

01



```
action at the end -add back the deselected mirror modifier object
ob.select=1
ext.scene.objects.active = modifier_ob
selected" + str(modifier_ob)) # modifier ob is the active ob
mirror.select = 0
 bpy.context.selected_objects[0]
a.objects[one.name].select = 1

if("please select exactly two objects, the last one gets the modifier unless its not a mesh")
```

OPERATOR CLASSES

```
types.Operator):
    "X mirror to the selected object"
    ob.mirror_mirror_x"
    "or X"
```

教学参考资料

信息技术

选择性必修 3 数据管理与分析



```
it's supposed to be
batch item is open
sf-sub-indicator
cart-menu .cart-icon-w
r.transparent header#e
menu > 14 .curren
sub-
n,
-1s
end-
ipo
is
2000
2000
```

普通高中
教学参考资料

信息技术

选择性必修 3
数据管理与分析

总主编：李晓明

副总主编：赵健 李锋

本册主编：郑骏

本册副主编：金莹 张洁

编写人员(按姓氏笔画排序)：

王肃 张洁 金莹 高峰 陶烨

责任编辑：曹祖红

美术设计：卢晓红 储平

普通高中 信息技术 选择性必修3 数据管理与分析 教学参考资料

上海市中小学(幼儿园)课程改革委员会组织编写

出版发行 华东师范大学出版社(上海市中山北路3663号)

印 刷 上海四维数字图文有限公司

版 次 2022年8月第1版

印 次 2025年1月第6次

开 本 890毫米×1240毫米 1/16

印 张 11

字 数 258千字

书 号 ISBN 978-7-5760-2953-6

定 价 24.00元

版权所有·未经许可不得采用任何方式擅自复制或使用本产品任何部分·违者必究

如发现内容质量问题,请拨打电话 021-60821714

如发现印、装质量问题,影响阅读,请与华东师范大学出版社联系。电话: 021-60821711

全国物价举报电话: 12315

说 明

《普通高中 信息技术 选择性必修 3 数据管理与分析 教学参考资料》根据教育部颁布的《普通高中信息技术课程标准(2017 年版 2020 年修订)》和高中信息技术教科书的内容及要求编写,与高中信息技术教科书配套,供高中二年级使用。

本书由华东师范大学、上海市信息技术教育教学研究基地(上海高校“立德树人”人文社会科学重点研究基地)主持编写,经上海市中小学教材审查委员会审查准予使用。

编写过程中,上海市中小学(幼儿园)课程改革委员会专家工作委员会、上海市教育委员会教学研究室、上海市课程方案教育教学研究基地、上海市心理教育教学研究基地、上海市基础教育教材建设研究基地等单位给予了大力支持。在此表示感谢!

欢迎广大师生来电来函指出书中的差错和不足,提出宝贵意见。出版社电话:021-60821711。

声明 按照《中华人民共和国著作权法》第二十五条有关规定,我们已尽量寻找著作
权人支付报酬。著作权人如有关于支付报酬事宜可及时与出版社联系。

目 录



第一章 数据管理与分析初步

一、本章学科核心素养的渗透 1

二、本章知识结构 2

三、本章项目活动设计思路 2

四、本章课时安排建议 3

第一节 数据价值 3

一、教学目标与重点 3

二、教学说明与建议 3

三、项目实施与评价 4

四、作业练习与提示 7

五、教学参考资源 7

六、教学参考案例 8

第二节 数据管理与分析技术的重要性 12

一、教学目标与重点 12

二、教学说明与建议 12

三、项目实施与评价 13

四、作业练习与提示 14

五、教学参考资源 15

六、教学参考案例 16

第三节 数据管理与分析方案 20

一、教学目标与重点 20

二、教学说明与建议	20
三、项目实施与评价	20
四、作业练习与提示	26
五、教学参考资源	27
六、教学参考案例	31

第二章 数据管理

一、本章学科核心素养的渗透	39
二、本章知识结构	40
三、本章项目活动设计思路	40
四、本章课时安排建议	41

第一节 数据分类与采集 41

一、教学目标与重点	41
二、教学说明与建议	42
三、项目实施与评价	43
四、作业练习与提示	44
五、教学参考资源	45
六、教学参考案例	46

第二节 数据模型设计 52

一、教学目标与重点	52
二、教学说明与建议	52
三、项目实施与评价	53
四、作业练习与提示	58
五、教学参考资源	59
六、教学参考案例	63

第三节 数据库的实施 67

一、教学目标与重点	67
二、教学说明与建议	68

三、项目实施与评价	68
四、作业练习与提示	72
五、教学参考资源	74
六、教学参考案例	78

第三章 数据安全

一、本章学科核心素养的渗透	85
二、本章知识结构	86
三、本章项目活动设计思路	87
四、本章课时安排建议	87

第一节 数据安全威胁与数据安全策略 88

一、教学目标与重点	88
二、教学说明与建议	88
三、项目实施与评价	88
四、作业练习与提示	90
五、教学参考资源	90
六、教学参考案例	93

第二节 数据备份与还原的实现 98

一、教学目标与重点	98
二、教学说明与建议	98
三、项目实施与评价	98
四、作业练习与提示	100
五、教学参考资源	100
六、教学参考案例	102

第四章 数据分析

一、本章学科核心素养的渗透	106
---------------	-----

二、本章知识结构	107
三、本章项目活动设计思路	107
四、本章课时安排建议	108

第一节 数据准备 108

一、教学目标与重点	108
二、教学说明与建议	109
三、项目实施与评价	109
四、作业练习与提示	112
五、教学参考资源	112
六、教学参考案例	121

第二节 数据分析方法与呈现 125

一、教学目标与重点	125
二、教学说明与建议	125
三、项目实施与评价	126
四、作业练习与提示	128
五、教学参考资源	130
六、教学参考案例	140

第五章 数据挖掘

一、本章学科核心素养的渗透	146
二、本章知识结构	147
三、本章项目活动设计思路	147
四、本章课时安排建议	147

第一节 数据挖掘过程 148

一、教学目标与重点	148
二、教学说明与建议	148
三、项目实施与评价	148
四、作业练习与提示	151

五、教学参考资源	151
六、教学参考案例	153

第二节 大数据时代下的数据管理与分析技术的发展 156

一、教学目标与重点	156
二、教学说明与建议	156
三、项目实施与评价	157
四、作业练习与提示	158
五、教学参考资源	159
六、教学参考案例	161

数据管理与分析初步

一、本章学科核心素养的渗透

1. 信息意识

选用生活中的实际案例,如交通路线规划、企业商品月销售量数据分析等,让学生在项目活动中了解数据的价值,发现不同生活场景中数据所蕴含的价值,了解数据管理与分析技术对获取有价值信息、形成正确决策的作用与意义,认识数据管理与分析技术对人类社会生活的重要影响,以及在促进数字经济发展,推动科技进步,加快数字中国建设进程中发挥的重要作用。在分析数据价值及数据管理与分析技术重要性的项目活动中,培养学生的信息意识。

2. 计算思维

以建立学生社团网站的数据管理与分析方案为项目活动,引导学生发现学习和生活中的数据管理与分析问题,通过项目活动的实施,了解建立数据管理与分析方案的基本过程,能对具体问题进行数据需求分析,设计合理的数据管理与分析方案,并能以合理性、完整性为出发点,对数据管理与分析方案进行分析和评价,选用恰当的策略和方法,对数据管理与分析方案进行优化或改进。在建立数据管理与分析方案的项目活动中,培养学生的计算思维。

3. 数字化学习与创新

培养学生的自主学习意识,为学生提供数字化学习资源,如网络学习平台、微课、数字化实验平台、学习评价系统等,使学生可以选择符合自身特点的方式进行学习,以满足不同层次学生的学习需求。培养学生能根据需要,主动选用合适的数字化平台、数字化学习工具和学习资源,开展自主或合作学习,提高学习质量。

引导学生进行自主学习、合作学习,创造性地解决实际问题。例如,在建立学生社团网站的数据管理与分析方案的项目活动中,鼓励学生通过讨论,发现学生社团网站中存在的其他数据需求,并根据发现的需求尝试建立适当的数据管理与分析方案。培养学生能合作解决学习中的实际问题,创新问题决策,反思与完善学习成果。

4. 信息社会责任

在发展数字经济的前提下,同时要捍卫数据主权,维护数据安全。数据安全是国家安全的重要组成部分,每一个公民都必须树立数据安全意识。在了解数据价值的过程中,指导学生认识到保护数据隐私的重要性。虽然数据蕴含了巨大的价值,但是人们需要树立维护数据隐私的意识,合法地使用数据。在建立数据管理与分析方案的过程中,指导学生识别数据来源的可靠性,在保证数据安全可靠的前提下采集和管理数据,培养学生的信息社会责任。

二、本章知识结构

本章遵循普通高中信息技术课程标准,依据学分和课时规定,紧扣学科概念体系,将内容分为三节,以“身边的数据价值以及数据管理与分析”为项目主题,围绕数据价值、数据管理与分析技术的重要性、建立数据管理与分析方案的基本过程展开设计。

第一节“数据价值”,从生活实例出发,使学生发现数据所蕴含的价值,了解数据价值的体现。

第二节“数据管理与分析技术的重要性”,从生活实例出发,使学生了解数据管理与分析技术的重要性。

第三节“数据管理与分析方案”,从如何建立学生社团网站的数据管理与分析方案出发,使学生了解建立数据管理与分析方案的基本过程,以及如何对方案进行分析、评价和优化。

三、本章项目活动设计思路

本章设计了三个项目任务,通过这三个项目任务,学生可以认识到数据是一种重要的资源,数据价值体现在哪些方面以及数据管理与分析的重要性,能根据实际问题进行业务需求分析,建立数据管理与分析方案,并对方案进行评价和优化。

项目任务 1:交通路线规划中的数据价值。

交通路线查询是人们生活中非常熟悉的一个应用场景,大多数学生都使用过交通路线查询软件或手机应用程序进行路线的查询和导航。本项目任务设计了一种交通路线规划方法,通过本项目任务,让学生认识到交通数据为人们的出行提供的便利服务,从而感受交通数据的价值。

项目任务 2:企业商品月销售量数据分析。

商品月销售量数据在企业生产管理系统中是非常重要的数据。一方面,通过查看历史月销售量,可以了解商品的历史销售情况,这是数据本身所蕴含的价值;另一方面,通过对历史月销售量进行数据分析(如利用简单移动平均法),可以预测未来月销售量,为企业更高效地组织生产或制定合适的营销策略提供决策支持,这是隐藏在数据中的信息,这些隐藏在数据中的信息需要应用数据管理与分析技术来挖掘,使数据体现出更多的价值。通过本项目任务,让学生理解数据管理与分析技术对获取有价值信息、形成正确决策的作用与意义。

项目任务 3:学生社团网站数据管理与分析方案制定。

学生社团网站是一个信息管理系统,也是贴近学生学习生活的一个应用场景。在本项目任务中,需要解决两个主要问题:(1)找到网站一周内发布的哪些文章的浏览量最高、评论量最高、转发量最高;(2)网站如何向用户进行文章的个性化推荐,方便学生更快地找到感兴趣的内容。学生以这两个问题为出发点,了解建立数据管理与分析方案的基本过程包括数据需求分析、数据管理、数据分析、方案评价和优化以及科学决策,对这两个问题进行数据需求分析,制定数据管理与分析方案,并对方案进行评价和优化。

四、本章课时安排建议

本章教学建议用4课时完成,具体参见表1.1。

表1.1 课时安排建议表

节名	建议课时
第一节 数据价值	1课时
第二节 数据管理与分析技术的重要性	1课时
第三节 数据管理与分析方案	2课时

第一节 数据价值

一、教学目标与重点

教学目标:

- 结合生活实际,认识到数据是一种重要的资源,通过科学管理与分析数据,可以使数据实现其应有价值。

教学重点:

- 结合具体实例,了解数据体现出的价值,能够发现生活中数据所蕴含的价值。

二、教学说明与建议

建议教师根据教科书中的项目活动先以学生为主,围绕交通路线规划,组织学生进行充分讨论,从而引导学生了解数据体现出的价值。然后,教师再讲解教科书上的知识内

容,在项目体验中总结知识要点,并结合学生身边真实场景适当补充更多案例,如图书借阅、在线打车、在线购物、在线社交网络等,引导学生发现生活中数据所蕴含的价值,从而加深学生对知识的记忆、认知和理解。

三、项目实施与评价

1. 核心概念精解

本节的核心概念是——数据价值。数据蕴含着巨大的价值,合理地使用数据是非常重要的。数据价值体现在信息社会的方方面面,主要有以下三个方面。

其一,在社会民生方面,数据可以为人们的生产生活提供服务和便利。例如:气象数据可以用于预测天气,为人们安排出行和生产生活提供方便;交通数据可以用于实时线路导航,为人们出行提供便利。

其二,在经济发展方面,数据可以帮助企业进行创新和决策以提高经济效益。例如:企业利用客户数据和销售数据可以对不同的客户群体进行有针对性的营销;企业可以利用库存数据和销售数据来安排企业未来的生产。

其三,在政府决策方面,数据可以为政府的科学决策提供支持。例如:公共卫生部门可以利用覆盖区域的居民健康档案数据和电子病历数据,快速检测传染病,进行全面的疫情监测;公交车管理部门可以利用公交车辆数据、地图数据和人们乘车数据,进行公交线路规划以及公交车班次调整等。

2. 项目活动的具体实施

步骤 1 对图 1.1 中的所有地点进行编号。在交通路线规划的数据管理中,地点信息一般是非常重要的数据,需要记录在数据库表中,但是为了区别地点数据,使其具有唯一性,一般不用地点名称来标识,而是采用每个地点定义一个唯一的编号来标识,比如字母 A、B、C、D……或者 001、002、003……。在本项目活动实施中,为每个地点定义一个字母作为编号,如表 1.2 所示。

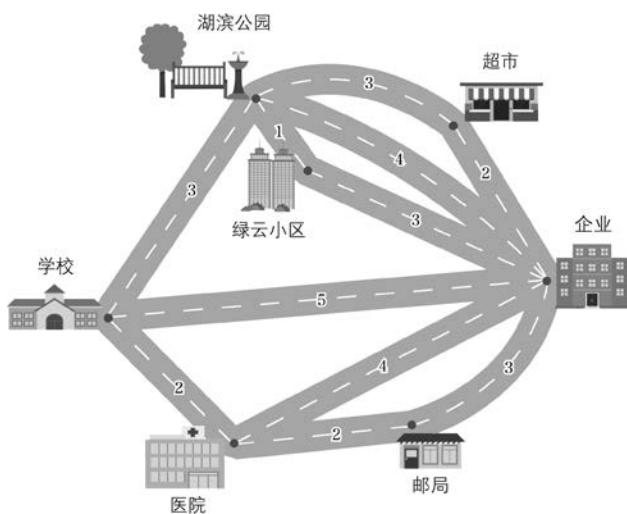


图 1.1 道路图

表 1.2 地点编号表

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

步骤2 根据图1.1将不同地点间直接到达(不经过其他地点)的道路长度(单位:千米)填入表1.3。图1.1中的数字表示道路长度(单位:千米)。此步骤记录地点间直接到达的距离,这也是交通路线规划中非常重要的数据。表1.3中的数字表示两个地点间直接到达的距离(即道路长度),“/”表示两个地点之间不能直接到达。

表1.3 地点间直接到达的道路长度表

	A	B	C	D	E	F	G
A	/	3	2	/	/	/	5
B	3	/	/	1	3	/	4
C	2	/	/	/	/	2	4
D	/	1	/	/	/	/	3
E	/	3	/	/	/	/	2
F	/	/	2	/	/	/	3
G	5	4	4	3	2	3	/

步骤3 计算出从学校(A)到企业(G)一共有6条可以到达的路线,并将相关数据填入表1.4。通过此步骤,学生能够了解到利用交通数据可以设计两个地点间的多条路线。教师引导学生对每条路线进行分析,可以让学生按照路线长度排序,并分析不同路线的特点。比如:路线1是直接到达路线且长度最短;路线2和路线3长度一样,但是路线2没有经过绿云小区。

表1.4 学校到企业的路线规划表

路线编号	路线	路线长度(千米)
1	A-G	5
2	A-B-G	7
3	A-B-D-G	7
4	A-B-E-G	8
5	A-C-G	6
6	A-C-F-G	7

步骤4 根据出行需求及路况,设计推荐路线并填入表1.5,格式为“路线(长度)”,其中路线长度以千米为单位。教师引导学生先设计不同的出行需求及路况(比如堵塞、修路等)并填入表1.5中第一列,再从表1.4中查找路线填入表1.5中第二列。

表 1.5 学校到企业的推荐路线表

出行需求及路况	推荐路线(可以有多条)	说明
路线长度最短	A - G(5)	根据路线长度可知
A - G 堵塞	A - C - G(6), A - B - G(7), A - B - D - G(7), A - C - F - G(7), A - B - E - G(8)	不能选择含有“A - G”的路线
A - G、B - G、D - G 修路, 道路不通	A - C - G(6), A - C - F - G(7), A - B - E - G(8)	不能选择含有“A - G”“B - G”“D - G”的路线
从学校到企业的路上需要经过绿云小区接某位同学	A - B - D - G(7)	路线中必须出现 D(绿云小区)
周末到湖滨公园的人很多, 经常发生道路拥堵	A - G(5), A - C - G(6), A - C - F - G(7)	路线中不出现 B(绕开湖滨公园)

通过本项目任务, 让学生感受到交通数据在交通路线规划中的价值, 即为人们的不同出行需要提供多条路线选择, 为人们的生活带来便利。教师可以以本项目任务为切入点, 引导学生发现交通数据在生活中的其他应用实例, 完成表 1.6。

表 1.6 交通数据的价值

应用场景	场景中的数据	数据价值
交通路线规划	地点位置、道路长度、路况等	为人们的不同出行需要提供多条路线选择, 为人们的生活带来便利
路况实时查询	地感监测数据、卫星监测数据、视频监控数据、车载导航数据等	为人们出行提供路况实时查询, 帮助交通部门了解实时路况并及时疏通拥堵道路, 帮助人们及时避开拥堵
公交车电子站牌	公交车实时位置、车速、站点位置、地图数据等	为人们提供公交车到站实时查询, 方便人们出行
在线打车	出租车信息、司机信息、乘客信息、出租车实时位置、乘客实时位置、地图数据等	为人们打车提供快捷方便的服务, 实现足不出户就可以叫车; 为出租车司机接单带来便利
.....

3. 项目活动的评价

本节的项目评价主要包括: 考查学生对数据价值等核心概念的掌握程度; 考查学生从生活实例出发, 发现数据所蕴含的价值, 了解数据价值的体现, 完成“交通路线规划中的数据价值”项目任务的情况。

评价建议: 对学生掌握核心概念的程度进行过程性评价, 可以采用分组讨论并提交报告等方式; 对学生参与任务的积极性、完成任务的情况进行过程性评价。

四、作业练习与提示

■ 题目描述

通信大数据行程卡可以分析手机信令数据,通过用户手机所处的基站位置获取用户的位置数据,通过对这些数据进行采集、管理与分析,准确定位用户的行程轨迹,形成用户的行程码,显示用户一段时间内的到达和途经城市,为疫情防控提供科学精准的技术支撑,发挥了重要的作用,这体现了_____。

- A. 数据思维
- B. 数据管理
- C. 数据分析
- D. 数据价值

■ 作业提示

D

五、教学参考资源

■ 参考资料:数据科学与工程的应用领域

数据科学与工程正在改变整个世界,重新定义人与自然的互动方式。它正在改变我们的学习方式、工作方式、生活方式、娱乐方式……下面是数据科学与工程的几个具有创新性、洞察力和激励性的典型领域。

1. 医疗:对生活的实际影响

大数据创新正在重塑医疗业的几乎各个方面。医学专家正在利用大数据来改变对患者的诊断方式;制药公司正在使用新的信号传感器来跟踪药物的副作用,将研发重点放在最有可能成功的药品上;疾病研究人员和流行病学家正在绘制疾病暴发的分布图,并将新细菌分解为遗传组分,从而阻止病毒的暴发;专业保险人员则正在筛选更加智能、受众面更广、索赔欺诈风险减少、激励机制更健全的医疗保险方案;还有基因医学的突破以及个性化药品的出现;等等。总之,大数据洞察将实现更健康的医疗保健体系。

2. 宇宙:解开宇宙奥秘,造福全人类

中国 2016 年建成的 500 米口径球面射电望远镜(FAST)是世界上最大口径的射电望远镜,全世界的科学家都将有望使用这样一个多学科基础研究平台所产生的海量数据,研究和揭示众多领域的科学奥秘,造福全人类。

3. 娱乐:由数据驱动的洞见

一直以来,音乐产业都是由具有丰富经验的音乐专家推动。签约和宣传哪些歌手,往往由具有出众的音乐判断能力的专家决定,他们能够凭借经验和直觉预测哪些歌手会成功,哪些歌手注定只能昙花一现或者默默无闻。基于大数据的创新成果将改变这一切。唱片公司收集并关联各种与歌手和歌曲相关的数据集(下载次数、外界评论、商品销售数量等),并将这些数据与空间地理和时间数据进行关联(如音乐会地点和日期以及电视播出次数等)。而音乐迷也经常会用各种褒贬不一的词汇来表达对歌手和音乐的评价,这一点可以通过对海量的文本进行分析得到。若要发现下一位有潜力的歌手,只需进行语义和文本挖掘就可以了。

4. 市民公用服务:智慧能源

能源供给是一个城市的关键领域。基于大数据的分析,可以使得能源供应商在用电高峰期或恶劣的天气情况下仍保持有效的用电供给。越来越多的城市都开始采用先进的数据分析平台来获得天气数据、实时传感器数据、连续用电计量数据等,然后使用强大的可视化和建模工具来预测电网中的供需状态。这样,能源供应商就可以预测出市民在未来每一时刻的用电需求,并在突发事件发生时,合理调配整个电网的能源供给。

5. 网络空间安全:大数据作为核心战略

网络空间安全是一个系统的全面的安全问题,需要确定国家层面的网络空间安全总体战略。大数据是网络空间安全战略的核心技术保障,通过对重要领域数据的寡头控制,特别是充分利用大数据领域的综合优势,注重基于大数据的整体网络安全态势的感知和掌控,以保证网络空间的行动自由和实际控制,实现对现实世界的掌控能力和攻击能力。大数据技术领域的竞争,将事关国家安全和军事安全。

——摘自《数据科学与工程导论》(有删改),王伟、刘垚,华东师范大学出版社

六、教学参考案例

■ 参考案例

数据价值

上海市七宝中学 吴美琴

(1课时)

1. 学科核心素养

在特定情境中,能综合分析获取的信息,挖掘数据的价值。(信息意识)

2. 《课程标准》要求

结合生活实际,认识到数据是一种重要的资源,通过科学管理与分析数据,可以使数据实现其应有价值。

3. 学业要求

- 学生能在特定的信息情境中,根据业务数据问题解决的需要,利用多种途径采集与甄别数据。

- 学生能够采用适当的方法提取数据。

4. 教学内容分析

本节课的核心概念是数据价值。数据本身蕴含巨大的价值,体现在生活中的不同方面,包括社会民生、经济发展和政府决策等,在教学中要引导学生发现身边的数据价值。在“交通路线规划中的数据价值”项目任务中,需要对出发地、目的地、道路长度、路况等多种交通数据进行分析,学生可以科学管理与分析交通数据,合理规划路线。

5. 学情分析

授课对象为高中学生,他们在前面的学习中了解了数据处理的一般过程和基本方法,

有一定的分析问题、解决问题的能力。通过生活中的实例,帮助学生理解身边的数据,通过路线规划分析了解数据价值。

6. 教学目标

- 认识到数据是一种重要资源,了解数据的价值,发现生活中数据所蕴含的价值。
- 认识到通过科学管理与分析数据,可以使数据实现其应有价值。

7. 教学重难点

- 教学重点:通过分析交通数据对路线进行合理规划。
- 教学难点:探索交通数据对人们生活的影响,感受数据所蕴含的价值。

8. 教学策略分析

本章设计的项目活动主题是“身边的数据价值以及数据管理与分析”,结合教科书中的项目情境,在教学中以“交通路线规划中的数据价值”作为本节课的项目任务,使学生通过“规划从学校到企业的路线”项目活动,学会提取特定情境的数据,感受交通数据的价值。

借助北斗卫星导航系统案例和纪录片《大数据时代》拓展学生的视野,激发学生对数据时代的好奇心,引导学生关注科技发展。

9. 教学环境

传统多媒体教室。

10. 教学过程设计(见表 1.7)

表 1.7 教学过程设计表

教学环节	教学内容	学生活动	设计意图
课前准备	发放教学资料(活动记录单),每组一份 我们身处大数据时代,我们的生活和数据息息相关,数据可以帮助我们了解各种信息,也可以帮助我们作出选择。这节课,我们来体验一下数据的价值。	浏览资料	初识任务,明确目标
情境导入	我校每年都会组织社会实践活动,今年我们要到一家研发智能手环的企业参观调研。 知道了企业的地址之后,我们可以用地图软件导航,到达该企业。 你们知道地图软件是如何实现智能推荐路线的吗?需要掌握哪些数据呢?	回顾生活经验,思考并回答问题	创设情境,引入项目主题——规划从学校到企业的路线
项目活动	发布项目任务,组织学生以小组为单位完成。学生根据活动记录单,规划学校到企业的路线,完成项目任务。 (1) 根据道路图对所有地点进行编号。 (2) 计算任意两个地点之间直接到达的路线长度。 (3) 分析从学校(A)到企业(G)的路线,计算路线长度。请1个小组分享答案,投影展示。 (4) 设计不同的交通情境,以及在这些情境下从学校(A)到企业(G)的推荐路线。请1个小组投影展示路线规划表,说说面对不同的出行需求及	根据项目活动需要解决的问题,小组共同分析,并进行汇报	学生从问题出发,通过提取数据对交通路线进行合理规划,感受数据的价值

教学环节	教学内容	学生活动	设计意图
	路况的推荐路线及原因,其他小组补充。 (5) 请1个小组分享本项目活动中用到了哪些数据,这些数据有什么价值,交通数据还能为我们的生活提供哪些便利		
活动小结	归纳交通数据在日常生活中的价值。 查阅北斗卫星导航系统官网,思考中国为什么要建设北斗卫星导航系统,思考交通数据对于国家的价值。日常生活中,还有哪些数据能为人们的生活提供便利?举例说明应用场景	回顾生活中的场景,思考并回答问题	发散思维,感受生活中的数据价值体现
教师总结	观看纪录片《大数据时代》,感受数据在日常生活、各行各业中的应用。 总结:信息社会的数据价值体现在生产生活以及各行各业中,例如气象数据、企业生产销售数据、政府数据等都蕴含巨大的价值。 我们要善于发现生活中数据所蕴含的价值	归纳总结	通过观看纪录片,了解大数据的发展,拓宽视野

【活动记录单】

交通路线规划中的数据价值

一、项目主题

规划从学校到企业的路线。

二、项目情境

下周,我们要开展一次社会实践活动——到一家研发智能手环的企业参观调研。现在,我们要规划从学校到企业的路线。为了完成这个任务,需要对出发地、目的地、道路长度、路况等多种交通数据进行分析,推荐合适的路线。

三、活动记录

1. 对图1.2中的所有地点进行编号,如表1.8所示。(图1.2中的数字表示道路长度,单位为“千米”)。

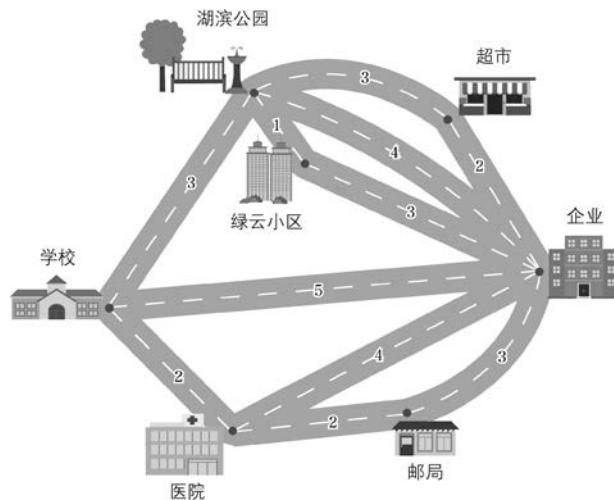


图1.2 道路图

表1.8 地点编号表

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

2. 根据图 1.2 将不同地点间直接到达(不经过其他地点)的道路长度(单位:千米)填入表 1.9。表 1.9 中的数字表示两个地点间直接到达的距离(道路长度),“/”表示两个地点之间不能直接到达。

表 1.9 地点间直接到达的道路长度表

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

3. 从学校(A)到企业(G)共有几条可以到达的路线? 请将路线及其长度填入表 1.10。

表 1.10 学校到企业的路线规划表

路线编号	路线	路线长度(单位:千米)
1	A-G	5

4. 根据出行需求及路况,设计推荐路线,设想更多的出行需求和路况(比如堵塞、修路等),填写在表 1.11 中。推荐路线的格式为“路线(长度)”,其中路线长度以千米为单位。

表 1.11 学校到企业的推荐路线表

出行需求及路况	推荐路线(可以有多条)
路线长度最短	A-G(5)
A-G 堵塞	
A-G、B-G、D-G 修路,道路不通	
周末到湖滨公园的人很多,经常发生道路拥堵	

续表

出行需求及路况	推荐路线(可以有多条)

5. 在本次交通路线规划中,使用了哪些数据,体现了什么样的数据价值? 交通数据还能为我们的生活提供哪些便利?

第二节

数据管理与分析技术的重要性

一、教学目标与重点

教学目标:

- 结合生活实际,感受数据管理与分析技术的重要性。

教学重点:

- 结合具体实例,感受数据管理与分析技术在实现数据价值中体现出来的重要性。

二、教学说明与建议

建议教师根据教科书中的项目活动先以学生为主,围绕企业商品月销售量数据分析,组织学生进行充分讨论,引导学生了解数据管理与分析技术的重要性。然后,教师再讲解教科书上的知识内容,在项目体验中总结知识要点,并结合学生身边真实场景适当补充更多案例,如图书借阅、在线打车、在线购物、在线社交网络等,引导学生讨论在这些场景中数据管理与分析技术的重要性体现在哪些方面,从而加深学生对知识的记忆、认知和理解。

三、项目实施与评价

1. 核心概念精解

本节的核心概念是数据管理与分析技术的重要性。数据本身蕴含着价值,通过数据管理与分析可以发现数据更多的价值,为科学决策提供重要依据,这就是数据管理与分析技术的重要性。

信息科技的发展已经使产生、存储、管理并实时地分析处理大量的数据得以实现。随着智慧城市、万物互连时代的到来,很多行业和企业一开始都认为数据积累得越多越好,有了大量的数据资源后,就能对行业发展、企业业务产生更大的价值。但是他们很快就发现自己陷入了“数据越多越有用”的误区。拥有大量数据不等于有能力利用好它们,面对海量的数据,想让它们为科学决策提供支持,就必须知道如何去发现更多的数据价值,这就需要数据管理与分析技术。通过有效的数据管理与分析可以发现更多的数据价值,为科学决策提供重要依据。需要强调的一点是:通过数据管理与分析技术可以发现更多的数据价值,但是反过来,认为数据价值只能通过数据管理与分析技术才能被发现是不对的。数据本身就蕴含了价值,但是还有一些价值是数据本身无法体现的,隐藏在数据内部,需要利用数据管理与分析技术才能发现这些价值。例如,在本节的项目任务“企业商品月销售量数据分析”中,从商品月销售量数据中可以看出每个月销售量的高低变化,企业可以了解到商品的销售情况,这是数据本身的价值。但是商品月销售量数据中还隐藏了更多的价值,比如可以根据这些数据预测未来的月销售量,这是只看数据本身看不出来的,需要通过数据分析才能实现。

2. 项目活动的具体实施

在“企业商品月销售量数据分析”项目任务中,教师向学生讲解简单移动平均法,学生按照计算公式进行计算并填写表 1.12。教师也可以引导学生将数据输入 Excel 表中,应用公式(平均值公式)、函数(average 求平均值函数、round 四舍五入函数)或者 Excel 中的移动平均数据分析工具(详见“企业商品月销售量预测移动平均示例.xlsx”)进行移动平均值的计算(计算结果保留到整数),将结果填入表 1.12 中。

表 1.12 今年智能手环月销售量预测表

月份	上一年度的月销售量(个)	n=2 的预测值	n=3 的预测值	n=4 的预测值
1	180	/	/	/
2	205	(205 + 180) / 2 = 193	/	/
3	220	(220 + 205) / 2 = 213	(220 + 205 + 180) / 3 = 202	/
4	243	(243 + 220) / 2 = 232	(243 + 220 + 205) / 3 = 223	(243 + 220 + 205 + 180) / 4 = 212

续表

月份	上一年度的月销售量(个)	$n=2$ 的预测值	$n=3$ 的预测值	$n=4$ 的预测值
5	234	239	232	226
6	257	246	245	239
7	260	259	250	249
8	285	273	267	259
9	290	288	278	273
10	300	295	292	284
11	305	303	298	295
12	288	297	298	296

3. 项目活动的评价

本节的项目评价主要包括:考查学生对数据管理与分析技术的重要性等核心概念的掌握程度;考查学生从生活实例出发,了解数据管理与分析技术的重要性,完成“企业商品月销售量数据分析”项目任务的情况。

评价建议:对学生掌握核心概念的程度进行过程性评价,可以采用分组讨论并提交报告等方式;对学生参与任务的积极性、完成任务的情况进行过程性评价。

四、作业练习与提示

■ 题目描述

- 张明收集了今年上海10月份每天的最高气温和最低气温,希望利用这一组数据求平均值,从而预测上海明年10月份每天的最高气温和最低气温。张明采用的数据预测方法是_____。
- 某企业今年的商品库存量如表1.13所示,当移动平均时期个数n取值为4时,该企业明年7月份的预计库存是_____台。(计算结果保留到整数)

表1.13 某企业今年的商品库存量

月份	1	2	3	4	5	6	7	8	9	10	11	12
库存量(台)	150	172	203	175	230	260	180	187	119	100	50	100

- 判断题:数据价值必须通过数据管理与分析技术才能体现。()

■ 作业提示

- 简单移动平均法。
- 计算方法:4~7月份4个月的库存量求平均值。
- 错误,因为数据本身就具有价值。

五、教学参考资源

■ 参考资料:大数据时代的思维变革

大数据的精髓在于我们分析信息时的三个转变,这些转变将改变我们理解和组建社会的方法。

第一个转变就是,在大数据时代,我们可以分析更多的数据,有时候甚至可以处理和某个特别现象相关的所有数据,而不再依赖于随机采样。19世纪以来,当面临大量数据时,社会都依赖于采样分析。但是采样分析是信息缺乏时代和信息流通受限制的模拟数据时代的产物。以前我们通常把这看成是理所当然的限制,但高性能数字技术的流行让我们意识到,这其实是一种人为的限制。与局限在小数据范围相比,使用一切数据为我们带来了更高的精确性,也让我们看到了一些以前无法发现的细节——大数据让我们更清楚地看到了样本无法揭示的细节信息。

第二个改变就是,研究数据如此之多,以至于我们不再热衷于追求精确度。当我们测量事物的能力受限时,关注最重要的事情和获取最精确的结果是可取的。如果购买者不知道牛群里有80头牛还是100头牛,那么交易就无法进行。直到今天,我们的数字技术依然建立在精准的基础上。我们假设只要电子数据表格把数据排序,数据库引擎就可以找出和我们检索的内容完全一致的检索记录。

这种思维方式适用于掌握“小数据量”的情况,因为需要分析的数据很少,所以我们必须尽可能精准地量化我们的记录。在某些方面,我们已经意识到了差别。例如,一个小商店在晚上打烊的时候要把收银台里的每一分钱都数清楚,但是我们不会、也不可能用“分”这个单位去精确度量国民生产总值。随着规模的扩大,对精确度的痴迷将减弱。

达到精确需要有专业的数据库。针对小数据量和特定事情,追求精确性依然是可行的,比如一个人的银行账户上是否有足够的钱开具支票。但是,在这个大数据时代,很多时候,追求精确度已经变得不可行,甚至不受欢迎了。当我们拥有海量即时数据时,绝对的精准不再是追求的主要目标。

大数据纷繁多样,优劣掺杂,分布在全球多个服务器上。拥有了大数据的我们不再需要对一个现象刨根究底,只要掌握大体的发展方向即可。当然,我们也不是完全放弃了精确度,只是不再沉迷于此。适当忽略微观层面上的精确度会让我们在宏观层面拥有更好的洞察力。

第三个转变因前两个转变而促成,即我们不再热衷于寻找因果关系。寻找因果关系是人类长久以来的习惯。即使确定因果关系很困难而且用途不大,人类还是习惯性地寻找缘由。相反,在大数据时代,我们无须再紧盯事物之间的因果关系,而应该寻找事物之间的相关关系,这会给我们提供非常新颖且有价值的观点。相关关系也许不能准确地告知我们某件事情为何会发生,但是它会提醒我们这件事情正在发生。在许多情况下,这种提醒的帮助已经足够大了。

如果数百万条电子医疗记录显示橙汁和阿司匹林的特定组合可以治疗癌症,那么找

出具体的药理机制就没有这种治疗方法本身来得重要。同样,只要我们知道什么时候是买机票的最佳时机,就算不知道机票价格疯狂变动的原因也无所谓了。大数据告诉我们“是什么”而不是“为什么”。在大数据时代,我们不必知道现象背后的原因,我们只要让数据自己发声。

——摘自《大数据时代》,维克托·迈尔-舍恩伯格、肯尼思·库克耶(著),盛杨燕、周涛(译),浙江人民出版社

六、教学参考案例

■ 参考案例

数据管理与分析技术的重要性

上海市七宝中学 吴美琴
(1课时)

1. 学科核心素养

- 能认识到数据管理与分析技术对于提高数据价值的重要性。(信息意识)
- 了解常用的数据管理与分析技术,根据需求选择合适的工具进行数据分析。(计算思维)

2. 《课程标准》要求

结合生活实际,认识到数据是一种重要的资源,通过科学管理与分析数据,可以使数据实现其应有价值,感受数据管理与分析技术的重要性。

3. 学业要求

• 能认识有效管理与分析数据对获取有价值信息、形成正确决策的作用与意义,认识数据管理与分析技术对人类社会生活的重要影响。

- 能正确选用数据分析方法和工具,分析并解释数据。

4. 教学内容分析

本节课的核心概念是数据管理与分析技术的重要性。在上一节课中学生了解了数据本身的价值,而通过数据管理与分析技术可以挖掘数据更多的价值,这是本节课的核心思想。在“企业商品月销售量数据分析”项目活动中,运用简单移动平均法挖掘月销售量数据的更多价值,通过已有销售数据对未来销售量进行预测,这是数据管理与分析技术重要性的体现。

5. 学情分析

通过前面的学习,学生了解了生活中的数据所蕴含的价值,知道了数据的重要性。学生在以前的学习中接触过一些常用的数据分析方法和工具,有一定的数据分析基础。

6. 教学目标

- 认识数据管理与分析技术的重要性。
- 了解四种常见的数据分析方法和特点。

- 了解简单移动平均法的作用,知道简单移动平均法的计算方法和适用场景,会使用简单移动平均法进行数据预测。

7. 教学重难点

- 教学重点:使用简单移动平均法进行数据预测。
- 教学难点:使用简单移动平均法进行数据预测,感受数据管理与分析技术的重要性。

8. 教学策略分析

- 通过销售领域的数据预测方法,理解并学会利用简单的统计方法(简单移动平均法)对数据进行预测,体验数据分析技术的价值,从而感受数据分析技术的重要性。

• 在项目活动中,设置了3个任务,难度逐步递增。任务1,通过自主阅读了解常用的四种数据分析类型;任务2,结合项目情境,引入简单移动平均法,了解简单移动平均法的适用场景,计算出 $n=2$ 时的预测值;任务3,通过对n值变化计算预测值,进一步巩固简单移动平均法的计算方法,探索n取不同的值时预测值的变化,从而理解选择平均移动法的作用。

9. 教学环境

计算机机房、表格软件。

10. 教学过程设计(见表1.14)

表1.14 教学过程设计表

教学环节	教学内容	学生活动	设计意图
课前准备	发放教学资料(活动记录单),每组一份	浏览资料	初识任务,明确目标
情境导入	上节课我们了解到数据本身蕴含着很大的价值,为我们的生活提供了方便,帮助我们进行科学决策。例如智能手环收集的血压值可以辅助进行健康预警。 数据本身不具有预警价值,但是对数据进行管理和分析可以挖掘其预警价值,让数据产生了更大的价值。 这节课,我们就来看看如何对数据进行管理和分析。 企业为我们提供了上一年度各月的销售量数据,请大家思考如何根据这些数据预测该商品今年的月销售量	回顾数据分析方法,思考并回答问题	创设情境,引入项目主题——企业商品月销售量数据分析
项目活动	在解决这个问题之前,我们先来了解一下数据分析的分类。 发布项目任务,组织学生以小组为单位完成。 (1)请1个小组分享任务1的答案,投影展示,结合具体案例阐述不同数据分析类型的用法。 (2)请1个小组分享任务2的答案,说明简单移动平均法的计算方法。 (3)请1个小组分享任务3的答案,计算不同的移动平均时期下的预测值,说说由此获得的启示。其他小组补充	根据项目活动需要解决的问题,小组共同分析,按活动记录单完成任务,并进行汇报	学生从问题出发,自主学习简单移动平均法的计算思路,感受数据管理与分析技术的重要性

续表

教学环节	教学内容	学生活动	设计意图
活动小结	归纳简单移动平均法的核心思想及对销售决策的价值。 观看《大数据时代》第1集中第20~25分钟视频片段，感受数据管理与分析技术在生活健康领域的应用	理解简单移动平均法的作用	感受数据分析技术的价值
教师总结	用思维导图总结本节课的重点	梳理本节课的核心概念	通过思维导图整体了解数据分析方法及其作用

【活动记录单】

企业商品月销售量数据分析

一、项目主题

企业商品月销售量数据分析。

二、项目情境

在企业进行调研时,我们了解了某种智能手环上一年度每个月的销售量(如表1.15所示)。

表1.15 某智能手环上一年度1~12月份的月销售量表

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

思考:如何根据这些数据预测该商品今年每个月的销售量?

三、活动记录

任务1:了解数据分析方法。

阅读教科书第19页的内容,也可查询网络资料,了解数据分析的四大分类,分析其特点,填写表1.16。

表1.16 数据分析的四大分类

类型	含义	主要分析方法	应用案例
描述性分析			
诊断性分析			
预测性分析			
规范性分析			

在本项目中,应采用_____进行分析。

任务 2：了解简单移动平均法。

简单移动平均法是一种常用的数据预测方法。阅读教科书第 8 页的内容，查阅资料，了解简单移动平均法的计算方法，回答下列问题。

(1) 简单移动平均法适合于预测()。

- A. 平稳序列
- B. 非平稳序列
- C. 有趋势成分的序列
- D. 有季节成分的序列

(2) 设 $n = 2$ ，利用简单移动平均法对今年的月销售量进行预测。(在 Excel 中实现)

任务 3：利用简单移动平均法，预测今年每个月的销售量。

按照表 1.17 设置的移动平均时期个数，用 Excel 计算今年的月销售量预测值，并填入表中。

表 1.17 今年某智能手环月销售量预测表

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

根据上面的数据思考：简单移动平均法有何作用？为什么数据管理与分析技术对于实现数据价值是非常重要的？

第三节

数据管理与分析方案

一、教学目标与重点

教学目标：

- 结合具体案例,初步了解分析业务需求、建立数据管理与分析问题整体解决方案的基本过程;尝试对既定方案进行分析、评价,发现问题并优化方案。

教学重点：

- 结合具体实例,了解建立数据管理与分析方案的基本过程,包括数据需求分析、数据管理、数据分析、方案评价和优化以及科学决策。
- 了解如何对方案进行分析、评价和优化。

二、教学说明与建议

建议教师根据教科书中的项目活动先以学生为主,围绕学生社团网站数据管理与分析方案的建立,组织学生进行充分讨论,从而引导学生了解建立数据管理与分析方案的基本过程,以及如何对方案进行分析、评价和优化。然后,教师再讲解教科书上的知识内容,在项目体验中总结知识要点,并结合学生身边真实场景适当补充更多案例,如图书借阅、在线打车、在线购物、在线社交网络等,引导学生在这些场景中建立数据管理与分析方案,从而加深学生对知识的记忆、认知和理解。

本节 2 课时安排建议:“数据需求分析”0.5 课时,“数据管理”0.5 课时,“数据分析”0.5 课时,“数据管理与分析方案的评价和优化”和“科学决策”共计 0.5 课时。

三、项目实施与评价

1. 核心概念精解

本节包含了 6 个核心概念,分别是建立数据管理与分析方案的基本过程、数据需求分析、数据管理、数据分析、方案评价和优化、科学决策。

(1) 建立数据管理与分析方案的基本过程

建立数据管理与分析方案的基本过程包括数据需求分析、数据管理、数据分析、方案评价和优化、科学决策 5 个环节。

(2) 数据需求分析

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

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

可以通过调研、咨询、讨论、收集资料等多种方式获取用户对数据的需求。

(3) 数据管理

数据管理是数据管理与分析过程中的重要环节。数据管理是利用计算机硬件和软件技术对数据进行有效采集、存储、处理和应用的过程,其目的在于充分有效地发挥数据的作用。

数据管理包括数据采集,即对数据需求分析中需要输入的数据进行采集。在进行数据采集时,首先需要明确数据来源,然后利用合理的方式,有目的地采集数据,这是保证数据管理与分析过程正确有效的基础。数据采集方法多样:既可以通过程序对数据库、服务器日志文件、网站进行数据采集;也可以通过设备,比如传感器、摄像头、麦克风、智能穿戴设备等进行数据采集;还可以通过人工进行数据采集,如调查问卷等。例如,需要采集学生社团网站一周发布的所有文章的数据,这些数据可以从该网站的数据库中导出。但是,如果没有权限,那么也可以编写网络爬虫程序从该网站上采集。采集数据时,应该在保证数据安全可靠的前提下,使采集到的数据尽可能全面、客观、具体、准确。

采集到的数据经过整理后一般需要重新组织并进行存储和管理。数据组织是指将具有某种逻辑关系的一批数据组织起来,按一定的存储表示方式配置在计算机的存储器中。

随着计算机技术的发展,数据管理经历了人工管理、文件系统、数据库系统三个发展阶段。在人工管理阶段,没有专门的软件用来管理数据,管理数据需要依赖应用程序本身来处理。数据和程序紧密联系,一组数据只能对应一个应用程序,数据不能共享、不具有独立性。在文件系统阶段,数据存储在文件中,由操作系统统一管理,数据共享性差、冗余度大,不具有独立性,并且安全性和完整性难以保证。到了数据库系统阶段,数据存储在数据库中,由专门的数据库管理系统对其进行管理。数据库系统相较于人工管理和文件系统,数据结构化并且能够共享,具有较高的独立性,可以保证数据的安全性和完整性。

目前,数据可以通过文件和数据库进行管理,主要的方式是数据库管理。数据库是长期储存在计算机内、有组织的、可共享的数据集合。数据库可以对数据进行操作、备份,并进行数据并发控制、安全管理。也可以通过文件系统对数据进行管理。例如,可以利用分布式文件系统管理大数据。分布式文件系统是指文件系统管理的数据不一定在本地计算机上,这些数据可能存储在通过计算机网络连接的其他计算机上。

随着信息技术的发展,数据库管理方法与技术也发生了重大变革,尤其是在大数据时代,由于大数据具有数据量大、数据增长快速、非结构化和半结构化数据多、价值密度低等特征,传统的关系数据库已不能满足大数据时代的数据存储要求,需要采取新的数据思维来管理数据,相应地涌现出了许多新型的非关系数据库,统称为 NoSQL(Not only SQL)数据库,主要有四大类型,即键值数据库、列族数据库、文档数据库、图数据库。非关系数

数据库具有更灵活的数据模型,不局限于固定的数据结构,可以减少时间和空间的开销,具有更好的并发性和可扩展性,更适合解决大规模数据集合、多重数据种类等大数据存储和管理。但是相较于关系数据库,NoSQL 数据库在面对复杂查询时,效率较低,并且无法保证数据的一致性和完整性。

(4) 数据分析

数据分析需要将采集到的数据进行整理、加工,然后再进行分析,帮助决策者进行科学决策。由于被分析的数据往往有多个来源,并且数据类型多种多样,因此在分析前需要对数据进行预处理和整理,然后设计合理高效的数据分析方法,再利用数据分析工具对数据进行深入分析,并将分析结果可视化,以图表形式直观、美观、清晰地展示给用户。数据分析具有较强的专业性。目前普遍应用的数据分析工具中,以开源软件为主的有 Python 语言、R 语言等。

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

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

(5) 方案评价和优化

方案的评价和优化贯穿于建立数据管理与分析方案的整个过程中,需要根据不同的应用,制定评价标准,按照评价标准去检验数据需求分析、数据管理以及数据分析等各个环节。需要注意的是,每个环节完成后,都应该进行评价和优化,而不是整个方案设计完成后才进行。如果整个方案设计完成后才进行评价和优化,那么一旦中间某个环节有问题,将会导致该环节以及其后各环节的方案都需要进行修改。例如,数据需求分析完成后应该随即进行评价和优化,如果发现问题,需要针对问题进行改进和优化,直至没有问题后再进行数据采集。

针对不同的环节,评价和优化的方法有多种,主要可以从以下几个方面进行评价:

① **数据需求目标评价**:数据需求分析是否可以解决需要解决的问题,是否可以达到既定目标。

② **数据真实性和有效性评价**:采集数据的目的是否明确;数据来源以及采集到的数据是否全面、是否真实可信、是否完整、是否合乎法律和伦理要求。

③ **方案合理和有效性评价**:数据管理方案是否合理、是否具有扩展性,数据库选择是否合适;数据分析方法是否正确高效、是否选择了有效的数据分析工具,分析结果是否可以为用户提供服务和决策支持。

④ 方案安全性和风险性评价:整个数据管理与分析方案是否将风险控制在可接受的范围内,是否符合相关法律法规、标准规范以及伦理要求。

在对数据管理与分析方案进行评价后,如果发现问题,需要及时对方案进行合理的改进和优化,从而更好地解决问题,指导人们进行科学决策。

(6) 科学决策

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

2. 项目活动的具体实施

(1) “学生社团网站数据需求分析”项目实施

进行数据需求分析时,首先要明确需要解决的问题。学生社团网站需要解决以下两个问题:①找到网站一周内发布的哪些文章的浏览量最高、评论量最高、转发量最高;②网站如何向用户进行文章的个性化推荐,方便学生更快地找到感兴趣的内容。

明确了需要解决的问题以后,根据这些问题分析需要输入哪些数据、输出什么结果,以及以什么样的方式输出结果。对于问题①,以浏览量为例,想要找到一周内发布的文章中浏览量最高的文章,首先要统计出一周内发布的每篇文章的浏览量,然后根据浏览量的高低对文章进行排序,就可以找到浏览量最高的文章。对于问题②,想要向用户推荐其感兴趣的文章,必须先了解用户的兴趣,用户的兴趣可以用多个指标来衡量,比如年龄、地区、爱好都可能会影响用户的兴趣,另一方面,用户的网站行为数据可以更有效地反映用户的兴趣,因为用户浏览、评论或转发的文章通常都是其感兴趣的。因此我们可以完成学生社团网站数据需求分析如表 1.18 所示。

表 1.18 学生社团网站数据需求分析表

需要解决的问题	需要输入的数据	输出的结果	输出方式
找到一周内每天浏览量最高的文章、评论量最高的文章、转发量最高的文章	一周内发布的所有文章的数据,包括文章的编号、标题、内容、发布时间、发布作者、浏览量、评论量、转发量等	一周内每天浏览量最高的文章、评论量最高的文章、转发量最高的文章	图表可视化方式
向学生推荐感兴趣的 文章	文章编号、学生的用户名、用户对文章的访问数据(是否浏览、转发、评论、收藏文章)	为不同的用户所推荐的文章	推荐文章链接列表

(2) “学生社团网站数据管理”项目实施

数据管理首先要对需要输入的数据进行采集,采集到的数据经过整理后一般还需要

重新组织并进行存储和管理,才能为信息管理系统和数据分析所用。目前最常见的数据库是关系数据库,它按一定的数据模型,即关系模型对数据进行组织描述和存储,关系模型中的关系就是二维表。数据管理方案需要确定数据用什么类型的数据库进行存储。学生社团网站的数据管理可以利用关系数据库进行存储和管理,设计文章数据表如表 1.19 所示。其中文章浏览次数、文章转发次数和文章评论次数需要通过用户对文章的访问数据表计算得到。表 1.20 是用户对文章的访问数据表的一个示例,这个表的数据由对数据库中的记录和服务器日志文件进行数据提取和计算得到。

表 1.19 文章数据表

1	2	3	4	5	6	7	8	9	10
文章 编号	文章 标题	文章 内容	文章所属 社团编号	文章作 者编号	文章发布 日期和时间	文章最后 一次修改 日期和时间	文章浏 览次数	文章转 发次数	文章评 论次数

表 1.20 用户对文章的访问数据表

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

续表

文章编号	用户编号	浏览	转发	评论	收藏	点赞
0003	e					
0004	e					
0005	e	是	是	是	是	是

通过对表 1.20 进行统计可以得到用户对每篇文章的评分,计算规则如下:用户对文章的一次行为记 1 分,行为包括浏览、转发、评论、收藏、点赞 5 种,因此用户对文章的评分最高值为 5,最低值为 0。0 表示用户没有任何行为,也就是说用户没有浏览过该文章。统计结果如表 1.21 所示。

表 1.21 用户文章评分表

	文章 0001	文章 0002	文章 0003	文章 0004	文章 0005
用户 a	0	3	4	5	0
用户 b	2	0	1	2	0
用户 c	1	0	3	1	4
用户 d	4	3	1	0	1
用户 e	2	4	0	0	5

(3) “学生社团网站数据分析”项目实施

要找到网站中浏览量、评论量、转发量分别最高的文章,首先需要统计文章的浏览量、评论量和转发量。以表 1.20 为例,从该表中统计出 5 篇文章的浏览量、评论量和转发量填入表 1.22,教师可以引导学生在数据分析软件中通过公式或者函数计算得到,例如在 Excel 中导入表 1.20 数据,把“是”替换为数字 1,然后使用 sumif 函数根据文章编号分别统计浏览量、评论量和转发量。统计完成后,分别对浏览量、评论量、转发量进行排序,就可以找到浏览量最高、评论量最高、转发量最高的文章编号。为了更直观地观察数据,可以采用图表可视化方法展示数据,如图 1.3 所示。以上都可以通过数据分析软件(如 Excel)完成,参考“学生社团网站数据分析方案设计活动示例.xlsx”文件。

表 1.22 文章关注度数据表

文章编号	浏览量(次)	评论量(条)	转发量(次)
0001	4	1	0
0002	3	3	1
0003	4	2	2
0004	3	1	1
0005	3	2	2

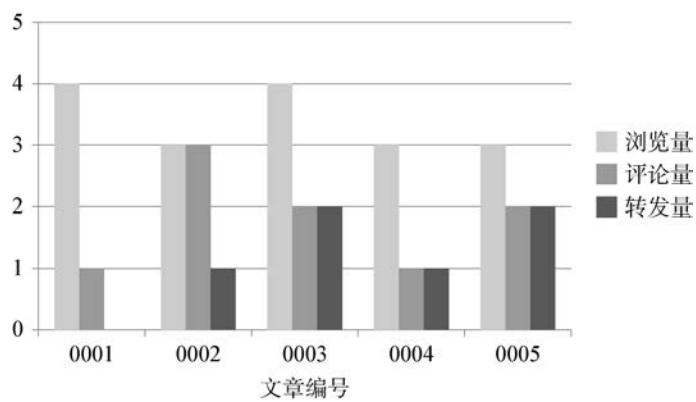


图 1.3 文章浏览量、评论量及转发量统计

上述活动完成后,可以选做以下活动:在采集数据时,通常还需要采集文章的发布日期和时间,这样可以分时间段统计出每天、每周、每月浏览量或评论量或转发量最高的文章。例如,对学生社团网站文章数据表(详见素材库)中的数据,运用数据分析工具统计出6月4日至6月10日间每天浏览量最大的文章,并通过图表可视化方式展现。教师可以引导学生在数据分析软件中完成该活动。以Excel为例:首先在数据表中利用MAX函数找到6月4日至6月10日间每天浏览量最高的文章编号、日期和浏览量,然后再插入图表进行数据可视化,详见“学生社团网站文章统计分析示例.xlsx”。

3. 项目活动的评价

本节的项目评价主要包括:考查学生对建立数据管理与分析方案的基本过程、数据需求分析、数据管理、数据分析、数据管理与分析方案的评价和优化、科学决策等基本概念的掌握程度;考查学生从生活中的实际问题出发,进行需求分析,建立数据管理与分析方案,并对方案进行分析、评价和优化,完成“学生社团网站数据管理与分析方案制定”项目任务的情况。

评价建议:对学生掌握基本概念的程度进行过程性评价,可以采用分组讨论并提交报告等方式,要重点对建立数据管理与分析方案的基本过程、数据需求分析的掌握程度进行评价;对学生参与任务的积极性、完成任务的情况进行过程性评价。

四、作业练习与提示

■ 题目描述

在信息社会,一切皆可数据化,包括学生的学习过程。请大家分组,针对在线学习系统中的某一个问题,设计数据管理与分析方案,并对其进行评价和优化。

■ 作业提示

本题为开放式题目,可以参考以下方案,但不局限于以下方案。建议学生分组(3~5人一组)查找资料,合作完成方案设计和优化。

1. 分析在线学习系统需要解决什么问题

在线学习系统是一个复杂的信息管理系统,如在线慕课平台、BB 平台等都是在线学

习系统。一般而言,在线学习系统的主要功能、产生的数据及数据管理与分析问题如表1.23所示。

表 1.23 在线学习系统的主要功能、产生的数据及数据管理与分析问题

主要用户	教师、学生、管理员		
主要功能、产生的数据及数据管理与分析问题	主要功能	产生的数据	数据管理与分析问题
	课程管理(新建、修改、删除)	课程数据	哪些课程学生访问量高、师生参与度高
	在线学习资源(课件、视频、其他文件)的共享和学习	学习资源数据、学生学习数据	学生学习情况分析,学生最常浏览的学习资源有哪些,学生学习哪些内容用的时间多、学习哪些内容用的时间少
	在线答疑交流	问题及回答数据	学生最常提问的内容是什么,哪些知识点学生疑问最多
	在线考试	考卷(考题)数据、学生答卷数据、学生成绩数据	学生成绩统计分析,卷面分析(题目得分统计,比如哪些题目平均分高,哪些题目平均分低)

2. 选择一个问题,进行在线学习系统的数据需求分析,设计数据管理与分析方案,并对其进行评价和优化

根据表1.23中列举的问题,每组学生可以选择一个进行数据需求分析,设计数据管理与分析方案,并对其进行评价和优化。学生也可以自己设计新的问题进行分析和方案制定。

五、教学参考资源

■ 参考资料1:什么是用户画像

什么是用户画像?从中文概念来讲,用户画像与用户角色非常相近,是用来勾画用户(用户背景、特征、性格标签、行为场景等)和联系用户需求与产品设计的,旨在通过从海量用户行为数据中“炼金”,尽可能全面细致地抽出一个用户的信息全貌,从而帮助解决如何把数据转化为商业价值的问题。从英文概念角度来讲,用户画像(user portrait)、用户角色(user persona)、用户属性(user profile)这三个概念其实都是各有侧重和容易混淆的。用户角色更倾向于业务系统中不同用户的角色区分。如在学校教务管理系统中,教师审核、设置课程,学生查看课程和成绩,那么教师、学生就是不同的用户角色。用户画像更倾向于对同一类用户进行不同维度的刻画。例如,对某个电商的买家进行用户画像设计,就是将买家进一步细分和具象,买家可能会被细分为闲逛型用户、收藏型用户、比价型用户、购买型用户等。用户属性则更倾向于对属性层面的刻画和描述,特别是基本属性的内涵居多,包括性别、年龄、地域等。根据以上描述,对于视频推荐业务来说,我

们将遵循以下概念使用：用户画像近似等同于用户角色，统一称为中文概念的用户画像，而用户属性则是用户画像的子集。

用户画像的应用是非常广泛的，很多领域和行业都有用户画像这个概念，它在视频推荐领域也得到广泛应用。其中一个主要原因是，用户画像是一种能将定性与定量方法很好结合在一起的载体。定性化的方法，通过对用户的生活情境、使用场景、用户心智进行分析，来对用户的性质和特征做出抽象与概括；定量化可以对特征做精细的统计分析与计算，获得对于用户较为精准的认识，便于在数值排序的基础上实现核心用户的发掘与突出。

——摘自《用户网络行为画像：大数据中的用户网络行为画像分析与内容推荐应用》，牛温佳等，电子工业出版社

■ 参考资料 2：“学生社团网站文章个性化推荐”项目实施

推荐系统中使用的推荐方法有很多种，其中基于用户的协同过滤推荐算法是最早诞生的，原理也较为简单。该算法于 1992 年被提出，最初用于邮件过滤和新闻过滤，后来演变为推荐系统中最著名的算法。

基于用户的协同过滤推荐算法通过分析用户感兴趣的物品来计算用户之间的相似度，再根据相似度来进行物品的推荐。

俗话说：“物以类聚，人以群分。”一般来说，相似度高的用户，可能感兴趣的内容也是相似的。当一个用户 A 需要个性化推荐时，可以先找到和他兴趣相似的用户 B，然后把 B 喜欢的、并且 A 没有听说过的物品推荐给 A，这就是基于用户的协同过滤推荐算法。例如，你喜欢《蝙蝠侠》《星球大战》等科幻电影，另外有个人也喜欢这些科幻电影，而且他还喜欢《变形金刚》《钢铁侠》，但是你还没有看过《变形金刚》《钢铁侠》，那么很有可能你也喜欢《变形金刚》《钢铁侠》，可以把这两部电影推荐给你。

根据上述基本原理，可以将基于用户的协同过滤推荐算法拆分为两个步骤：

步骤 1 计算用户间相似度，找到与目标用户兴趣相似的用户集合。

步骤 2 根据用户相似度和已知的物品评分数据，计算用户对未评分物品的预测评分，找到用户喜欢的、并且没有听说过的物品推荐给目标用户。

下面，我们应用基于用户的协同过滤推荐算法来完成教科书中的项目活动。 r_{ik} 和 r_{jk} 分别表示用户 u_i 和用户 u_j 对物品 I_k 的评分， n 表示物品总数。

首先采用余弦相似度方法计算用户间的相似度，计算公式如下：

$$sim(u_i, u_j) = \frac{\sum_{k=1}^n (r_{ik} \times r_{jk})}{\sqrt{\sum_{k=1}^n (r_{ik})^2 \sum_{k=1}^n (r_{jk})^2}}.$$

本项目活动中，5 个用户分别表示为 a、b、c、d、e，文章 0001、0002、0003、0004、0005 分别表示为 N1、N2、N3、N4、N5。根据表 1.21 用户文章评分表中的数据，用户 a 和其他用户的相似度计算结果如表 1.24 所示。

表 1.24 用户 a 和其他用户的相似度

用户相似度	计算公式	计算结果
用户 a、b 相似度 $sim(a, b)$	$\frac{4 \times 1 + 5 \times 2}{\sqrt{(3^2 + 4^2 + 5^2)(2^2 + 1^2 + 2^2)}}$	0.6600
用户 a、c 相似度 $sim(a, c)$	$\frac{4 \times 3 + 5 \times 1}{\sqrt{(3^2 + 4^2 + 5^2)(1^2 + 3^2 + 1^2 + 4^2)}}$	0.4627
用户 a、d 相似度 $sim(a, d)$	$\frac{3 \times 3 + 4 \times 1}{\sqrt{(3^2 + 4^2 + 5^2)(4^2 + 3^2 + 1^2 + 1^2)}}$	0.3538
用户 a、e 相似度 $sim(a, e)$	$\frac{3 \times 4}{\sqrt{(3^2 + 4^2 + 5^2)(2^2 + 4^2 + 5^2)}}$	0.2530

然后,根据用户相似度和评分数数据计算预测评分。以用户 u_i 为例,其未评分的物品为 I_k ,已知对物品 I_k 进行了评分的用户集合为 NU_k ,用户 u_i 对未评分物品 I_k 的预测评分计算公式如下:

$$p_{ik} = \frac{\sum_{j \in NU_k} [sim(u_i, u_j) \times c_{jk}]}{\sum_{j \in NU_k} sim(u_i, u_j)}.$$

根据以上公式,计算用户 a 对文章 N1 的预测评分如表 1.25 所示。

表 1.25 用户 a 对文章 N1 的预测评分

用户 a 对文章 N1 的预测评分	p_{a1}								
已知用户对文章 N1 的评分	<table border="1"> <tr> <td>c_{b1}</td><td>c_{c1}</td><td>c_{d1}</td><td>c_{e1}</td></tr> <tr> <td>2</td><td>1</td><td>4</td><td>2</td></tr> </table>	c_{b1}	c_{c1}	c_{d1}	c_{e1}	2	1	4	2
c_{b1}	c_{c1}	c_{d1}	c_{e1}						
2	1	4	2						
计算公式	$\frac{sim(a, b) \times c_{b1} + sim(a, c) \times c_{c1} + sim(a, d) \times c_{d1} + sim(a, e) \times c_{e1}}{sim(a, b) + sim(a, c) + sim(a, d) + sim(a, e)}$								
计算结果	2.1416								

同样方法,计算用户 a 对文章 N5 的预测评分:

$$p_{a5} = \frac{sim(a, c) \times c_{c5} + sim(a, d) \times c_{d5} + sim(a, e) \times c_{e5}}{sim(a, c) + sim(a, d) + sim(a, e)} = 3.2441.$$

通过以上计算可知,用户 a 对文章 N5 的预测评分比对文章 N1 的预测评分高,因此系统将首先考虑推荐文章 0005 给用户 a。

下面再利用基于物品的协同过滤推荐算法进行文章推荐,看看推荐结果有什么不同。

基于物品的协同过滤推荐算法认为,物品 A 和物品 B 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 B,因此该算法可以通过计算物品的相似度为用户推荐那些与他们感兴趣的物品相似的物品。比如,在某在线购书网站上,用户 K 购买过《人工智能》这本书,而《人工智能》和《机器学习》这两本书具有很大的相似度,因为购买过《人工智能》的用户大多也购买过《机器学习》,因此网站会为用户 K 推荐《机器学习》这本书。需要注意的是,基于物品的协同过滤推荐算法并不是利用物品的内容属性计算物品之间的相似度,它主要通过分析用户的行为数据计算物品之间的相似度。

基于物品的协同过滤推荐算法主要分为两步:

第 1 步 计算物品之间的相似度;

第 2 步 根据物品的相似度和已知的物品评分数据,计算用户对未评分物品的预测评分,找到用户喜欢的、并且没有听说过的物品推荐给用户。

下面,我们应用基于物品的协同过滤推荐算法来完成教科书中的项目活动。 r_{ki} 和 r_{kj} 分别表示用户 u_k 对物品 I_i 和 I_j 的评分, n 表示用户总数。

首先采用余弦相似度方法计算物品间的相似度,计算公式如下:

$$sim(I_i, I_j) = \frac{\sum_{k=1}^n (r_{ki} \times r_{kj})}{\sqrt{\sum_{k=1}^n (r_{ki})^2 \sum_{k=1}^n (r_{kj})^2}}.$$

本项目活动中,5 个用户分别表示为 a、b、c、d、e,文章 0001、0002、0003、0004、0005 分别表示为 N1、N2、N3、N4、N5。根据表 1.21 用户文章评分表中的数据,文章 N1 和其他文章的相似度计算结果如表 1.26 所示。

表 1.26 文章 N1 和其他文章的相似度

文章相似度	计算公式	计算结果
文章 N1、N2 相似度 $sim(N1, N2)$	$\frac{4 \times 3 + 2 \times 4}{\sqrt{(2^2 + 1^2 + 4^2 + 2^2)(3^2 + 3^2 + 4^2)}}$	0.6860
文章 N1、N3 相似度 $sim(N1, N3)$	$\frac{2 \times 1 + 1 \times 3 + 4 \times 1}{\sqrt{(2^2 + 1^2 + 4^2 + 2^2)(4^2 + 1^2 + 3^2 + 1^2)}}$	0.3464
文章 N1、N4 相似度 $sim(N1, N4)$	$\frac{2 \times 2 + 1 \times 1}{\sqrt{(2^2 + 1^2 + 4^2 + 2^2)(5^2 + 2^2 + 1^2)}}$	0.1826
文章 N1、N5 相似度 $sim(N1, N5)$	$\frac{1 \times 4 + 4 \times 1 + 2 \times 5}{\sqrt{(2^2 + 1^2 + 4^2 + 2^2)(4^2 + 1^2 + 5^2)}}$	0.5555

然后,根据物品相似度和评分数数据计算预测评分。以用户 u_i 为例,已评分的物品集合为 NI_i ,用户 u_i 对未评分物品 I_k 的预测评分计算公式如下:

$$p_{ik} = \frac{\sum_{j \in NI_i} [sim(I_k, I_j) \times c_{ij}]}{\sum_{j \in NI_i} sim(I_k, I_j)}.$$

根据以上公式,计算用户 a 对文章 N1 的预测评分如表 1.27 所示。

表 1.27 用户 a 对文章 N1 的预测评分

用户 a 对文章 N1 的评分	p_{a1}						
已知用户 a 对文章 N2、N3、N4 的评分	<table border="1"> <tr> <td>c_{a2}</td><td>c_{a3}</td><td>c_{a4}</td></tr> <tr> <td>3</td><td>4</td><td>5</td></tr> </table>	c_{a2}	c_{a3}	c_{a4}	3	4	5
c_{a2}	c_{a3}	c_{a4}					
3	4	5					
计算公式	$\frac{sim(N1, N2) \times c_{a2} + sim(N1, N3) \times c_{a3} + sim(N1, N4) \times c_{a4}}{sim(N1, N2) + sim(N1, N3) + sim(N1, N4)}$						
计算结果	3.5857						

同样方法,计算用户 a 对文章 N5 的预测评分:

$$p_{a5} = \frac{sim(N2, N5) \times c_{a2} + sim(N3, N5) \times c_{a3} + sim(N4, N5) \times c_{a4}}{sim(N2, N5) + sim(N3, N5) + sim(N4, N5)} = 3.5522.$$

通过以上计算可知,用户 a 对文章 N5 的预测评分和对文章 N1 的预测评分基本一样高,因此系统将文章 0001 和文章 0005 都推荐给用户 a。

通过以上两种方法的实现,可以发现利用基于物品的协同过滤推荐算法计算用户对文章 0001 和 0005 的预测评分与利用基于用户的协同过滤推荐算法计算出的结果有所差异,这些差异可能会影响推荐结果。不同的数据分析方法可能会产生不同的结果,从而影响用户的决策,因此,需要根据需求选择合适的数据分析解决方案。

六、教学参考案例

■ 参考案例

数据管理与分析方案

上海市七宝中学 吴美琴
(2课时)

1. 学科核心素养

- 针对复杂的问题进行需求分析,综合判断信息,确定解决问题的路径。(信息意识)
- 针对较为复杂的问题,区分问题解决中涉及的各种数据,采用适当的数据表进行存储,采用合适的数据分析方法进行分析并呈现结果。(计算思维)

2. 《课程标准》要求

• 结合具体案例,初步了解分析业务需求、建立数据管理与分析问题整体解决方案的基本过程。

- 尝试对既定方案进行分析、评价,发现问题并优化方案。

3. 学业要求

• 学生能在特定的信息情境中,根据业务数据问题解决的需要,利用多种途径采集与甄别信息。

• 学生能够确定学习和生活中的业务数据问题,能提出解决方案,评价其合理性、完整性以及分析方案优化或改进的可能性。

- 学生能根据需要,主动选用数字化工具开展自主或协作学习,创造性地解决问题。

4. 教学内容分析

本节的核心内容是建立数据管理与分析方案的基本过程,具体包括 5 个环节:数据需求分析、数据管理、数据分析、方案评价和优化、科学决策。其中,数据需求分析、数据管理和数据分析是重点,通过 3 个项目活动掌握基本的实现方法。科学决策、方案评价和优化对学生的要求不高,可以通过学生自主阅读和教师讲解的方式完成教学。

5. 学情分析

授课对象为高中学生,学生通过前两节的学习已经感受到了数据价值,知道了数据管理与分析技术的重要性,有一定的分析问题、解决问题的能力。本节通过制定学生社团网站数据管理与分析方案,让学生初步了解建立数据管理与分析方案的基本过程。

6. 教学目标

- 了解数据需求分析方法,能结合实际问题进行数据需求分析。
- 了解建立数据管理与分析方案的基本过程,能结合实际问题制定数据管理与分析方案。
- 能对制定的方案进行评价,针对发现的问题进行方案优化。

7. 教学重难点

- 教学重点:数据管理方法;数据统计技术的应用。
- 教学难点:数据分析方法的应用。

8. 教学策略分析

这是本章的第三个任务,也是最复杂的一个,活动主题是制定学生社团网站的数据管理与分析方案,包括数据需求分析、数据管理、数据分析、方案评估和优化、科学决策,其中数据管理和数据分析是重点。本节的内容较难,可以将本节的项目任务分解成 3 个小的项目活动(即活动记录单中的项目活动 1、项目活动 2、项目活动 3),在每一个小的项目活动开始前,教师可以结合学生已有经验对相关理论进行简单讲解,消除学生对未知技术的陌生感。通过教师引导,以活动记录单为支架,一步一步引导学生完成项目活动。在实际教学过程中,可以根据学情让学生选择性完成项目活动。

9. 教学环境

计算机机房、Excel 软件。

10. 教学过程设计(见表 1.28)

表 1.28 教学过程设计表

教学环节	教学内容	学生活动	设计意图
课前准备	<p>发放教学资料(活动记录单),每组一份; 将文章数据表发送至学生机桌面</p> <p>前面我们感受了数据的价值和数据管理与分析技术的重要性,这节课我们继续深入数据中,学习如何制定综合性的解决方案。 (展示华东师范大学官方微博截图) 看看这张图,大家能看到什么数据? 在信息社会,大家可以通过社交媒体了解身边发生的事情,还可以进行评论和交流。我们的社团网站也有大量的数据,如何通过分析这些数据来更好地为我们服务呢? 首先来看两个问题: 我们把上次调研活动的新闻发到了学生社团网站上,许多同学进行了转发、评论。我们想了解:学生社团网站一周内发布的哪些文章关注度最高?如何让学生社团网站像在线购物网站一样给用户推送其可能感兴趣的内容,实现个性化推荐?</p>	浏览资料	初识任务,明确目标
情境导入		回顾生活经验,思考并回答问题	创设情境,引入项目主题——制定学生社团网站的数据管理与分析方案
教师讲授	教师讲解建立数据管理与分析方案的一般过程:数据需求分析、数据管理、数据分析、方案评估和优化、科学决策	聆听教师讲授	了解建立数据管理与分析方案的一般过程
项目活动 1	<p>(1) 讲解数据需求分析的目的和内容。 (2) 发布项目任务,并组织学生以小组为单位完成:根据活动记录单,完成问题 1 和问题 2 的数据需求分析。 (3) 请 1 个小组投影展示,分享数据需求分析的结果,其他小组进行补充</p>	根据项目活动 1 需要解决的问题,小组共同分析,完成活动记录单中的相应问题,并汇报	学生从问题出发,体验数据需求分析的过程
活动小结 1	总结数据需求分析的重点	归纳总结	了解数据需求分析方法
项目活动 2	<p>(1) 引导学生填写数据管理的一般流程,分析解决问题 1 和问题 2 所需的数据表:文章数据表和用户文章评分表。 (2) 小组合作分析文章数据表所包含的字段,根据评分标准对数据进行整理,得到用户文章评分表。 (3) 请 1 个小组投影展示填写的文章数据表和用户文章评分表,并说明计算过程,其他小组补充</p>	根据项目活动 2 需要解决的问题,在教师的引导下,小组合作,完成活动记录单中的相应问题,并汇报	针对具体问题,对采集到的数据进行整理,形成合适的数据表
活动小结 2	<p>思考:实际应用中,用户对文章的访问数据表的数量是非常大的,大家有没有更快的方法可以得到最终的结果? 归纳数据管理的要点:数据采集、数据存储与管理</p>	回顾数据处理工具,回答问题	了解大量数据的处理方法

教学环节	教学内容	学生活动	设计意图
项目活动 3	(1) 提问:大家知道哪些常见的数据分析方法和数据呈现方式? (2) 小组合作,根据活动记录单,对一周的数据进行分析,统计出每天学生关注度最高的文章,并用可视化方式呈现。 (3) 请 1 个小组投影展示结果,并说明分析过程,其他小组补充	根据项目活动 3 需要解决的问题,在教师的引导下,小组合作,完成活动记录单中的相应问题,并汇报	利用传统的数据统计技术解决实际问题
活动小结 3	提问:对分析结果进行观察,思考浏览量最高的文章是否也是评论量最高或转发量最高的文章,三者之间是否有联系 归纳常见的数据统计方法和可视化工具。 拓展:阅读教科书第 16 页的“探究活动”,参考介绍网站推荐机制的视频,查阅相关资料,了解推荐算法在日常生活中的应用	思考并回答问题,查阅资料	总结传统数据统计技术,拓展数据挖掘技术
教师总结	讲授:数据管理与分析方案评价的角度。 归纳:总结建立数据管理与分析方案的基本过程和每一个环节的核心要点,并用思维导图进行总结	画思维导图	总结建立数据管理与分析方案的基本过程

【活动记录单】

制定学生社团网站的数据管理与分析方案

一、项目主题

制定学生社团网站的数据管理与分析方案。

二、项目情境

调研活动结束后,我们把活动报道发布到学生社团网站,让其他同学了解这次有意义的活动。大家对我们的活动非常感兴趣,有许多同学都在活动文章的评论区里发表了评论,还有许多同学转发了这篇文章。同样,我们也可以在网站上查看很多其他社团开展的丰富多彩的活动。学生社团网站上发布了这么多的活动,哪个活动的文章浏览量最高?哪个活动大家讨论得最热烈?哪个活动的文章转发量最高?我们如何从这么多的活动中找到自己感兴趣的活动?我们可以通过制定学生社团网站的数据管理与分析方案来解决这些问题。

三、活动记录

项目活动 1:学生社团网站数据需求分析。

数据需求分析是建立数据管理与分析方案的第一步,即对拟解决的问题进行详细分析,具体包括:需要输入的数据、最终输出的结果、输出结果的方式。

问题 1:分别找到一周内每天浏览量最高、评论量最高、转发量最高的文章。

需要输入的数据:

最终输出的结果：

输出结果的方式：

问题 2：向学生推荐其可能感兴趣的文章。

需要输入的数据：

最终输出的结果：

输出结果的方式：

项目活动 2：学生社团网站数据管理。

数据管理的一般流程：

阅读教科书第 11~12 页的内容，梳理数据管理的一般流程，并填入表 1.29。

表 1.29 数据管理的一般流程

流程	具体内容	常用方法
数据采集		
数据存储与管理		

根据需求分析，解决问题 1 需要了解文章数据，形成文章数据表；解决问题 2 需要了解用户对文章的兴趣程度，形成用户文章评分表。

(1) 文章数据表

用数据库对文章数据进行管理，将从学生社团网站中采集到的文章保存在一张二维表中，根据需求分析思考文章数据表应包含哪些列并填入表 1.30。

表 1.30 文章数据表

1	2	3	4	5	6
文章编号	文章标题					

(2) 用户文章评分表

表 1.31 为用户 a~e 对文章 0001~0005 的行为(浏览、转发、评论、收藏、点赞)数据，“是”表示进行了某种行为，空缺表示没有该行为。

请根据表 1.31 计算出用户 a~e 对文章 0001~0005 的评分并填入表 1.32, 算法如下: 用户对某篇文章的初始评分为 0, 用户对该文章的每个行为各计 1 分, 用户对某篇文章的最高评分为 5 分。

表 1.31 用户对文章的访问数据表

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

表 1.32 用户文章评分表

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

项目活动 3:学生社团网站数据分析。

数据统计技术:从数据中抽取样本,通过统计方法对数据进行排序、筛选、汇总、统计等处理,得出有意义的结论。

数据挖掘技术:利用算法帮助人们从大量的数据中提取隐藏的、人们事先不知道但是又潜在有用的信息。例如关联挖掘算法、协同过滤算法。

(1) 统计文章关注度

根据表 1.31,利用传统的统计技术统计文章 0001~0005 的浏览量、评论量、转发量,并填入表 1.33。(可以用 Excel 制作)。

表 1.33 文章关注度数据表

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

(2) 统计一周(6月 4 日至 6月 10 日)的文章关注度

打开电脑桌面上的文章数据表,运用数据分析工具分别统计出一周内(6月 4 日至 6月 10 日)每天评论量、转发量最高的文章,并通过图表可视化方式展现(参见图 1.4、图 1.5)。

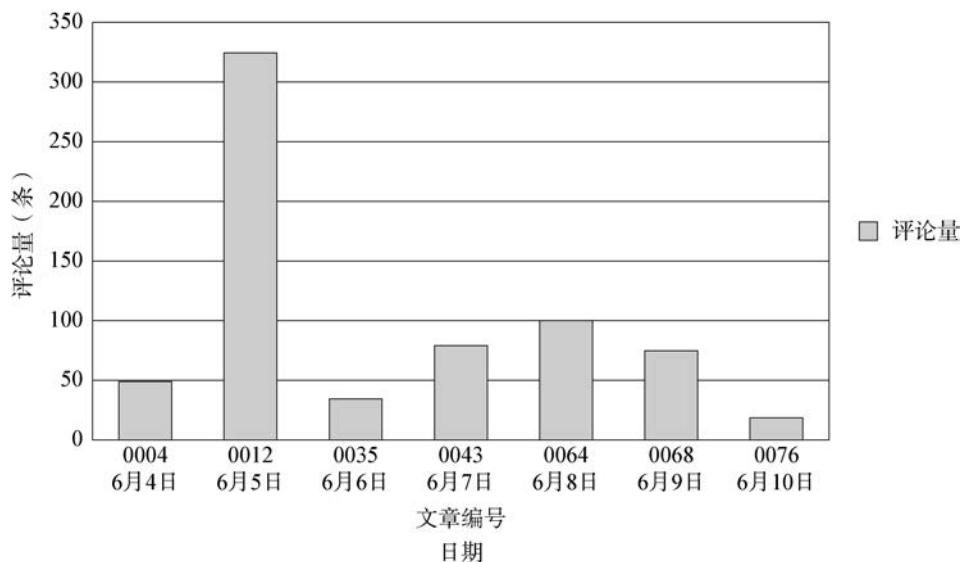
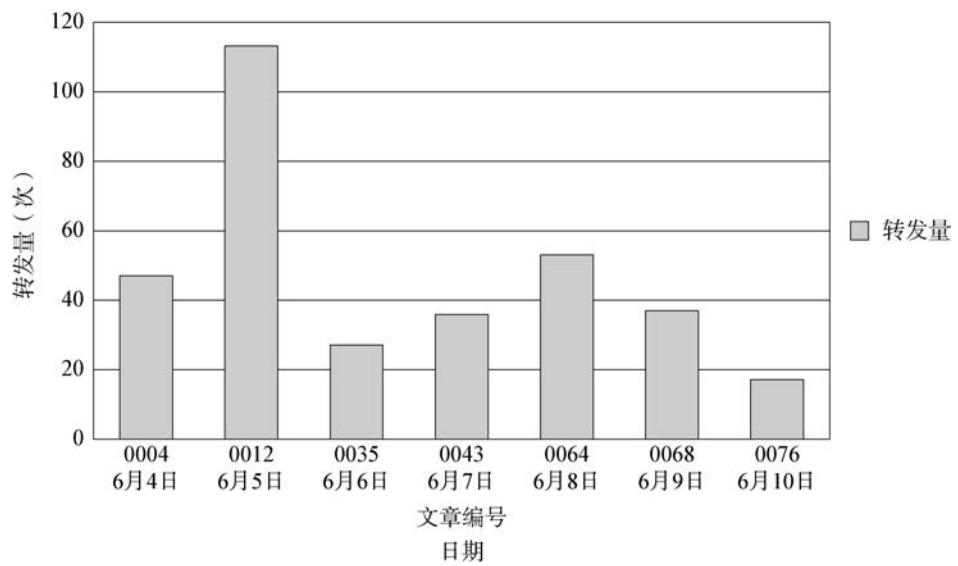


图 1.4 单日最高评论量对比图



数据管理

一、本章学科核心素养的渗透

1. 信息意识

选用生活中的数据管理项目案例,如网上购物管理、学生选课管理、学校运动会管理、学生用餐管理、社会实践活动管理等,让学生对项目中需要的数据进行分析,并通过合适的途径采集需要的数据。培养学生能在特定的信息情境中,根据业务数据问题解决的需要,利用多种途径采集与甄别数据。

2. 计算思维

引导学生对采集的数据进行分析、归类,并尝试使用数据库技术对数据进行管理,按照数据库设计的方法,设计出数据库的概念模型和数据模型。培养学生能对数据进行分类、抽象以及模型化处理。

选用一个具体的数据库管理系统,要求学生根据设计出的数据模型,建立关系数据库,并能使用SQL语言,对数据库进行创建、修改、删除、查询等操作。培养学生能按照特定数据管理的需求,使用数据库管理系统建立关系数据库,会选用恰当的策略与方法,对数据进行管理。随着我国在网络安全、自主可控方面一系列政策的出台,国产数据库的自主创新、助力科技强国已成为我国信息技术产业发展的重点。引导学生了解国产数据库及其发展现状,如人大金仓、南大通用、达梦、神舟通用等。

3. 数字化学习与创新

要培养学生的自主学习意识,为学生提供数字化学习的资源,如网络学习平台、微课、数字化实验平台、学习评价系统等,学生可以选择符合自身特点的方式进行学习,以满足不同层次学生的学习需求。培养学生能根据需要,主动选用合适的数字化平台、数字化学习工具和学习资源,开展自主或协作学习,提高学习质量。

引导学生进行自主学习、合作学习,创造性地解决实际问题。如对于数据采集,教师介绍常用的几种采集途径和采集案例,鼓励学生根据实际需求,探索其他的数据采集途

径。培养学生能协作解决学习中的实际问题,创新问题决策,反思与完善学习成果。

4. 信息社会责任

在数据采集过程中,要指导学生识别数据来源的可靠性,如通过官方网站、实地勘察等采集数据。培养学生能认识数据来源的准确性和可靠性的重要作用。

随着大数据时代的来临,数据的种类越来越多样化,数据管理的新技术不断出现。培养学生能积极学习数据管理的新理论和新技术。

二、本章知识结构

本章遵循普通高中信息技术课程标准,依据学分和课时规定,将内容分为三节,以“网上书店数据管理”为项目主题,围绕数据分类与采集、数据模型设计、数据库的实施展开设计。

第一节“数据分类与采集”,通过完成“网上书店数据采集”这一任务,使学生了解数据的分类、数据采集途径的多样性,了解使用网络爬虫采集数据的方法,并认识到采集到的数据中存在噪声数据的现象及其原因。

第二节“数据模型设计”,通过完成“网上书店数据库设计”这一任务,使学生掌握数据库设计的一般过程,了解数据库需求分析、概念模型、关系模型的基本概念,并能根据需求,设计简单的数据库概念模型和数据模型。

第三节“数据库的实施”,通过完成“网上书店数据库建立与查询”这一任务,使学生能使用一个具体的关系数据库管理系统,建立关系数据库,能对数据库中的数据进行增加、删除、修改、查询等基本操作,能使用结构化查询语言进行简单的数据查询。

三、本章项目活动设计思路

本章的项目活动要求学生设计一个简单的网上书店数据库,实现网上订书的数据管理。

建议学生首先可以从研究生活中的网上书店入手,如新华书店的“新华一城书集”网上书店,了解网上书店需要记录哪些数据,然后结合自己的需求,分析建立自己的网上书店需要哪些数据,并设计和创建自己的网上书店数据库。

项目任务1:网上书店数据采集。

网上书店数据库中需要记录图书、图书类别等数据,这些数据可以通过不同途径进行采集,如实体书店采集、网上书店采集等。在本项目任务中,主要为学生介绍使用网络爬虫在“新华一城书集”网上书店中采集图书相关数据的方法。

通过数据采集活动,让学生理解结构化数据、半结构化数据和非结构化数据的区别,了解采集到的数据中存在噪声数据的现象及其原因,并进一步学习其他的数据采集途径。

项目任务2:网上书店数据库设计。

在本项目任务中,学生将参照生活中的网上书店,着手设计自己的网上书店。按照数据库设计的一般步骤,首先需要对自己的网上书店数据库进行需求分析,然后建立数据库概念模型,再根据概念模型设计数据库关系模型。

建立数据库概念模型,要求学生能够分析出数据库中的实体、实体的属性和实体间的联系,可以让学生简单了解一下 E-R 图。数据模型主要介绍关系模型,要通过具体事例详细介绍关系模型中如何用二维表来表示实体间的三种联系。

在学生掌握了数据库设计的方法,并初步设计出自己的网上书店数据库后,教师可以指导学生在已有的数据库模型基础上进一步增加新的功能,修改和完善数据模型。

项目任务 3:网上书店数据库建立与数据查询。

在本项目任务中,学生使用 MySQL 关系数据库管理系统创建和使用网上书店数据库,并能使用 SQL 语句对数据库进行操作。

可以选用的关系数据库管理系统很多,教科书选用了开源的 MySQL 数据库管理系统作为学生实践操作的平台。

结构化查询语言(SQL)是关系数据库的标准语言,1987 年被国际标准化组织(ISO)采纳为国际标准,此后 SQL 语言又不断得到修改和完善。几乎所有的关系数据库管理系统都支持 SQL 语言。

要求学生能够使用 SQL 语言对数据库进行操作,在此基础上可以介绍一些使用图形操作界面操作数据库的方法。

四、本章课时安排建议

本章教学建议用 14 课时完成,具体参见表 2.1。

表 2.1 课时安排建议表

节名	建议课时
第一节 数据分类与采集	3 课时
第二节 数据模型设计	6 课时
第三节 数据库的实施	5 课时

第一节

数据分类与采集

一、教学目标与重点

教学目标:

- 理解不同结构化程度数据(结构化数据、半结构化数据和非结构化数据)的区别,以及它们在管理与应用上的特点。

- 了解数据采集途径的多样性,能利用适当的工具对数据进行采集和分类。能在特定的信息情境中,根据业务数据问题解决的需要,利用多种途径采集与甄别数据,能认识数据来源的准确性和可靠性的重要作用。
- 认识噪声数据现象和成因。
- 能根据需要,主动选用合适的数字化平台、数字化学习工具和学习资源,开展自主或协作学习,提高学习质量。
- 能协作解决学习中的实际问题,创新问题决策,反思与完善学习成果。

教学重点:

- 理解不同结构化程度数据(结构化数据、半结构化数据和非结构化数据)的区别。
- 了解数据采集途径的多样性,能利用适当的工具对数据进行采集和分类。

二、教学说明与建议

本节中的知识点“结构化数据、半结构化数据和非结构化数据”“数据采集”“噪声数据”都是“大数据”中的基础性知识,需要通过较多的实例才能让学生理解和掌握。

项目任务“网上书店数据采集”覆盖了本节中的大部分知识点,通过该项目任务让学生系统地理解本节知识。

本节3课时安排建议:“数据分类——结构化数据、半结构化数据和非结构化数据”1课时,“数据采集”1.5课时,“噪声数据”0.5课时。

建议通过分析网上书店中图书数据的格式使学生了解结构化数据、半结构化数据和非结构化数据。例如,网上书店中《现代汉语词典》的图书介绍如图2.1所示。图书的图



图2.1 《现代汉语词典》图书介绍

片属于非结构化数据;图书基本信息中包含半结构化数据;将《现代汉语词典》图书数据用二维表表示如表 2.2 所示,二维表中的数据属于结构化数据。

表 2.2 《现代汉语词典》图书数据

图书编号	书名	定价	作者	出版社	出版日期	折扣
1103018400	现代汉语词典(学生版·单色本)	35.00	商务国际辞书编辑部	商务印书馆国际有限公司	2017-07-01	0.55

数据采集的途径有很多,常用的包括人工采集、传感器采集、网络爬虫采集和数据库采集等,建议通过实例教学让学生理解。

完成数据采集后,紧接着需要考虑的问题是数据的质量是否满足要求,因此需要对数据质量进行分析。数据分析中的一项重要任务是检查是否有噪声数据,建议通过分析一些噪声数据的样例,让学生首先了解什么是噪声数据,然后再让学生对自己采集到的数据进行分析,去除噪声数据。

三、项目实施与评价

1. 核心概念精解

非结构化数据:非结构化数据是指不遵循统一的数据模型的数据。据估计,目前获得的数据有 80% 左右都是非结构化数据,并且其增长率要高于结构化数据。这种类型的数据可以是文本文件,也可以是二进制文件。文本文件如 Word 文件、PPT 文件等,二进制文件如图像、音频、视频等媒体文件。虽然文本文件和二进制文件都有根据文件格式本身定义的结构,但这属于文件的结构,与数据的结构含义是不同的。非结构化数据中“非结构化”的概念与包含在文件中的数据相关,与文件本身无关。

非结构化数据不能被直接处理或者用 SQL 语句查询,如果需要把它们存储在关系数据库中,只能以二进制大型对象(BLOB)的形式存储。随着文本、图形、图像、音频、视频等非结构化数据为主的信息急剧增加,如何存储、查询、分析、挖掘和利用这些海量信息资源就显得尤为关键。传统关系数据库擅长解决结构化数据管理问题,为了应对非结构化数据管理的挑战,出现了各种非结构化数据管理系统,NoSQL 数据库作为一个非关系数据库,能够用来同时存储结构化数据和非结构化数据。

半结构化数据:半结构化数据是结构化数据的一种形式,有一定的结构与一致性约束,但本质上不具有关系性,并不符合关系数据库或其他数据表的形式关联起来的数据模型结构,但包含相关标记,用来分隔语义元素以及对记录和字段进行分层。半结构化数据常常存储在文本文件中,常见的有 XML 数据、JSON 数据等。

2. 项目活动的具体实施

项目任务:网上书店数据采集,使用人工采集、网络爬虫采集等途径采集生活中的网上书店的图书数据。

活动 1: 使用人工采集, 在“新华一城书集”网上书店中采集你喜爱的图书的相关数据, 将数据填写在表 2.3 中(折扣 = 商城价 / 定价), 组织成结构化数据。

表 2.3 图书

图书编号	书名	定价	作者	出版社	出版日期	折扣
1103134637	大数据资源	80.00	朱扬勇	上海科学技术出版社	2018-01-01	0.8
1101046760	中国哲学史	72.00	冯友兰	华东师范大学出版社	2011-07-01	0.8
.....						

活动 2: 参照采集“计算机”类图书数据的方法, 使用网络爬虫采集其他类别图书的数据。

例如, 采集“新华一城书集”网上书店中“哲学”类图书的数据, 先查出“哲学”图书的类别编号为“5275”, 然后参考教科书上的爬虫程序编程采集数据。

3. 项目活动的评价

本节的项目评价主要包括: 考查学生对结构化数据、半结构化数据、非结构化数据、噪声数据等核心概念的掌握程度; 考查学生使用人工采集、网络爬虫采集等数据采集途径, 完成“网上书店数据采集”项目任务的情况。

评价建议: 采用纸笔测试等方式对学生掌握核心概念的程度进行终结性评价; 对学生参与任务的积极性、完成任务的情况进行过程性评价。

四、作业练习与提示

■ 题目描述

1. 请举出生活中结构化数据、半结构化数据和非结构化数据的例子。
2. 查阅 XML 数据和 JSON 数据的有关资料, 深入了解用这两种格式存储数据的方法。
3. 数据的采集途径有哪些? 请尝试使用不同的途径采集生活中的数据。
4. 什么是噪声数据? 请举例说明你在数据采集活动中遇到过哪些噪声数据。

■ 作业提示

1. (1) 结构化数据: 以二维表形式登记的学生名单、学生成绩等。
(2) 半结构化数据: 以 XML 和 JSON 格式存储的数据。可以参照教科书, 写一个以 XML 格式存储数据的实例。
(3) 非结构化数据: 学生的照片、教师录制的微客视频、PPT 文件、Word 文件等。
2. (1) XML 和 JSON 的有关资料参考本节的“教学参考资源”。
(2) 借助 XML、JSON“在线格式化工具”, 对用这两种格式存储的数据进行验证。
3. (1) 人工采集。发放问卷, 调查同学的兴趣爱好、特长等。
(2) 传感器采集。在物理 DIS 实验中, 使用传感器采集实验数据。

- (3) 网络爬虫采集。编写一个网络爬虫程序,抓取图书相关网站公开的图书数据。
 - (4) 数据库采集。从数据库中提取需要的数据。
4. 在采集到的数据中,一些不符合要求的、无意义的、错误或异常的数据通常称为噪声数据。例如:采集到的气温数据中超过正常值的数据;数据中没有意义的特殊符号;等等。

五、教学参考资源

■ 参考资料 1:XML

可扩展标记语言(eXtensible Markup Language,缩写为 XML)也是一种标记语言,类似 HTML,它被设计的宗旨是存储数据,而非显示数据。1998 年 2 月,万维网联盟(World Wide Web Consortium,缩写为 W3C)正式批准了可扩展标记语言的标准定义。

XML 标签没有被预定义,需要用户自行定义标签。XML 标签的书写形式有:

任何一个起始标签对应有一个结束标签,例如:<标签名></标签名>。

一个标签同时表示起始和结束标签,其语法是在大于符号之前紧跟一个斜杠(/),例如:<标签名/>。

一个标签中可以嵌套若干子标签,但所有标签必须合理嵌套,不允许有交叉嵌套。

■ 参考资料 2:JSON

JSON(JavaScript Object Notation)是一种轻量级的数据交换格式,用完全独立于编程语言的文本格式来存储和表示数据。

JSON 常见的语法格式:

表示对象:对象是一个无序的“键/值”对集合。一个对象以“{”开始,以“}”结束。每个“键”后面跟一个冒号,“键/值”对之间用逗号分割。例如:

```
{"name": "wangchen", "sex": "female"}
```

表示数组:数组是值的有序集合,一个数组以“[”开始,以“]”结束。值之间用逗号分割。例如:

```
["wangchen", "wukun"]
```

表示对象集合(对象和数组的合成),例如:

```
[{"name": "wangchen", "sex": "female"}, {"name": "wukun", "sex": "female"}]
```

■ 参考资料 3:网络爬虫

网络爬虫的实质是能获取网页并提取和保存数据的自动化程序。如果把互联网比作一张大网,那么网的节点就是一个个网页,爬虫爬到一个节点访问该页面并获取其信息,再通过这个网页继续获取其他网页,这样整个网的节点便可以被全部访问到,数据就可以被抓取下来了。

网络爬虫首先要做的是获取网页,即获取网页的源代码,源代码里面包含着有用的信息,我们可以从中提取需要的信息。Python 提供了许多库来帮助我们进行获取网页的操作,如 Urllib、Requests 等,可以利用这些库来帮助我们获得网页的源代码,实现用程序

来获取网页。

获取了网页源代码之后,接下来就是分析网页源代码,从中提取我们想要的数据。通用的方法是采用正则表达式提取。另外 Python 还提供了许多解析库,如 LXML、BeautifulSoup、PyQuery 等,使用这些库可以高效快速地提取网页信息,如节点的属性、文本值等内容。

提取信息后,需要对数据进行存储,存储的形式可以多种多样,最简单的是直接存储为文本文件,如 TXT、JSON 等,另外还可以存储到数据库中,如关系数据库 MySQL、非关系数据库 MongoDB 等。

■ 参考资料 4: 参考书

1. 张良均,王路,谭立云,苏剑林,等. Python 数据分析与挖掘实战[M]. 北京:机械工业出版社,2015.
2. 娄岩. 大数据技术概论[M]. 北京:清华大学出版社,2017.
3. 崔庆才. Python 3 网络爬虫开发实战[M]. 北京:人民邮电出版社,2018.

六、教学参考案例

■ 参考案例

数据分类

上海市南洋中学 陈敏

(1 课时)

1. 学科核心素养

能对数据进行分类、抽象以及模型化处理。(计算思维)

2.《课程标准》要求

- 结合案例,了解数据采集途径的多样性,能利用适当的工具对数据进行采集和分类。
- 理解不同结构化程度数据(包括结构化数据、半结构化数据和非结构化数据)的区别,以及在管理与应用上的特点。

3. 学业要求

能在特定的信息情境中,根据业务数据问题解决的需要,利用多种途径采集与甄别数据。

4. 教学内容分析

本节课教学内容为教科书第二章第一节中部分内容。采用项目学习的方式,通过对“志愿者服务管理”项目进行分析,引导学生归纳与提炼数据库中所需的数据,能够区分不同结构化程度数据(包括结构化数据、半结构化数据和非结构化数据),并且说明这三类数据在管理和应用上的特点。

5. 学情分析

本节课的学习主体是高中学生,学生逻辑思维能力趋于严密,已具备一定的探究并解

决问题的能力。在本节课之前,学生已完成第一章的学习,了解了数据的价值,认识了数据管理与分析技术的重要性。

本节课重点介绍三类不同结构化程度的数据,对学生来说这些知识和他们的生活距离比较远,理解起来会比较困难。

6. 教学目标

区分不同结构化程度的数据(结构化数据、半结构化数据和非结构化数据),说明这三类数据在管理与应用上的特点。

7. 教学重难点

• 教学重点:区分不同结构化程度的数据(结构化数据、半结构化数据和非结构化数据)。

- 教学难点:比较半结构化数据和结构化数据的特点。

8. 教学准备

设计教案、活动记录单、PPT,安排学生分组。

9. 教学策略分析

将学生熟悉的志愿者服务活动作为项目情境,引导学生从项目情境中提炼数据。以“网上书店”为例进行讲授,启发学生认识并能区分“志愿者服务管理”中的三类不同结构化程度的数据。

以小组合作探究形式,完成本节课的各项活动与体验。

10. 教学过程设计(见表 2. 4)

表 2. 4 教学过程设计表

教学环节	教学内容	学生活动	设计意图
情境导入	“志愿者服务管理”项目情境导入		引入项目活动,明确最终目标:完成“志愿者服务管理”数据库
活动 1	引导学生思考为开发“志愿者服务管理”数据库,需要采集哪些数据(如服务时长、评价等)	活动 1: 结合自己志愿者服务的经历,以小组为单位进行讨论,并进行交流; 完成活动记录单中的任务 1	对项目情境进行需求分析,归纳提炼数据
教师讲授	根据数据结构化程度的不同,可以将数据分为结构化数据、半结构化数据和非结构化数据		
活动 2	举例:网上书店中的《现代汉语词典(学生版·单色本)》图书数据(如表 2. 2 所示)。 说明:以二维表的形式表示,用关系数据库进行存储和管理	活动 2: 小组合作完成活动记录单中的任务 2; 小组间交流分享	通过教师举例和学生活动,引导学生了解结构化数据在管理与应用上的特点

续表

教学环节	教学内容	学生活动	设计意图
活动 3	<p>举例:</p> <p>1. 用 XML 格式存储图书数据</p> <pre><books> <bookID> 1103028726 </bookID> <title> 数据库技术 *** </title> <author> 朱 **、张 ** </author> <publisher> ** 教育出版社 </publisher> <price> 36.00</price> <pubdate> 2017-08-01 </pubdate> <discount> 0.8</discount> <comment> 这本书适合初学者 </comment> <recommend> ★★★★★ </recommend> </books></pre> <p>2. 在如上使用 XML 格式存储的数据中增加图书的简介,可以在标记<code><books></code>和<code></books></code>之间的任意位置增加如下内容: <code><abstract>数据库技术原理与设计分为三个部分:(一)基础原理;(二)方法与设计;(三)问题求解.....</abstract></code> 说明:有一定的结构,但又不符合关系数据库或数据表的表示形式</p>	<p>活动 3:</p> <p>1. 完成活动记录单中的任务 3 (如:个人荣誉最多填 5 项;有同学没有个人荣誉,浪费存储空间……)</p> <p>2. 使用半结构化数据进行改进 操作体验:完成活动记录单中的任务 4; 小组间交流分享</p>	通过教师举例和学生活动,引导学生了解半结构化数据在管理与应用上的特点
活动 4	<p>举例:网上书店图书的图片、音频等数据。</p> <p>说明:没有固定结构的数据,使用非关系数据库 NoSQL 进行存储和管理</p>	<p>活动 4:</p> <p>讨论交流:志愿者服务过程中需要采集哪些非结构化数据?</p> <p>完成活动记录单中的任务 5; 小组间交流分享</p>	通过教师举例和学生活动,引导学生了解非结构化数据在管理与应用上的特点
课堂小结	<p>结构化数据:以二维表的形式表示,用关系数据库进行存储和管理。</p> <p>非结构化数据:没有固定结构的数据,使用非关系数据库 NoSQL 进行存储和管理。</p> <p>半结构化数据:有一定的结构,但又不符合关系数据库或数据表的表示形式</p>		
课后思考	“志愿者服务管理”数据库中的数据,可以通过哪些途径进行采集?		为下节课数据采集做准备

附：“志愿者服务管理”项目活动指南

1. 项目主题

志愿者服务管理。

2. 项目情境

2015年上海市推出了《上海市普通高中学生志愿服务(公益劳动)管理工作的实施意见》(以下简称《实施意见》),其中规定每个学生参加校外实践活动的时间应不少于60学时。

《实施意见》颁布后,上海市教委、市青少年学生校外活动联席会议办公室广泛发动全市各场馆、基地通力合作,积极推动高中学生志愿服务、公益劳动工作。一方面,努力推动学生积极参加志愿服务,增强学生社会责任感。另一方面,完善了学生社会实践基地的管理,积极为高中生开展服务提供岗位,为学生的社会实践搭建更多平台。

校团委学生处为了更好地对同学们的志愿服务数据进行管理,希望能够快速地统计出每个同学的志愿服务经历和服务时长,查询出每个基地岗位设置及对志愿者的需求,并且能够通过相关经历作出综合评价。通过充分讨论后,学生会的同学准备开发一个“志愿者服务管理”数据库。

为了完成这个任务,同学们首先需要收集志愿者服务的相关数据,分析数据构成;然后设计一个简单的“志愿者服务管理”数据库,用以管理志愿服务过程中产生的数据;最后选择一个数据库管理系统来创建和使用这个“志愿者服务管理”数据库。

【活动记录单】

数据分类——志愿者服务管理

一、讨论

任务1:在“志愿者服务管理”数据库中,需要用到哪些数据?

二、结构化数据

任务 2:设计“志愿者报名信息汇总表”,将图 2.2 中的相关数据以二维表的形式进行整理汇总。

--

博物馆志愿者服务报名表

报名号	1001
姓名	李莉
班级	高一(3)班
个人荣誉	
1	校三好学生
2	**美术馆优秀志愿者
3	**基地优秀营员
4	**杯作文竞赛一等奖
5	**绘画比赛二等奖

博物馆志愿者服务报名表

报名号	1002
姓名	王志
班级	高二(4)班
个人荣誉	
1	校优秀团员
2	**美术馆优秀志愿者
3	
4	
5	

博物馆志愿者服务报名表

报名号	1003
姓名	赵力
班级	高一(7)班
个人荣誉	
1	
2	
3	
4	
5	

博物馆志愿者服务报名表

报名号	1004
姓名	钱小峰
班级	高一(8)班
个人荣誉	
1	
2	
3	
4	
5	

图 2.2 志愿者报名表

三、半结构化数据

任务 3:讨论如表 2.5 所示的二维表在设计上存在的问题。

表 2.5 志愿者报名信息汇总表

报名号	姓名	班级	个人荣誉 1	个人荣誉 2	个人荣誉 3	个人荣誉 4	个人荣誉 5
1001	李莉	高一(3)班	校三好学生	** 美术馆优秀志愿者	** 基地优秀营员	** 杯作文竞赛一等奖	** 绘画比赛二等奖
1002	王志	高二(4)班	校优秀团员	** 美术馆优秀志愿者			
1003	赵力	高一(7)班					
1004	钱小峰	高一(8)班					

任务 4:在如图 2.3 所示的使用 XML 格式存储的数据中增加如表 2.6 所示的志愿者报名信息。

```
<volunteers>
  <volunteer>
    <number> 1001</number>
    <name> 李莉</name>
    <class> 高一(3)班</class>
    <honour1> 校三好学生</honour1>
    <honour2> ** 美术馆优秀志愿者</honour2>
    <honour3> ** 基地优秀营员</honour3>
    <honour4> ** 杯作文竞赛一等奖</honour4>
    <honour5> ** 绘画比赛二等奖</honour5>
  </volunteer>
  <volunteer>
    <number> 1002</number>
    <name> 王志</name>
    <class> 高二(4)班</class>
    <honour1> 校优秀团员</honour1>
    <honour2> ** 美术馆优秀志愿者</honour2>
  </volunteer>
  <volunteer>
    <number> 1003</number>
    <name> 赵力</name>
    <class> 高一(7)班</class>
  </volunteer>
  <volunteer>
    <number> 1004</number>
    <name> 钱小峰</name>
    <class> 高一(8)班</class>
  </volunteer>
</volunteers>
```

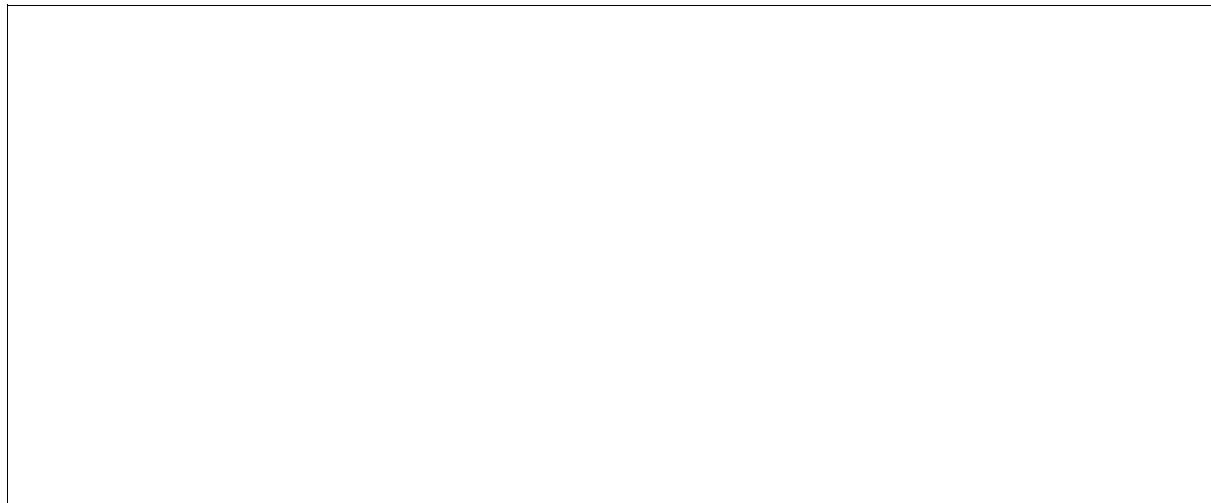
表 2.6 博物馆志愿者服务报名表

报名号	1005
姓名	孙琳
班级	高二(1)班
个人荣誉	
1	校优秀学生干部
2	校史博物馆优秀讲解员
3	
4	
5	

图 2.3 使用 XML 格式存储志愿者报名信息

四、非结构化数据

任务 5:志愿者服务过程中需要采集哪些非结构化数据? (如图片、视频、音频等数据)



第二节 数据模型设计

一、教学目标与重点

教学目标:

- 知道数据库设计的一般过程。
- 了解概念模型的基本概念;理解实体、属性、联系、关键字等概念;理解实体间联系的三种类型。
- 了解关系模型的基本概念;掌握建立关系模型的方法,学会用二维表表示实体与实体间的联系。
- 能根据数据管理的主题归纳出牵涉的事物和联系,体会从现实世界的事物及其联系抽象到实体和联系,进一步抽象到数据模型的过程。

教学重点:

- 理解实体间联系的三种类型。
- 掌握建立关系模型的方法,学会用二维表表示实体与实体间的联系。

二、教学说明与建议

本节中的知识点,包括实体、实体的属性、实体间的联系、关键字、关系模型等都是重要的基本概念,需要通过较多的实例让学生理解和掌握。

本节6课时安排建议：“数据库设计的一般过程”1课时、“需求分析”1课时、“概念设计”2课时、“逻辑设计”2课时。

在“一、数据库设计的一般过程”的教学中，简单介绍数据库设计的四个阶段，让学生对项目任务“网上书店数据库设计”的过程有个清晰的了解。在“二、需求分析”的教学中，让学生认真考察生活中的网上书店，了解网上书店提供了哪些功能，登记了哪些数据，鼓励学生对生活中网上书店登记的数据进行较全面的分析并做好记录。在“三、概念设计”的教学中，结合多个实例，让学生深刻理解实体、实体的属性和实体间的联系等基本概念。E-R图作为概念模型的一种表示方式，可以让学生简单了解，不需要用很多的时间去介绍E-R图的画法。“四、逻辑设计”部分是本节教学的重点，同时建立关系模型也是教学中的一个难点。建议可以先通过一些实例，让学生了解二维表之间的关系。突破了“二维表之间的关系”这个难点，学生再通过项目活动学习建立关系模型也会变得比较容易。

三、项目实施与评价

1. 核心概念精解

(1) 数据模型

在传统数据库领域中较常见的数据模型有：层次模型、网状模型、关系模型。目前使用最普遍的是关系模型。

层次模型是最早使用的一种数据模型，它通过链接的方式将相互关联的记录组织起来，形成一种层次关系。层次模型中表示实体及实体之间联系的数据结构是一种树型结构，树中的节点表示实体，父节点与子节点之间的关系表示实体之间的联系。

网状模型也是早期经常使用的一种数据模型，网状模型采用网状结构(图结构)表示实体和实体之间的联系，图中的节点记录表示实体，节点之间的关系表示实体之间的联系。

(2) 实体间的联系

① 一对联系(1:1)

例如：一位班主任只负责一个班级，一个班级只有一位班主任，班主任与班级之间是一对一联系。(如图2.4所示)

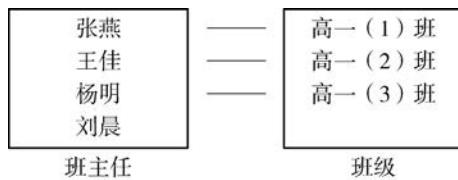


图2.4 建立“班主任”与“班级”之间的联系

② 一对多联系(1:N)

例如：一种图书类别包含有多本图书，一本图书属于一种图书类别，两者间属于“一对

多”联系。图书类别是“一方”,图书是“多方”。(如图 2.5 所示)

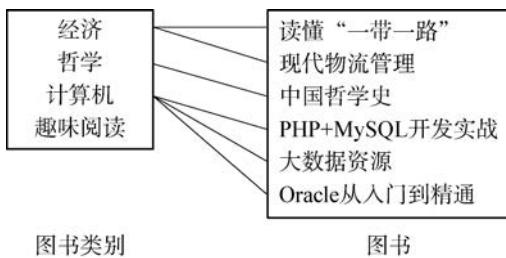


图 2.5 建立“图书类别”与“图书”之间的联系

③ 多对多联系(M : N)

例如:一张订单中可以订购多本图书,一本图书可以由多张订单订购,两者间属于“多对多”联系。(如图 2.6 所示)

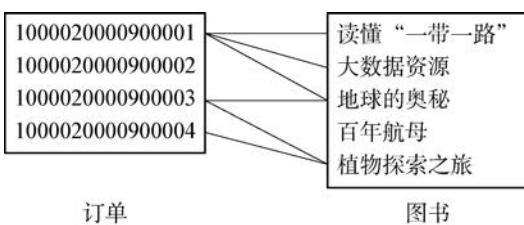


图 2.6 建立“订单”与“图书”之间的联系

(3) 数据库设计

数据库设计是一项非常复杂的工作,有着严格的理论基础,需要按照规范化的理论进行设计。按照规范化设计的方法,考虑数据库及其应用系统开发全过程,数据库设计分为以下六个阶段:需求分析阶段、概念设计阶段、逻辑设计阶段、物理设计阶段、数据库实施阶段、数据库运行和维护阶段。

概念设计阶段将建立数据库的概念模型。概念模型设计实质上就是要找出实体、实体的属性和实体间的联系。概念模型设计常用的方法是实体-联系方法,简称 E-R 方法。该方法直接从现实世界中抽象出实体和实体间的联系,然后用 E-R 图来表示。

2. 项目活动的具体实施

项目任务:网上书店数据库设计。对网上书店数据库进行需求分析,确定实体与实体间的联系类型,并建立关系模型,设计出一个简单的网上书店数据库。

(1) 活动 1:参考生活中的网上书店“新华一城书集”网,对构建自己的网上书店数据库进行数据需求分析

① 对“新华一城书集”网上书店中的数据进行分析,了解该网站登记了哪些数据

以“用户”数据为例,如图 2.7 所示,在“新华一城书集”网上书店的“用户设置”页面中,用户登记的数据有很多,如账户信息、账户安全、收货地址等。其中,账户信息中的基本信息又包括用户名、邮箱、真实姓名、性别、生日、所在地区等。



图 2.7 “新华一城书集”网上书店的“用户设置”页面

② 对构建自己的网上书店数据库进行数据需求分析

以“用户”数据为例，在自己构建的网上书店数据库中，同样也需要登记用户的数据，但不需要像“新华一城书集”网上书店那样登记得非常详细。可以根据自己的需求，确定需要记录用户的哪些数据。例如，需要登记的“用户”数据有：用户名、密码、邮箱、姓名、性别、生日、地址等。

(2) 活动 2：根据设计自己的网上书店数据库的需求分析，确定数据库中的实体与实体间的联系类型

① 分析网上书店中的实体及实体的属性

“用户”的属性：用户名、密码、邮箱、姓名、性别、生日、地址等。其中“用户名”具有唯一性，在注册时不允许重复，也不能为空，属于关键字。

“图书”的属性：图书编号、书名、定价、作者、出版社、出版日期、折扣等。其中“图书编号”具有唯一性，可以作为关键字。

订单中的订单号、下单时间、交易状态可以作为“订单”的属性，其中“订单号”是关键字。订单中列出的图书明细，如书名等信息应该记录在实体“图书”的属性中。

“图书类别”的属性：类别编号、类别名称等。例如，“01”表示“计算机”，“02”表示“哲学”。

网上书店中的实体及实体的属性如表 2.7 所示。

表 2.7 实体及其属性

实体的名称	实体的属性	关键字
用户	用户名、密码、邮箱、姓名、性别、生日、地址	用户名
订单	订单号、下单时间、交易状态	订单号
图书	图书编号、书名、定价、作者、出版社、出版日期、折扣	图书编号
图书类别	类别编号、类别名称	类别编号

② 确定实体间的联系类型

根据教科书中对网上书店中的用户、订单、图书、图书类别四个实体之间联系的分析，这些实体间的联系类型如表 2.8 所示。

表 2.8 实体间的联系类型

实体 A	实体 B	联系类型
用户	订单	一对多
订单	图书	多对多
图书类别	图书	一对多

(3) 活动 3: 建立网上书店关系数据模型(如图 2.8 所示), 即建立“用户”“订单”“图书”“图书类别”表及表之间的关系

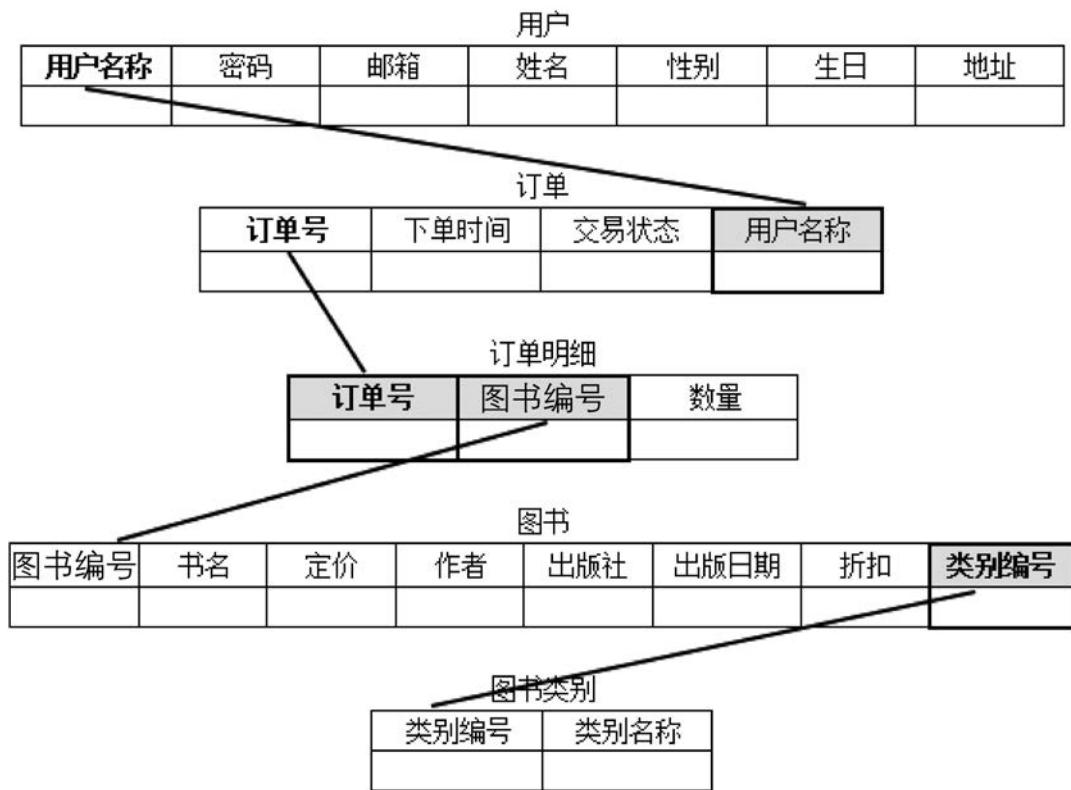


图 2.8 建立二维表及表之间的关系

“用户”表与“订单”表之间具有一对多关系,因此可以在“订单”表中加入“用户”表的关键字“用户名”,两张表之间通过公共属性“用户名”建立关系。

“订单”表与“图书”表之间具有多对多关系,因此,建立一张新的二维表“订单明细”,“订单”表与“图书”表通过“订单明细”表建立起关系。在“订单明细”表中加入“订单”表的关键字“订单号”和“图书”表的关键字“图书编号”。

“图书类别”表与“图书”表之间具有一对多关系,因此可以在“图书”表中加入“图书类别”表的关键字“类别编号”,两张表之间通过公共属性“类别编号”建立关系。

(4) 活动 4: 网上书店中还有哪些实体? 请分析这些实体的属性及实体间的联系类型

购物网站通常会记录客户购买或浏览的商品,并通过对这些数据的分析,为客户推送可能需要的商品。

在网上书店数据库中,记录用户浏览图书的信息,登记用户浏览图书介绍的开始时间和结束时间,可以这样实现:在“用户”表与“图书”表之间建立多对多关系,如图 2.9 所示,建立一张新的二维表“图书浏览”表,记录用户浏览图书的信息。



图 2.9 建立“用户”与“图书”表之间的关系

说明:为了实现网上书店的更多功能,可以根据需要再设计更多的实体。

3. 项目活动的评价

本节的项目评价主要包括:考查学生对概念模型、关系模型中的基本概念的掌握程度;考查学生在“网上书店数据库设计”项目任务中,完成数据库的需求分析、概念模型设计、关系模型设计的情况。

评价建议:设计相应的评价量表对学生掌握基本概念的程度进行过程性评价;针对数据库设计一般过程中的三个步骤——需求分析、概念设计、逻辑设计,分别设计过程性评价量表,对学生的活动表现、作业完成质量等方面进行评价。

四、作业练习与提示

■ 题目描述

1. 关系模型中用来管理和组织数据的方式是()。
A. 网状图 B. E-R 图 C. 二维表 D. 树状图
2. 在概念模型中,把具有共同特性和行为的对象称为()。
A. 个体 B. 实体 C. 事物 D. 记录
3. 在概念模型中,描述事物的某一特征称为()。
A. 特点 B. 实体 C. 字段 D. 属性
4. 如图 2.10 所示的实体联系类型属于()。
A. 多对多 B. 一对多 C. 多对一 D. 无法确定
5. 将数据库概念模型转换成数据模型,实体与实体间的联系可以转换成()。
A. 二维表 B. 关系图 C. 记录 D. 线条
6. 学校要举行校园运动会,需要开发一个运动会管理系统来管理比赛,该管理系统中的数据需求分析如下:

- (1) 登记参赛运动员的信息。记录运动员的号码、姓名、性别、班级等信息。
- (2) 登记比赛项目的信息。记录项目编号、项目名称、比赛时间、比赛地点等信息。
- (3) 一名运动员可以参加多个比赛项目,比赛结束后需要登记运动员的最终比赛成绩。

请根据以上需求分析,完成:

- (1) 设计概念模型。
- (2) 用三张数据表实现上述概念模型,要求既能反映实体及其属性,又能反映实体间的联系。
- (3) 需要记录裁判员的裁判编号、姓名、性别、联系电话、担任裁判的比赛项目等,且一名裁判员担任一个比赛项目的裁判,一个比赛项目可能需要多名裁判员。根据这一需求,修改概念模型。

■ 作业提示

1. C
2. B
3. D
4. A
5. A
6. (1) 如图 2.11 所示,用 E-R 图表示概念模型。
(2) 运动员 (运动员号码,姓名,性别,班级)

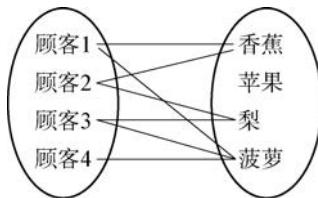


图 2.10

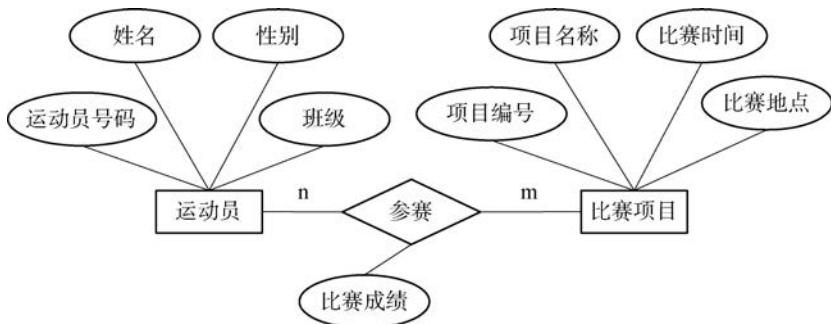


图 2.11 运动会管理数据库 E-R 图

比赛项目(项目编号,项目名称,比赛时间,比赛地点)

参赛表 (运动员号码,项目编号,比赛成绩)

(3) 如图 2.12 所示。

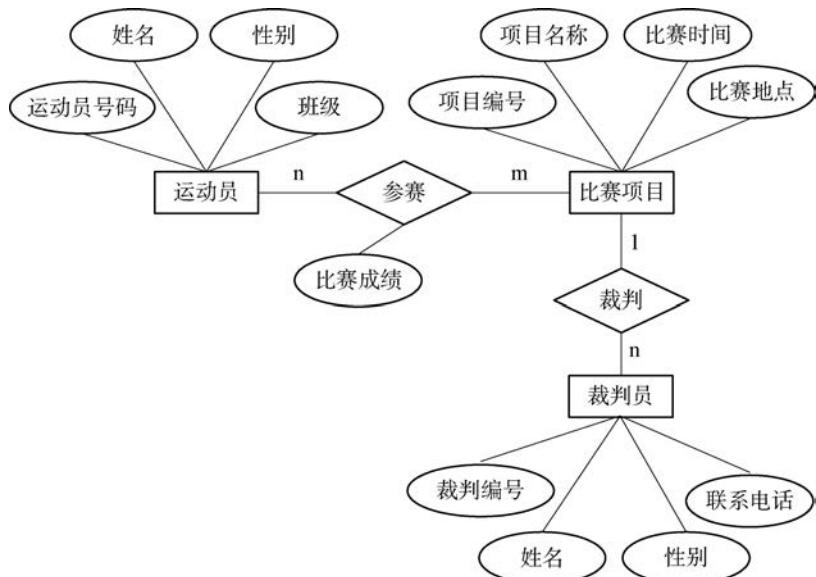


图 2.12 修改后的运动会管理数据库 E-R 图

五、教学参考资源

■ 参考资料 1:E-R 图

E-R 图,即实体-联系模型图,是将实体、属性、联系用图形的方式描述,使之更为直观。在 E-R 图中,用矩形框表示实体,用椭圆框表示属性,并用连线将实体和属性连接起来。实体间的联系用菱形框表示,并用连线和实体连接起来,在连线上注明联系类型。对于有些联系,其自身也会有某些属性,可以将这些属性与联系连接起来。如图 2.13 所示的 E-R 图中,实体是“供应商”,“供应商编号”“供应商名称”“联系人”“联系电话”是实体“供应商”的属性。

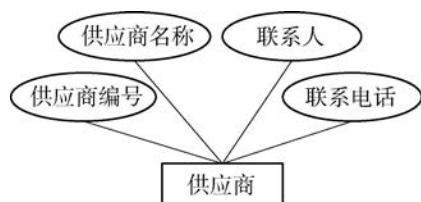


图 2.13 “供应商”E-R 图

教科书中的网上书店数据库的 E-R 图如图 2.14 所示：

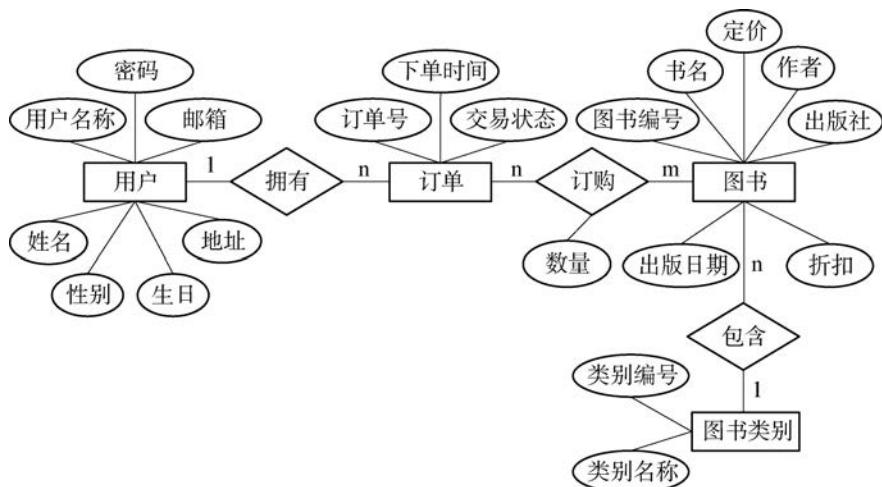


图 2.14 网上书店数据库 E-R 图

■ 参考资料 2：关系数据库规范化理论

数据库设计是数据库应用领域中的主要研究课题。数据库设计问题可以简单地理解为如果要把数据存储到数据库中,如何设计一个“好”的数据库结构、建立数据库及应用系统,使之能有效地存储和管理数据。对于关系数据库而言,关键就是如何设计一些“好”的数据表。

观察如表 2.9 所示的“学生选课”表,从该表中我们可以看到如下问题:姓名、性别、班级、班主任以及课程名称等数据存在重复存储的情况,造成数据冗余问题;如果课程名称“摄影”需要修改,则必须在多条记录中同时修改,否则会造成数据不一致。产生以上问题的原因是这个数据表设计得“不好”。

表 2.9 学生选课

选课编号	学号	姓名	性别	班级	班主任	课程号	课程名称	成绩
1	101	罗阳	女	高一(1)班	张慧	11	健美操	85
2	101	罗阳	女	高一(1)班	张慧	21	摄影	90
3	102	范英	女	高一(1)班	张慧	11	健美操	88
4	201	李丽	女	高一(2)班	王玲	11	健美操	85
5	201	李丽	女	高一(2)班	王玲	21	摄影	85
6	212	范英	女	高一(2)班	王玲	21	摄影	90

20 世纪 70 年代初,计算机科学家提出了关系数据库规范化理论,目的就是要设计“好的”关系数据库结构。规范化理论最初提出了三种范式,分别为第一范式(1NF)、第二范式(2NF)和第三范式(3NF)。随后又提出了一种更强的第三范式,被称为巴斯-科德范式(BCNF)。此后,又分别基于多值依赖和连接依赖陆续提出了第四范式(4NF)和第五

范式(5NF)。这几种范式之间的关系是:5NF \subset 4NF \subset BCNF \subset 3NF \subset 2NF \subset 1NF。通常可以通过判断数据表达到第几范式来评价规范化的程度,范式越高,规范化的程度越高,数据表就设计得越好。

对于规范化程度不高的数据表,可以通过分解,将一个低一级范式的数据表转换成若干个高一级范式的数据表,这就是规范化的过程。

1. 第一范式(1NF)

如果数据表中的每一列都不可再分,即已经分到最小,那么该数据表属于第一范式。

如图 2.15 所示,关系模式 A 不符合第一范式,关系模式 B 符合第一范式。

关系模式 A						
学号	姓名	班级	课程名称	成绩		
				平时成绩	期中成绩	期末成绩
关系模式 B						
学号	姓名	班级	课程名称	考试类型	成绩	

图 2.15

2. 第二范式(2NF)

如果数据表中的所有非主键字段都“完全函数依赖”(“不部分函数依赖”)于表中的每个关键字,即非主键字段必须由表中的每个关键字的全部字段决定,那么该数据表属于第二范式。

在图 2.16 中,“学生选课”表中的“选课编号”字段、“学号,课程号”组合字段都可以作为关键字。

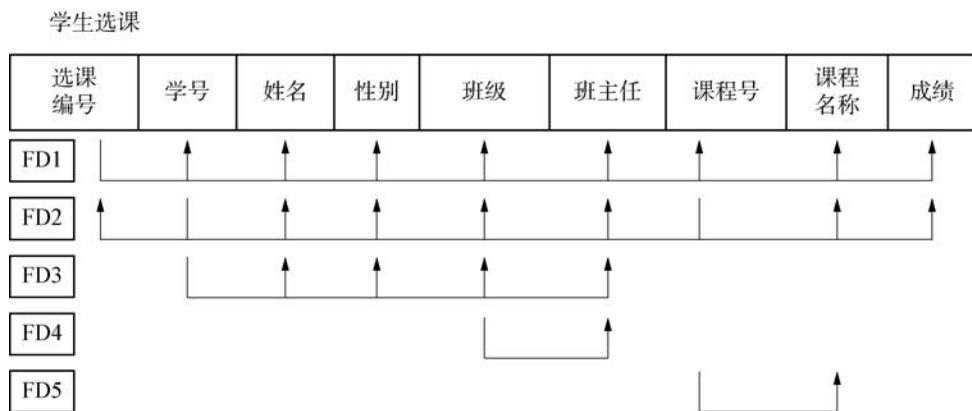


图 2.16

该表中存在如下函数依赖:

FD1: 选课编号 \rightarrow 学号,姓名,性别,班级,班主任,课程号,课程名称,成绩。

FD2: {学号,课程号} \rightarrow 选课编号,姓名,性别,班级,班主任,课程名称,成绩。

FD3: 学号 \rightarrow 姓名,性别,班级,班主任。

FD4:班级→班主任。

FD5:课程号→课程名称。

在FD3中,“学号”决定姓名、性别、班级和班主任,即存在关键字“学号,课程号”中的部分字段决定非主键字段的情况。因此姓名、性别、班级和班主任部分函数依赖于关键字“学号,课程号”。在FD5中,“课程号”决定课程名称,因此课程名称部分函数依赖于关键字“学号,课程号”。由于FD3和FD5的存在,“学生选课”表不符合第二范式,但符合第一范式。

为了把“学生选课”表规范化为第二范式,可以将其分解为如图2.17所示的三张表:把不符合第二范式的FD3、FD5中的字段拿出来,分别组成“学生”表和“课程”表,其他字段组成一张新的“学生选课A”表。

3. 第三范式(3NF)

如果一个数据表满足第二范式,并且不能有一个非关键字字段被另一个非关键字字段(或非关键字字段组合)决定,即不存在非关键字字段“传递依赖”于主键,那么该数据表属于第三范式。

图2.17中的“学生选课A”表和“课程”表都属于第三范式。但是在“学生”表中存在如下函数依赖:

FD3:学号→姓名,性别,班级,班主任。

FD4:班级→班主任。

主键“学号”决定“班级”,非主键“班级”决定“班主任”,即存在传递依赖:学号→班级→班主任。因此,“学生”表不符合第三范式。

为了把“学生”表规范化为第三范式,可以将其分解为如图2.18所示的两张表:把不符合第三范式的FD4中的字段拿出来,组成一张“班级”表,其他字段组成一张新的“学生A”表。

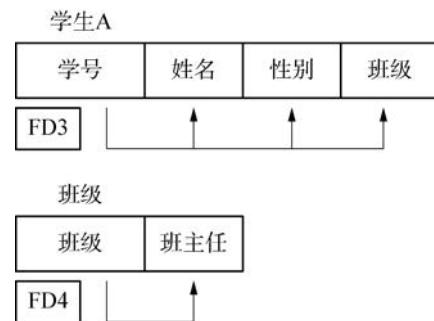


图 2.17

图 2.18

综上所述,可以将表2.9“学生选课”表分解为四张表:“学生选课A”表、“课程”表、“学生A”表、“班级”表,实现数据库设计的规范化。

六、教学参考案例

■ 参考案例

概念模型设计——志愿者服务管理

上海市南洋中学 陈敏

(1课时)

1. 学科核心素养

- 能确定学习和生活中的业务数据问题,提取问题的特征,提出解决方案,评价其合理性、完整性以及分析方案优化或改进的可能性。(计算思维)
- 能对数据进行分类、抽象以及模型化处理。(计算思维)

2.《课程标准》要求

结合案例,了解关系数据模型的基本概念,掌握设计简单关系数据库的逻辑结构的方法。

3. 学业要求

能按照特定数据管理的需求,使用数据库管理系统建立关系数据库,会选用恰当的策略与方法,对数据进行管理。

4. 教学内容分析

本课教学内容是教科书的第二章第二节中部分内容。采用项目学习的方式,通过对“志愿者服务管理”项目进行需求分析,培养学生从情境中提取信息、归纳出牵涉的事物的能力,进而确定事物、事物的属性及事物之间的联系;体会从现实世界的事物及其联系抽象到虚拟世界中的实体和联系,能够完成概念模型的设计,能为数据库的逻辑设计做好铺设。

5. 学情分析

在本节课之前,学生已经学习了数据的价值与重要性、数据管理与分析技术的重要性,知道不同结构化程度的数据,了解数据采集的多种途径。

由于学生接触数据管理与分析的时间非常短,所以他们对于数据管理与分析的信息意识与计算思维能力非常欠缺,在解决实际问题的时候往往会觉得无从下手。

6. 教学目标

- 复述数据库设计的一般过程。
- 复述概念模型建立的过程。
- 说明实体、实体的属性、实体的关键字、实体间的联系等概念。
- 区分实体间联系的三种类型:一对联系(1:1)、一对多联系(1:N)、多对多联系(M:N)。
- 能根据需求分析,建立“志愿者服务管理”数据库的概念模型。

7. 教学重难点

- 教学重点:概念模型建立的过程、确定关键字。
- 教学难点:实体间联系的三种类型,即一对联系(1:1)、一对多联系(1:N)、多对多联系(M:N)。

8. 教学准备

设计教案、活动记录单、PPT，安排学生分组。

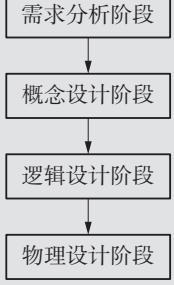
9. 教学策略分析

将学生熟悉的志愿服务活动作为项目情境，讨论项目情境的需求分析，引导学生从现实世界的事物、事物的属性、事物间的联系，抽象为虚拟世界的实体、实体的属性、实体间的联系，构建数据库设计中的概念模型。

小组合作探究，当活动情境发生变化时，在原有概念模型的基础上，添加新的实体、属性及相互的联系，继而将概念模型加以完善。

10. 教学过程设计(见表 2.10)

表 2.10 教学过程设计表

教学环节	教学内容	学生活动	设计意图
情境导入	播放学生参加志愿服务活动的视频	观看视频； 分享参加志愿者服务的经过	让有志愿者服务经历的同学重温过程，让没有经历的同学感受过程
活动 1	学生通过阅读，了解项目活动的最终成果是创建“志愿者服务管理”数据库。 数据库设计的一般过程，分为以下四个阶段：  本节课完成第二个阶段：概念设计阶段	完成活动记录单中的“任务 1”	
活动 2	问题：学生会需要“志愿者服务管理”数据库做什么？实现什么功能？	组内思考讨论问题。 完成活动记录单中的“任务 2”。 小组间交流分享	深入分析用户对此数据库的需求，分享交流，相互取长补短，提升数据库设计的质量
活动 3	引导：现实世界中有人、事、物，项目情境中的人、事、物是什么？ 现实世界中的事物，在虚拟世界中称为实体。 引出实体的概念	学生通过思考、交流得到实体（学生、志愿服务基地、岗位……）。 在活动记录单的“任务 4”中，用笔画出两个实体：学生、志愿服务基地。（注意间隔，可做简单修饰） 请一名学生在教室白板上完成此任务	图形符号可以让学生有更加直观的感受，帮助理解。 鼓励学生发挥想象，用不同的方式完成，形成不同的风格

续表

教学环节	教学内容	学生活动	设计意图
活动 4	<p>教师：每个实体都可以用一组数据来描述其特性，例如描述一个苹果（颜色、大小、品种、甜度……），又例如描述人（姓名、性别、年龄、职业、身高、体重……）。这些描述实体特性的数据就称为属性（属性概念）。</p> <p>当实体具有非常多的属性时，我们要筛选出符合题目需求的实体的属性。例如，突显教师身份的属性（姓名、性别、年龄、工号、教龄、工作单位等）。</p> <p>实体的关键字能够唯一地标识某一个实体的属性（或几个属性的组合）。</p> <p>例如教师（姓名、性别、年龄、工号、教龄、工作单位等），可以使用工号作为实体的关键字</p>	<p>学生思考“学生”和“志愿服务基地”两个实体的属性是什么。</p> <p>在活动记录单的“任务 4”中，在“学生”和“志愿服务基地”两个实体旁边，用笔画出它们各自的属性，在实体的关键字旁边用“*”标注。</p> <p>请一名学生在教室白板上完成此任务</p>	
活动 5	<p>在现实世界中，事物之间是有联系的。这些联系在信息世界中反映为实体间的联系，实体间的联系可以分为三类：一对一联系（1：1）；一对多联系（1：N）；多对多联系（M：N）。</p> <p>分析并举例：</p> <p>一对一联系：一个班级只有一位班主任，一位班主任只带一个班级，班级与班主任之间存在一对一联系。</p> <p>一对多联系：一个班级有多名学生，一名学生只属于一个班级，班级与学生之间存在一对多联系，班级是“一方”，学生是“多方”。</p> <p>多对多联系：一个班级有多位任课教师，每位教师可以任教多个班级，班级与教师之间存在多对多联系</p>	<p>学生思考“学生”和“志愿服务基地”两个实体的联系属于哪一类联系。</p> <p>在活动记录单的“任务 4”中，用线连接“学生”和“志愿服务基地”两个实体，在线上用笔标注它们属于哪一种联系。</p> <p>请一名学生在教室白板上完成此任务</p>	
成果展示		小组展示设计成果，相互交流分享	通过交流修正自己的设计，提升数据库设计的质量
思考讨论	<p>在“志愿者服务管理”数据库中，除了具有“学生”和“志愿服务基地”两个实体之外，还有其他的实体。</p> <p>之前有同学提到过“岗位”，它也是一个实体，请同学们思考：它在图中的位置在哪里？属性有哪些？和其他实体的联系是哪种类型？</p>	<p>学生思考讨论。</p> <p>在活动记录单的“任务 4”中，在之前概念模型设计的基础上，添加“岗位”实体，完善此概念模型</p>	课堂提升
成果展示		小组展示设计成果，相互交流分享	通过交流修正自己的设计，提升数据库设计的质量

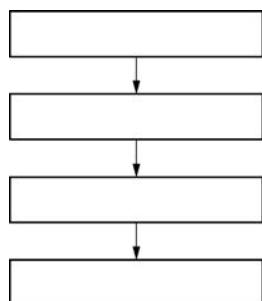
教学环节	教学内容	学生活动	设计意图
课堂小结	数据库设计一般需要哪些过程? 概念模型建立需要哪些过程? 什么是实体、实体的属性、实体的关键字、实体间的联系? 如何区分实体间联系的三种类型?		
课后思考	使用了“志愿者服务管理”数据库一段时间之后,学生会的同学发现,这个数据库只能对本校的志愿服务数据进行管理,若有其他学校的志愿服务数据加入进来,就无法实现该数据库的部分功能了。 这个问题应该怎么解决呢? 请同学们在原有概念模型设计的基础上继续加以完善		

【活动记录单】

概念模型设计——志愿者服务管理

一、数据库设计的一般过程

任务 1:请写出数据库设计的一般过程。



二、需求分析

任务 2:学生会需要数据库做什么? 实现什么功能?

三、设计概念模型

任务 3:选择合适的词语,填在下列横线上。

实体的关键字 实体的属性 实体间的联系 实体

- (1) _____是客观存在且相互区别的事物,也可以是抽象的概念或事件。
- (2) _____是描述实体特性的数据。
- (3) _____能够唯一地标识某一个实体的属性(或几个属性的组合)。
- (4) 在现实世界中,事物之间是有联系的,这些联系在信息世界中反映为_____。

任务4:写写画画。(实体用红色表示、实体的属性用蓝色表示、实体间的联系用黑色表示、实体的关键字用在旁边打“*”表示)

第三节 数据库的实施

一、教学目标与重点

教学目标:

- 了解数据库系统的概念,理解数据库、数据库管理系统、关系数据库、关系数据库管理系统等基本概念。
- 能使用数据库管理系统建立关系数据库,能对数据库中的数据进行增加、删除、修改等基本操作。
- 了解数据库基本的数据查询方法,如选择、投影、排序、统计等。
- 能使用结构化查询语言进行简单的数据查询。

教学重点:

- 能使用数据库管理系统建立关系数据库。
- 能使用结构化查询语言进行简单的数据查询。

二、教学说明与建议

本节的主要任务是：根据上节中已经设计完成的网上书店数据库关系数据模型，使用一个具体的关系数据库管理系统建立关系数据库，并对数据库进行查询等基本操作。本节中的知识点包括数据库系统、数据库、数据库管理系统、关系数据库、关系数据库管理系统等基本概念，需要结合理论学习和实践操作去理解与体会。

本节中使用的数据库管理系统是 MySQL，主要介绍使用结构化查询语言 SQL 进行数据库的创建和查询。为了让学生能更加直观地理解数据库的各种操作，教师可以借助 MySQL 提供的数据库工具 MySQL Workbench，或者第三方工具，如 Navicat Premium 等，使用图形操作界面操作数据库，同时让学生比较使用图形界面和使用 SQL 语言对数据库进行操作的区别。

本节 5 课时安排建议：“数据库系统”1 课时、“建立数据库”2 课时、“数据查询”2 课时。

三、项目实施与评价

1. 核心概念精解

数据模型除了包括用来定义数据库结构和约束的概念之外，还包括用来操作数据库的运算集合。关系模型的基本运算集合就是关系代数。用户运用关系代数的各种运算，来满足操纵数据库中数据的要求。

关系代数的运算可以分为两类：一类是传统的集合运算，包括并、交、差、笛卡尔积等；另一类是专门的关系运算，包括选择、投影、连接、除等。

(1) 选择

选择运算是从一个关系中，找出满足条件的记录，组成新的关系的操作。例如，从“用户”表中选择性别为“女”的记录，构成一个新的关系，运算结果如表 2.11 所示。

表 2.11 选择运算的结果示例

username	password	mail	fullname	sex	birthday	address
user002	*****	user002@mail.sh.cn	李明	女	1975-05-06	上海市徐家汇路
user005	*****	user005@mail.sh.cn	周慧	女	1965-06-05	上海市北京东路
user006	*****	user006@mail.sh.cn	刘实	女	1985-06-15	上海市龙华中路
user007	*****	user007@mail.sh.cn	马卫	女	1970-08-05	上海市四川北路

(2) 投影

投影运算是从一个关系中，找出包含指定字段的记录，组成新的关系的操作。例如，在“用户”表中，对“姓名”“性别”“出生日期”进行投影，运算结果如表 2.12 所示。

表 2.12 投影运算的结果示例

fullname	sex	birthday	fullname	sex	birthday
王成	男	1980-03-16	周慧	女	1965-06-05
李明	女	1975-05-06	刘实	女	1985-06-15
赵林	男	2000-09-01	马卫	女	1970-08-05
张海	男	1990-12-11	林夏	男	1985-12-09

2. 项目活动的具体实施

项目任务:网上书店数据库建立与查询。根据在第二节时完成设计的数据库关系模型,使用 MySQL 数据库管理系统创建网上书店数据库,并使用 SQL 语句在该数据库中查找需要的数据。

(1) 活动 1

根据网上书店数据库的关系模型,设计网上书店数据库中数据表“用户(users)”“订单(orders)”“订单明细(orderdetails)”“图书(books)”“图书类别(categories)”的结构,如表 2.13~2.17 所示。

表 2.13 用户(users)

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

表 2.14 订单(orders)

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

表 2.15 订单明细(orderdetails)

列名	数据类型	长度	说明
orderID	char	16	订单号
bookID	char	10	图书编号
quantity	int		数量

表 2.16 图书(books)

列名	数据类型	长度	说明
bookID	char	10	图书编号
title	char	20	书名
price	decimal	10,2	定价
author	char	20	作者
publisher	char	20	出版社
pubdate	date		出版日期
discount	float		折扣
categoryID	char	2	类别编号

表 2.17 图书类别(categories)

列名	数据类型	长度	说明
categoryID	char	2	类别编号
category	char	20	类别名称

(2) 活动 2

在 MySQL 数据库管理系统中, 使用 SQL 语句创建网上书店数据库、数据表及数据表之间的关系, SQL 语句如下:

```
CREATE DATABASE bookstore;
USE bookstore;
```

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

```

    birthday date,
    address char(50));

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

CREATE TABLE orderdetails
(
    orderID char(16),
    bookID char(10),
    quantity int,
    PRIMARY KEY (orderID,bookID));

CREATE TABLE books
(
    bookID char(10) PRIMARY KEY,
    title char(20),
    price decimal(10,2),
    author char(20),
    publisher char(20),
    pubdate date,
    discount float,
    categoryID char(2));

CREATE TABLE categories
(
    categoryID char(2) PRIMARY KEY,
    category char(20));

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

ALTER TABLE orderdetails ADD CONSTRAINT fk_orderID FOREIGN KEY
(orderID) REFERENCES orders(orderID);

```

```
ALTER TABLE orderdetails ADD CONSTRAINT fk_bookID FOREIGN KEY  
(bookID) REFERENCES books(bookID);
```

```
ALTER TABLE books ADD CONSTRAINT fk_categoryID FOREIGN KEY  
(categoryID) REFERENCES categories(categoryID)
```

(3) 活动 3

输入网上书店数据库中的数据(请参考本节的“五、教学参考资源”)。

在数据库中输入数据可以使用以下方法:

- ① 使用 INSERT 语句在数据表中添加数据。
- ② 使用复制/粘贴方法,将需要输入的数据复制后,在 Workbench 中打开数据表,用“Paste Row”命令将数据粘贴到数据表中,并保存。
- ③ 使用导入和导出方法,在 Workbench 中打开数据表,用“Export/Import”命令进行数据的导出和导入。注意导入的数据与 MySQL 中数据表数据采用的编码要保持一致,如都采用 UTF-8 编码。

3. 项目活动的评价

本节的项目评价主要包括:考查学生对数据库基本的数据查询方法,如选择、投影、排序、统计等基本概念的掌握程度;考查学生使用一个具体的数据库管理系统,完成创建网上书店数据库任务,并对该数据库进行增加记录、修改记录、删除记录,以及数据的查询等基本操作的情况。

评价建议:主要采用过程性评价。设计相应的评价量表,重点对学生设计数据库结构、使用 SQL 语句进行评价。在设计评价量表时,不仅要关注如何对学生的活动表现进行评价,更要对如何评价 SQL 语句的掌握程度进行详细设计。

四、作业练习与提示

■ 题目描述

1. 使用 SQL 语言,创建运动会管理数据库。数据库中包含运动员、比赛项目和参赛表三张数据表。

运动员(运动员号码,姓名,性别,班级)

比赛项目(项目编号,项目名称,比赛时间,比赛地点)

参赛表(运动员号码,项目编号,比赛成绩)

2. 使用 SQL 语言,在 bookstore 数据库中查询数据。
 - (1) 在“users”数据表中查找所有的用户信息,结果显示用户名、邮箱、姓名、地址。
 - (2) 在“users”数据表中查找姓名为“王成”的用户信息,结果显示用户名、邮箱、姓名、地址。
 - (3) 查找姓名为“王成”的用户订单,结果显示用户名、姓名、地址、订单号、下单时间、交易状态。
 - (4) 在“books”数据表中按出版社统计每个出版社的图书总数,结果显示出版社名称

和图书总数。

(5) 在“orderdetails”数据表中按订单号统计每张订单的订书总量，并显示订购 6 本以上图书的订单号和订书总量。

(6) 在“books”数据表中查找书名中包含“汉语词典”的图书信息，结果显示图书编号、书名、定价、作者、出版社，并按定价降序排序。

■ 作业提示

1. 参考答案如下：

```
CREATE DATABASE 运动会管理;
```

```
USE 运动会管理;
```

```
CREATE TABLE 运动员
```

```
(
```

```
运动员号码 char(4) PRIMARY KEY,
```

```
姓名 char(20),
```

```
性别 char(2),
```

```
班级 char(20));
```

```
CREATE TABLE 比赛项目
```

```
(
```

```
项目编号 char(4) PRIMARY KEY,
```

```
项目名称 char(20),
```

```
比赛时间 datetime,
```

```
比赛地点 char(20));
```

```
CREATE TABLE 参赛表
```

```
(
```

```
运动员号码 char(4),
```

```
项目编号 char(4),
```

```
比赛成绩 char(20),
```

```
PRIMARY KEY (运动员号码,项目编号));
```

```
ALTER TABLE 参赛表 ADD CONSTRAINT fk1 FOREIGN KEY (运动员号码)  
REFERENCES 运动员(运动员号码);
```

```
ALTER TABLE 参赛表 ADD CONSTRAINT fk2 FOREIGN KEY (项目编号)  
REFERENCES 比赛项目(项目编号)
```

2. 参考答案如下：

(1) SELECT username,mail,fullname,address

```

FROM users
(2) SELECT username,mail,fullname,address
FROM users
where fullname= '王成'
(3) SELECT users.username,fullname,address,orderID,ordertime,state
FROM users INNER JOIN orders ON orders.username= users.username
WHERE fullname= '王成'
(4) SELECT publisher,Count(bookID)
FROM books
GROUP BY publisher
(5) SELECT orderID,Sum(quantity)
FROM orderdetails
GROUP BY orderID
HAVING Sum(quantity)> 6
(6) SELECT bookID,title,price,author,publisher
FROM books
WHERE title LIKE '%汉语词典%'
ORDER BY price DESC

```

五、教学参考资源

■ 参考资料：教科书中的示例数据库“bookstore”的数据表

表 2.18 users

username	password	mail	fullname	sex	birthday	address
user001	*****	user001@mail.sh.cn	王成	男	1980-03-16	上海市晋元路
user002	*****	user002@mail.sh.cn	李明	女	1975-05-06	上海市徐家汇路
user003	*****	user003@mail.sh.cn	赵林	男	2000-09-01	上海市中山西路
user004	*****	user004@mail.sh.cn	张海	男	1990-12-11	上海市南京东路
user005	*****	user005@mail.sh.cn	周慧	女	1965-06-05	上海市北京东路
user006	*****	user006@mail.sh.cn	刘实	女	1985-06-15	上海市龙华中路
user007	*****	user007@mail.sh.cn	马卫	女	1970-08-05	上海市四川北路
user008	*****	user008@mail.sh.cn	林夏	男	1985-12-09	上海市淮海东路

表 2.19 orders

orderID	ordertime	state	username
1000020000823301	2018-07-13 10:09:20	交易完成	user001
1000020000823501	2018-07-13 20:10:11	交易完成	user002
1000020000826209	2018-07-15 11:20:16	交易完成	user003
1000020000827101	2018-07-16 15:52:50	交易完成	user001
1000020000828001	2018-07-17 18:06:33	交易完成	user002
1000020000829100	2018-07-18 12:00:26	交易完成	user004
1000020000829500	2018-07-18 21:15:53	交易完成	user005
1000020000830100	2018-07-20 10:15:12	交易完成	user006
1000020000830312	2018-07-21 19:20:08	交易完成	user005

表 2.20 orderdetails

orderID	bookID	quantity
1000020000823301	1103028726	1
1000020000823301	1101083927	1
1000020000823301	1103188334	1
1000020000823501	1103134637	1
1000020000823501	1103105268	2
1000020000823501	1102710848	1
1000020000823501	1103188334	1
1000020000826209	1101756339	1
1000020000826209	1102474857	3
1000020000826209	1103149922	1
1000020000826209	1102710848	1
1000020000826209	1212653156	1
1000020000827101	1103146032	1
1000020000827101	1102307520	1
1000020000827101	1103149922	2
1000020000827101	1102474857	1
1000020000827101	1102710848	1
1000020000828001	1101046760	1
1000020000828001	1103127445	1
1000020000828001	1102639657	1
1000020000829100	1103222096	1
1000020000829100	1103119488	1

续表

orderID	bookID	quantity
1000020000829100	1101046760	1
1000020000829100	1103264489	2
1000020000829500	1103149922	1
1000020000829500	1103123723	1
1000020000829500	1101046760	1
1000020000829500	1102757665	1
1000020000829500	1103264489	2
1000020000830100	1103113614	1
1000020000830100	1103151323	1
1000020000830100	1101046760	1
1000020000830100	1101983703	1
1000020000830100	1103228203	3
1000020000830312	1103168786	1
1000020000830312	1102639657	1
1000020000830312	1101083927	1
1000020000830312	1103123723	1
1000020000830312	1101983703	2

表 2.21 books

bookID	title	price	author	publisher	pubdate	discount	category ID
1103028726	数据 库 技 术——原 理与 设计	36.00	朱烨、张敏辉	高等教 育出 版社	2017-08-01	0.8	01
1103134637	大数 据资源	80.00	朱扬勇	上海科学技 术出版 社	2018-01-01	0.8	01
1101756339	PHP+MySQL 开发实 战	89.80	软件开发技 术联盟	清华大 学出 版社	2013-09-01	0.8	01
1103146032	Oracle从入 门到精 通	65.00	创客诚品、郑 彬彬、郑秋生	北京希望电 子出版 社	2017-10-01	0.8	01
1101046760	中国哲学史	72.00	冯友兰	华东师范大 学出版 社	2011-07-01	0.8	02
1103222096	大国崛起与国 际和平	90.00	何银	时事出版 社	2018-05-01	0.8	03

续表

bookID	title	price	author	publisher	pubdate	discount	category ID
1103149922	国际政治语言学	88.00	孙吉胜	世界知识出版社	2017-08-01	0.7	03
1103113614	法律的起源	39.00	刘春兴	中国政法大学出版社	2017-11-01	0.8	04
1103168786	数字公共图书馆著作权限制研究	39.00	赵力	知识产权出版社	2018-02-01	0.9	04
1101083927	百年航母	66.00	张召忠	广东经济出版社	2011-07-01	0.8	05
1103105268	现代物流管理	38.00	李焱	北京大学出版社	2017-11-01	0.7	06
1102474857	读懂“一带一路”	58.00	厉以宁等	中信出版社	2015-11-01	0.9	06
1102307520	中华文明的核心价值	38.00	陈来	三联书店	2015-04-01	0.8	07
1103127445	男孩女孩学习大不同	69.90	迈克尔·吉里安	浙江人民出版社	2017-12-01	0.8	08
1103119488	教师的真情与智慧	36.00	于春吉	中国轻工业出版社	2017-07-01	0.7	08
1103123723	地球的奥秘	36.00	嵇少丞	浙江教育出版社	2017-11-01	0.7	09
1103151323	中国古典诗词名篇诵读	36.00	季羨林	世界图书出版公司	2017-10-01	0.7	09
1102639657	仰望量子群星	78.00	魏凤文、高新红	浙江教育出版社	2016-03-01	0.7	09
1101983703	古代汉语词典(第2版)	119.90	商务印书馆辞书研究中心	商务印书馆	2014-03-01	0.7	11
1102710848	10000条成语大词典	32.80	说词解字辞书研究中心	华语教学出版社	2016-08-01	0.7	11
1102757665	现代汉语词典(第7版)	109.00	中国社科院语言研究所	商务印书馆	2016-09-01	0.8	11
1103228203	中国现代文学史	58.00	程光炜等	北京大学出版社	2011-10-01	0.8	12
1101488849	中国文脉	38.00	余秋雨	长江文艺出版社	2012-11-01	0.7	12
1103264489	自然哲学之数学原理	168.00	牛顿	北京大学出版社	2018-06-24	0.7	13

续表

bookID	title	price	author	publisher	pubdate	discount	categoryID
1212653156	植物探索之旅	80.00	桑德拉·纳普	长春出版社	2017-04-01	0.7	13
1103188334	丝绸之路考古论集	98.00	徐苹芳	上海古籍出版社	2017-12-01	0.8	14
1100456317	故宫陶瓷馆	480.00	故宫博物院	紫禁城出版社	2008-05-01	0.7	14
1212728631	室内植物装饰设计	39.00	庄夏珍	重庆大学出版社	2011-12-01	0.7	15

表 2.22 categories

categoryID	category	categoryID	category
01	计算机	10	中小学课本
02	哲学	11	语言文字
03	政治	12	文学
04	法律	13	自然科学
05	军事	14	历史地理
06	经济	15	美术设计
07	文化理论事业	16	趣味阅读
08	教育学	17	少儿读物
09	中小学课外阅读	18	竞技体育

六、教学参考案例

■ 参考案例

建立数据库——志愿者服务管理

上海市南洋中学 陈敏

(1课时)

1. 学科核心素养

- 能使用数据库管理系统建立关系数据库；能对数据库中的数据进行增加、删除、修改等基本操作。（计算思维）
- 能根据需要，主动选用数字化工具开展自主或协作学习。（数字化学习与创新）

2. 《课程标准》要求

使用数据库管理系统建立关系数据库。

3. 学业要求

能按照特定数据管理的需求,使用数据库管理系统建立关系数据库,会选用恰当的策略与方法,对数据进行管理。

4. 教学内容分析

本课教学内容为教科书的第二章第三节中部分内容。采用项目学习的方式,通过使用数据库管理系统创建“志愿者服务管理”关系数据库,使学生进一步理解数据管理的思想,进而理解数据管理技术;培养根据关系数据模型建立关系数据库的能力;体会使用数据库管理系统对数据库中的数据进行存储、处理和管理。

5. 学情分析

在本节课之前,学生已经学习了数据模型的设计,并按照数据库设计的一般步骤,设计了“志愿者服务管理”数据库的关系模型。

学生了解数据库、数据库管理系统、关系数据库、关系数据库管理系统等基本概念;知道常见的数据库管理系统,如 Access、MySQL、SQL Server 等,并能使用其中的某一个进行数据管理的简单体验操作。

6. 教学目标

- 了解创建、打开和删除数据库的方法。
- 理解数据表的相关概念;了解创建、删除和修改数据表的方法。
- 能够建立表之间的关系。
- 能对数据表中的数据进行增加、删除、修改操作。
- 能使用 SQL 语言对数据库和数据表进行操作。

7. 教学重难点

- 教学重点:创建数据表;对数据表中的数据进行增加、删除、修改操作。
- 教学难点:使用 SQL 语言创建数据表、建立表之间的关系。

8. 教学准备

设计教案、活动记录单、PPT,安排学生分组。

9. 教学策略分析

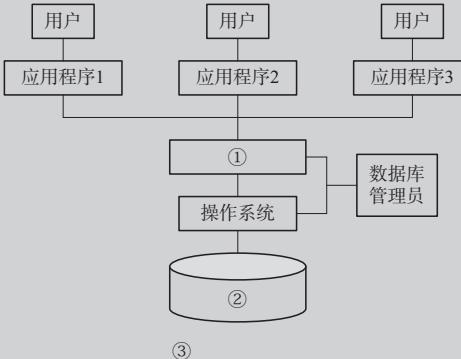
以教科书中的网上书店数据库为例,让学生理解数据库和数据表的相关基本概念,如表的结构(字段名、数据类型等)、表的记录、表之间的关系等。

以教科书中的网上书店数据库为例,介绍使用 SQL 语言创建数据库、数据表等操作。对于学习有困难的学生,可以先使用图形界面操作,再过渡到使用 SQL 语言。

提供微视频学习资源,帮助学生开展自主学习。

10. 教学过程设计(见表 2.23)

表 2.23 教学过程设计表

教学环节	教学内容	学生活动	设计意图																
情境导入	操作演示网上书店数据库、根据需求在此数据库中查找数据	观察并思考:在网上书店数据库中看到了什么?数据库可以实现哪些功能?	激发学生对使用数据库的兴趣																
教师讲授	1. 数据库系统; 2. 数据库; 3. 数据库管理系统	活动:将数据库系统、数据库、数据库管理系统分别填在图中序号处。 	让学生更好地理解数据库相关概念,以及它们之间的关系																
活动 1	操作演示: 1. 创建网上书店数据库 bookstore; 2. 打开网上书店数据库 bookstore	参考教科书,在 MySQL 数据库管理系统中,使用 SQL 语句创建、打开、删除网上书店数据库	掌握创建数据库的方法																
活动 2	1. 设计网上书店数据库中“用户(users)”数据表的结构 • 设计“用户(users)”数据表的结构。 <table border="1" data-bbox="325 1223 714 1470"> <thead> <tr> <th>列名</th><th>数据类型</th><th>长度</th><th>说明</th></tr> </thead> <tbody> <tr> <td>username</td><td>char</td><td>20</td><td>用户名</td></tr> <tr> <td>password</td><td>char</td><td>20</td><td>密码</td></tr> <tr> <td>.....</td><td></td><td></td><td></td></tr> </tbody> </table> • 使用 SQL 语言创建“用户(users)”数据表。 CREATE TABLE users (<列名 1><数据类型><列约束>, <列名 2><数据类型><列约束>,)	列名	数据类型	长度	说明	username	char	20	用户名	password	char	20	密码				思考:数据库中的数据表与概念模型之间的关系是什么? 学生活动: • 设计网上书店数据库中“订单(orders)”数据表的结构。 • 使用 SQL 语言创建“订单(orders)”数据表 学生活动: 使用 SQL 语言建立“用户(users)”数据表与“订单(orders)”数据表之间的关系	掌握创建数据表、建立表之间的关系的方法
列名	数据类型	长度	说明																
username	char	20	用户名																
password	char	20	密码																
.....																			

续表

教学环节	教学内容	学生活动	设计意图																												
活动 3	<p>1. 增加记录</p> <ul style="list-style-type: none"> • INSERT 语句格式 INSERT INTO 数据表名[(<列名表>)]VALUES(值列表)。 • 问题 如果需要插入的一条记录中只有部分数据,应该如何操作? 	<p>学生活动:</p> <ul style="list-style-type: none"> 在“用户(users)”数据表中插入一条记录: <table border="1"> <thead> <tr> <th>username</th><th>password</th><th>mail</th></tr> </thead> <tbody> <tr> <td>user001</td><td>*****</td><td>user001@mail.sh.cn</td></tr> </tbody> </table> <table border="1"> <thead> <tr> <th>fullname</th><th>sex</th><th>birthday</th><th>address</th></tr> </thead> <tbody> <tr> <td>王成</td><td>男</td><td>1980-03-16</td><td>上海市晋元路</td></tr> </tbody> </table> <ul style="list-style-type: none"> 参考教科书,在“用户(users)”数据表中插入如下记录: <table border="1"> <thead> <tr> <th>username</th><th>password</th><th>mail</th></tr> </thead> <tbody> <tr> <td>user002</td><td>*****</td><td>user002@mail.sh.cn</td></tr> </tbody> </table> <table border="1"> <thead> <tr> <th>fullname</th><th>sex</th><th>birthday</th><th>address</th></tr> </thead> <tbody> <tr> <td>张勤</td><td>女</td><td></td><td></td></tr> </tbody> </table>	username	password	mail	user001	*****	user001@mail.sh.cn	fullname	sex	birthday	address	王成	男	1980-03-16	上海市晋元路	username	password	mail	user002	*****	user002@mail.sh.cn	fullname	sex	birthday	address	张勤	女			掌握数据表的基本操作:增加记录、修改记录、删除记录
username	password	mail																													
user001	*****	user001@mail.sh.cn																													
fullname	sex	birthday	address																												
王成	男	1980-03-16	上海市晋元路																												
username	password	mail																													
user002	*****	user002@mail.sh.cn																													
fullname	sex	birthday	address																												
张勤	女																														
	<p>2. 修改记录</p> <ul style="list-style-type: none"> • UPDATE 语句格式: UPDATE 数据表名 SET <列名>=<表达式>[,<列名>=<表达式>][WHERE<条件>] • 介绍在 MySQL 中,条件表达式的算术运算符、关系运算符和逻辑运算符,以及模糊匹配 LIKE 	<p>学生活动:</p> <p>在“用户(users)”数据表中将用户名称为“user001”的用户姓名修改为“王晨”</p>																													
	<p>3. 删除记录</p> <ul style="list-style-type: none"> • DELETE 语句格式: DELETE FROM 数据表名[WHERE <条件>] 	<p>学生活动:</p> <p>删除“用户(users)”数据表中用户名称为“user001”的用户</p>																													
巩固提升		<ol style="list-style-type: none"> 完成活动记录单:思考如何建立“志愿者服务管理”数据库。 使用 SQL 语言建立“志愿者服务管理”数据库。 创建“学生”“博物馆”数据表,并建立两表之间的关系。 在数据库中进行增加、修改、删除记录等操作 	课堂巩固																												
成果展示		小组展示“志愿者服务管理”数据库,相互交流分享	通过交流修正自己的设计,提升数据库设计的质量																												

续表

教学环节	教学内容	学生活动	设计意图
课堂小结	<ol style="list-style-type: none">1. 如何创建、打开和删除数据库?2. 如何创建、删除和修改数据表?3. 怎样建立表之间的关系?4. 怎样对数据表中的数据进行增加、删除、修改操作?		
课后作业	在“学生”“博物馆”两张数据表的基础上,添加其他数据表,进一步完善数据库		

【活动记录单】

一、创建数据库

二、创建数据表

1. 数据表的结构

2. 数据表之间的关系

关系名称	数据表 1	外键	数据表 2	主键

3. SQL 语句

(1) 创建数据表

(2) 建立数据表之间的关系

--

三、编辑数据表

1. 增加记录

2. 修改记录

3. 删 除 记 录

数据安全

一、本章学科核心素养的渗透

1. 信息意识

随着人们对数据认识的日益加深,特别是大数据的兴起和快速发展,数据的价值不断被挖掘提升,从而也导致了因数据丢失、数据泄露而造成巨大经济损失,因此数据安全问题越来越受到人们的关注。

现在人们做各种事情和解决各种问题时,都会以数为源、以数为据、以数为宝,“用数据说话”已经成为现代人看待问题和作出决策的基本方式,数据意识也成为现代人必须具备的基本意识之一。但与此同时,如果在数据使用过程中不加注意,就会引发各种数据安全问题,大到国家安全、社会公共利益,小到组织在网络空间中的合法权益和个人的隐私信息都有可能受到影响。维护国家安全需要全面加强国家安全教育,增强全民国家安全意识和素养,筑牢国家安全的人民防线。因此,本章通过介绍数据安全的重要性和各种可能对数据安全造成威胁的因素,帮助学生建立数据安全的概念,了解保障数据安全的策略,知晓常用的数据安全防护手段,如数据加密、访问控制、安全审计、备份与还原等,从而牢固树立安全防范意识。除了了解数据库管理系统本身的安全、知道数据管理者需要具备较高的数据安全意识从而保证数据管理的安全外,因为我们每个人都是数据的使用者,所以要从自我做起,提升安全防范的意识,养成良好的数据使用习惯,加固软硬件系统,绕开病毒传播渠道,关闭可能成为黑客攻击的“后门”,捍卫自己的“信息疆域”。与此同时,做好数据备份工作,通过数据的备份与还原,将因各种主客观原因造成的数据损失降到最低。若人人如此,教科书中一开始所提到的病毒肆虐、黑客攻击等威胁数据安全的现象将大大减少或者影响将会大大降低,能使我们更多享受互联网带来的数据交换的便捷和更好地发挥数据的价值,从而建设更高水平的平安中国,以新安全格局保障新发展格局。

2. 计算思维

本章可以通过数据安全备份的内容使学生了解冗余与可靠的思想。冗余分两种,一

种是不需要的部分,一种是人为增加的重复部分。显然,为保证数据的完整性和可用性而采取的数据备份措施所产生的冗余属于后者。数据备份系统也可以称为冗余系统,指为增加系统的可靠性而采取两套或两套以上相同、相对独立配置的系统的设计。由于我们希望在采用冗余备份的机制后,如果一套系统出现故障,另一套系统能立即启动代替工作,从而让用户几乎感受不到故障的发生,因此我们在备份时不能只备份数据,而是要对包含数据及程序在内的整个系统进行备份,从而更加高效、便捷。

3. 数字化学习与创新

数字化学习与创新,需要学生能适应数字化学习环境,养成数字化学习与创新的习惯,掌握各种学习资源和工具,开展自主学习、协同工作、知识分享与创新。在数据安全方面,主要希望学生能掌握常用的确保数据安全的防护措施和手段,并能在实际应用中根据各种不同的实际需求,合理使用并有所创新。

4. 信息社会责任

信息社会责任要求每个个体都能在文化修养、道德规范、行为自律方面尽到自己的责任。

数据带来的价值和创新力不断提升,并已成为经济社会发展新的驱动力。大数据已经与农业时代的人口土地、工业时代的钢铁石油一样,成为一个国家的重要战略资源,正日益对国家治理能力、经济运行机制、社会生活方式以及各领域的生产、流通、分配、消费活动产生重要影响。世界各国对数据的依赖均快速上升,未来国家层面的竞争力将部分体现为一国拥有数据的规模、活性以及解释、运用的能力。因此现在的数据安全已不仅关系着个人和企业,更与国家安全和社会稳定息息相关。为了维护国家安全和社会公共利益,保护公民、法人和其他组织在网络空间中的合法权益,保障个人信息和重要数据安全,中国和包括英美、欧盟在内的很多国家和组织都制定了与大数据安全相关的法律法规和政策来推动大数据的利用和安全保护,在政府数据开放、数据跨境流通和个人信息保护等方面进行探索和实践,引导人们文明上网,合理合法地收集和使用数据。

本章通过对数据安全知识的学习,帮助学生增强数据安全意识,做到遵守信息安全相关法律法规,维护信息社会的伦理道德规范,在面对各种网络数据和信息时,能够理性判断,自觉规范自己的信息行为。

二、本章知识结构

本章遵循普通高中信息技术课程标准,依据学分和课时规定,紧扣学科概念体系,将内容分为两节,以“在线考试系统的安全维护”为项目主题,围绕数据安全威胁与数据安全策略、数据备份与还原的实现展开设计。

第一节“数据安全威胁与数据安全策略”,结合从新闻报道中的了解、自己及身边人学习与生活中的遭遇、教科书中的项目案例,了解数据安全的概念,知道数据安全包括数据本身的安全和数据防护的安全两方面的内容。对数据安全的控制需要从数据库管理系统本身的安全、人对数据的管理安全、对可以接触到数据的人的管理安全即内部安全三个层

面来实现；由于现有很多数据是保存在数据库系统中的，因此制定数据库系统的安全策略需要保证数据的保密性、完整性、可用性、可控性和不可否认性；数据安全防护手段包括数据加密、访问控制、安全审计、备份与还原，以及安装运行杀毒软件、防火墙，进行入侵检测等，更重要的是加强安全管理。

第二节“数据备份与还原的实现”，将数据安全防护手段中的备份与还原单独进行介绍，结合案例认识数据丢失的风险，利用全量备份、增量备份、差异备份、定时备份与实时备份等多种方法进行数据备份，守好数据安全的最后一道防线。

三、本章项目活动设计思路

本章的项目活动设计以学生很熟悉的考试为例，围绕无纸化在线考试系统的安全展开，让学生以在线考试系统管理员的身份思考相关安全问题。

项目任务1：在线考试系统的数据安全措施。

首先让学生思考，在考试前需要对系统做哪些准备工作，才能保证考试能在系统上顺利进行；接着让学生思考在考试过程中可能遇到哪些突发状况，并做好应急方案，以备不时之需；最后让学生考虑考试结束后，需要采用哪些措施维护考试数据。

项目任务2：不同备份方式的数据还原。

任何以预防为目的的保护措施，无论其多么全面周到、细致入微，都只能尽量地减少而不能完全杜绝灾难的发生。因此，数据备份是数据安全的最后一道防线。该项目活动带领学生通过实践掌握备份、还原的具体操作步骤，同时加深对不同的备份方式的理解，在操作过程中比较不同备份方式的特点和异同，帮助学生在不同的情境下作出最优的选择。例如：第一次备份的时候选择全量备份的方式，其他时候选择效率较高的差异备份方式；如果自己有空间足够的干净硬盘，可选择本地备份，如果拥有较好的网络带宽，可以选择云端备份；可以根据文件的重要性和修改保存的周期制定一个实时备份或定期备份的方案；除了可以对电脑中的各种数据文件和系统进行备份，也可以对手机中的各种文件和应用进行备份，选择适合不同系统的备份软件备份和还原数据。

四、本章课时安排建议

本章教学建议用3课时完成，具体参见表3.1。

表3.1 课时安排建议表

节名	建议课时
第一节 数据安全威胁与数据安全策略	2课时
第二节 数据备份与还原的实现	1课时

第一节

数据安全威胁与数据安全策略

一、教学目标与重点

教学目标：

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

教学重点：

- 结合案例认识数据丢失的风险,了解数据安全的重要性、威胁数据安全的主要因素和常用的防护手段。

二、教学说明与建议

建议根据教科书中的项目活动先以学生为主,围绕学生身边的相关内容,如在线考试系统的数据安全充分开展讨论,教师引导学生认识对数据安全造成威胁的外部因素和内部因素,有些因素很难避免,需要做好备份和应急预案,而有些因素可以通过各种措施加以控制。然后教师再讲解书上对应的知识内容,在项目体验中总结知识要点,还可以扩展案例,加深学生对知识的记忆、认知和理解。

本节2课时安排建议:“数据安全概念”0.5课时,“数据安全控制”0.5课时,“数据安全策略”0.5课时,“数据安全的防护手段”0.5课时。

三、项目实施与评价

1. 核心概念精解

(1) 数据安全

数据安全包含两方面的内容。一是数据本身的安全,主要是指采用复杂的加密算法对数据进行主动保护,如数据加密、访问控制等,以防止数据被泄露、篡改;二是数据防护

的安全,主要是指采用先进的信息存储手段对数据进行主动防护,如通过数据备份保证数据的安全,避免因数据丢失而造成损失。

数据安全应是全方位的,它涉及数据处理、存储、传输等各个环节的安全。数据处理的安全是指避免数据在录入、分析、统计、打印时由于一些突发因素,比如硬件故障、电脑断电、电脑死机、存储介质损坏、操作员误操作、不具备资格的人越权操作、管理不善、病毒入侵、黑客攻击以及自然灾害等,而造成的数据的损失或丢失现象。数据存储的安全保障主要是对存储介质中的数据提供高强度的加密保护,用户可将数据加密后存储,这样即便数据管理失控或网络遭受攻击,也能多一道防护屏障。数据传输的安全保障是指针对数据传输过程中可能遭受到的如中断、窃听、篡改、伪造等安全威胁而需要采取的相应措施。在全方位的数据安全中始终包含数据本身的安全和数据防护的安全两方面的内容。

从数据安全的概念中,引出威胁数据安全的因素:病毒感染、黑客入侵、自然灾害、软硬件故障以及人为失误等。

若要有效地保障数据安全就需要从数据库管理系统本身的安全、人对数据的管理安全以及对可以接触到数据的人的管理安全,即数据库管理系统的安全、数据管理安全(提高数据管理者的安全意识)、内部安全(提高内部数据使用者的安全意识)三个层面来实现。

同时,我们要清醒地认识到现实世界中不存在绝对的安全,因此只能在评估数据价值的基础上,选择合理的数据安全策略,不同价值的数据安全保护,采取的措施和为此付出的代价是不同的。

(2) 数据安全策略

数据安全策略是指为保证提供一定级别的安全保护所必须遵守的规则,是一组规定如何管理、保护和指派敏感信息的法律、规则及实践经验的集合。数据安全策略需要建立在授权的基础上,例如,对未经授权的用户,规定其不得访问、引用或使用数据,对不同等级的授权用户,规定其在相应的范围内使用数据。

由于现在有很多数据是保存在数据库系统中的,因此针对数据库系统的安全策略是保证数据的保密性、完整性、可用性、可控性和不可否认性。为保证数据库系统安全策略的实施,可以采取的安全防护手段包括数据加密、访问控制、安全审计、备份与还原、病毒查杀、安装防火墙、入侵检测以及加强对人的管理等。在数据加密这个最常用的防护手段中,对密码和密钥两个概念的理解需要特别注意:密码是指将明文与密文进行相互转换的算法;密钥则是在密码中使用且只有收发双方知道的信息。我们在日常生活中常说的加密文件的密码,实际对应的是密钥。

2. 项目活动的具体实施

项目任务 1:在线考试系统的数据安全措施。

考试前:主要考虑外部环境、计算机及网络硬件、考试系统等软件,以及考试内容和人员行为等相关的安全问题,以及可采用的应对办法。

考试中:主要考虑考试过程中可能出现的突发事件,以及产生的危害,了解事件产生的原因和可采用的应急办法,并了解应急预案的重要性。

考试后:考试结束后,主要是通过自动化或人工的方式批改试卷,并生成包括成绩信息在内的各种与考试相关的数据,可供教师和学生评价学习的效果。需要考虑如何妥善保存这些数据,以便后续分析,同时防止因各种原因导致的数据丢失和数据泄露。

3. 项目活动的评价

本节的项目评价主要包括:考查学生是否具备一定的数据安全意识,能否充分认识到数据安全的重要性;考查学生能否较为全面地考虑各种威胁数据安全的内外部因素,以及可以采取的防范措施;考查学生对数据安全、数据安全控制、数据安全策略、数据安全防护手段等核心概念的理解和掌握程度。

评价建议:设计评价量表,对学生参与任务的积极性、全面性进行过程性评价;采用纸笔测试等方式,对学生理解和掌握核心概念的程度进行终结性评价。

四、作业练习与提示

■ 题目描述

1. 日常生活中,你或周围的人有过数据丢失或泄露的经历吗?是什么原因造成了数据的丢失或泄露?产生了哪些影响?
2. 你采取过哪些保护数据安全的措施?
3. 应该怎样预防数据丢失或泄露事故的发生?
4. 现代社会中,手机已成为我们日常生活中必不可少的设备,其中存储着大量重要的数据。我们在使用手机中的各种应用程序时,该如何保障手机中数据的安全?当我们需要更换手机时,又需要注意哪些问题?

■ 作业提示

1. 该问题引导学生观察身边事物,了解什么是数据丢失、什么是数据泄露,以及它们造成了怎样的影响,从而深刻认识到数据安全的重要性,同时找到数据丢失或泄露的原因,为后续寻求解决办法奠定基础。
2. 该问题引导学生通过分析数据丢失和数据泄露的原因寻求保护数据安全的方法,便于后续学习时对常用的数据安全防护措施进行梳理和学习。
3. 该问题以预防数据丢失或泄露的方法为主,提醒学生注意,在数据安全方面,提前预防、未雨绸缪更为重要。
4. 该问题主要是让学生根据本节讲解的内容,总结知识,以手机数据安全保护为例,进行知识的应用,不仅掌握知识,还能动手动脑解决实际问题,进而提升学生的数据安全意识。

五、教学参考资源

■ 参考资料 1:从个人到国家的数据安全案例

从个人层面的数据安全到国家层面的数据安全,随着大数据时代的到来,数据安全显

得越来越重要。

1. 个人数据安全案例

当我们接到各种电话、短信、邮件烦不胜烦地滋扰时,可能就意味着我们的电话号码、邮箱、购物记录、收入水平等信息已经扩散出去了,而这可能就源于我们在注册某个网站时填报了个人信息。个人隐私问题在大数据时代显得越来越突出,因为稍不注意,我们就有可能成为“透明人”,一言一行完全暴露或为人掌握。

有关个人数据安全的案例,具有典型性的是“罗维邓白氏”公司案件,在 2012 年,中央电视台“3·15 晚会”曾经对该案件进行过报道。该公司对外号称拥有包含 1.5 亿条准确个人数据的数据库,该公司不仅从网络上广泛地收集个人数据,而且还向合作公司购买大量的个人数据,其个人数据信息库的内容涉及银行账户、住址、邮箱、消费记录、家庭成员、手机号码、浏览的网页、行为数据等极广泛和敏感的内容。

2. 公司数据安全案例

同样,通过对公司员工的行为数据进行分析,只要有足够的数据,目标公司也将毫无秘密可言。

现在很多公司通过其平台拥有大量的用户数据,如拥有交易数据和信用数据,拥有用户关系数据和基于此产生的社交数据,拥有用户搜索表征的需求数据等。这些数据一旦泄露,对用户而言也将成为巨大的安全隐患。公司数据安全隐患的典型案例有:2011 年 12 月 21 日,某软件官方发布了通知,称某网站 600 余万用户数据库发生泄密,几天后该网站承认确实发生泄密并公开予以道歉。几天以后,某社交平台 4000 万的用户资料又爆出泄露,账号、密码、邮箱竟然采用明文予以保存。发生泄露的用户数据竟然是采用明文存储的密码,如此漠不关心地对待用户的个人隐私,不能不说太不负责任。

3. 国家数据安全案例

在大数据时代,大数据已经成为国家的核心资产,是一种重要的“矿产资源”。国外的数据公司及情报部门无时无刻不在窥视我国的大数据资源,通过这些数据他们可以分析出我国政治、经济、社会的基本运行情况、社会舆论等我们可能根本就没有注意到的细节,并加以利用。从经济利益来看,这些可以帮助他们获得超额的利润,甚至可能控制国家的经济命脉。从政治因素来看,大数据也极可能被他们用于从事破坏国家安全、破坏领土完整、颠覆政权的活动。国外敌对势力一旦拥有了海量的大数据,并拥有了大数据分析挖掘的关键技术,就可能获得极有价值的情报。一些看起来毫无关联的数据可能产生极有价值的情报,因此大数据时代国家的数据安全隐患也急需充分重视。国家数据安全隐患的典型案例要数“棱镜门”事件了,该事件引起了全世界对数据安全的重视。2013 年 6 月,美国中情局曾经的职员斯诺登,曝光了“PRISM”监视项目,曝光了美国对全世界范围的几乎大部分国家进行了间谍活动。在该项目中,美国中情局通过监听民众的通信和网络数据,并分析杂乱无章的大数据来获取重要的情报信息,对民用设施进行了普遍的监听。该项目甚至可以形成及时识别安全威胁的能力,并在潜在威胁来临之前及时采取必要措施。

■ 参考资料 2: 大数据安全自身风险

风险 1 大数据加大了信息泄露风险,大数据囊括了大量的个人隐私,以及各种政府机构、公司行为的细节记录,数据集中存储增加了泄露风险。

风险 2 大数据的应用是人工智能、商业智能、数学算法、自然语言理解、信息技术等多个跨学科领域技术的集成应用,面临较高的技术和管理风险。

风险 3 大数据是更容易被“关注”的大目标,会吸引更多的潜在攻击者,使得黑客成功攻击一次就能获得更多数据,无形中降低了黑客的进攻成本,增加了攻击的“性价比”。

风险 4 数据集中存储会出现将数据乱放的情况,使数据的管理不合标准,影响到安全控制措施的良好运行,也加大了事后追溯的难度,这都将给数据安全带来威胁。

风险 5 黑客可以利用大数据挖掘和分析技术进行更加精准的网络攻击,搜集企业或个人的电话、家庭住址、企业信息防护措施等信息,提取网络攻击所需的情报。同时,大数据的价值密度低,黑客可以将攻击隐藏在大数据中,给安全预警分析带来了很大困难。

■ 参考资料 3: 与数据相关的法律法规

数据安全相关的立法行动在世界范围内已如火如荼地展开,全球已有上百个国家通过立法保护个人数据隐私。早在 1995 年,欧盟针对数据隐私的问题通过了一项立法法案——《数据保护指令》(DPD)。2016 年,欧洲议会通过了打磨 4 年的《一般数据保护法案》(GDPR),取代并且革新了《数据保护指令》,其适用范围既包括欧盟成员国境内企业的个人数据,也包括欧盟境外企业处理欧盟公民的个人数据。《一般数据保护法案》在 2018 年正式生效后一度被称为“史上最严个人信息保护法案”,至今仍是全球个人数据安全立法中极具标志性的一部法案。

美国迄今为止其实并没有统一的综合性个人信息保护法,不过高科技经济聚集地加州,曾在 2018 年通过了关于加州居民隐私权和消费者保护的州法规 CCPA,在其基础上,又于 2020 年再次通过了加州权法案 CPRA,二者共同构建了美国加州隐私保护法的主要制度框架,整体看来,这是北美版图上目前最具有典型意义的州隐私立法。^①

我国的数字化进程发展很快,我国对于数据安全的法规体系已逐步趋于完善,形成了以网络安全法、数据安全法和个人信息保护法为基础的法律框架体系,构建了统一协调的整体,并为全球数字治理贡献出中国方案。

2017 年 6 月 1 日起施行的《中华人民共和国网络安全法》是我国首部全面规范网络空间安全管理方面问题的基础性法律,包含的内容十分丰富,一共包括七章七十九条,包含网络运行安全、关键信息基础设施的运行安全、网络信息安全等内容。该法在数据(包括个人信息)安全与保护上也有诸多规定,诸如第四十至四十五条。

2021 年 6 月 10 日,经全国人民代表大会常务委员会审议,通过了《中华人民共和国数据安全法》,该法自 2021 年 9 月 1 日起施行。该法以总体国家安全观为指导,坚持统筹

^① 摘自《数据安全法:让数据真正成为数字经济发展的流动血液》,石承泰,《国际品牌观察》2021 年第 26 期。

发展与安全的原则,明确了一系列数据安全制度,规定了数据处理主体的数据安全义务,并就政务数据安全与开放提出了相关要求。此外,它还明确了主管部门的职责及违规的法律责任。该法是一部数据安全领域基础性、框架性的法律,为后续各类数据领域配套制度、规范及标准的制定提供了依据。

2021年8月20日,第十三届全国人民代表大会常务委员会第三十次会议表决通过了《中华人民共和国个人信息保护法》。该法自2021年11月1日起施行,共包括八章七十四条。在有关法律的基础上,该法进一步细化、完善个人信息保护应遵循的原则和个人信息处理规则,明确个人信息处理活动中的权利义务边界,健全个人信息保护工作体制机制。

六、教学参考案例

■ 参考案例

信息社会的数据安全策略

上海市北虹高级中学 冯聪

(1课时)

1. 学科核心素养

- 能够敏锐感觉到信息的变化,分析数据中所承载的信息,采用有效策略对信息来源的可靠性、内容的准确性、指向的目的性作出合理判断,提升数据安全意识。(信息意识)

- 掌握常用的数据安全的防护措施和手段,在实际应用中根据各种不同的需求,合理使用并有所创新。(数字化学习与创新)

- 遵守信息法律法规,信守信息社会的道德与伦理准则,在面对各种网络数据和信息时,能够理性判断,自觉规范自己的信息行为。(信息社会责任)

2.《课程标准》要求

- 认识到信息系统应用过程中存在的风险,熟悉信息系统安全防范的常用技术方法,养成规范的信息系统操作习惯,树立信息安全意识。

- 在日常生活与学习中,合理使用信息系统,负责任地发布、使用与传播信息,自觉遵守信息社会中的道德准则和法律法规。

- 结合案例,认识数据丢失的风险,利用实时备份与定时备份、全量备份、增量备份与差异备份等多种方法进行数据备份。

3. 学业要求

学生能预判可能存在的信息泄露等安全风险,掌握信息系统安全防范的常用技术方法;认识信息系统在社会应用中的优势及局限性,能够自觉遵守相关法律法规与伦理道德规范;能根据需要,主动选用数字化工具开展自主或协作学习,创造性地解决问题。

4. 教学内容分析

通过之前的学习,学生已经了解了数据的价值,并学会对数据进行初步的需求分析与数据管理,为本节课的学习准备了理论基础。同时,数据安全作为信息安全的重要组成部分,两者在学习内容上有一定的相似,学习了信息安全之后再学习本课内容或有温故而知新的效果。

5. 学情分析

在认知水平方面,学生已经具备了一定的观察、实践和思考分析能力,并且通过之前的学习对项目式学习已有初步的体验,能够按照学习要求选择合适的数字化工具、借助信息化资源开展自主探究学习和团队协作学习。

6. 教学目标

- 认识数据丢失和数据泄露的风险,在日常工作生活中采取措施降低风险发生的概率。
- 了解什么是数据安全以及威胁数据安全的主要因素,从而能尽量避免数据安全事故的发生。
- 了解数据安全策略。
- 掌握基本的数据安全防范措施,在不同的应用场景下采取相应的安全防范措施。
- 树立数据安全防范意识,养成良好的数据使用习惯,规范信息行为,遵守信息法律法规。

7. 教学重难点

- 教学重点:数据安全概念、威胁数据安全的因素、数据安全策略、数据安全的常用防护手段。

- 教学难点:制定合适的数据安全策略、数据加密。

8. 教学准备

准备教学课件、视频素材、学案、“密码安全鉴定器”等工具程序。

9. 教学策略分析

根据学科特点、教材内容和学生的认知规律,在整堂课的教学过程中主要采用教师授导与学生探究相结合的教学策略。

- 结合具体案例进行新课导入,同时以视频形式增强冲击感,激发内在学习动力。
- 以自主学习的方式完成对数据安全概念及相关知识内容的学习,并通过练习加以巩固。

- 围绕项目主题开展实践活动,通过小组讨论和上机实践的方式完成活动内容。
- 结合实际的生活体验,进行信息安全与道德教育,培养社会责任意识。

10. 教学环境

网络机房、广播软件。

11. 教学过程设计(见表 3.2)

表 3.2 教学过程设计表

教学环节	教学内容	学生活动	设计意图
情境导入	播放视频:某平台 5.38 亿用户数据在暗网出售。 热门盘点:2020 年上半年全球重大数据泄露事件。 提出问题:这些重大数据泄露事件会对社会造成哪些危害? 事件是如何造成的?	结合自己的认知,思考并回答问题,从中认识数据丢失和泄露的风险	通过具体案例,让学生认识到数据泄露的危害,从而提高数据安全意识,并激发进一步学习的欲望和兴趣
自主学习	组织学生自主学习数据安全的概念、威胁数据安全的主要因素、数据安全策略、数据安全的防护手段等内容	阅读教材,自主学习,并完成练习	通过自主学习,了解数据安全及相关的基本知识,为进一步学习打下基础
项目介绍	项目情境:在学习过程中,考试是一种常用的评价方式。随着无纸化考试的逐渐兴起,对数据安全也提出了越来越高的要求。为了确保考试的顺利完成,请大家帮助考试系统管理员一起分析影响在线考试顺利完成的安全问题,并制定合理的数据安全策略	结合在线考试系统的安全维护问题,分组讨论可能造成考试数据丢失和相关数据泄露的主要因素	运用生活场景,引导学生进行思考和分组讨论,提升安全意识
活动 1	制定在线考试系统的数据安全策略: (1) 考试前; (2) 考试中; (3) 考试后	根据活动要求,完成活动 1,并分享成果	通过思考和讨论,发挥学生的主观能动性,培养解决实际问题的能力
活动 2	数据加密: (1) 什么是数据加密; (教师演示解密过程) (2) “密码”和“密钥”的区别; (3) 检测自己常用的密码; (4) 设置一个高强度密码; (5) 归纳设置安全密码的原则	根据活动要求,完成活动 2,并分享实践结果	通过观察和实践,了解加密和解密的过程,提高对数据本身的安全防护意识,同时认识到未经允许的“解密”是一种违法行为
课堂总结	对整堂课的知识内容进行回顾与总结	梳理知识内容,分享学习体会	总结本节课所学知识

【活动记录单】

信息社会的数据安全策略

一、练习

1. 数据安全包含两方面的内容:

一是_____的安全,主要是指采用复杂的_____对数据进行主动保护,如身份认证等,以防止数据被泄露、篡改;

二是_____的安全,主要是指采用先进的_____手段对数据进行主动防护,如通

过磁盘阵列等,避免因数据丢失造成损失。

2. 威胁数据安全的主要因素:

第一,_____;
第二,_____;
第三,_____。

3. 数据安全策略的主要内容:(连线题)

数据的保密性	对数据的内容和传播具有控制能力
数据的完整性	数据在传送过程中不被修改、不被破坏和不丢失
数据的可用性	数据未经授权其内容不会显露
数据的可控性	数据发送方和接收方都需承认曾经完成的操作
数据的不可否认性	得到授权的用户可随时访问所需数据

4. 数据安全的防护手段:(连线题)

数据加密	对数据的访问者、访问时间、访问行为进行详细的审核和记录
访问控制	通过分布式存储、冗余和恢复来实现数据的容灾安全性
安全审计	安装运行杀毒软件、防火墙,进行入侵检测等
备份与还原	对原来为明文的数据按某种算法进行处理,使其成为不可直接读取并理解的乱码
防止病毒感染和黑客入侵的传统手段	对数据库、数据表的访问行为进行检测和判断
加强安全管理	不断完善和升级数据库系统,加强对数据的管理控制,以及规范对数据或数据库的操作等

二、项目活动

项目情境:在学习过程中,考试是一种常用的评价方式。随着无纸化考试的逐渐兴起,对数据安全也提出了越来越高的要求。为了确保考试的顺利完成,请大家帮助考试系统管理员一起分析影响在线考试顺利完成的安全问题,并制定合理的数据安全策略。

活动 1:制定在线考试系统的数据安全策略。

请完成表 3.3。

表 3.3 在线考试系统的数据安全策略

	影响数据安全的因素	采取的措施
考试前		
考试中		
考试后		

活动 2: 数据加密。

要求: 使用“密码安全鉴定器”工具, 完成此活动内容。

(1) 什么是数据加密?

(2) “密码”和“密钥”有何区别?

(3) 检测一个自己常用的密码。

被检测的密码中, 数字 _____ 个, 符号 _____ 个, 大写字母 _____ 个, 小写字母 _____ 个。

密码强度得分: _____。

(4) 设置一个高强度密码。

密码组成: _____;

密码强度得分: _____。

(5) 思考: 是否存在 100 分的密码? 设置安全密码的原则是什么?

第二节

数据备份与还原的实现

一、教学目标与重点

教学目标：

- 掌握数据备份与还原的基本方法。

教学重点：

- 能够利用实时备份与定时备份、全量备份、增量备份、差异备份等多种方法进行数据备份。

二、教学说明与建议

本节的主要内容是数据备份与还原,建议先由教师讲解数据备份的方法并进行比较,以一两个具体的备份案例和备份软件为主讲解几种常用备份方法的实际操作方法,然后让学生按照项目活动的要求进行不同备份方式的备份与还原操作。

三、项目实施与评价

1. 核心概念精解

本节的核心概念是:数据备份与还原。数据备份与还原是指通过分布式存储、冗余备份和恢复来实现数据的容灾安全性,是一种可用性机制,也就是保护数据在任何情况下都不会丢失,当需要时,获得授权的用户可以随时访问到所需要的数据。由于现实世界没有绝对的数据安全,因此数据备份与还原就成为人们防止数据丢失的最后一道防线。

数据备份是容灾的基础,是指为防止系统出现操作失误或系统故障导致数据丢失,而将全部或部分数据集合从应用主机的硬盘或阵列复制到其他的存储介质的过程。人们对数据备份都或多或少有所了解,但在认识上仍然存在一些误区,包括认为复制就是备份、以硬盘冗余备份代替备份、只备份数据文件、不重视备份数据的保管等,因此需要在概念理解的同时,带领学生走出误区。

从备份的数量角度看,数据备份可以分为全量备份、增量备份和差异备份三种。可以通过实践和比较掌握这三种备份方式的概念和特点。三种备份方式的对比如表 3.4 所示。

表 3.4 全量备份、增量备份和差异备份的对比

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

从备份的时间角度看,数据备份可以分为定时备份和实时备份。

一般为数据建立备份的过程如图 3.1 所示。

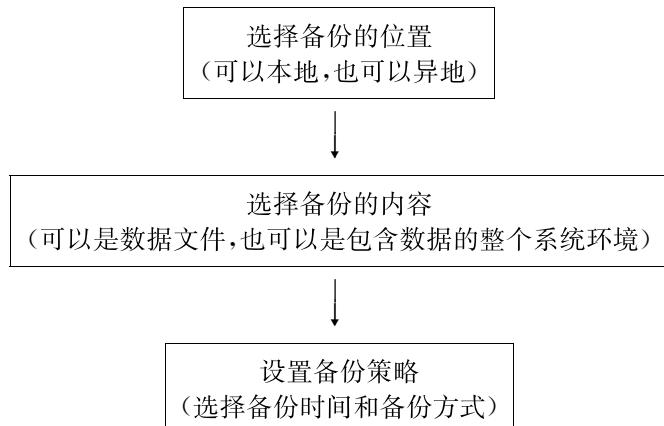


图 3.1 备份过程

数据还原就是通过技术手段,将保存在存储介质上的数据进行抢救和恢复的技术。

2. 项目活动的具体实施

项目任务 2:不同备份方式的数据还原。

该项目活动的开展,需要建立在充分认识数据备份的基础上:了解数据备份的重要性和必要性,纠正正在数据备份上的一些认识误区,掌握数据备份的不同方式及其差异,能够根据需要选择恰当的备份策略完成备份操作。在此基础上,实现用不同方式备份的数据的还原操作。

3. 项目活动的评价

本节的项目评价主要包括:考查学生能否运用具体的备份工具完成数据备份和还原的任务;考查学生对不同备份方式之间差异的掌握程度,能否根据具体情况,选择恰当的备份策略,实施备份和还原操作。

评价建议:主要采用过程性评价。在学生实施备份和还原操作的过程中,重点对学生备份方式的选择和实施进行评价。

四、作业练习与提示

■ 题目描述

1. 日常生活中,我们需要经常对哪些数据进行备份?
2. 备份数据有哪些方法?它们各有什么特点?在什么情况下使用?
3. 如何将备份的数据还原?
4. 如何运用实时备份与定时备份、全量备份和增量备份及差异备份等备份方式,防止在线考试系统中的数据丢失?
5. 对于手机中的联系人信息、拍摄的照片和视频、社交平台上的各种往来消息和文件,你都能利用相应的手段备份并在需要的时候还原吗?

■ 作业提示

1. 该问题主要引发学生对数据备份重要性的认知,并能结合日常生活,意识到哪些数据比较重要,引导学生直接将所学的知识与实际应用结合。
2. 该问题的目的是让学生带着问题学习后续的内容,对应的是全量备份、增量备份和差异备份三种方法的特点、异同和应用,以及实时备份和定时备份的方法。
3. 该问题主要结合后面学习的需要而提出,对应不同备份方法有不同的数据还原方式。
4. 该问题是让学生根据本节学习的内容,针对在线考试系统中的数据,采用不同的备份方式进行实际备份的操作,并进行比较。
5. 该问题是扩大数据备份和还原的应用面,让学生通过做基于手机平台的重要数据备份,进一步加深对几种备份方法的操作和理解。

五、教学参考资源

■ 参考资料:基于 Windows 系统的数据备份方法

建立备份的简单过程:

1. 一般建议将备份保存在外部存储设备上(如图 3.2 所示)。
2. 选择需要备份的数据内容,可包括数据文件,也可以是系统环境(如图 3.3 所示)。
3. 若要定时进行备份,可设置相应的备份计划(如图 3.4 所示)。



图 3.2 选择备份的位置

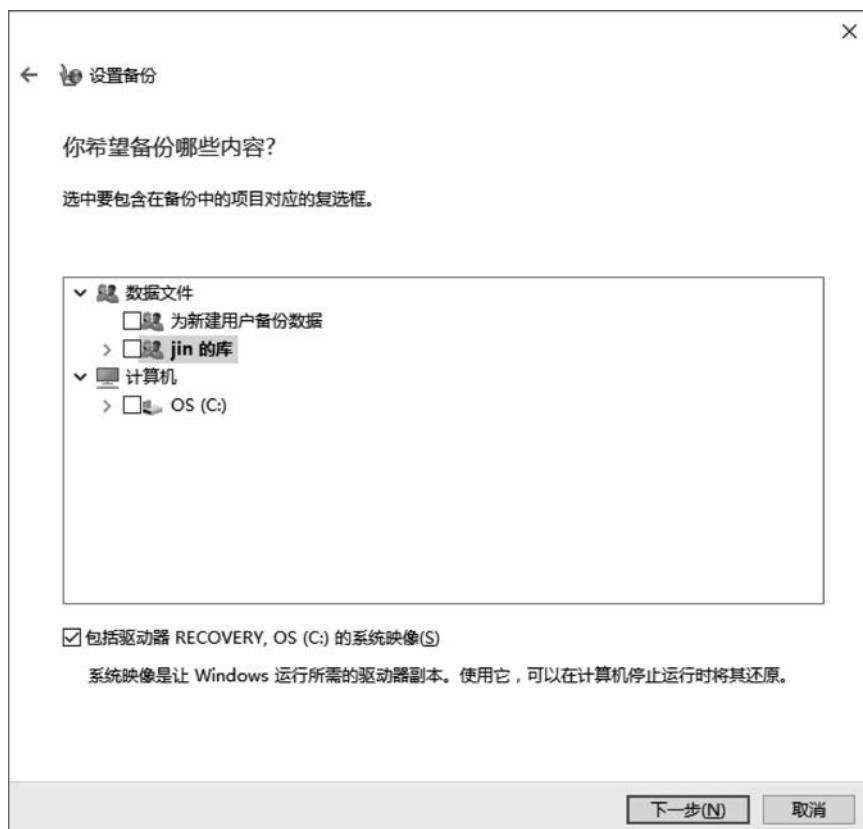


图 3.3 选择备份的内容

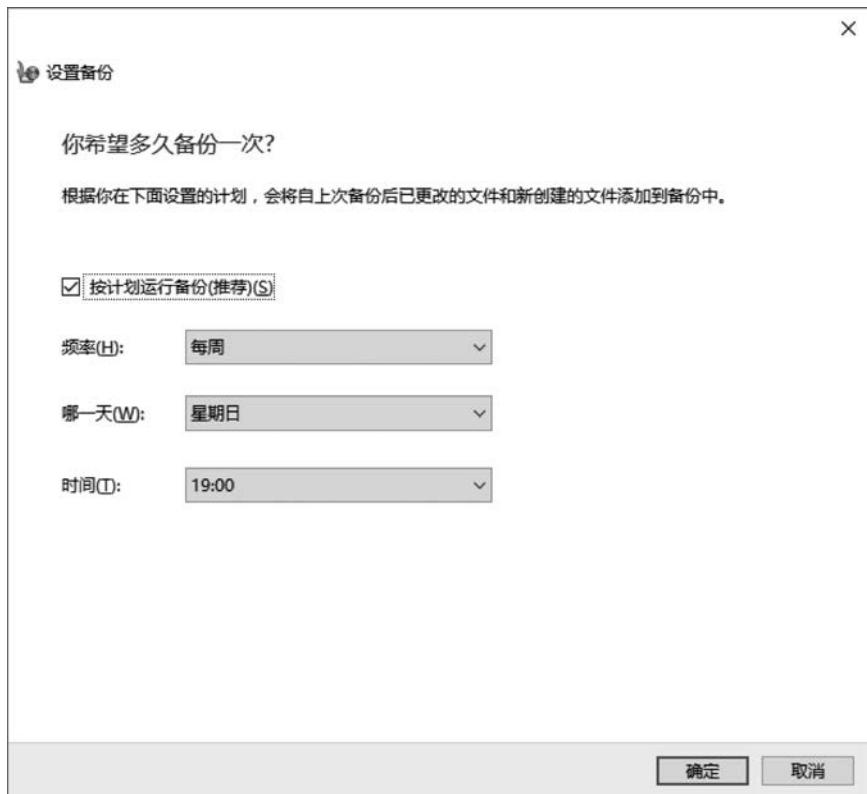


图 3.4 设置备份计划

六、教学参考案例

■ 参考案例

数据备份与还原的实现

上海市北虹高级中学 冯聪
(1课时)

1. 学科核心素养

- 掌握常用的数据安全的防护措施和手段,在实际应用中根据各种不同的需求,合理使用并有所创新。(数字化学习与创新)

- 通过判断、分析与综合各种数据资源,运用合理的算法形成数据安全备份的方案。(计算思维)

- 遵守信息法律法规,信守信息社会的道德与伦理准则,在面对各种网络数据和信息时,能够理性判断,自觉规范自己的信息行为。(信息社会责任)

2.《课程标准》要求

- 结合案例,认识数据丢失的风险,利用实时备份与定时备份、全量备份、增量备份与差异备份等多种方法进行数据备份。

- 认识到信息系统应用过程中存在的风险,熟悉信息系统安全防范的常用技术方法,养成规范的信息系统操作习惯,树立信息安全意识。

- 在日常生活与学习中,合理使用信息系统,负责任地发布、使用与传播信息,自觉遵守信息社会中的道德准则和法律法规。

3. 学业要求

学生能够认识数据备份的重要性,能根据需要及时备份与还原数据,确保数据安全;能根据需要,主动选用数字化工具开展自主或协作学习,创造性地解决问题。

4. 教学内容分析

通过之前的学习,学生已经了解了数据的价值,并学会对数据进行初步的需求分析与数据管理,为本节课的学习奠定了理论基础。同时,数据安全作为信息安全的重要组成部分,两者在学习内容上有一定的相似,学习了信息安全之后再学习本课内容或有温故而知新的效果。

5. 学情分析

在认知水平方面,学生已经具备了一定的观察、实践和思考分析能力,并且通过之前的学习对项目式学习已有初步的体验,能够按照学习要求选择合适的数字化工具、借助信息化资源开展自主探究学习和团队协作学习。

6. 教学目标

- 掌握数据备份与还原的基本方法。

- 了解不同的数据备份方式及各自的优缺点和应用范围,并根据实际情况选择最合适的备份方式。

- 树立数据安全防范意识,养成良好的数据使用习惯,规范信息行为,遵守信息法律法规。

7. 教学重难点

- 教学重点:数据备份、数据备份的方式、建立备份的过程、数据还原。

- 教学难点:数据的备份与还原。

8. 教学准备

准备教学课件、学案、AOMEI_Backupper_5.9_Portable 等工具程序。

9. 教学策略分析

根据学科特点、教材内容和学生的认知规律,在整堂课的教学过程中主要采用教师授导与学生探究相结合的教学策略。

- 以自主学习的方式完成对数据备份相关内容的学习。

- 围绕项目主题开展实践活动,以小组分工合作的形式完成活动任务。

- 结合实际的生活体验,进行信息安全与道德教育,培养社会责任意识。

10. 教学环境

网络机房、广播软件。

11. 教学过程设计(见表 3.5)

表 3.5 教学过程设计表

教学环节	教学内容	学生活动	设计意图
课堂导入	活动回顾： 活动 1：制定在线考试系统的数据安全策略。 活动 2：数据加密。 引出本节活动： 活动 3：数据备份与还原的实现	回顾上节课的活动，引出活动 3	承上启下，在加强安全教育的同时引入新课
自主学习	了解数据备份及相关的认识误区	阅读教材，自主学习	通过自主学习，了解数据备份的相关知识，为进一步开展活动打下基础
项目活动	数据备份与还原的实现。 (1) 阅读故事，归纳不同备份方式的特点； (2) 上机实践，比较不同备份方式的工作效率； (3) 思考对不同备份方式的选择	根据活动要求，分组合作完成活动 3，并分享实践结果	了解数据备份，学习和掌握防护数据安全的重要手段
活动小结	(1) 数据备份方式； (2) 建立备份的简单过程； (3) 数据还原	聆听，体会数据安全防护的重要性	提升对数据安全的防护意识
知识延伸	(1) 数据备份系统； (2) 冷备份与热备份； (3) 云备份	聆听与思考	了解更多有关数据备份的内容，拓展知识面
课堂总结	活动回顾与总结	梳理知识内容，分享学习体会	总结项目内容，融合信息安全教育

【活动记录单】

数据备份与还原的实现

活动 3：数据备份与还原的实现。

(1) 阅读故事，完成表 3.6。

从前有一位账房先生，他每天要记很多账单。为了保证账本的安全，他找来三个徒弟，交给他们一份“底账”，要求他们每天对账本做备份，并且每个月底交一份月总账，以供查账。

三个徒弟各展所长，分别采用了不同的做法：

大徒弟宅心仁厚、成熟稳重。他每天都规规矩矩把师父的账单重新抄录一份。这样做的好处就是每天都是一份完整的账本，每一个备份的账本都可以直接使用，坏处则是每天要花费很多时间进行抄录，并且需要费很多纸、墨水以及存账本的柜子。

二徒弟聪明伶俐、人小鬼大。他觉得大师兄的方法太累且耗时，不如每天只记录账本上新增的信息。于是他每天总是第一个写完，不仅节省时间，而且为店里节省了大量的纸张、墨水等成本。不过每到月底，他总是最忙碌的，他需要将每天的账目数据抄录到一起才能组成一个完整的账本。

三徒弟心思缜密、粗中有细。他借鉴了两位师兄的做法,进行折中创新,他只记录每天相比于“底账”变化的信息,比如周一买菜花了十文钱,那么周二就把“买菜花费十文”记录下来;周二买肉花了三十文钱,周三就把周二的“买肉花费三十文”和周一的“买菜花费十文”都记录下来;……。到月底的时候,他把最近一次记录的数据和“底账”合在一起,就是完整的账本了。这种方式在记账速度、纸墨使用量、账单查看等方面都是一种较为折中的方式。

账房先生见三个徒弟的做法各有千秋,后来为了便于记忆,便给这三种做法起了名字,分别叫全量备份、增量备份和差异备份。

表 3.6 三种数据备份方式的比较

	全量备份	增量备份	差异备份
备份内容			
备份速度			
还原速度			

(2) 使用 AOMEI_Backupper_5.9_Portable 工具,对电脑某一磁盘分区进行备份和还原,并完成表 3.7。(要求:每 4 人一组,各人选择不同的磁盘分区)

表 3.7 两种备份方式及还原效率的比较

	全量备份	增量备份
备份磁盘分区		_____ 盘
备份所需时间		
还原所需时间		

(3) 思考:在日常生活中,如何根据不同的情况选择不同的数据备份方式?

数据分析

一、本章学科核心素养的渗透

1. 信息意识

通过本章的学习,学生将掌握分析数据的基本过程与方法,能够根据特定问题解决的需要,坚持问题导向,选用合适的方法对数据进行分析与可视化表达,提取有用信息,形成结论;能认识有效分析数据对获取有价值信息的作用与意义,认识数据分析技术对人类社会生活的影响,认识数据分析对决策的重要价值。

2. 计算思维

本章教学增强学生的问题意识,要求学生能够确定学习和生活中的数据问题,坚持问题导向,提出解决方案,能根据不同的数据分析要求,采用合适的方法提取数据,能正确选用数据分析方法和分析工具分析数据,并能对分析结果进行合理解释和恰当呈现;采用 Python 程序语言实现数据提取以及数据分析的各种方法。学生在写程序过程中训练运用计算机处理问题、抽象特征、建立结构模型、合理组织数据等计算思维的能力。

3. 数字化学习与创新

本章教学提供数字化学习资源,学生将利用数字化学习工具完成学习任务,提高学习质量,协作解决问题,创新问题决策,反思与完善学习成果,并在此过程中进行自主或协作探究,能够根据需要合理选择常见的数字化资源与工具,提高学生自主学习的能力。

4. 信息社会责任

在本章的学习过程中,学生根据数据分析的目的和意图,选取真实、准确、合适的数据,选用合适的数据分析与可视化方法和工具,提高数据的识别度,使之更符合受众需求;懂得对数据分析时使用的数据要慎重对待处理。

二、本章知识结构

本章遵循普通高中信息技术课程标准,依据学分和课时规定,紧扣学科概念体系,以“上海市旅游景点数据分析”为项目主题展示数据分析的全过程,并在该过程中介绍常用的数据分析方法。内容分为两节,围绕数据准备、数据分析方法与呈现展开。

第一节“数据准备”,通过对某知名旅游网站上采集的上海 Top50 的景点数据进行数据预处理与数据提取,使学生了解数据准备的意义以及数据提取的常用方法。

第二节“数据分析方法与呈现”,通过对上海 Top50 景点数据进行分析,使学生掌握平均分析法、分组分析法、对比分析法、相关分析法等常用数据分析方法以及常用图表的绘制方法和应用场景,并且结合第一节的内容展示了数据分析的全过程。

三、本章项目活动设计思路

本章的项目活动设计为上海市旅游景点的数据分析:使用从某知名旅游网站上采集的上海 Top50 的景点数据,结合上海市文化和旅游局官网 A 级景点数据对其进行数据预处理,并为了解上海旅游景点的区属分布提取有效数据,采用常用的数据分析方法分析上海旅游景点数据,使用图表呈现分析结果,以此了解上海旅游景点的概况,并根据分析结果进行大致的推理,决定是否推荐游玩某景点。

本章项目活动设计思路如图 4.1 所示。

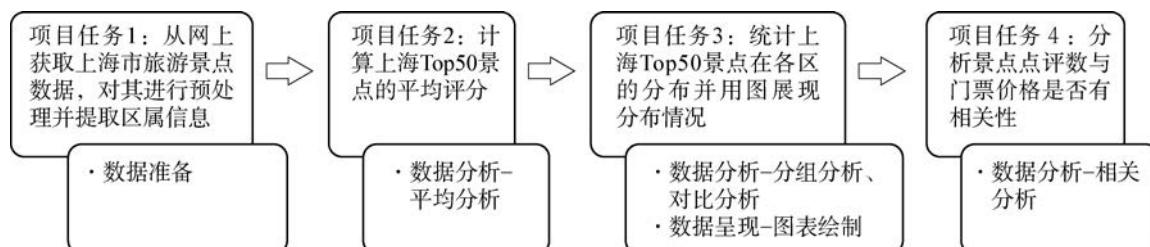


图 4.1 本章项目活动设计思路

项目任务 1:通过在分析前对获取的上海市旅游景点原始数据进行处理,了解数据质量的重要性,对缺失数据进行填充,整合数据并从地址信息中提取区属信息。

在该任务中,要求学生了解数据准备的意义,了解可能存在的数据质量问题,了解缺失值的处理方法,能根据需要选择恰当的方法对采集到的原始数据进行数据提取。

项目任务 2:使用平均分析法计算上海 Top50 景点的平均评分。

在该任务中,要求学生掌握平均分析法,了解平均数的分类,掌握简单算术平均数、加权算术平均数、中位数、众数等平均数的计算方法。

项目任务 3:使用分组分析法、对比分析法统计上海 Top50 景点在各区的分布并用图展示分布情况,以此了解上海旅游景点的概况。

在该任务中,要求学生掌握分组分析法和对比分析法,掌握分组标志的分类、组距的意义及划分方法;掌握对比的意义、横比与纵比的区别,了解统计指标的分类。另外,掌握各类图表的绘制及适用情境。

项目任务 4:通过计算了解上海地区景点点评数与门票价格之间是否有相关性,从而可以根据某景点的基本情况进行大致推理,以此决定是否推荐游玩该景点。

在该任务中,要求学生掌握相关分析法,区分正相关与负相关,掌握相关系数的意义,了解皮尔森相关系数的计算方法。

四、本章课时安排建议

本章教学建议用 7 课时完成,具体参见表 4. 1。

表 4. 1 课时安排建议表

节名	建议课时
第一节 数据准备	2 课时
第二节 数据分析方法与呈现	5 课时

第一节 数据准备

一、教学目标与重点

教学目标:

- 了解数据准备的意义,了解可能存在的数据质量问题,了解缺失值的处理方法。
- 根据需要,选择恰当的方法对采集到的原始数据进行数据提取。
- 了解常用数据分析工具的适用场景。

教学重点:

- 理解数据质量的优劣对数据分析结果的影响并能适当举例说明,了解数据质量的评估标准,了解缺失值填充的常用方法,能够发现在生活中采集到的数据中的数据质量问题,并对其进行处理。
- 了解数据提取的常用方法,掌握 Excel 和 Python 提取数据的方法,能够根据数据分析的任务提取相应的数据。

二、教学说明与建议

本节的主要内容为数据准备与数据提取,都与操作相关,需要在课堂上结合具体的软件进行教学。除了项目活动中涉及的数据质量问题,还可以举其他例子(也可以请学生举例),缺失值处理根据教科书简单介绍。本节中并未规定采用何种工具进行数据预处理,可以使用 Excel 对原始数据做处理。若教学时间不充足,可以简单地请学生先观察原始数据的问题再观察处理以后的数据;若教学时间充足,可以结合本节教学参考资源中体检数据的例子使用 Python 对数据做预处理。本节叙述了 3 种数据提取方法,除了完成探究活动以外,还可以采用 Excel 或 Python 对本节教学参考资源中的体检数据进行提取,可以提出不同的条件筛选出合适数据,或通过计算提取出 BMI 数据等。除了 Excel 和 Python 以外,建议对其他常用的数据分析工具做简单介绍(可以选择 1~2 个工具,对其界面进行截图展示)。

本节 2 课时安排建议:“数据预处理”1 课时,“数据提取”1 课时。

三、项目实施与评价

1. 核心概念精解

(1) 数据质量问题与评估标准

数据作为数据分析要处理的对象,本身的优劣对分析结果具有很大的影响。高质量的数据是分析结果准确的重要保证。

例如,A 企业需要重新调整员工的薪资结构,使员工的薪资结构在同行业中具备竞争力。假设该行业中研发人员的平均月薪为 4 500 元,研发人员的平均年龄为 35 岁。当企业研发人员的平均工资超出行业平均薪酬的 15%,并且低于 30% 时,可以认为该企业研发投入处于相对合理的区间;反之,则需要调整。当企业研发人员的平均年龄超过行业平均水平的 10% 时,则该企业需要加强年轻研发人员的培养。A 企业研发人员的年龄和薪资结构如表 4.2 所示。

表 4.2 A 企业研发人员工资单

姓名	性别	月薪(元)	年龄(岁)
张珊	男	5 500	37
王於	男	4 500	38
李偲	女	5 000	28
周姐	女	6 000	38
吴柳	男	7 000	25
郑歧	男	5 500	38
平均值		5 583. 333 33	34

根据表 4.2,A 企业研发人员年龄结构合理,薪资水平适中且偏上,无需任何调整。然而实际上由于人力资源部门的误操作,使上表中李偲与吴柳的年龄数据发生差错,真实的数据如表 4.3 所示。

表 4.3 A 企业研发人员工资单(实际)

姓名	性别	月薪(元)	年龄(岁)
张珊	男	5 500	37
王於	男	4 500	38
李偲	女	5 000	38
周姐	女	6 000	38
吴柳	男	7 000	45
郑歧	男	5 500	38
平均值		5 583. 333 33	39

纠正了两位员工的年龄错误后可以发现,实际上 A 企业的研发人员整体年龄偏大,急需补充新鲜血液,培养年轻研发人员。可见,准确的数据有利于得出准确的结果,错误的数据会影响分析的结果,进而影响决策,产生严重后果。

数据质量问题是指在数据的采集、数据的传输、数据的存储,甚至业务需求分析等方面都有可能导致质量问题。不同领域、不同业务、不同应用的数据质量需求不尽相同。评估数据是否达到预期设定的质量要求,可以通过以下六个方面来判断。

准确性问题:数据不正确或含噪声(包含错误或存在偏离期望的值)。例如社交软件中,如果不填写个人生日,那么生日默认值为 1 月 1 日,这个数据是不正确的。

唯一性问题:数据重复或属性重复。在一些拼接的数据表中经常会有这个问题。

完整性问题:数据不完整(缺少属性值或某些感兴趣的属性)。例如销售记录中,缺少顾客的年龄、性别。

一致性问题:数据记录不规范和数据不符合逻辑。例如:地区类的标准编码格式为“北京”而不是“北京市”;身高数据单位为米,“1. 67”是合理的,“167”则不合理。

时效性问题:数据过时。例如电商网站中用户的购买数据,使用时间过久的数据分析出的结论会有问题。

相关性问题:数据分析时需要使用的一些数据缺失,导致无法分析。

另外,数据质量问题还包括可信性、可解释性等其他问题。可信性问题指的是数据不可信赖,例如数据库中曾经有过一些错误,更正以后,用户可能还是不相信该数据库中的数据。可解释性问题指的是数据不可解释,例如某些数据使用行业内编码,其他行业不知如何解释这些数据。

(2) 数据预处理

为了提高数据的质量,应该在数据分析之前对数据进行预处理,使后续的数据分析得

到尽可能正确的结果。可以通过数据清洗来进行数据预处理。数据清洗有两个阶段。第一阶段是数据选取,原则是尽可能赋予属性名和属性值明确的含义、统一多数据源的属性值编码、去除重复、去除可忽略字段以及合理选择关联字段;第二阶段是去除数据中的噪音、填补缺失数据、光滑噪声数据、识别或删除离群点以及纠正不一致的数据。

项目活动中只涉及缺失值处理,除此以外还应该对所含噪声数据进行处理。可以运用历史数据或统计收集到的所有数据来清除噪声数据。最简单的方法是找到这些噪声数据的记录并删除,其缺点是如果只有一个属性上的数据需要删除或修正,将整条记录删除会丢失大量有用的、干净的信息。因此,也经常采用合理的数据“光滑”原始数据(称为数据光滑技术),去除噪声,例如使用平均值填写某个缺失或者离群的属性。

(3) 数据提取

“工欲善其事,必先利其器。”通常认为,数据是数据分析要处理的对象,数据分析技术流程应该从对数据的分析开始,需要根据特定应用的需求,从数据中抽取相关的有效数据,同时尽量摒除可能影响判断的错误数据和无关数据。

对于一个特定的应用来说,并不是数据越多越好,因此需要从数据源中抽取有效的数据,这个过程称为数据提取。数据提取本质上就是对数据源中的数据进行选择加工的过程。常用的数据提取方法有数据筛选、条件判断、函数提取、数据排序、数据透视等。

数据筛选是指在数据源的数据表中将满足查询条件的记录选择出来。查询条件一般通过逻辑表达式书写。

条件判断是指根据某一字段的设定判断条件、根据判定结果返回指定的字段或常量。

函数提取是指通过特定功能的函数对指定的一个或一组字段进行计算,生成新的字段。

数据排序是指根据数据记录的某一字段或某几个字段,将数据表中的记录重新排列成递增序列(升序排列)或递减序列(降序排列),形成的数据表是原数据表的重组。在数据处理中排序是最基础的处理方法,经常被使用。

数据透视可以从不同角度观察数据,从而进行数据分析,常见形式是数据透视表与数据透视图。以数据透视表为例,它可以让用户根据不同的分类、不同的汇总方式快速查看各种形式的数据汇总报表。数据透视表通常需要经过多次旋转(将行移动到列或将列移动到行)和其他操作(对数值数据进行分类汇总和聚合,创建自定义计算和公式,对最有用和最关注的数据子集进行筛选、排序、分组和有条件地设置格式等),直到符合数据分析的需求。

2. 项目活动的具体实施

项目任务 1 的具体实施如下:

首先,展示从某知名旅游网站上采集的上海 Top50 景点的原始数据,观察其中的缺失值与不规范数据。在此基础上,介绍数据质量问题。

然后,介绍缺失值的处理方法,填充缺失值、修正表格中有问题的数据,并结合上海市文化和旅游局官方网站 A 级景点数据,对原始数据进行整合。

最后,用 Python 提取各景点的区属信息(直接用书上的代码演示)。也可以在 Excel 中尝试进行区属信息的提取。

3. 项目活动的评价

本节的项目评价主要包括：考查学生对数据质量评估标准与数据质量问题的了解程度；考查学生对缺失值处理以及对数据提取的常用方法的掌握情况。

评价建议：主要采用过程性评价。对学生数据预处理、数据处理的步骤及结果进行评价。

四、作业练习与提示

■ 题目描述

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

■ 作业提示

可以让学生每人提交 2 本书的信息，包括书名、作者、出版社、字数、价格、本月销量等。设计数据表(如表 4.4 所示)让学生填写。

表 4.4 图书信息表

书名	作者	出版社	字数	价格	本月销量

学生提交后，可以让其汇总(人工或写程序都可以)。完成汇总后，让学生观察数据是否有质量问题，如果有，可以让学生进行数据预处理练习。因为数据条目不多，数据预处理可以使用人工方式或 Excel 或 Python 的 pandas 库(可能使用到的操作见“教学参考资源”)。

五、教学参考资源

■ 参考资料 1:Python 的相关操作

1. 导入、导出数据

(1) 使用 pandas 导入数据

`pd.read_csv(filename)`: 从 CSV 文件导入数据。

`pd.read_table(filename)`: 从限定分隔符的文本文件导入数据。

`pd.read_excel(filename)`: 从 Excel 文件导入数据。

`pd.read_sql(query,connection_object)`: 从 SQL 表/库导入数据。

`pd.read_json(json_string)`: 从 JSON 格式的字符串导入数据。

`pd.read_html(url)`: 解析 url、字符串或者 html 文件，抽取其中的表格。

`pd.read_clipboard()`: 从剪贴板获取内容。

`pd.DataFrame(dict)`: 从字典对象导入数据。

(2) 使用 pandas 导出数据

`df.to_csv(filename)`: 导出数据到 CSV 文件。

`df.to_excel(filename)`: 导出数据到 Excel 文件。

`df.to_sql(table_name, connection_object)`: 导出数据到 SQL 表。

`df.to_json(filename)`: 以 JSON 格式导出数据到文本文件。

(3) 示例

示例 1: 使用 pandas 读写 csv 文件。

```
import pandas as pd  
csvpd= pd.read_csv('score.csv')  
print(csvpd)  
csvpd.to_csv('score_backup.csv')
```

“score.csv”的内容如图 4.2 所示。

	name	Math	Physics	English	Python	PE
0	stu1	NaN	73.0	92	82	95
1	stu2	92.0	95.0	88	96	85
2	stu3	90.0	92.0	93	95	90
3	stu4	83.0	93.0	86	85	60
4	stu5	87.0	NaN	70	93	75
5	stu6	95.0	87.0	90	98	80

图 4.2

示例 2: 使用 DataFrame 数据存取-读写 Excel 文件。

```
import pandas as pd  
df= pd.read_excel('stu.xlsx')  
df.to_excel('stu.xlsx',sheet_name= 'stu')  
“stu.xlsx”的内容如图 4.3 所示。
```

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

图 4.3

2. 数据预处理

(1) 重复值检测与处理

```

import pandas as pd
df= pd.read_csv('score.csv')
print(df.duplicated())
print(df.duplicated('name'))
print(df.drop_duplicates())
print(df.drop_duplicates('name'))

```

“score.csv”的内容如图 4.4 所示。

name	Math	Physics	English	Python	PE
stu1		73	92	82	95
stu2	92	95	88	96	85
stu3	90	92	93	95	90
stu4	83	93	86	85	60
stu5	87		70	93	75
stu6	95	87	90	98	80
stu4	83	93	86	85	60
stu2	92	95	88	96	85
stu4	83	93	86	85	60

图 4.4

结果如图 4.5 所示。

```

0    False
1    False
2    False
3    False
4    False
5    False
6    True
7    True
8    True
dtype: bool
      name   Math  Physics  English  Python   PE
0  stu1    NaN     73.0      92     82    95
1  stu2   92.0     95.0      88     96    85
2  stu3   90.0     92.0      93     95    90
3  stu4   83.0     93.0      86     85    60
4  stu5   87.0      NaN      70     93    75
5  stu6   95.0     87.0      90     98    80

```

图 4.5

(2) 缺失值检测

```

import pandas as pd
df= pd.read_csv('score.csv')
print(df)

```

```
print(df.isnull())
```

示例如图 4.6 所示。

	name	Math	Physics	English	Python	PE
0	stu1	NaN	73.0	92	82	95
1	stu2	92.0	95.0	88	96	85
2	stu3	90.0	92.0	93	95	90
3	stu4	83.0	93.0	86	85	60
4	stu5	87.0	NaN	70	93	75
5	stu6	95.0	87.0	90	98	80

	name	Math	Physics	English	Python	PE
0	False	True	False	False	False	False
1	False	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	True	False	False	False
5	False	False	False	False	False	False

图 4.6

(3) 缺失值处理

可以使用 dropna 方法删除含有缺失值的行或列, 示例如图 4.7 所示。

In [25]:	df.dropna(axis=0)
Out[25]:	
	name Math Physics English Python PE
1	stu2 92.0 95.0 88 96 85
2	stu3 90.0 92.0 93 95 90
3	stu4 83.0 93.0 86 85 60
5	stu6 95.0 87.0 90 98 80
In [26]:	df.dropna(axis=1)
Out[26]:	
	name English Python PE
0	stu1 92 82 95
1	stu2 88 96 85
2	stu3 93 95 90
3	stu4 86 85 60
4	stu5 70 93 75
5	stu6 90 98 80

图 4.7

(4) 缺失值填充

可以使用 fillna 方法对缺失值进行填充, 填充时可以采用字符串、前一个数据、后一个数据、算术平均数、中位数等数值。示例如图 4.8 所示。

In [27]:	df.fillna('missing')
Out[27]:	
	name Math Physics English Python PE
0	stu1 missing 73 92 82 95
1	stu2 92 95 88 96 85
2	stu3 90 92 93 95 90
3	stu4 83 93 86 85 60
4	stu5 87 missing 70 93 75
5	stu6 95 87 90 98 80

```

In [28]: df.fillna(method='pad')
Out[28]:
   name  Math  Physics  English  Python  PE
0  stu1    NaN     73.0      92      82    95
1  stu2   92.0     95.0      88      96    85
2  stu3   90.0     92.0      93      95    90
3  stu4   83.0     93.0      86      85    60
4  stu5   87.0     93.0      70      93    75
5  stu6   95.0     87.0      90      98    80

In [29]: df.fillna(method='bfill')
Out[29]:
   name  Math  Physics  English  Python  PE
0  stu1   92.0     73.0      92      82    95
1  stu2   92.0     95.0      88      96    85
2  stu3   90.0     92.0      93      95    90
3  stu4   83.0     93.0      86      85    60
4  stu5   87.0     87.0      70      93    75
5  stu6   95.0     87.0      90      98    80

In [31]: df.fillna(df.mean())
Out[31]:
   name  Math  Physics  English  Python  PE
0  stu1   89.4     73.0      92      82    95
1  stu2   92.0     95.0      88      96    85
2  stu3   90.0     92.0      93      95    90
3  stu4   83.0     93.0      86      85    60
4  stu5   87.0     88.0      70      93    75
5  stu6   95.0     87.0      90      98    80

In [33]: df.fillna(df.median())
Out[33]:
   name  Math  Physics  English  Python  PE
0  stu1   90.0     73.0      92      82    95
1  stu2   92.0     95.0      88      96    85
2  stu3   90.0     92.0      93      95    90
3  stu4   83.0     93.0      86      85    60
4  stu5   87.0     92.0      70      93    75
5  stu6   95.0     87.0      90      98    80

```

图 4.8

(5) 异常值检测

可以使用箱形图分析, 箱体中的数据一般被认为是正常的, 箱体上下边界以外的数据, 通常认为是异常的。也可以使用 describe() 方法查看数据基本情况。

箱形图代码如下:

```

import pandas as pd
import matplotlib.pyplot as plt
df= pd.read_csv('score2.csv')

```

```
df= df.set_index('name')
df.boxplot()
plt.show()
```

数据如图 4.9 所示。

```
In [41]: df
Out[41]:
      Math  Physics  English  Python   PE
name
stu1    12       73      92     82   95
stu2    92       95      88     96   85
stu3    90       92      93     95   90
stu4    83       93      86     85   60
stu5    87      150      70     93   75
stu6    95       87      90     98   80
```

图 4.9

箱形图如图 4.10 所示。

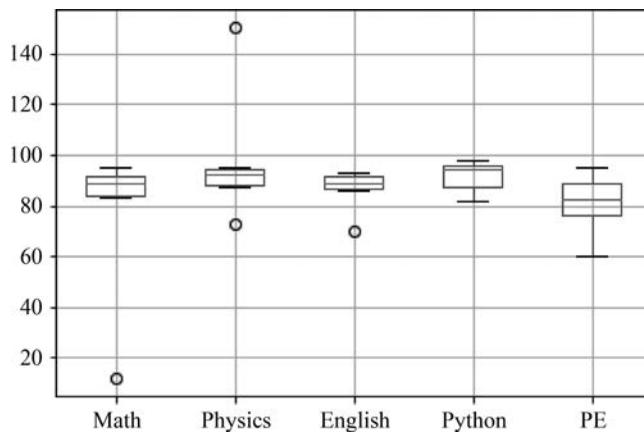


图 4.10

由箱形图可以发现其中有几个离群点。

使用 describe() 方法能查看数据基本情况, 示例如图 4.11 所示。

```
In [42]: df.describe()
Out[42]:
      Math      Physics      English      Python        PE
count  6.000000  6.000000  6.000000  6.000000  6.000000
mean   76.500000 98.333333 86.500000 91.500000 80.833333
std    31.866911 26.530486 8.479387  6.473021 12.416387
min   12.000000 73.000000 70.000000 82.000000 60.000000
25%   84.000000 88.250000 86.500000 87.000000 76.250000
50%   88.500000 92.500000 89.000000 94.000000 82.500000
75%   91.500000 94.500000 91.500000 95.750000 88.750000
max   95.000000 150.000000 93.000000 98.000000 95.000000
```

图 4.11

(6) 异常值过滤

```
import pandas as pd  
df= pd.read_csv('score2.csv')  
df= df.set_index('name')  
d1= df['Math']  
df['outer']= d1< d1.quantile(0.1)# 分位数  
print(df[df.outer== False])  
d2= df['Physics']  
df['outer']= d2> d2.quantile(0.9)  
print(df[df.outer== False])
```

示例如图 4.12 所示。

	Math	Physics	English	Python	PE	outer
name						
stu2	92	95	88	96	85	False
stu3	90	92	93	95	90	False
stu4	83	93	86	85	60	False
stu5	87	150	70	93	75	False
stu6	95	87	90	98	80	False

	Math	Physics	English	Python	PE	outer
name						
stu1	12	73	92	82	95	False
stu2	92	95	88	96	85	False
stu3	90	92	93	95	90	False
stu4	83	93	86	85	60	False
stu6	95	87	90	98	80	False

图 4.12

(7) 异常值修正

示例如图 4.13 所示。

```
In [60]: df.loc[df.Math< df.Math.quantile(0.1)]=df.Math.median()  
In [61]: df  
Out[61]:  
          Math  Physics  English  Python    PE  outer  
name  
stu1  88.5    88.5    88.5    88.5  88.5  88.5  
stu2  92.0    95.0    88.0    96.0  85.0  False  
stu3  90.0    92.0    93.0    95.0  90.0  False  
stu4  83.0    93.0    86.0    85.0  60.0  False  
stu5  87.0   150.0    70.0    93.0  75.0  True  
stu6  95.0    87.0    90.0    98.0  80.0  False
```

图 4.13

3. 数据提取

数据如下：

```
> > > df
```

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

(1) 选择行

例如,选择数据对象 df 中的前三行,可以写成 df['a':'c']或 df[0:3]或 df.head(3)。

(2) 选择列

例如,选择数据对象 df 中的“姓名”列,可以写成 df['姓名']或 df.姓名。

(3) 选择区域

可以使用标签(loc)、位置(iloc)。例如:df.loc['b':'d','语文':'英语'];df.iloc[1:4,1:4];df.loc['a':'c',];df.loc[:,['语文','数学']];df.iloc[:,[1,2,3]]。

(4) 筛选(条件选择)

可以使用筛选条件。例如,df[(df.index >= 'b') & (df.index <= 'd') & (df.数学 >= 90)]可以找出索引值在'b'~'d'之间(包括'b'和'd')并且数学成绩大于等于 90 的学生记录。pandas 还提供了更方便的条件查询方法,比如 query、between、isin、str.contains 等。

■ 参考资料 2:补充案例

1. 要求

对“体检数据.xlsx”作如下处理:

- (1) 导入文件。
- (2) 选取“编号”“性别”“年龄”“身高”“体重”“血压(舒张压/收缩压)”列前 200 行的数据生成新的 DataFrame 对象。
- (3) 检测“编号”是否含有重复值,如果有则过滤,将“编号”设为索引。
- (4) 将所有含有缺失值或者“未检”项的行过滤。
- (5) 将“血压(舒张压/收缩压)”列分为舒张压与收缩压两列。
- (6) 绘制箱形图查看每列异常值,并将异常值过滤。
- (7) 根据高血压的判断标准:收缩压(mmHg) ≥ 140 或者舒张压(mmHg) ≥ 90 ,添加“高血压”列,值为“Y”或“N”。
- (8) 数据写入文件“fdata.csv”中。

2. 参考答案

```
#对“体检数据.xlsx”作如下处理
```

```
#(1) 导入文件
```

```
import pandas as pd
```

```
data= pd.read_excel("体检数据.xlsx")
```

```
#(2) 选取“编号”“性别”“年龄”“身高”“体重”“血压(舒张压/收缩压)”列
```

```
#前 200 行的数据生成新的 DataFrame 对象
```

```

df= data.head(200).iloc[:,1:7]

#(3) 检测“编号”是否含有重复值,如果有则过滤,将“编号”设为索引
temp= df.编号.duplicated(keep= False)
print(temp[temp== True])

# 含有重复数据,保留第一条
df.drop_duplicates('编号',keep= 'first',inplace= True)

# 将“编号”设为索引
df.set_index('编号',inplace= True)

#(4) 将所有含有缺失值或者“未检”项的行过滤
df1= df.dropna(axis= 0)

for item in['性别','身高','体重','血压(舒张压/收缩压)']:
    df1= df1[df1[item].str.contains('未检') == False]

#(5) 将“血压(舒张压/收缩压)”列分为舒张压与收缩压两列
df2= df1['血压(舒张压/收缩压)'].str.split('/',expand= True)
df2.columns= ['收缩压','舒张压']

df= pd.merge(df1,df2,left_index= True,right_index= True)
df.drop('血压(舒张压/收缩压)',axis= 1,inplace= True)

#(6) 绘制箱形图查看每列异常值,并将异常值过滤
df= df[df['体重'].str.isdigit() == True]

import matplotlib as mpl
import matplotlib.pyplot as plt
mpl.rcParams['font.sans-serif']= ['SimHei']
df1= df[['年龄','身高','体重','收缩压','舒张压']]
df1.astype(float).boxplot()
plt.show()

# 异常值过滤
# 过滤身高异常值
d1= df['身高'].astype(float)
df['outer']= d1< d1.quantile(0.01)# 分位数
print('身高异常值\n',df[df.outer== True])
df= df[df.outer== False]

# 过滤体重异常值
d1= df['体重'].astype(float)
df['outer']= d1> d1.quantile(0.99)# 分位数
print('体重异常值\n',df[df.outer== True])
df= df[df.outer== False]

# 过滤舒张压异常值

```

```

d1= df['舒张压'].astype(float)
df['outer']= d1> d1.quantile(0.99)# 分位数
print('舒张压异常值\n',df[df.outer== True])
df= df[df.outer== False]
df.drop('outer',axis= 1,inplace= True)
# 绘制箱形图查看异常值过滤后的情况
df1= df[['年龄','身高','体重','收缩压','舒张压']]
df1.astype(float).boxplot()
plt.show()

#(7) 根据高血压的判断标准:收缩压(mmHg)>= 140 或者舒张压(mmHg)>= 90
#添加“高血压”列,值为“Y”或“N”
df['高血压']= (df.收缩压.astype(int)>= 140)|(df.舒张压.astype(int)>= 90)
df['高血压'].replace(to_replace= True,value= 'Y',inplace= True)
df['高血压'].replace(to_replace= False,value= 'N',inplace= True)
df.to_csv("fdata.csv",encoding= 'gb2312')

```

■ 参考资料 3:参考书

1. 零一,韩要宾,黄园园. Python 3 爬虫、数据清洗与可视化实战[M]. 北京:电子工业出版社,2018.
2. 张莉. Python 程序设计[M]. 北京:高等教育出版社,2019.

六、教学参考案例

■ 参考案例

数据准备

上海市南洋中学 陆栋樑

(1课时)

1. 学科核心素养

- 举例说明生活实际中的数据分析的典型案例。(信息意识)
- 理解数据分析对人们生产生活的作用与意义。(信息意识、信息社会责任)
- 解释数据质量评估标准和对数据进行质量分析。(信息意识)
- 运用适当方法与工具,对数据进行收集、整理和预处理。(数字化学习与创新)
- 学会从现有数据源中提取有效数据。(计算思维、数字化学习与创新)

2.《课程标准》要求

- 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解决中;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。

- 结合生活实际,认识到数据是一种重要的资源,通过科学管理与分析数据,可以使数据实现其应有价值,感受数据管理与分析技术的重要性。
- 结合案例,了解数据采集途径的多样性,能利用适当的工具对数据进行采集和分类;认识噪声数据的现象和成因;理解不同结构化程度数据(包括结构化数据、半结构化数据和非结构化数据)的区别,以及在管理与应用上的特点。

3. 学业要求

学生能够确定学习和生活中的业务数据问题,能提出解决方案,评价其合理性、完整性以及分析方案优化或改进的可能性。能认识有效管理与分析数据对获取有价值信息、形成正确决策的作用与意义,认识数据管理与分析技术对人类社会生活的重要影响;能在特定的信息情境中,根据业务数据问题解决的需要,利用多种途径采集与甄别数据。能按照特定数据管理的需求,使用数据库管理系统建立关系数据库,会选用恰当的策略与方法,对数据进行管理。能根据需要,主动选用数字化工具开展自主或协作学习,创造性地解决问题。

4. 教学内容分析

数据管理与分析,源于需求分析,始于数据准备,对需求分析和数据准备的具体教学内容分析如下:

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

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

5. 学情分析

高中学生对于信息与数据已经有了一定的认知,即学生具备一定的信息意识。信息作为日常生活中最常见的词汇之一,与人类的生活息息相关。人们通过获得、识别自然界和社会的不同信息来区别不同事物,得以认识和改造世界。信息是以文字、数字、符号、图像、图形、声音、情景、状态等形式传播的内容。数学家香农在题为《通讯的数学理论》的论文中指出:“信息是用来消除随机不定性的东西。”数据和信息之间是相互联系的,数据是反映客观事物属性的记录,是信息的具体表现形式;而信息需要经过数字化处理转变成数据才能存储和传输。可以说,数据是信息的载体,是描述客观事物的数、字符,以及所有能输入到计算机中、被计算机程序识别和处理的符号的集合。

学生能够接触到现实生活中的很多典型案例与数据,比如校园歌手大赛成绩管理、图书馆图书及借阅管理、社会实践调查问卷的管理与分析、早餐营养搭配管理、超市销售记录的管理与分析等。引导学生观察生活,发现生活,进而对这些数据进行收集准备,管理

分析再利用,就是数据管理与分析的目的。同时,为了做好数据分析等信息处理工作,必须先进行数据准备,这对学生来讲也是可以理解的。

6. 教学目标

- 通过分析生活中的典型案例,明确数据分析的意义及数据分析的过程和目的。
- 了解获取数据的方法,能对采集到的数据进行质量评估。
- 知道数据预处理的方法,能使用 Python 语言进行数据预处理操作。
- 学会从预处理的数据中,提取有效数据。

7. 教学重难点

- 教学重点:数据预处理、数据质量评估标准和数据质量问题。
- 教学难点:从预处理的数据中,提取有效数据。

8. 教学准备

准备教学设计、教学课件、练习资料、活动记录单和学案等参考资料。

9. 教学策略分析

- 教学方法采用项目教学法、观察法、讨论法、讲授法、实践法等。
- 以 PBL 项目学习为原则,引导学生观察生活,发现生活中的数据分析案例。
- 首先分析项目案例中涉及的数据,引出数据分析的概念和意义;然后说明对数据进行预处理的原因及预处理的具体方法,通过实践活动支持学生进行探究;最后将预处理的数据进行提取,为后续进行数据分析作数据准备。

10. 教学环境

计算机房、教学广播软件、投影、Excel 软件、Python 编程环境。

11. 教学过程设计(见表 4.5)

表 4.5 教学过程设计表

教学环节	教学内容	学生活动	设计意图
课前准备	将本课所需要的学习内容,发送到学生端	预习相关资料	明确任务目标
情境导入	国外姐妹学校来访上海,互动交流,通过数据分析向其介绍上海旅游景点的概况	思考讨论: 1. 上海有哪些景点如何获得数据? 2. 这些景点的评分怎么样? 3. 景点分布在哪里? 门票价格如何? 4. 上海的景点如何展示,怎么推荐?	发现身边的数据分析典型案例,使学生明确数据分析的意义及数据分析的过程和最终目的。 教师最后给出项目的任务 1、任务 2、任务 3、任务 4,贯穿整章学习内容
思考讨论	数据准备:数据是数据分析要处理的对象,数据分析的流程一般从数据准备开始。数据准备包括采集数据、数据预处理、提取相关有效数据	思考讨论: 1. 如何获得数据? 2. 数据质量如何? 3. 怎么处理数据?	使学生明确数据分析的数据来源、数据质量及数据预处理方法

教学环节	教学内容	学生活动	设计意图
活动 1	上海旅游景点数据预处理(数据质量评估和缺失值处理)	1. 使用 Excel 打开上海旅游景点原始数据; 2. 评估数据质量,发现数据问题; 3. 数据预处理(删除、修改、合并等); 4. 得到整理后的数据	实践探究数据预处理的过程,得到高质量的数据
思考探究	从现有数据源提取有效数据,对数据选择加工	1. 将数据导入数据库,查询有效数据; 2. 将数据导入软件进行数据筛选,提取有效数据; 3. 使用编程语言读入原始数据,提取有效数据	明确数据提取的一般方法和目的(提取有效数据,为进一步进行数据分析作准备)
活动 2	使用 Python 实现数据提取	从整合好的数据中提取景点区属信息	实践探究数据提取的过程
作业布置	班级参加学校数学考试的每位同学都有得分,将原始数据进行数据预处理和数据提取	1. 思考有哪些数据; 2. 分析这些数据的价值; 3. 对数据进行预处理; 4. 提取有效数据	进一步巩固数据分析过程中数据准备的过程和重要意义
巩固提升	网上书城销售信息数据预处理和有效提取	1. 获取原始数据; 2. 数据质量分析; 3. 对数据进行预处理; 4. 提取有效数据	巩固提高本课所学知识

【活动记录单】

数据准备

项目情境:国外姐妹学校来访上海,互动交流,通过数据分析向其介绍上海旅游景点的概况。

1. 上海有哪些景点如何获得数据?

2. 这些景点的评分怎么样?

3. 景点分布在哪里?门票价格如何?

4. 上海的景点如何展示,怎么推荐?

活动 1:上海旅游景点数据预处理。

请将对应的数据质量问题填写在表 4.6 中。

表 4.6 数据质量评估标准与数据质量问题

评估标准	准确性	完整性	唯一性	一致性	时效性	相关性
数据质量问题						

缺失值处理方法：

活动 2：使用 Python 实现数据提取(从整合好的数据中提取景点区属信息)。

作业布置：学校数学考试数据预处理和数据提取。

课外探究：网上书城销售信息数据预处理和有效提取。

附：编程参考

使用 Python 实现数据提取，对上海旅游景点数据进行预处理后，从整合好的数据中提取区属信息：

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

第二节 数据分析方法与呈现

一、教学目标与重点

教学目标：

- 了解常用的数据分析方法，如平均分析法、分组分析法、对比分析法、相关分析法等。
- 在实践中选用适当的数据分析工具，分析、呈现并解释数据。

教学重点：

- 掌握平均分析法、分组分析法、对比分析法、相关分析法的相关概念。
- 掌握常用图表的绘制方法及应用场景。

二、教学说明与建议

本节内容主要是常用的数据分析方法与常用的图表绘制。“一、平均分析法”的项目活

动中运行了 Python 代码来实现计算,教学中还可以尝试使用 Excel 进行计算;“二、分组分析法与对比分析法”的项目活动中使用了数据透视表对数据进行分组整合,教学中可以按照教科书的介绍使用 Python 和 Excel 两种工具操作;“三、数据可视化”教学中进行图表介绍时可以使用 Python 和 Excel 同步绘制;“四、相关分析法”的项目活动中既有相关系数的计算又有散点图的绘制,教学中除了运行 Python 代码以外,也可以使用 Excel 进行操作。

本节 5 课时安排建议:“平均分析法”1 课时、“分组分析法与对比分析法”1 课时、“数据可视化”2 课时、“相关分析法”1 课时。

三、项目实施与评价

1. 核心概念精解

(1) 平均分析法

涉及的概念有平均分析法、简单算术平均数、加权算术平均数、中位数、众数。可以使用简单的数据进行计算加深印象。

(2) 分组分析法

涉及的概念有分组分析法、分组标志的分类、数量标志分组时的组数与组距、等距分组时组距与组数的关系及组距的确定。

选择合适的分组标志应考虑研究问题的目的和任务,在分组之前也应对数据对象的特征和发展规律进行理论分析,选出能反映问题本质的标志,作出具体的分组。

(3) 对比分析法

涉及的概念有对比分析、横比与纵比、绝对数与统计相对数。

常用的统计相对数计算公式如表 4.7 所示。

表 4.7 常用的统计相对数计算公式

相对数类别	计算公式	说 明
计划完成程度相对数	$\frac{\text{实际完成数}}{\text{计划数}}$	用于实际完成值与目标值进行对比,反映完成目标的情况,经常用于统计公司业绩目标完成率
结构相对数	$\frac{\text{各组数值}}{\text{总体总数值}}$	用于计算具体某类数据占总体数据的比例,例如个人每月消费中食品支出额占消费支出总额的比例
利用程度相对数	$\frac{\text{实际利用数}}{\text{可能利用数}}$	用于说明人力、物力的利用程度,例如图书馆中图书的借出率
比较相对数	$\frac{\text{某现象数值}}{\text{同期另一同类现象的数值}}$	用于反映某种现象在同一时间内不同空间条件下的差异程度,例如不同地区商品价格对比
强度相对数	$\frac{\text{某现象的数值}}{\text{另一有联系现象的数值}}$	用于两个性质不同但有一定联系的指标对比,反映现象的强度、密度。例如人均国内生产总值、人口密度
动态相对数	$\frac{\text{报告期数值}}{\text{基期数值}}$	用于计算同一事物在不同时间上的数值比,说明现象在时间上发展变化的程度,例如发展速度、增长速度

(4) 相关分析法

涉及的概念有相关关系的特点、相关分析的分类(相关与不相关、正相关与负相关、线性相关与非线性相关)、相关系数的意义。需要了解皮尔森相关系数的计算。

(5) 数据可视化

数据可视化是研究如何将数据以图片或图形的方式展现的科学,基本思想是将数据库中每一个数据项作为单个图元元素表示,大量的数据集构成数据图像,同时将数据的各个属性值以多维数据的形式表示,可以从不同的维度观察数据,从而对数据进行更深入的观察和分析。

数据可视化使人们不再局限于通过关系数据表来观察和分析数据信息,还能以更直观的方式看到数据及其结构关系。

统计学家于 1973 年构造了如表 4.8 所示的四组数据。每一组数据都包括了 11 个 (x, y) 点。这四组数据的统计分析特性完全一样(如表 4.9 所示),但是绘制成散点图就不一样了(如图 4.14 所示)。

表 4.8 四组数据

第 1 组		第 2 组		第 3 组		第 4 组	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

表 4.9 四组数据的统计分析特性

性质	数值
x 的平均数	9
x 的方差	11
y 的平均数	7.5
y 的方差	4.1
x 与 y 之间的相关系数	0.8

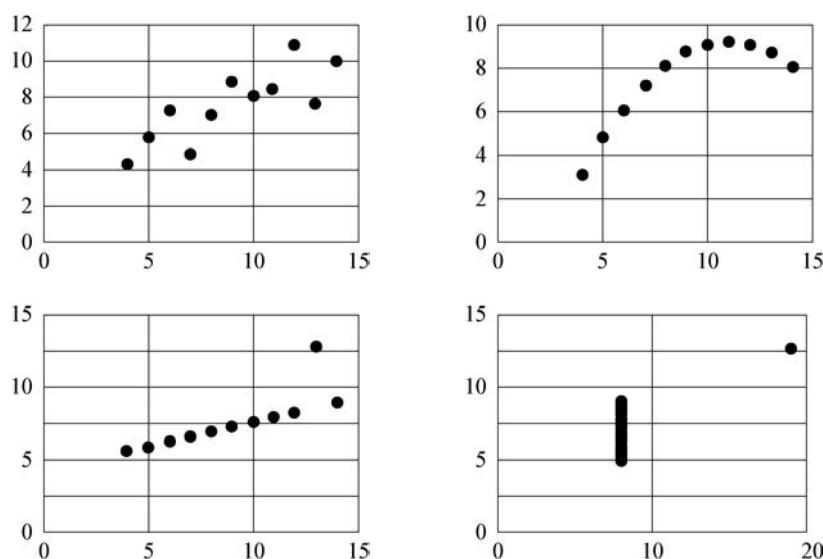


图 4.14 四组数据的散点图

这说明,使用图形的方式展示数据,数据的特征更容易被人们观测到。

(6) 常用图表

涉及柱形图、条形图、饼图、折线图、散点图、箱形图、雷达图、热力图、甘特图、词云。

2. 项目活动的具体实施

(1) 项目任务 2 的实施

本任务在介绍了平均分析法之后进行。使用教科书第四章第一节中准备好的数据,计算某知名旅游网站上海 Top50 景点评分的算术平均数,用 Python 编写程序。教学时可以先用 Excel 进行计算,再用 Python 编写程序实现计算。可以使用 Python 的 pandas 库,DataFrame 对象每列数据可以直接相乘,并且有 mean() 和 sum() 方法。

(2) 项目任务 3 的实施

“统计上海 Top50 景点在各区的分布”在介绍了分组分析法和对比分析法之后进行。将数据用区属分组,统计其中各级别景点的个数。活动采用了 3 种方法,学生能熟练掌握 1 种即可。“用图展示分布情况”在介绍了各种类型的图表之后进行。介绍柱形图、条形图、饼图、折线图、散点图、箱形图、雷达图时,可以使用 Excel 或 Python 绘制图表,热力图与甘特图建议直接使用 Excel 绘制,词云图可以使用 Python 绘制,用 wordcloud 库。活动中的堆积柱形图可以使用 Excel 绘制,或使用 Python 的 pandas 库绘制。

(3) 项目任务 4 的实施

本任务在介绍了相关分析的基本概念后进行。相关分析中正相关和负相关需要举例说明,需要掌握相关系数值的意义。

3. 项目活动的评价

本节的项目评价主要包括:考查学生对平均分析法、分组分析法、对比分析法、相关分析法等核心概念的掌握程度以及常用图表的应用情景的了解程度。

评价建议:采用终结性评价与过程性评价相结合的方式。采用纸笔测试等方式对核心概念的掌握程度进行终结性评价;让学生组队完成本节“四、作业练习与提示”中的题目,在实践中根据分析的任务选取合适的数据分析方法以及合适的图表进行展示,对学生在此过程中的分工表现进行评价。

四、作业练习与提示

■ 题目描述

1. 对教科书第四章第一节中提取区属信息后的数据使用平均分析、分组分析、对比分析、相关分析进行分析,并将分析结果以图表的形式进行展示。

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

■ 作业提示

1. 使用教科书中的 Python 程序进行平均分析、分组分析、对比分析、相关分析。绘制图表参考如下程序。

折线图使用 df.plot()。

柱形图使用 df.plot.bar()。

条形图使用 df.plot.bart()。

饼图使用 df.plot.pie()。

散点图使用 df.plot.scatter()。

箱形图使用 df.plot.box()。

在 Python 中绘制雷达图需要使用极坐标系, 设置稍复杂, 如果数学基础欠缺, 可以使用 Excel 绘制。

根据教科书第 92 页表 4.9 绘制雷达图的程序如下:

```
'''  
    语文 数学 外语 物理 化学  
甲班  94   95   84   64   90  
乙班  75   93   66   85   88  
丙班  86   76   96   93   67  
'''  
  
import numpy as np  
import matplotlib.pyplot as plt  
import matplotlib  
matplotlib.rcParams['font.family']= 'SimHei'  
matplotlib.rcParams['font.sans-serif']= ['SimHei']  
labels= np.array(['语文','数学','外语','物理','化学'])  
nAttr= len(labels)  
data= np.array([[94,95,84,64,90],[75,93,66,85,88],[86,76,96,93,67]])  
angles= np.linspace(0,2* np.pi,nAttr,endpoint= False)  
data= np.concatenate((data,data[:,[0]]),axis= 1)  
angles= np.concatenate((angles,[angles[0]]))  
fig= plt.figure(facecolor= "white")  
plt.subplot(111,polar= True)  
plt.plot(angles,data[0],'bo-',color= 'g',linewidth= 2,label= '甲班')  
plt.plot(angles,data[1],'bo-',color= 'r',linewidth= 2,label= '乙班')  
plt.plot(angles,data[2],'bo-',color= 'b',linewidth= 2,label= '丙班')  
plt.legend(loc= 'best')  
plt.thetagrids(angles* 180/np.pi,labels)  
plt.figtext(0.52,0.95,'成绩分析图',ha= 'center')  
plt.grid(True)  
plt.savefig('dota_radar.JPG')  
plt.show()
```

2. 需要安装 wordcloud 库。只需要收集书名和书的类别信息,由于数据量不大,人工整理即可,有基础的学生也可以使用网页解析来获得数据。整理好后参考《唐诗三百首》作者词云图程序,修改数据源与数据列即可。

《唐诗三百首》原始数据格式如表 4. 10 所示。

表 4. 10 《唐诗三百首》原始数据表(部分)

类别	诗 名	作者
五言古诗	西施咏	王维
五言古诗	送别	王维
五言古诗	送綦毋潜落第还乡	王维
五言古诗	青溪	王维
五言古诗	与高适薛据同登慈恩寺浮图	岑参
五言古诗	下终南山过斛斯山人宿置酒	李白
五言古诗	月下独酌	李白
五言古诗	春思	李白
五言古诗	梦李白·其一	杜甫
五言古诗	梦李白·其二	杜甫

作者词云绘制的程序如下:

```
import pandas as pd
from wordcloud import WordCloud
import matplotlib.pyplot as plt
df0= pd.read_csv('poetry_tang.csv',encoding= 'gb18030')
newslist= df0.作者
content= ' '.join(newslist)
wordcloud= WordCloud(font_path= 'simhei.ttf',
background_color= "grey",collocations= False,
max_words= 30).generate(content)
plt.imshow(wordcloud)
plt.axis("off")
wordcloud.to_file("tangshi.jpg")
plt.show()
```

五、教学参考资源

■ 参考资料 1:Matplotlib 绘图

1. 柱形图和条形图

```
import numpy as np  
import matplotlib.pyplot as plt  
plt.bar(range(7),[3,4,7,6,2,8,9])  
plt.show()
```

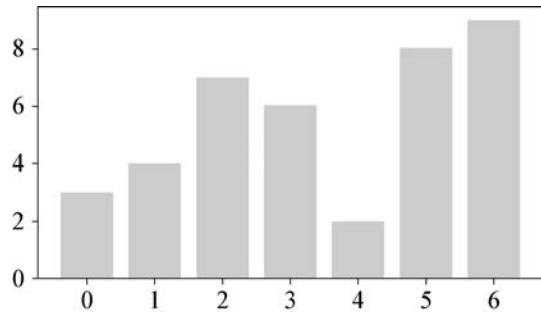


图 4.15 柱形图

```
import numpy as np  
import matplotlib.pyplot as plt  
plt.bartohoriz(range(7),[3,4,7,6,2,8,9])  
plt.show()
```

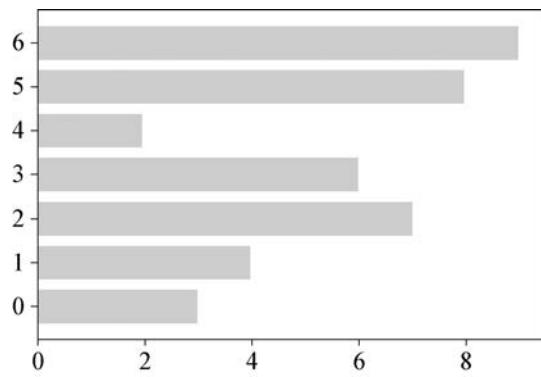


图 4.16 条形图

2. 饼图

```
import matplotlib.pyplot as plt  
plt.pie([0.15,0.6,0.25])  
plt.show()
```

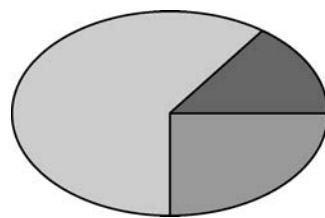


图 4.17 饼图

3. 直方图

```
import numpy as np  
import matplotlib.pyplot as plt  
x= np.random.randint(0,101,100)  
plt.hist(x)  
plt.show()
```

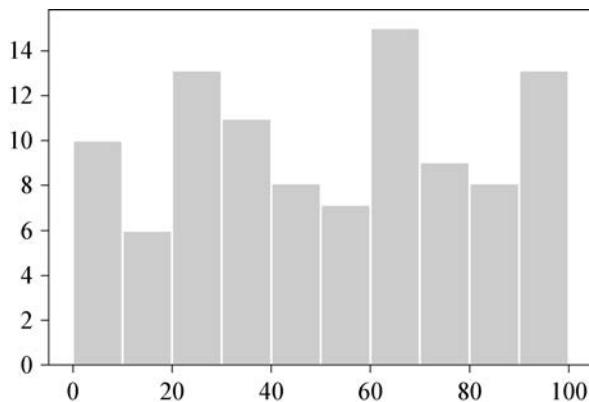


图 4.18 直方图

4. 多子图绘制

根据“体检数据.xlsx”中前 10 条绘图。

```
import matplotlib.pyplot as plt  
import matplotlib as mpl  
import pandas as pd  
df= pd.read_excel(r'体检数据.xlsx')  
df= df.head(10)[['编号','年龄','身高','体重']]  
df1= df.set_index('编号')  
df1= df1.astype('int32')  
mpl.rcParams['font.sans-serif']= ['SimHei']  
fig,axs= plt.subplots(2,2)  
plt.subplots_adjust(hspace= 0.5,wspace= 0.5)  
axs[0,0].plot(range(10),df1.身高,color= 'r',marker= 'o')  
axs[0,0].set_title('身高')  
axs[0,1].bar(range(10),df1.年龄,color= 'g')  
axs[0,1].set_title('年龄')  
axs[1,0].barh(range(10),df1.体重,color= 'b')  
axs[1,0].set_title('体重')  
import pandas as pd  
df= pd.read_excel(r'体检数据.xlsx')
```

```

df= df.head(10)[['编号','年龄','身高','体重']]
df1= df.set_index('编号')
# df1['身高']= df1['身高'].apply(pd.to_numeric)
df1= df1.astype('int32')
import matplotlib.pyplot as plt
from matplotlib.gridspec import GridSpec
fig= plt.figure()
gs= GridSpec(2,2)
ax1= fig.add_subplot(gs[0,0])
ax2= fig.add_subplot(gs[0,1])
ax3= fig.add_subplot(gs[1,:])
plt.subplots_adjust(hspace= 0.5)
ax1.plot(range(10),df1.身高,color= 'r',marker= 'o')
ax1.set_title('身高')
ax2.bar(range(10),df1.年龄,color= 'g')
ax2.set_title('年龄')
ax3.barh(range(10),df1.体重,color= 'b')
ax3.set_title('体重')
plt.show()

```

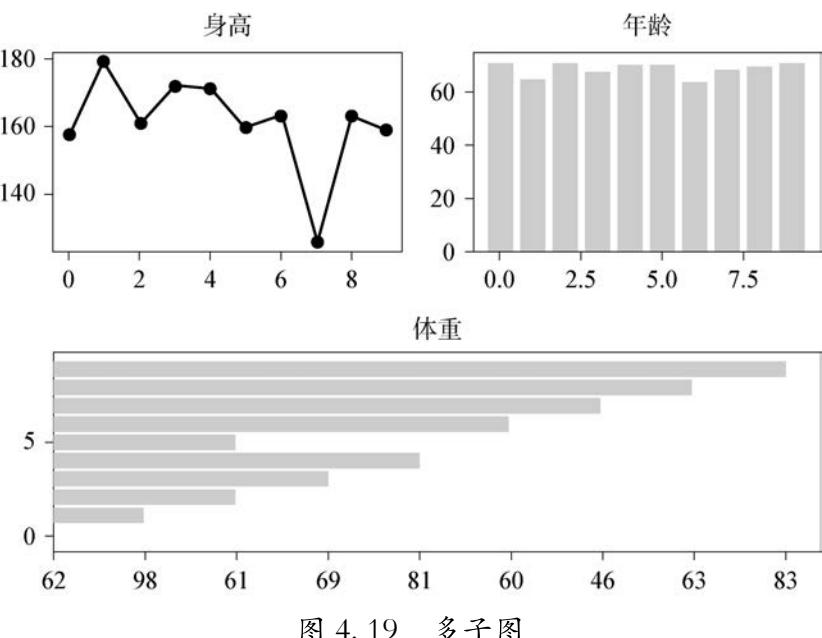


图 4.19 多子图

5. 散点图与气泡图

```

import matplotlib.pyplot as plt
import matplotlib as mpl

```

```

import numpy as np
a= np.random.randn(10)
b= np.random.randn(10)
s= np.power(10* a+ 20* b,2)
plt.scatter(a,b)
# plt.scatter(a,b,s) # 气泡图
plt.show()

```

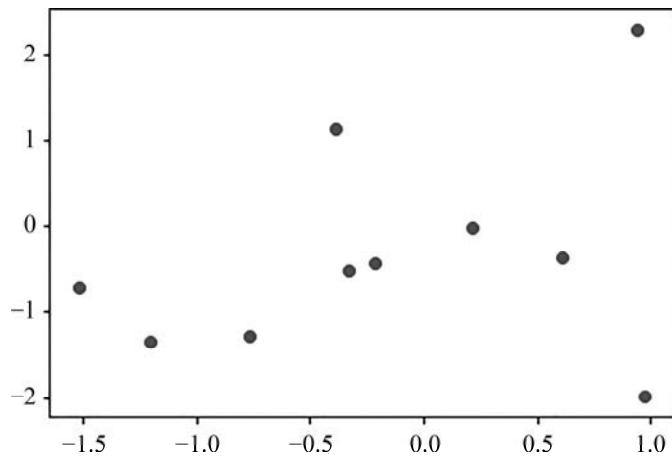


图 4.20 散点图

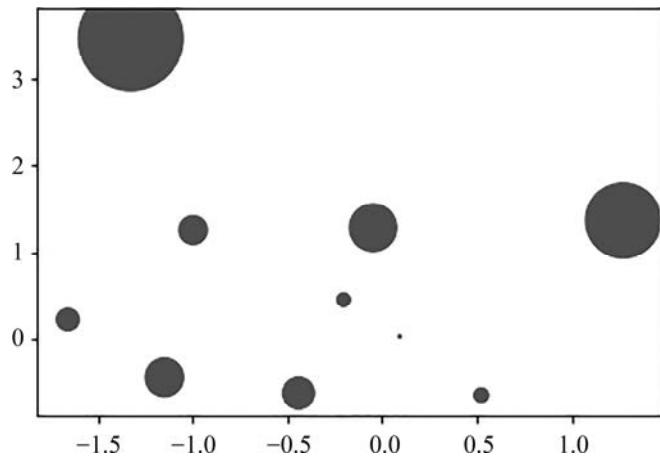


图 4.21 气泡图

6. 极线图(极坐标图)

```

import matplotlib.pyplot as plt
import numpy as np
theta= np.linspace(0.0,6* np.pi,100)
r= theta
plt.polar(theta,r,color= "red",linewidth= 4)
plt.show()

```

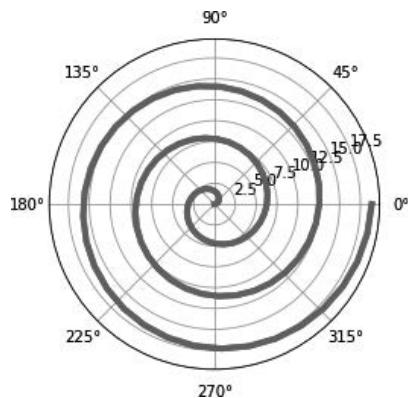


图 4.22 极线图(极坐标图)

7. 玫瑰线图

```
import matplotlib.pyplot as plt
import numpy as np
theta= np.linspace(0.0,2* np.pi,1000)
r= 2* np.sin(8* theta)
plt.polar(theta,r,color= "red",linewidth= 4)
plt.show()
```

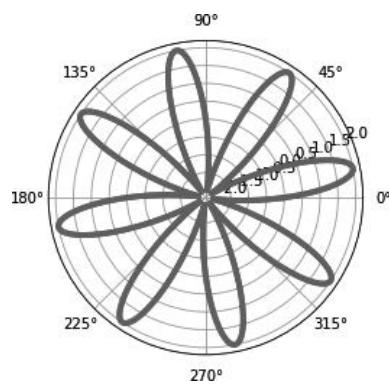


图 4.23 玫瑰线图

8. 棉棒图(火柴图)

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as mpl
mpl.rcParams['axes.unicode_minus']= False
x= np.linspace(0,10,20)
y= np.random.randn(20)
plt.stem(x,y,linestyle= 'r-',markerfmt= 'C0o',basefmt= '--',label=
'TestStem')
plt.legend()
```

```
plt.show()
```

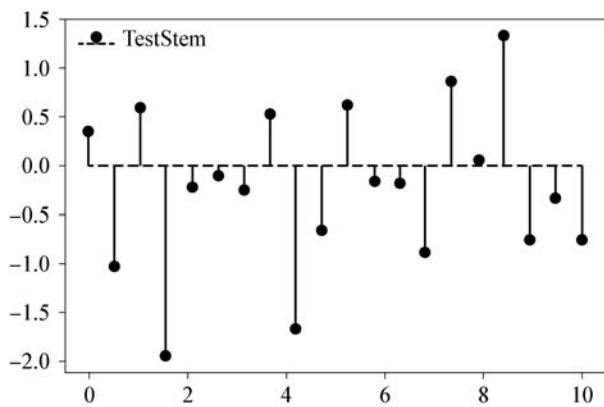


图 4.24 棉棒图(火柴图)

9. 图表设置坐标轴样式

```
import matplotlib.pyplot as plt
x_values= list(range(11))
y_values= [x* * 2 for x in x_values]
#用 plot 函数绘制折线图,线条颜色设置为绿色
plt.plot(x_values,y_values,c= 'green')
#设置图表标题和标题字号
plt.title('Squares',fontsize= 24)
#设置刻度的字号
plt.tick_params(axis= 'both',which= 'major',labelsize= 14)
#设置 x 轴标签及其字号
plt.xlabel('Numbers',fontsize= 14)
#设置 y 轴标签及其字号
plt.ylabel('Squares',fontsize= 14)
# MultipleLocator 类用于设置刻度间隔,把 x 轴的刻度间隔设置为 1,并存在变量里
x_major_locator= plt.MultipleLocator(1)
#把 y 轴的刻度间隔设置为 10,并存在变量里
y_major_locator= plt.MultipleLocator(10)
#ax 为两条坐标轴的实例
ax= plt.gca()
#把 x 轴的主刻度设置为 1 的倍数
ax.xaxis.set_major_locator(x_major_locator)
#把 y 轴的主刻度设置为 10 的倍数
ax.yaxis.set_major_locator(y_major_locator)
```

```

#把 x 轴的刻度范围设置为 0 到 11
plt.xlim(-0.5,11)
#把 y 轴的刻度范围设置为 0 到 110
plt.ylim(-5,110)
plt.axvline(color= 'black', linewidth= 2)
plt.axhline(color= 'black', linewidth= 2)
plt.show()

```

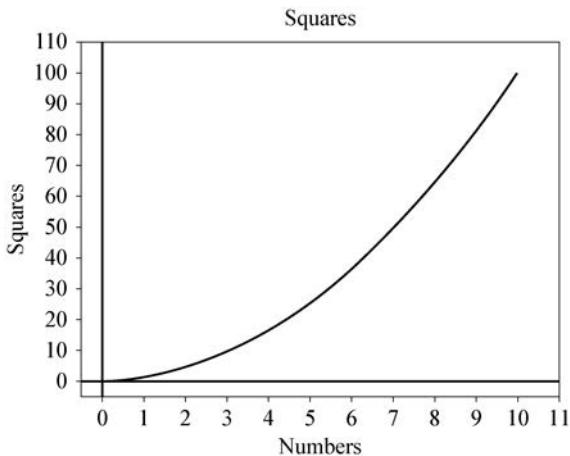


图 4.25

■ 参考资料 2: 补充案例

- 使用上海市地面气象数据(1981~2010 年)完成任务

任务 1 计算上海地区累年六月份的各气象要素日平均值并填写在表 4.11 中。

表 4.11

六月平均气温(摄氏度)	六月平均日最高气温(摄氏度)	六月平均日最低气温(摄氏度)	六月日平均水汽压(百帕)	六月 20~20 时日降水量(毫米)	六月日平均风速(米/秒)

任务 2 计算上海崇明区与徐汇区累年各月份的各气象要素日平均值, 填写在表 4.12 中, 并进行对比。

表 4.12

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
崇明区												
徐汇区												

任务3 根据任务2中的数据绘制折线图。

任务4 计算上海地区风速和降水量的相关系数,进行相关分析。

2. 对“fdata.csv”的数据进行平均分析练习

(1) 导入文件,计算数据的条数,并将前100条数据生成新的DataFrame对象,将“编号”设为索引。

- (2) 计算前100条数据中年龄的中位数。
- (3) 计算前100条数据中年龄的算术平均数。
- (4) 绘制前100条数据的年龄分布直方图。

参考答案:

```
#导入文件,计算数据的条数,并将前100条数据生成新的DataFrame对象,将“编号”设为索引
```

```
import pandas as pd
import matplotlib.pyplot as plt
data= pd.read_csv('fdata.csv',encoding= 'gb2312')
df= data.head(100)
df.set_index('编号',inplace= True)
#计算前100条数据中年龄的中位数
print("前100条数据中年龄的中位数:",df.年龄.median())
#计算前100条数据中年龄的算术平均数
print("前100条数据中年龄的算术平均数:",df.年龄.mean())
#绘制前100条数据的年龄分布直方图
plt.hist(df.年龄)
```

3. “图表格式设置”练习

绘制图4.26。

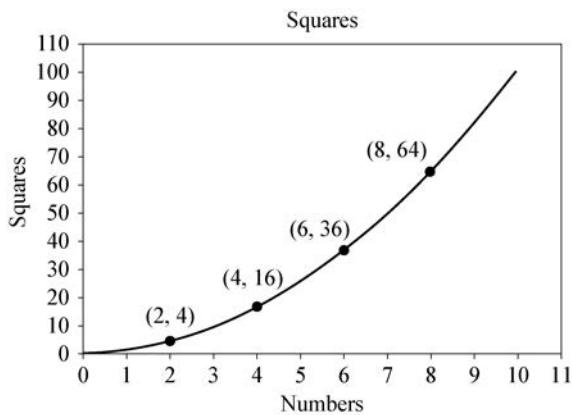


图 4. 26

参考答案:

```
import matplotlib.pyplot as plt
```

```

x_values= list(range(11))
y_values= [x* * 2 for x in x_values]
#用 plot 函数绘制折线图,线条颜色设置为绿色
plt.plot(x_values,y_values,c= 'green')
# 设置图表标题和标题字号
plt.title('Squares',fontsize= 24)
# 设置刻度的字号
plt.tick_params(axis= 'both',which= 'major',labelsize= 14)
#设置 x 轴标签及其字号
plt.xlabel('Numbers',fontsize= 14)
#设置 y 轴标签及其字号
plt.ylabel('Squares',fontsize= 14)
# MultipleLocator 类用于设置刻度间隔,把 x 轴的刻度间隔设置为 1,并存在变量中
x_major_locator= plt.MultipleLocator(1)
#把 y 轴的刻度间隔设置为 10,并存在变量中
y_major_locator= plt.MultipleLocator(10)
#ax 为两条坐标轴的实例
ax= plt.gca()
#把 x 轴的主刻度设置为 1 的倍数
ax.xaxis.set_major_locator(x_major_locator)
#把 y 轴的主刻度设置为 10 的倍数
ax.yaxis.set_major_locator(y_major_locator)
#把 x 轴的刻度范围设置为 0 到 11
plt.xlim(0,11)
#把 y 轴的刻度范围设置为 0 到 110
plt.ylim(0,110)
#点
x= np.array([2,4,6,8])
y= x* * 2
for x0,y0 in zip(x,y):
    plt.scatter([x0,], [y0,], s= 50,color= 'b')
    plt.text(x0+0.5,y0+0.5,r'({},{})'.format (x0, y0), fontdict =
{'size':16,'color':'r'})
plt.show()

```

■ 参考资料 3:参考书

1. 尼尔·J.萨尔金德. 爱上统计学(中译本 第2版)[M]. 史玲玲,译. 重庆:重庆大

学出版社,2011.

2. 杨轶莘. 大数据时代下的统计学[M]. 北京:电子工业出版社,2015.
3. 朝乐门. 数据科学理论与实践[M]. 北京:清华大学出版社,2017.
4. 科尔·努斯鲍默·纳福利克. 用数据讲故事[M]. 陆昊,吴梦颖,译. 北京:人民邮电出版社,2017.

六、教学参考案例

■ 参考案例

数据可视化

上海市南洋中学 陆栋樑

(1课时)

1. 学科核心素养

- 发现身边的一些数据可视化的具体表现与案例。(信息意识)
- 掌握一定的数据可视化的工具与方法。(计算思维、数字化学习与创新)
- 能根据数据可视化的结果进行分析预测。(信息意识、数字化学习与创新)
- 理解人工智能与大数据时代数据可视化的作用和意义。(信息意识、计算思维、信息社会责任)

2.《课程标准》要求

- 结合生活实际,认识到数据是一种重要的资源,通过科学管理与分析数据,可以使数据实现其应有价值,感受数据管理与分析技术的重要性。
- 了解常用的数据分析方法,如平均分析法、分组分析法、对比分析法和相关分析法等;在实践中选用适当的数据分析工具,分析、呈现并解释数据。
- 运用数字化学习方式,了解数据管理与分析技术的新发展;结合恰当的案例分析,认识数据挖掘对信息社会问题解决和科学决策的重要意义。

3. 学业要求

学生能认识有效管理与分析数据对获取有价值信息、形成正确决策的作用与意义,认识数据管理与分析技术对人类社会生活的重要影响;会采用适当的方法提取数据;能正确选用数据分析方法和工具,分析并解释数据;能根据需要,主动选用数字化工具开展自主或协作学习,创造性地解决问题。

4. 教学内容分析

广义上,数据可视化无处不在;狭义上的数据可视化,更多的是用纯图形去表示数据,也有很多分类。数据可视化旨在借助图形化手段,清晰有效地传达与沟通信息,它是关于数据视觉表现形式的科学技术研究。

数据分析的结果不仅可以用数据表形式表示,也可以使用图表表示。数据可视化,使得数据呈现方式更加灵活生动。数据可视化就是用图形来表示信息和数据。借助图表、

图形和地图等可视化元素,数据可视化工具可以使人们便捷地查看和了解数据中的趋势、异常值和模式。

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

在大数据领域,数据可视化工具和技术对于分析海量信息并制定数据驱动型决策而言至关重要。从艺术和广告到电视和电影,都是以可视形式传播文化的,我们的眼睛总是被颜色和图案吸引。我们与数据的交互方式应该反映这一现实。

本节课是本章的后半部分学习内容,在前面学习了数据准备和数据分析方法之后,学生已经对数据有了比较深刻的认识,但是难免感觉数据比较枯燥呆板和“冷冰冰”,本节课的目的就是让数据“热”起来,形象生动起来,借助工具和方法,把数据反映在用眼睛看得到的一些图表与图形上。

5. 学情分析

通过之前几节课的学习,学生对数据已经有了一定的认识,对数据管理与分析的方法有了一定的学习与探究。知道了数据分析有各种方法,并且通过已经学会的方法,能查询或筛选出自己想要获得的信息与数据。

学生已经在初中阶段学习了电子表格处理软件 Excel,对使用电子表格图表展示数据有一定的概念认知与实践操作,本节课结合学生对 Excel 图表的已有概念,用新工具 Python 编程绘制图表实现数据可视化,有一定基础的学生应该是可以接受和基本明白的。

实现数据可视化的常用方法是绘制各种图表。对于数据可视化的深刻理解,需要学生有一定知识基础,同时也具备一定数据分析与管理知识,学习由浅入深,循序渐进。

6. 教学目标

- 知道生活中的一些数据可视化典型案例。
- 掌握使用 Python 编程实现数据图表可视化。
- 学会在数据可视化后对数据结果进行分析预测。
- 理解人工智能与大数据时代数据可视化的意义。

7. 教学重难点

- 教学重点:用 Python 实现数据可视化分析与呈现并解释数据。
- 教学难点:数据可视化方法的应用场景和意义。

8. 教学准备

准备教学设计、教学课件、练习资料、活动记录单和学案等参考资料。

9. 教学策略分析

- 教学方法采用项目教学法、观察法、讨论法、讲授法、实践法等。
- 以 PBL 项目学习为原则,引导学生观察生活,发现生活中的数据分析和数据可视化典型案例,通过实践活动,初步理解大数据应用和数据可视化。
- 结合学生学过的知识,包括上一节课学习的数据分析方法等内容,鼓励学生发现数

据背后的故事，支持学生探究，对数据进行可视化呈现与展示。

- 通过数据分析和可视化，分析规律趋势，帮助思考与决策。

10. 教学环境

计算机房、教学广播软件、投影、Python 编程环境。

11. 教学过程设计(见表 4. 13)

表 4. 13 教学过程设计表

教学环节	教学内容	学生活动	设计意图
新课引入	播放视频《智慧城市大数据可视化》	观看视频并思考	初步感受大数据和数据可视化
互动讨论	谈谈你身边的数据可视化案例（炒股软件等）	观察生活，思考分析数据分析、数据可视化与生活	感受数据分析与可视化很常见
活动 1	使用 Python 编程实现数据可视化（柱形图、条形图……）	用 Python 实现数据分析可视化呈现	从已有经验知识入手，学习新知
活动 2	Python 编程实践：用图形展示分析结果（pandas/matplotlib 库应用）	使用 Python 呈现上海各区景点分布柱形图、条形图	通过代码填空题体验数据可视化图形
活动 3	Python 甘特图制作：制作学习任务进度表	跟着微视频完成；自主学习与实践	体验另一种数据可视化图表形式
阶段小结	总结各种可视化图表的功能和作用	理解图表数据可视化对数据分析与管理的意义	小结归纳图表数据可视化内容
学习新知	除图表外，也可以使用其他图形呈现数据	用数据讲故事，学生思考讨论分析（图形方法）	学习除图表外的其他数据可视化方法
观察体验	教师演示：打开相关网站，输入关键字“数据可视化”即刻生成词云	观察分析与思考： 1. 什么是图表可视化？ 2. 什么是数据可视化图形？	体验观察图形数据可视化的方法：词云
活动 4	Python 编程实践：词云（pandas/wordcloud 库应用）	用 Python 编写《唐诗三百首》中作者名字词云程序	学会用 Python 实现数据可视化图形绘制
课堂小结	分析总结数据可视化应用（大数据展示等）	学生分析总结数据可视化的场景与意义	总结本课知识点，深入体验数据可视化和大数据应用
课后拓展	某共享单车使用数据可视化分析	用 Python 实现数据可视化图形	拓展巩固本课学习内容

【活动记录单】

数据可视化

互动讨论：谈谈你身边的数据可视化案例。

活动 1 使用 Python 编程实现数据可视化(柱形图、条形图……)。

活动 2 Python 编程实践:用图形展示分析结果(pandas/matplotlib 库应用)。

活动 3 Python 甘特图制作:制作学习任务进度表。

活动 4 Python 编程实践:词云(pandas/wordcloud 库应用)。

课后拓展:某共享单车使用数据可视化分析。

附:编程参考

1. 绘制柱形图

```
# Python 柱形图制作  
import numpy as np  
import matplotlib.pyplot as plt  
plt.bar(range(7),[3,4,7,6,2,8,9])  
plt.show()
```

2. 绘制条形图

```
# Python 条形图制作  
import numpy as np  
import matplotlib.pyplot as plt  
plt.barh(range(7),[3,4,7,6,2,8,9])  
plt.show()
```

3. 绘制上海各区景点分布堆积柱形图

```
# 上海各区景点分布图制作
```

```
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif']= ['SimHei']
df= pd.read_csv('result.csv',encoding= 'gb2312')
df.index= df.区属
cht= df.plot.bar(title= '上海各区景点分布',stacked= True)
cht.set_ylabel('个数')
cht.set_xlabel('区属')
plt.show()
```

4. 绘制甘特图

```
# Python 甘特图制作
import plotly as py
import plotly.figure_factory as ff
pyplt= py.offline.plot
df= [dict(Task= "语文作业",Start= '2020-07-18',Finish= '2020-08-28'),
      dict(Task= "数学作业",Start= '2020-07-20',Finish= '2020-08-15'),
      dict(Task= "英语作业",Start= '2020-07-16',Finish= '2020-08-20'),
      dict(Task= "物理作业",Start= '2020-07-24',Finish= '2020-08-05'),
      dict(Task= "化学作业",Start= '2020-08-01',Finish= '2020-08-08')]
fig= ff.create_gantt(df)
pyplt(fig,filename= 'gantt.html')
```

5. 绘制词云

```
#《唐诗三百首》中作者名字词云制作
import pandas as pd
from wordcloud import WordCloud
import matplotlib.pyplot as plt
df0= pd.read_csv('poetry_tang.csv',encoding= 'gb18030')
newslist= df0.作者
content= ' '.join(newslist)
wordcloud= WordCloud(font_path= 'simhei.ttf',background_color= "grey",collocations= False,max_words= 30).generate(content)
plt.imshow(wordcloud)
plt.axis("off")
wordcloud.to_file("tangshi.jpg")
plt.show()
```

6. 绘制雷达图

```
# Python 雷达图制作
```

```

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

import numpy as np
import matplotlib.pyplot as plt
import matplotlib
matplotlib.rcParams['font.family'] = 'SimHei'
matplotlib.rcParams['font.sans-serif'] = ['SimHei']
labels= np.array(['语文','数学','外语','物理','化学'])
nAttr= len(labels)
data= np.array([[94,95,84,64,90],[75,93,66,85,88],[86,76,96,93,67]])
angles= np.linspace(0,2* np.pi,nAttr,endpoint= False)
data= np.concatenate((data,data[:,[0]]),axis= 1)
angles= np.concatenate((angles,[angles[0]]))
fig= plt.figure(facecolor= "white")
plt.subplot(111,polar= True)
plt.plot(angles,data[0],'bo-',color= 'g',linewidth= 2,label= '甲班')
plt.plot(angles,data[1],'bo-',color= 'r',linewidth= 2,label= '乙班')
plt.plot(angles,data[2],'bo-',color= 'b',linewidth= 2,label= '丙班')
plt.legend(loc= 'best')
plt.thetagrids(angles* 180/np.pi,labels)
plt.figtext(0.52,0.95,'成绩分析图',ha= 'center')
plt.grid(True)
plt.savefig('dota_radar.JPG')
plt.show()

```

数据挖掘

一、本章学科核心素养的渗透

1. 信息意识

通过大数据时代下数据挖掘过程的介绍,让学生学会根据解决问题的需要,分析数据中承载的信息,采用有效策略对信息来源的可靠性、内容准确性、指向的目的性作出合理判断,对信息可能产生的影响进行预期分析,为解决问题提供参考。大数据给人类社会所带来的影响和价值正是希望为解决问题提供有效的参考。

2. 计算思维

数据挖掘的过程必然涉及程序编写,本书采用 Python 程序语言实现数据挖掘的聚类算法。计算思维的训练需要学生学会运用计算机处理问题、抽象特征、建立结构模型、合理组织数据。所以在数据挖掘过程中,学生必须了解如何建立合适的模型、采用适当的算法,才能从数据集中发现有用的信息。

3. 数字化学习与创新

数字化学习与创新要求学生能适应数字化学习环境,养成数字化学习与创新的习惯,掌握各种学习资源和工具,开展自主学习、协同工作、知识分享与创新。数据挖掘和大数据时代下的数据挖掘,都需要学生充分开展自主学习、协同工作,从广阔的数字世界中汲取有用的知识。数据挖掘本身的技术并不是难点,难点是如何运用到各行各业中,发挥决策性作用,所以数字化时代的数据挖掘离不开数字化学习与创新。

4. 信息社会责任

信息社会要求每个个体都能在文化修养、道德规范、行为自律方面尽到自己的责任。数据挖掘本身就会牵涉个人隐私、虚拟空间等社会道德与伦理准则问题,因此在学生的学习过程中,教师必须正确引导学生对法律、伦理、道德准则的理解和掌握,必须有效维护信息社会中的社会安定及信息安全。国家安全是民族复兴的根基,社会稳定是国家强盛的前提。所以教师要对健全国家安全体系有深刻的认识,在教学中引导学生懂得强化经济、

重大基础设施、数据等安全保障体系建设的重要性。

二、本章知识结构

本章遵循普通高中信息技术课程标准,依据学分和课时规定,紧扣学科概念体系,将内容分为两节,以“电影数据的数据挖掘”为项目主题,围绕数据挖掘的过程、大数据时代下的数据挖掘展开设计。

第一节“数据挖掘过程”,通过对生活中常见的电影相关数据进行分析、挖掘,使学生了解数据挖掘的过程和基本算法。

第二节“大数据时代下的数据管理与分析技术的发展”,通过介绍大数据的基本概念、大数据实现的基础设施架构、数据管理与分析的发展,使学生了解大数据时代下数据存储、数据挖掘的特点,以及大数据对人类社会的影响等。

三、本章项目活动设计思路

本章的项目活动设计为电影数据的数据挖掘。通过该项目活动,使学生初步了解用好生活工作中遇到的各种数据对解决信息社会问题以及作出科学决策的指导意义。

首先使学生简单了解数据挖掘本身的含义,并且知道数据挖掘的过程以及基本的几种算法。然后以电影数据为例,介绍聚类算法中比较简单基础的 k -平均算法。同时,对收集的电影数据进行简化,便于学生开展活动。

项目任务 1:电影数据的聚类。

通过评分、评论数、票房 3 个字段对电影数据进行聚类。根据 k -平均算法的概念,此例中将 K 设置为 3,即预测聚类数量为 3 类,通过 Python 提供的聚类函数实现聚类,并用降维后的散点图展示聚类结果。最后,分析聚类结果与电影类型之间的一些关系。

项目任务 2:查询 3 个数据挖掘算法。

通过数字化学习方法查询挖掘算法的案例应用,了解数据挖掘在现实生活中的实际意义。

项目任务 3:大数据下的电影数据的数据挖掘。

如果待分析数据的数据量达到了大数据级别,则在数据挖掘过程中,会出现无法用传统方式在单机上实现数据存储,因此会涉及分布式存储,需要了解大数据的基础设施架构,包括大数据平台的操作系统和存储与传统小数据量用到的平台会有所不同。

四、本章课时安排建议

本章教学建议用 4 课时完成,具体参见表 5.1。

表 5.1 课时安排建议表

第一节 数据挖掘过程	3 课时
第二节 大数据时代下的数据管理与分析技术的发展	1 课时

第一节 数据挖掘过程

一、教学目标与重点

教学目标：

- 了解什么是数据挖掘,以及使用数据挖掘工具为信息社会的各行各业解决实际问题提供科学决策依据的意义。
- 初步了解数据挖掘的过程和基本算法。
- 了解数据管理与分析技术的新发展,特别是了解什么是大数据,以及在新发展环境下的数据挖掘的基本过程。

教学重点：

- 了解数据挖掘的过程和基本算法。

二、教学说明与建议

本节的内容是数据挖掘的过程,需要学生掌握数据挖掘的概念,而数据挖掘在现实中的应用可以通过项目活动和作业让学生自行查询资料来了解。本节结合生活中非常常见的电影数据,介绍聚类算法中的 k -平均算法,并对聚类结果进行简单分析。这个过程通过 Python 程序来实现,具体对学生的要求,可以根据学生前期对 Python 程序的熟悉程度来定。如果学生对 Python 程序不熟练,可以考虑只是简单模仿代码,直接运行即可;如果学生对 Python 程序有一定的掌握,可以考虑换一个数据集让学生练习。

本节 3 课时安排建议:“数据挖掘是什么”1 课时;“数据挖掘基本方法”2 课时。

三、项目实施与评价

1. 核心概念精解

(1) 数据挖掘

数据挖掘的目的是从数据中找出有用的信息或者知识。一般认为,数据挖掘是多领

域汇集的一种技术,包括统计学、人工智能、数据库,数据挖掘实现过程中还涉及并行技术、分布式技术等。

一般认为,数据挖掘能做的包括预测和描述两个方面。预测一般是指先训练一个模型,然后预测一个目标可能的属性值,这个训练出来的模型可以称为分类。例如,根据已知的花的特征预测其他花的种类;又如,根据买书或者不买书的用户的特征预测一个新用户会不会在网上书店买书。分类是用于预测离散目标对象的。而对于连续值目标对象,比如像股票的价格预测,则可以归为回归模型的预测任务。数据挖掘的描述任务一般被认为是探查性的,并且需要在数据挖掘结束之后进行一些数据解释和验证。这类任务中有这样的实例:一家商店收集销售数据之后,根据关联分析的模型,找到顾客经常同时购买的商品,找到关联规则,给商店提供商品销售的商机;每个持有信用卡的人都有信用记录,通过对产生欺诈行为的人的特征的分析,构造出一个合法交易的各种满足条件,如果一个新的交易与合法交易相比,有很多异常特征,那就可以标记这个交易为欺诈。

(2) 数据仓库

通常认为,数据仓库是在数据挖掘技术大量运用之前的一项技术,跟数据挖掘不同的是,数据仓库通常面向固定的主题,因而数据仓库的数据结构是相对稳定的。数据挖掘则对数据结构的要求没那么高,这是指数据挖掘的数据结构在数据预处理阶段会进行调整,在数据集成、选择、变换阶段可以将数据调整成适合数据挖掘的结构和形式。在数据挖掘过程中,数据仓库很可能是作为前期数据集成、数据选择等步骤中的数据存在形式,也就是说数据仓库是为了后期进行数据挖掘提供合适的数据形式。

数据仓库与传统数据库的区别在于:传统数据库一般存储的是当前的数据,并且为一些终端客户提供日常的查询操作,比如银行的客户进行存钱、取钱等操作;数据仓库一般面向主题,因而需要存储大量的历史数据,满足辅助决策者作决策。在使用的过程中,传统数据库可能会有比较频繁的查询操作,而数据仓库则通常不会频繁地进行操作,但是每次处理的数据量可能会比较大。

(3) 数据挖掘方法

数据挖掘一般会使用到的方法有:分类、聚类、关联规则。

分类是根据训练数据集和类标号属性,构建模型来分类现有数据,并用来分类新数据。实现数据分类一般要经过以下步骤:第一步,建立一个模型,描述预定数据类集和概念集。假定每个元组属于一个预定义的类,由一个类标号属性确定。其中包括训练数据集(由为建立模型而被分析的数据元组形成),其中的单个样本叫训练样本。学习模型可以用分类规则、判定树或数学公式的形式提供。第二步,使用模型,对将来的或未知的对象进行分类。首先评估模型的预测准确率,对每个测试样本,将已知的类标号和该样本的学习模型类预测比较,模型在给定测试集上的准确率是正确被模型分类的测试样本的百分比。测试集要独立于训练样本集,否则会出现“过拟合”的情况。典型分类算法包括决策树、贝叶斯分类、 k -近邻分类、支持向量机等。相应的连续值的预测同样需要根据已知模型估计未知的数据对象。典型的预测方法叫回归,包括线性回归、多元回归、非线性回归等。

聚类是将物理或抽象对象的集合分组成由类似的对象组成的多个类的过程。要最大

化类内的相似性和最小化类间的相似性。不像分类和预测分析标签类的数据对象,聚类分析数据对象不考虑已知的标签类。比如:对 Web 日志的数据进行聚类,以发现相同的用户访问模式;对城市进行规划,根据类型、价格、地理位置等来划分不同类型的住宅;在市场销售中帮忙决策,帮助市场人员发现客户中的不同群体,然后用这些知识来制定一个目标明确的市场计划。主要的聚类方法包括 k -均值法、层次聚类、基于密度的聚类(DBSCAN)等。

关联规则用于在交易数据、关系数据或其他信息中,查找存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构。比如购物篮分析。典型的方法有 Apriori、FP 增长算法等。

除此之外,还有异常检测(有时也叫离群点分析),有时在数据集中会出现与一般数据不一样的孤立数据,通常被称为“噪音”或异常,异常检测可以发现数据集中显著不同于其他数据的对象。可用于信用卡欺诈检测、移动电话欺诈检测、贷款审批、药物研究、气象预报、金融领域客户分类、网络入侵检测、故障检测与诊断等。

2. 项目活动的具体实施

(1) 项目任务 1:电影数据的聚类

① 展示数据集。数据本身都是真实数据,但是已经将电影名称替换为编号,可以将此视为数据预处理的一种体现。

② 介绍聚类算法的概念及应用场景。结合本数据集的特点,讨论聚类算法实施后可能的结果。

③ 取出将要进行聚类的数据(3 列数据),并介绍标称化数据的必要性。同时适当介绍 k -平均算法中的 K 值,该算法中的 K 值一般根据经验预估,但是也可能会在获得结果后重新调整 K 值。

④ 运行聚类的 Python 代码,将可视化结果呈现。代码先将数据聚类为 3 类,生成每类的概率密度图,了解每类数据可能的分布情况,可以比较明显地看到 3 类数据的值在不同区域内。再实现聚类结果的可视化。由于聚类数据有 3 列,所以需要降维处理成 2 维,才能以散点图的方式展现。

将聚类结果与电影数据集对照分析,分析聚类结果和原始数据中的类型之间的关系,通过分析,可以看出聚类有一定的科学性。聚类的结果对影院的排片有指导意义,而且可以用于后续的其他数据分析处理,比如分类。

(2) 项目任务 2:查询 3 个数据挖掘算法

学生自行查询数据挖掘算法,选择能理解的 3 个算法做成 PPT,在小组内进行交流。在此过程中不一定需要掌握代码实现,还是以理解算法为主。本活动考查学生数字化学习的能力和团队协作的精神。

3. 项目活动的评价

本节的项目评价主要包括:考查学生对活动中的代码实现的熟练程度;考查学生对常用算法的了解程度。

评价建议:采用过程性评价。对于实现活动中的 Python 代码的熟练程度进行评价;

让学生通过查询资料和讨论了解常用算法,对学生讨论中的表现进行评价。

四、作业练习与提示

■ 题目描述

商场想要推荐客户可能喜欢的商品。商场首先使用_____方法对客户群进行划分。根据物以类聚、人以群分的理念,同一类客户喜欢的商品可能比较相近,因此可以把同类客户平均购买次数或者购买数量最多的商品,推荐给这类客户中还没有买这些商品的人。

■ 作业提示

这是聚类方法比较典型的应用场景,根据场景的需要,由于不清楚客户有多少类,所以使用聚类方法把客户分群,对每一类客户标记上标签值。在此基础上,将同一类客户平均购买次数或者购买数量最多的商品推荐给该类客户中还没有购买这些商品的人,使推荐的成功率大大提高。

五、教学参考资源

■ 参考资料 1:数据挖掘的步骤

数据挖掘的步骤包括数据清洗、数据集成、数据选择、数据变换、数据挖掘、模式评估、知识表示。

数据清洗是为了处理采集到的数据中的一些噪声数据,比如重复数据、缺失的数据等,也就是说需要保证后续数据进一步处理的质量。这个过程也可能发生在数据集成之后。数据清洗,有时可以采取一些统计学方法进行处理。比如在“身高”这一字段中出现了负数,显然是错误的数据,又如,邮政编码中的数字颠倒了,导致该数字不可能是一个真实的邮编,这些都属于容易检测出来的,需要进一步考虑如何纠正。还有的是缺失了数据,也可以通过一些统计方法去填补,比如,如果是数值型数据,可以用这列字段数据的均值、中值或者众值去填充。另外,如果出现重复数据,显然需要去重。在处理这些噪声数据的过程中,很可能需要借助人工的力量完成。用人工智能的方式完全解决数据清洗的问题,这仍然是人工智能领域中的前沿热点技术。

数据集成、数据选择、数据变换是为了进行数据挖掘所做的数据准备中的几个步骤。数据集成是将多数据源进行融合,数据选择是在已经获取的数据中抽取与分析任务相关的数据,数据变换是为了将数据的格式统一成符合分析任务的格式。为了进行数据挖掘,采集的数据通常多源、异构、多维度,可能是数据库中的结构化数据,也可能是从多个新闻网站上爬取下来的文本数据即非结构化数据,而且数据类型也可能是多样化的。最终要进行数据挖掘的数据需要是结构化数据,因此将多数据源数据融合起来,势必需要统一数据的类型,并且填入结构化表中的时候,还要将数据填入正确的字段中。在这个过程中有一个难点:同一个对象的描述方法可能是不一致的,比如,对于“IBM 公司”“蓝色巨人”

“国际商业机器公司”，人可以辨别出这三者是指同一对象，但是要让计算机去判别三者是否一致是比较难的。这一难题有时也称为 DB Hard 问题。有时数据的量比较大，但是又比较稀疏，比如，用一个矩阵存储用户购买商品，显然用户不可能购买所有商品，所以这个矩阵中大量是 0，是一个稀疏矩阵，为了适合数据挖掘，还需要将稀疏数据进行压缩。最后交付给数据挖掘模型的数据应该是结构化的，并且是现有软硬件能够支撑进行计算的。

数据挖掘步骤与数据挖掘这个概念重合了，这里的步骤主要是指数据建模和数据分析过程。数据挖掘的建模过程，包括了分类、回归、聚类、关联规则等。教科书中对分类、聚类、关联规则均作了介绍。回归与分类的区别在前文中也进行了阐述。

知识表示这一步骤主要是为了实现数据挖掘的最终目的。数据挖掘通常从一个假设开始，也可能是在进行数据挖掘之后再进行假设并验证。而可视化的表示能让人更能被赋予思考的能力。比如，在新冠肺炎疫情期间，通过在网上查询到的以地图形式出现的各省的确诊人数图，我们可以一目了然地看到不同地区的疫情现状。同时，每日新增确诊人数的折线图被用于判断何时可以复工复产。

■ 参考资料 2：数据挖掘和机器学习

数据挖掘受到了很多学科领域的影响，其中数据库、机器学习、统计学无疑影响最大。粗糙地说，数据库提供数据管理技术，机器学习和统计学提供数据分析技术。由于统计学界往往醉心于理论的优美而忽视实际的效用，因此，统计学界提供的很多技术通常都要在机器学习界被进一步研究，变成有效的机器学习算法之后才能再进入数据挖掘领域。从这个意义上说，统计学主要是通过机器学习来对数据挖掘产生影响，而机器学习和数据库则是数据挖掘的两大支撑技术。

从数据分析的角度来看，绝大多数数据挖掘技术都来自机器学习领域。但能否认为数据挖掘只不过就是机器学习的简单应用呢？答案是否定的。一个重要的区别是，传统的机器学习研究并不把海量数据作为处理对象，很多技术是为处理中小规模数据设计的，如果直接把这些技术用于海量数据，效果可能很差，甚至可能无法使用。因此，数据挖掘界必须对这些技术进行专门的、不简单的改造。例如，决策树是一种很好的机器学习技术，不仅有很强的泛化能力，而且学得结果具有一定的可理解性，很适合数据挖掘任务的需求。但传统的决策树算法需要把所有的数据都读到内存中，在面对海量数据时这显然是无法实现的。为了使决策树能够处理海量数据，数据挖掘界做了很多工作，例如通过引入高效的数据结构和数据调度策略等来改造决策树学习过程，而这其实正是在利用数据库界所擅长的数据管理技术。实际上，在传统机器学习算法的研究中，在很多问题上如果能找到多项式时间的算法可能就已经很好了，但在面对海量数据时，可能连 $O(n^3)$ 的算法都是难以接受的，这就给算法的设计带来了巨大的挑战。

另一方面，作为一个独立的学科领域，必然有一些相对“独特”的东西。对数据挖掘来说，就是关联分析。简单地说，关联分析就是希望从数据中找出“买尿布的人很可能会买啤酒”这样看起来匪夷所思但可能很有意义的模式。如果在 100 位顾客中有 20 位购买了尿布，购买尿布的这 20 位顾客中有 16 位购买了啤酒，那么就可以写成“尿布 → 啤酒 [支

持度 = 20%，置信度 = 80%]”这样的一条关联规则。挖掘出这样的规则有很多用处，例如商家可以考虑把尿布展柜和啤酒展柜放到一起以促进销售。实际上，在面对少量数据时关联分析并不难，可以直接使用统计学中相关性的知识，这也正是机器学习界没有研究关联分析的一个重要原因。关联分析的困难其实完全是由海量数据造成的，因为数据量的增加会直接造成挖掘效率的下降，当数据量增加到一定程度，问题的难度就会产生质变。例如，在关联分析中必须考虑因数据量太大而无法承受多次扫描数据库的开销、可能产生在存储和计算上都无法接受的大量中间结果等，而关联分析技术正是围绕着“提高效率”这条主线发展起来的。

■ 参考资料 3：参考书

1. Pang-Ning Tan, Michael Steinbach, Vipin Kumar. 数据挖掘导论[M]. 范明, 范宏建, 等译. 北京: 人民邮电出版社, 2006.
2. Jiawei Han, Micheline Kamber, Jian Pei. 数据挖掘: 概念与技术(原书第 3 版) [M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2012.
3. 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

六、教学参考案例

■ 参考案例

数据挖掘过程

华东师范大学第二附属中学(紫竹校区) 郭威
(1课时)

1. 学科核心素养
 - 认识数据管理与分析技术对人类社会生活的重要影响。(信息意识)
 - 会采用适当的方法提取数据;能正确选用数据分析方法和工具分析并解释数据。(计算思维)
 - 能根据需要,主动选用数字化工具开展自主或协作学习,创造性地解决问题。(数字化学习与创新)
2. 《课程标准》要求
 - 结合生活实际,认识到数据是一种重要的资源,通过科学管理与分析数据,可以使数据实现其应有价值,感受数据管理与分析技术的重要性。
 - 结合恰当的案例分析,认识数据挖掘对信息社会问题解决和科学决策的重要意义。
3. 学业要求
 - 解释数据挖掘的概念。
 - 列举数据挖掘的一般过程和常用方法。
 - 处理数据,并分析数据,得到准确的结论。

4. 教学内容分析

本节课主要介绍数据挖掘技术的入门知识,阐述数据挖掘技术的概念与过程,使学生结合实例完成数据挖掘的应用与分析,既强调概念性知识的掌握,又注重技术应用,课程内容与生活中的实例息息相关。

5. 学情分析

通过之前的学习,学生已经了解了数据管理与分析的基础知识,认识到数据的重要性,以及数据管理与分析技术对人类社会生活的重要影响,具备了一定的数据采集、数据管理、数据分析和数据可视化的技能,能够通过案例实践,理解和掌握数据挖掘的过程。

6. 教学目标

- 了解什么是数据挖掘,以及数据挖掘对信息社会问题解决和科学决策的重要意义。
- 初步了解数据挖掘的过程和基本算法。

7. 教学重难点

- 教学重点:数据挖掘的意义、数据挖掘的概念、数据挖掘的方法。
- 教学难点:使用聚类方法进行数据挖掘的过程。

8. 教学准备

准备活动记录单、Python示例程序、示例数据表。

9. 教学策略分析

教学实施过程中,结合网络购物智能推荐、人工智能医生等生活实例,采用启发式教学策略,引导学生思考和感悟信息社会中数据和数据挖掘的意义与价值,通过创设“电影投资人”的问题情境,引导学生以小组为单位,使用聚类方法完成数据挖掘任务,分析数据,制定电影投资策略,与同学分享小组探究成果。

10. 教学环境

网络机房、广播教学软件、Anaconda集成开发环境。

11. 教学过程设计(见表5.2)

表5.2 教学过程设计表

教学环节	教学内容	学生活动	设计意图
课前准备	将活动记录单与示例程序、示例数据发给学生	浏览材料,初步了解课程内容,明确分组任务	初识任务、明确目标
情境导入	1. 介绍课程主题; 2. 播放视频:人工智能医生扫描眼底预测心脏病风险; 3. PPT展示案例:网络购物智能推荐、超市售货规划、影院排片计划。 提问:结合以上案例思考:数据给社会带来了什么影响?你还能举出哪些实例?	结合自己的认知,思考并回答问题,与同学分享	引发思考、引起兴趣、感悟数据挖掘的意义

续表

教学环节	教学内容	学生活动	设计意图
讲授新知	总结回答、提出数据挖掘的概念,阐释数据挖掘的意义,讲授数据挖掘的一般步骤	观看课程演示,理解新概念,学习新知识	学习新知
活动1 (自主学习)	布置小组学习任务: 小组合作,使用网络平台,搜索“数据挖掘的基本方法”,归纳整理出三种常见的数据挖掘方法,理解其概念,初步了解其原理和方法,整理材料与同学交流分享	小组合作,完成学习任务	采用小组自学的方式,激励学生主动学习,提高学习效果
讲授新知	归纳提炼学生的回答,讲授新知	巩固概念,掌握新知识	学习新知
活动2 (小组实践)	创设情境:每个小组的两位同学作为“电影投资人”,将要为未来的电影投资制定策略,现在有一份近半年的电影数据,请结合数据表,决定未来的投资方向,并阐述理由; 引导学生结合活动记录单,运行示例程序,进行数据挖掘,分析数据结果,制定策略并注明理由	小组合作编写程序,补充活动记录单,完成学习任务	通过实践练习,体验和理解聚类分析方法,学以致用,感悟数据挖掘的实际意义
活动3 (小组分享)	邀请两组学生以“电影投资人”的身份,和同学分享小组探究成果(投资策略)、探究过程和方法,点评并引导学生自我评价与互评	交流分享,相互点评	通过课堂分享和点评,增加课堂互动,活跃课堂氛围,提升课堂趣味,加深学生对知识的理解
课堂总结	梳理归纳学习内容,回顾与总结。 提问:谈谈对数据挖掘的学习感悟	归纳总结,交流学习感悟	巩固知识,加深学生对学习主题的理解

【活动记录单】

数据挖掘过程

1. 观看视频和课件案例,简要描述数据给社会带来了什么影响,并想想还能举出哪些实例。

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

3. 按照数据挖掘的一般步骤,对下列步骤进行排序:_____。(填写序号)

- ① 数据选择; ② 数据挖掘; ③ 数据清洗; ④ 知识表示; ⑤ 数据变换;
⑥ 模式评估; ⑦ 数据集成。

4. 自主学习:小组合作,使用网络平台,搜索“数据挖掘的基本方法”,归纳整理出三种常见的数据挖掘方法,简要描述其概念和原理。

① _____

② _____

③ _____

5. 数据挖掘基本方法。

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

(2) _____ 是找出描述并区分数据类的模型,以便能够使用模型预测一个新的对象的类标号。

(3) _____ 是根据在数据对象中发现的描述对象及其关系的信息,将数据对象分组并标记类标号。

6. 小组实践:你和你的同伴作为“电影投资人”,将要为未来的电影投资制定策略。现在有一份近半年的电影数据 film.csv,请使用 Anaconda 导入并运行示例程序 film.ipynb,根据程序运行后显示的图表内容,分析数据结果,决定未来的投资方向,并简要阐述决策理由。

第二节

大数据时代下的数据管理与分析技术的发展

一、教学目标与重点

教学目标:

- 了解大数据对社会产生了什么样的影响。
- 了解数据管理与分析技术的新发展。

教学重点:

- 理解大数据的概念。

二、教学说明与建议

本节的内容主要是大数据时代下数据挖掘的一些特点,以及大数据对社会产生的影响。

首先,介绍大数据的概念及特点,帮助学生了解大数据时代下数据挖掘与传统数据挖掘的一些区别,并简单介绍如果电影数据的数据量达到了大数据级别,可能产生的一些问题及可能解决的方法。

然后,简单介绍大数据平台,引导学生考虑大数据时代下的安全伦理并讨论。

最后,介绍数据管理与分析技术目前的发展情况,使学生能对数据管理与分析技术有一些全局观。

三、项目实施与评价

1. 核心概念精解

(1) 大数据

半个世纪以来,随着计算机技术全面融入社会生活,信息爆炸已经积累到了一个开始引发变革的阶段。它不仅使世界充斥着比以往更多的信息,而且其增长速度也在加快。互联网(社交、搜索、电商)、移动互联网(微博)、物联网(传感器、智慧地球)、车联网、GPS、医学影像、安全监控、金融(银行、股市、保险)、电信(通话、短信)都在疯狂产生着数据。

人工智能正在成为计算机领域中最被关注的领域之一,人工智能的出现早在 20 世纪 40 年代计算机诞生的时候就被关注,但是起起伏伏几十年,直到近些年才真正被重新重视,其中很重要的原因是:虽然计算机并不能很好地解决人工智能中的诸多问题,但是利用大数据突破性地解决了,其核心问题变成了数据问题。

大数据为何叫“Big Data”呢?在英语中,large、vast 和 big 都可以用于形容大,其中 big 更强调的是相对大小的大,是抽象意义上的大。所以大数据是抽象的大,是思维方式上的转变。在大数据时代,需要了解的是,数据量的暴增使量变带来质变。与此同时,思维方式、方法论都应该和以往不同。

大数据的三大思想是:第一,数据需要全集数据而非抽样数据;第二,数据处理的结果需要效率高,并能包容数据的混杂性,而不是注重于精确;第三,对于数据处理的结果分析关注的是相关性而不是因果性。例如 2009 年的时候一家搜索引擎公司曾经根据搜索关键词预测出当时的流感大流行,比疾控中心的结论要早两个星期,此为大数据技术的一个显著成就。然而,2013 年时,该公司同样的预测与真实结果相差甚远。这不仅是因为全集数据无法真正获得,该公司当年需要不断加入其他部门提供的数据去调整自己的预测,还因为大数据本身更注重相关性而非因果性的特点,使得大数据可能会忽略一些背景知识的影响。

正是由于大数据的思想是关注相关性而非因果性,所以往往分析结果的相关性缺乏人性伦理的一些思考,会造成对人们隐私和尊严的侵犯。这也是使用大数据技术的时候特别需要关注的点。

(2) 大数据平台

大数据分析用到的数据分析方法除了数据挖掘的经典方法外,还包括机器学习、神经网络、模式识别等具体的人工智能的方法。这些方法很多时候并不严格划分各自的地盘,

数据分析、数据挖掘、机器学习等概念之间的界限逐渐淡化,它们往往被统一称为数据挖掘。

大数据分析最终需要以计算的形态呈现出来,所以大数据分析除了掌握数据分析的技术,还要有工具,包括 SQL、Python、R、SAS、Scala、Java 等。这些工具在大数据分析中有相应的基础设施的支撑,比如 SQL 可以运作在 Cloudera 的 IMPALA 上、Apache 的 Hive 上,Python 可以运作在 Spark 上,R 和 SAS 可以运作在 Hadoop 的 MapReduce 和 Spark 上。

现在有一种流水线可以将数据采集、数据准备、数据存储、数据处理、数据查询等整合到一起,实现一站式的大数据分析。在大数据分析的各个阶段都可以在这个流水线中找到相应的工具去实现相应功能。更进一步地,现在有云技术,为所有的工具提供了云平台,使大数据分析的实现更为便利。

大数据平台中著名的 Hadoop 平台是由 Apache 开发的,它是一个分布式系统的基础架构,用于可靠高效地实现大数据的处理,其核心包含 HDFS 和 MapReduce。HDFS 为海量的数据提供了存储服务,而 MapReduce 则为海量的数据提供了计算服务。另外,Hadoop 还包括了数据仓库工具 Hive 和分布式数据库 HBASE,形成了 Hadoop 生态系统,常用于描述大数据分析过程中所有工具集合的平台。在大数据发展过程中,Hadoop 生态系统中逐步改进了一些工具,并添加了一些工具,Hadoop 从 Hadoop1.0 已发展到 Hadoop2.0。

还有一个流行的类似开源的大数据计算环境是 Spark,Spark 与 Hadoop 之间存在一些不同之处,Spark 在某些工作负载方面表现得更加优越。

2. 项目活动的具体实施

项目任务 3: 大数据下的电影数据的数据挖掘。

由于大数据平台的实现存在一定的难度,因此本项目任务,主要让学生从理论上了解大数据的存储方式即可,不一定要求具体实现。

3. 项目活动的评价

本节的项目评价主要包括:考查学生对大数据的特征的了解;考查学生对大数据处理问题过程中各种可能的情况的了解。

评价建议:采用过程性评价。主要引导学生进行思考和讨论,对学生在讨论过程中的表现进行评价。

四、作业练习与提示

■ 题目描述

请探讨大数据环境下的数据挖掘算法如何解决问题。

■ 作业提示

首先描述大数据的概念和特点;然后阐述数据挖掘算法有哪些常见的算法和应用,以及大数据环境下的数据挖掘与传统数据挖掘的不同之处;最后针对这些不同之处提出如

何解决问题。

建议从大数据的 4V 特征考虑。数据量大,则在空间时间复杂度问题上,数据规模的增长给算法的复杂度带来的影响使得必须考虑尽量选择简单模型,并且必须考虑数据的压缩。数据处理的时候用到的计算资源,在大数据的集成平台上实现的时候,都要采用分布式的形式,并且大数据分析的工具现在都在云平台上实现。

五、教学参考资源

■ 参考资料 1: 大数据时代下的隐私保护

1. 数据 vs 隐私

在大数据时代,数据成为了科学的研究的基石。我们在享受着推荐算法、语音识别、图像识别、无人驾驶等智能的技术带来的便利的同时,数据在背后担任着驱动算法不断优化迭代的角色。在科学的研究、产品开发、数据公开的过程中,算法需要收集、使用用户数据,在这个过程中数据就不可避免地暴露在外。历史上就有很多公开的数据暴露了用户隐私的案例。

2006 年 8 月,为了学术研究,一家互联网服务公司公开了匿名的搜索记录,其中包括 65 万个用户的数据,总共 20M 条搜索记录。在这些数据中,用户的姓名被替换成一个个匿名 ID,但是一家报社通过这些搜索记录,找到了匿名 ID 为 4417749 的用户在真实世界中对应的人。最后这家公司紧急撤下数据,发表声明致歉,但是已经太晚了。因为隐私泄露事件,这家公司遭到了起诉,最终向受影响的用户支付了高额赔偿金。

同样是 2006 年,一家影视公司举办了一项预测算法的比赛,比赛要求在公开数据上推测用户的电影评分。该公司把数据中唯一识别用户的信息抹去,认为这样就能保护用户的隐私。但是在 2007 年,两位研究人员表示通过关联该公司公开的数据和某网站上公开的记录就能够识别出匿名后用户的身份。3 年后,在 2010 年,该公司因为隐私原因宣布停止这项比赛,并因为隐私安全问题受到高额罚款。

近几年各大公司均持续关注用户的隐私安全。在大数据时代,如何才能保护我们的隐私呢?要回答这个问题,我们首先要知道什么是隐私。

2. 什么是隐私

我们经常谈到隐私泄露、隐私保护,那么什么是隐私呢?举个例子,居住在北京海淀区五道口的小明经常在网上购买电子产品,那小明的姓名、购买偏好和居住地址算不算隐私呢?如果某购物网站统计了用户的购物偏好并公开部分数据,公开的数据中显示北京海淀区五道口的用户更爱买电子产品,那么小明的隐私是否被泄露了呢?要弄清楚隐私保护,我们先要讨论一下究竟什么是隐私。

对于隐私这个词,科学的研究上普遍接受的定义是“单个用户的某一些属性”,只要符合这一定义都可以被看作隐私。我们在提“隐私”的时候,更加强调的是“单个用户”。那么,一群用户的某一些属性,可以认为不是隐私。在上文提到的例子中,针对小明这个单个用户,“购买偏好”和“居住地址”就是隐私。如果公开的数据显示,住在五道口的小明爱买电

子产品,那么这显然就是隐私泄露了。但是如果数据中只包含一个区域的人的购买偏好,就没有泄露用户隐私。进一步讲,如果大家都知道小明住在海淀区五道口,那么是不是大家就知道了小明爱买电子产品呢?这种情况算不算隐私泄露呢?答案是不算,因为大家只是通过这个趋势来推测,数据并不显示小明一定爱买电子产品。

所以,从隐私保护的角度来说,隐私是针对单个用户的概念,公开群体用户的信息不算是隐私泄露,但是如果能从公开数据中准确推测出个体的信息,那么就算是隐私泄露。

3. 隐私保护的方法

从信息时代开始,关于隐私保护的研究就开始了。随着数据不断地增长,人们对隐私越来越重视。隐私保护存在两种情况。

第一种情况是公司为了学术研究和数据交流开放用户数据,学术机构或者个人可以向数据库发起查询请求,公司返回对应的数据时需要保证用户的隐私安全。

第二种情况是公司作为服务提供商,为了提高服务质量,主动收集用户的数据,这些在客户端上收集的数据也需要保证隐私安全。学术界提出了多种保护隐私的方法和测量隐私是否泄露的工具,例如 k - 匿名化(k-anonymity)、差分隐私(differential privacy)、同态加密(homomorphic encryption)、零知识证明(zero-knowledge proof)等。

4. 实际案例

在实际应用中使用差分隐私时需要考虑的问题还有很多,我们在介绍差分隐私的时候假设所有的查询操作都由可信的数据库处理,数据库里存储着用户的原始数据。那么如果数据库被攻击了,包含用户隐私的原始数据就泄露了。

如果不收集用户的原始数据,在客户端上先做差分隐私,再上传给服务器,这个问题就解决了。比如互联网公司通过差分隐私的方法收集用户数据,并基于“随机应答”方法保护用户原始数据不被泄露。“随机应答”的流程如下:

(1) 当用户需要上报个人数据的时候,首先“抛硬币”决定是否上报真实数据。如果是正面,则上报真实数据。如果是反面,就上报一个随机的数据,再“抛一次硬币”决定随机数据的内容。

(2) 服务器收到所有的数据后,因为知道“抛硬币”是正面的概率,服务器就能够判断返回的数据是正确的概率。

这种“随机应答”的方法在理论上也被证明是服从差分隐私的。对于用户来说,隐私数据在上报给服务器之前就已经加了噪声,从而具有一定保证。对于公司来说,也能收集到有效的数据。

“随机应答”方法克服了之前只能回答简单查询语句的限制,现在可以上报包含字符串这类更加复杂的回答。在上报字符串信息的时候首先使用布隆过滤器(bloom filter)算法把字符串哈希到一个数组中,然后再加入噪声传给服务器。布隆过滤器不需要存储元素本身,并可以用于检索一个元素是否在一个集合中。通过使用这种方法,就可以对字符串数据添加噪声,保护用户的隐私。

上文介绍的模型都是先在本地做差分隐私,然后再上报给服务器,这种方法叫做本地

模式(local mode)。这种差分隐私的做法在上报数据可以相互关联的情况下还是存在隐私泄露的可能性。

另外,哈佛大学公开了一个名为 PSI(Ψ)的项目,提供了一个便捷的差分隐私工具。使用者通过上传数据,调整差分隐私的参数,就可以获得满足差分隐私的数据集。

5. 总结

从最开始的 k - 匿名化到现在的差分隐私,都是为了既保证用户的个人隐私安全,也能对实际应用和研究提供有价值的数据。在大数据时代,希望各公司在利用数据提供更好的服务的同时,能保护好用户的个人隐私。这是法律的要求,也是安全行业的追求。我们相信隐私保护技术会越来越受到重视,并从学术理论迅速投入工业界实战应用。

■ 参考资料 2:参考书

1. 维克托·迈尔-舍恩伯格,肯尼思·库克耶. 大数据时代[M]. 盛杨燕,周涛,译. 杭州:浙江人民出版社,2013.
2. 黄宜华. 深入理解大数据:大数据处理与编程实践[M]. 北京:机械工业出版社,2014.
3. 吴军. 智能时代:大数据与智能革命重新定义未来[M]. 北京:中信出版社,2016.

六、教学参考案例

■ 参考案例

大数据时代下的数据管理与分析技术的发展

华东师范大学第二附属中学(紫竹校区) 郭威

(1课时)

1. 学科核心素养

- 能认识有效管理与分析数据对获取有价值信息、形成正确决策的作用与意义,认识数据管理与分析技术对人类社会生活的重要影响。(信息意识)

- 能根据需要,主动选用数字化工具开展自主或协作学习,创造性地解决问题。(数字化学习与创新)

2. 《课程标准》要求

- 结合生活实际,认识到数据是一种重要的资源,通过科学管理与分析数据,可以使数据实现其应有价值,感受数据管理与分析技术的重要性。

- 运用数字化学习方式,了解数据管理与分析技术的新发展;结合恰当的案例分析,认识数据挖掘对信息社会问题解决和科学决策的重要意义。

3. 学业要求

- 解释大数据的概念,描述大数据的特征。
- 描述大数据环境下的数据挖掘过程。
- 举例说明大数据时代下数据管理与分析技术的发展。

4. 教学内容分析

本节课引入大数据的概念,探讨大数据时代下数据管理与分析技术的发展,引起学生对大数据时代下个人隐私等信息安全问题的思考。

5. 学情分析

通过上节课的学习,学生已经了解了数据挖掘的概念、意义和过程,掌握了使用聚类方法实现“电影数据分析”的数据挖掘过程,对数据管理与分析技术有了较为全面的认识,能够在大数据时代的背景下,探讨数据挖掘技术的应用、数据安全和大数据分析技术的发展。

6. 教学目标

- 了解数据管理与分析技术的新发展。
- 了解什么是大数据。
- 了解大数据时代下数据挖掘的基本过程。

7. 教学重难点

- 教学重点:大数据的概念、大数据时代下数据管理与分析技术的新发展。
- 教学难点:大数据对社会发展产生的影响。

8. 教学准备

准备辩论活动主题。

9. 教学策略分析

本节课内容以陈述性知识为主,需要学生了解大数据的概念、大数据挖掘、大数据处理平台、大数据分析的发展等内容,并且探讨大数据时代下的信息安全问题。课程设计采用数字化学习模式,引导学生通过网络工具查询课程中涉及的名词概念,提升学生的自主学习能力,同时,以“对个人而言,大数据时代下的数据挖掘利大于弊还是弊大于利”为主题,在课堂中组织自由式攻辩,通过辩论的形式激发学生对大数据时代下信息安全问题的深度思考。

10. 教学环境

网络机房、广播教学软件。

11. 教学过程设计(见表 5.3)

表 5.3 教学过程设计表

教学环节	教学内容	学生活动	设计意图
情境导入	播放视频:央视纪录片“大数据时代”片头	观看视频,思考“大数据”的含义	通过观看视频,激发学生的学习兴趣,引出学习主题
活动 1 (自主学习)	1. 布置学习任务:以“大数据”为搜索主题,使用网络平台,查询有关大数据的概念和特征,并列举 3 个和大数据紧密相关的应用实例; 2. 组织学生进行知识分享交流	使用网络平台,完成自学任务,并与同学交流分析	激励学生自主学习,提升学习效果

续表

教学环节	教学内容	学生活动	设计意图
讲授新知	讲授大数据的概念、大数据挖掘、大数据处理平台、大数据分析的发展等知识内容	观看课件,学习新知识	通过讲述,让学生了解大数据时代下的数据管理与分析技术的发展
活动 2 (辩论)	1. 公布辩论主题“对个人而言,大数据时代下的数据挖掘利大于弊还是弊大于利”,组织学生分为正反两方,各持一种观点进行辩论; 2. 引导学生使用网络资源组织论据; 3. 主持自由式攻辩,并维持纪律; 4. 点评与总结观点,引导学生辩证看待大数据时代下的信息安全问题	准备辩论材料、参与辩论环节、思考大数据时代下的信息安全问题	信息安全是大数据时代下的一个重要议题,通过辩论的形式,可以激发学生的深度思考,活跃课堂氛围,提升学习效果
课堂总结	梳理归纳学习内容,回顾总结	回顾总结	巩固知识,加深理解

经上海市中小学教材审查委员会审查
准予使用 淮用号 II-GJ-2022022

责任编辑：曹祖红



绿色印刷产品

ISBN 978-7-5760-2953-6

A standard EAN-13 barcode representing the ISBN number 978-7-5760-2953-6.

9 787576 029536 >

定价：24.00 元