



普通高中教科书

信息技术

必修 1

数据与计算



上海科技教育出版社

普通高中教科书

信息技术

必修 1

数据与计算



上海科技教育出版社

编写人员名单

主 编：郑 骏

分册主编：郑 骏

主要编写人员（以姓氏笔画为序）：

卢 源 朱晴婷 陈 凯

房爱莲 胡 杨 郭 骏

曹红霞

欢迎广大师生来电来函指出教材的差错和不足，提出宝贵意见。

上海科技教育出版社地址：上海市闵行区号景路 159 弄 A 座 8 楼

邮政编码：201101

联系电话：021-64702058

邮件地址：office@sste.com

亲爱的同学：

不知道你是否留意，我们几乎每天都能听到“数据”这个词。数据和每个人的生活都密切相关，它不仅是信息的载体，也是人们提取信息、作出决策的重要依据，并逐步成为社会发展的一项资源。人们合理选用技术工具处理数据，可以提高数据应用效能，发现其中隐含的信息，精准解决生活与学习中的问题。

在《数据与计算》的学习中，我们将带领你理解数据、信息和知识的相互关系，体验利用数字化工具处理数据和发现信息的过程；利用一种程序设计语言编写程序，实现简单算法，经历计算机解决问题的整个过程。你将在运用数字化工具的学习活动中，理解当今数字化世界的运转方式，提高利用信息技术解决问题的能力，发展信息意识和信息社会责任，养成数字化学习与创新的习惯。

为了让你在学习《数据与计算》的过程中获得更大的成功，请浏览本书的栏目介绍。



单元引言、学习目标和单元挑战

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



项目引言和学习目标

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



项目学习指引

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

核心概念和小贴士

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

思考与讨论??

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

数字化学习

引导你利用网络、数字化工具和数字资源进行学习。

活动

提出活动任务，并引导你运用所学知识，使用信息技术工具进行探究、总结和展示。



知识链接

系统整理和归纳本项目的知识要点，方便你学习。

拓展阅读

补充更丰富的阅读材料，开阔你的视野。

单元挑战

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

单元小结

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

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

编者

目 录



第一单元 数据与信息	1
项目一 探秘鸟类研究——认识数据、信息与知识	2
1. 采集鸟类活动的数据	3
2. 处理数据，获取信息	5
3. 利用大数据获取信息	7
知识链接	9
项目二 探究计算机中的数据表示——认识数据编码	12
1. 从树牌号认识编码	13
2. 了解数值数据和文本数据的编码	14
3. 了解声音和图像的数字化	16
知识链接	18
单元挑战 认识并制作二维码	25
单元小结	26
第二单元 数据处理与应用	27
项目三 调查中学生移动学习现状——经历数据处理的一般过程	28
1. 明确数据需求	29
2. 采集数据	30
3. 加工、分析和可视化数据	35
4. 撰写报告，提出数据应用建议	37
知识链接	38
项目四 认识智能停车场中的数据处理——体验数据处理的方法和工具	43
1. 探究停车引导中的数据处理	44
2. 计算停车费	48
3. 分析停车位使用数据	51
知识链接	56
单元挑战 采集与分析气象数据	64
单元小结	65
第三单元 算法和程序设计	67
项目五 描述洗衣机的洗衣流程——了解算法及其基本控制结构	68
1. 从洗衣流程认识算法	69
2. 描述“洗涤算法”	71

3. 分析洗衣流程的控制结构.....	72
知识链接.....	74
项目六 解决温标转换问题——认识程序和程序设计语言	79
1. 体验程序设计的一般过程.....	80
2. 了解程序的基本控制结构.....	83
3. 优化程序，判断输入有效性.....	84
知识链接.....	85
项目七 用计算机计算圆周率——设计简单数值数据算法	92
1. 设计算法实现用数学公式计算.....	93
2. 设计算法实现用随机投点法计算.....	95
知识链接.....	99
项目八 分析历史气温数据——设计批量数据算法	106
1. 用列表表示和计算平均气温.....	107
2. 用模块化设计批量计算平均气温.....	110
知识链接.....	115
单元挑战 探究密码安全问题	124
单元小结	125
第四单元 人工智能初步	127
项目九 了解手写数字识别——体验人工智能	128
1. 初识字符识别技术.....	129
2. 了解机器学习中的数据采集与预处理.....	130
3. 建立手写数字识别模型并进行验证.....	133
4. 评估手写数字识别模型并开展应用.....	135
知识链接.....	137
单元挑战 尝试用人工智能绘画	142
单元小结	143
附录 部分名词术语中英文对照	145

第一单元

数据与信息

在现实世界中，每个人每天会产生大量数据，如去过哪里、买过什么商品、走了多少路等。这些看似平凡的数据却蕴含了大量的信息，如果善加利用，会给社会创造意想不到的价值。例如，电商平台根据用户的浏览和购买记录，有针对性地向用户推荐商品，以提高商品销量；智能手环告诉佩戴者每天走了多少步、消耗了多少热量、深度睡眠有多长时间等，并提供保健建议，甚至推荐相应的健身产品；无人驾驶汽车使用摄像头、车载雷达、激光测距仪等设备采集数据，识别周围的交通状况，利用实时更新的地图进行自动导航，实现无人驾驶。

那么，究竟什么是数据？什么是信息？数据在计算机中是如何表示和处理的？本单元将带领大家揭开数据与信息的神秘面纱。



学习目标

- ◆ 通过具体实例，感知数据与信息，描述数据与信息的特征。
- ◆ 理解数据、信息与知识的关系，认识数据对人们日常生活的影响。
- ◆ 知道数值、文本、声音、图像等各类型数据的基本编码方式。

单元挑战

认识并制作二维码

项目一

探秘鸟类研究

——认识数据、信息与知识

为了解决各种问题，各行各业的人们都在做着采集数据、获取信息甚至构建知识的工作。例如，商店采集顾客购买的商品等数据，获取顾客购物喜好、商品畅销程度等信息，甚至构建以顾客为导向的市场营销战略知识，以更好地开展商品营销活动。又如，科学家长期在野外采集鸟类活动的数据（图 1-1），获取鸟类分布、鸟类对栖息地的选择等信息，从而构建鸟类与植物关系的知识，用于开展鸟类保护工作和生物多样性研究工作。



图 1-1 采集鸟类活动的数据

项目学习目标

在本项目中，我们将探究一群科学家的鸟类研究活动，了解科学家是如何采集数据、获取信息的。

完成本项目学习，须回答以下问题：

1. 什么是数据？什么是信息？数据和信息的特征有哪些？
2. 数据、信息与知识的关系是什么？
3. 什么是大数据？大数据的特征有哪些？

项目学习指引

1. 采集鸟类活动的数据

空中掠过几只鸟，转瞬消失在树林里，难觅踪迹。这些美丽的精灵栖息在哪里？它们喜欢怎样的林木环境？为回答这些问题，科学家在浙江某国家森林公园的一片实验林地里设立了一个国家野外科学观测研究站，并长期在那里采集各种数据。

通过观察、测量等工作，林地里的鸟类活动数据，如鸟的种类、数量、行为等，被定期记录了下来。这些描述鸟类活动的数据，有数值、文本、图形、图像等形式。为了更好地存储、处理这些野外采集来的数据，科学家将它们录入计算机中，如表 1-1、图 1-2 和图 1-3 所示。

核心概念

数据 (data) 是对客观事物属性的描述，是记录下来的某种可以识别的符号。在计算机科学中，数据是指所有能输入计算机中并能被计算机程序处理的符号的总称。

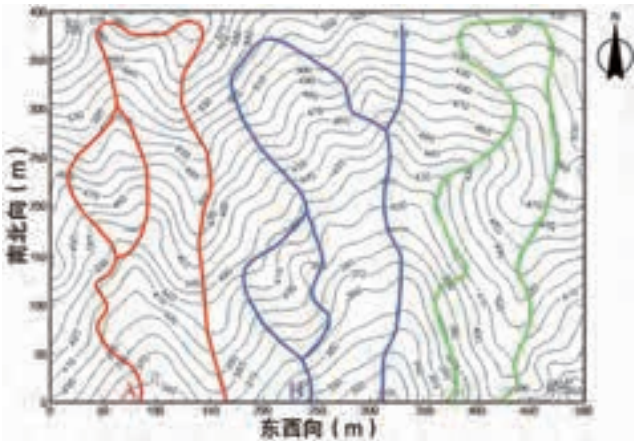
表 1-1 林地鸟类活动调查记录表

记录人：××× 记录时间：×××× 年 ×× 月 ×× 日

鸟的种类	数量(只)	栖息取食基层	行为	树牌号
灰眶雀鹛	5	灌丛	鸣叫	2130198
灰眶雀鹛	6	灌丛	鸣叫	2130123
白头鹎	5	冠中上	鸣叫	4080059
黄眉柳莺	2	冠中上	觅食	3080061
黑鹂	9	冠上	飞行	5060137
黄眉柳莺	1	冠中下	跳跃	1090013
红头长尾山雀	20	整个冠层	觅食，啄树干	2100030



图 1-2 用数码相机拍摄的灰眶雀鹛



(A、B、C 为林地调查线路)

图 1-3 用 GIS (地理信息系统) 绘制的林地调查路线图

小贴士

信息技术的发展使得人们采集和处理数据的手段不断加强，数据的内涵也逐渐丰富。在计算机发明前及发明初期，“数据”更多的是指数值型数据。随着计算机技术的发展，人们利用计算机处理的数据类型越来越丰富，涵盖了文本、声音、图形、图像、视频等非数值型数据。

近年来，人们利用各种信息技术工具，实现了自动采集数据。例如，在鸟类栖息地安装实时监控设备(图 1-4)，利用红外摄影机全天候拍摄视频数据，利用录音设备录制声音数据，这些数据可以直接保存到信息系统中，供人们分析、研究。



图 1-4 用实时监控设备记录鸟类活动

活动

1.1 近年来，随着信息技术的普及，国内外不少民间鸟类爱好者开始积极配合鸟类专业工作者，参与多项鸟类科学调查活动。一些观鸟网站和鸟类 App 都具有采集、整理、分享鸟类活动数据的功能，如图 1-5 和图 1-6 所示。查找并选择一个观鸟网站或鸟类 App，了解它向鸟类爱好者采集哪些鸟类活动数据。



图 1-5 某观鸟网站



图 1-6 某鸟类 App

2. 处理数据，获取信息

经过长年观测，这个研究站的工作人员采集了大量鸟类活动的数据。这些数据被多名科学家共享，他们对这些数据进行加工、分析，从而得出各种**信息**，为各自的科研服务。

现在人们越来越多地通过计算机来表示、组织和处理数据，从而可以获取并传播有价值的信息。

例如，科学家用计算机汇总 2010 年 10 月到 2012 年 10 月的数据后得知，在实验林地共观测到鸟类 44 种、4823 只次。其中，留鸟有 23 种，冬候鸟有 9 种，旅鸟有 8 种，夏候鸟有 4 种。进一步处理这些数据，能得出以下鸟类居留型种数的柱状图（图 1-7）及居留型比例的饼图（图 1-8）。

核心概念

信息（information）是数据中所包含的意义，是对数据进行加工的结果。把数据有组织、有规律地采集在一起就形成了信息。数据一方面承载着信息，另一方面也产生着信息。

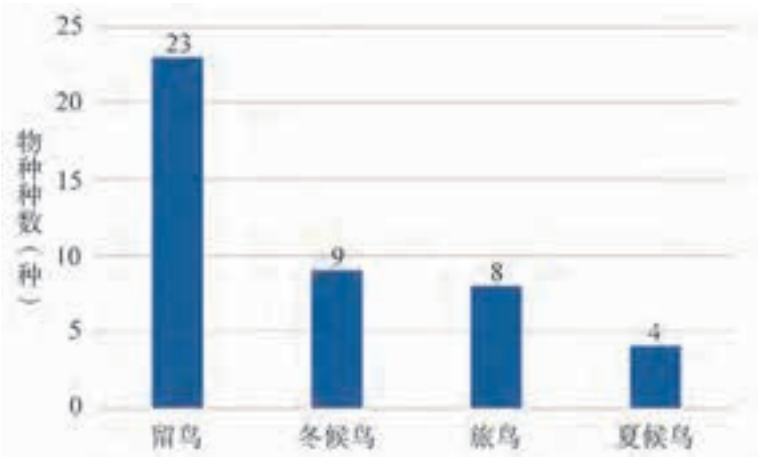


图 1-7 2010 年 10 月到 2012 年 10 月
林地内鸟类居留型种数柱状图

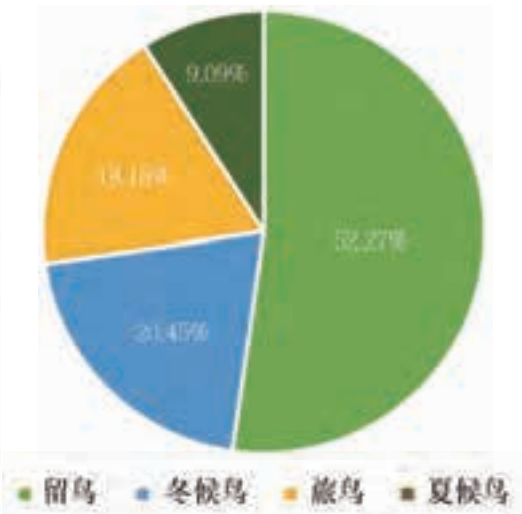


图 1-8 2010 年 10 月到 2012 年 10 月
林地内鸟类居留型比例饼图

这些数据包含着怎样的意义？从中能获取哪些信息？这些信息对人们有什么价值？

信息是对数据的解释，具有主观性。对同样的客观数据，不同的人会得到不同的信息。对于一般人而言，或许能从以上数据中得到“该林地鸟类众多”的信息，还可以得到“该林地的留鸟种数比候鸟多”的信息。对于鸟类爱好者来说，或许会得到“该林地是适宜的观鸟地点”这一信息。而科学家获得的信息或许是：该地区鸟类物种多样性高，且繁殖鸟（留鸟和夏候鸟）占总种数约六成，这表明“该林地是鸟类繁衍生息的重要场所”。

← 参见 P9 知识链接“数据
和信息”

数字化学习

上网查找并梳理国内外学者对信息的各种认识。

小贴士

知识 (knowledge) 是人们在改造世界的实践活动中所获得的可用于指导实践的认识、规律和经验,是归纳提炼出来的有价值的信息。

参见 P10 知识链接“数据、信息与知识的关系”

思考与讨论??

1. 该林地的野外数据采集工作是由多名工作人员共同完成的,但数据大家都可以使用。这反映了数据的什么特征?

2. 同一份鸟类研究数据,鸟类学家可以从中获取鸟类的生存状况与栖息地环境改变之间联系的信息,从而指导人们开展环境保护工作;卫生防疫部门可以从中获取候鸟迁徙路线的相关信息,从而指导人们开展禽流感防护工作。这反映了数据的什么特征?这说明数据和信息之间有怎样的关系?你能列举出类似的例子吗?

通过对大量数据、信息的归纳整理和反复验证,科学家完成了许多研究论文和学术著作,构建出不少关于鸟类的**知识**,如植物群落多样性与鸟类生存的关系等,为鸟类栖息地保护和物种多样性保护提供了理论依据(图 1-9)。



图 1-9 鸟类知识的构建

思考与讨论??

在班级里介绍自己知道的鸟类知识,说说自己是从哪里获取这些知识的。

活 动

1.2 (1) 利用活动 1.1 中的观鸟网站或鸟类 App 的查询统计功能(图 1-10 和图 1-11), 可以获取哪些数据? 利用这些数据, 鸟类爱好者和鸟类专业工作者可能会获取哪些信息? 这些数据和信息有着怎样的价值? 时效性如何?



图 1-10 某观鸟网站的鸟种统计页面

The image shows a table of bird species and their observation counts. The table has columns for '序号' (Serial Number), '鸟名' (Bird Name), and '记录数' (Record Count). The data is for the year 2016 in Tianjin. The table lists 10 species with their counts.

序号	鸟名	记录数
1	小鸦 Chalcophaps indica	19
2	喜鹊 Pica pica	17
3	灰喜鹊 Cyanopica cyaneus	16
4	灰喜鹊 Cyanopica cyaneus	15
5	灰喜鹊 Cyanopica cyaneus	14
6	灰喜鹊 Cyanopica cyaneus	13
7	灰喜鹊 Cyanopica cyaneus	12
8	灰喜鹊 Cyanopica cyaneus	11
9	灰喜鹊 Cyanopica cyaneus	10
10	灰喜鹊 Cyanopica cyaneus	9

图 1-11 2016 年天津地区观测到的鸟种及次数

(2) 利用思维导图, 整理这个网站或 App 提供的数据和服务, 并向大家介绍。

3. 利用大数据获取信息

当前, 随着信息技术的飞速发展、数据采集规模的快速增长和数据处理速度的突飞猛进, 大数据已深深影响了科学家开展科学研究和发现新知识的方式。

仍以鸟类研究为例, 如今, 摄像机、雷达乃至卫星等各种设备每天不停歇地自动获取规模大得不可想象的数据, 经过计算机的高速处理, 产生信息或知识。例如, 利用从多个

核心概念

大数据 (big data) 是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。

监测天气的雷达基站下载的海量图像数据,通过计算机的高速数据处理与分析,科学家获得了鸟类对于山川地理的认知地图,获得了它们感知地球磁场、确定飞行方位的内在机理,以及关于鸟类迁徙的更多知识。

图 1-12 所示的是 2010 年 9 月 10 日晚上某国东北部数百万只鸟的迁徙轨迹(圆圈的大小表示鸟的密度,颜色表示雷达的回波强度,箭头所指的是鸟儿的迁徙方向,箭头的长度表示鸟的飞行速度)。

参见 P10 知识链接“大
数据”

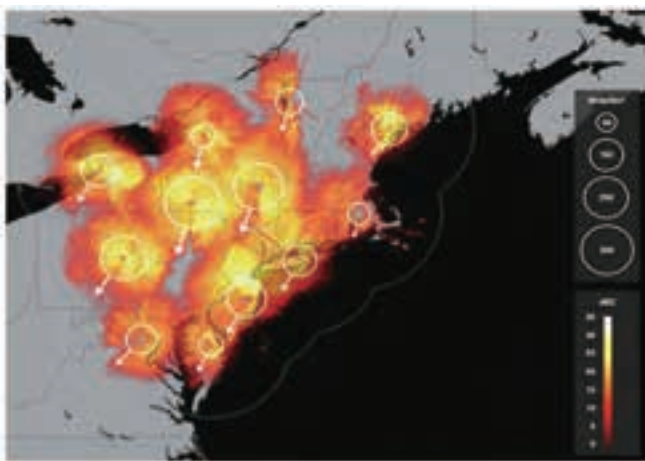


图 1-12 某夜某地上空鸟的迁徙轨迹图

思考与讨论??

除了科学发现,大数据对我们的日常生活也有着越来越深刻的影响。试着交流你所了解的大数据及其应用。

活 动

1.3 地球夜间灯光分布卫星图是根据卫星获得的数据制作的测绘地图,它展示了地球上入夜区域的城市灯光分布情况。

(1) 上网查找、观看中国夜间灯光分布卫星图。结合地理知识,分析我国各地夜间灯光分布情况,找出图中灯光最集中的城市,分析那里灯光强烈的原因。

(2) 有一项跟踪研究指出,夜间灯光与能源消耗、人口增长、GDP 增长紧密相关。你认同这一观点吗? 试着发掘图中信息,分析夜间灯光分布的规律及其原因。

(3) 结合活动体会,分析数据、信息与知识的关系,并利用恰当的工具绘制这三者的关系图。



知识链接

数据和信息

数据是对客观事物属性的描述，是记录下来的某种可以识别的符号。信息是数据中所包含的意义，是对数据进行加工的结果。

数据和信息之间有着固有的联系和区别，它们具有许多共同特征，同时又存在差异，具有一些不同的特征。

1. 数据和信息的不同特征

(1) 数据的载体性与信息的依附性

数据是信息的符号表示，是信息的载体；信息是数据的含义，是对数据的解释。两者密不可分。信息必须依附于某种载体，通过某种数据形式才能存储、表达和传播。相同的信息可以依附于不同的载体，其内容不会因载体形式的不同而发生变化。

例如，2017年7月，我国多地降暴雨，各地气象台发布了降雨量的数据，电视台播放了暴雨来袭的视频，广播台播放了语音报道，报纸做了文字和图片报道……这些文字、图像、声音、视频等都是数据，它们承载着“多地降暴雨”的信息。同时，“多地降暴雨”这一信息在存储和传播过程中依附于文字、图像、声音、视频等多种载体。虽然信息传递的渠道不同，所依附的载体形式不同，但信息的内容是相同的。

(2) 数据的孤立性与信息的联系性

数据是最原始的记录，与其他数据之间没有建立联系之前，是分散和孤立的。只有通过数据加工处理，与其他数据之间建立联系，才能形成针对某个特定问题的信息。例如，1、3、5、7、9、11、13、15是一组数值数据，孤立地看每一个数，无法知晓它代表什么。但如果将这组数据联系起来，可以发现这是一个公差为2的等差数列的开头一段，据此可以推断其后面的数应该是17、19、21……这样通过分析得出的结论便是信息。

(3) 数据的客观性与信息的主观性

数据是记录下来可以被识别的符号，是原始事实，具有客观性；信息是对数据的解释，是数据处理的结果，具有主观性。数据本身没有意义，只有经过加工和解释，才具有意义，从而转化为信息。例如，用粉笔在黑板上画一个圆圈，请被测试者回答这是什么，会得到许多答案，如“数字0”“英文字母O”“句号”“月亮”……这里，黑板上的圆圈是数据，是客观存在的一个符号，没有确定的含义，而“数字0”“英文字母O”等是人们解读这一数据得到的信息。同一数据，具有不同知识、经验的人从不同的角度解读，会得到不同的信息。

2. 数据和信息的共同特征

(1) 普遍性

数据是对客观事物属性的描述。事物是普遍存在的，因此，数据也无处不在，无时不有。考试的成绩、上课的铃声是数据，人们阅读的文章、观看的影片也是数据……有了数据，人们就会感知其中的意义，自觉或不自觉地获取信息。因此，信息也是普遍存在的。

（2）可处理性

对数据可以进行加工处理，生成新的数据。信息是数据加工的结果，同一数据经过不同的加工可以得到不同的信息。同时，对信息进行分析处理，可以得到更多的信息。例如，从某个人的身份证号码中提取第 7~12 位，得到数据 199006。根据身份证号码的编码规则，从中可以解读出信息——这个人的出生年月为 1990 年 6 月。作进一步加工处理，还可以从中解读出更多信息，如这个人的年龄、属相等。

（3）传递性与共享性

数据和信息是可以传递和共享的，同一数据或信息可以通过复制、传播，被多人重复使用。在传递和共享的过程中，数据和信息本身不会像物质和能源那样产生损耗。例如，同一新闻，可以通过报纸、电视、网络等多种渠道传播。在这一过程中，新闻通过各种渠道传递给多人，而新闻本身不会因传递和共享而有任何损失。

（4）价值相对性与时效性

数据和信息是有价值的，但其价值只有当数据或信息被利用时才能体现出来。数据和信息的价值具有相对性，是否有价值及价值的大小取决于使用者的需求，以及使用者对数据和信息的认知、理解和应用能力。例如，两家鞋厂分别派一名推销员到一个岛上推销鞋，他们上岛后共同感知的数据是“岛上居民一年四季都光着脚”。然而这两名推销员从中获取了不同的信息：第一名推销员认为“岛上无人穿鞋，没有市场”；第二名推销员认为“岛上无人穿鞋，市场潜力很大”。对数据和信息的不同理解，使他们做出了不同的选择，采取了不同的行动，从而获得了不同的结果。第一名推销员失望而归，第二名推销员请鞋厂速寄来 100 双鞋，把鞋送给岛上的居民，最终为鞋厂赢得了销售市场。数据和信息的价值也与个人需求有关。例如，“岛上居民一年四季都光着脚”这一数据及它承载的信息，对于想到岛上推销鞋的推销员是有价值的，但对于与此事不相关的人来说可能并没有什么价值。

数据和信息的价值还与时间有关，即具有时效性。例如，某商场今年 10 月 1 日至 7 日举办店庆活动，商品打折促销，如果消费者恰好想去该商场购物，并在 10 月 7 日之前获得了这个信息，那么，该信息对消费者是有价值的。但过了 10 月 7 日，该信息就无效了，其价值就降低了。

数据、信息与知识的关系

从数据到信息，再到知识，是一个从低级到高级的认知过程。数据是信息和知识的来源。无论信息还是知识，都来自数据，都是以数据为载体而存在的。信息是经过加工的数据，知识是经过人类归纳整理和反复验证后沉淀下来而呈现的规律。同时，相应的知识又是加工数据、提炼信息的基础，能帮助人们理解信息。由此可见，数据、信息与知识之间不存在绝对的界限，三者有着千丝万缕的联系。

大数据

信息技术与经济社会的交汇融合引发了数据量的迅猛增长，数据已成为国家基础性战略资源。大数据正日益对全球生产、流通、分配、消费活动，乃至经济运行机制、社会生活方式和国家治理能力产生重要影响。

大数据是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。大数据通常具有 4V 特征,也就是 Volume (数据量)、Velocity (处理速度)、Variety (多样性)、Value (价值性)。

(1) 数据量:大数据的体量很大,且数据集合的规模还在不断扩大。随着信息技术的大规模普及和应用,教育、商业、工业、科学研究、医疗等各行各业所产生的数据量都呈现出指数级增长的趋势。

(2) 处理速度:由于数据量增长速度快,大数据处理速度也必须快,且时效性要求高。大数据往往以数据流的形式动态地、快速地产生,需要在一定的时间限度内得到及时处理。

(3) 多样性:大数据来自多种数据源,数据种类和格式非常丰富。随着智能设备、社交网络等的流行,机器和传感器数据(如设备日志、地理位置数据)、社交数据(如网站用户行为记录数据等)等各种新类型数据越来越多。

(4) 价值性:大数据的价值性不仅指大数据能产生价值,更是指数据价值密度相对较低。海量的数据中可能发挥价值的仅是其中非常小的一部分。

拓展阅读

数据——信息社会的重要资源

随着人类跃进到大数据时代,数据不仅是新知识的来源,还是记录历史的最重要、最可靠、最好的方式。从今以后,人类所有的历史记录,无论是数字、文档、图片,还是音频和视频,都将以数据的形式存在,数据就是静态的历史,历史就是动态的数据。历史的碎片,就是游离的数据;历史的迷雾,就是模糊的数据;历史的盲点,就是缺失的数据。用数据构建的历史,因为精确的细节而永远鲜活,数据越丰富,后世的历史学家也就越能经由数据更好地再现当时的社会。

除了发现知识、记录历史,人类使用数据的巅峰形式,是通过数据训练机器,让机器获得智能,在不远的将来,无处不在的计算设备和网络将像有智商的人一样,为人类工作和服务。这意味着我们在向智能型社会迈进,在这个新的社会形态下,由于精准的计算和预测,整个社会的各个部分可以像无数个大大小小的轴承和齿轮一样,环环相扣,齿齿吻合。日常管理将通过数据得到优化,各种任务、合作可以无缝对接,社会运行的成本可大幅度降低。更重要的是,越来越多的工作将被计算机或者机器人代替。这既是进步,又是挑战。回望农业时代和工业时代,人类不断地开发我们赖以生存的自然环境,从地表到地下,物理性的资源终有耗尽的一天。而大数据将成为人类取之不尽、用之不竭的新资源,在这片资源之上,再通过软件和算法,人类将建设一个智能型世界。

数据,正在成为这个世界最重要的土壤和基础。

——摘自《数据之巅 大数据革命历史、现实与未来》

探究计算机中的数据表示

——认识数据编码

在鸟类研究过程中，科学家采集了各种各样、丰富多彩的数据。为了有效存储和处理这些数据，需要将它们数字化后存入计算机。

计算机是由逻辑电路组成的，逻辑电路通常只有高低两种电位状态，正好可以表示“0”与“1”，所以计算机采用二进制来存储和表示数据（生活中人们常用十进制数，二进制数和十进制数转换如图 1-13 所示）。因此，要想用计算机存储和处理现实中的数值、文本、图形、图像、声音和视频等数据，必须对数据进行二进制编码，即将其转化为由“0”和“1”组成的代码。数据的类型不同，编码的方法也不同。

二进制数	0	1	10	11	100	101	110	111	1000	1001
十进制数	0	1	2	3	4	5	6	7	8	9
二进制数	1010	1011	1100	1101	1110	1111	10000	10001	10010	10011
十进制数	10	11	12	13	14	15	16	17	18	19
二进制数	10100	10101	10110	10111	11000	11001	11010	11011	11100	11101
十进制数	20	21	22	23	24	25	26	27	28	29

图 1-13 二进制数和十进制数

项目学习目标

在本项目中，我们将通过探究一些鸟类活动数据的编码，了解数值数据和文本数据的编码方法，以及声音和图像的数字化过程。

完成本项目学习，须回答以下问题：

1. 编码的意义和作用是什么？
2. 数值数据编码的基本方法是怎样的？
3. 常用的文本数据编码方式是怎样的？
4. 声音数字化的基本方法是怎样的？
5. 图像数字化的基本方法是怎样的？

项目学习指引

1. 从树牌号认识编码

在项目一的林地鸟类活动调查中，科研人员以树为单位观察并记录每棵树上鸟的活动数据。为了清楚无误地区分和表示每一棵树，方便识别和交流，科研人员为林地中的每一棵树都设置了一个编号——树牌号，如图 1-14 所示。

树牌号
2130198
2130123
4080059

图 1-14 树牌号示例

给树编号的过程其实是一个**编码**的过程。为了给林地里的每一棵树设置一个唯一的树牌号，需要制定相应的编码规则。图 1-15 中的树牌号编码规则如下：每个树牌号由 7 位数字组成，第一位数字为一级区域编码（0~9，分别代表林地划分的一个一级区域），第二位和第三位数字为二级区域编码（01~20，分别代表该一级区域中的一个子区），第四位至第七位数字为树木编码（0001~9999，分别代表每个子区中的一棵树）。例如，树牌号 2130198，就代表 2 区 13 子区的第 198 棵树。



图 1-15 某树牌号的编码

生活中编码无处不在，如身份证号、银行卡号、邮政编码、学籍号、车牌号及条形码、二维码等，都是按照一定的规则产生的编码。

思考与讨论??

如果 2 区 11 子区中有 10023 棵树，以上的编码规则是否适用？

核心概念

编码（encoding）是指用预先规定的方法将文字、数字或其他对象转换成规定的符号组合，或将信息、数据转换为规定的脉冲电信号。在计算机中，编码一般是指用预先规定的方法将数字、文字、图像、声音、视频等对象编成二进制代码的过程。

← 参见 P18 知识链接“编码”

活 动

2.1 了解生活中的编码。

- (1) 了解身份证号的编码规则，分析一代身份证号与二代身份证号的区别，思考启用二代身份证号的原因。
- (2) 根据本校实际情况，设计适用的学籍号编码规则，保证每名同学拥有一个唯一的学籍号。
- (3) 在班级内分享自己的学籍号编码方案，说明如何保证无重码，以及在什么情况下需要修改编码规则、如何修改。

小贴士

十进制 (decimal system) 是生活中常用的数制，二进制 (binary system) 是计算技术中广泛采用的数制。

参见 P19 知识链接“数值数据的编码”

数字化学习

支秉彝，汉字编码和汉字信息处理和系统研究的开拓者。20 世纪 60 年代，支秉彝研究汉字信息字模，发明了“见字识码”编码方法。上网查找资料，了解并交流支秉彝等我国科学家为解决汉字进入计算机的难题而艰苦攻关、不懈奋斗的事迹。

2. 了解数值数据和文本数据的编码

要想用计算机存储和处理数据，必须先对它们进行编码，将它们转换成由“0”和“1”组成的二进制代码。对不同类型的数据，应采用不同的编码方法。

(1) 数值数据的编码

数值数据是一类常见数据，是可用于算术运算的具体数值。例如，鸟的数量是 21 只，这个数值数据在计算机中是如何表示的呢？

计算机中的数值数据是以补码的方式表示的，以十进制数 +21 和 -21 的 8 位编码为例，它们的二进制数、原码、反码和补码分别如下。

$(+21)_{10} = (+10101)_2$	$(-21)_{10} = (-10101)_2$
$[+10101]_{\text{原}} = 00010101$	$[-10101]_{\text{原}} = 10010101$
$[+10101]_{\text{反}} = 00010101$	$[-10101]_{\text{反}} = 11101010$
$[+10101]_{\text{补}} = 00010101$	$[-10101]_{\text{补}} = 11101011$

(2) 文本数据的编码

记录鸟类活动时需要记录鸟的名称，例如灰眶雀鹟的学名是 Alcippe Morrisonia。对这些由字母构成的数据，计算机是如何存储和表示的呢？

字母、数字、标点符号等，称为西文字符。计算机在存储和处理这些西文字符时，需要为每个字符规定一个由 0 和 1 组成的代码。目前，国际上普遍采用的西文字符编码标准是 ASCII 码 (American Standard Code for Information Interchange, 美国标准信息交换代码)。

标准 ASCII 码表如图 1-16 所示。

	000	001	010	011	100	101	110	111
0000	NUL	DLE	SP	0	@	P	~	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EOT	DC4	\$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

图 1-16 标准 ASCII 码表

思考与讨论??

常用汉字有近 5000 个，一个汉字的编码要用 2 个字节表示，而不是 1 个字节，这是为什么？

← 参见 P20 知识链接“文本数据的编码”

活 动

2.2 加密解密游戏。

(1) 以标准 ASCII 码表作为密码本，选出 4 名学生配合完成加密解密游戏。

模拟保密电文的发送和接收过程，角色分配及建议流程如下：

- 首长 1：拟电文（设计一段由字母、数字或符号组成的明文），传递给发报员。
- 发报员：对电文进行加密（将字母、数字或符号转换为 ASCII 码，成为密文），传递给接报员。
- 接报员：接收密文，进行解密（将 ASCII 码转换为字母、数字或符号），解出明文，并传递给首长 2。
- 首长 2：向首长 1 核对解密后的电文与原电文是否一致。若不一致，组织小组成员查找问题，并改正。

(2) 各小组自己设计编码方案和密码本，再玩一次加密解密游戏。

小贴士

标准 ASCII 码用 7 个二进制位表示 1 个字符，如，字母 A 的 ASCII 码是 1000001，符号 # 的 ASCII 码是 0100011。

由于标准 ASCII 码只能表示 128 个字符，无法满足西文字符编码的需要，后来又扩充了 128 个字符，称为扩展 ASCII 码。

3. 了解声音和图像的数字化

把自然界的鸟鸣声录制下来并转换为音频文件，经历了什么样的转换过程呢？自然界的鸟鸣声是一种连续的声波，为了用计算机存储和处理这些声音数据，需要将它们数字化，并记录成为音频文件。将模拟声音信号转换成数字声音信号，需要经历采样、量化和编码三个步骤，如图 1-17 所示。

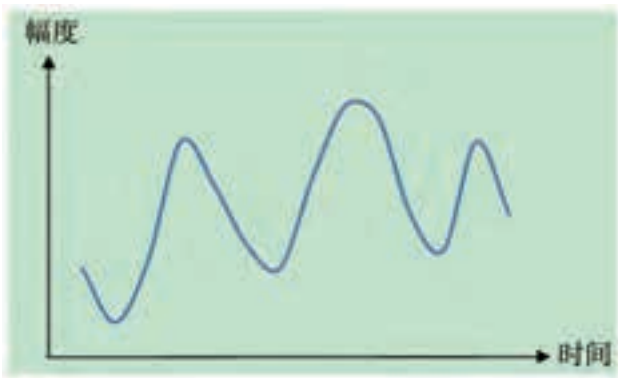


图 1-17 声音数字化的过程

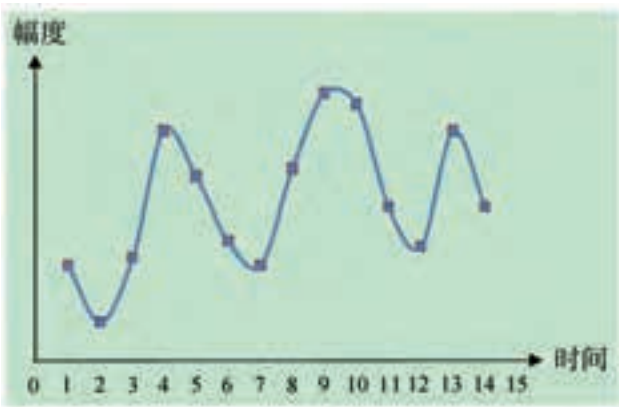
参见 P21 知识链接“声音数字化”

(1) 采样

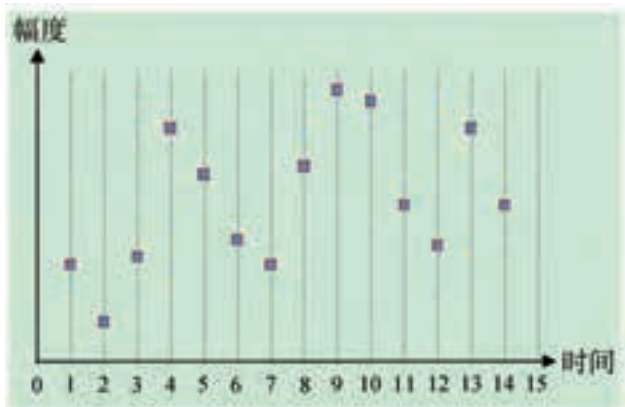
采样 (sampling) 即每隔一段时间在模拟声音信号的波形上采集一个幅度值。图 1-18 (a) 是一段鸟鸣声的模拟声音信号，对其采样时，在波形信号上按时间维度等距离地选取若干个离散的点，如图 1-18 (b) 所示。这些采样得到的幅度值被记录下来，如图 1-18 (c) 所示。



(a)



(b)



(c)

图 1-18 声音数字化的采样过程

(2) 量化

采样之后，要用二进制数将采样得到的幅度值表示出来，这就是量化(quantization)。例如，取**量化位数**为4，量化过程如下：

首先，确定量化位数为4。

然后，将声音信号的幅度值范围划分为 2^4 (16) 个量化级数。

第三，确定采样点的量化值。若采样得到的幅度值不在这些级数之内，则按照一定的规则将它近似到某个级数值上。如图 1-19 (a) 中，第 3 个采样点的真实幅度值约为 5.4，将其四舍五入近似到级数值 5；第 5 个采样点的真实幅度值约为 9.8，将其四舍五入近似到级数值 10。同理可将第 6、第 10 个采样点的真实幅度值近似到相应的级数值。量化结果如图 1-19 (b) 所示。

小贴士

量化位数：存储、记录声音幅度值所使用的二进制位数。

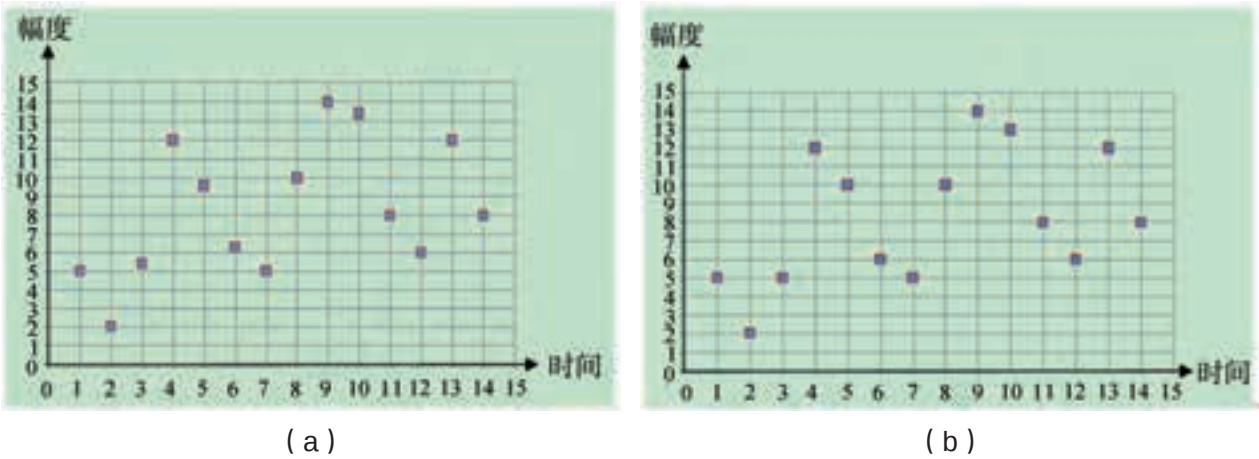


图 1-19 声音数字化的量化过程

最后，用二进制数表示这些采样点的量化值。例如，用 4 位二进制数来表示，第 1 个采样点的量化值为 0101，第 2 个采样点的量化值为 0010，第 3 个采样点的量化值为 0101……

(3) 编码

经过采样和量化，模拟声音信号转化为一组二进制数序列，再通过编码将其按照一定的规则记录下来。采用不同的编码方法，会形成不同格式的音频文件，如 WAV 格式、MP3 格式等。

通过手机、数码相机、数码摄像机等数字设备，可以拍摄鸟类的照片，得到图像文件。图像数字化的过程和声音数字化类似，都会经历采样、量化和编码三个步骤。

← 参见 P22 知识链接“图像数字化”

活 动

2.3 探究声音数字化参数对音频文件的影响。

(1) 探究采样频率对音频文件大小与音质的影响。

① 选择一种音频编辑软件,新建一个 WAV 文件,设置采样频率为 44.1kHz,量化位数为 32 位,声道数为 2 (立体声),录制一段声音,并将其保存为“录音 1.WAV”。

② 保持其他参数不变,修改采样频率为 11.025kHz,并将文件另存为“录音 2.WAV”,比较两个文件的大小及音质,分析原因。

(2) 模仿上述做法,分别探究量化位数和声道数对音频文件大小及音质的影响。

2.4 探究图像数字化。

(1) 开展数字化学习,了解图像数字化的知识,对比声音数字化与图像数字化的过程。

(2) 利用手机、数码相机等工具采集鸟类活动图片,选择一种图像处理工具,将鸟类活动的图像文件统一处理为相同分辨率和颜色深度的 BMP 文件。

(3) 参考活动 2.3 设计实验方案,探究图像分辨率和颜色深度对图像呈现效果和文件大小的影响,并设计表格记录实验数据及结论。

(4) 选择一种图像处理工具,探究不同文件格式对图像文件大小及质量的影响,并设计表格记录实验数据及结论。



知识链接

编码

编码是指用预先规定的方法将数字、文字或其他对象转换成规定的符号组合,或将信息、数据转换为规定的脉冲电信号。

编码一般具备以下功能和意义。

- 鉴别: 编码是对象的唯一标识。通过辨识编码可以找到其唯一对应的对象。例如,邮政编码对应的地区是唯一的,身份证号码对应的人是唯一的,包裹单上的条形码对应的包裹也是唯一的。

- 排序: 编码的符号都具有一定的顺序,比较容易进行排序。

- 专用含义: 编码一般都会包含一定的含义,例如,本项目的树牌号中包含着所在区和子区的信息,身份证号码中包含着出生日期的信息。

在计算机中,编码一般是指用预先规定的方法将数字、文字、图像、声音、视频等对象编成二进制代码的过程。

数值数据的编码

数值数据，又称为数字数据，是可用于算术运算的具体的数值。

1. 数制

数值数据通常采用数制来表达，如，1 打等于 12 个，用的是十二进制；1 小时等于 60 分钟，用的是六十进制；1 米等于 10 分米，用的是十进制。

生活中常用的是十进制数，它的基数为 10，由 10 个基本数码(0、1、2、3、4、5、6、7、8、9)组成，逢 10 进 1。例如，十进制数 328.56 中，3、2、8、5、6 所代表的数值大小分别如图 1-20 所示。其中， 10^0 、 10^1 等称为位权，以小数点为界，向左(整数部分)各位的位权依次为 10^0 、 10^1 、 10^2 ……向右(小数部分)各位的位权依次为 10^{-1} 、 10^{-2} ……

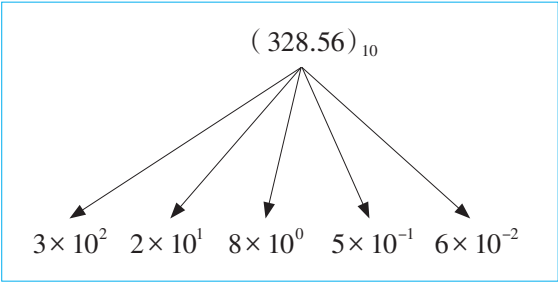


图 1-20 十进制数各位的位权

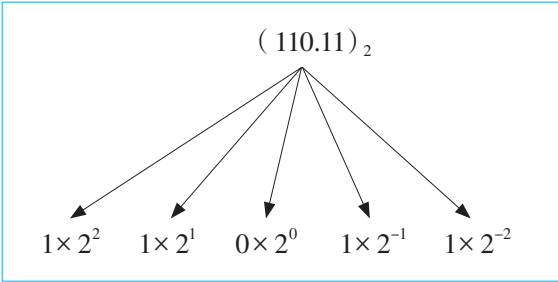


图 1-21 二进制数各位的位权

二进制是计算技术中广泛采用的一种数制，它的基数为 2。同样，以小数点为界，向左(整数部分)各位的位权依次为 2^0 、 2^1 、 2^2 ……向右(小数部分)各位的位权依次为 2^{-1} 、 2^{-2} ……例如，二进制数 110.11 中，各位数字所代表的数值大小分别如图 1-21 所示。

计算技术中常用的数制还有八进制和十六进制，见表 1-2。

表 1-2 常用数制

数制	基数	可用符号	位权	进位规则
十进制数	10	0、1、2、3、4、5、6、7、8、9	10^{n-1} 、 10^{n-2} …… 10^0 、 10^{-1} 、 10^{-2} ……	逢 10 进 1
二进制数	2	0、1	2^{n-1} 、 2^{n-2} …… 2^0 、 2^{-1} 、 2^{-2} ……	逢 2 进 1
八进制数	8	0、1、2、3、4、5、6、7	8^{n-1} 、 8^{n-2} …… 8^0 、 8^{-1} 、 8^{-2} ……	逢 8 进 1
十六进制数	16	0、1、2、3、4、5、6、7、8、9、A、B、C、D、E、F	16^{n-1} 、 16^{n-2} …… 16^0 、 16^{-1} 、 16^{-2} ……	逢 16 进 1
R 进制数	R	0、1……R-1	R^{n-1} 、 R^{n-2} …… R^0 、 R^{-1} 、 R^{-2} ……	逢 R 进 1

数值数据可用于算术运算，每种数制都有其运算规则。二进制数的算术运算规则如下。

加运算：0+0=0，0+1=1，1+0=1，1+1=10（逢 2 进 1）

减运算：1-1=0，1-0=1，0-0=0，10-1=1（向高位借 1 当 2）

乘运算： $0 \times 0=0, 0 \times 1=0, 1 \times 0=0, 1 \times 1=1$

除运算： $0 \div 1=0, 1 \div 1=1$

2. 数值数据的编码

数值数据的编码过程如图 1-22 所示。



图 1-22 数值数据的编码过程

(1) 转换

要用计算机存储和处理数值数据，首先要将其转换为二进制数。十进制数转换为二进制数，整数部分的转换方法是除 2 反向取余，小数部分的转换方法是乘 2 正向取整。如图 1-23 和图 1-24 所示，将十进制数 37.375 转换为二进制数，首先将其整数部分和小数部分分别转换为二进制数，然后再合并，得到 $(37.375)_{10} = (100101.011)_2$ 。



图 1-23 整数部分的转换

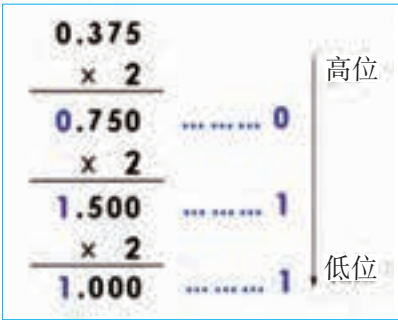


图 1-24 小数部分的转换

(2) 编码

计算机中数值数据的编码分为原码、反码和补码，其编码规则见表 1-3。通常情况下，计算机用一个数的最高位存放符号，即用 0、1 表示正负符号，正数为 0，负数为 1，这个二进制位称为符号位。

表 1-3 原码、反码和补码的编码规则

	原码	反码	补码
正数	符号位(0)+ 数字部分 (如果原数不足 n-1 位, 在高位补 0)	同原码	同原码
负数	符号位(1)+ 数字部分 (如果原数不足 n-1 位, 在高位补 0)	在原码的基础上, 符号位 不变, 其余各位取反	在反码的基础上 +1

注：n 为编码总位数

文本数据的编码

1. 西文字符的编码

应用最广泛的西文字符编码方案是 ASCII 码。ASCII 码是由美国国家标准学会 (American National Standard Institute, ANSI) 制定的通用单字节字符编码方案。

ASCII 码分为标准 ASCII 码和扩展 ASCII 码。标准 ASCII 码也叫基础 ASCII 码,使用 7 个二进制位来表示西文字符,包括所有的大写和小写字母、数字 0 到 9、标点符号,以及在美式英语中使用的特殊控制字符。扩展 ASCII 码用 8 个二进制位来表示字符,第 8 位用于确定附加的 128 个特殊符号字符、外来语字母和图形符号。

2. 汉字的编码

常用汉字有近 5000 个,这种信息容量要用 2 个字节长即 16 位二进制编码才能满足。1980 年,中国国家标准总局发布了中华人民共和国国家标准 GB2312—1980《信息交换用汉字编码字符集——基本集》,又称为国标码。国标码用 2 个字节表示一个汉字,其中每个字节的最高位为 0。例如,“大”字的国标码为 0011010001110011。

国标码在计算机内部存储和处理时会与 ASCII 码发生冲突,例如“4s”的 ASCII 码在计算机中的表示也是 0011010001110011。为了解决这个问题,汉字编码在计算机内的表示在国标码基础上稍做改动,将每个字节的最高位设为 1,这被称为机内码(简称内码)。例如,“大”字的机内码为 1011010011110011。机内码是用最高位均为 1 的 2 个字节表示一个汉字,是计算机内部存储、处理汉字所使用的统一编码。

3. Unicode

全世界有上百种语言,人们希望有一种编码,能将世界上所有的符号都纳入其中,每一个符号都给予一个独一无二的编码。Unicode 应运而生。

Unicode 是国际组织制定的可以容纳世界上所有文字和符号的字符编码方案。1990 年开始研发,1994 年正式公布。它为每种语言中的每个字符设定了统一且唯一的二进制编码,以满足跨语言、跨平台进行文本转换和处理的要求。目前的 Unicode 字符分为 17 组编排,每组称为 Plane(平面),每个 Plane 拥有 65536 个码位,共 1114112 个码位。

Unicode 一般用 2 个字节表示一个字符(非常偏僻的字符用 4 个字节)。但是,一篇英文文章,用 Unicode 编码比用 ASCII 编码需要多一倍的存储空间。于是,又出现了把 Unicode 编码转化为“可变长编码”的 UTF-8 编码。UTF-8 编码把一个 Unicode 字符根据不同的数字大小编码成 1~6 个字节,常用的英文字母被编码成 1 个字节,汉字通常是 3 个字节,只有很生僻的字符才会被编码成 4~6 个字节。

Unicode 的实现方式还有 UTF-16 和 UTF-32 等。

声音数字化

现实世界的声音是一种连续的波,称为声波。声音有两个参数:幅度和频率。要用计算机处理声音数据,必须把连续变化的波形信号转换成为离散的数字信号,以 0 和 1 编码的形式表示出来,这一过程称为声音数字化。声音数字化的过程包含采样、量化、编码三个步骤。

1. 采样

声音的采样是指每隔一段时间在模拟声音信号的波形上取一个幅度值。相隔时间相等的采样为均匀采样(又称为线性采样),相隔时间不相等的采样为不均匀采样(又称为非线性采样)。

计算机每秒钟在模拟声音信号的波形上采样的次数称为采样频率。常见的采样频率有 44.1kHz、22.05kHz、11.025kHz 等。采样频率越高,即采样的时间间隔越短,则在单位时间内得到的声音样本数据越多,对声音信号波形的表示越精确,声音的保真度越高。

2. 量化

声音的量化是用二进制数表示采样所得到的幅度值的过程。首先将幅度值范围划分为 2^n 个级数,每个级数对应一个幅度值,然后将采样得到的各个幅度值按一定的规则近似到某个级数值,并用二进制数表示,从而形成一组二进制数序列。这里的 n 称为量化位数。量化位数越大,划分的级数越多,采样结果近似到某个级数值时产生的误差就越小。因此,量化位数越多,数字化精度越高,声音就越保真。

3. 编码

声音的编码就是按照一定格式把经过采样和量化得到的离散数据记录下来,并在其基础上加入用于纠错、同步和控制的数据,最终转换成数字音频信号。不同的编码方法形成了不同格式的音频文件,如 WAV 格式、MP3 格式等。

采样频率、量化位数和声道数是数字化音频的技术指标,被称为声音数字化的三要素。它们直接影响数字化后音频的质量及其数据量的大小。一般情况下,未经压缩的音频文件的数据量可以按如下方法计算:

$$\begin{aligned}\text{数据量(单位:字节)} &= \text{数据率} \times \text{持续时间} \\ &= (\text{采样频率} \times \text{量化位数} \times \text{声道数}) \div 8 \times \text{持续时间}\end{aligned}$$

例如,一张 CD-ROM 中存放了 1 小时的数字音乐(未经压缩),则其数据量可按以下方法计算:

$$\begin{aligned}\text{数据量} &= (44100 \times 16 \times 2) \div 8 \times 60 \times 60\text{B} \\ &= 635040000\text{B} \\ &= 620156.25\text{KB} \\ &\approx 606\text{MB}\end{aligned}$$

标准 CD 格式的采样频率为 44.1kHz,量化位数为 16 位,声道数为 2(双声道)。数据量计算公式中的“ $\div 8$ ”是将位数转换成字节,一个字节由 8 个二进制位组成。1MB=1024KB,1KB=1024B。

图像数字化

从小小的商标到大型宣传海报,从各式各样的照片到风格迥异的图画,从茶杯上的图案到教材中的插图……凡此种种,从信息技术的角度看,都属于模拟图像,运用扫描技术或数字摄像技术可以将空间上连续的模拟图像转换成用 0、1 表示的数字图像,这一过程称为图像数字化。图像数字化的过程包含采样、量化、编码三个步骤。

1. 采样

图像的采样是按一定的空间间隔自左到右、自上而下提取画面信息,将一幅连续的模拟图像在空间上转换成若干个离散的像素点,每个像素点呈现不同的颜色(彩色图像)或亮度(灰度图像)。

一幅图像所包含的横向和纵向的像素点的数目称为图像分辨率。例如，一幅图像的分辨率为 640×480 ，表示该图像由横向 640 个像素点、纵向 480 个像素点，共 $640 \times 480 = 307200$ 个像素点组成。如果不考虑其他因素的影响，图像分辨率越高，采样的精度就越高，数字化后的图像就越清晰，同时图像所占的存储空间也越大。

2. 量化

图像的量化是将采样得到的每个像素点的颜色或亮度用若干位二进制数表示出来，其方法与声音数据量化的方法类似。首先确定颜色或亮度的取值范围，然后将近似的颜色划分成同一种颜色，每种颜色用一个二进制数来表示。例如，一幅黑白图像只有两种颜色，则每个像素点只需 1 个二进制位即可表示：1 表示黑色，0 表示白色。一幅 256 级灰度图像，每个像素点需要用 8 个二进制位来表示，表示 $2^8 = 256$ 个亮度层次；一幅 24 位真彩色的 RGB 图像，其三原色红、绿、蓝的每种光的强度被分成 256 个级别（0~255），需要用 8 个二进制位来表示，每个像素点有三种颜色，所以共需要用 24 个二进制位来表示，表示 $2^{24} = 16777216$ 种颜色。

记录每个像素点的颜色或亮度所需的二进制位数，称为颜色深度（也称色彩位数）。对于彩色图像来说，颜色深度决定了该图像可以使用的最多颜色数目，颜色深度越大，显示的图像色彩越丰富，画面越自然、逼真。但要注意，在图像分辨率相同的情况下，颜色深度越大，图像所占的存储空间也越大。

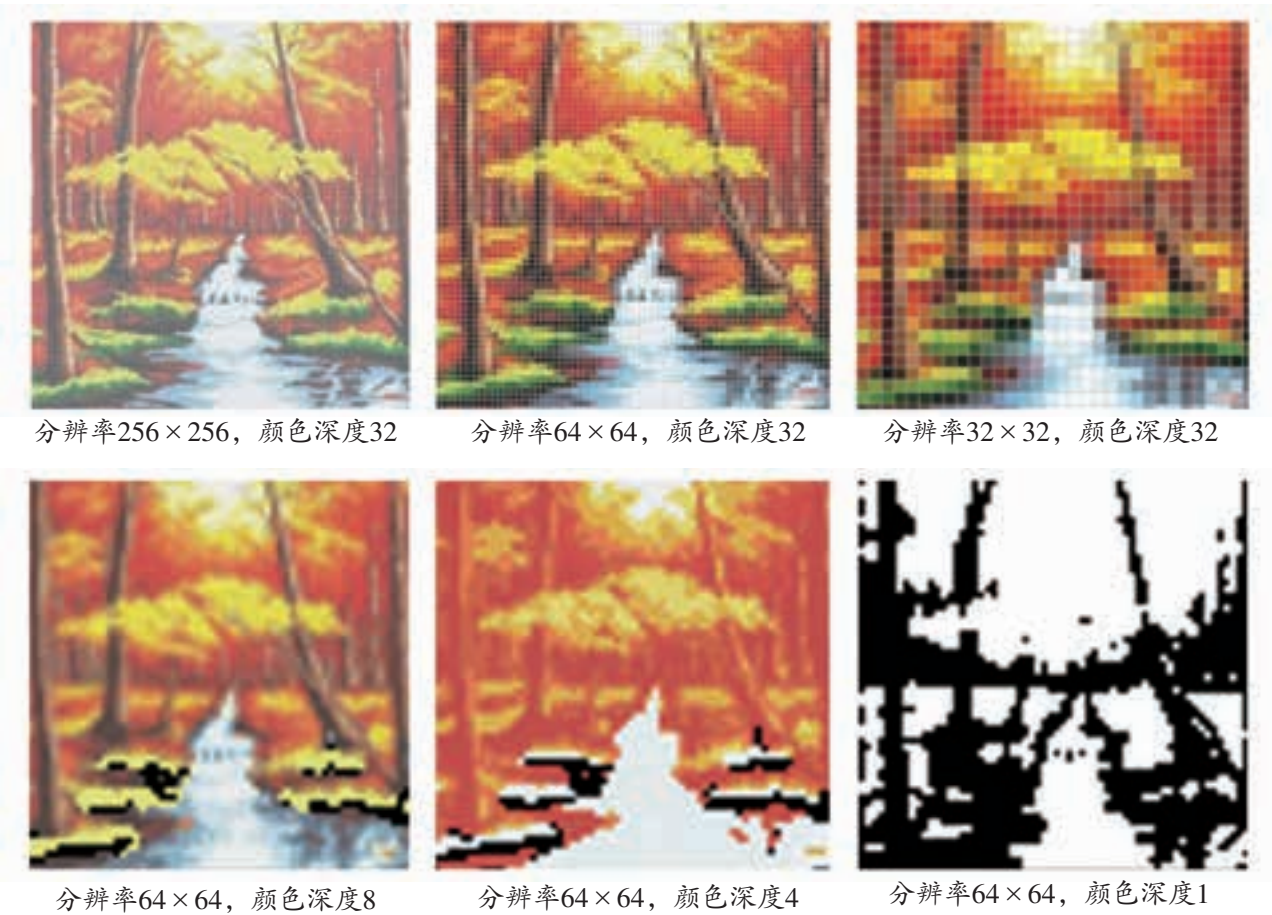


图 1-25 不同分辨率和颜色深度的位图比较

这种由纵横排列的像素点组成的图像称为位图(又称点阵图)。位图的质量主要由图像分辨率和颜色深度决定。图 1-25 呈现的是采用不同分辨率和颜色深度数字化后的位图。

未经压缩的位图图像的数据量(单位:字节)=图像分辨率×颜色深度÷8。

3. 编码

图像的编码就是按照一定的格式将位图上各个像素点的量化数据记录下来的过程。由于位图的数据量大,并且含有大量的重复数据,编码时一般采用数据压缩技术进行压缩和还原处理。不同的编码方法形成了不同格式的图像文件,如 BMP 格式、JPEG 格式等。

拓展阅读

计算机采用二进制的原因

1. 技术实现简单

人类早期设计的机械计算装置中主要用的是十进制。十进制数有 10 个基本符号,要用 10 种状态才能表示。使用电子器件的状态来表示 10 个基本符号过于复杂,而用电子器件的高电位与低电位或逻辑器件的开与关两种状态来表示两个基本符号就比较容易,所以二进制就成为电子计算机的数制基础。也就是说,电子器件的两种状态决定了电子计算机采用二进制来表示数据。

随着技术的发展,计算机的电子器件由电子管逐步变为晶体管、集成电路、大规模集成电路乃至超大规模集成电路,但电子器件的工作特点并没有改变。计算机是由逻辑电路组成的,逻辑电路通常只有两个状态,这两种状态可以表示“1”和“0”。这样的电路设计简单,而且只具有两种状态的电子器件容易找到,如继电器开关、灯泡、二极管等。所以,至今现代计算机仍然采用二进制来存储和表示数据。

2. 运算简单

二进制数的算术运算规则简单,有利于简化计算机内部结构,提高运算速度。同时,二进制只有两个数码,正好与逻辑代数中的“真”和“假”相吻合。因此,采用二进制可以简单方便地进行算术运算和逻辑运算。

另外,二进制数具有抗干扰能力强、可靠性高等优点,因为每位数字不是“0”就是“1”,当受到一定程度的干扰时,仍能可靠地分辨出它是“0”还是“1”。

单元挑战 认识并制作二维码

一、项目任务

二维码是近几年来移动设备上流行的一种编码方式。作为一种全新的数据存储、传递和识别技术，二维码的应用已渗透到人们生活的各个方面，如地铁广告、报纸、火车票、快餐店、电影院、团购网站及各类商品外包装上都能见到二维码，见图 1-26 和图 1-27。二维码被誉为“线上线下的一个关键入口”。



图 1-26 火车票上的二维码



图 1-27 手机扫描二维码

以小组为单位，了解二维码的组成结构、编码原理、功能、分类、特点及应用，学习二维码的制作方法，设计并制作小组的二维码。

二、项目指引

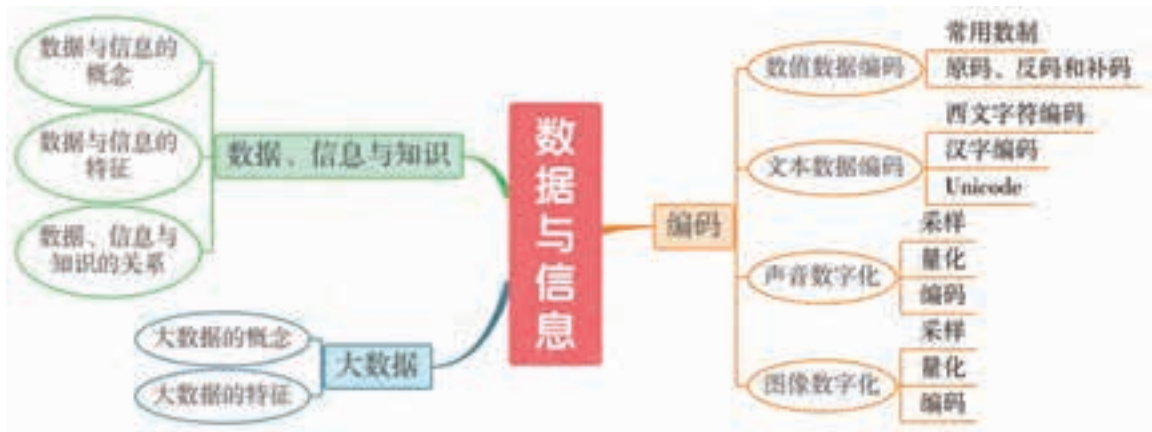
1. 以小组为单位开展数字化学习，收集资料，了解二维码的相关知识，并加以梳理和归纳，用思维导图呈现小组的学习成果。
2. 收集资料，了解二维码的制作工具和方法、二维码的类型及可以承载的内容；小组成员共同规划设计小组的二维码，确定本小组的设计目标和通过二维码承载的信息（如小组名、小组图标、电子邮箱地址等）；选择合适的工具和方法制作小组的二维码。

三、交流评价与反思

1. 各小组展示自己制作的二维码及介绍二维码的思维导图，分享学习体会。
2. 互相扫描各小组的二维码，对各小组制作的二维码、思维导图进行评价。
3. 小组成员共同回顾项目活动的过程，反思并交流本小组的收获与不足。

单元小结

一、主要内容梳理



二、单元练习

1. 某校正筹备运动会，需要为运动员编号。该校有高一、高二、高三共三个年级，每个年级 20 个班，每个班 45 ~ 55 名学生。请为该校设计运动员编码规则，保证每位运动员拥有一个唯一的编号，并能体现所在班级和性别。
2. 图书馆有一批纸质照片要转换成 JPG 格式的电子照片，刘楠小组承担了该任务。但刘楠发现大家提交的图像文件的大小和质量差距很大。排除摄影技术和环境因素，哪些参数会影响图像文件的大小和质量？为什么？请结合图像数字化的过程加以分析。

三、单元评价

评价内容	达成情况
能够通过实例分析，描述数据与信息的关系与特征（A、T、R）	
能够通过日常生活情景的分析，阐述数据对人们日常生活的影响（A、I、R）	
在运用数字化工具的学习活动中，理解数据、信息与知识的相互关系（A、T、I、R）	
知道大数据的概念、特征和常见应用领域（A、T、R）	
理解编码的意义和作用，知道数据编码的基本方式（T）	

说明：A—信息意识，T—计算思维，I—数字化学习与创新，R—信息社会责任

第二单元

数据处理与应用

信息社会里，人们的生产、生活越来越依赖于数据的处理与应用。企业管理者面向消费者开展市场调研，利用调研得到的数据和信息提高决策的科学性，降低企业经营风险；足球赛场上，教练组根据数据分析师的实时数据分析制定最佳应对策略，及时调整攻防模式；驾车出行时，人们会查询交通线路，获取当前道路的实时车流量数据，确定恰当的出行方案……现代社会中无处不在的数据处理与应用帮助人们了解现状，预测未来，制订措施和方案。而能否准确、灵活地使用信息技术，熟练地进行数据的采集、分析和可视化等，很大程度上会影响人们工作、学习的效率和质量。

在本单元中，我们将一起探究身边的数据处理工作，了解数据的采集、分析和可视化的基本方法，经历利用软件工具或平台处理数据的过程。



学习目标

- ◆ 了解数据处理及其作用，能够根据任务需求，选用恰当的软件工具或平台处理数据，完成分析报告。
- ◆ 了解采集数据的基本方法，掌握通过公式、函数进行数据计算的方法。
- ◆ 掌握数据分析的基本方法，学习数据可视化的表示方法，能够读懂图表传递的信息。
- ◆ 理解对数据进行保护的意义。

单元挑战

采集与分析气象数据

项目三

调查中学生移动学习现状

——经历数据处理的一般过程

在信息技术飞速发展的时代，信息、知识的数量以指数增长的迅猛趋势，给传统的学习带来了机遇与挑战，让人们强烈地意识到终身学习的必要性与紧迫性。随着科技的不断发展与完善，移动互联网技术渐趋成熟，各类移动设备层出不穷，让人们有了紧跟信息时代的“神器”。借助无线网络，利用移动设备随时随地获取学习资源，已成为一种新型的学习模式——移动学习（图 2-1）。移动学习



图 2-1 移动学习

代表着许多新的学习理念，如自主学习、个性化学习、终身学习等。专家预言，移动学习必将成为未来学习的一个重要方式。

当代中学生移动学习的现状如何？中学生对移动学习抱着怎样的认识和态度？中学生需要怎样的移动学习资源和移动学习环境？中学生移动学习的效率如何？中学生的移动学习存在哪些困难、哪些误区……这些是很多教育工作者、移动学习资源开发者和平台提供者迫切想要了解的问题。

项目学习目标

在本项目中，我们将围绕中学生移动学习现状，通过在线问卷调查采集数据，并借助软件工具完成数据分析和可视化，撰写调查报告，经历数据处理的全过程。

完成本项目学习，须回答以下问题：

1. 什么是数据处理？数据处理的过程是怎样的？
2. 如何认识数据处理的应用价值？
3. 什么是数据采集？什么是数据分析？数据分析的方法有哪些？
4. 什么是数据的可视化？它有哪些呈现方式？

项目学习指引

1. 明确数据需求

中学生移动学习现状调查是一项社会调查研究。社会调查研究是利用特定的方法和技术，从社会现实中采集研究需要的数据，通过数据的加工、分析和可视化，对社会现实进行描述或解释的认知活动。从技术角度看，社会调查研究就是围绕研究目的、需要研究问题而进行的一项数据处理工作（图 2-2）。

核心概念

数据处理（data processing）是对数据进行采集、存储、加工、分析和表达的过程。



图 2-2 数据处理工作是调查研究的核心

调查正式开始前，需要根据调查的目的和研究的问题来明确数据需求。不同的调查目的，需要不同的数据来支持。如果研究者想通过“中学生移动学习现状调查”了解中学生使用移动学习的基本情况，则只需要采集使用时间、使用频率、使用资源类型等方面的数据；如果想了解哪些因素影响中学生的移动学习，则需要采集中学生对移动学习的态度、认识、周围环境因素（如家长和教师是否支持、移动设备的拥有情况）等方面的数据；如果还想通过调研，为移动学习资源、产品的开发者提供决策依据，则需要采集中学生对现有移动学习资源、产品的需求数据，以及中学生期待的移动学习资源类型和设计特点等方面的数据。

← 参见 P38 知识链接“数据处理及其作用”

思考与讨论??

日常生活中，你知道哪些利用数据帮助判断、决策、解决问题的例子？

可以利用思维导图软件，进行头脑风暴，帮助自己理清研究思路，如图 2-3 所示。



图 2-3 利用思维导图软件工具梳理数据需求

思考与讨论??

面对中学生开展移动学习这一社会现象，不同的组织或个人，如教育部门、移动学习资源开发企业、家长，会面临或思考哪些问题？需要收集哪些数据帮助自己形成判断和决策？

活动

3.1 班级同学分成若干小组，分别扮演不同的角色，如移动学习资源开发企业主、学校教师、家长。

（1）针对“中学生移动学习现状”，各小组分别确定各自的调查目标、想要研究的问题以及希望通过调查收集的数据。

（2）各小组绘制思维导图，梳理本小组的调查目的及需要采集的数据。

参见 P38 知识链接“数据处理的一般过程”

2. 采集数据

（1）选择合适的数据采集方法和工具
传统的社会调查方法主要有发放纸质问卷开展调查、面

对面访谈调查、实地考察等。信息技术的飞速发展大大丰富了人们采集数据的方法和手段。随着互联网、移动用户的增多,通过网络开展在线调查已成为很多研究者的选择。人们推出了很多在线调查平台、网站及在线调查系统(图 2-4),提供从在线问卷设计、数据采集(data acquisition)到数据加工分析的“一站式服务”。研究者要根据自己的需求,选择最适合自己的数据采集方法和工具。



图 2-4 各种在线调查平台

思考与讨论??

- 1. 你了解并尝试使用了哪些在线调查平台? 你认为这些平台各自有什么特色?
- 2. 如何判断在线调查平台的设计是否专业、是否可信赖?

(2) 在线编辑发放问卷, 采集数据

问卷是为了达到调研目的和采集必要数据而设计的一系列问题。问卷设计的好坏,直接关系到数据采集工作质量的高低。

小贴士

“一站式服务”实质就是利用信息技术对服务进行集成、整合,因此也被认为是“提供整体解决方案”。它在提高服务效率的同时,也提高了服务提供者的竞争力。

小贴士

移动学习软件可以自动采集用户基础数据和行为数据,如图 2-5 所示。



图 2-5 某移动学习软件用户随时可以查看自己的学习数据

活 动

- 3.2 寻找适合自己研究的数据采集方法和工具。
- (1) 小组成员各自上网查找、了解可以利用的在线调查平台，并尝试使用。
- (2) 小组交流讨论，对各种在线调查平台进行评估：
- 是有偿使用还是免费使用？
 - 提供的服务有哪些？能否满足需要？
 - 问卷编辑是否方便？能否方便地导入？
 - 问卷发放有哪些渠道？数据导出格式有哪些？是否支持自选分析工具？
 - 使用过该平台的人如何评价？
- (3) 小组成员协商、确定本小组选用的在线调查平台。

数字化学习

请利用网络开展学习，了解问卷的组成，明确问卷的设计原则及注意事项。分析几份调查问卷实例，掌握问卷的各种题型及其使用场合。

当前许多在线调查平台都提供多种问卷创建方式，一般有三种模式，一种是设计好纸质问卷并直接导入，另一种是在平台上根据提示自由创建问卷(图 2-6)，还有一种是直接使用平台提供的模板通过简单编辑生成问卷(图 2-7)。

小贴士

在很多在线调查平台上编辑问卷时，可在屏幕上直观地得到即将打印到纸张上的效果，故也称可视化操作。这是一种“所见即所得”技术。



图 2-6 某在线调查平台的问卷编辑页面

思考与讨论??

在线调查平台在编辑问卷时，要求先明确每个问题的类型，如单选题、多选题、排序题等。这是为什么？



图 2-7 某在线调查平台提供的模板

在线问卷可以利用平台中的相关设置来控制数据采集的目标人群和问卷发放数量等(图 2-8)。

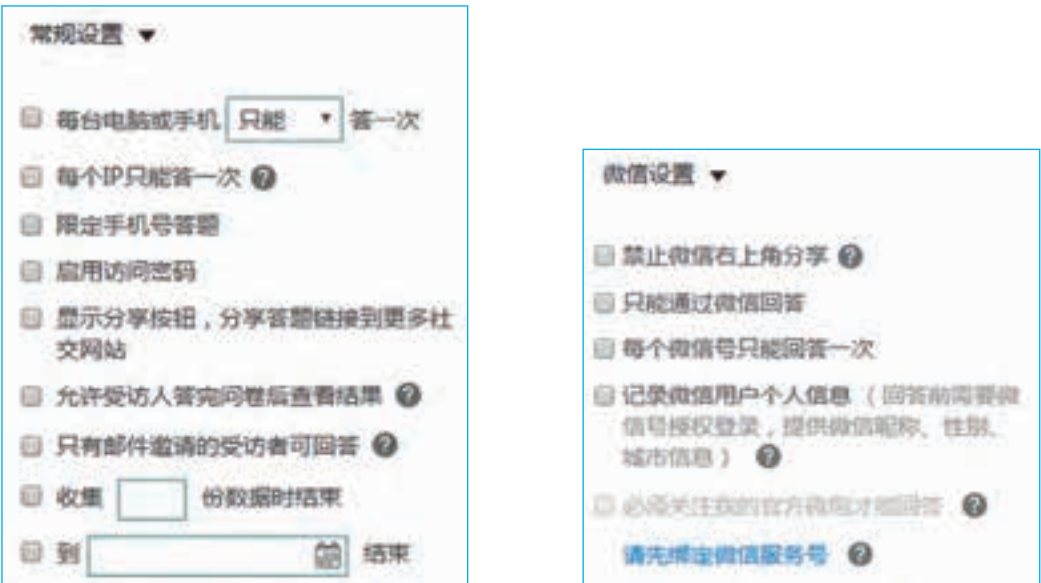


图 2-8 某在线调查平台的数据采集设置

思考与讨论??

- 1. 小组选择的平台提供了哪些数据采集设置？每条设置会对调查取样产生怎样的影响？
- 2. 如果只想将问卷链接发送给本地区几所指定中学的学生，有哪些方法？采用这些方法会采集到大量非目标人群的数据吗？

(3) 获取表格数据

传统纸质问卷回收后，需要研究者设计表格，手工录入纸质答卷上的数据。在线问卷则可以利用调查平台的相关功能，如“导出数据”，直接下载本次调查的答卷文件、表格数据(图 2-9)。

1	A	B	C	D	E	F	G	H	I
2	序号	提交答卷时间	所用时间	来源	来源详情	来源IP	1. 性别：	2. 你的年级：	3. 你平时住校还是走读？
3	1	2017/2/14 20:19:45	66秒	手机提交	直接访问	1.189.139.58(黑龙江-哈尔滨)	女	高二	住校
4	2	2017/2/14 20:20:49	116秒	手机提交	微信	101.85.178.104(上海-上海)	男	高三	住校
5	3	2017/2/14 20:28:03	163秒	手机提交	直接访问	222.69.88.7(上海-上海)	女	高二	住校
6	4	2017/2/14 20:28:03	91秒	手机提交	微信	1.189.139.60(黑龙江-哈尔滨)	男	高一	住校
7	5	2017/2/14 20:32:07	87秒	手机提交	直接访问	117.136.8.73(上海-上海)	男	高一	住校
8	6	2017/2/14 20:33:26	68秒	链接	http://192.168.	1.189.139.59(黑龙江-哈尔滨)	男	高三	走读
9	7	2017/2/14 20:33:26	59秒	手机提交	直接访问	114.89.185.164(上海-上海)	男	高二	住校
10	8	2017/2/14 20:34:47	84秒	手机提交	直接访问	114.89.248.170(上海-上海)	男	高二	住校
11	9	2017/2/14 20:35:25	97秒	链接	http://192.168.	1.189.139.59(黑龙江-哈尔滨)	女	高二	住校
12	10	2017/2/14 20:35:45	92秒	手机提交	直接访问	101.98.127.227(上海-上海)	男	高一	住校
13	11	2017/2/14 20:38:32	174秒	手机提交	微信	101.85.152.68(上海-上海)	女	高一	住校
14	12	2017/2/14 20:38:39	70秒	链接	http://192.168.	1.189.139.60(黑龙江-哈尔滨)	男	高三	住校
15	13	2017/2/14 20:44:53	101秒	手机提交	微信	114.81.254.163(上海-上海)	女	高二	住校
16	14	2017/2/14 20:55:10	87秒	手机提交	直接访问	58.40.172.92(上海-上海)	男	高三	走读
17	15	2017/2/14 21:08:27	126秒	手机提交	直接访问	180.160.55.47(上海-上海)	男	高三	住校

图 2-9 某在线调查平台提供的答卷文件

思考与讨论??

- 1. 传统纸质问卷通过“不记名”达到保护个人隐私的目的，但在线调查通过 IP 地址就可以跟踪数据来源。在线调查应该如何保护个人隐私？如何对数据进行保密？
- 2. 某同学在公交车站等车时，一名市场调查人员请他用手机扫一个二维码，说是在网上简单回答一些问题，即可获得一份小礼品。这名同学是否应该接受这名调查人员的提议呢？

活 动

3.3 各小组设计并发放自己的“中学生移动学习现状调查”问卷，采集数据。

(1) 小组同学合作设计一份问卷。

(2) 在小组选择的平台注册，并在平台上完成问卷的创建和编辑。问卷设计好后，各组自评或互评问卷，根据所提意见修改完善问卷。

(3) 确定问卷的发放范围、途径、数量及调查结束时间，发放问卷，将链接发送给目标人群。

(4) 问卷收集结束后，下载答卷文件，获取表格数据。

3. 加工、分析和可视化数据

调查问卷被收回意味着数据采集工作结束。接下来就要进行数据加工、数据分析和数据可视化了。

在线调查平台会在与用户的交互中，完成数据的加工、分析和可视化，用户可以直接查看结果，选择想要的结果呈现类型(图 2-10)。



图 2-10 单选题统计结果表格和饼状图示例

在观察了关于每个问题的统计数据后，如果发现有些数据还需要进一步挖掘，可以利用平台提供的分析工具，如分类统计、交叉分析等，进一步完成一些比较复杂的数据分析工作。例如，若调查结果显示，75% 的中学生每周移动学习时间少于 1 小时，那么可进一步问，是否住校、不同性别所得

核心概念

数据分析 (data analysis) 是指用适当的分析方法与工具，对采集的数据进行分类整理，提取与发现其中有价值的信息，以形成结论的过程。

数据可视化 (data visualization) 是指将数据分析的结果通过表格、图表、图形等形式显示出来。

数字化学习

请利用网络开展学习，了解利用电子表格软件制作图表的方法。

到的结果有没有差异？为了验证自己的假设，再对数据做交叉分析（图 2-11）。可以发现，是否住校对中学生移动学习的时间有比较大的影响。

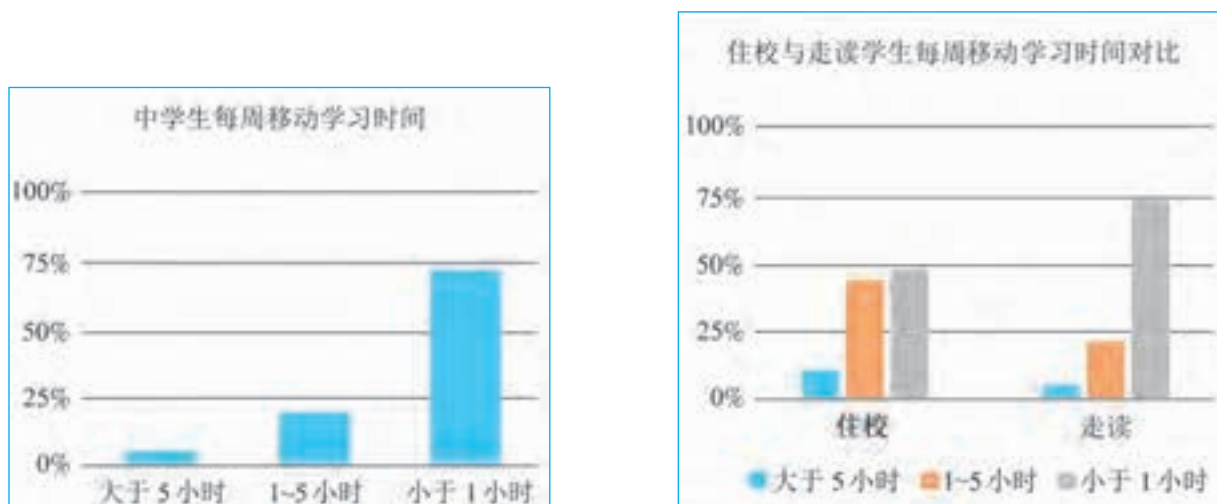


图 2-11 交叉分析及可视化示例

如果调查者认为平台提供的数据分析功能无法满足某些个性化或复杂的需求，还可以利用下载的答卷文件，运用专业的统计分析软件进行数据加工、分析和可视化。

思考与讨论??

1. 问卷调查中常见的单选题、多选题、排序题，分别适合哪些统计分析方法？分别用什么类型的图表呈现效果较好？
2. 调查采集到的数据需要备份吗？为什么？如果打算备份数据，则应采用何种方法？

活 动

3.4 分析和可视化“中学生移动学习现状调查”数据。

- (1) 查看在线调查平台中每个问题的统计结果，选择合适的统计方法和呈现方式。
- (2) 观察数据及其初步统计的结果，展开小组讨论，提出新的假设，使用在线调查平台的交叉分析工具对问卷中的一些数据进行交叉分析，挖掘更多的信息。
- (3) 选择合适的可视化工具，将数据分析结果用图表等形式表达出来。

4. 撰写报告, 提出数据应用建议

调查报告要清楚、准确地报告研究者为解决所研究的问题而做的一切工作。调查报告一般包括五个部分: 研究背景和研究目的、调查对象和调查方法、调查结果、调查结论、意见和建议。进行汇报时要包括以下内容:

- 数据需求的产生(研究背景和研究目的)。
- 数据来源和采集数据的方法(调查对象和调查方法)。
- 数据的分析和可视化结果(调查结果)。
- 数据背后隐藏的信息(调查结论)。
- 数据应用(意见和建议)。

社会调查研究的最终目的是应用研究结果, 为相关人员提供信息, 帮助他们更好地决策。因此, “中学生移动学习现状调查”项目最后的工作就是选择合适的受众, 向他们宣传自己的研究(图 2-12)。研究者也可以将自己的研究通过互联网与更多人共享。

小贴士

研究人员应该坚持实事求是、认真严谨的态度。研究报告应该全面而诚实地报告研究过程和研究结果, 不能编造数据来支持自己的观点, 更不能因为商业利益而在报告中宣传某些产品。



图 2-12 通过演示与他人分享研究成果

思考与讨论??

1. 与传统的纸质问卷调查相比, 在线问卷调查有哪些优势和劣势?
2. 对于现有的在线问卷调查平台, 你有哪些改进意见和建议?

活 动

3.5 撰写“中学生移动学习现状调查报告”。

- (1) 了解报告的结构、数据的显示方式、结论与建议等。
- (2) 各小组撰写调查报告。
- (3) 全班互评调查报告。
- (4) 从数据来源和问卷设计等方面，反思本次问卷调查的数据对本小组观点的支持度和有关结论的可信度。



知识链接

数据处理及其作用

数据并不是一堆枯燥无味的、单纯的数字，深挖数据背后的价值，可获取更多有用的信息。数据处理是从大量的原始数据中抽取有价值信息的过程，即数据转换成信息的过程。它是对输入的各种形式的数据进行加工整理，这一过程包含对数据的采集、存储、加工、分析和表达。

数据处理的作用体现在三个方面：现状分析、原因分析和预测分析。

数据处理的应用无处不在，举不胜举。例如，某高校对学生在食堂刷卡吃饭的数据进行分析，确定受扶助学生的名单和资助金额，“偷偷”给这些学生的饭卡充钱，精准地扶贫；基于对用户搜索行为、浏览行为、评论历史和个人资料等数据的分析，某互联网企业向用户推荐他们可能喜欢的书籍、电影、美食或近期可能要购买的商品；某互联网公司大数据部上线的“疾病预测”，利用用户的搜索数据，并结合气温、湿度变化等因素建立预测模型，实时提供流感等疾病的活跃度、流行指数；交通部门基于用户和车辆的定位数据，分析道路拥堵的原因，从而针对不同时间点、不同道路的车流量进行智能车辆调度或采用潮汐车道。

数据处理的一般过程

数据处理的一般过程如图 2-13 所示。



图 2-13 数据处理的一般过程

1. 明确目标

明确目标是指明确数据处理的目的，确立分析思路。首先，要思考开展数据处理的原因，即要解决什么问题。然后，要梳理数据分析的思路，搭建分析框架，确定使用哪些分析方法和工具。

2. 数据采集

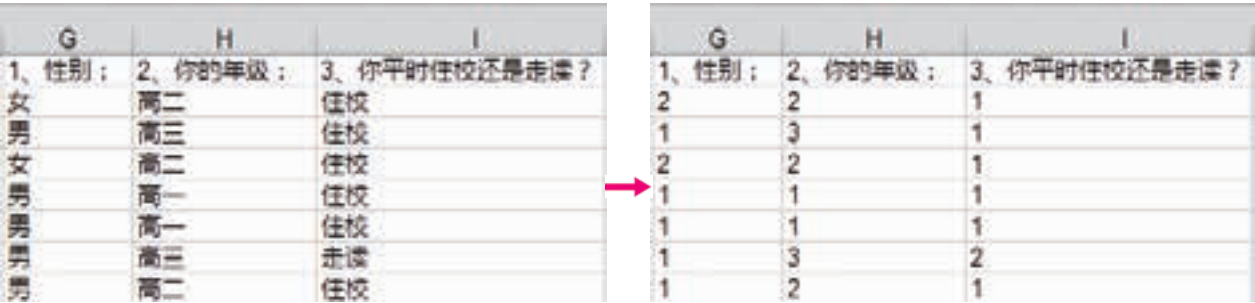
数据采集是指人们根据需要获取数据，它是确保数据处理过程有效的基础。技术工具的发展使得数据采集方式日趋多样。目前数据采集的来源主要有以下几种：

- 人工输入的观察、调研数据；
- 利用技术工具（例如传感器）直接采集的数据；
- 各种数据库中的数据；
- 利用搜索引擎工具在网络上快速获取的数据；
- 通过网络调查问卷采集的数据。

3. 数据加工

数据加工是指通过数据编码、数据清洗、数据重组等一系列过程，使采集到的数据符合数据分析的需求。

在本项目中，直接使用平台提供的功能对采集到的问卷数据进行统计分析，省略了数据加工环节。但在现实工作中，采集到的数据大多不能立即用于数据分析，还须使用恰当的工具和方法进行加工。以问卷调查为例，须剔除出现答案残缺不全、重复填写、数据错误等问题的答卷；还有些数据须进行编码，如性别、年级、是否住校等选择题答案（图 2-14）；而开放式问题得到的回答会非常多，显得杂乱无章，须分类并确定代码。



G	H	I
1、性别：	2、你的年级：	3、你平时住校还是走读？
女	高二	住校
男	高三	住校
女	高二	住校
男	高一	住校
男	高一	住校
男	高三	走读
男	高二	住校

G	H	I
1、性别：	2、你的年级：	3、你平时住校还是走读？
2	2	1
1	3	1
2	2	1
1	1	1
1	1	1
1	3	2
1	2	1

图 2-14 未经编码的数据表（左）和经过编码的数据表（右）

4. 数据分析

数据分析是指用适当的分析方法与工具，对采集到的数据进行分类整理，提取与发现其中有价值的信息，形成结论的过程。数据分析的目的是从描述研究对象的数据中发现其内在特征和规律。数据分析有对比、细分和预测三大类，它们又各自对应不同的具体分析方法。

在日常工作和现状研究中，运用最多的是描述性分析方法，如对比分析法、平均分析法和交叉分析法。

(1) 对比分析法

对比是人们认识客观世界的基本方法。通过将两个或两个以上的数据进行对比，分析它们的差异，可以分辨数据的性质、变化、发展等个性特征。对比分析法可以分为横向比较和纵向比较(图 2-15)。横向比较是同一时间不同总体指标的对比，纵向比较是不同时间同一总体指标的对比。

(2) 平均分析法

利用计算平均数的方法，可以反映总体在一定时间、地点下数据特征的一般水平。平均分析法可以分为位置平均数和数值平均数(图 2-16)。其中，运用得最多的是算术平均数。



图 2-15 对比分析法

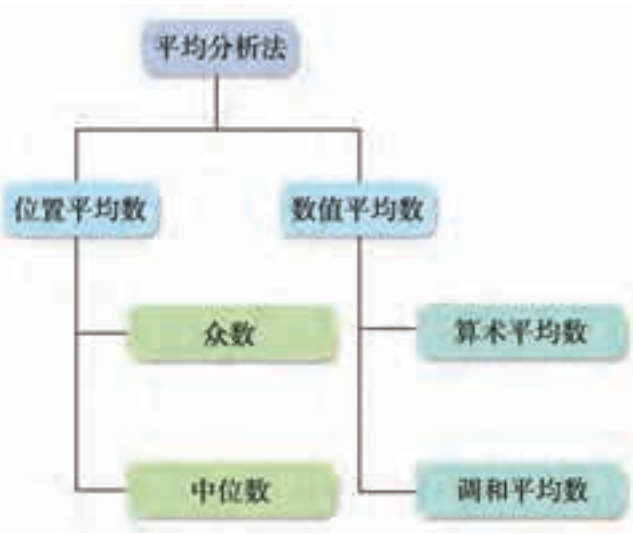


图 2-16 平均分析法

(3) 交叉分析法

这是一种立体分析法，它从横向和纵向两个方向来计算两个或多个有联系的变量在交叉点的统计值。

5. 数据可视化

数据可视化是指将数据分析的结果通过表格、图表、图形等形式显示出来，还可以通过这些形式对分析结果进行一些交互处理。利用人对形状、颜色等特性的感官敏感性，数据可视化能更清晰、有效地帮助人们发现数据之间的关系、规律和趋势，传递数据背后的信息，如图 2-17 所示。

常见的数据图表包括条形图(图 2-18)、折线图、饼图(图 2-19)、柱状图(图 2-20)、面积图、散点图、雷达图等，使用图表工具还可以得到交互图表(图 2-21)、漏斗图、帕累托图、旋风图、矩阵图等，数据可视化图形则包括地图、词云(图 2-22)、热力图(图 2-23)、树图、网络图等，甚至可以是动图、动画。



图 2-17 利用数据可视化工具呈现的某游乐园游客实时数据

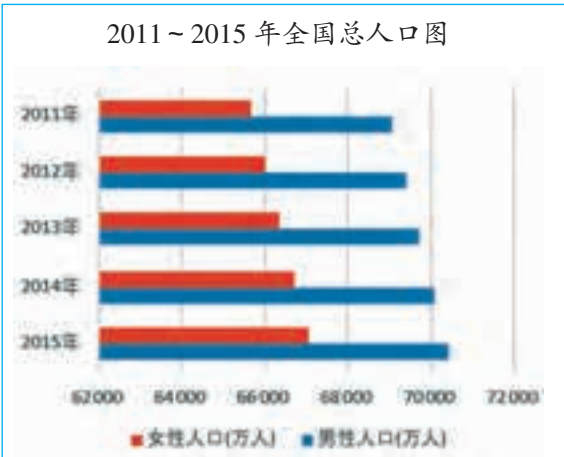


图 2-18 条形图示例



图 2-19 饼图示例



图 2-20 柱状图示例



图 2-21 交互图表示例



图 2-22 词云示例

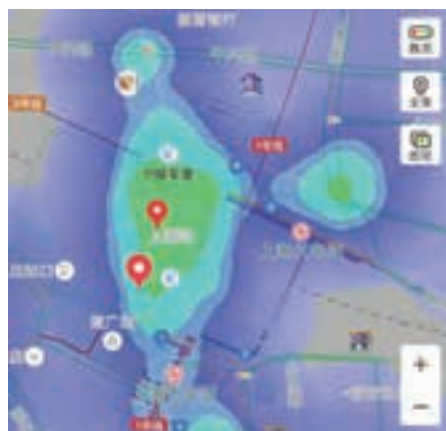


图 2-23 热力图示例

最常用的图表工具是 WPS 表格或 Excel 等电子表格软件，在互联网上有大量优秀的数据可视化工具，如 RAW、Infogram、Tableau 等。除了利用现成的工具，也可以按自己需要编程实现。

6. 报告撰写

报告撰写是对整个数据处理过程的总结。通过报告,将数据处理的目的、过程、结果及方案完整地呈现出来,为决策提供参考或依据。报告的种类很多,但不管采用怎样的呈现方式,都应做到清晰可读,尤其要注重数据可视化,以便于阅读者正确、迅速地理解报告内容。此外,报告不仅要发现问题,更要有依据科学、严谨的数据分析过程推导出来的结论和建议。

认识智能停车场中的数据处理

——体验数据处理的方法和工具

在面积较大、空闲率低的停车场，尤其是层数较多的地下停车场，找到空闲车位往往是让驾驶员头疼的事。为了帮助驾驶员快速找到空闲车位，避免因盲目找车位造成停车场内部通道的堵塞，近年来，人们利用信息技术设计研发了智能停车场管理系统，如图 2-24 所示。在智能停车场中，停车引导和车辆收费是两项重要的工作。不同于传统的由人工登记进出车辆相关数据的做法，智能停车场能自动记录停车位的使用数据和车辆的进出数据，并自动计算车辆的停车时间和费用。有些智能停车场，还会对多年积累的车位使用数据、缴费数据等进行分析，为停车场的科学管理提供依据。在有些城市，市中心不少智能停车场的的数据都被接入城市智能交通系统，再通过街上的引导屏，告知周边停车场的空闲车位数，引导车辆找到停车位。



图 2-24 某智能停车场

项目学习目标

在本项目中，我们首先通过探究智能停车场的车辆引导和停车费计算这两项工作，了解数据采集、组织和计算的基本方法。然后学习使用一种数据处理工具，对智能停车场数据库中记录的停车位数据进行处理，获取隐藏在这些数据中的信息。

完成本项目学习，须回答以下问题：

1. 采集数据的具体方法和工具有哪些？
2. 数据的组织方式是怎样的？
3. 数据的存储方式有哪些？
4. 表格数据的加工方法有哪些？



图 2-25 户外车位引导屏

项目学习指引

1. 探究停车引导中的数据处理

智能停车场往往在停车场入口、各层各区域的交叉路口设立引导屏，显示行车路线指引信息和空闲车位数（图 2-25）。同时，在每个车位的上方，根据车位的使用情况控制车位指示灯显示不同的颜色——绿色为“空闲”，红色为“占用”。驾驶员在几十米外即可看到指示灯，方便他们快速找到空闲车位。

思考与讨论??

要统计出图 2-26 户外车位引导屏中的空闲车位数，需要采集每一个车位的哪些数据？与一个停车场中的车位相关的数据还有哪些？



根据户外车位引导屏指示
进入车库



根据区域引导屏指示
进入有空闲车位的区域



根据车位指示灯
快速找到空闲车位

图 2-26 停车引导过程

小贴士

数据的自动采集是指利用技术手段（如传感器、摄像头）从系统外部直接采集数据并输送到系统内部接口的过程。

要实现停车引导，首先应该了解车位占用情况，再将采集到的数据保存下来，并进行分析。

（1）车位占用情况数据的自动采集

大型停车场通常采用多层多区域的组织方式，车位占用情况数据（空闲或占用）不是用人工记录，而是通过传感装置自动采集。图 2-27 是使用超声波传感器采集车位占用情况的示意图。

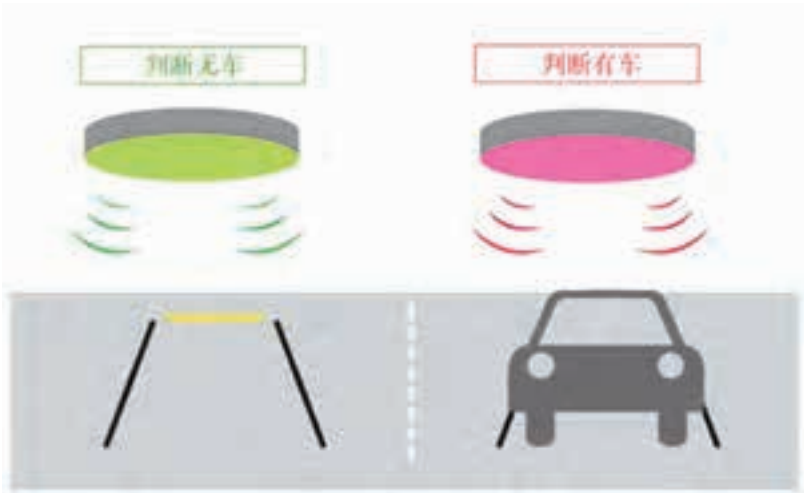


图 2-27 超声波传感器采集车位占用情况

安装在车位上方的超声波传感器自上而下发出超声波，并据此探测、分析物体或地面的反射波，精确测量出反射面到传感器的距离，由此准确地检测出每个车位的占用情况，从而实时采集到车位占用情况数据，同时将该数据用车位指示灯直观呈现。

思考与讨论??

某居民小区拟在出入口安装车牌识别摄像头，但部分车主担心自己的个人信息、车牌信息、停车数据等隐私被泄露。你觉得他们的顾虑有道理吗？

← 参见 P56 知识链接“数据采集的方法和工具”

← 参见 P57 知识链接“数据的保护”

活 动

4.1 调查某个居民小区停车管理的数据采集方式。

有些大型居民小区已对小区的停车实现了智能化管理。小区停车位有在地面的，有在地下的；有出售给业主的，有租赁给业主的，还有供临时停放的。有固定车位的业主不用在每次进出小区时缴费，临时停放在小区的车辆离开时需要按规定缴费。智能停车管理系统会在小区入口实时采集进入车辆的数据，在出口对离开车辆进行识别，并对临时停放的车辆根据车辆出入时间计算停车费。

（1）选择一个居民小区开展调查，了解其停车收费管理中需要用到哪些数据，以及这些数据是如何采集的。

（2）了解该居民小区在入口采集车辆数据的方式、采集数据所使用的硬件设备，以及获取的数据。

方式 1：_____ 数据采集设备：_____ 获取的数据：_____

方式 2：_____ 数据采集设备：_____ 获取的数据：_____

(2) 停车位数据的组织

智能停车场引导屏上的空闲车位数是按停车场的层、区域甚至方位汇总得到的。停车引导系统采集好车位占用情况数据后，除了改变空闲车位数，还要输送车位的所在层、所在区域等数据。

停车位数据到底包括哪些内容？这个问题关系到如何有效构建数据。任何一个事物都包含许多属性，事物的全部属性可按需求选择性地呈现。要解决的问题不同，对同一事物要呈现的属性也不同。在应用系统中，由若干属性构成的数据称为结构化数据(structured data)。

停车引导工作的需求是汇总不同层、不同区域的空闲车位数，统计时往往要用到车位占用情况、层、区域等属性，而无须关心车位尺寸等属性。例如，某停车场的停车位使用实时数据包括采集时间、层、区域、编号、车位占用情况、状态(对应该停车场的管理情况，如将车位分为开放、关闭、内部)，其结构图如图 2-28 所示。

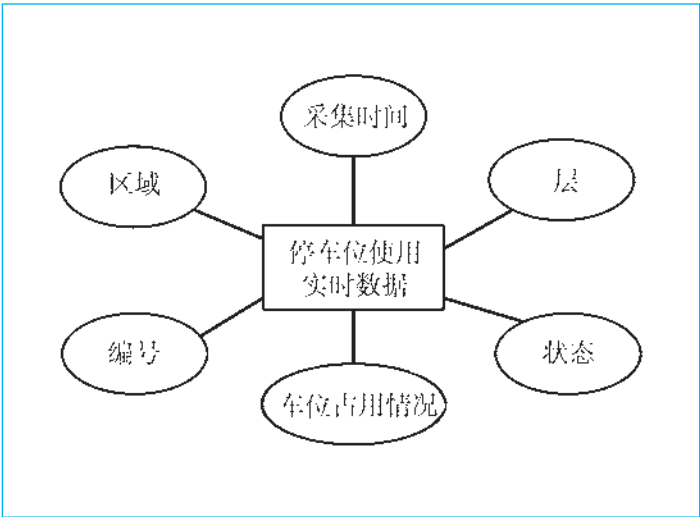


图 2-28 停车位使用实时数据结构图

小贴士

关系表(relational table)：一种规范化的二维表，符合关系模型的要求。

在大多数的数据处理中，通常以关系表的形式组织数据，如图 2-29 所示的停车位使用实时数据表。表格的第一行是标题，列出“采集时间”“编号”等属性名称。从第二行开始，每一行是一条记录(record)。每一列是一个属性，称为表格的字段(field)。行列交叉处是一个单元格，存放一条记录的某个字段值。

序号	采集时间	层	区域	编号	车位占用情况	状态
1	2016-12-8 14:30:00	B1	A	10	0	内部
2	2016-12-8 14:30:00	B1	A	11	0	内部
3	2016-12-8 14:30:00	B1	B	1	1	开放
4	2016-12-8 14:30:00	B1	C	1	1	开放
5	2016-12-8 14:30:00	B1	D	1	1	开放
6	2016-12-8 14:30:00	B1	D	2	0	开放
7	2016-12-8 14:30:00	B2	A	1	0	关闭
8	2016-12-8 14:30:00	B2	A	2	0	关闭

图 2-29 停车位使用实时数据表

(3) 停车位数据的存储

如图 2-30 所示，停车引导工作本质上是一个数据处理过程。停车位使用实时数据表就存储在停车场服务器的数据库中，供停车场管理者查询实时数据或分析历史数据。

要将数据存储到数据库中，首先要创建关系表的结构，然后将记录添加到关系表中。创建关系表结构时要确定关系表中每一个属性的数据类型。为了保证数据的有效性，还要设置属性值的数据约束，包括属性值是否唯一、是否可以为空、是否要在一定的数值范围内等。

小贴士

数据库(database)是按照特定的数据结构(data structure)来组织、存储和管理数据的、建立在计算机存储设备上的仓库。不同类型数据在计算机内存储和处理的方式不相同，因此数据库中的数据必须明确其数据类型。

参见 P58 知识链接“数据的组织和存储”

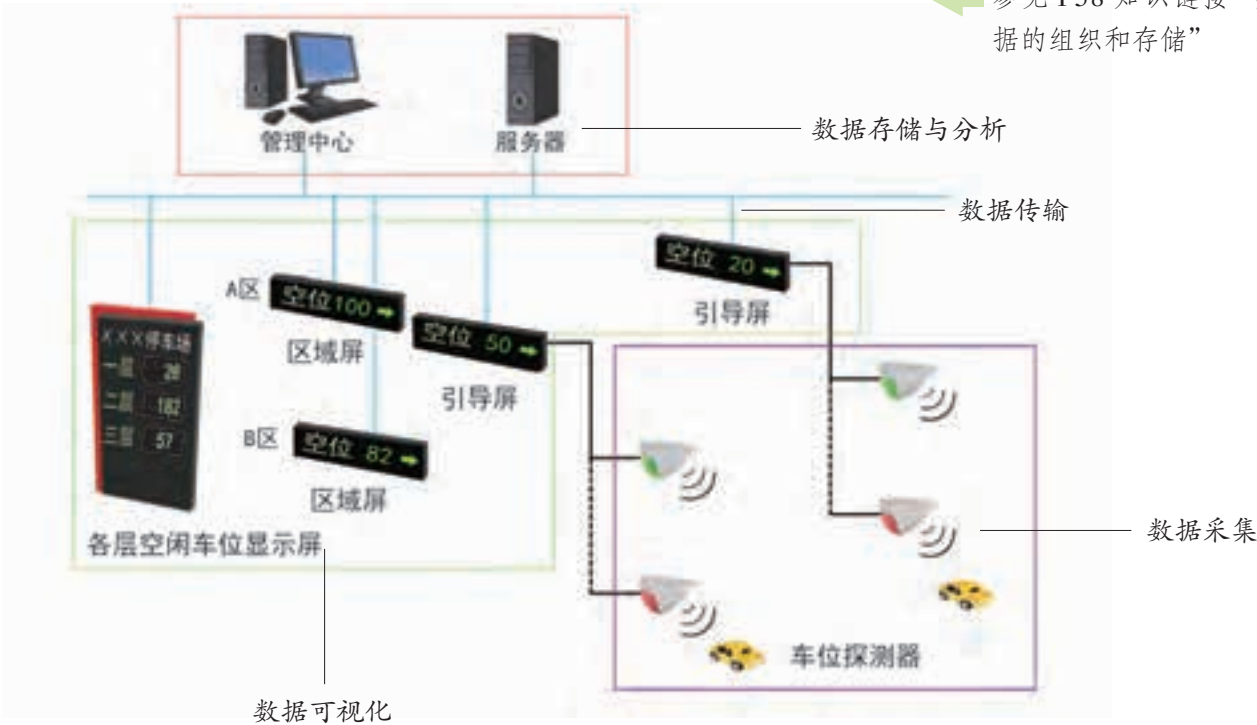


图 2-30 停车引导中的数据处理

活 动

4.2 设计居民小区停车位管理表。

居民小区的停车管理是小区物业的重要工作。请为活动 4.1 中所调查的居民小区设计一张停车位管理表，方便物业管理人员了解停车位的分配情况（已出售、已租赁、待租赁、临时），以及已租赁停车位的月租费缴纳情况。

- （1）设计停车位管理表。
- （2）使用一种电子表格软件创建停车位管理表。
- （3）分析表中的各项属性值可以从哪里采集。

4.3 表 2-1 是某停车场的停车位使用实时数据表的字段构成。请用“数值”“文本”“日期/时间”或“逻辑”填写各字段的数据类型，用“唯一”或“非空”填写数据约束，也可以设定属性的合理数值范围，便于出错时进行检查。

表 2-1 停车位使用实时数据表的字段构成

属性	序号	采集时间	层	区域	编号	车位占用情况	状态
数据类型							
数据约束							

2. 计算停车费

在停车场管理中，除了停车引导，计算停车费也是一项重要工作。为了计算停车费，停车场大多在入口处采集车辆的驶入时间，在出口处采集车辆的驶出时间。根据驶入时间和驶出时间，计算车辆在停车场的停留时间，再根据停车收费规定计算停车费。

在智能停车场中，停车费的计算工作是由系统自动完成的。例如，按照图 2-31 的停车收费规定，利用电子表格软件这一常见的数据处理工具，可以模拟停车费的计算工作。

电子表格软件一般以工作表来组织数据，一张工作表由若干个单元格构成。每个单元格可以存储一个数据，单元格的值可以直接手工输入，也可以通过公式计算得到。利用查询等功能，可以从停车场数据库中导出数据，再在电子表格软件中导入数据，得到类似图 2-32 所示的表格。



图 2-31 某停车场收费规定

	A	B	C	D
1	序号	车牌	驶入时间	驶出时间
2	1	沪A87358	2016/02/28 11:00	2016/02/28 11:25
3	2	沪A39498	2016/02/28 07:50	2016/02/28 09:45
4	3	沪C03818	2016/02/28 22:00	2016/02/29 05:30
5	4	沪A43987	2016/02/28 10:23	2016/03/03 13:43

单元格 A1 区域 B3:C4

图 2-32 从停车场数据库中导出的表格数据

每个单元格都有一个引用名称，例如 A1 表示第 1 行第 1 列的单元格。多个单元格在行列方向上连续排列而构成的矩形称为区域，引用名称由左上角单元格和右下角单元格构成。例如 B3:C4 表示从第 3 行第 2 列到第 4 行第 3 列的矩形区域，包含 4 个单元格。

← 参见 P58 知识链接“表格数据的加工方法”

思考与讨论??

- 1. 在上表的 C 列和 D 列中新插入一列后，序号 1 的驶出时间单元格的引用名称仍是 D2 吗？为什么？
- 2. 讨论上表中各个字段的计算特性，确定每个字段的数据类型。

根据停车收费规定，可以在以上导出表格的基础上添加辅助列和结果列，并利用软件内置的公式和函数计算停车费用，如图 2-33 所示。

电子表格软件中的公式 (formula) 是由常量、单元格、运算符、函数等构成的表达式，由 “=” 引出。

函数 (function) 则是软件内置的一些常用功能模块，在使用的时候可以代入参数，获得相应的计算结果。图 2-33 的公式中使用的 INT 函数的作用是取整数。

时间 / 日期数据在计算机中表示为一个数值型的编码，其中日期是长整型，时间则是小于 1 的实数。时间 / 日期数据可以进行日期、时间的加减运算。两个时间 / 日期数据做减法，得到的差值单位为天；若要以小时进行计算，可以再乘以 24。

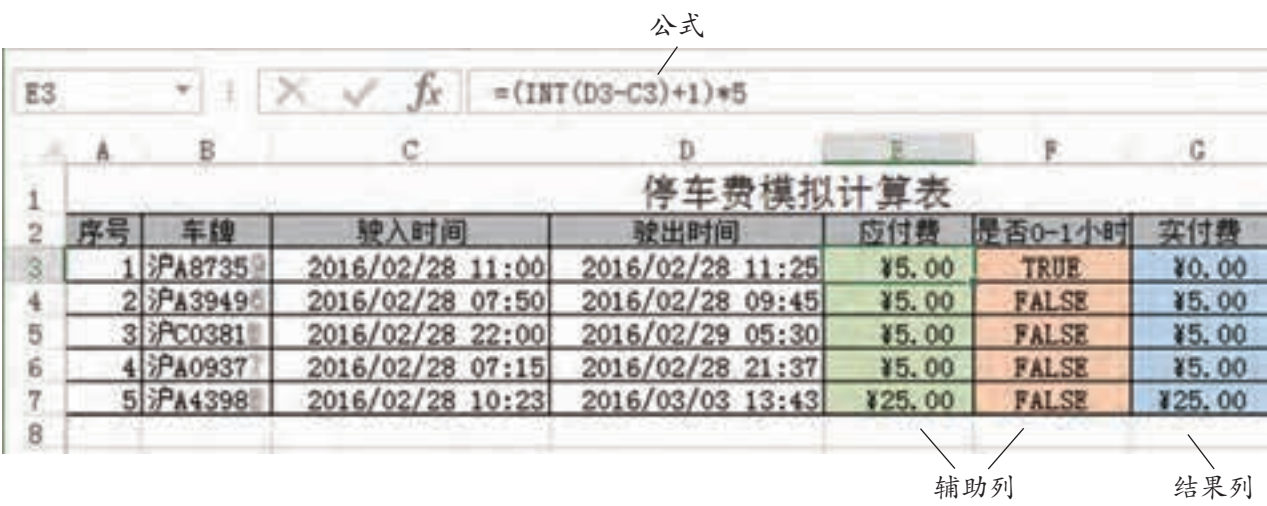


图 2-33 停车费模拟计算表

思考与讨论??

1. 图 2-33 的公式中为什么会出现“+1”？
2. 在 E3 单元格中输入公式后，如何将公式应用于其他单元格（E4 至 E7）？你能想到几种方法？
3. 若在“驶入时间”列前插入一空列，则原来的“驶入时间”列的单元格 C3: C7 变成了 D3: D7，这会 影响辅助列和结果列的值吗？为什么？

活 动

4.4 为某停车场计算停车费。

某停车场收费规则为：半小时以内免费，超过半小时不到 1 小时收费 5 元，之后每小时加收 2 元，但 24 小时内最高收费限额为 20 元，超过 24 小时则重新按上述规定计费。

（1）按照以上收费规则，设计停车费计算表，补充表 2-2 的表头。

表 2-2 停车费计算表

序号	车牌	驶入时间	驶出时间					

（2）选择一种电子表格软件，导入本项目中从数据库导出的数据。

（3）创建停车费计算表，并利用公式和函数计算停车费。

3. 分析停车位使用数据

停车引导和车辆收费等数据处理工作是智能停车场的日常业务活动。日积月累，数据库中会存储大量的数据。例如，按一定时间间隔（半分钟甚至更短）采集每个车位的占用情况数据，一天下来，数据库中可能存储上万或几十万条记录，一周、一个月、一年的数据累计量更是庞大。而对这些数据进行分析 and 挖掘，可以得到新的信息，为停车场的决策提供支持服务。

(1) 分析某一时刻各停车位实时使用情况

例如，某停车场接到通知：下周四 14:30，停车场附近有临时性的展览活动，预计会有大量车辆驶入。停车场管理人员想了解：该时间段停车场的接纳量大概是多少？哪些区域空闲车位多？该时间段如何开展停车引导？

① 针对此任务，数据分析人员从数据库中导出类似时刻（如采集时间为 2016 年 12 月 8 日周四 14:30）的所有记录，得到相应的数据表，并在电子表格软件中打开，如图 2-34 所示。

序号	采集时间	层	区域	编号	车位占用情况	状态
1	2016-12-8 14:30:00	B1	A	10	0	内部
2	2016-12-8 14:30:00	B1	A	11	0	内部
3	2016-12-8 14:30:00	B1	B	1	1	开放
4	2016-12-8 14:30:00	B1	C	1	1	开放
5	2016-12-8 14:30:00	B1	D	1	1	开放
6	2016-12-8 14:30:00	B1	D	2	0	开放
7	2016-12-8 14:30:00	B2	A	1	0	关闭
8	2016-12-8 14:30:00	B2	A	2	0	关闭

图 2-34 从数据库中导出的某时刻停车位使用实时数据表

② 对数据表中的数据进行汇总统计。数据透视表工具是一种快速汇总大量数据的交叉分类统计工具。利用数据透视表工具可以对数据表中的数据在行、列方向上重新布局 **分类字段**，在行列交叉处按照设定的统计方法（如求和、计数、最大值、最小值、方差）计算统计值。使用数据透视表工具进行交叉分类统计，选择图 2-35 中的行、列、数值组合及值字段设置，就可得到某时刻停车位数据透视表，如图 2-36 所示。

小贴士

为更好地反映某时刻的车位占用情况，一般还需要导出更多数据来减少误差，如导出近三个月每周四 14:30 的数据，或者近一个月每个工作日 14:30 的数据等。

小贴士

分类字段的数值应有区分性，一个值表示一个类别，且数值范围是一个有限集合。如停车位使用实时数据表中的字段“层”“区域”“状态”都可作为分类字段。



图 2-35 某电子表格软件的数据透视表设置

采集时间	层	区域	车位总数			已占车位数			车位总数汇总	已占车位数汇总
			关闭	开放	内部	关闭	开放	内部		
2016-12-8 14:30:00	B1	A			25			11	25	11
2016-12-8 14:30:00	B1	B		20		20			30	20
2016-12-8 14:30:00	B1	C		26		23			30	23
2016-12-8 14:30:00	B1	D		30		11			30	11
2016-12-8 14:30:00	B1	E		25		23			25	23
2016-12-8 14:30:00	B1	汇总		121	25	80	11		146	91
2016-12-8 14:30:00	B2	A		20		5			20	5
2016-12-8 14:30:00	B2	B		31		11			31	11
2016-12-8 14:30:00	B2	C	30			0			30	0
2016-12-8 14:30:00	B2	D	32			0			32	0
2016-12-8 14:30:00	B2	E	32			0			32	0
2016-12-8 14:30:00	B2	汇总	94	61		0	16		155	16
2016-12-8 14:30:00	汇总		94	182	25	0	96	11	301	107
总计			94	182	25	0	96	11	301	107

图 2-36 某时刻停车位数据透视表

从图 2-36 所示的数据透视表中，可以读到该时刻每一层每一区域的车位总数和已占车位数，以及在不同状态（关闭、开放、内部）的统计数据。

思考与讨论??

1. 在图 2-36 所示的数据透视表中，整个停车场分为几个区域？如何准确描述一个区域？
2. B1 层 A 区域内部车位的使用情况如何？B2 层 B 区域所有车位的使用情况如何？
3. B1 层有多少空闲车位？整个停车场有多少空闲车位？有多少内部车位？

小贴士

排序是对表中的一列或多列数据按指定顺序（升序或降序）重新显示。多关键字排序时，对表中数据先按主要关键字排序，对主要关键字值相同的数据再按次关键字排序，以此类推。

③ 对图 2-36 所示的数据透视表，使用**排序**工具按已占车位数汇总的数据顺序进行重新排列（图 2-37），可以清楚地查看某一时刻各个停车区域的车位饱和情况，并找出各层中相对空闲的区域，如图 2-38 和图 2-39 所示。

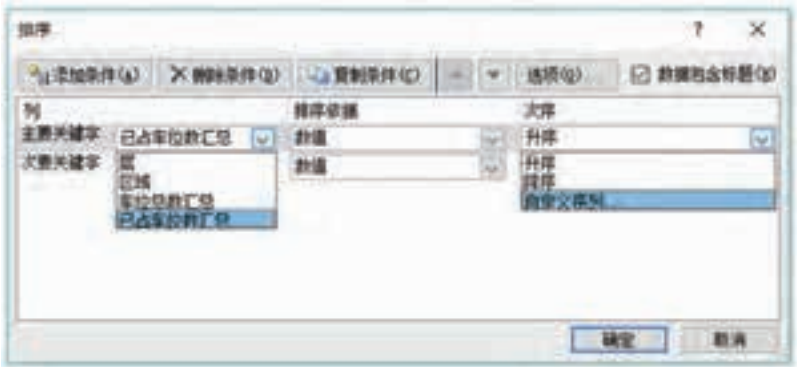


图 2-37 排序工具

层	区域	车位总数汇总	已占车位数汇总
B1	A	25	11
B1	B	20	20
B1	C	30	23
B1	D	20	14
B1	E	25	23

图 2-38 数据透视表中 B1 层的汇总数据(源数据表)



层	区域	车位总数汇总	已占车位数汇总
B1	A	25	23
B1	B	20	20
B1	C	30	14
B1	D	20	11
B1	E	25	23

图 2-39 排序后的目标数据表

④ 对图 2-36 所示的数据透视表，使用筛选工具查看某一时刻各层关闭的车位总数，可供管理人员对活动日当天是否将关闭的保留车位对外开放进行决策，如图 2-40 和图 2-41 所示。

小贴士

筛选就是留下符合条件的数据。很多数据处理软件都提供了筛选工具，数据透视表工具的每一列(除汇总数据)也都支持筛选操作。



图 2-40 筛选工具

采集时间	层	区域	车位总数	已占车位数
2016-12-8 14:30:00	B2	C	30	0
2016-12-8 14:30:00	B2	B	32	0
2016-12-8 14:30:00	B2	E	32	0
2016-12-8 14:30:00	B2	汇总	94	0
2016-12-8 14:30:00	汇总		94	0
总计			94	0

图 2-41 某一时刻 B2 层关闭的车位总数

(2) 挖掘停车位历史数据

例如，某停车场接到通知：下周四 6:30 到 22:00，在停车场附近将安排一个一小时左右的活动，预计会有大量车辆驶入。为解决停车问题，希望停车场管理人员给出建议：下周四哪个时间段开展活动较好？为此，停车场管理人员需要了解平时工作日 6:30 到 22:00 各层、各区域的使用率，进而预测活动当天停车场的使用低谷时段。

① 针对此任务，数据分析人员从停车场数据库中导出类似的某天 6:30 到 22:00 的停车位使用实时数据表，筛选出以 10 分钟为间隔的记录，并按时间、层、区域分类统计已占车位数，再计算出车位占用率(已占用车位数 / 该区域车位总数)，得出各层各区域车位占用率数据表，如图 2-42 所示。

② 利用各层各区域车位占用率数据表中的数据，可以绘制出多种折线图，如图 2-43 和图 2-44 所示。

时间	层	区域	已占车位数	车位占用率
6:30	B1	A	0	0%
6:40	B1	A	0	0%
6:50	B1	A	0	0%
7:00	B1	A	2	8%
7:10	B1	A	4	16%
7:20	B1	A	5	20%
7:30	B1	A	5	20%
7:40	B1	A	5	20%
7:50	B1	A	5	20%
8:00	B1	A	5	20%
8:10	B1	A	6	24%
8:20	B1	A	6	24%
8:30	B1	A	7	28%
8:40	B1	A	7	28%
8:50	B1	A	10	40%
9:00	B1	A	10	40%
9:10	B1	A	12	48%
9:20	B1	A	12	48%
9:30	B1	A	13	52%
9:40	B1	A	13	52%
9:50	B1	A	13	52%
10:00	B1	A	13	52%

图 2-42 各层各区域车位占用率数据表(部分)

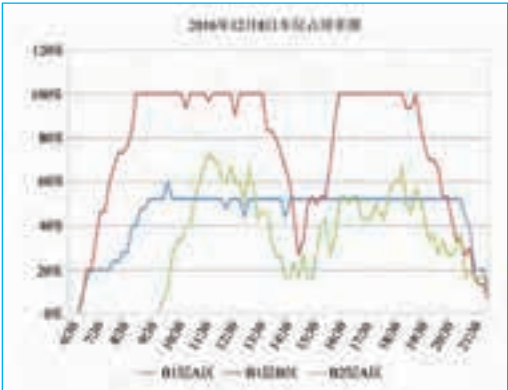


图 2-43 部分区域车位占用率图

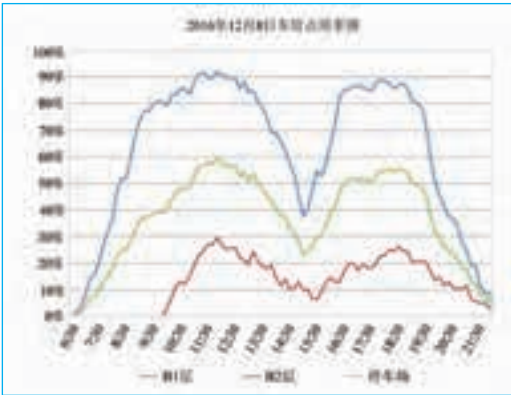


图 2-44 各层车位占用率图

从图 2-43 和图 2-44 中，可以明显看出各层各区域车位占用率的高低，以及车位占用高峰时间段，不同区域的占用率呈现出不同的趋势状态。

思考与讨论??

1. 如果要得到如图 2-44 所示的各层车位占用率图，该如何处理数据表中的数据？
2. 能不能从图 2-44 中得出结论“9:00~13:30、16:30~19:30 是该停车场的高峰时期，建议 9:00 开放 B2 层”？

通过分析各层各区域车位占用率数据表，停车场管理人员可以考虑实施动态优惠停车方案（在停车场的空闲时段实施车费优惠，车辆在该时段进入并离开停车场即可享受该优惠）等进一步的决策。

此外，生活中还有许多类似的数据处理工作，如图书馆的新书入库、借书、还书等日常业务活动，以及读者借书趋势、特定图书借阅频度等数据分析工作。图书馆管理人员也可以运用类似的数据分析方式来工作，为相关图书馆业务决策提供参考。

活 动

4.5 帮助停车场管理人员统计某一时刻停车场的空闲车位，计算停车场的实时车位占用率。

（1）利用以上项目中从数据库导出的某时刻停车位使用实时数据表，制作数据透视表，计算每层、每个区域的空闲车位数，统计整个停车场的空闲车位数，计算停车场的实时车位占用率（图 2-45）。

层	区域	总车位数	空闲车位数			车位占用率
			关闭	开放	总计	
B1	A					
B1	B					
B1	C					
B1	D					
B1	E					
	B1小计					
B2	A					
B2	B					
B2	C					
B2	D					
B2	E					
	B2小计					
	总计					

图 2-45 停车位使用实时统计表

（2）使用一种图表工具，选择合适的图表类型作图：①显示 B1 层的各个区域的车位占用率，②显示 B1 层、B2 层和整个停车场的车位占用率。

（3）从管理员的角度观察上面的图表数据，对是否开放或关闭某停车区域或停车层作出决策，实施分流措施。在小组内交流自己的决策及依据。



知识链接

数据采集的方法和工具

数据采集是数据处理工作的前提和基础。采集数据时，须运用适当的方法和工具。数据采集的常用方法和工具如表 2-3 所示。

表 2-3 数据采集的常用方法和工具

采集方法	人工获取					自动采集		
	调查	访谈	观察	实验	文献调研	物联感知	视频监控	网络平台
采集工具	问卷		表格		文献检索工具	传感器	摄像机	采集软件 网络爬虫 移动 App

1. 人工获取数据

人工获取数据指人直接从社会现象、自然现象或文献中获得数据。

（1）社会科学研究经常通过调查或访谈的方法获得一手数据。例如关于中学生移动学习现状的研究，可以从中学生的个体特质、移动学习的特点、家庭影响等角度展开探讨，设计相关问题，制作并发布问卷，从群体或个人获得一手数据。问卷是调查、访谈时常见的数据采集工具。相比传统纸质问卷，很多网络工具都可以帮助调查者更快捷、更精准地获得调查数据。

（2）在观察自然现象和进行科学实验时，需要设计各种表格，科学地记录通过观察或实验得到的各种数据。

（3）文献数据既包括正式出版、发行的纸质书刊、报表、年鉴，也包括政府机构、职能部门网站定期发布的公报、统计信息、研究报告等，还包括企业、机构网站上免费或有偿提供的数据库数据。文献数据可以通过文献检索工具获取，如利用搜索引擎工具在网络上搜索专业的数据库。

2. 自动采集数据

信息社会中，在信息技术的支持下，各种终端设备、网络数据库中记录存储着日益增长的海量数据。自动采集数据的方法通常有物联感知采集、视频监控采集、网络平台采集以及从已有数据库中采集等。

（1）物联感知采集，是指对于在物联网中使用电子标签或无线终端标识的智能化物体，通过传感器感知它们的数据变化，采集相关数据，并利用各种通信技术上传至网络信息中心存储。例如，通过佩戴相关设备，可以实时记录佩戴者的运动状态、呼吸量、血压、运动量、睡眠质量等生理状态数据，再利用无线或蓝牙技术，就可以将数据传送到网络信息中心或个人智能移动终端。

(2) 视频监控采集是指借助不同监控点的摄像机采集监控区域的数据。例如, 高清电子警察系统利用动态视频检测触发技术对车辆违规行为进行抓拍并完成车牌识别, 清晰、完整地记录车辆违章过程, 以及违章车辆的车型、车身颜色、车牌号码等数据。

(3) 网络平台采集主要是指用户在访问网站或使用 App 时, 网站服务器上安装的采集软件自动采集用户的各种行为数据。如一个学习平台可以采集学生浏览了哪些视频或课件, 看了多长时间, 重复观看了哪些课件, 是否快进观看, 以及观看课件的顺序等, 这些行为都被完整地记录在系统日志文件中。通过日志搜索分析技术, 可以筛选出有用的数据, 用于判断学生的学习行为模式。

采集互联网数据的工具还有很多, 如网络爬虫、移动 App 等, 而且这样的工具还在不断地发展中。网络爬虫是一个自动下载网页的计算机程序或自动化脚本, 是搜索引擎的重要组成部分。网络爬虫类产品如八爪鱼采集器、网络矿工采集器等, 在数据采集领域有着广泛的应用, 可以定期实时采集各大门户网站的数据。近年来, 随着移动终端和通信技术的发展, 移动 App 技术逐渐成为移动过程中数据采集的主导技术, 采集方式更加灵活、多样。例如, 学生可以通过无线网络, 使用移动终端与云端学习平台进行互动。结合移动终端的定位技术, 利用传感器、视频监控等设备, 通过网络平台实时采集学习者的学习地点、学习时间、学习内容 & 学习状态等数据, 可以让教师实时了解学生的学习情况, 进而实现个性化智能辅导。

数据的保护

数据在采集、存储、管理与使用的过程中面临诸多安全风险。大数据时代, 人们对数据的依赖性不断增强, 数据安全和隐私保护问题更加突出。

1. 数据备份

数据在传输、存储、交换的过程中会面临导致丢失或损坏的各种风险因素, 如自然灾害、信息攻击、设备故障、误操作等。为避免风险, 通常需要进行数据备份。数据备份是周期性地数据以某种方式制作一个或多个备份, 并将其存放在专门设备上加以保护, 以便在数据丢失或损坏时能够有效地进行数据恢复。

个人数据的备份主要通过文件的复制完成, 由用户对重要的数据文件在不同的存储介质上归档保存。

企业数据是企业的重要资产, 如果缺失数据备份措施, 数据的安全性就得不到保障, 可能导致数据丢失或损坏, 对企业产生无法弥补的损失, 甚至带来灾难性后果。企业要制定数据备份策略, 明确数据备份内容、数据备份时间和数据备份方式等。企业的信息管理系统一般都包含数据备份的功能, 以自动、全面、高效地在服务器上进行数据备份。

2. 数据的隐私保护

在大数据的背景下, 人们在互联网上的一言一行都会被自动记录: 在网上阅读电子图书, 阅读习惯会被记录; 在网上聊天, 与好友的联络情况会被记录; 在网上购物, 购物喜好会被记录; 发送电子邮件, 联络方式会被记录; 在网上搜索, 搜索习惯会被记录……多项案例说明, 即使看似无害的数据, 被大量采集分析后, 也会暴露个人隐私。

不少网络企业既是数据的生产者,又是数据的存储者、管理者 and 使用者,如果对用户数据的采集、存储、管理与使用等缺乏规范和监管,用户就无法确保自己隐私数据的安全。

每个人都应有权决定自己的数据如何被利用,决定自己的数据何时以何种形式披露,或者何时被销毁。

数据的隐私保护需要从立法、技术、管理等多方面给予保障。对个人来说,需要不断地提高和加强自身的隐私保护意识和防范能力。如在使用某一种系统和服务时,要考虑对方要求自己提供的数据是否与服务相关;在被要求提供身份证号码、电话号码等相关敏感数据时,要考虑对方是否正规机构;在一些社交媒体公布自己的生活照时,要注意是否涉及敏感数据。

为对数据隐私做好保护,一些技术便应运而生。数据隐私保护技术包括:数据采集时的隐私保护,如数据精度处理;数据共享、发布时的隐私保护,如数据的匿名处理、人工加扰等;数据分析时的隐私保护;数据生命周期的隐私保护;隐私数据的可信销毁。

数据的组织和存储

数据以文件或数据库的形式永久存储在外存储器中。按照数据的组织和编码方式,文件可以分为不同的类型,可以由文件的后缀名加以区分。数据库按照特定的数据结构来组织、存储和管理数据,它相当于建立在计算机存储设备上的仓库。数据库有很多种类型,从最简单的存储各种数据的表格,到能够进行海量数据存储的大型数据库系统,都有着十分广泛的应用。

结构化数据通常存储在关系型数据库或表格文件中。关系型数据库是现代信息系统中最流行的一种数据存储结构。关系型数据库中的关系也称表(table),一个关系型数据库由若干个二维表组成。非结构化数据(unstructured data)主要以多媒体格式文件存储,例如各种格式的视频文件、音频文件、图像文件、文本文件等。

随着大数据时代的到来,数据量急速增长。为了满足大数据的海量存储、快速查询、安全兼容的需求,一些新型的非关系型数据库应运而生,以应对大规模数据集合和多重数据种类带来的挑战,尤其是解决一些大数据应用难题。

表格数据的加工方法

数据分析中最常遇见的数据是表格数据,表格数据的加工方法主要包括数据的计算、排序、筛选和分类汇总。

1. 数据的计算

(1) 数值数据

数值数据一般由阿拉伯数字、小数点和正负号构成,一些数据处理软件还提供分数、百分比、货币、科学记数法等表示形式,其写法如表 2-4 所示。

数值数据的计算包括算术运算和关系运算。算术运算包括加法(+)、减法(-)、乘法(*)、除法(/)、乘方(^)。关系运算也称为比较运算,包括等于(=)、大于(>)、小于(<),

表 2-4 数值数据表示形式示例

正数	负数	小数	分数	百分比	货币	科学计数法
1233	-534	23.5	2/3	78%	¥12.00	1.21212E+11
123	-9012	0.909	10 5/7	78.00%	\$566.00	2.31E-11
213.534	-90.32	212.64	2 1/11	0.90%	€ 34.00	3.23E+11

大于等于(>=)、小于等于(<=)、不等于(<>)。关系运算的结果为 TRUE 或者 FALSE，属于逻辑数据。

常见的数值数据函数主要有：用于数值统计的 SUM（求和）、COUNT（计数）、AVERAGE（求平均值）、MAX（求最大值）、MIN（求最小值）等；用于计算的 SQRT（求平方根）、MOD（求余数）、POWER（求乘幂）等；以及三角函数、数值舍入取整函数、随机数函数等。

例如，区域 A24:A132 中存放着一组实数，求它们总和的公式为 SUM(A24:A132)，求它们平均值的公式为 AVERAGE(A24:A132)，求它们的最大值和最小值之差的公式为 MAX(A24:A132)-MIN(A24:A132)。

（2）文本数据

文本数据一般是字母、汉字等字符，但也可以是完全由数字构成的文本数据，例如邮政编码、身份证号码、工号等。

文本数据一般只有连接(&)运算，即将两个操作数据连接在一起，构成新的文本数据。例如，如果 B2 单元格存储的数据是“沈萧”，那么公式 B2&“同学”，就是将 B2 单元格的数据与文本数据“同学”连接，得到“沈萧同学”。

文本数据函数用于对文本数据进行操作，主要有求字符串长的 LEN，求子串的 LEFT（从左边截取字符串）、RIGHT（从右边截取字符串）、MID（截取指定子串），查找字符串的 FIND（返回字符串的位置），删除空格的 TRIM 等。

例如，要取出身份证号码中表示性别的第 17 位字符，可使用 MID 函数，截取源字符串中从指定位置开始的指定长度的子串。

（3）日期 / 时间数据

日期 / 时间数据包括日期和时间两部分，输入方式为 YYYY-MM-DD HH:MM:SS。时间和日期的显示方式非常丰富，不同地区、不同应用场合都有所不同。日期 / 时间数据是通过数值编码的，日期以 1 天为单位编码，时间以 $\frac{1}{24 \times 60 \times 60}$ 天（1 秒）为单位编码。系统规定 1900/1/1 是第一天，编码为 1，其他的日期按照递增方式编码，所以日期 / 时间数据支持有效范围内的加减运算。例如，求两个日期相距的天数可以用减法完成，求一个日期后的第 n 天是哪一天可以用加法完成。

常见的日期 / 时间数据函数有 TODAY（返回当天日期）、NOW（返回当前日期和时间）、DATEDIF（返回两日期间相差的实足年数、月数和天数）、DATE（构造一个日期）、TIME（构造一个时间）等，还有从日期中提取年、月、日信息的函数——YEAR、MONTH、DAY，以及从时间中提取小时、分、秒信息的函数——HOUR、MINUTE、SECOND。

日期 / 时间数据在公式中不能直接书写，要用 DATE 函数构造，用法为：DATE(year, month, day)，返回的是日期数据。例如，求现在距离 2029 年国庆节还有多少天的公式为：DATE(2029,10,1) – TODAY()。

（4）逻辑数据

逻辑数据只有两个：TRUE（真）、FALSE（假）。逻辑数据没有运算，但是关系运算和逻辑函数都会产生逻辑数据。

最典型的逻辑数据函数是 IF 函数，它可以根据条件是否满足返回不同的结果。例如，IF(A1>60,"合格","不合格")。

IF 语句支持多个条件判断，可以解决复合判断问题。例如，身高体重指数 BMI（BMI= 体重 ÷ 身高²）可以从一个方面反映人的健康情况，如表 2-5 所示。

表 2-5 BMI 指数

BMI 指数范围	评价
BMI <18.5	体重轻
18.5<=BMI<24	健康
24<=BMI<28	超重
BMI>=28	肥胖

可以设计如表 2-6 所示的数据表，并计算出对每组数据的评价。

表 2-6 体检表

序号	体检号	身高(米)	体重(千克)	BMI	评价
1	10100510228	1.56	63.1	25.93	超重
2	10100720214	1.60	48.7	19.02	健康

例如，序号 1 的评价由嵌套的 IF 公式给出（假设序号 1 的“BMI”单元格为 E2）：
IF(E2<18.5,"体重轻",IF(E2<24,"健康",IF(E2<28,"超重","肥胖")))

2. 数据的排序

排序工具可以对表格中的一列或多列数据按指定顺序重新显示。排序有助于快速地组织和查找所需的数据，也有助于更好地理解数据，是数据处理中不可缺少的常见操作。

排序有升序和降序两种基本方式。所谓升序，就是从小到大排列数据，降序则正好相反。数值数据按数值的大小排序；文本数据按 ASCII 码值的大小排序；逻辑数据的 FALSE 相当于 0，TRUE 相当于 1；汉字有两种排序方式，一是按拼音的字典顺序排序，二是按笔画的多少逐字排序。

3. 数据的筛选

当需要从表格中找出满足一定条件的几行或几列数据时，就需要用到数据筛选功能。数据筛选仅仅是将不符合条件的数据隐藏起来，只显示那些满足条件的数据。筛选条件通常是针对某一列进行设定。筛选可以累加，进行多列筛选时，后一次的筛选是在前一次的基础上完成的。

筛选条件的具体设定方法随着数据类型不同而不同，见图 2-46。数值数据的筛选可针对数据的值域范围、平均值、最大值、最小值等来设定；文本数据的筛选可针对所包含的字符或字符串实现模糊查找；日期/时间数据的筛选支持按年、月、日分级选择，筛选条件可根据日期的大小和范围来设定；逻辑数据只有两个值，只需选择相应的值进行筛选。



图 2-46 筛选条件的设定示例

4. 数据的分类汇总

表格数据可以按照不同的类别进行汇总统计，汇总统计包括求总和、计数、求平均值、取最大值、取最小值、求偏差、求方差等。

数据处理软件通常会提供两种汇总工具，一种是单方向分类汇总，另一种是交叉分类汇总。

(1) 单方向分类汇总

如图 2-47 所示，可以对表格“某地 2013 年天气数据”进行分类汇总，分类字段为“天气类别”，汇总各类天气的天数。

日期	最高气温	最低气温	风向	风力	天气类别
2013/1/1	7		1 北风	3-4级转4-5级	晴
2013/1/2	3		1 西北风	4-5级	阴
2013/1/3	3	-2	北风	4-5级转3-4级	雪
2013/1/4	5		0 北风	3-4级	雪
2013/1/5	6		2 北风	3-4级	雪
2013/1/6	5		1 西北风	3-4级	雪
2013/1/7	6		1 西风	3-4级	雪
2013/1/8	6		2 东风	4-5级	阴
2013/1/9	6		0 西北风	3-4级	晴
2013/1/10	9		1 北风	3-4级	晴
2013/1/11	10		2 西风	3-4级	阴
2013/1/12	8		4 西北风	3-4级转4-5级	中雨
2013/1/13	10		3 西北风	3-4级	多云

图 2-47 某地 2013 年天气数据(部分)

首先将表格中的数据按分类字段排序,如图 2-48 所示,然后利用“分类汇总”工具对汇总方式、汇总项等进行设定,如图 2-49 所示。



图 2-48 按分类字段排序

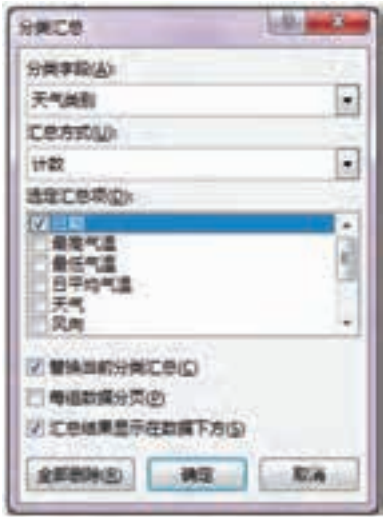


图 2-49 单方向分类汇总设置

设定完成后,每一个部分的结束行后面会出现汇总行,汇总出每个天气类别的计数,如图 2-50 所示。

	1	日期	最高气温	最低气温	风向	风力	天气类别
+	126	124					晴 计数
+	194	67					多云 计数
+	254	59					阴 计数
+	361	106					雨 计数
+	371	9					雪 计数
-	372	365					总计数

图 2-50 单方向分类汇总结果

单方向分类汇总支持多级分类汇总,例如,可以对表格“某地 2013 年天气数据”先按“天气类别”再按“风向”进行多级汇总,汇总前需要将表格先按主关键字“天气类别”再按次关键字“风向”进行排序。

(2) 交叉分类汇总

数据处理软件提供的交叉分类汇总工具通常是数据透视表。数据透视表是一种功能强大、操作简单的数据分析工具。进行交叉分类汇总时，选择表格中不同的行列组合，可以得到不同的统计数据。

表 2-7 学生档案数据表

学号	年级	班级	姓名	性别	出生日期	中考成绩

以表 2-7 所示的某高中的学生档案数据表为例，按图 2-51 的行列字段组合进行设置，得到的数据透视表是统计各年级各班的男生人数和女生人数，行汇总可以得到每个班的人数，列汇总可以得到全校的男生人数、女生人数和全校人数。若在行标签中删去“班级”字段，则得到的统计数据是各年级的男生人数和女生人数。



图 2-51 交叉分类汇总设置

拓展阅读

“手机导航 + 智能停车”服务

身在都市的有车族一般都有被寻找停车位、缴纳停车费困扰的经历。某地图 App 厂家，瞄准用户的这一痛点，与两家智能停车企业开展合作，为用户提供一体化的停车场电子支付服务。用户在手机地图 App 的导航下到达目的地时，该 App 会给用户推荐周边的智能停车场，并引导用户进行实时停车费查询和手机支付停车费等智能停车服务。

通过手机寻找停车场、缴纳停车费，有两个便利：一是可以通过导航到达停车场，这对不熟悉道路的车主很有用；二是可以直接在线完成支付，省去找零钱的麻烦。在手机地图 App 中，进入“发现周边服务”，可以在“车主服务”中找到“停车场”功能。“停车场”功能可以显示附近停车场的地图和列表，选择其中一个停车场便可以得到该停车场的地理位置。车主进入停车场后，可以直接在手机地图 App 里完成“找车位—进场—在线支付—快速出场”一系列动作。此外，车主还可以实时查询停车账单和预存停车费。

许多移动互联网用户已经习惯将手机地图作为生活服务的入口，充分使用“位置服务 + 生活服务”带来的便利。“手机导航 + 智能停车”将会为众多车主提供智能、便利的“行 + 停”无缝出行体验。

单元挑战 采集与分析气象数据

一、项目任务

气象数据是反映天气的一组数据，气象站采集的地面气象观测数据如温度、湿度、气压和风力、风向等，是气象数据的重要组成部分。随着技术的发展，人工气象站正逐渐被自动气象站所取代（图 2-52）。

查找资料，了解并比较人工气象站和自动气象站采集气象数据的方法和过程。重点关注自动气象站是如何自动采集数据的，借助了哪些技术手段或工具。

上网采集当地或某地近几年 7 月和 8 月的历史气温数据，并对气温数据进行统计和分析，了解近几年该地区 7 月和 8 月高温天气的走势。



图 2-52 自动气象站

二、项目指引

1. 以小组为单位，查找资料或访问气象站，了解常见的地面气象观测数据及其人工采集方法和自动采集方法。

2. 每个小组确定一个城市，上网查找该城市近几年 7 月和 8 月的历史气温数据，并选择一种电子表格软件设计、创建表格，输入数据。

3. 制作合适的图表，对比不同年份 7 月和 8 月气温的走势。

4. 使用函数计算或数据分析工具等方法，参考高温预警信号的等级分类定义（图 2-53），统计该城市近几年 7 月和 8 月各级高温天气的天数，进行对比分析，并选择适当的工具可视化数据。

5. 根据以上分析数据，了解该城市 7 月和 8 月的高温天气呈现怎样的趋势，思考背后的原因，并撰写报告。



图 2-53 高温预警信号

三、交流评价与反思

各小组派出代表，用自己熟悉的信息表达工具（如演示文稿等）制作电子作品，通过网络或课堂，展示交流小组的报告，并对其他小组的报告进行评价。

单元小结

一、主要内容梳理



二、单元练习

1. 某学校拟举办校园歌手大赛，有 20 名选手报名参加。比赛时共有 6 名评委评分（0~10 分），每名选手的最终成绩为 6 位评委所评分数的平均分。

（1）设计选手成绩管理表，并使用电子表格软件创建表，然后将一些模拟数据输入表中。

（2）利用电子表格软件计算各选手的最终成绩，并按成绩从大到小排序。

2. 为了解全班学生的身体健康状况和每周锻炼情况，请以小组为单位，采用多种方式和工具采集数据，并对数据进行分析。

（1）利用多种数据采集方式和工具，获取全班学生的身体健康数据和每周锻炼情况的数据。

（2）对采集到的数据进行加工，分析班级学生的身体健康状况与每周锻炼情况。

（3）利用可视化的方式展示分析结果，撰写分析报告。

3. 教材配套资源中的“订单表”记录了某网上书店过去一年的订单数据。尝试利用电子表格软件对该订单表进行分析。

（1）筛选出过去一年中销售量最高的 10 本图书。

（2）分析一年中不同月份图书销售量的变化。

三、单元评价

评价内容	达成情况
了解数据处理的概念和过程(A、T)	
能够认识数据处理的作用和应用价值(A、R)	
了解数据采集的方法和工具(A、T)	
能够根据实际情况,选择合适的数据采集方法(A、T)	
了解数据分析的概念和基本方法(A、T)	
知道数据的组织方式(A、T)	
知道数据的存储方式(A、T)	
能够根据给定的任务,使用并设计二维表进行数据存储(A、T)	
能够掌握通过公式、函数进行数据加工的方法(T)	
能够根据给定的任务需求,选用恰当的软件工具或平台处理数据,完成分析报告(A、T、I)	
能够使用合适的可视化方法,表示数据的含义(A、T、I)	
能够根据完成的数据分析报告,读懂数据背后隐藏的信息(A、T、I、R)	
能够理解对数据进行保护的意义(A、R)	

说明: A—信息意识, T—计算思维, I—数字化学习与创新, R—信息社会责任

第三单元

算法和程序设计

信息社会里，计算机及由计算机控制的智能系统几乎渗透到了人们工作、生活的所有领域。课堂上，教师使用智能软件实时收集、分析学生的学习情况；工厂里，流水线上的机械手自动执行着加工、装配等任务；仓库里，仓储机器人有条不紊地分拣、运送着货物；医院里，病人在挂号机上自助挂号、缴费，医生则用计算机辅助诊断；道路上，司机通过导航了解路况，交警借助智能交通监控系统了解瞬息万变的实时交通状况并及时疏导拥堵；住宅里，扫地机器人自动清洁着地面，人们用计算机上网冲浪……至于“神舟”上天、“蛟龙”号入海探测等尖端技术项目，更是离不开计算机的支持。

计算机何以如此神奇？这主要归功于算法及实现算法的计算机程序。在本单元中，我们将带领大家逐步通过生活中的实例认识算法，学会描述算法的方法，并尝试使用程序设计语言实现简单算法，编程解决问题。



学习目标

- ◆ 理解算法的概念和特征，能运用恰当的描述方法和控制结构表示简单算法。
- ◆ 了解程序设计语言产生与发展的过程及不同种类程序设计语言的特点。
- ◆ 掌握一种程序设计语言的基本知识，能使用程序设计语言实现简单算法。
- ◆ 体验程序设计的基本流程，掌握程序调试与运行的方法。

单元挑战

探究密码安全问题

项目五

描述洗衣机的洗衣流程

——了解算法及其基本控制结构

洗衣服是生活中常见的家务活动。对许多人而言，洗衣服是一种繁重的重复性体力劳动。洗衣机的发明将人类从洗衣劳动中解放了出来，它被誉为历史上 100 个最伟大的发明之一。自 1858 年问世以来，洗衣机逐渐发展到半自动洗衣机和全自动洗衣机（图 3-1），且正在向智能化洗衣机的方向发展。

洗衣机模拟人洗衣的过程，自动执行洗衣流程，节省了大量的人力，洗衣效率非常高。这主要归功于由算法控制的机器设备。



图 3-1 模拟手工洗衣流程的全自动洗衣机

项目学习目标

在本项目中，我们将通过对洗衣机洗衣流程的观察，了解什么是算法，以及算法应具有哪些特征，同时掌握算法的三种描述方法，并学会用流程图描述三种基本控制结构。

完成本项目学习，须回答以下问题：

1. 什么是算法？算法有哪些特征？
2. 算法有哪些描述方法？这些描述方法分别有哪些优缺点？
3. 算法的三种基本控制结构是什么？如何用流程图来表示？

项目学习指引

1. 从洗衣流程认识算法

常见的手工洗衣流程可以用如图 3-2 所示的“故事板”描述。

步骤一	步骤二	步骤三	步骤四
在盛有衣服的盆中加入适量的水和洗衣粉，搅拌均匀。	搓揉衣服。	用清水漂洗。	将衣服从盆里拿起，拧干。
			

图 3-2 手工洗衣流程“故事板”

以上“故事板”中，这个有序的、能够完成洗衣任务的流程，可以称为“手工洗衣算法”。

算法中的每一个步骤能被人或机器等计算装置执行。对计算装置而言，算法通常是以一步接一步的方式来详细描述如何将输入或者预定初始值转化为所要求的输出的过程，或者说，算法是对计算装置中执行的计算过程的具体描述。

全自动洗衣机是一种能够自动执行固定洗衣流程的电器，其中嵌入了**单片机**——一块集成了微型计算机功能的芯片。嵌有单片机的设备在生活中非常多见，例如微波炉、电梯、红绿灯等。

洗衣机的洗衣流程实际上是对手工洗衣流程的模拟，是对“洗衣算法”的实现。以上“故事板”中描述的“手工洗衣算法”是无法直接让机器完成的。机器能够实现的算法，必须具有一定的特征。例如，算法中的每一个步骤必须有明确的定义。

以某型号的洗衣机为例，它的“快洗”模式包括一次“洗涤”和一次“漂洗”操作，洗衣流程如图 3-3 所示。

核心概念

算法 (algorithm)：在有限步骤内解决问题所使用的方法。

小贴士

单片机全称单片微型计算机，又称微控制器，是采用超大规模集成电路技术把计算机中的 CPU、内存等重要部件集成到一块硅片上而构成的一个小而完善的微型计算机系统，在工业控制和智能设备中应用十分广泛。

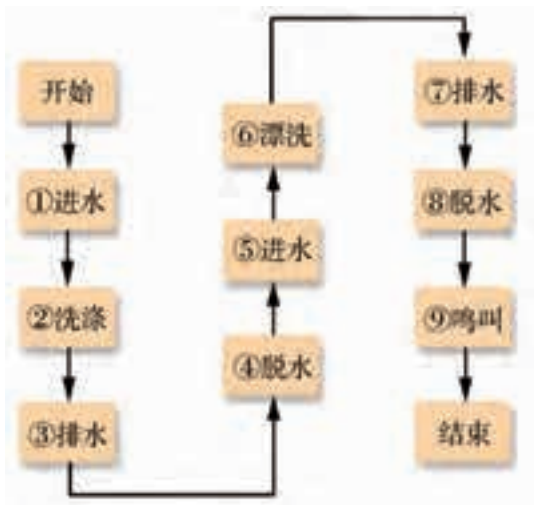


图 3-3 洗衣机“快洗”模式的洗衣流程

核心概念

算法的特征：一个算法必须具有有穷性、确定性、有零个或多个输入、有一个或多个输出、可行性这五个特征。

算法是解决“做什么”和“怎么做”的具体步骤的描述。**算法的特征**主要包括有穷性、确定性等。所以上图中的流程还不能称为洗衣机的“洗衣算法”，因为其每一步没有定义洗衣机可以明确执行的细节。以第②步“洗涤”为例，洗衣机的电动机带动波轮（或桶）旋转，产生水流搅动衣物。通过电动机正转、停、反转、停的反复循环，完成洗涤任务。为这一步骤单独设计的“洗涤算法”可以用自然语言描述如下：

- 2.1 电动机正转 30 秒；
- 2.2 停 2 秒；
- 2.3 电动机反转 30 秒；
- 2.4 停 2 秒。

这样，“洗涤算法”中的每一步，洗衣机都能够理解和正确执行，并且在有限的时间内结束。

此外，一个算法有零个或多个输入，以刻画运算对象的初始情况，所谓零个输入是指算法本身设定了初始条件。如对整个“快洗”模式的算法来说，进水的时间、洗涤的次数都有初始设定，不需要输入。一个算法有一个或多个输出，以反映算法执行的结果。“快洗”模式的算法结束时，鸣叫就是一种输出，宣告算法运行正确结束。

算法确定之后，需要将其“转告”算法的执行者，也就是编写洗衣机可以理解执行的机器指令。这些机器指令在洗衣机出厂前就提前写入了单片机的存储器中。

参见 P74 知识链接“算法”和“算法的特征及评价”

活 动

5.1 医院挂号是病人就诊的第一道手续。为了缓解挂号窗口排长队的现象，越来越多的医院引进了自助挂号机。病人在自助挂号机上挂号非常方便，选用本人有效证件，投入钱币或刷卡付款，就可以直接选择专科、专家、专号。请尝试使用“故事板”画出医院“自助挂号算法”，再用自然语言进行描述。注意在算法描述的过程中，考虑如何体现算法的特征。

2. 描述“洗涤算法”

上面对“洗涤算法”采用了自然语言的描述方式。自然语言是人们日常使用的语言，这种描述方法对于人来说非常容易理解。

除自然语言外，也可以采用框图和箭头组成的流程图来描述“洗涤算法”，如图 3-4 所示。流程图是最常用的算法的描述方法。流程图的主要特点是清晰直观。但如果要用流程图来描述整个“快洗”模式的算法，图就有些庞大。

还有一种可以用来描述算法的方法是伪代码。伪代码可以较容易地由开发人员转换为能被机器执行的指令。“洗涤算法”的伪代码描述如下。

```
i=0
WHILE i < 5
    电动机正转 30 秒
    停 2 秒
    电动机反转 30 秒
    停 2 秒
    i=i+1
```

思考与讨论??

洗衣机的“洗涤算法”是对手工洗衣中搓揉衣服的模拟，它们有何异同？

小贴士

算法的描述方法有自然语言、流程图和伪代码等。

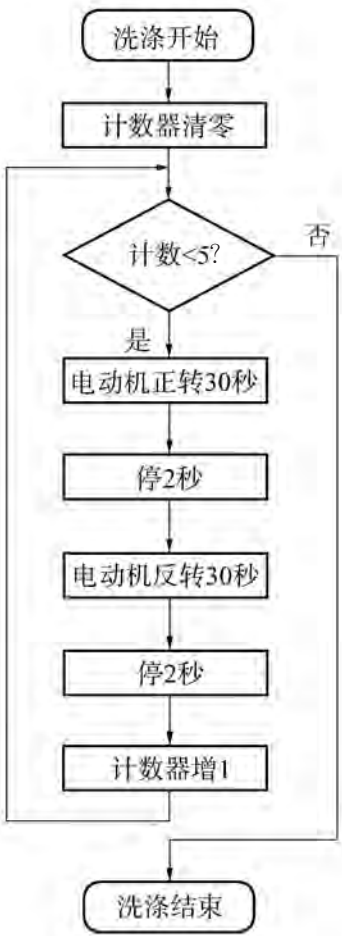


图 3-4 “洗涤算法”的流程图

← 参见 P75 知识链接“算法的描述方法”

活 动

5.2 分别使用流程图和伪代码，来描述医院“自助挂号算法”，比较不同算法描述方法的优缺点。

小贴士

算法的基本控制结构包括顺序结构(sequence structure)、选择结构(selection structure)和循环结构(loop structure)。

参见 P77 知识链接“算法的基本控制结构”

3. 分析洗衣流程的控制结构

不难发现，洗衣机中的不同洗衣模式大多是围绕浸泡、进水、洗涤、漂洗、排水、脱水这几个步骤组合而成的，每个步骤所用的时间和操作细节有所不同。

“快洗”模式的洗衣流程一共有 9 个步骤，洗衣机从第①步“进水”顺次执行到第⑨步“鸣叫”，这种自上而下、顺序执行的算法的基本控制结构称为顺序结构。

除了不同的洗衣模式，一些全自动洗衣机还能对不同种类衣物预设不同的洗衣流程，这些预设的洗衣流程的不同之处主要体现在洗涤次数、洗涤时长和水位等方面。例如，有的洗衣机可以按照选定的衣物种类，进行相应时长的洗涤（图 3-5）。如果是“牛仔裤”，则洗涤 12 分钟；如果是“丝织物”，则洗涤 8 分钟；否则，洗涤 10 分钟。这种先判断某些条件，再根据判断的结果来控制算法流程的结构称为选择结构。

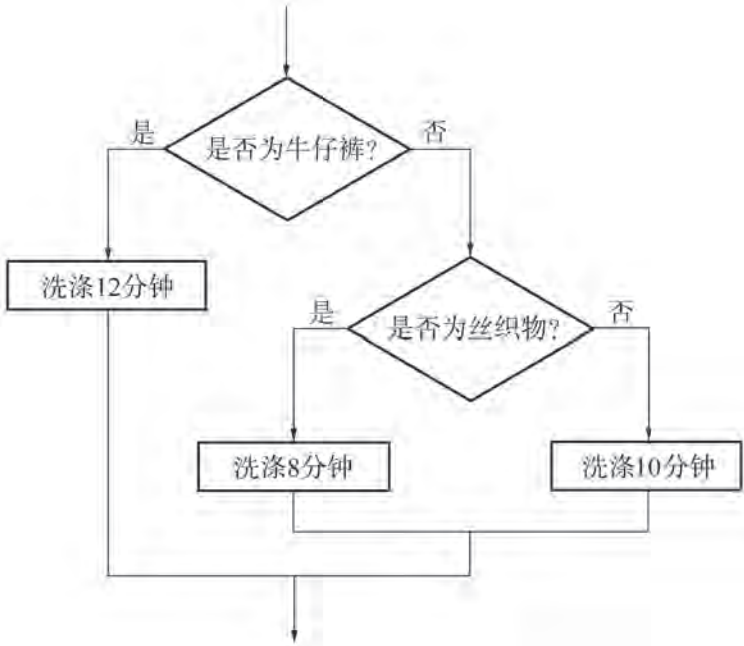


图 3-5 设定洗涤时长的流程图

循环结构是一种描述重复执行某段算法的控制结构，它可以减少算法步骤重复书写的工作量。例如，“洗涤算法”中

的 2.1~2.4 要反复执行 5 次,那么可以使用一个计数器计数 5 次,来控制它们的重复执行。

实际上,无论哪种洗衣模式,都可以用上述顺序结构、选择结构、循环结构这三种基本结构或其组合画出流程图。

思考与讨论??

是否有可能存在第四种控制结构?为什么?

数字化学习

观看配套资源中的“ATM 取款实例演示”动画及“取款算法”的流程图,结合自己的生活经验,描述图书馆借书算法或烧水泡茶算法等,并选用适当的数字化工具进行交流。

活 动

5.3 全自动洗衣机通常会提供多种洗衣模式,如图 3-6 所示。请选择其中一个洗衣模式,观察其洗衣流程,并尝试绘制相应流程图。

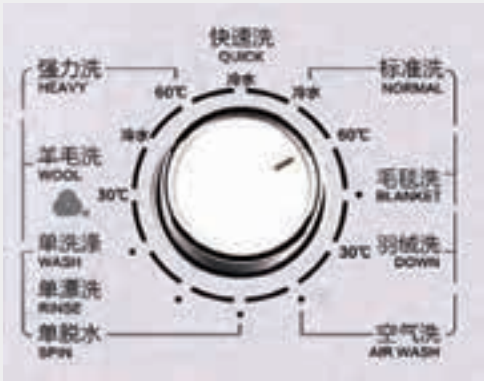


图 3-6 选择洗衣模式

5.4 非洲棋 Mancala 是一种古老的棋类游戏(图 3-7),游戏目的是想办法赢得较多的棋子。上网搜索这个游戏的主要游戏规则,并使用流程图将游戏步骤表示出来。



图 3-7 非洲棋 Mancala 的棋盘



知识链接

算法

通俗地说,算法是在有限步骤内解决某一问题所使用的方法。算法在生活中普遍存在。算法中的每一个步骤能被人或机器等计算装置执行。随着计算机的产生和发展,很多复杂的算法思想有机会变成了现实。例如,银行自动提款机可以模拟银行柜台人员的工作,用户只需要按照各种提示进行操作,即可实现自动提款。又如,工业机器人可以自动执行生产流水线上的各种操作步骤,而且动作更精确,既提高了生产质量,又降低了成本。此外,把下棋的规则与方法借助计算机来实现,就可以让计算机与人对弈。

算法的特征及评价

1. 算法的特征

算法是解决“做什么”和“怎么做”的具体步骤的描述,这些步骤应能够在有限时间内完成,并且不可以含有会导致步骤无法继续进行下去的歧义性语句。一个算法必须具有有穷性、确定性、有零个或多个输入、有一个或多个输出、可行性这五个特征。

(1) 有穷性

一个算法在执行有限步骤后在有限时间内能够实现,这就是算法的有穷性。也就是说,算法不能是无限循环的。例如,洗衣机不会无休止地洗涤下去,它最终总会完成洗涤任务。在现实中,任何算法的实现还需要考虑运行时间的合理性。所谓合理,取决于实际情况。例如,洗衣任务应在可接受的洗涤时间内结束,而不会让单个洗衣任务执行一整天。

(2) 确定性

一个算法中的每一个步骤的表述都应该是确定的、没有歧义的语句。例如,“快洗”模式的算法中对“洗涤”步骤有明确的“电动机正转 30 秒”等时间设定,让洗衣机可以按照确定的时间来执行。

(3) 有零个或多个输入

输入就是算法在执行时从外界获得的数据,它们是算法执行的某些初始状态。输入可以是零个,也可以是多个。零个输入代表这个算法没有外部输入,其初始设定被隐藏在算法中。例如,“洗涤算法”中并不需要输入电动机正转、反转的次数,而是直接进行了初始设定。

(4) 有一个或多个输出

输出就是算法执行的结果,也就是按照算法对数据进行处理所得到的结果。没有输出的算法是没有意义的。算法的输出,可以是数值,可以是图形,还可以是控制信号。例如,启动蜂鸣器鸣叫告知用户洗衣任务已完成,就是全自动洗衣机洗衣算法的一种输出形式。

(5) 可行性

算法的可行性就是指每一个步骤都能够有效地执行,并且得到确定的结果,同时能够

用来方便地解决一类问题。可行性意味着算法可以被机器按部就班地执行,并最终得到正确的结果。

2. 算法的评价

评价一个算法的好坏,主要考虑以下几个指标。

(1) 正确性

算法首先必须是正确的,即对于任意一组输入,包括合理的输入与不合理的输入,总能得到预期的输出。

(2) 可读性

可读性是指一个算法可供人们阅读的难易程度。算法应当是可读的和可以理解的,人们能方便地对它进行分析、修改和使用,让它实现。

(3) 高效性

好的算法还应该具备时间效率高和存储量低的特点。时间效率是针对算法的执行时间而言的。对于同一个问题,如果有多个算法能够解决,那么执行时间短的算法时间效率高,执行时间长的算法时间效率低。存储量指的是算法在执行过程中需要的最大存储空间,主要是指所占用的内存或外部硬盘存储空间。一台可执行算法的机器,其存储空间受到物理上的限制,并不是无穷尽的,所以算法的存储量要低。

算法的描述方法

算法是解决问题的方法,是对问题解决过程的精确描述,所以需要有恰当的方式对算法进行描述。算法的描述方法很多,主要有自然语言、流程图和伪代码三种。

1. 自然语言

自然语言是人们日常生活中使用的语言。用自然语言描述的算法通俗易懂,但是描述用词要做到准确、统一,无二义性,且算法的控制结构和嵌套层次要清晰明确。

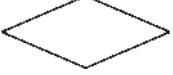
例如,有些洗衣机会根据投入衣物的重量来自动设定水位,其算法用自然语言可描述如下:

如果衣物重量属于范围 1,则设定为第一挡水量;
否则,如果衣物重量属于范围 2,则设定为第二挡水量;
否则,如果衣物重量属于范围 3,则设定为第三挡水量;
否则,报警提示衣物超重。

2. 流程图

流程图由一些简单的框图组成,框图用来表示算法的步骤。流程图所使用的符号遵循国家颁布的标准,常用的流程图符号、名称及作用见表 3-1。

表 3-1 流程图符号及其作用

符号	名称	作用
	起止框	表示一个算法的开始或结束。
	处理框	表示要处理的内容，该框有一个入口和一个出口。
	输入 / 输出框	表示数据的输入或结果的输出。
	判断框	表示条件判断的情况。满足条件，执行一条路径；不满足条件，执行另外一条路径。
	连接框	用于连接因画不下而断开的流程线。
	流程线	指出流程控制方向。

用流程图描述算法，形象直观、清晰简洁，算法结构表达明确。只是当控制结构和嵌套层次较复杂时，对应流程图所占篇幅会比较大，影响可读性，也不易于修改。

例如，洗衣机根据投入衣物的重量来自动设定水位，其算法流程图如图 3-8 所示：

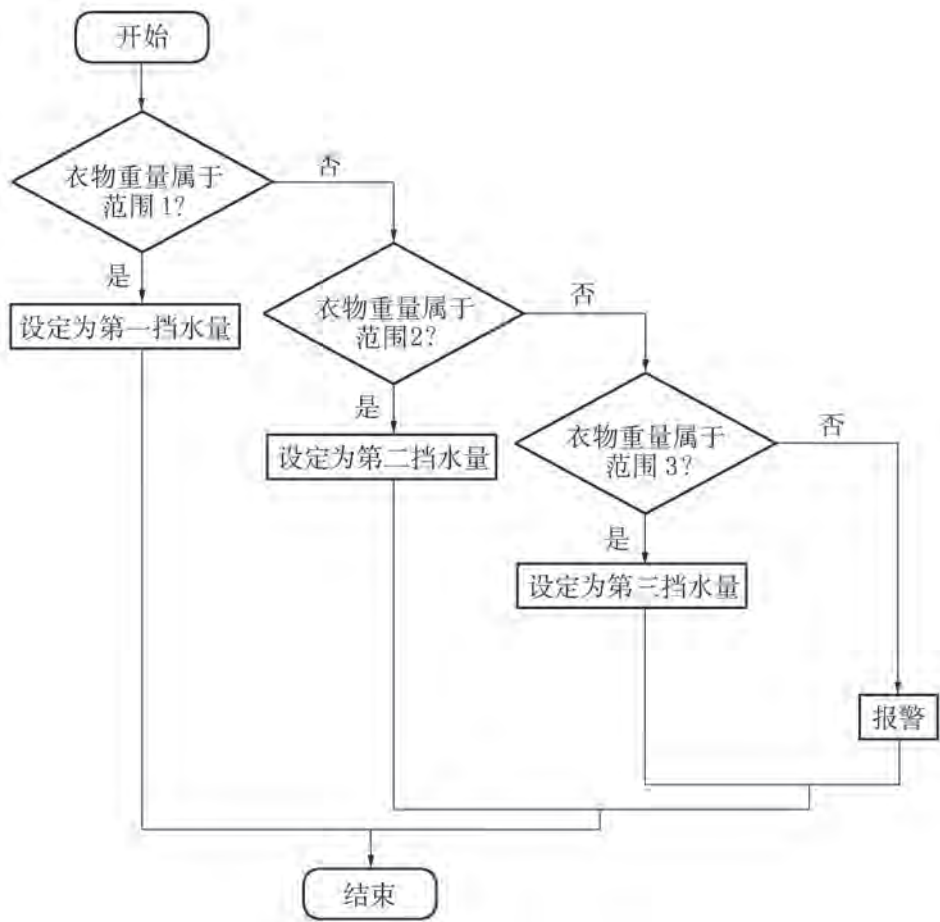


图 3-8 自动设定水位的流程图

3. 伪代码

伪代码是一种算法描述语言，它是用介于自然语言和程序设计语言之间的人工语言来描述算法。伪代码书写方便、格式紧凑、表达精炼，比较容易被开发人员理解。

例如，洗衣机根据投入衣物的重量来自动设定水位，其算法用伪代码可描述如下：

```
IF 衣物重量 IN 范围 1
    设定 水位 = 第一挡水量
ELSEIF 衣物重量 IN 范围 2
    设定 水位 = 第二挡水量
ELSEIF 衣物重量 IN 范围 3
    设定 水位 = 第三挡水量
ELSE 报警提示衣物超重
END IF
```

算法的基本控制结构

算法的基本控制结构包括顺序结构、选择结构和循环结构。任何一个算法，都可以由若干个基本结构或其组合构成。构造一个具体算法的时候，只要遵守这三种基本结构的规范，就可以画出任何算法的流程图。

1. 顺序结构

顺序结构是最为简单的一种基本结构，就是由上至下，按先后顺序依次执行（图 3-9）。

2. 选择结构

选择结构也称为分支结构，是根据给定的条件进行判断而作出选择的一种结构（图 3-10）。选择结构的流程图中必定包括一个判断框，满足条件时执行一个处理框，不满足条件时执行另一个处理框。

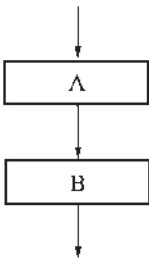


图 3-9 顺序结构

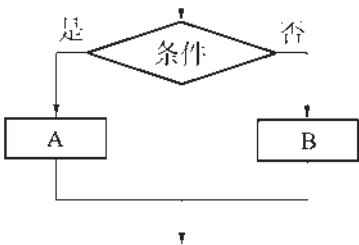


图 3-10 选择结构

3. 循环结构

循环结构是描述重复执行操作的控制结构，它解决了重复操作多次或按条件进行控制的通用性问题。循环结构充分利用了计算机运算速度快和自动执行的优点。典型的循环结构包括当循环和直到循环。

当循环采取先判断表达式、后执行循环体语句的方式(图 3-11)。当判断框中的条件为“是”时,执行循环结构中的循环体语句,如此往复;当判断框中的条件为“否”时,结束循环。

直到循环采用先执行循环体语句、后判断循环条件是否成立的方式(图 3-12)。当判断框中的条件为“否”时,返回重新执行循环体语句,直到判断框中的条件为“是”时结束循环。

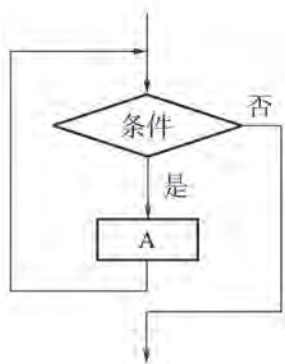


图 3-11 当循环

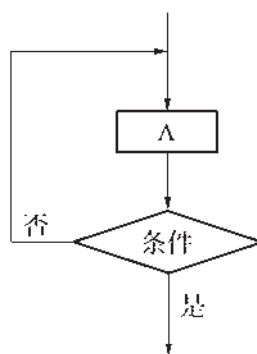


图 3-12 直到循环

解决温标转换问题

——认识程序和程序设计语言

温度是表示物体冷热程度的物理量。温度对自然环境、物理性质、人体都有着重要的影响。海明威的小说《一天的等待》中，小主人公，一位9岁的小男孩，听到医生说他的体温达到102度时，非常担心自己会死去，因为他得知人发烧到40度就要活不成了。这是怎么回事呢？

原来，生活中计量温度的常用标准有两种：华氏温标和摄氏温标。小男孩把它们搞混了。人们常常需要将华氏温标和摄氏温标相互转换，使用摄氏华氏温度计能够方便地解决这个问题（图3-13）。但是如果没有摄氏华氏温度计，就需要利用公式人工计算，或者执行一个温标转换程序了。用户输入任意一个华氏温标下的温度（华氏温度），计算机通过执行一个温标转换程序，就能立即输出这个温度在摄氏温标下的温度（摄氏温度），用户不需要知道具体的实现方法。



图 3-13 摄氏华氏温度计

项目学习目标

在本项目中，我们将使用一种程序设计语言，编程解决温标转换问题。

完成本项目学习，须回答以下问题：

1. 什么是程序？什么是程序设计语言？
2. 程序设计的一般过程是怎样的？
3. 程序的基本控制结构有哪几种？
4. 程序设计语言的产生与发展历史是怎样的？

项目学习指引

1. 体验程序设计的一般过程

(1) 分析问题

华氏温度和摄氏温度之间的换算公式是：摄氏温度 = $5 \times (\text{华氏温度} - 32) \div 9$ 。使用换算公式，可以列式计算一个特定华氏温度所对应的摄氏温度。

用程序解决这个问题过程是：首先接收用户输入的华氏温度，然后使用换算公式计算出相应的摄氏温度，最后输出摄氏温度的值。

本问题涉及两个数据：已知数据——华氏温度，所求的结果数据——摄氏温度。

(2) 设计算法，编写程序

首先需要确定算法，然后选择一种程序设计语言来编写程序。各种程序设计语言广泛应用于科学计算、数据处理、事务处理、系统软件、人工智能等领域。不同的程序设计语言，其适用领域和特点也不相同。目前较热门的程序设计语言有 Java、C/C++、C#、Python、PHP、VB 等。程序设计语言的发展很快，新的语言不断产生，且越来越接近人类的自然语言。图 3-14 给出了算法流程图，以及分别使用 Python 语言和 C 语言编写的程序。

核心概念

程序 (program)：一组能被计算机理解并执行的指令序列，是计算机执行算法的操作步骤。

程序设计语言 (programming language)：人们编制程序所使用的计算机语言。

参见 P85 和 P86 知识链接“程序及其特征”和“程序设计语言的产生和发展”

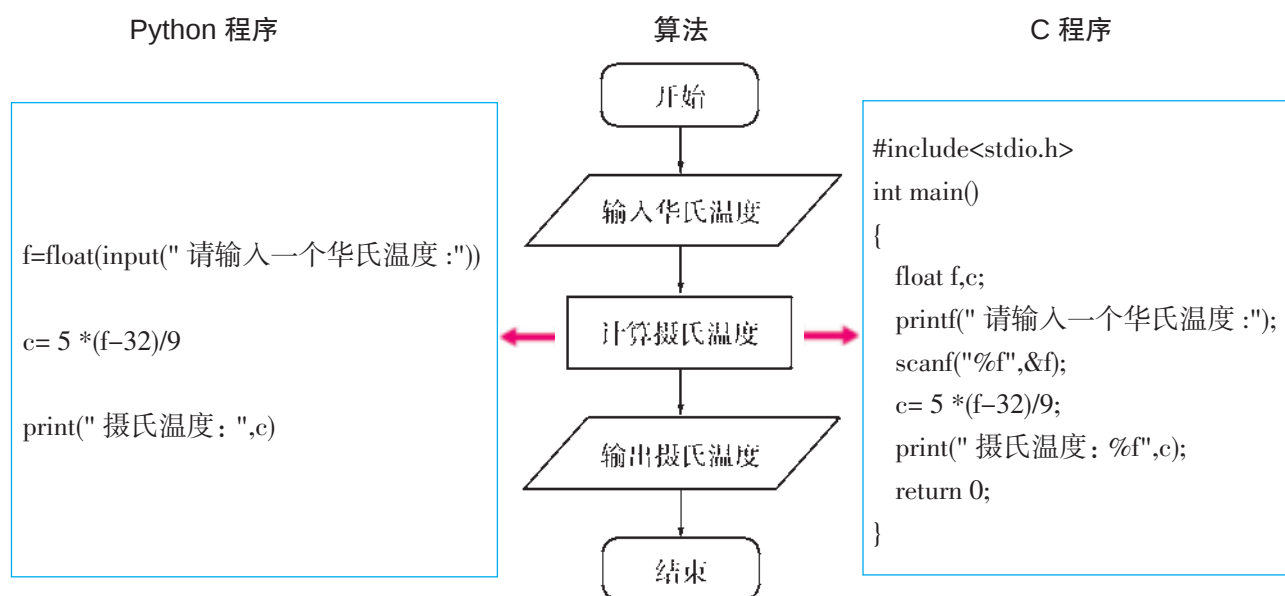


图 3-14 “温标转换”算法和程序间的对应关系

对比两种程序设计语言,可以发现 Python 程序的代码行数更少,更简洁。

程序是由语句构成的。Python 语言使用回车来表示一条语句的结束。程序中出现的 c 和 f 是变量, c 表示摄氏温度, f 表示华氏温度。

公式由算术表达式构成,其中包含常量、变量、运算符等要素。公式中的 “*” “-” “/” 是算术运算符,注意它们与数学中的符号在书写上的差异。

“=” 表示赋值运算,在 Python 语言中表示将 “=” 右边的数据对象绑定到其左边的变量,根据数据对象的类型确定变量的类型。

`input`、`float`、`print` 是 Python 语言提供的系统函数。`input` 函数的作用是接收一个文本数据,双引号内的字符串会显示在计算机屏幕上,作为输入操作的提示文本。华氏温度是浮点数据,所以用 `float` 函数将输入的文本(字符串数据)转化为浮点数据。`print` 函数负责输出结果,语句 `print("摄氏温度:",c)` 首先输出文本内容:“摄氏温度:”,然后输出变量 c 的值。`print` 输出后默认换行。

思考与讨论??

你还听说或使用过哪些程序设计语言? 用它们解决了哪些问题?

(3) 运行、调试程序

程序发生错误后,寻找产生错误的原因、位置,并改正错误的过程称为调试程序。

① 在 Python Shell 中逐条输入语句,观察变量的值和最终输出结果。

```
>>> f=float(input('请输入一个华氏温度:'))
请输入一个华氏温度:102
>>> f
102.0
>>>
```

② 在 Python 提供的 IDLE 窗口中利用菜单命令创建、编辑一个程序文件(图 3-15)。

核心概念

变量(variable): 引用在程序中可能会变化的数值,通过变量名可以访问存储在内存中的数据。

小贴士

Python 程序有两种运行模式:

- 交互模式。启动 Python,在 Python Shell 窗口的提示符 `>>>` 后,直接键入指令,Python 会立即执行该指令。

- 文件模式。打开 Python 的 IDLE 集成开发环境,可以创建(File)、编辑(Edit)程序文件,还可以检查语法错误(Check Module)、运行程序(Run Module)。程序文件以 “py” 为后缀名。

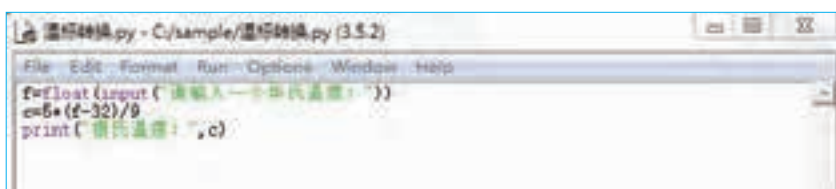


图 3-15 “温标转换”程序编辑窗口

参见 P88 知识链接“程序设计及其一般过程”

③ 检查语法错误，无误后运行程序（如图 3-16 所示），查看运行结果是否正确。



图 3-16 “运行程序”菜单

数字化学习

参考配套资源或上网学习，了解 Python 更多知识。

运行、调试程序时常会出现两类错误：一类是语法错误，即程序的书写格式不正确，不符合语言的语法要求，对于这类错误，语言系统能够自动检查并报错、指出出错位置；另一类是逻辑错误，即程序可以运行，但是在运行的过程中会出问题，或程序的结果不符合预期设计。

思考与讨论??

1. 若将第一条语句改为：`f=input("请输入一个华氏温度: ")`，运行程序，会发生什么？为什么？
2. 若将第二条语句改为：`c=5*(f-32)/9`，运行程序，会发生什么？为什么？

活动

6.1 尝试使用 Python 编写程序，计算并打印华氏 100 度到 105 度所对应的摄氏温度。

- (1) 分析问题，确定已知数据、所求数据和解决方法。
- (2) 设计算法，并用自然语言或流程图描述算法。
- (3) 创建并编辑程序。
- (4) 运行、调试程序，观察运行结果，分析程序有无错误。若有，尝试解决。

华氏温度	摄氏温度
100	37.78
101	38.33
102	38.89
103	39.44
104	40.00
105	40.56

2. 了解程序的基本控制结构

运行下面两个“打印华氏、摄氏温度对应表”程序(程序1和程序2),可以得到相同的运行结果。对比两个程序,可以发现循环结构的程序2表述明显更简洁。

```
print("    华氏    摄氏 ")
f=100
c=5*(f-32)/9
print("%8.0f%10.2f"%(f,c))
f=101
c=5*(f-32)/9
print("%8.0f%10.2f"%(f,c))
f=102
c=5*(f-32)/9
print("%8.0f%10.2f"%(f,c))
f=103
c=5*(f-32)/9
print("%8.0f%10.2f"%(f,c))
f=104
c=5*(f-32)/9
print("%8.0f%10.2f"%(f,c))
f=105
c=5*(f-32)/9
print("%8.0f%10.2f"%(f,c))
```

程序 1 (顺序结构)

```
print("    华氏    摄氏 ")
for f in range(100,106):
    c=5*(f-32)/9
    print("%8.0f%10.2f"%(f,c))
```

程序 2 (循环结构)

程序1中,第二至第四条语句的作用是设定f的值,计算c的值,输出f和c的值。重复使用这三条语句6次,每次只需修改f的值即可。

程序2实现的是一个计数循环算法,即对程序1中的第三至第四条语句重复执行6次。f的值从100变化到105,每次增加1,作用相当于一个计数器。

输出语句中的格式设定为%10.2f,f表示float浮点数,10表示总的输出宽度为10,数据长度不足在左边补充空格,.2表示小数点后保留2位,四舍五入。

range(start,stop,step)函数可以产生一个整数序列,从start开始,到小于stop的最大整数结束,间隔为step。step省略值为1。因此range(100,106)产生序列[100,101,102,103,104,105]。

← 参见 P88 知识链接“程序设计语言的功能”

小贴士

print 语句可以使用格式控制符将变量的值按一定的格式输出,使表格数据对齐。格式为:

格式控制串 "%(值序列)"
双引号内是格式控制串,格式控制串中%引出的是格式控制符。值序列中的值以逗号分隔,可以是变量的值,也可以是表达式的值。

for 语句用于实现循环结构。for 语句的一般形式如下:

for 变量 in 序列:
语句块

变量引用序列中每一个值,每个变量值执行一次语句块。

活动

6.2 编写程序，打印指定范围的华氏、摄氏温度对应表，间隔为 10，其中上限和下限由用户输入。

(1) 用两个变量分别表示下限和上限，增加输入语句获取下限和上限的值。(注：输入任意整数可用语句 `int(input())`。)

(2) 修改 for 语句。

```
for f in range(_____, _____, _____):
```

```
    c=5*(f-32)/9
```

```
    print("%8.0f%10.2f"%(f,c))
```

(3) 运行、调试程序，观察运行结果，分析程序有无错误。若有，尝试解决。

(4) 对比本程序和程序 2，讨论变量和输入语句在提高程序通用性方面所起的作用。

输入下限 =0	
输入上限 =101	
华氏温度	摄氏温度
0	-17.78
10	-12.22
20	-6.67
30	-1.11
40	4.44
50	10.00
60	15.56
70	21.11
80	26.67
90	32.22
100	37.78

思考与讨论??

1. 在程序 2 的算法中，哪一个变量是循环控制变量？它的初值、终值是多少？它是如何有规律地变化的？
2. 如果打印一张华氏、摄氏温度对应表，其中华氏温度从 50 变化到 100（含 100），间隔 10，该如何修改程序 2？循环体中的语句将被重复执行几次？

小贴士

if 语句用于实现选择结构，其一般形式如下：

```
if 条件表达式:
```

```
    语句块 1
```

```
else:
```

```
    语句块 2
```

如果满足条件表达式则执行语句块 1，否则执行语句块 2。

3. 优化程序，判断输入有效性

运行活动 6.2 的程序时，如果输入如下数据：

```
输入下限 =100
输入上限 =0
```

运行结果会是打印一张只有表头的空表。这是一种程序的逻辑错误，即程序没能控制不合理的输入数据。

解决这一问题，需要用到选择结构，即对输入的上限和下限进行检查，判断输入是否有效。如果输入有效，就打印温度对应表；如果输入无效，则给出相关提示，结束程序的运行。算法流程图（图 3-17）和程序如下页所示。

活动

6.3 体温计的合理范围是华氏 90 度到 110 度。修改程序 3，打印供医务人员使用的指定范围的华氏、摄氏温度对应表。若输入的上下限超出人体温度的合理范围，则给予提示。

- (1) 查阅 if...elif...else 语句的使用方法，完成程序修改。
- (2) 设计几组输入数据，运行程序，检查程序是否无误。

运行数据 1：有效数据；运行数据 2：l>=h；运行数据 3：l 小于 90 度；运行数据 4：h 大于 90 度；运行数据 5：边界值数据 (l=90 或 h=110)。

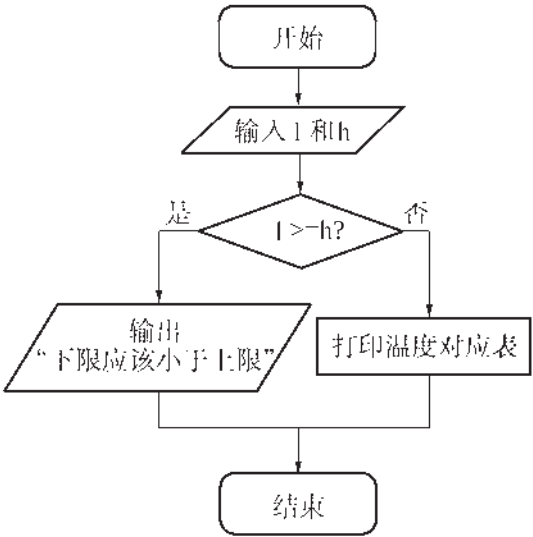


图 3-17 打印温度对应表的流程图

```
l=int(input(" 输入下限 ="))
h=int(input(" 输入上限 ="))
if l >= h:
    print(" 输入的下限应该小于上限 ")
else:
    print(" 华氏 摄氏 ")
    for f in range(l,h):
        c=5*(f-32)/9
        print("%8.0f%10.2f"%(f,c))
```

程序 3 (选择结构)

知识链接

程序及其特征

程序实际上是一组操作指令或语句序列，是计算机执行算法的一系列操作步骤。计算机是通过运行该组指令完成预期任务的。

一般来说，程序具有以下一些特征：

- (1) 解决问题的算法的每一个步骤对应程序中的一个或多个语句，每个语句实现一个或多个操作。
- (2) 在程序中被施行操作的对象称为数据，不同类型的数据有不同的属性、取值范围和运算方法。
- (3) 除了遇到特殊的程序转向控制语句之外，程序都是从第一条语句开始顺序执行。有时语句要求执行时作出判断：在某种情况下执行一条或一组语句，否则执行另一条或

另一组语句。有时一条或一组语句可能需要执行一次以上，此时必须给出重复的次数或终止条件。

程序设计语言的产生和发展

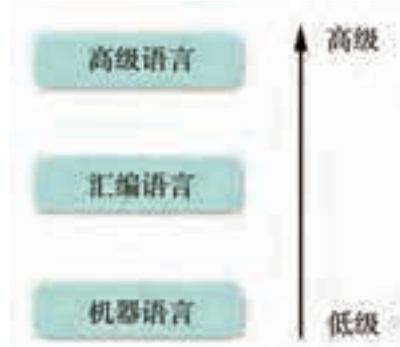


图 3-18 程序设计语言的发展

程序要能被计算机理解和执行，必须精确描述执行所有动作的细节过程，不能有一点错误和含糊之处，承担这一描述任务的就是程序设计语言。程序设计语言是人们编制程序所使用的计算机语言。随着计算机的发展，程序设计语言经历了从机器语言到高级语言的发展历程，如图 3-18 所示。

1. 机器语言

计算机主要由电子元器件组成的电路构成。由于电子元器件的特性，计算机只能识别二进制的代码。早期的程序设计语言就是由二进制代码指令组表示的，称为机器语言 (machine language)。通常，不同的计算机，其指令系统会有所不同。每一条机器指令一般包含两个主要部分：操作码 (规定了指令的功能) 和操作数 (规定了被操作的对象)。有的指令没有操作数，如停止指令。在这些指令的控制下，计算机可以实现最基本的算术运算和逻辑运算。

以 5+12 运算为例，若某型号计算机指令采用 16 位二进制表示，则用机器语言编写的一个完成 5+12 运算的程序如表 3-2 所示：

表 3-2 用机器语言编写的程序 (5+12)

指令序号	机器指令	指令说明	指令功能
1	1011 0000 0000 0101	操作码：表示数据传输功能。 左操作数 (4 位) 表示 0 号寄存器，右操作数 (4 位) 是 5。	把加数 5 送到 0 号寄存器中。
2	0000 0100 0000 1100	操作码：表示加法功能。 左操作数表示 0 号寄存器，右操作数是 12。	把 0 号寄存器中的内容与另一数相加，结果存在 0 号寄存器中 (即完成 5+12 的运算)。
3	1111 0100	停止功能	停止操作

说明：计算机的运算器中有多种用于存放参与运算的操作数的部件，如运算器执行加法运算时，必须有两个分别存放加数与被加数的寄存器。

机器语言与计算机相关，不同型号的计算机有着不同的指令集，由不同的硬件组成。使用机器语言编写的程序可以直接执行，执行速度快，但是在不同型号的计算机上不能通用。

2. 汇编语言

用机器语言编写程序，对大多数使用者来说非常难。20 世纪 50 年代中期，程序开发人员开始采用一种类似英语缩略词并带有助记符号的语言，替代复杂的二进制代码指令和操作数来编写程序，这就是汇编语言 (assembly language)。仍以上述完成 5+12 运算的程序为例，改用汇编语言编写的程序如表 3-3 所示：

表 3-3 用汇编语言编写的程序 (5+12)

指令序号	汇编语言指令	指令功能
1	MOV AL, 5	把加数 5 送到累加器 AL 中。
2	ADD AL, 12	把累加器 AL 中的内容与另一数相加, 结果存在累加器 AL 中 (即完成 5+12 的运算)。
3	HLT	停止操作

说明：累加器是运算器中的一种寄存器，用于存放计算结果。

用汇编语言编写的程序，让程序开发人员可以方便地辨识指令。但是计算机不能理解这种程序，需要使用翻译程序将每条语句翻译成机器语言，计算机才能执行，如图 3-19 所示。这种翻译程序被称为汇编程序，也叫汇编器。



图 3-19 汇编过程

3. 高级语言

汇编语言仍然是与计算机相关的语言，特定的计算机指令集对应特定的汇编语言。使用汇编语言编写的程序不可移植。20 世纪 60 年代中期，接近于人类自然语言的高级语言 (high-level language) 问世。用高级语言编写的程序更加简洁易懂。

仍以上述运算为例，用 Python 语言编写的程序语句如下：

5+12

和汇编语言一样，用高级语言编写的程序也不能直接被计算机理解，必须经过转换才能被执行。高级语言按转换方式可分为解释类和编译类两类。

• 解释类：执行方式类似于日常生活中的“同声翻译”，应用程序源代码由相应语言的解释器“翻译”成目标代码 (机器语言)，边“翻译”边执行，如图 3-20 所示。由于不能生成可独立执行的文件，应用程序不能脱离其解释器。这类语言有 Python、BASIC 等。其中 Python 语言的解释器就是 Python Shell，它可以逐条计算表达式的值并执行命令。

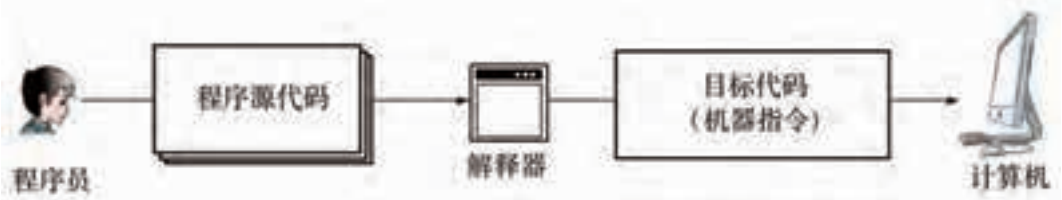


图 3-20 逐条解释执行

• 编译类：编译 (compile) 是指在源程序执行之前，就将程序源代码通过编译器一次性地“翻译”成目标代码 (机器语言) 文件，如图 3-21 所示。因此其目标程序可以脱离语言环境，独立执行。

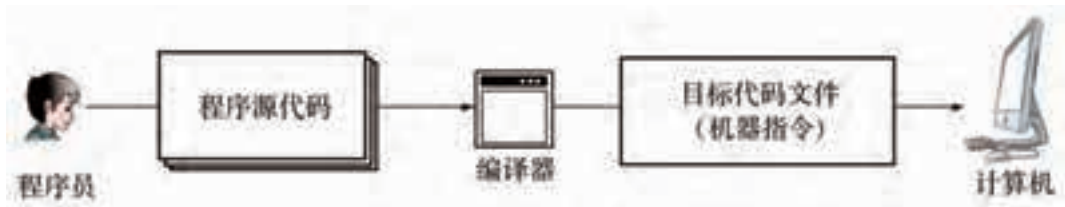


图 3-21 成批编译执行

高级语言与机器语言之间是一对多的关系，即一条高级语言语句对应多条机器语言语句。高级语言独立于计算机，用其编写程序无须顾及与计算机相关的实现细节，具体工作由与不同计算机相匹配的解释器和编译器完成。

综上所述，机器语言是特定计算机使用的语言，它能够直接操控硬件，运行速度快，占用内存少。计算机可以执行的就是使用机器语言编写的程序。汇编语言比较接近机器语言，实际应用于底层硬件操作和高要求的程序优化场合，如嵌入式操作系统、工业控制等。高级语言自诞生至今，经历了多次更新换代，新的语言和更高效的程序设计方法不断涌现，使得程序的开发过程更简单、更有效率，便于应对快速软件开发的要求，为计算机的应用开辟了更广阔的前景。

程序设计及其一般过程

计算机的所有操作都是按照人们预先编制好的程序进行的。因此，若需要运用计算机帮助解决问题，就必须把具体问题转化为计算机可以执行的程序。在问题提出之后，从分析问题、设计算法、编写程序，一直到运行调试程序的整个过程就称为程序设计（programming）。在程序设计过程中，分析问题、设计算法、编写程序和运行调试是一个不断反复的过程。本项目中，通过对“温标转换”算法的实现，可以看出程序设计的基本流程如图 3-22 所示。

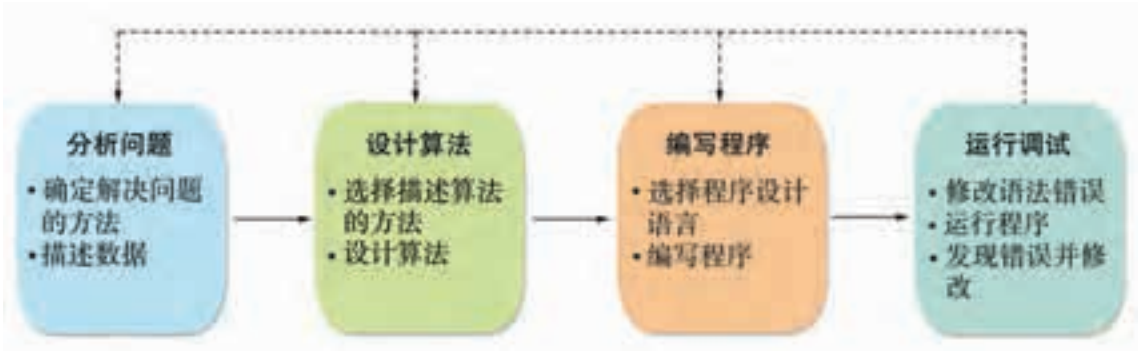


图 3-22 程序设计的基本流程

程序设计语言的功能

程序设计语言是人与计算机交流的语言，人需要使用程序设计语言告诉计算机需要处理的数据和处理数据的流程，所以程序设计语言必须具有数据表达和流程控制的能力（图 3-23）。以 Python 语言为例，它是一门面向对象的程序设计语言，它用对象类型描述数据，使用三种基本控制结构和函数等提供数据处理的控制手段。



图 3-23 程序设计语言的功能

1. 数据表达

(1) 数据对象和类型

现实生活中的数据是多种多样的，程序设计语言首先要将纷繁复杂的数据对象抽象为若干种数据类型。数据类型定义了该类型数据的属性和范围是什么，可以通过该类型的对象做什么操作或运算。例如，整型所包含的数据对象是 {...,-3,-2,-1,0,1,2,3,...}，而 +、-、*、/ 是作用于整型类型上的运算。

在程序设计语言中，一般都有好几种定义数据类型的语句，供程序员直接使用。Python 语言的数据类型如图 3-24 所示，主要分为基本数据类型和组合数据类型。基本数据类型主要是数值型（包括整型和浮点型）、布尔类型和其他语言中不多见的复数类型。组合数据类型可以应用于一次处理多个对象的场合，包括字符串、元组、列表、集合类型和字典类型。

基本数据类型	组合数据类型
• 整型 int • 浮点型 float • 布尔类型 bool • 复数类型 complex	• 字符串 str • 元组 tuple • 列表 list • 集合类型 set • 字典类型 dict

图 3-24 Python 语言的数据类型

(2) 常量和变量

程序中的数据有两种表示方式：常量和变量。

常量（constant）是直接给定的、在程序中不可以改变的数值。常量区分为不同的数据类型，可以从书写形式上区分，例如整数 389，浮点数 23.56，字符串 "hello"。

变量描述的是存储空间的概念，用于引用在程序中可能会变化的数值。数据存储在内存中，要引用存储在内存中的数据，需要用一个名称来访问内存空间，这个名称就称为变量名。在程序运行的过程中，变量对数据的引用是可以发生变化的。例如，可以通过赋值语句改变变量的引用对象。

在 Python 语言中，变量的数据类型由引用的数据对象的类型决定，如下所示。

```
>>> x=12.3
>>> type(x)
<class 'float'>
>>> x='red'
>>> type(x)
<class 'str'>
```

为了增加程序的可读性，编写程序时变量的名称应尽量体现数据的意义。Python 语言中变量的命名规则为：由大小写英文字母、数字或下划线组成，以英文字母或下划线为首字符，长度任意，大小写敏感，但不能与 Python 的关键字（又称保留字，是指语言里事先定义的、有特别意义的标识符）同名。

在交互模式下，可以使用 `help("keywords")` 查阅 Python 的关键字。

2. 流程控制

程序设计语言除了要能表达各种各样的数据之外，还必须提供表达数据处理过程的手段，即程序的流程控制。流程控制包括语句级的程序基本控制结构，以及单位级的函数定义和调用。

（1）程序的基本控制结构

程序有三种最基本的控制结构：顺序结构、选择结构和循环结构。每一个程序都是由若干个基本结构组成的，一个基本结构中可以包含一条或若干条语句。

除了程序转向控制语句之外，顺序结构描述的程序都是从第一条语句开始顺序执行，到最后一条语句结束。例如，“温标转换”程序一共有三条语句，依次按顺序执行，这就是一个顺序结构。常见的顺序结构语句有输入语句、输出语句和赋值语句。

选择结构会根据对条件判断的结果来执行不同的语句块。在本项目的程序 3 中，通过检验条件表达式 $l \geq h$ 成立与否来决定是输出提示“输入错误”还是打印温度对应表。Python 语言使用 `if` 语句来实现选择结构。

循环结构的作用是控制一段程序语句重复多次地执行。本项目的程序 2 中，通过控制华氏温度的取值范围，重复执行循环体中的两条语句。Python 语言用于实现循环结构的语句有 `for` 语句和 `while` 语句。

（2）模块和函数

当要解决的问题比较复杂时，一种典型的程序设计方法是将复杂的程序划分为若干个功能相对独立的部分，再分而治之。这可以降低程序的设计难度，提高程序的可读性和可维护性。Python 语言提供了函数来实现这些独立部分的功能，函数既可以是系统函数也可以是用户自定义函数。

Python 语言的系统函数由标准库中的很多模块提供用户选择使用。标准库中的模块又分成内置模块和非内置模块。

内置模块 `builtin` 中的函数称为内置函数，可以直接使用。内置函数包含了使用频繁或最基本的操作，如数学函数 (`abs`, `round`)、类型转换函数 (`int`, `float`, `str`, `repr`)、集合操作函数 (`len`, `max`, `min`, `sum`)、输入输出操作函数 (`input`, `print`)。

函数的调用方式与数学上的函数类似，即函数名加上相应的参数值，多个参数值之间以逗号分隔：

< 函数名 > (参数值序列)

非内置模块要先导入模块再使用。Python 标准库中有 100 多个非内置模块。输入输出函数包含在 `sys` 模块中；通常的数学函数，如三角函数、对数函数等包含在数学模块中。

例如，计算 $\log_2 136$ 的方法如下。

```
>>> import math
>>> print(math.log(136, 2))
7.08746284125034
```

自定义函数是允许程序员在程序中自行定义的函数，它是完成独立功能的程序段。函数定义后可以在程序中调用，其使用方法和系统函数的使用方法相同。

Python 语言通过加载第三方开发的模块扩充类库，可以得到更为丰富的数据表达类型和相应处理数据的功能函数。这也使得 Python 语言迅速获得了广泛应用。

拓展阅读

Python 语言概述

Python 语言由荷兰人吉多·范罗苏姆 (Guido van Rossum) 在 1989 年发明，并于 1991 年公开发布第一个版本。

Python 语言是开源项目的优秀代表，其解释器的全部代码都是开源的，可以在 Python 语言的主网站上自由下载。这促使世界上出现了围绕 Python 程序设计的大型开源社区，那里提供了数量可观的开源函数库，为使用 Python 语言进行快速开发提供了强有力的支持。

Python 语言有着很多区别于其他程序设计语言的特点，其中最重要的有以下几点：

(1) 语法简洁：实现相同功能，Python 语言的代码行只相当于其他语言的 1/10~1/5。

(2) 与平台无关：程序无需修改，便能在安装了解释器的 Windows、Linux、Unix、MacOS 上跨平台使用。

(3) 类库丰富：Python 标准库中提供了几百个内置类和函数库。此外，全世界的程序员为开源社区贡献了十几万个第三方函数库，几乎覆盖了计算机技术的各个领域。

(4) 扩展性强：能够把其他语言 (如 C++、Java) 制作的模块很轻松地连接在一起。

(5) 多模式：Python 语言虽然是一门面向对象的语言，但却同时有面向过程和面向对象两种编程方式，为使用者提供了灵活的工作方式。

基于上述优点，Python 语言被广泛地应用于科学计算、数据处理、网站开发、图形处理、人工智能等领域。

项目七

用计算机计算圆周率

——设计简单数值数据算法

圆周率 π 是一个在数学及物理学中常用的数学常数，它是圆的周长与直径的比值，也等于圆的面积与半径平方之比。古今中外，一代代数学家为探求圆周率的值贡献出了自己的智慧和辛劳。魏晋时期的刘徽（图3-25）在公元263年提出了著名的割圆术（用圆的内接或外切正多边形来逼近圆的周长），得出 $\pi \approx 3.14$ 。南北朝时期的祖冲之进一步得出精确到小数点后7位的圆周率值，这一成就在世界上领先了一千多年。

计算机发明之后，人们开始借助计算机来计算圆周率，并为之设计了许多算法。随着计算机性能的快速提升，计算机计算出的圆周率精确位数不断刷新。



图 3-25 数学家刘徽

项目学习目标

在本项目中，我们将对用计算机计算圆周率一探究竟，利用 Python 语言编程实现两种计算圆周率的算法，进而探究程序的控制结构，并感受不同算法的效率。

完成本项目学习，须回答以下问题：

1. 数值数据的常用运算有哪些？
2. 如何设计并实现选择结构算法？
3. 如何设计并实现循环结构算法？

项目学习指引

1. 设计算法实现用数学公式计算

几千年来，经过不断探索，数学家们发现圆周率经常出现在各种数学公式中。1735 年，数学家欧拉得到了如下的公式：

$$\frac{\pi^2}{6} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots$$

以下我们就尝试用 Python 语言，根据上述欧拉发现的公式来计算圆周率。

(1) 分析问题

观察上面的数学公式可知，它是一项一项连续相加的，并且每一项都可由前一项经过有规律的变化得到。有规律的数据项连续相加的问题，称为累加问题，用计算机算法处理就是逐项累加。上述累加计算的算法可以用自然语言描述，具体如下。

- 1. 设置累加器变量 s 的初值为 0;
 - 2. $s=s+1/(1 \times 1)$;
 - 3. $s=s+1/(2 \times 2)$;
 -

← 参见 P99 知识链接“数值数据及其运算”

累加可以用一个公式来表示： $s=s+item$ 。这个式子称为累加的通项公式。对上述欧拉发现的公式来说，item 的值是 $1/i^2$ ，即 $1/(i \times i)$ ，i 从 1 开始，每次增加 1。解决本问题需要重复进行的工作是：

- ① 累加 item 到 s，即 $s=s+item$;
- ② 构造 item，即 $item=1/(i \times i)$ ， $i=i+1$ 。

完成公式的计算就是重复执行以上步骤，直到满足某一个循环控制条件。循环控制条件可以是计数执行 n 次，也可以是某一个变量达到指定值。先根据循环控制条件确定循环控制变量，进而确定循环控制变量的初值、终值和如何变化，如图 3-26 循环结构的流程图所示。

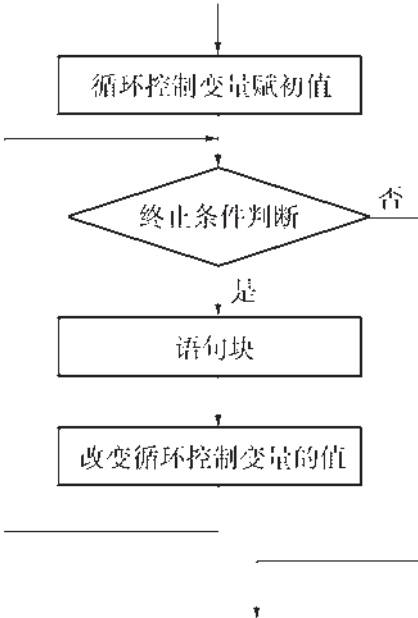


图 3-26 循环结构的流程图

思考与讨论??

按以下要求该如何分别控制循环的执行? 循环控制变量是哪一个? 初值、终值是怎样的? 如何变化?

- 累加 n 个 $item$;
- 圆周率的精确度达到小数点后 4 位。

参见 P99 和 P101 知识链接“循环结构的算法构建”和“实现循环结构的语句”

(2) 设计算法, 编写程序

如果以累加项 $item$ 的值小于 0.000 001 为循环终止条件, 那么用自然语言描述上述问题的算法如左下所示。在 Python 语言中, 利用 `while` 语句可以实现该循环结构的算法, 程序如右下所示。这里用 π 表示圆周率 π 。

算法

1. 令 $i=1$, $item=1.0$, $s=0$;
2. 循环当 $item$ 大于等于 0.000001;
 - 2.1 $s=s+item$;
 - 2.2 $i=i+1$;
 - 2.3 $item=1/(i \times i)$;
3. $\pi = \sqrt{6 \times s}$;
4. 输出 π 的值。

程序

```
from math import sqrt
i=1
s=0
item=1
while item>0.000001:
    s=s+item
    i=i+1
    item=1/(i*i)
pi=sqrt(6*s)
print("pi 的值是 ", pi)
```

小贴士

在 Python 语言中, 当只需引用外部模块中一个特定的系统函数时, 可以使用 `from` 命令。`from math import sqrt` 语句的作用是从 `math` 库中载入数学函数 `sqrt`, 这个函数用于求一个数的平方根。

思考与讨论??

1. 循环开始前, $item$ 的值是多少? 循环结束后, $item$ 的值是多少?
2. 如果将第 2.1 步调至循环体的最后一行, 该如何调整算法步骤来保证程序结果的正确? 请上机运行检验。
3. 如果按累加 n 项的方法控制循环, 该如何修改程序? 如何分别用 `while` 语句和 `for` 语句实现?

(3) 运行、调试程序

在 Python 中创建并运行程序, 检查运行结果。

活 动

7.1 使用沃利斯公式计算圆周率。

$$\frac{\pi}{2} = \frac{2}{1} \times \frac{2}{3} \times \frac{4}{3} \times \frac{4}{5} \times \frac{6}{5} \times \frac{6}{7} \times \frac{8}{7} \times \frac{8}{9} \times \dots$$

(1) 分析问题：

- 以阶乘运算 $5! = 1 \times 2 \times 3 \times 4 \times 5$ 为例，思考累乘问题的通项公式及累乘变量的初值。

- 观察公式，找出变化规律，确定 item 的构成，并构造由当前 item 计算下一项 item 的计算公式。

- 确定循环控制的方法。

(2) 写出用自然语言或流程图描述的算法。

(3) 根据循环控制的方法，选择 for 语句或 while 语句完成程序。

(4) 运行、调试程序。

2. 设计算法实现用随机投点法计算

随机投点法是通过在正方形内随机投点，用随机模拟的方法来估算圆周率。

随机在正方形里撒一把豆子，每一颗豆子落在正方形内任何一点的可能性都是相等的，落在每个区域中的豆子数与这个区域的面积近似成正比。落在每个区域中的豆子数是可以数出来的，设落在扇形内的豆子数为 hits，总豆子数为 darts，则：

$$\frac{\text{hits}}{\text{darts}} \approx \frac{S_{\text{扇}}}{S_{\text{正}}}$$

由圆周率等于圆的面积与半径平方之比可知：

$$\pi = \frac{S_{\text{圆}}}{R^2}$$

观察图 3-27 可知，单位圆的面积等于第一象限中扇形面积的 4 倍，单位圆半径的平方正好是图中单位正方形的面积。因此，

$$\begin{aligned} \pi &= 4 \times \frac{S_{\text{扇}}}{S_{\text{正}}} \\ &\approx 4 \times \frac{\text{hits}}{\text{darts}} \end{aligned}$$

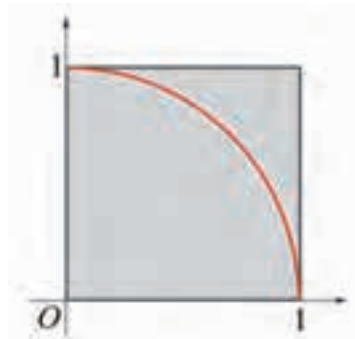


图 3-27

这样,就可以通过数出落在每个区域中的豆子数得到 π 的近似值。豆子撒得越多,结果就越精确。

(1) 分析问题

使用计算机模拟随机投豆(也就是投点)的整个过程如下。

首先,确定总的投点数 darts。

然后,在边长为 1 的正方形中投点,并判断这个点是否在扇形区域内,如果是,则扇形区域的投点数 hits 增加 1。(每个投点的 x 坐标和 y 坐标都是 [0,1) 中的随机数;若 $x^2+y^2<1$, 则判断该点投在扇形区域内。)

最后,在所有点投完后,计算圆周率的近似值 pi。

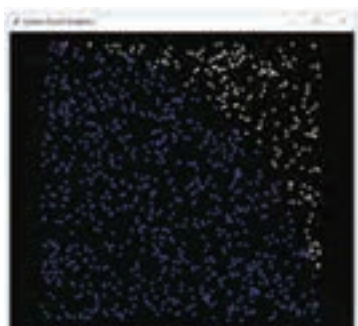
$$pi=4 \times \frac{hits}{darts}$$

这里涉及的数据变量有:总的投点数 darts, 每个点的坐标值 x、y, 扇形区域的投点数 hits, 圆周率的近似值 pi。

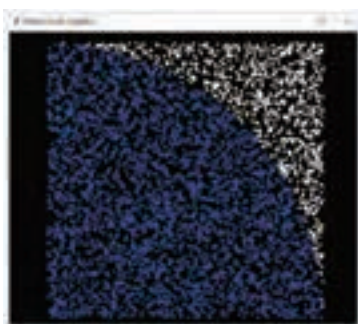
(2) 设计算法, 编写程序

用随机投点法计算圆周率的算法如下:

1. 输入总的投点数 darts;
2. 令 hits=0;
3. 循环 i 从 0 到 darts-1, step1;
 - 3.1 获取 x, y 的随机值;
 - 3.2 如果 $x^2+y^2<1$, 则 hits 增加 1;
4. 计算圆周率的近似值;
5. 输出圆周率的近似值。



投入 1000 个点



投入 5000 个点

图 3-28 利用随机投点法
计算圆周率时的投点图像

思考与讨论??

算法中哪几步在重复执行?

该算法的流程图(图 3-29)和用 Python 语言编写的程序见下页,请把它们补充完整。程序中用到了 Python 标准库 random 模块提供的系统函数 random, 它可随机产生一个 [0,1) 中的实数。

(3) 运行、调试程序

在 Python 中创建并运行程序,检查运行结果(图 3-28)。可以发现,当总投点数为 1000 时,pi 值的精确度还是很低的。通过增加总投点数,可以提高 pi 值的精确度。

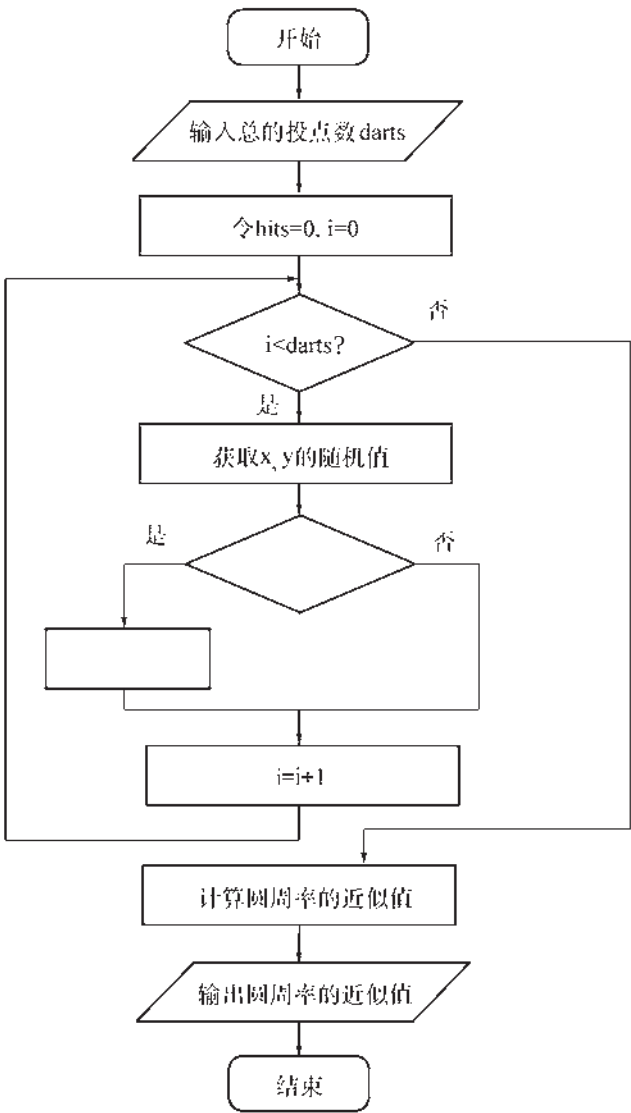


图 3-29 计算圆周率的流程图

← 参见 P102 和 P103 知识链接“条件表达式”和“实现选择结构的语句”

```
from random import random

darts= int(input(" 请输入总的投点数 :"))

hits=0

for i in range (darts):

    x, y = random(), random()

    if _____:

        _____

pi = 4 * (hits/darts)

print("pi 的值是 ",pi)
```

计算圆周率的程序

思考与讨论??

1. 分别输入 1000、5000、10 000，得到圆周率的近似值分别可以精确到第几位？

2. 多次输入同样的总投点数 1000，会得到不同的结果，这是为什么？如何减少这种不确定性？

圆周率 π 是一个无限不循环小数，无论是用数学公式计算，还是用随机投点法计算，本质上都是对 π 的精确值的无限逼近。选择合适的计算方法，并设计出高效的算法，正是用计算机解决类似问题的一般思路。

活 动

7.2 对比两种计算圆周率的算法。

（1）改写数学公式法程序，输入精确度要求，并输出运行时间（程序如下，其中#后文字为行注释）。多次运行程序，改变 limit 的值，得到不同的圆周率近似值，在表 3-4 中记录每次的运行结果。

```
from math import sqrt
from time import clock           # 引入 clock 函数
limit=float(input(" 输入精确度: "))
clock()                          # 启动计时器
i=1
s=0
item=1
while item>=limit:
    s=s+item
    i=i+1
    item=1/(i*i)
pi=sqrt(6*s)
print("pi 的值是 ", pi)
print(" 运行时间是 ", clock(), " 秒 ")  # 输出程序运行时间
```

表 3-4 程序运行结果记录表

limit	pi 的值	准确的小数位数	运行时间 (秒)
0.0001	3.1320765318091053	1	0.008484839644298251
0.00001			

（2）改写随机投点法程序，输出运行时间。多次运行程序，改变总投点数，得到不同的圆周率近似值，记录每次的运行结果。

（3）对比两种算法计算圆周率的运行情况，填写表 3-5。

表 3-5 两种算法的对比

比较项目 \ 算法	数学公式法		随机投点法	
计算结果的精确度	☐ 较高	☐ 较低	☐ 较高	☐ 较低
程序运行时间	☐ 较长	☐ 较短	☐ 较长	☐ 较短
对算法的理解	☐ 容易理解	☐ 不易理解	☐ 容易理解	☐ 不易理解



知识链接

数值数据及其运算

数值数据分为整数 (int) 和浮点数 (float)。浮点数是带小数的数值。例如, 1 是整数, 1.0 是浮点数。在计算机中, 两种数据的存储方式是不同的。

数值数据的运算符包括标准的算术符号, 如表 3-6 所示。其中 +、-、*、/ 运算符的含义和用法和数学符号是一样的。

表 3-6 常用的算术运算符

运算符	含义	示例	运算结果
+	加法	53+21	74
-	减法	63.1-2.2	60.9
*	乘法	300*1.5	450.0
/	除法	1/8	0.125
//	整除	1//8	0
**	次幂	2**3	8
%	求余数	17%3	2

运算符 // 执行的是整除运算, 它会产生一个小数部分为 0 的结果, 小数部分会被舍掉。运算符 ** 执行的是次幂运算, $a**b$ 等价于数学运算 a^b 。运算符 % 执行的是求余数运算。这 3 种运算符的用法举例如下。

```
>>> 4.8//1.2
4.0
>>> 5//1.2
4.0
>>> 7//2
3
>>> 3//4
0
```

```
>>> 2**3
8
>>> 16**0.5
4.0
>>> 1.6**2.5
3.238172324012421
```

```
>>> x=34
>>> x%2
0
>>> x%3
1
>>> x%17
0
>>> x%4
2
```

循环结构的算法构建

要构建一个循环结构, 需要解决两个问题: 一是重复要做的是什么事, 二是如何控制循环。

对重复要做的是什么事, 可以从分析问题入手解决。而对循环操作的控制, 一般是通过使用一个循环控制变量来实现的。实现循环控制有三个要素: ①循环控制变量的初值; ②循环控制变量的终值; ③在循环体中使循环控制变量由初值向终值变化的语句。

例 1: 求 $50!$ 。

• 分析问题

$50! = 1 \times 2 \times 3 \times \cdots \times 50$ ，和累加问题类似，计算机算法的处理是逐项将 $1, 2, 3, \cdots, 50$ 累乘到一个变量中。与累加问题不同的是，累乘变量的初值应为 1。

• 设计算法

算法一

1. 令 $s=1$;

2. $s=s \times 1$;

3. $s=s \times 2$;

.....

51. $s=s \times 50$ 。

观察算法一，可以发现每一步都是在做乘法。虽然乘的数据不同，但是这些数据的变化是有规律的，可以得到通项公式： $s=s \times i$ 。接下来要确定循环控制变量。求阶乘的循环控制是由通项公式的执行次数决定的，一共执行 50 次。此时循环控制变量相当于一个计数器，从 1 开始，到 50 结束，每一次递增 1。循环控制变量的变化和通项公式中的变量 i 的变化规律一样，所以将 i 作为循环控制变量。

算法二

1. 令 $s=1$;

2. 令 $i=1$;

3. 循环当 i 小于等于 50;

3.1 $s=s \times i$;

3.2 $i=i+1$ 。

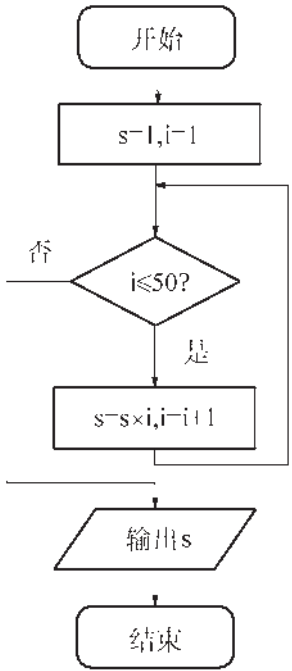


图 3-30 算法二流程图

值得注意的是，在循环体中一定要有一句使该循环能趋于结束的语句，如这里的 $i=i+1$ ，这样循环控制变量才能从初值变化到终值，最终结束循环。

利用循环结构解决问题时，也会遇到一些循环执行次数不确定的问题，如例 2。

例 2: 输入一个任意正整数，求它的逆序数。

• 分析问题

求一个正整数的逆序数的方法，是每次取该数的最后一位数字，构造出一个新的整数。例如，726 的逆序数的求解过程如下。

第一次	第二次	第三次	结束
• $726 \% 10 = 6$ • $0 * 10 + 6 = 6$ • $726 // 10 = 72$	• $72 \% 10 = 2$ • $6 * 10 + 2 = 62$ • $72 // 10 = 7$	• $7 \% 10 = 7$ • $62 * 10 + 7 = 627$ • $7 // 10 = 0$	

构造循环控制时,可以将表示输入数的变量作为循环控制变量。输入初始值后,每次整除 10,直到该数为 0,循环结束。重复执行的操作是将原数的最后一位取下,通过位权值构造到新的逆序数中。

- 设计算法

```

设输入的数为 num, 逆序数为 reverseNum。
1. 输入一个正整数到 num;
2. 令 reverseNum=0;
3. 循环当 num>0;
   3.1 reverseNum=reverseNum*10+num%10;
   3.2 num=num//10;
4. 输出 reverseNum。
  
```

在本例中,循环的次数在程序执行前是未知的,由数 num 是否为 0 决定循环是否停止执行。循环体中通过一个整除操作来改变 num 的值,使它不断向终值 0 变化,最终结束循环的运行。这里的循环控制变量 num 相当于一个标志量,根据它的值决定循环是否结束。

实现循环结构的语句

根据上面的分析,按循环执行次数的确定性划分,循环可以分为确定次数循环和不确定次数循环。

1. 确定次数循环

确定次数循环指循环结构对循环次数有明确的定义,循环次数采用遍历结构中元素的个数来体现。Python 语言中通过 for 语句来实现确定次数循环。

```

for < 循环控制变量 > in < 遍历结构 >:
    < 语句块 >
  
```

for 语句简化了循环控制变量计数的过程,用遍历结构提供循环控制变量的所有值的集合。每一次循环,循环控制变量依次取集合中的一个值。遍历结构可以是 range 函数产生的序列、组合数据对象(列表、字符串等)或文件对象。

用 for 语句实现求 50! 的语句块如下:

```

s=1
for i in range(1,51):
    s=s*i
print(s)
  
```

如果要将 1~50 之间所有的奇数累加,则可以用 range 函数的第三个参数设定循环控制变量变化的步长,其语句块如下:

```
s=0
for i in range(1,50,2):
    s=s+i
print(s)
```

2. 不确定次数循环

不确定次数循环会一直保持循环操作，直到特定循环条件不被满足时才结束，不需要提前知道循环次数。Python 语言中通过 while 语句来实现不确定次数循环：

```
while <条件>:
    <语句块>
```

用 while 语句实现例 2 求逆序数的语句块如下：

```
num=int(input(" 请输入一个数 :"))
reverseNum=0
while(num):
    reverseNum=reverseNum*10+num%10
    num=num//10
print(" 它的逆序数是 :",reverseNum)
```

while 语句也可以用于完成确定次数循环，如用它实现求解 50! 的语句块如下：

```
s=1
i=1
while i<=50:
    s=s*i
    i=i+1
print(s)
```

条件表达式

在选择结构中需要根据判断条件是否被满足来执行不同的分支语句，在循环结构中则需要根据判断条件是否被满足来决定是否继续执行循环体语句。程序中表示判断条件的式子称为条件表达式，它由关系运算或逻辑运算构成，执行的结果为布尔值 True 或 False。

1. 关系运算符

关系运算符也称为比较运算符，它的作用是对两个数据的值进行比较，比较的结果是一个布尔值。用关系运算符连接的表达式称为关系表达式。如果表达式成立，其值为 True；如果表达式不成立，其值为 False。Python 语言中的 6 种常用的关系运算符如表 3-7 所示。

表 3-7 常用的关系运算符

Python 运算符	数学符号	名称	示例 (dist=0.87)	结果
<	<	小于	dist<1	True
<=	≤	小于等于	dist<=1	True
>	>	大于	dist>1	False
>=	≥	大于等于	dist>=1	False
==	=	等于	dist==1	False
!=	≠	不等于	dist!=1	True

特别要注意的是，Python 语言使用“=”表示赋值语句，使用“==”表示“等于”。此外，Python 语言支持连续比较，例如，判断一个数 x 是否为 100 到 200 之间的一个数，其条件表达式为：100<=x<=200。

字符和字符串也可以进行比较运算，字符的比较对应字符 ASCII 码值的大小；字符串的比较则按照字典顺序进行，举例如下。

```
>>> "python"=="python"
True
>>> "Python">"python"
False
>>> "C"<"Python"
True
```

2. 逻辑运算符

在实际问题中，对一些复杂的条件，需要用几个关系表达式组合起来表示，这种组合操作就是逻辑运算。常用的逻辑运算符包括与 (and)、或 (or)、非 (not)。这三种运算符的运算顺序是先 not，再 and，最后 or，但可以使用括号来调整运算的先后顺序。

例如，判断一个学生是否为体重 (weight) 大于 50 千克或者身高 (height) 小于 155 厘米的女生 (gender:Female)，条件表达式可写为：

```
(gender=="Female") and (weight>50 or height<155)
```

表达式中的括号可以用于定义运算的优先级，这与数学中括号的作用相同。上式中先执行括号里的 or 运算，再执行 and 运算。如果不用括号，使用了如下的条件表达式：

```
gender=="Female" and weight>50 or height<155
```

那么表达式先执行 and 运算，再执行 or 运算，其含义改变为判断一个学生是否为体重大于 50 千克的女生，或者身高小于 155 厘米的学生 (不论男女)。

实现选择结构的语句

Python 语言用 if 语句实现选择结构，它包括以下三种基本形式 (表 3-8)：

表 3-8 实现选择结构的三种语句

单分支语句	双分支语句	多分支语句
if 条件表达式: 语句块 1	if 条件表达式: 语句块 1 else: 语句块 2	if 条件表达式 1: 语句块 1 elif 条件表达式 2: 语句块 2 else: 语句块 n

语句块是 if 语句的条件被满足时执行的一个或多个语句序列。语句块中的语句通过与 if、else、elif 等关键字所在行形成缩进格式来表示包含关系。

1. 单分支语句

if 语句首先运算条件表达式，如果结果为 True，则执行语句块里的语句序列；如果结果为 False，则语句块里的语句会被跳过。单分支结构流程图如图 3-31 所示。

例如，求 x 的绝对值的语句如下。

```
if x<0:
    x=-x
```

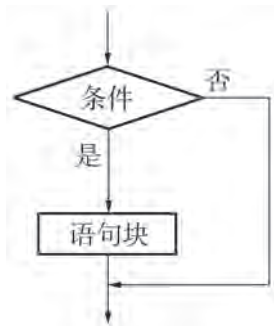


图 3-31 单分支结构

2. 双分支语句

Python 语言用 if-else 语句来实现双分支结构，语句块 1 是 if 语句的条件被满足时执行的一个或多个语句序列，语句块 2 是 if 语句的条件不被满足时执行的一个或多个语句序列。

双分支语句针对条件的两种可能结果 True 或者 False，分别形成两条执行路径。双分支结构流程图如图 3-32 所示。

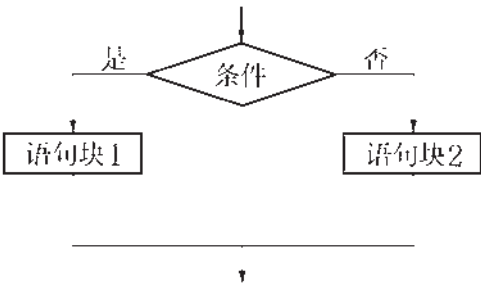


图 3-32 双分支结构

例如，编写一个让小学生练习 100 以内正整数减法的程序，程序随机产生两个正整数 x、y，然后向学生提问类似“92-14=”这样的问题，学生回答问题后，程序显示答案是否正确。

这个程序要完成两次判断。第一次，产生的两个随机数x和y，应满足 $x \geq y$ 才能得到非负整数的结果，如果不是，出题目的时候应该改为 $y-x$ 。第二次，要将学生输入的答案与正确答案进行比较，给出“正确”或“错误”的反馈。这两个判断都是双分支结构。程序中要用到随机函数randint(a,b)，产生一个[a,b]中的整数。具体语句如下。

```
import random
x,y=random.randint(1,100),random.randint(1,100)
if(x>y):
    print("%d-%d=%d"%(x,y))
    answer=x-y
else:
    print("%d-%d=%d"%(y,x))
    answer=y-x
youranswer=int(input())
if youranswer==answer:
    print("正确")
else:
    print("错误")
```

3. 多分支语句

Python 语言用 if-elif-else 语句来实现多分支结构，多分支结构是双分支结构的扩展。程序运行时依次判断 if 语句和 elif 语句的条件是否被满足，如果结果为 True，则执行该条件下的语句块，语句块执行结束后跳出整个 if-elif-else 结构，执行后面的语句；如果结果为 False，则转向下一个条件进行判断。如果没有任何条件被满足，则执行 else 下面的语句块。else 子句是可选的。多分支结构流程图如图 3-33 所示。

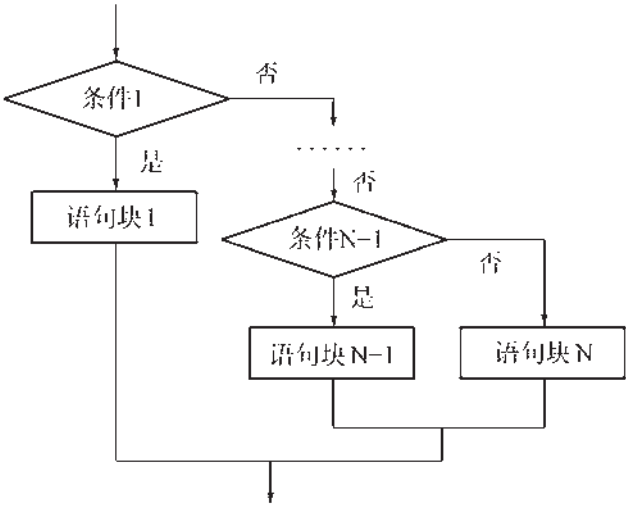


图 3-33 多分支结构

例如，编写程序，输入身高和体重，计算身体的质量指数 BMI 值，并给出健康信息。这个问题可以用两种分支结构来实现，语句分别如下。

```
height = eval(input("请输入身高(米): "))
weight = eval(input("请输入体重(千克): "))
bmi = weight / pow(height, 2)
print("BMI 数值为: %.2f"%bmi)

if bmi < 18.5: # < 18.5
    print("BMI 指标为: 偏瘦")
elif bmi < 25: # 18.5 <= bmi < 25
    print("BMI 指标为: 正常")
elif bmi < 30: # 25 <= bmi < 30
    print("BMI 指标为: 偏胖")
else:
    print("BMI 指标为: 肥胖")
```

多分支语句实现

```
height = eval(input("请输入身高(米): "))
weight = eval(input("请输入体重(千克): "))
bmi = weight / pow(height, 2)
print("BMI 数值为: %.2f"%bmi)

if bmi < 18.5:
    print("BMI 指标为: 偏瘦")
if 18.5 <= bmi < 25:
    print("BMI 指标为: 正常")
if 25 <= bmi < 30:
    print("BMI 指标为: 偏胖")
if bmi >= 30:
    print("BMI 指标为: 肥胖")
```

单分支语句实现

上面程序中 eval 函数的作用是将字符串参数转化为有效的表达式，求值并返回计算结果，具体到此处为自动识别输入的身高字符串和体重字符串，并转化为数值数据。两个程序的运行结果是相同的，但是它们的算法效率有明显差异。例如，当计算出的 BMI 值为 18 时，对比它们的执行过程：若执行多分支语句，进入第一个分支，输出结果后，程序执行结束；若执行单分支语句，第一个 if 条件满足，输出结果后，还要继续依次执行后三条 if 语句，并运算条件表达式的值。所以多分支语句的算法效率更高。

分析历史气温数据

——设计批量数据算法

气温是与人们日常生活息息相关的一个物理量，它表示空气的温度。在中国，用摄氏温标($^{\circ}\text{C}$)表示气温。气温在一天中有着高低变化，气象观测站一般每天观测4次，分别为2时、8时、14时、20时四个时点。有些观测站每个小时观测一次气温，甚至每5分钟观测一次。除了气温，反映天气的气象数据还包括气压、风力、风向、降水量等诸多数据。现今，随着综合观测系统的建设和完善，人们获取和利用的气象数据量已呈指数级增长。

日积月累，一个地区的历史气象数据会越来越多。利用计算机日益强大的数据处理能力，人们对海量的历史气象数据进行统计和建模分析，开展天气预报工作及各类气象服务(图3-34)，并了解人类活动对气候变化的影响。



图 3-34 天气预报

项目学习目标

在本项目中，我们将围绕网上发布的历史气温数据展开研究，利用程序设计语言编写程序，存储和计算平均气温等批量数据。

完成本项目学习，须回答以下问题：

1. 批量数据如何在程序中表示？
2. 访问批量数据的常用程序模式有哪些？
3. 如何定义和使用函数？
4. 如何理解“自顶向下、逐步求精”的程序设计方法，以及嵌套、模块化设计的思想？

项目学习指引

1. 用列表表示和计算平均气温

气象台每日发布的气象数据中，常常会出现最高气温、最低气温和日平均气温。最高气温是一日内气温的最高值，一般出现在 14 时左右；最低气温是一日内气温的最低值，一般出现在凌晨。日平均气温的计算方法不一，可以是最高气温和最低气温的平均值，也可以是一天多次观测值的平均值。而一个月的日平均气温的平均值就是月平均气温。

(1) 计算日平均气温

问题：已知气象台某日观测得到的 4 次气温值为：10℃、13℃、18℃、12℃，求该日的日平均气温。

大部分高级语言都提供了对**批量数据**在程序中进行表示的方法，最常见的是数组。Python 语言则提供了更加便利的列表类型，可以应用于一次处理多个数据对象的场合。

创建一个列表，存储该日观测得到的气温数据的语句如下：

```
>>> t_Day=[10,13,18,12]
```

计算平均值就是把一系列数值数据累加后再除以数据的个数。实现这一计算的方法很多，可以通过算术表达式计算，也可以使用内置函数计算。

方法 1: 使用算术表达式计算

```
>>> t_Day=[10,13,18,12]
>>> avg=(t_Day[0]+t_Day[1]+t_Day[2]+t_Day[3])/4
>>> avg
13.25
```

方法 2: 使用内置函数计算

```
>>> t_Day=[10,13,18,12]
>>> avg=sum(t_Day)/len(t_Day)
>>> avg
13.25
```

在 Python 语言的列表中，各数据项顺序排列，并允许按下标访问。下标是一个整数，有效范围从 0 到列表的长度减 1。`t_Day[0]` 表示列表中的第一个数据项，`t_Day[i]` 表示列表中的第 $i+1$ 个数据项。通过访问数据项可以读取、修改数据项的数据。

核心概念

批量数据：一组数据类型相同的数据对象。数据类型可以是简单类型，也可以是复杂的构造类型。

← 参见 P115 知识链接“批量数据及其表示和操作”

小贴士

Python 语言中使用列表来表示批量数据。列表用“`[]`”表示。其中每一项可以是任意数据类型的数据项，各数据项之间用逗号分隔。通过赋值运算可以创建列表：

`L=[]` # 创建一个空的列表 L

内置函数 `sum` 的作用是求列表中数值数据的累加和，内置函数 `len` 的作用是求列表中数据项的个数。有统计功能的内置函数还有 `max`、`min` 等。

(2) 计算月平均气温

问题：已知某市 2013 年 8 月的 31 个日平均气温数据为：34,32,33,33,35,36,36,36,35,34,33,32,32,32,31,32,32,30,31,30,30,31,32,29,27,29,31,32,30,26,26，求该市 2013 年 8 月的月平均气温。

使用 Python 语言的列表可以存储以上 31 个气温数据。

```
>>> t_Day=[34, 32, 33, 33, 35, 36, 36, 36, 35, 34, 33, 32, 32, 32, 31, 32, 32, 30, 31, 30, 30, 31, 32, 29, 27, 29, 31, 32, 30, 26, 26]
```

计算月平均气温，既可以使用内置函数，也可以用循环结构先求得这个月的日平均气温总和再完成计算。访问列表中的每一个数据项时，通常有两种不同的访问方法：下标访问和序列迭代访问。

小贴士

假设列表为 L。

- 下标访问方法是通过变量 i 表示下标值，该值从 0 到 len(L)-1 变化，依次访问数据项 L[i]。

```
for i in range(0,len(L)):
```

访问 L[i]

- 序列迭代访问方法是迭代访问列表中的每一个数据项 x。

```
for x in L:
```

访问 x

```
>>> s=0
>>> n=0
>>> for i in range(len(t_Day)):
>>>     s=s+t_Day[i]
>>>     n=n+1

>>> avg=s/n
>>> avg
31.677419354838708
```

下标访问方法

```
>>> s=0
>>> n=0
>>> for x in t_Day:
>>>     s=s+x
>>>     n=n+1

>>> avg=s/n
>>> avg
31.677419354838708
```

序列迭代访问方法

思考与讨论??

1. 如何使用电子表格软件计算月平均气温？
2. 从数据准备和求解方法等方面谈谈自己对 Python 语言和电子表格软件这两种工具的认识。

(3) 使用用户自定义函数计算平均值

许多数据分析工具软件都会提供求平均值的系统函数，用来直接计算一组数据的平均值。Python 语言没有提供求平均值的内置函数，但可以通过**用户自定义函数**进行扩充。

```
>>> def getAvg(L):
    return sum(L)/len(L)

>>> avg=getAvg(t_Day)
>>> avg
31.677419354838708
```

(4) 读取文件中的数据到列表

以上问题中的 31 个日平均气温数据，若要长期存储并反复使用，通常可将其保存在文件中，再通过程序指令来读取文件中的数据。文本文件是存储程序数据的常用文件格式。

首先，新建文本文件“201308tq.txt”，将 31 个日平均气温数据依次输入，每行一个数据。

然后，在 Python Shell 中读取存储在文件“201308tq.txt”中的气温数据到列表 L 中。

```
>>> L=list(open("c:\\sample\\201308tq.txt"))
>>> L
['34\n', '32\n', '33\n', '33\n', '35\n', '36\n', '36\n', '36\n', '35\n', '34\n', '33\n', '32\n', '32\n',
 '32\n', '31\n', '32\n', '32\n', '30\n', '31\n', '30\n', '30\n', '31\n', '32\n', '29\n', '27\n', '29\n',
 '31\n', '32\n', '30\n', '26\n', '26\n']
>>> for i in range(len(L)):
    L[i]=int(L[i])

>>> L
[34, 32, 33, 33, 35, 36, 36, 36, 35, 34, 33, 32, 32, 32, 31, 32, 32, 30, 31, 30, 30, 31, 32, 29, 27, 29,
 31, 32, 30, 26, 26]
```

Python 提供了文件的快速访问方式，可将文本文件中的数据按每行一个字符串的形式读入一个列表中，其格式为：

列表 =list(open(文件名))

列表 L 中得到的是 31 个包含“\n”的字符串数据，因此，需要使用 int 函数逐个将字符串数据转化为整型数据。

核心概念

用户自定义函数：由用户自己编写的、能够实现独立功能的语句块。

小贴士

函数定义的一般形式：

def 函数名 (< 参数 >):
函数体

每个函数可以返回一个值，return 语句的作用为返回表达式的值。

← 参见 P117 知识链接“函数的定义和调用”

← 参见 P118 知识链接“文本数据的表示和操作”

活 动

8.1 编写程序, 计算文件“201308tq.txt”中存储的 31 个日平均气温数据的平均值。

(1) 编写并测试函数 readList: 从文件中读取数据到列表, 并将列表中的数据项转化为整型数据后, 返回列表。

1. 读入文件数据到列表 L;
2. 将 L 中的数据项转化为整数;
3. 返回 _____。

```
def readList(filename):
    _____
    for i in range(len(L)):
        L[i]=int(L[i])
    _____
```

- ① 分析函数, 分析需要传入函数的参数和函数的计算结果。
- ② 设计函数的算法。
- ③ 编程实现函数。
- ④ 调用测试函数(调用 readList, 读取数据到列表 L, 并显示 L)。

(2) 编写程序, 实现计算一组日平均气温数据的平均值, 并使用 getAvg 函数和 readList 函数完成相关功能。

① 设计算法(提示: 完整的算法应该包括创建表示一个月的日平均气温序列的列表, 实现计算和结果的输出)。

② 创建程序文件, 实现算法(提示: 在程序文件中, 应先完成两个函数的定义, 再开始编写程序代码, 调用定义好的函数)。

小贴士

csv 文件采用的是以逗号分隔数值的存储方式。这是一种通用的、相对简单的文本文件格式, 在电子表格软件中会显示为二维表的形式, 被广泛应用于在程序之间转移表格数据。

2. 用模块化设计批量计算平均气温

问题: 已知某地 2013 年 365 天的日平均气温数据(已按一月一行的形式存放在“2013avgtq.csv”文件中, 图 3-35 显示了前两个月的数据), 批量计算该地 2013 年 12 个月的月平均气温。



图 3-35 某地 2013 年日平均气温数据的文件

(1) 分析问题

依次计算各个月的月平均气温, 即得到月平均气温的序

列。解决本问题的关键在于如何组织每个月的日平均气温数据,不同的数据组织所对应的算法设计也不同。

在 Python 语言中,可利用**嵌套**列表组织以上数据:将每个月的日平均气温构成的列表作为一个列表的数据项——子列表,构成嵌套的列表,其形式如下:

[[], [], [], [], ...]

使用下标可以方便地访问嵌套列表中的每一个子列表。

在 Python 语言中使用下列语句块,可以从“2013avgtq.csv”文件中逐月读取日平均气温数据,把它们转化为整型数据,并以子列表形式存储在嵌套列表 yL 中。该嵌套列表中包含 12 个子列表,一个子列表就是一个月的日平均气温序列。

```
f=open("c:\\sample\\2013avgtq.csv")
yL=[]
for line in f:
    line=line.replace("\n"," ")
    L=line.split(",")
    for i in range(len(L)):
        L[i]=int(L[i])
    yL.append(L)
f.close()
```

Python 语言对文件采用统一的操作步骤,即“打开—操作—关闭”。open 函数用来打开文件(默认用只读方式打开),并返回文件对象。for line in f 是逐行访问文件对象 f, line 表示一行字符串。line.replace("\n", " ")是将每一行结束处的换行符“\n”用空串替代,相当于删除换行符。line.split(",")是按照逗号将 line 字符串切割为若干个字符串,返回一个列表。yL.append(L)是将列表 L 追加到列表 yL 中,作为一个数据项。close 函数用来关闭文件。

思考与讨论??

1. 执行for循环后得到的列表L的数据项是什么类型? yL列表的数据项又是什么类型?
2. 12个月的数据如何读取并存入列表中? 重复步骤是哪几步? 循环控制模式是怎样的?

核心概念

嵌套: 两个物体有装配关系时,将一个物体嵌入另一物体,这种方法叫做嵌套。嵌套的思想在计算机程序设计中应用广泛。典型的嵌套包括嵌套控制结构和嵌套数据结构。

核心概念

“自顶向下、逐步求精”的设计方法：先进行顶层设计，再对各组成部分进行细化。

小贴士

程序模块图是模块化设计的结果。在模块化设计中，按照需求分析将程序分割成许多模块，建立以模块为单位的多层次结构体系。模块实现的具体形式为函数或语句块。

核心概念

模块化：解决一个复杂问题时自顶向下逐层划分为若干个模块的过程。程序设计中的模块化就是把一个复杂庞大的程序按照某种依据（如功能）划分成若干个相对独立的程序块，以方便程序的编写、调试和对程序的了解。

(2) 设计算法

批量计算一年中 12 个月的月平均气温，可以使用“自顶向下、逐步求精”的方法来设计算法。

• 顶层算法设计

1. 创建日平均气温列表；
 2. 计算 12 个月的月平均气温；
 3. 输出 12 个月的月平均气温。

上述算法的顶层程序模块图如图 3-36 所示。



顶层算法中的三步，并不能直接对应一条明确的程序语句，还需要进行细化设计，将顶层算法模块化，每一步为一个独立的功能模块。

• 设计模块 1 的算法——创建日平均气温列表

- 第 1 步算法细化

 - 1.1 打开文件对象 f；
 - 1.2 初始化新列表 dL；
 - 1.3 逐行读入文件数据，将一行字符串切割，转化为整型数据后，作为子列表追加到列表 dL 中；
 - 1.4 关闭文件。

第 1.3 步算法细化

- 1.3.1 按逗号切割字符串 line 中的数据到列表 L;
- 1.3.2 循环 i 从 0 到 len(L)-1;
 - 1.3.2.1 将 L[i] 转化为整数;
- 1.3.3 将 L 追加到列表 dL 中。

小贴士

算法步骤 1.3 是一个循环结构, 1.3.2 也是一个循环结构, 在一个循环结构中出现了一个循环结构, 这称为循环的嵌套结构, 也称为多重循环。

- 设计模块 2 的算法——计算 12 个月的月平均气温
嵌套列表 dL 中包含 12 个子列表, 每个子列表是一个月的日平均气温序列。对每个子列表求平均值, 结果存放在列表 avgy 中。

第 2 步算法细化

- 2.1 初始化 avgy;
- 2.2 迭代访问 dL 中每一个子列表 m;
 - 2.2.1 求列表 m 的平均值 avg;
 - 2.2.2 将 avg 追加到列表 avgy 中。

类似的, 还须完成模块 3 的算法设计——输出 12 个月的月平均气温, 其具体设计与对屏幕显示的要求相关。

思考与讨论??

如何设计一个一年中 12 个月的月平均气温数据的输出样式? 算法如何设计?

(3) 函数实现

模块化设计中的子模块既可以用函数实现, 也可以直接用程序块实现。下面以模块 1 创建日平均气温列表为例, 完成其函数实现。

在 Python Shell 中编写并测试函数 readNList, 该函数功能为从参数中获取指定文件名, 读入文件数据到列表, 处理后得到包含 12 个月的日平均气温子列表的新列表, 并返回新列表。

- 分析函数

本问题所涉及的数据对象包括函数的输入参数(文件名 filename)、存放日平均气温数据的嵌套列表 dL、文件对象 f。

函数的计算结果是包含 12 个月的日平均气温子列表的新列表,通过 return 语句返回。

- 定义函数

参照模块 1 的算法设计,在 Python Shell 中定义函数 readNList。

```
>>>def readNList(filename):
    f=open(filename)
    dL=[]
    for line in f:
        line=line.replace("\n","")
        L=line.split(",")
        for i in range(len(L)):
            L[i]=int(L[i])
        dL.append(L)
    f.close()
    return dL
```

参见 P121 知识链接“自顶向下、逐步求精”的程序设计方法和“模块化设计的思想”

- 调用函数

调用 readNList, 获取一年的日平均气温数据列表 tqy, 并显示 tqy。

```
>>> filename=input("请输入csv文件名:")
请输入csv文件名: c:\sample\2013avgqtq.csv
>>> tqy=readNList(filename)
>>> tqy
```

类似的, 模块 2 与模块 3 也可以用函数实现。

思考与讨论??

以下函数调用中, 存储文件名的变量为 file, 与函数定义时输入参数的名称 filename 不一致。这种函数调用正确吗? 为什么?

```
>>> file="c:\\sample\\2013avgqtq.csv"
>>> tqy=readNList(file)
>>> tqy
```

活 动

8.2 编写程序，利用文件“2013avgtq.csv”中的 2013 年日平均气温数据，分别计算 12 个月的月平均气温，按如下格式显示在屏幕上。



(1) 基本要求：定义并调用 calAvg 函数和 readNList 函数，实现模块 2 计算一年 12 个月的月平均气温，按如下算法编写程序。

- 1. 读入数据文件名；
- 2. 调用 readNList 获取嵌套列表 tqy；
- 3. 调用 calAvg 函数，访问 tqy，计算 12 个月的月平均气温并存储在列表 avgy 中；
- 4. 输出列表 avgy 的数据。

(2) 拓展要求：定义并调用 saveList 函数，实现模块 3 输出 12 个月的月平均气温，并将计算结果写入一个用户指定的 csv 文件中。

(3) 绘制完整的程序模块图。

(4) 交流：若将 366 个或 365 个日平均气温直接作为数据项放在一个列表中，如何设计算法？（提示：每个月的天数是不一样的，2 月份的天数在闰年和平年还不相同，故在取某一个月的数据时须计算该月数据在列表中的起始位置和结束位置。）

知识链接

批量数据及其表示和操作

在实际的计算机处理问题的过程中，程序要处理的对象大多是大批量的相同数据类型的数据，如一次科学实验中获得的大量实验数据、关键字搜索时获得的大量网页中所包含的单词和一个 BMP 图像中包含的像素点等，它们分别对应大量数值的集合、大量文本的集

合和大量像素点对象的集合。程序设计语言支持批量数据的存储,用统一的名称管理一批数据,在内存上的存储空间是连续的。

对批量数据中数据项的访问可通过下标。例如, `a[1]`, `a[i]`。下标的含义是与第一个数据项的偏移量,通常从 0 开始。

例如, `Color=["red","green","blue"]`, `Color[0]` 的值是 "red", `Color[1]` 的值是 "green", `Color[2]` 的值是 "blue"。

与单变量数据相比,批量数据的优势在于:

- (1) 一批批量数据只需定义一个名称,程序的通用性更强。而一个单变量只可以控制一个数据,使用单变量,程序可控制的数据的个数是固定的。
- (2) 使用方便,可以组织循环控制结构,通过控制下标的值控制一批数据。

高级语言都提供了对一批同类型的数据在程序中的表示方法,最常见的是数组。Python 语言提供了更加便利的列表类型,其形式更灵活,操作功能更强。

Python 语言中提供了对列表的一些通用操作,以实现对序列的索引、连接、复制、检测成员等,见表 3-9。

表 3-9 Python 列表的基本操作

操作	描述
<code>x1+x2</code>	连接列表 <code>x1</code> 和 <code>x2</code> , 生成新列表
<code>x*n</code>	将列表 <code>x</code> 复制 <code>n</code> 次, 生成新列表
<code>x[i]</code>	引用列表 <code>x</code> 中下标为 <code>i</code> 的列表成员, <code>i</code> 从 0 开始计数
<code>x[i:j]</code>	引用列表 <code>x</code> 中下标从 <code>i</code> 到 <code>j-1</code> 的子列表
<code>x[i:j:k]</code>	引用列表 <code>x</code> 中下标从 <code>i</code> 到 <code>j-1</code> 、间隔为 <code>k</code> 的子列表
<code>len(x)</code>	计算列表 <code>x</code> 中成员的个数
<code>max(x)</code>	列表 <code>x</code> 中最大数据项
<code>min(x)</code>	列表 <code>x</code> 中最小数据项
<code>v in x</code>	检测 <code>v</code> 是否在列表 <code>x</code> 中, 返回布尔值
<code>v not in x</code>	检测 <code>v</code> 是否不在列表 <code>x</code> 中, 返回布尔值

Python 语言还提供了如表 3-10 所示方法,可以执行对列表的操作。

表 3-10 Python 列表的方法

方法	描述
<code>list.append(x)</code>	把一个数据项添加到列表的结尾, 相当于 <code>a[len(a):] = [x]</code>
<code>list.extend(L)</code>	将一个列表中的所有数据项都添加到另一个列表中, 相当于 <code>a[len(a):] = L</code>
<code>list.insert(i, x)</code>	在指定位置插入一个数据项, 如 <code>a.insert(0, x)</code> 会插入到整个列表之前, 而 <code>a.insert(len(a), x)</code> 相当于 <code>a.append(x)</code>
<code>list.remove(x)</code>	删除列表中值为 <code>x</code> 的第一个数据项; 如果没有这样的数据项, 就返回 <code>False</code>

函数的定义和调用

在设计一个大的应用程序时，经常会发生实现类同功能的程序块在程序中多次反复出现的情况。如果单独编写这些程序块，供程序中各部分甚至各应用程序共享使用，就可以减少编写代码的工作量，而且也便于修改和查看。Python 语言提供了函数来实现有独立功能的程序块。

1. 函数的定义

构造一个函数的思路与分析数学应用题有些类似，要知道已知什么和求什么。将已知条件放在参数中，在调用时通过参数传递得到，而所求的值可通过函数的返回值返回。

函数定义的一般格式如下：

```
def 函数名 (< 参数 >):  
    函数体
```

函数的定义中一般都包含 `return` 语句，`return` 语句的作用是结束函数的执行，返回一个值作为函数的返回值。返回值可以是简单的整型数据、浮点型、文本数据等，也可以是复杂的列表等，还可以没有返回值，仅仅结束函数的执行。

在本项目中，`getAvg` 函数的返回值是一个表示数据系列的平均值的浮点值，`readNList` 函数的返回值则是一个存放了一个月的日平均气温值的列表。没有返回值的函数往往是执行一个操作，例如输出 12 个月的月平均气温的函数不需要返回值。

2. 函数的调用

调用 `getAvg` 函数时，需要告诉函数计算哪些数据的平均值，所以需要有一个参数来接收数据序列。函数最后返回的是求得的平均值，在调用函数时，还要考虑接收函数的返回值。

函数调用的一般形式如下：

```
函数名 (< 参数值列表 >)
```

因为大部分函数是有返回值的，所以调用时常和赋值语句一起使用，由变量引用返回值，其调用形式为：

```
变量名 = 函数名 (< 参数值列表 >)
```

例如，要计算 $7! + 11! - 10!$ ，可以定义一个函数，用于计算 $n!$ 。要计算 $n!$ ，必须有一个参数接受 n 的值，并有一个返回值作为计算结果。

定义函数的代码如下：

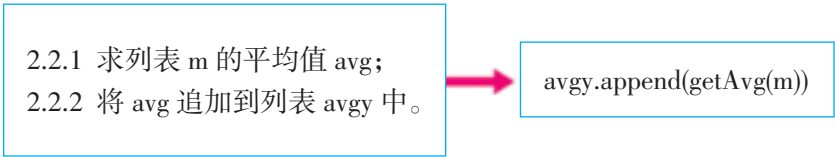
```
>>> def fact(n):  
    p=1  
    if n<0:  
        return -1  
    for i in range(1,n+1):  
        p=p*i  
    return p
```

计算 $7! + 11! - 10!$ 时, 需要调用 3 次 fact 函数, 调用语句如下:

```
>>> result=fact(7)+fact(11)-fact(10)
>>> result
36293040
```

调用函数的方式是十分灵活的, 可以单独调用, 也可以将它包含在一个表达式中, 让函数返回值直接参加表达式运算, 还可以将它作为另一个函数调用的参数。

例如, 本项目模块化设计的算法中, 步骤 2.2.1 和 2.2.2 可以用一个函数调用完成。函数 append 的参数是 getAvg 函数的返回值。



文本数据的表示和操作

1. 文本数据的表示

计算机中表示文本的最基本单位是字符, 包括可打印字符和不可打印的控制字符。可打印字符包括: ①英文大小写字母 a-z 和 A-Z; ②数字字符 0~9; ③标点符号和一些键盘上的常见符号。控制字符包括回车、制表符、退格等, 在程序中要用转义字符表示这些控制字符。Python 语言中的转义字符以 \ 为前缀, 如表 3-11 所示。

表 3-11 Python 的转义字符

转义字符	描述	转义字符	描述
\\	反斜杠符号	\t	横向制表符
\'	单引号	\r	回车
\"	双引号	\n	换行
\a	响铃	\(在行尾时)	续行符
\b	退格 (Backspace)	\f	换页
\e	转义	\oyy	八进制数 yy 代表的字符, 如: \o12 代表换行
\000	空	\xyy	十六进制数 yy 代表的字符, 如: \x0a 代表换行

字符串是字符的序列表示。字符串常量可以使用单引号、双引号、三引号封装, 但前后必须一致。其中单引号、双引号可以表示单行字符串, 两者作用一致。使用单引号时, 双引号可以是字符串的一部分; 使用双引号时, 单引号可以是字符串的一部分; 三引号可以表示单行或多行字符串。举例如下。

```
>>> print('I'm a programmer.')
I'm a programmer.
>>> print('She said:"you are right".')
She said:"you are right".
>>> print('三引号中可以有"单引号"
可以有"双引号"
还可以换行')
三引号中可以有"单引号"
可以有"双引号"
还可以换行
```

2. 字符串类型数据的基本操作

Python 语言提供了对字符串类型数据的一些通用操作，如表 3-12 所示，以实现对字符串的连接、复制、包含、索引等。

表 3-12 Python 字符串类型数据的操作符

操作符	描述
x+y	连接两个字符串 x 和 y
x*n	复制 n 次字符串 x
x in s	包含判断，判断字符串 x 是否为字符串 s 的子串，是则返回 True，否则返回 False
str[i]	索引操作，返回下标位置的一个字符
str[N:M]	切片，返回从 str[N] 到 str[M-1] 的子串
len(x)	返回字符串 x 的长度
str(x)	返回任意类型 x 所对应的字符串

下面一组操作示例了连接、复制、包含操作符的使用。

```
>>> s1="hello "
>>> s2="Python "
>>> s3=s1+s2
>>> print(s3)
hello Python
>>> s4=3*s2
>>> print(s4)
Python Python Python
>>> s2 in s4
True
>>> s1 in s4
False
```

下例为获取月份英语缩写的程序，输入一个表示月份的整数（1~12），输出对应的月份字符串名称。

```
monthStr='JanFebMarAprMayJunJulAugSepOctNovDec'
monthId=eval(input('请输入月份 (1-12):'))
pos=(monthId-1)*3
print(monthStr[pos:pos+3])
```

程序的运行结果如下：

```
===== RESTART: C:/sample/getMonthname.py =====
请输入月份 (1-12): 8
Aug
```

要注意的是，可以通过下标访问字符串中的指定字符或子串，但不能改变字符串的内容。举例如下。

```
>>> print(s4[7:13])
Python
>>> s4[7:13]='Hello'
Traceback (most recent call last):
  File "<pyshell#48>", line 1, in <module>
    s4[7:13]='Hello'
TypeError: 'str' object does not support item assignment
```

Python 语言还提供了丰富的字符串操作方法，如表 3-13 所示(s 表示一个字符串对象)。

表 3-13 字符串对象的常用方法

常用方法	描述
s.capitalize()	返回首字符大写后的字符串，s 对象不变
s.lower()	返回所有字符改小写后的字符串，s 对象不变
s.upper()	返回所有字符改大写后的字符串，s 对象不变
s.strip()	返回删去前后空格后的字符串，s 对象不变
s.replace(old,new)	将 s 对象中所有的 old 子串用 new 子串代替
s.count(sub[,start[,end]])	计算子串 sub 在 s 对象中出现的次数，start 和 end 定义起始位置
s.find(sub[,start[,end]])	计算子串 sub 在 s 对象中首次出现的位置
s.join(iterable)	将序列对象中所有字符串合并成一个字符串，s 对象为连接分隔符
s.split(sep=None)	将 s 对象按分隔符 sep 拆分为字符串列表，默认为空格

字符串对象的常用方法示例：

```
>>> s='  hello Python  '
>>> t=s.strip()
>>> s
'  hello Python  '
>>> t
'hello Python'
>>> t.find('he')
0
>>> s.count('h')
2
>>> L=t.split()
>>> L
['hello', 'Python']
>>> t1='.'.join(L)
>>> t1
'hello.Python'
```

“自顶向下、逐步求精”的程序设计方法

本项目在解决“批量计算 12 个月的月平均气温”问题时，先进行顶层设计，再对各部分进行细化。这种程序设计方法使程序的结构变得更加清晰，程序的调试也更加简单。

这种程序设计方法称为“自顶向下、逐步求精”。该方法是把整个设计过程分出层次来，逐步加以解决。每一步都是在前一步的基础上进行的，是对前一步设计的细化和具体化。这好比画一幅画，先整体规划，确定这幅画由哪些内容组成，勾勒出大致的轮廓，然后考虑各部分的形状是怎样的，最后进行细节润色。

许多程序从逻辑上可以分为 3 个阶段：输入数据并做相应调整的处理阶段、计算得到结果的阶段、显示结果并保存数据的阶段。在自顶向下的程序设计方法中，最顶层仅仅描述了程序的功能，它虽然是程序的一个完整表示，但很少提供足够的信息，不能直接用程序语句表示，因此需要进行细化处理。对每一部分进行细化，遇到不能清楚解释的步骤再向下一层延伸。依次进行第二层、第三层的细化，直至每一步都是一个基本操作为止。每次分步处理，都是算法在一个层次上的细化。

这样考虑问题，思路比较清晰，有条不紊，而且把一个大问题分解为若干个小问题考虑，既降低了问题的复杂性，也便于程序实现。

模块化设计的思想

大多数解决现实问题的程序都比本书中介绍的程序要大得多，对于一个复杂的问题通常需采取“分而治之”的策略，也就是模块化设计的思想。运用软件工程的思想来组织软件开发工作，一般要经历需求分析、设计程序、编写程序等阶段。需求分析所做的事情是，根据用户的要求确定程序应当具备哪些功能，达到什么性能。设计程序阶段大致又可再分为两个子阶段，即程序的模块化设计和模块内的逻辑设计。在模块化设计中，要按照需求分析把程序分割成许多模块，建立以模块为单位的多层次结构体系，各个模块之间形成调

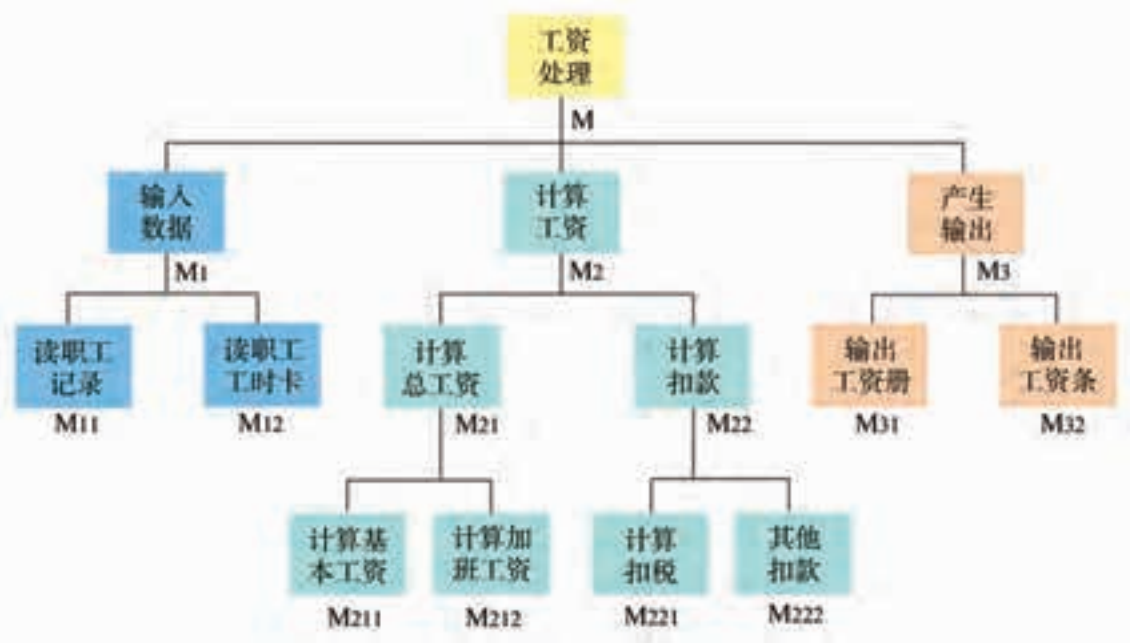


图 3-37 模块化设计示例

用关系。模块化之后还要确定每一模块内的具体算法，并准确表达这些算法。整个程序设计完成之后，就可以使用特定的程序设计语言来编写程序。

一个程序的算法设计可以从三个基本步骤——“输入”“处理”“输出”开始，形成顶层模块，再对每一个模块继续向下细化，直至每一个模块都相对独立地完成一个基本功能。例如，图 3-37 演示了一个工资处理程序的模块化设计的过程。

常见算法——枚举法

枚举法在生活中非常普遍，是处理问题最常用的算法思想之一。枚举法的基本原理是根据已知条件，在给定的范围内对所有可能的答案按某种顺序进行逐一枚举和检验，从中找出那些符合要求的答案。在日常生活中，人们在解决问题时，如果需要枚举的范围比较小，采用枚举法会显得比较直观、合理；但当枚举范围比较大时，则会十分繁琐、容易出错。然而，对于计算机来说，它的高速运算能力保证了枚举法的可行性，而且实现相对简单。

枚举法的一般模式可以总结如下：

- 确定范围：问题所涉及的情况有哪些，情况的种数是否可以确定。
- 验证条件：分析出来的这些情况，须满足什么条件才能成为问题的答案。
- 尽可能缩小搜索范围，减少求解时间。

例如，用枚举法解决以下问题：有一张由 5 位数构成的单据号码，其千位和百位上的数字模糊不清了，已知这个数能被 57 或 67 整除，请使用计算机求这张单据号码。

• 问题分析：假设这个 5 位数为 abcde，则 $abcde = a \times 10000 + b \times 1000 + c \times 100 + d \times 10 + e$ 。本问题中，已知万位、十位、个位上的数字 a、d、e（可以由用户输入），而千位和百位上的数字 b、c 是待求的，因此可采用枚举法列举 b、c 所有可能的数字（0~9），每列举一次，判断构造的号码能否被 57 或 67 整除，若能则输出该号码。

• 设计算法：

1. 输入 a、d、e；
2. 循环 b 从 0 到 9，step1；
 - 2.1 循环 c 从 0 到 9，step1；
 - 2.1.1 $number = a \times 10000 + b \times 1000 + c \times 100 + d \times 10 + e$ ；
 - 2.1.2 如果 number 能被 57 或 67 整除，则输出 number。

拓展阅读

Python 语言实现网络爬虫应用

随着互联网的发展,海量的数据资源蕴藏在全球网站提供的海量网页之中。如何有效地获取和利用这些数据,决定了解决数据分析问题的效率,于是就产生了“网络爬虫”。

网络爬虫是一个自动提取网页的程序。其应用一般分为两个步骤:①通过网络链接获取网页内容,②对获得的内容进行处理。

Python 语言的简洁性和脚本特点非常适合链接和网页的处理,所以 Python 开源社区中支持 URL 和网页处理的第三方库很多,包括 urllib、urllib2、urllib3、wget、scrapy、requests 等。网络爬虫爬取回来的网页内容,可以通过 re、beautifulsoup4 等函数库来处理。

例如,教材配套资源中的程序“爬取气象数据.py”使用 request 库和 beautifulsoup4 库从天气网的历史数据中爬取某市 2013 年 8 月的气象数据。稍作修改就可以按用户要求爬取不同地区不同时段的数据。

单元挑战 探究密码安全问题

一、项目任务

如今,人们的生活已经被密码层层包围。密码关系着用户的重要信息、隐私甚至财产。设置一个安全的密码,其重要性不言而喻。遗憾的是,不少用户并不重视密码的安全性。

枚举法是破译密码程序的典型算法。尝试利用枚举法,编写一个“三位数字密码锁”程序,并以小组为单位,分析如何设置密码才能加强密码安全性,共同完成密码安全问题报告。

二、项目指引

1. 利用枚举法,编写一个“三位数字密码锁”的程序,实现用户输入三位数字密码,输出计算机猜中密码的尝试次数。

(1) 根据枚举法的一般模式,思考如何枚举“三位数字密码锁”的密码。

(2) 选用恰当的方式(如流程图)描述算法,并在小组内交流、优化算法,减少猜测次数。

(3) 编写程序,实现输入一个三位数字密码,输出计算机猜中密码的尝试次数。观察怎样的密码,计算机猜测的次数多。

2. 小组交流:

(1) 与三位数字密码锁(图 3-38)相比,五位字母密码锁(图 3-39)的安全性是否提高?原因是什么?

(2) 在设定密码时,应避免使用与个人信息有明显联系的数据,如生日、电话号码等,这是为什么?

(3) 现在很多密码验证环节还需要输入验证码,这是为什么?

(4) 应对使用枚举法的破译密码程序,该如何设置密码?实际应用中还可以采用哪些措施来提高密码的安全性?

3. 撰写小组报告,总结密码安全问题的探究过程,对“弱密码”使用者提出警示和合理建议。



图 3-38 三位数字密码锁



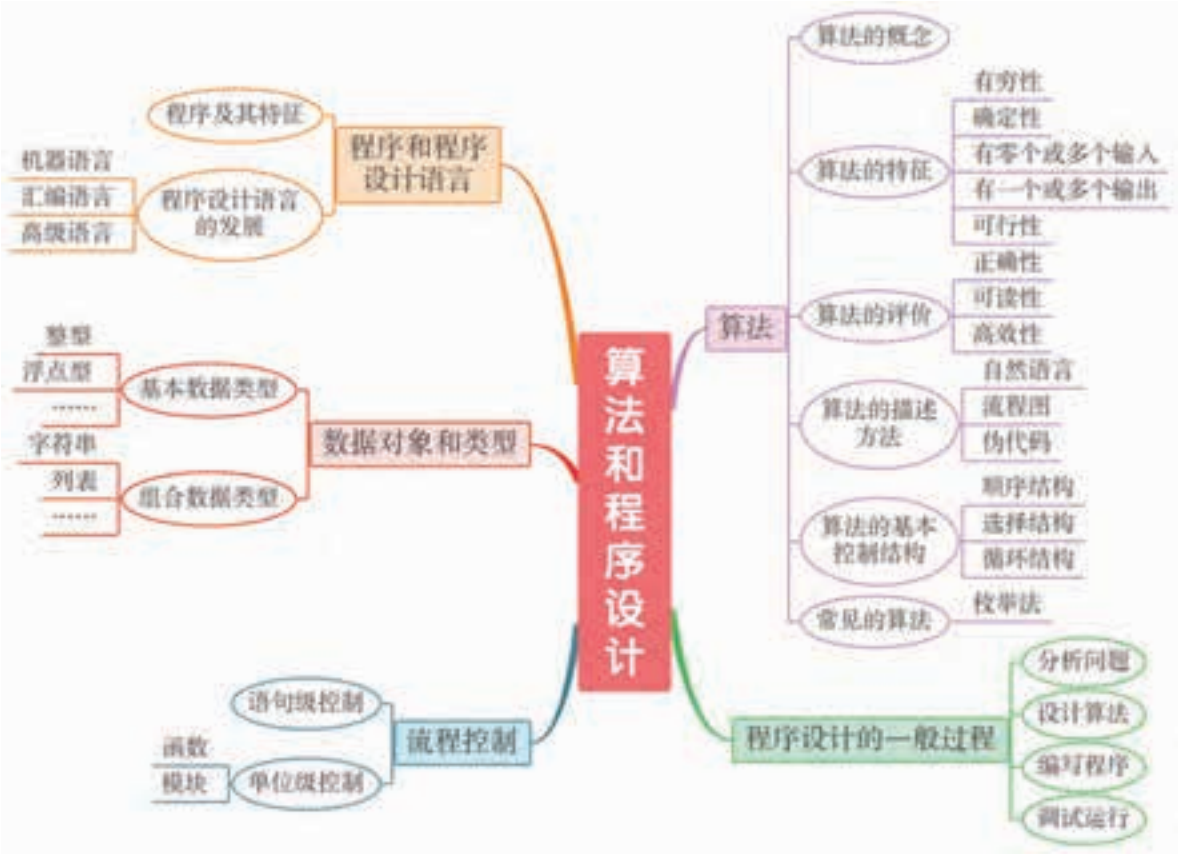
图 3-39 五位字母密码锁

三、交流评价与反思

各小组派出代表,以自己熟悉的信息表达工具(如演示文稿等)制作电子作品,通过网络或课堂展示、交流本小组的算法及程序,并对其他小组的结果进行评价。

单元小结

一、主要内容梳理



二、单元练习

1. 共享单车为用户提供的是一种单车分时租赁服务。以某共享单车为例，使用手机 App 扫描单车二维码后，车锁会自动打开，手机 App 开始记录用户的行程，在用户到达目的地并关闭车锁后，手机 App 自动结束行程记录，计算车费（每小时 1 元）后从用户账户扣款。请根据共享单车的租赁流程，用恰当的方式描述单车租赁算法。
2. 假设有 1g、2g、3g、5g、10g、20g 的砝码（其总重 $\leq 1000\text{g}$ ），每种重量的砝码个数如表 3-14 所示。

表 3-14 砝码重量及个数

砝码重量	1g	2g	3g	5g	10g	20g
个数	1	1	0	2	1	2

- (1) 用这些砝码能称出多少种不同的重量(除去一个砝码也不用的情况)? 尝试使用枚举法编程解决这个问题。
- (2) 如果使用的砝码个数不定, 每次需由用户输入砝码个数, 该如何修改程序?
3. 在输入英文的时候, 很多文本编辑软件都有首字母自动大写的功能。请编程模拟这项功能: 如果输入几句英文, 每句以“.”结束, 则会输出这几句英文, 且首字母大写。

三、单元评价

评价内容	达成情况
能从生活实例出发, 概述算法的概念, 能区别生活中的算法和计算机算法(A、T)	
能举例说明算法在人们生活、工作和学习中起到的作用(A、R)	
能解释算法的特征(T)	
知道算法的描述方法和其优缺点, 能根据实际情况选择合适的描述方法(A、T)	
能使用流程图的方式描述算法的三种控制结构(T)	
能使用三种控制结构来表示解决实际问题的算法(T)	
知道算法评价的指标, 并能对不同的算法进行正确的评价(T)	
知道程序的定义和功能, 能根据实例描述程序设计的基本流程(A、T)	
掌握 Python 程序设计语言的基本知识, 包括数据类型、常量和变量、运算符、分支语句、循环语句、函数(T)	
能理解枚举法的一般过程, 能使用枚举思想设计实际问题的解决方法(A、T)	
能掌握程序调试与运行方法, 理解程序调试在程序设计过程中的重要性(A、T)	
能根据给定的问题, 使用程序设计的方法解决问题(A、T)	
知道程序设计语言产生的必要性, 能描述程序设计语言从机器语言到高级语言的发展历程(T)	
能通过网络学习了解几种不同的常用程序设计语言, 了解其特点(T、I)	
能通过自主学习掌握更多的 Python 语言使用方法(T、I)	

说明: A—信息意识, T—计算思维, I—数字化学习与创新, R—信息社会责任

第四单元

人工智能初步

让机器像人那样思考和行动，是人类在远古时期就有的梦想，这也是人工智能（Artificial Intelligence，AI）的终极目标。

人工智能是一门新兴的学科，自 1956 年人们提出这一概念至今，仅 60 多年的时间。早期的人工智能已经能够解决很多对于人类智力来说非常困难的问题，如大规模数据的运算。但对人工智能的真正挑战在于解决那些对于人来说非常简单而机器却很难理解的问题。例如，一个果篮里放了几个苹果，几个橘子？这个问题人一看便知答案，而机器要识别苹果和橘子却并非易事。

人工智能的发展起起落落。近年来，有了大数据的支持，智能算法的效果得到了极大的体现。同时，人工智能的不断创新，又促使大量“沉睡”的数据被唤醒、被挖掘，释放出巨大价值。人机围棋对弈、无人驾驶、智能识图、机器翻译等新应用不断出现，对人们的生活产生了深远影响。在当前诸多的成熟分支中，应用比较广泛的是机器学习。

在本单元中，我们将一起进入大数据背景下的人工智能世界，了解机器是如何学习的，感受人工智能的魅力，以及人工智能对当今社会的作用和影响。



学习目标

- ◆ 通过机器学习的典型案例，了解人工智能及其强大功能和应用潜力。
- ◆ 认识人工智能在信息社会中的重要作用。

单元挑战

尝试用人工智能绘画

项目九

了解手写数字识别

——体验人工智能

提到人工智能,或许不少人会联想到战胜人类世界围棋冠军的AlphaGo,会做家务甚至会后空翻的人形机器人,工厂流水线上自动作业的工业机器人,以及各国竞相研发的无人驾驶汽车……人工智能看似遥不可及,其实它已经出现在许多日常生活应用中。

由于人工智能通常藏在各种应用的背后,很多人并不知道这些应用已经由人工智能来支持了。例如,早期的购物网站也有商品推荐栏目,但大部分都是由编辑人员人工进行配置推送的;随着人工智能的发展及应用,具有智能算法的机器逐渐替代了人工,达到了“千人千面”的“个性化”推荐效果。又如,邮政分拣“智能机器人”的问世,大幅提升了邮政业务的工作效率(图4-1)。

那么,该如何认识和理解人工智能?当前人工智能的热点之一机器学习又是怎样的?机器已经能够像人那样进行学习了吗?或许,从手写数字识别这个较简单的机器学习案例中,我们可以得到部分答案。



图 4-1 邮政分拣“智能机器人”

项目学习目标

在本项目中,我们将通过机器学习来解决手写数字识别问题,初步认识和感受人工智能。

完成本项目学习,须回答以下问题:

1. 什么是人工智能?
2. 什么是机器学习? 机器学习的一般过程是什么?
3. 计算机如何进行手写数字识别?
4. 人工智能的发展对社会的作用和影响是什么?

项目学习指引

1. 初识字符识别技术

作业不会做时，用手机启动某学习软件 App，拍下想要解决的问题并上传（图 4-2），就会收到帮助信息（图 4-3）；不认识外文菜单时，把手机摄像头对准菜单上的外文菜名，屏幕上会实时显示出翻译好的中文菜名；戴上智能眼镜逛超市，看着商品上的文字，就能自动搜索出生产商情况及在不同电商平台的价格……这些应用场景的背后都有**人工智能**中 OCR 技术的应用。OCR（Optical Character Recognition，光学字符识别）技术是指，由扫描仪、数码相机等设备捕获图像，然后用字符识别的方法将图像中的文字辨认出来，成为计算机能够处理的对象。通俗地说，OCR 就是利用光学设备去捕获图像并识别文字，将人的视觉和阅读能力延伸到机器上。



图 4-2 拍下问题并上传

OCR 技术诞生于 20 世纪 50 年代，人们利用它来实现各类文档的数字化。早期的 OCR 设备庞大而复杂，且只能识别干净背景下的某些印刷字体。20 世纪 80 年代，平板扫描仪诞生，OCR 技术可以识别的印刷字体数量也增多了，但它仍然不能辨认手写字体。智能手机普及后，越来越多的人用手机拍摄所看到的场景，此类场景中的文字识别难度远远高于平板扫描仪时期，即便拍摄的是印刷字体，也不能得到很高的识别率，更别说手写字体了。

过去几十年里，人们想出了很多字体识别的方法，手写字体的识别技术也得到了发展，但与识别支票等要求高可靠性和高识别率的现实应用仍有一定距离。近些年，随着人工智能和大数据分析的快速发展，出现了各种新技术和新算法，手写字体识别的研究及应用迎来了新的契机。

核心概念

人工智能是利用计算机或者计算机控制的机器，模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

← 参见 P137 知识链接“人工智能及其发展、应用”



图 4-3 收到帮助信息

数字化学习

《列子·汤问》记载，公元前 900 年，偃师花了数月时间，为周穆王打造了一款“会唱歌表演的艺人”——能倡者。这是我国古代人民对人工智能的早期幻想。通过网络收集资料，了解并交流我国古代人民对智能机器曾有过哪些想像和创造。

核心概念

机器学习(machine learning)是一门研究学习算法的学问,它利用已有的数据来建立模型,再用模型去解决未知的问题。

随着大数据时代的来临,大规模数据的积累被提升到了战略高度。有了长期积累的、特征丰富的数据,才能够训练出性能优越的模型,用于解决未知的问题。人工智能应用通常会不断地从经验学习中改善算法的性能,使训练出的模型变得越来越智能,越来越强大。这类经验驱动的人工智能方法就是机器学习的研究对象。

活 动

9.1 选择一种可识别文字的 App, 试用其功能。

2. 了解机器学习中的数据采集与预处理

机器学习的一般过程为:用采集到的数据进行训练,以建立一个模型,再对模型进行验证和评估,然后投入应用,如图 4-4 所示。

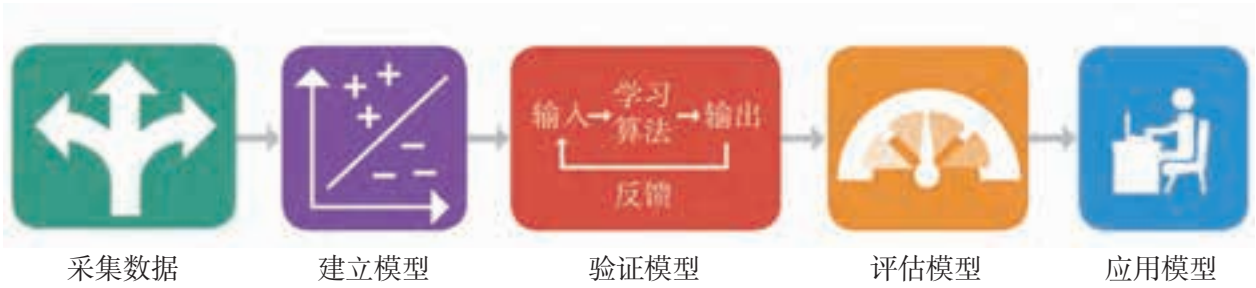


图 4-4 机器学习的一般过程

参见 P139 知识链接 ➡
“机器学习”

小贴士

标注是对采集来的数据进行类别(分类)或数值(回归)的指定。通常,这个工作须由人工完成。

(1) 数据的采集

在手写字体识别中,最简单的当属手写数字识别,但这对计算机来说却仍然颇具挑战性。针对手写数字识别,人们开发了可用于机器学习的 MNIST 数据集。作为机器学习的入门学习者,我们将直接使用现成的 MNIST 数据集。MNIST 数据集采集了 70000 张从 0 到 9 这 10 个数字的样本图片,由 250 个人手写的不同数字构成。所有图片样本都由人工完成标注,即标明它是哪个数字。

建立模型、验证模型和评估模型,都需要使用样本图片,所以采集到的 70000 张图片被分成两部分,其中 60000 张图片作为训练样本,10000 张图片作为测试样本。如图 4-5 所示,训练样本又可以再细分为 Train(训练)集合和 Validation(验证)集合。其中 Train 集合包含 55000 张图片,用于训练以建立模型。Validation 集合包含 5000 张图片,用于粗略判断训练的

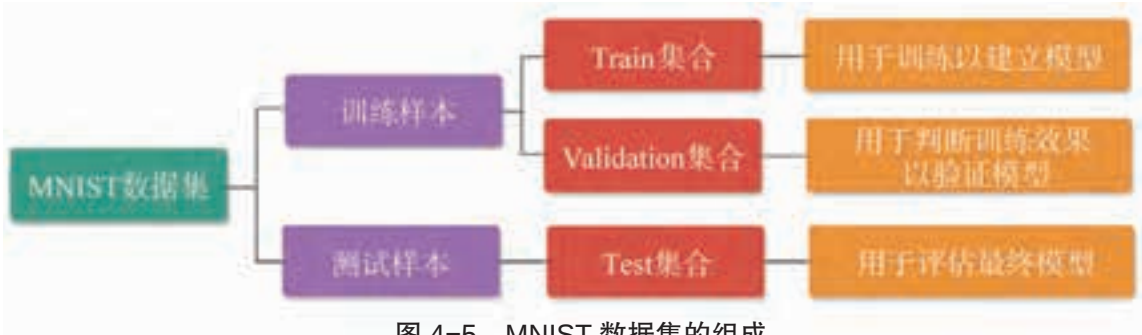


图 4-5 MNIST 数据集的组成

效果以验证模型。测试样本即 Test (测试) 集合, 用于评估最终模型的优劣, 在建立模型的训练过程中是不使用的。

(2) 数据的预处理

采集到的 70000 张样本图片, 每一张都代表了从 0 到 9 中的一个数字, 所有的数字都出现在图片的正中间 (图 4-6)。

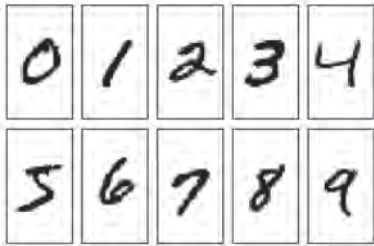


图 4-6 样本图片样例

这些图片并非二值图像, 而是灰度图像。每一张图片都由 28×28 个像素点组成, 每个像素点用一个灰度值表示。但是这样的灰度图像并不能直接用于机器学习, 必须先要对它进行向量化。

小贴士

二值图像是像素点的值只能为 0 或 1 的图像。

灰度图像是用 0 来表示白色, 用 1 来表示黑色, 中间各级别的灰色用 (0, 1) 中的相应小数来表示的图像。

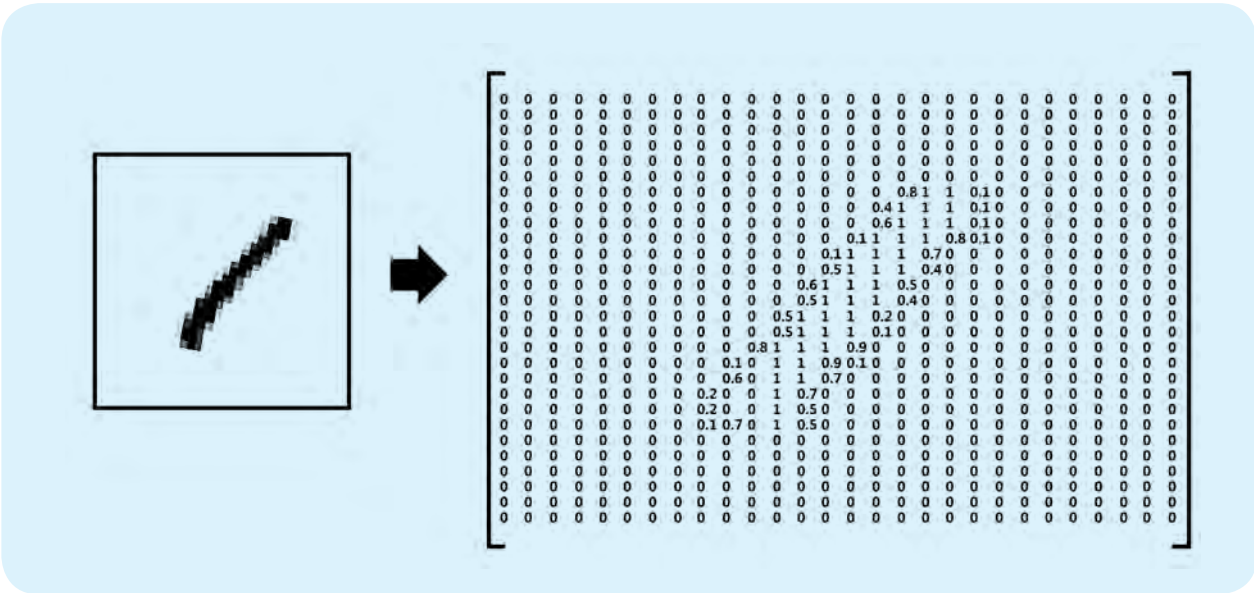


图 4-7 把灰度图像转换成矩阵

小贴士

MNIST 数据集中的图片并不是以图像格式存储的，而是被保存在二进制文件中。

数字化学习

按照配套资源上的指导，下载并安装一种第三方工具。

如图 4-7 所示，把各种灰度的像素点转换成具体的数值，就可以得到一个矩阵。把矩阵中除第一行外的每一行依次接在前一行之后，可以得到一个长度为 784 (28×28) 的一维数组，该数组中的每一个元素与图片像素矩阵中的每一个数字是一一对应的。这样就完成了图片的向量化。对所有的训练样本和测试样本进行同样的向量化处理，就得到了手写数字识别算法所需要的输入样本。

MNIST 数据集可直接在 MNIST 数据集官方网站下载。在 Python 语言中，有一些第三方工具，可用来方便地读入 MNIST 数据集，并直接完成向量化。

通过以下这行代码，可以读取 MNIST 数据集。注意先要将下载的数据集存放在指定路径中。

```
MNIST=input_data.read_data_sets("MNIST_data", one_hot=True)
```

`input_data.read_data_sets` 函数会自动将 MNIST 数据集划分为 Train、Validation 和 Test 三个数据集。设置 `one_hot=True`，可以将标注信息转化为本项目中所使用的手写数字识别算法的标准格式。

思考与讨论??

1. 为什么要将矩阵转化为一维数组？
2. 如果是二值图像，该如何完成向量化？

活 动

9.2 在 Python 语言中，运行配套资源中的相关代码，将读入的 MNIST 数据集中的二进制文件数据转换成图像格式文件，观察两者之间的对应关系，体会图像是如何被向量化的。

3. 建立手写数字识别模型并进行验证

建立手写数字识别模型的过程，就是对 Train 集合进行训练的过程，建立的模型其实就是从样本数据中学习得到的结果。建立模型时，首先要选择合适的学习算法，然后利用学习算法对样本进行训练。手写数字识别属于分类问题。对于分类问题来说，得到的这个模型就是通常所说的识别器。

机器学习发展至今，众多学者已经提出了非常多的算法，如决策树、逻辑回归等。近年来，由于大数据的不断积累和计算机性能的提升，基于多层神经网络的深度学习算法表现出更强劲的性能，在学术界和工业界受到了广泛关注。但各种算法其实并没有绝对的好坏之分，应该根据不同的应用场合，选择较适合的算法。

深度学习 (deep learning) 是机器学习的一个分支，在提取特征方面表现得很强大。它可以从数据中学习得到一些由简单特征组合而成的复杂特征，使后面的学习更加有效。AlphaGo 战胜人类职业围棋棋手的原因就在于其拥有的强大深度学习和搜索运算能力 (图 4-8)。



图 4-8 拥有强大深度学习和搜索运算能力的 AlphaGo

深度学习的主要算法之一——卷积神经网络 (Convolutional Neural Network, CNN) 在图像分类上有非常出色的表现。因此，这里将用 Python 语言通过卷积神经网络来建立和验证模型。

训练过程的主要代码如下所示。

```
for i in range(10000):
    batch = MNIST.train.next_batch(100)
    if i%200 == 0:
        print(" 当前迭代次数为 %d, 训练正确率为 %g"%(i, accuracy.eval
            (feed_dict={x:batch[0], y_: batch[1], keep_prob: 1.0})))
        train_step.run(feed_dict={x: batch[0], y_: batch[1], keep_prob: 0.5})
```

小贴士

对机器进行训练前，大部分学习算法须设置一些参数。参数设置不同，经训练样本而建立的模型有可能效果截然不同，因此须反复调整参数。

小贴士

深度学习中，小批量的训练样本被不断地送入到卷积神经网络中进行训练，然后利用验证样本测试当前迭代次数下的识别正确率，并根据结果调整参数，以寻求更优的模型。

第 1 行代码，设置了一个参数，即总的迭代次数为 10000 次；第 2 行代码，设置了另一个参数，即每次送入卷积神经网络中进行训练的图片数为 100 张；第 3 至 5 行代码，设置了第三个参数，即每间隔 200 次迭代，在控制台输出当前迭代次数下模型的识别正确率；最后一行代码，`train_step.run` 是最关键的进行训练的代码，它针对当前批次的样本进行训练，得出相应模型。训练结果如图 4-9 所示。

当前迭代次数为 0, 训练正确率为 0.08
当前迭代次数为 200, 训练正确率为 0.88
当前迭代次数为 400, 训练正确率为 0.94
当前迭代次数为 600, 训练正确率为 0.95
当前迭代次数为 800, 训练正确率为 0.95
当前迭代次数为 1000, 训练正确率为 0.96
当前迭代次数为 1200, 训练正确率为 0.96
当前迭代次数为 1400, 训练正确率为 0.99

图 4-9 某次训练的前 1400 次迭代的正确率

从图 4-9 可以看到，随着训练的不断继续，模型的识别正确率也不断得到提升。当前迭代次数下模型的识别正确率是通过 Validation 集合计算得到的，这个过程就是验证模型。

由于卷积神经网络模型训练过程中选取训练样本的随机性，每次得到的结果不会完全相同。本代码经过 10000 次迭代，可以得到能识别未知数字的最终模型。

思考与讨论??

1. 为什么建立模型的过程需要进行反复迭代？
2. 预设的迭代次数与最终模型之间有什么关系？

活 动

9.3 以小组为单位，利用配套资源中的相关代码，尝试设置不同的参数，运行程序并观察控制台输出的结果，研究不同的参数设置对训练出来的模型的正确率有何不同影响。

4. 评估手写数字识别模型并开展应用

整个训练结束之后,可以用 Test 集合评估最终模型的识别正确率。每个测试样本送入模型后,模型会识别出一个结果。如果这个结果和标注信息一致,就是正确的;反之,就是错误的。通过以下代码可以输出最终模型对 10000 张测试图片进行识别的正确率。迭代次数为 10000 次时,最终模型的识别正确率通常在 99% 以上。

```
print(" 测试正确率为 %g"%accuracy.eval(feed_dict={  
    x: MNIST.test.images, y_: MNIST.test.labels, keep_prob: 1.0}))
```

识别正确率经评估达到要求的模型,就可以作为识别器,应用于对未知样本的识别。具体过程如图 4-10 所示。

小贴士

未知样本也需要经过同样的向量化才能够被识别。

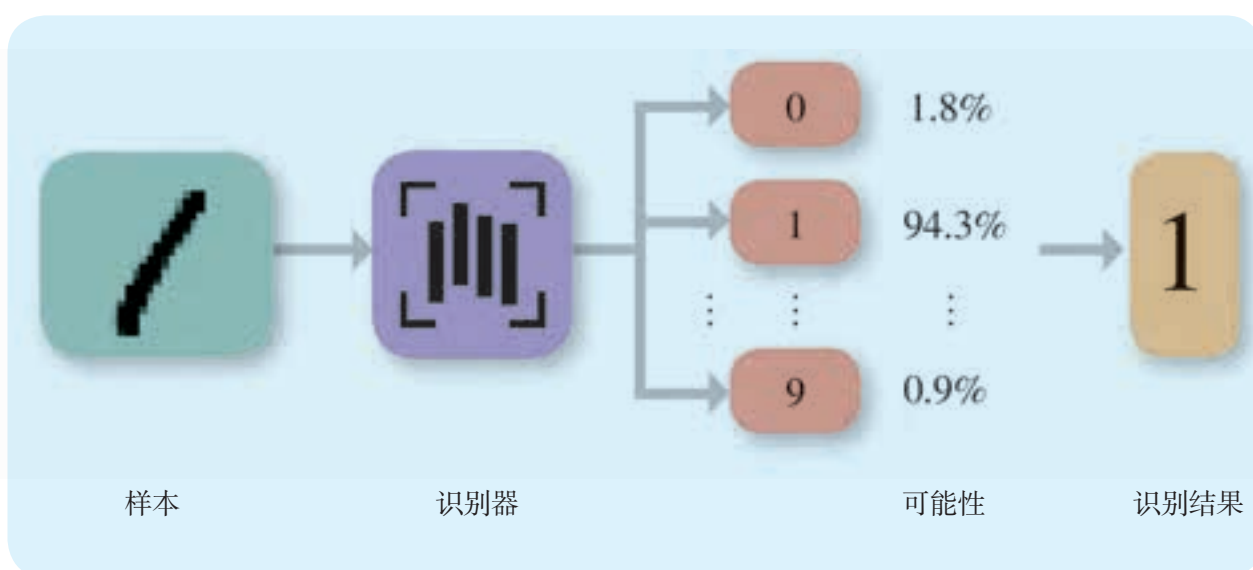


图 4-10 未知样本的识别过程

思考与讨论??

1. 为什么需要对样本进行标注? 训练样本和测试样本的标注各有什么用处?
2. 除了识别正确率外,还有什么指标可以评价模型的优劣?
3. 用 Test 集合评估出的正确率就是现实情况中的正确率吗?



图 4-11 指纹识别



图 4-12 花卉识别

参见 P140 知识链接
“人工智能的作用和影
响”

手写数字识别是一个分类问题，一共有 10 个类别。在深度学习的卷积神经网络中，一般不会直接给出该样本是哪个数字的结论，而是会给出该样本是各个数字的可能性即概率分别有多大，然后根据实际需要去判定结果。一般来说，概率最大的那个类别就会被认为是当前样本的识别结果。

手写数字识别已经应用于邮政编码、财务报表、银行票据等的自动识别，极大地节省了人力和物力，提升了工作效率。手写汉字识别也取得了重大突破，如可识别手写汉字地址的“基于多源信息融合的邮件分拣关键技术及应用”就荣获了 2015 年上海市科技进步一等奖，被广泛应用于邮件的自动分拣。

近年来，除了手写数字识别和手写汉字识别外，人脸识别、指纹识别、物体识别等图像识别技术也均有了较大发展，识别准确性得到大幅提升，如图 4-11 和图 4-12 所示。这些识别技术在导航、家居安防、生理病变研究、天气预报、环境监测等许多领域都有着广泛应用，既推动了相关领域的智能化跃升，也悄然改变着人们的生活方式和思维方式。

思考与讨论??

1. 日常生活中，你还知道哪些应用或产品的背后有着人工智能的支持？
2. 人工智能在推进社会发展的同时，也带来了挑战。作为学生，应该如何规划人工智能时代的未来生活？应该注重哪些技能、素养的学习和培养？可以采用哪些有效的学习方式？

活 动

9.4 探究无人驾驶汽车的伦理问题。

无人驾驶汽车的发展，除了技术因素外，还面临着很多伦理挑战。例如，车辆在无法避免事故的极端条件下，是优先保护车内乘客还是优先保护路上行人？如果车辆在无人驾驶过程中发生事故，应该由制造商、车内乘客还是车主来承担责任？请以小组为单位，选择一个伦理问题开展探究。每名小组成员代表一种身份（如汽车制造商、车主、车内乘客、路上行人、交通警察、保险公司员工等），从各自的身份和立场出发，阐述自己对该问题的理解，并提出解决方案。

知识链接

人工智能及其发展、应用

1. 什么是人工智能

人工智能的定义有很多，根据中国《人工智能标准化白皮书(2018)》中的定义，人工智能是利用计算机或者计算机控制的机器，模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

尽管不同专家对人工智能的认识有一定的差异，但目标都是希望造出像人类那样，甚至具有比人类更好的感知、思维、学习和行为能力的机器。

2. 人工智能的发展阶段

从技术角度，通常把人工智能的主要发展阶段分为运算智能、感知智能和认知智能，如图 4-13 所示。早期的人工智能是运算智能，即具有快速运算和记忆存储的能力。现今，由于算法的优化和硬件技术的进步，计算机的快速运算能力和记忆存储能力早已超过人类。



图 4-13 人工智能的发展阶段

当前大数据时代的人工智能是感知智能，即具有“视觉”“听觉”等感知能力，并具有与周围环境进行互动的能力。随着计算机处理能力的不断提升和大数据的不断积累，再加上深度学习算法在数据训练上取得的进展，在运算能力、数据、算法三大要素发展的推动下，计算机在某些领域的感知智能已越来越接近于人类，在医学影像读片等特殊任务中甚至已经超越了人类。迄今为止，感知智能在语音识别和机器视觉等方面研究较多，成果较显著。

从感知智能向认知智能方向迈进，已成为人工智能未来的发展趋势。尽管认知智能方面目前存在大量难以解决、有待探索的问题，但可以期待的是，到了认知智能阶段，机器将能够主动思考并采取行动，辅助甚至替代人类的工作，完全独立操作的无人驾驶汽车、自主行动的机器人等将会出现。

3. 当前人工智能的主要应用领域

现今，人工智能已成为新一轮科技革命和产业革命的核心驱动力，催生了大量新的技术、产品、产业。在语音识别、机器视觉、数据挖掘等多个领域，人工智能已走进了实际应用场合。人工智能与各行业领域的深度融合，将重塑甚至改变传统行业。据统计，当前人工智能的主要应用集中在个人助理、安防、自动驾驶、医疗健康、电商零售、金融、教育等方面，如图 4-14 所示。

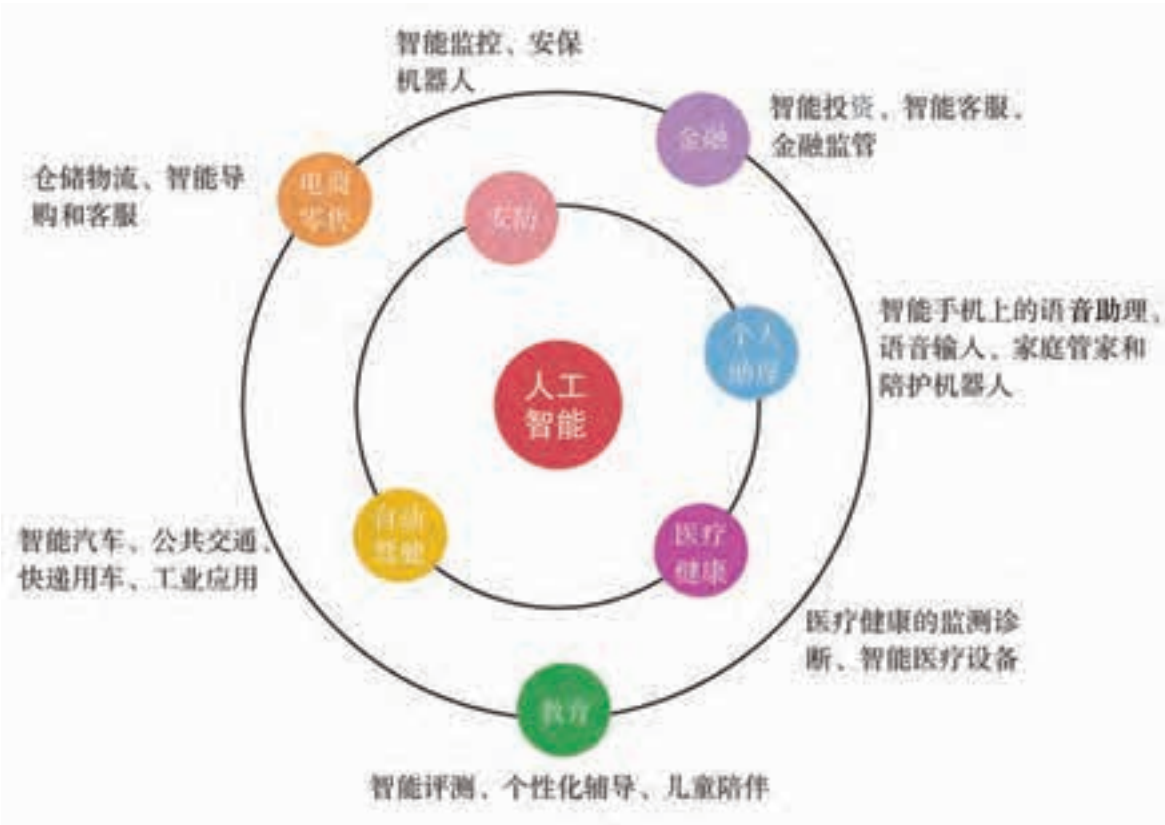


图 4-14 当前人工智能的主要应用领域

机器学习

机器学习是人工智能的核心研究领域之一，它研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构，从而不断改善自身的性能。

人类学习的机理，一般认为是通过积累并利用经验来构造或修改对于所经历事物的判断。机器学习正是从人类学习中受到启发：机器要学会某一个概念，可以通过从数据中获取一些规则性的东西，来形成一个可用于判断未知事物的模型。实际上，机器学习是一门研究学习算法的学问，它利用已有的数据来建立模型，再用模型去解决未知的问题。和一般计算机程序中的算法不同的是，学习算法具有不断改善自身性能的能力，从而使机器的能力变得越来越强，甚至算法设计者本人也不能预测机器的能力可以达到何种程度。特别是在大数据的背景下，用大量长期积累的数据学习得到的模型，已经表现出了优越的性能。数据为王，未来的机器其能力将以数据为依托，提升空间巨大。

近年来，机器学习在许多应用领域中均发挥了重要作用，极大地推动了社会的进步，如图 4-15 所示。



图 4-15 机器学习在不同产业中的应用

在机器学习中，要解决的问题可以分为分类（识别）和回归（预测）两大类。

分类问题是指，给定一个新的模式，根据训练集推断它所对应的类别是什么，这是一种定性输出，也叫离散变量预测。指纹识别、人脸识别、图像分类等都属于分类问题。

回归问题是指，给定一个新的模式，根据训练集推断它所对应的输出值是多少，这是一种定量输出，也叫连续变量预测。天气预报、市场预测、金融分析等都属于回归问题。

根据已有数据的不同,机器学习又可以分为监督学习和无监督学习。针对已被标注的数据,采用监督学习;针对未被标注的数据,采用无监督学习。目前,监督学习已在工业界取得了许多成功应用。无监督学习是众多科研工作者一直在努力的研究方向。如果在无监督学习方向能有所突破,将会极大地推动机器学习的发展,产生巨大的社会效益。

人工智能的作用和影响

人工智能是引领未来的战略性技术,是国际竞争的新焦点。世界主要发达国家都已将发展人工智能作为提升国家竞争力、维护国家安全的重大战略。中国也已围绕人工智能出台了相关规划和政策,对人工智能核心技术、顶尖人才、标准规范等进行了部署,以加快促进人工智能的技术和产业的发展。

人工智能将人从枯燥的劳动中解放出来,越来越多简单、重复、危险的任务可以由机器完成,它还能够比人做得更快、更准确。一些程式化、重复性、仅靠记忆与练习就可以掌握的工作,未来将逐步被智能机器取代或改变。例如,在一些医院已经开始由手术机器人为病人做手术(图 4-16)。又如,当机器翻译取得根本性突破后,大多数人类翻译工作可能会被机器部分或完全取代。人工智能在教育、医疗、金融、养老、环境保护、城市运行等领域的广泛应用,能够全面提升人们的生活品质,引发经济结构的重大变革,实现社会生产力的整体跃升,并给人们的生产生活方式和思维模式带来革命性变化。

但是,人工智能又是一种影响面极广的颠覆性技术,它可能带来改变就业结构、冲击法律与社会伦理、侵犯个人隐私、挑战国际关系准则等问题。例如,设计者和生产者在开发人工智能产品的过程中,并不能准确预知被开发产品可能存在的风险。又如,用于安防或金融的人工智能系统一旦遭到攻击,其学习算法或决策结果被破坏或篡改,后果可能是毁灭性的。近期人工智能的发展建立在大数据的信息技术应用之上,这又不可避免地涉及个人信息的合理使用问题,而且人工智能技术的发展也让侵犯个人隐私的行为变得更为便利。如何保障个人隐私不被侵犯,已写入议事日程。



图 4-16 人工智能将改变未来医疗保健的方式

拓展阅读

上海：全面实施 AI@SH 行动，向智能未来城进发

算法突破、数据爆发和计算能力提升，让计算机视觉、自然语言理解等走出实验室，全球加速迈进智能时代。

2017 年 11 月 14 日，上海推动新一代人工智能发展实施意见正式出台。上海将发挥比较优势，围绕“AI 优先”这一科技创新“新引擎”发力。



图 4-17 全球第一块用于智能手机的 AI 中国“芯”——麒麟 970

数据是人工智能的“燃料”，上海也有比较优势。上海已建成覆盖 2400 万常住人口、200 多万家企业以及涵盖全市域的人口、法人、空间地理基础数据库；医联数据共享系统拥有 250 亿条数据，交通数据流量每日新增 30GB；大数据交易中心日均数据交易量 3000 万条，占全国 50% 左右。

就信息化基础而言，上海结合智慧城市建设，着力建设全球领先的千兆宽带接入网和覆盖全城的窄带物联网，依托智能制造、工业互联网、智能网联汽车、“互联网+政务服务”等，构建起各类丰富的感知系统和应用场景。

上海将全面实施“智能上海（AI@SH）”行动，形成应用驱动、科技引领、产业协同、生态培育、人才集聚的新一代人工智能发展体系，推动人工智能成为上海建设“四个中心”和具有全球影响力的科技创新中心的新引擎，为上海建设卓越的全球城市注入新动能。

——摘自新华社 2017-11-15 电讯

单元挑战 尝试用人工智能绘画

一、项目任务

想不想把自己的照片变成梵高画作的风格？神经网络科学家发现，利用卷积神经网络可以让计算机模仿任何画家的绘画风格，重新绘制一张图片，这种方法叫作“图像风格迁移”（Neural Style）。在本项目中，让我们利用 Neural Style，制作出各种画风的图片。



图 4-18 原图



图 4-19 模仿的图



图 4-20 结果图

二、项目指引

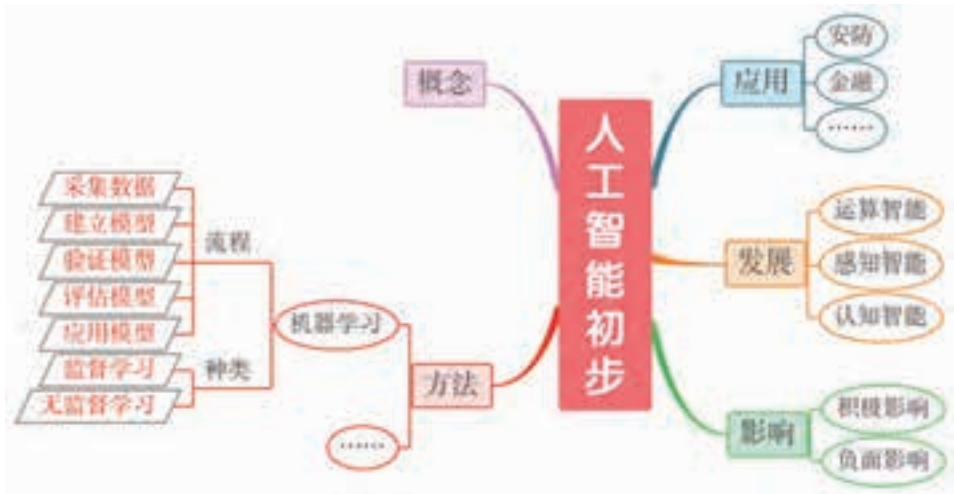
1. 以小组为单位，收集几张用于转换风格的图片（类似图 4-18），以及几张不同流派画家的作品图片（类似图 4-19）。
2. 从网上下载相关的 Python 代码及 Pre-trained VGG network 模型文件，并在 Python 环境中安装相关的第三方工具。
3. 准备就绪后，运行相关代码，处理图片。
4. 尝试设置不同的迭代次数，得到不同的效果。比比谁制作的图片的风格更接近所模仿画家的绘画风格。
5. 设置一些不同时间段的确认点，保存中间处理过程中的图片，看看到底发生了什么。

三、交流评价与反思

向同学展示自己制作出来的图片，谈谈心得与体会，并讨论如何才能做出风格更接近被模仿画家的图片。探究不同图片效果产生的原因，并讨论为什么处理过程所需时间会较长。如果条件允许，可以在具备 GPU 的计算机中再次运行代码，看看会有什么不一样。

单元小结

一、主要内容梳理



二、单元练习

1. 随着手机支付的日益流行，人们对智能手机的安全性也越来越重视。保障手机安全的最基本措施是为手机设置密码。目前，除了数字密码、手势密码外，指纹识别作为密码工具已在很多手机上出现。

- (1) 尝试在手机上设置自己的指纹密码，并用指纹密码解锁手机。
- (2) 与数字密码、手势密码相比，指纹密码的优点是什么？缺点又是什么？
- (3) 结合本项目中的手写数字识别技术，思考指纹识别技术是如何实现的。

2. 聊天机器人是一个用来模拟人类对话或聊天的程序。随着人工智能技术的不断发展，聊天机器人变得越来越智能，可以与人们进行正常对话，听从人们的指令做各种事情。例如，智能手机或智能音箱中的聊天机器人，能按人的指令播放歌曲、控制家中的智能家电等。

- (1) 调查目前几种主流的聊天机器人，完成一份调查报告，列出各聊天机器人的功能和特点，指出这些功能背后采用了什么样的技术，在实际应用中存在怎样的问题。
- (2) 为了提升效率、节省人力成本，聊天机器人已成为各大平台客服系统中不可或缺的一部分。访问几个主流平台，与客服机器人对话，完成一份体验报告，对比使用体验及聊天机器人的智能程度。

三、单元评价

评价内容	达成情况
能理解人工智能的概念(A、T)	
能列举生活中人工智能的应用实例,剖析人工智能在该应用中的具体表现(A、T)	
了解人工智能的发展阶段(A、I)	
知道机器学习的概念和机理(A、T)	
能描述机器学习的一般过程(T)	
能理解智能信息处理的巨大进步和应用潜力(A)	
能通过资料收集和实例分析,知道人工智能与大数据之间的关系(A、T、I)	
能正确分析人工智能发展对社会的作用和影响(A、R)	

说明: A—信息意识, T—计算思维, I—数字化学习与创新, R—信息社会责任

附录

部分名词术语中英文对照

(以汉语拼音字母次序为序)

ASCII 码	American Standard Code for Information Interchange	结构化数据	structured data
编码	encoding	卷积神经网络	Convolutional Neural Network, CNN
编译	compile	量化	quantization
变量	variable	人工智能	Artificial Intelligence, AI
表	table	深度学习	deep learning
采样	sampling	十进制	decimal system
常量	constant	数据	data
程序	program	数据采集	data acquisition
程序设计	programming	数据处理	data processing
程序设计语言	programming language	数据分析	data analysis
大数据	big data	数据结构	data structure
二进制	binary system	数据可视化	data visualization
非结构化数据	unstructured data	数据库	database
高级语言	high-level language	顺序结构	sequence structure
公式	formula	算法	algorithm
关系表	relational table	调试	debug
函数	function	信息	information
汇编语言	assembly language	选择结构	selection structure
机器学习	machine learning	循环结构	loop structure
机器语言	machine language	知识	knowledge
记录	record	字段	field

致 谢

在本书的研究、编写过程中，得到了许多专家学者、学校领导、教学研究人员和广大教师的大力帮助与支持。黑龙江省实验中学、黑龙江省哈尔滨第三中学、上海市上海中学组织了对本册教材的试用，并提出宝贵意见和建议。王伟、冯聪、孟云、陆花、杨杰、施金花等为本书提供了丰富的配套数字资源，唐思贤、束炯等为本书提供案例和指导。还有一些专家在审稿过程中提出了许多宝贵意见，对本书的完善做出了贡献。在此，特向关心本书及为本书提供帮助的所有同志表示诚挚的感谢。

编写组

2018年2月

普通高中教科书
信息技术 必修1
数据与计算

上海科技教育出版社有限公司出版发行

(上海市柳州路218号 邮政编码200235)

湖南省新华书店经销 湖南长沙鸿发印务实业有限公司印刷

开本 890×1240 1/16 印张9.5

2020年8月第1版 2021年8月第2次印刷

ISBN 978-7-5428-7320-0/G·4292

定价:11.75元

批准文号:湘发改价费[2017]343号 举报电话:12315



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



致力于用榜样的力量提升学生成绩的共享家教平台

中国家庭教育学会荣誉会员单位

985/211 大学生 1对1 上门辅导

找家教就像叫“代驾”一样简单
家长们都在偷偷用的家教预约神器

记得拍照留存哦



扫码关注 预约上门

关注送200元优惠券

小初高全科辅导

学霸云集任您挑

学历真实可担保



与优秀大学生同行，激发孩子无限潜能



微信搜索公众号：365优教网

咨询热线：4000-711-365

YOUJ 优教

既是找老师，更是找榜样

家教老师全国招募中