国家教材委员会专家委员会审核通过 2019

普通高中教科书

信息技术

必修1

数据与计算

浙江教育出版社
ZHEJIANG EDUCATION PUBLISHING HOUSE

普通高中教科书

信息技术

必修1

数据与计算

闫寒冰　主编

浙江教育出版社·杭州

主　　编：闫寒冰
副 主 编：赵　健　魏雄鹰

执行主编：吴建锋
编写人员（按姓氏笔画排列）：
吴建锋　何海源　陈　跃　陈军辉
邵红祥　夏祎华

前　言

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

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

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

具体而言，本套教材体现了如下特点：

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

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

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

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

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

本册教材为必修《数据与计算》，是信息技术课程后续学习的基础。通过本教材的学习，期待同学们能认识到数据在信息社会中的重要价值，合理处理与应用数据，掌握算法与程序设计的基本知识，根据需要运用数字化工具解决生活与学习中的问题，认识到人工智能在信息社会中越来越重要的作用，逐步成为信息社会的积极参与者。

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

本册教材在编写过程中得到了各方面的大力支持。北京大学计算机系李晓明教授、浙江大学计算机学院卜佳俊教授和翁恺教授、北京航空航天大学欧阳元新副教授在百忙之中对书稿内容进行了审阅并提出了宝贵的修改意见。

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

目 录
MULU

第五章 人工智能及应用

第一章 1

数据与信息

在日常生活中，人们每天都要与数据打交道，数据也时刻影响着人们的选择。早晨醒来，过低的气温数据意味着该适当地添加衣物；出门远行，交通工具的选择可能取决于目的地的距离和预计到达的时间。

随着人类进入信息社会，数据已融入社会的各个方面。面对日益庞大的各种数据，传统的人工处理方式已力不从心，更多时候需要依靠计算机来处理这些数据，计算机已成为处理数据最主要的工具。

互联网的普及与传感器的大规模应用，使得人们可以从更多的途径获取数据。数据量的快速增长、数据形式的多元化、数据的时时变化，催生了大数据技术。现在，大数据正影响着人们的生活，人类社会已经进入了大数据时代。

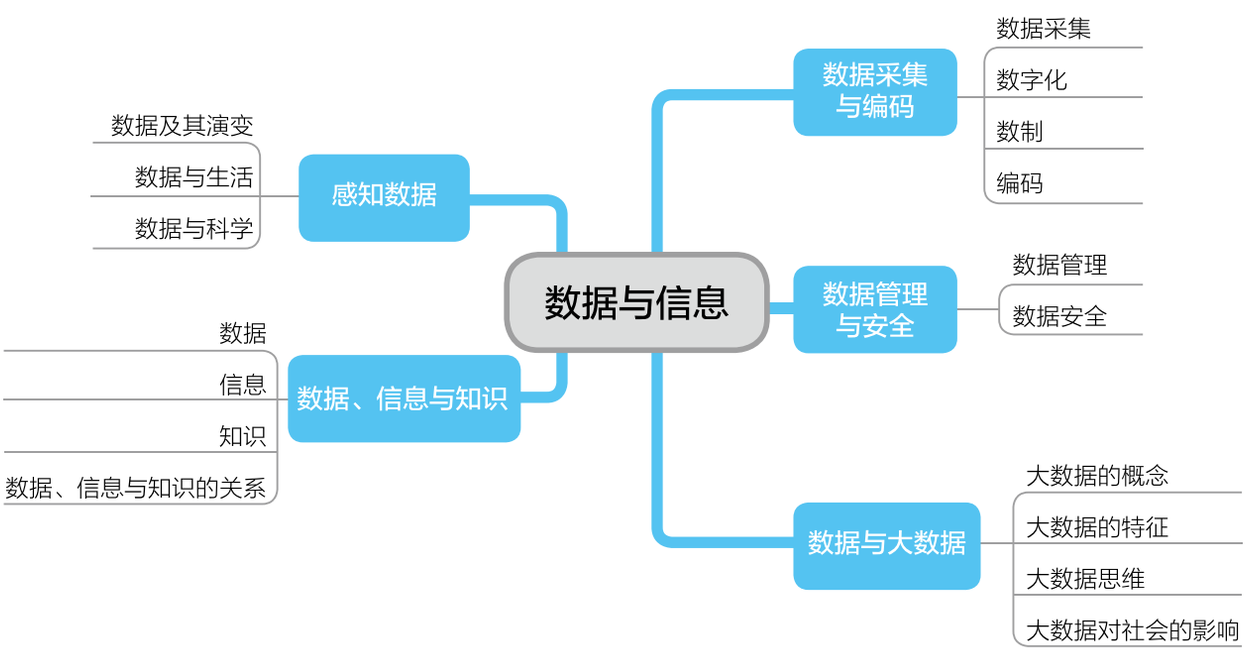
问题与挑战

- 数、数字、数据、信息、知识、智能、智慧，这些常见的名词之间是怎样的关系？它们与技术的发展又有怎样的关系？
- 电影《火星救援》中的一个片段：主人公Mark Watney被困火星，他想与地球取得联系，于是找到了美国于1996年发射到火星的探测器——火星探路者，Watney巧妙利用十六进制与ASCII码，成功地与地球建立了联系。那么十六进制与平时熟悉的十进制之间是怎样的关系？ ASCII码又是如何表示信息的？
- 千百年来，人类在不断摸索天气变化规律，在长期的生活实践中形成了各种有关天气的谚语。这种对天气的预测，主要依据经验做出判断，其准确性往往不高。随着社会的发展，人们可以轻易通过各种媒体获取天气预报信息，甚至能知道几分钟后的天气情况，不仅如此，人们发现天气预报的准确率也在提高。那么，天气预报准确率的提高，数据起到了怎样的作用？

学习目标

1. 认识数据对人们生活的影响。
2. 能辨别数据与信息，描述数据与信息的特征。
3. 理解数据、信息与知识的相互关系。
4. 感受数字化工具和资源对学习和生活的作用。
5. 知道数据编码的基本方式。
6. 了解数据管理的基本方法，理解对数据进行保护的意义。

内容总览



1.1 感知数据

当今社会，数据体现出了前所未有的价值。每时每刻都有各种数据被人们发现、分析、利用，并创造出巨大的财富。数据改变着人们的生活、学习、工作方式，而数据的种类与形式也在不知不觉中变化着。

1.1.1 数据及其演变

早在远古时代，人们在长期的社会实践中就逐渐形成了数的概念。为了记数，居住在洞穴中的原始人就用石器或骨器在墙壁上刻画图案，这些图案就是最原始的“数据”。后来，人们发明了结绳记事的方式来记事或记数（如图1.1.1）。据《周易·系辞下》记载：“上古结绳而治，后世圣人易之以书契，百官以治，万民以察。”结绳记事的方法现已失传，但通过《易九家言》中“事大，大结其绳，事小，小结其绳，结之多少，随物众寡”的描述，大致可以了解到通过“大事大结，小事小结”来记录各种不同的事件和数量的情形。这些大小不一、数量不同的绳结也是“数据”。

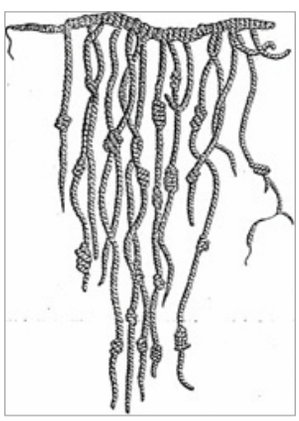


图1.1.1 结绳记事

随着文字与数字的出现，数据以更加明确的形式被记录下来。图1.1.2所示的是一块公元前3000多年的泥板，为当时生活在美索不达米亚地区的苏美尔人所遗留，泥板上以楔形文字记载的内容为“29086单位大麦37个月库辛”。这句话的意思明显与数据有关。

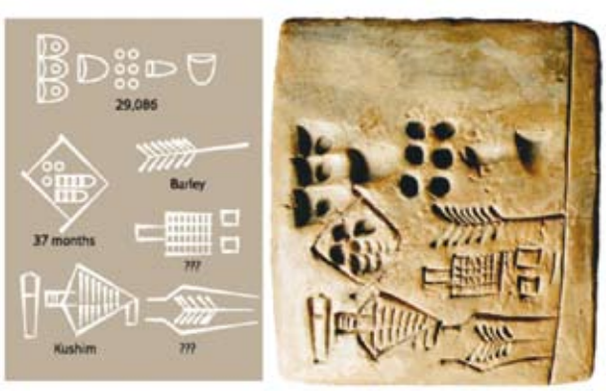


图1.1.2 楔形文字记载的内容

人类对于数据的利用在一定程度上依赖于载体，造纸术与印刷术的发明，使得数据的记录与传输变得便捷高效，直至今日，这些技术还在发挥着重要作用。

技术的发展为人们提供了更多记录数据的载体，除了文本形式的数据，图像与声音的记录方式也在发生变化。最早的图像由手绘而成，画家可以凭借高超的绘画技艺画出逼真的场景。摄影术的发明使得保存真实的图像变得更加便捷，某一时刻的真实情景可以快速地以照片的形式被记录下来。相对于图像，声音的保存则比较困难。直到19世纪，爱迪生发明了留声机，才得以将声音记录下来。

到了现代，数据的记录形式越来越多样化，数据量也不断地增长，手工处理数据的方式已经无法满足数据处理的需求，于是人们发明了各种各样的工具来协助处理。这些工具中最具代表性的是诞生于20世纪40年代的电子计算机，短短的几十年，计算机已成为数据处理的主要工具。互联网技术的发展，加速了数据的传输与处理；随着移动网络与传感器的普及，大数据进入了人们生活的方方面面。

1.1.2　数据与生活

在日常生活中，人们每天都在使用数据。比如在超市购物，结账后一般会收到购物清单，上面列出了本次购买的商品与价格；购买的火车票上有时间、目的地、身份证号码等数据（如图1.1.3）。

图1.1.3　纸质火车票

相比于这些传统数据，在互联网时代，人们生活中的数据形式也在发生改变。云计算、物联网、大数据等技术陆续融入生活。数据的采集技术迅速发展，数据的表现形式也越来越多样化，给人们的生活习惯带来了巨大的转变。例如，以前人们每到一个新的城市，可能会购买一册当地最新版的地图，循着纸质地图上的路线来熟悉这个陌生的城市。现在则可以通过电子地图，预先了解目的地及周边的相关情况，通过电子地图的全景模式，可以全方位观察周围的环境，获得身临其境的体验，如图1.1.4所示。

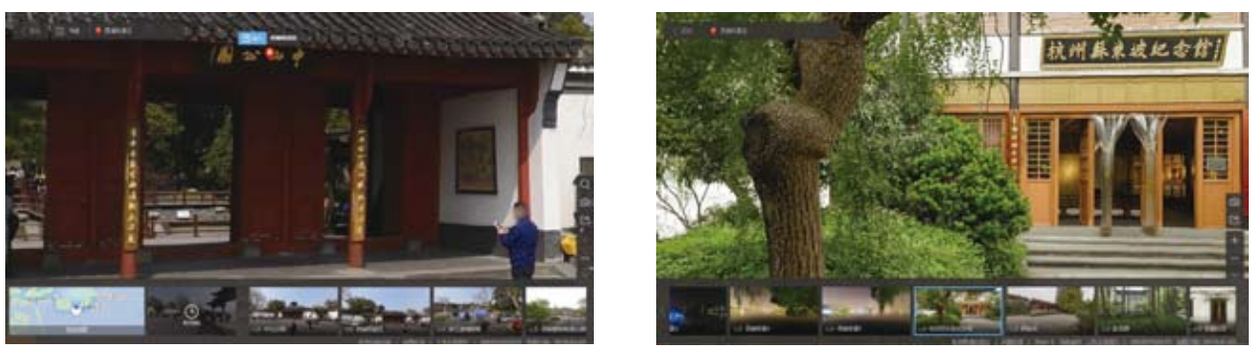


图1.1.4　景点的全景地图

移动网络的升级大大扩展了手机的应用领域，手机的功能已不再局限于通话与短信，智能手机的各种应用给生活带来极大的便利。人们通过手机上的购物平台与电子支付系统，足不出户就可以购买到各种商品。同时，各种票据如机票、车票的形式也在转变，电子票据使得人们可以不必手持纸质票据就能乘坐交通工具。城市中的出租车服务体系也变得多样化，除了传统的巡游出租汽车，网约车（网络预约出租汽车）作为另一种营运模式，为乘客提供了更多的选择，人们通过手机上的各种网约车APP（如图1.1.5），可以方便地完成预约、支付等操作。

图1.1.5　网约车APP

大多数数据会随着时间的推移而变化，为了更直观地了解数据的变化情况，不同的领域会以各自独特的形式予以展示。如股票交易中心的股票走势图，通过曲线、颜色等元素，使得股票的涨跌情况一目了然。

人们在利用数据的同时，自身的行为也在产生数据。如每年的春运都会有巨大的客运流量，根据这些乘客的出行情况制作出的迁徙图，可以形象地反映春运客流情况。

互联网加速了数据的传递，通过与传统行业的深度融合，发展成了一种新的社会经济形态——“互联网+”，给传统行业注入了活力。如“互联网+农业”就是其中一个案例。

互联网+农业

我国是一个农业大国，农业作为关系着国计民生的基础性产业，在与“互联网+”的深度融合下，农资准备及农产品生产、流通和消费等各个环节都有新技术的应用，其模式发生了巨大的变化。

在此之前，由于产销信息不对称，销售途径单一，我国曾多地出现鲜活农产品滞销、买贵卖难的现象。一方面各种农产品轮番涨价，另一方面部分农产品丰产滞销现象频现，“滞销、卖难、买贵”形成一个怪圈，使得农民“丰产”难增收。互联网技术为农产品销售搭建交易平台，将产销之间的距离大大拉近，减少层级，降低成本，形成扁平化交易模式，同时也实现了产销的充分对接，大大减少了生产的盲目性，拓宽了销售的视野。电商与物流的结合使得农产品不再难卖，“订单式种植”可以规避各种风险，这一切都得益于数据的充分利用。移动互联网的发展更使得农产品的购买与销售可以随时随地进行。

“互联网+”不仅仅改变了农业的生产与销售模式，也在影响着社会的各行各业。“互联网+”充分发挥互联网在社会资源配置中的优化和集成作用，提升了全社会的创新力和生产力。

1.1.3 数据与科学

自古以来，人们通过观察与实践，获得了大量数据，这些数据不仅在生产与生活中发挥了作用，而且为一些早期科学成果的取得打下了基础，如古人根据月相变化和季节更替的规律逐渐形成了我国特有的传统历法——农历，直至今日，在人们的生活中也经常用到农历。

科学研究离不开数据。科学强调证据，而数据的客观性正好为科学研究提供了可靠的依据。如天王星被发现后，天文学家发现它的运行轨道总是偏离根据万有引力定律计算出的路线，经过仔细计算，从而推算出了影响天王星的那颗未知星体——海王星。

现在各国都很重视高精尖实验室的建设，花费巨额经费来购买和研发实验设备，就是为了获取某些数据。例如，世界上最大的粒子物理学实验室——欧洲核子研究组织（通常简称为CERN），为高能物理学的研究提供宝贵的实验数据。CERN把大量的实验数据进行全球共享，让全世界的科学家和公众一起研究。迄今为止，CERN已经取得了多项巨大的科学成就。

拓展链接

欧洲核子研究组织

欧洲核子研究组织是万维网的发源地。它成立于1954年9月29日，总部位于瑞士日内瓦西北部郊区，目前有二十几个成员国。作为科学实验基地，每天为科学界提供大量的实验数据。

CERN 也被用来称呼它的实验室，其主要功能是满足高能物理学研究的需要，提供粒子加速器和其他基础设施，以进行许多国际合作的实验。同时也设立了资料处理能力很强的大型计算机中心，不仅可以协助分析实验数据，还可以供其他地区的研究员使用，形成了一个庞大的网络中枢。

思考与练习

1. 在成长的道路上，我们已经留下了很多痕迹，这些痕迹大都可用数据来记录。哪些数据可以大致描述你的成长轨迹呢？

2. 除了数字，在日常生活中还有哪些形式的数据？

1.2 数据、信息与知识

数据来源的不同，决定了数据的多样性。从各个途径获取的大量的、庞杂的数据，需要经过一定的处理，才能从中提取有意义、有价值的内容。在人类发展的历史长河中，人们通过处理数据、分析数据，从中寻找规律，积累了丰富的知识，成为人类社会的宝贵财富。

1.2.1 数据

数据是对客观事物的符号表示，如图形符号、数字、字母等。其中，数字是最简单的一种数据，是对数据的一种传统和狭义的理解。

单纯的数据是没有意义的，因为数据的表现形式还不能完全表达其内容，经过解释，数据才变得有意义，数据和关于数据的解释是密不可分的。

随着人类社会进入数字化时代，计算机被广泛使用。数据的种类与表现形式也越来越多样化，数据在采集的方式、处理的速度等方面都有了质的飞跃，数据的含义也得到了扩展。

在计算机科学中，数据是指所有能输入到计算机并被计算机程序处理的符号总称，是用于输入到计算机中进行处理，具有一定意义的数字、字母、符号和模拟量等的通称。其表现形式可以是文字、图形、图像、音频、视频等。例如，人们在网上预订车票时，余票的数量是数据，座位等级也是数据；观看在线影视时，点播的视频就是数据；而一个U盘、一张光盘，其存储的文件也统称为数据。

1.2.2 信息

信息自古就有，人类的生活一直与信息密切相关，人类通过了解信息来认识自然，利用信息进行发明创造。

1. 信息的定义

到目前为止，信息还未有统一的定义，出于研究目的、观察角度的不同，不同的学科往往有自己的定义。信息论的奠基者克劳德·艾尔伍德·香农（Claude Elwood Shannon）在《通信的数学理论》中提出：“信息是用来消除随机不确定性的东西。”这一定义常被人们看作是经典性定义并加以引用。

尽管不同的人对信息的定义可能有所不同，但所指向的都是同一对象，这些对象有些能被直接感受，有些需要借助设备或其他事物才能被感受。

比如向朋友介绍某款新车，对品牌、颜色、功率、内饰等属性描述得越多、越准确，朋友对该款新车的认识就越全面，即消除的随机不确定性越多。

2. 信息的特征

（1）载体依附性

信息是不能独立存在的，必须依附于一定的载体。如果存储信息的载体遭到破坏，那么其承载的信息就会消失。历史上，好多珍贵文献没有流传下来，究其原因，是这些文献的载体遭到破坏。如秦始皇的“焚书令”使得当时大量书籍被烧毁，这些书中的信息自然就丢失了。其中比较知名的是“四书五经”中的《尚书》，由此产生了今文与古文两个版本，引发了后世的真伪《尚书》之争。

同一信息也可以依附于不同的载体，因此人们获取信息的途径与方法也可以不同。例如，某场球赛的最终结果，人们可以通过电视直播获悉，也可以通过网络查询，还可以在与朋友交谈时得知。信息依附于载体也体现了信息的可存储性与传递性。

（2）时效性

信息往往反映的是事物某一特定时间内的状态，它会随着时间的推移而变化。及时掌握最新信息，人们才能更好地利用它。例如，强台风来临之前，要时刻关注台风的走向，在台风袭击之前做好人员撤离与设施加固等工作，尽量减少台风带来的影响。但是在台风过后，有关这次台风的信息对于本次防御的重要性就降低了。

（3）共享性

信息是可以共享的，同一种信息可以同时被不同的接收者获取，人们也可以重复利用信息。与物质、能源不同的是，信息不会因为被别人获取而发生损耗。正如萧伯纳（George Bernard Shaw）所说：“你有一个苹果，我有一个苹果，彼此交换一下，我们仍然各有一个苹果；但你有一种思想，我有一种思想，彼此交换，我们就都有了两种思想，甚至更多。”

（4）可加工处理性、真伪性

信息是可以加工处理的。信息经过加工、处理、分析后，可以更好地被人们所使用。这一特征使信息具有真伪性，如两军交战，双方总是想尽办法迷惑对方，让对方做出错误的决策，从而取得胜利。第二次世界大战期间，盟军利用种种虚假情报，诱使德军做出错误判断，造成了诺曼底地区的兵力空虚，使得盟军在诺曼底成功登陆。

（5）价值性

信息具有价值性，信息的价值包括显性价值与隐性价值。显性价值指的是信息内容本身具有的价值，一般可被人们直接了解或体会。如根据紫外线指数的预报，人们可以做好外出前的个人防护。而隐性价值指的是除信息内容外的价值，包括与信息紧密相关的所有价值，如人们利用所学知识和技能，通过收集、整理和总结获得的其他价值。

信息的价值也是相对的，对于不同的人群、不同的时间，其价值可能有所不同。例

如，天气预报中有关海浪高度的信息，对于出海的渔民与居住在内陆的居民，其价值是不一样的。

1.2.3 知识

知识，这个词语是每个人都耳熟能详的，“知识就是力量”等名言传诵至今。人们从小开始学习各种知识，从课堂上、书本上获取知识，也从生活中、社会上获取知识。

知识是人类在社会实践中所获得的认识和经验的总和，也是人类在实践中认识客观世界（包括人类自身）的成果，它包括对事实、信息的描述以及在教育和实践中获得的技能。

知识是可以继承和传递的。牛顿的名言“站在巨人肩膀上”可以理解为“站在前人的肩膀上”，前人的研究成果已经成为现在研究的基础，人类的知识就这样一代一代地传承下去。

1.2.4 数据、信息与知识的关系

当人们孤立地看3.14时，它仅仅是一个数据；当人们在讨论圆的特性时，3.14是圆周率的近似值，这是信息；当人们用3.14乘以半径的平方来计算圆的面积时，这就是知识。

数据可以是数字、文字、图像、符号等，如上面的3.14，单独写在纸上，其本身没有明确的意义。将数据放在某个语境中，或在某个真实场景中使用，数据就有了意义，这就是信息。如3.14在上面的语境中是圆周率，如果在某个要填写日期的文本中写上3.14，那就不是圆周率，而是表示3月14日。

信息是数据经过储存、分析及解释后所产生的意义，通常是在某一特定情境脉络下的具体呈现。人们通过归纳、演绎、比较等手段对信息进行挖掘，将万千信息中有价值的部分沉淀下来，与已存在的人类知识体系相结合，形成知识。

与数据和信息相比，知识更接近行动，它与决策相关。当人们说某人掌握了某种知识时，不仅指的是他（她）“知道是什么”（Know-what），而且“知道为什么”（Know-why）以及“知道怎么做”（Know-how）。这就是为什么当人们在搜索引擎的帮助下可以获得海量信息，但这并不说明他们拥有海量的知识。而知识的丰富也不仅仅是依靠简单的检索、积累和存储。

知识的获得，是人利用自身已有的知识对信息进行加工，进而将新的信息纳入自己的知识结构的过程。这不仅仅是对信息的积累，还是对信息进行分析、判断、确认、归纳、演绎或比较等一系列的认知过程。这个过程与知识建构者个人的经验储备、所处情境和反思能力有关，因此，即使面对同样的信息，不同的人会因理解不同，形成不同的知识。所谓“一百个人心中有一百个哈姆雷特”，即当人们接收同样的信息后，所建构的知识也是有区别的。

智慧是一种更高层次的综合能力，主要表现为收集、加工、应用、传播知识的能力，以及对事物发展的前瞻性看法。它是在知识的基础之上，通过认识的累积，而形成的对事

物的认识、远见，体现为一种卓越的判断力。数据、信息、知识、智慧的关系如图1.2.1所示。

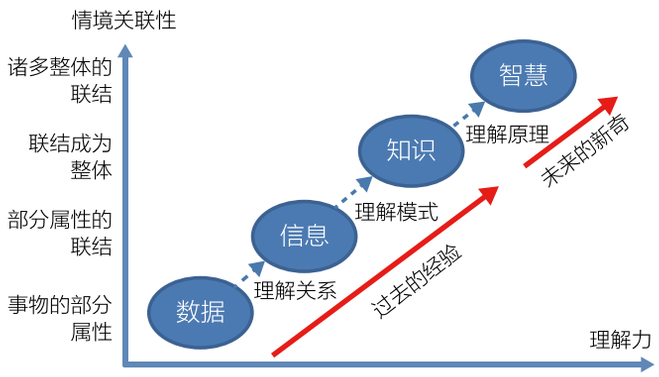


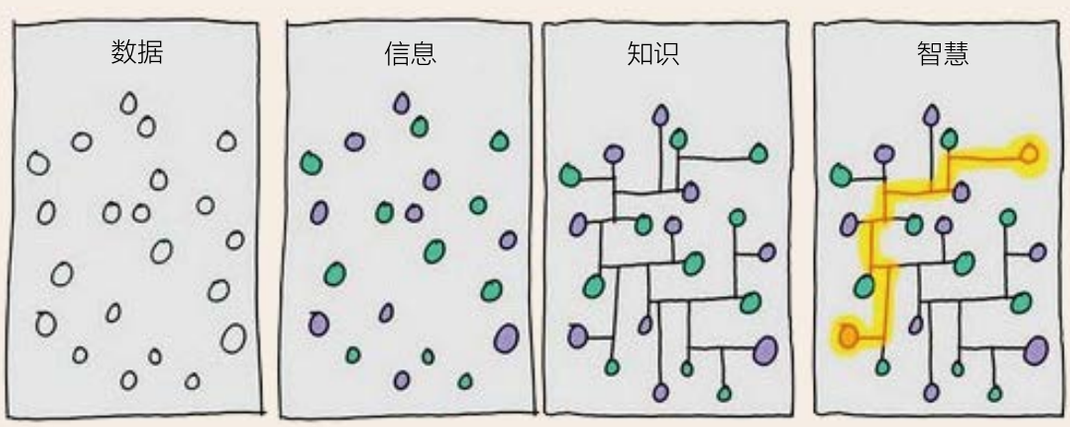
图1.2.1　数据、信息、知识、智慧的关系

问题与讨论

诗人艾略特（T. S. Eliot）的诗句：Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?（智慧迷失在知识中，知识迷失在信息中）你认为这两句诗表达了信息、知识和智慧之间怎样的关系？你能举出智慧迷失/没有迷失在知识中、知识迷失/没有迷失在信息中的四类例子吗？

思考与练习

1. 根据下图说明数据、信息、知识与智慧的关系。



2. 既然数据是对客观事物的符号表示，那么为什么我们会经常看到“虚假数据”这一词语？如何才能保证数据的真实性？

1.3 数据采集与编码

计算机技术的应用，使得数据的处理方式发生了巨大的转变，数据的处理效率也得到了极大的提高。不同的采集方式使得所获取的数据形式多种多样，要用计算机处理这些数据，需要对采集到的数据进行一定的转换。

1.3.1 数据采集

在开展研究时，研究工作者往往需要收集大量的数据。早期一般是通过观察、实验等人工方式得到数据，并将其记录下来。数据量较小时，可以用传统方式进行处理。比如统计某项活动的报名情况，可以先用纸笔方式记录，再用手工方式输入到计算机中进行处理。

现在，互联网、物联网的发展使得数据的获取方式变多、获取速度变快。传感器随时获取来自自然信源的数据，网络爬虫可在短时间内获取大量网络数据。由此，数据的获取方法已逐渐以机器获取为主。

传感器是一种能感受被测量并按照一定的规律转换成可用输出信号的器件或装置，通常由敏感元件和转换元件组成。在科研、生产和日常生活中，常需要利用传感器对环境中的物理量、化学量和生物量等进行感知与测量，并转换成电信号，进行适当处理后形成数据。

传感器可以持续不断地采集数据。如“环境空气颗粒取样器”可以实时监测着大气中PM2.5的浓度变化，地感线圈可以记录道路上车辆的通行数据（如图1.3.1），安装在野生动物身上的GPS追踪设备可以真实记录动物的活动轨迹。

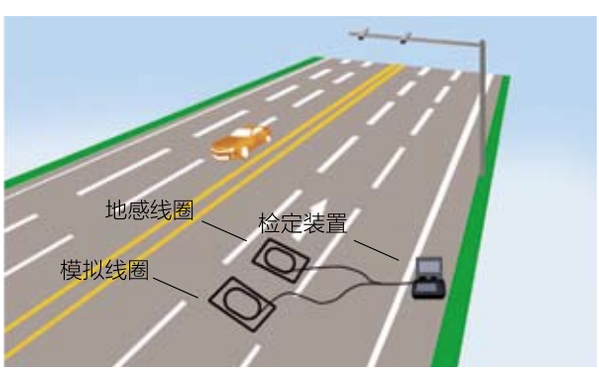


图1.3.1　地感线圈采集交通数据

传感器已广泛应用于各行各业中。例如，在现代农业生产中传感器得到了大量部署，用来记录农作物生长的环境温度、空气湿度、风力强度、土壤中的营养成分的动态变化、农机设备的使用状态等，由此获得丰富的数据。跟踪和分析这些数据，以便及时采取相应

的措施，提高农产品的产量和质量。

人们日常使用的设备中也包含传感器，如智能手机中用到了多个传感器，借助这些传感器，不仅可以获取声音、图像等数据，还可以获取地理位置、海拔高度、运动速度等数据，既丰富了手机功能，又改善了用户体验。又如各种可穿戴设备中通过传感器记录使用者的运动、心率、睡眠等数据，使人们的日常行为量化成一个个数据，来实现用户感知拓展、运动娱乐、健康监测等功能，为用户提供个性化的服务，并在一定程度上改善人们的行为方式。

除此之外，互联网也成了人们日常所需数据的主要来源。如互联网上有许多向公众开放的数据服务，用户可以通过应用程序接口（Application Programming Interface，简称API）采集这类数据，如国家气象局网站提供了气象数据API服务，用户根据API调用规则即可采集气象数据；也有一些专业的数据平台，将收集到的数据整理后出售，用户根据项目开发需要进行购买。充分利用这些数据，从中挖掘出信息，可以为下一步的决策提供依据，从而产生更大的经济价值和社会价值。

拓展链接

从互联网采集数据

经过多年的发展与应用，互联网上已经积聚了海量数据，这些数据覆盖了社会的各个领域。互联网上包含各类数据，每时每刻又有新的数据产生。人们可以随时从网上获取所需的各种数据用于日常生活，如在线预订火车票、机票，在电子地图上查找周边的学校、餐饮店、加油站等。专业人员则运用各种技术，从互联网上采集大量数据，用于研究、分析、决策等，如通过采集上网用户的相关数据，分析网民行为，以便推送精准的个性化服务。

从互联网采集数据有很多方法，如借助网络爬虫来获取数据。网络爬虫是一种按照一定的规则，自动地抓取网页上数据的程序或脚本。与人在浏览网页时的行为相似，网络爬虫也是通过网页中的超链接在网页间跳转，根据需求按特定的关键字获取某一方面的网页数据，然后对这些数据进行处理、存储等操作，并可用专门软件对这些数据进行分析。

1.3.2 数字化

信息可用模拟信号或数字信号表达。模拟信号以连续变化的物理量存在，如水银温度表呈现的温度值，电流表指针指向的电流值等。模拟信号经过采样量化后可以得到数字信号。数字信号在取值上是离散的、不连续的信号，在信息技术中，这种信号表示的数据是指可被计算机存储、处理的二进制数据。

模拟信号与数字信号可相互转换，如将语音通过计算机的麦克风、声卡等设备存储在计算机中，这一过程实现了模拟信号转换成数字信号，其中用到的主要设备是模数转换器（ADC）。模拟信号与数字信号可以相互转换，将模拟信号转换成数字信号的过程称为数字

化。自然界中的数字、文字、图像、声音等各种模拟信号，通过采样定理都可以用0和1来表示，即通过数字化工具将模拟信号转换成数字信号，这样才能用计算机来进行处理。从某种意义上说，数字化是信息社会的技术基础。

信息论的奠基者香农指出：在一定条件下，用离散的序列可以完全代表一个连续函数，这是采样定理的基本内容。在数字信号处理领域，采样定理是将连续信号（模拟信号）转换成离散信号（数字信号）的理论依据。它确定了信号带宽的上限，也确定了捕获连续信号时所允许的采样频率下限。

将模拟信号转换成数字信号一般需要经过采样、量化与编码。模拟信号先由采样器按照一定时间间隔采样获得时间上离散的信号，再经模数转换器（ADC）在数值上进行离散化（量化），经过编码转换成数字信号，如图1.3.2所示。

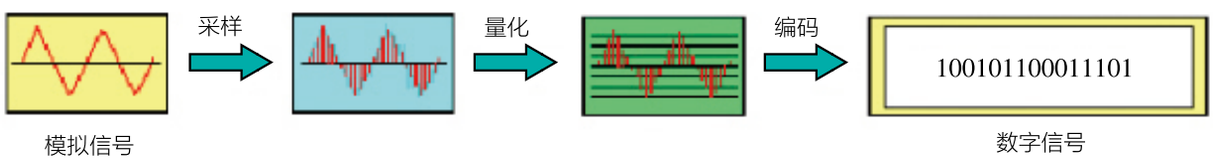


图1.3.2　模拟信号的数字化过程

1. 采样

在信号处理领域，采样是将信号从连续时间（空间）域上的模拟信号转换到离散时间（空间）域上的离散信号的过程，通过采样器实现。

对于同一模拟信号，采样的时间间隔越小，采集到的信号样本数量越多。每秒的采样样本数叫作采样频率，单位用赫兹（Hz）表示。在相同的时间内，采样频率越高，采集的样本数量越多。对于某一模拟信号，通过采样定理可以确定其采样频率下限。

将模拟信号转换成数字信号，会引起失真，影响信号保真度的一个因素是采样频率。一般而言，在对模拟信号采样时提高采样频率能提高保真度。

对于基于时间域的模拟信号，采样其实就是按一定的时间间隔取值。例如，图1.3.3甲所示为一段声音模拟信号，对其在时间轴上进行分隔，就能确定采集多少个样本或采样次数，如图1.3.3乙所示。

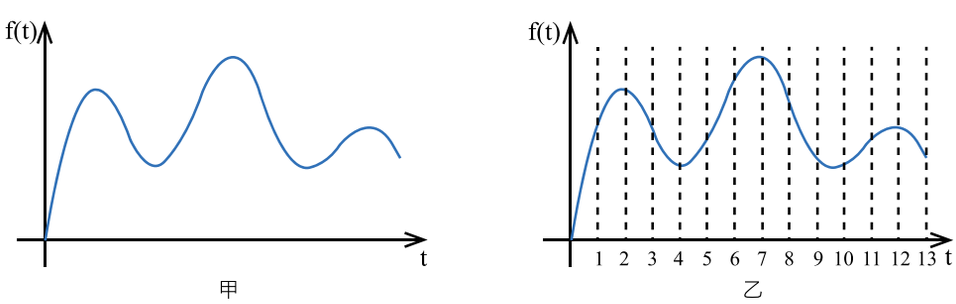


图1.3.3　声音模拟信号的采样

2. 量化

在数字信号处理领域，量化指将信号的连续取值近似为有限个离散值的过程。连续信号经过采样成为离散信号，离散信号经过量化后可用数值表示。

量化就是将采样到的信号用数字表示出来，即将模拟信号的波形转换为数字，量化的过程是先将整个幅度划分成有限个小幅度的集合，把落入某个范围内的样值归为一类，并赋予相同的量化值。

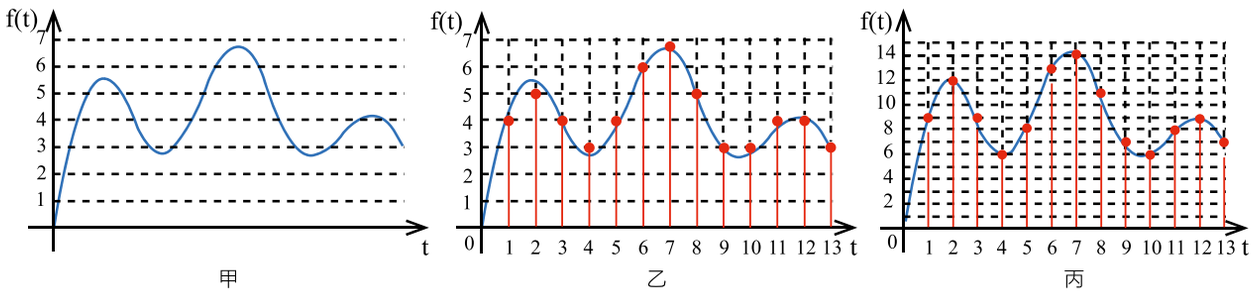


图1.3.4　离散信号的量化

对于图1.3.3 中的声音模拟信号而言，量化就是在纵坐标上进行划分（如图1.3.4甲），然后结合每个采样点将这些数据表示出来（如图1.3.4乙）。纵坐标划分得越细，量化就越精细，与实际数据也越接近，如图1.3.4丙所示。

问题与讨论

既然计算机只能处理数字信号，那么是不是意味着不再需要模拟信号了？请结合生活实例予以说明（如音频的录制与回放、答题纸的扫描与阅卷等）。

1.3.3　数制

数据在计算机内部是以二进制方式进行存储和处理的。与日常使用最多的十进制一样，二进制也是一种常见的数制。

1. 数的进制

进制是一种记数方式，亦称进位计数法或位值计数法。利用这种记数法，可以使用有限种数字符号来表示所有的数值。任何一种数制都包含两个基本要素：基和权。基又叫基数，是组成该数制的数码个数，一般来说，k进制的基数就是k，包含k个数字；权又叫权值，是指每一个数位上的1对应的数值，可以表示为基数的若干次幂。十进制数的基数为10，十进制数234中2的权值是10^2，3的权值是10^1，4的权值是10^0，所以十进制数234还可表示为：$2\times10^2+3\times10^1+4\times10^0$。

在信息技术中，人们通常采用二进制、八进制、十进制、十六进制来表示信息。为了区别各种进位制的数码，通常用一个下标来表示该数的进制（十进制数可以省略），也可以在该数的最后以字母来表示，见表1.3.1。

表1.3.1　进制的标识

进位制	二进制	八进制	十进制	十六进制
标识	B	O	D	H

2. 二进制

世界上有很多事物只存在两种状态，如逻辑上的“真”与“假”，黑白图像中的“黑”与“白”等，这些事物状态可以用“1”和“0”两个数来表示，如电路中可用1来表示开关的合，用0表示开关的断。这种只有两个数码的数制称作二进制，由17世纪德国数学家莱布尼茨（Leibniz）发明，到20世纪40年代应用于电子计算机中。

二进制数的特点是：

①有两个基本数码：0，1。

②采用逢二进一的进位规则。

例如，$1101.01B = 1\times2^3 + 1\times2^2 + 0\times2^1 + 1\times2^0 + 0\times2^{-1} + 1\times2^{-2}$。其中B表示二进制，$2^3$，$2^2$，$2^1$，$2^0$，$2^{-1}$，$2^{-2}$是不同位置上的权值。

3. 十六进制

二进制数在实际使用中，由于位数太长，不便于书写和记忆，所以人们常采用十六进制数来表示。

十六进制数的特点是：

①由十六个基本数码组成，即0，1，2，…，9，A，B，C，D，E，F。

②采用逢十六进一的进位规则。

例如，$B574H = 11\times16^3+5\times16^2+7\times16^1+4\times16^0$。与二进制相类似，H表示十六进制，$16^3$，$16^2$，$16^1$，$16^0$是不同位置上的权值。

其实十六进制不是现在才有的，在中国古代已经采用了十六进制，当时的1斤为16两，成语“半斤八两”就源于此。十进制、十六进制和二进制之间的对应关系如表1.3.2所示。

表1.3.2　十进制、十六进制、二进制之间的关系

十进制	十六进制	二进制	十进制	十六进制	二进制
0	0	0	3	3	11
1	1	1	4	4	100
2	2	10	5	5	101

续表

十进制	十六进制	二进制	十进制	十六进制	二进制
6	6	110	11	B	1011
7	7	111	12	C	1100
8	8	1000	13	D	1101
9	9	1001	14	E	1110
10	A	1010	15	F	1111

1.3.4 编码

编码（Encoding）是信息按照某种规则或格式，从一种形式转换为另一种形式的过程。解码是编码的逆过程。

计算机对信息进行存储、加工、传递等处理，实际上是对信息的载体——数据进行处理。数据的表现形式可以是文本、图形、图像、声音、视频等，但不管是哪种形式的数据，最终存储在计算机中的都是经过一定规则编码后的二进制数字。

1. 字符编码

常见的字符编码有ASCII、Unicode及各种汉字编码。

（1）ASCII码

ASCII（American Standard Code for Information Interchange，美国信息交换标准代码）是一套基于拉丁字母的计算机编码系统，主要用于显示现代英语和其他西欧语言。它由电报码发展而来，是现今最通用的单字节编码系统。基本的ASCII码共有128个，用1个字节中的低7位编码。二进制范围为00000000～01111111，即十六进制的00～7F。基本的ASCII码由33个控制字符、10个阿拉伯数字、26个英文大写字母、26个英文小写字母与一些标点符号、运算符号组成。

ASCII码值及对应的字符见附表。

拓展链接

数据的存储容量单位

计算机中存储容量最小的单位是比特（bit），1位二进制数码表示1个bit，但由于1bit所能表示的值太小，实际上计算机中以8bit为一个基本单位，称为字节（Byte）。常见的单位还有KB、MB等，它们之间的换算关系是：

1KB=1024B　　1MB=1024KB　　1GB=1024MB　　1TB=1024GB

1PB=1024TB　　1EB=1024PB　　1ZB=1024EB

（2）汉字编码

计算机中的汉字也是采用二进制进行编码的。汉字编码分为外码、交换码、机内码和字形码。其中，外码也叫输入码，是用来将汉字输入到计算机中的一组键盘符号。常用的输入码有拼音码、五笔字形码等。

根据国标码的规定，每一个汉字都有确定的二进制代码，在计算机内部汉字代码都用机内码，在磁盘上记录汉字代码也使用机内码，在早期的GB2312字符集中，1个汉字在计算机中用2个字节表示。如图1.3.5中的“中国China”这几个字符，其中“中国”两个汉字的内码为D6 D0 B9 FA，用二进制表示就是11010110 11010000 10111001 11111010；而英文字符“China”是ASCII字符，其中每个字母都用1个字节编码表示。

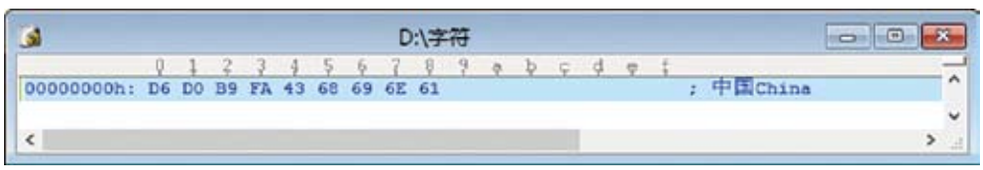


图1.3.5　字符内码

2. 条形码

条形码（barcode）是将宽度不等的多个黑条和白条，按照一定的编码规则排列，用以表达一组信息的图形标识符。条形码技术最早产生于20世纪20年代，常见的条形码是由反射率相差很大的黑条（简称条）和白条（简称空）排成的平行线图案。条形码可以标出物品的生产国、制造厂商、商品名称、生产日期、图书分类号、邮件起止地点、类别、日期等信息，因而在商品流通、图书管理、邮政管理、银行系统等领域广泛应用。

目前国际上有多种条形码编码方式，图1.3.6为我国普遍采用的EAN13条形码。

图1.3.6　商品条形码

这种条形码由13位数字组成，前3位数字表示国家代码，图中的“690”表示中国大陆地区。最后一位叫校验码，用来检查扫描到的数字是不是有错误，这个数字由前12位数字按一定规律计算得到。

3. 二维码

二维条码 / 二维码（2-dimensional bar code）是用某种特定的几何图形按一定规律在平面上（二维方向）分布的黑白相间的图形记录数据符号信息，如图1.3.7所示。相对于一维的条形码，二维码的信息存储量更大，功能也更加强大。随着智能手机的普遍使用，手机已成为个人用户扫描二维码读取信息的常用工具。现在的生活环境中随处可见各种各样的

二维码，人们可以通过扫码即时完成支付和信息阅读等操作，也可以自己制作二维码分享给他人。

二维码在为人们的生活提供便利的同时，也带来了一定的安全隐患。不随意扫描非官方的二维码或安装未经验证的应用，是信息社会的基本常识。

图1.3.7　二维码

4. 声音编码

声音是振动产生的声波，通过介质（空气、固体或液体）传播并能被人或动物的听觉器官所感知的波动现象。

声音的频率一般以赫兹表示，记为Hz，指每秒周期性振动的次数。人耳可以感知到的声音，其频率范围在20～20000Hz。分贝是用来表示声音强度的单位，记为dB。

将模拟声音数字化需要经过采样、量化、编码三个过程。比如对语音进行数字化时，声音信号通过麦克风转换成电流信号，通过声卡上的模数转换器（ADC）将电流信号转换成一串数字信号，采样、量化后对其进行编码，存储到计算机中。重放时，这些数字信号送到声卡上的数模转换器（DAC）还原为模拟信号，放大后送到扬声器发声。

在音频信号数字化过程中，声音的保真度不仅受到采样频率的影响，也依赖于量化值。量化值一般用二进制数表示，其二进制位数决定了量化的精度，也称作量化位数。量化位数越大，量化精度也越高。在图1.3.4乙中，量化值取值范围是0～7 这8个数，需要用3位二进制表示，即量化位数为3位（2^3=8）；当量化位数提高到4 位时（2^4=16，量化值取值范围为0～15 这16 个数），如图1.3.4丙所示，图中第2个采样点的量化值从101B变成1100B，采样点与实际数据的差异明显减少。

模拟音频信号经过采样和量化以后，形成一系列的数字信号。将这些数字信号按一定的方式进行编码，以文件的形式存储在计算机的外部设备中。根据声音的存储格式可分为不同的文件类型，常见的声音文件类型有Wave、MP3、WMA等。

Wave格式音频文件的存储容量可以通过下面的公式进行计算：

存储容量=采样频率（Hz）× 量化位数（bit）× 声道数 × 时长（s）（单位：位）

5. 图像编码

图像是人对视觉感知的物质再现。图像可以由光学设备获取，如照相机、镜子、望远镜及显微镜等；也可以人为创作，如手工绘画。图像可以记录、保存在纸质媒介、胶片等对光信号敏感的介质上。随着数字采集技术和信号处理理论的发展，越来越多的图像以数字形式存储。

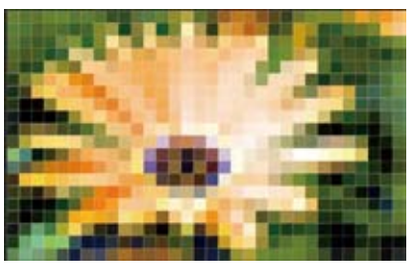
数字图像包括矢量图形与位图图像。

在计算机图形学中，矢量图形是指用点、直线或者多边形等基于数学方程的几何图元表示的图像。矢量图形保存的文件大小一般比位图要小，并且文件大小与图形的大小无关，在图像处理软件中任意放大矢量图形，不会丢失细节或影响清晰度，因为矢量图形与

分辨率是无关的。

位图图像又称栅格图或点阵图，将图像数字化也需要经过采样、量化、编码等环节。图像的采样就是把一张图像分解成一个一个大小相同的点，这些点称作像素，是组成位图图像的基本单位。图1.3.8甲为512×320像素的图像，也就是水平方向有512个像素，垂直方向有320个像素，而图1.3.8乙由32×20像素组成。可以直观地看出，点越多，图像越真实，越能体现细节，同时也需要更多的存储空间。

甲　512×320像素



乙　32×20像素

图1.3.8　不同像素的图像

图像的量化是指要使用多大范围的数值来表示图像采样之后每个像素的颜色信息。一般用二进制数来表示，其长度也称为颜色的位深度。如256种颜色的图像，它的位深度为8位。

图像存储容量可以根据像素与颜色位深度进行计算：

存储容量＝总像素 × 颜色位深度（单位：位）

对量化后的数据按一定规则进行编码后，数字图像将以文件方式存储于计算机中。根据不同的编码方式，可分为多种图像文件格式，如BMP、JPEG、GIF、PNG等。

6. 视频编码

静态的图像连续播放就形成视频，如早期的模拟电视中，PAL制式的视频每秒播放25帧图像，而NTSC制式的视频每秒播放30帧图像。目前，我国已经完成了由模拟电视向数字电视的转换。与传统的模拟电视相比，数字电视采用了数字传输和存储技术，具有高清晰度、双向交互、多功能多业务等优势。

视频数据由于数据量大，不便于存储与传输，往往需要对其进行压缩。视频的编码一般是指通过特定的压缩技术对视频进行压缩。常见的视频编码方式有MPEG-1、MPEG-2、MPEG-4等。

随着信息技术的发展，更多专业服务系统产生，如可用于卫星定位的GPS系统、可用于位置服务的POI数据库等，这些专业服务所需的编码需要各自特定的编码方式。

实践与体验

字符乱码之谜

在计算机上查询资料时，我们有时会看到某篇文章或某段文字内容出现乱码；用计算机中的“记事本”软件保存的文档，偶尔也会出现乱码现象（如图1.3.9）。在学了本节内容后，让我们来探究一下，字符出现乱码与字符的编码有怎样的关系。

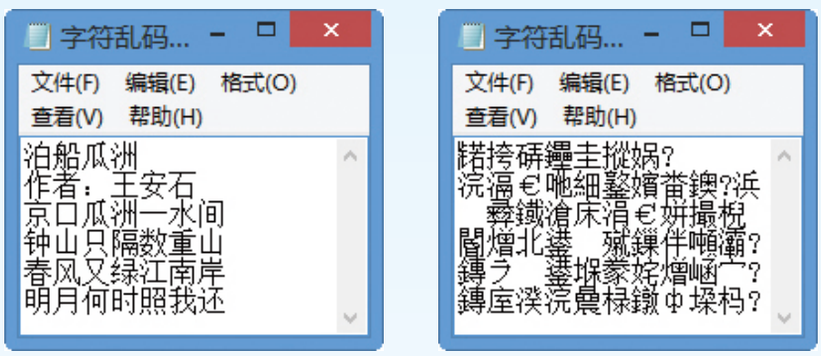


图1.3.9　乱码现象

实践内容：

1. 观察字符在不同字符集中的内码值，分别找出ASCII码字符与汉字在这些字符集中的编码规律。

2. 总结由编码因素引起乱码的原因。

实践步骤：

1. 准备软件。

Windows附件中“记事本”软件；UltraEdit软件。

2. 查阅资料。

通过互联网或相关图书资料，查找Unicode字符集的有关知识。根据资料，分析产生乱码的可能原因。

3. 上机实践。

在“记事本”软件中分别以不同的编码保存某个文本文件，然后用UltraEdit软件打开，观察字符内码，寻找规律并填写如下表格。

字符	ANSI		Unicode		UTF-8		说明
	内码	字节	内码	字节	内码	字节	
							ASCII字符
							汉字

结果呈现：

1. 将发现的规律与得出的结论形成文档，通过多媒体广播或投影形式向同学展示研究成果。

2. 思考：导致字符乱码的原因除了实践中发现的以外，还有哪些？

思考与练习

1. 在生活中，除了二进制、十进制与十六进制，我们常用的还有哪些进制？

2. 将1000个苹果放入10个箱子。要取走1～1000中任意个数的苹果，要求不拆开箱子。应如何装箱？结合二进制思想，说明其原理。

3. 为了提高声音的保真度，是否可以无限制地提高采样频率与量化位数？

1.4 数据管理与安全

随着技术的发展，数据量的增长速度越来越快，如何有效管理数据和保证数据安全成为各行各业都面临的问题。

1.4.1 数据管理

数据管理是利用计算机硬件和软件技术对数据进行有效收集、存储、处理和应用的过程，其目的在于充分、有效地发挥数据的作用。计算机数据的管理已经经历了人工管理、文件管理和数据库管理三个阶段。

在人们日常使用的计算机中，数据一般以文件的形式存储。根据编码规则的不同，文件的格式也不相同，用以区分不同类型的存储数据，如文本、图像、音频等。在目前使用较广的Windows操作系统中，可以用文件扩展名来表示某些特定的文件类型，如网页文件的文件扩展名为.htm或.html，而JPEG图像的文件扩展名为.jpg等。

计算机一般采用树形目录结构来管理文件，如图1.4.1所示。在Windows系统中，则采用了更为形象的文件夹来管理文件，如图1.4.2所示。

图1.4.1 树形目录结构

图1.4.2 文件夹

问题与讨论

“记事本”软件默认保存的是.txt文件，Word软件默认保存的是.docx文件，这两种文件有哪些区别？

相对于普通用户对数据的日常管理，软件开发人员在编写具体应用软件时，需要将与应用有关的所有数据收集在一个文件中，然后对该文件进行管理。每个应用程序都有自己的数据文件，数据可能在多个文件中重复出现，造成数据冗余。同时数据文件之间的相互关联，需要大量的人工干预，给数据的维护与更新造成不便。这些问题直到数据库系统的出现才得以解决。

生活中的数据库应用随处可见，不论是超市购物，还是银行存取款，都是在数据库系统支撑下进行的。传统数据库技术基于结构化数据开发，凭借其数据独立性、数据可共享等特点，已经成为现代社会数据管理的主要方式，成功应用在政府、军工、教育、电力、金融、农业、卫生、交通、科技等诸多行业。

随着数据采集技术的提高，数据量急剧增长，大量半结构化、非结构化数据被源源不断地采集起来。对于这些数据，已经很难用传统的数据库技术进行管理。现在，借助云计算、大数据等技术，数据管理水平正不断提高。

拓展链接

结构化、半结构化和非结构化数据

结构化数据，也称作行数据，是由二维表结构来进行逻辑表达和实现的数据，严格地遵循数据格式与长度规范，主要通过关系型数据库进行存储和管理。

非结构化数据是数据结构不规则或不完整，没有预定义的数据模型，是不方便用数据库二维逻辑表来表现的数据。包括各类格式的办公文档、文本、图片、HTML、各类报表、图像、音频、视频等。

半结构化数据，就是介于结构化数据和非结构化数据之间的数据，具有一定的结构性。

1.4.2 数据安全

信息化正在快速地改变着这个世界，数字化、网络化、智能化已经上升至国家战略。大到全球经济发展格局，小到每个人的日常工作生活，都与数据息息相关，数据的安全问题也变得越来越重要。

威胁数据安全的因素有很多，如硬盘驱动器损坏、操作失误、黑客入侵、感染计算机病毒、遭受自然灾害等，都有可能造成计算机中数据的损坏。

数据存储在特定的介质上，保护数据的安全也需要保护存储数据的介质。对于政府

部门或企业的数据，目前主要采用主动防护的手段，如通过磁盘阵列、数据备份、异地容灾等手段，保证数据的安全。

对于数据安全，不仅要做好防护上的安全，还应提高数据本身的安全，如通过数据加密、数据校验等方法来提高数据的保密性和完整性。数据加密，是指通过加密算法和加密密钥将明文转变为密文，而解密则是通过解密算法和解密密钥将密文恢复为明文。数据校验，是为保证数据的完整性进行的一种验证操作，通常用一种指定的算法对原始数据计算出一个校验值，接收方按同样的算法计算出一个校验值，如果两次计算得到的校验值相同，则说明数据是完整的。常见的数据校验方法有MD5、CRC、SHA-1等。

在信息社会中，数据安全关乎国计民生，每个人都应提高数据安全意识，增强法律意识，采取必要的安全防范措施，及时备份重要数据，这样才能保障个人数据和财产安全。

拓展链接

容灾系统

容灾系统是指在相隔较远的异地，建立两套或多套功能相同的IT系统，互相之间可以进行健康状态监视和功能切换，当一处系统因意外（如火灾、地震等）停止工作时，整个应用系统可以切换到另一处，使得该系统可以继续正常工作。

拓展链接

文件的MD5校验

文件的MD5校验是将整个文件当作一个大文本信息，通过其不可逆的字符串变换算法，产生唯一的MD5信息摘要并提供给用户。用户下载完文件以后，通过专用程序计算下载文件的MD5校验码，比对前后的校验数据，判断下载文件是否完整。

问题与讨论

现在，人们在生活中越来越依赖于智能手机，手机中不仅存储了联系人、照片、视频等个人数据，还有电子银行、支付宝以及微信钱包等账户信息，如果存有这些信息的手机丢失，将会造成很大的损失。如何才能将因手机丢失而造成的损失降到最小？

思考与练习

1. 如何管理生活中的各类数据？

2. 密码根据其组成字符的复杂度可以分为强密码与弱密码。需要多种字符组合且符合一定长度的密码称为强密码。强密码尽管很难被破解，但也带来了记忆上的困难。应该如何合理设置各类密码？

1.5 数据与大数据

互联网、移动网络、物联网等每天都产生着大量数据，这些数据规模巨大、格式多样，已经很难用传统的方式进行处理。于是，大数据技术应运而生，通过分析、挖掘这些数据，发现其蕴藏的价值。

1.5.1 大数据的概念

继移动互联网、云计算之后，大数据正在引发信息科技产业新的变革，并已经开始对社会的组织结构、国家治理模式、企业决策架构以及个人生活方式产生深刻影响。

早在几十年前，在“信息爆炸”“知识爆炸”的表述中，就已经对数据的大规模增长有所认识。20世纪90年代末，随着数据处理技术的发展，“大数据”的概念首次由美国硅图公司（SGI）的一位科学家正式提出。2016年，数据科学家将大数据正式定义为：大数据代表着信息量大、速度快、种类繁多的信息资产，需要特定的技术和分析方法将其转化为价值。也就是说，大数据之“大”，不仅指规模、速度和种类的特征，还意味着它超出以往常用的数据采集、组织、管理和加工等软件的处理能力，要求新型集成技术从多元、复杂和巨量规模的数据集里洞察规律。

1.5.2 大数据的特征

数据量大并不一定就是大数据，用传统算法和数据库系统可以处理的海量数据不能算“大数据”。符合大数据概念的数据一般具有数据规模大、处理速度快、数据类型多、价值密度低四个特征，可以用4个V来概括，即数量（Volume）、速度（Velocity）、多样（Variety）和价值（Value）。

第一，数据体量巨大。大数据收集和分析的数据量非常大。现在，传感器、互联网、智能终端等每天都在源源不断地产生海量数据，人类社会的数据量在不断刷新一个个新的量级单位，已经从TB、PB级别跃升至EB、ZB级别。可以通过下面这个例子简单感受1EB（$1EB=2^{60}B$）的数据量：一本《红楼梦》约有87万个字（含标点），每个汉字占两个字节，即1个汉字=2B，由此得出1EB约等于6626亿部《红楼梦》。这个数据量必将随着大数据处理能力的发展而不断扩大。

第二，速度快。速度快有两种含义，一是数据产生的速度快。有的数据是爆发式产生的，例如欧洲核子研究组织的大型强子对撞机在工作状态下每秒产生PB级的数据；有的数据是累积产生的，比如微博、微信中的数据，每个用户产生的数据量可能不大，但是由于用户众多，短时间内产生的数据量依然非常庞大。二是数据处理的速度快。在信

息社会中，数据往往实时变化，数据的价值也会随着时间的推移而变化，只有高效率的数据处理技术才能充分发挥数据的价值，例如，通过气象卫星等设备采集到的数据，只有及时处理才能满足天气预报的需求。

第三，数据类型多。大数据的数据来源多，既有人工产生的，如人们日常使用智能手机，短信、微信、视频、语音、电子邮件等会产生各种数据；也有机器自动产生的，如各种传感器在生产监测、环境监测、交通监测、安防监测等过程中也会产生大量数据。正因为大数据来自多种数据源，其数据种类和格式不可能保持一致，各种结构化、半结构化和非结构化数据共存是大数据的普遍现象。

第四，价值密度低。大数据蕴含着巨大的价值，但因其数据量庞大，可能发挥价值的仅是其中非常小的部分，价值密度相对较低。以当前广泛应用的监控视频为例，在连续不间断的监控过程中，大量的视频数据被存储下来，其中有许多冗余数据。比如某起交通事故的视频画面，有效的部分可能仅仅只需要几秒钟，大量不相关的视频信息会增加获取有效数据的难度。价值密度的高低与数据总量的大小成反比，“提纯”大数据，让其发挥更大的价值，是人们一直在努力的目标。

1.5.3 大数据思维

大数据是一场变革，改变的不仅是数据，还有人们的思维。

首先，大数据要分析的是全体数据，而不是抽样数据。以往对于某项研究中的数据，限于技术等因素，人们无法进行全样本分析，往往会随机抽取部分样本进行研究，以此推论全体情况。抽样数据分析的方式效率较高，经常被人们采用，但这种方式取决于抽取样本的随机性，在某些情况下，不同的样本可能会得出截然不同的结论。在大数据时代，人们不仅可以获得研究所需的直接数据，而且还能对与之有关联的所有数据进行分析。分析数据已经不再依赖于采样，从而带来更全面的认识，也能更清楚地发现抽样数据无法揭示的详尽信息。

其次，对于数据不再追求精确性，而是能够接受数据的混杂性。对于传统的数据库，数据有严谨的结构，人们追求数据的准确性，通过各种技术或人工手段，来保证每个数据准确无误。而在大数据处理过程中，数据的来源多种多样，这些数据可以是结构化的、半结构化的，也可以是非结构化的。当数据量大到一定程度时，个别数据的不准确就显得不那么重要。

再次，不一定强调对事物因果关系的探求，而是更加注重它们的相关性。在传统的思维方式中，人们往往执着于现象背后的因果关系，试图通过有限样本数据来剖析其中的内在机理。这种思维方式有一定的局限性，此外，有限的样本数据也无法反映出事物之间的相关关系。在大数据时代，比如电商的个性化推荐，不必知道人们购买某些商品的原因，只要找到商品之间的关联性，就能为客户提供精确的推荐。

1.5.4 大数据对社会的影响

大数据已渗透到各行各业，成为重要的生产因素。作为全球网民数量最多的国家、重要的电子信息产品生产基地和最具成长性的信息消费市场，中国已经成为重要的大数据资源集聚地和大数据应用市场，大数据产业快速发展，产业链加速形成，大数据正在对经济社会发展发挥着越来越重要的作用。

大数据让生活更便利。例如，人们可以通过城市热力图了解一个区域的人流量及拥挤情况，如图1.5.1所示，绿色部分显示的是人流量小或稀疏的地理区域。城市热力图通过手机基站来定位区域中的手机用户，根据用户数量渲染地图的颜色，来展示该区域的人流密度，为人们的出行提供参考。又如，在电子商务领域，每天有数以万计的交易在淘宝上进行，与此同时，相应的交易时间、商品价格、购买数量会被记录，更重要的是，这些信息可以与买方和卖方的年龄、性别、地址甚至兴趣爱好等个人特征信息相匹配。通过大数据服务，商家可以了解商务平台上的行业宏观情况、自己品牌的市场状况、消费者行为情况等，并据此进行生产、库存决策，而与此同时，更多的消费者也能以更优惠的价格买到心仪的物品。

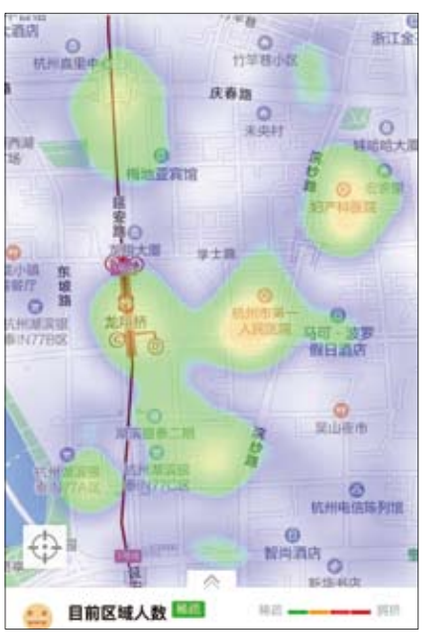


图1.5.1 城市热力图

大数据让决策更精准。大数据支持动态跟踪与全样本采集，为各种决策提供了第一手的材料，再加上可视化技术的应用，提高了数据分析的即时性，可以帮助管理者及时发现问题，进行即时干预。例如，江西省上饶市教育局利用大数据动态收集农村孩子入学、辍学、父母陪伴等信息，及时发现留守儿童的学习问题，开展精准助学与帮扶；又如，北京在共享单车运行一年后，重新调整了公交路线，正是共享单车所产生的大数据让交通部门发现了部分线路的公交站点缺失，从而精准地确定了更为利民的交通路线图。

大数据带来新的就业需求。随着大数据的发展，与之相关的职业需求也急剧增长。例

如，系统研发工程师、应用开发工程师、数据可视化工程师和数据分析师等职业，带来了成千上万的岗位。全国各大高校也纷纷开设与之相关的专业，培养大数据行业的专业人才，以满足社会发展的需要。

大数据带来新的社会问题。大数据给生活带来便利的同时，也带来如信息泄露、数据安全、个人隐私甚至伦理道德等方面的社会问题。用户在网上注册、网上购物等过程中，会留下个人信息。大数据的汇集不可避免地加大了用户隐私数据信息泄露的风险，如何保护个人的隐私成为亟待解决的问题。由于数据中包含大量的用户信息，在对大数据开发利用的过程中很容易涉及公民隐私，使得恶意利用公民隐私的技术门槛大大降低。在大数据应用环境中，数据呈现动态特征，面对数据库中属性和表现形式不断随机变化，基于静态数据集的传统数据隐私保护技术面临挑战。各领域对于用户隐私保护有多方面的要求和特点，数据之间存在复杂的关联性和敏感性，针对传统关系型数据的隐私保护模型和算法，大部分都不能直接将其移植到大数据应用中。

问题与讨论

大数据为生活带来便利的同时也带来了安全隐患，各种信息泄露事件时有发生。请结合实例，探讨可能引发信息泄露的原因以及由此产生的危害。我们应该如何预防？

思考与练习

1. 学生学籍系统中存放着大量的学生数据，这些数据是否属于大数据？为什么？

2. 在处理数据时，往往会选择“抽样数据”或“全样本数据”进行分析，请比较这两种分析方式的特点，并举例说明。

巩固与提高

1. 物质、能源、信息是人类赖以生存的基础。控制论的创始人诺伯特·维纳（Norbert Wiener）认为：信息就是信息，不是物质，也不是能量。结合实例说明物质、能源、信息三者之间的关系。

2. 十进制数10的二进制值为1010B，十进制数20、40、80的二进制值分别是多少？

3. 二维码因其使用方便，在社会多个领域中广泛应用，但与此同时也存在着各种安全问题。请查阅相关资料，分析、总结如何安全使用二维码技术。

4. 大数据已经进入了人们的生活，并改变着人们的思维方式。请以具体事例说明大数据在生活中的应用。

5. 录制一段时长为4分零5秒的双声道音频，采样频率为44.1kHz，量化位数为2个字节。若不进行压缩，则存储容量约为多少MB？

6. 将一张分辨率为1024×768、256色位图的JPEG图像文件通过“画图”程序另存为24位位图的BMP文件。文件的大小为多少？此操作能否提升图像质量？

项目挑战

哪些技术影响了你——年度最具影响力的技术TOP5

当今社会，信息技术的广泛应用，加快了社会生产力的发展，引发了社会各个领域、各类行业的深刻变革。信息技术与社会的深度融合，也改变着人们的学习、工作、生活方式。信息技术促进了新技术的研发，每年都会涌现出大量新的设备、新的应用。在这些技术中，你认为哪些技术正在影响着你，甚至影响着世界？

从当今的新技术中找出你认为最具影响力的前五项。

项目任务

以小组合作的形式，借助数字化学习平台，上网搜索相关信息，了解信息技术的最新发展动态，结合幻灯片或动画，讲述你们组评出的新技术TOP5。具体要求如下：

1. 通过网络搜索，了解信息技术的发展与应用情况。
2. 根据技术的先进性与应用情况，分析其对社会的影响力。
3. 分析技术的潜力与发展趋势，预测其在未来的成长程度。
4. 结合技术的影响力、发展速度、市场前景、行业口碑、未来潜力等方面，形成分析报告，展示年度最具影响力的技术。

注：所找出的新技术未必是本年度发明或推出的，主要考查它们在本年度所表现出的影响力。

过程与建议

小组成员分工明确，任务落实到每个成员；理解和整理网上查找到的相关资料，小组内达成共识，制作成精美的演示文稿，边展示边讲解。具体步骤如下：

1. 落实任务

每个小组由组长负责，统一将具体任务落实到每个成员，成员之间有明确的分工与合作。小组成员的分工可根据各成员的特长，也可根据任务模块。小组成员分工表要体现在展示成果的作品中。

资料搜索可以先搜索所有类别的新技术，也可以根据各个领域中的年度最佳进行搜索，然后从中筛选出与信息技术密切相关的部分作为候选。对于往年的技术，根据其在本年度的影响力，也应作为候选。

2. 搜索并整理资料

通过互联网搜索相关资料，按技术的应用领域或研究方向进行分类。从收集的资料中选择有代表性的图片、动画、视频等，选择合适的软件处理资料。将处理好的资料按报告内容或文件类别进行存储。

3. 制作演示文稿

将处理好的资料按设计的思路制作成演示文稿，形成汇报材料。其中应包括如下内容：

（1）推出的年度TOP5技术。

（2）选择这些技术的依据（理由）。

（3）对每种技术的简介：概念、应用、所产生的影响力、未来展望等。

（4）团队的分工情况。

4. 讲述与评价

每个小组派代表展示演示文稿，并结合文稿内容进行讲述。

展示完毕后，汇总所推出的技术，全班票选，选出班级认同的“年度最具影响力的技术TOP5”。

评价标准

根据项目所涉及内容的严谨性及实际展示效果，对自己完成项目的情况进行客观的评价，并思考后续完善的方向。将评价结果和完善方案填写在下面的表格中。

评价条目	说明	评分（1～10分）	评分主要依据阐述	后续完善方向
选择技术	所选择的技术具有先进性，对社会和技术的发展有重要影响			
选择依据	小组共同形成的技术选择，依据清晰且具有说服力			
小组合作	小组分工合理、协作密切、合作有成效			
演示文稿	演示文稿制作精美、内容清晰、逻辑性强，包含所有要求的内容			
展示效果	在规定时间内有条理地、清晰地介绍研究成果			

拓展项目

Base64编码是计算机中常见的一种编码方式，规则是把3个字节（24位）的数据按6位1组分成4组（24 ÷ 6=4），然后将每组数据分别转换为十进制，根据表1.5.1将这些十进制数所对应的字符连接，即为Base64编码。

表1.5.1　Base64编码表

索引	字符	索引	字符	索引	字符	索引	字符	索引	字符	索引	字符	索引	字符
0	A	10	K	20	U	30	e	40	o	50	y	60	8
1	B	11	L	21	V	31	f	41	p	51	z	61	9
2	C	12	M	22	W	32	g	42	q	52	0	62	+
3	D	13	N	23	X	33	h	43	r	53	1	63	/
4	E	14	O	24	Y	34	i	44	s	54	2		
5	F	15	P	25	Z	35	j	45	t	55	3		
6	G	16	Q	26	a	36	k	46	u	56	4		
7	H	17	R	27	b	37	l	47	v	57	5		
8	I	18	S	28	c	38	m	48	w	58	6		
9	J	19	T	29	d	39	n	49	x	59	7		

以编码字符“Web”为例，如表1.5.2所示，字符“Web”对应的ASCII编码分别是87，101，98，分别转换为8位二进制数，按6位二进制数分组后再转换成十进制，查找它们的对应字符，得到“Web”的Base64编码为“V2Vi”。

表1.5.2　Base64编码方法

文本	W								e								b							
ASCII编码	87								101								98							
二进制位	0	1	0	1	0	1	1	1	0	1	1	0	0	1	0	1	0	1	1	0	0	0	1	0
索引	21				54				21				34											
Base64编码	V				2				V				i											

若某字符的Base64编码为“QW55”，则其原文是什么？查找相关资料，说明当字符数不是3的倍数时，剩下的字符应如何编码？字符“award”的Base64编码是什么？

第二章 2

算法与问题解决

“算法”指的是解决某个问题的一组步骤。人们在解决问题时都会经历一个“怎么做”的阶段，而思考“怎么做”的过程，就是“算法设计”的过程。设计算法并用一定的方式准确地描述算法后，算法执行者（人或者机器）就能按照描述的算法分步处理并最终解决问题。

用计算机解决问题时，通常先设计算法，然后将算法用合适的计算机程序设计语言表示，计算机就能按照人们设计的计算机程序进行高速、准确的自动化处理，从而帮助人们解决问题。

问题与挑战

- 越来越多的人通过走路来锻炼身体，为了激发人们的锻炼热情，可以开发一个名为“动动有奖”的手机APP，并通过阶梯“奖金”的形式给步行者不同的“奖金”。这样步行者就可以每天通过手机来衡量自己的锻炼状况并获得相应的“奖金”。如何为这个手机APP设计一个根据走路步数统计“奖金”的算法？

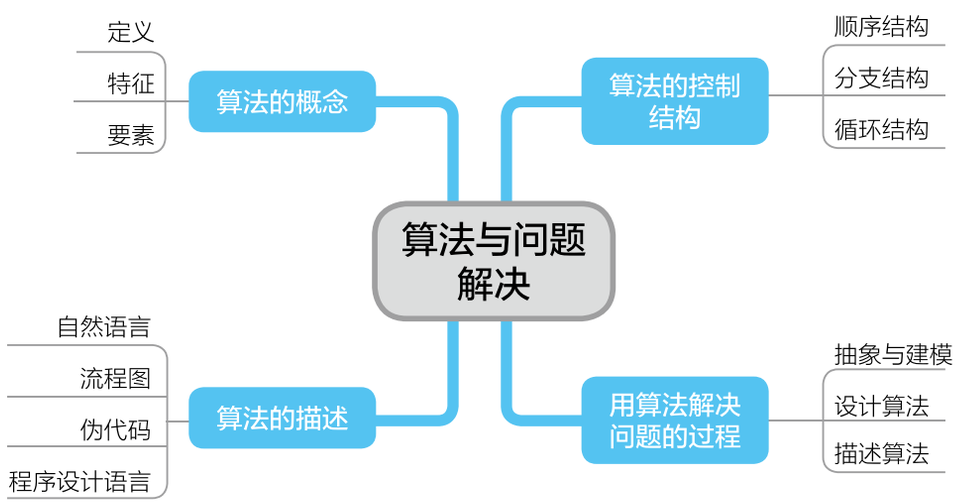
- 假期一起去自由行旅游，旅途中有的费用每人可以均摊（比如住宿、吃饭、交通等费用），但有的费用并不能均摊（比如品尝美食、游乐项目等消费，各人不同）。如果旅途中每次消费都集中付款，最后对产生的总费用按照每个人的实际消费进行快速结算，可以设计一个怎样的算法来解决这个问题？

- 超市通过关联分析找出关联度较大的商品，并将这些商品集中陈列，以达到提升商品销售量的目的。为了实现关联分析，首先要统计关联次数。如果已经收集了超市前期的销售数据，设计怎样的算法可以实现商品的关联次数统计？

学习目标

1. 能从生活和学习中发现算法，理解算法的内涵和外延。
2. 能根据实际问题进行抽象与建模，并完成算法的设计与描述。
3. 初步认识算法的多样性。

内容总览



2.1 算法的概念及描述

算法可以帮助算法执行者高效地解决问题。只有掌握了算法的定义，设计出符合算法特征的有效算法，并围绕算法要素加以准确描述，才能运用针对性的算法解决问题。

2.1.1 算法的概念

新学期开学，为了方便高一新生完成注册、缴费等事宜，学校在校园入口处摆放如图2.1.1所示的“高一新生报到流程”示意图。

这个“高一新生报到流程”示意图就是一个算法，该算法可以帮助高一新生解决“入学报到”问题。现实生活中处处存在着解决各种问题的步骤，如计算两个正整数最大公约数的步骤，无人汽车利用摄像头和雷达获得数据进行计算与决策，并发出操控汽车指令的过程等。这些都属于算法。

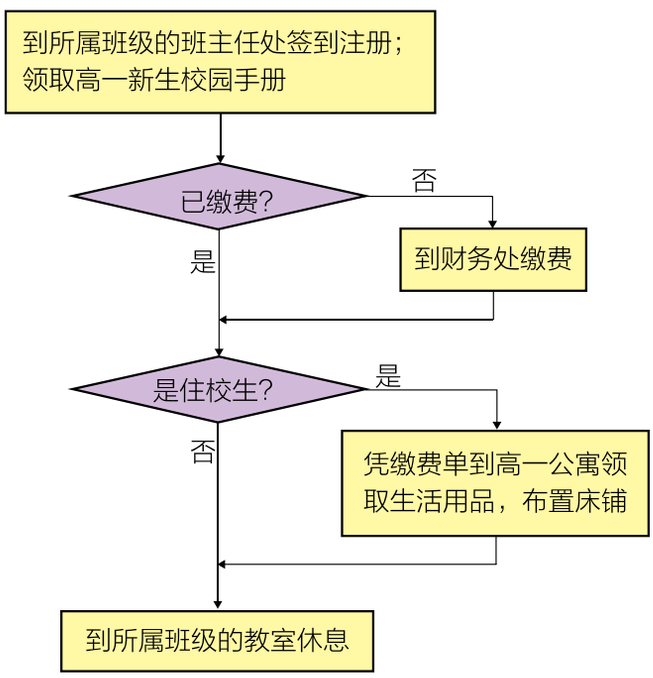


图2.1.1 高一新生报到流程

1. 算法的定义

算法自古有之，古代数学家欧几里得在《几何原本》中提出的“辗转相除法”（也称欧几里得算法）就是算法，利用该算法可以求出任意两个正整数的最大公约数。用“辗转相除法”计算正整数m和n的最大公约数的步骤如下：

①输入两个正整数m和n。

②若m＜n，则交换m和n的值。

③以m除以n，相除得到的余数为r。

④若r＝0，则输出n的值，算法结束；否则，执行步骤⑤。

⑤令m＝n，n＝r，返回步骤③继续执行。

我国古代数学专著《九章算术》中也记载了求两个正整数最大公约数的算法——更相减损术。除《九章算术》外，我国古代还出现过其他各类有关算法的著作，如《孙子算经》《数书九章》等。

古代的算法主要指的是“算术”，即数值的算术运算。随着科学技术的发展，“算法”的外延和内涵逐渐发生着变化。

广义地讲，“算法”指的是解决问题或完成任务的一系列步骤。广义的算法中，需要解决的问题不仅仅指传统意义上的计算任务（算术），也可以是社会生活中各种事务的处理。比如，假期的自由行旅游方案、一道菜的烹饪过程、洗衣机的操作步骤等。这些算法的执行者往往是人，人们按照算法的要求逐步执行，最终解决问题。

在计算机科学领域内，“算法”指的是用计算机解决问题的步骤，是为了解决问题而需要让计算机有序执行的、无歧义的、有限步骤的集合。这些需要解决的问题不仅包含了数值计算，还包含了非数值计算的数据处理。例如，在包含上万人信息的数据中查找某人的数据、导航程序中两个地点之间最短路线的规划等。解决这些问题的算法执行者是计算机，为了让计算机理解算法中的步骤，需要用计算机能理解的语言来描述算法并将其输入到计算机中，这个过程就称为计算机程序设计。

如果只是凭概要的方法去解决问题，往往会影响问题的解决。而将方法细化成算法执行者能理解的、更明确的步骤，使之成为算法后，算法执行者就能按照算法要求逐步解决问题。例如，只要将“辗转相除法”细化为前面所述的算法，那么只会基本算术运算的人就能够计算出两个正整数的最大公约数。

拓展链接

百钱买百鸡

“鸡翁一，值钱五；鸡母一，值钱三；鸡雏三，值钱一；百钱买鸡百，问翁、母、雏各几何？”这是我国古代数学家张丘建在《算经》中提出的经典问题。同时，他还在书中给出了解决该问题的算法“鸡翁每增四，鸡母每减七，鸡雏每益三，即得”。

拓展链接

穷举算法

穷举算法也称枚举算法，指的是在求解过程中，先按照一定的顺序一一列举所有可能的解，然后用条件判断列举出的可能解是否为正确解。穷举法一般适合解决解集为离散的且范围明确的问题。

再如，“用求根公式求解一元二次方程的实数根”是一个方法，它给出了求一元二次方程解的总则和方向，但没有说明具体如何做，对于一个不知道求根公式具体运算过程的人或机器来说，没有任何实际意义。如果将该方法分步骤具体描述（例如，先根据方程系数计算 $\Delta=b^2-4ac$ 的值，然后根据求根公式 $\frac{-b\pm\sqrt{\Delta}}{2a}$ 分别计算两个实数根），使之成为一个算法，则具有基本计算能力的人或计算机就可以按照该算法完成求解任务。

在用计算机解决问题时，同样需要将解决问题的方法细化成计算机能理解的各个步骤，并通过输入设备告诉计算机，计算机才能按照算法解决问题。

2. 算法的特征

根据算法的定义，算法具有下列特征：

①有穷性。一个算法的处理步骤必须是有限的。无论具体需要执行的操作步骤有多少，这个数量必须是确定的。例如，“计算斐波那契数列的前n个元素的过程（n为一个确定的正整数）”就符合有穷性，因为无论n有多大，求数列元素的次数肯定不会超过n；而“计算斐波那契数列的所有元素”就不符合有穷性，因为这里没有明确元素的个数，计算次数就无法确定（无限）。

②可行性。一个算法中的每一步操作与要求都应该是算法执行者（人或者机器）可以实施的，同时在现实环境中能做到并且能在有限的时间内完成。

③确定性。算法中对于每个步骤的执行描述必须是明确的。例如，“取区间的中点”就不符合确定性，因为没有说明在哪个区间；而“取区间[100，200]的中点”或“取区间[i，j]的中点（i，j为某一确定的值）”就符合确定性的要求。

④0个或多个输入。算法被执行者实施时，一般需要从外部获取可变的数据。如果问题求解时所有数据都是不变且已知的，则所需数据包含在算法中，不必再在执行时输入数据；如果一些初始数据需要在算法执行时临时获取以适应不同情形的问题，则算法需要包含一个或多个输入。

⑤1个或多个输出。算法必须包含至少一个输出，以告诉外界问题求解的结果。如果一个算法没有输出，那么这个算法就没有意义，因为算法的核心价值就是解决问题，而解决问题的终极目标就是需要知道结果究竟如何。

问题与讨论

为防止用户账户被盗，在用户登录账户时，有些信息系统会限制用户尝试输入密码的次数（如图2.1.2），一旦超出限定的次数，系统就会禁止输入并要求进行注册账户验证。下面为某系统验证用户输入密码正确与否的算法：

图2.1.2　用户登录界面

①密码输入错误次数初始化为零。

②接受用户输入的密码。

③将用户输入的密码与原来设置的密码比较，若相同则转⑦，否则转④。

④密码输入错误次数增加1。

⑤若密码输入错误次数少于5，输出信息“密码错误，请再次输入密码！”，然后转⑥；否则，输出信息“密码输入错误已达5次，请通过注册邮箱找回密码”，然后转⑧。

⑥接受用户输入的密码，然后转③。

⑦密码正确，进入系统。

⑧密码验证算法结束。

请结合上述算法，谈谈算法的特征在其中的具体体现。比如，该算法体现了“可行性”特征，因为算法中的“判断密码正确性”“密码输入错误次数统计”等处理都是现实中确实可以实现的。

3. 算法的要素

用计算机解决问题，本质上是以“数据运算”的方式来实现的。各种“运算”有时需要依次进行，有时需要根据条件选择一部分进行，有时又需要重复执行某些“运算”，这些“运算”顺序的调控就需要借助控制转移来实现。因此，通过算法让计算机解决问题时，数据、运算及控制转移就成为算法的要素。

（1）数据

用算法解决问题时，必须明确参与运算的初始数据、运算时产生的中间数据以及代表问题解决的结果数据。如图2.1.3所示，在洗衣机执行洗衣算法前，必须进行洗涤时间、漂洗次数、脱水时间、每次洗涤所加水量的设置，并将这些设置产生的数据输入到算法中，洗衣机才能按照需求工作。

图2.1.3　全自动洗衣机操作面板

（2）运算

在对数据进行运算时，必须明确每一步的运算是什么、对哪些数据进行运算等。例如，在洗衣机的控制算法中必须包含“洗涤时间的计时”“漂洗次数的统计”以及“判断加水是否到达50升”等运算。

（3）控制转移

在算法执行过程中，有时需要根据数据或运算结果的特点进行不同的处理，这时就需要运用控制转移来执行不同的操作。例如，在洗衣机控制算法的进水过程中，如果水量达到50升则关闭进水阀，否则不关闭进水阀，这个环节就采用了一种称为分支结构（也称选择结构）的控制转移。再如，漂洗过程中，当漂洗次数未达到2次时，需要继续加水到50升，然后重复原来的漂洗处理，这种需要实现重复执行某些操作的控制转移称为循环结构。

问题与讨论

很多设备的“自动”功能，都是内部算法控制的结果。比如，在夏天把空调温度设定在26℃（如图2.1.4），每当空调内部的温度传感器测得室内温度小于或等于26℃时，算法就会“告诉”空调已经到达目标温度，可以暂停工作，空调就会“自动”暂时关闭压缩机的运行。这样，既确保了室内温度，又实现了节能环保。

图2.1.4　壁挂式空调内机

还有很多设备用算法来帮助设备实现自动化。与同学讨论交流，哪些设备采用算法实现了自动化？并尝试说出这些设备实现自动化控制的算法。

2.1.2　算法的描述

一个作曲家想让一个钢琴家演奏他创作的新作品，首先他要写出琴谱，然后钢琴家才能根据琴谱进行演奏。同样地，设计出一个解决问题的算法，也需要用能被算法执行者理解的形式加以呈现，才能被算法执行者理解并执行。算法的这种呈现就称为算法的描述。

掌握各种算法的描述方法，在解决问题过程中选择恰当的方式合理地描述算法，是解决问题的一个重要环节。

常见的算法描述方式有自然语言、流程图、伪代码、计算机程序设计语言等。

1. 用自然语言描述算法

自然语言是人们在日常生活中交流使用的语言，如汉语、英语、德语、日语等。用自然语言描述算法通俗易懂，且不需要进行专门的学习和训练。

当然，人们在长期用自然语言描述算法的过程中也逐渐形成了一些约定俗成的规则，了解这些规则有助于快速应用自然语言描述算法。例如，在算法执行过程中如果需要记住某些数据，且这些数据可能会发生改变，可用“变量”（往往用由字母、数字、下划线等组成的一串字符来表示）来表示这些数据；算法执行过程中某些数据会根据问题特征以不同的值参与计算，这些数据可以通过“输入”来获得。

停车场车位探测中的算法

某停车场每个车位的上方都装有传感器（车位探测器）、前方装有车位指示灯（空车位显示绿色，否则显示红色），如图2.1.5所示。车位上方的传感器探测下方的车位是否为空，然后根据探测结果控制车位指示灯的颜色并向区域控制器发送该车位的状态信息（“空车位”或“非空车位”）。

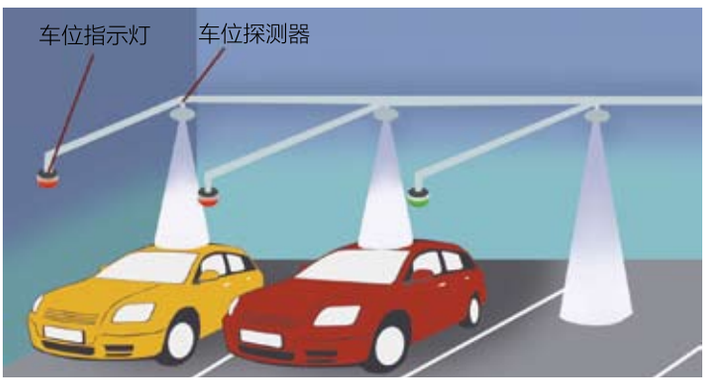


图2.1.5　停车场中的车位探测

为了根据传感器的探测结果进行相应处理，可以将传感器回传的数据作为输入数据，控制设备就可以根据该输入数据，控制车位指示灯的颜色并向区域控制器输出该车位的状态信息。下面用自然语言来描述该数据处理过程中的算法。

分析：传感器每隔一段时间对车位进行探测，在某个时刻，若传感器测得结果为空车位，则将指示灯设置为绿色，同时向区域控制器输出“空车位”的信息；否则，将指示灯设置为红色，同时向区域控制器输出“非空车位”的信息。

将传感器回传的数据作为输入数据并进行数字化设定，若测得空车位，则用输入数值1表示，否则用输入数值0表示。因为该数据会发生改变，所以用变量flag保存该输入数据。

根据上述分析，某个时刻对车位进行数据处理的任务可以界定为以下问题：

输入flag的值，根据flag的值设置车位上方指示灯的颜色，并输出车位状态（“空车位”或“非空车位”）。

解决本问题的算法可以用自然语言描述如下。

（1）输入变量flag的值。

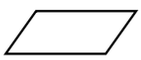
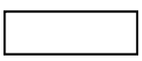
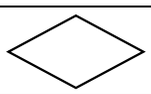
（2）若flag的值为1，则设置指示灯为绿色，输出“空车位”；否则，设置指示灯为红色，输出“非空车位”。

2. 用流程图描述算法

用自然语言描述算法虽然通俗易懂，但也存在难以避免的问题。例如，在描述某些操作时容易出现歧义（面对同样的文字描述，不同的人产生不同的理解）；在描述根据条件进行不同处理的算法时比较烦琐。此时，采用流程图来描述会显得比较直观和易于理解。

流程图用一些图形符号表示规定的操作，并用带箭头的流程线连接这些图形符号，表示操作进行方向。流程图描述算法结构清晰、寓意明确。常用的流程图基本图形及其功能如表2.1.1所示。

表2.1.1　常用的流程图基本图形及其功能

图形	名称	功能
	开始／结束符	表示算法的开始或结束
	输入／输出框	表示算法中数据的输入或输出
	处理框	表示算法中数据的运算处理
	判断框	表示算法中的条件判断
	流程线	表示算法中的流向
	连接点	表示算法中的转接

根据表格所示的基本图形，前面用自然语言描述的车位探测中的算法可用流程图描述如下（如图2.1.6）。

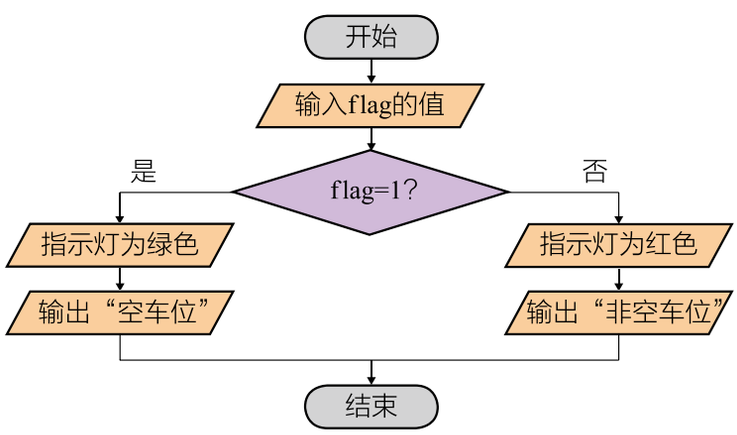


图2.1.6　用流程图表示车位探测中的算法

拓展链接

N–S图

“N-S图”是由美国学者纳西（Nassi）和斯奈德曼（Shneiderman）提出的一种在流程图中完全去掉流程线，全部算法写在一个矩形框内的算法描述方式。相比于原来的流程图描述，结构性显得更好，也更有助于高效地编写程序。前面车位探测中的算法，可用N-S图表示成如下形式。

输入flag的值	
是　flag=1?　否	
指示灯绿色	指示灯红色
输出“空车位”	输出“非空车位”

问题与讨论

结合实例，与同学一起分组讨论：为什么算法必须要有“输出”，但有时却可以没有“输入”？

3. 用伪代码描述算法

流程图虽然直观易懂，但当分支增多时会出现流程线相互交叉而影响算法理解的情况。而且，自然语言和流程图描述的算法要转化为计算机能理解的计算机程序时，中间还需要较多的语义解释和格式转换工作。由此，人们想出了用伪代码来描述算法。

伪代码指的是一种比较直观简洁的、符号接近计算机程序代码的算法描述方式，其风格很像计算机程序设计语言，但又不是真正的可以被计算机理解的代码。伪代码的表示方法没有统一、严格的规定，只要定义合理、表达正确即可。

伪代码由于语法比较接近计算机程序设计语言，所以描述的算法更加紧凑简练，也便于进一步转化为相应的计算机程序。

为了今后使用的方便，在本书中对伪代码表示做如下的语法约定：

（1）条件判断语句

格式1：If 条件 then
　　（语句序列1）
　Else
　　（语句序列2）

语义：若条件成立，则执行语句序列1（由一个或多个语句组成），否则执行语句序列2。如果在该条件基础上还需要做进一步的条件判断，那么可以进行条件判断语句的嵌套，在If语句中继续放入另一个If语句。

当条件成立需要执行特定的语句序列1，而条件不成立不需要执行特定的处理时，则采用下列格式：

格式2：If 条件 then
　　（语句序列1）

（2）循环语句

格式：while 条件
　　（循环体）

语义：循环体由一个或多个语句组成。循环语句执行时，先判断条件是否成立，若条件成立则执行循环体，循环体执行完后再次判断条件是否成立，如此重复，直到某次条件不成立，则结束循环语句，接着去执行循环语句后面的语句。

根据上面的约定，车位探测中的算法可用伪代码表示如下：

```
flag←车位探测结果;          #将测得的车位当前状态值输入给变量flag
If  flag=1  then
   (指示灯绿色
   输出“空车位”)
Else
   (指示灯红色
   输出“非空车位”)
```

4. 用计算机程序设计语言描述算法

无论是自然语言描述的算法，还是流程图或者伪代码描述的算法，计算机都无法理解并执行。为了让计算机帮助人们真正解决问题，需要将算法用某种计算机程序设计语言来描述，这个过程称为程序编写（或称代码编写）。

世界上有很多计算机程序设计语言，实际工作中可以根据问题特点选择恰当的程序设计语言来描述算法。前面车位探测中的算法可用C++程序设计语言描述如下：

```
void MainWindow::on_pushButton_clicked()
{
      int flag = ui->lineEdit->text().toInt();
      if (flag == 1){
            ui->label_4->setStyleSheet("color:green;");
            ui->label_4->setText("绿色");
            ui->label_5->setText("空车位");
      } else {
            ui->label_4->setStyleSheet("color:red;");
            ui->label_4->setText("红色");
            ui->label_5->setText("非空车位");
      }
}
```

图2.1.7　当前车位无车时的输出信息

图2.1.8　当前车位有车时的输出信息

该程序的运行结果分别如图2.1.7和图2.1.8所示。

该算法还可以用Python程序设计语言描述如下：

```
flag=int(input("输入车位状态值：  "))
if flag==1:
    print("绿色")
    print("空车位")
else:
    print("红色")
    print("非空车位")
```

拓展链接

计算机程序设计语言

计算机程序设计语言经历了“机器语言→汇编语言→高级语言”的发展历程。机器语言中的指令由“0”“1”二进制码组成，机器执行效率高但可读性、维护性差。为了提升编程的效率，科学家用特定的符号（助记符）来表示各个机器指令，发明了汇编语言。科学家后来又发明了高级语言，用接近人类日常用语的符号来表示各类指令。常见的高级语言有Basic、C、C++、Java、Python、Ruby等。

问题与讨论

1. 某智能停车场车位引导系统中，通过一个区域控制器来统计、显示该区域空车位情况。当该区域控制器接收到每个车位发送的状态信息（“空车位”或“非空车位”）后，它会统计该区域当前的空车位总数，并将该信息通过引导屏呈现在停车库入口处（如图2.1.9），引导驾驶员有方向地寻找空车位。

图2.1.9　车位引导系统中的引导屏

与同学讨论，该区域控制器可用怎样的算法来解决空车位的统计和显示问题？

2. 与同学一起讨论，是否可以设计出除本节介绍的四种算法描述方式之外的其他方式？并将第1题的算法用自己设计的方式加以描述。

实践与体验

体验算法的多样性

现实中解决一个问题的算法往往具有多样性，即可用多种不同的算法来解决同一个问题。比如，求两个正整数的最大公约数问题，既可用“辗转相除法”解决，也可用“更相减损术”解决。

实践内容：

在《九章算术》中记载的“更相减损术”算法可以求任意两个正整数的最大公约数。了解“更相减损术”算法，并比较“辗转相除法”和“更相减损术”在求解两个正整数的最大公约数问题中的特点。

实践步骤：

1. 确定“更相减损术”算法的信息源。
2. 根据找到的算法介绍，用伪代码描述该算法。
3. 比较“辗转相除法”和“更相减损术”在计算效率和数学本质上的异同点。

结果呈现：

围绕下列两个问题进行思考，选择合适的平台和形式发布思考结果。

1. 两个算法的计算效率哪个更高？为什么？
2. 当两个正整数呈现何种趋势特点时，两个算法的计算效率差异愈加明显？

思考与练习

1. 用自然语言描述“高一新生报到流程”（如图2.1.1）对应的算法。

2. 用智能电饭煲烧饭时，在算法的控制下，当饭烧熟时，智能电饭煲会自动停止高热烧饭，转为低热保温。这是因为锅底的温度传感器每隔一定时间（比如200毫秒）会将温度数据传送给算法，一旦发现温度达到103°C（此时锅中水被蒸发完），算法就会控制继电器释放触点，让电饭煲停止烧饭，转入低热保温模式。

图2.1.10所示的流程图描述了某个时刻智能电饭煲根据输入的温度数据进行判断、处理的算法，则在流程图中①、②标记处应填入的内容分别为①_____，②_____（选填“变为低热保温”或“继续高热烧饭”）。

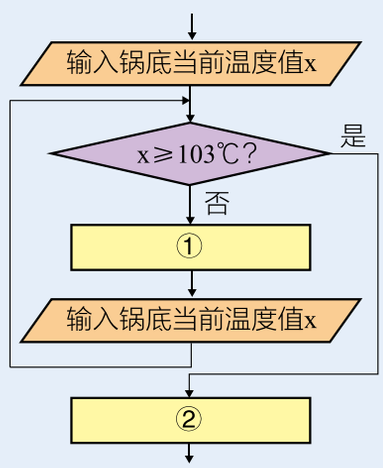


图2.1.10　智能电饭煲烧饭算法流程图

2.2 算法的控制结构

玩过积木的人都知道，即使很复杂的积木作品，都是由最基本的积木块（不妨称为基本结构）通过各种组合构成的。类似地，无论内容怎样复杂、功能如何强大的算法，也都由基本的结构组合而成，这些基本的结构称为算法的控制结构。算法的控制结构有三种，即顺序结构、分支结构和循环结构。

2.2.1 顺序结构

在网上购买火车票时，必须严格按照顺序依次进行各步操作（如图2.2.1），具有这种特点的算法结构称为顺序结构。

顺序结构指的是算法中各个步骤按照先后顺序依次执行的结构。如图2.2.2所示，首先执行“第一个操作”，然后按照顺序再依次执行“第二个操作”“第三个操作”。

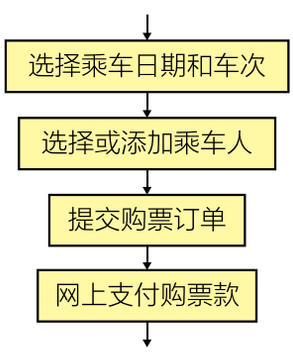


图2.2.1 网上购票的算法

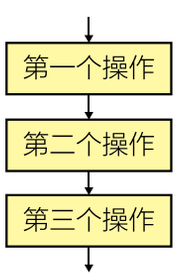


图2.2.2 顺序结构算法的一般结构

由上面的例子可知，顺序结构的算法执行时，必定具有下列特点：

①每个步骤按照算法中出现的顺序依次执行。

②每个步骤一定会被执行一次，而且只执行一次。

2.2.2 分支结构

一个一元二次方程是否存在实数根，需要根据条件“$b^2-4ac\geqslant 0$”是否成立来判断。如果条件成立就输出“有实数根”，否则就输出“无实数根”（如图2.2.3）。这种先进行条件判断，再根据判断结果分别执行不同处理的控制结构就称为分支结构（也称选择结构）。

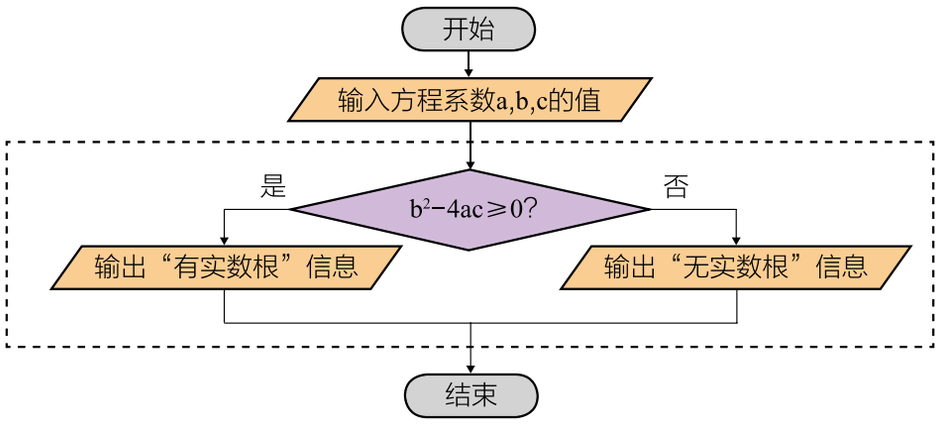


图2.2.3　判断方程是否有实数根的算法流程图

从上面的流程图可知，分支结构在执行时具有下列特点：

①首先进行条件判断，根据条件满足与否来决定执行哪个分支。

②在一个分支结构中，必定有一个分支被执行，其他的分支则被忽略。

用流程图描述分支结构的算法时，流程线会从条件判断框（菱形）上面的角进入，在进行条件判断后，从条件判断框的左、右或者下面的角走向各个分支。

在解决问题的一个完整算法中，有时需要几种控制结构的协同才能完整地表示解决问题的全部过程。如图2.2.4所示，如果对判断方程是否有实数根不做细化，那么整个算法是顺序结构，即先执行第一步“输入方程系数a，b，c的值”，然后再按照顺序执行第二步“根据系数判断方程是否有实数根并赋值给变量f”。如果将算法进一步细化，那么原来第二步处理就需要用分支结构来实现。

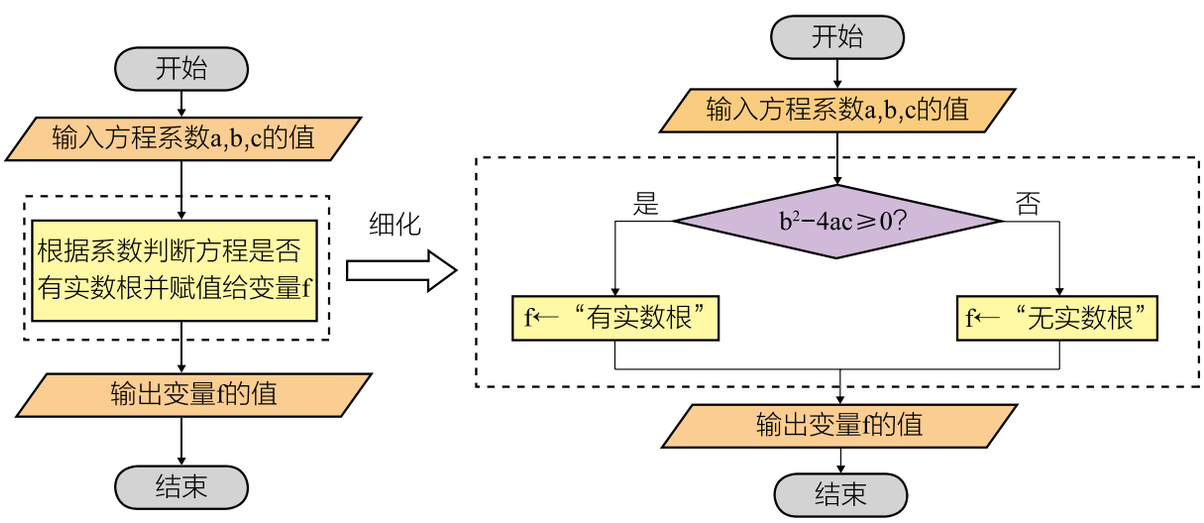


图2.2.4　从粗到细的算法细化

2.2.3　循环结构

在智能电饭煲烧饭的过程中，如果当前所测温度没有达到103°C，则算法会控制电饭

煲继续高热烧饭，此时采用的是循环结构。

算法执行过程中，在条件控制下，某些操作步骤需要重复执行（循环）的控制结构称为循环结构。图2.2.5所示的循环结构中，算法会先判断循环条件是否满足。若满足则进入循环，执行循环体，然后再次判断循环条件是否满足，若满足则再次进入循环，执行循环体，然后再次判断循环条件是否满足……直到某次循环条件不满足，退出循环。

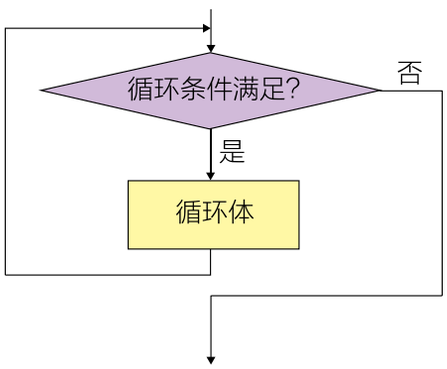


图2.2.5　循环结构算法的流程图

循环结构的重复执行（循环）并不是没有限制的，而是在条件控制下的一种可控的重复。当需要重复处理的条件不满足时，重复处理必须能及时结束。这样才符合算法的有穷性特征。

图2.2.5所示的算法在执行时，如果循环条件始终满足，那么循环体永远会被执行，此时算法陷入“死循环”，也就违背了算法的有穷性特征。因此，算法在设计时应避免此类情形的发生。

与分支结构的算法流程图描述类似，用流程图描述循环结构的算法时，通常算法会从条件判断框（菱形）上面的角进入，在进行条件判断后，会从条件判断框的左、右或者下面的角走向需要重复执行的循环体或者退出循环结构。

超市收银系统

具有一定规模的超市，收银通常由超市管理系统来完成。收银时，收银员用扫描仪逐个扫描商品上的条形码，随着一连串的“嘀”声，收银员可以快速地完成顾客所购商品的费用结算（如图2.2.6）。

为了开发一个超市管理系统中的收银子系统，需要针对收银员的收银过程设计一个算法来解决上面所述的收银问题。


图2.2.6　大型超市的信息化收银台

分析：一个完整的超市管理系统是由通信网络、计算机（包括服务端和客户端）、计算机程序等组成的综合系统。商品在超市上架之前会有一个数据录入环节，通过数据录入，每个商品的编码、数量、价格等数据会保存到超市管理系统的数据库中。收银过程中，扫描仪扫描条形码，本质上是将条形码对应的商品编码输入到计算机中，然后再以输入的编码为关键字，从已有的数据库中查找与之相符的商品数据（如商品名称、价格）并进行金额统计结算。另外，在通常的超市收银子系统中，收银员可按“结算”键来告诉系统，商品数据输入结束并开始结算。

根据上述分析，可以确定算法需要解决以下问题：

输入每个商品的编码，即时显示商品的相关信息（编码、商品名称、价格）。根据所有输入商品的数据统计顾客的应付总金额并输出结算清单。

根据问题要求，算法应先根据输入的商品编码在数据库中找到对应商品的数据。然后显示当前商品的数据，将当前商品的价格累加到应收总金额中，同时即时显示当前应收总金额。处理完当前商品后，需要继续输入下一个商品的编码，并重复上述处理。这个重复过程可以用循环结构来实现。当按“结算”键时，算法应能退出循环，并输出结算清单。

为方便上述算法的描述，可以事先做出一些符号化的设定。用code表示商品的编码，用sum表示顾客应付的总金额，用x表示每个商品的价格。

综上所述，解决该问题的算法可用如图2.2.7所示的流程图来描述。

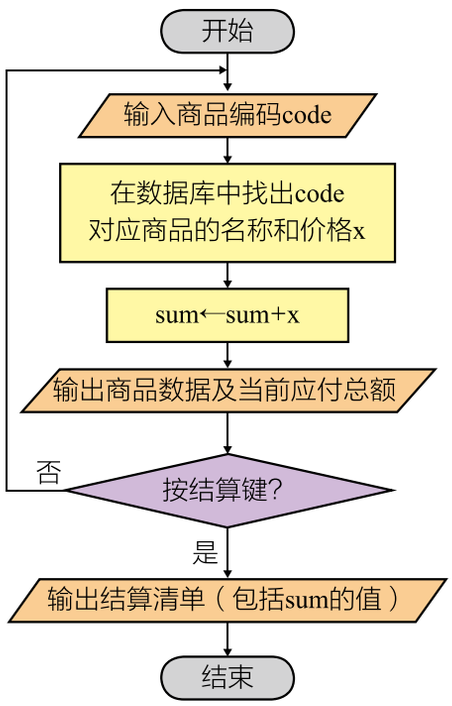


图2.2.7　解决超市收银问题的算法流程图

拓展链接

程序设计中的“累加器”

“累加器”指的是算法执行过程中对同类事物或数据进行统计计算的实现技术。上述算法中的“sum←sum+x”就起到了累加的作用。

思考与练习

智能农业大棚通过传感器、控制器、网络设施和计算机程序等来实现大棚的自动化管理（如图2.2.8）。例如，自动温度控制系统中的温度传感器每隔一定时间采集大棚中的温度，一旦温度超过预设的最高温度40℃，控制系统会启动通风和喷水系统实现降温；如果温度低于预设的最低温度18℃，控制系统会启动加热器，给大棚升温。

图2.2.8　智能农业大棚

（1）自动温度控制系统进行温度控制的算法用流程图描述如图2.2.9所示，请完善该流程图，在①、②处填入合适的内容。

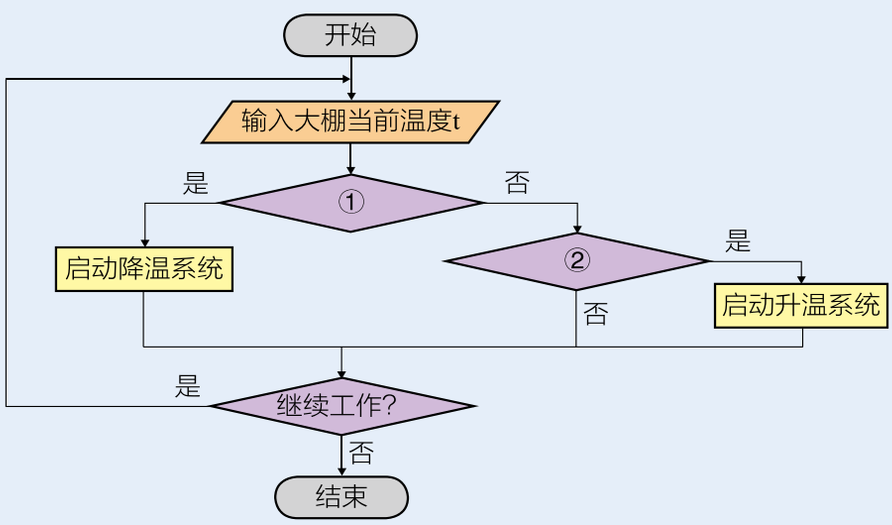


图2.2.9　智能农业大棚温控系统的算法流程图

（2）请用自然语言描述上述算法，并尝试比较用流程图和自然语言描述算法的特点。

2.3 用算法解决问题的过程

用计算机解决问题时，由于实际问题情境的复杂性，需要先对实际问题进行抽象与建模，再根据建立的计算模型设计算法，并将算法用合适的方式加以准确描述。

“动动有奖”是某手机走路计步器程序（程序界面如图2.3.1所示），它能根据系统传递给它的走路步数给运动者奖励，运动者可以用累计的“奖金”去换取软件开发商提供的各种体育用品。具体的奖励规则如下：

1. 每天走路的前1000步奖励0.3金，之后每2000步奖励0.1金（不足2000步没有奖励），每天最高奖励不超过3金。

2. 每天必须到计步器页面点击“领奖”按钮，才能领取昨日走路奖金。

3. 如果连续3天领奖成功，从第4天起走路奖金翻1倍（乘以2），每天最高奖励不超过6金。翻倍期间若有1天没有领奖（即连续每天领奖行为中断），则翻倍权益取消，重新连续3天领奖成功才能继续翻倍。

图2.3.1 “动动有奖”APP系统界面

任何一个信息应用系统都是硬件和一系列应用程序的有机结合，上述“动动有奖”APP系统就需要在手机、运动传感器等硬件基础上结合计算机程序来实现。其中计算机程序不仅需要从传感器获得每天的走路步数，还需要根据每天是否成功领奖来确定“奖金”是否翻倍。

下面就在抽象与建模的基础上，尝试设计“动动有奖”的算法。在后续学习了计算机程序设计的知识后，还可以在计算机上实现“动动有奖”算法的程序。

第一步：抽象与建模

抽象与建模指的是从现实项目的真实情境中提炼出核心的要素并加以确定或假设，最终定义出一个有明确已知条件和求解目标的问题，并用数学符号描述解决该问题的计算模型。

对于本问题，可以依次通过下列步骤逐步分析出计算模型。

1. 提炼核心要素并加以确定或假设

本问题的已知数据包含了每天走路的步数，以及每天是否成功领取前一天奖金的标记。因为这些数据在事先都是不确定的，所以需要通过输入将数据传递给算法，不妨用变量 X 来表示每天走路的步数，用 F 表示是否成功领取了每天的奖金（1表示成功领取，0表示没有领取）。

为了使建立的问题模型具有一般性，可以认为需要统计的走路天数是不定的，所以用变量 n 来表示这个可变的数据。

2. 用数学符号描述解决问题的计算模型

明确了问题的已知条件后，需要明确问题的解决目标。这个问题的解决目标比较直接，就是统计 n 天过去后，该用户一共拥有的“奖金”总数。

基于上述分析，可以得出解决该问题的计算模型如下：

已知n（$1 \leqslant n \leqslant 30$）组数据：$X_i$，$F_i$（$1 \leqslant i \leqslant n$），计算“奖金”总和total。

$$\text{其中 total}=\sum_{i=1}^{n} S_i,\quad S_i=\begin{cases} 0 & (F_i=0) \\ t & (F_i=1\text{且}F_{i-1}，F_{i-2}，F_{i-3}\text{不全为}1) \\ 2t & (F_i=1\text{且}F_{i-1}，F_{i-2}，F_{i-3}\text{全为}1) \end{cases}$$

$$t=\begin{cases} 0 & (X_i<1000) \\ 0.3 & (1000 \leqslant X_i<3000) \\ 0.3+\lfloor (X_i-1000)\div 2000 \rfloor \times 0.1 & (3000 \leqslant X_i \leqslant 55000) \\ 3 & (55000<X_i) \end{cases}$$

注：“$\lfloor\ \rfloor$”表示对表达式的值向下取整。

如果有下列4组数据：

$X_1=4500$，$F_1=1$；$X_2=9870$，$F_2=1$；$X_3=12890$，$F_3=0$；$X_4=57890$，$F_4=1$。

则根据上述计算模型得到的“奖金”总和为4.1金。

第二步：设计算法

有了计算模型后，就可以遵循算法的特征、围绕算法的要素设计算法。

对任何数据的处理，总体上都需要经历下列三个步骤：

①输入数据。

②处理数据。

③输出处理结果。

本问题需要输入的数据是数据数量规模n以及n组X_i，F_i的值。

处理数据时，需要根据计算模型对每组X_i，F_i（$1 \leqslant i \leqslant n$）依次进行处理。由于每天

处理数据的规律是相同的，所以数据处理部分可用循环结构来解决，每执行一次循环体就处理一天的数据。按照“自顶向下、逐步细化”的结构化程序设计思想，对前面的算法进行如下细化：

①输入总天数n。

②表示天数的变量i初始化为1。

③若i≤n，则转④，否则转⑦。

④输入第i天的数据（包括第i天走路步数X_i，是否成功领取第i天“奖金”的标记F_i）。

⑤根据当前输入的数据X_i，F_i，统计该天领取的奖金并累加到总奖金total中。

⑥表示天数的变量i增加1，然后转③。

⑦输出变量total的值。

至此，解决“动动有奖”APP的算法已经基本形成，接下来在一些细节上做进一步的细化，使得算法的各个处理步骤变得更具体、更清晰。

第三步：描述算法

由于计算时涉及较多的条件判断，所以可以用图2.3.2所示的流程图来进一步描述解决该问题的算法。为了判断“奖金”是否翻倍，用变量c保存连续成功领奖的天数。

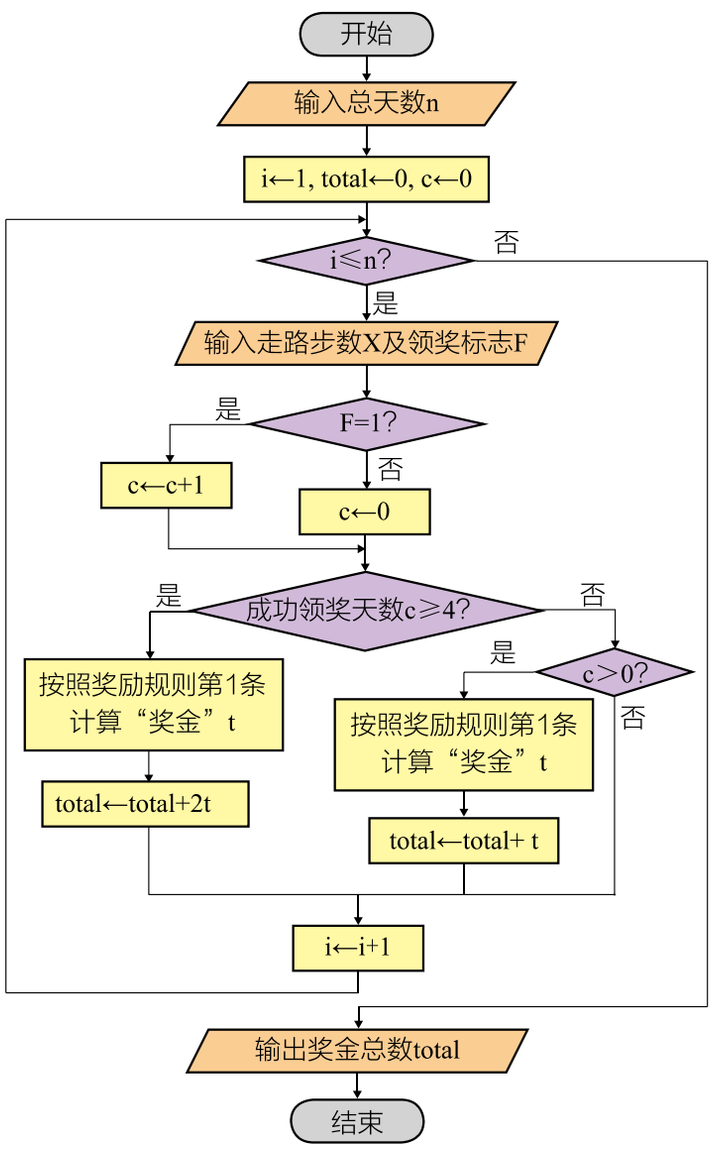


图2.3.2 “动动有奖”APP的算法流程图

拓展链接

常用算法介绍

本例采用模拟策略来设计算法，即根据现实事务的实际流程和要求逐步进行处理，以达到数据处理的目标。计算机科学家根据各种问题的模型特征提出了各种针对性的算法设计策略，如穷举算法、顺序查找算法、对分查找算法、冒泡排序算法、深度优先搜索法以及动态规划等。

思考与练习

上述算法中，“按照奖励规则第1条计算‘奖金’t”在两个环节中出现，请根据算法功能完成下列练习。

（1）改进算法，使得算法中只有一个环节出现“按照奖励规则第1条计算‘奖金’t”。

（2）请进一步细化原算法中的“按照奖励规则第1条计算‘奖金’t”，并用流程图进行描述。

巩固与提高

1. 任意输入三个数，输出其中最小数的值。设计解决该问题的算法并用流程图描述该算法。

2. 任意输入三个数，按从小到大的顺序依次输出这三个数。设计解决该问题的算法并用流程图描述该算法。

3. 比较表中三种算法描述方式的特点，并填写下表。

算法描述方式	优点	不足
自然语言		
流程图		
伪代码		

4. 学校举行迎新年大合唱比赛，每个班级合唱结束后，主持人会当场宣读10个评委的打分，然后再统计出总评分作为该班级的最终得分。每个班级表演所得的总评分计算规则如下：

（1）在10个评委的打分中去掉一个最高分和一个最低分。

（2）剩余8个分数的平均分即为总评分。

为了在比赛现场能快速根据主持人所宣读的10个分数计算出总评分，小伟需要为比赛编写一个计算机程序。请你设计一个解决该问题的算法，并用合适的方式描述。

项目挑战

为超市寻找关联次数最多的商品

人们通过研究发现，将某些不同商品（比如休闲食品和饮料）陈列在一起销售，能使相关商品的销售量增长20%～30%。为了寻找这些能相互促进销量的商品，就需要进行商品的关联分析。“支持度”是反映商品关联性的一个重要度量值，为了统计相关商品的支持度，需要先统计相关商品的关联次数。关联次数指的是不同商品同时出现在同一个购物篮中的次数。

图2.3.3　超市商品陈列

表2.3.1　购物篮中的关联商品

购物篮	购物流水号	商品
购物篮1	201609270027	x_1，x_2，x_3，x_4，x_5，x_6
购物篮2	201609270028	x_1，x_4，x_7，x_8，x_9
购物篮3	201609270029	x_2，x_5，x_6，x_7，x_9

如表2.3.1所示，商品x_1和x_4的关联次数是2（这两个不同商品在购物篮1和购物篮2中同时出现）；商品x_2，x_5和x_6的关联次数是2（这三个不同商品在购物篮1和购物篮3中同时出现）；商品x_5和x_6的关联次数是2（这两个不同商品在购物篮1和购物篮3中同时出现）。

学校超市想通过商品的关联分析来改进商品的陈列，从而方便同学们购物。为了帮助超市进行商品的关联分析，设计怎样的算法可以解决不同商品间的关联次数统计问题？

项目任务

根据超市某个时期内的流水记录，找出超市内关联次数最多的一对或多对商品（这里只统计两个不同商品之间的关联次数，即两个不同商品如果同时出现在同一个购物篮中，则称这对商品关联1次）。具体要求如下：

1. 抽象与建模。明确问题的已知条件和求解目标，建立一个可行的计算模型。

2. 设计算法，并选择合适的方式进行描述，为后阶段用计算机程序求解提供支撑。

过程与建议

你可以借鉴前面为“动动有奖”APP设计算法的过程与方法来完成这个项目任务，具体步骤如下：

1. 抽象与建模

要解决该问题，需要明确一些条件。比如，数据由哪些构成？数据之间的关系如何？如果关联次数最多的不同商品不止一对，求解目标应如何确定？等等。

为此，我们需要进一步分析问题，尽量使问题在一个明确的范围内求解，最终用自然语言清晰地表述问题并建立计算模型。比如，对于上述问题，一种可参考的分析思路如下：

已知超市一段时期内流水账的数据，且数据已经通过整理保存在Access数据表中，该数据表的结构及内容如表2.3.2所示（流水号相同的商品属于同一个购物篮）。

表2.3.2　保存于Access数据表中的超市流水账

货号	品名	数量	金额	流水号
398626	散装鸡蛋	0.485	13.15	201607290093
986712	带柄西蓝花	0.505	6.05	201607290093
486134	老酸奶	4	20	201607290093
419803	鲜奶	3	13.95	201607290093
999088	散装番茄	0.51	2.63	201607290093
954798	牛里脊	0.216	23.76	201607290093
992927	杀白秋鸭	1.425	37.62	201607290093
…	…	…	…	…
965423	青菜	0.51	1.02	201608020032
486134	老酸奶	2	10	201608020032
954798	牛里脊	0.32	35.2	201608020032

根据表中的数据，通过抽象与建模，明确一个“统计关联次数最多的商品对”的问题，并用清晰、准确的语言描述该问题，然后再建立计算模型。

2. 设计算法

依托前面建立的计算模型，根据已知数据及数据之间的关系，分析通过已知数据如何

解决该问题的算法，并描述你的分析过程。

3. 描述算法

选用恰当的方式（自然语言、流程图或者伪代码等）来描述你的算法，帮助算法执行者顺利地按照你的算法逐步执行（如果要让计算机来实施处理，需要用计算机程序设计语言准确地描述该算法）。

4. 展示交流

为了推广你的项目，你需要将项目及解决方案在一定的平台、以一定的形式展示给大家，通过展示实现交流，并基于交流对项目做出进一步的完善。为了使展示达到较好的效果，你需要就“展示形式”（网页、PPT、文字稿、视频等）、“展示平台”（互联网、校园网、计算机教室局域网、现场讲解等）等方面做出规划，并简单阐明理由。

评价标准

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

评价条目	说明	评分（1～10分）	评分主要依据阐述	后续完善方向
抽象与建模	确定的问题明确，建立的计算模型正确			
设计算法	设计的算法正确，且能输出符合实际的正确结论			
算法描述	描述方式的选择或组合有助于正确理解算法			
展示方式	项目成果展示方式的选择利于他人较好地了解项目的意义以及问题求解的特点			

拓展项目

1. 小谢刚从大学毕业，来到某城市一家IT公司从事软件开发工作。目前他和别人合租在一套公寓中，为了给自己更多的自由，他决定贷款买一套面积在50平方米左右的单身公寓。小谢看中了一套售价大约为1.5万元/平方米的现房。他目前的经济状况如下：

（1）第一年实习期间每月收入约7000元，第二年转正后每月收入大约1.2万元。今后，随着经验的丰富和职位的提升，收入应该还会有20%～50%的提升空间。

（2）每月开支：房租费600元，手机通信费约100元，水电费约50元，餐费约1200元，其他费用约500元。

（3）小谢所在的公司愿意为他提供30万元的无息贷款。

请设计一个算法，计算贷款买房的费用，并帮助小谢制订一个贷款买房计划。

2. 兵马俑、华山、壶口瀑布等景点是众多旅游者热捧的对象。作为一名“驴友”，小刚打算在暑假组织几名同学一起去陕西旅游。旅途中有的消费项目产生的费用是每人可以均摊的（比如住宿、吃饭、交通等费用），但有的消费费用并不能均摊（比如品尝美食、游乐项目等消费，各人不同）。而且，一般情况下肯定是集中付款比较方便。因此，旅途中每次消费集中付款，最后根据每个人的消费情况统一结账，这个过程的实施方案就成为小刚这个组织者必须事先要解决的问题。请你为小刚设计一个算法，用于解决旅途中记账以及旅游结束后统计每人应承担费用的计算问题。

3. 饮食结构很大程度上决定了人的健康状况，一种较为健康的饮食结构的总体荤素比例应为1∶4，即鱼虾、肉、蛋等荤食每1份需要搭配由蔬菜、水果和主食（主要指米饭、面食等）组成的4份素食。进一步地，在荤食中鱼虾、肉、蛋的比例能达到2∶2∶1，而素食中蔬果和主食的比例能达到2∶3。请从学校食堂收集一周的相关数据，设计算法进行统计，并根据统计结果向学校食堂提出饮食结构合理化的改进建议。

第三章 3

算法的程序实现

用计算机程序解决问题时，需要将算法用某种计算机程序设计语言精确描述（也称“编写程序”），并在计算机上调试运行直至正确，才能最终解决问题。

在计算机科学中，常见的程序设计语言有Python、C++、Java、Ruby、Visual Basic等。同一个算法可以用不同的程序设计语言来实现。尽管不同的程序设计语言特点不同，语法规则也可能不同，但是程序设计方法基本相同。本章将以Python程序设计语言为基础，介绍将算法进行程序实现的一般过程与方法。

问题与挑战

- 在数学研究中，哥德巴赫猜想被称为近代三大数学难题之一，直至今日，数学家们仍未能完整证明哥德巴赫猜想。数学家们力求在数值上对哥德巴赫猜想进行验证，但仅凭人工的计算验证根本无法达到所期望的数值。如何借助计算机用尽可能大的数值来验证哥德巴赫猜想？

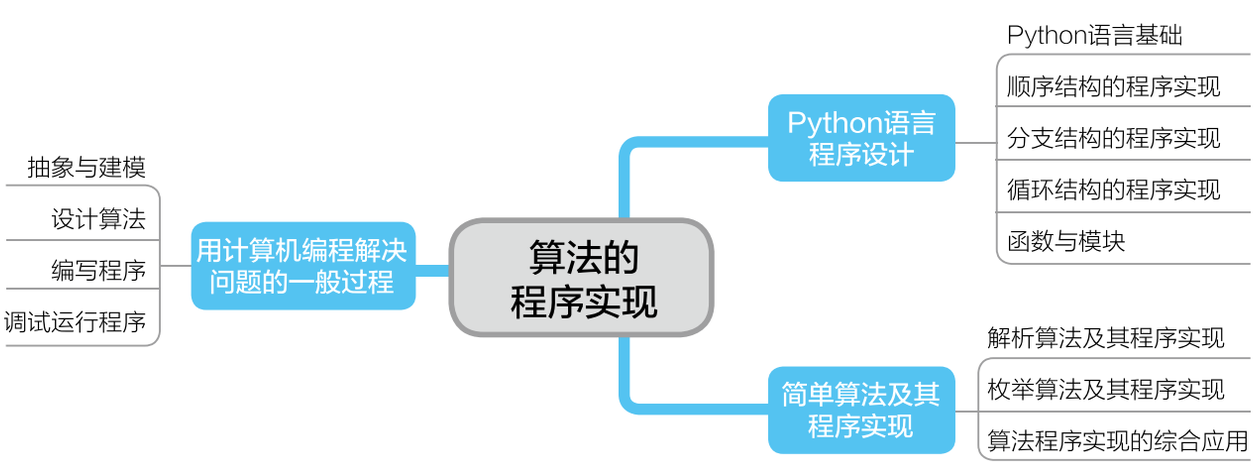
- 答题卡识别技术大大减少了阅卷、统计、分析的工作量，机动车牌自动识别和处理技术在各停车场收费、道路测速等方面大大提高了车辆管理的效率，这些技术是通过程序实现的。那么，这些程序的算法是如何设计的？程序又是如何实现的？

- 根据超市销售数据进行商品的关联分析，可以挖掘出顾客的购物喜好，进而提高超市的经营管理水平。在利用超市销售数据，完成寻找关联次数最多的一对商品的算法设计后，如何将该算法进行程序实现，并进一步拓展到统计所有商品间的关联次数？

学习目标

1. 了解用计算机编程解决问题的一般过程。
2. 掌握Python语言的基本知识，体验程序设计的基本流程。
3. 能用程序实现简单算法，掌握程序调试与运行的方法，感受算法的效率。

内容总览



3.1 用计算机编程解决问题的一般过程

计算机已成为人们解决问题的重要工具。例如，用Word解决文字处理的问题，用Excel解决一般的数据计算、统计的问题等。但由于现实问题的多样性，并不是所有的问题都可以用现成的计算机程序来解决。因此，针对这些问题，需要通过抽象与建模、设计算法、编写计算机程序来解决。

下面以编写计算机程序绘制一个正多边形为例，了解用计算机编程解决实际问题的一般过程。

1. 抽象与建模

正多边形的各边边长相等，各内角度数也相等。因此，绘制一个正多边形，可以通过“画一条边，旋转一定角度后再画一条边”的重复操作来完成。例如，图3.1.1呈现的是绘制一个正六边形的过程。

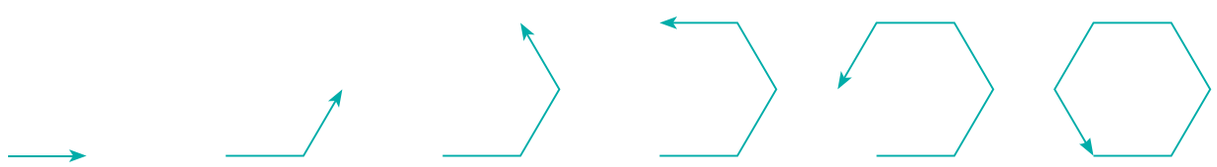


图3.1.1　绘制正六边形的过程

绘制正多边形，除了要知道它的边数n和边长a，关键是要计算出每次旋转的角度。因此，解决这个问题的计算模型可以表示如下：

假设正多边形的边数为n，边长为a。

则内角度数d的值为：$d=(n-2)\times 180\div n$。

每次旋转的角度为：$180-d$。

2. 设计算法

基于问题的抽象与建模，绘制一个正多边形的算法可以做如下描述：

① 输入要绘制的正多边形的边数n和边长a。

② 计算正多边形的每个内角度数d，其中$d=(n-2)\times 180\div n$。

③ 将以下过程重复执行n遍：画一条长度为a的线段，再将画笔方向向左（逆时针）旋转（$180-d$）度。

3. 编写程序

要让计算机按照预先设计的算法进行处理，需要将该算法用计算机程序设计语言描述，形成计算机程序。绘制正多边形的算法用Python语言描述如下：

```
import turtle
n=int(input("请输入正多边形的边数n："))
a=int(input("请输入边长a："))
d=(n-2)*180/n
t=turtle.Pen()
for i in range(n):                    #重复执行n遍
    t.forward(a)                      #向前绘制长度为a的线段
    t.left(180-d)                     #向左旋转（180-d）度
turtle.done()
```

4. 调试运行程序

通过运行程序，计算机会自动执行程序中的命令。但是，在将算法进行程序实现时，可能会因为录入错误、语法错误、逻辑错误等原因，导致程序不能正常运行或输出错误的结果。此时，需要对程序进行调试，以便发现错误并进行修正。例如，字母大小写的疏忽可能直接决定程序能否正常运行，程序中参数的调整可能影响输出图形的形状。

问题与讨论

在用计算机编程解决问题的过程中，算法与程序两者之间的关系如何？

思考与练习

1. 请描述用计算机编程验证“哥德巴赫猜想”的一般过程。

2. 为了分析某网站中不同中文词的出现频率，利用现有的工具软件无法满足所有的统计要求，需要自己编写程序来解决。请结合本事例，通过流程图来描述用计算机解决该问题的一般过程。

3.2 Python语言程序设计

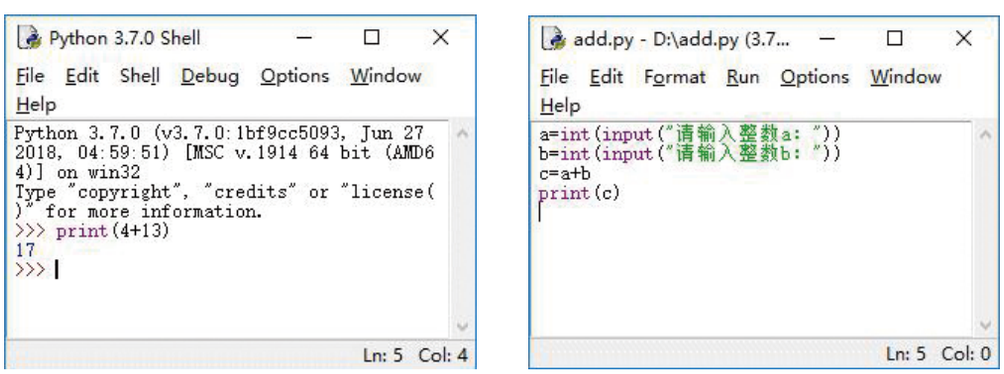
Python是一种面向对象、解释型的计算机程序设计高级语言，其语法简洁清晰，方便对数据进行组织和处理；具有丰富和强大的库，可以支持很多日常问题的程序实现。因其解释性语言的本质，Python在大多数平台上都是一种理想的脚本语言，特别适合应用程序的快速开发。

3.2.1 Python语言基础

使用Python语言编程解决问题时，需要严格遵守Python语言的语法规则，并选择合理的程序运行环境运行程序。

1. 编程环境

编写Python程序比较方便的方式是使用集成开发环境（Integrated Development Environment，简称IDE），如图3.2.1所示界面为用于Python程序开发的IDE：IDLE。



甲　解释器操作界面　　乙　编辑器操作界面

图3.2.1　IDLE的操作界面

IDLE是开发Python程序的基本IDE。打开IDLE，会出现交互式解释器Python Shell，如图3.2.1甲所示，可以通过它在IDLE内部执行Python命令，也可以在Python Shell的提示符“>>>”后输入任意的语句、表达式或者一小段代码进行测试。

如在Python Shell中直接输入：print(4 +13)，按回车键后，就可以得到4和13相加的结果：17。

```
>>> print(4 +13)
17
```

如果在Python Shell中输入：print("Hello"+" Python!")，按回车键后，将显示“Hello Python!”。

```
>>>print("Hello"+" Python!")
Hello Python!
```

除此之外，IDLE还带有一个编辑器，如图3.2.1乙所示，可以用来编辑Python程序。通过Python Shell菜单：“File”—“New File”，打开编辑器，输入相应的Python程序。例如，求两个整数和的程序如下：

```
a=int(input("请输入整数a："))
b=int(input("请输入整数b："))
c=a+b
print(c)
```

通过编辑器菜单：“Run”—“Run Module”，运行程序。程序运行时，在解释器Python Shell的交互界面中输入相应数据，可得到如下结果：

```
>>>
请输入整数a：4
请输入整数b：13
17
```

拓展链接

集成开发环境

集成开发环境（IDE）是提供程序开发环境的应用程序，一般包括代码编辑器、调试器和图形用户界面工具。目前，用于编写Python程序的IDE较多，如IDLE、Spyder、Wing、PyCharm等。如图3.2.2是Spyder的操作界面。

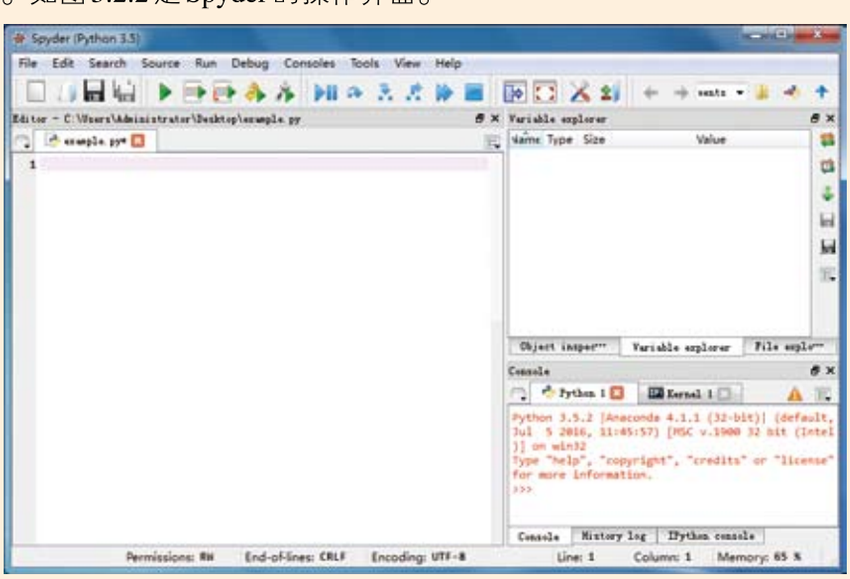


图3.2.2　Spyder的操作界面

2. 数据类型与表达式

数据按照其本身特征可以分为若干种不同的类型，常见的Python基本数据类型如表3.2.1所示。

表3.2.1　Python常见数据类型

数据类型名	数据表示形式
整型	数学中的整数，如1，–8080，0等 十六进制数（用0x前缀），如：0xff00，0xa5b4c3d2等
实型	数学中的实数，如3.14，–9.01等 用科学记数法表示的实数，如0.000012可以写成1.2e–5等
字符串型	用单引号、双引号或三引号表示，如：'这是一个字符串！'、"This is a string!"、'''X'''等
布尔型	只有两种值：True和False。布尔型数据可以进行not、and和or等逻辑运算

通过程序的执行对数据进行加工处理，基本运算是数据处理中最常用的手段。Python的基本运算包括算术运算、关系运算和逻辑运算三大类。变量、常量、运算符和圆括号等按一定的规则组合构成一个表达式，可以用来描述数据的计算过程或各种条件的判断等。算术运算是运用算术运算符进行数的加、减、乘、除等数学运算，表3.2.2所示的是Python中常用的算术运算符。

表3.2.2　Python算术运算符

运算符	表达式	描述	示例	优先级
**	x**y	求x的y次幂	5**2结果为25	1
*	x*y	将x与y相乘	5*2结果为10	2
/	x/y	用x除以y，产生实数值	5/2结果为2.5	2
//	x//y	用x除以y，取整数部分	5//2结果为2	2
%	x%y	用x除以y，取余数	5%2结果为1	2
+	x+y	将x与y相加	5+2结果为7	3
–	x–y	将x减去y	5–2结果为3	3

Python中的算术运算存在着优先级顺序，优先程度最高级别为1，级别数字越大，优先级越低。在同一个表达式中，如果有一个及以上的运算符，那么先执行优先级高的运算，同优先级的基本运算按照自左向右的顺序执行。例如，表达式“123–123//100*100”的运算结果为23。

关系运算的结果是一个布尔值，若两个数据之间指定的关系成立，则计算的结果值为真（True），否则为假（False）。例如，3==5就是运用关系运算符“==”（等于）对数字3和

5进行比较，其值为False。在Python中，常用的关系运算符如表3.2.3所示。

表3.2.3　Python关系运算符

运算符	表达式	描述	示例
>	x>y	x大于y	5>2结果为True
<	x<y	x小于y	5<2结果为False
>=	x>=y	x大于等于y	5>=2结果为True
<=	x<=y	x小于等于y	5<=2结果为False
==	x==y	x等于y	5==2结果为False
!=	x!=y	x不等于y	5!=2结果为True
in	x in y	x是y的成员	"5"in"2"结果为False

其中，“in”成员资格运算符用来检查一个值是否包含在指定的序列中，以下示例使用了成员资格运算符分别检查“w”和“x”是否出现在字符串“rw”中。

```
>>>"w" in "rw"
True
>>>"x" in "rw"
False
```

逻辑运算符经常用于描述复杂情况的判断。在Python中，常用的逻辑运算符如表3.2.4所示。

表3.2.4　Python逻辑运算符

运算符	表达式	描述	示例
and	x and y	布尔“与”	True and False结果为False
or	x or y	布尔“或”	True or False结果为True
not	not x	布尔“非”	not False结果为True

3. 变量和赋值语句

程序设计时，有些数据是未知或可变的，为了更灵活地使用这些数据，可以使用变量来存储。为了能对变量进行访问，需要对变量进行命名。在Python中，变量名可以包括字母、数字和下划线，但不能以数字开头，而且字母区分大小写。所以，Plan9是合法变量名，而9Plan不是；变量名teacher和TEACHER表示两个完全不同的变量。由于Python是动态类型语言，因此在使用前不需要预先声明变量的数据类型。例如：

```
>>>degrees_cel=26.0
>>>degrees_cel
26.0
>>>degrees_cel="26.0"
>>>degrees_cel
'26.0'
```

上例中，语句“degrees_cel=26.0”创建了一个名为degrees_cel的变量，变量的类型是实型且值为实数26.0。而下面的语句“degrees_cel="26.0"”执行后，变量degrees_cel的类型变成了字符串型。因此，在Python中，变量的值和类型都可以改变。

类似“degrees_cel=26.0”的语句称为赋值语句，“=”为赋值符号。如：

```
>>> number=0
>>> number=number+1
>>> print(number)
1
```

上面的赋值语句“number = number + 1”也可以写成“number += 1”，其功能是先将变量number值加1，然后将计算结果赋值给变量number。其中，“+=”为运算符“+”和赋值符号“=”的组合。类似的赋值运算符还有“-=”“*=”“/=”和“%=”等。

4. 基本数据结构

程序设计时，需要根据数据之间的逻辑关系和处理任务的要求，将各种数据组合成具有一定结构的复合体。例如，在超市购物清单中，一条商品信息由“编号”（字符串型）、“名称”（字符串型）、“单价”（实型）、“数量”（实型）等数据项组成，在Python中可用列表来组织和存储；而在处理学生基本档案信息时，可以用字典组织数据。列表、字典都是Python中常用的数据结构。

（1）字符串和列表

字符串和列表都是由一些数据元素共同组成的一个序列整体。字符串是由0个或多个字符组成的序列，如字符串“Hello”的第一个字符是“H”，第二个字符是“e”……类似地，列表也是由0个或多个元素组成的序列，其中的元素可以是数字、字符串等混合类型的数据，甚至是其他的列表。也就是说，不同类型的元素可以存在于同一列表中。列表一旦创建，就可以添加或删除其中的元素。列表用方括号“[]”来表示，元素之间以逗号“,”分隔。例如，某个商品信息的“编号”“名称”“数量”数据项的值分别为：BH60018、苹果、50，若要利用列表来组织这些数据，可创建如下名为info的列表：

```
>>>info=["BH60018","苹果",50]
```

字符串和列表在创建以后都可以进行某些特定的操作，如提取序列中的一部分元素、判断某个元素是否为序列的成员等。

字符串、列表中的元素都是通过索引来定位的。如图3.2.3所示，第一个元素的索引是0，第二个元素的索引是1，以此类推不断递增。

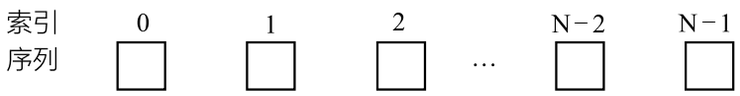


图3.2.3　包含N个元素的序列索引

info列表的索引如图3.2.4所示。

图3.2.4　info列表索引

字符串或列表中的元素可以通过索引进行访问，如下所示：

```
>>> info=["BH60018","苹果",50]
>>> info[2]
50
>>> s="Hello"
>>> s[1]
'e'
```

上面的实例中，info[2]表示的是列表的第三项，该项元素为50。所有序列都可以通过这种方式进行索引查找，来获取某个元素。但在索引查找时，不能访问一个不存在的元素，比如程序访问info[10]时，程序就会报错，提示索引值越界了。

若要访问的不是单个元素，而是一定范围内的多个元素，可以通过冒号“:”间隔的两个索引参数（开始元素序号、结束元素序号的后一个序号）来实现。

```
>>> info[0:2]
['BH60018', '苹果']
>>> s[1:4]
'ell'
```

上面的实例中，info[0:2]表示从列表索引为0的元素开始取，一直取到索引为1的元素。

（2）字典

字典和列表类似，可包含多个元素。字典中的每个元素包含两部分内容：键和值。键通常用字符串或数值来表示，值可以是任意类型的数据。键和值两者一一对应，且每个键只能对应一个值。类似于现实中的字典，可以通过查到某个特定的字（键），从而找到它的注解（值）。如要利用字典组织和存储四种文具的名称和数量信息，键（名称）和值（数量）的对应关系如图3.2.5所示。

键和值在字典中以成对的形式出现，并以如下方式标记：d={key1:value1,key2:value2,…}。

键-值对用冒号分隔，各个对之间用逗号分隔，所有这些都包括在花括号“{}”中。字典中的元素是没有顺序的，引用元素时以键为索引。例如：

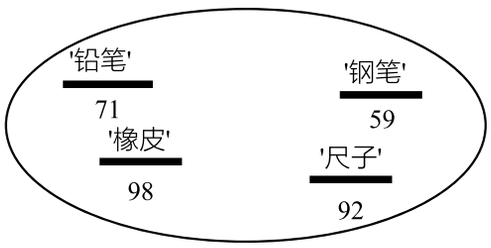


图3.2.5　字典结构示例

```
>>> dic={"铅笔":71,"钢笔":59,"橡皮":98,"尺子":92}
>>> print(dic["铅笔"])
71
```

上述程序第一行语句创建了一个名为dic的字典，共有4个元素，第1个元素包含了键"铅笔"和值71，第2个元素包含了键"钢笔"和值59 ……第二行输出字典dic中键"铅笔"对应的值。

问题与讨论

1. 通过网络学习，了解各种程序设计语言的特点，通过比较得出Python语言的优缺点。

2. 请列举日常生活中所接触到的数据（如通讯录、成绩表等），并说明它们在使用Python语言描述时适用的数据类型或数据结构。

3.2.2　顺序结构的程序实现

在编写顺序结构算法的程序时，应按照算法中的顺序逐步实现。

例如，将两个整型变量a、b的值互换的算法流程图如图3.2.6所示，使用Python语言来实现此算法的程序如下：

```
a=int(input("请输入整数a的值："))
b=int(input("请输入整数b的值："))
c=a          #语句1
a=b          #语句2
b=c          #语句3
print("a=",a)
print("b=",b)
```

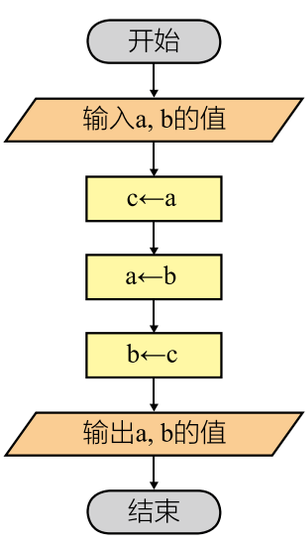


图3.2.6　交换两个变量值的流程图

问题与讨论

如果调换上述程序中语句1、语句2或语句3的位置，程序的运行结果将会如何变化？

本例程序中，“#”后面的内容为注释，在程序运行时不执行。注释主要用于对程序代码进行说明，便于程序的理解和维护。

上述程序中的int、input、print等都是Python的内建函数。其中，input函数实现了用户和计算机程序的交互输入，当用户根据提示“请输入整数a的值：”，输入合适的数据并按回车键后，input函数会将用户输入的数据以字符串型接收到程序中；int函数将接收到字符串型数据转换为整型数据；print函数实现计算结果输出。Python中有许多内建函数，常见的内建函数如表3.2.5所示。

拓展链接

两个变量值的直接交换

在Python中，两个变量值的互换可不借助第三个变量而直接进行，如：a,b=b,a。

表3.2.5　Python常见内建函数

函数	描述
print(x)	输出x的值
input([prompt])	获取用户输入
int(object)	将字符串和数字转换成整型
float(object)	将字符串和数字转换成实型
abs(x)	返回x的绝对值
help()	提供交互式帮助
len(seq)	返回序列的长度
str(x)	将x转换成字符串
chr(x)	返回x对应的字符
ord(x)	返回x对应的ASCII值
round(x[,n])	对x进行四舍五入（如果给定n，就将数x转换为小数点后有n位的数）
max(s,[,args...])	返回序列的最大值（如果给定多个参数，则返回给定参数中的最大值）
min(s,[,args...])	返回序列的最小值（如果给定多个参数，则返回给定参数中的最小值）

3.2.3　分支结构的程序实现

算法进行程序实现时，分支结构可以用if语句来实现。

1. if 语句

一般格式是：

```
if<条件>:
    <语句块1>
else:
    <语句块2>
```

条件是一个表达式，它的值可以是真（True）或假（False）。当条件为真时，执行语句块1中的语句，否则（条件为假）执行语句块2中的语句。如果程序只需要对条件为真的情况做出处理，那么if语句可省略else及语句块2部分，格式变成：

```
if<条件>:
    <语句块>
```

拓展链接

语句块缩进

在Python中，行尾冒号的作用是告诉Python接下来要创建一个新的语句块。因此，只要某一行以冒号结尾，它接下来的内容就应该有缩进。Python中有一个惯例：总是将语句块缩进4个空格。

if语句中的冒号表示下方紧接着一个语句块。在Python中，语句块是一行或放在一起多行的语句，一般通过行缩进来标识。同一个if语句中，if、else下方的语句块必须采用相同的缩进。

区间测速

目前，国内很多高速公路都启用了区间测速。所谓区间测速，是在同一路段上布设两个监测点，基于车辆通过前后两个监测点的时间来计算车辆在该路段上的平均行驶速度，并依据该路段上的限速标准判定车辆是否超速，如图3.2.7所示。

现有一段长为25千米的测速区间，小车的限速是100千米/时。数据中心需要编写一段程序，用来判断某辆小车在此测速路段是否超速。

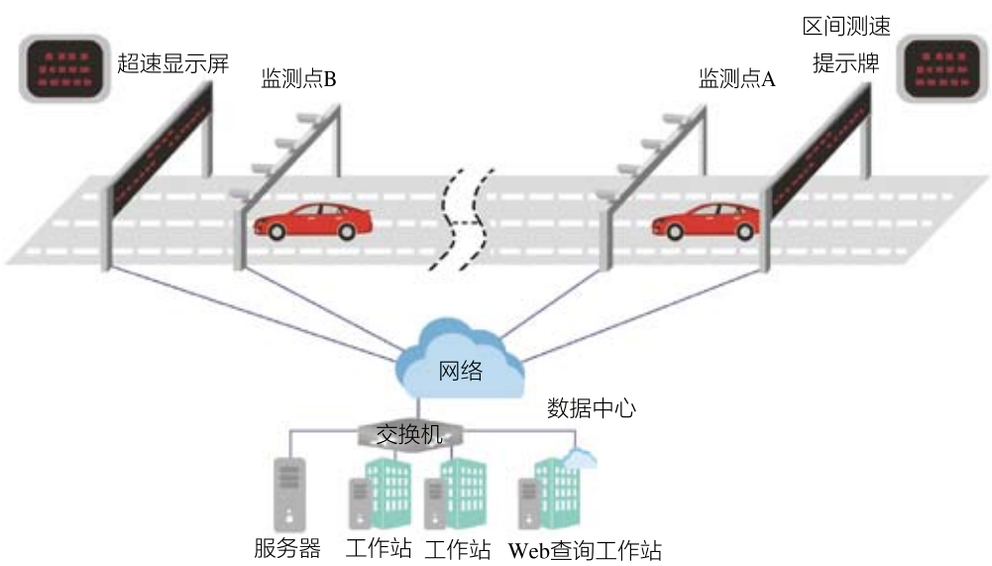


图3.2.7　区间测速示意图

（1）抽象与建模

要判断某辆小车在该测速区间内是否超速，首先要求出该小车的平均车速，然后与限速100千米/时进行比较，若平均车速小于等于限速，属于正常；否则就判定为超速。而求平均车速时需要提供区间距离和行驶时间。区间距离已经明确，小车的区间行驶时间可通过小车经过前后两个监测点的时间差来计算，为简化问题，可将时间差作为行驶时间的输入数据。各个数据相应的数据类型及变量名如表3.2.6所示。

表3.2.6　区间测速数据分析表

变量名	数据类型	含义
s	数值型——整型	区间距离（千米）
t	数值型——整型	用时（秒）
v	数值型——实型	平均车速（千米/时）

通过上述的问题抽象，可建立如下计算模型：

判断结果=$\begin{cases}\text{“正常”}（v\leqslant 100）\\ \text{“超速”}（v>100）\end{cases}$，其中$v=\frac{s\times 3600}{t}$。

（2）设计算法

根据上述计算模型，解决问题的关键是根据v值做出判断，可采用分支结构设计算法。该算法的流程图如图3.2.8所示。

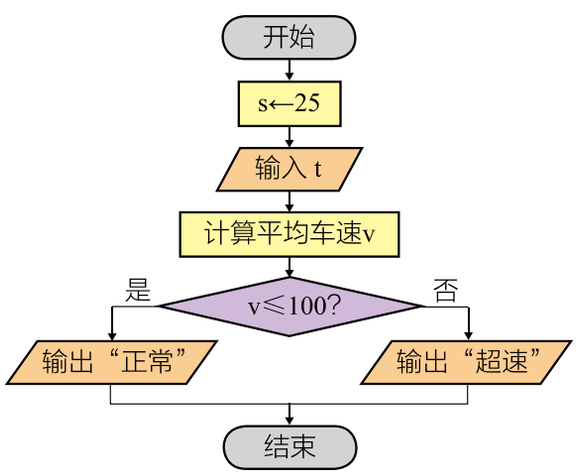


图3.2.8　超速判断算法流程图

（3）编写程序

根据算法编写的程序及测试结果如下：

程序	测试结果

s=25
t=int(input("请输入用时（秒）："))
v=s*3600/t
if v<=100:
　　print("正常")
else:
　　print("超速")
	请输入用时(秒)：800
	超速

问题与讨论

分析下面两段代码，找出两者的区别。

代码一	代码二

t=int(input("请输入用时（秒）："))
v=25*3600/t
if v<=100:
　　print("正常")
else:
　　print("平均车速",round(v,1))
print("超速")
	t=int(input("请输入用时（秒）："))
	v=25*3600/t
	if v<=100:
	　　print("正常")
	else:
	　　print("平均车速",round(v,1))
	　　print("超速")

2. if–elif语句

如果上面项目中要求在超速的情况下再区分超速的程度，根据“超过规定时速且不足20%”“超过规定时速20%以上且不足50%”“超过规定时速50%以上且不足70%”“超过规定时速70%以上”等标准进行分类，此时就需要对多个条件进行判定。在Python中，可用带有elif子句的if语句来实现，其格式是：

```
if<条件1>:
    <语句块1>
elif<条件2>:
    <语句块2>
……
elif<条件N>:
    <语句块N>
else:
    <语句块N+1>
```

一个if语句可以包含多个elif子句，最后一个else子句是可选的。elif子句仅当其if语句中的条件为假时才执行。如果if语句和elif子句中的条件都不为真时，末尾的else子句的语句块就会被执行。因此，带有elif子句的if语句有一个很重要的特性：只要某个条件为真，计算机就会执行其所对应的语句块，然后就退出该语句。

在上述“区间测速”的基础上，如果某辆小车超速，数据中心能同步显示超速的程度，如“超过规定时速且不足20%”“超过规定时速20%以上且不足50%”“超过规定时速50%以上且不足70%”“超过规定时速70%以上”。如何用程序来实现这个目标？

（1）抽象与建模

根据超速标准，设定如表3.2.7所示的判断条件。

表3.2.7　超速标准及其判断条件对应表

超速标准	判断条件
超过规定时速且不足20%	100＜v＜120
超过规定时速20%以上且不足50%	120≤v＜150
超过规定时速50%以上且不足70%	150≤v＜170
超过规定时速70%以上	v≥170

（2）设计算法

当条件v≤100不成立时，需要参照表3.2.7进一步对平均车速进行判断。算法对应的流程图如图3.2.9所示。

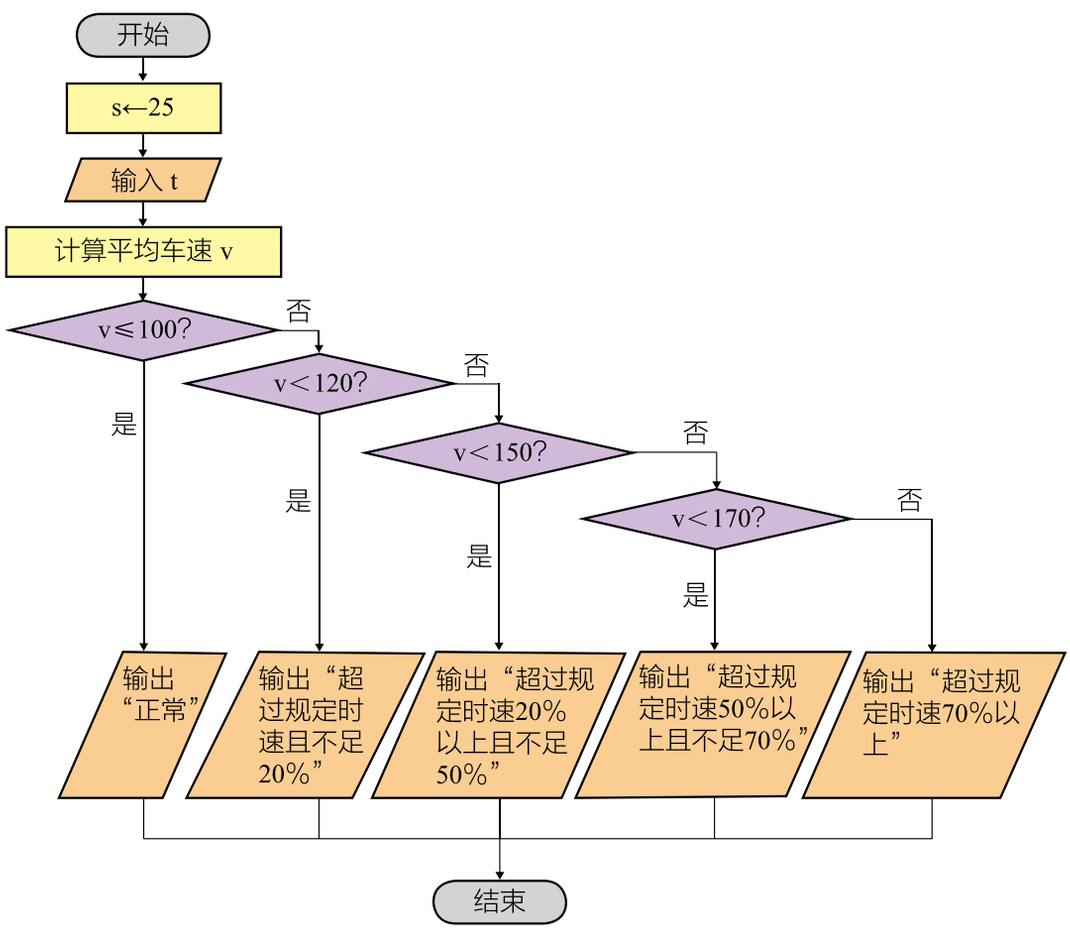


图3.2.9　超速程度判断算法流程图

（3）编写程序

根据算法编写的程序及测试结果如下：

程序	测试结果

s=25
t=int(input("请输入用时(秒)："))
v=s*3600/t
print("平均车速",round(v,1))
if v<=100:
print("正常")
elif v<120:
print("超过规定时速且不足20%")
elif v<150:
print("超过规定时速20%以上且不足50%")
elif v<170:
print("超过规定时速50%以上且不足70%")
else:
print("超过规定时速70%以上")
	请输入用时(秒)：720
	平均车速 125.0
	超过规定时速20%以上且不足50%

3.2.4 循环结构的程序实现

循环结构的算法可以通过for语句和while语句来实现。

1. for语句

在Python中， for语句的格式为：

```
for  <变量> in <序列>:
     <循环体>
[else:
     <语句块>]
```

for语句通过遍历序列中的元素实现循环，序列中的元素会被依次赋值给变量，然后执行一次循环体。当序列中的元素全部遍历完时，程序会自动退出循环，继续执行else子句中的语句块（该else子句可选）。若循环过程中执行了循环体中的break语句，则程序会中途退出for语句，转而去执行for语句后面的语句（即使有else子句，该子句也不会被执行）。如要依次显示某名学生的兴趣爱好（篮球、羽毛球、看书、旅游、音乐），可通过下面的语句来实现：

```
hobby=["篮球","羽毛球","看书","旅游","音乐"]
for x in hobby:
    print(x)
```

若序列中的元素为有序整数，则可利用内建函数range来实现。如下列循环语句：

```
for num in range(0,10,1):
    print(num,end=' ')
```

该语句执行后，输出的结果是：

0 1 2 3 4 5 6 7 8 9

range 函数由三个参数（起始值、终值、步长值）来决定序列中元素的个数和范围。如上例中的range(0,10,1)，生成0～9这10个整数序列。若起始值缺省，则默认值为0。步长值是序列中的每个元素之间的差，若缺省，则默认值为1。例如，range(0,10,1)也可写成range(10)。

拓展链接

end=' '的作用

代码“print(num,end=' ')”中的“end=' '”表示将print()函数的结束值设置为一个空格。这样，下一次对print()的调用结果将会直接从空格的右边开始。而print()函数默认以换行符作为其结束值。

热量消耗

人体运动时，热量的消耗取决于多方面因素。进行同样的运动，体重越重所消耗的热量就越高。运动项目、运动强度、运动量等因素的不同也会导致所消耗的热量有较大的差异。

请查阅相关资料，估算某一天你的主要运动所消耗的热量，并编程计算总量。

（1）抽象与建模

要计算某一天中主要运动所消耗的总热量，可根据个人体重、运动项目、运动强度和持续的时间等因素，列表格并估算各运动项目所消耗的热量，然后累加各运动项目消耗热量的和。表3.2.8所示的是某学生一天中主要运动消耗热量的情况。

表3.2.8　某学生一天中主要运动消耗热量表

运动项目	慢走	骑自行车	打羽毛球	爬楼梯	跳绳	慢跑
消耗热量（单位：大卡）	95	100	122	180	245	221

要计算总热量，可将各项运动消耗的热量进行累加。计算模型如下：

$s=\sum_{i=0}^{n} a_i$（其中s为总热量，a_i为各项运动消耗的热量，n=5）

（2）设计算法

要将每项运动消耗的热量进行累加，可采用循环结构来实现。具体算法的流程图如图3.2.10所示。

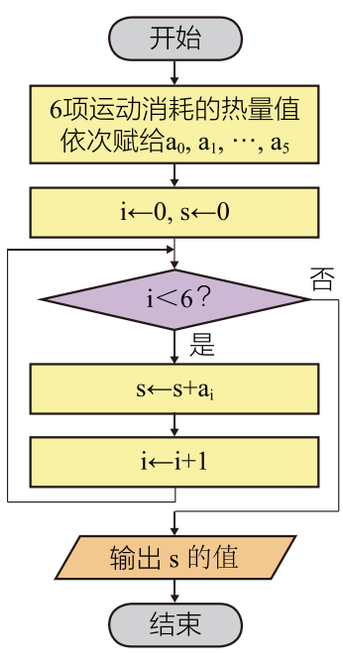


图3.2.10　计算总消耗热量流程图

（3）编写程序

用列表a存储每项运动所消耗的热量数据，变量s存储总热量值，利用for循环结构来实现。示例的程序及测试结果如下：

程序	测试结果
a=[95,100,122,180,245,221] s=0 for j in a: 　　s=s+j print(" 总消耗热量为：",s)	总消耗热量为：963

2. while语句

在许多情况下，当一个循环执行之前，可能并不知道它需要执行的次数。这时，就可以使用while循环。其常见格式如下：

```
while <条件>:
    <循环体>
```

while循环在执行时，首先会判断条件是否为真，如果条件为真，执行一次循环体，再次判断条件是否为真，如果仍为真，那么再执行一次循环体，以此类推，直到条件为假时退出while语句。

猜数游戏

编程实现一个“猜数游戏”。给定一个数让用户猜，用户输入猜测的数字，计算机给出相应提示：“偏大”“偏小”或“正确”。若所猜数字正确，则游戏结束；否则继续猜数。

（1）抽象与建模

游戏中首先要确定一个数number，然后将用户猜测的数guess与number比较，直到相等为止。

通过上述的问题抽象，建立如下模型：

$$\text{猜数结果}=\begin{cases}\text{“正确”（guess = number），游戏结束。}\\ \text{“偏小”（guess < number），继续猜数。}\\ \text{“偏大”（guess > number），继续猜数。}\end{cases}$$

（2）设计算法

要将用户输入的数与给定数（如number值为23）进行反复比较，用布尔型变量running表示猜数是否正确，其值为True表示猜数正确，游戏结束；反之，继续猜数。具体算法的流程图如图3.2.11所示。

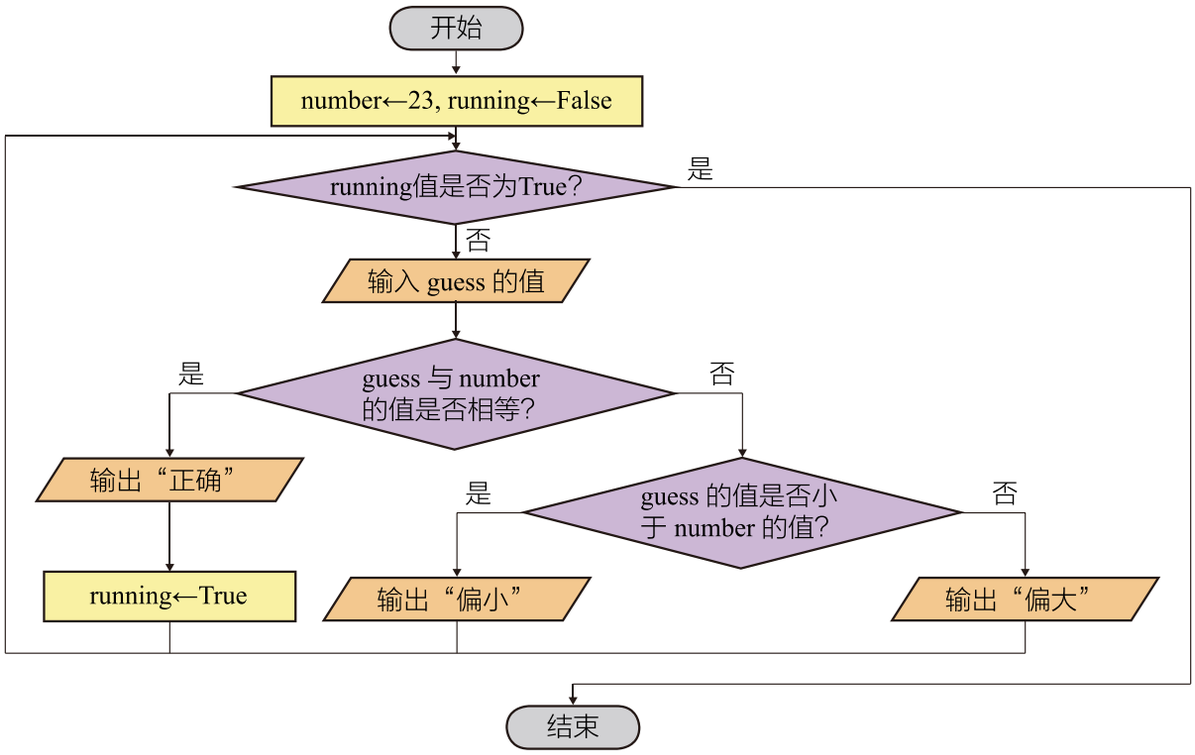


图3.2.11　猜数游戏流程图

（3）编写程序

根据算法编写的程序及测试结果如下：

程序	测试结果

number=23
running= False
while not running:
　　guess=int(input("请输入猜测的数："))
　　if guess==number:
　　　　print("正确")
　　　　running=True
　　elif guess<number:
　　　　print("偏小")
　　else:
　　　　print("偏大")
	请输入猜测的数：300
	偏大
	请输入猜测的数：20
	偏小
	请输入猜测的数：100
	偏大
	请输入猜测的数：25
	偏大
	请输入猜测的数：23
	正确

while循环在条件为假时结束，for循环在遍历完序列后结束。当循环条件为真或序列没有遍历完的时候，可以用break语句实现中途退出循环。在循环结构中，允许在一个循环体里面嵌入另一个循环。

问题与讨论

在“猜数游戏”中，若不引入布尔型变量（本例中的running），程序该如何实现？

3.2.5 函数与模块

在用算法解决问题的过程中，经常采用模块化程序设计思想，将问题分解成若干个子问题，并用相对独立的程序段来针对性地解决各个子问题，提高程序设计的效率。对于一些常用的程序代码，以模块化的形式进行保存，需要时可重复调用。

在Python中，主要利用函数、模块等方式实现模块化程序设计。

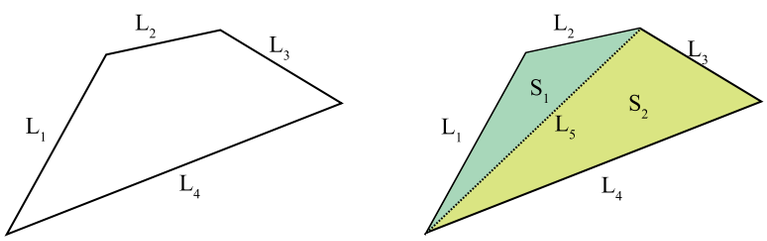
1. 函数的构造及应用

Python中的内建函数能实现许多功能，但在实际程序设计中，并不是所有的功能都有内建函数来直接提供支持，有时候需要根据实际情况自己构造函数以实现常用代码的模块化。定义函数的语法如下：

```
def 函数名(参数集合):
    <函数体>
    [return 函数值]
```

函数名的命名规则和变量名一样。完成函数的构造后，在程序中就可以根据需要调用该函数。

例如，某地块示意图如图3.2.12甲所示，地块边长分别为L_1、L_2、L_3、L_4。要想计算其面积，可通过如下算法来完成：先将此地块划分成如图3.2.12乙所示的两个三角形，只要再丈量出L_5的长度，就可以利用海伦公式分别计算出这两个三角形的面积S_1和S_2，从而得到此地块的总面积S（$S=S_1+S_2$）。



甲　划分前　　乙　划分后

图3.2.12　某地块示意图

在上述算法的程序实现过程中，将两次使用海伦公式求三角形的面积，这可以通过构造一个函数来解决代码重复的问题。

下面的Python程序段构造了一个利用海伦公式求三角形面积的函数Area：

```
def Area(a,b,c):
    p=(a+b+c)/2
    s=(p*(p-a)*(p-b)*(p-c))**0.5
    return s
```

主程序在调用Area函数时，需要将三角形三条边的边长值分别传递给Area函数的参数a、b、c。例如，主程序中有函数调用表达式Area(3,4,5)，则函数表达式传递给边长a、b、c的值分别为3、4、5。计算如图3.2.12所示不规则地块面积的表达式为：S=Area(L_1,L_2,L_5)+Area(L_3,L_4,L_5)。

2. 模块的导入及应用

在编写程序的时候，经常需要引用其他模块，这些模块包括Python内置的模块和来自第三方的模块。Python模块补充了许多功能强大的函数，在使用import语句或from-import语句将函数所在的模块导入后，就能使用其中的函数。比如使用math模块中的sqrt函数可以实现数学的开方运算，在Python Shell中导入math模块，并调用该模块中的sqrt函数，可采用以下两种方法：

方法一	方法二
>>> import math >>>math.sqrt(9) 3.0	>>> from math import sqrt >>>sqrt(9) 3.0

下面简要介绍几种Python内置模块的使用：

（1）math模块

math模块中的常用常数和部分函数如表3.2.9所示。

表3.2.9　math模块中的常用常数与函数

名称	含义
math.e	自然常数e
math.pi	圆周率π
math.ceil(x)	对x向上取整，比如x=1.2，返回2
math.floor(x)	对x向下取整，比如x=1.2，返回1
math.pow(x,y)	指数运算，得到x的y次方
math.log(x)	对数运算，默认基底为e
math.sin(x)	正弦函数
math.cos(x)	余弦函数
math.tan(x)	正切函数
math.degrees(x)	弧度转换成角度
math.radians(x)	角度转换成弧度

在编程求圆面积时，公式$s=\pi r^2$中的π和r^2可以分别调用math模块中的圆周率常数math.pi和函数pow(r,2)来完成。其完整的程序如下：

程序	测试结果
import math r=float(input("请输入圆的半径r：")) pi=math.pi s=pi*pow(r,2) print("圆面积是：",str(s))	请输入圆的半径r：5 圆面积是： 78.53981633974483

（2）random模块

random模块用来生成随机数，其常用函数如表3.2.10所示。

表3.2.10　random模块中的常用函数

名称	含义
random.random()	随机生成一个[0,1)范围内的实数
random.uniform(a,b)	随机生成一个[a,b]范围内的实数
random.randint(a,b)	随机生成一个[a,b]范围内的整数
random.choice(seq)	从序列的元素中随机挑选一个元素 比如random.choice(range(10))，从0到9中随机挑选一个整数
random.sample(seq,k)	从序列中随机挑选k个元素
random.shuffle(seq)	将序列的所有元素随机排序

例如有高一年级的（2）班、（3）班、（5）班、（8）班、（9）班共5个班的学生参加大合唱比赛，为了公平起见，需要随机安排他们的出场顺序。使用random模块，可以完成此任务，相应的Python程序如下：

```
import random
cla = ["（2）班", "（3）班", "（5）班", "（8）班", "（9）班"]
random.shuffle(cla)
for x in cla:
    print(x)
```

（3）Image模块

Image模块是PIL库（Python Imaging Library）中的重要模块，引用它可以完成对图像的一些常用操作，比如获取图像尺寸和像素颜色、旋转图像或改变图像格式等。下面的程序利用了Image模块完成对图像相关信息的获取和操作：

```
from PIL import Image
im = Image.open("school.jpg")        #打开school.jpg图像文件
print(im.format)                     #获取图像文件格式
print(im.size)                       #获取图像尺寸大小（以像素为单位表示图像的宽度和高度）
print(im.mode)                       #获取图像的颜色模式
im.rotate(45).show()                 #将图像旋转45°后显示
```

除了上述提到的模块，Python还包括了大量的其他模块，它们的功能涉及系统管理、科学计算、图形处理等各个领域。比如，用于实现部分操作系统功能（可用于文件、目录等操作）的os模块，与时间处理有关的time模块，可以实现科学计算、数据可视化的numpy和matplotlib，用于多媒体开发和游戏软件开发的pygame模块，支持图形处理的tkinter等。

实践与体验

编程实现图像的简单处理

用Python程序设计语言可编写具有图像处理功能的应用程序，如图像的大小和颜色调整、图像的合成、图像的滤镜添加等。下面，我们通过调整图像的颜色来体验Python语言在图像处理上的功能。

实践内容：

位图图像由像素组成。现提供一幅RGB模式的彩色位图，通过将其中每个像素的颜色值进行调整，使之成为一幅黑白图像。即设定某一特定值（如128），当像素值大于特定值时，该像素的RGB值变为1，否则变为0。适当调整特定值，观察黑白图像的效果。

实践步骤：

1. 导入模块。

为了调整图像的颜色，需要引用Image、numpy、matplotlib三个模块，其在此例中的功能分别如下：

Image：对图像的基本操作，如打开图像文件等。

numpy：将图像每个像素的RGB值以矩阵形式存储。

matplotlib：根据调整后的像素生成新的图像。

```
from PIL import Image
import numpy as np
import matplotlib.pyplot as plt
```

2. 打开图像并转换成数字矩阵。

```
img=np.array(Image.open('lena.jpg').convert('L'))
```

3. 调整每个像素的值。

```
rows,cols=img.shape                  #图像尺寸分别赋值
for i in range(rows):                #依次取每个像素的坐标
    for j in range(cols):
        if (img[i,j]>128):           #像素值大于128，赋值1，否则为0
            img[i,j]=1
        else:
            img[i,j]=0
```

4. 生成新的图像并显示。

```
plt.figure("lena")                   #指定当前绘图对象
plt.imshow(img,cmap='gray')          #显示灰度图像
plt.axis('off')                      #关闭图像坐标
plt.show()                           #弹出包含了图片的窗口
```

结果呈现：

原始图　　特定值=50　　特定值=128　　特定值=188

思考与练习

1. 写出下列Python表达式或程序语句的值。

（1）123%100

（2）len("Hello Leo!")

（3）abs（–12）

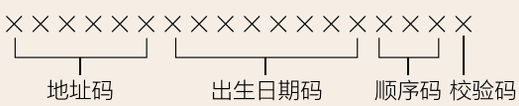
（4）data=[172,9,165,29,156,21]
　　max(data)

2. 编程求几个数中的最大值。

（1）分别输入两个数，输出它们中的最大值。

（2）分别输入三个数，输出它们中的最大值。

3. 十八位居民身份证号码由六位数字地址码、八位数字出生日期码、三位数字顺序码和一位校验码组成（其中倒数第二位是性别代码，男单女双），其格式如下：



请编程识别身份证号码中所包含的出生日期和性别信息，输出格式如下：

您的出生日期为：××××年××月××日，性别为×

4. 编程求1～100的所有偶数的和。试采用两种不同的方式实现，并考虑程序的效率。

5. 检测字符串是否全是数字。编程实现：对输入的一串字符串进行检测，输出“该字符串包含非数字字符”或“该字符串全是数字字符”的信息。

6. 运行以下程序，观察所呈现的图形，思考每行语句的功能。

```
import turtle
t = turtle.Pen( )
turtle.bgcolor("white")
colors = ["red","green","blue","yellow"]
for x in range(100):
    t.pencolor(colors[x%4])
    t.circle(x)
    t.left(91)
```

7. 修改“猜数游戏”实例中的程序，增加用户猜测次数（如允许用户最多猜5次）的限制功能。

8. 修改“猜数游戏”实例中的程序，使给定数由程序随机生成。

3.3 简单算法及其程序实现

解决问题的算法有很多，当算法设计完成后，就可以用程序设计语言来描述算法。本节将介绍如何使用Python程序设计语言来描述一些简单的算法。

3.3.1 解析算法及其程序实现

解析算法的基本思想是指根据问题的前提条件与所求结果之间的关系，找出求解问题的数学表达式，并通过表达式的计算来实现问题的求解。

在解析算法的程序实现过程中，首先要确保数学表达式的正确性，然后在程序中正确描述该数学表达式。如第二章中的“动动有奖”项目，需要在具体的奖励规则中抽象出“奖金”的数学计算模型，再编写程序来实现。

下面将通过一个答题卡填涂识别的项目实例来体验解析算法的程序实现过程。

答题卡填涂识别

如图3.3.1所示的答题卡常用于标准化考试、选举和调查。答题卡一般采用2B铅笔填涂，填涂好的答题卡经过扫描后得到相应的数字化图像，再通过光学识别，完成答题卡信息数据的采集、分析与统计。那么，计算机是如何判断答题卡中哪些信息点被填涂了呢？

图3.3.1 答题卡部分区域样例

分析：答题卡上的信息点填涂会导致该信息点的像素颜色发生变化（如填涂前为白色，填涂后为黑色）。因此，判断某信息点是否被填涂，可以从判断一个像素的颜色开始。具体步骤如下：

（1）抽象与建模

由于扫描得到的答题卡图像可能是彩色图像或者是灰度图像，为了提高填涂内容的识别准确率，需要先将图像统一转换为黑白图像。以彩色图像（RGB颜色模式）为例，可以先按照如下数学模型将彩色图像中每个像素的R、G、B值转换成灰度值：

灰度值=0.299×红色颜色分量+0.587×绿色颜色分量+0.114×蓝色颜色分量

在此基础上，再根据像素的灰度值，依据一定的颜色判断标准（如灰度值小于132，判定为黑色，否则判定为白色），将灰色近似判定为黑色或白色。

（2）设计算法

基于问题的抽象与建模，判定一个像素的颜色可以用解析算法，算法描述如下：

①给定颜色初值：输入某像素在RGB颜色模式下的各颜色分量。

②转换颜色模式：将彩色（RGB颜色模式）值转化成灰度值。

③判定黑、白颜色：若灰度值小于132，则判定为黑色；否则判定为白色。

（3）编写程序

用变量R、G、B分别存储某像素红色、绿色、蓝色的颜色分量，Gray_scale是灰度值，判定某像素（颜色值为RGB(43,10,241)）为黑色或白色的Python程序及测试结果如下：

程序	测试结果

R=43
G=10
B=241
Gray_scale=0.299*R+0.587*G+0.114*B
if Gray_scale<132:
　　print("黑色")
else:
　　print("白色")
	黑色

3.3.2 枚举算法及其程序实现

枚举算法的基本思想是把问题所有可能的解一一列举，然后判断每一个列举出的可能解是否为正确的解。在日常生活中，有很多问题可以使用枚举算法来解决，如求大面值纸币等额兑换成若干小面值纸币的方案、检测两篇文章的相似度等。

在枚举算法的程序实现中，逐一列举出每一个可能解，判断其是否为正确解的过程可采用循环结构来实现。而在利用问题提供的约束条件筛选、判断解的过程中则需要用到分支结构。

例如，在求解某整数x的所有因子（不包含x本身）的问题中，可以一一列举[1,x−1]范围内的所有整数，如果x能被这个范围内的某个整数整除，那么这个数就是整数x的因子。实现此算法的Python程序及测试结果如下：

程序	测试结果

x=int(input("请输入整数x："))
i=1
while i<=x−1:
　　if x%i==0:
　　　　print(i)
　　i=i+1
	请输入整数x：243
	1
	3
	9
	27
	81

在上述算法的程序实现过程中，由于整数x无法被从$\lfloor\frac{x}{2}\rfloor+1$到x−1范围内的整数整除，因此，枚举范围可缩小至从1到$\lfloor\frac{x}{2}\rfloor$，即循环进行的条件为$i\leqslant\lfloor\frac{x}{2}\rfloor$。在枚举的过程中，当数据规模较大时，减少枚举的次数，将大大提高解决问题的效率。因此，在设计枚举算法时，既不能遗漏任何一个正确解，又要尽可能地缩小解的列举范围，以提高算法的效率。

在前面的项目学习中，我们已经解决了答题卡信息点中某个像素是否为黑色的判断问题（即判断该像素是否被涂黑）。但要完整判定某信息点是否被填涂，还需要对该信息点区域中的所有像素进行判断。

（1）抽象与建模

判断某信息点是否被填涂与该信息点区域中的黑色像素数量有关，当黑色像素数量达到一定比例（如黑色像素的数量不少于该信息点区域内所有像素数量的64%），则认定该信息点被填涂。因此，判断某信息点是否被填涂，首先需要对该信息点区域中的所有像素逐一进行判断，然后统计所有黑色像素的总数。统计黑色像素个数的计算模型如下：

$$count=\sum_{i=1}^{n} S_i,\quad S_i=\begin{cases}0 & (Gray\_scale[i]\geqslant 132)\\ 1 & (Gray\_scale[i]<132)\end{cases}$$

其中，count为n个像素中的黑色像素总数，Gray_scale[i]为某个像素的灰度值。

（2）设计算法

基于问题的抽象与建模，统计某信息点中黑色像素的个数可以用枚举算法，算法描述如下：

①逐一列举某信息点中的各个像素。

②如果当前枚举的像素是黑色，那么黑色像素的数量加1。

③输出该信息点中黑色像素总数。

（3）编写程序

下面以5个像素为例，判定它们当中哪些是黑色像素并且统计黑色像素的个数，实现此功能的程序及测试结果如下：

程序	测试结果

count=0
Gray_scale=[47,178,146,185,116]
for i in range(0,len(Gray_scale)):
　　if Gray_scale [i]<132:
　　　　print(" 第 "+str(i+1)+" 个像素为黑色；")
　　　　count=count+1
print(" 黑色像素总共有："+str(count)+" 个。")
	第 1 个像素为黑色；
	第 5 个像素为黑色；
	黑色像素总共有：2 个。

程序中，变量count用于统计黑色像素的个数，列表Gray_scale存储了5个像素的灰度值。

如果上述5个像素的颜色用R、G、B的颜色分量表示，为了方便枚举，可以将它们的R、G、B颜色分量值用列表p_color进行组织：

p_color=[[84,24,70],[229,160,145],[133,161,107],[200,176,200],[201,80,85]]

再使用以下语句读取各个像素的R、G、B颜色分量值（变量i为列表p_color的索引）：

```
R=p_color[i][0]
G=p_color[i][1]
B=p_color[i][2]
```

从而计算出各个像素的灰度值并实现黑白像素判断。

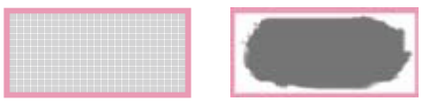


图3.3.2　填涂示意图

如果要判断如图3.3.2所示信息点中的300个像素（25×12）的颜色情况，那么需要对300个像素的颜色数据（RGB各颜色分量值）逐一进行列举并判断。下面的Python程序实现了从某文件中读取300个像素的颜色数据，并将判断后的填涂结果写入文件。

```
fname = input("请输入文件名称: ")
f = open(fname,"r+")                          # 以读写的方式打开文件
count=0

line = f.readline()                           # 从文件中读取一行
while line:                                   # 当line非空（从文件读取到了数据）
      line = line.split()                     # 把空白字符去除，变成包含三个str的list
      R, G, B = map(int, line)                # 把line中三个str转化成int并赋值给R, G, B
      if  0.299 * R + 0.587 * G + 0.144 * B < 132:
            count = count + 1
      line = f.readline()                     # 继续读取一行

if count >= 300 * 0.64:
     f.write("\n已填涂！")
else:
     f.write("\n未填涂！")
f.close()
```

拓展链接

Python的文件读写操作

读写文件是计算机中常见的输入输出操作，读写文件时会请求操作系统打开一个文件对象，然后通过操作系统提供的接口从这个文件对象中读取数据（读文件），或者把数据写入这个文件对象（写文件）。

Python内置了读写文件的函数。读文件时，可以使用内置的open()函数打开由参数指定的文件对象，并通过参数指定打开方式。如：

```
>>> f = open('test.txt', 'r')
```

上述命令的作用是以读文件模式（参数'r'）打开文件test.txt，如果文件打开成功，可以用read()方法将文件中的全部内容读取到内存。如果文件test.txt内容为“Hello, world!”，那么命令的执行结果为：

```
>>>f.read()
>>>'Hello, world!'
```

由于调用read()会一次性读取文件的全部内容，为避免读取文件过大，影响计算机运行性能，可采用多次调用read(size)方法，每次最多读取size个字节的内容。调用readline()可以每次读取一行内容，并按行返回list。另外，如果是配置文件，调用readlines()可以一次性读取整个文件内容，并按行返回到list。因此，可根据实际需要灵活调用这些方法。如：

```
for line in f.readlines():
    print(line.strip())                    #把末尾的'\n'删掉
```

文件使用完毕后必须关闭。关闭文件的方法如下：

```
>>>f.close()
```

调用open()函数写文件时，用参数'w'表示写文本文件模式；'r+'模式则表示在打开一个文本文件时同时允许读和写。例如，将“Hello,world!”写入test.txt，可使用下列命令：

```
>>>f = open('test.txt', 'w')
>>>f.write('Hello, world!')
>>>f.close()
```

将数据写入一个文件后，必须调用f.close()来关闭文件。因为写文件时，操作系统往往不会立刻把数据写入磁盘，而是放到内存缓存起来，空闲的时候再写入。只有在调用close()方法时，操作系统才会把内存中待写入的数据全部写入磁盘。

问题与讨论

请结合枚举算法的学习经历，谈谈枚举算法的一般程序结构特点。

3.3.3 算法程序实现的综合应用

在本节答题卡填涂识别的项目中，也可以通过编写函数来实现原有程序的模块化设计。下面的程序段创建了函数bw_judge，能够根据某彩色像素的R、G、B三种颜色分量值，通过计算进而“识别”该像素的颜色情况。

```
def bw_judge(R,G,B):
    Gray_scale=0.299*R+0.587*G+0.114*B
    if Gray_scale<132:
        color="黑色"
    else:
        color="白色"
    return color
```

调用bw_judge函数时，需将R、G、B的值传递至bw_judge函数的参数表。例如，主程序中有函数调用语句bw_judge(43,10,241)，则返回的函数值将是彩色像素RGB(43,10,241)被识别后的结果。如果在程序的其他位置也要检测黑、白颜色的像素，那么只需调用bw_judge函数即可。

在本项目中，为了方便读取图像中的像素颜色，并快速实现分析、判断，可以使用Python中的Image模块。实现此项目的Python程序如下：

```
from PIL import Image
im = Image.open("RGB.bmp")
pix = im.load()
width = im.size[0]                          # size中有两个参数，第1个参数为图像宽度值
height = im.size[1]                         # 第2个参数为图像高度值
count=0
for x in range(width):
    for y in range(height):
        R,G,B = pix[x, y]                   # 根据像素坐标获得该点的 RGB 值
        if  bw_ judge(R,G,B)=="黑色":       # bw_ judge函数用于判断黑、白像素
            count=count+1
if  count>=width*height*0.64:
    print("已填涂！")
else:
    print("未填涂！")
```

在本节答题卡填涂识别的项目中，已经实现了判断一个信息点填涂情况的程序编写。但在答题卡的填涂判断中，往往需要对一批信息点进行检测。如图3.3.3所示的准考证号填涂区中有9列、10行，共90个填涂信息点。这样，就需要对90个信息点进行一一检测，才能确定所填涂的准考证号。

图3.3.3　准考证号填涂区域

（1）抽象与建模

如图3.3.3所示的90个信息点，每个信息点在图像中都有固定的坐标值，根据坐标可以确定它们的位置。图3.3.4所示的是图3.3.3中准考证号填涂区左上角的4个信息点的坐标。

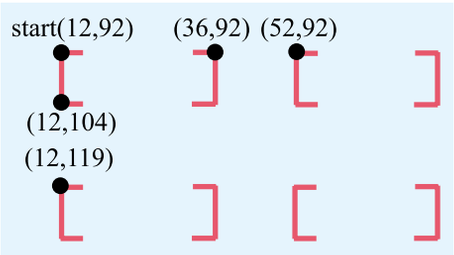


图3.3.4　部分填涂信息点的坐标示意图

答题卡在模板设计时就指定了每个信息点的格式与参数。如图3.3.5所示，信息点中的每个像素坐标可表示为：（水平方向值，垂直方向值）。其中，“start”为准考证号填涂区第一行第一列的信息点起始位置，其坐标记作（x_start, y_start）。另外，信息点的宽度（fill_width）、高度（fill_height）以及信息点之间的间隔距离（space_width、space_height）都相同，且同一列的信息点水平方向坐标值和同一行的信息点垂直方向坐标值也都相同。

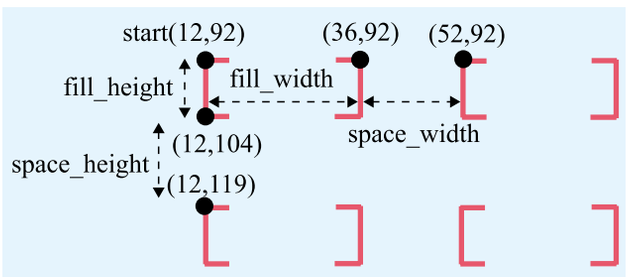


图3.3.5　信息点坐标分析示意图

为了更方便地取得相邻信息点（水平向右或者垂直向下方向）的起始位置坐标，可以将一个标准信息点的宽度、高度与信息点之间的间隔距离组合成一个整体。该整体的参数为：

水平方向总宽度：total_width = fill_width + space_width

垂直方向总高度：total_height = fill_height + space_height

这时，答题卡中的任意一个信息点的起始位置水平方向坐标值x可表示为：

x = x_start + total_width * col

起始位置的垂直方向坐标值y可表示为：

y = y_start + total_height * row

其中col为列号，row为行号；col∈[0,8]范围内的整数，row ∈ [0,9]范围内的整数。

根据每个信息点的起始位置坐标，就可以找到每个信息点的填涂区域，从而对其进行填涂识别。

（2）设计算法

基于问题的抽象与建模，识别准考证号填涂的算法描述如下：

①初始化数据：读取填涂区图像、输入信息点的相关参数。参数包括：起始位置“start”坐标（x_start, y_start）、信息点宽度（fill_width）和高度（fill_height）、信息点之

间间隔的宽度（space_width）以及信息点之间间隔的高度（space_height）。计算一个标准信息点的宽度、高度与对应方向（向右或向下）填涂区域之间间隔距离的和，存储到total_width、total_height中。

②确定信息点位置：按从左到右（按列）、从上到下（按行）的顺序对每个信息点进行检测，其中列（col）为[0,8]范围内的整数，行（row）为[0,9]范围内的整数。每个信息点的起始位置坐标(x,y)的值为：x = x_start + total_width * col，y = y_start + total_height * row。

③检测信息点填涂：若该信息点被填涂，则将填涂处信息（即row的值）连接至准考证号（number）；若当前列不存在被填涂的信息点，则当前列的填涂信息用“#”标记，并将其连接至准考证号（number）。

④输出填涂的准考证号。

（3）编写程序

实现此算法的Python程序如下：

```
from PIL import Image
x_start = 12                                    # 起始点坐标
y_start = 92
fill_width   = 24                               # 信息点宽度
fill_height  = 12                               # 信息点高度
space_width  = 16                               # 间隔宽度
space_height = 15                               # 间隔高度
num_length = 9                                  # 准考证号长度

def  bw_ judge(R, G, B):                        # bw_ judge用于判断一个像素的填涂情况
     Gray_scale = 0.299 * R + 0.587 * G + 0.114 * B
     return Gray_scale < 132

def  fill_ judge(x, y):                         # fill_ judge用于判断信息点的填涂情况
     count = 0
     for i in range(x, x + fill_width + 1):
          for j in range(y, y + fill_height + 1):
               R, G, B = pixels[i, j]
               if  bw_ judge(R, G, B) == True:
                    count = count + 1
     if  count >= fill_width * fill_height * 0.64:
          return True

total_width  = fill_width + space_width
total_height = fill_height + space_height
image = Image.open("fill.bmp")
pixels = image.load()
number = " "
```

```
for col in range(num_length):                    # x从左至右，y从上至下对填涂区进行检测
    for row in range(10):
        x = x_start + total_width  * col
        y = y_start + total_height * row
        if  fill_ judge(x, y) == True:
            number = number + str(row)
            break
    else:                                        # 10个信息点检测完毕后未发现有填涂
        number = number + "#"
print(number)
```

问题与讨论

分析本节中的准考证号填涂识别算法及其程序实现，你认为在提高填涂识别的准确性及合理性等方面还可以做哪些完善？相应的程序又该如何实现？

实践与体验

图像字符画

字符画是一种由字母、标点、汉字或其他字符组成的图画。复杂的字符画通常利用占用不同数量像素的字符代替图像上不同明暗的点，用纯文字拼出该图像所对应的黑白图，可以由程序制作而成。下面，我们将利用Python程序来体验字符画的制作。

实践内容：

如何将图像中的像素转换成字符？

我们可以把字符看作是比较大块的像素，一个字符表现为一种颜色。每个字符因为其笔画的复杂度都有对应的“视觉亮度”。字符的种类越多，可以表现的颜色也越多，图像也会更有层次感。下面是我们选择的字符集：['@', '#', '$', '%', '&', '?', '*', 'o', '/', '{', '[', '(', '|', '!', '^', '~', '-', '_', ':', ';', ',', '.', '`', ']。

实践步骤：

1. 准备图像素材“boy.jpg”。

2. 在同一目录下新建如下 Python 程序文件，运行并观察测试结果。

```
from PIL import Image
serarr=['@','#','$','%','&','?','*','o','/','{','[','(','|','!','^','~','-','_',':',';',',','.','`',"]
count=len(serarr)

def  toText(image_file):
        asd =''                                                  # 储存字符串
        for h in range(0, image_file.size[1]):                   # 垂直方向
            for w in range(0, image_file.size[0]):               # 水平方向
                r,g,b =image_file.getpixel((w,h))
                gray =int(r* 0.299+g* 0.587+b* 0.114)
                asd=asd+serarr[int(gray/(255/(count-1)))]
            asd=asd+'\r\n'
        return asd

image_file = Image.open("boy.jpg")                               # 打开图片
image_file=image_file.resize((int(image_file.size[0]*0.9), int(image_file.
size[1]*0.5)))
                                                                 # 调整图片大小

tmp=open('boy.txt','a')
tmp.write(toText(image_file))
tmp.close()
```

结果呈现：

思考与练习

请编制Python程序解决下列问题：

1. 某城市现有80万人口，如果每年人口增长率为1.2%，问：多少年后该城市人口数达到100万？

2. 某企业在第1年初购买了一台价值为120万元的设备，该设备的价值在使用过程中逐年减少。已知从第2年到第6年，每年初的价值比上年初减少10万元；从第7年开始，每年初的价值为上年初的75%。问：第n年初该设备的价值是多少？

3. 推算某一天是星期几，可使用蔡勒公式来计算。

蔡勒公式：w=y+[y/4]+[c/4]−2c+[26(m+1)/10]+d−1。相关参数说明如下：

w：星期(w对7取模得：0—星期日，1—星期一，2—星期二，3—星期三，4—星期四，5—星期五，6—星期六)；c：世纪（年份前两位数）；y：年（年份后两位数）；m：月（在公式中，某年的1、2月要看作上一年的13、14月来计算，比如2003年1月1日要看作2002年的13月1日来计算）；d：日；[ ]代表取整，即只取整数部分。

输入年、月、日，求出这一天是星期几。

4. 某加密算法要求对输入的小写英文字符串（明文）做如下处理：按照英文字母“a”“b”……“z”的排列顺序，将字符串中的每个字符都取其前一个字符后重组得到密文(其中规定字符“a”的前一个字符取“z”)。例如，明文“student”加密后的密文是“rstcdms”。请编写计算机程序实现该加密算法。

5. 民间流传着“韩信点兵”的故事。韩信带1500名士兵打仗，战死四五百人，剩下的士兵排队：站3人一排，多出2人；站5人一排，多出4人；站7人一排，多出6人。韩信马上说出人数：1049。请编写计算机程序验证结果。

6. 我国古代数学家张丘建在《算经》中提出了如下的数学问题：鸡翁一，值钱五；鸡母一，值钱三；鸡雏三，值钱一；百钱买百鸡，问翁、母、雏各几何？请编写计算机程序解决该问题。

巩固与提高

1. 篮球赛是高得分的球类比赛，比分领先的一方可能很快又会被反超。某体育专家发明了一种算法，用于预测篮球比赛中怎样的比分领先优势是不可超越的（也称为“安全的”），从而在比赛结束前就能大致预测哪支球队会获胜。

此算法用自然语言描述如下：

①取领先球队的分数。

②此分数减去3分。

③如果目前是领先球队控球，则加上0.5分，否则减去0.5分。若此数字小于0则分数变成0。

④计算此分数平方后的结果。

⑤如果得到的结果比当前比赛剩下的时间秒数更大，那么这个领先是“安全的”。

试编制程序实现此算法。

2. 反弹高度。某小球从100米高度自由落下，每次落地后反弹回原高度的一半，再落下。编程求出小球在第10次落地时，共经过多少米。第10次的反弹高度是多少？

3. 角谷猜想。以一个正整数n为例，当n是奇数时，下一步变为3n+1；当n是偶数时，下一步变为n/2。不断重复这样的运算，经过有限步后，一定可以得到1。请编程验证角谷猜想，随机生成一个正整数并输出验证的完整过程。

4. 文本文件“score.txt”中保存着30个学生某次测试的成绩，编写一个计算机程序，从该文件中读取每个学生的分数，统计并输出各等级的学生人数。根据分数判断其所属等级的标准如表3.3.1所示。

表3.3.1　分数与等级的对应关系

分数段	90～100	80～89	70～79	60～69	60以下
等级	A	B	C	D	E

5. 鸡兔同笼。今有鸡兔同笼，上有三十五头，下有九十四足，问：鸡兔各几何？编程输出鸡、兔的数量，并考虑如何提高程序的效率。

项目挑战

“寻找关联次数最多的商品”问题之算法实现

在第二章的项目挑战中，我们已经完成了“为超市寻找关联次数最多的商品”问题的算法设计。学习了Python程序设计知识后，今天将用Python程序来实现算法，并进一步思考该算法的优化方法。

图3.3.6　超市商品陈列

项目任务

根据第二章“项目挑战”中设计的算法，采集超市某个时期内的流水账记录，将已有的算法进行程序实现，找出超市内关联次数最多的一对商品（“关联次数”的具体叙述详见第二章的“项目挑战”）。具体要求如下：

1. 采集商品交易数据，采用合适的文件格式进行存储。
2. 根据算法特点，预处理数据并能采用合理的数据结构进行组织。
3. 编写计算机程序，找出关联次数最多的商品。

过程与建议

在采集超市一个时期商品销售数据的基础上，首先对数据进行预处理，并用二维数组（Python中可用列表实现）组织预处理结果，其中的a[i,j]表示商品i和j的关联次数。然后用二重循环枚举所有不同商品之间的关联次数，运用求最大值的方法求得关联次数最多的那一对商品。具体步骤如下：

1. 采集数据

为了提高统计的可信度，需要采集尽可能大的数据样本。商品交易数据的来源比较广泛，可以通过网络或者实体超市等途径采集到，也可收集自己家里一段时间的超市购物小票。为了实现数据的结构化，需要将采集到的数据进行整理并保存到文本文件中。

2. 预处理数据

编写Python程序，从上面所述文本文件中读取数据并将数据进行初步统计，保存到二维数组a中。如下表所示，a[i,j]表示“商品i”和“商品j”的关联总次数。例如，本账

单中同时出现“商品2”和“商品5”，则将这两类不同商品的关联次数增加1，即a[2,5]←a[2,5]+1。

i \ j	商品1	商品2	商品3	商品4	商品5	商品6	……
商品1		a[1,2]	a[1,3]	a[1,4]	a[1,5]	a[1,6]	
商品2			a[2,3]	a[2,4]	a[2,5]	a[2,6]	
商品3				a[3,4]	a[3,5]	a[3,6]	
商品4					a[4,5]	a[4,6]	
商品5						a[5,6]	
商品6							
……							

3. 程序实现

（1）数据结构设计

用max保存最大关联次数，用ans1和ans2保存关联次数最多的两个商品的编号，用n表示商品总数，用二维数组a保存预处理结果（该步骤在预处理阶段进行）。

（2）算法框架

对二维数组a中的数据进行下列处理，即可找出关联次数最多的商品。

```
max←0
i←0
while i＜n:
    j←0
    while j＜n:
        如果 a[i,j]＞max    那么:
            max←a[i,j]
            ans1←i
            ans2←j
        j←j+1
    i←i+1
输出max、ans1、ans2的值
```

运行调试程序直至结果正确，进一步思考是否还有其他算法。

4. 展示交流

（1）以一定的形式（如网页、PPT、文字稿、视频等）在某个群体内展示。通过展示

交流，进一步思考项目成果的完善方向。如数据是否足够多、分析方法是否科学等，然后加以完善。

（2）将项目研究成果形成报告，并将其推荐给相关的超市。

评价标准

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

评价条目	说明	评分（1～10分）	评分主要依据阐述	后续完善方向
数据样本	采集到的数据来源可靠，数量较多			
数据整理	整理产生的数据结构化较好，有利于数据组织和问题求解			
程序设计	编写的程序正确，能输出合理的结果			
项目成果	项目研究报告观点明确，能较好地为实际工作提供决策支持，数据能较好地支撑观点			

拓展项目

1. 在某网上书店购买一批图书，当前该书店有一些优惠活动：满100减30，满200减70，满300减120。请制订购书方案，使购买图书的费用最少。试编制计算机程序，实现自动计算所购图书的费用。

2. π作为经常使用的数学常数，对它的近似计算已经持续了2500多年，至今依然在进行着，其间涌现出许多计算方法，它们各有千秋，如数值积分法、泰勒级数法、蒙特卡罗法、韦达公式法、拉马努金公式法、迭代法等。请你选择其中两种方法，用计算机编程求π的近似值，体验计算过程，并比较两种方法的精确度。

3. 聊天机器人是一个用来模拟人类对话或聊天的程序，程序设计者可以将自己感兴趣的回答事先放到文本文档中，当一个问题提交给聊天机器人时，它会通过算法，从数据库中找到最贴切的答案，给予回复。试编写一个聊天机器人的程序，使其具有数据筛选、学习等能力。

第四章 4

数据处理与应用

数据正逐渐成为现代社会基础设施的一部分，就像公路、铁路、电网和通信网络一样不可或缺。传感器、卫星导航系统、社交网络等时刻产生新的数据，通过数据处理平台，可以对数据进行收集、加工、储存、分析，并应用到社会的各个领域，为人们的判断、预测、决策提供有力的依据。

问题与挑战

- 进入高中，同学们都面临着学科选择，即从思想政治、历史、地理、物理、化学、生物等科目中，选择三门作为高考选考科目。选课前，同学们需要了解哪些大学开设了自己喜欢的专业、这些专业对选考科目提出了哪些要求等信息，以便更好地进行学业规划和未来职业生涯规划。那么，该从哪儿着手，又该如何做呢？

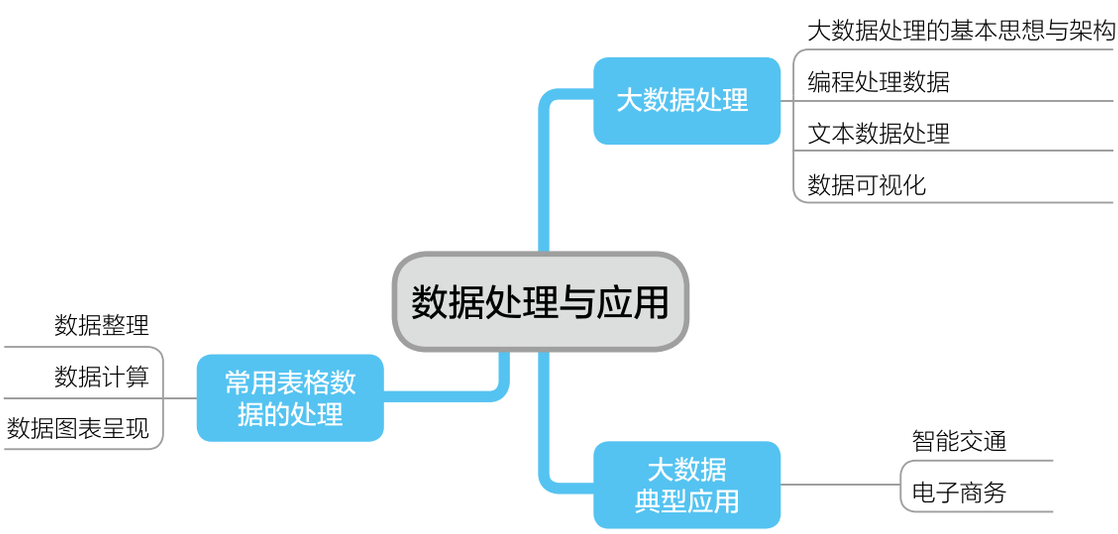
- 随着电子商务的发展，网络购物已成为人们生活中的一部分。很多人都有这样的购物体验：登录常去的购物平台时，平台会在页面的明显位置向用户推送“猜你喜欢”之类的商品信息；将某件商品加入购物车后，平台会向用户推送类似或相关的商品信息……这些信息是怎么来的呢？

- 智能导航系统，可以利用卫星定位系统提供的位置、速度及时间等交通数据，配合高精度电子地图，为用户准确、实时地在电子地图上规划出行路线。在用户输入出发地、目的地后，导航系统会快速计算出可能的出行路线，同时推荐符合用户偏好的路线；还会在行进途中用语音提示方向、实时显示用户位置及周围设施，为人们的出行提供了极大的便利。那么，导航系统是如何运用交通大数据为用户提供这些服务的呢？

学习目标

1. 能根据实际需求，对表格数据进行处理。
2. 了解大数据处理的基本思想与架构。
3. 能使用Python进行简单数据处理，解决实际问题。
4. 能根据实际需求，选用合适的工具和方法对数据进行可视化。
5. 能从实际生活中发现数据应用的价值，认识到有效的数据处理对于提高数据价值的重要意义。

内容总览



4.1 常用表格数据的处理

在生产生活中，人们经常遇到以二维表方式组织存储的数据，如成绩数据、商品销售数据、家庭收支数据等。这些基于表格的数据常常需要进行计算、排序、筛选、图表呈现等处理。

数据处理的核心是数据，数据的质量直接影响数据分析的结果。但获取到的数据并不都是优质的，常常存在缺失、重复、错误、数量级不同等问题。因此，在数据分析和数据挖掘前，通常先对数据进行整理。

4.1.1 数据整理

数据整理的目的是检测和修正错漏的数据、整合数据资源、规整数据格式、提高数据质量。常见的数据问题有数据缺失、数据重复、数据异常，还有逻辑错误、格式不一致等。

数据缺失问题是数据集中普遍存在的问题，最简单的处理方法是忽略含有缺失值的实例或属性。但这样处理可能造成数据集不完整，致使后续的统计分析结果出现偏差。因此较好的方法是根据数据间的关联性估计较准确的缺失值，并通过合适的方法对缺失值进行填充。通常采用平均值、中间值或概率统计值来填充缺失值。

数据重复问题在多数据源进行合并集成时经常出现。重复数据会导致数据冗余，浪费存储空间和网络带宽，在数据分析中还可能会误导用户。重复数据的检测可以分为基于字段和基于记录两个方面，需要根据实际情况采用不同的算法进行检测。对于重复数据，可以在进一步审核的基础上进行合并或删除等处理。

异常数据指数据集中不符合一般规律的数据对象，它可能是要去掉的噪声，也可能是含有重要信息的数据对象。

逻辑错误问题指数据集中的属性值与实际值不符，或违背业务规则或逻辑。如某人的生日为“2000/13/25”，月份数据超出了月份的最大值，通过检测字段中各属性有效数据值的范围可以判断该值错误。

不同来源的数据可能存在格式不一致的情况，这就需要进行数据转换，以便形成一个适合后续分析和挖掘的描述形式。数据转换通常包括属性数据类型的转换、根据已有属性集构造新属性的转换、将不同来源的相同属性的定义及其值进行统一标准化表达的转换等。

问题与讨论

两个不同来源的数据集A、B如图4.1.1、4.1.2所示。若要合并这两个数据集以对比分析两个球员的技术情况，将遇到哪些问题？该如何处理？

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	詹姆斯-哈登																
2	年度	球队	场数	先发	分钟	投篮%	三分%	罚球%	进攻	防守	篮板	助攻	抢断	盖帽	失误	犯规	得分
3	2016	火箭	30	30	36.7	43.7	34.9	83.9	1.3	6.7	8	11.7	1.4	0.3	5.3	2.5	27.8
4	2015	火箭	82	82	38.1	43.9	35.9	86	0.8	5.3	6.1	7.5	1.7	0.6	4.5	2.8	29
5	2014	火箭	81	81	36.8	44	37.5	86.8	0.9	4.7	5.6	7	1.9	0.7	3.9	2.6	27.4
6	2013	火箭	73	73	38	45.6	36.6	86.6	0.8	3.9	4.7	6.1	1.5	0.4	3.6	2.4	25.4
7	2012	火箭	78	78	38.3	43.8	36.8	85.1	0.8	4.1	4.9	5.8	1.8	0.4	3.7	2.3	25.9
8	2011	雷霆	62	23	1.44	9.1	39	84.6	0.5	3.6	4.1	3.7	1	0.2	2.2	2.4	16.8
9	2010	雷霆	82	52	6.74	3.6	34.9	84.3	0.5	2.6	3.1	2.1	1.1	0.2	1.2	2.5	12.2
10	2009	雷霆	76	2	2.94	0.3	37.5	80	80.6	2.6	3.2	1.8	1	0.2	1.3	2.6	9.9

图4.1.1　数据集A

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	斯蒂芬-库里																
2	赛季	球队	出场	首发	时间	投篮	三分	罚球	前篮板	后篮板	总篮板	助攻	抢断	盖帽	失误	犯规	得分
3	16-17	勇士	30	30	33	46.8%	39.6%	93.4%	0.7	3.6	4.2	5.9	1.8	0.2	2.7	2.5	24
4	15-16	勇士	79	79	34	50.4%	45.4%	90.8%	0.9	4.6	5.4	6.7	2.1	0.2	3.3	2	30
5	14-15	勇士	80	80	33	48.7%	44.3%	91.4%	0.7	3.6	4.3	7.7	2	0.2	3.1	2	24
6	13-14	勇士	78	78	37	47.1%	42.4%	88.5%	0.6	3.7	4.3	8.5	1.6	0.2	3.8	2.5	24
7	12-13	勇士	78	78	38	45.1%	45.3%	90.0%	0.8	3.3	4	6.9	1.6	0.2	3.1	2.5	23
8	11-12	勇士	26	23	28	49.0%	45.5%	80.9%	0.6	2.8	3.4	5.3	1.5	0.3	2.5	2.4	15
9	10-11	勇士	74	74	34	48.0%	44.2%	93.4%	0.7	3.2	3.9	5.8	1.5	0.3	3.1	3.1	19
10	09-10	勇士	80	77	36	46.2%	43.7%	88.5%	0.6	3.9	4.5	5.9	1.9	0.2	3	3.2	18

图4.1.2　数据集B

4.1.2　数据计算

数据计算是数据处理的常用方法之一。日常简单的数据处理可以使用Excel软件完成，专业的数据处理和统计分析工具有SPSS、SAS、MATLAB等，也可以通过R、Python、Java等计算机语言编程进行数据处理。

拓展链接

常用的数据处理和统计分析工具

Excel软件是微软公司推出的Microsoft Office系列套装软件中的组成部分，是一个简单易用的电子表格软件，可以进行数据的处理、统计分析和辅助决策操作，广泛应用于文秘办公、财务管理、市场营销、行政管理和协同办公等事务。

SPSS是IBM公司推出的一款统计分析软件，具备数据收集、准备、分析、描述、解释和

展现的功能。SPSS 提供丰富的统计算法，并且操作简便、功能强大、扩展性强，但需要使用人员具备一定的数理统计学知识背景，比较适合专业分析、研究等人员使用。

SAS 是 SAS 软件研究所开发的一套大型集成应用软件系统，共有三十多个功能模块，具有数据访问、数据管理、数据分析、数据呈现等功能。SAS 系统从大型机上的系统发展而来，其操作以编程为主。系统地学习和掌握 SAS，需要花费一定的精力，比较适合统计专业人员使用。

MATLAB 是 MathWorks 公司推出的一种科学计算语言和编程环境，主要应用于数据分析、无线通信、深度学习、计算机视觉、量化金融与风险管理等领域。MATLAB 将适合迭代分析和设计过程的桌面环境与直接表达矩阵和数组运算的编程语言相结合，为分析数据、开发算法和创建模型等提供了便于探索和发现的环境，深受工程师和科学家的青睐。

在 Excel 软件中，可以应用公式进行数据的计算。公式是以“=”开头，由常数、函数、单元格引用和运算符组成的式子。公式不仅用于计算，更重要的是构建计算模型。

单元格引用是指对工作表中的单元格或单元格区域的引用。默认情况下，单元格引用是相对的，如 A1；单元格绝对引用，如 A1；连续的单元格区域引用，如 A2:D5；不连续的单元格区域引用，如 A2:A5，D2:D5。

算术运算符有 ^、%、*、/、+、−，用于进行基本的数学运算。比较运算符有 =、>、<、>=、<=、<>，用于比较两个值，结果为逻辑值 TRUE 或 FALSE。文本连接运算符“&”，可以连接一个或多个文本字符串，生成一段文本。

函数是预定义的公式，通过使用参数按特定顺序或结构进行计算。求和、平均值、最大值、最小值的函数语法如下：

函数语法

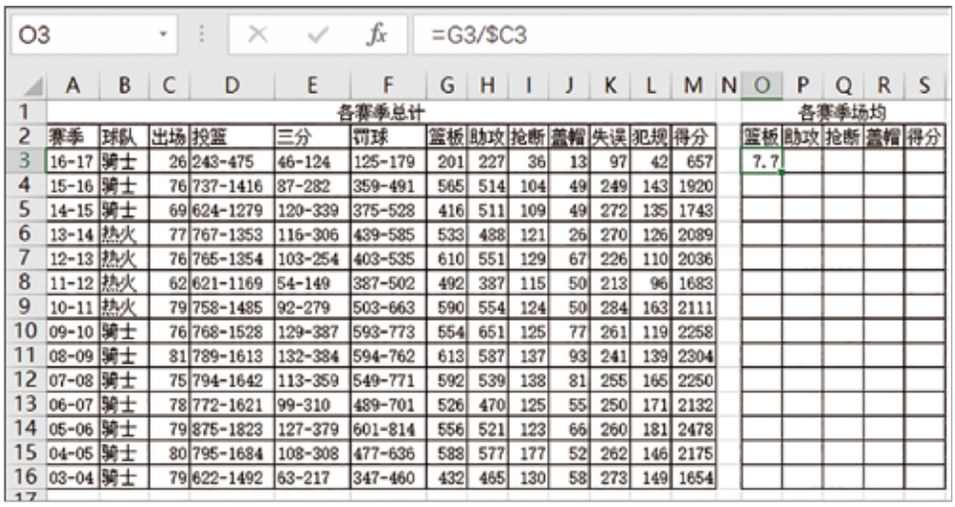
SUM（number1，[number2]，...）	求参数的和
AVERAGE（number1，[number2]，...）	求参数的平均值
MIN（number1，[number2]，...）	返回参数列表中的最小值
MAX（number1，[number2]，...）	返回参数列表中的最大值

↓

参数可以是数字、单元格或单元格区域

使用Excel软件进行数据计算

某球员的各赛季数据如图4.1.3所示，使用Excel软件统计其各赛季场均情况。



O3 =G3/$C3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	各赛季总计														各赛季场均				
2	赛季	球队	出场	投篮	三分	罚球	篮板	助攻	抢断	盖帽	失误	犯规	得分		篮板	助攻	抢断	盖帽	得分
3	16-17	骑士	26	243-475	46-124	125-179	201	227	36	13	97	42	657		7.7				
4	15-16	骑士	76	737-1416	87-282	359-491	565	514	104	49	249	143	1920						
5	14-15	骑士	69	624-1279	120-339	375-528	416	511	109	49	272	135	1743						
6	13-14	热火	77	767-1353	116-306	439-585	533	488	121	26	270	126	2089						
7	12-13	热火	76	765-1354	103-254	403-535	610	551	129	67	226	110	2036						
8	11-12	热火	62	621-1169	54-149	387-502	492	387	115	50	213	96	1683						
9	10-11	热火	79	758-1485	92-279	503-663	590	554	124	50	284	163	2111						
10	09-10	骑士	76	768-1528	129-387	593-773	554	651	125	77	261	119	2258						
11	08-09	骑士	81	789-1613	132-384	594-762	613	587	137	93	241	139	2304						
12	07-08	骑士	75	794-1642	113-359	549-771	592	539	138	81	255	165	2250						
13	06-07	骑士	78	772-1621	99-310	489-701	526	470	125	55	250	171	2132						
14	05-06	骑士	79	875-1823	127-379	601-814	556	521	123	66	260	181	2478						
15	04-05	骑士	80	795-1684	108-308	477-636	588	577	177	52	262	146	2175						
16	03-04	骑士	79	622-1492	63-217	347-460	432	465	130	58	273	149	1654						

图4.1.3　某球员的各赛季数据

- 分析数据

数据采用电子表格格式组织和存储，表中数据包含了赛季、球队、出场、投篮、三分等数据。各项场均与各项总计、出场次数的关系为：各项场均=各项总计/出场次数。

- 计算各赛季场均篮板、助攻、抢断、盖帽、得分

①在O3单元格中输入公式 =G3/$C3。

②拖曳“填充柄”自动填充公式到R16单元格，完成各赛季场均篮板、助攻、抢断、盖帽的计算。

③在S3单元格中输入公式=M3/C3，自动填充至S16单元格，完成各赛季场均得分的计算。

- 查看、分析计算结果

观察数据表中的数据，重点检查各赛季场均数据的计算是否正确、完整。通过分析各赛季比赛的场均数据，可以了解该球员在各赛季比赛中的技术发挥和表现情况。

问题与讨论

在公式填充过程中，公式中的相对引用和绝对引用有何区别？

4.1.3 数据图表呈现

图表是用视觉形式向人们展示数据的一种方法。常见的图表类型有柱形图、折线图、饼图、雷达图、散点图、气泡图等。在运用图表表现数据、传递信息时，通常依据数据间的关系选择相应的图表类型。

问题与讨论

雷达图、散点图、气泡图等分别适合展现何种数据关系？

使用Excel软件创建图表

2012～2016年部分省市区国内生产总值的数据（单位：亿元）如图4.1.4所示，使用Excel软件创建图表，分析和展现2012～2016年北京市、天津市、上海市三地国内生产总值的变化情况。

2012－2016 年部分省市区国内生产总值

地区	2012年	2013年	2014年	2015年	2016年
北京市	17879.4	19800.81	21330.83	23014.59	25669.13
天津市	12893.88	14442.01	15726.93	16538.19	17885.39
河北省	26575.01	28442.95	29421.15	29806.11	32070.45
山西省	12112.83	12665.25	12761.49	12766.49	13050.41
江苏省	54058.22	59753.37	65088.32	70116.38	77388.28
浙江省	34665.33	37756.59	40173.03	42886.49	47251.36
广东省	57067.92	62474.79	67809.85	72812.55	80854.91
上海市	20101.72	21010.15	23567.7	25123.45	20170.65

图4.1.4　2012～2016年部分省市区国内生产总值数据

- 分析数据

数据以电子表格的形式进行组织和存储，其中，2012～2016年北京市、天津市、上海市三地国内生产总值的数据包含时间趋势和大小比较的两层关系，因此图表类型可以选用折线图。

- 创建图表

①选择要在图表中展示的数据区域A2:F4，A10:F10。

②单击“插入”选项卡上的“插入折线图或面积图”按钮，选择“折线图”，生成的折线图如图4.1.5所示。

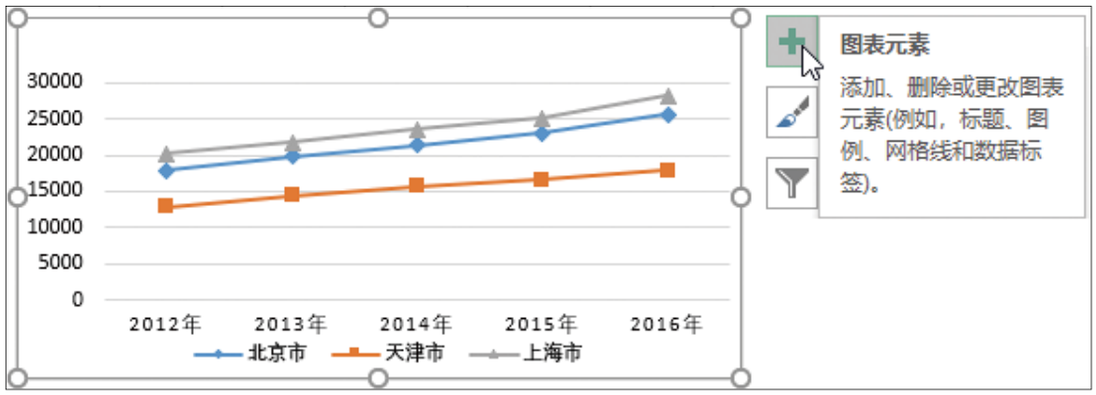


图4.1.5　2012 ～ 2016年北京市、天津市、上海市生产总值折线图

③使用图表右上角附近的“图表元素”按钮，可添加坐标轴标题和数据标签等图表元素，使用“图表样式”按钮可自定义图表的外观，使用“图表筛选器”按钮可更改图表中显示的数据。

另外，选中图表，使用“设计”和“格式”选项卡可自定义图表的外观，美化图表。

- 检查图表

查看图表中数据的展现是否完整，将鼠标移到折线相应的数据点上，查看数据卡显示的数据与表格中的数据是否一致。观察折线的走势，发现北京市、天津市、上海市三地国内生产总值在2012～2016年间一直呈上升趋势。

思考与练习

浏览各省市统计局、国家统计局、国家数据等网站的数据，收集你感兴趣的数据，使用Excel软件进行分析。分析建议如下：

问题	建议	过程记录
找到了哪些感兴趣的数据	浏览数据，寻找感兴趣、可分析的数据，确定分析目标	分析目标：
收集了哪些数据	使用合适的方法收集相关数据，使用Excel软件进行组织和存储	收集数据的方法： 文件名：
哪些数据是本次分析所必需的	整理数据，使之符合分析需要	最后保留的数据及格式：
如何分析这些数据	采用合适的方法分析数据，创建图表呈现数据	采用的分析方法：
发现了什么	记录分析的结果和形成的结论	结果： 结论：
有哪些心得体会	记录在数据收集与分析过程中的心得体会	体会：

4.2 大数据处理

大数据具有数据量大、数据来源与类型多样、处理速度快等特点，简单的表格处理软件已经无法满足大数据的处理需求，同时，大数据技术、理论和处理方法也在不断发展，为大数据的处理提供了越来越有力的支持。

4.2.1 大数据处理的基本思想与架构

处理大数据时，一般采用分治思想。分治，字面上的解释是“分而治之”，就是把一个复杂的问题分成两个或更多相同或相似的子问题，找到求这几个子问题的解法后，再找出合适的方法把它们组合成求整个问题的解法。如果这些子问题还难以解决，可以再把它们分成几个更小的子问题，以此类推，直至可以直接求出解为止。

处理大数据的分治思想

某公司搜集了过去一年发布的所有微博数据，需要统计其中出现频率最高的100个词。

统计文件filename中各单词出现的频率，用Python编程实现的部分代码如下：

```
wordcount ={}
for word in open(filename,'r').read():
    wordcount[word] += 1
```

在数据量较小的情况下，程序的处理速度是非常快的。如果数据量、单词词汇量非常大（数十亿），那么运行这个程序、处理数据的速度将变得非常慢。

- 假设有10台计算机，每台计算机可以处理1000M数据。每台计算机处理数据后，将计算结果汇总到一台主控计算机上，由主控计算机根据中间计算结果汇总统计出最终计算结果，并输出出现频率最高的单词，这样就可以处理10G的词汇数据。
- 假设有100台计算机，按理应该可以处理100G词汇数据。但又有新的问题，100台计算机同时向主机传输数据可能会遇到主控计算机网络传输带宽的瓶颈。这时，可对网络结构进行改造，每10台分为一组分别汇总，最后提交给主控计算机完成最后的统计。
- 如果是1000台、1万台或者10万台计算机，这种处理模式就行不通了。随着计算机数量的增加，发生机器故障、网络故障的风险不断增加。即使只有1台计算机出现了问题，整个的计算都将是不成功的。可用的办法是：将同一份数据分发给不同的计算机，假设发给了3台计算机，当其中1份数据发生计算故障时，剩下的2份备份数据的计算结果还能相互验证，保证最终结果的正确性。这就需要一台或多台计算机负责管理，并运行专

门的软件检测计算过程中的故障，在检测到故障时能重新安排计算任务。这种“分治”的思想就是处理大数据的基本思路。

拓展链接

分布式计算与并行处理

分布式计算（Distributed Computing）是把一个需要非常巨大的计算能力才能解决的问题分成许多小部分，然后把这些部分分配给许多计算机进行处理，最后把这些计算结果综合起来得到最终的结果。例如，利用分布在世界各地成千上万台闲置计算机的计算能力，分析来自外太空的电讯号，探索可能存在的外星智慧生命。

并行处理（Parallel Processing）是计算机系统中能同时执行两个或更多处理的一种计算方法。并行处理的主要目的是节省大型和复杂问题的处理时间。

目前，大数据处理按照类型可划分为对静态数据的批处理、对流数据的实时计算和对图结构数据的图计算，如图4.2.1所示。静态数据指在处理时已收集完成、在计算时不会发生改变的数据，一般采用批处理方式；流数据是指不间断地、持续地到达的实时数据，随着时间的流逝，流数据的价值也随之降低，通过实时分析计算可以得到更有价值的分析结果；现实世界中的许多数据，如社交网络、道路交通等数据，可采用图计算模式进行处理。

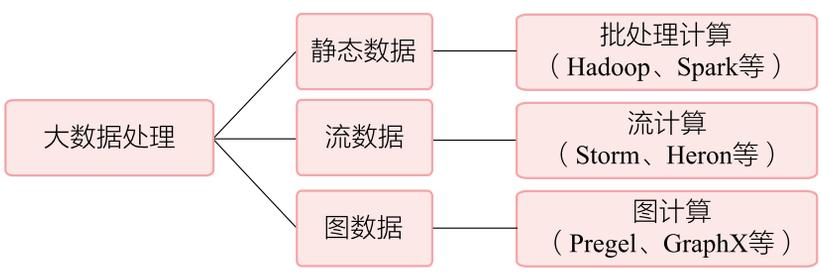


图4.2.1　大数据处理类型

1. 批处理计算

Hadoop是一个可运行于大规模计算机集群上的分布式系统基础架构，适用于静态数据的批处理计算。借助Hadoop，程序员可以在不了解分布式底层细节的情况下，轻松编写分布式并行程序，将其在计算机集群上运行，完成海量数据的存储与分析。Spark是一种与Hadoop相似的、应用较广的开源分布式计算架构。Spark启用了内存存储中间结果，运行速度比Hadoop快很多。

Hadoop计算平台主要包括Common公共库、分布式文件系统HDFS、分布式数据库HBase、分布式并行计算模型MapReduce等多个模块，其主要组成结构如图4.2.2所示。

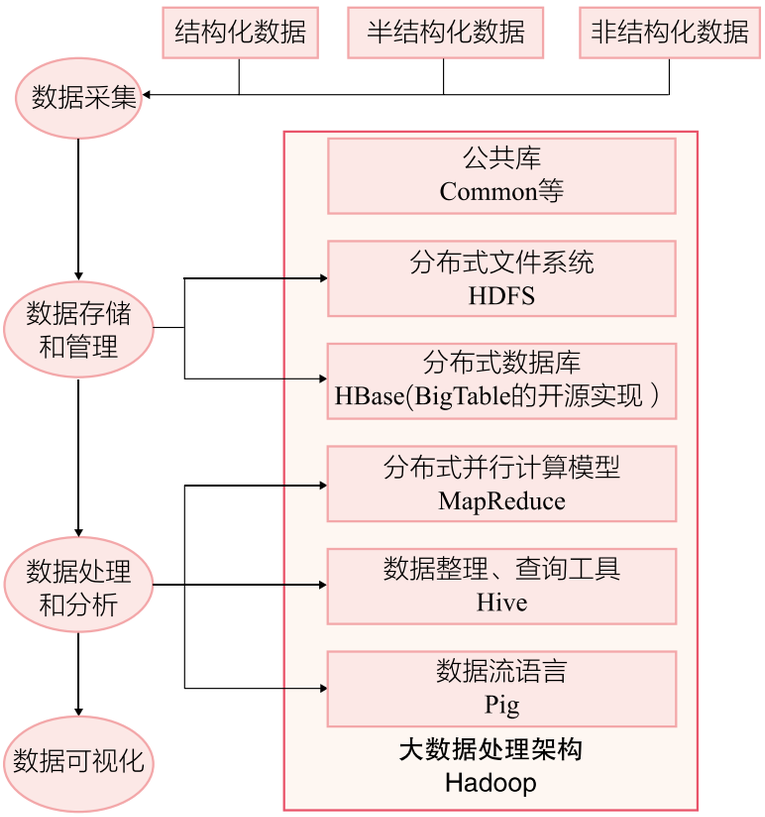


图4.2.2　Hadoop的组成

（1）分布式文件系统HDFS

Windows的文件系统采用FAT32或NTFS，Linux的文件系统为Ext2/Ext3/Ext4，这些文件系统均不能满足分布式文件的管理需求。Hadoop分布式文件系统（Hadoop Distributed File System，简称HDFS）是谷歌文件系统（Google File System，简称GFS）的开源实现。它的主要功能是将大规模海量数据以文件的形式、用多个副本保存在不同的存储节点中，并用分布式系统进行管理。HDFS是一个高度容错性的系统，适合部署在廉价的机器上。目前，云盘、网盘的底层一般采用HDFS实现。

（2）分布式数据库HBase

HBase是一个高可靠、高性能、可伸缩、分布式的列式数据库，是谷歌BigTable数据库的开源实现。与传统关系型数据库采用基于行的存储形式、用于管理表格类的结构化数据不同，HBase建立在HDFS提供的底层存储基础上，采用基于列的存储方式，主要用来存储非结构化数据和半结构化数据，具有良好的横向扩展能力，可管理PB级的大数据。

（3）分布式并行计算模型MapReduce

MapReduce是一种分布式并行编程模型，能够处理大规模数据集的并行运算，主要由Map（映射）和Reduce（归纳）2个函数构成。HDFS提供了分布式计算时每个节点服务器对数据的访问，HDFS 与MapReduce 的结合，使得在处理大数据的过程中计算性能、数据容错性得到了保障。

当数据量很大时，一台服务器的处理能力无法满足需求，这时，MapReduce分布式并行计算的优势就体现出来了，它的核心处理思想是将任务分解并分发到多个节点上进行处理，最后汇总输出。如图4.2.3所示，大数据集拆分为多个分片数据后分发到多个服务器中，Map函数把处理要求映射为多个map任务在节点服务器进行计算处理，节点任务处理完成后由Reduce函数归纳计算结果并输出。

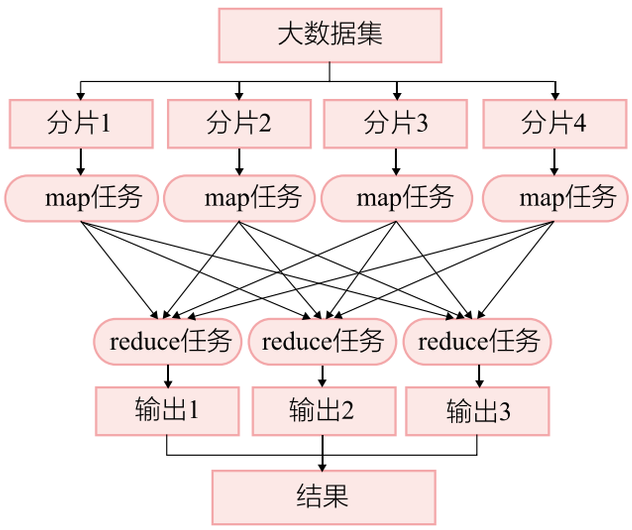


图4.2.3　MapReduce的工作流程

拓展链接

Hadoop的发展历史

Hadoop最早起源于Nutch项目。Nutch是一个开源的网络搜索引擎，由Doug Cutting于2002年创建。随着网页数量的增加，项目组遇到了数十亿网页的存储和索引问题。

2003年底，谷歌发表了关于谷歌分布式文件系统的论文。该论文描述了谷歌搜索引擎网页相关数据的存储架构，该架构可解决Nutch 遇到的网页抓取和索引过程中产生的超大文件存储需求问题。由于谷歌仅开源了思想而未开源代码，Nutch项目组便根据论文开源实现了Nutch 的分布式文件系统（NDFS）。

2004年，谷歌发表了关于谷歌分布式计算框架MapReduce的论文，该框架可用于处理海量网页的索引问题。Nutch的开发人员依据论文完成了MapReduce的开源实现。

2006年初，NDFS和MapReduce从Nutch项目分离，Doug Cutting用儿子的棕黄色大象玩具的名字为项目起名为Hadoop。同年2月，Apache Hadoop项目正式启动以支持MapReduce和HDFS 的独立发展。

2008 年1月，Hadoop 成为Apache顶级项目，迎来了它的快速发展期。

2. 流计算

Hadoop的设计初衷是面向大规模的批量处理，适用于处理静态数据，在流数据实时处理时明显性能不足，比如大型购物网站的广告推荐、社交网络的个性化推荐、根据交通路况实时更新导航线路等应用场景。随着数据处理量及实时性要求的提高，诞生了专门处理流数据的计算平台，如图4.2.4所示。

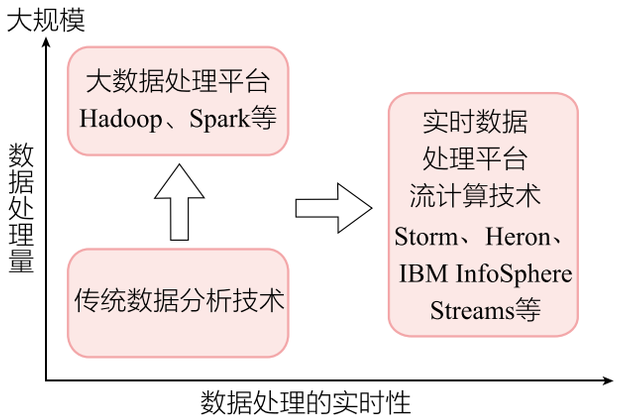


图4.2.4　流计算的发展

通过流计算系统，可以简单、高效、可靠地实现实时数据的获取、传输和存储，在与数据库、Hadoop、编程语言等整合后可开发出功能强大的实时计算与分析应用。典型的应用如Twitter的社交网络数据处理，采用了如图4.2.5所示的分层数据处理架构，每天可实时处理数十亿事件的数据。

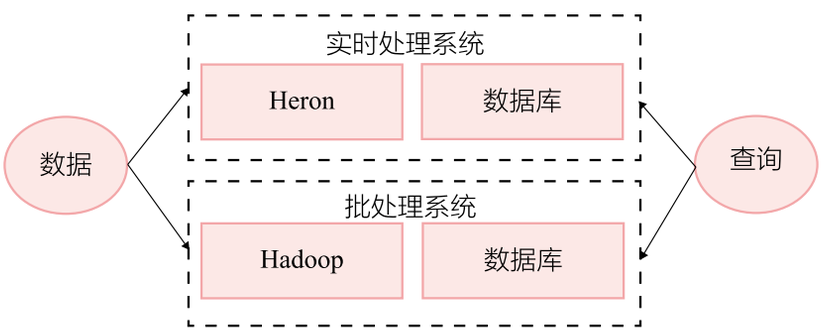


图4.2.5　Twitter的分层数据处理架构

拓展链接

主要的流计算软件系统

目前，处理流数据的软件系统主要有IBM InfoSphere Streams、Twitter Storm、Yahoo! S4、银河流数据处理平台（淘宝）、Facebook Puma等。Storm和S4是目前较为流行的开源分布式实时计算系统。Heron是Storm的替代产品，其外部接口和Storm保持兼容，在流数据处理性能方面与Storm相比有了大幅提升。

3. 图计算

现实世界中的很多数据是以图的形式呈现的，或者是可以转换为图以后再进行分析的，如社交网络、网络浏览与购买行为、传染病的传播路径等。大规模的图往往有数十亿的节点和数千亿的边（节点之间关系的连线），节点之间的关系错综复杂，如图4.2.6所示

的蛋白质激素构成图。传统的Hadoop架构在处理大型图计算的问题时性能上明显不足，专业的图计算软件应运而生。目前通用的图处理软件主要包括两类：一类是图数据库，如Neo4j、InfiniteGraph、OrientDB等；另一类是并行图处理系统，如Google Pregel、Apache Giraph、卡内基梅隆大学的GraphLab、运行于Spark平台的GraphX等。

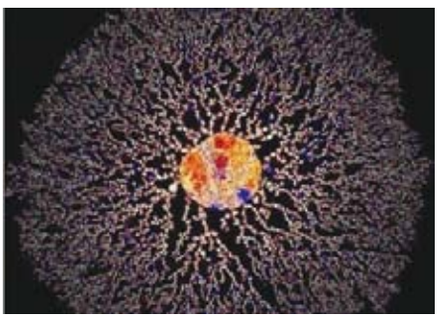


图4.2.6　蛋白质激素构成图（由2.7万个节点和794万条边组成）

4. 实时处理与批处理的整合

2014年9月，Twitter开源了其大数据处理系统Summingbird，该系统实现了批处理和流计算在一个平台架构下的整合（Hadoop+Storm）。开发者在同一个平台既可以做批处理，也可以做流计算，还可以进行两种模式的混合使用。平台的整合缩短了批处理与流处理之间的切换延时时间，有利于减少系统的开销，降低使用成本。

问题与讨论

结合生活实践，查找资料，列举静态数据、流数据处理实例。

拓展链接

Hadoop应用实例：
北京城市数据映像——流动的城市

“北京城市数据映像”项目采集了北京市地铁一卡通数据、出租车GPS定位轨迹数据、移动手机基站定位、地理位置微博数据、工商业POI地点等约2TB的数据。数据计算平台采用了服务器集群、Hadoop和HBase架构。

通过收集北京市各相关行业的数据，运用大数据分析和可视化表达技术，将城市的发展和变化过程变得直观、透明和可视。大数据分析为城市管理提供了技术支撑，是发现、分析城市问题的新思维和技术方法。

图4.2.7分析显示了市民居住地与工作地的分布情况，图4.2.8展示了北京地铁系统不同线路的流量与换乘情况。

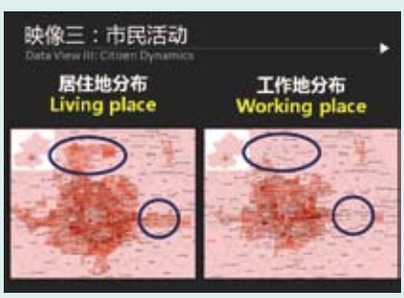


图4.2.7　市民活动

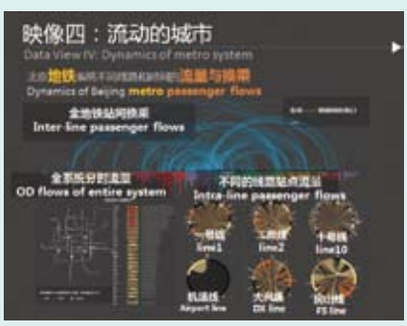


图4.2.8　流动的城市

4.2.2 编程处理数据

使用计算机语言编程，可以更加灵活、深入地进行数据分析和挖掘。选用Python语言编程进行数据处理，可以调用Python的扩展模块，常用的扩展模块有numpy、scipy、pandas和matplotlib等。

numpy模块是Python中做科学计算的基础库，主要提供科学计算中常用的随机数、数组运算等基础函数。

scipy模块是基于numpy构建的一个模块，增强了在高等数学、信号处理、图像处理、统计等方面的处理能力。

pandas模块基于numpy实现，主要用于数据的处理和分析。它提供了大量处理数据的函数和方法，能方便地操作大型数据集。

下面以pandas为例，介绍使用计算机程序设计语言编程进行数据处理的方法。

1. 利用pandas模块处理数据

pandas提供了Series和DataFrame两种数据结构。使用这两种数据结构，可完成数据的整理、计算、统计、分析及简单可视化。

在Python中引入pandas模块的方法如下：

```
import pandas as pd
```

（1）Series

Series是一种一维的数据结构，包含一个数组的数据和一个与数据关联的索引（index），索引值默认是从0起递增的整数。列表、字典等可以用来创建 Series 数据结构，与列表不同的是，Series的索引可以指定，类型可以为字符串型。

例1　创建1个Series结构类型的对象s1，存储3名同学的身高值。

```
s1=pd.Series([166,178,180])
print(s1)
运行结果：
0    166
1    178
2    180
dtype: int64
```

左列：index

右列：values

```
#创建Series对象时指定索引
s2=pd.Series([166,178,180],index=["s01","s02","s03"])
print(s2)
运行结果：
s01    166
s02    178
s03    180
dtype: int64
```

通过索引可以选取Series对象中的值，通过赋值语句可以修改Series对象中的值。如：s1[0]=168、s2["s01"]=168，可将s1、s2对象中的“166”改为“168”。

Series对象常用属性如表4.2.1所示。

表4.2.1　Series对象常用属性

属性	说明
index	Series的下标索引，其值默认是从0起递增的整数
values	存放Series值的一个数组

●●● 例2　查看例1中s1对象的index、values属性值。

```
for i in s1.index:
    print(i)
运行结果:
0
1
2
```

```
for i in s1.values:
    print(i)
运行结果:
166
178
180
```

```
for i in s1:
    print(i)
运行结果:
166
178
180
```

（2）DataFrame

DataFrame是一种二维的数据结构，由1个索引列（index）和若干个数据列组成，每个数据列可以是不同的类型。DataFrame可以看作是共享同一个index的Series的集合。创建DataFrame对象的方法很多，通常用一个相等长度的列表或字典来创建。

●●● 例3　使用相等长度列表的字典构建一个DataFrame对象df1，存储3名同学的姓名、性别、图书借阅次数数据。

```
import pandas as pd
data={"姓名":["王静怡","张佳妮","李臣武"],"性别":["女","女","男"],"借阅次数":[28,56,37]}
df1=pd.DataFrame(data,columns=["姓名","性别","借阅次数"])
print(df1)
运行结果:
    姓名  性别  借阅次数
0  王静怡   女    28
1  张佳妮   女    56
2  李臣武   男    37
```

设定df1中数据列的顺序

index

可以直接读取二维数据文件创建DataFrame对象。如使用read_excel()函数，读取Excel文件创建DataFrame对象，也可以使用to_excel()函数，创建Excel文件保存数据。

●●● 例4　读取Excel文件“test.xlsx”中的数据，创建DataFrame对象df。

```
import pandas as pd
df=pd.read_excel("test.xlsx")
print(df)
```

运行结果：

	地区	规格	单位	价格	采价点	采集时间
0	北京市	红富士 一级	元/500克	2.98	超市2	11月中旬
1	北京市	红富士 一级	元/500克	4.88	超市1	11月中旬
2	天津市	红富士 一级	元/500克	5.00	超市1	11月中旬
3	天津市	红富士 一级	元/500克	5.00	超市2	11月中旬
4	石家庄市	红富士 一级	元/500克	3.98	超市1	11月中旬
5	石家庄市	红富士 一级	元/500克	3.98	超市2	11月中旬

DataFrame对象常用属性如表4.2.2所示。DataFrame中的索引、列标题及值可以通过属性来显示。

表4.2.2　DataFrame对象常用属性

属性	说明
index	DataFrame的行索引
columns	存放各列的列标题
values	存放值的二维数据
T	行列转置

●●● 例5　查看df1对象的索引、列标题、值，并将行、列转置。

```
for i in df1.index:
    print(i)
运行结果：
0
1
2
```

```
for i in df1.columns:
    print(i)
运行结果：
姓名
性别
借阅次数
```

此处可简写为df1:

```
for i in df1.values:
    print(i)
运行结果：
    ['王静怡' '女' 56]
    ['张佳妮' '女' 52]
    ['李臣武' '男' 68]
```

```
print(df1.T)      #转置行、列
运行结果：
          0     1     2
姓名    王静怡  张佳妮  李臣武
性别      女     女     男
借阅次数  56    52    68
```

和Series对象一样，DataFrame对象中的一列可以通过字典记法或属性来检索，列可以通过赋值来修改。

●●● 例6　分别检索df1对象中“姓名”“借阅次数”列数据，并修改“借阅次数”列数据。

```
print(df1.姓名)
#通过属性检索列
运行结果：
0   王静怡
1   张佳妮
2   李臣武
Name:姓名,dtype: object
```

```
print(df1["借阅次数"])
#通过字典记法检索列
运行结果：
0   28
1   56
2   37
Name:借阅次数, dtype: int64
```

```
df1.借阅次数=[30,52,68]
print(df1)
运行结果：
    姓名  性别  借阅次数
0  王静怡   女     30
1  张佳妮   女     52
2  李臣武   男     68
```

可以通过布尔型数据选取满足条件的行。如通过df1[df1["借阅次数"]>30]，可以检索df1 对象中“借阅次数”大于30 的数据行。使用at[]方法可以根据行标签和列标签选取单个值，如通过df1.at[0,"姓名"]，可以选取df1对象中第1行、“姓名”列的值。

DataFrame数据结构提供了丰富的函数，这些函数可以用来进行行、列编辑和统计计算等。DataFrame常用函数如表4.2.3所示。

表4.2.3　DataFrame 常用函数

函数	说明
count()	返回非空（NaN）数据项的数量
sum()、mean()	求和、求平均值，通过axis=0/1 确定行列
max()、min()	返回最大、最小值
describe()	返回各列的基本描述统计值，包含计数、平均数、标准差、最大值、最小值及4分位差
head()、tail()	返回DataFrame的前n个、后n个数据记录
groupby()	对各列或各行中的数据进行分组，然后可对其中每一组数据进行不同的操作
sort_values()	排序，通过axis=0/1 确定行列
drop()	删除数据，通过axis=0/1 确定行列
append()	在指定元素的结尾插入内容
insert()	在指定位置插入列
rename()	修改列名或者索引
concat()	合并DataFrame对象
set_value()	根据行标签和列标签设置单个值
plot()	绘图

① DataFrame对象中行、列的编辑。DataFrame中，新增列、删除列、重命名列可以通过insert()、drop()、rename()等函数完成；追加数据行可以通过append()函数完成；使用set_value()函数可以根据行标签和列标签设置单个值。

●●● 例7　对df对象中的数据进行以下编辑：在最后追加一行数据；删除“规格”列数据；删除第1行数据。

```
#添加1行数据
df_add=df.append({"地区":"石家庄市","规格":"红富士 一级","单位":"元/500克","价格":4.00,"采价点" :"集市3","采集时间":"11月中旬"},ignore_index=True)
df_delc=df.drop("规格",axis=1)              #删除"规格"列数据
df_delr=df.drop(0)                          #删除第1行数据
```

说明：append()、drop()函数均不改变原有df对象中的数据，而是通过返回另一个DataFrame对象来存放改变后的数据。如本例中df_del=df.drop("规格",axis=1) 不改变df对象中的数据，删除后的数据存放在df_del对象中，del df["规格"] 会永久删除df 对象中"规格"列数据。

② DataFrame对象中数据的统计与计算。使用groupby()函数，可以对DataFrame对象各列或各行中的数据进行分组，然后对其中每一组数据进行不同的操作。

例8　将df对象中的数据按“地区”分组，并计算分组后各组数据的平均值。

```
g=df.groupby("地区",as_index=False)
print(g.mean())                                    #计算每组价格数据的平均值
#分组、求平均的代码，也可以写作：g=df.groupby("地区",as_index=False).mean()
运行结果：
          地区          价格
0        北京市          3.93
1        天津市          5.00
2       石家庄市         3.98
```

③ DataFrame对象中数据的排序。DataFrame对象中，按索引排序可以使用sort_index()函数，按值排序可以使用sort_values ()函数。通过选项axis=0/1 确定排序的轴向，axis默认值为0，纵向排序；通过选项ascending=True/False确定升/降序，ascending默认值为True，升序排序。排序结果返回一个新DataFrame对象。

例9　对df对象中的数据，按“价格”值降序排序。

```
df_sort=df.sort_values("价格",ascending=False)  #按价格值降序排序
print (df_sort)
运行结果：
          地区          规格          单位          价格     采价点     采集时间
2        天津市      红富士 一级      元/500克      5.00     超市1     11月中旬
3        天津市      红富士 一级      元/500克      5.00     超市2     11月中旬
1        北京市      红富士 一级      元/500克      4.88     超市1     11月中旬
4       石家庄市     红富士 一级      元/500克      3.98     超市1     11月中旬
5       石家庄市     红富士 一级      元/500克      3.98     超市2     11月中旬
0        北京市      红富士 一级      元/500克      2.98     超市2     11月中旬
```

2. 利用matplotlib模块绘图

matplotlib是一个绘图库，使用其中的pyplot子库所提供的函数可以快速绘图和设置图表的坐标轴、坐标轴刻度、图例等。常用绘图函数如表4.2.4所示。

表4.2.4　常用绘图函数

函数	说明
figure()	创建一个新的图表对象，并设置为当前绘图对象 注：不创建figure对象，直接调用plot等绘图函数进行绘图，matplotlib会自动创建一个figure对象
plot()	绘制线形图
bar()	绘制垂直柱形图
barh()	绘制水平柱形图
scatter()	绘制散点图
title()	设置图表的标题
xlim()、ylim()	设置X、Y轴的取值范围
xlabel()、ylabel()	设置X、Y轴的标签
legend()	显示图例
show()	显示创建的所有绘图对象

在Python中引入matplotlib的pyplot子库的方法为：

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

●●● 例10　绘制正弦曲线图。

```
import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(0, 10, 1000)
y1 = np.sin(x)
y2 = np.sin(x**2)
plt.figure(figsize=(8,4))                          #创建图表对象

plt.title("sin(x) and sin(x**2)")                  #设置图表标题文字
plt.plot(x,y1,label="sin(x)",color="r",linewidth=2)  #绘制线形图
plt.scatter(x,y2,label="sin(x**2)")                #绘制散点图

plt.ylim(-1.5,1.5)                                 #设置y坐标轴的取值范围
plt.xlim(0,10)                                     #设置x坐标轴的取值范围
plt.legend()                                       #显示图例

plt.show()
```

运行程序，上述代码中figsize参数指定figure对象的宽度和高度；color指定线条的颜色；linewidth指定线条的宽度；label给线条指定一个标签名称，该标签显示在图例中，绘制的图表如图4.2.9所示。

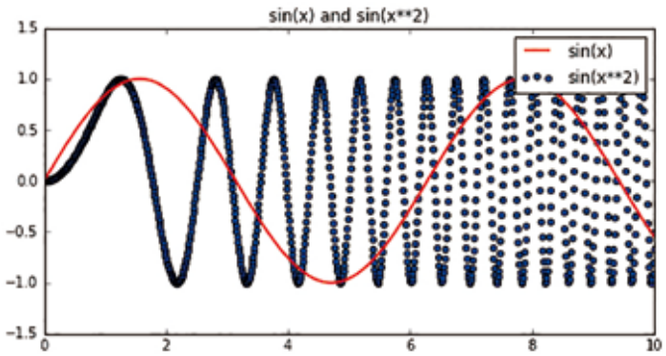


图4.2.9　正弦图

3. 利用 Python 分析数据实践

下面以“身边的百家姓”项目为例，学习和体会编程处理数据的方法和优势。

身边的百家姓

通过统计某地的姓名数据，分析当地姓氏的构成情况。

- 分析数据

如图4.2.10所示，姓名数据以CSV文件格式组织和存储，以UTF-8格式编码。xm.csv文件大小为26.7 MB，共有2594178条姓名数据，内容包含姓氏和名字。本次数据处理的目的是统计全部数据中不同姓氏的人数，并通过排序和图表进一步分析。

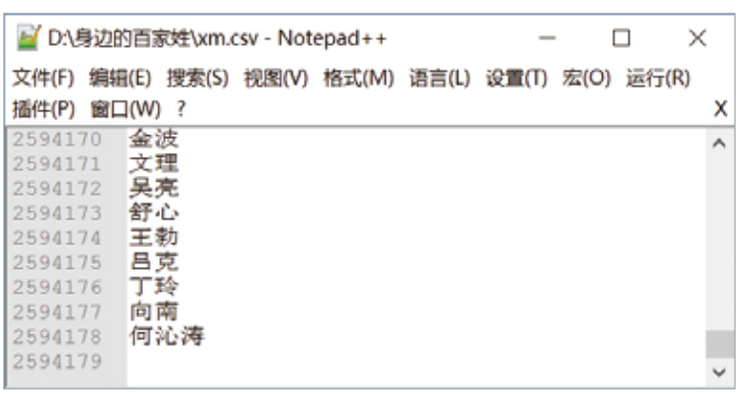


图4.2.10　姓名数据

- 编制程序

使用Python编程统计、分析当地居民姓氏的构成情况，程序如下：

```
import pandas as pd
import matplotlib.pyplot as plt
import codecs                                                    #处理中文utf-8编码
from matplotlib.font_manager import FontProperties               #显示中文字体

file = codecs.open("xm.csv","r","utf-8")                         #打开文件
# 定义复姓 list
fx=["欧阳","太史","端木","上官","司马","东方","独孤","南宫","万俟","闻人","夏侯","诸葛","尉迟","公羊","赫连","皇甫","濮阳","公冶","申屠","公孙","慕容","钟离","长孙","宇文","司徒","鲜于","司空","闾丘","子车","亓官","宰父","谷梁","拓跋","轩辕","令狐","百里","呼延","东郭","南门","羊舌","公仪","西门","第五"]
```

```
xing=[]
for line in file:
    if line[0:2] in fx:                                    #取复姓
        xing.append(line[0:2])
    else:                                                  #取单姓
        xing.append(line[0:1])
data={"xing":xing,"renshu":0}                              #构造 DataFrame 数据结构
df=pd.DataFrame(data)
s= df.groupby("xing").count()                              #按"xing"分组计数
s=s.sort_values("renshu",ascending=False)                  #按"renshu"降序排序
ax=s[0:20].plot(kind="bar",rot=0)                          #对前20绘图
#显示中文标签
font = FontProperties(fname=r"c:\windows\fonts\simsun.ttc", size=12)
for label in ax.get_xticklabels() :
    label.set_fontproperties(font)
plt.show()                                                 #显示图形
print(s)
```

- 查看结果

运行上述Python程序，结果如图4.2.11所示。观察图表，发现在2594178条姓名数据中，人数前五的姓氏依次为：王、李、张、刘、陈；王姓的人数最多，有102400人。

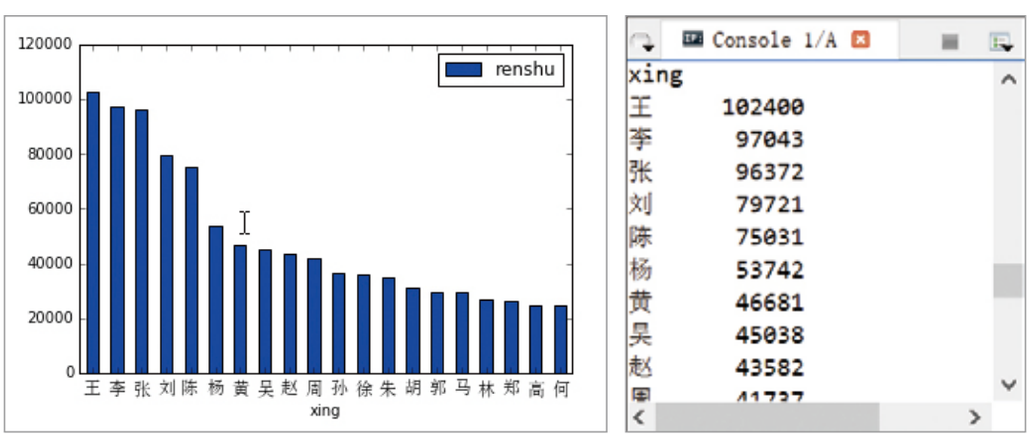


图4.2.11　姓氏统计结果

拓展链接

用Hadoop处理姓氏数据

当xm.csv文件的数据量增长到GB、TB时，单台计算机将不适于处理这类问题，这时就需要采用适用于处理静态大数据的Hadoop架构，编写Map和Reduce函数来处理。在Map函数中统计每个分片数据中各个姓的人数，统计结果再作为Reduce函数的输入，在Reduce函数中汇总每个姓的总计人数。在Hadoop服务器中运行MapReduce任务，系统会自动把任务分配

到各个计算机中运行。如图4.2.12所示，Hadoop分配了n个Map任务和m个Reduce任务实现姓氏统计。

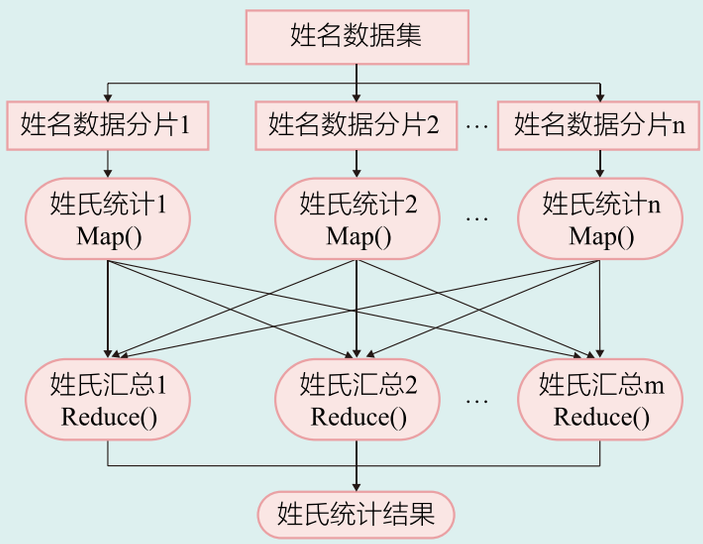


图4.2.12　姓氏统计的MapReduce示意图

4.2.3　文本数据处理

文本数据处理是大数据处理的重要分支之一，目的是从大规模的文本数据中提取出符合需要的、感兴趣的和隐藏的信息。目前，文本数据处理主要应用在搜索引擎、情报分析、自动摘要、自动校对、论文查重、文本分类、垃圾邮件过滤、机器翻译、自动应答等方面。

1. 文本数据处理的一般过程

文本内容是非结构化的数据，要从大量的文本中提取出有用的信息，需要将文本从无结构的原始状态转化为结构化的、便于计算机处理的数据。典型的文本处理过程主要包括：分词、特征提取、数据分析、结果呈现等，如图4.2.13所示。

图4.2.13　典型的文本处理过程

（1）中文分词

中文分词是中文文本信息处理的基础，机器翻译、全文检索等涉及中文的相关应用中都离不开中文分词。分词是将连续的字序列按照一定的规范重新组合成词序列的过程，也

就是将一个汉字序列切分成一个一个单独的词。因为英文词语与词语之间有明显的空格，分词不涉及复杂的关键词提取方法，而中文词与词之间是紧密相连的，分词方法相当复杂，目前的分词算法还不能实现完全准确的分词。常用的中文分词算法可分为如下三类：

①基于词典的分词方法，也称作基于字符匹配的分词方法，即在分析句子时与词典中的词语进行对比，词典中出现的就划分为词。如图4.2.14所示的是Python中文分词模块jieba中词典（dict.txt）的截图。

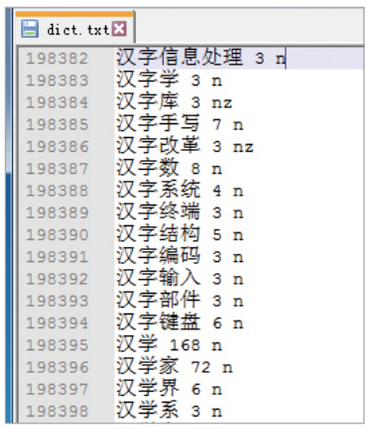


图4.2.14　jieba分词模块中的词典

②基于统计的分词方法，统计分词的思想是依据上下文中相邻字出现的频率统计，同时出现的次数越高就越可能组成一个词。在实际应用中，一般是将其与基于词典的分词方法结合使用。

③基于规则的分词方法，通过让计算机模拟人的理解方式，根据大量的现有资料和规则进行学习，达到对文字进行分词的效果。由于中文语言知识的笼统性、复杂性，这种分词方法目前还处于试验阶段。

目前常见的公开提供服务的分词系统如表4.2.5所示。

表4.2.5　常见的分词系统

名称	简介
jieba分词	Python开源项目
IKAnalyzer	Java开源分词工具包
NLPIR	北京理工大学大数据搜索与挖掘实验室，非商业应用免费
语言云	哈尔滨工业大学社会计算与信息检索研究中心，在线API接口调用
BosonNLP	玻森中文语义开放平台，在线API接口或库调用

（2）特征提取

在中文文本分析中可以采用字、词或短语作为表示文本的特征项。相比较而言，词的切分难度比短语的切分难度小且更能表达文本的含义。目前，大多数中文文本分析中都采用词作为特征项，这种词称作特征词。

通常可直接用分词算法和词频统计得出的结果作为特征词，但对于稍大一些的文本，提取出的特征词数量将非常大，其计算处理过程的效率非常低，计算结果的准确性也很难令人满意。因此，必须找出最具代表性、最有效的文本特征，通常的办法是通过特征提取来减少特征词的数量，提高文本处理的速度和效率。

特征提取一般采用的方式为根据专家的知识挑选有价值的特征，或者用数学建模的方法构造评估函数自动选取特征等。目前大多采用评估函数进行特征提取的方式，评估函数大多是基于概率统计设计的，这就需要用庞大的训练数据集才能获得对分类起关键作用的特征。随着深度学习、大数据分析等技术的发展，文本特征提取将更加准确、科学。

2. 文本数据分析与应用

在取得特征词后，对文本的分析就需要根据项目的需求，确定解决问题的路径，选取合适的工具、设计算法抽取出文本中隐含的价值。下面以标签云、文本情感分析等为例感受文本数据的处理。

（1）标签云

标签云用词频表现文本特征，将关键词按照一定的顺序和规律排列，如频度递减、字母顺序等，并以文字大小的形式代表词语的重要性，如图4.2.15所示。标签云广泛应用于报纸、杂志等传统媒体和互联网。

标签云是文本可视化的一种方式。文本可视化将文本中复杂的或者难以通过文字表达的内容和规律以视觉符号的形式表达出来，使人们能够利用视觉感知能力快速获取文本数据中所蕴含的关键信息，为更好地理解文本和发现知识提供了新的有效途径。

图4.2.15　标签云

图4.2.16　城市心情

（2）文本情感分析

文本情感分析是指通过计算机技术对文本的主观性、观点、情绪、极性进行挖掘和分析，对文本的情感倾向做出分类判断。文本情感分析作为一个多学科交叉的研究领域，涉及自然语言处理、信息检索、机器学习、人工智能等领域。

文本情感分析根据分析的粒度不同，分为词语级、语句级、整篇文章级三类。词语级是在分词的基础上，根据情感词典进行特征提取与分类，再分别给特征词赋予权重进行统计分析。特征词的权重，例如，满意+5；差–5等。

文本情感分析主要应用于网络舆情监控、用户评论分析与决策、信息预测等众多领域。

“北京城市数据映像”项目通过采集北京地区的微博数据进行了情感分析的研究，用不同的颜色表示心情，在区域地图上展示了不同地域人们在不同时间点情绪的变化，如图4.2.16所示。

拓展链接

语文作文机器自动评分

语文作文试卷的评分一般可从字迹工整度、词汇丰富性、句子通顺性、文采、篇章结构、立意等多个维度综合评估。某语文作文智能评分系统处理流程如图4.2.17所示，智能阅卷的步骤主要包括：试卷图文转写、内容相似检测及拒评处理、专家定标评分、评分模型训练、智能评分等几个主要部分。

图4.2.17　某语文作文智能阅卷流程

1. 试卷图文转写。首先将学生的试卷扫描转换成电子图片存储于计算机中，然后识别待评阅试卷图片中的文字，将学生手写的文字进行高精准度的识别，转换为计算机可处理的文本格式。一般而言，书写越工整，识别越准确，过于潦草的，计算机会拒识。

2. 内容相似检测及拒评处理。除了由于字迹潦草而被拒识的情况外，计算机还可检测作文内容的相似度情况，将高相似的作文做拒评处理，交由人工裁定和评分。

3. 专家定标评分。从待评作文集中抽取适量作文形成定标作文集，由专家对这些作文进行评分，这个过程称为定标。定标集应选择有代表性的作文，覆盖各个分数段。计算机可利用文本分析等技术，将作文按照内容相似性和大概的写作水平进行分类。

4. 评分模型训练。训练是从已知数据寻找模型参数的过程。计算机可从定标集中提取出作文的内容材料特征，用一个或多个机器学习算法对评分模型参数进行调整、验证，训练出最终的评分模型。

5. 智能评分。计算机可从多个评分维度对一篇作文的质量进行评价。如同我们可以通过身高、皮肤、五官、头发、胡须、衣着、动作、嗓音等特征来判断一个人的年龄，计算机智能评分则是通过字迹工整度、词汇丰富度、句子通顺性、是否切题、立意高低、篇章结构、文采等多个维度对一篇作文进行综合评分。

以最直观的线性回归算法为例，针对一篇作文，按照上述评分维度抽取若干数值化的特征，比如使用了x_1个成语、全篇涂改了x_2处、立意高低得分为x_3……作文的分数y可以按照如下公式计算：

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

其中，w_1，w_2，…，w_n称为各维特征的权重。机器学习算法可以从已知得分y的定标集数据中估算出这些权重，一组权重即构成一个评分模型。对于未知分数的一篇作文，同样抽取数值化特征，分别乘以模型中对应的权重，再累加即可得到分数。

实践与体验

中文分词与标签云

教材配套的“搜索抓取、中文分词与标签云生成”软件，其界面如图4.2.18所示。

图4.2.18　搜索抓取、中文分词与标签云生成

实践内容：

使用“搜索抓取、中文分词与标签云生成”软件，收集感兴趣的数据，如家乡的旅游景点、美食等，进行分词、词频统计并以标签云方式展现结果。

实践步骤：

1. 启动软件，在关键词框中输入需要检索的关键词（如数据），选择微软Bing或百度搜索引擎，单击“检索”按钮启动搜索引擎，抓取搜索结果网页中的文本内容。

2. 单击“分词”按钮，对抓取到的网页内容进行分词。

3. 单击“词频统计”按钮，统计分词后每个词语的出现次数。

4. 单击“选择模板”按钮，选择创建标签云的模板图片。

5. 单击“标签云”按钮，创建标签云。

结果呈现：

1. 观察生成的标签云，并进行分享。

2. 在“学习编程”菜单中，给出了软件实现的核心源代码。下面是实现标签云的部分代码，感兴趣的同学可参照编写Python程序来实现。

```
import os
import numpy as np
from PIL import Image
import matplotlib.pyplot as plt
from wordcloud import WordCloud,STOPWORDS
d = os.path.dirname(__file__)                          #d 取当前文件路径
pic="alice_mask.png"                                   #pic存放图片名称
pic_mask = np.array(Image.open(os.path.join(d,pic)))
wc = WordCloud(background_color="white", max_words=6000, mask=pic_mask,
stopwords=STOPWORDS,font_path="fonts/simhei.ttf")
wc.fit_words(wf)                                       #生成标签云，wf存放词语及词频
plt.imshow(wc)                                         #显示图片
```

4.2.4 数据可视化

数据可视化是将数据以图形图像等形式表示，直接呈现数据中蕴含信息的处理过程。随着数据数量的不断增加和结构的多元化，直接从数据中获取信息变得困难，将数据以可视化方式展现出来，使用户可以通过直观、交互的方式浏览和观察数据，发现数据中隐藏的特征、关系和模式。

1. 可视化的作用

（1）快捷观察与追踪数据

利用可视化技术，可以将处于不断变化中的数据生成实时变化的可视化图表，帮助人们快捷地发现各种数据的动态变化过程。如百度地图提供的实时路况服务，可以实时查询各大城市的路况信息；中国天气网提供的临近预报服务，可以实时查询全国各地降水、温度、风力等天气实况。

（2）实时分析数据

利用可视化技术，可以实时将数据转换为图像呈现给用户，帮助用户分析数据的内涵和特征。同时，用户还可以根据自己的实际需求，通过改变可视化系统的设置，交互式地从不同角度对数据进行解读和分析。如图4.2.19，是利用百度指数分析全国某段时间搜索关键词“数据可视化”的情况，通过交互，用户可以选择从趋势研究、需求图谱、舆情洞察、人群画像等多个角度进行分析。

图4.2.19　利用百度指数分析关键词“数据可视化”的搜索情况

（3）增强数据的解释力与吸引力

利用数据图表，直观、动态地呈现新闻、研究报告等内容，可以帮助人们在短时间内了解内容、理解数据背后的含义，同时增强数据的吸引力，提高人们的阅读兴趣。已经有

越来越多的新闻、研究报告等使用可视化的方式进行播报和发布。如图4.2.20所示是国家统计局利用可视化方式分析我国大陆人口情况。

图4.2.20　国家统计局分析大陆总人口情况

2. 可视化的基本方法

（1）有关时间趋势的可视化

不同的数据类型决定了可视化的表现形式。万事万物都随着时间的推移而变化，如天气在变化、人口在迁移、经济在发展……人们通过时间序列数据来观察这些事物变化的过程和趋势，如某个变化量是上升还是下降，是否存在周期性变化等。展现这类时间数据可采用柱形图、折线图等。

（2）有关比例的可视化

面对一系列总和为1的比例数据，人们常常关心各部分的大小及其占总体比例的情况，如衣服面料中各组成成分的比例，投票结果中赞成、反对、弃权的情况等。展现这类比例关系的数据可以采用饼图、环形图（也称面包圈图）等。

（3）有关关系的可视化

实际生活中，人们常遇到这样的问题：当某个对象的数量增加时，另一个数量是否会变化？如全民的平均身高增高了，平均体重也会随之增长，这是一种简单的、成正比的关联。关联性意味着当一件事情变化时，另一件事情也可能会发生某种变化。关联性可以帮助人们根据某一已知指标来预测另一指标。要想探究这种数据的分布关系，可以使用散点图、气泡图等。

散点图用于表现2～3个变量之间的关系，以圆点的多少或疏密展示成对的数和它们所代表的趋势之间的关系。如果两个指标是正相关的，在从左往右读图表时，点的位置会越来越高。相反，如果是负相关，从左往右点的位置会越来越低。有时会通过增加颜色维度来表示第三个变量。如图4.2.21所示，某快递公司用户满意度与收货天数关系图的两个维度分别为用户满意度和收货天数。

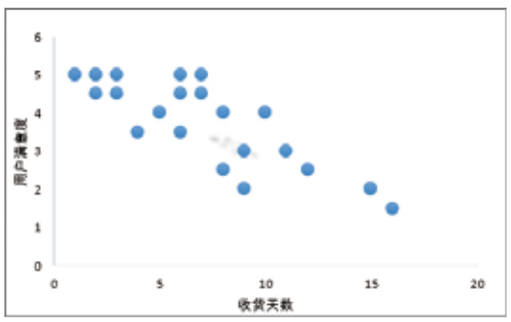


图4.2.21　用户满意度和收货天数关系图

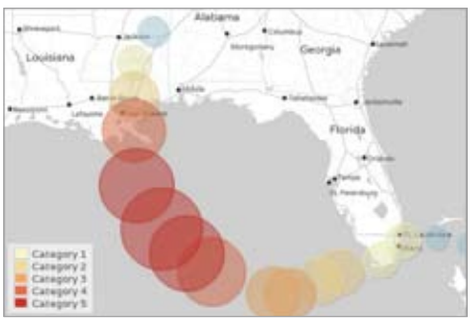


图4.2.22　卡特里娜飓风路径图

气泡图一次比较3～4个变量，x轴和y轴各表示一个变量，第三个变量通过气泡的面积大小来表示，第四个变量通过气泡的颜色来体现。如图4.2.22所示，卡特里娜飓风路径气泡图的四个维度分别为经度、纬度、强度和风力等级，点的面积代表强度，点的颜色表示风力等级。

（4）有关差异的可视化

当数据中包含多种变量，要将所有对象进行分组，然后分析每一个变量及所有变量之间的差异，找出其中的异常值。如两个篮球运动员的场均得分可能是天壤之别，但他们的场均篮板、抢断和盖帽却可能非常接近。要探寻包含多种变量的对象与同类之间的差异和联系，可以采用雷达图。

雷达图有多条轴，每一条轴代表一个变量，从正中心开始，等距平分圆周摆放，每相邻两个变量的终点之间有一条连接线。正中心表示各个变量的最小值，而轴末端的终点代表最大值。雷达图反映数据相对中心点和其他数据点的变化情况，如图4.2.23所示。

图4.2.23　NBA球队赛季成绩雷达图

（5）有关空间关系的可视化

在人们的日常活动中，常常会关心“我们在哪里”“我们周边有什么”“我们如何到达目的地”等问题，这些信息都与地理位置有关。随着传感器、移动终端等设备的普及，带有经度、纬度标签的空间数据成为大数据中的重要数据类型。地理数据或者基于地理数据的分析结果可以运用不同颜色或图表直接表现在地图上进行展示。

问题与讨论

举例说明如何选择合适的可视化方法分析和展示数据。

3. 可视化的工具

数据可视化工具软件很多，常见的数据分析软件中一般包含创建可视化图表功能。主要用于数据可视化的工具有大数据魔镜、Gephi、Tableau等，也可以使用Python、R等计算机语言编写程序实现数据的可视化。此外，还有一些优秀的可视化工具库，如基于JavaScript的D3.js、Highcharts、Google Charts等，基于Python的matplotlib等。

Tableau主要用于实时可视化分析。它可以连接本地或云端数据，包括文件、SQL 数据库、Web 数据，生成柱形图、饼图、基本地图等多种图形。还可以连接动态数据源，将各种图形混合搭配形成定制视图，或者通过仪表盘视图实时关注数据状态。

D3.js是运行在JavaScript上的数据可视化开源工具库。它使用数据驱动的方式，结合强大的可视化组件，可以创建实时交互的网页。

Highcharts是一个用纯JavaScript编写的、基于HTML5技术的开源图表库，支持移动端，能够简单便捷地在Web网站或是Web应用程序中添加动态、交互性的图表。Highcharts 的图表类型丰富，其中很多图表可以集成在同一个图形中形成混合图。它可以免费用于个人学习、个人网站和非商业用途。

Google Charts是为浏览器与移动设备定制的交互式图表开发包，用于在Web上可视化数据。Google Charts功能强大，容易使用，提供了从饼图、时间序列到多维交互矩阵等大量的可视化类型，生成的交互式图表既可以实时输入数据，也可以使用仪表板进行控制。

4. 可视化的典型案例

（1）风、气象、海洋状况的全球地图

“风、气象、海洋状况的全球地图”是一个对全球天气进行可视化的网站。该网站将全球的海洋流动、天气变化和风向、风速等的动态数据，在地图上进行可视化展示，如图4.2.24所示。在这个交互的动画地图上，可以查看现在地球表面的风速流动方向、气象和海洋状况等信息。鼠标拖曳可以移动、改变观察位置。风速越强，地图上线条流动就越快；温度升高，地图颜色就会转为暖色系。有台风出现时，还能清楚地看到台风的结构状况，如圆形的台风眼等。

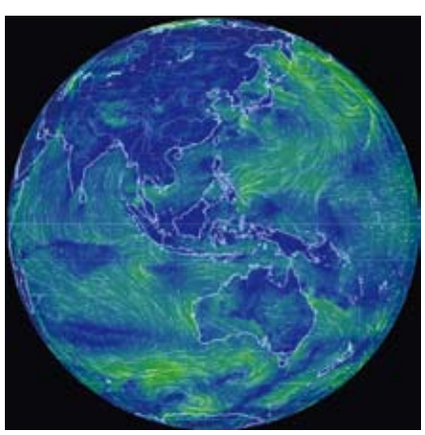


图4.2.24　风、气象、海洋状况的全球地图

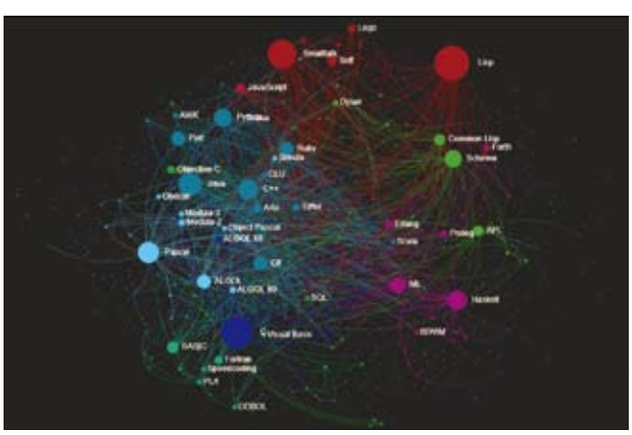


图4.2.25　编程语言之间的影响力关系图

（2）编程语言之间的影响力关系图

图4.2.25显示编程语言之间的影响力关系，它是德国软件工程师Ramio Gómez制作的交互式关系网图。该图的数据来自 Freebase网站维护的编程语言表，其中包含3900多万个主题、2011 种类型和3万多个属性。图中的每个节点代表一种编程语言，点与点之间的连线表示语言之间有影响关系。影响关系多的语言，其节点在关系网中也越大，如Lisp、C、Smalltalk、Java等。单击各节点，可以查看受其影响的其他语言。如单击Python语言的节点，可以查看受Python影响的其他语言。

（3）“双十一”全网销售直播图

如图4.2.26所示为（星图数据）2016年“双十一”网购狂欢节电商全网销售实时直播图。直播过程中，对海量的实时销售数据，采用了可视化方式进行展现。观察图中信息，可以发现各平台占比采用条形图呈现，贸易往来关系采用地图展现，交易过程中产生的包裹数量采用折线图呈现，各平台海外销售额占比采用柱形图呈现，此外还采用了环形图及其他个性化展现方式。

图4.2.26　“双十一”全网销售直播图

（4）航班飞行实时跟踪地图

图4.2.27是一家航班跟踪数据公司（Flight Aware）基于全球数以千计的实时数据源提供的航班跟踪地图。在地图上实时显示当前区域中的航班飞行状态，单击某航班，可以查看其已飞航线、航路计划、飞行时间、准点情况、机型、周围航班等信息。

图4.2.27　航班飞行实时跟踪地图

（5）微博热词趋势图

微指数是新浪微博提供的数据分析工具，它通过对海量微博数据、用户行为数据的整理与挖掘，呈现热词整体趋势、实时变化、地域解读、人群属性分析等结果，反映微博舆情。其中的热词指数通过统计关键词的每日微博热议度，分析其在微博平台中的长期热议趋势，并采用折线图进行展现。如图4.2.28所示，对比“粽子”“龙舟”两个关键词在端午节前后1个月的微博热议趋势发现，“粽子”“龙舟”均在端午节当天热议指数达到顶峰，但是“粽子”的提及度要远高于“龙舟”，说明“粽子”在端午节比“龙舟”更受人们的关注。

图4.2.28 “粽子”“龙舟”微博热议趋势图

热词趋势分析的折线图反映出搜索词近期的热度。实时趋势分析可以实时地展现这个词此时此刻最新的微博搜索数据，结果同样展现为折线图。地域解读部分，将不同地域的关键词搜索热度，在地图上通过地域的颜色深浅来展现。属性分析部分，主要是对用户群体的性别、年龄、标签、星座进行分析，并采用柱形图、雷达图等展现。

思考与练习

1. 上网查找Hadoop处理大数据的应用实例，制作演示文稿并向同学介绍。

2. 参照“身边的百家姓”实例，统计分析当地近年来居民姓名中常用名的情况。

3. 使用Python中文分词模块jieba，体验中文分词（以下为示例代码）。

```
import jieba                                    #引用jieba分词模块
text= open("filename.txt","r").read()           #读入文本文件
seg_list = jieba.cut(text, cut_all = True)      #全模式分词
print("全模式分词:", " ".join(seg_list))          #输出分词结果
seg_list = jieba.cut(text)                      #默认模式分词
print("默认模式:", " ".join(seg_list))            #输出默认模式分词结果
```

4. 列举文本数据处理的应用领域以及将来可能的应用。

4.3 大数据典型应用

随着大数据在各行业的应用，数据成为核心资产，数据规模以及运用数据的能力成为各行业发展的推动力。目前，大数据广泛应用于金融、交通、环境、医疗、能源、农业等行业，极大地促进了各行业的发展。本节以智能交通和电子商务为例进行分析。

4.3.1 智能交通

在交通运输领域中，随着移动互联网、物联网、云计算、大数据等技术的发展，智能交通的发展进程正逐渐加快。人们越来越多地感受到智能交通带来的便利。例如，铁路部门推出的网络订购火车票服务，让人们通过个人计算机、智能手机就能随时随地查看火车车次和购票；民航提供的网络订票、在线值机等服务，让出行者可以便捷地查看航班动态、购票值机；智能公交系统，让人们没出家门就知道即将乘坐的公交车到哪儿了、何时进站。人们的日常出行越来越离不开导航系统、打车软件。那么，智能交通系统是如何实现这些服务的呢？

智能交通将数据通信传输技术、电子控制技术、计算机处理技术等应用于交通运输行业，通过对交通数据的实时采集、传输和处理，借助各种科技手段和设备，对交通情况进行协调和处理，从而使交通设施得以充分利用，提高交通效率和安全性，最终使交通运输服务和管理智能化。

智能交通整合了物联网、大数据、云计算、人工智能等技术，其基本架构如图4.3.1所示。GPS、卡口、视频检测、浮动车、地感线圈等产生的交通流监测数据、视频监控数据、系统数据、服务数据等构筑了交通大数据。交通数据采集的广度、深度和数据量随着智能交通的发展不断扩大，数据贯穿在智能交通的感知、处理、应用等各个环节。交通大数据是智能交通中“智能”的基础。

图4.3.1 智能交通架构图

云计算使千亿数据的检索实现了秒级返回，为大数据的分析应用提供了速度保障。基于深度学习的智能分析算法，为大数据的分析应用提供了有力的支撑。交通大数据的分析，为交通管理、规划、决策、服务和主动安全防范等提供了更加有效的支持。

智能交通主要通过交通信息服务、交通管理、公共交通、车辆控制、货运管理、电子收费、紧急救援等服务子系统为用户提供服务。以下简要介绍其中的三个子系统。

（1）交通信息服务系统

交通信息服务系统建立在完善的信息采集、处理和传输系统上。交通参与者通过安装在道路、车上、换乘站、停车场以及气象中心等地的传感器和传输设备，向交通信息中心提供各地的实时交通数据。交通信息服务系统获得这些数据，经过处理后，实时向交通参与者提供道路交通、公共交通、换乘、交通气象、停车场等出行相关信息，并能根据车辆目的地、行驶习惯、路面情况推荐行驶路线。出行者可以根据这些信息确定出行方式和路线。

（2）交通管理系统

交通管理系统主要提供给交通管理者使用，用于检测、控制和管理公路交通，在道路、车辆和驾驶员之间提供通信联系。它与交通信息服务系统共用信息采集、处理和传输系统。交通管理系统对道路系统中的交通状况、交通事故、气象状况和交通环境等进行监视，获得实时交通数据，利用大数据技术辅以智能研判，对交通进行优化调控，如优化红绿灯配时，实时发布诱导信息，及时进行道路管制、事故处理与救援等。

（3）电子收费系统

电子收费系统通过安装在车辆挡风玻璃上的车载器与收费站电子收费系统车道上的微波天线之间的微波专用短程通信，利用计算机联网技术与银行进行后台结算处理，使车辆通过路桥收费站时不需要停车即可交费。同时所交纳的费用经过后台处理后直接清分给相关的收益业主。在现有的车道上安装电子收费系统，可以使车道的通行能力大大提高。

问题与讨论

智能交通为人们的出行提供了哪些便利？

4.3.2 电子商务

电子商务企业利用电子设备和网络技术进行商务活动。大型电商企业拥有大量用户数据，同时，在交易、营销、供应链、仓储、配送和售后等环节也产生了大量数据。这些数据通过电商企业的数据平台，为其电子商务平台上的商户和客户提供精准营销、供应链管理、智能网站等多种数据服务。

（1）精准营销

精准营销基于用户购买行为的大数据，使用推荐算法深度挖掘出用户的行为偏好，智能地向用户展示符合其兴趣偏好和购买意图的商品，实现个性化推荐，帮助用户快速地找到所需商品，提高网购效率。精准营销的主要方式是网站推荐、短信等。

（2）供应链管理

在仓储管理中，根据商品的销售情况和市场预期数据，依靠预测模型，在库存量达到某一个阈值时自动生成订单发给供货商，实现了商品自动补货。在物流配送领域，供应链管理通过分析物流人员、仓库以及用户之间的地理关系数据，为物流人员提供最优配送路径，提高配送速度，提升用户体验。

（3）智能网站

基于大数据挖掘和分析，网站变得越来越智慧。例如，牙膏等商品具有被重复购买的特点，购买之后会在可预期的一段时间内用完。通过分析用户两次购买此类商品的平均时间，在下一次购买时间到来之前，推荐系统向用户推介相应的商品，提升用户的体验，提高商品的转化率。

实践与体验

出租车轨迹可视化分析

采集本市出租车的运行数据，示例数据格式如图4.3.2所示。

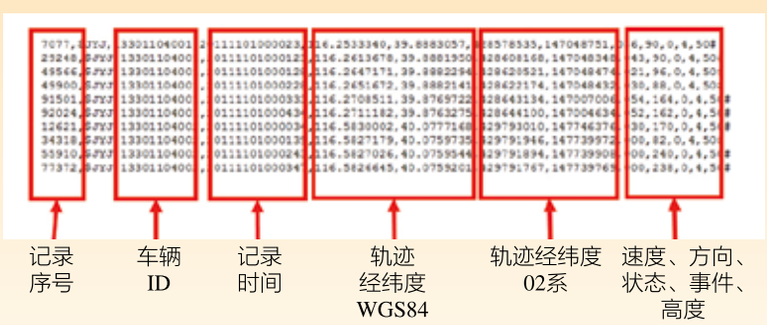


图4.3.2　出租车轨迹数据示例

实践内容：

对采集到的出租车轨迹数据进行可视化分析。

实践步骤：

1. 分析数据。

出租车轨迹数据采用TXT格式文件组织和存储，以UTF-8格式编码，内容包含“记录序号”“车辆ID”“记录时间”“轨迹经纬度WGS84”“轨迹经纬度02系”

“速度、方向、状态、事件、高度”。本次数据分析的目的是对轨迹数据进行可视化，并进一步分析、挖掘可视化轨迹所体现的规律和特征。

2. 编写程序。

本次分析采用Python编程完成，出租车轨迹数据的可视化程序如下：

```
import matplotlib.pyplot as plt

def plot_file(file):                          #绘制每个文件的GPS坐标轨迹
    jd=[]                                     #经度
    wd=[]                                     #纬度
    for line in open(file):
        #切分行数据
        splitline=line.split(',')
        #取轨迹坐标
        x = float(splitline[4])
        y = float(splitline[5])
        jd.append(x)
        wd.append(y)
    plt.plot(jd,wd)                           #画点

filename='xyz.txt'
plot_file(filename)
plt.show()
```

结果呈现：

1. 运行上述Python程序，观察程序生成的出租车轨迹图，说说你的发现，并尝试分析其背后的原因。

2. 进一步理解出租车轨迹数据，做一次你感兴趣的研究分析，如各时间段出租车的速度、载客与不载客时出租车的速度比较等。

3. 截取该市地图并保存为图片，尝试将运行轨迹显示在该图片上。

思考与练习

列举3种以上智能交通中用于采集交通数据的设备，指出它们可以采集的数据。

巩固与提高

1. 2017年1～6月全国旅客运输量数据（单位：万人）如图4.3.3所示。计算各月份旅客运输量的总数，以及不同运输方式在1～6月运输旅客人数的平均值、最大值、最小值；创建图表，分析铁路、公路连续6个月的旅客运输量变化情况。

	A	B	C	D	E	F
1	时间	铁路	公路	水运	民航	合计
2	2017年1月	24756	127396	1751	4393	
3	2017年2月	25525	137718	2321	4279	
4	2017年3月	22624	121678	2057	4431	
5	2017年4月	26504	118528	2367	4402	
6	2017年5月	26397	121334	2580	4498	
7	2017年6月	24077	116395	2482	4374	
8	平均					
9	最大值					
10	最小值					

图4.3.3　全国旅客运输量数据集示例

2. 简述Hadoop的组成和基本功能。

3. 现有50亿个32位正整数存储在文本文件intfile.txt中，每行1个数字。若内存限制为4GB，如何采用分治思想找出这些数的中位数？写出你的思路。

4. 某市普通高中选课数据如图4.3.4所示，学生从地理、化学、生物等科目中选择三门作为高考选考科目，“1”表示已选择的选考科目。使用Python编程分析每所学校各科目选考的总人数、全市各科选考总人数及其占比。

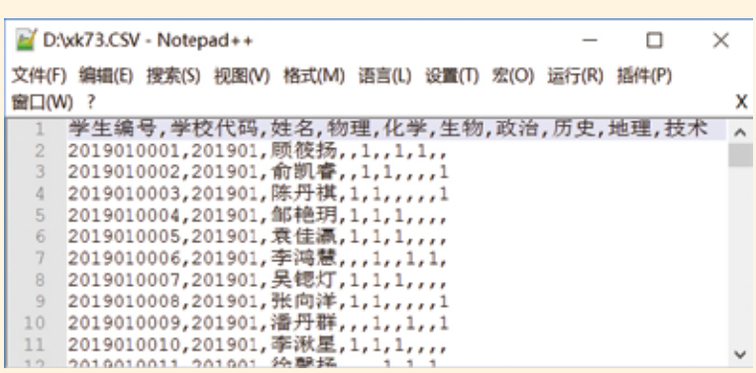


图4.3.4　某市普通高中选课数据集

5. 简述文本数据处理的一般过程。

6. 简述数据可视化的作用，并通过实例进行说明。

项目挑战

助力公益，用数据普惠民生

农副产品价格波动，一方面对种植户、养殖户、经营者的收入和积极性产生直接影响，另一方面又关乎百姓的日常生活和切身利益，是社会广泛关注的问题。最近，某志愿者团队发起了一项关注民生的公益活动，希望通过及时关注各地农副产品的价格数据，分析并发现价格变化规律，动态监测价格波动，以便为政府的精准决策与及时干预提供依据。现在，你应邀加入这个团队，希望你能通过这个活动，为社会尽一份力。

项目任务

根据本地政府公布的农副产品价格数据，发现其变化规律并做出未来预测。具体要求如下：

1. 了解本地农副产品价格的总体情况。
2. 分析近期农副产品价格的涨跌情况，探讨农副产品价格涨跌与节假日的关系。
3. 形成可以动态监测价格波动的可视化图表以及支持可视化的数据计算模型。
4. 基于近期价格及多方面的影响因素，形成分析报告，在相关农副产品的生产、流通和居民消费引导等方面向相关管理部门提出合理建议。

过程与建议

为了顺利开展本项目的研究，建议你组建研究小组，在充分理解活动要求的基础上，确定你要研究的地域范围和主要目标商品，然后分工协作，共同开展本次研究。

1. 收集数据

根据你确定的地域范围和主要目标商品，选择可靠的数据源采集目标数据（注：各省官方定期发布包括农副产品价格在内的民生数据）。

2. 数据整理

对采集到的数据，确定满足研究需求的数据结构并在此基础上采用恰当的方法处理数据。在确定数据结构时，可以考虑但不限于如下数据信息：

（1）每周各市场农副产品的最高价格、最低价格、平均价格等。

（2）近期各市场中主要农副产品平均价格的环比增幅、累计增幅等。

3. 数据分析与可视化

选择合适的工具软件分析数据，发现价格数据中隐含的信息和规律。

一般来讲，对数据进行可视化可以有效地发现数据中的规律。如对主要农副产品的环比增幅可视化，可以直观地分析出是否出现“多连涨”“多连降”的产品；对近期市场中主要农副产品的累计增幅进行可视化，可以直观地分析出是否出现价格大幅涨跌的产品。

思考一下，如果你是一个监控农副产品价格的管理者，你最想通过可视化图表得到哪些信息？请依据现有数据将这些可视化图表（如四张图）制作出来，并给出支撑每个图表的数据计算模型（注：以原始数据为起点来呈现计算模型）。

4. 分析其他关联因素（可选）

调用与上述调查的时间和地域范围相一致的天气预报信息（或其他你觉得有关联的数据），将其与农副产品价格的数据信息进行关联，深入分析价格波动背后的原因（如季节、高温、台风、霜冻等）。

5. 撰写分析报告

基于以上工作，撰写分析报告，其中应包括如下内容：

（1）研究背景与目标。

（2）研究过程。

（3）数据分析。

（4）可视化图表及数据计算模型。

（5）其他关联要求分析（可选）。

（6）结论与建议。

6. 交流分析报告

在全班范围内交流你们小组的分析报告，如果报告的价值得到教师和同学的认可，可以考虑通过一定的途径推广分析报告。

评价标准

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

评价条目	说明	评分（1～10分）	评分主要依据阐述	后续完善方向
项目理解	从项目开展和分析报告中展现出对项目目标、内容与任务的正确理解			
小组协作	小组分工合理，协作紧密，合作有成效			
数据采集	数据来源可靠，数据获取方法合理			
数据分析	数据分析所采用的方法恰当，能够有效地回应本项目任务的关切			
数据可视化	所呈现的可视化图表对于管理者能够起到支持决策的作用；支持可视化图表的数据计算模型科学，未来可利用原始数据进行动态计算			
分析报告	分析报告包含所有要求内容，内容详实，结论与建议有价值			
展示交流	受众分析正确，展示经过精心准备，表达清晰，便于受众理解			

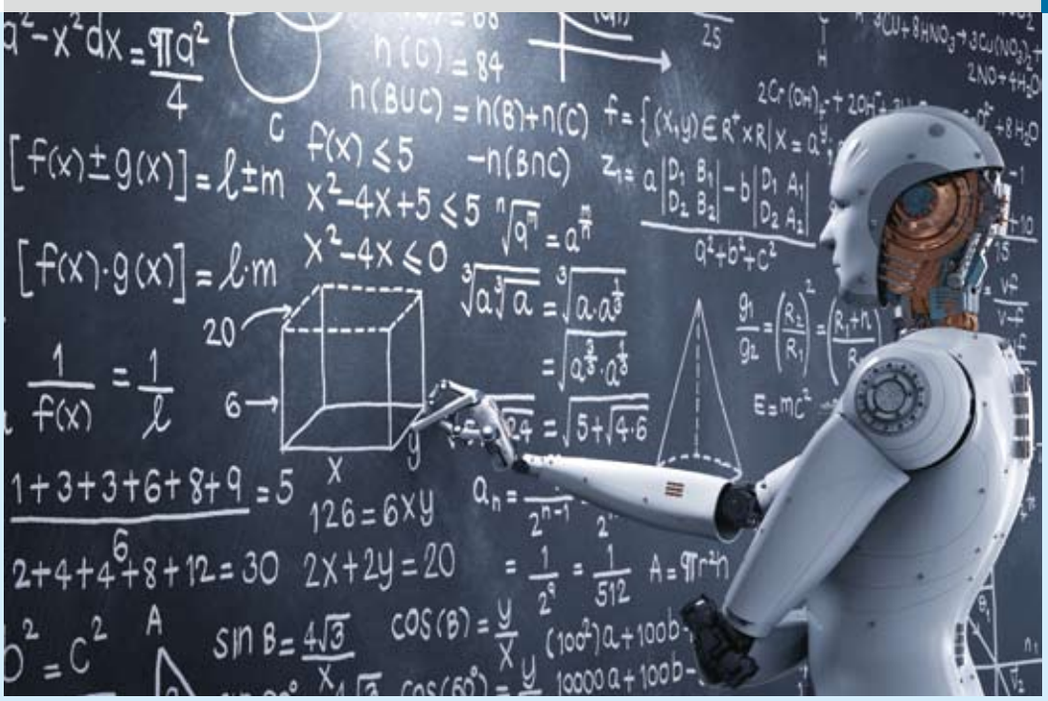
拓展项目

1. 某学校开设《中学生职业生涯规划》课程，引导、帮助学生规划自己的学习和职业生涯。在课程开设过程中，要求学生根据自己的个性特征、兴趣爱好、职业倾向等进行模拟选课，即从思想政治、历史、地理、物理、化学、生物等科目中，选择三门作为高考选考科目。选课前，同学们需要了解哪些大学院校开设了自己喜欢的专业，这些专业又指定了哪些选考科目等信息。请和同学一起收集相关数据，通过数据分析，结合自身实际，进行一场模拟选课之旅。

2. 环境为人类的生存和发展提供了必需的资源和条件。随着社会经济的发展，环境污染、生态恶化等环境问题日益突出，关注环境问题、保护环境关乎每个人。浏览中华人民共和国环境保护部网站，了解空气、水等环境状态。根据你的关注方向，查看网站数据中心的有关数据，理解数据特征、确定分析的角度，收集、整理、分析数据，并完成一篇分析报告。

第五章 5

人工智能及应用



人工智能（Artificial Intelligence，简称AI）是以机器为载体所展示出来的智能。1955年，人工智能就按照模仿人类部分功能这一目标开始了其漫长征途，如研制能够进行定理证明的算法来模仿解题者、研制机器翻译系统来模仿翻译家等。在这个过程中，形成了符号主义人工智能、联结主义人工智能和行为主义人工智能等代表性方法。

目前，人工智能进入了又一个发展高峰期，越来越深入地影响着人们的学习、工作和生活，如人工智能在象棋和围棋等博弈领域战胜人类选手，通过人工智能来识别人脸从而完成商品支付或车票查验等。人工智能在以技术手段改变人类生活的同时，也引发了一些社会问题，如机器取代了人类部分工作、个人隐私被泄露等。因此，在科学理性地利用人工智能造福人类时，也必须考虑如何应对人工智能应用可能带来的个人和公共安全、法律伦理和世界和平等新的挑战。

问题与挑战

- 2018年1月5日，阿里巴巴研制的人工智能算法在机器阅读理解竞赛（SQuAD machine reading comprehension challenge）中战胜了人类选手。在这个竞赛中，要求人工智能程序和人类选手在阅读完文档后，根据文档内容来回答与该文档相关的问题，根据回答问题的准确度来评判人工智能算法和人类选手的理解能力。机器在此比赛中战胜了人类选手，是否可以认为机器的阅读能力超越了人类？

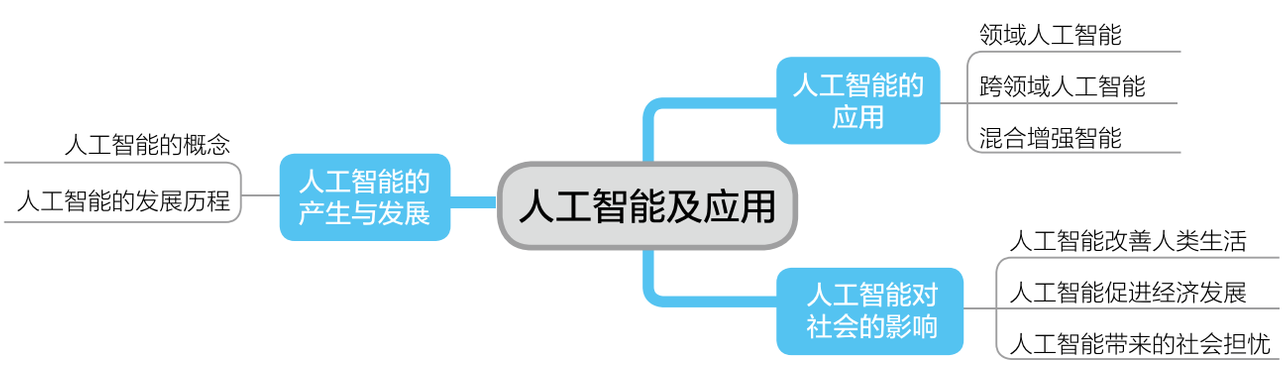
- 无人驾驶汽车是人工智能的一个重要研究方向。无人驾驶汽车把驾驶员从驾驶座上解放出来的同时，也引起了人们的担忧。如，公交车、出租车、运输车等司机将失业；如果无人驾驶汽车突然出现故障，发生交通事故，责任该如何界定等。那么，无人驾驶汽车应不应该推广呢？

- 人工智能会代替人类从事一些危险、重复的工作，如火山口地质探测和汽车装配等。但是，在相当长的时间内，机器智能和人类智能将会一起完成机器或人类无法单独完成的任务，如达芬奇外科手术机器人和人机混驾系统。那么，哪些工作在机器智能和人类智能相互结合的情况下，其效率会进一步提升？

学习目标

1. 理解人工智能的概念，能结合实例辨别人工智能技术。
2. 了解人工智能产生及其发展历史，能初步辩证地看待人工智能的发展。
3. 了解人工智能在各个领域的应用，感受人工智能对促进社会发展的巨大作用。
4. 能科学地看待人工智能及其应用，认识到人工智能与人类和谐共处的必要性。

内容总览



5.1 人工智能的产生与发展

近年来，人工智能已经深刻而广泛地影响着人们的生活。从刷脸支付到交通出行预测，从扫地机器人到无人飞机，从“深蓝”到AlphaGo，都渗透着人工智能的创新应用。人们依靠智能导航出行，通过语音与机器互动，应用智能工具搜索知识信息……已自觉或不自觉地处于人工智能的环境中。

5.1.1 人工智能的概念

在历史的发展过程中，人类依靠自身的智慧，发明和创造了许多机器，使得人类从繁重的体力和脑力劳动中解放出来。利用机器，人们可以上天入海，可以遨游宇宙，也可以窥探分子世界，还可以“智能”地做很多事情。人类从未放弃过对人工智能的追求与探索。

所谓人工智能，是指以机器（计算机）为载体，模仿、延伸和扩展人类智能，其与人类或其他动物所呈现的生物智能有着重要区别。

人工智能作为一门多学科广泛交叉的前沿科学，不仅涉及计算机科学，还涉及控制科学、认知科学、心理科学、脑及神经科学、生命科学、语言学、逻辑学、行为科学、教育科学、数理科学等众多学科领域，其学科结构如图5.1.1所示。

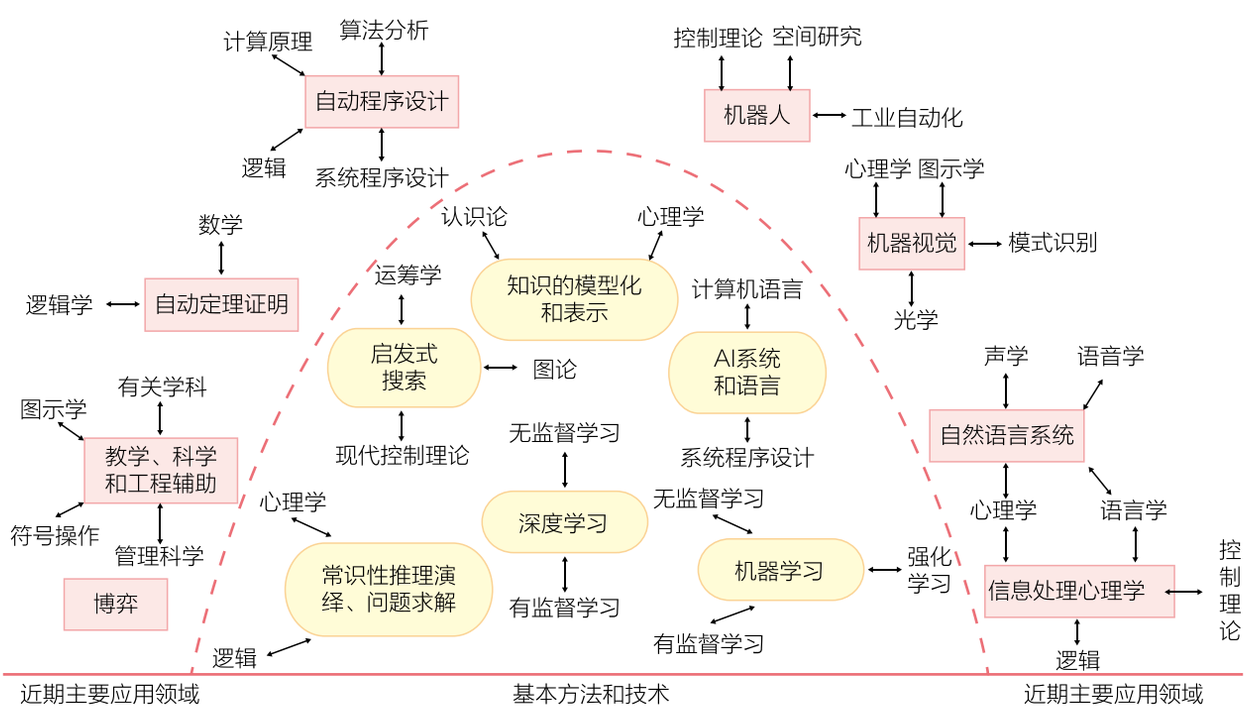


图5.1.1 人工智能学科结构

在长期的研究过程中，不同专家学者通过从不同方面来实现人类智能，形成了人工智能的三种主要方法，即符号主义、联结主义和行为主义。

符号主义（Symbolicism），又称逻辑主义、心理学派或计算机学派，认为学习或者其他的智能特征原则上均可以被符号精确地描述，从而被机器仿真。因此，在符号主义人工智能中，智能行为就是对符号的推理和运算。在这种方法中，每个符号反映了其在客观世界中的语义，如IsCar（A）表示A这个符号为“小车”或者“不为小车”。但是，人类某些语义并不能被精确描述，如“仁义”“微笑”等语义就很难直接用符号来描述。

联结主义（Connectionism），又称仿生学派或生理学派，通过模仿人类大脑中神经元之间的复杂交互来进行认知推理。在这种方法中，需要从海量数据出发，学习神经网络中成千上万的神经元之间的关联关系，这些关联关系通过神经元之间的链接权重来刻画。如果两个神经元之间的权重很大，表示输入端数据激活了这两个神经元。于是，所有被激活的神经元以逐层递进的形式来一起刻画数据中所蕴含的模式（即概念或知识）。目前性能表现优越的深度学习就是这种学习方法的典型代表。

行为主义（Actionism），又称进化主义或控制论学派，这一方法从“交互—反馈”角度来刻画智能行为，认为智能体可以在与环境的交互中不断学习，从而提升自己的智能水平。如将一台扫地机放入一个会场，其事先并不知道会场中桌椅的摆放形式。于是，扫地机在运动中不断从环境中学习，如墙壁挡路则避让、桌椅空隙过窄难以通过则后退等，经过一段时间的交互，扫地机就通过学习获悉了环境的全貌，从而提升自身智能水平而自如地执行清扫任务。

毫无疑问，对人工智能的研究将使计算机呈现出更为高级的“智能”，并对人类社会的发展产生深远的影响。

问题与讨论

请列举通过机器为载体所实现的人工智能与人类智能的不同之处。

5.1.2 人工智能的发展历程

早在19世纪中期，科学家就萌发了让机器进行自动计算的思想。如1842年，艾达·洛夫莱斯（Ada Lovelace）为当时的一台分析机编写了历史上第一个计算机程序，该程序可自动计算伯努利数（Bernoulli number）。但是，她指出“分析机不能自命不凡，以为自己无论什么问题都能解决。其实它只能完成我们告诉它如何做的事情”。实际上，这道出了大多数“智能系统”只能“机械执行预先设定指令”，无法完成超越指令所指定的任务的实质。

人工智能自1955年登上历史舞台后，在视觉计算、语音识别、机器翻译、问答助理、商品推荐和无人系统等领域蓬勃发展。

1. 从计算到智能测试

20世纪初，人们发现有许多问题经过长期研究，仍然找不到有效的算法。如，无法以清晰的步骤一步一步地证明或证伪著名的费马定理（即当n大于2时，关于x，y，z的方程$x^n+y^n=z^n$没有正整数解）。于是人们开始怀疑，是否对某些问题来说，根本就不存在算法，即这些问题是不可计算的。人们开始思考，计算的本质是什么，如何去定义计算。

20世纪30年代，三种计算机制相继被提出，它们分别是原始递归函数、lambda演算和图灵机。已经证明，这三种计算机制在性能上是等效的，即任何一种计算机制所能完成的计算任务均可被其他两种计算机制同样完成；另一方面，如果某一计算任务不能被某一计算机制完成，那么这一任务也无法被其他两种计算机制完成。

由于图灵机可以通过最简单、最基本和最确定的方法，一步一步机械地完成计算任务（如图5.1.2），图灵机成了现代计算机的理论模型，而其发明人阿兰·图灵也被誉为“现代计算机理论之父”，计算机界的最高奖“图灵奖”就是用其名字来冠名的。

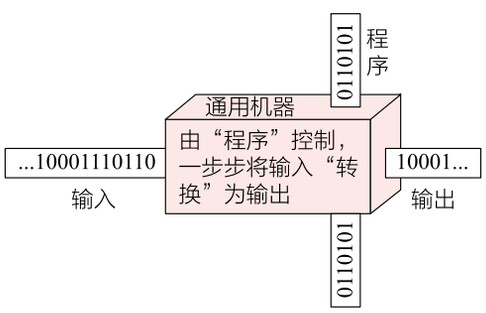


图5.1.2　图灵机模型

由于三种计算机制是等效的，因此我们可定义所谓的可计算任务就是图灵可停机任务，即这一任务可通过图灵机一步一步处理，在处理结束（即程序停机）时输出计算结果。图灵在1937年发表的论文《论可计算数及其在判定问题中的应用》，证明了“不可计算数”的存在，即存在一些“不可计算（不可停机）”的任务。

计算机的诞生为通过机器来模拟人类智能提供了无限的想象空间，促进了人工智能的发展。图灵对“智能测试”这一问题也进行了深入思考，1950年提出了著名的“图灵测试”（Turing Test）。

图灵测试是测试机器是否具有智能的一种方法。在这个测试中，一台机器和一个人被安排在两个彼此隔离的房间中。假设一名法官给机器和人出了10道题目（如人生的意义是什么、节假日中城市景区交通如何等）。如果我们把机器和人的答案收集起来，法官无法区分哪个答案是机器回答、哪个答案是人回答的，那么图灵测试认为机器具有了智能。虽然在2014年图灵测试竞赛上，一个智能程序通过了图灵测试，但是有科学家认为这种测试并不能真正评估人工智能是否具有人类心智，因此又提出了“视觉图灵测试”等概念。

2. 人工智能登上历史舞台

1955年8月，约翰·麦卡锡（John McCarthy）、马文·明斯基（Marvin Lee Minsky）、

香农以及纳撒尼尔·罗切斯特（Nathaniel Rochester）四位学者联名向美国洛克菲勒基金会提交了一份名为“人工智能达特茅斯夏季研讨会”的项目申请书，在这份申请书中首次提出了“人工智能”的术语，从此AI踏入了人类历史长河。这份项目申请书中指出“人工智能”的研究目标是实现能模拟人类的机器，机器能使用语言，具有概念抽象和理解能力，完成人类自身才能完成的任务，并且不断提高自身的各种能力。申请书中同时希望洛克菲勒基金会能够于1956年夏季资助十余人在美国达特茅斯学院工作两个月，研究人工智能领域面临的七个问题：自动计算机、机器编程、基于神经网络的概念理解、计算复杂性、自我学习与提高、抽象能力、直觉能力。

1956年，这个研讨会在美国达特茅斯学院如期召开，这标志着人工智能作为一门新兴学科正式诞生。人工智能从其诞生之日起，开始了漫长的征途，逐渐形成了符号主义人工智能、联结主义人工智能和行为主义人工智能等代表性方法。

3. 以符号主义表达与推理为代表的人工智能

符号主义人工智能方法认为学习或者其他的智能特征原则上都可以被精确地描述（一般以逻辑形式描述），其包含知识库和推理引擎两个部分。在这种方法中，先要将所有知识以逻辑形式表达，然后依靠推理引擎，去验证命题或谓语正确与否，或者学习推导出新规则、新知识。如IBM研制的“沃森”（Watson）和卡耐基梅隆大学研制的NELL（Never-Ending Language Learning，永不停息的语言学习）。

1965年，在斯坦福大学化学专家的配合下，爱德华·费根鲍姆（Edward Feigenbaum）成功研制了第一个专家系统DENDRAL。DENDRAL是化学领域的一位“专家”，在输入化学分子式和质谱图等信息后，通过分析推理来判断有机化合物的分子结构，其分析能力已经接近甚至超过了有关化学专家的水平。DENDRAL专家系统为早期人工智能的发展树立了典范，虽然它只是在实验室内试验，但其意义远远超出了它在实用中创造的价值。

1976年，斯坦福大学的肖特列夫（Shortliff）开发了医学专家系统MYCIN（这个系统在知识工程领域被视为“专家系统的设计规范”）。在MYCIN的知识库里，存放着约450条判别规则和1000条细菌感染方面的医学知识。MYCIN通过文字形式一边与患者对话，一边进行病情诊断。它通过分支结构规则进行病情诊断，显示患者可能性最高的病因，并给出用药建议。

1977年，费根鲍姆在第五届国际人工智能大会上提出了“知识工程”的概念，为规则驱动或知识驱动的人工智能指明了方向。所谓知识工程，即尽可能对人类知识进行逻辑编码，然后通过推理引擎对编码知识进行操作，形成某一领域的“专家系统”。

在这种方法中，知识的精确化编码是阻碍符号主义人工智能发展的一个瓶颈问题。如推理引擎会从“所有的鸟都会飞”和“鸵鸟是鸟”这两条知识出发，推导出“鸵鸟会飞”这条错误知识。造成这个结果的原因是“所有的鸟都会飞”这条知识表述得不严密和不精确。人类很多知识（如视觉知识或灵感等）是无法通过符号编码的，如很难想象如何用符号来表达“仓廪实而知礼节”所蕴含的丰富内涵。

基于规则学习的人工智能方法解释性强（与人类逻辑推理过程相符），但其可拓展性较弱，难以构建完备的知识库和完善的推理方法。

4. 数据驱动的人工智能方法

“手工构造知识库+推理引擎”的“专家系统”虽然能够帮助人们解决一些实际问题，但随着需要解决的问题越来越复杂，这种方法需要手工构造越来越多的知识和规则，导致知识库越来越庞大且无比复杂，维护知识库变得越发困难（如难以增加新知识或难以删除不合适的知识点），同时使得专家系统本身也越发笨拙。有计算机科学家就指出“如果机器人掌握了除学习以外的所有能力，人类很快就会抛弃它”，这句话很好地说明了“专家系统”的局限性，也为人工智能的进一步发展提供了启示，即如何让计算机能从数据本身进行知识学习，而不是单纯依赖专家来手工构造知识。

在这一方面，深度学习成为数据驱动人工智能方法的佼佼者。深度学习是一种对原始数据所蕴含的特征模式进行学习的算法模型。最常用的深度学习模拟人类大脑处理数据的机制，逐层抽象对原始数据进行学习。多层神经网络（包含输入端、隐藏层和输出端）是一种典型的深度学习模型，如图5.1.3所示。

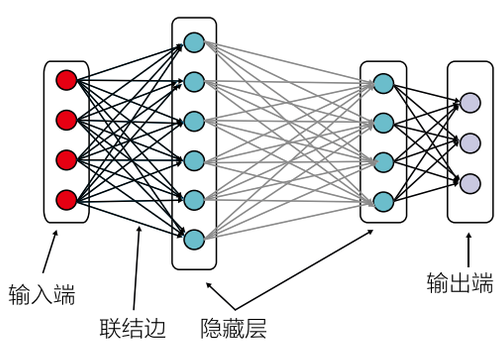


图5.1.3　多层神经网络

在深度学习中，一旦给定标注数据，根据“端到端（即输入端到输出端）”机制从数据出发，即可学习数据中蕴含的概念或模式，而不需要事先手工构造知识库。例如，给定数万张人脸图像，深度学习从这些人脸图像出发，学习挖掘人脸图像中像素点的空间分布模式，然后就能基于学习得到的像素点空间分布模式来进行人脸识别。在这种方法中，不用事先定义人脸中“左右眼睛对称、眼睛在鼻子上方”等知识，而是直接从数据出发，从数据中学习隐性知识或隐含模式，来指导原始数据的识别和分类。这一数据驱动学习模式不但将我们从手工定义知识的烦冗工作中解放出来，而且解决了某些视觉形象难以通过符号文字来定义的难题。

在深度学习中，一般会构造包含若干层的神经网络，每一层中有若干神经元，前后相邻层中的神经元彼此联结。一旦给定海量数据，就可以学习神经元之间的链接权重。赫布理论（Hebbian theory）指出，神经元与神经元之间的链接权重会在持续重复的刺激

下增加。因此，可将神经网络中神经元之间的链接理解为一种“记忆”，即针对数据中所蕴含的知识而言，神经网络“记忆”了这种知识的模式。

目前，深度学习这一数据驱动方法在自然语言处理、知识图谱构建、图像分类、语音识别和视频运动提取等领域表现出良好的性能，如微软公司研究人员通过152层的“深层残差卷积网络”在2015年ImageNet计算机视觉识别挑战赛中取得了96.43%的图像分类正确率；谷歌公司研发的围棋软件AlphaGo从人类选手棋局中利用深度神经网络学习（由策略网络和价值网络构成）和蒙特卡洛树搜索，初步具备了下棋能力，然后再结合强化学习来进一步提升棋力，于2016年3月战胜了围棋九段棋手李世石。

5. 问题引导下的人工智能学习方法

人工智能典型方法中还存在另外一种学习方式，即问题引导下的试错学习。在这种学习方法中，学习者事先不知道最终答案，而是在学习过程中不断尝试各种解决问题的可能途径，然后根据结果反馈来调整相应的学习方法，这一学习机制叫强化学习。强化学习体现了一种自我学习的能力，即从过去的经验中不断学习，提升能力。如围棋人工智能系统AlphaGo Zero不依赖人类棋手数据而在自我博弈中不断提升棋力，卡耐基梅隆大学研制的德州扑克人工智能Libratus则在与人类选手博弈中不断提高牌技。

问题与讨论

东汉马融在《围棋赋》中说：“三尺之局兮，为战斗场。”可见围棋是一个模拟战场决策的博弈竞智。AlphaGo击败了围棋世界冠军，请同学们讨论：是否可以将AlphaGo的算法直接应用于现实世界中复杂的战场博弈？

实践与体验

利用神经网络解决分类问题

分类问题是人工智能领域的经典问题，它在日常生活中具有十分广泛的应用，例如天气预测、新闻分类、邮件分类、性格测试、指纹识别等。如气象学家通过测定气温、湿度等指标，预测风、晴、雨、雪等天气情况；医生通过观察病人的症状，判断病人的患病类型。

在人工智能领域，解决分类问题的方法有很多，其中神经网络是一种应用非常广泛的人工智能方法。一个解决特定分类问题的神经网络通常以特定的数据特征作为输入，选择适当的分类方法输出分类结果。以天气预报为例，气温、湿

度、风力以及前一天的天气状况都可以作为预测未来天气的数据特征，而最终预测的天气状况就是分类的结果。

在神经网络中，不同的网络结构适用于不同的分类问题。神经网络的深度与宽度是决定网络结构的重要因素。网络的深度指网络的层数，网络的宽度指每层中的隐节点个数。下面将借助Tensorflow Playground平台提供的功能，直观体验如何构造一个神经网络对不同颜色的数据进行分类（同一种颜色数据可视为归属一个类别）。

实践内容：

1. 利用Tensorflow Playground平台，了解神经网络的基本要素。
2. 构造神经网络，解决不同复杂度的分类问题。

实践步骤：

1. 了解神经网络的基本构造。

登录Tensorflow Playground网站，如图5.1.4所示，该平台包含了数据、网络结构、训练控制以及输出效果四大块内容。

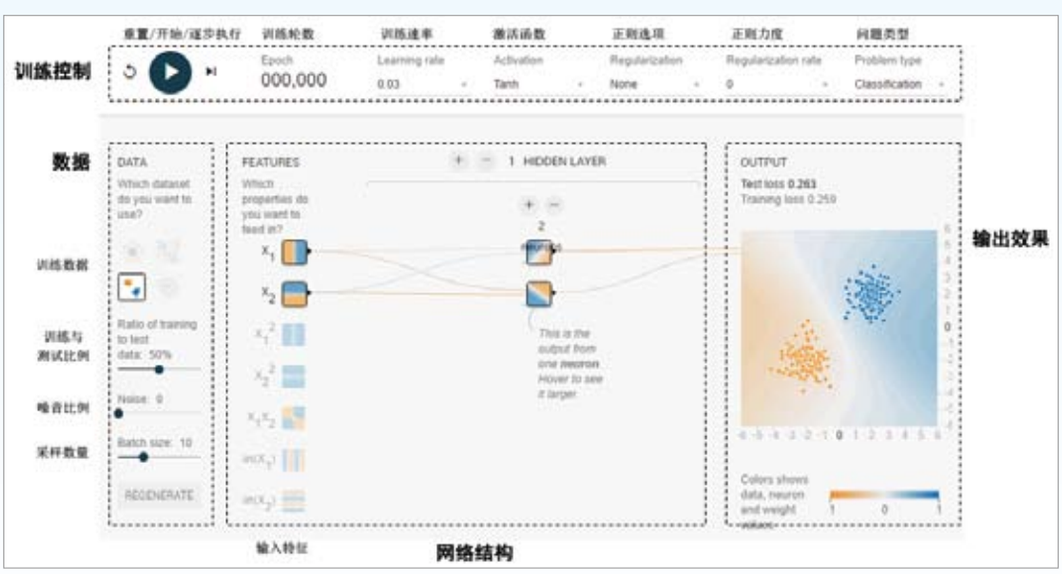


图5.1.4　Tensorflow Playground神经网络实验平台

（1）设置数据。平台提供了四组“训练数据”（训练数据由一团蓝色和一团橘色的数据点组成）。如图5.1.4所示，先选择第三组数据，然后选择“训练与测试比例”为50%（50%的数据将用于训练，50%的数据将用于测试），将“噪音比例”设为0（为了把两团数据按照颜色完美地区分开），将“采样数量”设为10（神经网络训练过程中每次采样的数据量为10）。完成设置后，就可以从“输出效果”中观察到目前数据在平面上的分布情况。

（2）设置网络结构。平台中包含了两个维度（x_1, x_2）的“输入特征”可供选择，在实际的人工智能场景当中，可供使用的“输入特征”远不止这些。如图5.1.5所示，选择网络的“输入特征”包含x_1与x_2两种特征，通过“+/–”调节网络结构，使用一个包含2个隐节点的隐藏层。

（3）设置训练控制。一个神经网络需要经过训练才能够满足特定的需求。如图5.1.4所示，选择“训练速率”为0.03、“激活函数”为Tanh、“正则选项”为None、“正则力度”为0、“问题类型”为Classification，然后点击“开始”按钮开始网络训练。

（4）观察效果输出。一个神经网络应用于某个特定问题的误差值是评价这个网络的重要指标。在训练过程中，随着“训练轮数”的不断增加，可以看到输出误差快速地从0.5下降到0.0，同时数据点被准确分类到蓝色与橘色两个区域当中。

完成训练后，点击“重置”按钮，使用“逐步执行”观察网络输出的变化，并记录误差值减小到0.0时所经过的“训练轮数”。

2. 准确构造神经网络，解决不同复杂度的分类问题。

在上一个实践环节中，通过在平面上找到一条合适的直线，将两种颜色的点准确分开。接下来尝试解决稍微复杂的问题。

（1）选择“训练数据”中的第一类数据，如何修改已经构造的神经网络使其能够将第一类数据准确区分？

可以在“网络结构”中添加x_1^2与x_2^2两种特征，再重新训练神经网络，就可以将数据准确区分，如图5.1.5所示。

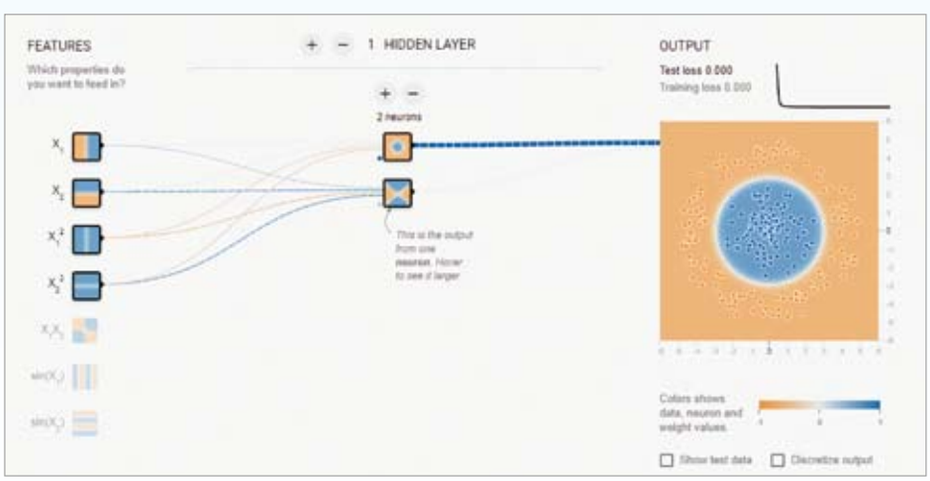


图5.1.5　第一类数据的分类网络

（2）如何区分“训练数据”中的第二类数据？

可以使用x_1和x_2两个“输入特征”。通过“+/–”将网络的深度扩展为3层隐藏层，其中第一层设置4个隐节点，第二层设置3个隐节点，第三层设置2个隐

节点，然后重新开始训练，如图5.1.6所示。

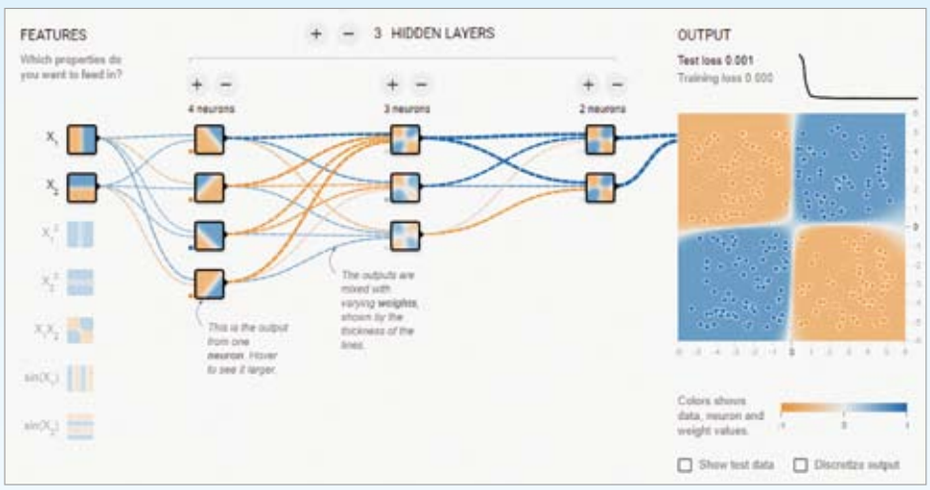


图5.1.6　第二类数据的分类网络

（3）如何解决比较复杂的分类问题？

“训练数据”中的第四类数据，已经无法使用简单的几条直线或者规则的圆形区域进行分类，构造什么样的神经网络才能将这些数据准确区分？

结果呈现：

完成上述实践任务，呈现结果并思考以下问题。

1. 针对“训练数据”中的第一类数据，如果构造其他“输入特征”进行训练能否得到相同的效果？

2. 修改如图5.1.6所示的第二类数据分类网络中的网络结构，分析不同网络结构对分类性能造成的影响。

3. 使用不同的训练速率，如将“训练速率”修改为3或0.003，然后重新训练网络，得到的训练结果以及训练过程所需要的轮数是否相同？

4. 综合上述影响神经网络分类性能的因素，尝试构造出能够解决“训练数据”中第四类数据分类的神经网络。

思考与练习

1. 请举出若干个难以通过文字符号或逻辑来编码表达的人类知识的例子。

2. 请指出通过图灵测试来测试机器是否有“智能”的不足之处。

5.2 人工智能的应用

随着互联网的普及、传感网的渗透、大数据的涌现和信息社区的崛起，数据和信息在彼此融合的信息空间、物理空间和人类社会中传播，人类迈向“三元空间”（Cyber-Physical-Society，简称CPS），新技术、新产业和新业态不断涌现，使得人工智能迅速发展，在众多领域发挥着巨大的作用。

5.2.1 领域人工智能

依赖于领域知识和数据的人工智能被称为领域人工智能。这类机器具有强大的存储、记忆和搜索功能，因此，如果机器遍览某一领域的数据和知识，它就可以在这一特定领域表现出较强的能力，如专门用于下国际象棋的超级计算机“深蓝”（如图5.2.1）和用于人机对话的系统“沃森”（如图5.2.2），以及应用于交通领域预测的人工智能系统等。

“深蓝”存储了100多年来200多万局象棋棋谱，遍览国际象棋各种对弈棋局。1996年，“深蓝”首次挑战国际象棋世界冠军加里·卡斯帕罗夫，以失败告终。比赛结束后，研究小组对“深蓝”加以改良。1997年，“深蓝”在正常时限的比赛中首次击败了卡斯帕罗夫，机器的胜利开启了国际象棋历史的新时代。而人机对话系统“沃森”在问答竞赛中展示了较强能力，依靠其强大的记忆能力和快速搜索手段，在益智游戏中战胜了人类选手。

图5.2.1 超级计算机“深蓝”

图5.2.2 人工智能应用平台“沃森”

在智慧交通领域，人工智能（尤其是深度学习方法）在车牌识别上取得了较大成功。此外，人工智能在车辆颜色与车辆厂商标志识别、无牌车检测、非机动车检测与分类、车头车尾判断、车辆检索、人脸识别等方面的应用也比较成熟。

5.2.2 跨领域人工智能

跨领域人工智能指智能系统从一个领域快速跨越到另外一个领域。跨领域人工智能不仅依赖于已有数据和已有规则，而且专注于知识和技能的获取，能够举一反三、触类旁通，开展深度推理。

如谷歌公司的AlphaGo在围棋领域表现出了超越围棋选手的能力，其后，谷歌公司将AlphaGo使用的机器学习算法用于控制其数据中心风扇、制冷系统和窗户等120个变量，帮助谷歌公司在电力使用效率上提升了15%。作为一种跨领域人工智能的应用，AlphaGo从围棋人工智能跨界到电力控制领域。

再如，IBM将“沃森”的智能能力从益智游戏领域移植到了医疗领域。在医疗领域，“沃森”收录了肿瘤学研究领域的42种医学期刊、60多万条临床试验医疗证据和200万页文本资料。“沃森”可在几秒钟之内筛选出数十年癌症治疗历史中的150万份患者记录，包括病历和患者治疗结果，并为医生提供可选择的治疗方案。2016年，“沃森”在日本东京仅用了10秒钟，便对一名60多岁患罕见白血病的女性患者做出诊断并给出治疗方案。

跨领域人工智能研究难度较大，虽然人类擅长举一反三式的跨域学习，但是对人工智能算法而言，尚缺乏一条清晰推进跨领域人工智能的思路。专家们认为，在跨领域人工智能的研究过程中，需要从特殊技能到泛化技能、从单一知识到多源知识、从易到难，永不停息地学习。

5.2.3 混合增强智能

混合增强智能是多种智能体的混合形式，它将人的作用或人的认知模型引入人工智能系统，形成“混合增强智能”的形态。如人、机器、物联网和互联网的结合可形成智能城市这样复杂的智能系统。在混合增强智能中，不同智能发挥自己的长处，相互协调，形成了超越任何一种智能能力的增强智能。但是，需要注意的是，在智能叠加协调的回路中，人类智能是智能回路的总开关。

在医疗领域，因为医疗关系到人的生命健康，人们对错误决策的容忍度极低，人类疾病也很难用规则去穷举，不仅需要人类医生智能，也需要机器智能，因此需要发展人机交互的混合增强智能系统。如图5.2.3所示的达芬奇外科手术机器人，其设计的理念是通过使用微创的方法来实施复杂的外科手术。达芬奇外科手术机器人由外科医生控制台、床旁机械臂系统、成像系统三部分组成。人类医生坐在机器后面操纵机器人用灵巧的手臂完成高端、复杂的外科手术。在这个过程中，如果缺少了人类医生，又或者缺少了这种复杂、高端的从事临床外科手术的机械臂，都不能完成一个高难度

图5.2.3 达芬奇外科手术机器人

的手术任务。人与机器智能的结合使得在狭窄空间内的手术操作更加精确，在实际手术治疗中已得到广泛应用。

在电商平台上，人工智能机器客服与人类客服一起合作来回答顾客购物过程中出现的各种问题；通过用户手机搜索记录和位置移动的数据来感知城市中人群的流动，预测关键景点的拥堵情况等。这些都是人和机器共同参与的混合智能应用。

另外，在产业风险管理、刑事司法中应用人工智能系统时，也需要引入人类智能，在机器智能中以交互形式利用人的知识和智慧，最优地平衡人的智力和计算机的计算能力。人机协同的混合增强智能在军民领域中也发挥着重要作用，如人机协同的太空机器人、协同作战机器人，已成为人类智能的自然延伸和增强，并以人机群组协同的方式解决社会、经济发展面临的重大问题。

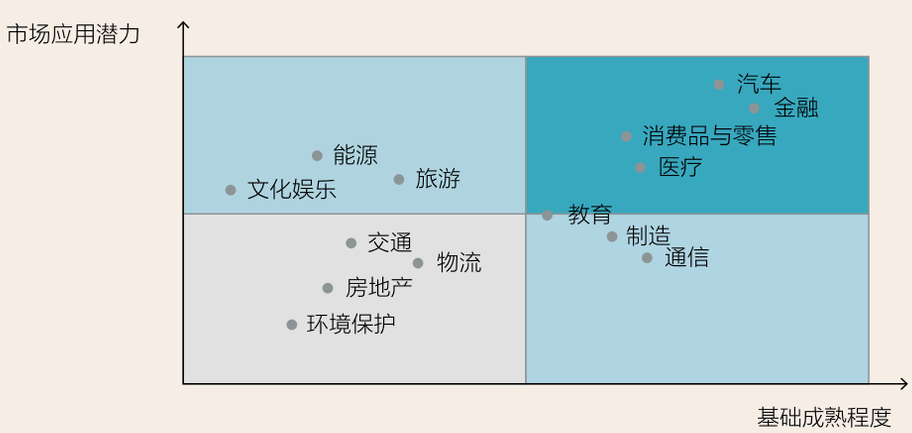
问题与讨论

在相当长的时间内，人类智能和机器智能将彼此协调、相互取长补短。为什么说人类智能是“智能回路”的总开关？

思考与练习

1. 人工智能已在社会各个领域得到应用，它对目前的教育行业会产生怎样的影响？

2. 某机构对各行业人工智能发展基础与市场应用潜力的评估如图5.2.4所示。结合图中各行业人工智能的应用情况，分析造成这种发展差异的原因。



资料来源：中国人工智能学会、罗兰贝格分析

图5.2.4　各行业人工智能发展基础与市场应用潜力

5.3 人工智能对社会的影响

随着人工智能技术的快速发展，各类智能应用和产品不断问世，改变着人类生活、促进着经济发展，同时也引发了一些新的社会问题。

5.3.1 人工智能改善人类生活

近年来，人工智能技术正快速融入人们的生活，使人们居家、出行、购物、医疗等日常活动越来越便捷。

智能家居。智能冰箱可以对用户膳食的合理性进行分析，自动统筹安排，推荐食谱，提示需要补充的食材，如果与生鲜电商联网，还可以自动采买食材；智能温控设备能够通过不断地观测、记录、学习用户习惯的舒适温度、湿度等，动态调整室温；智能安防能对人物暴力、跌倒等危险动作行为和越界、徘徊等可疑运动轨迹进行精确检测、分析，自动提供预警信号；智能扫地机器人（如图5.3.1）可以通过对障碍物、地形等环境的感知、记录，智能学习家中布局，合理规划路线，完成清扫、吸尘、擦地工作。未来，还可能出现人形机器人，来帮助人们处理日常家务，陪伴家人。这些智能设备使人们的家居生活变得更加安全、舒适、便捷，而且节能环保。

图5.3.1 智能扫地机器人

智慧城市。城市大脑参与城市管理，使人们的出行更通畅、安全，交通事故的救援更及时。阿里云ET城市大脑已经在杭州、苏州等部分城区落地。城市大脑融合数字地图、交警微波和视频数据感知交通事故，并触发相应机制进行智能处理。内核人工智能技术，可以对整个城市的全局进行实时分析，自动调配公共资源，修正城市运行中的错误与漏洞。杭州城市大脑在感知交通事故方面，准确率已达92%，120救护车到达现场的平均时间大大缩短；由城市大脑接管信号灯的路口，很大程度上减少了通行的时间。随着人工智能的发展，城市管理者将能够更加及时、精确地掌握城市运行脉搏，把握城市运行规律，实现更加精准的城市运行管理。

智能出行。未来，人们或可搭乘无人公交、无人汽车，让出行更加简单安全。在深圳福田保税区，搭载“阿尔法巴（Alphabus）智能驾驶公交系统”的电动公交车，已在固定线路上试运行。“阿尔法巴智能驾驶公交系统”通过工控机、整车控制器、CAN网络分析路况环境，实现自动驾驶下的行人与车辆检测、减速避让、紧急停车、障碍物绕行、变道等功能。此外，“阿尔法巴智能驾驶公交系统”还包含客流统计分析、智慧调度、安全评估与应急响应、智能充电、智能维保等子系统。

智能购物。通过对商品、场景、消费的数字化、智能化，人工智能技术正在催生无人超市、无人酒店、无人餐厅等。无人超市，消费者无须排队买单，可以自动结算，购物更为快捷、便利；无人酒店，将入住过程化繁为简，住客无须押金、无须登记身份证，通过人脸识别、扫码开门等，即可自行入住；无人餐厅，顾客扫码入座后可以自助点餐、取餐，吃完后自动扣款。

随着人工智能技术的发展及与各个行业的深度融合，生活中的人工智能将无处不在，人工智能必将为人们呈现一个安全、便捷的智能社会。

问题与讨论

2011年，在美国最受欢迎的智力问答节目《危险边缘》中，人机对话系统“沃森”一举打败了人类智力竞赛的冠军。“沃森”的智能每年都在提高，将来可能给人们的生活带来哪些改变？

5.3.2 人工智能促进经济发展

在国家层面，各国政府正在不遗余力地推进人工智能技术的发展。人工智能技术在经济建设以及国家战略层面的作用日益重要。

人工智能的发展可以为人类社会带来巨大的经济效益。人工智能作为全新的生产要素，创造了一种虚拟的劳动力，能够胜任需要适应性和敏捷性的复杂工作，从而提高实体经济运行的效率，降低生产成本，促进经济增长。互联网平台模式通过减少信息不对称，降低了传统经济活动中的交易成本。机器学习的引入，能够更准确、更快速地进行相关数据的分析和学习，实现更精准的服务匹配，优化资源分配，进一步降低经济活动中的交易成本。

通过人工智能技术提高生产力、创造全新的产品和服务，是经济竞争和升级的迫切需求。人工智能并不是单一的技术，它将融入现有的生产中，在垂直领域加深数字化的影响。国内很多制造企业纷纷向“智能化”转型。2017年，格力电器宣布向智能制造转型，结合人工智能技术，在智能装备领域坚持自主研发生产。其智能装备产品包括数控机床、工业机器人、伺服机械手、智能仓储装备等。与此同时，人工智能技术也带动了一系列的新兴产业，包括智能软件、智能机器人、智能运载工具、智能终端、物联网基础器件等。

推动人工智能与实体经济结合，是加快实体经济转型升级的必然发展方向。人工智能对传统产业的转型升级有着强大的驱动作用。例如，某煤业集团与百度公司进行战略合作，综合应用现代传感技术、工业物联网技术、自动化技术、云计算技术、大数据技术、智能化技术等先进技术，共同设计建造智能化工厂，实现复杂环境下生产运营的高效、节能和可持续发展，从而全面提高煤化工生产的安全性、环保性，推动煤炭工业转型升级。在各行业引入人工智能是一个渐进的过程，从最基础的感知能力，到对大数据的分析能力，再到理解与决策，人工智能将逐步改变各领域的生产方式，推进结构转型。

拓展链接

国家新一代人工智能开放创新平台

2017年7月，国务院发布了《新一代人工智能发展规划》，为我国在抢抓人工智能发展的重大战略机遇，构筑我国人工智能发展的先发优势，加快建设创新型国家和世界科技强国方面打下了重要基础。随后，科技部陆续确定建设自动驾驶、城市大脑、医疗影像、智能语音、智能视觉5个国家新一代人工智能开放创新平台。这5个平台分别依托百度、阿里云、腾讯、科大讯飞、商汤科技5家公司来建设，以在汇聚创新资源、促进众创共享方面发挥更大的作用。

5.3.3 人工智能带来的社会担忧

人工智能发展在就业、安全、伦理等方面可能带来新的挑战，人们一方面希望人工智能和智能机器能够代替人类从事各种劳动，另一方面又担心它们的发展会引起新的社会问题。

1. 人工智能技术将人类从繁复工作中解脱出来的同时，也会取代一些工作岗位

随着人工智能技术的发展和智能机器成本的降低，一些领域中的工人会被自动化所取代。如机器人技术和工艺的日益成熟，使其成本越来越低，甚至低于人工成本。这对劳动力市场产生了巨大的影响，在生产和服务业等一些领域，机器人取代人工渐成趋势，如图5.3.2所示，快递分拣机器人已代替人工进行部分快递分拣工作。

图5.3.2 快递分拣机器人

从发展角度看，在可预测环境中从事高度重复或按部就班工作的人员，如电话销售员、仓库工作人员、收银员、火车司机、厨师、律师助理等，将会逐步被机器人取代。同时，也会产生新的工作岗位，如研发和维护智能销售系统、生产仓管机器人等岗位。而且从长期来看，科技带来的就业远大于失业。

因此，面对人工智能所带来的就业问题，社会层面上，需要在对人工智能可能带来的就业影响进行判断的基础上，定向调整人才培养方案，从而使未来劳动力供给与经济社会发展需求更加匹配；个体层面上，人们需要改变自己的思维和工作方式，学会与智能机器和谐共处，以适应这种变化。

2. 人工智能技术推动人类社会进步的同时，也可能威胁人类安全

科学技术推动人类社会进步的同时，也会给人类安全造成一定威胁。新技术的最大危险在于人类对它失去控制，正如化学科学成果被用于制造化学武器，生物学科学成果被用于制造生物武器，核物理研究成果被用于制造核武器。核武器、生物武器、化学武器至今仍对人类安全构成重大威胁，为此人们制定了《不扩散核武器条约》《禁止生物武器公约》《禁止化学武器公约》等国际防扩散领域的重要条约，各领域的科学研究人员共同遵守这些约定，安全地发展和应用相关科学研究成果，造福人类。

人工智能理论和技术的快速发展和不断突破，尤其是在一些方面超越人类的表现，使人们开始担忧是否会对它失去控制或是一旦落入反人类的社会成员手中会被用于反人类和危害社会的犯罪。对此人们必须保持高度警惕，同时人类要有足够的智慧和信心，研制出防范和侦破各种智能犯罪活动的措施。正如美国著名科幻作家阿西莫夫（Asimov）提出的“机器人三守则”：①机器人必须不危害人类，也不允许它眼看人类受害而袖手旁观；②机器人必须绝对服从人类，除非这种服从有害于人类；③机器人必须保护自身不受伤害，除非为了保护人类或者人类命令它做出牺牲。

人工智能技术对经济发展、社会进步都有巨大的推动作用。随着技术的进步，这种影响将越来越大、越来越明显。也许有些影响现在还难以预测，但可以肯定，未来人类与智能机器必定可以安全、和谐地相处，人工智能必将对人类的物质文明和精神文明产生深远影响。

问题与讨论

人们一般认为，只有人类才具有情感和意识，并以此与机器相区别。如果有一天，机器也能够思考和创作，有人类的情感和意识，甚至机器的智能超过人类的自然智能。那时，人类将如何与智能机器共处？

思考与练习

1. 人工智能将怎样改变我们的生活？

2. 无人汽车的研发涉及哪些人工智能技术？无人汽车真正上路需要具备哪些条件？可能会产生哪些问题？

巩固与提高

1. 举例说明：为什么说“人工智能是一门交叉学科”？

2. 1959年，毕业于我国西南联大的华裔科学家王浩在IBM704计算机上证明了《数学原理》中全部150条一阶逻辑以及200条命题逻辑定理。请根据人工智能三种主要方法的特点，分析该人工智能技术应用属于哪种方法。

3. 通过各种途径了解当前人工智能在各个领域的应用，将这些应用与人类的智力活动进行比较，分析当前人工智能的局限性。

4. 无人超市的出现刷新着消费者的认知，但也面临着较大的风险和挑战。分析无人超市当前面临的问题并设想应对措施。

5. 为什么当年IBM“深蓝”的棋力不如AlphaGo？试从两者所采用的主要技术展开原因分析。

6. “贝叶斯公式”是一种从先验概率计算后验概率的方法。如果没有贝叶斯公式，人工智能的“智商”就等于零了。在机器翻译、视觉分类和语音识别等人工智能应用中，均会用到贝叶斯公式。请了解贝叶斯公式，并简要概述其在人工智能方面的应用。

项目挑战

“奇点”到来了吗

2015年，谷歌和微软同时宣布自家的算法在图像识别方面已胜过人类；2017年，在继AlphaGo完胜人类棋手后，其升级版本Alpha Zero又以100∶0完胜AlphaGo，并且Alpha Zero可以事先无任何人类输入，从“零”开始自学围棋（当然，需要事先告诉Alpha Zero围棋的基本规则和胜负标准）；2018年1月，微软和阿里巴巴同时宣布已开发出“在阅读上胜过人类”的人工智能软件，他们的软件都在SQuAD（Stanford Question Answering Dataset）比赛中超过了人类平均水平，图5.3.3所示为微软和阿里巴巴宣称该成就所依据的机器阅读比赛成绩排名榜。

Rank	Model	EM	F1
1 Jan 05, 2018	SLQA+ (ensemble) *Alibaba IDST NLP*	82.440	88.607
1 Jan 03, 2018	r-net+ (ensemble) *Microsoft Research Asia*	82.650	88.493
2 Dec 17, 2017	r-net (ensemble) *Microsoft Research Asia* http://aka.ms/rnet	82.136	88.126
2 Dec 22, 2017	AttentionReader+ (ensemble) *Tencent DPDAC NLP*	81.790	88.163
3 Nov 17, 2017	BiDAF+Self Attention+ELMo(ensemble) *Allen Institute for Artificial Intelligence*	81.003	87.432
4 Jan04, 2018	{EAZI} (ensemble) *Yiwise NLP Group*	80.436	86.912
5 Jan05, 2018	r-net+ (single model) *Microsoft Research Asia*	79.901	86.536
6 Dec 05, 2017	SAN (ensemble model) *Microsoft Business AI Solutions Team* http://arxiv.org/pdf/1712.03556.pdf	79.608	86.496
6 Dec 28, 2017	SLQA+ (single model) *Alibaba IDST NLP*	79.199	86.590
7 Oct 17, 2017	Interactive AoA Reader+ (ensemble) *Joint Laboratory of HIT and iFLYTEK*	79.083	86.450

图5.3.3　2018年机器阅读比赛排名榜

人工智能正在一个接一个的特定领域逐渐超越人类，特别是语言与阅读这个被认为最能代表人类智能的领域，现在也被机器超越，于是有人惊叹“奇点终于来了”（奇点指的是人工智能超越人类的那个时刻）。但也有人在担心人工智能的发展，认为总有一天智能设备会控制人类……

项目任务

为了理性回答“奇点”是否到来这个问题，你需要完成以下任务：

1. 分析、比对人类智能和当前人工智能的覆盖范围，回答人工智能在哪些领域还未能超越人类。

2. 以“机器阅读”为切入点，了解此类人工智能的基本原理与方法，对它的现状与未来有一个相对清晰的判断。

3. 思考如何正确看待人工智能与人类的关系。

过程与建议

1. 比对人类智能和当前人工智能的覆盖范围

很多人工智能技术公司都推出了相对成熟的服务，如百度AI推出了语音识别、语音合成、语音唤醒、文字识别、机器翻译、智能营销、增强现实、知识图谱、舆情监控等服务项目。除了上述项目，现实中还有哪些人工智能的服务？请广泛了解不同的服务所依赖的技术分别是什么。然后，试着列出依靠人类智能所能做的服务。比较人类智能与人工智能所能覆盖的范围，并分析：

（1）当前人工智能不能替代人类的是哪些部分？

（2）人工智能的优势是什么？弱势是什么？

2. 了解机器阅读人工智能的基本原理与方法

机器阅读人工智能软件在SQuAD比赛中超过了人类水平。此类人工智能软件的基本原理与方法是什么？试着分析：

（1）SQuAD比赛考查的是不是人类语言与阅读的全部？如果不是全部，哪些部分未被考查？

（2）以现在机器阅读类人工智能软件的方法，是否有可能在语言与阅读领域全面超越人类？为什么？

3. 思考人工智能与人类的关系

人类研发智能技术的初衷是为了借助机器来帮助人类更好地生活，但随着机器的“智能”越来越发达，人类却有了越来越多的担心。请分析人工智能研究的价值及意义，思考为了让人工智能真正有益于人类，应该在哪些方面做努力。

（1）心态上：我们应该以怎样的心态来看待人工智能的发展？

（2）自身发展上：人类应该如何发展自己，才能在诸多工作可能被人工智能所替代的社会中幸福、快乐、有价值地生活？

（3）制度保障上：为了避免人工智能的发展走上歧路，应该制定怎样的相关法律法规？

（4）其他：还可以做哪些努力？

4. 形成观点，分享交流

将自己收集的数据、展开的分析以及形成的观点以合适的形式整理并呈现（PPT、Word、微信、微博等），你的呈现应能很好地回答以下问题：

（1）对于“奇点终于来了”这个观点，你的看法是什么？

（2）你为什么会有这样的看法？有哪些证据（数据、专家观点、研究成果、比对分析

结果等）可以支持你的观点？

（3）面对人工智能的快速发展，如果需要对人类的应对提出建议，你有哪些建议？提出这些建议的原因是什么？

评价标准

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

评价条目	说明	评分（1～10分）	评分主要依据阐述	后续完善方向
主要观点	观点的描述清晰明了			
支撑材料	支撑材料科学、可靠、丰富且针对性强，能够很好地支撑观点			
论证过程	表述简明扼要，内容之间的逻辑关系清晰			
后续建议	面对人工智能的发展，所提出的建议思考全面、理性、有价值			

拓展项目

1.“图灵测试”是测试机器是否具有智能的一种方法。你对该测试方式有何改进意见？设想你是图灵测试中的询问者，你会向被测试的人和机器提出什么样的问题？（至少列出三个问题）

2.“中文房间”实验。美国哲学家约翰·希尔勒（John Searle）提出，假设他被锁在一个房间里，房间里有很多中国书法作品。他并不懂中文，甚至无法将汉语与日语或其他毫无意义的字区分开来。希尔勒在房间中发现了一套规则（即人工智能算法），这些规则可将中文翻译为英文。随后，每当屋外的人用中文向希尔勒提问，希尔勒通过这些规则将中文问题翻译为英文，又用这些规则将自己回答的英文答案翻译为中文，进而将翻译得到的中文反馈给屋外的提问者（如图5.3.4所示）。过了一会儿，希尔勒逐渐熟悉了这项任务——尽管他仍然不清楚自己操作的这些符号到底是什么。希尔勒问，这种情况下能否说房间内的人“懂”中文？答案是否定的。你从这个“中文房间”实验中得到了哪些启示？

图5.3.4　“中文房间”实验

3. 智慧城市。人工智能技术正致力于改变我们的城市。1964年，也就是纽约世界博览会举办的那一年，英国建筑师罗恩·赫伦（Ron Herron）提出了“行走的城市”的概念。如同在前卫的建筑期刊《阿基格拉姆学派》中所描述的，赫伦支持建造巨大的人工智能移动机器人平台，这种平台就像是背着摩天大楼在地球漫步的蜘蛛。这些四处行走的城市可以在地球上无国界地生存，可以随意前往它们需要获取资源或制造能量的任何地方。赫伦的城市甚至还有相互连接以创造更大的“行走的城市”的能力。这种城市不仅可以自给自足，而且由于人工智能的突破性发展，还能够自治。请你以此展开想象，人工智能将在哪些方面改变我们的城市？

附　表

ASCII码表

十进制	十六进制	字符	十进制	十六进制	字符	十进制	十六进制	字符	十进制	十六进制	字符
0	00	NUL	32	20	（空格）	64	40	@	96	60	`
1	01	SOH	33	21	!	65	41	A	97	61	a
2	02	STX	34	22	"	66	42	B	98	62	b
3	03	ETX	35	23	#	67	43	C	99	63	c
4	04	EOT	36	24	$	68	44	D	100	64	d
5	05	ENQ	37	25	%	69	45	E	101	65	e
6	06	ACK	38	26	&	70	46	F	102	66	f
7	07	BEL	39	27	'	71	47	G	103	67	g
8	08	BS	40	28	(	72	48	H	104	68	h
9	09	HT	41	29	)	73	49	I	105	69	i
10	0A	LF	42	2A	*	74	4A	J	106	6A	j
11	0B	VT	43	2B	+	75	4B	K	107	6B	k
12	0C	FF	44	2C	,	76	4C	L	108	6C	l
13	0D	CR	45	2D	–	77	4D	M	109	6D	m
14	0E	SO	46	2E	.	78	4E	N	110	6E	n
15	0F	SI	47	2F	/	79	4F	O	111	6F	o
16	10	DLE	48	30	0	80	50	P	112	70	p
17	11	DC1	49	31	1	81	51	Q	113	71	q
18	12	DC2	50	32	2	82	52	R	114	72	r
19	13	DC3	51	33	3	83	53	S	115	73	s
20	14	DC4	52	34	4	84	54	T	116	74	t
21	15	NAK	53	35	5	85	55	U	117	75	u
22	16	SYN	54	36	6	86	56	V	118	76	v
23	17	ETB	55	37	7	87	57	W	119	77	w
24	18	CAN	56	38	8	88	58	X	120	78	x
25	19	EM	57	39	9	89	59	Y	121	79	y
26	1A	SUB	58	3A	:	90	5A	Z	122	7A	z
27	1B	ESC	59	3B	;	91	5B	[	123	7B	{
28	1C	FS	60	3C	<	92	5C	\	124	7C	|
29	1D	GS	61	3D	=	93	5D	]	125	7D	}
30	1E	RS	62	3E	>	94	5E	^	126	7E	~
31	1F	US	63	3F	?	95	5F	_	127	7F	DEL

注：标准ASCII码共有128个，其中32～126为可显示字符，其他位置为控制字符。