药，减轻患者在治疗过程中的痛苦；科研人员能研发出更有针对性的药物；医院能优化内部管理，改善患者就医体验；政府相关部门能更好地监管医疗体系。

在加强安全保障和隐私保护的前提下，越来越多的政府部门和公司将数据资源开放共享，这使得更多的组织可以利用这些数据资源，充分挖掘其价值。同时，各行各业也在积极推动行业内及不同行业间的数据资源整合，加强数据资源的发掘运用。

数据作为信息社会的重要资源，它的价值来源于数据本身、技术和思维三个层面。数据本身是数据价值的起点，只有拥有数据或能够接触到数据才能开启数据的价值。数据的拥有者需要借助于各种技术，特别是数据管理与分析技术，获取数据中隐含的信息，在具

体的业务中体现数据的价值。数据思维就是提出数据的创新性用途，挖掘数据的新价值。有些看似毫不相关却非常重要的数据需要依靠人类的智慧不断分析，通过数据思维创新性地实现数据的价值。

数据管理与分析技术

数据资源的开发利用离不开数据管理与分析技术。数据管理技术可以存储、管理数据，而数据分析技术可以探寻数据间的关系，获取有价值的信息。通过数据管理与分析技术，能从数据中挖掘信息和知识。目前，数据管理与分析技术已经渗透到各个领域之中。因此，建立在大量真实数据的管理与分析基础上的行为和决策，不仅维护了数据的安全和秩序，而且大大提高了生产、生活的效率和质量。

1. 数据管理技术

数据管理技术发展至今，经历了以下几个阶段：

（1）人工管理阶段

时间：20 世纪 50 年代中期以前。

功能：计算机主要用于科学计算。当时没有磁盘等直接存取数据的设备，只有纸带、卡片、磁带等外部存储设备；软件只有汇编语言，没有操作系统和管理数据的专门软件。数据处理的方式基本是批处理。

特点：① 数据不保存。② 系统没有专用的软件对数据进行管理。每个应用程序都要包括数据的存储结构和存取方法等。程序员在编写应用程序的同时，还要安排数据的物理存储，负担很重。③ 数据不共享。数据是面向程序的，一组数据只能对应一个程序。④ 数据不具有独立性。程序依赖于数据，如果数据的类型、格式、输入 / 输出方式等逻辑结构或物理结构发生变化，则必须对应用程序作相应的修改。

（2）文件系统管理阶段

时间：20 世纪 50 年代后期至 60 年代中期。

功能：计算机不仅用于科学计算，还在信息管理方面发挥着作用。随着数据量的增加，数据的存储、检索和维护成为迫切需要解决的问题，数据管理技术迅速发展起来。磁盘、磁鼓等直接存取设备开始普及，这一时期的数据管理技术是把计算机中的数据组织成相互独立的、被命名的数据文件，并可按文件的名字来进行访问，对文件中的记录进行存取。

特点：① 数据可以长期保存。由文件系统管理数据，可以对数据进行反复处理，并支持文件的查询、修改、插入和删除等操作。② 文件的形式多样化，数据具有一定的独立性。③ 文件系统实现了记录内的结构化，但从文件的整体来看却是无结构的。其数据面向特定的应用程序，因此数据的共享性、独立性差，冗余度大，管理和维护的成本很高。

（3）数据库管理阶段

时间：20 世纪 60 年代后期以来。

功能：数据库系统克服了文件系统的缺陷，提供了对数据更高级、更有效的管理。这个阶段的程序和数据的联系通过数据库管理系统来实现。

特点：① 数据结构化。在描述数据时不仅要描述数据本身，还要描述数据之间的联

系。数据结构化是数据库的主要特征之一，也是数据库系统与文件系统的本质区别。② 数据共享性高、冗余少且易扩充。数据不再针对某一个应用，而是面向整个系统，数据可被多个用户和多个应用共享使用，而且容易增加新的应用。③ 数据独立性高。④ 数据由数据库管理系统统一管理和控制。数据库为多个用户和应用程序所共享，对数据的存取往往是并发的，即多个用户可以同时存取数据库中的数据，甚至可以同时存取同一个数据。

（4）大数据背景下的数据管理技术

时间：21 世纪初期以来。

功能：在大数据时代下，可以用于分析的数据变得非常多，有时甚至可以处理和某个现象相关的所有数据，不再依赖于随机采样，因此对数据的精确度要求也有所减弱。同时，通过大数据的分析与挖掘，可以找出事物之间的相关关系，从而体现出数据的巨大价值。

特点：大数据的 4V 特征是 Volume（数据量）、Velocity（处理速度）、Variety（多样性）、Veracity（真实性）。

常用方式：

• 并行计算

大数据处理的传统方法是使用并行数据库系统。并行数据库系统是在大规模并行处理系统和集群并行计算环境的基础上建立的高性能数据库系统。

• NoSQL 数据库

NoSQL 数据库是指数据模型定义不明确的非关系数据库。NoSQL 数据库具有灵活的数据模型、高可扩展性和较好的发展前景。它是突破了关系数据库在处理大数据问题上局限性的一种新型数据库。

• 云数据库技术

云数据库技术是云计算的一个重要分支，是对云计算的具体运用。云数据库是部署在虚拟化云计算环境中的数据库。它极大地增强了数据库的存储能力，消除了人员、硬件和软件的重复配置，让软硬件升级变得更加容易，同时也虚拟化了许多后端的功能。

2. 数据分析技术

数据分析是数学与计算机科学相结合的产物。数据分析是指用适当的统计分析方法对采集来的大量数据进行分析，提取有用信息和形成结论，并对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，但直到计算机的出现才使得数据分析的实际操作成为可能。在现实生活中，数据分析可帮助人们作出判断，以便采取适当行动。

在统计学领域，有些人将数据分析划分为描述性数据分析、探索性数据分析以及验证性数据分析。

（1）描述性数据分析：对调查对象总体所有变量的有关数据作统计性描述，主要包括数据的频数分析、数据的集中趋势分析、数据离散程度分析、数据的分布以及一些基本的统计图形。

（2）探索性数据分析：通过绘制统计图形、编制统计表格、计算统计量等方法来探索数据的主要分布特征，揭示其中可能存在的规律，为选择合适的方法分析数据奠定基础。

（3）验证性数据分析：利用相关数据对已有假设进行证实或证伪。

随着大数据时代的到来，数据在加速地增长，用传统的方法已很难有效地分析大数据，因此数据分析的工具、技术和分析方法也在不断发展，以满足海量数据存储、管理和实现其价值的诉求。大数据是“全数据”分析，数据来源广、类型多、数据量大，而传统的数据分析是一种抽样数据分析，一般针对少量的数据。大数据分析主要利用分布式数据库或者分布式计算集群来对存储于其内的海量数据进行分析。传统的数据分析更侧重统计上的分析，而大数据的数据分析核心方法是数据挖掘。数据挖掘一般没有预先设定好的主题，主要是在数据上运行各种数据挖掘算法，从而发现规律或异常，满足一些高级别数据分析的需求。

项目二

了解网络购物数据的管理与分析

——经历数据管理与分析的流程

现今，网络购物已成为消费者购物的主要方式之一，从进入网上商店查询到选定并购买商品的一系列过程产生了大量的数据，对这些数据的分析可以帮助商家了解消费者的购物习惯，从而将更多适合消费者喜好的商品推荐给他们（图 1-7）。此外，对这些网络购物数据的分析还可以指导商家的营销和新商品上架等工作。其中，对网络购物数据中订单数据的分析是一种常见的分析。通过对消费者订单数据的分析，可以发现消费者购买商品中的隐含规则，据此设计促销方案。

在本项目中，我们将通过寻找订单数据中的隐含规则，了解业务需求分析、建立数据管理与分析问题整体解决方案的基本过程；了解如何对既定的方案进行分析、评价，发现问题并优化方案。

图 1-7　网络购物

项目学习目标

完成本项目学习，须回答以下问题：

1. 分析业务需求、建立数据管理与分析问题整体解决方案的基本过程是什么？

2. 如何对既定方案进行分析、评价，发现问题并优化方案？

项目学习指引

1. 分析业务需求

核心概念

业务需求是为了实现商业目的而产生的需求，它通常描述组织为什么要去执行相应的任务。

网上购物发展至今，如何提升销售额一直是网上商店经营者的主要业务需求。一般情况下可以通过对消费者订单数据进行分析，找出消费者购买的商品之间的关系，了解消费者的购买行为，有针对性地制定销售方案。如根据消费者的购买行为向其推荐符合其购买偏好的商品，分析消费者购买的商品之间的关系来制定捆绑销售策略，以及针对消费者的消费心理和购买量开展相应的促销活动等，以此提升网上商店的销售额。

参见 P19 知识链接“数据管理与分析问题整体解决方案”

本项目主要对消费者购买的商品之间的关系进行分析，从而帮助商家制定有效的捆绑销售和推荐策略，达到提升销售额的目的。

活　动

2.1　假设你与小伙伴合作运营一家网上文具店，试分析影响文具店销售额的因素有哪些，并提出文具店的业务需求，尝试对业务需求进行分析，查阅互联网上的资料，撰写业务需求分析文档。

2. 管理网上商店订单数据

（1）数据的采集与存储

网上商店的商品数据、消费者数据等都保存在相应的数据库中。当消费者完成购买行为之后，网上商店自动生成该消费者的订单数据，以二维表的形式保存在数据库中。例如，某网上商店的订单数据存储在订单表、订单明细表等多张表中，订单表中含有订单编号、会员编号、会员名、付款金额、订购日期、是否付款、收货地址等数据，订单明细表中含有订单编号、商品编号、商品名称、订购数量等数据。

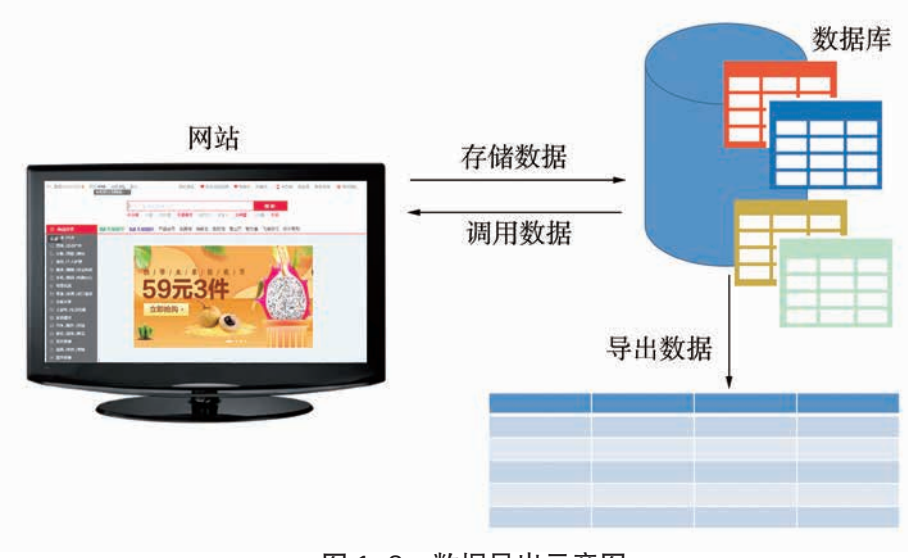


图 1-8　数据导出示意图

为了对数据库中的订单数据进行分析，先要从数据库的大量数据中选择并导出所需数据（图 1-8）。例如，根据业务需求分析可知，本项目计划分析的是订单中商品之间的相关关系，因此需要的数据是每张订单购买了哪些商品，可以从数据库中转换导出订单表，如表 1-6 所示。

表 1-6　转换导出的订单表

订单编号	会员编号	商品名称
20911000364424100	A1	连衣裙，外套，饼干
23834390588606299	A2	手机，外套，T 恤
20462982543279927	A3	连衣裙，手机，外套，T 恤
20471829726608330	A4	手机，T 恤
23443211525636382	A5	连衣裙，外套
19875820480768012	A6	**，&&

思考与讨论

导出的订单表中能否出现会员名、收货地址等数据？为什么？

（2）数据预处理

仔细检查导出的订单表，可以发现，订单编号为“19875820480768012”的订单中存在异常数据，其“商品名称”出现了异常。这种异常数据被称为噪声数据，可能会对后续的数据分析产生影响，导致分析结果不准确，因此需要

小贴士

数据预处理是在数据分析之前对数据进行的一些处理。数据预处理方法有很多种，主要包括数据清理、数据集成、数据变换、数据归约等。数据清理的目的是清除有错误或有问题的数据。

进行数据预处理。这里由于数据较少，可以通过肉眼观察找出噪声数据，而实际处理数据时，往往通过程序自动实现噪声数据的查找。程序查找不仅速度快，方便应对大量数据的预处理，而且可以根据设置的范围查找，不容易错漏。

在无法根据现有的数据推测出异常值内容的情况下，可以将包含异常值的这条记录删除。对异常值的处理在可推测的情况下，也可以通过其周围的数据进行推算，如周围都是数值型数据，则可通过求平均值等方法推测替换。

发现并清除噪声数据后，还可对订单表作如下预处理：首先，为了方便分析，将连衣裙、手机、外套、饼干、T 恤等五种商品分别用编号“1”“2”“3”“4”“5”表示，如表 1-7 所示。然后，对表中的数据进行统计转换，转换结果如表 1-8 所示（为简化本例，以下仅以导出的订单表的前四条记录为例）。

表 1-7　商品名称—编号对应表

商品名称	编号
连衣裙	1
手机	2
外套	3
饼干	4
T 恤	5

表 1-8　会员编号—已购商品编号对应表

会员编号	已购商品编号
A1	1,3,4
A2	2,3,5
A3	1,2,3,5
A4	2,5

活　动

2.2　打开配套资源中网上文具店的订单数据表，观察订单数据表中是否有噪声数据，并进行数据预处理。

3. 分析订单数据

分析订单数据，找出商品之间的关系，可以使用的数据分析方法有很多。以下将利用 Apriori 算法对订单数据进行分析，寻找商品之间的关联规则。

① 为了寻找商品之间可能存在的购买关系，可以从会员编号—已购商品编号关系对应表的数据里分析出所有可能存在的关系，并用项集表示，结果如表 1-9 所示。

表 1-9　会员编号—已购商品编号关系对应表

会员编号	已购商品编号关系
A1	{1,3} {1,4} {3,4} {1,3,4}
A2	{2,3} {2,5} {3,5} {2,3,5}
A3	{1,2} {1,3} {1,5} {2,3} {2,5} {3,5} {1,2,3} {1,3,5} {2,3,5} {1,2,3,5}
A4	{2,5}

② 利用 Apriori 算法，计算每种关系出现的次数，即支持度计数，结果如表 1-10 所示。

表 1-10　支持度计数表

商品编号关系	支持度计数
{1,2}	1
{1,3}	2
{1,4}	1
{1,5}	1
{2,3}	2
{2,5}	3
{3,4}	1
{3,5}	2
{1,2,3}	1
{1,3,4}	1
{1,3,5}	1
{2,3,5}	2
{1,2,3,5}	1

③ 找出表 1-10 中支持度计数大于 1 的关系：{1,3}{2,3}{2,5}{3,5}{2,3,5}。

小贴士

Apriori 算法是一种最有影响的探求数据之间关联规则的算法。

项集即若干个项的集合。这里消费者购买的一件或多件商品即可作为一个项集。

支持度计数是指候选项集在记录中出现的频数。

小贴士

本项目中分析的是非常少量的订单数据。一般情况下，由于订单中的数据量非常大，不可能通过人工利用以上的方式寻找商品之间的关联规则，而是利用 Python 等软件编写程序，自动处理订单数据。

④ 以关系 {1，3} 为例，通过计算置信度，可以分别抽象出两条规则，如表 1-11 所示。

表 1-11　抽象出的两条规则

规则	置信度	解析
1 → 3	100%	购买商品 1 的用户，有 100% 的概率购买商品 3
3 → 1	67%	购买商品 3 的用户，有 67% 的概率购买商品 1

小贴士

置信度表示一个事物出现，另一个事物同时出现的概率。

A 对 B 的置信度，表示在 A 出现的前提下 B 出现的概率，利用公式可以表示为：

$$\text{置信度} = \frac{\text{A，B 同时出现的支持度计数}}{\text{A 出现的支持度计数}}$$

活　动

2.3　打开配套资源中的 apriori.py 程序，调用已经预处理过的订单数据表，获取网上文具店全部订单中商品之间的关联规则。

4. 完成科学决策

从以上分析结果可以发现，对于 {1,3}，消费者购买商品 1 之后再购买商品 3 的概率为 100%，因此在设计营销方案时，可以将商品 1 和商品 3 进行捆绑销售。消费者购买商品 3 之后再去购买商品 1 的概率为 67%，那么，在设计营销方案时可以考虑在消费者购买了商品 3 之后，再向其推荐商品 1。

小贴士

本分析结果并不能代替整体订单数据的分析结果，实际分析时应使用完整数据。

活　动

2.4　根据活动 2.3 的数据分析结果，试着为网上文具店制定营销方案。

5. 评价、优化整体方案

在实际工作中，利用 Apriori 算法编写的程序分析数据时会遇到以下问题：当分析的数据量很大时，往往关系也会非常多，从而导致复杂度增加，计算机所消耗的资源与时间呈指数递增，计算的结果也会受影响。

因此，当要分析的数据量较大时，可以根据实际需求对 Apriori 算法进行优化，提高分析效率。例如，在订单数较多时可采用 Apriori 的优化算法——En-Apriori 算法。

数字化学习

上网查找资料，了解 Apriori 算法和 En-Apriori 算法。

活　动

2.5　尝试对网上文具店的营销方案进行分析、评价，并优化方案。

知识链接

数据管理与分析问题整体解决方案

在各行各业中，大到跨国公司，小到微店、微商，其日常业务涉及诸多环节。随着业务的发展，会不断产生新的问题和需求。整体解决方案就是为了解决这些新问题或需求而设计的一个全面系统的综合性解决方案，它是在对数据进行深入分析之后，在充分满足业务需求的基础上形成的系统化的解决方案。整体解决方案是一种“量体裁衣”式的综合性方案，在不同的行业中它的形式不完全一样。尽管如此，整体解决方案的设计，一般都要经过如下几个步骤，如图 1-9 所示。要注意的是，在整个过程中的每一个步骤都离不开方案优化。

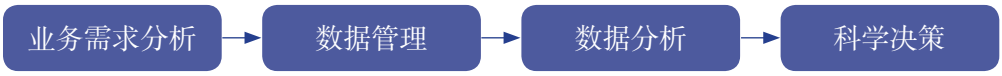


图 1-9　设计整体解决方案的一般过程

1. 业务需求分析

业务需求分析最重要的是确定方案目标。开展工作之前确定目标，有助于抓住工作重点，确保工作顺利完成。一个全面、系统的整体解决方案往往会涉及诸多领域、流程，也可能需要和多个部门、客户打交道。为了防止决策的偏差，一般需要通盘考虑各方面的因素。因此，在设计整体解决方案之前，需要全面了解现实情况，汇总来自各方面的“诊断”信息，找出当前问题的症状及原因，明确需要解决的具体问题。

2. 数据管理

数据管理是一个对数据进行有效采集、存储、处理和应用的过程。确定了需求、明确了任务后，首先需要着手寻找“原料”——数据。数据采集是根据需求采集数据，从而使数据分析有的放矢。数据采集的方法有很多，有问卷调查、资料查阅、传感器采集、智能设备采集、网络爬虫采集、从已有数据库中采集等。采集到的数据通常通过数据库进行存储、处理和应用。

随着用户需求的提升，传统的关系数据库已无法支撑大规模、形态结构各异、支持决策分析的数据业务，因此出现了非关系数据库。随着数据采集、存储和分析技术的飞跃式发展，人们可以更进一步地利用海量、类型多样和来源各异的数据，而不再是少量的样本数据，数据管理进入了大数据时代。

海量的数据难免会包含噪声数据、空缺数据和不一致性数据，因此需要通过数据预处理技术提升数据质量。数据预处理的方法包括数据清理、数据集成、数据变换和数据归约。数据清理可以去掉数据中的噪声，纠正不一致的数据。数据集成可以将来自多个数据源的数据整合成一致的数据进行存储。数据变换则是将数据变换成适于数据分析挖掘的形式。数据归约用于简化数据集的表示，降低数据规模。

3. 数据分析

对于数据规模较小的简单数据分析任务而言，可以通过 Excel、Access、MySQL 等软件完成数据分析和可视化任务。然而，对于大规模的数据和更为复杂的数据分析任务而言，需要对数据进行加工转换，然后利用专业化软件对数据进行深入的分析提炼，从而发现数据背后的秘密，为决策等提供重要依据。分析数据时，要能准确全面地反映实际需求，从而保证设计的方案合理和实用。

4. 科学决策

科学决策是指决策者在科学的决策思想指导下，遵照科学的决策规律，借助各种科学的分析手段和方法，在调查研究、充分掌握有关信息的基础上，依据一定的程序选择最优方案。科学决策是对经验决策、盲目决策以及其他一些不规范、容易造成较大失误的决策方式的否定和改进，避免非理性的决策后果。在大数据时代，随着数据意识的不断提升，对数据的管理与分析将大大增强决策的科学性。

网上商店的数据分析结果，不仅可以帮助商家制定有效的营销方案，也可以作为精准投放广告策略制定的依据。行业数据分析对于行业的科学决策有重大的意义。以电信与金融服务业为例，电信业经由数据分析能够设计出不同的服务组合以扩大利润；保险业能通过数据分析侦测出可能不寻常的投保组合并作预防；在医疗领域，对病人进行疗程组合时，数据分析的结果能作为这些疗程组合是否会导致并发症的判断依据。

当然，任何方案都不一定是最佳方案。方案设计完成后，能否满足用户的需求？是否适应发展的需要？是否安全稳定？是否经济适用？是否为最现实可行的方案？这就需要对方案进行评价，从而发现其中的问题，进行改进和优化。方案优化需要从整体上对业务需求分析、数据管理与分析问题的整体解决方案进行分析、优化。例如，重新分析业务需求，增加数据采集的范围，采用其他算法对数据进行分析，制定新的营销策略等。

在经济快速发展的今天，整体解决方案已成为组织提升科学化管理水平、实现现代化服务的必然产物。根据企业的需求，结合客户业务现状和未来发展需要为企业设计整体解决方案，将助力企业优化业务流程、提升运作效率。

单元挑战　调查校园数据管理现状

一、项目任务

以小组为单位，分别走访学校的教务处、体育室、图书馆等部门，调查学校是如何管理学生学籍和成绩、体质健康测试、图书借阅情况等数据的，了解各部门使用数据管理与分析技术的现状。

二、项目指引

1. 在班级里成立小组，各组分别选取学校的一个部门为调查对象。

2. 设计调查方案。

调查方案（参考样例）

调查目的：了解学校体育室数据管理和分析现状。

调查对象：学校体育室。

调查内容：

- 日常工作涉及哪些数据？
- 这些数据分别是通过怎样的渠道和工具采集到的？
- 有无应用计算机管理数据？使用了哪些软件来管理和分析数据？
-
-

3. 开展调查，完成小组的调查报告。

4. 汇总各组的调查结果。

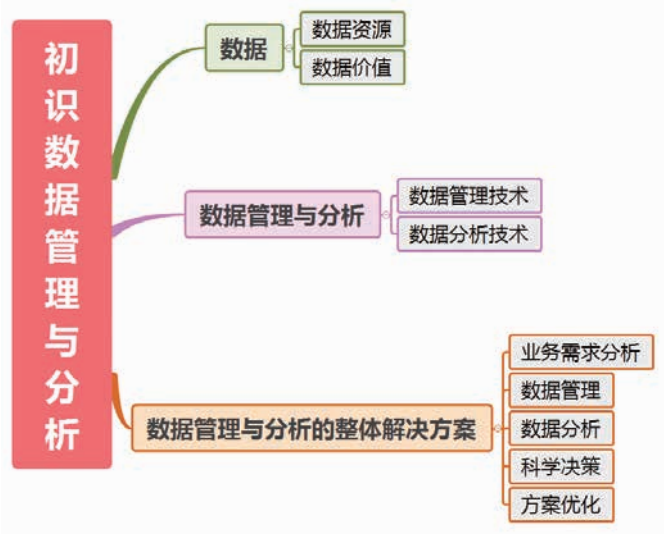
调查部门	工作涉及的数据	数据采集渠道和工具	管理和分析数据的手段和工具	发现的问题及建议

三、交流评价与反思

在班级里展示各组的调查报告，交流数据管理与分析技术的重要性。反思在完成项目的过程中，小组或个人遇到了哪些困难，又是如何克服困难完成项目任务的。

单元小结

一、主要内容梳理



二、单元练习

1. 为了充分挖掘数据资源的价值，交通运输、健康医疗、教育等各个行业都在促进行业数据资源的开放共享，同时也在不断完善数据资源开放共享的安全机制，以确保开放数据资源的安全。围绕数据资源开放共享与数据安全问题写出你的观点。

2. 某校为了有效管理学生的成绩，计划设计一个成绩管理系统。

（1）在设计成绩管理系统时，如何开展业务需求分析？

（2）为成绩管理系统采集数据的时候，需要采集哪些数据？可以使用哪些数据采集工具？

（3）成绩管理系统需要采集学生的家庭地址、家庭成员身份号码等数据吗？个人或企业能否随意采集公民的身份号码等数据？为什么？

三、单元评价

评价内容	达成情况
能认识到数据是一种重要的资源（A，R）	
知道通过数据管理与分析技术，可以使数据实现其应用价值（A，R）	
能够感受数据管理与分析技术的重要性（A，R）	
初步了解分析业务需求、建立数据管理与分析问题的整体解决方案的基本过程（T，I）	
能够尝试对既定方案进行分析、评价，发现问题并优化方案（T，I）	

说明：A—信息意识，T—计算思维，I—数字化学习与创新，R—信息社会责任

第二单元

数据管理

数据是一种重要的资源，但要利用各行各业所产生的庞大数据，首先要将这些原本看起来杂乱无章的数据采集、存储起来，并进行精心的组织和管理。也就是说，需要利用数据管理技术对数据进行有效管理。

生活中常见的车站售票、银行存取业务、超市收银，以及网上书店、旅行 App 等应用背后，都有一种先进的数据管理技术——数据库技术支撑着。当你通过网页访问一家网上书店时，其实你正在访问存储在某个数据库中的数据，同时你的访问记录也可能被存储到数据库中；当你提交订单后，订单数据也被存储到数据库中……我们的现代生活已离不开数据管理技术。

在本单元中，我们将结合案例了解数据管理工作，并掌握和体会数据的分类、抽象、提取和查询等思想和方法。

学习目标

- 了解数据采集的途径与工具，能利用适当的工具对数据进行采集和分类。
- 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。
- 认识噪声数据的现象和成因。
- 了解关系数据模型的基本概念，掌握设计简单数据库的逻辑结构的方法。
- 能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。

单元挑战

建立年级作业评价数据库

项目二

了解健身数据的采集与分类

——认识数据的结构化

随着人们的健康意识不断增强，以健身为目的的各类运动渐渐成为时尚。由于个体情况存在差异，每个人适合的运动项目和所能承受的运动负荷是不同的。因此，科学地安排运动的内容显得格外重要和必要。人们常常会选择去健身俱乐部，请健身教练帮助自己规划有针对性的运动项目。

针对不同的健身者，健身教练如何推荐合适的运动项目？要解决这一问题，首先要采集健身者的基本健康与运动数据（图 2-1）。准确、全面、可靠地采集到健身者的数据，去伪存真，并分类整理，这是健身者获得合适的运动项目推荐的基础和保障。

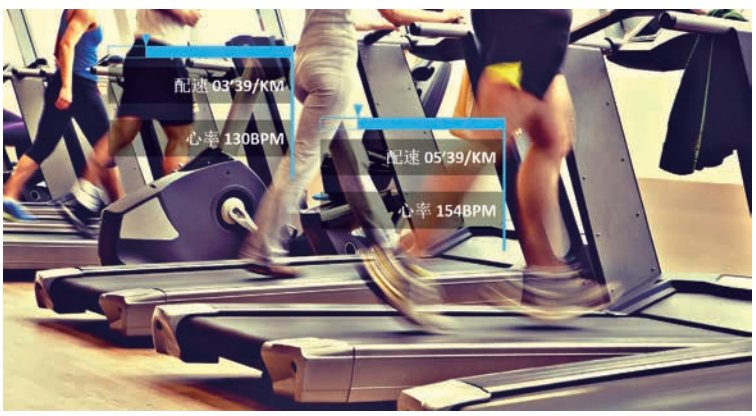


图 2-1　运动数据的采集

项目学习目标

在本项目中，我们将结合某健身俱乐部采集会员健身数据的实例，了解如何利用多种途径对数据进行采集和分类；认识噪声数据的现象与成因，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。

完成本项目学习，须回答以下问题：

1. 数据采集的途径和工具有哪些？如何利用恰当的工具对数据进行采集和分类？

2. 不同结构化程度的数据有何区别？在管理与应用上各有哪些特点？

3. 什么是噪声数据？噪声数据产生的原因有哪些？

项目学习指引

1. 采集会员健身数据

贸然训练往往会对会员身体造成伤害，因此健身教练为会员规划训练项目前，要先了解会员身体的基本情况，并进行相应的数据采集工作，即采集会员个人基本数据、健康数据和运动数据，为会员建立个人身体运动档案。在对这些数据进行分析后，才能提供合适的训练建议。

（1）人工采集数据

个人基本数据主要是会员的身份数据，包括姓名、性别、出生年月和联系方式等。健身教练会通过询问或让会员填写表格、问卷等方式，采集会员的个人基本数据。

核心概念

数据采集是指按照既定的目的，通过人工或利用设备，获取客观世界中相关的数据，并输入计算机进行存储的过程。

思考与讨论

个人身份数据大多是一些敏感数据，一旦泄露会带来很多负面影响。日常生活中应该如何保护自己的个人身份数据？

参见 P30 知识链接“数据采集”

（2）利用设备采集数据

除了采集会员的个人基本数据外，一般还需要了解会员的健康数据和运动数据等。这类数据的采集途径和工具较多，视健身俱乐部及会员的实际情况而定。

例如，有些健身俱乐部会利用人体成分分析仪（图 2–2）来采集会员的健康数据。在人体成分分析仪的面板上输入会员的年龄、性别，人体成分分析仪通过采集，得到会员的身高、体重、身体各部位体脂率、肌肉量和目前的基础代谢量等数据。采集到的数据可以传输到与仪器相连的计算机上。需要注意的是，人体成分分析仪仅对人体各部位体脂率、肌肉量等进行评估，不能完全作为会员健康程度的依据。如果要对身体健康状况进行总体评估，还需要会员提供最近一年的医学体检数据。

小贴士

人体成分分析仪采集到数据后，还可以通过相关的健康分析软件，形成人体成分健康分析报告，如图 2–3 所示。

有些健身俱乐部会让会员当场做一些体能测试，并在

图 2-2　人体成分分析仪

图 2-3　人体成分健康分析报告

思考与讨论

回忆自己在医院体检的经历，说说哪些体检数据是通过医生人工采集的，哪些是通过医疗仪器自动采集的。

小贴士

通过运动捕捉技术，可以在人们训练过程中实时追踪和记录其动作数据。如，在篮球比赛馆内安装多个摄像头，以每秒 25 帧的速度收集篮球以及每位球员的移动数据。动作捕捉技术采集的球员运动数据，可在赛后生成动画效果模拟比赛，并供球队进行技术和战术分析。

体能测试时利用摄像设备为会员录制视频、拍摄照片，这类数据可直观反映会员的运动状态，也便于对比训练前后的效果。还有些健身俱乐部会通过运动捕捉技术，对会员体能测试时的动作数据进行实时、精确、定量的连续采集（图 2-4）。

此外，目前还有很多便携的可穿戴设备用于日常运动数

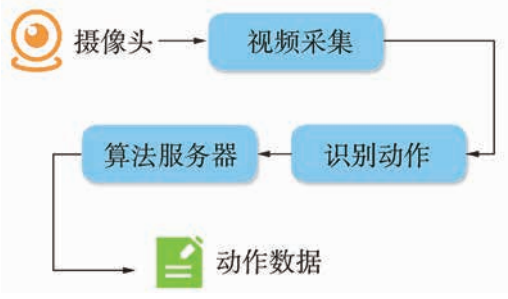


图 2-4　某种运动捕捉设备的核心技术

据的采集，例如运动手环（图 2-5）。健身教练也可以参考会员利用这些设备采集的日常运动数据（图 2-6），了解会员运动情况。

图 2-5　运动手环

图 2-6　运动手环采集的数据

思考与讨论

1. 你平时有无采集自己的日常运动数据？生活中采集运动数据的工具和途径有哪些？

2. 常见的可穿戴设备有哪些？它们分别用来采集哪些数据？

活　动

3.1　尝试佩戴运动手环或利用手机上的运动 App 记录自己的日常运动数据，并了解其采集了哪些数据。

3.2　查阅资料，了解采集网络数据的途径和方法。

核心概念

结构化数据（structured data）是指存储在数据库里，可以用二维表结构来逻辑表达实现的数据。包括预定义的数据类型、数据格式和数据结构，其模式使其易于搜索。

非结构化数据（unstructured data）是指不方便用二维表结构来逻辑表达实现的数据，没有固定的结构。包括所有格式的文本、图像、音频和视频数据等。

半结构化数据（semi-structured data）是指介于结构化数据和非结构化数据之间的数据。半结构化数据往往具有一定的结构性，一般为有识别模式的文本数据文件，支持语法分析，如 XML 文件。

参见 P30 知识链接“不同结构化程度的数据”

2. 分类存储会员健身数据

健身教练采集到的个人基本数据、健康数据、运动数据都是零散的，而且格式各异，只有对数据进行分类整理、存储之后才方便后续的管理和应用。

根据数据结构化程度的不同，可以将数据分为结构化数据、非结构化数据和半结构化数据。不同结构化程度的数据，其存储方式各不相同。

通过表格、问卷等方式采集到的个人数据一般可以用二维表结构来表达和存储，如图 2-7 所示，这些数据属于结构化数据。

编号	姓名	性别	出生年月	联系方式

图 2-7　会员表示例

视频和图片等记录会员体能测试的运动数据不方便用二维表结构来表达和存储，一般以文件的形式放入文件夹中进行存储，如图 2-8 所示，这些数据属于非结构化数据。

图 2-8　非结构化数据的存储

思考与讨论

每次旅行时拍摄的照片和视频，你是如何存储的？怎样存储能方便查找？

除了以上两类数据，还有一类半结构化数据。这类数据既不方便利用二维表结构来表达和存储，也不能简单地像视频和图片一样以文件形式单独存储。例如，健身教练可以将会员的个人基本数据存储在一张二维表中，但如果还需要记录会员健康状况，包括是否有慢性病以及有什么慢性病，就难以用二维表存储了。半结构化数据可以利用表格（非二维表）结构来表达和存储，也可以用 XML 文件来存储。

活　动

3.3　无论是何种结构的数据，在存储之后都会面临管理与应用的问题。利用互联网查阅资料，总结不同结构化程度的数据在管理与应用上的特点。

3. 认识噪声数据

在采集数据过程中，由于误操作等原因，可能会产生错误或异常数据，这种有问题的数据就是噪声数据。例如，健身教练在采集会员个人数据时，将某会员原本的年龄“26”误写为“9”，就出现了噪声数据。又如，使用人体成分分析仪时，由于机器发生故障，也可能会产生噪声数据。

由于噪声数据的存在，采集到的数据有可能无法准确地反映会员的身体状况，这会导致依据数据分析出的会员身体状况评估不准确，进而导致制定的运动方案出现偏差。正是由于噪声数据会影响数据价值的获取和科学决策，因此，一般需要对采集到的数据进行预处理，采用一定的技术或方法来检查并修正数据，以保证数据的质量。例如，通过检测数据类型或利用数据完整性约束规则查找出噪声数据，然后删除噪声数据或利用数据推算替换等方式清除噪声数据。

清除噪声数据后，健身教练可以对数据进行分析，为会员提供恰当的健身建议。采集到准确、可靠、真实的数据是数据管理和分析工作顺利进行的前提和基础，对解决数据业务问题起着关键作用。

核心概念

噪声数据（noisy data）是指在数据采集过程中产生的错误的、异常的、不完整的或无意义的数据。

小贴士

项目二中采集数据库的订单数据时，也产生了噪声数据，我们通过删除异常订单的方式清除了噪声数据。

参见 P32 知识链接“噪声数据”

活　动

3.4　在数据的采集、存储过程中都有可能产生噪声数据。上网查阅资料，了解噪声数据产生的各种原因；查找一些因噪声数据影响决策的案例，说明数据准确性、可靠性和真伪性的重要作用。

3.5　假设你是一个网上书店开发者，要创建一个网上书店平台在线销售图书，需要采集哪些数据？可以通过什么途径和工具？上网浏览已有的各种网上书店，回答以上问题，并在班级里交流自己的收获和想法。

知识链接

数据采集

根据不同的数据来源，可以选择不同的数据采集途径。需要调查人员参与采集的数据称为人工采集数据。例如，健身教练请会员填写表格，采集会员基本信息；交警向驾驶员询问交通事故发生的情境；心理咨询师与来访者交流，了解对方的心理状况等。人工采集数据通常使用观察法、访谈法、测验法等调查方法。

除了人工采集数据之外，还可以利用设备采集数据。例如，利用录音机、摄像设备等数据采集设备采集现场音频、图像、视频等数据；利用各种传感器实时采集温湿度、压力、速度等数据。

此外，由于互联网技术的迅速发展，互联网已经成为一种重要的数据来源。互联网上的海量数据可以使用自动化的采集软件获取。例如，网络爬虫（或称网页蜘蛛）就是一种按照一定规则自动抓取网络数据的程序或脚本。

如今，随着技术的发展，数据采集的途径和工具越来越多，数据泄露、数据侵权问题也日益严重。一方面，我们在使用某些技术和工具采集数据时，需要遵循相应的道德规范，维护良好的网络环境和秩序。例如，在使用网络爬虫软件前，需要了解哪些网页数据可以抓取，哪些不能，否则有可能侵权；抓取的数据若属于他人的隐私或商业秘密，应及时停止抓取并删除。另一方面，我们在日常生活中要注意保护自己的个人数据不要被他人悄然或恶意地采集。例如，应谨慎使用社交网络平台，你填写的个人信息有可能会透漏你的真实身份；应妥善处理存储有个人账户资料的废旧手机，更换之前务必做好彻底清理工作。

不同结构化程度的数据

根据结构化程度的不同，数据可以分为结构化数据、非结构化数据和半结构化数据。

1. 结构化数据

能够用二维表结构来逻辑表达实现的数据属于结构化数据，如图 2-9 所示。结构化数据就是行数据，严格地遵循数据格式与长度规范，主要通过关系数据库进行存储和管理。结构化数据具有任何一列数据都不可以再细分，任何一列数据都是相同的数据类型等特征。

结构化数据的应用很多，例如航空预订系统、库存控制、销售事务等。

编号	姓名	出生日期	性别	民族	籍贯	教育程度
1	张元	1988.10.06	男	汉族	湖南衡阳	本科
2	王红娟	1981.01.08	女	汉族	江苏苏州	硕士

图 2-9　结构化数据示例

2. 非结构化数据

相对于结构化数据而言，无法用二维表结构来逻辑表达实现的数据称为非结构化数据，例如全文文本、图像、音频、视频等数据，它们一般以文件形式存储在文件夹中，或通过非关系数据库进行存储和管理。例如，图像数据就属于非结构化数据。图 2-10 是一张应聘者照片在计算机中的部分图像数据，它与某企业应聘人员基本信息表中用二维表表示的结构化数据截然不同，没有固定的结构。

非结构化数据的典型案例包括医疗影像系统、教育视频点播、视频监控、国土地理信息系统、文件服务器等。

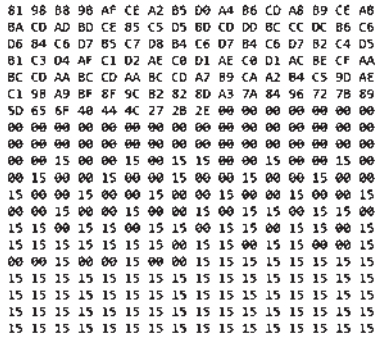


图 2-10　某应聘者照片的部分图像数据

3. 半结构化数据

半结构化数据是介于结构化数据和非结构化数据之间的数据，例如，个人简历中的工作经历就属于半结构化数据。由于每个人的经历不同，例如教育背景、工作经历、技术技能等，有的简历比较简单，有的简历则较为复杂。通常要完整地保存这些数据并不容易。因此，在存储半结构化数据时，可以将半结构化数据转化为结构化数据，或者利用 XML（eXtensible Markup Language，可扩展标注语言）文件进行存储。将简历中的半结构化数据转化为结构化数据存储时，一般会采用对个人简历中每一类别建立子表的方式，例如建立基本情况子表、求职意向子表、自我评价子表、工作经历子表等，并在简历主表中加入一个备注字段，将不能归结到以上子表的内容保存到备注中。

半结构化数据的常见存储方式是 XML 文件。XML 是一种标准通用标记语言的子集，是一种用于标记电子文件使其具有结构性的标记语言。它可以用来标记数据、定义数据类型，是一种允许用户对自己的标记语言进行定义的源语言。它非常适合万维网传输，提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据。利用 XML 文件保存简历时可将简历中不同类别的数据保存在 XML 文件的不同节点中，如图 2-11 所示。

半结构化数据的典型应用有邮件系统、教学资源库、数据挖掘系统、档案系统等。

将数据根据结构化程度进行划分，可以方便对不同结构化程度的数据进行管理与应用。例如，结构化数据可使用传统的关系数据库进行管理与应用，允许使用结构化查询语言查询；非结构化数据和半结构化数据则不使用关系数据库进行管理，而是利用专门管理非结构化数据和半结构化数据的数据库系统，如 NoSQL 数据库等。

```xml
<?xml version="1.0" encoding="utf-8"?>
<person name="A" age="22" major="计算机科学" gender="male">
  <!--教育背景-->
  <educations>
    <item> 2014年 向日葵高中毕业</item>
    <item> 2018年 华东师范大学毕业</item>
  </educations>
  <!--外语情况-->
  <languages>
    <item> 英语</item>
    <item> 日语</item>
  </languages>
```

图 2–11　XML 文件保存的简历

噪声数据

噪声数据是在数据采集过程中产生的错误的、异常的、不完整的或无意义的数据。

产生噪声数据的原因很多，一般可以归结为两类：

（1）人为原因：人工记录数据或者将数据录入计算机时产生的错误。例如，人工录入公民身份号码时，可能会出现某个数字的误输入，导致身份号码无法与相应的人对应；注册某个网站时，当用户不希望提交真实的个人信息时，可能故意输入错误的数据，如故意将年龄填为 300 岁；问卷调查中，有些被调查者认为身高、体重是隐私，拒绝填写相关数据等；修改存储在多处的同一数据时，可能没有进行同步修改。

（2）设备原因：采集设备发生故障时产生的数据。例如，在利用指纹机采集指纹时，由于机器的故障没有采集到完整的指纹；在利用扫描仪扫描文稿时，出现文字乱码；由于存储数据的设备突然断电等情况导致数据缺损等。

噪声数据未必增加原本数据的存储空间，但它可能会影响数据分析的结果。尤其是一些对噪声敏感的算法，噪声数据可能会导致分析结果出现比较大的偏差。所以，在数据分析之前需要对数据进行必要的预处理，将噪声数据去除或修复，提高数据质量。

项目四

建立简易网上书店数据库

——了解关系数据库的建立

如今，无论是手机上的旅游 App、天气查询 App，还是互联网上的网上书店、网上商城等，各个信息系统都离不开数据的存储和访问。例如，在旅游 App 中订飞机票，离不开各航空公司的航班数据、会员数据以及机票订单数据等各种数据；又如，从网上书店信息系统（以下简称网上书店）中找书、购书，网上书店必须存储图书数据、会员数据以及图书订单数据等（图 2-12）。

那么，如何有效地使用和保存旅游 App、网上书店等应用中的数据呢？以上这些应用中产生的各种数据通常存储在数据库中。关系数据库是一种常用的数据库，它把数据组织为二维表的形式。利用成熟的数据库技术，人们可以创建关系数据库。例如，创建网上书店数据库，将图书、会员、订单等数据以二维表的形式组织并存储起来，以满足网上购书的需要。

图 2-12　网上书店

项目学习目标

本项目通过创建网上书店数据库，带领大家了解关系数据库设计与创建的一般过程，了解关系数据模型的基本概念，掌握设计简单关系数据库逻辑结构的方法，并使用数据库管理系统建立关系数据库。

完成本项目学习，须回答以下问题：

1. 什么是关系数据模型？
2. 如何描述实体集及其联系？
3. 设计、创建数据库的一般过程是怎样的？
4. 如何使用数据库管理系统建立数据库？

项目学习指引

1. 分析数据库设计需求

核心概念

数据库（database）是按照数据结构来组织、存储和管理数据的仓库，是长期储存在计算机内、有组织的、可共享的数据集合。

关系数据库（relational database）是建立在关系数据模型上的数据库，一个关系模型的逻辑结构是一张二维表，由行和列组成。这个二维表就叫关系。关系数据库用二维表来组织和存储数据。

设计一个信息系统一般要先做需求分析，也称软件需求分析、系统需求分析或需求分析工程等，它是开发人员经过深入细致的调研和分析，准确理解用户和项目的功能、性能、可靠性等具体要求，将用户的需求表述转化为完整的需求定义，从而确定系统必须做什么的过程。

要创建一个网上书店，首先需要对网上销售图书的整个业务活动作全面、详细的需求调查，并分析哪些业务计算机可以完成，哪些业务计算机不能完成。然后，确定网上书店这个信息系统能提供哪些功能和服务，并分析信息系统的数据要求，即系统需要输入什么数据，要得到什么结果，最后应输出什么数据，也就是确定网上书店数据库要存储什么数据，使得信息系统可以方便地处理这些数据，同时也要体现这些被存储的数据之间有什么关系。还要确定数据库的类型，一般都选择使用关系数据库。

参见 P41 知识链接“数据库设计的一般步骤”

就数据库设计需求分析而言，网上书店的开发人员通过与某网上书店需求方（不同部门的相关人员）进行充分沟通和交流，了解该网上书店的业务活动如下：网上书店是以网站作为图书交易平台。书店职员将图书的基本信息发布到网页中；会员通过网页浏览、查询图书，提交订单，实现图书的在线订购；订单提交后，书店职员会对订单进行处理，通知配送公司送书等。深入分析网上购书这个核心业务，其主要业务流程如图 2-13 所示：

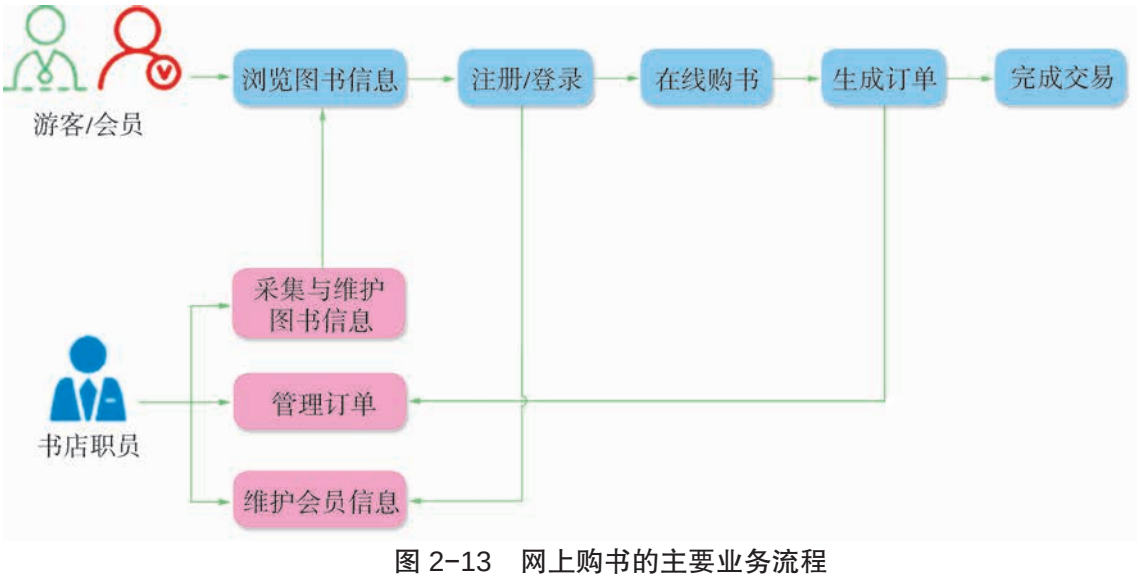


图 2-13　网上购书的主要业务流程

可以发现，该网上书店需要支持的用户有多种，主要为会员、书店职员。会员可以浏览和检索图书信息、从网上订书、在线支付、完成订单查询与修改等；书店职员可以维护和发布图书信息、处理订单等。

各类用户要在网上书店中完成以上业务活动，离不开数据库的支持。以会员用户为例，其主要业务活动所涉及的数据大致包括：

• 会员的注册数据。会员注册时填写登录名、真实姓名、登录密码、性别、出生日期、收货地址、电话等数据。数据经网上书店信息系统检查无误后，产生会员编号数据，并存入数据库。

• 图书数据。会员浏览或检索图书时，系统会提供图书编号、图书名称、类别、出版社、出版日期、作者、ISBN、定价、封面图等数据（数据由书店职员发布）。

• 会员的订单数据。会员将计划购买的图书放入购物车并填写购买数量、选择支付方式后，以上数据会存入数据库，生成订单。每张订单要记录订单编号、会员编号、图书编号、购买数量等数据。

活　动

4.1　选取一个知名的网上书店，体验购书流程，分析该网上书店为用户提供哪些功能和服务，列出会员在购书过程中涉及哪些数据。

2. 建立实体集和联系

在设计数据库的过程中，通常需要从创建数据库的业务需求中抽象出客观存在的实体集及描述每个实体集的特征，即属性；然后找出实体集之间的联系以及描述这种联系的属性。

设计网上书店的实体集和联系，首先要确定数据库中所有的实体集和属性，然后确定主关键字、定义实体集之间的联系。

核心概念

实体（entity）是现实世界的数据对象在信息世界的反映。在信息世界中，客观存在并且可以相互区别的人、事、物可以称为实体。实体集（entity set）是指具有相同类型及相同性质的实体集合。例如，学校里的每个学生都是一个实体，所有学生的集合可定义为学生实体集。

属性（attribute）是指实体集中每个实体都具有的特性描述。

（1）确定网上书店业务需求中的实体集及其属性

通过对网上书店业务需求的分析，可以知道其中涉及的实体集有职员、会员、图书、订单等。例如，对于职员，它的属性有编号、登录名、密码、真实姓名、电话、电子邮箱等。

（2）确定主关键字

对于网上书店来说，会员注册时常常会出现相同的姓名，职员中也可能会出现相同的姓名。为了能唯一标识每个实体，可以设置部分属性为主关键字。例如，对于职员来说，每一个职员会有一个编号，不会与其他职员重复，因此，编号可以作为职员这个实体集的主关键字。

小贴士

主关键字（primary key）又称为主键，是实体的一个或多个属性，它用于唯一地标识某一个实体的值。

思考与讨论

1. 图书名称属性能否作为图书实体集的主关键字？为什么？

2. 职员实体集有一个属性为登录名，在注册时系统会要求登录名不能重复。那么，登录名能否作为职员实体集的主关键字？为什么？

（3）定义联系

确定实体集后，需要进一步发现实体集之间的联系。

例如，会员购书时，一张订单中可以有很多种图书，而一种图书也可以在多张订单中出现。因此，订单实体集和图书实体集之间存在多对多的联系；又如，一个会员可以在网上书店里有多张订单，而一张订单只能属于一个会员。因此，会员实体集和订单实体集之间存在一对多的联系。

核心概念

实体集之间的相互关联称作联系（relation）。实体集之间的联系有一对一、一对多、多对多三种类型。

思考与讨论

会员实体集与图书实体集之间存在什么联系？

活　动

4.2 确定本网上书店中的实体集，并确定各实体集之间的联系。

3. 建立数据模型

将现实世界中的事物及其联系转换成实体集和联系后，还需要将其转换成计算机世界中的数据模型。数据模型中应用最广泛的是关系数据模型。设计关系数据模型的过程是将实体集、属性和联系转换成二维表的过程。

参见 P40 知识链接“数据模型及关系数据模型”

（1）将实体集及属性转换成二维表

将网上书店的所有实体集转换成表，实体集的属性转换成字段。例如，对于图书实体集，它的属性有图书编号、图书名称、类别、出版社等，其转换成的二维表如图 2-14 所示。

属性名转换成字段名　　实体集名转换成表名

图书表

图书编号	图书名称	类别	出版社	出版日期	作者	ISBN	定价	简介	封面图	库存数	书评

图 2-14　二维表示例

其中，实体集名“图书”转换成表名“图书”，属性名转换成字段名。图书表可以存储书店中所有图书的数据，每一行存储一条图书数据。这一条条数据被称为记录，它是关于书店中每一本图书的描述。

（2）优化表

有时，设计出的数据模型可能会存在数据冗余、删除异常和修改困难等问题。例如，分析以上图书表，可以发现以下问题：

① 在图书表中，每个会员的书评占用一条记录，同一本书可能有多条评论，因此，除书评外的字段会大量重复出现在图书表中，造成数据冗余。

② 当某种图书类别中只有一本图书时，如果删除该图书的记录，对应的图书类别也随着一同被删除，出现删除异常。

针对上述问题，需要对表进行优化，常用的优化方法是对表进行拆分，消除表中存在的不合理的地方。

例如，解决图书表中的书评数据冗余问题，可以将原图书表拆分成以下的图书表和图书评价表，如图 2-15 所示。

核心概念

数据模型（data model）是数据特征的抽象。数据模型一般包括三个部分：数据结构、数据操作、数据约束。关系数据模型的数据结构是关系（二维表），数据操作包括插入、删除、查询、更新等，数据约束包括实体完整性、参照完整性以及用户自定义完整性约束。

小贴士

字段（field）是指在关系数据库的表中的每一列，每个字段表示实体的一个属性。每个字段都有一个唯一的字段名。

记录（record）是指在关系数据库的表中的每一行，它记录了关于一个实体的属性值。

小贴士

数据冗余：同一个数据大量重复地出现在数据库的表中。

删除异常：删除表中某一条记录，需要保留的数据也会随之消失。

修改困难：欲修改表中的一个数据，需要连同修改多处（字段、记录）数据。

图书表

图书编号	图书名称	类别	出版社	出版日期	作者	ISBN	定价	简介	封面图	库存数

图书评价表

图书编号	评价编号	书评

图 2-15　图书表和图书评价表

小贴士

正确的数据库设计不是一蹴而就的，而是一个循序渐进和反复设计的过程。

思考与讨论

为解决图书表的图书类别删除异常问题，应该如何优化图书表？

（3）确定主关键字

每一个实体集都有一个主关键字，当实体集转换成表，属性转换成字段时，同样需要为表确定主关键字。例如，图书实体集的主关键字是“图书编号”，那么“图书编号”字段就可以作为图书表的主关键字。又如，图书评价表的主关键字为“图书编号”和“评价编号”的组合。在表示时，在表的主关键字下标注横线，如图 2-15 所示。

小贴士

这里的主关键字是在关系数据库的表中唯一确定每个记录的一个字段或一组字段。

思考与讨论

优化前的图书表，“图书编号”字段能作为主关键字吗？为什么？

小贴士

如果表 T1 和表 T2 有相同的属性 A，属性 A 在表 T1 是主关键字，那么属性 A 被称为表 T2 的外关键字，又称为外键。外键表示了两个表之间的相关联系。以另一个表的外键作主关键字的表被称为主表，具有此外键的表被称为主表的从表。

（4）建立表间联系

通过表与表之间的相同属性，建立表间联系。例如，在如图 2-16 所示的图书表和图书类别表这两表中，通过图书表的“类别编号”和图书类别表的“类别编号”这两个属性相同的字段，建立起了一对多的联系（一本图书对应一个类别，一个类别可以有多本不同的图书）。图书表中的“类别编号”是图书表的外关键字，图书类别表是主表，图书表是图书类别表的从表。

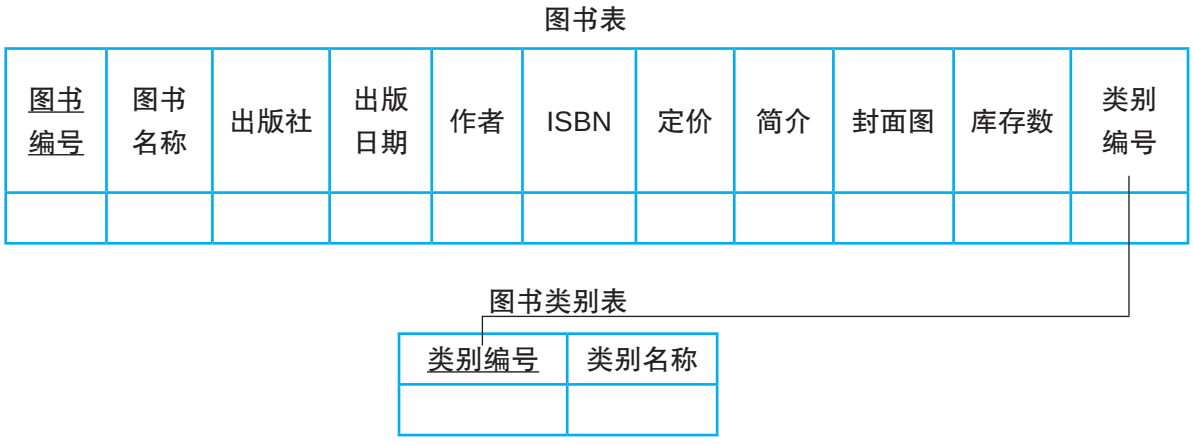


图 2-16　表间联系示意

活　动

4.3 根据已经确定的网上书店的实体集和联系，建立数据模型。

4. 创建数据库

创建数据库需要使用数据库管理系统。数据库管理系统有很多种，这里以 MySQL 数据库管理系统为例，创建网上书店数据库。

① 安装并启动 MySQL 数据库管理系统。

② 创建网上书店数据库。

在 MySQL 中输入如下命令，按回车键，创建一个名为“netbook”的空数据库。

```
CREATE DATABASE netbook;
```

③ 创建表结构。

表是数据库最重要的组成部分之一，是数据库真正存储数据的地方。一个数据库管理系统中可以存在多个数据库，所以在创建数据库的表时，首先要确定将表建立在哪个数据库中。

例如，在 MySQL 中输入如下命令，按回车键，打开“netbook”数据库。

```
USE netbook;
```

核心概念

数据库管理系统（Database Management System，DBMS）是一种开发、使用、维护数据库的管理软件。它为用户提供了各种对数据进行操纵的工具，帮助业务管理者对数据进行有效的组织、存储和管理。

参见 P42 知识链接“数据库管理系统”

数字化学习

利用配套资源，学习安装并启动 MySQL 数据库管理系统。

打开数据库后，便可在数据库中创建表。在创建表时要为每个字段确定数据类型。

例如，输入如下命令，创建职员表。

```
CREATE TABLE 职员 (
编号 tinyint NOT NULL,
登录名 varchar(20) UNIQUE,
密码 varchar(32) NULL,
真实姓名 varchar(20) NULL,
电话 varchar(11) NULL,
电子邮箱 varchar(30) NULL,
PRIMARY KEY (编号)
);
```

小贴士

为了规范数据的使用和存储，在数据库中常使用数据类型来约束数据。数据类型规定了对数据的允许取值和取值范围的说明，它是数据的基本属性。

因为表的每一个字段都只能存放单一数据类型的数据，因此在创建表的结构时，可以根据需要为字段设置数据类型。数据类型定义合适，可以正确表达数据，定义不合适则会造成数据丢失或存储空间的浪费。

思考与讨论

1. 以上语句中 NOT NULL、UNIQUE 的作用分别是什么？

2. MySQL 中提供了哪几种常用的数据类型？和 Python 中的数据类型比，有哪些不同？

类似的，可以用同样的方式创建数据库中的其余表，完成网上书店数据库的创建。

活　动

4.4 根据建立的数据模型，利用 MySQL 数据库管理系统，创建网上书店数据库。

知识链接

数据模型及关系数据模型

现实世界非常复杂，计算机不可能直接处理其中的具体事物，因此必须使用相应的手段将具体事物转换成计算机能够处理的数据。数据模型的主要任务就是将现实世界中的具

体事物转换成计算机能识别和处理的数据。数据模型所描述的内容包括数据结构、数据操作和数据约束。创建数据模型的具体方法是：把现实世界中存在的客观对象抽象为某一种不依赖于具体计算机系统的数据结构，然后将其转换成计算机系统所支持的数据模型。数据模型是直接面向计算机系统（即数据库）的数据的逻辑结构。目前成熟地应用在数据库技术中的数据模型有三种：层次数据模型、网状数据模型和关系数据模型。

关系数据模型是应用最广泛的一种数据模型，它由许多以某种条件联系在一起的二维表组成。在关系数据模型中用二维表描述实体集、属性以及实体集之间的联系。

关系数据模型由关系数据结构、关系操作集合和关系完整性约束三大要素组成。

（1）关系数据结构：关系模型把数据库表示为关系的集合（关系模型中数据的逻辑结构是一张二维表）。

（2）关系操作集合：关系模型中常用的关系操作包括查询操作和插入、删除、更新操作两大部分，其中查询操作可以分为选择、投影、连接等。

（3）关系完整性约束：数据库的数据完整性是指数据库中数据的正确性、相容性和一致性。关系完整性约束包括三方面内容：① 实体完整性，即主关键字的主属性不能为空，不能重复，如会员编号作为会员表的主关键字，不能为空，不能重复；② 参照完整性，即外键取值只能取被参照关系（即主表）中已经存在的主关键字值或者空值，如图书表某个图书记录中的类别编号只能是图书类别表中已经存在的类别编号，或者不指定类别编号，即为空值，但是不能是一个不存在的非法的编号；③ 用户自定义完整性，需要根据用户的实际需求定义，如性别只能是“男”或“女”。

数据库设计的一般步骤

数据库是一个长期存储在计算机内、有组织的、可共享的、可统一管理的数据集合。数据库设计要经历一个从现实世界到信息世界再到数据世界的逐步抽象过程。数据库的设计需要经历以下几个阶段。

1. 需求分析

需求分析是指针对业务所处的现实世界进行调查与分析。需求分析一般是对整个数据库应用系统所要处理的对象进行全面的了解，明确业务管理的目的。需求分析的主要任务是分析并归纳应用系统应该具有的功能要求和对数据的处理、存储、输入与输出的要求。

设计者只有熟悉相关的业务，才能设计出符合实际需求的数据库。在设计数据库之前，设计者一方面要深入实地开展业务调查，采集相关业务数据，了解行业业务现状和具体业务流程；另一方面要详细分析采集到的各种数据，明确用户的各种要求，总结归纳出需要数据库管理的数据信息和管理信息。

2. 建立实体集和联系

数据库设计者要对原始数据进行综合，抽象出所要研究的数据，将现实世界中的事物及其联系转换成信息世界中的实体集及实体集间的联系。

建立实体集和联系的一般过程是：

（1）确定实体集和属性。现实世界中，一组具有某些共同特性和行为的对象就可以抽

象为一个实体集。例如在学校的选课系统中，张珊、李偲、王於等学生对象可以抽象为学生实体集。对象的成分和特性可以抽象为该实体集的属性。例如学生的学号、姓名、班级、选修的课程等可以抽象为学生实体集的属性。

（2）确定主关键字。在实体集的属性中往往可以找到唯一标识该实体集的属性，那么这个属性被称为主关键字，主关键字可以由一个或几个属性组成。

（3）定义联系。实体集之间的联系是现实世界中客观事物之间的固有关系的反映，分为一对一、一对多、多对多三种类型。例如在网上书店中，一个会员可以购买多种图书，一种图书也可以被多个会员购买，因此，图书实体集与会员实体集之间存在多对多的联系。

3. 建立数据模型

将实体集和联系转换成数据世界中的数据及其联系，并用数据模型进行描述。对关系数据模型来说，就是要定义表及表间联系等。

（1）定义表。将实体集与属性转换成表与字段，其中实体集名转换成表名，属性名转换成字段名。

（2）确定主关键字。在表的多个字段中，能唯一确定每条记录的一个字段或一组字段，即为表的主关键字。

（3）建立表间联系。将实体集间的联系转换成表间的关联关系，并确定外关键字。

数据库管理系统

数据库管理系统是一组实现对存储于计算机存储器中的数据执行统一管理操作，如读出、写入、查询、修改、删除等操作的程序的集合，简称 DBMS，通过它数据库开发和管理人员可以和数据库进行交互。

数据库管理系统的目标是让使用者能够更方便、更有效、更可靠地建立数据库和使用数据库中的信息资源。数据库管理系统一般不是设计成直接面向用户的形式，它是为使用该数据库的各种应用程序提供数据管理服务。例如，要实现网上书店的功能，还需要网上书店信息系统来直接面向用户。

目前，常见的数据库管理系统有 Access、SQL Server、Oracle、MySQL 等。其中，MySQL 是一种开放源代码的关系数据库管理系统，因其体积小、速度快、总体拥有成本低，受到数据库开发者的热捧，成为当前非常流行的数据库管理系统之一。

项目五

管理网上书店数据库

——使用结构化查询语言

网上书店数据库建立并投入使用后，新的数据会不断地被添加或更新到数据库中。例如，会员注册成功后，需要将新会员的数据添加到会员表中；会员提交订单并完成支付后，需要同步更改图书的库存数据；书店新购进一批书后，需要将新书数据批量导入图书表中。

除了数据存储和数据更新外，用户还会做一些数据查询工作（图 2-17）。例如，会员购书时，可能会查找书名含有“大数据”关键词的图书，或将少儿类图书按销量从大到小排序。又如，书店职员可能会查找库存量小于警戒数的图书，并分别统计每种书前几个月的销量，以便考虑哪些图书需要补货，该补多少册。

对数据库的管理与维护，既可以利用数据库管理系统的结构化查询语言编写命令，直接操作数据库，也可以利用信息系统中由开发人员事先编写好的程序去操作数据库。在本项目中，我们仅涉及使用结构化查询语言直接操作数据库。

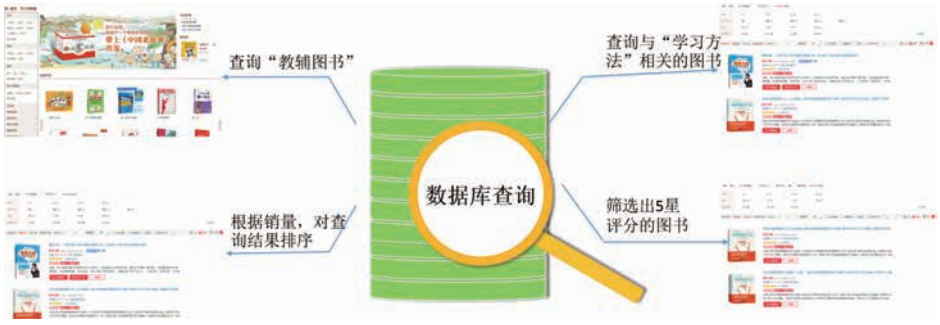


图 2-17　查询网上书店数据库

项目学习目标

在本项目中，我们将使用结构化查询语言对网上书店数据库进行添加、查询、更新和删除等操作，了解数据库的基本数据查询方法。

完成本项目学习，须回答以下问题：

1. 什么是结构化查询语言？

2. 如何使用结构化查询语言进行数据添加、查询、更新和删除等操作？

项目学习指引

1. 添加数据

网上书店的运营过程中会产生大量数据，需要频繁地对网上书店数据库进行操作。通过在数据库管理系统中使用结构化查询语言编写 SQL 语句，可以方便地对网上书店数据库进行添加、查询、更新和删除数据等操作。

项目四创建的网上书店数据库还没有存储数据，因此，可以向数据库的各表中添加数据。少量数据可以采用逐条添加的方法，较多数据可以采用批量添加的方法。

以职员表为例，要将如图 2-18 所示的数据添加至职员表中，可以采取以下两种方法。

编号	登录名	密码	真实姓名	电话	电子邮箱
1	zhangliang	12ab56	张亮	8658051	zl@163.com
2	wangwei	Ww#123	王伟	8605202	ww@163.com

图 2-18　职员表新增数据

方法一：逐条添加。例如，在 MySQL 中输入如下 SQL 语句，按回车键，向职员表中添加一条记录。类似的，再输入下一条语句，添加下一条记录。

```
INSERT INTO 职员 VALUES(1,"zhangliang","12ab56"," 张亮 ",
"8658051","zl@163.com");
```

方法二：批量添加。例如，在 MySQL 中输入如下 SQL 语句，按回车键，向职员表中添加两条记录。

```
INSERT 职员 VALUES(1, "zhangliang","12ab56"," 张亮 ",
"8658051", "zl@163.com"), (2,"wangwei","Ww#123"," 王伟 ",
"8605202","ww@163.com");
```

思考与讨论

1. 如果数据量高达几百条，那么批量添加的方式就不那么高效，有无其他方式可以高效地将数据存储到数据库中？具体方式是怎样的？

2. 用户浏览网上书店的书目或下单时，会改变网上书店数据库中各表的记录。但对于用户来说，他们并不直接操作数据库，那么，数据库中的记录是如何被改变的呢？

核心概念

结构化查询语言(Structured Query Language，简称 SQL)，是一种数据库查询语言，用于存取数据以及查询、更新和管理关系数据库系统。

参见 P47 知识链接“结构化查询语言”

小贴士

添加数据语句的格式是：
INSERT [INTO]〈表名〉
[(字段名 1,…)]
VALUES (值 1,…);
[ ] 中的内容为可选项。

活　动

5.1　网上书店最近新采购了一批图书，请将这些图书的数据批量添加至数据库的图书表中（新购图书数据见配套资源中的“新购图书表 .xls”文件）。

2. 查询数据

数据添加完成后，可以使用 SQL 语句查询数据。

① 查看职员表中的数据，查询语句如下：

```
SELECT * FROM 职员；
```

② 如果需要查询职员表中登录名与其真实姓名的对应关系，可以在 SELECT 后指明需要查询的字段名，即投影操作，该查询语句如下：

```
SELECT 登录名，真实姓名 FROM 职员；
```

③ 如果要根据职员的真实姓名对查询结果进行排序，可以在 SELECT 语句的 ORDER BY 后加上排序的依据，即排序操作，该查询语句如下：

```
SELECT 登录名，真实姓名 FROM 职员 ORDER BY 真实姓名；
```

④ 如果只查询某个职员的部分数据时，可以在 SELECT 后指明需要查询的内容，在 WHERE 后指明查询条件，即选择操作。以查询真实姓名为“王伟”的职员为例，查询语句如下：

```
SELECT 登录名，真实姓名，电话，电子邮箱 FROM 职员
WHERE 真实姓名 =" 王伟 ";
```

小贴士

查询数据语句的功能很强大，其语句简单的格式如下：

SELECT < 字段名 1, …> 或 * FROM < 表名 >

[WHERE< 条件表达式 >];

* 代表查询所有的字段。

小贴士

数据库的查询通常由选择、投影、连接、聚集、排序等操作构成。投影操作的目的是对查询结果的属性进行筛选。当 SELECT 语句仅指定被查询表中的部分属性时，数据库管理系统会执行投影操作。

排序操作的目的是将查询结果按照某个属性从大到小或从小到大排列。从小到大称为升序，从大到小称为降序。

选择操作的目的是在一张表中筛选出符合某些条件的对象。在 SQL 查询语句中，一旦出现了由 WHERE 指定的选择条件，则是在告诉数据库管理系统需要执行选择操作。

⑤如果要统计姓李的职员有几位，可以使用 COUNT 函数实现统计查询，查询语句如下：

```
SELECT COUNT( 真实姓名 ) FROM 职员 WHERE 真实姓名 LIKE " 李 %";
```

小贴士

统计查询是 SQL 查询的一项强大功能。通过使用函数（如 COUNT、SUM、AVG 等），可以在查询结果集上进行统计计算（如统计总数、求和、求平均值等），获得统计结果。

思考与讨论

1. “%”是一个通配符，用于表示任意长度的一段文字。如果想查找书名中含有“大数据”一词的书，条件表达式是怎样的？

2. 以上查询都是针对一张表的查询。能否从两张表或更多表中查询出需要的信息？具体查询方法是怎样的？

活　动

5.2　使用查询语句，随机查询已经添加至图书表中的图书数据，核对添加的数据是否有误。再完成以下数据的查询工作。

（1）查找作家莫言的所有图书的书名和出版日期。

（2）查找作家莫言在 2009 年出版的图书的所有信息，并按出版时间排序。

（3）查找书名中包含“哲学”的图书。

（4）统计书名中包含“哲学”的图书的数量。

3. 更新数据

已经存在于数据库中的数据，可以使用 UPDATE 语句对满足条件的记录进行更新。以更新密码为例，登录名在职员表中唯一，因此可以将职员表中登录名为“liming”的密码更改为“LiMing@456”，更新语句如下：

```
UPDATE 职员 SET 密码 ="LiMing@456"
WHERE 登录名 ="liming";
```

小贴士

更新数据语句的格式是：

UPDATE ＜表名＞

SET＜字段名 1＞=＜值 1＞

[,＜字段名 2＞=＜值 2＞,…]

[WHERE＜条件表达式＞];

活　动

5.3 完成以下的数据更新工作。

（1）类别编号为“1”的图书销售量不错，故网上书店对类别编号为“1”的图书均分别补了 100 本，请更新相关图书的库存量。

（2）某职员发现职员表数据有误，请根据配套资源中的“职员.xls”更新数据库中的职员数据。

4. 删除数据

当某个职员（如王伟）离职后，需要注销其个人账号，就要从网上书店数据库的职员表中删除该职员的数据。删除语句如下：

```
DELETE FROM 职员 WHERE 登录名 ="wangwei";
```

思考与讨论

在具体操作中，UPDATE 语句与 DELETE 语句对数据有何影响？

小贴士

删除数据语句的格式是：

DELETE FROM< 表名 >

[WHERE< 条件表达式 >];

使用 DELETE 语句，一次可以删除一条记录，也可以删除多条记录。

活　动

5.4 有些图书网上书店不再出售，需要清除数据（假设这些图书均未产生订单），请从数据库中删除满足下列条件的数据。

（1）库存量为 0 的图书。

（2）出版日期在 2000 年前的图书。

知识链接

结构化查询语言

结构化查询语言简称 SQL 语言。SQL 语言作为一种十分重要的标准数据库语言，其功能强大，简单易学，被广大数据库开发人员普遍使用。关系数据库系统大多采用 SQL 语言

作为共同的数据库操作语言，尽管各个数据库管理系统使用的 SQL 版本不同，但都具有标准 SQL（ANSI SQL）的功能，包括数据定义、数据查询、数据操作和数据控制四个方面。在本项目中，用 SQL 语言进行了数据查询、添加、更新和删除等基本操作。下面以 MySQL 提供的 SQL 语言为例，介绍 SQL 语言的常用语句。

1. 数据查询

数据查询是数据库的核心操作，其功能是根据用户的需要以一种可读的方式从数据库中提取所需数据。SQL 的查询语句只有一条 SELECT 语句，但是它几乎能完成各种查询任务，如选择查询、投影查询、多表查询、数据统计、结果排序等。在 MySQL 中，所有的查询都是由 SELECT 语句实现的。MySQL 中 SQL 查询语句的基本结构如下：

```
SELECT * 或 < 字段名 1,…> 或 < 表达式 1,…> 或 < 函数 1,…>
FROM < 表名 > 或 < 视图 > 或 < 嵌套 SELECT>
[WHERE< 条件表达式 >]
[GROUP BY< 分组字段 >]
[HAVING< 分组筛选条件 >]
[ORDER BY< 排序字段 排序选项 >];
```

SELECT 语句的功能是从 FROM 子句指定的表中，选择满足条件（由 WHERE 子句指定）的数据，并对它们进行分组、统计和排序，形成查询结果集。

其中，< 字段名 1,…> 指要查询的字段，若查询所有字段，则 < 字段名 1,…> 可用“*”代替。< 表名 > 指要查询的表，WHERE 子句中的 < 条件表达式 > 用来限定查询的条件。GROUP BY 子句的作用是将结果按 < 分组字段 > 的值进行分组，即将字段值相等的记录分为一组，实现数据的分组统计。HAVING 子句用来限定分组必须满足的条件。ORDER BY 子句用来对结果按 < 排序字段 > 的值进行排序，默认是按照该字段值的升序排序，如果需要按照该字段值的降序排序，则“排序选项”可以指定关键字 DESC。

2. 数据操作

数据操作语句包括 INSERT、UPDATE 和 DELETE 三种基本形式，分别适用于实现添加、更新和删除数据的操作。

（1）INSERT 语句

INSERT 语句主要用来向表中添加数据，使用 INSERT 语句可以一次向表中添加一条记录，也可以一次向表中添加多条记录。INSERT 语句的基本格式如下：

```
INSERT [ INTO] < 表名 > [ ( 字段名 1,…) ] VALUES( 值 1,…);
```

其中，< 表名 > 指新记录将要插入的表；VALUES(值 1,…) 用来指明新添加记录的各字段的值。使用此格式向表中添加记录时，一定要以表中的字段为准，字段名和值必须一一对应。

（2）UPDATE 语句

UPDATE 语句主要用于修改表中字段的值，可以一次修改一个字段的数据，也可以同时修改多个字段的数据。UPDATE 语句的基本格式如下：

```
UPDATE  < 表名 >  SET  < 字段名 1>=< 值 1> [,< 字段名 2>=< 值 2 >,…]
[WHERE< 条件表达式 >];
```

其中，< 表名 > 用于指明要修改的表，SET 子句用于指明要修改的 < 字段名 > 及具体的值，WHERE 子句用于限定满足修改条件的记录。UPDATE 语句的功能是对表中满足 WHERE 子句指定条件的记录进行修改，SET 子句指定新的值取代相应字段原来的值。

（3）DELETE 语句

DELETE 语句也叫数据删除语句，其基本格式如下：

```
DELETE  FROM < 表名 >  [WHERE< 条件表达式 >];
```

其中，< 表名 > 用于指明要删除的数据所在的表，WHERE 子句表示要删除的数据需要满足的条件。如果不写 WHERE 子句则表明清空当前表，即表的结构还在，但是里面包含的所有数据都被删除了。

单元挑战　建立年级作业评价数据库

一、项目任务

任课教师会对学生每次作业给予不同的评分与评价。如果不存储与管理这些数据，教师对学生每次作业的评分与评价将会随着作业本的丢弃而遗失。在学习了数据管理相关内容后，你能否创建一个年级作业评价数据库管理这些数据？

尝试以小组为单位，分工合作，设计并创建年级作业评价数据库；在数据库创建完成后，以小组互评的形式评价各组的数据库。

二、项目指引

1. 在班级里开展调查，分析年级作业评价数据库有哪些应用和需求，有哪些业务流程。

2. 根据业务需求，确定年级作业评价数据库的实体集和属性，并定义主关键字及实体集间的联系。

3. 建立年级作业评价数据库的关系数据模型，确定表、主关键字和表间联系。

4. 选择合适的数据库管理系统，根据设计的表，创建年级作业评价数据库，并录入一些数据至数据库中。

表名	内容					
	字段名					
	数据类型					
	字段大小					
	字段名					
	数据类型					
	字段大小					

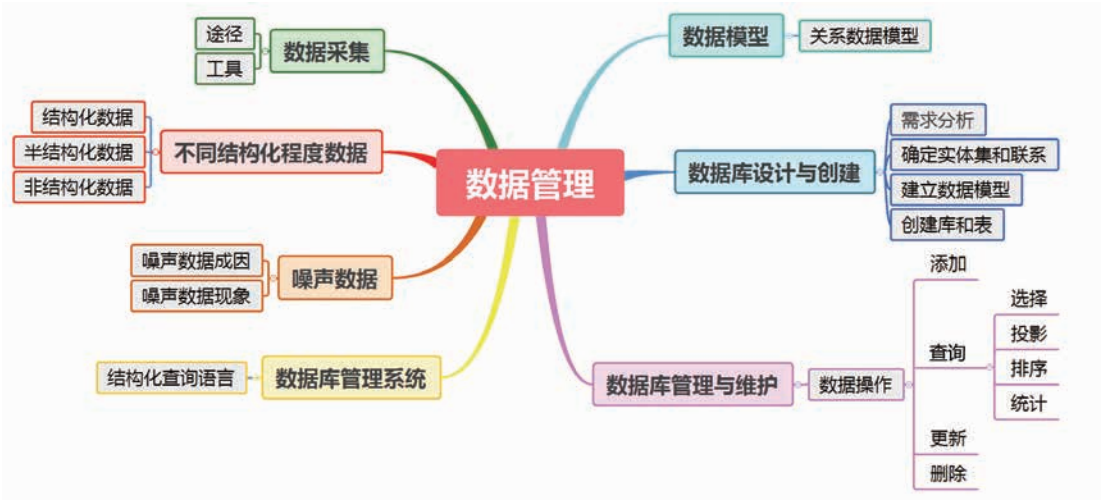
5. 请其他小组提出一些需求或数据查询任务，使用结构化查询语言查询数据，根据任务完成情况进行互评。

三、交流评价与反思

在班级里展示并交流小组创建数据库过程中所完成的需求分析、确定实体集和联系、建立数据模型等工作，以及根据其他小组所提需求开展的数据查询工作。对本组的工作进行自评并对其他小组进行评价。

单元小结

一、主要内容梳理



二、单元练习

1. 网上书店其实是一个比较庞大的信息系统，本单元中仅简单介绍了其中的一部分功能。尝试完成以下任务。

（1）如果需求分析中增加了职员管理、订单配送或其他功能，应该考虑设计哪些表？选择一种新增的功能重新设计数据库。

（2）确定新增的各表的主关键字及表间联系。

2. 在实际工作中，书店职员或数据库管理员对网上书店的操作是非常多样且复杂的。尝试通过上网学习，回答以下问题。

（1）当职员离开工作岗位时，应该对数据库表进行什么操作？

（2）在系统中对于数据的查询是否要限制权限？为什么？

3. 数据库系统的数据安全设计非常重要，尝试回答以下问题。

（1）对于已经产生订单的某本图书，能否在图书表中直接删除该图书数据？为什么？

（2）数据库表中的密码项（如会员的登录密码）是否要通过加密运算后再保存到数据库表中？为什么？

三、单元评价

评价内容	达成情况
了解数据采集的途径与工具（A、I、R）	
能利用适当的工具对数据进行采集和分类（A、T、R）	
认识噪声数据的现象和成因（A、I）	
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（A、I）	
了解关系数据模型的基本概念（A、T）	
掌握设计简单数据库的逻辑结构的方法（A、T）	
能使用数据库管理系统建立关系数据库（A、T）	
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（A、T）	

说明：A—信息意识，T—计算思维，I—数字化学习与创新，R—信息社会责任

第三单元

数据分析

互联网和数据库等技术的高速发展，让获取和存储数据变得越来越容易，导致数据存储量呈现爆炸式增长。如何充分利用这些数据，从中提取有用的信息，进而让它们发挥出更大的价值，正是数据分析工作的重要意义之所在。交通管理部门想要了解道路交通拥堵状况，企业管理者想要提高经营决策的科学性，网上商城想要对用户进行精准的商品推荐……所有的管理和经营活动都需要借助数据分析得出的结论来作出正确的决策。

在本单元中，我们将一起探究身边的数据分析工作，了解数据分析的基本方法，学会选择适当的分析工具，经历分析、呈现、解释数据的过程，认识数据挖掘的重要意义。

单元挑战

分析在线社交平台用户情况

学习目标

- 了解数据分析的过程、目的和作用。
- 根据任务需求，正确选用适当的数据分析工具，分析、呈现并解释数据。
- 掌握常用的四种数据分析方法（对比分析法、分组分析法、平均分析法和相关分析法）。
- 运用数字化学习方式，了解数据管理与分析技术的新发展。
- 了解数据挖掘，认识数据挖掘对信息社会问题解决和科学决策的重要意义。

项目六

分析城市交通拥堵状况

——了解常用的数据分析方法

城市交通是城市经济发展和现代化建设的纽带。作为城市发展的主要动力，交通对生产要素的流动、城镇体系的发展有着决定性的影响。近年来，随着中国经济的快速增长和城市化进程的不断加快，城市人口和车辆急剧增长，城市交通拥堵等问题不断显现。

利用车辆定位数据、行驶轨迹数据等，可以对城市的交通状况进行粗略分析，了解城市交通拥堵的现状（图 3-1）。通过进一步分析，找出造成交通拥堵的相关因素，可以为城市交通发展规划与决策提供有力的数据支撑。

图 3-1　某市实时路况信息

项目学习目标

在本项目中，我们将通过分析城市交通拥堵状况及其相关因素，掌握主要的数据分析方法，包括平均分析法、分组分析法、对比分析法和相关分析法。

完成本项目学习，须回答以下问题：

1. 常用的数据分析方法有哪些？
2. 不同数据分析方法的区别是什么？
3. 在何种情况下可以采用何种方法进行数据分析？

项目学习指引

1. 了解城市道路交通拥堵状况

随着信息技术与互联网技术的发展，很多车辆都装备了定位导航设备，这为分析城市道路交通拥堵状况提供了新的手段。对大量车辆不断产生的定位数据进行数据分析，可以得到路段、区域、城市乃至更大范围的实时和历史路况数据信息。这些信息可以帮助人们了解并掌握整个路网的交通拥堵状况和通行效率，发现城市道路交通存在的问题，为交通规划提供良好的决策支持。

路段上的车辆行驶速度是城市道路交通状况的最直接反映。通过计算同一路段上多辆车的平均行驶速度，可以了解该路段的交通拥堵状况。再将各路段的交通拥堵状况分组分析，就可以得到城市整体交通状况。若将其与历史数据进行比较，并结合对交通状况有影响的其他因素的相关分析，还可以发现交通拥堵状况出现的规律。

（1）分析车辆行驶情况

装备有定位导航设备的车辆在行驶过程中会产生一系列定位数据。定位数据通常包括定位设备的 ID 号、定位时间戳、定位点经纬度等关键信息，如图 3-2 所示。

设备ID	时间戳	经度	纬度
pjyju7da2_8ce5eDprhrlzgh9e56da6n	1477969150	104.07513	30.72702
pjyju7da2_8ce5eDprhrlzgh9e56da6n	1477969154	104.07504	30.72672
pjyju7da2_8ce5eDprhrlzgh9e56da6n	1477969156	104.07497	30.72630
pjyju7da2_8ce5eDprhrlzgh9e56da6n	1477969159	104.07497	30.72582
pjyju7da2_8ce5eDprhrlzgh9e56da6n	1477969162	104.07496	30.72544
pjyju7da2_8ce5eDprhrlzgh9e56da6n	1477969168	104.07489	30.72487
pjyju7da2_8ce5eDprhrlzgh9e56da6n	1477969171	104.07476	30.72456
pjyju7da2_8ce5eDprhrlzgh9e56da6n	1477969174	104.07457	30.72434
pjyju7da2_8ce5eDprhrlzgh9e56da6n	1477969177	104.07434	30.72406
pjyju7da2_8ce5eDprhrlzgh9e56da6n	1477969180	104.07422	30.72379

图 3-2　某车辆定位数据

单条定位数据记录车辆在某一时刻的坐标位置，一系列连续的数据记录形成车辆行驶的轨迹数据。通过计算某两个定位点之间的距离和时间差的比值，可以求得车辆在这两个定位点之间的平均行驶速度。定位设备产生定位数据的时间间隔通常为 5~10 秒，因此两个连续定位点间的平均行驶速

核心概念

数据分析（data analysis）是指用适当的统计分析方法对数据进行分析，从中提取有用信息并形成结论的过程。

小贴士

时间戳用于记录产生定位数据的时间。各类计算机语言都有处理时间戳的函数，直接调用就能将它转换成需要的时间格式。

小贴士

轨迹数据体现了个体的活动规律，蕴含个体的行为模式、活动方式、活动范围及社会网络关系等特征，在城市规划、城市交通、公共安全等领域都能得到广泛的应用。

度基本上可反映车辆的实时速度。根据定位点的经纬度信息，调用地图软件的 regeocode 回调函数，反查得出定位点在地图上所对应的路段信息（图 3-3），由此可推断出车辆在什么时间以何种速度通过了哪个路段。根据各个定位点在时间上的顺序关系，可以确定车辆的行驶方向。将车辆行驶轨迹数据与地图数据结合起来，还可以知道车辆在某一具体路段的行驶方向、行驶速度等信息。

参见 P62 知识链接“数据分析”

图 3-3　某车辆轨迹示意图

通常情况下，车辆在交通状况良好的路段上行驶速度较快。但是由于车辆状况和个体驾驶习惯等因素的影响，不同车辆在同一时间段、同一路段上所表现出的行驶速度有可能存在较大差异，仅依据单辆车的行驶速度来评价交通状况会出现较大的偏差。因此，要了解某一时间段、某一路段、某通行方向的交通状况，需要采用平均分析法，通过求得某一路段上多辆车的平均行驶速度，来得到反映路段通行速度的更准确的结果。根据多辆车的行驶轨迹数据，可以计算得到如表 3-1 所示的多个数据组。

将表中同一时间段、同一路段、相同通行方向的数据组提取出来，用平均分析法进行综合分析，即可求得反映路段交通状况的平均行驶速度。例如，通过调用编程语言如 Python 的 mean 函数等方式，可以计算得到 2018 年 4 月 25 日 8:00~8:05 时间段某市延安高架由西向东方向车辆的平均行驶速度为 40.535km/h。

（2）分析道路拥堵变化情况

道路的交通拥堵状况可以通过道路交通状态指数即 TSI 来描述，但各地区的 TSI 计算方法不尽相同。譬如上海市的

核心概念

平均分析法（average analysis）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。

小贴士

道路交通状态指数（Traffic State Index，TSI）主要用于中观、宏观区域的路网拥堵评价分析，一般分为畅通、基本畅通、轻度拥堵、中度拥堵、严重拥堵五个级别。

表 3-1　车辆行驶轨迹数据

车辆 ID	时间段	路段	通行方向	行驶速度
1	2018-04-25 8:00~8:05	延安高架	由西向东	35.9km/h
1	2018-04-25 8:05~8:10	延安高架	由西向东	32.1km/h
1	2018-04-25 8:10~8:15	南北高架	由南向北	25.1km/h
…				
2	2018-04-25 8:00~8:05	瑞金一路	由北向南	27.1km/h
2	2018-04-25 8:05~8:10	瑞金二路	由北向南	27.5km/h
…				
1051	2018-04-25 9:15~9:20	延长路	由北向南	23.6km/h
1051	2018-04-25 9:20~9:25	汶水路	由东向西	28.7km/h
…				

TSI 是由车辆的实际行驶速度与车辆的自由行驶速度综合计算得出，即：

$$\text{TSI}=\frac{V_f-V_i}{V_f}=\frac{\text{自由行驶速度}-\text{实际行驶速度}}{\text{自由行驶速度}}\times 100$$

车辆的自由行驶速度可以用非高峰时段的车辆实际行驶速度来确定。如通过计算得到某市延安高架上车辆的自由行驶速度为 60km/h，那么 2018 年 4 月 25 日 8:00~8:05 时间段延安高架由西往东方向的 TSI 为：(60-40.535)/60×100=32.44，说明其处于基本畅通的状态。

通常可以将一个城市划分为不同区域，先综合各区域所包含路段的交通状况来衡量区域交通状况，再采用分组分析法对不同拥堵状况的区域进行分组统计，以掌握该城市的整体交通状况。如某市根据快速路网的交通特性，结合道路网设施的交通分流、合流节点位置，将快速路划分为 42 个指数区域，将外环线以内地面道路划分为 68 个指数区域。只要掌握这些指数区域的交通状况，就可以基本了解该市的整体交通状况。

利用上述方法先计算出各路段每个时间段（如 1 小时）的 TSI，然后对区域所包含路段的 TSI 求平均，即可计算得到如表 3-2 所示的区域 TSI。

核心概念

分组分析法（group analysis）是在分组的基础上，对现象的内部结构或现象之间的依存关系从定性或定量的角度做进一步分析研究，认识所要分析对象的不同特征、不同性质及相互关系。

表 3-2　区域 TSI

路段编号	时间段	TSI	交通状况
1	2018-04-25 8:00~9:00	86	严重拥堵
2	2018-04-25 8:00~9:00	54	轻度拥堵
3	2018-04-25 8:00~9:00	32	基本畅通
4	2018-04-25 8:00~9:00	16	畅通
5	2018-04-25 8:00~9:00	67	中度拥堵
6	2018-04-25 8:00~9:00	88	严重拥堵
...			
42	2018-04-25 8:00~9:00	18	畅通
1	2018-04-25 9:00~10:00	76	中度拥堵
...			

通过调用编程语言如 Python 的 groupby 函数等方式，可以统计出该市快速路各时间段交通状况的分组分布情况，如表 3-3 所示。

表 3-3　各时间段交通状况分组分布情况

时间段	交通状况分组路段统计				
	严重拥堵	中度拥堵	轻度拥堵	基本畅通	畅通
2018-04-25 8:00~9:00	5	12	15	8	2
2018-04-25 9:00~10:00	1	3	6	19	13
...					
2018-04-25 20:00~21:00	0	1	2	6	33

根据分组统计结果可以知道：早高峰期间，该市快速路的交通压力普遍较大，大部分路段都有不同程度的拥堵，只有少部分路段处于畅通状况；而在 20 点以后，该市的整体交通状况基本处于畅通状况。

思考与讨论

1. 如何根据所获得的定位数据计算两个或多个连续点之间的车辆行驶速度？

2. TSI 对于人们的交通出行具有哪些参考意义？

活　动

6.1　利用配套资源提供的某市交通轨迹点数据集，分析静安寺商圈在某一周末不同时间段（以小时为考察对象）的车辆数量变化情况。

6.2　利用配套资源提供的网上书店数据集，根据商品的订单归属地，分组比较不同地区的人均图书购买量，探讨不同地区人群的读书习惯。

2. 分析造成城市道路交通拥堵的相关因素

从时间维度对城市交通数据进行分析，可以得到关于城市交通变化过程的有用信息，利用历史数据进行统计，可以总结和发现交通变化趋势，对交通状况进行预测；从空间维度对城市交通数据进行分析，可以对城市内部交通状况进行横向对比，得出影响最大的局部区域、路段、节点等信息，从而更清晰地认识城市关键节点的交通状况，便于有针对性地采取措施去提高城市交通效率或出行效率。

城市交通具有较明显的周期性规律，如星期天和星期一的情况差别就非常大。以一星期七天作为考察周期，采用对比分析法，将同一路段某天各个不同时间段（如 5 分钟）的 TSI 与历史数据进行对比，考察其匹配情况，从而简单揭示交通拥堵规律。历史数据可以通过选取过去某一足够长时间段（一年或者两年）内所有与对比日的星期几相同的日子，将同样时间段的 TSI 求平均得出。

若某时间段的 TSI 与历史数据的相差比例在某一阈值范围（如 5%）以内，则定义为匹配，否则定义为不匹配。若一天内所包含的与历史数据相匹配的时间段数比例超过某一阈值（如 90%），则定义为全天匹配，否则定义为非全天匹配。

对历史数据的匹配情况进行分析可以为预测交通状况提供有效帮助：对于匹配较好的路段，依据历史交通状况进行预测能得到较为理想的结果；对于匹配较差的路段，就要多考虑其他因素的影响。图 3–4 显示了某天某两个路段的实时交通数据与历史交通数据的对比结果。

从图 3–4 中可以看出，路段 1 的实时交通数据与历史交通数据匹配较好，相似度非常高。相比较而言，路段 2 的实时交通数据与历史交通数据匹配得不太好，这说明影响路段 2 的交通状况的随机因素更多。

核心概念

对比分析法(comparative analysis)也称比较分析法，其特点是通过准确、量化的数据，直观地反映事物某方面的变化或差距。

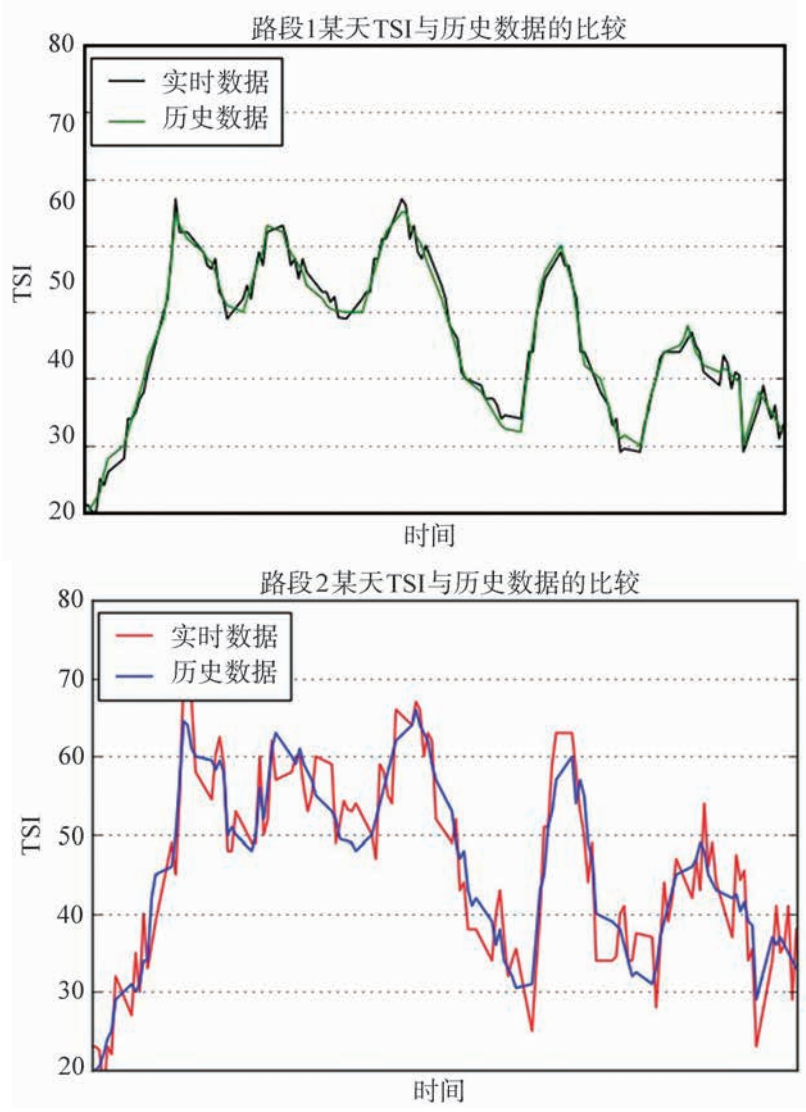


图 3-4　不同路段交通状态指数与历史数据比较

在实际生活中，道路交通状况还受到路段宽度（车道数）、路道长度、交通灯数、车辆数、交叉路口数等因素的影响。采用相关分析法对路段交通状况受上述因素的影响程度进行分析，可以进一步了解城市道路交通拥堵出现的规律。

利用 Pearson 相关系数计算公式，可以得到每一个路段的 TSI 与路段上车辆数的相关系数。

表 3-4　不同时间段 TSI 及车辆数

路段	时间段	TSI	车辆数
路段 1	时间段 1	30	80
路段 1	时间段 2	40	100
…			
路段 n	时间段 m−1	40	70
路段 n	时间段 m	35	65

核心概念

相关分析法（correlation analysis）是研究现象之间相关关系的规律，并据此进行预测和控制的分析方法。

小贴士

相关系数由统计学家卡尔·皮尔森（Karl Pearson）提出，用于反映变量之间相关关系的密切程度。相关系数的绝对值越大，相关性越强。

表 3-4 所示为某市某些路段某段时间（90 天）内不同时间段的平均 TSI，以及这些路段在不同时间段的车辆数。

先计算同一路段在所有时间段的平均 TSI 和平均车辆数，然后利用相关系数计算公式进行计算，得到每一个路段的 TSI 与路段上车辆数的相关系数，如表 3-5 所示。

通过结合地图数据，还可以获取不同路段的路段宽度、路道长度、交通灯数、交叉路口数等数据，用于进一步计算路段 TSI 与这些数据的相关性。如路段 TSI 与路段宽度具有极强的正相关性，增加车道数可以改善路段交通状况；交通灯数过多会严重影响通行效率；交叉路口数过多导致的车辆频繁分流和汇合也会对路段交通状况产生较强的负面影响。

平均分析法、分组分析法、对比分析法及相关分析法都是较常用的数据分析方法，广泛适用于对各种应用场景进行数据分析。比如通过对电子商务平台所产生的用户购物数据进行分析，可以了解人们的在线购物行为习惯；通过对各种外卖平台所产生的订单数据进行分析，可以了解人们的在线订餐与用餐行为特点；通过对在线旅行平台所产生的机票、酒店等数据进行分析，可以了解人们的出行行为偏好。平均分析法、分组分析法、对比分析法可以对用户行为进行整体分析，相关分析法则可以进一步分析用户行为背后所关联的相关因素，进而为引导用户行为提供数据支持。

表 3-5　路段 TSI 与车辆数的相关系数

路段	相关系数
路段 1	0.86
路段 2	0.95
…	
路段 n-1	0.74
路段 n	0.88

参见 P62 知识链接“常用数据分析方法”

思考与讨论

现实生活中能接触到的数据，哪些属性之间可能存在相关性？哪些属性是独立的，相互之间不存在相关性？

活　动

6.3　利用配套资源提供的网上书店数据集，完成对比分析和相关分析。

（1）以图书类别进行区分，比较不同产品在不同地区的受欢迎程度。

（2）对区域图书销量与区域用户数之间的相关性，以及图书销量与图书价格之间的相关性进行分析。

知识链接

数据分析

数据分析是指用适当的统计分析方法对数据进行分析，从中提取有用信息并形成结论的过程。常用的数据分析方法有平均分析法、分组分析法、对比分析法、相关分析法等。数据挖掘则是一种特殊的数据分析方法，专注于建模和知识发现，用于预测而非纯粹的描述目的，可以对数据中所隐藏的信息做进一步的挖掘和利用。

常用数据分析方法

1. 平均分析法

平均分析法是一种利用平均指标对现象进行分析的方法。平均指标又称为平均数，是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。平均分析法使用的平均数是一个抽象化的数值，用来说明总体的集中趋势，它的值介于最小值和最大值之间。常用的平均数有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的，最简单的算术平均数计算公式如下：

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

还有加权算术平均数，计算公式如下：

$$\overline{x}=\frac{x_1\times f_1+x_2\times f_2+\cdots+x_n\times f_n}{f_1+f_2+\cdots+f_n}$$

其中$f_1, f_2, \cdots, f_n$为$x_1, x_2, \cdots, x_n$的权值。

此外，常用的数值平均数还有调和平均数、几何平均数、平方平均数等。

位置平均数是根据数据的大小顺序或数据出现的次数得出的，是总体趋势的代表值。常用的位置平均数有中位数和众数。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为合适。众数是总体中出现次数最多的数据，有时在一组数中会有好几个众数，即出现次数最多的数据不止一个。

2. 分组分析法

分组分析法是对现象的内部结构或现象之间的依存关系从定性或定量的角度做进一步分析研究，认识所要分析对象的不同特征、不同性质及相互关系，以便寻找事物发展的规律，正确地分析问题和解决问题。分组分析法的关键在于正确选择分组标志，以提高统计分析结果的科学性和真实性。

分组时，应根据研究的目的和客观现象的内在特点，按某个或某几个标志把被研究的

总体划分为若干个不同性质的组，使组内的差异尽可能小，组间的差异尽可能大。分组必须遵循两个原则：穷尽原则和互斥原则。穷尽原则，就是使总体中的每一个单位都有组可归；互斥原则，就是在特定的分组标志下，总体中的任何一个单位只能归属于某一个组，不能同时归属于几个组。

根据作用的不同，分组分析法可以分为结构分组分析法和相关关系分组分析法，其中结构分组分析法又可再细分为按品质标志分组和按数量标志分组。品质标志是指姓名、性别等性质属性，可以用来分析现象的类型特征和规律性。数量标志是指年龄、身高、体重等数量特征，可以用来分析现象总体的内部结构及其变化。相关关系分组分析法可以用来分析社会经济现象之间的相关关系。这些分组分析法在实际操作中常常被结合在一起使用。

3. 对比分析法

对比分析法也称比较分析法，是自然科学、社会科学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。

根据实际需要，对比分析法主要有两种形式：横向比较和纵向比较。横向比较是指在同一时期对不同事物进行比较，例如比较不同地区某个季度的销售额。纵向比较是指在不同时期对同一事物进行比较，例如比较不同季度某个地区的销售额。

在数据对比分析中，除了使用绝对数进行比较外，也可以使用相对数进行比较。相对数是两个有联系的现象数值的比率，例如每月食品支出额占消费支出总额的比例、图书馆中图书的借出率等。

4. 相关分析法

相关分析法是研究现象之间相关关系的规律，并据此进行预测和控制的分析方法。通过对大量数据的观察和研究可知，利用相关分析法可以发现各因素之间确实存在某种相关关系，并且有的关系强，有的关系弱，程度各有差异。这种相关程度的差异可以利用绘制散点图的方式呈现。由于散点图自身的限制，它并不能够准确描述各因素之间的相关关系，此时可以通过计算各因素之间的相关系数，更加精确地描述它们之间的相关程度。

各因素之间的相关性可以分为三种类型：单相关，指两个因素之间的相关关系，即研究时只涉及一个自变量和一个因变量；复相关，指三个或三个以上因素的相关关系，即研究时涉及两个或两个以上的自变量和因变量；偏相关，指在某一现象与多种现象相关的场合，当假定其他变量不变时，其中两个变量之间的相关关系。

相关系数用来描述两个因素之间的相关性程度，常用字母“r”表示。常用的有 Pearson 相关系数，计算方式为：

$$r=\frac{\sum(x-\bar{x})\times\sum(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\times\sqrt{\sum(y-\bar{y})^2}}$$

相关系数的取值范围在 –1 和 1 之间，即 $-1 \leqslant r \leqslant 1$。其中：

- 若 $0 < r \leqslant 1$，表明变量之间存在正相关关系，即两个变量的变动方向相同；

• 若 $-1 \leqslant r < 0$，表明变量之间存在负相关关系，即两个变量的变动方向相反；

• 若 $|r| = 1$，其中一个变量的取值完全取决于另一个变量，两者为函数关系；当 $r = 1$ 时，表明变量之间完全正相关；当 $r = -1$ 时，表明变量之间完全负相关；

• 若 $r = 0$，表明变量之间不存在线性相关关系，但并不排除变量之间存在非线性相关关系。

根据经验，可将变量之间的相关程度分为以下几种情况：

• 若 $|r| \geqslant 0.8$，视为高度相关；

• 若 $0.5 \leqslant |r| < 0.8$，视为中度相关；

• 若 $0.3 \leqslant |r| < 0.5$，视为低度相关；

• 若 $|r| < 0.3$，表明变量之间的相关程度极弱，可视为不相关。

项目七

揭示网上书店图书销售情况
——分析、呈现并解释数据

随着互联网技术和物流业的发展，以电子技术为手段、把传统销售购物渠道转移到互联网的电子商务获得了快速发展。电子商务打破了国家与地区有形无形的壁垒，使生产企业达到全球化、网络化、个性化、一体化。

电子商务的兴起产生了庞大的商业数据，如商家在线销售的各种商品数据、用户的购买记录数据和用户的评价数据等。通过使用高效的可视化技术，商家可以直观地看到销售数据的分析结果（图3-5），了解商品的销售情况，发现用户和商品变量之间的相关关系等，从而调整商品库存和销售策略，以获得最大的经济效益。

图 3-5　可视化销售数据

项目学习目标

在本项目中，我们将围绕网上书店图书销售数据，用数据分析及图形可视化等技术，完成网上书店的销售数据分析和结果呈现，帮助商家了解图书的销售情况，发现用户和图书之间的相关性关系。

完成本项目学习，须回答以下问题：

1. 数据分析及图形可视化工具有哪些？如何选择恰当的工具？
2. 如何用可视化图形了解数据质量，并揭示数据的分布？
3. 什么是相关性？如何通过分析发现变量之间的相关关系？

项目学习指引

1. 分析并呈现网上书店图书销售情况

网上书店因其便捷性，每天都会产生大量的交易记录数据，这些数据承载着网上书店的图书销售信息。对网上书店的管理人员来说，及时对采集到的图书销售数据进行分析、呈现和解释，并根据结果调整图书商品的分布结构，在销售旺季前做好备货，对销量不理想的图书进行营销宣传等，可以让书店获得更大的经济效益。

数据分析的基本工作流程如图 3-6 所示。

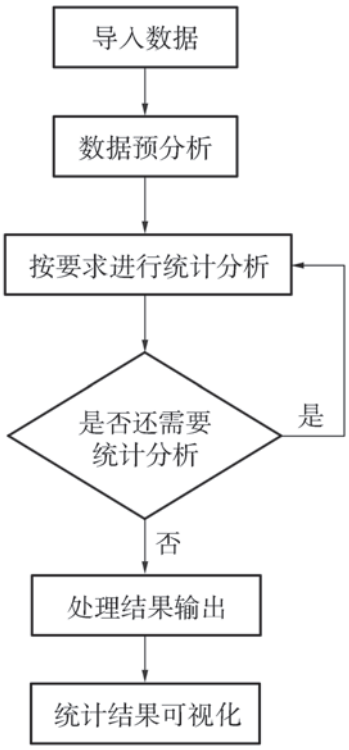


图 3-6　数据分析的基本工作流程

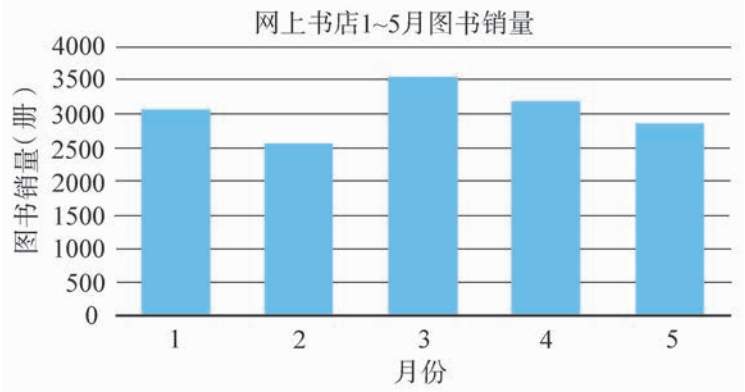
核心概念

数据可视化(data visualization)是指将数据分析的结果通过表格、图形或图像等形式显示出来，再进行交互处理的理论、方法和技术。

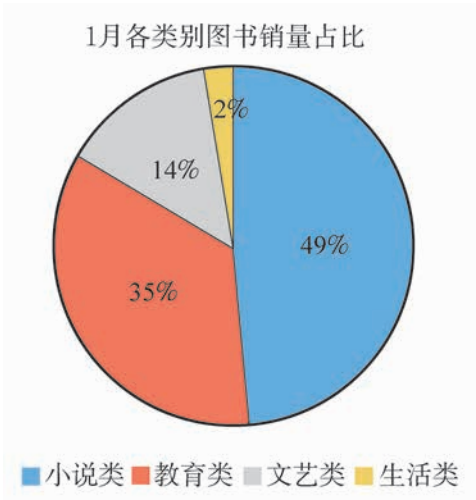
为了直观呈现图书的销售情况，通常需要借助数据可视化技术。数据可视化是数据处理的重要技术之一，是对数据分析结果的呈现。它综合运用计算机图形学、数字图像处理、计算机视觉、计算机辅助设计及人机交互技术等领域中的相关技术，从复杂的多维数据中产生图形，还可为后续更深层次的分析解释打下基础。根据不同的算法，可以得到不同的可视化图形，反映不同的内在规律。

比如网上书店各月图书销量可以用柱状图显示，某月各类别图书的销量占比情况可以用饼图显示，如图 3-7 所示。类似这样的简单数据分析任务，可以直接使用电子表格软件完成。电子表格软件操作简单、上手快，适用于财务、金融等一般数据量和简单数据分析任务的处理。

参见 P75 知识链接“数据可视化”



柱状图



饼图

图 3-7　某网上书店基本销售情况

对于一些复杂的数据分析任务，则需要使用专业数据分析软件。这些软件集成了近 20 年发展起来的一些新的数据分析方法，具有强大的数据可视化功能。有一些软件类似于傻瓜相机，界面友好、使用简单，可以解决绝大部分数据分析问题，适合初学者；还有一些软件类似于单反相机，可以通过编程来实现数据分析任务，功能强大，适合中高级用户。

市场上大多数数据分析软件是商业软件，虽然简单易用且较专业，但一般需要付费使用，而且对于较复杂的数据分析任务，处理起来并不方便。如商家想有针对性地确定各类图书的采购量并制定营销策略，就需要了解平均销量、销量变化范围以及容易出现的异常情况等信息。类似这样较复杂的数据分析任务，可用 Python、R 等编程语言来完成。如通过使用 Python 的 numpy、matplotlib、scipy、scikit-learn 等扩展库，就可以进行基本的数据分析，并完成可视化任务。

参见 P75 知识链接“常用数据分析工具”

以某网上书店四种不同类别图书的部分销售数据为例，2017 年 1 月到 5 月共 151 天的图书销量（册）如表 3-6 所示。

表 3-6　四种类别图书销售数据

日期	小说类	教育类	文艺类	生活类
2017 年 1 月 1 日	51	35	14	2
2017 年 1 月 2 日	49	30	14	2
2017 年 1 月 3 日	47	32	13	2
…				
2017 年 5 月 29 日	65	30	52	20
2017 年 5 月 30 日	62	34	54	23
2017 年 5 月 31 日	59	30	51	18

商家可以根据不同的图书销量分布情况选择不同的备货和销售策略，如对平均销量高的图书多备货，或者将销量高的图书与销量低的图书捆绑打折销售等。这时，可以尝试分析图书销量分布的主要数字特征，如图书销量的四分位数和中位数。箱线图能够直观呈现这些数据，反映以上四种不同类别图书的销量分布情况。

小贴士

在统计学中，把所有数值由小到大排列并分成四等份，处于三个分割点位置的数值就是四分位数。上四分位数等于样本中第 75% 的数，中位数等于样本中第 50% 的数，下四分位数等于样本中第 25% 的数。

箱线图（box plot）是一种用于显示一组数据分散情况的统计图，可直观简洁地展现数据分布的主要特征。

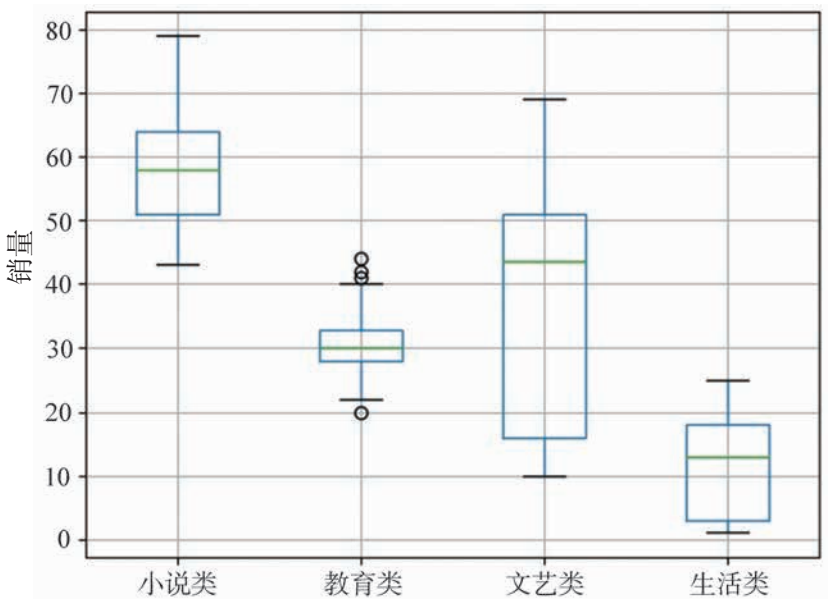


图 3-8　四种类别图书的销量箱线图

如在 Python 中，可以调用 boxplot 函数绘制箱线图（图 3-8）。在绘制得到的箱线图合集中分别列示了“小说类”“教育类”“文艺类”和“生活类”图书的销量箱线图。

从图 3-8 中可以看出，小说类图书的销量中位数最大（图中绿色的线），文艺类图书的销量变化范围最大（图中长方形上面的线为上四分位数，下面的线为下四分位数）。箱线图还可以反映出异常值，让商家了解到某些特殊情况对图

书销量的影响，以便制定更有效的应对策略。从图 3-8 中可以看出，教育类图书的销量最容易出现异常情况（箱线图上方和下方均出现异常值，以圆圈表示）。

如果商家想要根据图书的销售情况调整各种图书的库存量，就需要进一步了解在某一段时间内图书销量的总体分布情况，如不同图书销量对应的天数占总天数的比率等，这时可以使用密度图。

如在 Python 中，可以调用 plot 函数并将其 kind 参数设为“density”来绘制四种不同类型图书在 151 天内每天销量的密度图（图 3-9），其中横轴表示每天卖出的图书的册数，纵轴表示卖出该册数的频率。

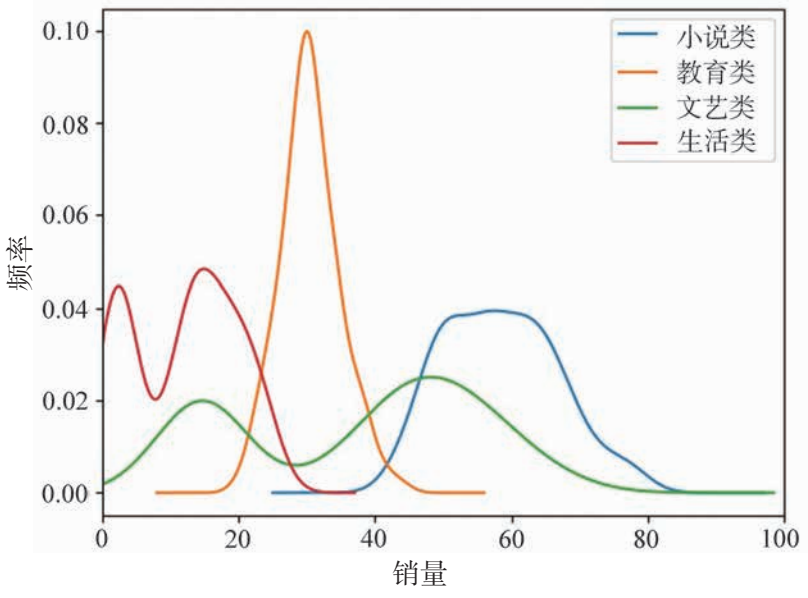


图 3-9　四种类别图书销量对应天数的密度图

从图 3-9 中，可以清晰地看出各类别图书每天的销量分布数据。其中，小说类和教育类图书只有一个销量数值出现的频率最高，文学类和生活类图书则各有两个销量数值出现的频率较高。

如果商家想要根据图书销量的波峰或波谷来调整促销策略，则需要了解图书具体销量对应的天数在特定销量区间的分布，这时可以使用直方图。

如在 Python 中，可以调用 plot 函数并将其 kind 参数设为“hist”，或调用 hist 函数绘制直方图。图 3-10 的直方图以小说类图书为例，从中可以看出，2017 年 1 月至 5 月这段时间，小说类图书卖出 55 册的天数最多，卖出 75 册的天数最少。

小贴士

密度图（density plot）常用于显示数据在连续时间段内的分布情况。

小贴士

直方图（histogram）是用一系列高度不等的纵向条纹或线段来表示数据分布的情况。

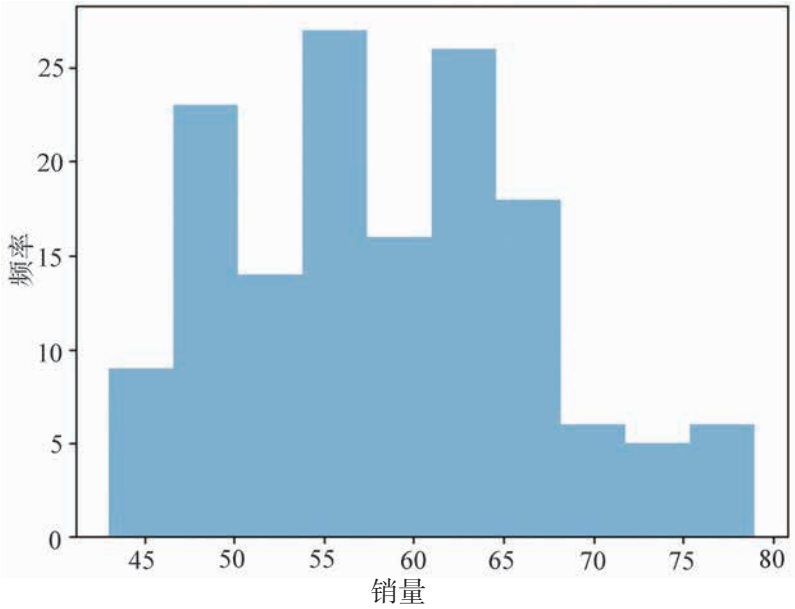


图 3-10　小说类图书每天销量直方图

如果商家想要制定相应的捆绑营销策略，就需要知道各类别图书销量之间的关系，这时可以使用散点图。

如在 Python 中，可以调用 plot 函数并将其 kind 参数设为“scatter”，或调用 scatter 函数绘制散点图。以小说类和教育类图书为例，绘制两种类型图书销量密度之间关系的散点图，图中用红色点表示前面 1~50 天的散点，黄色点表示中间 51~100 天的散点，蓝色点表示最后 101~151 天的散点，如图 3-11 所示。

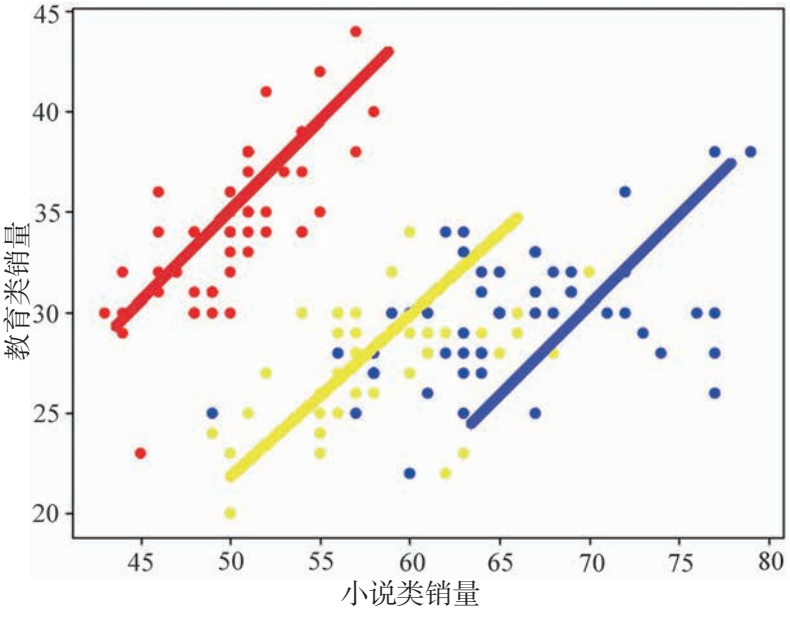


图 3-11　小说类和教育类图书销量对应关系的散点图

可以明显地看出，不论在哪个周期内，散点都近似地分

小贴士

散点图（scatter plot）是指在回归分析中，数据点在直角坐标系平面上的分布图。

小贴士

在可视化绘图中，可以尝试将不同工具进行结合。如图 3-11 中的三条直线是后来用其他工具添加上去的，不是软件自动生成的。

布在一条直线（图中三条不同颜色的直线）左右，因此可以判断出，小说类图书的销量与教育类图书的销量基本成正比，捆绑销售能够同时促进两者的销量。

两类图书销量关系的散点图还可以转换成六边形箱式图，帮助商家更直观地了解这两类图书销量的关系。如在 Python 中，可以调用 plot 函数并将其 kind 参数设为“hexbin”，或调用 hexbin 函数绘制六边形箱式图（图 3-12）。

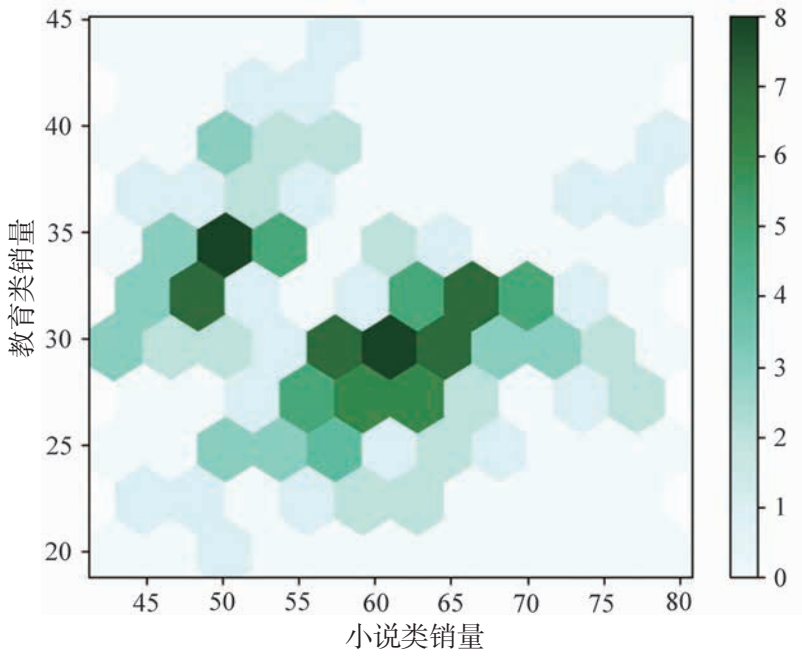


图 3-12　小说类和教育类图书销量对应关系的六边形箱式图

从图 3-12 中可以看出，小说类和教育类图书在（60,27）和（50,35）这两个值附近的散点数最多（六边形颜色越深表示散点数越多），说明这两类图书在这两个销量附近相关程度最高。

思考与讨论

怎样的数据适合用散点图进行分析？散点图分析结果能够说明隐藏在数据中的哪些规律？

小贴士

六边形箱式图（hexagonal boxplot）是一种直观了解数据之间关系的图形，它是基于散点图得到的。

活　动

7.1　全班同学分成几个小组，合作完成以下可视化图形。

（1）绘制上个项目中车辆运行时间分布密度的直方图。

（2）绘制上个项目中车辆运行时间和行驶轨迹两类数据之间对应关系的散点图和六边形箱式图。

2. 发现用户数据的相关性

通过以上的数据分析及可视化呈现，商家可以发现图书销量之间是存在相关性的，但这种认识不够具体和准确，不足以为决策提供足够的置信度。商家若想通过制定更精确的经营策略来提高图书销量，还需要进一步对一些用户数据的相关性进行分析。

下面选取与提高图书销量有关的几个关键用户数据，包括图书的购买数、收藏数、好评数和差评数，如表 3-7 所示。

表 3-7　几个关键用户数据

月份	购买数	收藏数	好评数	差评数
1	18609	19125	3107	2740
2	22633	22512	4602	2930
3	19257	18052	3995	3857
…				
10	23178	20873	6724	4327
11	15935	17497	3551	4777
12	18823	16516	4174	3067

小贴士

折线图（line chart）是用直线段将各数据点连接起来而组成的图形，以折线方式显示数据的变化趋势。

首先绘制出 1~12 月图书购买数、收藏数、好评数和差评数随月份变化的折线图（图 3-13），然后探寻这四个变量之间的相关关系。如在 Python 中，可以调用 plot 函数绘制折线图。

从图 3-13 中可以看出，购买数、收藏数和好评数的变

参见 P76 知识链接“常见可视化图形”

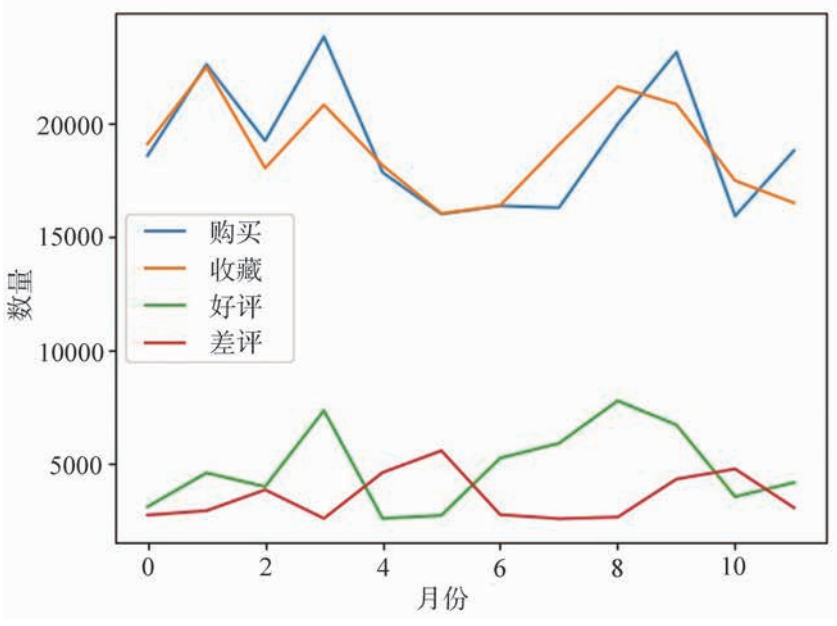


图 3-13　图书购买数、收藏数、好评数和差评数随月份变化的折线图

化规律相似，当其中一个数据增加时，另外两个数据也随之增加，反之亦然。我们称购买数、收藏数、好评数之间是正相关关系。差评数与另外三个数据的变化规律正好相反，我们称差评数与购买数、收藏数、好评数都是负相关关系。商家可以通过好评送礼物、收藏送折扣券等策略，提高好评数、收藏数，降低差评数，最终提升图书的销量。

折线图虽然能直观地反映变量的变化规律和变量间的相互关系，但是不能具体地表明两个变量之间的相关程度。商家若想更准确地针对不同数据设计具体的营销策略，还需要准确掌握两个变量之间的相关程度，为此需要引入一些数值化的计算结果，即相关系数。

相关系数有很多种，除了上个项目介绍的最常见的 Pearson 相关系数，还有 Spearman 相关系数和 Kendall Tau 相关系数等。Python 的 pandas 库提供了多种相关系数的计算方法，直接调用函数即可计算出这些相关系数。图书购买数、收藏数、好评数和差评数四个变量之间的三种相关系数，如表 3-8 所示。

表 3-8　三种相关系数

Pearson 相关系数	购买数	收藏数	好评数	差评数
购买数	1.000000	0.786494	0.563608	-0.317979
收藏数	0.786494	1.000000	0.618589	-0.434319
好评数	0.563608	0.618589	1.000000	-0.564893
差评数	-0.317979	-0.434319	-0.564893	1.000000
Spearman 相关系数	购买数	收藏数	好评数	差评数
购买数	1.000000	0.727273	0.594406	-0.363636
收藏数	0.727273	1.000000	0.545455	-0.475524
好评数	0.594406	0.545455	1.000000	-0.671329
差评数	-0.363636	-0.475524	-0.671329	1.000000
Kendall Tau 相关系数	购买数	收藏数	好评数	差评数
购买数	1.000000	0.515152	0.424242	-0.242424
收藏数	0.515152	1.000000	0.363636	-0.303030
好评数	0.424242	0.363636	1.000000	-0.515152
差评数	-0.242424	-0.303030	-0.515152	1.000000

不同数据之间的相关系数是不同的。正值表示正相关，负值表示负相关，数值越大表明两个数据之间的相关程度越高。不同相关系数得到的结果也是不同的，应当根据数据的类型，选择合适的系数作为设计策略的标准。

参见 P77 知识链接“三种相关系数”

对于销售数据来说，较适合的是 Pearson 相关系数。在 Pearson 相关系数下，图书购买数与收藏数、好评数、差评数的相关系数分别为 0.786494、0.563608、–0.317979。因此，商家应着重考虑如何提高收藏数，在降低差评数上则可采用一般策略。

数据分析的基本思想是：以数据为根本，可视化为手段，描述真实世界，探索隐藏在数据背后的客观规律。由此可见，数据分析不仅可以在网上书店等商业应用上大显身手，在其他诸如政治、经济、军事、科技、生活等领域也都能发挥举足轻重的作用。人们每天都可以看到一些借助数据分析及可视化呈现实现的新奇应用，由此来获取真正有价值的信息，如通过社会关系可视化、地理信息可视化等，很多组织和个人都感受到数据分析的强大力量。

随着海量数据时代的到来，数据驱动已成为必然趋势，通过对海量数据进行分析将极大地推动人类文明的发展。

思考与讨论

散点图和相关系数都可以用于发现数据间存在的相关关系，它们分别适用于什么情况？

活　动

7.2　绘制收藏数与购买数、好评数与差评数的散点图，并结合 Pearson 和 Spearman 相关系数表格，说明散点图中数据的分布与相关系数之间的关系。

7.3　绘制上个项目中路段交通状况与路段宽度、路道长度、交通灯数等因素的变化曲线，并计算它们之间的 Kendall Tau 相关系数和 Spearman 相关系数。

知识链接

数据可视化

数据可视化是指将数据分析的结果通过表格、图形或图像等形式显示出来，再进行交互处理的理论、方法和技术。数据可视化的本质是视觉对话，它旨在借助图形化手段，清晰有效地传达与沟通信息。数据是信息的载体，信息的质量很大程度上依赖于数据的表达方式。对数据分析结果进行可视化呈现，可以帮助人们更好地从数据中提取信息，从信息中收获知识和价值。数据可视化的优势包括以下方面。

1. 传递信息速度快

人脑对视觉信息的处理速度要比书面信息快 10 倍，使用表格、图形或图像来总结和展示复杂的数据，可以提高人们对各种关系的理解速度。

2. 数据显示多维性

数据可视化可以将数据每一维的值分类、排序、组合并呈现，让人们可以看到表示对象或事件的数据的多个属性或变量。

3. 展示信息更直观

数据可视化能够用一些简洁的图形来体现复杂信息，甚至只需要单个图形就能帮助人们轻松地解释各种不同的数据源。

4. 克服大脑记忆能力限制

观察物体时，人脑有长期记忆和短期记忆。很多研究表明，图像更容易被理解，更有趣，也更容易被记住。

常用数据分析工具

数据分析工具有很多，在实际应用中，可以根据应用场景和数据分析任务的复杂程度，选择不同的数据分析工具。各种数据分析工具的优缺点参见表 3-9。

表 3-9　各种数据分析工具的优缺点

工具名	优点	缺点
WPS 表格 / Excel	极易使用	运行效率较低、样本量受限、统计学功能不完善
SPSS	易学易用，统计学功能全面	收费，运行效率不高
Stata	易学易用，在统计方面优势突出	收费
SAS	统计学功能强大，适合大样本分析	收费，需要学习相关编程语言
Python	免费的通用编程语言，有很多扩展库支持	需要学习相关编程语言
R	免费的通用编程语言，有很多扩展库支持	较难学习

当数据量小、分析任务较简单时，可以选用 WPS、Excel 等电子表格软件以提高数据分析的效率；当分析任务较为复杂时，可以选用专业数据分析软件，如社会科学应用通常选用 SPSS，商务应用通常选用 Stata、SAS；有一定的程序设计基础或考虑开源软件应用开发时，可以选用 Python、R 等编程语言以支持个性化的数据分析。

常见可视化图形

1. 直方图

直方图是用一系列高度不等的纵向条纹或线段来表示数据分布的情况，例如图 3-10。一般用横轴表示数据类型，纵轴表示分布情况。将数据取值的范围分成若干区间（一般是等间隔的），每个区间的长度称为组距。考察数据落入每一区间的频数与频率，在每个区间画一个矩形，其宽度是组距，高度可以是频数、频率或频率 / 组距。当高度是频率 / 组距时，每个矩形的面积恰好是数据落入区间的频率，这种直方图可以用于估计总体的概率密度。

2. 折线图

折线图是用直线段将各数据点连接起来而组成的图形，以折线方式显示数据的变化趋势，例如图 3-13。折线图可以显示随时间而变化的连续数据，因此非常适用于显示相等时间间隔下数据的变化趋势。在折线图中，类别数据沿横轴均匀分布，数值数据沿纵轴均匀分布。在折线图中，数据是递增还是递减、增减的速率、增减的规律（周期性和规律性等）、峰值等特征都可以清晰地反映出来。

3. 密度图

密度图又称密度曲线图，常用于显示数据在连续时间段内的分布状况，例如图 3-9。这种图形是直方图的变种，使用平滑曲线来绘制数值水平，从而得出更准确的分布。密度图的峰值显示数值在该时间段内最为集中的位置。密度图不受所使用分组数量（典型直方图中所使用的条形）的影响，所以能比直方图更好地界定分布形状。

4. 箱线图

箱线图是一种用于显示一组数据分散情况的统计图，因形状如箱子而得名，可直观简洁地展现数据分布的主要特征，例如图 3-8。箱线图的构造方法如下：画一个矩形，上下两端分别是上四分位数和下四分位数；矩形中间画一条横线，表示中位数的位置；从矩形上下两端向外各画一条竖直线段，直到非异常值的最远点，然后在两个最远点处各画一条横线，标示异常值截断点；异常值用“圆圈”表示，在异常值截断点以外画出来。

5. 散点图

散点图是指在回归分析中，数据点在直角坐标系平面上的分布图，例如图 3-11。散点图反映了因变量随自变量变化的大致趋势，可用于展示数据的分布和聚合情况。根据散点图，可以选择合适的函数对各数据点进行拟合。

6. 六边形箱式图

六边形箱式图是一种直观了解数据之间关系的图形，它是基于散点图得到的，例如图 3-12。首先将绘制图形的整块区域划分成很多六边形的小区域，然后根据六边形小区域

内散点的数目确定六边形的颜色。散点数目越多，颜色越深，表示在这段区间内两数据之间的相关性越高。

三种相关系数

1. Pearson（皮尔森）相关系数

Pearson 相关系数又称积差相关系数或简单相关系数，用于衡量连续变量的相关指标。它一般适用于两个变量呈线性相关的情况，通常用字母 r 表示。当两个变量的标准差都不为零时，Pearson 相关系数才有定义，其适用范围包括：

（1）两个变量之间是线性关系，且都是连续数据；

（2）两个变量的总体是正态分布，或接近正态的单峰分布；

（3）两个变量的观测值是成对的，每对观测值之间相互独立。

2. Spearman（斯皮尔曼等级）相关系数

Spearman 相关系数又称秩相关系数，是利用两个变量的秩次大小作线性相关分析，对原始变量的分布不作要求，属于非参数统计方法，适用范围较广。可以计算 Pearson 相关系数的数据，亦可计算其 Spearman 相关系数，但统计效能要低一些。Spearman 相关系数的计算公式可以完全套用 Pearson 相关系数的计算公式，只需将公式中的 x 和 y 用相应的秩次代替即可。

Spearman 相关系数对数据条件的要求没有 Pearson 相关系数严格。只要两个变量的观测值是成对的等级评定资料，或者是由连续变量观测资料转化得到的等级资料，不论两个变量的总体分布形态、样本容量的大小如何，都可以用 Spearman 相关系数进行研究。

3. Kendall Tau（肯德尔等级）相关系数

Kendall Tau 相关系数是用于反映分类变量相关性的指标，适用于两个变量均为有序分类的情况。Kendall Tau 相关系数经常用希腊字母 τ（tau）表示。Kendall Tau 相关系数与 Spearman 相关系数对数据条件的要求相同。

项目八

探索网上书店图书推荐
——认识数据挖掘的重要意义

随着互联网的普及和发展，互联网上的信息量急剧增长。人们需要耗费巨大的时间和精力去寻找自己真正需要的东西。作为一种有效的信息过滤手段，推荐系统成为解决当前信息过载问题及实现个性化信息服务的好方法。

推荐系统采用的是一种数据挖掘方法，它能主动收集用户的特征资料，通过研究用户的喜好和兴趣，为用户推荐其可能需要的各种资源，从而大大降低用户搜寻信息的成本。推荐系统在电子商务、社交网络、数字化图书馆、视频 / 音乐点播等领域都已得到了广泛的应用（图 3–14）。

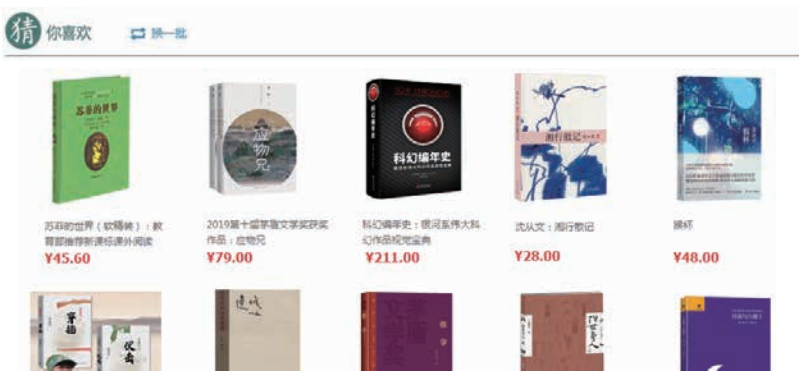


图 3–14　图书推荐系统

项目学习目标

在本项目中，我们将通过分析网上书店在线用户的历史行为数据，挖掘其潜在的阅读兴趣，从而为其推荐合适的图书，借此初步体验数据挖掘的过程，同时了解数据管理与分析技术的新发展。

完成本项目学习，须回答以下问题：

1. 数据管理与分析技术有哪些新发展？
2. 什么是数据挖掘？数据挖掘的基本过程是怎样的？
3. 如何认识数据挖掘的应用价值？

项目学习指引

1. 了解数据管理与分析技术的新发展

互联网、物联网和云计算技术等的迅猛发展，使得大量数据充斥整个世界。面对海量的数据，人们不得不花费更多的时间和精力去搜寻对自己有帮助的信息，信息过载现象越来越严重。数据作为一种重要的资源，亟待人们对其进行合理、高效、充分的利用，使它能够给学习、生活、工作带来更大的效益和价值。

传统的数据管理和分析技术已不适应海量数据的处理，人们迫切需要发展新的技术以应对这种挑战。为此，科学家对计算机进行了纵向与横向扩展。纵向扩展即提高计算机的存储和计算能力，由此产生了大型主机甚至超级计算机。但是这种方式硬件成本高昂，只适用于少数特定场合。横向扩展即产生分布式系统，该系统由一系列通过网络进行通信、为了完成共同任务而协调工作的计算机节点组成，能够利用更多的廉价机器，处理更多的数据。

大数据的出现给数据管理和分析技术带来了新的挑战，同时催生出新的技术。除了分布式系统外，数据管理和分析技术在数据存储与组织、计算方法及用户接口技术等方面都有了新的发展。在数据分析方面，除了用统计方法分析数据的表层规律之外，人们还使用新技术挖掘隐藏在数据内部的有价值信息，这些技术统称为数据挖掘。

通过数据挖掘得到的信息可有效地指导生产实践活动。如某市在承办大型活动期间，通过收集闭路电视摄像头记录、公共交通卡刷卡记录、移动电话基站连接等海量数据，对城市客流量进行实时监控和流动趋势预测分析，制定合理的客流疏导策略，保障活动正常、有序地进行。

思考与讨论

大数据与传统数据存在哪些不同之处？

小贴士

信息过载（information overload）是指真正需要、真正感兴趣的信息，被淹没在其同类信息的海洋里，为了找到它需要耗费巨大的时间和精力。

数字化学习

上网查找资料，了解数据管理与分析技术的新发展。

核心概念

数据挖掘（data mining）是以海量数据作为挖掘对象，通过使用一系列的方法、工具或者算法，发现数据中隐含的、未知的有价值信息。

参见 P86 知识链接“数据管理与分析技术的新发展”

活　动

8.1　班级同学分成若干小组，上网查阅资料，了解数据管理与分析技术的新进展。

（1）任选一个主题，可以是数据存储、计算方法、数据分析等。

（2）收集该主题的相关资料，总结其发展历史及前沿进展。

（3）制作演示文稿并进行汇报。

2. 挖掘用户阅读兴趣

读者在选购图书时会面对种类繁多的书目，即便限定在某一类别，也会有大量的图书可供选择。如果网上书店能够为用户个性化地推荐其喜欢阅读的书目，就能够缩短用户的寻找时间，提升用户的使用体验，从而提升商家的收益。如何找到用户喜欢阅读的图书并推荐给他们呢？

平台向用户推荐的图书必须是用户“感兴趣”的，因此先要知道用户的阅读兴趣，这正是需要数据挖掘解决的问题。

在网上书店的平台上，人们一般通过注册成为用户，然后登录平台搜索、浏览、购买图书；对于暂时不想买但将来可能需要的图书可以进行收藏；对于已购买的图书还可以发表评论并进行评分。用户的搜索、购买、收藏、评分等行为数据都记录在网上书店数据库中，对这些行为数据进行整理、分析，就可以了解用户感兴趣的图书，以及用户性别、年龄、地理位置、阅读兴趣等有价值的信息，甚至进一步挖掘出图书之间的关联关系。

根据用户搜索、购买、收藏过的图书，可以为用户推荐与其相似的其他图书。而要找出相似的图书，可以利用关联规则。关联规则是发现商品之间关系的常用方法，也是非常经典的数据挖掘方法之一。利用关联规则，可以从交易记录中挖掘出图书之间的关联关系。比如，通过调研发现，30%的用户会同时购买图书 A 和 B，而在购买 A 的用户中有 80%的人购买了 B，这就表示存在一种隐含关系：$A \Longrightarrow B$，也就是说，购买 A 的用户会有很大可能购买 B。这种关系既反映了图书之间的内在联系，又反映了用户的购买习惯，因此可以

核心概念

关联规则（association rules）就是发现数据背后存在的某种规则或者联系，它描述了一个事物中某些属性同时出现的规律和模式。

利用关联规则挖掘出图书间的关联关系，进一步推测出用户的阅读兴趣。

但使用这种方式只能知道用户的部分阅读兴趣，比如用户只购买了文学方面的图书，却并不代表他对数学不感兴趣。很多用户往往仅在网上书店平台上留下少量行为，不足以凭此准确推断出他的阅读兴趣，这就是数据挖掘中经常遇到的数据稀疏问题。有没有其他办法去挖掘用户的阅读兴趣呢？

不妨换一个角度思考：阅读兴趣相同的用户，显然会去搜索或购买同类的图书。行为相似的用户间存在相同的兴趣点，这正是协同过滤推荐方法的基本思想。

协同过滤推荐也是一种数据挖掘方法，一般由收集用户行为、发现相似用户和推荐对象三个步骤组成。收集用户行为是分析和挖掘用户兴趣和偏好的基础，行为数据越多，越能准确地反映用户的兴趣和偏好，推荐结果也越准确。因此，可以先找出与目标用户相似的用户，再把相似用户感兴趣的图书推荐给目标用户。在向用户推荐图书时，基于用户的协同过滤推荐方法依然基于用户的阅读兴趣，但是它巧妙地避开了“用户的阅读兴趣是什么”这样的问题，而是基于“用户具有阅读兴趣”这一事实，使用相似用户的阅读兴趣去隐性地表达目标用户的阅读兴趣。

思考与讨论

利用关联规则推荐图书存在哪些不足之处？

小贴士

数据稀疏（data sparsity）是指在数据集中存在很多数值缺失的数据。稀疏数据绝对不是无用数据，只不过是信息不完全，通过适当的手段可以挖掘出大量有用信息。

核心概念

协同过滤推荐（collaborative filtering recommendation）：“物以类聚、人以群分”，该方法利用志趣相投或者经验类似的群体喜好，推荐用户感兴趣的信息、商品或服务等。

活　动

8.2　收集某超市的交易数据，并挖掘商品之间的关联规则。

（1）人工收集某超市的购物小票。

（2）统计每张购物小票中出现的商品。

（3）利用关联规则挖掘商品之间的关联关系。

3. 用协同过滤推荐方法推荐图书

想要用基于用户的协同过滤推荐方法来推荐图书，关键是寻找到相似用户，也就是计算出用户之间的相似性。关于这一问题，可以有不同的解决方法，其中较常用的是基于用户评分的解决方法。

用户购买图书之后，可能会在平台上对图书进行评分，这种评分称为显性评分。不妨假设评分的取值范围为 1 到 5，分数越高表明用户对图书越有“好感”。另外还有一种隐性评分，是通过将用户对图书的操作行为进行量化来转换得到的。例如，用户对图书的搜索、购买等行为，都隐性反映出用户对图书的喜爱程度。综合用户对图书的显性评分和隐性评分，可以得出用户对某一本书的综合评分，评分越高，代表用户越喜欢这本书。

通过分析用户对图书的评分，可以挖掘出用户之间的相似度。比如 5 个用户对两本图书的评分如表 3-10 所示：

小贴士

这里暂不考虑用户因商家服务质量等因素对评分产生的影响，认为评分仅是针对图书的。

表 3-10　5 个用户对两本图书的评分

用户编号	图书编号	
	B1	B2
U1	3.3	5
U2	4	2.5
U3	3.4	4.8
U4	3.2	4.5
U5	4.5	2.2

从表中很难直观发现 5 个用户间的联系，但是将这些评分用散点图表示出来后，用户间的关系就容易发现了。在图 3-15 的散点图中，横轴是用户对图书 B1 的评分，纵轴是用户对图书 B2 的评分。

通过用户兴趣的分布可以发现，用户 U1、U3、U4 距离较近，形成了一个群体；用户 U2 和 U5 形成了另一个群体。同一个群体一般拥有相似的阅读兴趣，因此可以向用户 U1 推荐用户 U3 和 U4 喜欢的图书。

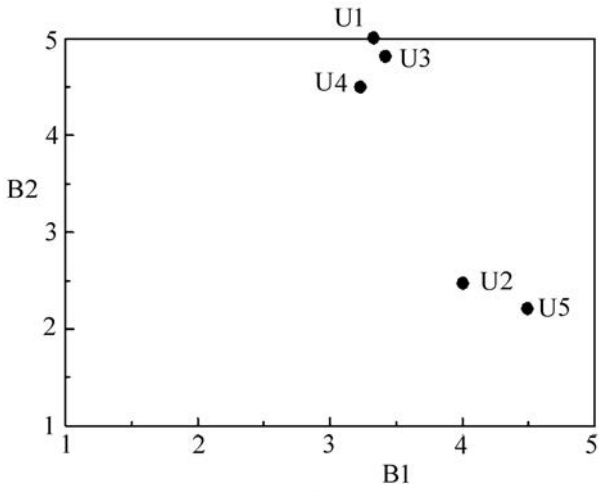


图 3-15　用户评分散点图

对多本图书进行分析时，会产生高维数据，用户之间的相似关系虽然依然成立，但需要高级可视化技术才能直观显示出来。解决这一问题的替代方法是对用户之间的相似度进行数值计算，并依据计算结果来完成图书推荐。

有两种常用的计算相似度的方法：欧氏距离法和 Pearson 相关系数法。假设两个用户 X 和 Y 对 n 本书的评分分别为 $x=(x_1, x_2,\cdots,x_n)$ 和 $y=(y_1,y_2,\cdots,y_n)$，则用户 X 和 Y 之间的欧式距离可由以下公式计算得到：

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

其中 x_i 和 y_i 分别表示两个用户对第 i 本书的评分，$i=1,2,\cdots,n$。使用以下公式可以将距离 d 转换为相似度 s：

$$s = \frac{1}{1+d}$$

可见，用户间的距离越小，用户的相似度就越大。这与上述二维情况下的用户兴趣分布分析是一致的。

Pearson 相关系数法前面已经介绍过，其系数的取值范围为 [−1,1]，取值为正表明两个用户正相关，取值为负表明两个用户负相关，数值越大表明两个用户的相似度越高。

假设 5 个用户对 5 本图书的评分如表 3-11 所示：

小贴士

对高维数据，需要先使用降维技术，将数据降到二维或三维，然后再用可视化技术进行显示。

表 3-11　5 个用户对 5 本图书的评分

用户编号	图书编号				
	B1	B2	B3	B4	B5
U1	3.3	5	2	2.4	3.9
U2	4	2.5	2.2	3.6	2.7
U3	3.4	4.8	2.2	2.8	2.3
U4	3.2	4.5	2.9	2.6	2.9
U5	4.5	2.2	3.4	3.8	2.2

利用 Pearson 相关系数法，计算得到用户之间的相似度如表 3-12 所示：

表 3-12　用户之间的相似度

	U1	U2	U3	U4	U5
U1	1	–0.14	0.76	0.83	–0.64
U2	–0.14	1	0.07	–0.30	0.74
U3	0.76	0.07	1	0.91	–0.20
U4	0.83	–0.30	0.91	1	–0.47
U5	–0.64	0.74	–0.20	–0.47	1

由相似度计算结果可知，用户 U1、U3、U4 的相似度较高，他们属于同一个群体，拥有相似的阅读兴趣，因此可以向用户 U1 推荐 U3 和 U4 喜欢的图书。但是用户 U1 已经对图书 B1 ~ B5 进行了评分，所以不能将这些图书再推荐给他，重复推荐是没有意义的，要推荐的是用户 U1 还没有评过分的图书。

假设找到了图书 b，用户 U1 没有对它评分，而用户 U3 和 U4 已经对它进行过评分，那么 U1 对该书的评分可近似估算为相似用户 U3 和 U4 对该书评分的加权平均，权重取为用户之间的相似度占比，具体计算公式为：

$$m_{1b}=\frac{s_{13}\times m_{3b}+s_{14}\times m_{4b}}{s_{13}+s_{14}}$$

其中 m_{ib} 表示用户 i 对图书 b 的评分，s_{ij} 表示用户 i 与用户 j 之间的相似度。

如果找到了多本这样的图书，就可以按照估算评分的大小顺序向用户 U1 进行推荐了。例如，有两本图书 B6 和 B7，用户 U1 没有对它们进行评分，用户 U3 和 U4 已经对它们进行过评分，如表 3-13 所示。

参见 P87 知识链接“数据挖掘”

表 3-13　两名用户与目标用户的相似度及对图书的评分

	与 U1 的相似度	对 B6 的评分	对 B7 的评分
U3	0.76	3.8	4.5
U4	0.83	4	3.2

按以上公式，可以估算出用户 U1 对图书 B6 的评分为 3.9，对图书 B7 的评分为 3.82。因此，可以按估算评分的顺序，依次向用户 U1 推荐图书 B6 和 B7。

随着互联网的发展，在线平台一方面为用户提供了便捷的服务，另一方面为改善用户体验也采集了许多用户的在线行为。利用平台采集的这些行为数据，商家可以用协同过滤推荐的方法完成电影、音乐、在线产品、手机 App 和服务等的个性化推荐，为人们的生活提供更大的便利。

思考与讨论

隐性评分是用数字去量化用户对图书的感兴趣程度，从而方便算法的实施。如何对用户的搜索、收藏等行为设定不同的权重以体现这种感兴趣程度？

活　动

8.3　探讨并制定针对某视频网站的电影推荐解决方案。

（1）从用户和电影的角度，分别列出需要采集哪些数据，并在选定的视频网站上采集所需的数据。

（2）讨论并制定电影推荐策略，给出具体计算方法。

（3）选择部分用户作为测试样例，将这些用户看过的电影中约 1/3 的电影去掉，对这些用户进行电影推荐，对比推荐的电影中有哪些与去掉的电影是一致的。

知识链接

数据管理与分析技术的新发展

随着信息化程度的不断提高，人们获取数据的渠道越来越多，致使全球数据存储量呈现爆炸式增长。从文明出现到 2003 年，人类总共才创造了 5EB（5×10^{18} 字节）的数据，但是我们现在仅用两天就能创造出相同的数据量。截至 2015 年，全球的数字数据量已达到 8ZB（8×10^{21} 字节）。当前，数十亿台传感设备，包括个人计算机、智能手机、摄像头等，每时每刻都在产生海量的数据。为处理这些数量庞大、形式多样的数据，大数据时代下的数据采集、数据管理和数据分析等技术正在不断发展。

1. 大数据时代下的数据采集

在计算机出现前，采集数据的工作主要由人工完成，不仅效率低，而且成本高、时间长。计算机出现后，为获取自然界、生产领域和生活中的信息，各种类型的传感器开始扩展人们感知世界的方式。互联网出现后，各种在线应用又成为一个个“社会传感器”，捕捉并感知着人们的在线行为。随着人类认知世界能力的提升，势必会有更多微型化、数字化、多功能化、网络化和智能化的新型数据采集方式出现。

2. 大数据时代下的数据管理

数据库是实现数据管理的主要技术。自 20 世纪 70 年代第一款关系数据库诞生后，关系数据库逐渐成为各行各业信息化必不可少的组成部分。然而，随着互联网和物联网技术的发展，不仅数据量在不断增大，而且数据形式也越来越多样化，海量的结构化、非结构化和半结构化数据不断产生。一方面，传统的关系数据库难以存储海量的数据；另一方面，关系数据库管理非结构化和半结构化数据显得力不从心，而非结构化和半结构化数据所占比例目前已接近 80%。针对非结构化和半结构化数据的管理，提倡运用非关系模型来存储数据的 NoSQL 数据管理技术得到了急速发展。

为了应对数据量和用户服务请求的爆炸性增长，一些互联网公司设计并开发了分布式数据库系统。分布式数据库将数据分散地存储在不同的存储单元上，不仅提升了传统数据库的数据存储能力，而且改善了用户体验。

面对海量的、形式多样的数据，企业往往需要同时部署关系型和非关系型数据库系统，这大大增加了数据管理的成本。为降低数据管理成本，一些云服务提供商设计并开发了云端数据库。云端数据库不仅可以免去企业自建数据库的硬件投入和运营维护成本，而且具有部署快和可扩展性好等优点。

3. 大数据时代下的数据分析

传统的数据分析技术专注于对单一数据源的分析。随着信息化的推进，特别是互联网应用的不断涌现，结构化和非结构化数据越来越呈现出碎片化的特征，具体表现为零散细碎、来源各异、形式多样和逻辑关联。然而，数据的碎片化会导致以下几个问题：

- 质量低：从单个数据源来看，数据常常出现缺失、表达不完整或者不准确等问题，甚

至数据本身可能存在真实性问题；从多个数据源来看，来自不同数据源的数据可能存在冲突或不一致。

• 异构性：形形色色的数据库及互联网服务平台构成了庞大且异构的数据源。这些数据种类多样、来源各异，可能包含位置信息、照片、音频、视频等结构化和非结构化数据，几乎囊括了所有的数据类型。

• 关联缺：现实生活中，各种对象之间存在着千丝万缕的联系。但是有些数据在不同平台独立存储、独立维护，隔断了彼此之间的相互关联，形成数据孤岛，致使人们难以发现它们之间的关联关系、相关性或因果结构。

为应对数据碎片化带来的问题，数据分析必须能够连接数据孤岛，分析挖掘数据对象之间的关联，融合异构数据源中的数据，提升数据质量，最终体现数据的价值。

数据挖掘

数据挖掘是将记录下来的海量数据作为挖掘对象，通过使用一系列的方法、工具或者算法，发现数据中隐含的、未知的有价值信息（如各类规则、规律或者趋势等），用来帮助人们解决现实生活中的问题，或者为相关领域的决策提供支持。从技术上来说，数据挖掘起源于数据库中的知识发现，是传统的人工智能与数据库管理系统的交叉领域。数据挖掘任务一般可以分两类：描述性挖掘任务和预测性挖掘任务。描述性挖掘任务用于刻画数据库中数据的一般特性；预测性挖掘任务则是在当前数据的基础上进行推断。

1. 常用数据挖掘方法

用于数据挖掘基础分析的方法主要有关联规则、分类、聚类等。

（1）关联规则

关联规则就是发现数据背后存在的某种规则或者联系，它描述了一个事物中某些属性同时出现的规律和模式。关联规则的经典算法是 Apriori 算法。在大数据时代，数据量越来越大，项集越来越多，因此又出现了许多针对 Apriori 算法进行改进的并行优化算法，如基于划分、哈希、采样等的一些算法。

（2）分类

分类是指找出描述并区分数据类或概念的模型（或函数），以便能够使用模型预测类标记中未知的对象类。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可准确描述的模型，并利用这些模型预测新数据所属的类。

（3）聚类

聚类是指将对象集合分成由类似对象组成的多个类的过程。与分类不同的是，聚类对已知的数据不提供类标记，对象根据最大化类内的相似性、最小化类间的相似性原则进行聚类或分组，形成簇。由聚类所组成的簇是一组数据对象的集合，这些对象与同一簇中的对象彼此类似，与其他簇中的对象相异。聚类任务的一般步骤如图 3-16 所示。

图 3-16　聚类任务的一般步骤

2. 数据挖掘的一般过程

数据挖掘的一般过程通常分为业务需求分析、数据准备、数据挖掘分析、结果分析与评估四个部分（图 3-17），也可以使用整体解决方案来进行。

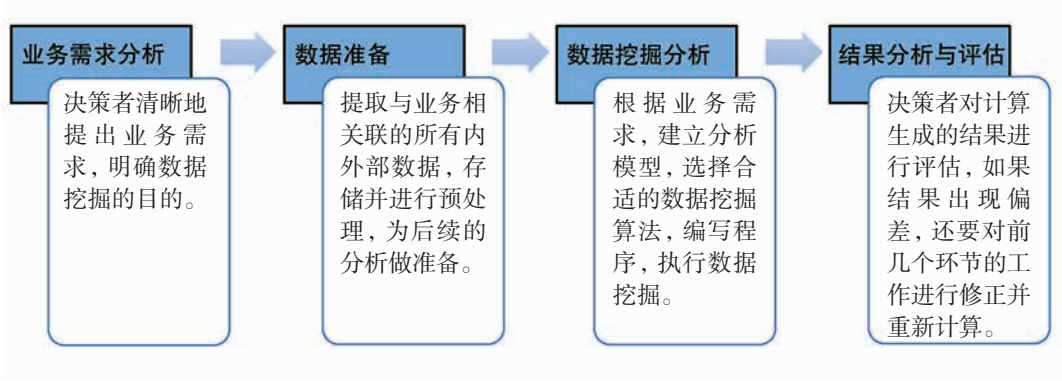


图 3-17　数据挖掘的一般过程

3. 协同过滤推荐

协同过滤是推荐算法中最经典的一种类型，用来预测目标用户对特定商品的喜好程度，系统根据这一喜好程度对目标用户进行推荐。协同过滤推荐也是一种数据挖掘方法，分为三种类型，分别是基于用户的协同过滤推荐、基于物品的协同过滤推荐和基于模型的协同过滤推荐。

（1）基于用户的协同过滤推荐

基于用户的协同过滤推荐，其主要思想可以表述为“和你喜好相同的人喜欢的物品，你也很有可能喜欢”，通过对用户之间信息的处理，挖掘出与当前用户相似度高的用户集合，然后将用户集合中评分高且当前用户没有评分的物品推荐给当前用户。基于用户的协同过滤推荐，其算法步骤可以描述为：

步骤一：采集用户数据

采集可以代表用户兴趣的数据。根据平台累积的用户行为数据完成用户对物品的评分。

步骤二：针对当前用户的相似用户搜索

利用相似度算法对平台用户的数据进行处理，计算用户之间的相似度。

步骤三：产生推荐结果

有了相似用户的集合，就可以对当前用户的兴趣进行预测，产生推荐结果。依据推

荐目的的不同，可以进行不同形式的推荐。较常见的推荐结果有 Top-N 推荐和关系推荐。Top-N 推荐针对个体用户，对每个人产生不一样的结果。关系推荐则是对最相似用户的记录进行关系规则挖掘。

（2）基于物品的协同过滤推荐

基于物品的协同过滤推荐，其主要思想可以表述为“能够引起用户兴趣的物品，必定与其评分高的物品相似”，通过对平台上积累的物品数据进行处理，挖掘出与当前物品相似的物品集合，再结合用户对当前物品的评分，向用户推荐与他们评分高的物品相似度高且用户没有评分的物品。基于物品的协同过滤推荐，其算法步骤可以描述为：

步骤一：采集用户数据

与基于用户的协同过滤推荐相似，在进行挖掘之前需要采集与用户兴趣相关的数据。

步骤二：针对当前物品的相似物品搜索

利用相似度算法对平台上物品的数据进行处理，挖掘出与当前物品相似度高的若干物品。

步骤三：产生推荐结果

结合用户对当前物品的评分，向用户推荐与用户评分高的物品相似度高的其他物品。

（3）基于模型的协同过滤推荐

基于模型的协同过滤推荐是目前最主流的协同过滤推荐类型，也是比较复杂的一类。它的主要思想可以概括为：基于样本用户的喜好数据，训练出一个推荐模型，根据实时的用户喜好数据进行预测推荐。以视频网站为例，根据视频网站中已有用户和视频数据可以训练出一个推荐模型。新用户注册时，虽然在网站平台上该用户对视频的操作数据是空白，但是视频网站仍然可以利用训练出的模型向新用户推荐视频。

单元挑战　分析在线社交平台用户情况

一、项目任务

随着互联网的发展和普及，人们越来越多地使用网络进行信息传递和交流，新的在线社交平台不断涌现（图 3-18）。人与人之间的网络连接形成的复杂在线社交网络，在社会结构研究中具有举足轻重的地位。

以小组为单位，收集一个在线社交平台的用户数据，综合运用四种基本分析方法分析平台用户的特征，并使用数据挖掘方法分析活跃用户的特征，以可视化的形式呈现分析结果。

图 3-18　社交平台

二、项目指引

1. 采集数据：选择一个开放的在线社交平台，利用爬虫工具或手工收集平台用户的相关数据（如用户 ID、所在地区、性别、发帖数、粉丝数等）。

2. 数据分析及可视化：

• 按天统计用户的平均发帖数，并选择适当的可视化形式呈现。

• 按所在地区或性别等用户属性，分组统计发帖数，并进行对比分析。

• 计算并可视化呈现用户发帖数与各用户属性间的相关性，由此分析出哪些用户属于活跃用户。

• 通过挖掘用户之间的关系（如好友、关注、粉丝）建立连接，构建社交网络，并进行可视化呈现。

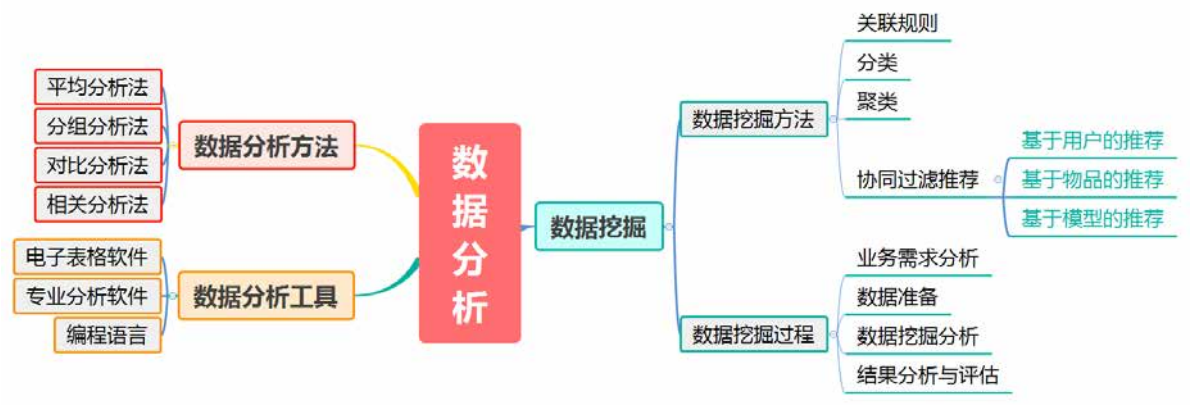
3. 撰写报告：完成小组报告，呈现数据分析的目的、过程及结果。

三、交流评价与反思

各小组派出代表，通过网络或课堂展示交流小组报告，并从选用的数据分析工具是否恰当、能否准确呈现分析结果等方面对其他小组的报告进行评价。

单元小结

一、主要内容梳理



二、单元练习

1. 为落实“健康中国 2030”国家战略，某市医疗改革小组打算对该市的医疗服务进行改革创新，让患者得到更加优质的服务。假设你是改革小组成员，请尝试针对以下问题给出具体解决方案。

（1）选择合适的数据分析方法对就医高峰期（天、月、年）进行分析，并收集外来患者的就医路线，合理安排交通布局。

（2）分析各医院科室接待患者人次及其所占比重，根据现有医疗资源状况制定人才引进方案。

（3）根据患者的实际情况和医疗记录，用协同过滤算法为患者推荐合适的医院和医生。

2. 在数据分析的过程中，分析结果的呈现形式通常与分析任务的目标密切相关，请上网收集资料，对两者的关系进行总结，然后尝试通过图形的组合（如箱线图和柱状图的组合），对分析结果进行更深入的呈现。

3. 在教育领域，推荐系统可以应用在个性化精准教育资源推荐上，请考虑以下问题，并尝试设计一个推荐系统。

（1）针对教育资源推荐这一应用场景，哪种推荐算法较为合适？为什么？

（2）基于上面选择的推荐算法，需要收集哪些数据？

（3）请设计一个简单可行的推荐流程。

三、单元评价

评价内容	达成情况
掌握四种基本数据分析方法（A、T）	
认识数据分析的基本流程（A、T）	
掌握箱线图、密度图、直方图、散点图、六边形箱式图的可视化方法（A、T）	
掌握利用折线图展示数据相关性的方法（A、T、I）	
了解三种相关系数的计算方法（A、T）	
了解大数据时代数据管理与分析技术的新发展（A、R）	
认识关联规则、分类、聚类等常用数据挖掘方法（A、T）	
了解数据挖掘的一般过程（A、T）	
了解三种类型的协同过滤推荐（A、T、I）	

说明：A—信息意识，T—计算思维，I—数字化学习与创新，R—信息社会责任

第四单元

数据备份与数据安全

作为信息化的基础，无论是社会治理、科学研究还是商业应用，都离不开数据。但是数据在存储、处理、传输等过程中会面临各种安全风险。一旦发生数据丢失、数据泄露等数据安全问题，轻则损害企业利益，重则影响到国家安全。因此，保证数据安全不仅是企业的责任，而且是全社会的责任，与我们每个人都息息相关。

数据备份是实现数据安全的一种重要策略，可以运用数据备份和数据还原手段来防止数据丢失。在实际应用中，应该选择合适的备份方法，或者选用合理的数据库系统容灾方案来应对数据安全问题。

在本单元中，我们将认识数据丢失的风险和数据安全的重要性，掌握常用的数据备份方法，学会衡量不同备份方法的优劣，并了解应对数据安全问题的常用策略。

单元挑战

探索 MySQL 数据库的实时备份

学习目标

- 结合案例，认识数据安全和数据丢失风险。
- 利用实时备份、定时备份、全备份、增量备份和差异备份等方法进行数据备份。
- 了解如何优化数据库系统容灾方案。

项目九

探秘网上书店数据库系统容灾方案
——应对数据丢失风险

在信息时代，数据即财富。守护好数据、防止数据丢失成为数据安全问题的重中之重。在实际工作中，难免会遇到数据误删除、误修改等情况，软硬件故障和一些自然因素也会导致数据丢失。尤其是对企业来说，数据丢失会带来严重后果，造成巨大经济损失，甚至可能导致破产倒闭（图 4-1）。

对企业来说，在数据丢失后的数据恢复能力将决定它的生死存亡。没有数据库的备份，就没有数据库的恢复。因此，企业应当重视数据备份工作，为其系统选择合适的备份设备和技术，从而避免可能发生的重大损失。

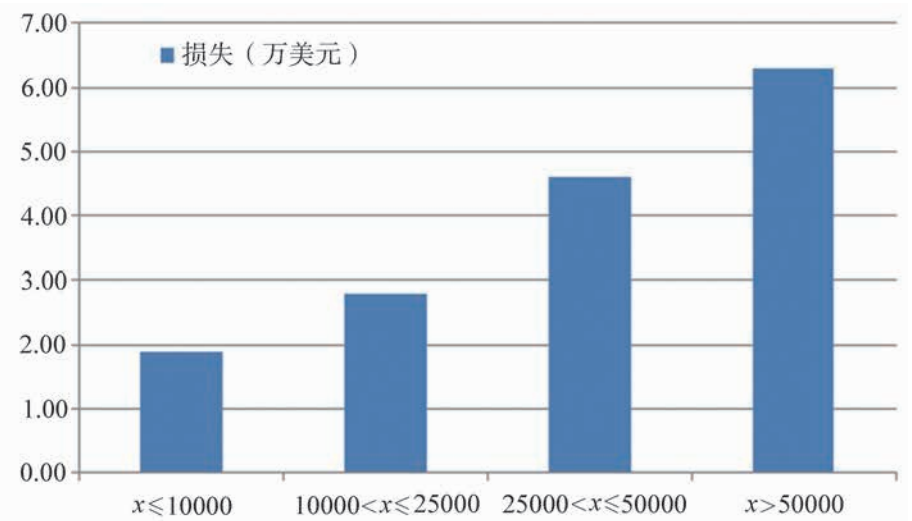


图 4-1　数据记录丢失数量对平均损失的影响

项目学习目标

在本项目中，我们将围绕数据丢失现状，通过数据丢失案例分析，认识数据丢失的原因，并探索应对数据丢失的容灾方法。

完成本项目学习，须回答以下问题：

1. 数据丢失的风险有哪些？
2. 如何实现实时备份、定时备份、全备份、增量备份和差异备份？
3. 如何衡量和评价不同数据库系统备份方法的优劣？
4. 大数据时代，如何优化数据库系统容灾方案？

项目学习指引

1. 了解数据丢失风险

随着计算机网络的飞速发展，数据安全的重要性日趋明显。在大数据时代，人们以从前无法想象的方式从海量数据中分析出有价值的信息，成为合理决策的有力武器，但海量数据实时、跨机构的流动，也导致传统的数据安全保障模式被打破，数据安全事件经常出现。数据安全不仅关系到国家的政治经济安全、国防安全和文化安全，也与每个人的工作和生活息息相关。因此，维护数据安全是全社会的责任，不仅需要政府、企业、社会组织的参与，也需要普通民众的共同参与。

核心概念

数据安全（data security）是指系统中的数据受到保护，不因偶然或恶意的因素而遭到破坏、更改或者泄露。

参见 P100 知识链接“数据安全”

数据安全问题涉及数据来源安全、数据存储安全、数据传输安全等。只要发生数据存储、数据传输和数据交换，就有可能产生数据安全问题，这时如果没有恰当的恢复手段和措施，就会导致数据丢失。

核心概念

数据丢失（data loss）是指信息系统中由于信息存储、传输或者处理过程中出现的故障导致数据不可用。

造成数据丢失的原因主要包括自然因素、硬件故障、软件损坏及人为错误等几个方面。经历过敏感数据丢失的企事业单位往往面临着非常大的风险，轻则造成经济损失，重则直接宣告破产。

据某国际知名咨询公司的调查数据显示，在经历了数据完全丢失而导致系统宕机的企业中，有 2/5 再也没能恢复运营，余下的企业也有 1/3 在两年内宣告破产。也就是说，六成企业因数据完全丢失而倒闭。另一份国际咨询公司的研究报告中提到，经历数据丢失的企业将导致其客户量及相关收入降低 8%；对于上市企业而言，每股股价会下跌 8%；平均每丢失一个客户记录便会造成 600 多元的额外损失。

参见 P100 知识链接“数据丢失”

思考与讨论

你知道哪些国内外数据丢失的案例？通过这些案例，你能总结出哪些数据丢失的风险？

活　动

9.1　分析数据丢失可能造成的影响。

（1）分析数据丢失对企事业单位可能造成哪些影响。

（2）结合自己的生活经验，分析个人数据丢失可能造成哪些影响。

2. 备份网上书店数据

核心概念

数据备份（data backup）是指为防止计算机系统出现操作失误或系统故障导致数据丢失，而将全部或部分数据集合从应用主机的硬盘或阵列中复制到其他存储介质上的过程。

通常的保护措施都只能尽量减少而不能完全杜绝数据安全问题。为保证数据安全，可以运用数据备份手段防止数据丢失，或者通过授权访问方式防止数据泄露、丢失。数据备份是为了在发生数据丢失后，能够通过之前的备份文件完整、快速、简捷、可靠地恢复原有数据和系统。备份时，通常将数据库系统中的数据复制到其他存储介质上。当发生数据丢失时，通过调用保存在其他存储介质上的数据，即可部分或者完全恢复丢失的数据。

常用的数据备份方法有全备份、增量备份和差异备份。下面通过研究销售数据库中图书的库存变化，了解各种不同备份方法的区别。

上架图书从周一到周四的库存变化都记录在销售数据库中，数据库中存储的库存表如图 4-2 所示。其中浅绿色表示初始状态，白色表示库存没有发生变化，黄色表示库存被更新，红色表示库存记录被删除，紫色表示新增库存记录。

编号	周一	周二	周三	周四
A	24	22	18	15
B	40	40	48	48
C	35	35	35	
D		106	106	104
E			15	15
F				24

图 4-2　库存表

编号为 A 的图书，其库存经历了周二 A1 版本，周三 A2 版本，周四被更新为 A3 版本；编号为 B 的图书，其库存周二

没有变化，周三被更新为 B1 版本，周四没有变化；编号为 C 的图书，其库存周二、周三没有变化，周四被删除；编号为 D 的图书，其库存记录在周二新增，周三没有变化，周四被更新为 D1 版本；编号为 E 的图书，其库存记录在周三新增，周四没有变化；编号为 F 的图书，其库存记录在周四新增。

假定周一做了一次全备份，以后每天都做一次备份。选用不同的备份方法时，周一到周四的备份文件中包含的数据记录如图 4-3 所示：

图 4-3　数据库不同备份方法比较

周四的数据库中包含了编号为 A、B、D、E、F 的五本图书的库存信息。以该天为例，用三种备份方法分别对库存信息进行备份。

全备份将这五本书的库存信息都做了备份，数据还原效率高。但是全备份产生的备份文件中会包含很多冗余信息，备份耗时长、效率低。

增量备份克服了全备份的缺点，它只备份自上一次备份之后发生变化的数据。相对于周三，图书 A 和 D 的库存有更新，新书 F 上架，因此增量备份文件中只包含 A3、D1、F 三个库存信息，备份耗时短、效率高。但是数据还原时，需要使用之前所有的增量备份文件，耗时较长。

差异备份是将所有自上一次全备份之后发生变化的数据都进行备份，数据还原效率较高。相对于上一次全备份的周一，图书 A 和 B 的库存有更新，图书 C 被删除，而且有三本新书 D、E、F 上架，因此差异备份文件中包含 A3、B1、D1、E、F 五个库存信息。

以上三种备份方法都属于定时备份，即在规定的时间执行数据备份。通过选择适当的备份时机，定时备份可以降低对业务停滞时间的影响。另外还有一种备份策略叫实时备

小贴士

数据还原（data recovery）又称数据恢复，是指通过技术手段，将保存在磁带、磁盘等存储介质上的数据进行抢救和恢复的技术。

参见 P101 知识链接“数据备份与数据还原”

份，即只要数据记录发生变化，便会实时、准确地执行数据备份任务，从而实现连续的数据保护。

为了便于备份数据库，绝大部分数据库产品都提供了用于数据库备份的工具。以 Windows 环境下的 MySQL 数据库备份为例，编写一个名为 MySQL_backup.bat 的批处理文件，在不关闭数据库的情况下定时备份数据库。使用 Windows 的“计划任务”，可以每天定时执行该批处理文件，备份文件可以按每天的日期来区分。

目前，一些互联网企业推出“云备份”服务，来帮助企事业单位防止数据丢失。云备份服务支持多平台管理，让数据更加安全，数据加密传输更放心，而且不受空间和设备限制。云备份大大减少了各企事业单位在信息化过程中的硬件投资和机房能耗等成本，提高了人力资源使用效率，为各企事业单位节约资源、提高效率、保证数据安全发挥了积极作用。

数字化学习

请自主学习如何在 Windows 环境下设置计划任务，以完成数据库定时备份的功能。

小贴士

云备份（cloud backup）主要指以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服务。

思考与讨论

1. 增量备份的效率最高，它是不是最好的备份策略？全备份、全备份 + 差异备份、全备份 + 增量备份这三种方案各有哪些优缺点？

2. 磁带、光盘等备份介质分别有什么特点？在选择备份介质时要考虑哪些因素？

活　动

9.2　了解并运用 MySQL 数据库的数据备份与还原方法。

（1）了解数据库备份命令及恢复命令。

（2）尝试备份并还原由 mysqldump 命令产生的数据库备份文件。

9.3　比较不同的数据备份方法，分析备份方法选择策略。

（1）选择什么数据备份方法能实现备份介质中数据损失量尽可能小的目标？

（2）不同数据备份方法对业务停滞时间分别有多大影响？

（3）除了备份介质中的数据损失量和业务停滞时间，选择备份策略还需要考虑哪些因素？

3. 优化数据丢失防范方案

目前，很多大型互联网公司和金融企业都采用分布式的数据中心来存储数据。可一旦数据库发生故障，哪怕只是恢复数据中心的一部分数据，也要花费几小时到几天时间。更为甚者，若是发生地震、水灾或火灾等，备份数据也可能会丢失。因此，如何优化数据丢失防范方案成为一个亟待解决的问题。

（1）实现数据冗余

大型互联网公司的数据中心由成千上万个节点组成。随着数据中心节点数量的增加，它们发生故障的可能性也大大增加。如果数据只存储一份，那么一个节点的故障就会导致大量数据丢失或者损坏。而且数据中心越大，数据损坏或者丢失的可能性也就越大。

数据冗余是常用的防止数据丢失的方法，指同一个数据在系统中多次重复出现。最简单的实现数据冗余的方法是复制每个节点中的数据。如果一个节点发生故障，那么可从另外一个节点中恢复数据。而两个节点同时发生故障的概率非常低，几乎可以忽略不计。

（2）实施异地多活容灾方案

为了保证数据不丢失，且能在发生故障时缩短业务停滞时间，规模稍大的互联网企业一般都对数据中心实施了异地多活容灾方案。

异地多活容灾方案中的“异地”通常指在不同城市建立的独立数据中心，不同数据中心之间保持数据一致；“多活”则是指这些数据中心会分担日常业务，当主数据中心发生故障时，异地数据中心能够迅速接管业务。考虑到断电、网络中断、自然灾害可能导致同一个地点的所有节点同时宕机，异地多活容灾方案通常会选择在距离足够远的不同城市部署多个数据中心。

异地多活容灾方案实现了较高的容灾等级，通过运用数据冗余技术保证了不同数据中心之间数据的远程同步。该方案除了能避免自然灾害外，更重要的是通过数据复制几乎可以实现数据零丢失；而且通过“多活”使得多个数据中心能够同时对外提供服务，将业务停滞时间降到分钟级甚至秒级，从而保证了数据库系统的持续服务能力。

小贴士

数据中心（data center）是用于组织内部或者组织间协作的特定网络设备，在该网络设备上可以分享、展示、计算、存储数据信息。

思考与讨论

通过数据复制的方式实现数据冗余会对业务停滞时间产生什么影响？

活　动

9.4　了解异地多活容灾方案。

（1）比较异地多活容灾方案与传统的备份方法，分析异地多活容灾方案的优缺点。

（2）结合自己的生活经验，分析异地多活容灾方案的适用范围。

知识链接

数据安全

数据安全是指系统中的数据受到保护，不因偶然或恶意的因素而遭到破坏、更改或者泄露，系统能够连续、可靠、正常地运行，信息服务不中断。

数据安全通常包含数据本身的安全和数据防护的安全。数据本身的安全主要是指采用现代密码算法对数据访问进行控制，如数据保密性、数据完整性、双向强身份认证等。数据防护的安全主要是指采用现代信息存储手段对数据进行主动防护，如通过磁盘阵列、数据备份、异地容灾等手段保证数据的安全。

数据丢失

数据丢失是指信息系统中由于信息存储、传输或者处理过程中出现的故障导致数据不可用。

造成数据丢失的原因主要包括以下几个方面：

（1）自然因素：计算机、服务器或数据中心遭受如火灾、水灾、地震、战争等灾难后宕机，致使数据丢失。

（2）硬件故障：计算机、服务器或数据中心因突然断电、磁盘失效、电压不稳、网络连接中断等故障，致使数据丢失。

（3）软件损坏：软件原因造成的数据丢失现象一般表现为操作系统丢失、磁盘读写错误、找不到所需要的文件、文件打不开、文件打开后乱码、硬盘没有分区、硬盘被锁等。

（4）人为错误：人为错误主要表现为随意更改存储系统和应用程序信息的注册表单，以及更改系统文件的属性和存储位置，导致数据丢失和系统崩溃，从而无法对外提供服务。

数据备份与数据还原

1. 数据备份

数据备份是容灾的基础，是指为防止计算机系统出现操作失误或系统故障导致数据丢失，而将全部或部分数据集合从应用主机的硬盘或阵列中复制到其他存储介质上的过程。为了保障生产、销售、开发的正常运行，用户应当采取有效措施对数据进行备份，防患于未然。

2. 数据备份的主要方法

（1）定时备份与实时备份

在备份时机的选择上，有定时备份和实时备份两种策略。人们大都采用定时备份的方法，即在规定的时间对数据进行备份。但是，只要数据备份存在时间间隔，一旦发生数据丢失，备份间隔之内的数据极易丢失。数据备份的时间间隔越大，丢失的数据也就越多。实时备份会对数据库进行自动监控，只要数据记录发生变化，便会实时、准确地执行数据备份任务，从而实现连续的数据保护。实时备份会占用系统资源，对业务停滞时间会造成一定影响。

（2）全备份、增量备份与差异备份

按备份内容分，定时备份主要有全备份、增量备份与差异备份三种方法。

全备份是指对整个系统（如组成服务器的所有卷）或用户指定的所有文件数据进行全面的备份。全备份的好处是直观、易理解，出现数据丢失等问题时可以只使用一份备份文件就快速完成数据还原。其不足之处也很明显：因为要备份所有的数据，每次的工作量都很大，且需要占用大量备份介质。如果频繁进行全备份，备份文件中会有大量重复数据，占用大量存储空间，增加备份成本。如果需要备份的数据量较大，读写操作所需的时间也会较长。因此，全备份通常每隔一段较长的时间进行一次。一旦发生数据丢失，只能使用上一次的备份数据，还原到当时的状况，在备份间隔之内更新的数据有可能丢失。

增量备份是指只备份上一次备份后新产生或更新的数据。在特定的时间段内一般只有少量的数据发生改变，因此增量备份中没有重复的备份数据，既节省了存储空间又缩短了备份时间。这种备份方法比较经济，可以频繁进行，但是一旦发生数据丢失或文件误删除，数据还原工作会比较麻烦。一般需要找出最后一次全备份之后的所有增量备份文件，将记录在一次或多次增量备份中的改变全部恢复，才能完成数据还原。增量备份的备份文件就像链子一样一环套一环，其中任何一个备份文件的磁带、磁盘出现问题，都会导致整条链子脱节，可靠性较低。数据还原过程中需要使用全备份的数据，因此，所有的增量备份都是在最近一次全备份以后执行的。

差异备份是指只备份上一次全备份后新产生或更新的数据。采用差异备份方法时，数据还原只涉及两份备份文件，可以简化还原的复杂性，同时避免了全备份和增量备份两种方法的缺陷。因为无需频繁进行全备份，差异备份工作量较小，备份所需时间较短，能够

节省存储空间。虽然差异备份的工作量比增量备份略大，但是它的数据还原相对简单。系统管理员只需要对上一次的全备份文件和最近一次的差异备份文件进行恢复，就可以完成数据还原。

3. 数据还原

数据还原又称数据恢复，是指通过技术手段，将保存在磁带、磁盘等存储介质上的数据进行抢救和恢复的技术。如果做过数据备份，那么一旦发生数据丢失，通过备份进行数据还原将十分简单。大部分数据备份系统都包含有数据还原的功能，甚至可以自动完成对数据及其运行环境的还原。

拓展阅读

信息系统灾难恢复能力等级

信息系统灾难恢复能力等级与恢复时间目标（RTO）和恢复点目标（RPO）具有一定的对应关系，各行业可根据行业特点和信息技术的应用情况制定相应的灾难恢复能力等级要求和指标体系。

某行业 RTO/PRO 与灾难恢复能力等级的关系示例见表 4-1。

表 4-1　RTO/RPO 与灾难恢复能力等级的关系

灾难恢复能力等级	RTO	RPO
1	2 天以上	1 天至 7 天
2	24 小时以上	1 天至 7 天
3	12 小时以上	数小时至 1 天
4	数小时至 2 天	数小时至 1 天
5	数分钟至 2 天	0 到 30 分钟
6	数分钟	0

——摘自《信息系统灾难恢复规范》（GB/T 20988-2007）

单元挑战　探索MySQL数据库的实时备份

一、项目任务

配置主从 MySQL 服务器系统，实现 MySQL 数据库系统的双机实时备份（图 4–4）。

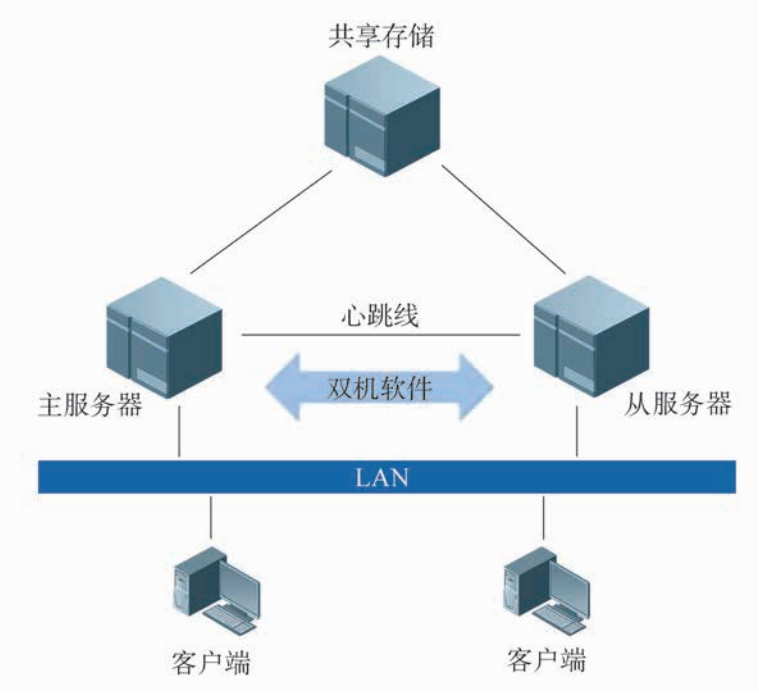


图 4–4　双机实时备份

二、项目指引

1. 以小组为单位查找资料，了解 MySQL 数据库的主从架构，学习配置主从 MySQL 数据库系统。

2. 理解实现 MySQL 数据库实时备份的原理。

3. 通过操作实现主从 MySQL 数据库系统的双机实时备份。

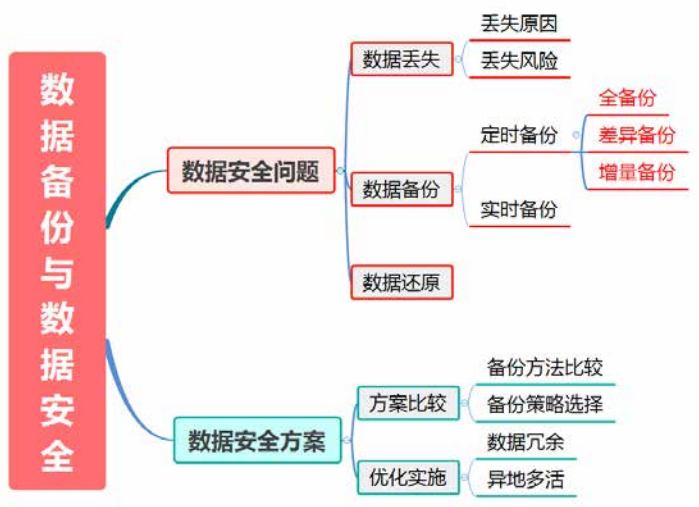
4. 验证主从 MySQL 数据库系统的双机实时备份。

三、交流评价与反思

各小组用演示文稿等方式制作电子作品，通过网络或课堂展示研究成果，并对其他小组的作品进行评价。

单元小结

一、主要内容梳理



二、单元练习

1. 数据库备份的介质主要有哪些？请比较这些备份介质的优缺点。

2. 假设销售数据库记录了上架图书从周一到周三的库存变化，存储在 MySQL 数据库中，如图 4-5 所示。其中浅绿色表示初始状态，白色表示库存没有发生变化，黄色表示库存被更新，红色表示库存记录被删除，紫色表示新增库存记录。假定在周一时，做了一次全备份，以后每天都做一次备份。

（1）每天的差异备份、全备份和增量备份分别怎么做？

（2）请分析比较差异备份、全备份和增量备份的优缺点。

编号	周一	周二	周三
A	24	22	18
B	40	40	48
C	35	35	
D		106	106
E			15

图 4-5　库存表

三、单元评价

评价内容	达成情况
认识数据丢失风险（A、T、R）	
理解数据丢失的原因（A、R、I）	
知道防止数据丢失的方法（A、T）	
理解不同数据备份方法的差别（A、T）	
认识数据安全的重要性（A、R）	
能利用多种方法进行数据备份（A、T、I）	
具备数据备份与还原意识，明确在数据安全中的社会责任（A、R）	
知道数据库容灾等级及异地多活容灾方案（A、T、I、R）	

说明：A—信息意识，T—计算思维，I—数字化学习与创新，R—信息社会责任

附 录

部分名词术语中英文对照

（以汉语拼音字母次序为序）

半结构化数据　semi-structured data
对比分析法　comparative analysis
非结构化数据　unstructured data
分组分析法　group analysis
关联规则　association rules
关系数据库　relational database
记录　record
结构化查询语言　Structured Query Language, SQL
结构化数据　structured data
可扩展标注语言　eXtensible Markup Language, XML
联系　relation
六边形箱式图　hexagonal boxplot
密度图　density plot
平均分析法　average analysis
散点图　scatter plot
实体　entity
实体集　entity set
属性　attribute
数据安全　data security
数据备份　data backup
数据丢失　data loss
数据分析　data analysis
数据还原　data recovery
数据可视化　data visualization
数据库　database
数据库管理系统　Database Management System, DBMS
数据模型　data model
数据挖掘　data mining
数据稀疏　data sparsity
数据中心　data center
相关分析法　correlation analysis
箱线图　box plot
协同过滤推荐　collaborative filtering recommendation
信息过载　information overload
云备份　cloud backup
噪声数据　noisy data
折线图　line chart
直方图　histogram
主关键字　primary key
字段　field

PUTONG GAOZHONG JIAOKESHU
XINXIJISHU

普通高中教科书
信息技术　选择性必修3
数据管理与分析
上海科技教育出版社有限公司出版发行
（上海市闵行区号景路159弄A座8楼　邮政编码201101）
湖南省新华书店经销　湖南长沙鸿发印务实业有限公司印刷
开本890×1240　1/16　印张7
2021年1月第1版　2021年12月第3次印刷
ISBN 978-7-5428-7400-9/G·4340
定价：8.98元
批准文号：湘发改价费〔2017〕343号　举报电话：12315

此书如有印、装质量问题，请向印厂调换
印厂地址：长沙黄花印刷工业园三号　电话：0731-82755298