普通高中教科书

信息技术

选择性必修3

数据管理与分析

浙江教育出版社
ZHEJIANG EDUCATION PUBLISHING HOUSE

普通高中教科书

信息技术

选择性必修3

数据管理与分析

闫寒冰 主编

浙江教育出版社·杭州

主　　编：闫寒冰
副 主 编：赵　健　魏雄鹰

本册主编：何海源
编写人员（按姓氏笔画排列）：
吴海忠　何海源　陆海丰　陈　跃
周亚利　虞颖健

前　言

信息技术作为当今先进生产力的代表，已经成为我国经济发展的重要支柱和建设网络强国的战略支撑。在这样的大背景下，教育部全面修订并颁布了《普通高中信息技术课程标准（2017年版）》，为这门课程设定了与新时代相符的育人目标：帮助学生掌握信息技术基础知识与技能、增强信息意识、发展计算思维、提高数字化学习与创新能力、树立正确的信息社会价值观。

本套教材依据《普通高中信息技术课程标准（2017年版）》编写，包括两本必修教材《数据与计算》《信息系统与社会》，六本选择性必修教材《数据与数据结构》《网络基础》《数据管理与分析》《人工智能初步》《三维设计与创意》《开源硬件项目设计》，两本选修教材《算法初步》《移动应用设计》。

本套教材的编写组汇集了来自信息技术、课程与教学、教育技术等领域的高校学者与教学一线专家。编者们通力合作，从课程内容、教材体例、技术选择、教学方法、学习方法等方面精心打磨，期待以最专业的样态帮助学生达到课程预期的育人目标。

具体而言，本套教材体现了如下特点：

1. 体例上——为核心素养的培养创造空间和条件：将核心学习内容与支持学习的方法有机融合在一起，支持学生在自主、合作、探究的学习情境下发展核心素养。

2. 内容上——体现概念、内容与方法的精准与专业：在增强教材可读性的同时，精炼提升综合素养所必需的核心内容，强调所有概念、内容与方法的精准与专业。

3. 活动上——着力提升学生的高级思维能力：精心设计与布局教材中的练习、思考、讨论、实践与项目学习，追求对高级思维能力的培养。

4. 案例上——体现信息科技的多层需求与多维格局：把案例的呈现作为开阔视野的重要手段，帮助学生理解信息技术对于社会发展所具有的价值与意义。

5. 技术上——引领学生拓宽视野与发展思维：将每种具体应用软件都作为解决某些问题的一条路径来看待，期待学生通过具体的技术操作体验，理解其背后的原理与格局、特点与局限，拓宽视野、发展思维。

本册教材为选择性必修《数据管理与分析》。通过本教材的学习，学生应了解数据管理与分析技术，能根据需求分析形成解决方案；能选择一种数据库工具对数据进行管理，从给定数据中提取有用信息并应用于实际问题的解决；在活动过程中形成对数据特征、数据价值、数据管理思想与分析方法的认识。

就教材本身所讲述的知识内容而言，我们相信，只要同学们潜心自学就可以基本掌握。但“知识内容”只是发展信息技术核心素养的基础部分，所以，我们希望同学们不要仅满足于对具体知识与具体技术的掌握，还要重视教材中的各类学习活动，与老师和学友一起，更多地去创造、研究、解决问题、制作、交流、合作和评价，唯有如此，同学们才能藉由这门课程的学习全面地提升信息素养，增强在信息社会的适应力与创造力，为实现中华民族伟大复兴的宏伟目标做出更大贡献！

本册教材在编写过程中得到了各方面的大力支持。北京大学计算机系李晓明教授、浙江大学计算机学院卜佳俊教授和翁恺教授、北京航空航天大学欧阳元新副教授在百忙之中对书稿内容进行了审阅。范建农、杨琦霞、钟明金、管永根、滕春毅、徐建强等多位高中信息技术教师对书稿内容提出了宝贵的修改意见。

由于水平有限，本书可能还存在不足之处。希望大家在教材使用过程中，能够及时将意见和建议反馈给我们，对此，我们深表谢意。

目 录
MULU

第一章 1

数据管理与分析概述

数据正在以指数级的速度增长。对数据进行科学的管理与分析，可以挖掘出数据潜在的价值，帮助人们提高办事效率、优化服务能力、提高决策水平。电子商务、社交网络、移动通信等各种数据应用，正以一种前所未有的方式改变着人们的生活。

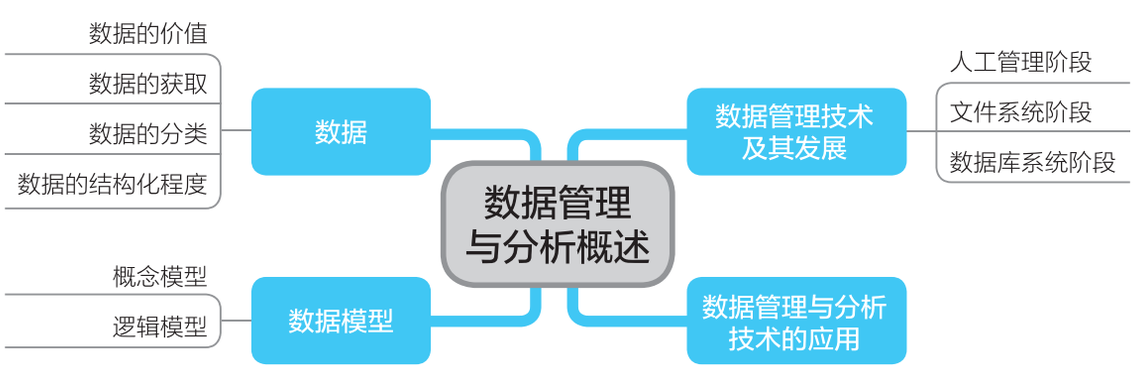
问题与挑战

- 在生活中，人们常常使用手机APP寻找附近的风景名胜、交通线路、餐厅或者超市。若要寻找某特色餐馆，相关APP会呈现哪些信息？这些数据是从哪里来的？
- 当当、亚马逊、中国图书网等网上书店，这些网站页面上向用户呈现了书目、价格、库存等数据。网上书店是如何管理这些数据的？在运营过程中还会产生哪些有价值的数据？
- 校园微课视频网站借助网络平台，以视频为载体，提供数字化学习服务。教师可以上传微课视频、教案、习题等学习资源，学生可以自主学习与检测。校园微课视频网站如何管理账号、密码、教案、视频、评论等数据？

学习目标

1. 能根据实际需求，选择适当的方式采集数据。
2. 能区别不同结构化程度的数据，理解其管理和应用上的特点。
3. 了解常见的数据模型及其应用。
4. 了解数据管理技术及其发展。
5. 了解数据管理与分析技术的应用及价值。

内容总览



1.1 数据

数据以各种形式存在于人们生活和工作的方方面面，已经成为一种新型的资源。随着社会信息化程度不断提高，数据来源的渠道越来越多，获取数据的手段也逐渐多样化，数据潜在的价值不断地被释放。

1.1.1 数据的价值

大数据时代，数据在不同的领域发挥着重要的作用。在太空中，有上千颗人造地球卫星，这些卫星对地球进行全方位的观测，并将观测到的数据传回地球，用于科学探测、通信传播等。在大气层中，每时每刻都有大量的飞机在飞行，每架飞机都会不断产生数据，空中交通管制系统通过处理大量的飞行数据，保障安全和维护空中交通秩序。在地面上，有数以千万计的传感器每分每秒都会产生数据，这些数据被广泛应用于工业、农业、医疗卫生、公共安全等领域。在地面下，有很多地下管线，包括具有探测作用的传感器，这些传感器能实时采集和传输数据，如图1.1.1所示。在人们的日常活动中，网上购物、网络社交、交通出行等一系列行为都会产生数据。这些数据可以用于分析用户行为，实现精细化运营。

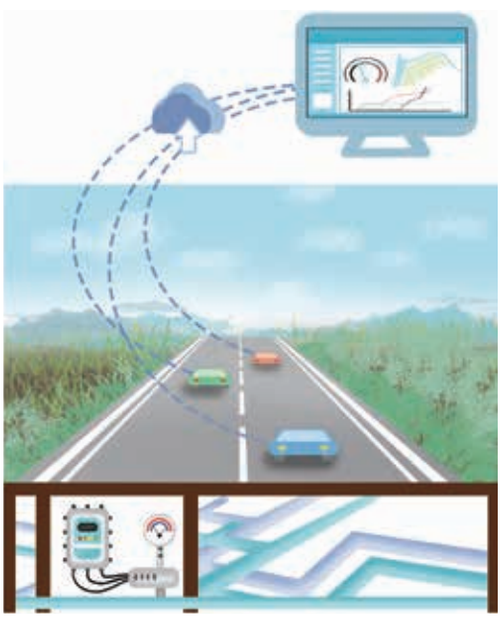


图1.1.1 地下管网数据

数据犹如“矿石”一般，蕴藏着价值。在电子商务平台上，买家个人登录信息、搜索、浏览、收藏、交易、评价等行为数据都被网站收集。分析某个买家的用户名登录信息、IP地址数据和收货地址，就可能获取到该买家的性别、年龄、居住地等基本信息，甚至可以根据一段时间内购买的商品信息和消费金额，推断出其喜好、月收入水平等。例如，某买家近期开始关注某一类商品，电子商务系统可以根据他以往的消费水平推荐相应价位的商品。掌握的买家信息越精确，系统推荐就能越精准，成交的可能性也就越大。

大规模的数据中，蕴藏的价值不可估量。在电子商务平台上，某个店铺对所有买家信息进行统计分析，可以得出本店商品主要面向人群的年龄、职业以及喜欢的商品风格取向等，根据这些信息商家可以掌握买家人群定位，并推断出其消费水平、消费需求和消费观念，对店铺商品选款和广告的精准投放起决定性作用，从而形成稳定的客户群。通过预测和掌握贸易变化，商家和制造商可以及时调整经营模式，如收集某电子商务平台中所有商

品点击量和购买量多年的历史数据进行统计、分析，预测需求变化，电商企业可以根据分析结果采取相应的库存策略，以达到库存成本最小化的目的；收集某种商品的所有买家收货地址数据进行统计、分析，得出地理位置分布图，电商企业可以优化仓储地点分布，以降低送货的物流成本。不仅在电子商务方面，数据在其他领域的应用也在不断拓展。如医疗方面，穿戴式的医疗传感器把检测数据传送到医疗中心，医生可以实时监控病人病情的变化。可穿戴助听器式的动态心电图装置，通过耳内检测仪将声音放大，同步采集心电图等数据，并对数据进行处理、分析，医护人员可在远程计算机上查看实时动态心电图。当检测到心率异常时，检测仪自动将心电图数据发送至检测中心，以便及时诊治。

问题与讨论

“数据是数字时代的石油”“通过把数据卖给需要的人，就是实现了数据价值”请对以上观点发表你的看法，分小组讨论数据价值的真正含义。

1.1.2 数据的获取

人们根据解决问题的需要，获取相应的数据。数据获取的途径多种多样，早期一般通过观察、测量、调查、实验等方式。随着信息技术的发展，传感器和互联网成为获取数据的主要来源。

1. 使用设备采集数据

传感器在特殊环境下能够连续进行检测，获得数据。在大规模农作物种植基地，土壤温湿度传感器常常被安装在作物根部土壤中，每单位时间将检测到的土壤温度数据、水分含量数据传输到计算机，便于及时和适量浇灌。智能手机也有很多传感器，其中光线传感器、温度传感器可以采集环境光亮度与室温数据，距离传感器通过红外测距获得手机距离数据，重力传感器可以获取手机姿态的数据，还有加速度传感器、磁场传感器、指纹传感器等。在实际生活中，传感器的应用已经很普遍，如人脸识别设备、雷达生命探测仪、无人驾驶车辆、城市监控摄像头、运动手环、智能手表等。

数码相机拍照时，光学镜头把光线聚焦到影像传感器上，可将捕捉到的景物光信号转换成电信号，通过模/数转换器转换成数字信号，以数据的形式进行存储。半导体指纹采集设备就是根据手指表面指纹凹凸不平，按压时与接触面实际距离不同而产生不同的电容数据，实现指纹数据的采集。高考期间，很多省市采用了指纹识别进行考生身份验证。高考报名时，学校设立信息采集点，考生将手指在指纹采集器的采集模块上按压，把模拟信号转换成数据进行保存，完成指纹数据采集，在高考入场时通过刷指纹进行比对验证身份，如图1.1.2和图1.1.3所示。

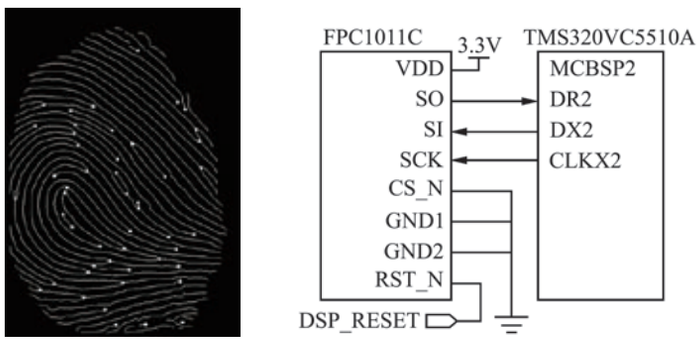


图1.1.2　指纹与指纹传感器模块

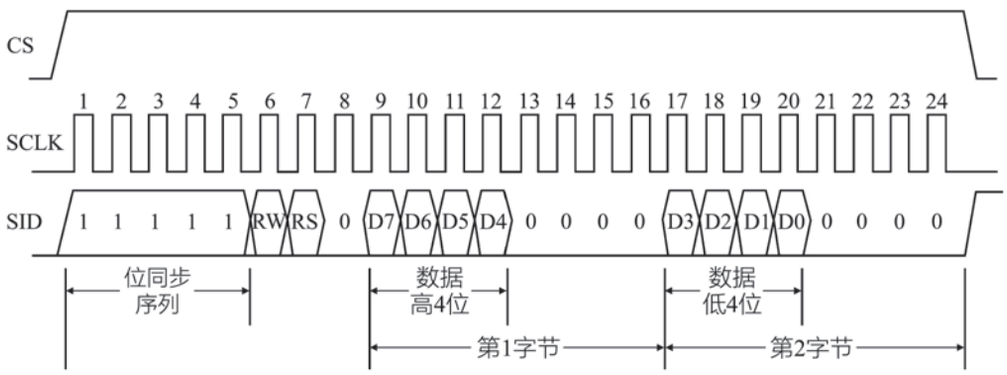


图1.1.3 指纹传感器人机交互模块

问题与讨论

很多驾驶员在不熟悉的路段行驶时喜欢使用手机导航APP，因为导航软件不但能“指路”，而且还能提供实时路况。哪里拥堵，哪里发生事故，这些信息都具有很强的实时性。在导航过程中，手机导航APP是如何获取实时路况数据的？

2. 从互联网上收集数据

人们可以从互联网上直接下载数据，政府机构、综合门户、社交平台等网站提供大量开放共享的数据。通过网络搜索关键词获取数据的同时，搜索使用的关键词数据也在被网站收集。例如，2008年，谷歌公司在《自然》杂志上发表了基于谷歌搜索引擎的45个与流感疫情最为相关的关键词，准确地预测了全美及其9个地区的流感趋势，这个预测与官

拓展链接

数据开放共享

数据开放共享是大数据时代的必然趋势，有很多数据可以直接从开放的数据库中获取。国家数据网，数据来源于国家统计局，共享我国经济民生等多方面的数据；全球经济数据库（CEIC）网站提供了全球超过128个国家的经济数据，其中“中国经济数据库”收编了超过300000条时间序列数据，数据内容涵盖宏观经济数据、行业数据和地区经济数据；中国统计信息网汇集了海量的全国各级政府各年度的国民经济和社会发展统计数据。

方数据的相似度高达97%。

人们通过使用网络爬虫类软件高效、快速地获取某个网站的数据。常见的爬虫类软件有“八爪鱼”采集器、“火车头”采集器、“集搜客”等。例如，用“八爪鱼”采集器收集电影评论数据。登录“八爪鱼”官方网站，完成网站用户注册，安装并启动“八爪鱼”采集器软件。打开“豆瓣电影”影评网页，将网页地址复制到“八爪鱼”采集器软件的采集网址框中，并对采集页面进行设置，把电影名称、影评内容添加到“配置抓取模板”中，然后运行任务，开始数据采集，如图1.1.4所示。最后选择需要的导出方式，就可完成数据的收集与保存，如图1.1.5所示。

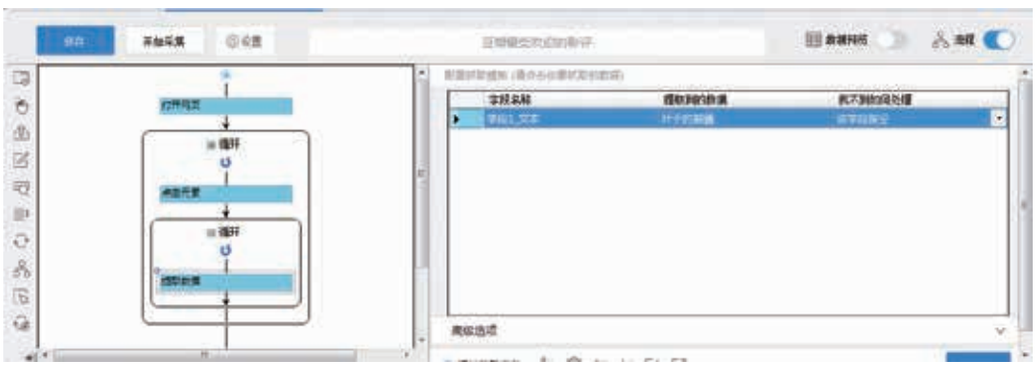


图1.1.4 “八爪鱼”采集器软件数据采集样式呈现界面

	用户名	电影名	影评内容
1	火堂	红海行动	这部华语军事片，是春节档质量最好的电影……
2	小睿	战狼2	伟大祖国在危难时刻不抛弃不放弃任何一个中国人
3	aha	唐人街探案2	剧中有许多的喜剧元素，挺好看的……
4	Enjels	我不是药神	这确实是一部年度作品，有望顶今年国产片口碑榜之势
5	海轩	西虹市首富	本片的剧本还是挺有料的……
6	秋水	捉妖记2	作为贺岁档，中规中矩，但也有亮点……
7	思肖	后来的我们	表演非常优秀，优秀到如果单独从中截出场景来看……
8	火堂	一出好戏	让人联想起《蝇王》……
9	小岛	无问西东	生命有限，浩瀚无穷，唯有立德立言，无问西东……
10	shawsh	影	对人物内心的刻画真实细腻，叩击人心

已采集:40条 已用时：22秒 平均速度：104条/分钟 导出数据

图1.1.5 “八爪鱼”采集器软件导出数据界面

问题与讨论

爬虫类软件能否抓取任何网站上的数据？

噪声数据是指数据中存在错误、异常的数据。通常情况下，使用设备采集数据时，采集到的数据往往夹杂着一些不需要的、随机的噪声数据，这些数据是由周围的干扰或者测试误差引起的。以录音为例，录音时周围环境产生的干扰可能会生成噪声数据。另外，音

响设备信号也会产生设备噪声和放音环境噪声，如电声系统中由于音频电缆屏蔽不良、设备连接不实等原因产生的“嗡嗡”交流声，录音媒介、放大器产生的“嘶嘶”声等。在音频数据文件中，某一段没有内容只有噪声的波形就是纯噪声数据。而噪声数据往往与录音内容存在于同一段波形中，会破坏音频的质量，如图1.1.6所示。

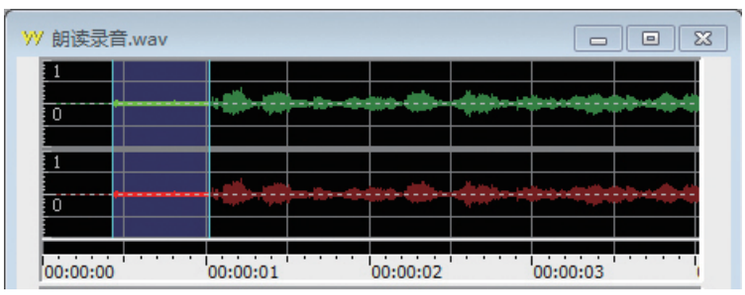


图1.1.6　音频数据文件

类似地，从网上收集数据也会遇到同样的情况。如用爬虫软件收集到的数据中会夹杂着与主题无关的内容，也可能会遇到“缺失值”数据、“重复值”数据、“格式异常”数据等情况。

噪声数据直接影响到数据处理的复杂度，如果处理不当，会对数据分析的结果产生不利的影响，甚至导致错误的结论。

例如，图1.1.7所示为某大学部分学生的年龄图，中间孤立点的年龄超过了60岁，而其他数据点均在20岁上下浮动。一般认为偏离期望值的孤立点是噪声数据，这可能是人工输入数据的操作失误引起的，当然也有可能是真实的数据，需要进一步分析。

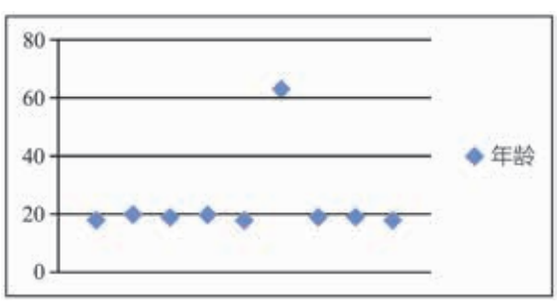


图1.1.7　某大学部分学生年龄图

1.1.3　数据的分类

人们在使用数据的过程中，往往会把收集到的数据按照其特征进行分类。常见的数据类型有数值型、文本型、音频、图形图像、视频、Point of Interest等。

1. 数值型数据

数值型数据由数字、小数点、正负号等组成，用于表示数量，可以直接用算术方法进行运算，是日常生活中经常使用的数据类型。

2. 文本型数据

文本型数据由字符组成。文本中使用的字符来自于字符集。常见的字符集有ASCII码字符集、GB2312汉字编码字符集、Unicode字符集等。

3. 音频数据

数字化之后的声音数据常被称作音频数据。常见的音频数据文件格式有WAV、MIDI、MP3、CDA等。

4. 图形图像数据

图形是以几何线条、几何符号等形式表示物体的轮廓，一般由计算机软件生成，多为矢量图。图像由像素组成，在计算机中保存的是它的像素数据。常见的图形图像数据文件格式有AI、WMF、JPEG、BMP、PNG、TIFF等。

5. 视频数据

视频数据是随时间变化的图像流，是一组连续的图像序列，能更丰富、直接、生动地表达信息和内容。常见的视频数据文件格式有AVI、MPEG等。

6. POI（Point of Interest）数据

每个POI数据包括名称、类别、经度和纬度四个方面的信息，蕴含位置信息，在电子地图和位置服务中应用广泛。如图1.1.8所示，在杭州武林广场附近搜索停车场，电子地图提供7个停车场的POI数据点。

图1.1.8　杭州武林广场附近停车场POI数据图

1.1.4　数据的结构化程度

在描述具体的事物时，单一类型的数据往往不能完整地反映事物的实际情况，需要采用多种类型数据的集合体。数据从结构化程度的角度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

结构化数据也称作行数据，是由二维表结构逻辑表达和实现的数据，严格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。如表1.1.1中的数据就是典型的结构化数据。

表1.1.1　学生信息表

学号	姓名	性别	出生日期	出生地	入学年份
G1801001	李春芳	女	2001年10月11日	浙江杭州	2017
G1801002	方刚	男	2002年3月2日	浙江杭州	2017
…	…	…	…	…	…
G1801045	刘晓丽	女	2001年12月4日	浙江嘉兴	2017

2. 非结构化数据

相对于结构化数据，文本、网页、图像、视频等数据没有严格的逻辑结构，这些数据称为非结构化数据。

3. 半结构化数据

介于结构化数据和非结构化数据之间，有一定的结构性，但结构变化较大，描述数据的属性可以根据实际情况扩充，数量不定且可以重复，这类数据称为半结构化数据。

以存储员工信息为例。每个员工的信息大不相同，有的员工信息很简单，只包括教育情况；有的员工信息却很复杂，包括工作经历、户口档案、特长技能等情况，还有一些事先无法预料的信息。

若使用二维表形式存储，首先对现有员工信息进行粗略的统计整理，总结出信息的所有类别；然后建立一个主表，同时对每个类别分别建立一个子表，如教育情况子表、工作情况子表等，并在主表中加入一个备注字段，将其他信息和初始时没有考虑到的信息保存在备注中。这样存储的优点是查询统计比较方便，缺点是不能适应数据扩展的需要，不能对扩展的信息进行检索。

常见的半结构化数据格式有XML和JSON。用XML表示部分员工信息数据，可以描述如下：

```
<员工简历>
    <编号 id="001">
      <姓名>李芳</姓名>
      <学历>大学本科</学历>
      <毕业院校>浙江大学</毕业院校>
    </编号>
    ……
    <编号 id="005">
      <姓名>陈泽翔</姓名>
      <学历>硕士研究生</学历>
      <教育经历>
```

```
        <阶段1 学历="本科">
            <开始时间>2011年8月</开始时间>
            <结束时间>2015年6月</结束时间>
            <院校名称>浙江大学</院校名称>
            <专业>汉语言文学</专业>
        </阶段1>
        <阶段2 学历="硕士">
            <开始时间>2015年8月</开始时间>
            <结束时间>2018年6月</结束时间>
            <院校名称>北京大学</院校名称>
            <专业>汉语言文学</专业>
        </阶段2>
    </教育经历>
</编号>
</员工简历>
```

实践与体验

比较不同结构化程度数据的组织方式

为了对比结构化数据、非结构化数据和半结构化数据的组织方式，使用Excel软件导出CSV格式文件观察结构化数据，使用Notepad++软件观察JSON格式的半结构化数据，通过IE浏览器观察网页源代码非结构化数据。

实践内容：

1. 选择合适的软件，查看结构化数据和半结构化数据文件的数据组织方式。
2. 打开任意一个网页，观察非结构化数据的组织方式。

实践步骤：

1. 查看结构化数据文件。

①新建一个Excel文件，输入结构化数据（如图1.1.9所示），并保存为“班级藏书.csv”。

	A	B	C	D
1	书名	作者	收藏日期	价格
2	平凡的世界	路遥	2017/3/1	36
3	白鹿原	陈忠实	2017/3/1	31
4	桐花季节	李国文	2017/3/2	29

图1.1.9　Excel文件中的结构化数据

②使用记事本软件打开“班级藏书.csv”文件，观察数据组织的呈现方式。

2. 查看半结构化数据文件。

①准备好半结构化数据文件“beijing_bus.json”。

②安装notepad++软件，通过“打开文件”方式打开“beijing_bus.json”，如图1.1.10所示。

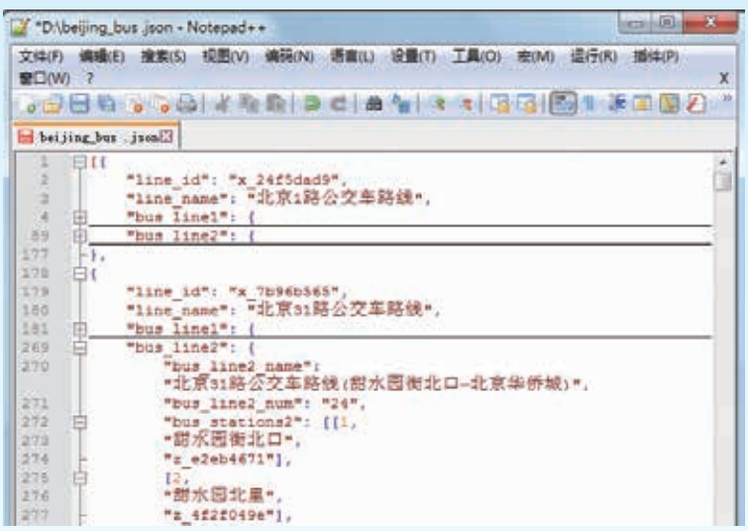


图1.1.10　json格式文件截图

③观察半结构化数据组织的呈现方式。

3. 查看非结构化数据。

①利用IE浏览器打开任意一个网页，查看网页源文件。

②观察源代码的数据组织方式。

结果呈现：

将观察结果填写在下面的表格中。依据观察结果，分析结构化数据和半结构化数据在实际应用中的特点。

数据	文件格式	数据描述的字段与值情况	数据特点
结构化数据			
半结构化数据			
非结构化数据			

思考与练习

1. 通过“中国POI数据”网站，查找并记录你感兴趣的景点的POI坐标数据。

2. 使用爬虫软件，收集网络上关于“2018俄罗斯世界杯”的数据。

1.2 数据模型

模型是对现实世界中某个对象特征的模拟和抽象。常见的模型有桥梁结构模型、汽车模型、房屋建筑模型等。数据模型是对现实世界数据特征的抽象，用来描述和组织数据。针对不同的应用目的，数据模型有不同的分类，常用的有概念模型和逻辑模型。

1.2.1 概念模型

概念模型是人们对现实世界事物抽象的结果，能比较真实地模拟现实世界。它按照用户的观点对数据进行建模，不依赖于具体的计算机系统。它是数据库设计人员和用户之间进行交流的语言，简单、清晰、易于用户理解。

表示概念模型最常用的方法是“实体—联系法”。“实体—联系法”在表示概念模型的时候需要使用实体、属性、实体集、码和联系等基本概念。实体是客观存在并可相互区别的事物，一个人、一辆车、一个苹果都是实体。实体也可以是一个抽象的概念，比如一次考试。一个人有身高、体重、肤色等特征，这些描述实体的特征称为属性。人都具有相同的属性，属于同一类型，多个同一类型实体的集合称为实体集，全人类是实体集。每个人都是独一无二的，每个人都有一个或多个属性能与他人区分，比如指纹、DNA，这样的唯一标识实体的一个或多个属性称为码。现实世界中的人并不是孤立的，人和人之间可能存在某种联系，比如亲属关系、朋友关系。人和其他事物也会有联系，比如某辆汽车属于某个人。实体与实体之间存在某种联系。

问题与讨论

除了实体与实体之间存在联系，实体的属性之间是否也存在联系？请举例说明。

1.2.2 逻辑模型

逻辑模型是按计算机系统的观点对数据进行建模，主要用于数据库管理系统的实现。现有的数据库系统均是基于某种逻辑模型的，因此，要使用数据库系统来管理数据就必须先建立该数据库系统支持的逻辑模型。要得到逻辑模型，先要把现实世界的事物抽象为概念模型，然后把概念模型转化为逻辑模型，如图 1.2.1 所示。

图1.2.1　现实世界事物抽象为逻辑模型过程

常见的逻辑模型有关系模型、列族模型、键值模型、文档模型、图模型、层次模型和网状模型等。

1. 关系模型

关系模型是一种重要的数据模型，它建立在严格的数学概念的基础上。在关系模型中，无论实体还是实体之间的联系都用关系表示，每个关系的数据结构是一张规范的二维表。关系模型结构简单、清晰，可以较好地描述结构化数据。关系术语与现实生活中的表格所使用术语的对比如表1.2.1所示。

表1.2.1　关系术语与现实生活中的表格使用术语粗略对比

关系术语	一般表格的术语
关系名	表名
关系模式	表头（表格的描述）
关系	（一张）二维表
元组	行
属性	列
属性名	列名
属性值	列值

表1.2.2所示的学生基本信息表就是一个关系。表中的一行（某个学生信息）即为一个元组。表中的一列即为一个属性，给每一个属性起一个名称即属性名。表1.2.2有6列，对应6个属性：学号、姓名、性别、出生日期、政治面貌、班级。学号是学生的唯一标识，可以唯一确定一个学生，像这样可以唯一确定一个元组的一个或多个属性称为码。表头对表格的描述称为关系模式。关系模式是对关系的描述，一般表示为：关系名（属性1，属性2，属性3，…，属性n）。表1.2.2学生基本信息表的关系可以描述为：学生（学号，姓名，性别，出生日期，政治面貌，班级）。

表1.2.2　学生基本信息表

学号	姓名	性别	出生日期	政治面貌	班级
3010607	唐杰	男	1992年7月2日	共青团员	高三（6）班
3010828	邵军	男	1991年10月10日	共青团员	高三（8）班
3010201	王佳	女	1992年1月18日	共青团员	高三（2）班

2. 列族模型

列族模型是2006年谷歌发布的BigTable数据库支持的逻辑模型。谷歌的很多数据，包括Web索引、卫星图像数据等在内的海量结构化和半结构化数据，都存储在BigTable中。

列族模型由很多表格组成，每个表格包含很多行，每行通过一个行键唯一标识，每行又包含很多列。某一行的某一列构成一个单元，在单元中存储数据。

列族模型把相关列组合起来成为一个列族。组成列族的所有列的数据存储在磁盘的同一块区域，所以访问相关数据的效率比传统的关系数据库要高。如表1.2.3所示，列“省”“市”“县（区）”“详细地址”组成列族“家庭住址”；列“总收入”“总支出”组成列族“财务状况”。

表1.2.3　列族模型示例1

	姓名	家庭住址				财务状况		好友		
		省	市	县（区）	详细地址	总收入	总支出	…	…	…
行键=1	刘波	浙江	杭州	上城区	文庭雅苑小区1单元121室	87582.0	42662.0	…	…	…
行键=2	李刚	江苏	南京	玄武区	天和人家小区2单元242室	85291.0	44550.0	…	…	…
行键=3	李海	湖南	长沙	天心区	达龙骏景小区4单元418室	89209.0	32501.0	…	…	…
行键=4	张勇	广东	广州	白云区	阳光上东小区6单元611室	85612.0	43537.0	…	…	…
行键=5	王军	四川	成都	青羊区	北苑家园小区3单元356室	90317.0	40861.0	…	…	…
行键=6	张伟	山东	济南	市中区	龙泽苑小区4单元468室	90670.0	42208.0	…	…	…
…	…	…	…	…	…	…	…	…	…	…

关系模型中关系的列定义后就无法再增加或删除列，而列族模型中列族包含的列是不需要预先定义的，可以动态增加或删除列族中的列，非常适合表示半结构化数据。如表1.2.4所示，“好友”列族中每个列的列名是用户的行键，对应的单元中保存的是好友的联系方式。当用户需要保存一个新的好友联系方式，如果“好友”列族中还没有以这个新好友行键为列名的列，在“好友”列族中再添加一列以这个新好友行键为列名的列，然后保存这个新好友的联系方式即可。

列族模型相比关系模型的另一个特点是适合存储稀疏数据。稀疏数据是指表格中大部分单元不存放任何数据。如表1.2.4所示，并不是任意两个用户之间都存在好友关系，所以大部分单元都没有存放任何数据。“谷歌地球”在表示地理位置信息的时候使用的也是列族模型，地图的某个经度是一个行键，地图的纬度是列名，总共有15000个不同的行键和15000个不同的列名。几乎大部分单元都不包含数据。

列族模型还有一个特点是可以保存不同时间的数据，这些不同的数据版本通过时间戳来区分。如表1.2.4所示，行键“1”对应的列族“财务状况”中的列“总收入”有多个版本，这样我们就可以查询王军在不同时间段的总收入情况。

表1.2.4 列族模型示例2

	姓名	家庭住址	财务状况	好友						
		…	…	1	2	3	4	5	6	…
行键=1	刘波	…	…		lg@qq.com					…
行键=2	李刚	…	…	lb@qq.com				wj@163.com		…
行键=3	李海	…	…		lg@qq.com		zy@qq.com			…
行键=4	张勇	…	…			lm@sina.com				…
行键=5	王军	…	…							…
行键=6	张伟	…	…							…
…	…	…	…					…	…	…

总收入	总支出
时间戳=2017/02/12 15:48:36, 值=67582	…
时间戳=2017/03/13 22:41:30, 值=68659	…
时间戳=2017/04/22 08:20:10, 值=70316	…
…	…

目前，基于列族模型的数据库产品有HBase、Cassandra等。

3. 键值模型

键值模型由一组键值对组成。键由不重复的字符串或数值组成，键的值可以是任意类型的数据，如字符串、图像、声音、视频等。键值模型与Python中的字典类似。

Python中的字典{0: '零', 'Pi': 3.1415926, 'lang': ['Python', 'Ruby', 'JavaScript', 'Lisp'], 'student': {'name': 'tom', 'id': 6, 'birth': 19820124}}对应的键值模型如表1.2.5所示。

表1.2.5 键值模型示例

键	值
0	'零'
'Pi'	3.1415926
'lang'	['Python', 'Ruby', 'JavaScript', 'Lisp']
'student'	{'name':'tom','id':6,'birth':19820124}

键值模型较关系模型、列族模型都要简单，只支持简单的应用场景的数据模型建模。比如，用键值模型对互联网上的网页建模，可以用URL作为键值对的键，URL对应的网页内容作为键值对的值。通过查询键（URL）来定位URL对应的网页的内容。

目前，基于键值模型的数据库产品有redis、riak等。

4. 文档模型

文档模型是一个树形、多层嵌套的结构。如图1.2.2所示，该结构开始于一个根节点，并且包含一些子分支，而这些子分支也能再嵌套包含子分支，数据保存在叶节点。

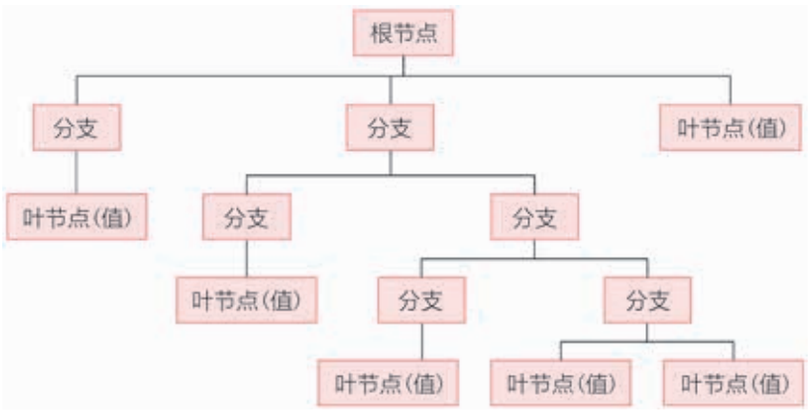


图1.2.2　文档模型

文档模型和关系模型一样也可以表示结构化数据，但大部分应用场景是用文档模型表示半结构化数据，比如XML、JSON。JSON文档{"id": "5197227", "name": "陆议", "address": {"province": "浙江", "city": "杭州"}}表示为文档模型如图1.2.3所示。

图1.2.3　学生文档模型

最典型的文档模型是医院电子病历建模。医疗卫生行业的数据复杂度较高，电子病历的数据几乎覆盖了所有的数据类型。电子病历中的数据包括表格类型、正文文本、影像、处方、化验单等，影像又包括X光、B超、CT、核磁共振等。一份完整的电子病历基本上是嵌套层次结构的半结构化数据。

目前，基于文档模型的数据库产品有MongoDB、CouchBase等。

5. 图模型

图模型是基于图论的逻辑模型，它重点关注实体之间的联系，图论中的知识对于分析实体之间的联系非常有用。在图模型中将实体表示为节点，将联系表示为边。

图1.2.4描述社交网络的好友关系，节点代表“用户”实体，节点之间的连线代表“朋友关系”联系的边。节点和边都有自己的属性，比如“用户”实体有性别、年龄、姓名等属性，边有“亲密度”属性，它代表了两个“用户”实体之间联系的强弱。

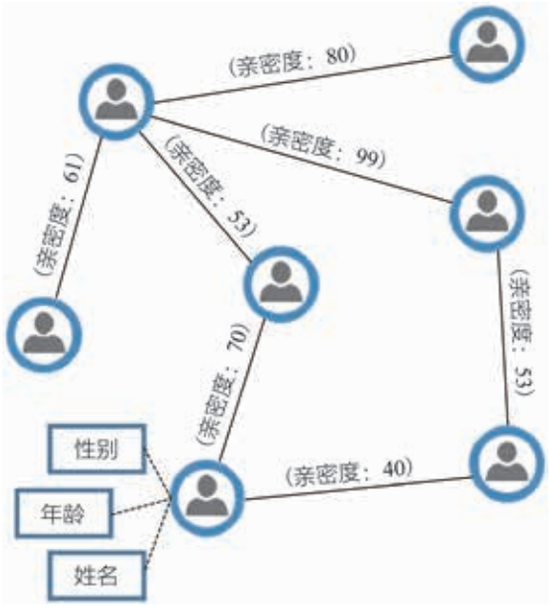


图1.2.4　社交网络用户朋友关系图模型

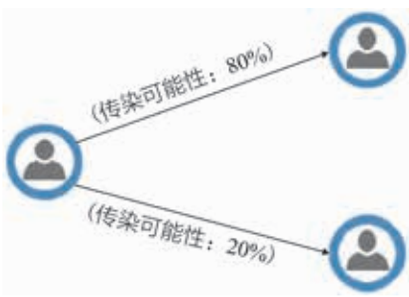


图1.2.5　疾病传染可能性图模型

在图1.2.4中，代表“朋友关系”联系的边没有方向，是无向图。有些情况下，代表联系的边也需要有方向。如图1.2.5所示，一个患有传染病的病人和其他人接触以后，有可能将疾病传染给他人，代表“传染”联系的边是有方向的，称为有向图。联系“传染”这条有向边有“传染可能性”属性，若接触的人已经患过该传染病，则得病的可能性较低。

图模型最流行的应用场景有社交网络数据之间的关系的建模、社交网络数据分析等。

目前，基于图模型的数据库产品有Neo4j、FlockDB等。

拓展链接

层次模型

现实世界中很多事物是按层次组织起来的。层次模型的提出，首先是为了模拟这种按层次组织起来的事物。层次模型是一种用树形结构描述实体及其之间联系的数据模型。在这种结构中，每一个实体集都用节点表示，实体集之间的联系则用节点之间的有向线段来表示。每一个父节点可以有多个子节点，但是每一个子节点只能有一个父节点。这种结构决定了采用层次模型只能处理一对多的实体联系。由IBM公司于1968年推出的IMS（Information Management System）数据库管理系统是第一个层次模型数据库管理系统，也是最典型的一个，曾得到广泛应用。学校办公系统层次模型如图1.2.6所示。

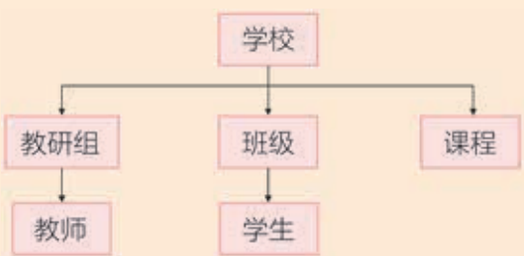


图1.2.6　学校办公系统层次模型

拓展链接

网状模型

网状模型最早由美国的查尔斯·巴赫曼发明。现实世界中事物之间的联系更多的是非层次关系的，用层次模型表示这种关系很不直观，而网状模型可以清晰地表示这种非层次关系。网状模型允许一个节点可以同时拥有多个父节点和子节点。同层次模型相比，网状模型更具有普遍性，能够直接地描述现实世界的实体；也可以认为层次模型是网状模型的一个特例。在图1.2.6所示的学校办公系统层次模型中，学生和课程之间的关系无法正确表达，但是利用网状模型，可以将课程实体集作为学生实体集的子节点，这样就能表示一个学生和多门课程的联系，如图1.2.7所示。

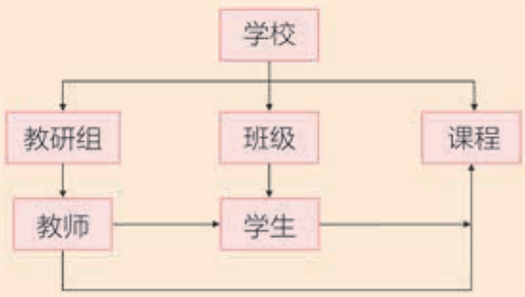


图1.2.7　学校办公系统网状模型

思考与练习

1. 知乎提供了“一人提问，众人解答”的功能。该事例中包含哪些实体？这些实体有哪些属性？哪些实体之间存在联系？

2. 如果要存储的数据的数据模型在系统上线后可能会发生变化，可以考虑使用哪些数据模型呢？

3. 如果要将身边好友关系建立逻辑模型，模型重点关注的是好友关系，用哪种模型比较合适？

1.3 数据管理技术及其发展

科学地管理数据，可以方便人们快速检索、提取、更新和分析数据，获取大量有价值的信息。在应用需求的推动下，在计算机硬件、软件发展的基础上，数据管理技术的发展历程大致可以分为人工管理阶段、文件系统阶段和数据库系统阶段。

1.3.1 人工管理阶段

20世纪50年代中期，计算机主要用于科学计算。当时外存只有纸带、卡片、磁带，没有磁盘等直接存取的存储设备，也没有操作系统，更没有专门管理数据的软件。人工管理阶段具有以下特点：

①数据不保存。由于当时计算机主要用于科学计算，一般不需要将数据长期保存。

②应用程序管理数据。数据由应用程序设计、定义和管理，没有相应的软件系统负责数据的管理工作。

③数据不共享。数据是面向应用程序的，一组数据只能对应一个程序。

④数据不具有独立性，数据完全依赖于应用程序。

1.3.2 文件系统阶段

20世纪50年代后期到60年代，计算机硬件、软件快速发展，存储设备出现了磁盘、磁鼓等，也有了操作系统支持下的专门管理数据的软件，即文件系统。

文件系统管理的对象有文件、目录、存储空间。文件是文件系统管理的直接对象；目录包含文件名和该文件所在的物理地址，是对文件存取和检索的关键。有效管理文件和目录所占的存储空间，不仅可以提高外存的利用率，还能提高文件的存取速度。

文件的逻辑结构可以理解为文件内容的组织形式，包括流式文件和记录式文件。流式文件是相关的有序字符的集合，直接由一连串信息组成，是无结构的文件，如源程序、可执行文件、库函数等；记录式文件是一种有结构的文件，是一组连续顺序的记录集合。以省教育部门管理学生信息为例。Excel软件很难满足几百万学生信息量的管理需求。若使用文本文件存储，打开“学生名单.txt”文件可以看到数据存储方式如下：

```
170404050572   陆朱丹  高中一年级（13）班     女
170404050202   钱留洋  高中一年级（05）班     男
……
```

该文件可以从头到尾以顺序的方式进行访问，好像文件中的信息都排成一行，这样的

文件称为顺序文件。音频文件、视频文件、文本文件等都属于顺序文件，事实上，大多数由计算机用户创建的文件都是顺序文件。

若要在“学生名单.txt”文件中查找某同学的信息，通常借助文件索引来提高检索效率。建立文件索引可以快速确定逻辑记录的位置，文件索引包含存储在该文件中的键的列表和指示包含每个键的记录存储位置的项。如图1.3.1，对顺序记录的数据文件建立了索引，如要检索数据文件中某个编号的数据，直接访问数据文件相对比较耗时，而建立索引文件后，可以快速检索已经排好序的编号，然后在数据文件中根据该记录的行号即可实现快速检索。

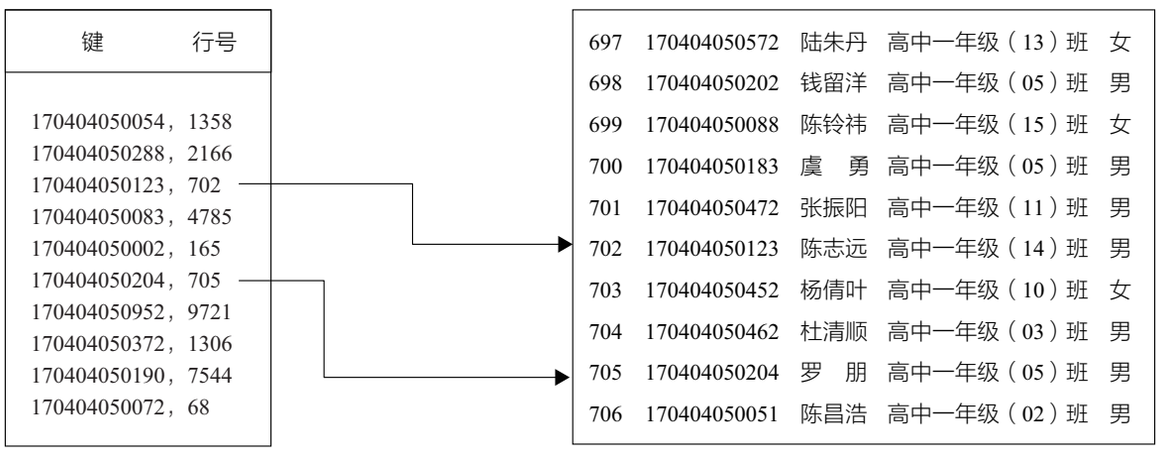


图1.3.1 索引文件与数据文件

用文件系统管理数据具有以下特点：

①数据可以长期保存。很多重要的文件数据可以长期保留在外存上反复进行查询、修改、插入和删除等操作。

②文件系统管理数据。文件系统把数据组织成相互独立的数据文件，通过“按文件名访问，按记录进行存取”的方式，提供了对文件进行打开与关闭、对记录进行读取和写入等存取方式。文件系统中的数据不能同时访问、读取和存储。

③数据共享性差。在文件系统中，一个文件基本上对应一个应用程序，即文件仍然是面向应用的。当不同的应用程序具有部分相同的数据时，必须建立各自的文件，而不能共享相同的数据，因此数据的冗余度大，浪费存储空间。

④数据独立性差。文件系统中的文件是为某一特定应用服务的，当数据的逻辑结构改变时，应用程序中文件结构的定义必须做相应的改变。因此，应用程序依赖于数据文件，缺乏独立性。

1.3.3 数据库系统阶段

20世纪60年代后期以来，计算机管理的对象规模越来越大，应用范围越来越广，数

据量急剧增长。此时，大容量、低成本的磁盘出现了，而编制和维护系统软件及应用程序所需的成本相对增加。为了解决多用户共享数据的需求，使数据为尽可能多的应用服务，数据库技术便应运而生。

数据库是指有组织、动态地存储在辅助存储器上的，能为多个用户共享的、与应用程序彼此独立的一组相互关联着的数据集合。数据库系统管理数据具有以下特点：

①数据结构化。数据库中的数据不再仅仅针对某一个应用，而是面向整个组织；不仅仅数据内部是结构化的，而且整体也是结构化的，数据之间是具有联系的。因此，数据库系统实现了整体数据的结构化，这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数据共享性好、冗余度低。数据库系统从整体角度看待和描述数据，数据可以被多个用户、多个应用共享使用。这样，大大降低了数据的冗余度，节省了存储空间。

③数据独立性高。数据与程序的独立把数据的定义从程序中分离出去，从而简化了应用程序的编制，减少了应用程序的维护和修改。

④数据由数据库管理系统统一管理。数据库的共享会面临多个用户同时存取数据库中的数据，甚至同时存取同一个数据的情况，所以数据库管理系统还提供了数据安全性保护、完整性检查、并发控制和数据库恢复等功能。

数据库管理系统是为了建立、使用和维护数据库而设计的数据管理软件。在计算机系统中，它介于操作系统和用户之间，负责对数据库资源进行统一的管理和控制，所有用户或程序发出的有关数据库方面的操作命令，都通过数据库管理系统来实现，如图1.3.2所示。常见的数据库管理系统有PostgreSQL、MySQL、Oracle、SQL Server等。

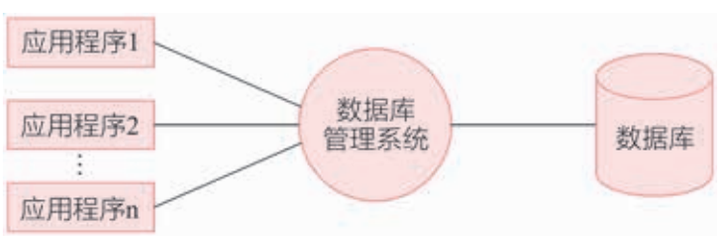


图1.3.2　应用程序、数据库管理系统及数据库的关系

随着大数据时代的来临，科学研究、企业应用、网络行为等都在源源不断地产生新的数据，其数据种类繁多，除了结构化数据以外，更多的是非结构化数据，包括邮件、音频、视频、微信、微博、位置数据、日志数据等。如此数量庞大、类型复杂的数据，传统的关系数据库已经不能满足管理的需求，越来越多的数据被存储在非关系数据库中。常见的非关系数据库有键值数据库、列族数据库、文档数据库和图数据库等。

非关系数据库具有灵活的可扩展性、灵活的数据模型以及与云计算紧密融合等特点，解决了大规模数据集和多重数据带来的问题。因此，从数据结构化程度来看，结构化的数据一般采用关系数据库进行管理；半结构化数据可采用非关系数据库进行管理，也可以转化为结构化数据后采用关系数据库进行管理。

思考与练习

比较三种不同数据管理技术的特点，并完成下面表格。

特点	人工管理阶段	文件系统阶段	数据库系统阶段
数据的管理者	用户（程序员）	文件系统	数据库管理系统
数据面向的对象	某一应用程序	某一应用程序	一个组织（如部门、学校、企业等）
数据共享程度			
数据冗余程度			
数据的独立性			
数据的结构化			

1.4 数据管理与分析技术的应用

随着新兴的开源数据管理工具以及硬件的发展，尤其是内存计算方面的硬件发展，数据管理与分析技术应用领域不断拓宽，数据价值不断被挖掘。有效利用数据库技术，可以简化业务流程，提高工作效率；通过分析行为数据，可以提升企业服务品质，扩大企业经营效益；综合运用大数据技术，可以整合资源配置，引领创新驱动发展。

1. 打造“数据平台”，提高业务效率

数据库系统在各行各业中应用广泛，它根据用户需求对大量的数据进行存储和管理，利用数据库提供的查询、插入、修改、删除和统计等功能，可以快速便捷地管理和使用数据库中的数据。数据分析平台可以模拟场景，揭示关键影响因素，提供实时洞察，驱动业务增长。

以单位档案信息管理为例，单位需要管理的数据涉及各个方面，包括质量管理档案、财务档案、市场营销档案、技术创新档案、产品档案、人力资源档案等。通过数据库管理系统建立多个专题信息数据表，并且按单位管理工作流程建立每个信息数据表之间的联系，可随时增加、处理信息，还可以进行多表查询，生成新的报表，动态管理所有数据。这不仅可以提高管理效率，还能在一定程度上节约管理成本。

对企业来说，更需要分析与产品相关的各种类型数据。以某饮料公司为例，全国近万名业务员每天对所管辖营业点的饮品摆放与购买场景进行拍照或者摄像，并将数据传

回公司总部，每月产生约3TB的数据。公司试图从中找出很多问题的答案，如“怎样摆放饮品堆更能促进销售？”“什么年龄的消费者在饮品堆前停留更久，他们一次购买量多大？”……“气温的变化对销售有怎样的影响？”“竞争对手的新包装对销售产生了怎样的影响？”这些问题的答案，以前多数要靠经验总结出来，而生产、销售、成本等数据报表需要很长时间才能统计得到，这种滞后给公司的发展带来极大不利。2003年，公司与数据分析团队合作，开始尝试对这些非结构化数据进行分析，并逐渐关注运输环境数据，如收集高速公路的收费、道路等级、天气、配送中心辐射半径、不同市场的售价、不同渠道的费用、各地的人力成本等数据，形成物流、资金流和信息流彼此关联的实时统计报告，精准地管控物流成本。公司不仅准确掌握了生产和销售的平衡数据，做到“要多少、送多少”，还将400多家办事处、30多个配送中心纳入到体系中，实时掌握精准的数据。在强大的数据分析技术支持下，饮品销售量以30%～40%的年增长率增加，2016年公司的国内市场业绩遥遥领先。

2. 描绘“用户画像”，提升服务品质

销售、通信、金融、旅游等众多行业，通过数据分析用户行为和业务数据之间的关联，勾勒用户画像，提升用户体验，提供个性化服务，实现精细化运营。用户画像是建立在一系列数据之上的目标用户模型。通过用户画像，可以从多个角度精细化地刻画用户特征，帮助商家发现用户关注点，改进产品或服务，实现精准营销。它经常应用于内容推送、应用推荐等个性化服务领域。

拓展链接

用户画像标签建模

用户画像标签建模主要包括四个步骤：首先获取原始数据，包括历史交易数据和用户的基本信息数据，另外还收集互联网数据，这部分数据主要通过网络爬虫等技术，对用户行为数据进行爬取；其次对原始数据进行统计分析，得到事实标签，如年龄分布、性别比例、购买频率等；然后对事实标签进行分析，得到模型标签，如人口属性、产品购买偏好、用户关联关系等；最后预测用户行为。

以旅游情境化网络推荐服务为例，每一名用户登录旅游网站时都需要填写个人信息，网站就能收集到用户的属性信息，包括用户的姓名、性别、年龄、教育程度、消费情况、家庭住址、联系方式等。通过爬虫技术收集该用户的网络行为数据，也可以获取更多的信息，包括使用的移动设备、联网方式、消费能力、风格喜好、历史旅游记录、评价信息等。网站统计所有用户信息数据，可以得到年龄段、消费能力、选择景点之间的关系。针对某用户，系统检测到登录信息时，通过模拟情景推荐景点。例如，某用户的信息标签如表1.4.1所示，系统根据基本信息将其描绘为“一个精力旺盛、爱好旅行的年轻工科男”，由此系统为其匹配最为相似的部分用户数据，然后结合登录地点标签初步给出景点推荐结果为：西溪湿地公园、瑶琳仙境、虎啸峡漂流、东天目山风景区、浙西大

峡谷、桐庐山湾湾激流探险、浙西凉源峡漂流、千岛湖等。但是对比其行为标签与情景标签，其中有些推荐与实际情况不符。比如行为标签中的风格喜好为“游泳”，价格偏好为“100～500元”，而系统给出的东天目山风景区、浙西大峡谷等为爬山项目。于是，系统再次根据时间、温度等情景标签与用户行为标签中的“游泳”这一偏好特征进行比对，最终给出景点推荐结果为：虎啸峡漂流、桐庐山湾湾激流探险、浙西凉源峡漂流等。

表1.4.1　某用户信息标签

数据类别	标签	用户属性
基本信息	用户标识	旅行的小鹿
	姓名	王小山
	性别	男
	年龄	23
	年级	大三
行为信息	使用设备	智能手机
	联网方式	4G
	价格偏好	100～500元
	风格喜好	游泳
	行为状态	运动
情景信息	时间	秋季9月
	地点	杭州汽车站
	天气	晴朗
	温度	28℃

3. 深度“数据挖掘”，催生研究创新

数据价值的挖掘是无止境的，从互联网创业公司到金融机构，从工农业到国防部，数据科学家通过大数据项目做创新驱动，不断推动着行业研究与创新，包括科学研究、生物技术研发等。

以智慧城市为例。通过互联网将无处不在的遍布于城市各处的智能传感器连接起来，实现对城市的全面感知，利用云计算等处理技术对海量数据进行处理和分析，实现网上城市数字空间与物联网的融合，并发出指令，对包括政务、民生、环境、公共安全等在内的各种需求做出智能化响

拓展链接

数据科学家

数据科学家是能够采用科学方法、运用数据挖掘工具寻找新的数据洞察的工程师，需要具备数据提取与综合能力、统计分析能力、数据洞察与信息挖掘能力、软件开发能力、网络编程能力、数据可视化表达能力等，知识涉及计算机科学、数理统计学、图形设计学、人机交互学等多门学科。目前，社会对数据科学家的需求相对迫切，国家开始注重数据科学方面专业人才的培养。

应和决策支持。其中，面向智慧城市数据的深度分析、预警与预测，需要借助数据挖掘技术实现，包括聚类、分类、关联规则、协同过滤等机器学习算法，实现特定数据集的挖掘与分析。

智慧城市中的大数据采集、共享与分析涉及多个行业，通过一体化的思路，处理各行业在大数据应用中的共性问题。如公安等安全部门掌握着视频监控图像、情报信息等数据，其他社会组织（如交通、城管、街道、居委会等）也掌握着大量的数据，还有大型商场、楼宇等各类社会资源数据。将这些数据输入城市一体化公共信息服务平台，进行统一管理，如图1.4.1所示。

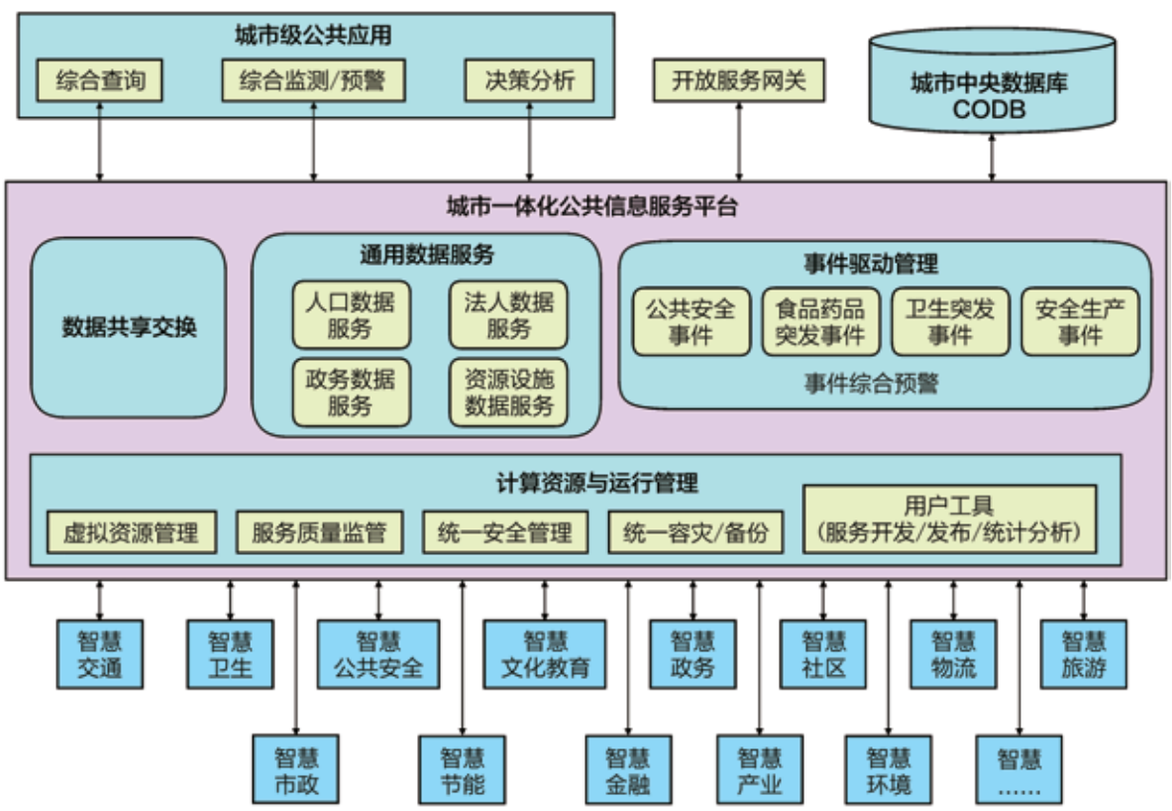


图1.4.1　城市信息共享架构图

这些数据通过深入的数据分析和挖掘，与业务深度融合，可以为人们提供决策分析服务，并应用于交通、医疗卫生、社会管理、农业、商业、金融等领域。如创新社会管理，实现社会管理一体化模式，其过程如下：

- 建立创新社会管理一体化平台概念模型

①动态感知。对人员、场所、设备设施、活动、舆论等数据进行动态采集。

②互联互通。实现区、街道、居委会等社会综合管理部门的数据共享，并与市级公安、司法、住房等社会管理相关部门信息整合。

③应用智能。构建社会管理业务模型，对采集的大量数据进行分析处理，形成趋势判断，为社会综合治理提供预警和处理的智能化手段。

④管理创新。形成社会管理综合应用。

● 建立创新社会管理总体框架

创新社会管理总体框架包括全面联动的社会管理机构，以人为本的社会建设和保障体系，完整有效的社会管理服务和协调机制，一体化整合的社会管理信息化支撑平台，以及多方位协同的社会管理信息化综合应用。

● 建设创新社会管理内容

①社会管理一体化平台。它包括综合数据管理、信息共享服务、应用集成服务、运行管理服务。

②社会管理一体化应用服务。它包括社会管理信息动态汇聚、城市常态化协同联动、城市应急响应处置、多维度领导辅助决策和多渠道信息综合服务。

③社会管理关联系统整合与集成。与公安、司法、工商、房地产、网格化管理中心等社会管理部门进行系统整合，实现包括视频共享、数据整合、跨部门信息调阅、协同处置等功能。

④社会管理一体化系统配套体系研究。它包括社会管理一体化运营模式和一体化系统标准规范。

智慧城市以“数字化、智能化、网络化、互动化、协同化、融合化”为主要特征，通过对城市内人、物及其行为的全面感知和互联互通，大幅优化并提升城市运行的效率和效益，实现生活更加便捷、环境更加友好、资源更加节约的可持续发展。

思考与练习

结合生活实际，分析总结数据对你个人的日常生活产生了哪些影响。

生活中的事例	运用的数据管理与分析技术	作用与影响

巩固与提高

1. 使用合适的数据收集方式，了解数据在我国的发展历史与现状，并填写下表。

描述获取的数据	所属数据类型（画“✓”）		该数据类型常见的文件格式	结构化程度（画“✓”）		
				结构化	半结构化	非结构化
	数值					
	文本					
	网页					
	图形图像					
	POI数据					
	其他					

2. 利用互联网数据资源，了解数据管理与分析技术在不同领域的应用，总结数据管理与分析对日常生活或社会发展的重要意义。具体要求如下：

（1）分组，建议四名同学为一个小组。

（2）确定特定领域。

（3）确定具体的数据管理与分析应用实例。

（4）分析运用的具体数据管理与分析技术。

（5）总结该实例中数据管理与分析的重要作用。

（6）制作成PPT进行展示交流。

项目挑战

校园微课网站的数据管理

近年来，基于计算机网络技术和多媒体数字技术的网络教育蓬勃发展。大规模开放的在线视频课程得到广大学习者的青睐，它借助网络平台，以视频、文本等素材为载体，提供数字化学习服务。越来越多的中小学校也开始设计自己的微课网站，开放本校教师制作的数字资源。

项目任务

某中学的教师制作了不同学科的微课，为了便于学生在线学习，计划将这些视频都“搬”上网站。请你利用已学的数据管理知识，为这个微课网站设计良好的数据管理方式，以实现有价值的功能。

注：一般情况下，微课包括教案、微课视频、练习与评估等内容。

过程与建议

1. 明确微课网站需要具备的功能

广泛了解当前各在线视频网站的功能，将有助于微课网站的功能设计。在这一步里，请浏览多个知名在线视频网站，利用下表梳理和分析它们现有的功能（至少要分析三个网站）。

注：某网站具备某项功能时，请在相应的单元格里画“✓”并备注相关内容，否则画“×”。如发现更多网站或更多功能，可在其他项中添加。

网站功能		中国大学MOOC	爱课程	网易公开课	其他：	其他：
面向发布者的功能	发布者注册					
	发布内容					
	浏览评论					
	其他：					
	其他：					

续表

网站功能		中国大学MOOC	爱课程	网易公开课	其他：	其他：
面向学习者的功能	搜索课程					
	课程排序					
	评论课程					
	其他：					
	其他：					
	其他：					

2. 确定学校微课网站的功能和所涉及的数据

通过对知名课程网站的浏览、梳理与分析，对本校微课网站的功能有了初步的认识。在这一步里，请明确本校微课网站的功能，并对这些功能的意义与必要程度（用五星标识，五颗星是非常必要，一颗星是一般必要）做出标注。

微课网站功能		涉及的数据	所属的数据类型	必要程度
面向发布者的功能	发布者注册	用户名、密码等	文本	☆☆☆☆☆
				☆☆☆☆☆
				☆☆☆☆☆
				☆☆☆☆☆
				☆☆☆☆☆
面向学习者的功能	浏览微课	相关教案、相关介绍、视频截图、视频、对微课的评价信息	文本、图像、视频	☆☆☆☆☆
				☆☆☆☆☆
				☆☆☆☆☆
				☆☆☆☆☆
				☆☆☆☆☆

3. 分析并明确微课网站的数据管理方式

数据		数据形式	结构化程度（结构化、半结构化、非结构化）	管理方式（关系数据库、文件系统管理）
用户信息	用户名	数值、字符	结构化数据	关系数据库管理
	密码			
	照片			
	所属班级			

续表

数据		数据形式	结构化程度（结构化、半结构化、非结构化）	管理方式（关系数据库、文件系统管理）
微课信息				
其他：				

4. 讨论用户生成数据的价值

在用户使用网站时，会生成大量数据，重视并管理这些数据，将会产生更大的数据价值。请与学友分组讨论：根据目前设计的功能，该校园微课网站在使用过程中会生成哪些数据？分析这些数据可能存在的价值。

想象一下，如何对这些数据进行分析和挖掘，以便产生更大的价值。

5. 讨论与分享

各组展示自己的数据管理方式，然后辨析不同的方式，最后确定校园微课网站中所涉及数据的管理方式。

评价标准

请根据项目实施的过程、效果以及成果展示交流的结果，对自己完成项目的情况进行客观的评价，并思考后续完善的方向。把评价结果和完善方案填写在下面的表格中。

评价条目	说明	评分（1～10分）	评分主要依据阐述	后续完善方向
知名网站功能总结	收集到较为全面的知名视频网站资料，并能科学归纳总结其功能			
网站功能分析	能全面列出校园微课堂视频学习网站具备的功能			
用户与权限分析	能正确分析用户使用的身份及相关的功能权限			
用户名与密码样例	能合理设置用户名和密码数据样例			

续表

评价条目	说明	评分（1～10分）	评分主要依据阐述	后续完善方向
数据管理方式选择	能正确分析视频、教案、练习等数据的结构化程度，并选择合适的数据管理方式			
讨论分享	讨论展示内容与项目研究目标一致，对微课堂视频学习网站搭建有建设性作用			

拓展项目

1. 利用互联网搜索引擎，查找最近一年北京地区每月的汽油价格，分析该网站管理了哪些数据。选择合适的软件对12个月的汽油价格进行数据分析，由此你可以得出怎样的结论？

2. 教学质量是学校发展的生命线，教师的教学行为与学生的学习成效密不可分。学校计划设计“网上评教系统”，阶段性反馈学生对教师上课进度、教学方式、作业布置等方面的评价，达到以评促教的目的。你认为“网上评教系统”需要收集哪些数据？采用哪种数据管理技术比较合适？针对某教师连续一年内的评教数据，一般可以得出怎样的科学评价？通过对评教数据的管理与分析，这个系统对教师的教学、学生的学习将会产生哪些作用与影响？

第二章 2

需求分析与方案设计

一个数据管理与分析项目开始后，首先要进行需求分析，明确这个项目需要哪些数据，如何进行管理，具备哪些功能。方案设计是在需求分析的基础上寻找实现项目的可行方案，主要研究数据怎么进行管理与分析。通俗地讲，需求分析考虑“做什么”，方案设计考虑“怎么做”。

问题与挑战

- 微信的群聊和朋友圈的使用频率很高，非常受用户欢迎。如果你是微信的开发人员，要给微信添加一些新功能，你认为哪些功能会受大家的欢迎？

- 学校准备开发一个互动式网络问答社区系统。在系统中注册账号并登录后，学生可以把遇到的问题发布到社区中，老师或者其他的同学可以予以解答。这个问答社区系统中的数据应该以何种结构保存？数据之间的关系又是如何的？

- 某共享汽车租赁公司准备开发一个管理系统。该系统能够存储员工、客户、汽车、租赁等数据，实现公司业务的信息化管理，提升公司的工作效率，这个系统要如何设计呢？

学习目标

1. 结合具体案例了解需求分析的任务和方法。
2. 掌握E-R图。
3. 初步了解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。

内容总览

2.1 需求分析

数据管理与分析项目的开始阶段，先要对项目进行需求分析。开发团队需要和用户进行沟通，收集用户提出的要求并对这些要求进行归纳、整理、分析，明白用户想要解决什么问题，项目必须做什么，即明确项目完整、准确、清晰的需求。需求分析的主要任务包括业务流程分析、功能需求分析、数据需求分析和非功能需求分析四个部分。

不同专业背景的人做需求分析采用的方法是不同的，不同类型的项目采用的分析方法也是不同的。本节以“云课堂学习平台”项目为例，简要介绍如何对数据管理分析项目进行需求分析。

例 1

某教育培训机构要开发一个“云课堂学习平台”来提供在线教育培训服务。传统的教育培训方式是在固定的时间、固定的地点由讲师对学员进行培训，该教育培训机构希望通过构建“云课堂学习平台”改变传统的培训模式，打造一个随时随地在线学习的平台，除了通过平台进行在线教育培训盈利，还希望通过建设在线教育平台节省传统培训模式中的人力、时间、场所等成本。平台采用网络视频点播服务的方式实现网络教学与培训，学习视频由讲师制作并发布到平台上，学员可以在平台上观看视频进行学习。

2.1.1 业务流程分析

业务流程分析是功能需求分析、数据需求分析的基础，通过业务流程分析厘清项目需要的功能，以及需要保存、处理、分析的数据。

业务流程分析普遍采用以下方法：

①调研法。需求分析人员与项目相关人员沟通了解业务活动的情况，识别项目的业务流程。调研法需要项目的全体人员参与，通过与相关人员的交流、头脑风暴等，分析项目的业务流程。

②借鉴法。如果要进行的项目已经有类似的案例，可以先通过分析该项目，得到项目的基本业务流程，然后分析人员与项目相关人员一起讨论完善这些业务流程。

问题与讨论

除了调研法与借鉴法，还可以通过哪些方法进行业务流程分析？

业务流程分析分为以下几个步骤：

①针对项目相关的人员，识别相关人员发起的主体业务流程和变体业务流程。主体业务流程是相关人员的主要业务流程，变体业务流程是主体业务流程中存在的一些变化。

②识别项目的支撑业务流程。支撑业务流程是为了更好地服务客户或者辅助支持业务的业务流程。在“互联网+”时代，人们需要的服务更加注重个性、灵活、优质、精准，这些服务往往是通过支撑业务来实现的，因而支撑业务已经成为评价项目实施优劣的关键因素。

③识别管理流程。管理流程是为了控制业务开展、规避风险、控制结果的业务流程。

④对识别出的业务流程进行分析，绘制业务流程图。

学员进入“云课堂学习平台”，主体业务流程是“学习课程”，如图2.1.1所示。

在这个主体业务流程上还存在一些变体，即变体业务流程。比如，学员学习完课程之后要考试，考试未通过，学员可以选择终止学习；学员课程学习未完成，也可以选择终止学习。这些都是主体业务流程“学习课程”的变体，把这些变体业务流程添加到主体业务流程中，得到如图2.1.2所示的“学员学习课程”变体业务流程图。

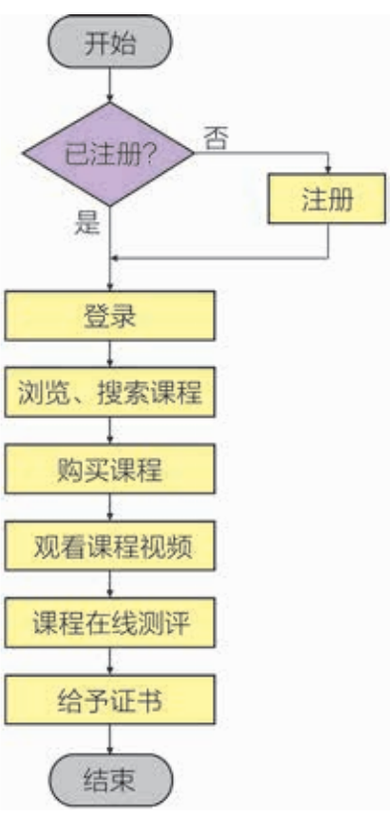


图2.1.1 “学员学习课程”主体业务流程图

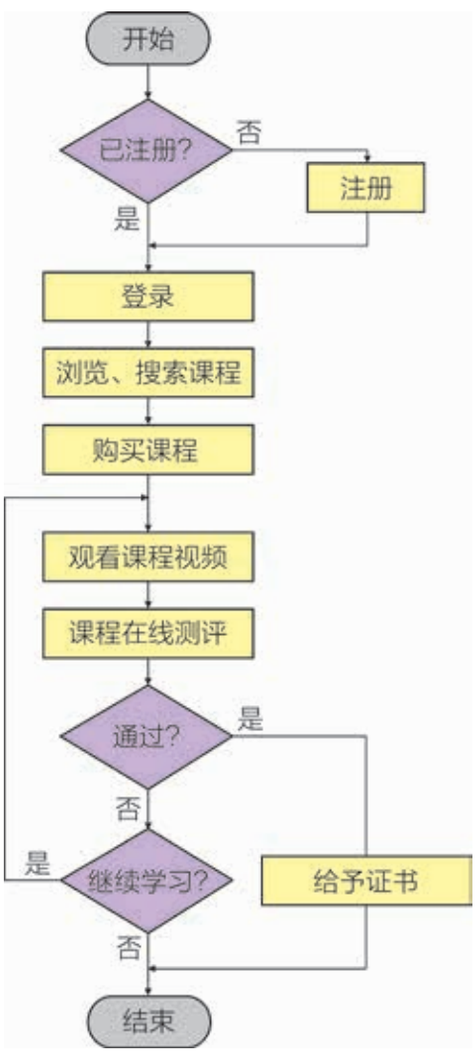


图2.1.2 “学员学习课程”变体业务流程图

“学员学习课程”业务流程中的“观看课程视频”实际上是子业务流程，在这个业务流程中，为了方便学员学习，可以加入一个继续观看的支撑业务流，方便学员从上次学习的位置继续往下学习。视频观看完毕，可以添加视频评价、打分的支撑业务流，用于评价课程视频的质量。图2.1.3所示是“学员观看课程视频”业务流程图。

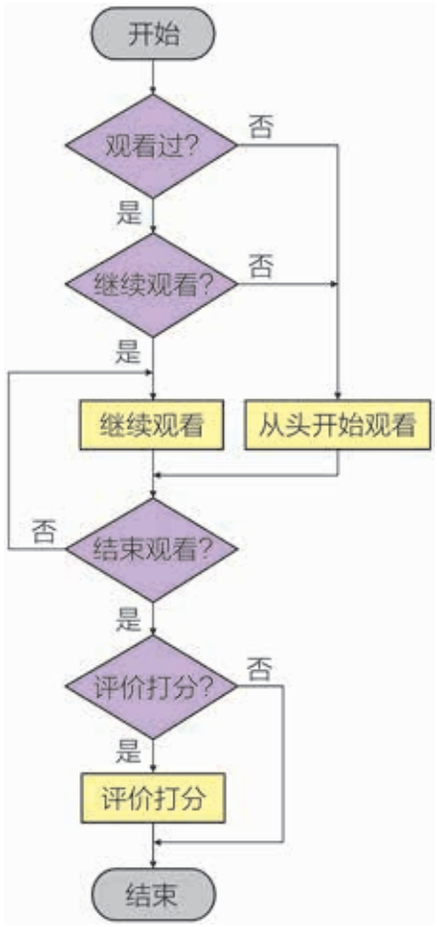


图2.1.3 “学员观看课程视频”业务流程图

业务流程通常不仅仅涉及一个员工或者一个部门，而是由多个员工或部门协作来完成的。比如，“学员购买课程”是“学员学习课程”主体业务流程的子流程，这个流程除了学员以外，还涉及系统，其业务流程图如图2.1.4所示。

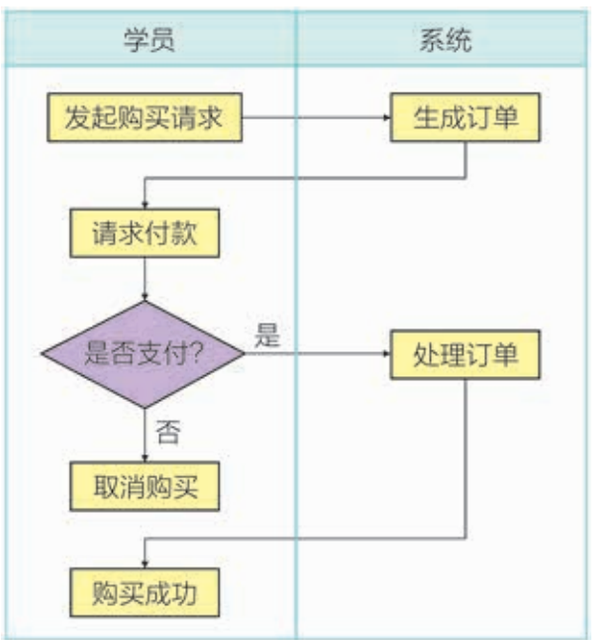


图2.1.4 “学员购买课程”业务流程图

讲师进入“云课堂学习平台”，主体业务流程是“发布课程”，其业务流程如图2.1.5所示。

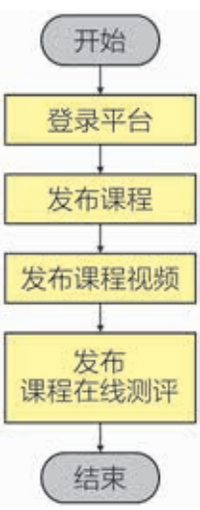


图2.1.5 “讲师发布课程”业务流程图

支撑业务流程也包含特定的时间、状态发起的业务流程。当平台录入了一门新的课程，根据学员以往学习情况分析的结果，这门课程是学员感兴趣的，系统发送一条提醒消息，这就是特定状态发起的业务流程。

在讲师正式发布课程的学习视频、在线测评前，管理员要对课程内容进行审核，只有审核通过的课程才能正式发布。“审核课程”业务流程

图如图2.1.6所示。为了监控课程的质量，需要通过相应的数据分析得到反馈信息。为了吸引更多的学员购买课程，增加盈利，还需要分析学员对哪些课程感兴趣，并且多研发类似的课程。这些业务流程属于管理流程。

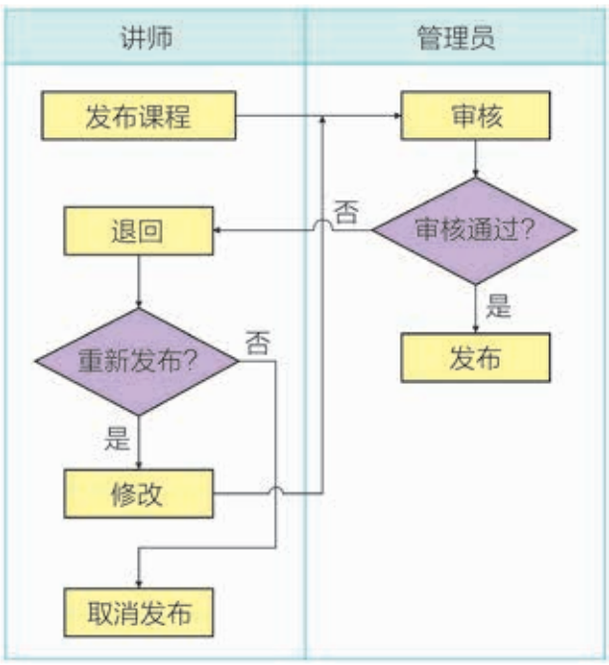


图2.1.6 “审核课程”业务流程图

问题与讨论

讲师“发布课程”主体业务流程上存在哪些支撑业务流程？

序号	支撑业务流程
1	课程视频断点续传
2	
3	
4	

2.1.2 功能需求

对于一个数据管理与分析项目，功能需求描述项目为用户解决什么问题，需要完成哪些功能。功能需求的任务是得到项目需要完成的所有功能。

业务流程分析完成以后，需要识别这些流程中存在的项目相关的业务场景。业务场景是以部门或人员等角色为主完成的、相对独立的、可以汇报的业务活动。通过分析业务场景，可以导出项目的功能。

在“云课堂学习平台”中涉及的角色包括学员、讲师、管理员。

分析图2.1.1 所示的“学员学习课程”主体业务流程图，学员完成的业务活动有：注册、浏览课程、搜索课程、购买课程、学习课程。

分析图2.1.5所示的“讲师发布课程”业务流程图，讲师完成的主要业务活动有：发布课程、上传课程视频、上传课程在线测评。

管理员的主要业务活动包含学员管理、讲师管理、课程审核、课程质量监控、课程难易度监控、学员兴趣点分析。

通过对“云课堂学习平台”业务场景的识别与分析，整理出该项目的功能图，如图2.1.7所示。

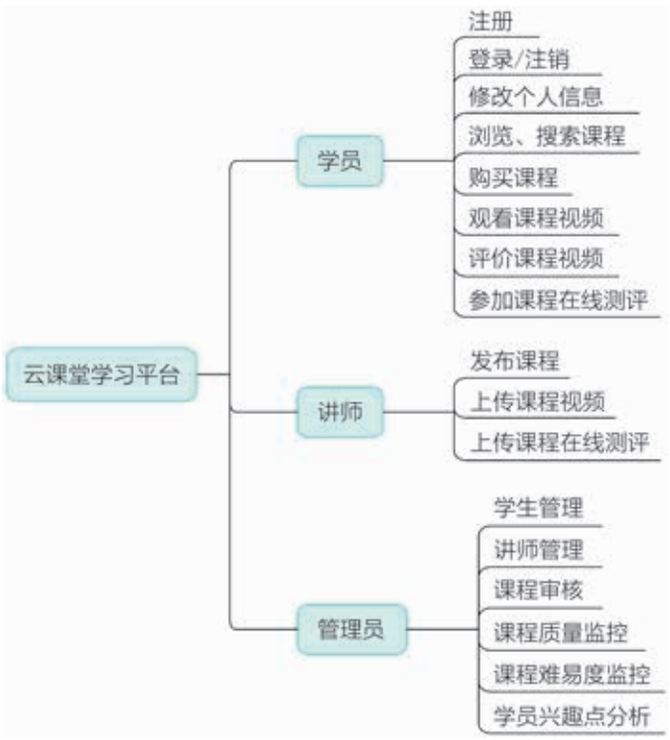


图2.1.7 “云课堂学习平台”功能图

有了功能图，还需要绘制数据流图对系统功能和功能之间的数据流动进行建模。

拓展链接

数据流图

数据流图以图形化的方式描绘系统中的数据在流动的过程中所经历的加工和处理，能够方便、直观地表示系统的功能和行为。数据流图理解相对容易，是开发团队和用户之间进行交流和沟通的有效手段。

对于一个复杂的系统，一张数据流图无法将整个系统的数据加工处理过程表达清楚，需要对问题逐层分解，用分层的数据流图来进一步反映整个系统数据的加工处理过程。图2.1.8是分解之后的学员搜索课程数据流图。

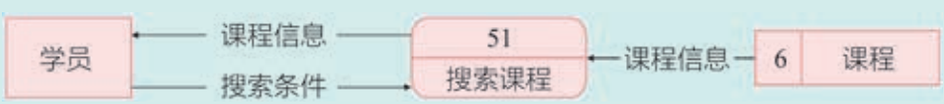


图2.1.8 学员搜索课程数据流图

一个“学员”实体将数据流“搜索条件”输入到处理程序“搜索课程”，处理程序“搜索课程”从存储“课程”中得到数据流“课程信息”，随后处理程序“搜索课程”将数据流“课程信息”返回给“学员”实体。处理程序“搜索课程”上的数字“51”代表分层数据流图的编号，存储“课程”中的数字“6”代表存储的编号。

2.1.3 数据需求

数据需求的主要工作是描述项目涉及的数据以及数据之间存在的关系。在数据管理与分析项目中，数据是整个项目的核心，包括需要处理的数据和产生的数据，其准确性很大程度上决定了项目实施的成败。

数据需求分析通常需要建立一个能正确反映项目数据和数据之间关系的概念模型。概念模型通常采用图形化的方式来表示，以便数据的描述更加直观、易于理解。常用的概念模型的表示方法有E-R图和UML类图。

UML（统一建模语言）是一种能够描述问题和解决方案，起到沟通作用的语言。它是一种用文本、图形和符号的集合来描述现实生活中的各类事物、活动及其之间关系的语言。UML中的类图能够很好地描述实体、实体集以及实体之间的联系，可以用来表示概念模型。图2.1.9是表示班级与学生关系的UML类图，班级端的“1”代表一个学生必定属于一个班级，学生端的“1..*”表示一个班级必须有一到多个学生。

图2.1.9　班级与学生关系的UML类图

1. E-R图

E-R图是一种实体联系数据模型的图形化表示方法，是当今概念模型设计中应用最广泛的表示概念模型的方法之一。E-R图的基本组成是：用矩形表示实体集，用菱形表示实体之间的联系，用椭圆形表示实体集或联系集的属性，如图2.1.10所示。按照实体、联系和它们的属性之间的现实关系，用线段把它们连接起来，构成数据库设计所需要的一幅或多幅E-R图。

图2.1.10　E-R图符号

（1）实体间的联系

实体之间的联系分为一对一联系、一对多联系和多对多联系。

①一对一联系（1∶1）。如果对于实体集A中的每一个实体，实体集B中至多有一个（也可以没有）实体与之联系，反之亦然，就称实体集A与实体集B具有一对一联系，记为1∶1。例如，一个汽车车牌号只属于一辆汽车，一辆汽车只能拥有一个汽车车牌号。

②一对多联系（1∶n）。如果对于实体集A中的每一个实体，实体集B中有n（n≥0）

个实体与之联系，反之，对于实体集B中的每一个实体，实体集A中至多只有一个实体与之联系，就称实体集A与实体集B有一对多联系，记为1∶n。例如，一个学校可以有多个班级，但是每个班级都只能从属于一个学校，则学校和班级是一对多联系。

③多对多联系（m∶n）。如果对于实体集A中的每一个实体，实体集B中有n（n≥0）个实体与之联系，反之，对于实体集B中的每一个实体，实体集A中也有m（m≥0）个实体与之联系，就称实体集A与实体集B具有多对多联系，记为m∶n。例如，一个学生可以选修多门课程，一门课程也可以被多个学生选修，则学生和课程之间是多对多联系。

实体之间的联系（一对一联系、一对多联系、多对多联系）用E-R图表示的方法如图2.1.11所示，具体实例如图2.1.12所示。

一对一联系　　一对多联系　　多对多联系

图2.1.11　实体之间的三种联系

图2.1.12　实体之间的三种联系的实例

（2）实体和联系的属性

如果学生实体有学号、姓名、性别、出生日期、共青团员、电话、QQ号等属性，那么用E-R图表示如图2.1.13所示。除了实体有属性以外，联系也可以有属性。比如，学生选修课程的多对多联系，学生学习课程会有一个成绩，这个成绩就是联系“选修”的属性，如图2.1.14所示。

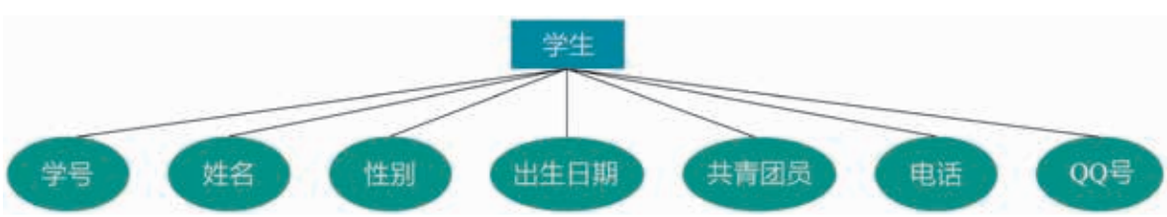


图2.1.13　学生实体及属性E-R图表示

图2.1.14　联系及属性举例E-R图表示

2.“云课堂学习平台”E-R图

分析“云课堂学习平台”业务流程图可以得到项目相关的数据，这些数据如表2.1.1所示。

表2.1.1　“云课堂学习平台”实体及属性

实体	实体的属性
学员	账号、密码、身份证、姓名、性别、头像、手机、邮箱
讲师	账号、密码、姓名、个人简介
课程	课程编号、名称、简介
学习视频	视频编号、名称、视频地址
在线测评	在线测评编号、名称、测评内容

该平台实体与实体之间的联系如表2.1.2所示。

表2.1.2　“云课堂学习平台”联系及属性

联系类型	联系名称	联系的属性	备注
一对一	课程—包含—在线测评	—	
一对多	讲师—发布—课程	—	
	课程—包含—学习视频	—	
多对多	学员—购买—课程	价格	
	学员—学习—学习视频	观看进度、评价、评分	“观看进度”是为了实现继续观看的功能，“评价”和“评分”是为了实现对课程学习视频质量的监控
	学员—答题—在线测评	答案、评分	“答案”与“评分”分别为学员在线测评的作答情况与得分

通过表2.1.1和表2.1.2可以得到该平台的E-R图如图2.1.15所示。

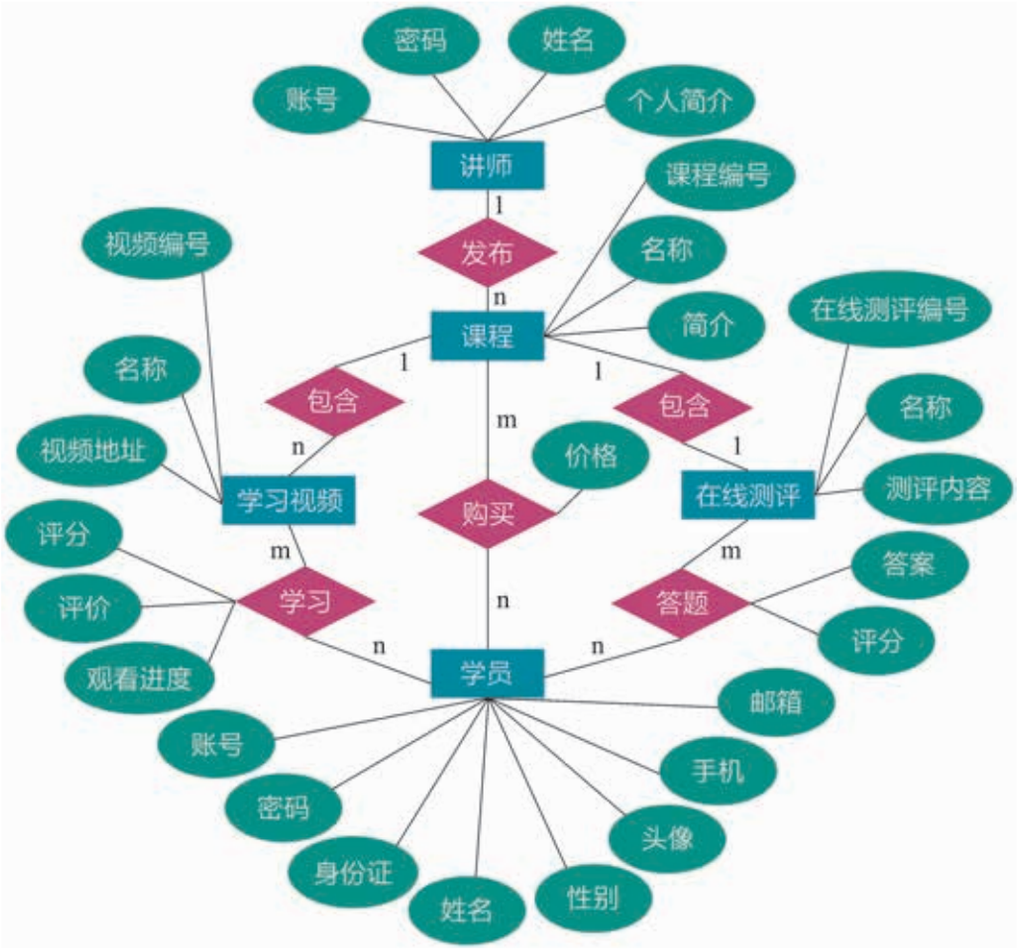


图2.1.15 “云课堂学习平台”E-R图

2.1.4 非功能需求

除了功能需求、数据需求以外，还有描述项目质量要求的非功能需求。非功能需求确保项目能够实际投入使用。如果在需求分析时不考虑这些因素，项目很有可能无法满足实际应用的需要。非功能需求主要包含安全性、可靠性、易用性、性能、可维护性、可移植性等。

实践与体验

“学校网络问答社区平台”需求分析

学校学生会计划建立一个学校网络问答社区平台，通过建设该平台帮助广大同学解决生活与学习中的困惑。

实践内容：

1. 分析“学校网络问答社区平台”的业务流程。
2. 得到“学校网络问答社区平台”的功能需求。
3. 得到“学校网络问答社区平台”的数据需求。

实践步骤：

1. 与在校学生交流关于“学校网络问答社区平台”应该提供的主要业务，绘制业务流程图，并讨论业务流程图的完整性。

2. 从业务流程图识别、分析业务场景，得到“学校网络问答社区平台”的功能。

3. 分析业务流程图得到“学校网络问答社区平台”过程中涉及的数据，并分析其中包含的实体及属性，然后绘制E-R图。

结果呈现：

1. 绘制“学校网络问答社区平台”的业务流程图和功能图。
2. 绘制“学校网络问答社区平台”的E-R图。

思考与练习

1. 公司的员工与公司之间的联系属于哪一类？画出公司与员工及其联系的E-R图。

2. 银行客户与其银行账号的联系属于哪一类？画出银行客户与银行账号及其联系的E-R图。

3. 购物网站上的商品和订单的联系属于哪一类？画出商品与订单及其联系的E-R图。

2.2 方案设计

通过需求分析，明确了数据管理与分析项目必须“做什么”，接下来要考虑项目应该“怎么做”，即设计数据管理与分析项目的方案。方案设计阶段的目标是解决“项目应该如何实现”的问题，主要包括数据管理方案和数据分析方案。

多年来，许多研究者对数据管理与分析技术进行了大量的研究，沿着不同的道路进行有益的探索，形成了各自的知识体系，内容非常丰富，这使得方案设计没有既定的模式。不同的项目管理与分析数据的方式、方法不同，方案设计也会不同。开发人员的专业背景、已有的经验和积累的资源均会影响方案的设计。本节以“云课堂学习平台”项目为例来阐述方案设计。

2.2.1 数据管理方案设计

数据管理与分析项目要管理大量数据。建立一个良好的数据管理体系，使整个系统可以安全、迅速、方便、准确、完整地管理与使用数据，是数据管理方案设计的主要任务。数据的管理方式有数据库与文件两种方式。数值、文本、日期等类型的数据通常采用数据库管理；音频、图像、视频等数据类型既可以用数据库管理，也可以用文件系统管理。

1. 数据库管理方案

数据库管理方案包括管理的数据及其类型、逻辑模型选择和数据库选择。

（1）管理的数据及其类型

分析本章“云课堂学习平台”项目需求分析的E-R图中的实体集、联系及属性，可以得到数据库管理的实体、联系的属性的数据类型如表2.2.1所示。

表2.2.1 “云课堂学习平台”实体、联系的属性的数据类型

实体/联系	属性	属性的数据类型
讲师	账号	文本
	密码	文本
	姓名	文本
	个人简介	文本

续表

实体/联系	属性	属性的数据类型
学员	账号	文本
	密码	文本
	身份证	文本
	姓名	文本
	性别	文本
	头像	图像
	手机	文本
	邮箱	文本
课程	课程编号	数值
	名称	文本
	简介	文本
学习视频	视频编号	数值
	名称	文本
	视频地址	文本
在线测评	在线测评编号	数值
	名称	文本
	测评内容	文本
学员—购买—课程	价格	数值
学员—学习—学习视频	观看进度	时间
	评分	数值
	评价	文本
学员—答题—在线测评	答案	文本
	评分	数值

分析表2.2.1，学员的头像是一张照片。图像容量较小，可以直接存放在数据库中，也可以采用文件系统存放，在数据库中存储图像文件的文件名。学习视频容量较大，不适合直接存放在数据库中，通常采用文件形式存储，在数据库中保存视频的文件名。

（2）逻辑模型选择

在选择逻辑模型的时候，要考虑管理的数据的结构化程度和特点。管理的数据是结构化的，数据与数据之间存在联系，通常选择关系模型。管理的数据是半结构化的，通常选择对半结构数据支持良好的逻辑模型，比如文档模型。除此之外，开发人员的已有经验也是需要考虑的。如果开发人员擅长关系数据库，要管理的数据是半结构化的，也可以将半

结构化数据转化为结构化数据，再使用对结构化数据支持良好的逻辑模型建模。如果开发人员擅长文档数据库，可以直接将结构化数据建模转化为对半结构化数据支持良好的逻辑模型。对于“云课堂学习平台”，逻辑模型的选择方案如表2.2.2所示。

表2.2.2 “云课堂学习平台”数据模型的选择

<table>
<tr><td>数据的结构化程度</td><td>☑结构化</td><td>□半结构化</td><td>□非结构化</td></tr>
<tr><td>数据的特点</td><td colspan="3">大部分数据与数据之间存在联系</td></tr>
<tr><td>开发人员经验</td><td colspan="3">擅长使用MySQL、PostgreSQL、SQL Server等关系数据库
擅长使用MongoDB文档数据库</td></tr>
<tr><td rowspan="2">适合的逻辑数据模型</td><td>☑关系模型</td><td>□列族模型</td><td rowspan="2">□键值模型</td></tr>
<tr><td>☑文档模型</td><td>□图模型</td></tr>
</table>

根据数据的结构化程度和特点，选用合适的数据模型。管理的数据是结构化的，大部分数据与数据之间存在联系，可以采用关系模型。

（3）数据库选择

数据库的选择主要从以下几个方面考虑：

①逻辑模型。不同数据库存储数据对应的逻辑模型不同。关系数据库对应关系模型，文档数据库对应文档模型。

②数据量。如果管理的数据量较小，普通的桌面型数据库（如Microsoft Access）就可以胜任；如果管理的数据较大，可以选用MySQL、SQL Server等企业级数据库；如果需要管理的数据量很大，要考虑使用支持大数据管理的分布式数据库，比如Hadoop、MongoDB、CouchBase等。

③并发要求。小或中的并发量，用一般的数据库就可以；如果数据量巨大，同时访问、使用的用户很多，这个时候要考虑支持高并发的数据库。比如淘宝，后端数据存储层采用的是基于MySQL的分布式关系数据库集群MyFOX和基于HBase的NoSQL存储集群Prom。

④开发人员的经验。开发人员的经验是需要考虑的一个很重要的因素。选用的数据库应该是开发人员熟悉的数据库，这样可以尽量避免开发过程中出现各种和数据库相关的问题，也可以缩短开发周期。

⑤成本。在满足要求的情况下，可以选用开源的数据库。购买商业数据库需要更多的资金投入，但商业数据库提供了很多后续的支持和服务。

对于“云课堂学习平台”，具体的数据库选择方案如表2.2.3所示。综合以上多方面的考虑，可以选择甲骨文公司的MySQL社区版或企业版。

表2.2.3 “云课堂学习平台”数据库选择

考虑因素	说明
逻辑模型	关系模型
数据量	1. 平台讲师数量最多200人 2. 平台学员数量最多2500人 3. 平台课程数量最多1000门，每门课程不超过30个课程视频，总课程学习视频数不超过30000个
并发要求	1. 平台高峰期时预估最多有500名学员同时在线学习课程视频，对课程学习视频进行评价、打分，完成在线评价反馈 2. 平台最多有500名学员同时在线并对数据库中存储的数据进行读/写操作
开发人员经验	有10年以上关系数据库开发的经验，擅长MySQL、Oracle、SQL Server数据库开发
成本	MySQL企业版单服务器提供1年服务的价格大约是3万元；提供7×24小时的技术支持服务；提供企业级备份可为数据库提供联机“热”备份，从而降低数据丢失的风险

2. 文件管理方案

文件管理方案设计的主要任务是根据文件管理需求选择最适合的文件存储方式。文件存储的选择主要考虑以下因素：文件的大小、文件的数量、并发量、性能。在文件数量较少、并发量和性能要求不高的情况下，使用计算机操作系统自带的文件管理系统就能胜任。如果文件数量大、并发量和性能要求高，可选择分布式文件系统来存储文件。

“云课堂学习平台”中需要保存每名学员的头像，如果人数不多，可以使用操作系统的文件系统直接保存这些照片；如果人数较多，就需要考虑使用分布式的小文件存储了。如果平台使用人数不多，平台上传的学习视频可以直接保存在操作系统的文件系统中；否则，还要考虑大文件被用户访问所需的带宽，不仅要使用支持大文件的存储系统，还要有足够的网络带宽。

对于“云课堂学习平台”，具体的数据库选择方案如表2.2.4所示。

表2.2.4 “云课堂学习平台”文件管理方案

数据	文件大小	数量	并发	存储方式
学员照片	200～500KB	2500	1. 最多50 2. 除了管理员，很少有学员会看其他学员照片	文件系统或数据库
学习视频	100～200MB	30000	1. 最多500 2. 同时可能有500人在线观看 3. 学习视频以平均码率2Mbps计算，需要1Gbps带宽	阿里云视频存储

拓展链接

TFS

TFS（Taobao File System）是淘宝开源的海量小文件存储项目。淘宝上有海量的小文件，比如产品图片，但是它们的大小都不超过1MB。为了提供高可靠性、高并发的存储访问，满足淘宝对小文件存储的需求，淘宝开发了TFS并将其开源，TFS被广泛地应用在淘宝各项应用中。

2.2.2 数据分析方案设计

数据分析方案主要包含分析要求、数据收集、数据预处理、分析方法、分析工具与结果呈现方式6个方面。

1. 分析要求

一般情况下，数据分析是为了控制业务开展、规避风险、控制结果。

“云课堂学习平台”项目的分析要求如表2.2.5所示。

表2.2.5 “云课堂学习平台”项目分析要求

分析要求	说明
学员兴趣点分析	教育培训机构研发“云课堂学习平台”是为了通过平台进行盈利，这就需要更多的学员购买在线课程
课程质量监控	学员也许对课程的内容感兴趣，但是如果课程的质量不行，学员也不会购买课程，所以要监控课程的质量
课程难易度监控	保证课程的难易度是学员能够接受的：太难，学员无法继续学习；太简单，学员就会失去兴趣

2. 数据收集

对于“云课堂学习平台”项目，要挖掘学员兴趣点，可以使用爬虫软件到互联网上的慕课网站获取其开设的课程的学习情况与课程评价数据来分析学员较为感兴趣的课程。也可以直接从“云课堂学习平台”的数据库中提取学员购买课程的情况，来分析学员较为感兴趣的课程。要监控课程的质量，可以从“云课堂学习平台”数据库中提取课程学习视频的评分，通过分析评分，确保平台推出的课程质量。要监控课程难易度，可以从“云课堂学习平台”数据库中提取在线测评的成绩，通过分析成绩，监控课程的难易度。

3. 数据预处理

数据预处理描述对数据进行加工、整理的方法，使其能够符合数据分析的要求和规范。在加工和整理的时候，必须在方案中指明具体的数据处理方法。

“云课堂学习平台”为了分析学员兴趣点，需要从互联网上获取其他慕课网站开设课程的情况。爬虫软件爬取的数据主要是开设的课程信息（课程类别等）以及课程的学习情况（学习人数、浏览人数等）。数据爬取下来后，首先要对重复的课程类别进行去重，相似的课程视为同一课程类别；去除那些学习人数、浏览人数不完整的课程；对于课程的学习人数、浏览人数，爬取下来的数据有可能不是数字，比如学习人数的数据为“1280人次”“一千”或“1k”，需要通过转换得到对应的学习人数。

4. 分析方法

分析方法描述了对数据进行怎样的分析。

对于“云课堂学习平台”，要监控课程质量，需要对课程学习视频的评分进行平均分析与对比分析。要监控课程的难易程度，需要对在线测评的成绩进行平均分析。要挖掘学员兴趣点，需要进行对比分析。

5. 分析工具

对于从慕课网站爬取的相关开设课程的学习情况与课程评价的数据，可以使用Excel进行分析。从“云课堂学习平台”数据库提取的数据，则可以直接编写SQL语句查询进行分析。

6. 结果呈现

呈现是为了把数据分析的结果直观地表现出来。呈现一般采用表格和图形的方式，图形比较直观，常用的图形呈现方式有柱形图、折线图、饼图、条形图等。

对于“云课堂学习平台”，可以使用标签云来呈现学员关注的兴趣点，使用柱形图来展示不同课程的平均分，体现不同课程的难易度。

2.2.3 方案的评价与优化

方案的评价与优化主要包括数据管理方案和数据分析方案的评价与优化。

1. 数据管理方案的评价与优化

数据管理方案的评价主要考虑以下几个方面：

①数据管理的方式是否合适，数据是应该放在数据库中进行管理还是用文件系统来管

理。在“云课堂学习平台”项目的方案设计中，学习视频属于大文件，适合用云视频存储来管理，如果存放在数据库中，就会对数据库的访问造成巨大的压力。而对于学员头像的图像数据，容量比较小，可以存储在数据库或文件系统中。

②数据库管理方案中的逻辑模型选择是否合理。逻辑模型能正确地表示和反映项目管理的数据，符合项目管理的数据的特点是合理的。在“云课堂学习平台”项目的方案设计中，管理的数据是结构化的，根据开发人员经验选择关系模型或文档模型都是合理的。

③数据库管理方案中的数据库选择是否合适。能满足项目的存储要求、并发要求，符合开发人员的经验和习惯，数据库的选择就没有问题。

④文件系统管理数据的方式是否合适，主要考虑的因素是并发和性能。“云课堂学习平台”项目的方案设计中要管理的学员照片，其数量相对来说不是很多，普通的文件系统就可以胜任项目的要求。

数据管理方案的优化主要考虑以下几个方面：

①数据库管理方案中的数据库选择的优化，主要从并发和性能方面考虑。项目实施后，管理的数据如果数据量增长比较快、并发要求比较高，要选用支持大数据、支持分布式、具有良好并发和性能的数据库产品。

②文件管理方案的优化。项目实施后，管理的文件数据量增长比较快，文件系统有可能无法胜任，比如“云课堂学习平台”中的学员照片，后期如果学员人数越来越多，传统的文件系统会不合适，就要改用分布式的小文件存储。学员增多以后，文件访问的带宽会不够，文件存储方式需要升级为专业的支持高性能和并发的大数据存储。

2. 数据分析方案的评价与优化

数据分析方案的评价主要考虑以下几个方面：

①数据收集的方式是否合理、有效。

②收集的数据预处理后是否符合数据分析的要求和规范。

③分析方法的选择是否正确。采用的数据分析方法得到结果是准确的，以及分析结果具有指导意义。

④分析工具的选择是否正确。分析工具简单且功能强大，符合分析人员的使用经验和习惯，这样的分析工具选择是合理的。

⑤结果呈现是否反映分析结果。比如用图形来呈现分析结果的时候，选择的图表类型是否适合。

数据分析方案的优化主要考虑以下几个方面：

①数据收集在保证合理、有效的前提下，尽量通过多个渠道进行收集，收集的数据量越多越好。“云课堂学习平台”要分析学员的兴趣点，还可以增加一个途径，通过网络问卷调查的方式去收集学员兴趣点的数据。

②采用分析结果更加精准的分析方法。“云课堂学习平台”中学员兴趣点的分析，还可以对不同年龄段、不同职业的学员进行分组分析，得到针对特定人群的更加精准的分

析结果。

③采用更好的结果呈现方式。有些结果数据用表格无法直观地呈现，可以采用图形化的方式进行呈现；有些数据是动态变化的，用静态图无法较好地呈现，可以采用动态图的方式进行呈现。

思考与练习

1.“学校网络问答社区平台”项目中管理的数据是什么类型的数据？哪些数据可以存储在数据库中，哪些需要以文件的形式存储？

2. 开发人员有多年MongoDB、MySQL数据库开发经验，现在需要管理半结构化的电子病历数据，用什么逻辑模型来表示电子病历数据，选用什么数据库来管理电子病历数据比较合适？

3. 要开发一个在线文档存储服务，其存储的文件主要是文档（Word、Excel文档），每个文档的大小不超过5MB，应该如何管理这些文档数据？

4. 简述数据分析方案设计的主要内容及其注意点。

巩固与提高

1. 某酒店需要开发一个酒店管理系统，希望通过该系统来提高酒店管理的效率。该酒店管理系统包含一个住宿管理子系统，主要是为了方便管理住店客人的数据和住宿流程。

（1）住店客人的主体业务流程是“住宿流程”，绘制其业务流程图。

（2）讨论主体业务流程“住宿流程”有哪些变体业务流程？

（3）讨论主体业务流程“住宿流程”有哪些支撑业务流程？

（4）对主体业务流程“住宿流程”及其变体业务流程和支撑业务流程进行分析，得到这些业务流程相关的数据及相互间的关系，并对这些数据进行建模得到其E-R图。

2. 某学校想要开发一个选修课管理系统来管理选课流程以及学生、选修课和选修课考试成绩等与选修课相关的数据。通过需求分析，得到该系统数据需求的E-R图如图2.2.1所示。

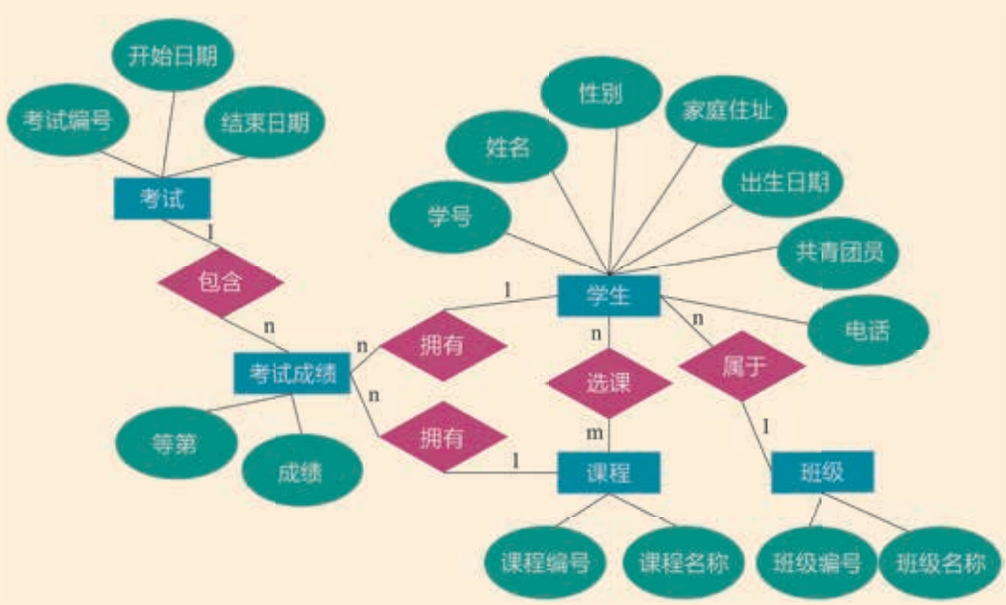


图2.2.1　选修课管理系统数据需求的E-R图

请完成以下任务：

（1）分析该系统管理的数据及其类型。

（2）列举管理该系统数据适用的数据模型。

（3）列举该系统的数据可以做哪些数据分析，涉及的数据有哪些。

项目挑战

共享汽车租赁管理系统数据管理与分析方案设计

共享汽车租赁是指由公司提供电动汽车，市民可随时在指定地点租赁，方便市民出行的一项服务。为了推进公司的信息化管理，提升公司的工作效率，某共享汽车租赁公司需要一个共享汽车租赁管理系统，能够存储客户、汽车等信息，以便系统、科学、安全和方便地管理公司的各项业务。要想开发这样一个管理系统，首先需要设计该管理系统的数据管理与分析方案。

项目任务

对“共享汽车租赁管理系统”项目进行业务流程分析、功能需求分析、数据需求分析等，最终设计完成该项目的数据管理与分析方案设计。

过程与建议

为了顺利开展本项目的研究，建议组建研究小组，在充分理解具体工作要求的基础上，分工协作，共同开展本次任务。

1. 项目的业务流程分析

与小组成员交流“共享汽车租赁管理系统”应该提供的主要服务，绘制业务流程图并讨论业务流程图的完整性。从主体业务流程、变体业务流程、支撑业务流程、管理流程多方面进行讨论。

思考：为了让这个项目能够提供更好的服务，在流程上要做哪些方面的努力。如何提供更加个性化、灵活、精准的服务，比如用户账号可以绑定支付宝或者微信，用户租车的时候可以直接使用支付宝或微信付款。

2. 功能需求分析

从业务流程图识别、分析出业务场景，得到“共享汽车租赁管理系统”的功能，并绘制功能图。

（1）识别、分析业务场景。

角色	业务活动
用户	租车、还车、_____

续表

角色	业务活动

（2）绘制功能图。

3. 数据需求分析

（1）分析业务流程图，得到“共享汽车租赁管理系统”项目涉及的数据。

实体/联系	实体/联系的属性
用户	账号、密码、姓名、身份证、住址、手机号码、机动车驾驶证号
车辆	车牌、品牌、车型、车龄、发动机号

（2）绘制项目的E-R图。

（3）梳理需要分析的数据。

数据	数据的意义
投放点汽车日租赁量	调整投放点汽车数量
用户满意度	

4. 撰写数据管理与分析方案

（1）数据管理方案。包括数据库管理方案和文件管理方案。数据库管理方案包含管理的数据及其类型、逻辑模型选择与数据库选择。

（2）数据分析方案。包括分析要求、数据收集、数据预处理、分析方法、分析工具与结果呈现。

5. 展示与交流

将完成的数据管理与分析方案制作成演示文稿，小组派代表进行展示，展示中应突出各自方案设计中的亮点。展示与交流完毕后，各小组相互之间进行评价、探讨，进一步完善方案设计。

评价标准

请根据项目实施的过程、效果以及成果展示交流的结果，对自己完成项目的情况进行客观的评价，并思考后续完善的方向。把评价结果和完善方案填写在下面的表格中。

评价条目	说明	评分（1～10分）	评分主要依据阐述	后续完善方向
业务流程图	绘制的业务流程图业务流程合理、流程图完整			
服务支持	支撑业务流程图的业务流程能提供更加个性化、灵活、精准的服务			
功能	功能完整			
E-R 图	绘制的E-R 图能准确地反映项目的数据			
数据管理与分析方案	方案完整，表述清晰			
小组合作	小组分工合理、职责明确			

拓展项目

许多刚刚踏入校园的高一新生对学校不是很了解，例如，他们不了解学校各幢楼的分布，不知道教室、操场、食堂等的位置，因此也无从规划从一个地点到另一个地点的最短路径。为了让高一新生能够快速地熟悉校园环境，学校决定开发一个校园导航系统，它应该包含如下功能：

- 能显示自己当前在校园中的位置。
- 输入目的地，马上能够在导航中显示当前位置到目的地的最短路径。
- 能查看校园内的各个位置，并给出位置的附加信息。
- 能够分析哪些地点是同学们经常去的。

学校希望学生会来负责校园导航系统的开发工作，并招募了一些同学作为开发人员。现在需要完成以下工作：

（1）对校园导航项目进行需求分析。

（2）建立校园导航项目的数据模型。

第三章 3

数据管理

数据已经完全融入人们的日常生活。通过高效的数据管理，能合理提取数据，获取有效信息，为数据分析和决策提供依据。随着人们对数据价值的认识进一步提高，数据安全与备份的意识也在逐步加强。

问题与挑战

- 自动驾驶汽车通过摄像头、激光雷达、毫米波雷达等传感器收集汽车及其周围环境的数据。在汽车自动驾驶过程中，收集到的数据是如何存储与管理的？

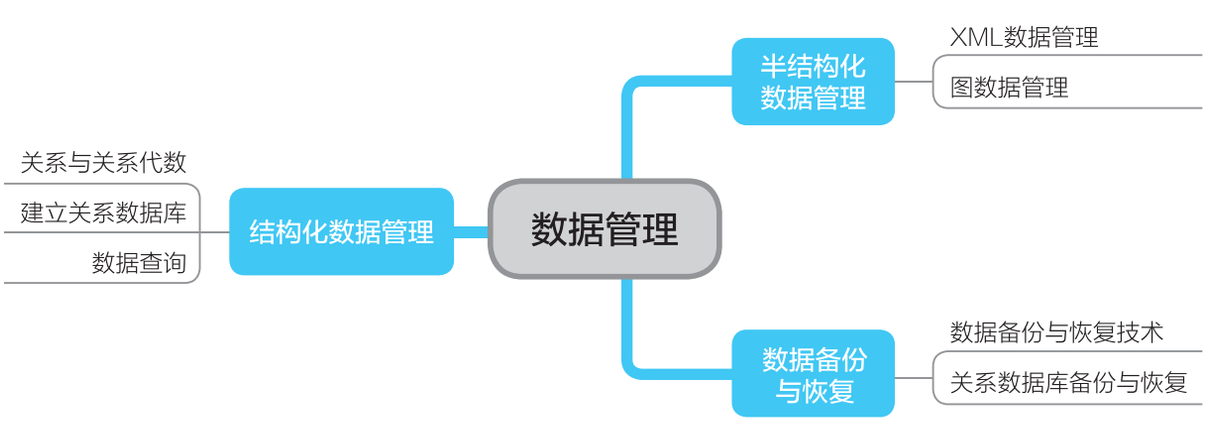
- 若共享汽车租赁项目已完成E-R图，计划使用关系数据库管理数据。为了提升服务，提高共享汽车系统管理效率，达到个性、精准、优质灵活的管理目标，如何提取客户信息、车辆信息、动态租赁情况、用户满意度等数据？

- 百度公司收集了互联网上的大量数据，处理并存储后，通过搜索引擎提供检索服务，为用户呈现所需的数据。这些收集到的海量数据是如何管理的？

学习目标

1. 掌握简单关系数据库的逻辑结构，能建立关系数据库。
2. 能根据实际需要，使用SQL语言查询数据、提取信息。
3. 知道不同的数据管理技术及其特点。
4. 认识数据丢失的后果，能选用合适的方法备份数据。

内容总览



3.1 结构化数据管理

关系数据库是最常用、最普遍的数据库，使用关系数据库管理数据实现了数据结构化、数据与程序分离以及数据统一管理与控制，数据拥有了更高的独立性，从而使数据具有高共享、低冗余和易扩充的特性。使用标准的查询语言，可以高效地查询数据、获取信息。

3.1.1 关系与关系代数

关系数据库对数据的操作建立在关系代数的基础上，这使得关系数据库能支持复杂的关系运算，从而能高效地提取数据。

1. 关系

关系模型的基本数据结构就是关系，通俗地讲，关系模型的逻辑结构就是一张扁平的二维表。二维表的每行对应一个元组，每列是同质的，同列的每个值是同一类型的数据。不同列的数据类型可以不同，每列的名称称为属性。

关系数据结构具有简单、灵活、存储效率高、数据类型一致和数据不可再分割等特点，因此在结构化数据组织过程中得到了广泛的应用。

2. 关系操作

关系模型中常用的关系操作包括查询操作和插入、删除、修改操作两大部分。关系的查询表达能力很强，是关系操作中最主要的部分。查询操作主要有选择、投影、连接、除、并、差、交、笛卡儿积等，其中选择、投影、并、差、笛卡儿积是5种基本操作。关系操作的特点是集合操作方式，即操作的对象和结果都是集合。

3. 关系代数

关系代数是一种抽象的查询语言，用对关系的运算来表达查询，是研究关系数据语言的数学工具。关系代数的运算对象是关系，运算结果也是关系。关系代数的运算按运算符的不同可分为传统的集合运算和专门的关系运算两类。其中，传统的集合运算包含并、交、差、广义笛卡儿积4种运算。专门的关系运算包括选择、投影、连接等运算。数据集R与数据集S如图3.1.1所示。

图3.1.1 数据集R、S

并运算，返回属于R或属于S的元组集合。
交运算，返回属于R且属于S的元组集合。
差运算，如R与S的差，返回属于R而不属于S的元组集合。
数据集R和S的并、交、差运算的结果如图3.1.2所示。

图3.1.2　数据集R和S的并、交、差关系运算维恩图

集合A和集合B的数据情况如表3.1.1，集合A和集合B的交运算结果如表3.1.2。

表3.1.1　集合A和集合B

集合A			集合B		
id001	手机	60	id001	手机	60
id002	笔记本	68	id002	笔记本	68
id003	台式电脑	45	id004	平板电脑	24
id004	平板电脑	24	id005	数码相机	23
id005	数码相机	23	id007	投影仪	23

表3.1.2　集合A和集合B的交集

id001	手机	60
id002	笔记本	68
id004	平板电脑	24
id005	数码相机	23

笛卡尔积运算对数据集R与S的列数不做要求，可以不同，如R为n列，S为m列，数据集R和S的笛卡儿积返回一个（n+m）列元组的集合，每个元组前n列是R的一个元组，后m列是S的一个元组。若R有k个元组，S有j个元组，则数据集R和S的笛卡尔积有k×j个元组，如图3.1.3所示。

R

1	1-1
2	2-1

S

a	a1	a2
b	b1	b2
c	c1	c2

笛卡尔积

1	1-1	a	a1	a2
2	2-1	b	b1	b2
1	1-1	c	c1	c2
2	2-1	a	a1	a2
1	1-1	b	b1	b2
2	2-1	c	c1	c2

图3.1.3　数据集R和S的笛卡尔积

3.1.2 建立关系数据库

借助计算机和数据库技术可以科学地保存数据，高效地管理数据。

1. 数据库管理系统

使用关系数据库管理数据，首先要选择和安装关系数据库管理系统软件，这样数据才能保存到数据库，达到数据独立、共享和统一管理等目标。

问题与讨论

数据库管理系统产品繁多、各具特色，常见的有Oracle、DB2、SQL Server、MySQL、Access等。通过网络检索相关信息，讨论几种数据库管理系统的特点。

2. 创建关系数据库

安装数据库管理系统软件后，可以使用数据库管理系统创建数据库，也就是在数据库系统中划分一块存储数据的空间。关系模型是关系数据库支持的逻辑模型，在创建数据表之前要完成从E-R图到关系模型的转化，随后再根据关系模式建立数据表。

（1）E-R图中的实体转化为关系模式

在图2.1.15所示的“云课堂学习平台”系统的E-R图中，学员实体的名称“学员”是关系模式的关系名，“学员”实体的属性“账号”“密码”“身份证”“姓名”“性别”“头像”“手机”“邮箱”是关系的属性，学员实体的属性“账号”是关系的码，学员实体的关系模式为：

学员（账号，密码，身份证，姓名，性别，头像，手机，邮箱）

E-R图中的实体转化为关系模式时一般是将实体转化为一个关系模式、实体的属性转化为关系的属性、实体的码转化为关系的码。转化后“云课堂学习平台”系统中的讲师、课程、学习视频、在线测评的关系模式如下：

讲师（账号，密码，姓名，个人简介）
课程（课程编号，名称，简介）
学习视频（视频编号，名称，视频地址）
在线测评（在线测评编号，名称，测评内容）

（2）E-R图中的联系转化为关系模式

E-R图中的实体转化为关系模式后，实体之间的联系也需要转化为关系模式，具体有一对一联系、一对多联系和多对多联系到关系模式的转换。

一对一联系转化为关系模式时，一个1∶1联系可以与任意一端对应的关系模式合并，

在该关系模式的属性中加入另一个关系模式的码和联系本身的属性。例如，在“云课堂学习平台”中，课程实体与在线测评实体的联系是一对一联系。课程编号唯一标记课程，是课程关系的码，在线测评关系模式中加一个码“课程编号”，“包含”联系被转换到了其中的一端，得到新的在线测评关系模式：

在线测评（在线测评编号，名称，测评内容，课程编号）

一对多联系转化为关系模式时，一个1∶n联系可以与n端对应的关系模式合并，1端关系的码成为n端实体的属性。例如，在“云课堂学习平台”中，课程与学习视频之间是一对多联系，要将这个联系转化为关系模式，需要在学习视频实体中添加新的属性“课程编号”，学习视频关系模式变化为：

学习视频（视频编号，名称，视频地址，课程编号）

多对多联系转化为关系模式时，与m∶n联系相连的各实体的码以及联系本身属性均转化为关系的属性，各实体的码组成关系的码或关系码的一部分。例如，在“云课堂学习平台”中，学员与学习视频的联系是m∶n多对多联系“学习”，“学习”联系转化为一个单独的关系模式“学员学习视频”时，两端实体的码是这个关系模式的属性。两端的实体是“学员”和“学习视频”，学员实体的码“账号”和学习视频实体的码“视频编号”是独立关系模式“学员学习视频”的属性，如下：

学员学习视频（账号，视频编号）

联系“学员学习视频”包含属性“评价”“评分”“观看进度”，因此最终的“学员学习视频”关系模式如下：

学员学习视频（账号，视频编号，评价，评分，观看进度）

创建数据库需要设置表、索引、用户等数据库对象。关系数据库中数据的结构化是由数据库对象来约束的。关系模式建立完成后，可以在数据库中创建相应的数据表。数据库中的表对应的是相应的关系模式，也是由行和列组成的。

列对应的是关系的属性，又称为字段，每列的标题称为字段名。每个字段的取值必须使用相同的数据类型，常见的有字符型、文本型、数值型、逻辑型和日期型。不同的数据库管理系统支持的数据类型不完全相同。字段选取哪种数据类型要根据实际情况而定。一般从两个方面考虑，一是取值范围，二是要做哪些运算。如保存人年龄的字段，数据类型可采用范围较小的整数。

行由若干字段的值组成，一行数据称为一条记录或元组，是有一定意义的信息组合。表由一条或多条记录组成，没有记录的表称为空表。每个表中通常都有一个主关键字，它的值用于唯一地标识表中的一条记录。

下面以“云课堂学习平台”系统的数据库为例，学习和体验建立关系数据库的过程。

例1　使用图形化客户端创建数据库

数据管理系统采用MySQL，使用前需下载并安装MySQL，通过客户端连接后管理数据库。如图3.1.4所示的phpMyAdmin是一款图形化客户端，连接数据库后，在网页“数据库名”输入框处输入“cclp”，点“创建”按钮即可。

图3.1.4　phpMyAdmin图形化管理软件

数据库一般以目录和文件形式储存，必须符合相应命名的限制，确保规范性和唯一性。

“云课堂学习平台”系统的关系模式“讲师”有4个属性：账号、密码、姓名和个人简介，创建数据表时需要指定数据表名和相应的4个字段“id”“password”“name”和“introduction”，具体对应关系如图3.1.5所示。使用phpMyAdmin建立数据库和数据表后，还可进行后续修改。

关系模式：讲师	数据表：teachers
账号	id
密码	password
姓名	name
个人简介	introduction

图3.1.5　关系模式与对应的数据表

拓展链接

图形化管理软件phpMyAdmin

phpMyAdmin是一个以PHP语言编写的架构在网站主机上的应用系统，用于管理MySQL数据库。phpMyAdmin处理数据的导入与导出较为方便，可以远程管理MySQL数据库，方便创建、修改、删除数据库及数据表。

3. 数据导入数据库

数据库的数据表建立后默认都是空表，通过导入数据等方式添加数据，实现数据查询等后续操作。导入数据主要通过phpMyAdmin等客户端的导入功能或者程序语言导入数据等方式完成。

导入的原始数据可以是结构化数据、半结构化数据或是有一定规律的非结构化数据，根据结构化程度的差异可以采用不同的方式导入。

下面以“云课堂学习平台”系统的数据库为例，学习和体验导入数据的过程。

例2　phpMyAdmin客户端的数据导入

将“云课堂学习平台”的讲师信息表中的数据导入到数据库相应数据表。源数据保存于Excel文件中，部分如图3.1.6所示。

	A	B	C
1	name	password	sex
2	范康潇	83867162	男
3	张瑞耶	50471781	男
4	程忆宁	38628797	女
5	俞冰凤	77099032	女
6	陆诗琦	36794985	女
7	俞商思	16745443	女

图3.1.6　结构化数据导入

1. 选用phpMyAdmin图形化方式实施数据导入。

2. 原始数据格式预处理。删除标题行、调整列顺序与数据表字段顺序保持一致、转换特殊的数据类型、保存为CSV文件格式如“teachers.csv”，文件编码与数据库保持一致。

除了用phpMyAdmin图形化方式导入结构化的数据，还可使用Python或其他程序语言编程导入数据。用Python语言导入结构化数据的代码如下：

```
import pandas as pd
from sqlalchemy import create_engine
data = pd.read_csv('teachers.csv')  #读取数据
connect = create_engine('mysql+pymysql://root:pwd@localhost:3306/cclp?charset=utf8')
#连接数据库cclp
data.to_sql('teachers', connect, index=False, chunksize=10000, if_exists='replace')
#创建数据表teachers并导入数据
```

数据库导入的数据不局限于简单的结构化数据，在需求分析的基础上，使用程序语言等方式对原始数据预处理，使处理后的结构化数据符合数据模型的需求，导入关系数据库后方便数据管理。

下面以普通高校招生计划数据为例，学习和体验原始数据预处理和导入的过程。

例3　程序语言导入数据

原始数据为Excel文件。由于每行数据结构差异性较大，在Excel中也较难开展数据查询或筛选操作，需要使用Python程序语言进行预处理，生成包含院校（专业）名称、科类、录取人数、平均分、最低分和名次号的结构化数据，导入数据库后管理数据。部分原始数据如图3.1.7所示。

	A	B	C	D	E	F
1	2016年浙江省普通高校招生文理科录取专业平均分、最低分和名次号情况					
2	文理科第一批					
3	院校代码、院校(专业)名称、所在地	科类	录取人数	平均分	最低分	名次号
4	【浙江】					
5	0187浙江大学(农科大类)(浙江·杭州)					
6	*应用生物科学(农学)	理	294	665	660	7773
7	*应用生物科学(生工食品)	理	70	672	667	5787
8	*工科试验班(海洋)	理	60	670	664	6622
9	*海洋科学	理	24	667	662	7204
10	0001浙江大学(浙江·杭州)					
11	*人文科学试验班【外国语言文学(非英语)】	文	29	663	661	878
12	社会科学试验班	文	73	678	675	322
13	人文科学试验班	文	115	669	665	685

图3.1.7　部分原始数据

Python程序语言对原始数据的预处理代码：

```
import pandas as pd
file = "2016专业平均分、最低分和名次号情况（普通类）.xlsx"
df = pd.read_excel(file, names=['院校（专业）名称', '科类', '录取人数', '平均分', '最低分', '名次号'])
#读取数据并自定义列名
df['学校代码'] = df['学校名称'] = df['批次'] = 'nothing'
#增加列并赋初值
df['年份'] = file[:4]
info = {'xxdm': 7, 'xxmc': '', 'pici': ''}
#学校代码（源数据值没有7）、学校名次、批次
for i in range(len(df)):
    sf = df.iat[i, 0]
    if sf.endswith('批'):
        info['pici'] = sf[-3:]                    #批次赋值
    elif sf[0:4].isnumeric():
        info['xxdm'] = sf[0:4]                    #学校代码赋值
        xm = abs(sf.find('(') * sf.find('('))     #校名定位
        info['xxmc'] = sf[4:xm]                   #学校名称赋值
        wz = abs(sf.rfind('(') * sf.rfind('('))   #位置定位
    elif info['xxdm'] != 7:                       #学校代码、学校名称、批次赋值
        df.iat[i, 6] = info['xxdm']
        df.iat[i, 7] = info['xxmc']
        df.iat[i, 8] = info['pici']
dfo = df[df.科类.isin(['文', '理'])]              #筛选数据
dfo = dfo.reset_index(drop=True)
dfo = dfo['年份', '批次', '学校代码', '学校名称', '院校（专业）名称', '录取人数', '名次号']   #筛选列数据
dfo.to_excel('out-%s.xlsx' % file[:4], index=False)      #保存
```

Python导入数据库代码：数据库名'gkzs'，数据表名'zsjh'。

```
import numpy as np
import pandas as pd
from sqlalchemy import create_engine
datatype = {"年份" : np.int64,                                    #列设置类型
            "批次" : np.object,
            "学校代码" : np.object,
            "学校名称" : np.object,
            "院校（专业）名称" : np.object,
            "录取人数" : np.int64,
            "名次号" : np.int64
           }
data = pd.read_excel('out-2016.xlsx',dtype=datatype)              #读取数据
connect = create_engine('mysql+pymysql://root:pwd@localhost:3306/gkzs?charset=utf8')
#连接数据库
data.to_sql('zsjh', connect, index=False, chunksize=10000, if_exists='replace')
#创建数据表zsjh并导入数据
```

数据表'zsjh'内容如图3.1.8：

年份	批次	学校代码	学校名称	院校（专业）名称	录取人数	名次号
2016	第一批	0187	浙江大学	*应用生物科学(农学)	294	7773
2016	第一批	0187	浙江大学	*应用生物科学(生工食品)	70	5787
2016	第一批	0187	浙江大学	*工科试验班(海洋)	60	6622
2016	第一批	0187	浙江大学	*海洋科学	24	7204
2016	第一批	0001	浙江大学	*人文科学试验班［外国语言文学	29	878
2016	第一批	0001	浙江大学	社会科学试验班	73	322
2016	第一批	0001	浙江大学	人文科学试验班	115	685
2016	第一批	0001	浙江大学	人文科学试验班(传媒)	84	565
2016	第一批	0001	浙江大学	人文科学试验班(外国语言文学)	26	410
2016	第一批	0001	浙江大学	工科试验班(中外合作办学，ZJU-	15	2727
2016	第一批	0001	浙江大学	工科试验班(竺可桢学院文叉创新	20	270
2016	第一批	0001	浙江大学	社会科学试验班	158	2337
2016	第一批	0001	浙江大学	工科试验班(信息)	304	2058
2016	第一批	0001	浙江大学	工科试验班(机械与能源)	95	2403
2016	第一批	0001	浙江大学	工科试验班(材料与化工)	144	2651
2016	第一批	0001	浙江大学	工科试验班(建筑与土木)	112	2602
2016	第一批	0001	浙江大学	工科试验班(电气与自动化)	110	1346
2016	第一批	0001	浙江大学	工科试验班(航空航天与过程装备)	65	2785
2016	第一批	0001	浙江大学	科技与创意设计试验班	46	2214
2016	第一批	0001	浙江大学	理科试验班类	146	2013
2016	第一批	0001	浙江大学	理科试验班类(生命、环境与地学)	95	2784
2016	第一批	0001	浙江大学	医学试验班(八年制)	36	681

图3.1.8　数据导入数据库后的数据情况

问题与讨论

结合上述案例的学习，通过网络查询相关材料后讨论以上两种数据导入方式的异同点以及应用场景。

3.1.3　数据查询

数据查询是数据库的核心操作，利用关系数据库数据的结构化这一特点可以使用查询语言定位目标数据，多个数据集合之间通过关系操作完成更大数据范围内的数据查询。

1. 结构化查询语言（Structured Query Language, 简称SQL）

SQL是关系数据库的标准语言，也是一种通用的、功能极强的关系数据库语言。除了查询功能，还包括数据库创建，数据库数据的插入与修改，数据库安全性、完整性定义与控制等一系列功能。

SQL提供了SELECT语句实现数据查询，SELECT语句语法灵活、功能丰富。依靠高度结构化的数据，能对查询后返回的记录集再次查询，结合关系代数运算，开展选择、投影、排序和统计等操作。

SELECT语句一般格式为：

```
SELECT [DISTINCT]
    select_expr [, select_expr ...]          /*指定返回列，投影，列操作*/
    FROM tb_name                             /*操作的数据集*/
    [WHERE where_condition]                  /*选择，行操作*/
```

```
[GROUP BY {col_name | expr}]                    /*分组统计*/
[ORDER BY col_name [ASC | DESC], ...]           /*排序*/
```

2. SQL简单数据查询

“云课堂学习平台”数据库的数据表如图3.1.9所示。下面通过此案例，学习和体验使用SQL完成简单的数据查询与数据操作。

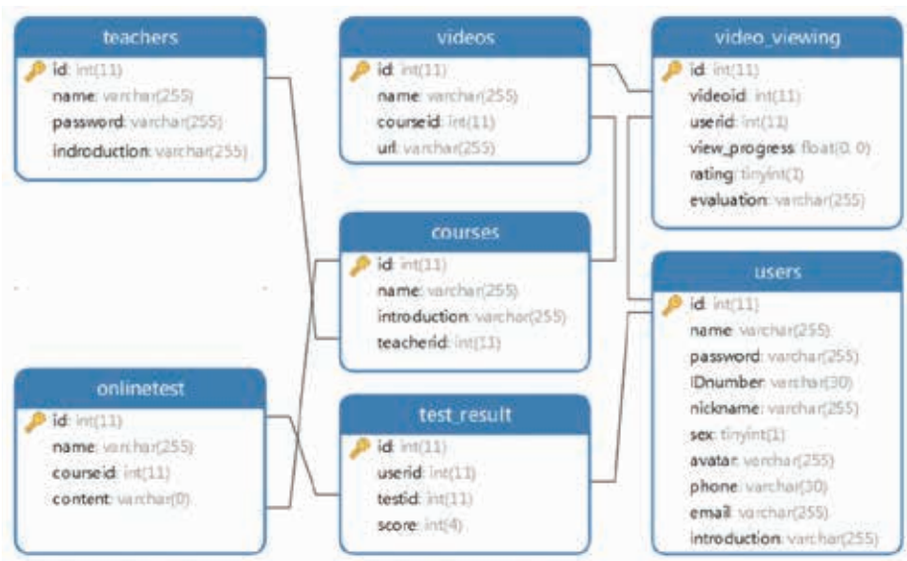


图3.1.9 “云课堂学习平台”系统数据库的数据表

（1）投影

投影查询是从关系中选择若干属性列组成新的关系。投影操作是从列的角度的运算。例如，查询讲师所有列信息时使用的SQL语句是：

```
SELECT * FROM teachers;
```

●●● 例4 投影查询讲师姓名与性别列信息并去除重复行数据。

```
SELECT DISTINCT name, sex FROM teachers;
```

（2）选择

查询结果只返回满足指定条件的记录时可以使用WHERE子句，实现关系运算中的选择运算。选择又称为限制，是在关系中选择满足给定条件的元组集合。其中选择条件是一个逻辑表达式，取逻辑值“真”或“假”。

●●● 例5 选择查询讲师记录中女性讲师的数据。

```
SELECT * FROM teachers WHERE sex = 0 ;          /*0为女性，1为男性*/
```

（3）排序

查询后返回的记录集很可能与预期结果不匹配，如未按指定要求排列，此时可以使用ORDER BY子句，指定一个或多个字段的升序或降序完成排序，默认是升序排序。

●●● 例6　查询所有课程的编号、课程名称和授课讲师编号并按授课讲师编号升序返回查询结果。

```
SELECT id,name,teacherid FROM courses ORDER BY teacherid;
```

（4）统计

在使用查询语言时经常要做一些简单的统计计算，如统计数据集的元组数、对关系中依据某属性的值分类统计求和或求平均值。关系数据语言中提供了COUNT、SUM、MAX、MIN、AVG等聚合函数来完成这些数据统计操作，聚合函数经常配合GROUP BY子句使用。

●●● 例7　按授课讲师编号统计每名讲师授课数量。

```
SELECT teacherid,count(*) FROM courses GROUP BY teacherid;
```

总而言之，投影是对列的操作，选择是对行的操作，排序是对集合的操作，统计是在集合分组后对列的操作。

除了聚合函数，SQL还有数学函数（如表3.1.3）、字符串函数（如表3.1.4）和日期函数（如表3.1.5）等。

表3.1.3　数学函数

函数名	作用
ABS()	返回绝对值
MOD(N,M)	返回N被M除后的余数
POWER(X ,Y)	返回X的Y乘方的结果值
ROUND()	返回值接近于最近似的整数

表3.1.4　字符串函数

函数名	作用
LOWER()	将字符串参数值全转换为小写字母
UPPER()	将字符串参数值全转换为大写字母
SUBSTR(str,pos[,len])	从源字符串str中的指定位置pos开始取一个长度为len的字符串
LENGTH()	字符串的存储长度
REPLACE(str, f_str,t_str)	在源字符串str中使用替代字符串t_str替换f_str
LEFT(str, len)	最左边的len长度的子字符串
RIGHT(str, len)	最右边的len长度的子字符串

表3.1.5　日期函数

函数名	作用
NOW()	返回服务器的当前日期和时间
DATE()	返回日期
TIME()	返回时间
YEAR()	返回年
MONTH()	返回月
DAY()	返回天
HOUR()	返回时
DATEDIFF(expr1, expr2)	返回两个日期（expr1−expr2）差的天数

●●● 例8　查询讲师数据表中编号大于等于195的记录，按编号降序返回编号、姓名数据。

```
SELECT id,name
       FROM teachers
       WHERE id >= 195
       ORDER BY id DESC;
        id          name
       198         金泓飞
       197         沈菲华
       196         陈　楠
       195         方仕仪
```

（5）连接查询

涉及两个及以上表的查询，称为连接查询，连接是从两个关系的笛卡儿积中选取属性之间满足一定条件的元组。

●●● 例9　使用笛卡儿积运算查询所有可能的讲师与课程组合数。

```
SELECT COUNT(*) FROM teachers,courses;
```

笛卡儿积的实用性较弱，只有多表连接加上限制条件时，才会有实际意义。连接查询时WHERE子句中的运算符为“=”，称为等值连接，其他运算符称为非等值连接。若在等值连接中把目标列中重复的属性列去掉，则为自然连接。不仅两个表之间可以连接，表与表自身也可以连接，称为表的自身连接。

●●● 例10　查询讲师授课数据。

```
SELECT teachers.name AS 授课讲师, courses.name AS 课程
```

```
FROM teachers, courses
WHERE teachers.id = courses.teacherid;
```

授课讲师	课　程
臧立续	3d Max基础课程
沈晨昱	60分钟搞定西班牙语
查颖欣	90分钟会读韩文字
卢可琪	AE教程超级合辑
陆心杰	App Inventor趣味编程
……	

在查询讲师授课信息时，若有讲师未安排授课任务，仍要在查询结果中呈现（结果会填空值Null），此时就需要用外连接，左外连接列出左边数据表中的所有记录，右外连接列出右边数据表中的所有记录。

●●● 例11　查询所有讲师的授课数据，没有授课的讲师的数据也要返回，此时可以使用左外连接查询。

```
SELECT teachers.id AS 讲师编号, teachers.name AS 讲师姓名, courses.name AS 课程名称
    FROM teachers
    LEFT JOIN courses
    ON teachers.id = courses.teacherid;
```

讲师编号	讲师姓名	课程名称
100	范康潇	游戏设计三部曲3d Max建模
101	张瑞耶	(NULL)
102	程忆宁	(NULL)
103	俞冰凤	影视后期AE特效与PR剪辑超级合集
104	陆诗琦	(NULL)
……		

SELECT查询语句返回的是记录集合，多个SELECT语句的查询结果可以实施集合操作，集合操作主要包括并操作UNION、交操作INTERSECT和差操作EXCEPT，三个集合操作语法相似。注意：参加集合操作的各查询结果具有相同的结构，列数相同，且对应的数据类型必须兼容。

●●● 例12　查询编号大于195的男讲师和编号小于110的女讲师的记录集。

```
SELECT id AS 讲师编号, name AS 姓名, sex AS 性别
    FROM teachers
    WHERE id>195 AND sex=1
UNION
SELECT id, name, sex
    FROM teachers
    WHERE id<110 AND sex=0;
```

讲师编号	姓　名	性　别
196	陈　楠	1
197	沈菲华	1
198	金泓飞	1
102	程忆宁	0
103	俞冰凤	0
104	陆诗琦	0
105	俞商思	0

3. 数据操作

数据库可以使用INSERT语句向指定表中插入记录，DELETE语句删除记录，UPDATE语句更新记录。

●●● 例13　在数据表videos中插入课程“SPSS数据分析应用”的一条记录。

```
INSERT INTO videos (name,courseid,url) VALUES(' SPSS数据分析应用 第一讲',27, 'spss001.mp4');
```

ID字段为数据表的主键，且有自增长属性，即ID字段的值不为空也不重复，新增一条记录时，ID字段的值自动加1。INSERT语句中可以省略ID列的赋值。

●●● 例14　删除学员数据表中编号为20的记录。

```
DELETE FROM users WHERE id = 20;
```

●●● 例15　删除学员数据表中全部记录。

```
DELETE FROM users;
```

●●● 例16　更新学员数据表中id等于3的记录，设置姓名为“快乐是福”。

```
UPDATE users  SET name ='快乐是福' WHERE id = 3;
```

●●● 例17　更新学员数据表中全部记录。

```
UPDATE users  SET name ='快乐是福';
```

例17的结果是学员数据表中所有姓名都被更新为同一个值。需注意的是，无论哪种删除数据操作和更新数据操作均应谨慎对待，特别是多条记录的删除或更新操作，失误后的修复工作远大于操作本身。对于关系数据库而言，INSERT、UPDATE、DELETE操作均有可能破坏数据完整性，操作时应特别关注。

问题与讨论

通过查阅相关书籍与网络搜索，讨论笛卡儿积、等值连接、自身连接、左外连接、右外连接的差异。

实践与体验

股价分析报告

股票交易数据可用于描述股票行情。使用恰当的软件管理股票交易数据，通过分析股价，从中提取有效信息，可为进一步了解股票提供依据。

实践内容：

1. 通过网络搜索下载股票交易数据并将其导入数据库。
2. 完成股票股价的分析报告。

实践步骤：

1. 通过网络搜索下载某股票交易数据，格式可参考图3.1.10。

日期	开盘价	最高价	最低价	收盘价	涨跌额	涨跌幅(%)	成交量(手)	成交金额(万元)	振幅(%)	换手率(%)
09-15	5.86	5.86	5.75	5.77	-0.11	-1.87	2,024,035	117,272	1.87	0.08
09-14	5.92	5.97	5.86	5.88	-0.05	-0.84	1,762,812	104,455	1.86	0.07
09-13	5.91	5.95	5.90	5.93	0.00	0.00	1,218,445	72,351	1.01	0.05
09-12	5.88	5.94	5.83	5.93	0.04	0.68	1,446,317	84,902	1.87	0.05
09-11	5.92	5.98	5.88	5.89	-0.01	-0.17	1,756,429	104,118	1.69	0.07
09-08	5.87	5.94	5.85	5.90	0.02	0.34	1,087,328	64,011	1.53	0.04
09-07	5.90	5.91	5.84	5.88	-0.02	-0.34	1,158,945	67,989	1.19	0.04
09-06	5.93	5.97	5.88	5.90	-0.05	-0.84	1,233,432	72,852	1.51	0.05
09-05	5.87	6.00	5.86	5.95	0.07	1.19	2,339,242	139,324	2.38	0.09
09-04	5.77	5.89	5.76	5.88	0.09	1.55	2,116,664	123,881	2.25	0.08
09-01	5.86	5.92	5.76	5.79	-0.11	-1.86	3,399,657	198,796	2.71	0.13
08-31	5.95	5.98	5.84	5.90	-0.07	-1.17	2,531,553	149,354	2.35	0.09
08-30	6.08	6.11	5.89	5.97	-0.11	-1.81	2,629,251	157,712	3.62	0.10
08-29	6.03	6.08	5.98	6.08	0.03	0.50	1,712,719	103,664	1.65	0.06
08-28	6.14	6.18	6.00	6.05	-0.11	-1.79	3,507,362	213,370	2.92	0.13

图3.1.10　部分股票交易数据

2. 建立数据库及数据表，并导入数据。
3. 使用数据库SQL语言配合Python程序语言完成股价分析。

结果呈现：

撰写一篇股价分析报告，内容可包含均价、成交金额和换手率分析等。

思考与练习

某学校准备开发选修课程管理平台，系统的部分E-R图如下：

（1）1∶1联系　班主任与班级

（2）1∶n联系　班级与学生

（3）m∶n联系　学生与课程

请回答下面的问题：

（1）将上面3个局部E-R图转化为关系模式。

（2）关系模式转换后，在MySQL中创建相应的数据库和数据表。

（3）选择合适的方式将原始数据导入数据库。

（4）分组统计各班男生与女生人数。

（5）查询选修信息技术课程的学生人数。

3.2 半结构化数据管理

数据库应用领域的扩展以及描述事物的数据格式的多样化，对数据管理提出了更高的要求。传统的关系模型对复杂对象的表示有一定的局限性，较难满足社交媒体、基于位置服务等应用的数据类型的处理要求。为此，人们提出并发展了许多新的数据模型，其中半结构化数据模型因其良好的数据结构动态修改特性受到广泛应用。

1. XML 数据管理

可扩展标记语言（XML）以其半结构化数据的格式特征成为网上数据交换的标准。通过 XPath 或 XQuery 语言可像关系数据库的 SQL 一样查询 XML 文件中相应的数据。如 XML 数据“book.xml”描述的书籍信息如下：

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
    <book>
        <title lang="chs">红楼梦</title>
        <price>36.9</price>
    </book>
    <book>
        <title lang="chs">三国演义</title>
        <price>23.4</price>
    </book>
</bookstore>
```

XML 节点树模型中使用 XPath 在 XML 文档中选取节点。节点是通过路径或者序号选取的。“/”是从根节点选取，“//”是选择所有匹配选择条件的节点，“[1]”表示第一个匹配的节点，“..”返回当前节点的上级节点，“@”选取属性。例如，/bookstore/book[1] 是从根节点选取 bookstore 子节点的第一个 book 节点。//book/title 选取所有 book 节点下的 title 节点。//@lang 选取所有属性名为 lang 的属性值。HTML 文件在描述网页数据时与 XML 文件的描述极为相似，因此在数据获取时也可使用 XPath 语言，常用于爬虫程序中。

下面以“book.xml”文件为例，体验使用 XPath 对 XML 文件进行数据查询。

```
from lxml import etree
strx = etree.parse('book.xml')
title = strx.xpath('//bookstore/book[1]/title/text()')
print('第一本书的书名是%s' % title[0])
```

```
lang = strx.xpath('//bookstore/book[1]/title/@lang')
print('第一本书的语言是%s' % lang[0])
price = strx.xpath('//title[text()="三国演义"]/../price/text()')
print('三国演义这本书的价格是%s元' % price[0])
运行结果:
第一本书的书名是红楼梦
第一本书的语言是chs
三国演义这本书的价格是23.4元
```

拓展链接

XML路径语言XPath

XPath即XML路径语言（XML Path Language），是一种用来确定XML文档中某部分位置的语言。XPath能基于XML的树状结构，提供在数据结构树中找寻节点的能力。

纯XML数据库系统采用基于XML节点树模型的方式，能够较自然地支持XML数据的管理，关系数据库可以通过扩展支持XML数据的管理。目前纯XML数据库有eXist，Berkeley DB XML等。XML数据管理以其结构规范、可扩展强等特点，广泛应用于证券交易、医疗行业电子病历与健康档案等领域。

2. 图数据管理

图数据库是一种非关系数据库，它应用图理论存储实体之间的关系信息。相对于关系数据库，图数据库更易于表达事物间丰富的关系，查询速度也更快。图数据库最常应用于社交网络中，描述人与人之间的关系，关系数据库用于存储与处理社交网络数据时效果并不好，查询复杂、响应缓慢，而图数据库的独特设计恰恰弥补了这个缺陷。

在图数据库中移动访问由关系连接的节点的操作，称为遍历。它是图数据库数据检索的一个基本操作，也是图模型中所特有的。遍历的重要概念在于检索本身的局域化，遍历在查询数据时仅仅用到所必需的数据，不需要像关系数据库中连接操作时对所有的数据集实施耗时的连接分组等操作。

图数据库只需要标明节点之间存在的关系即可查询数据。关系数据库中需要创建关联表来记录不同实体间的多对多联系，系统中实体之间联系越多、越复杂，创建的关联表也越多，查询代价也越大。

下面以电影数据为例，学习和体验基于Neo4j的图数据管理。

例1　Neo4j查询电影数据

电影数据格式为JSON，每一部电影的数据包含影片编号、影片名、年份、导演、演员

等信息。数据结构较为一致，关系较简单，但数据量大，较难开展复杂查询。保存为JSON数据格式如下：

```
{
 "id": "200110",
 "name": "女篮5号",
 "year": "1958",
 "directors": [
  "谢晋"
 ],
 "actors": [
  "秦怡",
  ……
  "陆晶荪",
 ]
},
```

Neo4j图数据库非常擅长处理复杂的查询请求。如查询与“谢晋”有关的电影数据，或是更为复杂的查询，如查询与“谢晋”有关的电影以及与这些电影相关的人员曾经导演过的电影。使用浏览器连接Neo4j后，通用Cypher查询语句查询，返回的结果可以是Python能操作的JSON格式数据，方便后续处理，有些还可以是图形化返回并显示。

结合该电影的原始JSON数据，使用Python连接Neo4j可将数据导入，其中关键的Cypher操作语句如下：

```
CREATE (p:person {name:'谢晋', id:'pid_15091'})                    //添加人员
CREATE (m:movie {name:'女篮5号', id:200110, year:1958})            //添加电影
//完成添加人员与电影节点数据后，还要添加节点与节点的关系
match (p),(m)
where p.id='pid_15091' and m.id=200110
create (m)-[:导演]->(p)
```

查询与“谢晋”有关的电影的Cypher语句如下：

```
match(p:person{name:'谢晋'})<-[]-(m:movie) return p,m
```

返回JSON格式的数据部分如下，可供Python等程序语言继续处理。

```
[[["p","m"],2,[{"identity":{"low":21067,"high":0},"labels":["person"],"properties":{"name":"谢晋",
"id":"pid_15091"}},{"identity":{"low":1075,"high":0},"labels":["movie"],"properties":{"name":
"鸦片战争","year":"1997","id":"201076"}}],……]]
```

返回的图形化结果显示如图3.2.1。

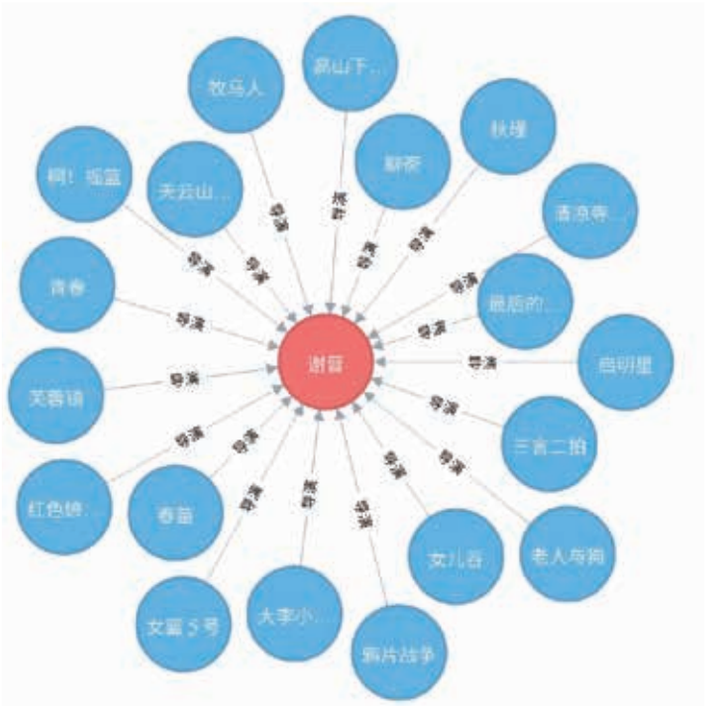


图3.2.1　Neo4j查询与“谢晋”有关的电影

查询与“谢晋”有关的电影工作人员导演过的电影，其Cypher语句如下：

```
match (p:person{name:'谢晋'})<-[]-(m:movie)-[r]->(p2:person)<-[r2:rel{property1:'导演'}]-(m3:movie)
return p,m,r,p2,m3
```

返回的图形化结果显示如图3.2.2。

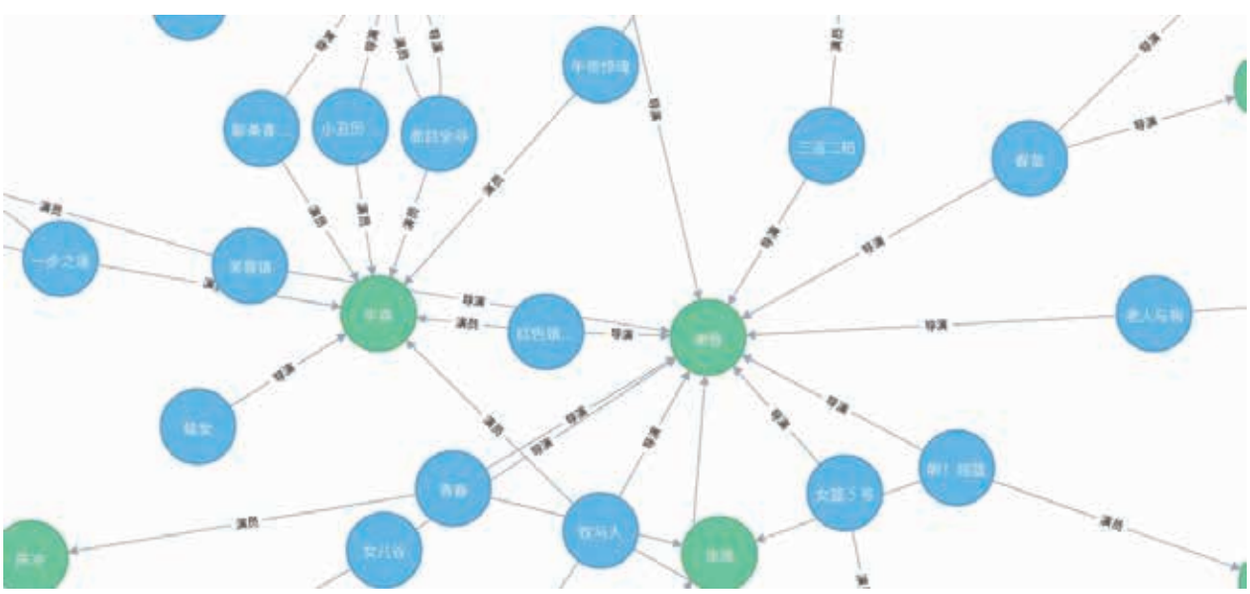


图3.2.2　Neo4j间接查询

拓展链接

Cypher Query Language

Cypher是一种在图数据库中使用的声明性、开源的图查询语言，能查询并返回节点、关系和路径等数据，语法简练，可读性高，功能强大。最初在图数据库Neo4j中使用，后被其他图数据库采用，正成为图数据库的标准化查询语言。

Cypher包含多种子句，最常见的是MATCH和WHERE。MATCH用于描述基于关系的搜索模式的结构，WHERE用于为添加其他约束。CREATE和DELETE用于创建和删除节点与关系。SET和REMOVE用于修改节点上的标签。RETURN返回查询结果。

拓展链接

其他数据管理技术

1. 对象关系数据管理

对象关系数据库是关系数据库与面向对象数据库的结合。对象关系数据库既继承了关系数据库已有的技术，又能支持面向对象模型和对象管理。PostgreSQL是对象关系数据库的典型代表，因为它在可靠性、稳定性、数据一致性等方面有良好的特性，所以在银行金融、通信运营、医疗保险、互联网、汽车生产、政府部门和物流等领域广泛使用。

2. 地图数据管理

数字地图有别于纸质地图，是在一定坐标系统内具有确定的坐标和属性的地面要素的数据集合。数字地图信息丰富多样，在各个行业广泛运用，极大地改善了人们的生活。

车载导航领域已成为21世纪汽车电子工业发展最为迅速的新兴产业，与之相应的各类涉及导航和道路信息数据的应用也在不断发展。为了科学、全面地描述这些信息，实现信息资源共享，制定了以道路交通信息为主的，服务于汽车导航系统的数字地图数据模型和交换格式的标准。地图数据有GDF、KIWI、NavTech等格式，其中GDF应用较为普遍。

3. 内存数据管理

内存数据库已经成为大数据时代数据库系统的一个新方向。内存数据库是指将数据库的全部或大部分数据放在内存中的数据库系统，内存作为常规的数据存储设备，磁盘是数据的永久存储及后备存储。常见的内存数据库有redis、MySQL的MEMORY存储引擎、Microsoft SQL Server Compact等。

内存数据库具有优异的数据存储访问性能、较高的数据访问带宽和数据并行访问能力等特性，广泛应用于金融、电信、电子商务等实时响应性能要求较高的领域。如电信行业的话费实时查询，可取得实时话费累计数据，使后付费和预付费的融合成为可能，实现实时预警、停机、防欺诈等功能。

4. 分布式与并行数据管理

分布式数据库是通用计算机网络将物理上分散的多个数据库单元连接起来组成的一个逻辑统一的数据库。每个被连接的数据库单元称为站点，分布式数据库有统一的数据库管理系统。分布式数据库的基本特点包括物理分布性、逻辑整体性和站点自治性。

并行数据库系统是在并行机器上运行的具有并行处理能力的数据库系统。并行数据库系统能充分发挥多处理和I/O并行性，是数据库技术与并行计算技术相结合的产物。

谷歌公司开发的MapReduce技术，作为面向大数据分析和处理的并行计算模型，发布后便引起了工业界和学术界的广泛关注。Hadoop是MapReduce的一个实例，百度公司用Hadoop处理海量数据，搜索日志分析和网页数据挖掘工作；中国移动研究院基于Hadoop 开发了“大云”（Big Cloud）系统，不但用于相关数据分析，而且还对外提供服务；淘宝的Hadoop系统用于存储并处理电子商务交易的相关数据。

问题与讨论

通过查阅相关书籍与网络搜索，讨论“双11”期间，为响应上百亿的支付请求，电商可以采用哪些数据管理技术。

实践与体验

自定义起点与终点公交站查询途经公交站点数据

公交出行，既能节能减排，又对环境影响小，是一种典型的绿色出行方式。在导航软件或在线地图等软件上，输入行程起点和终点的公交站信息，就可以方便地查询到途经的公交站点数据。

尝试通过图数据库方式管理公交线路及站点数据，并查询途经站点数据，体验关系数据库与图数据库在数据查询上的差异。

实践内容：

1. 了解图数据库中的节点与节点的关系。
2. 能初步使用Cypher查询数据。
3. 体验关系数据库与图数据库在数据查询上的差异。

实践步骤：

1. 获取当地公交线路及站点数据，并导入Neo4j图数据库。

2. 使用Neo4j图数据库的Cypher查询公交线路数据。

例如，尝试以“天安门东”为起点公交站，以“密云北站”为终点公交站，查询途经公交站点数和途经公交站点名称。

```
match pa = shortestPath( (f:busstation {name:'天安门东'})-[*]->(t:busstation {name:'密云北站'}) )
return length(pa) AS len, extract(x IN nodes(pa) | x.name) AS path
| len | path |  ------站点数 | 途经站点
| 14 | ["天安门东", "东单路口西", "永安里路口西", "光华桥南", "亮马桥", "三元桥", "静安庄", "西坝河", "太扬家园", "密云少年宫", "密云西大桥加油站", "密云大剧院", "密云沿湖小区", "密云沙河", "密云北站"] |
| 15 | ["天安门东", "王府井", "东单路口东", "北京站口西", "北京站东", "北京站口北", "东直门", "左家庄", "西坝河", "太扬家园", "密云少年宫", "密云西大桥加油站", "密云大剧院", "密云沿湖小区", "密云沙河", "密云北站"] |
```

3. 使用Python程序语言连接Neo4j，提取查询结果。

结果呈现：

对比百度地图公交线路查询结果与图数据库方式查询结果，体验图数据库的数据管理。

思考与练习

1. 使用XML文件格式描述自己的学校。

2. 使用Python结合XPath获取中国新闻网“即时新闻精选”栏目下8条新闻的标题。

3.3 数据备份与恢复

数据是宝贵的财富，人们对数据价值的认识也在逐步深化，无论因何种原因导致的数据丢失，都可能造成不可挽回的损失，因此数据备份和恢复是数据管理的重要环节。

3.3.1 数据备份与恢复技术

数据丢失随时可能发生，应该及时、合理地备份数据，在需要的前提下实施数据恢复。

1. 数据备份与恢复

造成数据丢失和毁坏的原因可能来自以下几个方面：数据处理或软件平台故障、操作系统或软件漏洞、系统的硬件故障、人为的操作失误、网络攻击或供电故障等。

数据备份就是将数据保留到其他的存储介质，以便在系统遭受破坏或其他特定情况下，重新加以利用恢复的过程。数据备份的重要性不言而喻，计算机里的数据资料，不论是对企业还是对个人，都是至关重要的。因此，对数据应当采取先进、有效的措施，合理地备份，防患于未然。

数据备份和数据恢复是不可分割的。数据备份的目的就是防止发生数据灾难，或当发生灾难时能及时有效地恢复数据。数据恢复是指在自然或人为灾害后，重新启用信息系统硬件设备与软件上的数据，恢复正常运作的过程。所以，信息化程度越高，相应的数据备份和数据恢复措施就越重要。

2. 数据备份策略

日常生活与工作中经常会涉及系统数据与文件数据备份。系统数据备份主要针对计算机系统中的操作系统、设备驱动程序、系统应用软件及常用软件工具等数据，常用的备份软件有Ghost等。文件数据备份是针对具体应用程序和用户产生的数据，将用户的重要数据存储备份，常用的软件有FileGee等。图3.3.1所示的是Ghost与FileGee的软件界面。

图3.3.1 Ghost（上）与FileGee（下）软件界面

备份策略是指确定备份的内容、备份时间和备份方式。根据备份实施的时间与内容，目前备份策略主要有实时备份、定时备份、全备份、增量备份和差异备份，应根据实际情况制订不同的备份策略。

实时备份是连续数据保护。通过实时备份技术对数据自动监控，连续捕获数据变化备份数据，可以使数据丢失的间隔误差达到秒级，只要数据发生变化，便实时、准确地备份数据。它的优点是数据安全性高，缺点是较消耗设备的性能。

定时备份是在固定的时间间隔备份数据。它不能保证数据零丢失。

全备份是备份系统中的所有数据。它的优点是数据恢复时间短且操作方便，缺点是数据量大，可能导致空间消耗大、备份时间长，缺失最后一次备份到恢复期间产生的数据。

增量备份是仅备份上一次备份以后更新的所有数据。它的优点是数据量小、备份时间短，缺点是恢复操作复杂。

差异备份是备份上一次全备份以后更新的所有数据。差异备份策略在避免了全备份和增量备份策略的缺陷的同时，具备它们的所有优点。因无须每次全备份，所以备份所需时间短，并节省磁盘空间，数据恢复也很方便。

3.3.2 关系数据库备份与恢复

关系数据库在运行过程中也需要对数据进行备份与恢复，以应对因各种原因产生的数据损失，确保数据库处于正确状态。关系数据库在数据备份与恢复过程中数据库事务与日志能确保数据库处于正确状态。

1. 数据库事务与日志

事务是用户定义的一个数据库操作序列，这些操作要么全做，要么全不做，是一个不可分割的工作单位。一个事务可以是一条SQL语句、一组SQL语句或整个程序。

事务的开始与结束可以由用户显式控制。如果用户没有显式地定义事务，那么由数据库管理系统按默认规定自动划分事务。在MySQL中，定义事务的语句一般有三条：START TRANSCTION | BEGIN开始一个事务、COMMIT事务确认、ROLLBACK事务回滚。事务可以串行执行，也可以并行执行。

例如，银行转账事务，事务把一笔金额从一个账户甲转给另一个账户乙。

```
START TRANSCTION
读取账户甲的余额到变量BALANCE
BALANCE = BALANCE−AMOUNT;  /* AMOUNT为转账金额 */
IF(BALANCE<0) THEN
{
打印金额不足不能转账
ROLLBACK;
 }
```

```
ELSE
{
读取账户乙的余额BALANCE1
BALANCE1 = BALANCE1+ AMOUNT;
写回BALANCE1;
COMMIT;
}
```

这个例子中的更新操作要么都做，要么全部不做，否则就会使数据库处于不一致状态。事务具有4个特性，简称ACID特性：原子性、一致性、隔离性和持续性。

原子性（Atomicity），由于事务是数据库的逻辑工作单位，事务中包括的诸多操作要么都做，要么都不做。

一致性（Consistency），确保事务执行的结果必须是使数据库从一个一致性状态变到另一个一致性状态。

隔离性（Isolation），指一个事务的执行不能被其他事务干扰，即一个事务的内部操作及使用的数据对其他并发事务是隔离的。

持续性（Durability），也称永久性，指一个事务一旦提交，它对数据库中数据的改变是永久性的。

日志是用来记录事务对数据库的更新操作的文件。不同数据库系统采用的日志文件格式不完全一样。日志文件主要有两种格式：以记录为单位的日志文件和以数据块为单位的日志文件。

对于以记录为单位的日志文件，日志文件中需要登记的内容包括各个事务的开始标记、结束标记以及所有更新操作。对以数据块为单位的日志文件，日志记录的内容包括事务标识和被更新的数据块。

日志文件在数据库恢复中起着非常重要的作用，可以用于多种故障引发的数据恢复。为保证数据库是可恢复的，登记日志文件时必须遵循两条原则：一、登记的次序严格按照并发事务执行的时间次序；二、必须先写日志文件，后写数据库。

2. 数据库的备份与恢复

因为数据库存在各种各样故障的可能性，因此需要相应的数据备份与恢复方案。

（1）数据库中的各种故障

事务内部的故障，有的是可以通过事务程序本身发现的，更多的是非预期的，不能由事务程序处理，数据库因此可能处于不正确状态。

系统故障指的是造成系统停止运转的任何事况，如硬件错误、操作系统故障、系统断电等，使内存中的数据丢失，数据库很可能处于不正确状态。

介质故障指的是外存故障（如磁盘损坏等），它会破坏数据库，发生的可能性较小，但破坏性最大。

计算机病毒是一种人为的破坏计算机功能或者破坏数据的程序代码，既是计算机系统的主要威胁，也是数据库系统的主要威胁。例如，2017年5月12日起勒索病毒在网络上肆虐，被“袭击”的计算机所在的组织和机构包括高校、火车站、加油站和政府办事终端等，多项公共服务受此影响而暂停服务。

（2）数据库的数据备份

数据库备份与恢复涉及两个方面：一、如何备份数据库数据，即建立冗余数据；二、如何利用这些冗余数据恢复数据库。建立冗余数据最常用的技术就是数据转储和登记日志文件，数据转储就是将整个数据库复制保存到磁盘或其他存储介质。通常在一个数据库系统中，这两种方法是一起使用的。

（3）数据库的数据恢复

当数据库发生故障后，可以将保存的备份数据重新装入，从而让数据库恢复到故障发生前的状态。

不同故障引发的数据库恢复操作也是不同的。事务故障的恢复可以利用日志文件撤销此事务已对数据库的修改，恢复操作是自动完成的，对用户是透明的。系统故障的恢复要先撤销未完成的事务，重做已完成的事务，由系统在重新启动时自动完成，不需要用户干预。介质故障的恢复方法是重装数据库，重做已完成的事务。

MySQL可以使用mysqldump命令将数据库数据备份成一个SQL文本文件。具体语法为：shell> mysqldump [options] > dump.sql。例如，导出“zs”数据库到“zs.sql”文件中，“zs.sql”文件包含数据库与数据表的SQL创建语句、表数据等。具体命令如下：

```
mysqldump -h 127.0.0.1 -u root -p --databases zs > zs.sql
```

mysqldump命令的工作原理很简单，先查出需要备份的数据库、表的结构，再在文本文件中生成相应的CREATE语句，将表中的所有记录转换成一条INSERT语句，运行该文件即可完成恢复。具体语法为：

```
mysql -u root -p --databases zhaosheng < zs.sql
```

思考与练习

1. 尝试对“云课堂学习平台”数据库“cclp”进行数据备份与恢复。
2. 通过网络查询，了解MySQL数据库的异地双机热备的工作方式。
3. 阐述数据备份与数据容灾的差异。

巩固与提高

1. 应用数据库管理数据可以使校运会管理工作更加科学、规范、高效。学校使用的运动会管理系统的数据库有以下数据表：比赛项目数据表 match_item(item_id, item_name, item_sex, start_time)，字段分别表示项目编号、项目名称、参赛人员性别和比赛时间；运动员数据表 athlete(athlete_id, stu_id, athlete_name, sex, class_id)，字段分别表示运动员编号、学号、姓名、性别和班级编号；年级数据表 grades(grade_id, grade_name)，字段分别表示年级编号和年级名称；班级数据表 (class_id, class_name, grade_id)，字段分别表示班级编号、班级名称和年级编号；项目成绩数据表 match_score(athlete_id, item_id, result, position)，字段分别表示运动员编号、项目编号、成绩和位次。

请利用学校运动会管理系统的数据库查询数据完成下列问题：

（1）查询各项运动的最佳成绩和相应的运动员信息。

（2）查询下届运动会高三年级 100 米径赛项目最具实力的 3 名男运动员。

（3）统计各班拿到前 3 名的项目数量。

2. 使用 Python 爬取中药处方的数据，使用图数据库查询使用频率最高的药材。

3. 尝试对手机中通讯录、短信记录、通话记录等数据进行备份与恢复。

项目挑战

共享汽车创业计划——数据管理提升服务品质

共享汽车租赁项目已完成了系统的E-R图，经过逻辑结构设计，创建数据库后，相关数据将实时保存至数据库。如何通过数据管理提高共享汽车系统管理效率，从而为更多的客户提供个性、精准、优质、灵活的服务？

项目任务

依据已完成的E-R图选用合适的数据库管理系统软件，创建相应数据库和数据表，对客户信息、车辆情况、动态租赁情况、用户满意度等进行统计分析，以实现动态管理与配置，提高管理效率。

过程与建议

为了顺利开展本项目的研究，建议组建研究小组，在充分理解具体工作要求的基础上，小组分工协作，共同开展本次任务。

1. E-R图转化为关系模式

转化为关系模式时，在遵循转化规则的前提下，结果可能会有多个，合理的数据模式对于数据库的创建与数据管理都有影响。

完成转化后，请填写以下表格。

序号	实体名称	关系模式名	关系模式属性	关系的码
1				
2				
3				

序号	联系名称	关系模式名	关系模式属性	关系的码
1				
2				
3				

2. 相关数据导入数据库，测试数据库

浏览原始数据，比对关系模式要求，调整原始数据，选用恰当的方式将数据导入数据库。

3. 理解现有数据并能提取数据

使用查询语言提取数据后，可以有针对性地发现数据中包含的信息。请做以下（但不限于此）数据查询。

- 投影查询“汽车”数据表的所有信息。
- 选择查询本年度各租赁点租赁车辆的信息。
- 排序查询系统使用以来租赁次数最多的前50名客户信息。
- 统计查询上年度月平均用户满意度。

保存查询结果并与同学讨论，你从中发现了什么？

4. 分析提高管理效率的方式方法

系统投入使用后，数据是实时更新的，作为管理者，需要根据获得的数据做出决策，实现更高效的管理。那么，哪些数据可以为管理者的决策提供更具体、更有针对性的依据呢？

管理目的	需要哪些信息	哪些数据可以提供
提高车辆使用率	发现闲置车辆	车辆租赁时间、里程、油耗和租金等

5. 确定查询内容并实施查询

通过上面的分析，对某些数据的分析和呈现将有助于提升管理效率。请至少确定3个查询内容，使用SQL语言查询数据并呈现结果，以此作为提供给管理者的决策依据。

6. 小组成果的展示与交流

总目标：_____

查询内容与相应目的：_____

实现情况：_____

反思与进一步优化的思考：_____

评价标准

请根据项目实施的过程、效果以及成果展示交流的结果，对自己完成项目的情况进行客观的评价，并思考后续完善的方向。把评价结果和完善方案填写在下面的表格中。

评价条目	说明	评分（1～10分）	评分主要依据阐述	后续完善方向
关系模型转化	E-R模型转化的完整度、正确性			
数据导入	数据预处理后的匹配度、导入的完成度			
提高效率关注点	关注点设置的合理性与可行性			
程序实现	数据导入与数据查询时，程序编写的完成度			
展示	项目展示的完整性与展示获得的反馈度			

拓展项目

医院信息系统的建设日新月异，支付宝推出了“未来医院”，微信推出了“智慧医院”，都是通过线上、线下相结合，统一使用数据库管理数据，优化医疗资源配置，提高医院内部管理效率，实现挂号、候诊、缴费等全流程移动就医服务，提升患者就医体验。

目前已经获取某医院的信息系统挂号数据表的备份文件“m.csv”，共有110万多条记录，包含的列有编号、科室名称、挂号日期、收费日期和总金额，列名分别是bh、KSMC、GHRQ、SFRQ、ZJE。部分数据如下：

```
bh,KSMC,GHRQ,SFRQ,ZJE
1,耳鼻咽喉科,2017-04-01 15:37:24, 2017-04-01 15:50:38, 295.46
2,眼科,2017-04-02 08:39:44, 2017-04-02 08:40:28, 68.79
3,肺病科,2017-04-02 08:46:40, 2017-04-02 09:09:23, 32.11
……
```

请使用数据库管理这些数据，并通过查询提取数据，获取相关信息。

（1）创建数据库及数据表进行数据导入。

（2）使用SQL查询年度最忙碌的科室。

（3）使用SQL查询平均就诊时间最长的科室。

（4）自选角度，使用SQL查询数据。

（5）根据各查询结果，使用数据描述此医院的挂号就诊情况，并分析查询结果所反映的问题，尝试提出问题的解决方案。

第四章 4

数据分析

数据是一种重要的资源，如何有效利用工具从数据中提炼信息、发现知识，是数字化社会公民需要具备的一种重要能力。进行数据分析是为了最大化地开发数据资源，发挥数据的作用。通过数据分析，可对数据进行简化和抽象。数据可视化能还原并增强数据固有的信息，可视化也是数据分析的重要手段。在实际应用中，推荐系统能有效解决信息过载问题，实现信息消费者和信息生产者的双赢；利用复杂网络分析工具能有效挖掘隐藏在数据中的共性和联系。

问题与挑战

- 品牌手机都有自己的应用市场，供用户下载APP软件。这些手机应用市场都会列出每个APP软件的名称、安装次数、类别等数据。如果收集应用市场中排名靠前的热门APP软件的数据，并进行相关分析，那么能发现许多有用的信息。你会从哪个角度去分析这些数据，用来表达你的观点或去验证你的假设？

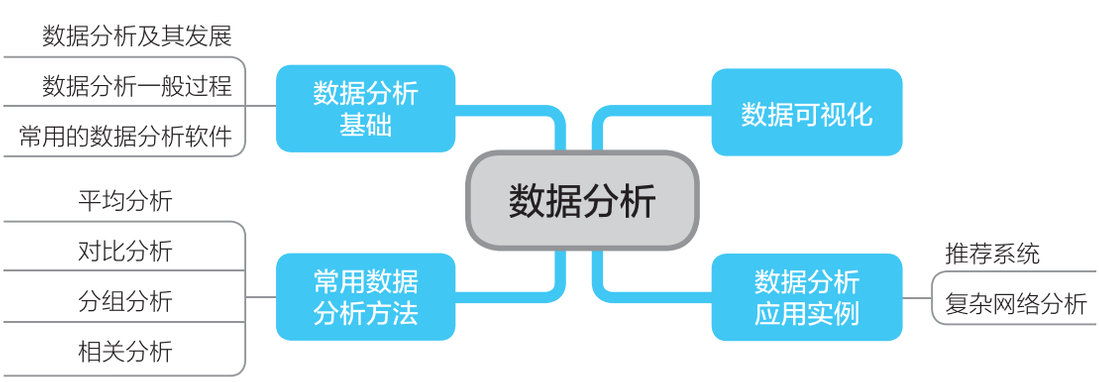
- 互联网上有些企业利用大数据进行电影票房预测，为电影投资方、观众提供参考，为院线排片提供依据。互联网电影资料网站提供了大量的电影信息和用户评论，从知名电影资料网站上获取这些数据，对它们进行数据分析和挖掘，可以得到有价值的信息。那么如何从这些网站上获取所需的数据，又可以运用什么工具和技术进行分析与挖掘？

- 人们经常需要对某些事物进行评价或打分。一些系统能准确预测出用户要打的分值并默认填写好，这样一方面可以提高工作效率，另一方面可以改善用户体验。若某系统给出了一个系统内用户的历史评分数据集，以此作为训练数据，如何实现在原有基础上新增一个评分预测模块？

学习目标

1. 知道常用的数据统计指标、数据分析方法及数据分析工具。
2. 能根据实际需要选用合适的分析方法、工具对数据进行分析并解释。
3. 了解不同数据类型的可视化表现形式，能选择适当的工具对数据进行简单可视化分析并呈现。
4. 了解推荐系统、复杂网络分析方法，知道PageRank算法的原理和用途。

内容总览



4.1 数据分析基础

数据分析的目的是从一大堆数据中提炼出信息，探索数据对象的内在规律，常用于现状分析、原因分析、预测分析等。随着数据量的急剧增长，数据分析的方法、工具和技术也在不断地发展。

4.1.1 数据分析及其发展

数据分析就是对数据进行分析。任何对数据进行计算、处理从而得出一些有意义的结论的过程，都属于数据分析。从数据分析发展的角度，数据分析可以分为统计分析、数据挖掘、大数据分析等。

1. 统计分析

统计分析是指通过统计学方法对数据进行处理，提取有用信息，形成结论的过程。统计分析的作用是将繁杂的数据进行简化和抽象，以便抓住事物的本质和特征。分析者通过分析报告表达观点和立场，为决策提供支持。

进行统计分析时，一般先用描述性统计的方法计算出数据的集中趋势、离中趋势和相关系数等指标，然后在此基础上，以样本信息推断总体情况，并分析和推测总体的特征和规律。常用的统计分析方法有平均分析、对比分析、分组分析、相关分析、回归分析等。

统计分析有明确的目标和思路，先做假设或判断，采用已知模型，通过数据统计、数值计算来验证假设是否成立，从而得到相应的结论。

2. 数据挖掘

数据挖掘是为了改进传统分析方法的不足，针对大规模数据的分析处理而产生的。它基于数据库、统计学、模式识别、机器学习、人工智能、可视化等技术，从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、未知的、潜在有用的信息和知识。

数据挖掘以发现隐藏于数据背后的规律或数据间的关系为目标，常见的任务主要有预测建模、关联分析、聚类分析、异常检测等。数据挖掘从一开始就是面向应用的，随着数据挖掘技术的日趋完善，它广泛应用于科学研究、商业应用、金融投资、互联网应用等领域。

拓展链接

关联分析和聚类分析

关联分析是找出描述数据项之间存在的关联关系的规则，即发现隐藏在数据间的关联或相互关系。关联分析广泛用于购物车分析、找出具有相关功能的基因组等。如通过挖掘顾客购物车中商品之间的关联，分析顾客的购物习惯，从而帮助零售商制订相应的营销策略。

聚类分析是把一组数据按照相似性和差异性分为几个类别，使同一类中的成员彼此相似，而与其他类别的成员不同。聚类分析可以用来对相关的顾客分组、压缩数据、识别模式和处理图像等。如根据文章中出现的词的相似性，对文章进行分组。

问题与讨论

查阅相关资料，讨论归纳出统计分析与数据挖掘的主要区别，并填写下表。

比较条目	统计分析	数据挖掘
数据量	不大	极大
分析对象	数值	
目标	明确	
任务		预测性
功能	验证假设，得到结论	
结果	解释	

3. 大数据分析

大数据分析需要解决的难题是海量数据在多台机器上的存储以及如何对存储在多台机器上的数据进行计算分析。MapReduce是最具代表性的批处理模式，其核心设计思想是将问题分而治之。目前主流的三大分布式计算系统为Hadoop、Spark和Storm。对已有的统计分析和数据挖掘算法加以改进，迁移到这些分布式计算系统上，就能实现大数据分析。

因此，大数据的分析方法是基于常规统计分析或数据挖掘算法，很多分析方法都是对原有算法的改进，将原来单机实现的算法改成多台机器的分布式计算。可视化是大数据分析的重要手段，也是探索和理解大数据最有效的途径。可视化将计算机自动处理和可视分析紧密结合起来，通过对数据结果和数据模型的交互可视化，让人脑介入大数据的分析过程。

大数据分析的一个重要应用是情景感知，它根据收集到的数据对行为进行细致的“猜测”，帮助用户完成日常工作。例如，智能电表每隔五分钟或十分钟收集一次数据，通过这些数据可以预测用户的用电习惯，从而推断出整个电网在未来2～3个月的用电量。

4.1.2 数据分析一般过程

数据分析过程通常可以分为六个步骤：制订方案、收集数据、数据预处理、分析数据、数据可视化和报告撰写。这六个步骤的顺序不是固定的，视实际需要而定。例如，若先有数据，则根据数据特点制订分析目标和分析思路；若数据呈现复杂、非结构化，数据分析和数据可视化同时进行。

（1）制订方案

在接收到数据分析的任务时，首先需要分析目的，理清具体的分析思路，搭建分析框架。搞清数据分析需要从哪几个角度来进行，采用怎样的分析方法最有效，最后制订出具体的数据分析方案。

（2）收集数据

数据分析的核心是数据。收集数据是为数据分析提供直接的素材和依据；全面、准确地收集数据是科学开展数据分析的前提和保障。尽可能获取一手数据，如原始数据。

（3）数据预处理

在获取数据后，需要对数据进行审查、验证、清洗、转换、分组等操作，将数据整理成适合数据分析的样式。比如，根据实际情况对噪声数据进行删除或转化，对缺失数据进行删除或预估，对重复数据进行合并，对错误数据进行修改或删除等。去噪声会引起信息损失，并且不同的去噪方法造成的信息损失各不相同。

（4）分析数据

选择合适的分析方法及工具，对预处理过的数据进行分析，提取有价值的信息，形成有效结论。在这一过程中，可以采用数据统计、数值计算、信息处理等方法，采用已知的模型分析数据，计算与数据匹配的模型参数。

（5）数据可视化

俗话说“一图胜千言”，数据可视化能有效、直观地传递分析人员要表达的观点。可视化也是数据分析的重要手段。数据分析后得到的数据往往是原始数据的简化和抽象，可视化借助人眼快速的视觉感知和人脑的智能认知能力，直接提高对信息认识的效率，起到清晰有效的传达和沟通的作用。可视化也能引导用户分析和推理出有效信息。

（6）报告撰写

在完成数据分析后，需要展示分析结果并形成分析报告。数据分析报告是对数据分析过程的总结和归纳，需要描述出分析的目的和思路、数据的来源、分析的过程、分析的结论和要点。一份好的数据分析报告，需要有一个好的分析框架，层次明晰、图文并茂，能够让阅读者一目了然。数据分析报告必须有明确的结论、建议或解决方案。

问题与讨论

1. 举出生活中运用到数据挖掘的一个例子，并阐明为什么它既不属于统计分析，也不属于大数据分析。

2. 查阅各类数据报告资料，讨论并总结常见的数据报告的形式及其特点。

4.1.3 常用的数据分析软件

数据分析与自然语言处理、数值计算、认知科学、计算机视觉等结合，衍生出不同种类的分析方法和相应的分析软件。如科学计算领域的MATLAB，机器学习领域的Weka，自然语言处理领域的SPSS/Text、SAS Text Miner，计算机视觉领域的OpenCV，图像处理领域的Khoros、IRISExplorer，为教育和研究应用领域提供统计的软件Minitab，以及能完成一些简单的数据分析任务的办公软件Excel等。

（1）SPSS

SPSS是一系列用于统计学分析运算、数据挖掘、预测分析和决策支持任务的软件产品及相关服务的总称。SPSS集数据录入、整理、分析功能于一身，用户可以根据实际需要和计算机的功能选择模块，广泛应用于教育学、心理学、经济学、生物、地理、医学等学科领域。

SPSS基本功能包括数据管理、统计分析、图表分析、输出管理等。统计分析过程包括描述性统计、均值比较、一般线性模型、相关分析、回归分析、对数线性模型、聚类分析、数据简化、生存分析、时间序列分析、多重响应等。SPSS也有专门的绘图系统，可以根据数据绘制各种图形，图4.1.1为SPSS绘制的散点图。

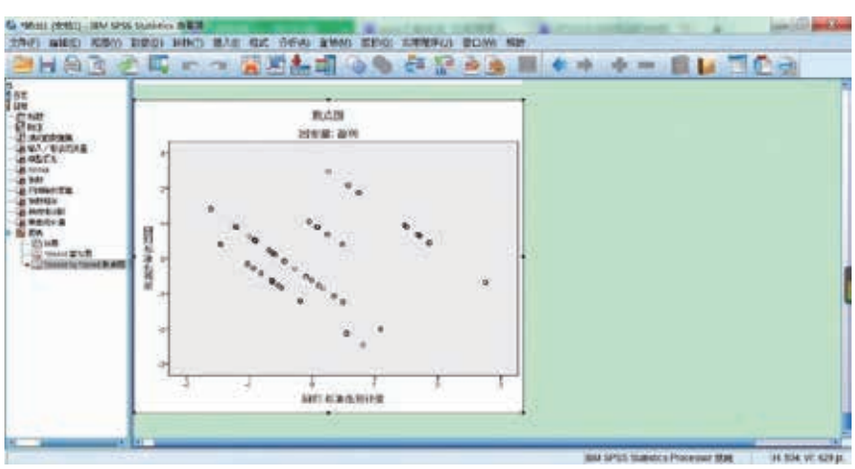


图4.1.1　SPSS应用

（2）SAS

SAS是一款广泛应用于化学、生物、心理学、农业、医学等领域的统计分析软件。它

由数十个专用模块构成，功能包括数据访问、数据储存及管理、应用开发、图形处理、数据分析、报告编制、运筹学方法、计量经济学与预测等。SAS主要完成以数据为中心的四大任务：数据访问、数据管理、数据呈现和数据分析。

（3）MATLAB

MATLAB（矩阵实验室）是一款用于算法开发、数据可视化、数据分析以及数值计算的商业数学软件。利用MATLAB提供的脚本语言，可以创建用户界面、调用其他语言编写的程序。图4.1.2是在MATLAB中用脚本语言绘制的一个三维曲面图。

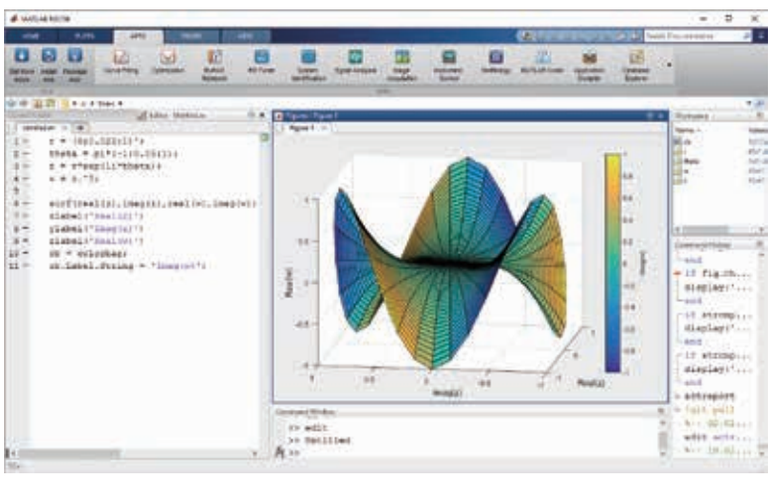


图4.1.2　MATLAB应用

（4）Minitab

Minitab提供了各种统计工具和质量工具，是一款操作较为简单的数据分析软件，应用于工学、社会学等研究领域。它具有强大的图表功能，有简易的可视化操作特点，支持的图表类型有散点图、柏拉图、鱼骨图、直方图、控制图、箱线图等。

（5）Tableau

Tableau是用于可视分析数据的商业智能工具。用户可以创建和分发交互式、可共享的仪表板，实现快速分析、可视化和分享。Tableau容易上手，将数据拖放到数字“画布”上，就能创建各种图表，将数据运算与美观的图表完美集成在一起。Tableau可连接到文件、关系数据源和大数据源以获取和处理数据，允许数据混合和实时协作。

（6）R语言

R语言是一种编程语言与操作环境，主要用于统计分析、绘图、数据挖掘。R语言可以通过安装包（Packages，具有一定功能的类或接口）增强功能。用R语言可以进行分子生物学数据分析和基因图谱分析。

4.2 常用数据分析方法

有了明确的分析目的和思路后，就可以选择合适的数据分析方法，从数据的发展趋势、与其他数据的对比、细分等方面进行分析。常见的数据分析方法有平均分析法、对比分析法、分组分析法、相关分析法等。

4.2.1 平均分析

平均分析法运用计算平均指标的方法反映总体在一定条件下某一数量特征的一般水平。平均指标有算术平均数、加权平均数、几何平均数、移动平均数等。

（1）算术平均数

算术平均数是一种数值平均数，主要用于反映整体数据的一般水平。例如，有一组记录年龄的数据：68，66，84，74，72，78，70，77，83，67，79，76。这组数据的算术平均值为74.5，说明这组年龄数据分布在74.5附近。

算术平均数掩盖了数据间的差异，容易受到极端值的影响。为了消除极端值对平均数的影响，可根据实际情况去掉极端值。例如，在歌手大奖赛对选手计分时，去掉一个最高分和一个最低分，然后计算其算术平均数。

（2）加权平均数

日常生活中经常会碰到各种成分占整体的比例不同，在整体中的重要性不同。为了突出差异和重要性，在计算时就要给予数据不同的权重。加权平均数的计算如下：将各数乘以相应的权重，然后求和得到总计值，再除以总的单位数。

权重大小是依据一定的理论或实践经验确定的。比如，世界大学排名中心每年公布一次全球大学排行榜，总分组成包括教育质量、就业情况、科院质量、研究成果、出版数量、影响力以及论文引用等方面，各组成在计算时的权重比例如图4.2.1所示。

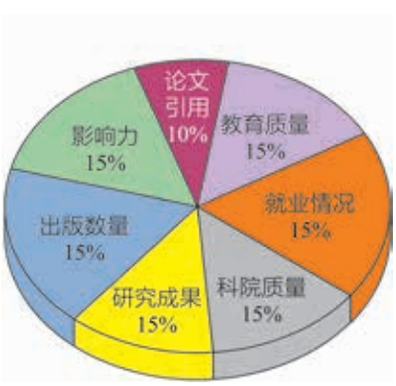


图4.2.1　总分各组成的权重

图4.2.2　“计算器”窗口中设置计算表达式

在数据分析软件中，可以通过计算表达式的方法计算加权平均数。例如，在Minitab中世界大学排名“总分”列的计算表达式设置如图4.2.2所示。

利用平均指标对比某些现象在不同历史时期的变化，能说明其发展趋势和规律；利用平均指标对比同类现象在不同地区、不同行业、不同类型单位等之间的差异程度，比用总量指标对比更具有说服力。

平均数只能代表总体的一般水平，掩盖了个体间的差异。因此，在使用平均指标分析问题时，必须与离中趋势指标相结合，才能对整体做出更全面、更深刻的描述。一般来说，离中趋势指标越小，平均指标的代表性越大。平均分析法也往往要结合各种分组进行对比分析，提示整体内部结构、现象间的依存关系。比如，分析不同地区的平均从业人数、不同行业的平均营业收入等。

拓展链接

集中趋势和离中趋势

一组数据向某一中心值靠拢的倾向称为集中趋势，如图4.2.3所示。一组数据集中趋势的代表值或中心值，能概括反映整体的某些特征和一般水平，利用它们可以对几组数据进行整体概括和比较，如图4.2.4所示。集中趋势度量指标有算术平均数、加权平均数、中位数、众数等。

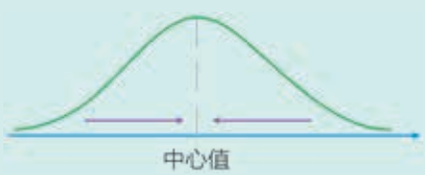


图4.2.3　集中趋势示意图

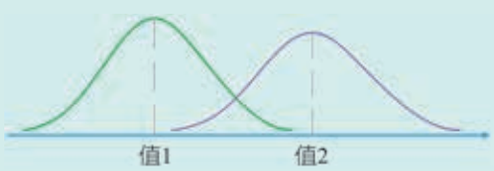


图4.2.4　中心值比较

离中趋势是指一组数据中各数据值以不同程度的距离偏离中心值的倾向，反映集中趋势指标值所概括数值的代表性大小，如图4.2.5所示。离中趋势指标主要有四分位距、方差、标准差等。

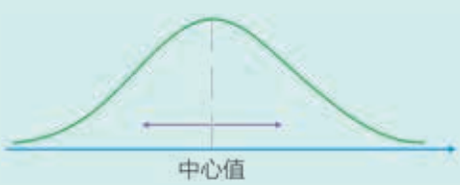


图4.2.5　离中趋势示意图

4.2.2　对比分析

任何事物都既有共性特征又有个性特征，通过对比，可以分辨出事物的个性特征，从而更深刻地认识事物的本质和规律。数据分析也是如此，对庞大、复杂的数据单独做分析，通常很难发现其意义。通过汇总及对比，可以深入发现数据的意义。

对比分析将两个或两个以上的数据进行比较，分析它们的差异，从数量上展示和说明研究对象规模的大小、水平的高低、速度的快慢，以及各种关系是否协调，从而揭示

这些数据所代表的事物发展变化情况和规律性。对比分析可分为横向对比和纵向对比。横向对比是在同一时间条件下对不同总体指标的比较。纵向对比是在同一总体条件下对不同时期指标的比较。进行对比分析时，可以单独使用横向对比或纵向对比，也可结合使用；可以灵活使用总量、平均数等指标进行对比；可以选择与预定目标、不同时期、同级对象、行业典型、活动效果、经验或理论标准等进行对比。在实际生活中，经常使用同比、环比指标来反映数据的变化。

（1）同比和环比

同比指与历史同期进行比较得到的数值，该指标主要反映事物发展的相对情况。环比是指与前一个统计期进行比较得到的数值，该指标主要反映事物逐期发展的情况。环比和同比的最大区别在于，环比体现的数据是短期里的连续性，直观地反映每一期相比上期的效果；同比突出每年在这一期的时间里面有什么变化。

例如，图4.2.6是2015年4月和2014年4月某地区商品房成交量的同比柱形图，图4.2.7是某市2017年商品房成交量的环比柱形图，折线显示的是当月相对于上月的增长率。

图4.2.6　成交量同比柱形图

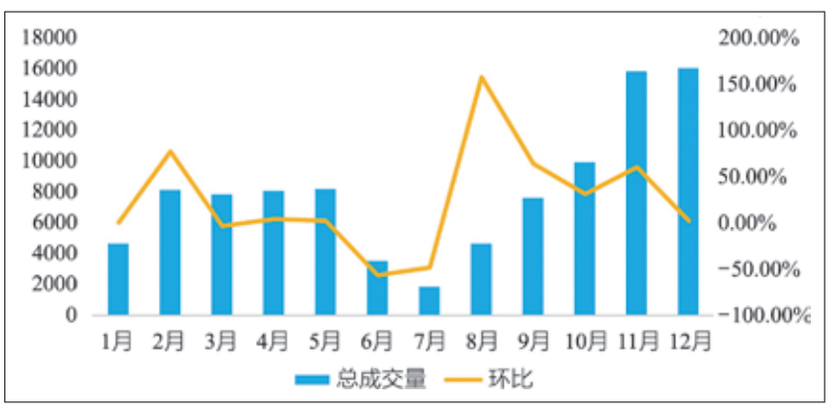


图4.2.7　成交量环比柱形图

（2）数据标准化

对比分析在使用过程中必须遵守可比性原则。指标的内涵和外延要可比，指标的时间范围要可比，指标的计算方法要可比，总体性质要可比等。当数据不能直接进行对比时，就需要对数据进行变换处理，最常用的数据变换方法是数据标准化。

例如，某学生物理成绩为65分，英语成绩为89分，不能说明该学生英语好、物理差。某学生英语前一次成绩为78分，后一次成绩为82分，也不能说明该学生进步了，因为两次考试题目的难度、平均分可能不同。

Z-score标准化是一种比较常用的标准化方法。Z值等于一个数与平均数的差再除以标准差。

$$Z=\frac{x-\bar{x}}{S}$$

其中x为某一具体数，$\bar{x}$为数组的平均数，S为数组的标准差。

经过标准化后，数据具有以下特点：一是符合标准正态分布，平均数为0，标准差为

1；二是不会改变原始分数的分布形状和分布顺序；三是标准分会出现负值，有99.7%的数值在+3与−3之间。

因Z值的范围多在[−3,3]之间，为了方便理解，一般再进行线性变化将Z值转换成T值。若T值采用计算公式T = 70 + 10Z，则转换后的T值的平均数为70、标准差为10。

例如，在Minitab软件中利用计算器和内置函数，根据原始分计算出相应的T分数，如图4.2.8所示。在计算器中T分数的计算公式如图4.2.9所示。

	C1-T	C2	C3	C4	C5
	考号	信息	通用	信息T分数	
1	410100104	37.0	43.5	74.6187	
2	410100129	35.0	40.0	72.2772	
3	410100137	33.5	36.0	70.5210	
4	410100140	27.0	33.5	62.9109	
5	410100232	43.0	40.0	81.6434	
6	410100233	45.0	30.5	83.9850	
7	410100238	37.0	33.0	74.6187	
8	410100325	30.5	39.0	67.0087	
9	410100406	45.0	37.0	83.9850	
10	410100424	38.0	32.5	75.7895	

图4.2.8　工作表

图4.2.9　“计算器”窗口

利用Minitab中的时间序列图，可以观察数据变换前后的变化。在制作时间序列图向导中，选择“简单”时间序列图，序列为“信息”和“信息T分数”，“多图形”窗口中选择“重叠在同一图形”。制作的“信息”和“信息T分数”时间序列图如图4.2.10所示。由图可知：“信息T分数”和“信息”的折线非常相似，“信息T分数”相比“信息”在Y轴方向进行了整体的提升，数据的数值范围也发生了变化。

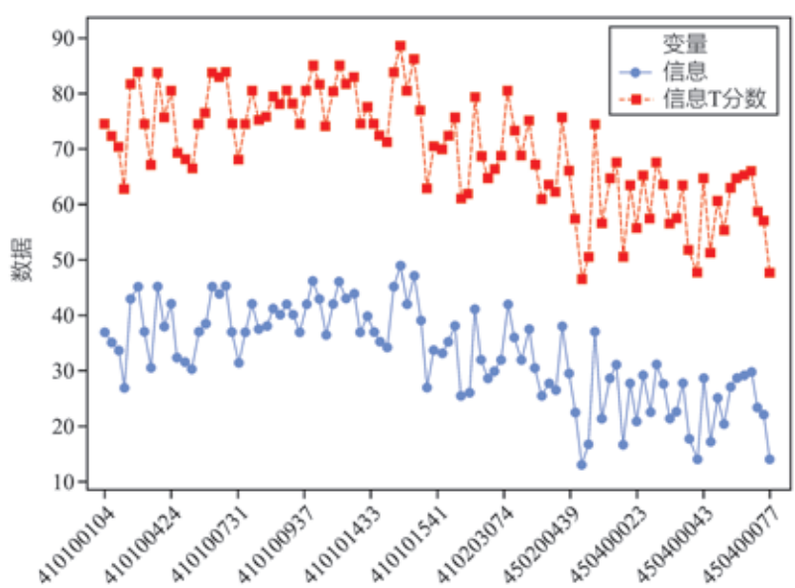


图4.2.10　原始分与T分数时序图

（3）对比分析的应用

利用集中趋势指标直接进行对比。常用的对比指标除了算术平均值以外，还有中位数。例如，比较不同企业间的员工薪水，中位数更具说服力。

比较不同时期集中趋势指标的发展变化。例如，对某市不同时期平均工资进行对比，反映该市职工工资水平的变化趋势。

用样本指标推算总体指标，用推算出的总体指标进行对比分析。例如，以某农作物抽样测量所得的亩产量，推断该农作物的总体亩产量。以推算出的该地区的总产量，进行不同地区或不同时期的比较。

比较离中趋势指标。离中趋势反映集中趋势指标值所概括数值的代表性大小。通过离中趋势指标比较，使得分析更加有说服力、更加严密。

问题与讨论

在进行对比分析时，如何合理地使用同比或环比？

4.2.3 分组分析

平均分析和对比分析可以对总体数量的特征和关系进行分析，要进一步反映数据内部间的关系可进行分组分析。分组分析根据分析对象的特征，按照一定的指标，把分析数据划分为不同的部分和类型来进行研究，以揭示其内在的联系和规律性。

分组时区分不同性质的对象，合并相同性质的对象，保持各组内对象属性的一致性、组与组之间属性的差异性，然后进行对比分析。比如，在分析某地区人口的结构比重时，按年龄大小划分为0～14岁、15～64岁、65岁以上三个组，即少年儿童组、成年组和老年组。

分组分析的步骤：①确定组数。根据数据本身的特点确定分组的数量。组数太少，数据分布就会过于集中，失去分组的意义；组数太多，数据的分布就会过于分散，不便于观察数据的分布特征和规律。②确定各组的组距。组距是一个组的最大值与最小值之差。组中最大值与最小值的平均数称为组中值。③据组距大小，对数据进行分组整理，划分到相应组内。

例如，在Minitab中将某次考试成绩数据按学校分组进行分析，并制作分析图。在“显示描述性统计量”窗口中，变量选择“信息”，按变量分组选择“学校”，如图4.2.11所示。

图4.2.11 选择分组变量

单击“统计量”按钮，在后续操作中，统计量选择“均值”“标准差”“最小值”“最大值”“下四分位数”“中位数”和“上四分位数”，统计结果如图4.2.12所示。图形选择“数据箱线图”，在弹出的箱线图中再添加均值连接，

如图4.2.13所示。可以看出，学校7的均值和中位数都最大，并且中位数大于均值；学校7的中位数线最高，中位数线高于均值点，且均值点也最高；学校8的标准差最小，其在箱线图中的边缘线较短、箱体较矮；学校4不仅均值点较低，而且箱体较高、边缘线较长，说明其学生成绩非常分散，高分段和低分段人数都很多。

描述性统计量: 信息

变量	学校	均值	标准差	最小值	下四分位数	中位数	上四分位数	最大值
信息	学校1	38.491	5.456	25.500	35.000	38.000	43.000	49.000
	学校2	24.771	6.372	11.000	20.500	24.500	29.500	43.000
	学校3	31.33	6.68	18.00	29.00	30.50	36.00	42.00
	学校4	24.872	8.837	8.000	17.625	23.500	32.000	45.000
	学校5	31.68	7.18	15.00	27.75	31.00	37.00	44.00
	学校6	33.44	5.77	26.00	28.88	32.00	39.75	42.00
	学校7	39.36	5.28	30.50	35.25	40.00	44.00	47.00
	学校8	30.88	4.74	25.50	26.75	30.00	36.13	38.00

图4.2.12　不同学校统计结果

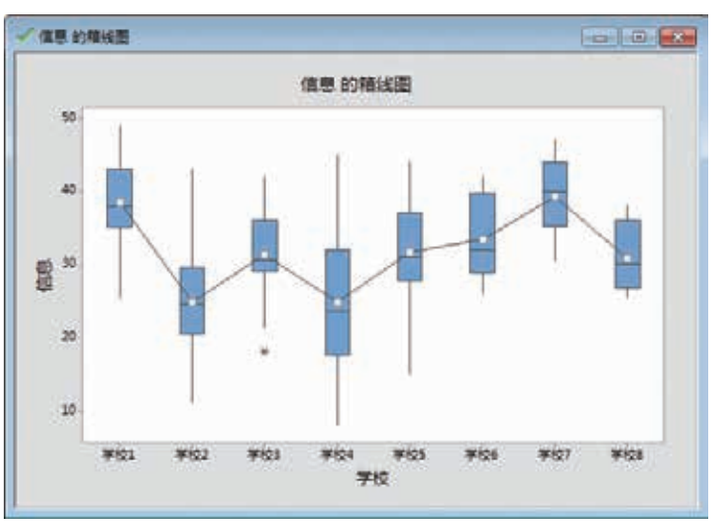


图4.2.13　不同学校箱线图

拓展链接

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，如图4.2.14所示。箱线图主要包含六个数据点：上边缘、上四分位数、中位数、下四分位数、下边缘和异常值。上边缘也称上界，上边缘值 = Q3+1.5(Q3−Q1)，其中Q3为上四分位数，Q1为下四分位数。下边缘也称下界，下边缘值 = Q1−1.5(Q3−Q1)。异常值是指超出上界或下界的数据点，超出3倍四分位数差(Q3−Q1)距离的异常值为极端异常值，用“*”表示；处于1.5～3倍四分位数差距离的异常值为温和异常值，用“·”表示。

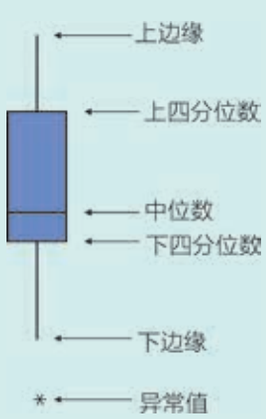


图4.2.14　箱线图

4.2.4 相关分析

有统计显示，游泳死亡人数越高，冰淇淋卖得越多。换句话说，游泳死亡人数和冰淇淋销售量之间存在相关性，但不能由此得出吃冰淇淋就会增加游泳死亡的风险。相关性并不意味着因果性。两个变量A和B具有相关性，其原因是有很多种的，并非只有A→B或者B→A这样的因果关系。如C→A并且C→B，A和B也会表现出明显的相关性，但并不能说A→B或者B→A。

变量间的关系一般分为两类：确定性关系和非确定性关系。

确定性关系，也称为函数关系，是指在两个变量中，一个变量值确定后，另一个变量值也就完全确定了，即确定性关系往往可以表示成一个函数形式。例如，圆面积和半径的关系式$S=\pi r^2$。

非确定性关系，也称为相关关系，是指在两个变量中，当给定一个变量值后，另一个变量值可以在一定范围内变化。例如，子女身高与其父母身高的关系。从遗传学角度看，父母身高比较高，其子女身高一般也比较高。但实际情况并不完全这样，还有其他因素会影响后代子女的身高。

相关分析是研究现象之间是否存在某种依存关系，并对具体有依存关系的现象探讨相关方向及相关程度的方法。关系包括两个数据之间的单一相关关系和多个数据之间的多重相关关系。相关方向有正相关和负相关，两个变量共同变化的紧密程度——相关系数。

典型的相关分析步骤为：第一步，绘制两个变量的散点图；第二步，计算变量之间的相关系数；第三步，相关系数的显著性检验。

1. 散点图的绘制

散点图呈现变量之间是否相关及其相关程度，它是将变量X和Y的观察值（x_i,y_i）（$i=1,2,\cdots,n$）看成平面直角坐标系中的一个点，并在图中标出这n个点。有时分析者很难从大量的数据中发现核心问题，通过可视化的方式，分析者可快速地发现数据间的关系。通过散点图，可以初步判断数组间是否具有相关关系及其相关方向。

2. 相关系数的计算

统计学上经常使用相关系数反映变量之间相关关系的密切程度。常用的相关系数有皮尔逊相关系数、斯皮尔曼相关系数和肯德尔相关系数。

如果散点图中的n个点基本在一条直线附近，但又不完全在一条直线上，就可以使用皮尔逊相关系数表示变量之间相关关系的密切程度。

皮尔逊相关系数的经验解释如下：

①当$r=\pm1$时，各个点完全在一条直线上，这时两个变量是完全线性相关。

②当$r=0$时，两个变量不相关，这时散点图上的n个点可能毫无规律。

③当$r>0$时，两个变量为正相关；当$r<0$时，两个变量为负相关。

④当|r|≥0.8时，两个变量为高度相关；当0.5≤|r|<0.8时，两个变量为中度相关；当0.3≤|r|<0.5时，两个变量为低度相关；当|r|<0.3时，两个变量之间的相关程度极弱，可视为不相关。

例如，某省普通话水平测试试卷构成如下：读单音节字词10分，读双音节词语20分，短文朗读30分，说话40分。该省的普通话水平测试方式为：前三项采用机器打分，第四项采用人工打分（两名测试员独立打分，然后取平均值）。计算机根据大量语料对语音的波形进行因素边界切分，将波形分成音节，然后得出波形对应的真正读音是什么。实验发现机器打分与专家打分很接近，这说明人工智能在这方面已经能够胜任。图4.2.15为某次普通话测试的机器打分和人工打分。

→	C1-T	C2-T	C3-T	C4-T	C5	C6	C7
	编号	考生姓名	出生年份	性别	机器打分	人工打分	最终分
1	201701010001	徐*卫	1973	男	44.7	27.0	71.7
2	201701010002	周*超	1991	男	51.8	28.2	80.0
3	201701010003	钱*鱼	1996	男	49.5	28.5	78.0
4	201701010004	武*林	1979	女	50.6	28.5	79.1
5	201701010005	郭*豪	1992	男	55.5	30.2	85.7
6	201701010006	孟*洁	1992	女	51.6	30.7	82.3
7	201701010007	曹*芸	1991	女	50.1	30.0	80.1
8	201701010008	张*	1992	男	52.0	31.0	83.0

图4.2.15　普通话测试成绩

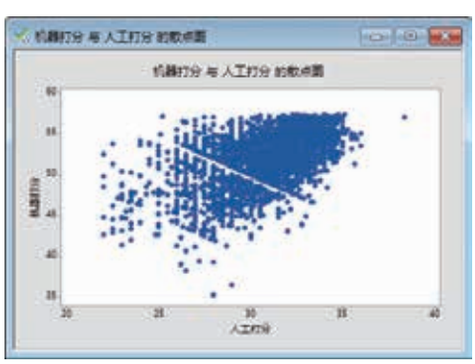


图4.2.16　机器打分和人工打分散点图

图4.2.17　机器打分和人工打分的相关系数

由图4.2.16和图4.2.17可以看出，机器打分和人工打分呈现中度相关的关系。一般来说，一个考生的前三项成绩和第四项成绩应该是高度相关的，机器打分和人工打分也应该是一致的。但是，人在长时工作时容易疲劳且很难保持前后标准的统一，人工打分不可避免出现误差波动；机器不会因为长时工作而疲劳且能保持前后标准统一，但识别技术可能造成准确性偏差；当然一个人在说话和朗读时运用普通话可能也有差别。因此，机器打分和人工打分呈现中度相关是由很多因素造成的。

问题与讨论

高铁速度快，运营密集，发送人数和运营次数多。一列高铁有4万多个零部件，其中任何一个小的零部件出现问题，都是一个大的安全隐患。高铁仅在夜间0：00～4：00停止运行，对各类设备进行检修。如何提前预知哪些零部件存在安全隐患，是需要解决的大问题。请分析高铁列车远程预警系统是如何利用传感器、大数据、相关性分析来实现“提前预知”的。

实践与体验

中国城镇化研究

城镇化是人类社会发展的客观趋势，是国家现代化的重要标志。城镇化水平又称为城镇化率，是衡量城镇化发展程度的数量指标，也是衡量一个国家和一个地区社会经济发展水平的重要标志。中国正在加快推进新型城镇化建设，计划在2020年实现1亿左右农业转移人口和其他常住人口在城镇落户，常住人口城镇化率达到60%，户籍城镇化率达到45%。

实践内容：

1. 运用数据、图表描述中国城镇化的重大意义（城镇化率与人均国内生产总值相关性）、发展现状和发展态势。

2. 从城镇化率和城镇化率变动两个方面，将中国与城镇化率位居世界前列的国家进行对比分析。

3. 绘制出优化后的城镇化空间布局和城镇规模结构分布图。

实践步骤：

1. 通过数字化学习了解城镇化相关知识，从互联网收集中国与世界各国城镇化率的相关数据。

2. 根据收集数据制作中国城市化水平发展折线图、城镇化率与人均国内生产总值散点图、城镇化率位居世界前列的国家的城镇化率变动折线图等。

3. 根据发展趋势制作中国城镇化空间布局和城镇规模结构分布图。

4. 利用图表信息，分析中国城镇化水平发展特点、趋势和挑战，完成“中国城镇化研究”报告。

结果呈现：

以Word文档或PPT演示文稿的形式提交“中国城镇化研究”报告。报告应包含城镇化及其意义、中国城镇化水平发展趋势图和空间布局图等内容。

思考与练习

1. 某购物网站随机抽取若干注册用户，将他们11月份的购物数制作成直方图，如图4.2.18所示。在分析该月人均购物数时，可以选择哪个集中趋势指标的数值作为体现该月购物数量的一般水平？

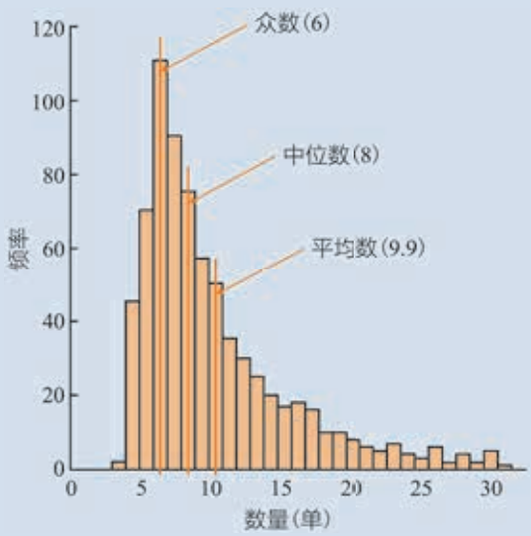


图4.2.18　购物数统计直方图

2. 在进行对比分析时，为了全面地比较每组数据，经常使用集中趋势度量指标进行对比，并计算每组数据的离散趋势指标，说明集中趋势度量指标的代表性。如果在某次分析时已经确定使用标准差作为分析每组数据的离散趋势度量指标，那么相应的集中趋势度量指标应选择什么？

3. 两组数据具有正相关，是否就存在因果关系？举例说明。

4. 大数据下得出的两个变量具有相关性，与小样本下得出的两个变量具有相关性，意义上有什么不同？

5. 为校园十佳歌手大赛活动编写一个Python程序，实现输入8个评委的评分，去掉一个最高分和一个最低分，输出平均得分（选手的最终得分）。同时要对已有成绩按从高到低的顺序进行排名，对每次评委评分也进行排名，计算标准差并反馈给评委。

4.3 数据可视化

数据可视化综合运用计算机图形学、图像处理、人机交互等技术，以图形符号、图像、视频或动画的方式直接呈现数据中蕴含的信息。将数据以可视化方式展现出来，用户可以通过直观交互的方式观察数据、分析数据和探索数据，发现数据中隐藏的特征、关系和模式，有助于决策或预测。因此，可视化既是数据分析结果的重要呈现方式，也是分析和探索数据的有效途径。

问题与讨论

浏览ECharts网站中的可视化实例（如图4.3.1），散点图为各国人均寿命与GDP关系演变动态图，你对这些实例的数据可视化能力有何评价？对不同类型图表特征如何理解？

图4.3.1　ECharts可视化实例

1. 基本图表可视化

基本图表是最早的数据可视化形式之一，至今仍然被广泛地使用。常见的图表类型有柱形图、折线图、饼图、雷达图、散点图、气泡图、箱图等。在运用基本图表来展现数据、传递信息时，关键是选择合适的图表类型。基本图表的使用建议如图4.3.2所示。

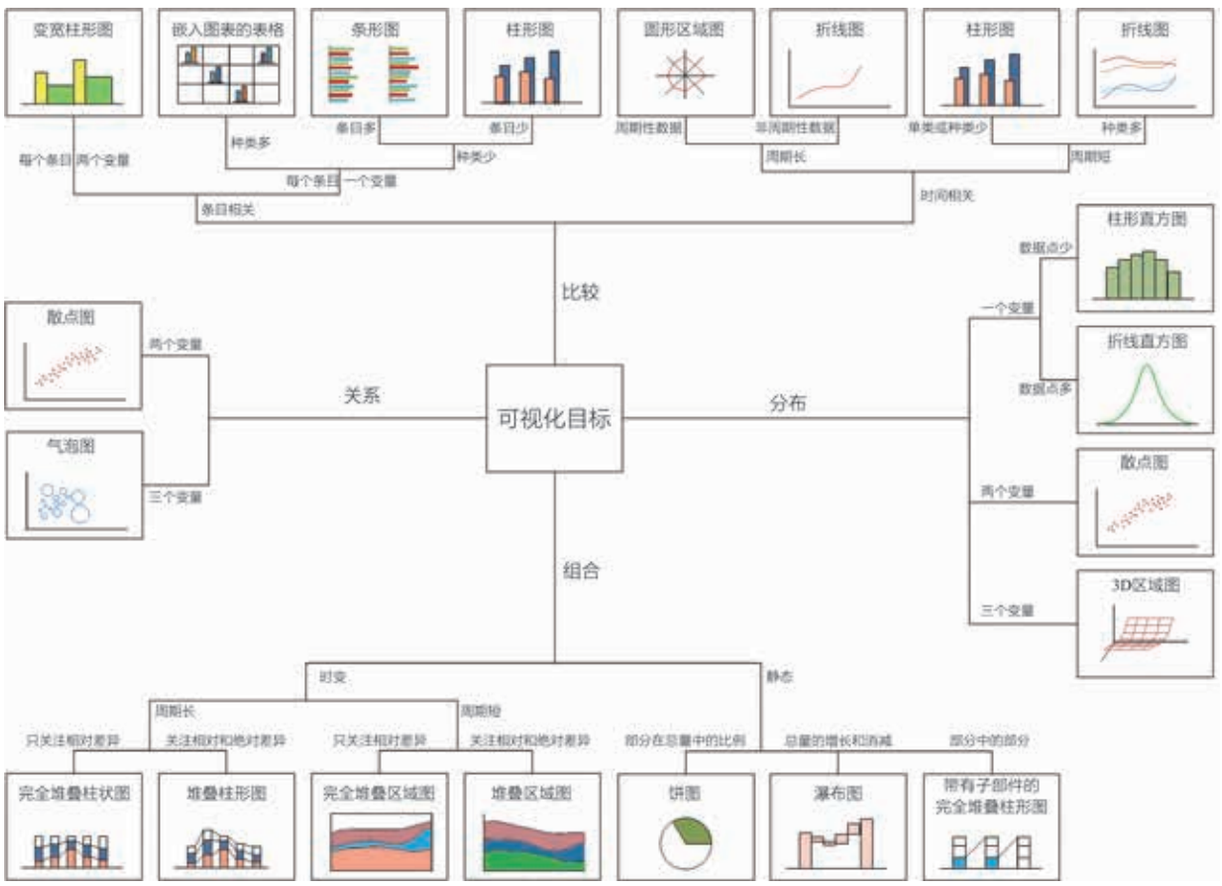


图4.3.2　基本图表使用建议

图4.3.3所示为某地某年度蒸发量、降水量、平均气温可视化图表，其中采用折线图呈现了年平均气温的变化趋势，采用柱形图展现了各月蒸发量、降水量数据，同时也展现了三者年度变化趋势的对比情况。

图4.3.3　蒸发量、降水量、平均气温分析

问题与讨论

结合自己的实践经验，说说如何制作图4.3.3所示的图表。

2. 位置数据可视化

地理位置数据的可视化分为点数据的可视化、线数据的可视化和区域数据的可视化等。点数据的可视化是将位置点（经度和纬度坐标）直接标识在地图上。在地理空间数据中，线数据通常指连接两个或更多地点的线段或者路径。最基本的线数据可视化通常采用绘制线段来连接相应地点的方法，如图4.3.4所示。地理空间中的一个区域是由一系列点所标识的一个二维的封闭空间。可视化区域的目的是为了表现区域的属性，最常用的方法是采用不同颜色表示这些属性的值，如图4.3.5所示。

图4.3.4　路径规划

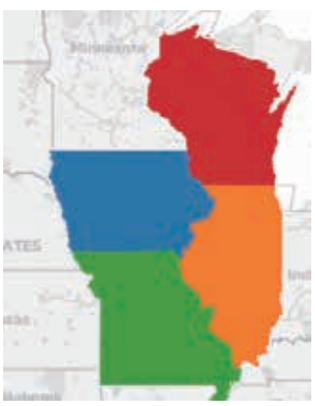


图4.3.5　区域标识

3. 文本数据可视化

文本可视化技术是将文本中复杂的或者难以通过文字表达的内容和规律以视觉符号的形式表达出来，同时向人们提供与视觉信息进行快速交互的功能，使人们能够利用与生俱来的视觉感知的并行化处理能力快速获取大数据中所蕴含的关键信息。

文本可视化的关键是选择合适的视觉编码呈现文本信息的各种特征。例如，词语的频度通常由字体的大小表示，不同的命名实体类别用不同的颜色加以区分。文本可视化中的交互方式有高亮、缩放、动态转换、关联更新、焦点加上下文等。

文本可视化主要有基于文本内容的可视化、基于文本关系的可视化、基于多层面信息的可视化。

基于文本内容的可视化主要关注的是如何快速获取文本内容的重点，主要分为基于词频的可视化和基于词汇分布的可视化。“标签云”是最典型的基于词频的可视化形式，如图4.3.6所示。

图4.3.6　标签云

基于文本关系的可视化研究文本内外关系，帮助人们理解文本内容和发现规律。句法分析是识别句子中的“主谓宾”“定状补”等语法成分，并分析各成分之间的关系。语义依存分析是分析句子各个语言单位之间的语义关联，并将语义关联以依存结构呈现。 不

同的表达方式如果表达同一个语义信息，那么它们的语义依存结构是相似的。如句子“小明在学习大数据分析”，从句法上分析：“学习”是整个句子的核心，“小明”和“学习”是主谓关系，“分析”和“学习”是并列关系，“在”是“学习”的状语，“数据”是“分析”的前置宾语，“大”是“数据”的定语；从语义依存分析：“学习”是整个句子的根节点，“小明”和“学习”是施事关系，“在”是“学习”的时间标记，“分析”和“学习”是顺承关系，“数据”和“学习”是客事关系，“大”是“数据”的具体描述。具体分析如图4.3.7所示，图中英文缩写含义见表4.3.1。

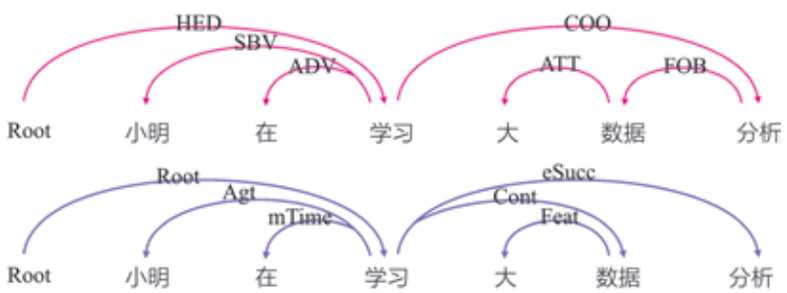


图4.3.7　句法分析和语义依存分析

基于多层面信息的文本可视化，主要研究如何结合信息的多个方面帮助用户更深层次地理解文本数据，发现其内在规律。例如，包含时间信息的文本可视化，微博上热门关键字随时间的动态变化等。

表4.3.1　句法分析和语义依存分析中英文缩写含义

标签	关系类型	完整描述
Root	根节点	Root
HED	核心关系	head
COO	并列关系	coordinate
SBV	主谓关系	subject-verb
ADV	状中结构	adverbial
FOB	前置宾语	fronting-object
ATT	定中关系	attribute
Agt	施事关系	Agent
mTime	时间标记	Time
eSucc	顺承关系	event Successor
Cont	客事关系	Content
Feat	描写角色	Description

4. 层次数据可视化

层次数据着重表达个体之间的包含和从属关系。层次数据的可视化可采用节点—链

接法，将单个个体绘制成一个节点，节点之间的连线表示个体之间的层次关系。布局有正交布局（如图4.3.8）和径向布局（如图4.3.9）。机器学习中的决策树也可表示为层次关系，每一个节点就是一个问题，不同答案对应不同的分支，叶节点对应最后的决策。

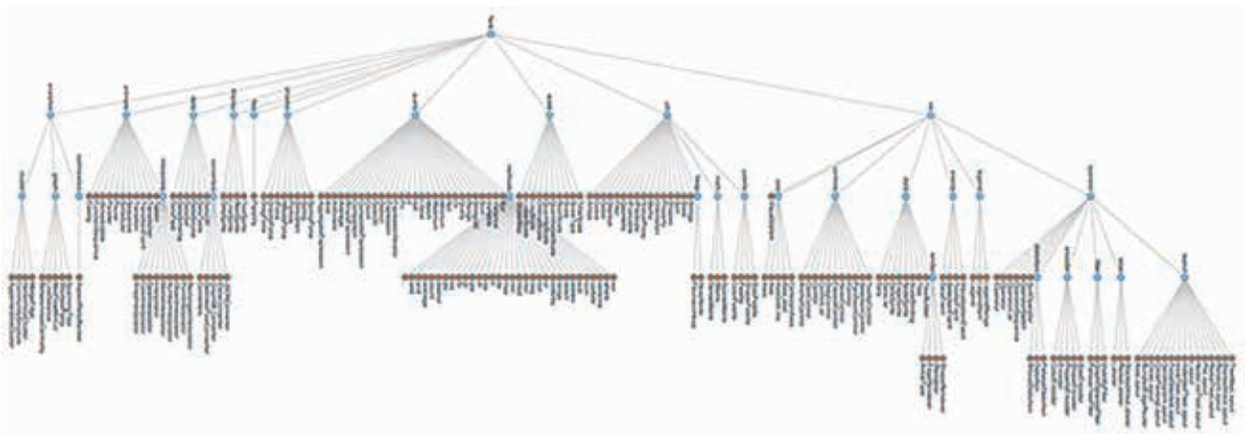


图4.3.8　正交布局

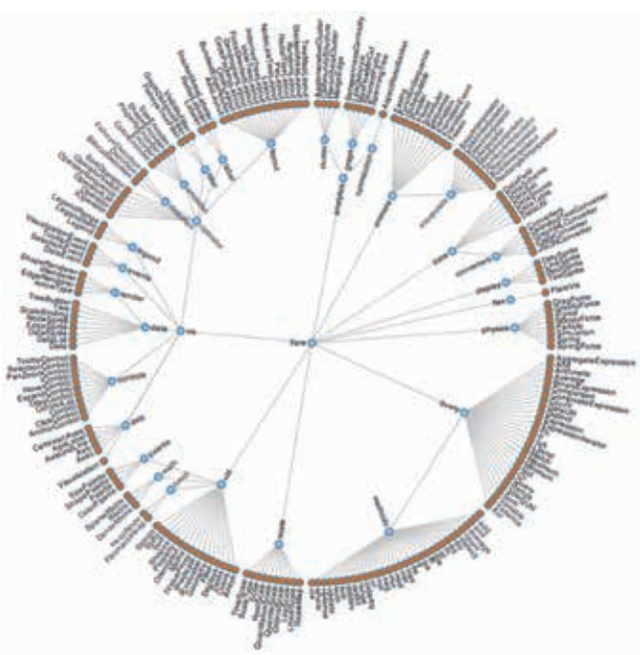


图4.3.9　径向布局

5. 网络数据可视化

网络数据并不具有自底向上或自顶向下的层次结构，表达的关系更加自由和复杂，网络通常用图来表示。线性表和树可以看成是图的简化。网络布局确定图的结构关系，布局可视化最常用的方法是节点—链接法。节点—链接法是最自然的可视化布局表达方法，它用节点表示对象，用线（或边）表示关系。对于具有一定内在层次结构的网络数据，可以采用树型布局算法或扩展为回路图。例如，地铁交通大量采用了正交的布局方式（即网格型布局）。

节点—链接布局方法主要有力引导布局和基于距离的多维尺度分析布局两种方法。力引导布局方法借用弹簧模型模拟布局过程：用弹簧模拟两个点之间的关系，受到弹力的作用后，过近的点会被弹开而过远的点会被拉近；为了减少布局中边的交叉，尽量保持边的长度一致，可以通过不断地迭代使整个布局达到动态平衡，趋于稳定，如图4.3.10所示。节点—链接法的一个变种是弧长链接图，节点沿某个线性轴或环状排列，圆弧表达节点之间的链接关系，如图4.3.11所示。

图4.3.10　力引导效果图

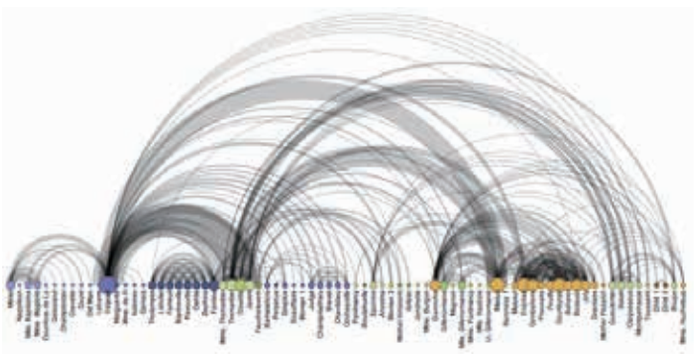


图4.3.11　弧长链接效果图

对复杂网络数据进行可视化，有助于呈现或解释复杂网络数据和模型，从而发现数据中不易发现的模式、特点和关系。网络数据可视化不仅能够挖掘节点（数据实例）属性，而且能展现并揭示节点间的联系。常见的网络数据可视化工具有Pajek、Gephi、NetworkX等。

6. 时序数据可视化

时间是一个非常重要的维度和属性，时序数据是按时间顺序记录的统一指标的数据序列，如个人摄像机采集的视频序列、各种传感器设备获取的监测数据、股市股票交易数据等。在实际应用中，时序数据量大、维数多、类型丰富，为数据分析、挖掘带来了巨大的挑战。采用合适的数据可视化方法展现原始时序数据或分析后的结果，有利于观察和发现与时间相关的变化规律和趋势，有效地揭示时序数据中隐藏的特征模式。

在时序数据可视化的过程中，可以将时间属性当成时间轴变量，其他属性采用不同的可视化通道表达。对时间属性的表达可以采用线性时间、周期时间等方式。线性时间是假定一个出发点并定义从过去到将来的数据元素的线性时域；周期时间则采用循环时间域，一般通过圆环或螺旋的方式布局时间轴。

图4.3.12展示了上海市2018年7月的天气预报，反映了未来1个月内历史均值高温、历史均值低温、预报高温、预报低温的变化趋势，其横轴表示线性时间（时间点和时间间隔），纵轴表示时间域内的气温属性。线性时间的表达方式善于表现数据元素在线性时间域中的变化。

图4.3.12　上海市2018年7月的天气预报

图4.3.13展示了周期下的时序数据的可视化效果。图中采用螺旋的方式布局时间轴，将时间序列沿螺旋排列，一个回路代表一个周期，这种方式很好地体现了时序数据的周期结构及特征。

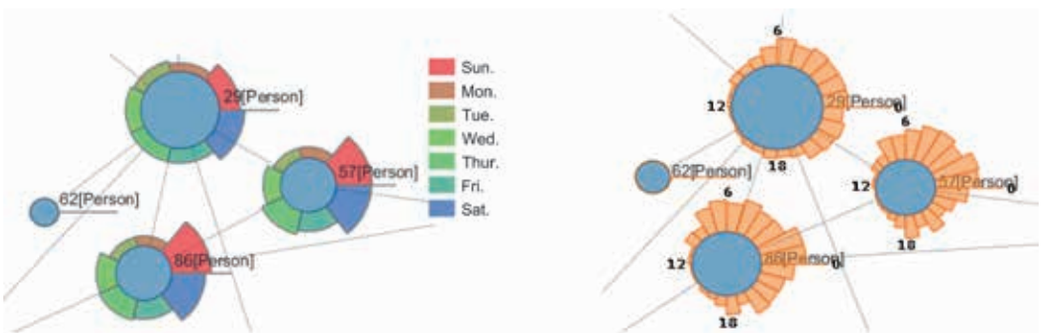


图4.3.13　采用环状表示一周（左）和一天（右）中手机用户活动的时间分布

除了上述可视化方法外，还有流数据可视化、动态网络数据可视化、跨媒体数据可视化、空间向量场数据可视化等。同时，随着信息技术的发展，可视化技术也在不断发展。数据可视化呈现出由低维数据可视化向高维数据可视化、由静态可视化向动态可交互可视化发展的趋势。

实践与体验

在地图上呈现数据

Tableau有自己的地图服务器，当构建地图视图时，具有地理角色的位置名称会自动引用存储在 Tableau Map Service 中的几何图形。当为某个字段（如“省/市/自治区”）分配地理角色时，Tableau将创建一个“纬度（生成）”字段和一个“经度（生成）”字段，并在内置的地理编码数据库中查找，将纬度和经度值分配给该字段中的每个位置。Tableau 根据纬度和经度信息，在地图上绘制图形、布置字符。

实践内容：

1. 打开电子表格数据源，将省市区名称字段转换为地理角色字段。

2. 利用经度、纬度、省/市/自治区字段，在地图上绘制出省市区形状，并将销售量标注到每个区域上。

实践步骤：

1. 在Tableau中将数据源连接到“节日销售成绩.xlsx”文件，并切换到工作表工作区。鼠标右击维度窗格中的“省/市/自治区”字段，在弹出的菜单中选择“地理角色”→“省/市/自治区”。将度量窗格中新生成的“经度（生成）”拖动到列功能区，“纬度（生成）”拖动到行功能区。

2. 将维度窗格中的“省/市/自治区”字段拖放到标记卡中的“标签”项上，将“销售额”拖动到标记卡中的“颜色”项上，单击标记卡中的“颜色”项，编辑颜色，将色板选择为红色。

结果呈现：

在中国地图上以不同颜色标识各省/市/自治区，且销售额越大，颜色越深。同时在各省/市/自治区上标出相应的销售额。

思考与练习

1. 使用手机导航软件寻找你周边的美食，选择一个喜欢的美食地点，规划从当前位置出发到达目标位置的路径。观察、分析这一过程中地理位置数据可视化的方法，并说明位置数据可视化的意义。

2. 选择一个感兴趣的主题，收集有关数据，使用Tableau创建可视化作品。

3. 百度指数是以网民行为数据为基础的数据分享平台。打开百度指数搜索网页，选择感兴趣的关键词，如“数据可视化”“大数据”等，研究这些关键词在一段时期内的搜索趋势、网民需求的变化、搜索人群的特征等。

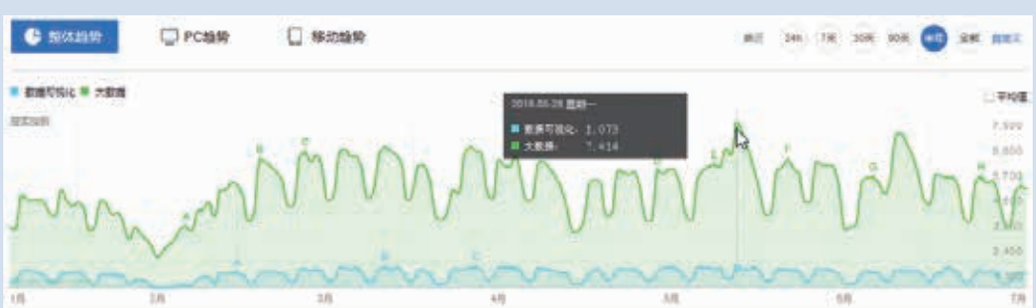


图4.3.14　关键词搜索趋势研究

4.4 数据分析应用实例

在大数据时代，人们面临的主要问题是信息过载和复杂关联的数据网络。数据的共性、网络的整体特征隐藏在数据网络中，需要高效的算法、强大的工具去发现和挖掘这些复杂网络中隐含的信息，提炼和发挥它们的价值。

4.4.1 推荐系统

信息的飞速增长使得人们面临信息过载的困境，如何从海量信息中高效地获得所需的信息变得越来越困难，于是推荐系统应运而生。推荐系统是解决信息过载最有效的方式之一。

1. 推荐系统的工作原理与组成

推荐系统帮助用户在没有明确目的时发现感兴趣的新内容，它依赖于用户的行为数据。在许多网站中都可以看到推荐系统的应用，针对不同用户提供个性化页面、推送个性化广告和新闻等。运用推荐系统较广的领域有电子商务、在线视频、在线音乐、社交网络、基于位置的服务等。

推荐系统将用户模型中的兴趣需求信息和推荐对象模型中的特征信息进行匹配，同时使用相应的推荐算法进行计算筛选，找到用户可能感兴趣的推荐对象，然后推荐给用户。推荐系统通用模型流程如图4.4.1所示。

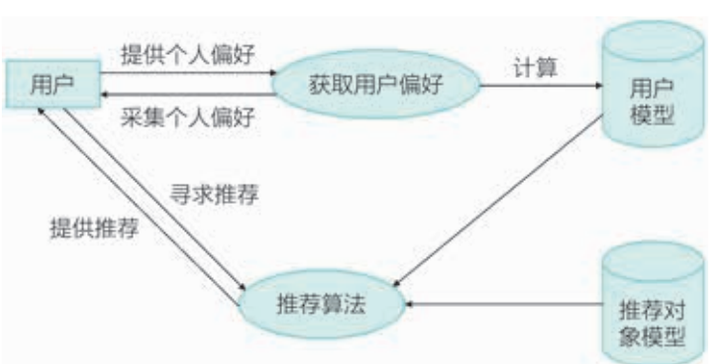


图4.4.1 推荐系统通用模型

推荐系统接收的主要数据源为：①推荐对象数据，如颜色、价格等。②用户的基本数据，如性别、年龄等。③用户偏好数据，如用户对物品的评分、浏览记录、购买行为等。

组成推荐系统的模块有：用户建模模块、推荐对象建模模块、推荐算法模块。用户建模模块是指对用户进行建模，根据用户行为数据和用户属性数据分析用户的兴趣和需求。

推荐对象建模模块是指根据对象数据对推荐对象进行建模。推荐对象描述文件与用户描述文件通常采用相同的表达方法。推荐算法模块是指基于用户特征和物品特征，采用推荐算法计算得到用户可能感兴趣的对象，并根据推荐场景对推荐结果进行一定的调整，将推荐结果最终展示给用户。

2. 常见的推荐策略

推荐系统一方面帮助用户发现对自己有价值的信息，另一方面让信息呈现在对它感兴趣的用户面前，实现信息消费者和信息生产者的双赢。常见的推荐策略有基于内容的推荐和协同过滤推荐。

基于内容的推荐主要应用在信息检索系统中，它根据用户过去喜欢的产品，为用户推荐类似的产品，如在百度的搜索框中出现的候选关键字。

协同过滤算法通过对用户历史行为数据的挖掘，发现用户的偏好，基于不同的偏好对用户进行群组划分并推荐品味相似的商品。

基于用户的协同过滤推荐的基本思想是兴趣相似的用户往往有相同的物品喜好。它的主要工作是先找到和目标用户有相似兴趣的用户群体，然后将这个用户群体喜欢的且目标用户还没有的物品推荐给目标用户。例如，用户甲、丙都喜欢物品A和物品C，如图4.4.2所示。因此，可以认为这两个用户是相似的用户，于是将用户丙喜欢的物品D推荐给用户甲。基于用户的协同过滤算法的关键是计算用户与用户之间的兴趣相似度。

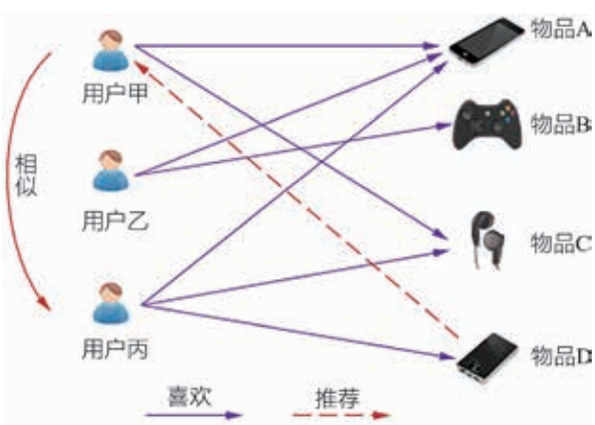


图4.4.2　基于用户的协同过滤

基于物品的协同过滤算法是指给用户推荐和他们之前喜欢的物品相似的物品。不过，该算法并不利用物品的内容属性计算物品之间的相似度，而是通过分析用户的行为记录计算物品之间的相似度。该算法认为，物品A和物品C具有很大的相似度，是因为喜欢物品A的用户大都也喜欢物品C。例如，用户甲、丙都购买了物品A和物品C，因此，可以认为物品A和物品C是相似的或相关的，如图4.4.3所示。而用户乙只购买了物品A，没有购买物品C，所以推荐算法为用户乙推荐物品C。基于物品的协同过滤算法是目前业界应用最多的基础算法。在京东购买商品时，在加入购物车的商品下方会展示一些相关商品。这

些相关商品与购物车中的商品被许多用户共同购买或浏览过。

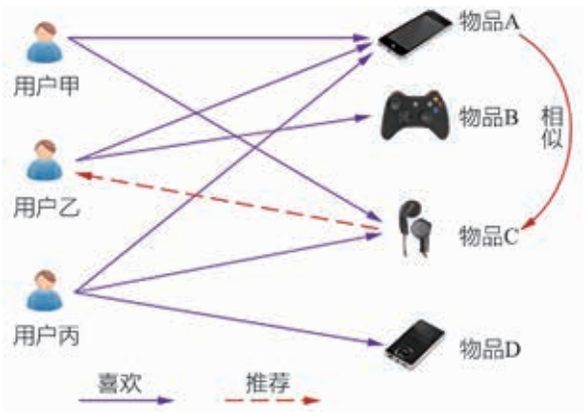


图4.4.3　基于物品的协同过滤

预测准确度是推荐系统领域的重要指标。好的推荐系统能够准确预测用户的行为，帮助用户发现那些他们可能感兴趣，却不容易发现的东西。网站在提供推荐服务时，一般是给用户一个个性化的推荐列表，这种推荐叫作TopN推荐。很多提供推荐服务的网站还都有一个让用户给物品打分的功能，根据用户对物品的历史评分，从中习得用户的兴趣模型，预测将来他看到一个没有评过分的物品时，会给这个物品评多少分。预测用户对物品评分的行为称为评分预测，它一直是推荐系统研究的热点和核心，绝大多数推荐系统的研究都是基于用户评分数据的评分预测。例如，Netflix把推荐系统任务抽象为评分预测问题，并举办Netflix大奖赛——关于机器学习和数据挖掘的比赛，旨在解决电影评分预测问题。

问题与讨论

大数据时代，人们在频繁的互动中创造数据，同时这些数据也在清晰地描绘出人的特征。请从推荐系统对人的影响角度分析：如果数据发生泄露，将会对人们的生活和社会产生什么影响？举例加以说明。

4.4.2　复杂网络分析

自然界中存在的大量的复杂系统都可以通过网络加以描述。各种复杂系统得到的数据，表面看往往是一个个孤立的数据和分散的链接，但如果将反映相互关系的链接整合起来，就会形成一个网络。数据的共性、网络的整体特征隐藏在数据网络中。因此，要理解数据，就要对数据后面的网络进行深入分析。

1. 复杂网络

在生产和生活中，人们为了反映事物之间的关系，常常在纸上用点和线画出各式各样

的示意图。要研究复杂网络，首先需要一种描述网络的工具，这种工具在数学上称为图。例如，图4.4.4描述长江经济带上部分城市间的合作关系。

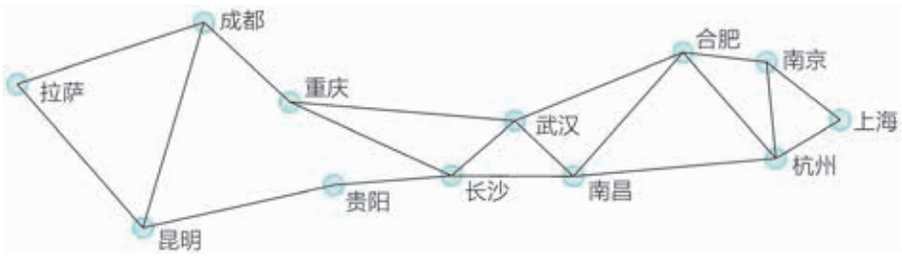


图4.4.4　城市合作示意图

直观来说，复杂网络就是呈现高度复杂性的网络，其节点数目巨大，结构非常复杂。生活中大量的真实网络是复杂网络，常见的有互联网、电力网、朋友关系网、引文网络、交通运输网络、生态网络和神经网络等。在复杂网络中，网络上的节点和连接是不断变化的。如图4.4.5所示为某地流行病毒的传播图。

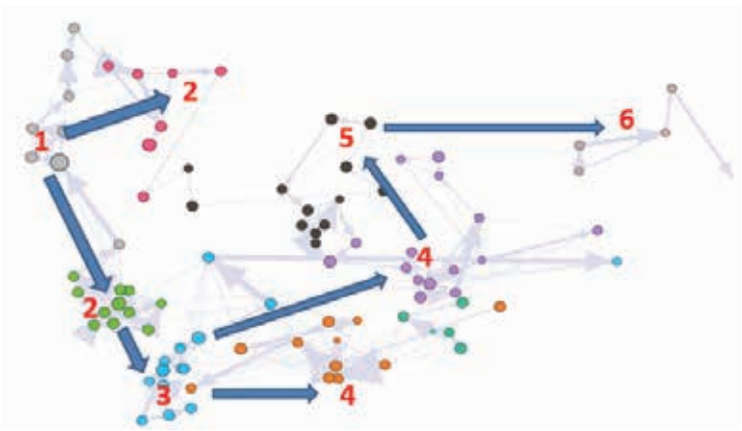


图4.4.5　流行病毒传播图

2. PageRank算法简介

互联网虽然很复杂，但实际就是一张大图，每一个网页是一个节点，网页中的超链接是连接节点的边。对搜索引擎的网络爬虫来说，就是遍历这张大图，从一个网页出发，用图的遍历算法，自动地访问每一个网页并保存它们。对于用户的查询，搜索引擎就是通过排序，把用户最想看到的结果排在前面。

搜索结果的排名取决于两个因素：网页的质量和这个查询与每个网页的相关性。在谷歌出现以前，曾出现过许多通用或专业领域搜索引擎。谷歌最终能击败所有竞争对手，很大程度上是因为它解决了最大的难题：如何对搜索结果按重要性排序。而解决这个问题的算法是PageRank。PageRank算法是由谷歌的创始人拉里·佩奇和谢尔盖·布林于1998年发明的，2001年9月被授予美国专利。它的出现使得谷歌搜索的相关性有了质的飞跃，成功解决了以往网页搜索结果中排序不好的问题。在学术界，这个算法也被公认为是文献检索中最大的贡献之一。PageRank将网页的重要性等级分为1～10级，值越高，说明该网页越受欢迎，也就越重要。

3. PageRank原理与实现

PageRank算法的核心思路是根据网站的外部链接和内部链接的数量和质量来衡量这个网站的价值。如果一个网页被很多网页所链接，说明它受到普遍承认和信赖，那么它的排名就高。对于某个网页来说，该网页PageRank的计算基于以下两个假设：①数量假设。如果一个页面节点接收到的其他网页指向的入链数量越多，那么这个页面越重要。②质量假设。质量高的页面会通过链接向其他页面传递更多的权重，所以越是质量高的页面指向一个页面，该页面越重要。

例如，图4.4.6所示的每个圆代表一个网页，圆的大小表示PageRank值的大小。由于指向网页B和E的链接较多，所以B和E的PageRank值较大。另外，虽然指向C的网页很少，但是最重要的网页B指向了C，所以C的PageRank值比E还要大。

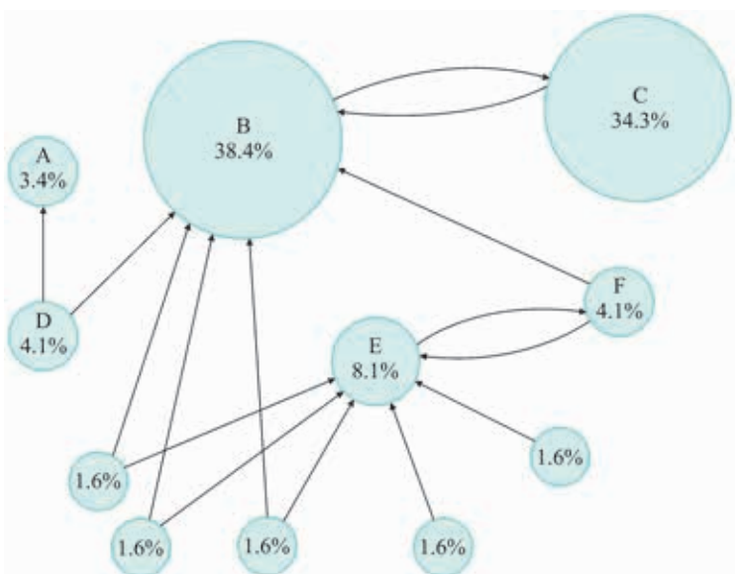


图4.4.6　PageRank示意图

一个网页的PageRank等于指向该网页的其他网页的加权PageRank之和，权重为其他网页的外链数的倒数。那其他网页的PageRank又分别是多少，如何计算？ PageRank算法将这个问题变成一个二维矩阵相乘的问题，并且用迭代的方法加以解决。首先假设刚开始所有网页的排名是相同的，并且根据这个初始值算出各个网页的第一次迭代排名；然后根据第一次迭代排名算出第二次的排名。理论上可以证明这个算法不论初始值如何选取，最终网页排名的计算值都能收敛到固定值。

由于互联网上网页的数量是巨大的，这个二维矩阵从理论上有网页数量的二次方的元素数量。假设有十亿个网页，那么这个矩阵就有一百亿亿个元素，这么大的矩阵相乘，计算量是非常大的，PageRank算法巧妙地运用稀疏矩阵简化了计算量。互联网网页数量的增长使得PageRank计算量越来越大，必须利用许多台服务器才能完成，并且更新一遍所有网页的PageRank周期很长。2003年，谷歌的工程师发明了MapReduce并行计算工具，大大缩短了计算时间，使得网页排名的更新周期缩短了许多。

4. PageRank 应用于社交网络关系分析

社交网络作为一个互联网交友平台与信息传播平台，每天都有海量数据生成。社交网络也是由许多节点构成，每个节点都代表了现实生活中的一个人或者一个组织，节点之间的好友关系即现实社会中的社会关系。根据“六度分隔理论”，一个用户的消息最多只需转发六层，网络上绝大部分的人都可以看到这条消息。PageRank 算法能计算一个社交网络中每一个用户的社会影响力排名。社交网络的信息传播与影响力的研究已经存在于各个领域，主要目的在于挖掘网络用户行为以及商业价值的应用价值等。

拓展链接

Gephi 简介

Gephi 是一款网络分析领域的数据可视化处理软件，用于处理任何能够表示为节点和边的网络数据，能进行复杂网络分析，实现交互可视化与探测。利用 Gephi 将网络数据图形化后，就可以用图的术语进行描述和计算，对图进行网络特性的统计分析，进行不同方式的布局等。

描述一个节点的属性：Id（节点编号）、Label（节点标签）等。保存节点的 CSV 文件，第一行用 Id、Label 等表示该列是节点的哪种属性。

描述一条边的属性：Source（源节点编号）、Target（目标节点编号）、Type（边的类型，Directed 表示有向边，Undirected 表示无向边）、Id（边编号）、Label（边标签）等。保存边的 CSV 文件，第一行用 Source、Target 等表示该列是边的哪种属性。

单击“数据资料”界面的“输入电子表格”按钮，打开“输入电子表格”窗口。利用该窗口可导入节点表格数据和边表格数据。通过“外观”窗口可以设置节点、边以及它们标签的颜色、大小。在“布局”窗口可以选择某种策略对节点和边进行排布，使图形更加合理和易于视觉识别。在“统计”窗口可对网络、节点、边、动态四部分列出的项目进行计算分析，如网络的平均度、网络直径、PageRank 等。在“滤波”窗口中选择适当的过滤工具和参数，可以显示满足条件的节点和边，隐藏其他节点和边。

实践与体验

Excel 计算网页 PageRank

谷歌利用稀疏矩阵和 MapReduce 计算整个互联网上每个网页的 PageRank（以下简称为 PR），并将 PR 按重要性分为 1～10 级。但事实上 PR 就是一个概率，很多软件计算出的 PR 也都是小于 1 的正数。无论采用什么方法来计算 PR，核心仍然是不断的迭代计算，当 PR 趋于稳定时停止计算。图 4.4.7 所示的是一个抽象化的网页链接关系图，通过利用 Excel 迭代计算，体验 PR 计算过程。

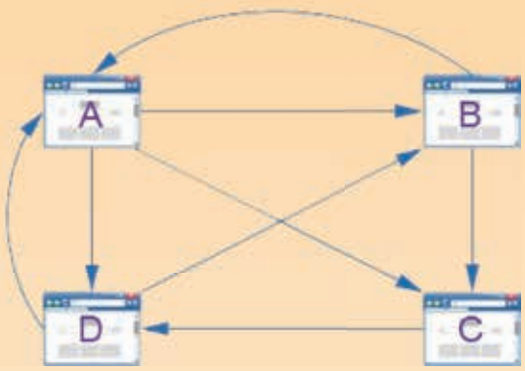


图4.4.7　Web抽象结构图

实践内容：

1. 写出每个网页的PR计算公式。

2. 将每个网页的计算公式用Excel计算公式来表示。

3. 在Excel中将每个网页的PR重复计算50次，观察收敛和稳定趋势。

实践步骤：

PR算法的基本思路：开始时默认所有节点的PR总值为1，且各个节点均分总值；一个节点的PR均分给出边指向的节点，即一个节点的PR等于所有入边连接的节点均分过来的PR之和。

1. 观察图4.4.7，如果用a、b、c、d分别表示A、B、C、D各节点的PR值，写出各节点PR值的计算公式。

2. 打开Excel软件，在新工作表中输入数据，如图4.4.8所示。在B3：E3单元格中分别输入计算公式。

	A	B	C	D	E
1		节点A	节点B	节点C	节点D
2	初始值	0.25	0.25	0.25	0.25
3	第1轮值				
4	第2轮值				
5	第3轮值				
6	第4轮值				
7	第5轮值				
8	第6轮值				
9	第7轮值				
10	第8轮值				
11	第9轮值				
12	第10轮值				

图4.4.8　数据工作表

3. 利用自动填充功能，计算出前50轮各个节点的PR值。

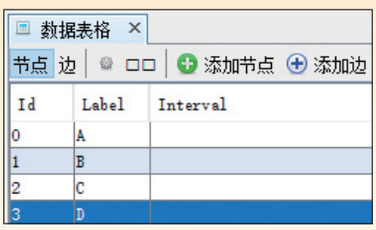
4. 将计算后的表格数据制作成一张带数据标志的折线图，观察数据的变化规律。

5. 观察数据表，看看迭代多少次后计算结果达到稳定，数据不再发生变化。

6. 利用Excel迭代计算PageRank后，讨论以下问题：

（1）对于只有4个节点的网络，计算量已经比较大了，试猜想MapReduce工具在计算谷歌PR时的计算量及其重要性。通过数字化媒体，了解稀疏矩阵和MapReduce相关知识。

（2）将图4.4.8的节点和边信息输入到Gephi数据表格中，如图4.4.9和图4.4.10所示。计算PR值，并与Excel计算结果进行对比。



Id	Label	Interval
0	A	
1	B	
2	C	
3	D	

图4.4.9　节点数据

Source	Target	类型	Id	Weight
0	1	有向的	0	1.0
0	2	有向的	1	1.0
0	3	有向的	2	1.0
1	0	有向的	3	1.0
1	2	有向的	4	1.0
2	3	有向的	5	1.0
3	0	有向的	6	1.0
3	1	有向的	7	1.0

图4.4.10　边数据

（3）进一步了解对于特殊情况的网页链接（互相有指向关系，而没有指向到其他节点的；存在有节点没有入度的），PageRank算法是如何修正的。

结果呈现：

提交体验所用的Excel工作簿文件，工作表中应该包含前50行PR的计算过程、折线图，迭代稳定后不变的数据行设置为浅色。

思考与练习

1. 打开智能手机找一找，哪些系统内置应用、APP软件使用了推荐系统？它们推送的内容是否对你有用？

2. 在学习PageRank排名算法后，如果你是一名网站管理员，会采用什么方法来提升网站的PR排名？

3. 很多手机APP软件在首次使用时，要求用户允许它读写通讯录、短信、位置、相机等权限。从推荐系统使用、隐私安全等角度考虑，如何合理设置这些APP的访问权限？

4. 某系统中包含5个网页，它们的链接关系如图4.4.11所示。请将各个节点的PR计算公式填写在下面的表格中。

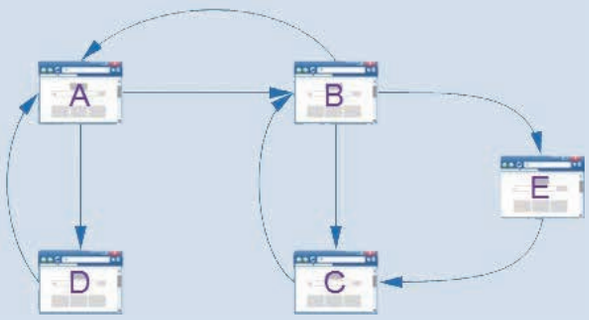
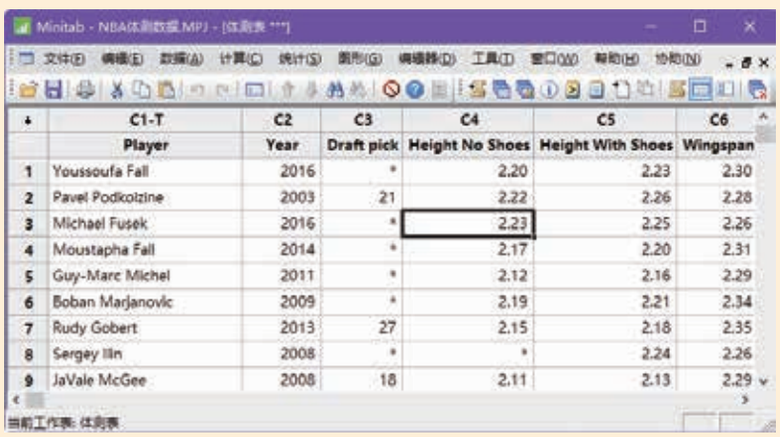


图4.4.11　某系统链接关系

	节点A	节点B	节点C	节点D	节点E
当前值	a	b	c	d	e
PR公式					

巩固与提高

1. 近几年参加篮球选秀球员的体测数据如图4.4.12所示，通过计算球员身高的平均数、中位数、众数，制作身高分布直方图，从整体上了解篮球运动员的身高特征。按年份进行分组，计算不同年份参加篮球选秀球员的身高和体重的平均数和众数，了解球员身高和体重发展趋势。



	C1-T	C2	C3	C4	C5	C6
	Player	Year	Draft pick	Height No Shoes	Height With Shoes	Wingspan
1	Youssoufa Fall	2016	*	2.20	2.23	2.30
2	Pavel Podkolzine	2003	21	2.22	2.26	2.28
3	Michael Fusek	2016	*	2.23	2.25	2.26
4	Moustapha Fall	2014	*	2.17	2.20	2.31
5	Guy-Marc Michel	2011	*	2.12	2.16	2.29
6	Boban Marjanovic	2009	*	2.19	2.21	2.34
7	Rudy Gobert	2013	27	2.15	2.18	2.35
8	Sergey Ilin	2008	*	*	2.24	2.26
9	JaVale McGee	2008	18	2.11	2.13	2.29

图4.4.12　体测数据

2. 地球每天要接受大约5万吨的外太空的物体，但大多数在距地面5～20千米的高空烧毁，很少能坠落到地球表面形成陨石。图4.4.13所示的工作表中保存了大约4.5万块的陨石信息，请将这些陨石按地理位置和质量大小在地图上标注出来，并描述陨石分布的特点。

	A	B	C	D	E	F	G
1	名称	标识	类型	质量(g)	时间	纬度	经度
2	Aachen	1	L5	21	01/01/1880 12:	50.775	6.08333
3	Aarhus	2	H6	720	1951/1/1 0:00	56.18333	10.23333
4	Abee	6	EH4	107000	1952/1/1 0:00	54.21667	-113
5	Acapulco	10	Acapulcoi	1914	1976/1/1 0:00	16.88333	-99.9
6	Achiras	370	L6	780	1902/1/1 0:00	-33.1667	-64.95
7	Adhi Kot	379	EH4	4239	1919/1/1 0:00	32.1	71.8
8	Adzhi-Bog	390	LL3-6	910	1949/1/1 0:00	44.83333	95.16667
9	Agen	392	H5	30000	01/01/1814 12:	44.21667	0.61667
10	Aguada	398	L6	1620	1930/1/1 0:00	-31.6	-65.2333
11	Aguila Bl	417	L	1440	1920/1/1 0:00	-30.8667	-64.55
12	Aioun el	423	Diogenite	1000	1974/1/1 0:00	16.39806	-9.57028
13	Aïr	424	L6	24000	1925/1/1 0:00	19.08333	8.38333
14	Aire-sur-	425	Unknown		01/01/1769 12:	50.66667	2.33333

图4.4.13　部分陨石信息

3. 如果你是一个网站系统中负责推荐系统模块的开发人员，对于一个刚注册的新用户，怎么向他推荐适合的物品或信息？

4. 判断图4.4.14能否一笔画出，并说明理由。

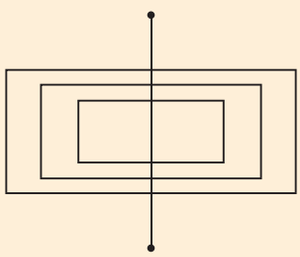


图4.4.14

5. 某网络图如图4.4.15所示，每个节点的度是否相同？这个网络图的度和阶数分别是多少？

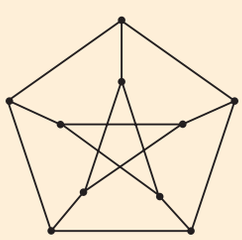


图4.4.15　某网络图

6. 给定一个社交网络数据集，其中包含了200多个联系人及他们之间的联系。问：如何找出这个社交网络中的核心人物？

项目挑战

影评数据分析

以前人们只能通过电影海报获取电影信息，当前可以通过一些专业的电影资料网站获取电影介绍和评论数据，这些专业网站的观众参与度高、数据量大，具有代表性。通过对影评数据的分析可以得到很多信息，如热映电影、高评分电影的特征以及演员间的人脉关系，甚至票房预测和投资方向等。

项目任务

学校影评协会向全校学生征集一份影评数据分析报告，期望这份报告能够尽量多地挖掘出影评数据背后的信息，如我国电影业的发展状况、趋势、电影人之间的关系及影响等。

过程与建议

1. 了解可获得数据的途径

专业的电影资料网站往往包含了多种数据，浏览这些网站，了解它们分别提供了哪些信息，并填写下面的表格。

收集数据名称	豆瓣	IMDb	1905电影网		
片名					
评分					
上映时间					
题材类型					
出品方					
演职员					
同类排名					
评论					

注：网站具有所列收集数据的画“√”，不具有的画“×”。如果要分析更多网站，请在空白列写出网站名称；表格中所列收集数据只是示例，在空白行填写数据名称，尽可能多地拓展收集数据。

2. 预测可以分析出的结果

在了解了当前能够获得的影片数据后，应该对可能的分析结果做深入的研讨和预测，以便有针对性地进行数据收集。

例1：

可能的结果：过去十年间的影片类型及未来发展趋势预测。

依据数据：影片类型、影片上映时间、数量、评分情况等。

例2：

可能的结果：演员之间的关系和影响。

依据数据：演职员表、影片时间等。

其他：

可能的结果：_____

依据数据：_____

根据预测的分析结果，如果发现有些数据无法从当前的影评网站上获得，请考虑是否有其他途径帮助我们达到预测的分析结果。

3. 获取数据

从互联网获取相应的影片数据，可以搜索下载现存的数据集，也可以从专业的影评网站上获取。当从专业的影评网站上获取数据时，可以手工逐条复制，也可以通过编程的方法自动采集。某些网站可能对单位时间内的请求数有限制，请制订应对的策略或方法。

注：如果通过编程的方法自动采集，可以先采集部分数据，观察数据是否符合要求，然后完善程序处理噪声数据，程序试用无误后，再进行全面采集。

4. 数据分析及可视化

针对要预测的结果，利用已采集的数据，采用数据分析软件进行统计分析。反复讨论、观察、调整分析方法和软件，从中发现规律，得出分析结论。

将数据分析的结果通过可视化的方式表达出来，如：

- 通过趋势图发现不同类型影片的发展趋势，进行预测。
- 利用复杂网络分析软件对演职员表进行探索性分析，发现演员间的人脉关系，寻找高票房与演职人员的关系。

5. 数据分析结果的检验

邀请小组之外的其他人（如老师、家长或其他同学等）浏览你们的关键分析结果、数据依据和可视化呈现的效果。请他们从以下几个方面进行评价：

（1）分析结果的价值。

（2）数据依据的说服力。

（3）可视化效果的解释力。

根据评价反馈，修订数据分析结果、数据依据及可视化呈现方式。

6. 撰写数据分析报告

小组合作，完成数据分析报告，报告中必须包括以下内容：

（1）目标：此数据分析报告的背景与意欲达到的目标。

（2）研究过程：描述小组成员为了达到预期目标所采用的方法与步骤。

（3）数据来源：简述支撑数据分析报告的数据来源及采集方法。

（4）关键发现：这一部分是数据分析报告的关键，由若干关键发现组成。每一个关键发现都应该包括发现、数据依据、可视化图片、此分析的特点和优缺点等。

（5）结语：简述此报告的意义、潜在问题和未来进一步探究的方向。

7. 交流展示

制作演示文稿，将数据分析报告中的关键信息进行讲解和呈现，重点介绍数据分析报告中最有价值的关键发现与依据。

评价标准

请根据项目实施的过程、效果以及成果展示交流的结果，对自己完成项目的情况进行客观的评价，并思考后续完善的方向。把评价结果和完善方案填写在下面的表格中。

评价条目	说明	评分（1～10分）	评分主要依据阐述	后续完善方向
收集数据	数据来源、收集方法、数量大小、技术难度			
处理数据	参与程度、格式符合、清洗技术、输出种类			
分析数据	结论科学、角度多样、技术应用、自主探究			
呈现方式	多样准确、报告形式、技术难度、发布方式			
个人能力	技术掌握、学习方式、编程能力、交流能力			

拓展项目

推荐的最终目的是成功引导用户进入推荐的引流。个性化的推荐系统和预测评分都需要对用户进行预测，协同过滤是两者用得最多的算法。协同过滤算法有基于用户的协同过滤和基于物品的协同过滤。现有某网站的评分数据集，包含观众观影后的评分记录，其数据格式如下所示：

“张文”:{“王牌保镖”:7.0,“追击”:7.3,“绣春刀Ⅱ”:7.4,……},
“鲁建”:{“王牌保镖”:7.2,“追击”:6.9,……},
“刘思义”:{“追击”:7.5,“绣春刀Ⅱ”:7.0,……},
“李京一”:{“王牌保镖”:6.9,“绣春刀Ⅱ”:7.4,……},
……

将上述数据集作为实验数据，采用基于物品的协同过滤算法，用Python语言编写一个简易的评分预测程序，实现从键盘上输入某个已经注册用户的姓名和电影名，输出预测这个用户给这部电影的打分。