普通高等教育“十一五”国家级规划教材

电子信息与电气学科规划教材·电子信息科学与工程类专业

信息论与编码

（第2版）

陈 运 主编

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY http://www.phei.com.cn

普通高等教育“十一五”国家级规划教材
电子信息与电气学科规划教材·电子信息科学与工程类专业

信息论与编码

（第2版）

陈　运　主编

周　亮　陈　新　陈伟建　李　飞　副主编

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内 容 简 介

本书系统介绍和论述了信息的基本概念；信息论的起源、发展及研究内容；香农信息论的三个基本概念：信源熵、信道容量和信息率失真函数，以及与这三个概念相对应的三个编码定理；解决通信系统有效性、可靠性和安全性的三类编码：信源编码、信道编码和安全编码——密码的基本方法等内容。为了便于教学和读者自学，每章后面都附有习题。

本书不追求高深的数学理论，尽可能以通俗易懂、形象生动的语言强化物理概念的描述，特别适合于初学者。已掌握工科高等数学和工程数学的读者都能读懂本书。

本书适合作为高等院校电子信息类相关专业高年级本科生的教材，也可作为低年级研究生的教学参考书，还可供从事信息科学与技术的科研人员和工程技术人员参考。

图书在版编目（CIP）数据

信息论与编码/陈运主编. —2 版. —北京：电子工业出版社，2007.9
电子信息与电气学科规划教材·电子信息科学与工程类专业
ISBN 978-7-121-04458-8

Ⅰ. 信… Ⅱ. 陈… Ⅲ. ①信息论—高等学校—教材 ②信息论—高等学校—教材 ③信息编码—编码理论—高等学校—教材 Ⅳ. TN911.2

中国版本图书馆 CIP 数据核字（2007）第 128504 号

责任编辑：王传臣
印　　刷：北京市李史山胶印厂
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×1 092　1/16　印张：16　字数：409.6 千字
印　　次：2007 年 9 月第 1 次印刷
印　　数：5 000 册　　定价：22.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

第 2 版前言

1948 年，美国科学家香农（C. E. Shannon）发表了题为“通信的数学理论”的学术论文，这篇划时代论文的问世，宣告了信息论的诞生。从信息的度量开始，信息的概念和研究范围在不断扩大和深化并迅速地渗透到其他相关学科领域，如无线电技术、自动控制、人工智能、神经网络、信号与信息处理、网络技术、计算机技术、生命科学、材料学、密码学、心理学、质量管理、市场营销、信息经济、美学等。信息论的研究领域从自然科学扩展到经济、管理科学甚至人文社会科学，从狭义信息论发展到如今的广义信息论，成为涉及面极广的信息科学。

微电子技术、通信技术、计算机技术和网络技术的迅猛发展，加速了社会信息化的进程。21 世纪，国际社会已进入信息化时代。信息论作为信息科学和技术的基本理论，犹如信息科学大厦的地基，在信息社会中占据越来越重要的地位。

信息论是信息科学中最成熟、最完善的一部分，它的创立者在通信理论的研究中独辟蹊径，终于抓住了通信的本质。信息论与其他学科的交叉和融合，促进了许多新兴学科的生长，这也是信息论最具生机和最有魅力之处。

信息论为计算机和远程通信奠定了坚实的理论基础，是 20 世纪产生的对人类最伟大的贡献之一。它不仅在方法论的层面上解决通信的有效性、可靠性和安全性问题，而且在认识论层面上帮助人们认识事物的本质。学完信息论之后，再重新审视一下周围的事物，我们会有许多新的看法和认识。对此，作者有深切的体会。在从事科学研究的过程中，常会遇到错综复杂的问题、面临“山重水复疑无路”的境地，此时跳过方法的死结，用信息论的观点重新审视一下症结所在，很多时候会有“柳暗花明又一村”的惊喜发现。用信息论可以宏观地认识某些政治问题，也可以定量地解决某些经济问题，还可以分析、解释外语学习中存在的问题……总之，信息论是高层次信息科学和技术人才必不可少的基础知识。

然而，作者在多年的教学中发现，由于信息论涉及众多学科，需要广泛的数学基础，许多学生虽然认识到信息论的重要性，但在繁杂的数学公式面前只好望而却步。针对这种情况，作者在《信息工程理论基础》和《信息理论与编码》两本讲义的基础上，根据多年的教学经验，在 2002 年编著出版了《信息论与编码》，把信息论涉及的数学知识限制在工科高等数学和工程数学的范畴内，重新证明了某些性质和定理，并尽量以通俗形象的语言强化物理概念的描述，使读者读得懂，愿意学，有兴趣学。

《信息论与编码》出版四年多来，收到来自国内数十所院校师生的电子邮件和信函，就书中的有关问题进行交流，或对本书提出建议和批评。作者在此感谢国内广大读者对本书的关注和厚爱！作者在此次修订过程中，更正了第 1 版中的错误疏漏，针对几年来教学过程中出现的问题，对书的结构进行了调整，使逻辑性更强；在描述方法上进行了较大改进，充实了示例，更易于读者理解；增加了密码的安全性测度一章，补充了较多信源编码的内容，其余章节内容也进行了不同程度的更新，更注重实用性和先进性。

本书共分 7 章，遵照由浅入深、循序渐进的教学规律，系统地组织教学内容。第 1 章：概论，介绍信息的基本概念和定义、信息论的起源、发展和研究内容；第 2 章：信源熵，介绍熵的概念、性质、定理，信息冗余度的概念以及离散无失真信源编码定理等；第 3 章：信道容量，介绍信道容量的定义、计算和信道编码定理；第 4 章：信息率失真函数，介绍信息率失

真函数的概念、计算、应用以及保真度准则下的信源编码定理；第 5 章：信源编码，介绍各种常见和实用的信源编码方法；第 6 章：信道编码，介绍信道编码的概念、思路和三类常见的信道编码方法；第 7 章：密码体制的安全性测度，简要介绍了密码的基本知识和信息熵测度密码安全性的基本概念。其中，第 1、3、4 章由成都信息工程学院陈运教授编写；第 2 章由陈运和郑州轻工业学院陈新教授联合编写；第 5 章 5.1 节由陈新编写，5.2～5.4 节由电子科技大学陈伟建副教授编写；第 6 章由电子科技大学周亮教授编写；第 7 章 7.1～7.3 节由成都信息工程学院李飞教授编写，7.4 节由陈运编写。全书由陈运教授审阅统稿。

为了方便教学，本书开发了配套的多媒体课件和部分习题答案或题解，免费提供给使用本教材授课的教师，授课教师可向电子工业出版社索取；同时，还将开通课程网页为广大师生提供方便的交流平台，有关事宜，读者可发邮件至 wzhen@cuit.edu.cn，与吴震老师联系。

作者在此特别感谢庞宏、吴震讲师，陈俊博士，硕士生谭丽娟、郭兰、李铮铮、何云程、廖志强、蒋定德、曾旭、李家勋、孙德强、盛茂和陈增荣，他们先后参与了课件制作、网页和网上题库开发或习题解答以及录入等相关工作。

由于时间紧迫和作者的知识水平所限，书中错误疏漏之处在所难免，热忱希望广大读者批评指正。

作　者

2007 年于成都

目　录

第1章 概 论

1.1 信息的一般概念

信息科学、材料科学和能源科学一起被称为当代文明的“三大支柱”。一位美国科学家说：“没有物质的世界是虚无的世界，没有能源的世界是死寂的世界，没有信息的世界是混乱的世界。”可见信息的重要性。随着社会信息化进程的加速，人们对信息的依赖程度会越来越高。

花朵开放时的色彩是一种信息，它可以引来昆虫为其授粉；成熟的水果会产生香味，诱来动物觅食，动物食后为其传播种子，所以果香也是一种信息；药有苦味，这种信息是味觉感知的；听老师讲课可以得到许多知识，知识也是信息……可见信息处处存在，人的眼、耳、鼻、舌、身都能感知信息。

那么信息究竟是什么呢？

信息自古就有，但是古代社会文明程度很低，信息传递手段落后，获取信息困难，人们没有意识到信息的存在。随着人类社会的不断进步，人们才意识到信息的存在。对信息的认识随着社会文明程度的提高不断提高和深入。然而，信息学科毕竟还是一门年轻的学科，人们对信息还没有一个全面的、系统的、准确的、一致的认识。从不同的学科、不同的角度、不同的方面、不同的层次、不同的深度，对信息有不同的认识。

信息的概念十分广泛，不同的定义在百种以上。例如，“信息是事物之间的差异”，“信息是事物联系的普遍形式”，“信息是物质和能量在时间和空间中分布的不均匀性”，“信息是物质的普遍属性”，“信息是收信者事先所不知道的报道”，“信息是用以消除随机不确定性的东西”，“信息是负熵”，“信息是作用于人类感觉器官的东西”，“信息是通信传输的内容”，“信息是加工知识的原材料”，“信息是控制的指令”，“信息就是数据”，“信息就是情报”，“信息就是知识”……

数学家认为“信息是使概率分布发生改变的东西”，哲学家认为“信息是物质成分的意识成分按完全特殊的方式融合起来的产物”……

1928年，美国数学家哈特莱（Hartley）在《贝尔系统电话杂志》上发表了一篇题为“信息传输”的论文，把信息理解为选择通信符号的方式，并用选择的自由度来计量这种信息的大小。他认为，发信者所发出的信息，就是他在通信符号表中选择符号的具体方式。例如，从符号表中选择了这样一些符号：“I am well”，他就发出了“我平安”的信息；如果选择了“I am sick”这些符号（包括空格），他就发出了“我病了”的信息。发信者选择的自由度越大，所能发出的信息量也就越大。此外，哈特莱还注意到，选择的具体物理内容是无关紧要的，重要的是选择的方式。也就是说，不管符号代表的意义是什么，只要符号表的符号数目一定，“字”的长度一定，那么，发信者所能发出的信息的数量就被限定了。所以他认为“信息是选择的自由度”。

时隔20年，另一位美国数学家香农（C. E. Shannon）在《贝尔系统电话杂志》发表了题为“通信的数学理论”的长篇论文。这篇论文以概率论为工具，深刻阐述了通信工程的一系列基本理论问题，给出了计算信源信息量和信道容量的方法和一般公式，得到了一组表征信息传递重要关系的编码定理，从而创立了信息论。但是香农并没有给出信息的确切定义，他认为“信息就是一种消息”。

后来，随着认识的进一步深化，人们把信息理解为广义通信的内容。美国数学家、控制论的主要奠基人维纳（Winner）在1950年出版的《控制论与社会》一书中写道："人通过感觉器官感知周围世界"，"我们支配环境的命令就是给环境的一种信息"，因此，"信息就是我们在适应外部世界，并把这种适应反作用于外部世界的过程中，同外部世界进行交换的内容的名称"，"接收信息和使用信息的过程，就是我们适应外界环境的偶然性的过程，也是我们在这个环境中有效地生活的过程"。在这里，维纳把人与外部环境交换信息的过程看做是一种广义的通信的过程，认为"信息是人与外界相互作用的过程中所交换的内容的名称"。

这些定义都或多或少地从某种程度上描述了信息的一些特征，但是都不够全面、系统和准确。例如，消息、信号、数据、情报和信息都是在通信系统中传送的东西，但是这些概念之间有着原则的区别。消息是信息的外壳，信息则是消息的内核。同样多的消息，所包含的信息量可能差异很大；反之，不同形式的消息可能包含同样多的信息。信号也不等同于信息，信号只是信息的载体，信息是信号所载荷的内容。至于数据，它只是记录信息的一种形式，而且不是唯一的形式，因此不能把它等同于信息本身。"情报"一词在日语中的确就是信息，但是在汉语中，情报只是一类专门的信息，是信息的一个子集。

维纳对信息的认识也不够准确。因为在人与外界相互作用的过程中，参与内容交换的不仅仅是信息，还有物质和能量。后来维纳自己也认识到"信息既不是物质又不是能量，信息就是信息"。这句话起初被人批评为唯心主义，也有人笑话维纳"说了等于没说"。但是人们后来才意识到，正是维纳揭示了信息的特质，即信息是独立于物质和能量之外存在于客观世界的第三要素。

上述定义虽然各不相同，实质内容并无太大的差异，主要差异在于侧面不同、详略不同、抽象的程度不同和概括的层次高低不同。根据不同的条件，区分不同的层次，可以给信息下不同的定义。最高的层次是最普遍的层次，也是无约束条件的层次，定义事物的信息是该事物运动的状态和状态改变的方式。我们把它叫做"本体论"层次。在这个层次上定义的信息是最广义的信息，使用范围也最广。每引入一个约束条件，定义的层次就降低一点，使用的范围就变窄一点。

例如，引入一个最有实际意义的约束条件：认识主体，即站在认识主体的立场上定义信息，这时本体论层次的信息定义就转化为认识论层次的信息定义。即信息是认识主体（生物或机器）所感知的或所表述的相应事物的运动状态及其变化方式，包括状态及其变化方式的形式、含义和效用。其中认识主体所感知的东西是外部世界向认识主体输入的信息，而认识主体所表述的东西则是其向外部世界输出的信息。

虽然认识论比本体论的层次要低一些，所定义信息的使用范围也要窄一些，但是信息概念的内涵比本体论层次要丰富得多。因为认识主体具有感觉能力、理解能力和目的性，能够感觉到事物运动状态及其变化方式的外在形式和内在含义，并能够判断其效用价值。对认识主体来说，这三者之间是相互依存、不可分割的关系。因此，在认识论层次上研究信息的时候，"事物的运动状态及其变化方式"就不再像本体论层次上那样简单了，它必须同时考虑到形式、含义和效用三个方面的因素。

事实上，认识主体只有在感知了事物运动状态及其变化的形式，理解了它的含义，判明了它的效用之后，才算真正掌握了这个事物的认识论层次信息，才能做出正确的决策。我们把同时考虑事物运动状态及其变化方式的外在形式、内在含义和效用价值的认识论层次信息称为"全信息"，而把仅仅考虑其中形式因素的部分称为"语法信息"，把考虑其中含义因素的部分称为"语义信息"，把考虑其中效用因素的部分称为"语用信息"。换句话说，认识论层次的信息是同时考虑语法信息、语义信息和语用信息的全信息。

香农信息论仅考虑了事物运动状态及其变化方式的外在形式，实际上研究的是语法信息。从这个角度出发，可以对信息下这样的定义：信息是对事物运动状态和变化方式的表征，它存在于任何事物之中，可以被认识主体（生物或机器）获取和利用。从数学观点出发研究香农信息论，可以认为信息是对消息统计特性的一种定量描述。

信息存在于自然界，也存在于人类社会，其本质是运动和变化。可以说哪里有事物的运动和变化，哪里就会产生信息。

信息必须依附于一定的物质形式存在，这种运载信息的物质，称为信息载体。

人类交换信息的形式丰富多彩，使用的信息载体非常广泛。概括起来，有语言、文字和电磁波。语言是信息的最早载体；文字和图像使信息保存得更持久，传播范围更大；电磁波则使载荷信息的容量和速度大为提高。

信息本身既看不见，又摸不着，没有气味，没有颜色，没有形状，没有大小，没有重量……总之，它是非常抽象的东西。但信息又处处存在，呼之塞耳，示之濡目。它既区别于物质和能量，又与物质和能量有相互依赖的关系。

综合起来，信息有如下重要性质：

（1）存在的普遍性。信息的本质是事物的运动和变化，只要有事物的存在，就会有事物的运动和变化，就会产生信息。绝对静止的事物是没有的，因此，信息普遍存在。

（2）有序性。信息可以用来消除系统的不确定性，增加系统的有序性。认识论层次的信息是认识主体所感知和表述的事物运动的状态和方式。获得了信息，就可以消除认识主体对于事物运动状态和方式的不确定性。信息的这一性质对人类有特别重要的价值，要使一个系统从无序变为有序，必须从外界获取信息。

（3）相对性。对于同一个事物，不同的观察者所能获得的信息量可能不同。

（4）可度量性。信息虽然很抽象，但它是可以度量的。信息的多少用信息量表示。

（5）可扩充性。信息并非一成不变。随着时间的推移，大部分信息将得到不断的扩充。例如，人类对于宇宙的认识就是不断扩充的，人们对信息的认识也在不断地扩充。香农创立信息论之前，很少有人意识到信息的客观存在，如今人们对信息的研究已经非常广泛和深入。

（6）可存储、传输与携带性。信息依附于信息载体而存在，而任何物质都可以成为信息的载体。既然物质可以存储、传输和携带，所以信息可通过信息载体以多种形式存储、传输和携带。

（7）可压缩性。人们得到信息之后，并非原封不动拿来应用，往往要进行加工、整理、概括、归纳，使信息更加精练、可靠，从而浓缩。信息论研究的主要问题之一就是信息的压缩。

（8）可替代性。信息能替代劳力、资本、物质材料甚至时间，正确、及时、有效地利用信息，可创造更多的物质财富，开发或节约更多的能量，节省更多的时间，收到巨大的经济效益。

（9）可扩散性。信息可以在短时间内较大范围地扩散开来。如广播、电视信息，顷刻之间即传遍全球。

（10）可共享性。信息与实物不同，可以大家共享。甲传递一件东西给乙，乙得到，甲便失去。但信息持有者传递一条信息给另一个人的时候，他自己所拥有的信息并不会丧失。正像教师把知识传授给学生一样，学生掌握了知识，但教师并不会成为“白痴”。信息的这种特性对人类具有特别重要的意义。可以说没有信息的共享性就没有人类社会的发展和进步。

（11）时效性。信息以事实的存在为前提。它不是一成不变的死东西，可以随着事实的不断扩大而增值，也会随着事实的过去而衰老，从而失去本身的价值。因此，信息是有“寿命”的。

信息在信息化程度越来越高的社会中将起到越来越重要的作用，是比物质和能量更为宝贵的资源。全面掌握信息的概念，正确、及时、有效地利用信息，能够为人类创造更多的财富。

1.2 信息的分类

前面一节关于信息概念和性质的讨论，使我们对信息有了定性的认识。但要全面、准确地掌握信息的概念，必须对信息有定量的认识。这就要求首先能够确切地描述信息。

由前可知，信息是一种十分复杂的研究对象。要找到一种通用的方法来描述各种各样的信息以及用统一的方法来恰如其分地描述信息的方方面面，显然是非常困难的。要清楚、具体地认识信息，必须对信息进行分类。

信息分类有许多不同的准则和方法。

按照性质，信息可以分成语法信息、语义信息和语用信息。

按照地位，信息可以分成客观信息和主观信息。

按照作用，信息可以分成有用信息、无用信息和干扰信息。

按照应用部门，信息可以分成工业信息、农业信息、军事信息、政治信息、科技信息、文化信息、经济信息、市场信息和管理信息等。

按照携带信息的信号的性质，信息还可以分成连续信息、离散信息和半连续信息等。

……

我们研究信息的目的，就是要准确地把握信息的本质和特点，以便更有效地利用信息。因此，在众多的分类原则和方法中，最重要的就是按照信息性质的分类。

按照性质的不同可以把信息划分成语法信息、语义信息和语用信息三个基本类型。其中最基本也最抽象的类型是语法信息。它是迄今为止在理论上研究得最多的类型。

语法信息考虑的是事物运动状态和变化方式的外在形式。根据事物运动状态和方式在形式上的不同，语法信息还可以进一步分成有限状态和无限状态；其次，事物运动状态可能是连续的，也可能是离散的，于是，又可以分成连续状态语法信息和离散状态语法信息；再者，事物运动状态还可能是明晰的或者是模糊的，这样，又可以分成状态明晰的语法信息和状态模糊的语法信息。

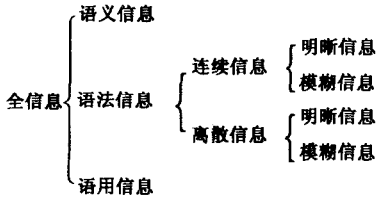

图 1.2.1 不同性质的信息分类

当然，按照事物运动的方式，还可以把信息进一步细分为概率信息、偶发信息、确定信息和模糊信息。香农信息论主要讨论的是语法信息中的概率信息，本书也以概率信息为主要研究对象。

上述分类可以用图 1.2.1 直观地表示。

1.3 信息论的起源、发展及研究内容

信息论自诞生到现在不到 60 年时间，在人类科学史上是相当短暂的，但它的发展对学术界及人类社会的影响是相当广泛和深刻的。信息作为一种资源，如何开发、利用、共享，是人们普遍关心的问题。

在人类历史的长河中，信息传输和传播手段经历了五次重大变革，正是在不断的变化中，人们逐渐认识到信息的存在及重要作用。第一次变革是语言的产生。人们用语言准确地传递感

情和意图，使语言成为传递信息的重要工具。第二次变革是文字的产生。不久又发明了纸张，人类开始用书信的方式交换信息，使信息传递的准确性大为提高。第三次变革是印刷术的发明。它使信息能大量存储和大量流通，并显著扩大了信息的传递范围。

第四次变革是电报、电话的发明，开始了人类电信时代，通信理论和技术迅速发展。1924 年，奈奎斯特（Nyquist）解释了信号带宽和信息速率之间的关系。20 世纪 30 年代，新的调制方式，如调频、调相、单边带调制、脉冲编码调制和增量调制的出现，使人们对信息能量、带宽和干扰的关系有了进一步的认识。1936 年，阿姆斯特朗（Armstrong）指出增大带宽可以使抗干扰能力加强，并根据这一思想提出了宽频移的频率调制方法。1939 年，达德利（Dudley）发明了带通声码器，指出通信所需带宽至少同待传送消息的带宽应该一样。声码器是最早的语言数据压缩系统。这一时期还诞生了无线电广播和电视广播。通信技术的进步使人们更深入地考虑问题：究竟如何定量地研究通信系统中的信息？怎样才能更有效和更可靠地传递信息？现有的各种通信体制如何改进？等等。

1928 年，哈特莱首先提出了用对数度量信息的概念。哈特莱的工作给香农很大的启示，他在 1941～1944 年对通信和密码进行深入研究，用概率论和数理统计的方法系统地讨论了通信的基本问题，得出了几个重要而带有普遍意义的结论。他阐明了通信系统传递的对象就是信息，并对信息给予科学的定量描述，提出了信息熵的概念。指出通信系统的中心问题是在噪声下如何有效而可靠地传递信息，以及实现这一目标的方法是编码，等等。这些成果 1948 年以“通信的数学理论”（A mathematical theory of communication）为题公开发表，标志着信息论的正式诞生。与此同时，维纳（Winner）在研究火控系统和人体神经系统时，提出了在干扰作用下的信息最佳滤波理论，成为信息论的一个重要分支。

20 世纪 50 年代，信息论在学术界引起了巨大反响。1951 年，美国无线电工程师协会（IRE）成立了信息论组，并于 1955 年正式出版了信息论汇刊。这一时期，包括香农本人在内的一些科学家做了大量工作，发表了许多重要文章，将香农的科学论断进一步推广，同时信道编码理论有了较大的发展。信源编码的研究落后于信道编码。1959 年，香农在发表的“保真度准则下的离散信源编码定理”（Coding theorems for a discrete source at the fidelity criterion）一文中系统地提出了信息率失真理论（rate-distortion theory），为信源压缩编码的研究奠定了理论基础。

20 世纪 60 年代，信道编码技术有了较大发展，成为信息论的又一重要分支。它把代数方法引入到纠错码的研究中，使分组码技术达到了高峰，找到了可纠正多个错误的码，并提出了可实现的译码方法。其次是卷积码和概率译码有了重大突破，提出了序列译码和维特比（Viterbi）译码方法。

1961 年，香农的重要论文“双路通信信道”开拓了多用户信息理论的研究。

第五次变革是计算机技术与通信技术相结合，促进了网络通信的发展。宽带综合业务数字网（B-ISDN, Broad-Integrated Service Digital Network）的出现，给人们提供了除电话服务以外的多种服务，使人类社会逐渐进入了信息化时代。信息理论的研究得到进一步的发展，多用户理论的研究取得了突破性的进展。至此，香农的单用户信息论已推广到多用户信息论。20 世纪 70 年代以后，多用户信息论——即现在所说的网络信息论成为中心研究课题之一。

后来，随着通信规模的不断扩大，人们逐渐意识到信息安全是通信系统正常运行的必要条件。于是，把密码学也归类为信息论的分支。随着计算机和网络病毒以及黑客的泛滥，信息安全已是各国政府、企业、个人共同关心的问题。

人们对信息的认识越来越深入，先后提出了加权熵、动态熵等概念，建立在模糊数学基础之上的模糊信息的研究也取得了一定的进展。信息论不仅在通信、广播、电视、雷达、导航、计算机、自动控制、电子对抗等电子学领域得到了直接应用，还广泛地渗透到诸如医学、生物学、心理学、神经生理学等自然科学的各个方面，甚至渗透到语言学、美学等领域。

从20世纪60年代开始，一些社会学家在研究社会问题和社会现象时，先后提出了后工业社会和信息社会的概念，信息论开始向经济学和社会科学领域渗透。1977年，美国经济学家马克·波拉特发表了长达九卷的《信息经济》报告，用信息论的基本概念研究经济现象和社会现象，将信息论的研究从自然科学领域正式移植到经济学和社会科学领域。另一方面，随着量子理论的发展，逐渐形成了量子信息论。信息论迅速发展成为涉及范围极广的广义信息论——即信息科学。

信息论的研究对象是广义通信系统。不仅电子的、光学的信号传递系统，任何系统，只要能够抽象成通信系统模型，都可以用信息论研究，如神经传导系统、市场营销系统、质量控制系统等。关于信息论的研究内容，一般有以下三种解释。

1. 信息论基础

亦称香农信息论或狭义信息论。主要研究信息的测度、信道容量、信息率失真函数，与这三个概念相对应的是香农三定理以及信源和信道编码。

2. 一般信息论

主要研究信息传输和处理问题。除了香农基本理论之外，还包括噪声理论、信号滤波和预测、统计检测与估计理论、调制理论。后一部分内容以美国科学家维纳为代表。虽然维纳和香农等人都是运用概率和统计数学的方法研究准确或近似再现消息的问题，都是通信系统的最优化问题，但他们之间有一个重要的区别。维纳研究的重点是在接收端，研究消息在传输过程中受到干扰时，在接收端如何把消息从干扰中提取出来。在此基础上，建立了最佳过滤理论（维纳滤波器）、统计检测与估计理论、噪声理论等。香农研究的对象是从信源到信宿的全过程，是收、发端联合最优化问题，重点是编码。香农定理指出：只要在传输前后对消息进行适当的编码和译码，就能保证在有干扰的情况下，最佳地传送消息，并准确或近似地再现消息。为此，发展了信息测度理论、信道容量理论和编码理论等。

3. 广义信息论

广义信息论是一门综合性的新兴学科，至今并没有严格的定义。概括说来，凡是能够用广义通信系统模型描述的过程或系统，都能用信息基本理论来研究。不仅包括一般信息论的所有研究内容，还包括如医学、生物学、心理学、遗传学、神经生理学、语言学、语义学，甚至社会学和经济管理中有关信息的问题。反过来，所有研究信息的识别、控制、提取、变换、传输、处理、存储、显示、价值、作用和信息量的大小的一般规律以及实现这些原理的技术手段的工程学科，也都属于广义信息论的范畴。

总之，人们研究信息论的目的，是为了高效、可靠、安全、经济并且随心所欲地交换和利用各种各样的信息。

第2章 信 源 熵

2.1 单符号离散信源

信息论是在信息可以度量的前提下，研究有效地、可靠地、安全地传递信息的科学。信息的可度量性是建立信息论的基础。

信息度量的方法有：结构度量、统计度量、语义度量、语用度量、模糊度量等。最常用的方法是统计度量。它用事件统计发生概率的对数来描述事物的不确定性，得到消息的信息量，进而建立熵的概念。熵的概念是香农信息论最基本、最重要的概念。

如果甲告诉乙说："你考上了研究生，"那么乙就得到了信息。如果丙又告诉乙同样的话，那么对乙来说，此次他只是得到了一条消息，并没有得到其他任何信息。其实乙得到信息还有一个前提条件，就是乙参加了研究生考试。如果乙根本没有参加研究生考试，也就不可能考上研究生。那么甲的话对乙来说就没有任何信息。

在这个事件当中，"考上了研究生"是对考试结果的一种描述，而考试的结果不止一种，可见乙在得到消息之前具有不确定性。在得到消息之后，只要甲没说错，乙的不确定性就消除了，也就获得了信息。如果我们把考试结果看成是事物的一种状态，把各种不同的结果看成是事物状态运动的方向，那么信息就是对事物运动状态（或它的存在方式）的不确定性的一种描述。不确定性即随机特性，可以用研究随机现象的数学工具——概率论与随机过程来描述信息。

我们再来看一下上面的例子。当乙再次被告知考上研究生的消息时，事件是完全可信的，这相当于概率为1的情况，这种消息不含有不确定性，因此不含有任何信息。同样，若乙根本没有参加研究生考试，那么他被告知的消息是完全不可信的，这相当于概率为0的情况，从理念上来说这种消息同样不应该含有任何信息。不过我们在后面的学习中将会看到，当考察随机事件的单次实验结果和平均实验结果时，得到的结论是不一样的。

从随机变量出发来研究信息，正是香农信息论的基本假说。

信息是由信源发出的，在量度信息之前，首先要研究一下信源。

2.1.1 单符号离散信源的数学模型

信源发出消息，消息载荷信息，而消息又具有不确定性，所以可用随机变量或随机矢量来描述信源输出的消息，或者说用概率空间来描述信源。

一类信源输出的消息常常以一个个符号的形式出现，例如文字、字母等，这些符号的取值是有限的或可数的，这样的信源称为离散信源。有的离散信源只涉及一个随机事件，有的离散信源涉及多个随机事件，分别称为单符号离散信源和多符号离散信源，可分别用离散随机变量和随机矢量来描述。另一类输出连续消息的信源称为连续信源，可用随机过程来描述。

对于离散随机变量 X，取值于集合

$$\{a_1, a_2, \cdots, a_i, \cdots, a_n\}$$

其中 n 可以是有限正整数，也可以是可数无限大整数，即 $n \in I$（整数域），$X \in \{a_i, i = 1, 2, \cdots, n\}$。

规定集合中各个元素的概率为 $p(a_i)$，即

$$p(a_i)=P(X=a_i)$$

其中 $P(X=a_i)$表示括号中随机事件 X 发生某一结果 a_i 的概率。单符号离散信源的数学模型可表示为

$$\begin{pmatrix} X \\ P(X) \end{pmatrix}=\begin{Bmatrix} a_1, & a_2, & \cdots, & a_i, & \cdots, & a_n \\ p(a_1), p(a_2), \cdots, p(a_i), \cdots, p(a_n) \end{Bmatrix} \tag{2.1.1}$$

其中 $p(a_i)$满足

$$0\leqslant p(a_i)\leqslant 1,\quad \sum_{i=1}^{n}p(a_i)=1 \tag{2.1.2}$$

公式(2.1.2)表示信源的可能取值共有 n 个：$a_1, a_2, \cdots, a_i, \cdots, a_n$，每次必取其中之一。

需要注意的是，大写字母 X, Y, Z 代表随机变量，指的是信源整体，带下标的小写字母 a_i, b_j, c_k 代表随机事件的某一结果或信源的某个元素。两者不可混淆。

2.1.2 自信息和信源熵

在以下的讨论中常用到概率论的基本概念和性质。我们先对这些概念和性质进行简要的复习。

随机变量 X, Y 分别取值于集合$\{a_1, a_2, \cdots, a_i, \cdots, a_n\}$和$\{b_1, b_2, \cdots, b_j, \cdots, b_m\}$。$X$ 发生 a_i 和 Y 发生 b_j 的概率分别定义为 $p(a_i)$和 $p(b_j)$，它们一定满足 $0\leqslant p(a_i)$，$p(b_j)\leqslant 1$ 以及 $\sum_{i=1}^{n}p(a_i)=1$ 和 $\sum_{j=1}^{m}p(b_j)=1$。如果考察 X 和 Y 同时发生 a_i 和 b_j 的概率，则二者构成联合随机变量 XY，取值于集合$\{a_ib_j \mid i=1,2,\cdots,n,\ j=1,2,\cdots,m\}$，元素 a_ib_j 发生的概率称为联合概率，用 $p(a_ib_j)$表示。有时随机变量 X 和 Y 之间有一定的关联关系，一个随机变量发生某结果后，对另一个随机变量发生的结果会产生影响，这时我们用条件概率来描述两者之间的关系。如 X 发生 a_i 以后，Y 又发生 b_j 的条件概率表示为 $p(b_j/a_i)$，代表 a_i 已知的情况下，又出现 b_j 的概率。当 a_i 不同时，即使发生同样的 b_j，其条件概率也不相同，说明了 a_i 对 b_j 的影响。而 $p(b_j)$则是对 a_i 一无所知情况下 b_j 发生的概率，有时相应地称 $p(b_j)$为 b_j 的无条件概率。同理，b_j 已知的条件下 a_i 的条件概率记为 $p(a_i/b_j)$。相应地，$p(a_i)$称为 a_i 的无条件概率。例如，集合 X 表示球类活动，含有三个元素$\{a_1, a_2, a_3\}$，分别代表篮球、排球、乒乓球活动。集合 $Y=\{b_1, b_2, b_3\}$代表喜欢篮球、排球和乒乓球运动的同学。在不知道有哪种球类活动的情况下，假设三个同学参加活动的可能性各占 1/3，即 $p(b_1)=p(b_2)=p(b_3)=1/3$。但如果已知举行的是乒乓球活动，则同学 b_3 参加的可能性就比较大，而同学 b_1 和 b_2 参加的可能性就比较小，即 $p(b_3/a_3)$大，$p(b_1/a_3)$和 $p(b_2/a_3)$小。说明 a_3 发生后对 b_1，b_2 和 b_3 的影响，它与 a_3 发生前 b_1，b_2 和 b_3 本身的概率是不同的。这就是条件概率和无条件概率之间的区别。

无条件概率、条件概率、联合概率满足下面一些性质和关系：

（1）$0\leqslant p(a_i), p(b_j), p(b_j/a_i), p(a_i/b_j), p(a_ib_j)\leqslant 1$

（2）$\sum_{i=1}^{n}p(a_i)=1$，$\sum_{j=1}^{m}p(b_j)=1$，$\sum_{i=1}^{n}p(a_i/b_j)=1$，$\sum_{j=1}^{m}p(b_j/a_i)=1$，$\sum_{j=1}^{m}\sum_{i=1}^{n}p(a_ib_j)=1$

（3） $\sum_{i=1}^{n} p(a_i b_j) = p(b_j)$， $\sum_{j=1}^{m} p(a_i b_j) = p(a_i)$

（4） $p(a_i b_j) = p(a_i) p(b_j / a_i) = p(b_j) p(a_i / b_j)$

（5）当X与Y相互独立时， $p(b_j / a_i) = p(b_j)$， $p(a_i / b_j) = p(a_i)$， $p(a_i b_j) = p(a_i) p(b_j)$

（6） $p(a_i / b_j) = \dfrac{p(a_i b_j)}{\sum_{i=1}^{n} p(a_i b_j)}$， $p(b_j / a_i) = \dfrac{p(a_i b_j)}{\sum_{j=1}^{m} p(a_i b_j)}$

一、信息量

1. 自信息量

一个随机事件发生某一结果后所带来的信息量称为自信息量，简称自信息。定义为其发生概率对数的负值。若随机事件发生 a_i 的概率为 $p(a_i)$，那么它的自信息量 $I(a_i)$为

$$I(a_i) = -\log_2 p(a_i) \tag{2.1.3}$$

自信息量的单位与所用对数的底有关。在信息论中常用的对数底为 2，信息量的单位为比特（bit, binary unit 的缩写）。在信息论的公式推导中，为方便起见，常取自然对数，即以 e 为底的对数，信息量的单位为奈特（nat, nature unit 的缩写）。当随机事件的概率很小时，$I(a_i)$是一个相当大的正整数。为了运算方便，可以取 10 作为对数底，信息量的单位是笛特（Det, Decimal Unit）或哈特（Hart, Hartley），以纪念科学家哈特莱首先提出用对数值来度量信息。这三个信息量单位之间的转换关系如下：

$$1\ \text{Nat} = \log_2 e \approx 1.433\text{bit}$$

$$1\ \text{Hart} = \log_2 10 \approx 3.322\text{bit}$$

$$1\text{bit} \approx 0.693\ \text{Nat}$$

$$1\text{bit} \approx 0.301\ \text{Hart}$$

由式(2.1.3)可知，一个以等概率出现的二进制码元(0,1)所包含的自信息量为 1bit。因为当 $p(0) = p(1) = \dfrac{1}{2}$ 时，有

$$I(0) = I(1) = -\log_2 \frac{1}{2} = \log_2 2 = 1\ (\text{bit})$$

需要注意的是，信息量是纯数，信息量单位只是为了标示不同底数的对数值，并没有量纲的含义。

【例 2.1.1】 某地二月份天气的概率分布统计如下：

$$\begin{pmatrix} X \\ P(X) \end{pmatrix} = \begin{Bmatrix} a_1(\text{晴}), & a_2(\text{阴}), & a_3(\text{雨}), & a_4(\text{雪}) \\ \frac{1}{2}, & \frac{1}{4}, & \frac{1}{8}, & \frac{1}{8} \end{Bmatrix}$$

这四种气候的自信息量分别为 $I(a_1) = 1\text{bit}$，$I(a_2) = 2\text{bit}$，$I(a_3) = 3\text{bit}$，$I(a_4) = 3\text{bit}$。

容易证明，自信息量 $I(a_i)$具有下列性质。

（1）$I(a_i)$是非负值

定义式(2.1.3)中的 $p(a_i)$代表随机事件发生的概率，在闭区间[0,1]上取值。根据对数的性质，$\log_2 p(a_i)$ 为负值，故 $-\log_2 p(a_i)$ 恒为非负值。这一性质从对数的几何图形上也很容易理解（见