研究生教学用书

教育部研究生工作办公室推荐

信息论与编码

Information Theory & Coding

(第三版)

姜 丹 编著

中国科学技术大学出版社

研究生教学用书

教育部研究生工作办公室推荐

信息论与编码

Information Theory & Coding

（第三版）

姜 丹 编著

中国科学技术大学出版社

内 容 简 介

本书系统论述香农信息论的基本理论，介绍编码的基本方法. 全书共分 12 章. 内容包括：信息的定义、信息论的基本思路；单符号离散信源与信道、信息熵、互信息、信道容量、数据处理定理、加权熵、效用信息熵；多符号离散信源与信道、极限熵、独立并列信道的信道容量；连续信源与信道、相对熵、高斯白噪声加性信道的信道容量；无失真信源编码定理、霍夫曼编码方法；抗干扰信道编码定理、线性分组码、汉明码与扩展汉明码；限失真信源编码定理、信息率-失真函数、数据压缩原理、信息价值、广义信息率-失真函数；信源-信道编码定理；网络信息理论等.

本书可作为高等院校、科研院所相关专业的研究生、高年级本科生的教材或教学参考书，也可供从事信息理论、信息技术和信息科学的教学、科研和工程技术人员参考.

图书在版编目(CIP)数据

信息论与编码/姜丹编著. —3 版. —合肥：中国科学技术大学出版社，2009. 12
ISBN 978-7-312-02329-3

Ⅰ. 信… Ⅱ. 姜… Ⅲ. ①信息论 ②信源编码—编码理论 ③信道编码—编码理论
Ⅳ. TN911. 2

中国版本图书馆 CIP 数据核字(2008)第 061728 号

中国科学技术大学出版社 出版发行

(安徽省合肥市金寨路 96 号，邮编：230026)
合肥学苑印务有限公司印刷
全国新华书店经销

开本：787 mm×960 mm 1/16 印张：50. 75 字数：1222 千
2001 年 8 月第 1 版 2009 年 12 月第 3 版 2009 年 12 月第 5 次印刷
印数：18001—20000 册
定价：70. 00 元

第三版前言

作者根据《信息论与编码》(第二版)出版以来几年中的研究生教学实践和学科发展动态,对《信息论与编码》(第二版)的章、节结构和内容安排,又作了进一步的调整、充实和完善,现以《信息论与编码》(第三版),由中国科学技术大学出版社出版.

真诚欢迎广大读者对书中的错误和不当之处予以批评指正.

姜 丹

2008年8月于北京

再版前言

作者编著的《信息论与编码》一书，于2001年由中国科学技术大学出版社首次出版发行，并于2003年入选教育部研究生工作办公室推荐的“研究生教学用书”.

作者鉴于在研究生教学实践中的体会和经验，对首次出版的《信息论与编码》部分章、节作了修订，进一步充实、完善了教材的内容和结构. 现以《信息论与编码》(第二版)，由中国科学技术大学出版社再版发行.

真诚欢迎广大读者对书中的错误和不当之处予以批评指正.

姜 丹

2004年2月于北京

前　言

随着科学技术，特别是信息技术的发展，信息理论在通信领域中发挥着越来越重要的作用，显示出解决通信领域中有关问题的有力工具的本色. 同时，由于信息理论解决问题的思路和方法的独特、新颖和有效，在当今信息时代，信息理论已渗透到其它相关的自然科学，甚至社会科学领域，与电子技术、自动控制、计算机网络以及管理科学、生物医学工程、遗传工程、人工智能、心理学等学科密切结合，显示出它的勃勃生机和不可估量的发展前景. 信息论是信息科学中最成熟、最完整、最系统的重要组成部分，它是信息科学的发展起源与基石. 信息论是信息与通信学科的基础理论.

本书以香农(Claude E. Shannon)信息论为基础，论述近代信息理论的基本概念和主要结论.

作者鉴于近 30 年的教学经验，为了便于读者正确认识通信领域中信息的定义和本质，理解信息论解决问题的思路和方法，在“引言”中归纳、提炼出香农信息论的三大理论支柱. 为了便于读者建立信息流通的完整系统概念，把信息论的基础理论部分由传统的“信源一条线”、“信道一条线”的“纵向结构”，改变成由“单符号离散通信系统”(第一章、第二章)、“多符号离散通信系统”(第三章)、“单维连续通信系统”(第四章)、“多维连续通信系统”(第五章)等四个“横向教学板块”组成的“横向结构”，由简单到复杂、由浅入深、循序渐进地安排教学内容. 信息论是一门具有严密的数学演绎体系和高度抽象性、概括性的科学理论. 为了帮助读者排除学习信息论过程中经常遇到的数学分析方面的困难，结合有关内容，系统而简明地介绍必要的数学基础知识，给出导致重要结论的数学推演过程，提供不同的证明方法和途径. 为了帮助读者正确理解有关结论的物理含意，提供通俗易懂、富有哲理的诠释.

本书在全面系统地论述信息论基础理论的基础上，严密论证了“无失真信源编码定理”、“抗干扰信道编码定理”、“限失真信源编码定理”和“信源－信道编码定理”等信息论中的关键定理和结论，深入阐明剖析了信息率—失真函数的定义、数学特性及其内涵. 按照理论联系实际的原则，介绍“霍夫曼(Huffman)码”、“线性分组码”、“汉明(Hamming)码”和“扩展汉明码”等实际编码方法. 使读者既能掌握、理解信息论的总的结论，看到实现有效而可靠的通信系统的光明前景，又能掌握实现通信系统的“最优化”的某些实际编码方法和技巧.

本书对如何构建“加权熵”、“效用信息熵”；如何运用信息率—失真理论定义“信息价值”；如何凝练“信息率—失真函数”的数学精髓，构建“广义信息率—失真函数”，估算通信系统的有关指标界限等问题，进行了探索性的讨论. 以“多用户信道”的容量界限为重点，对网络信息传输的有关特性，作了初步探讨，给读者提供探究当今正在蓬勃兴起的互联网通信理论的初步基础知识.

本书的一个鲜明特色，是具有较强的理论性. 通篇贯穿了一条主轴，这就是用数学模型描述要讨论的问题，用严密的数学理论分析，导致讨论问题的结论，用完整而系统的数学推演论证定理.

本书可作为高等院校、科研院所相关专业的研究生(或高年级本科生)的教材或教学参考书,也可供从事信息理论、信息技术和信息科学的教学、科研以及工程技术人员参考.

作者在撰写本书过程中,得到中国科学院电子学研究所陈宗骞教授、中国科学院研究生院钱玉美教授的热情指导和帮助,在此一并表示衷心感谢.

热忱希望广大读者对书中的错误和不当之处予以批评指正.

姜 丹

2000 年 11 月于北京

目　录

(注:有*符号的章节和习题,可作为参考内容,不列入教学计划)

引　言

《信息论》是人们在长期通信实践活动中，由通信技术与概率论、随机过程、数理统计等学科相结合而逐步发展起来的一门新兴交叉学科，它是解决通信问题的理论基础和有力工具. 美国科学家香农(C. E. Shannon)于 1948 年发表的著名论文《通信的数学理论》(Claude E. Shannon, A Mathematical Theory of Communication)，奠定了《信息论》的理论基础.

香农信息论的起点与基石，是解决信息的度量问题.

人们经常习惯地把“消息”不加区分地就称之为“信息”，以致“信息”这个词汇得到越来越广泛的应用. 当人们收到由电话、传真、电子邮件、广播、电视等媒体传播的“消息”后，往往说成获得了“信息”. 一般以数字、数据、图表、曲线等形式出现的计算机通信、运算和处理所需要的条件、内容和结果，人们也习惯称之为“输入信息”、“输出信息”、“补充信息”、“反馈信息”等等. 人们通常也把由眼、耳、鼻和口等感觉器官直接感觉到的颜色、声音、气味、味道、气温、湿度等说成感知到了某种外界“环境信息”. 在这种把“消息”和“信息”不加区分地混为一谈的经验性的、习惯性的感性认识下，信息的度量将是十分困难的. 例如某人收到一封来信，谈的是同学最近的工作、学习情况. 同时，又收到一封家信，谈的是家人的健康情况. 显然，他从这两封信中都获得了信息. 但若要问：他从哪一封信中获得了更多的信息？ 也许，按某种想当然的感觉，他会给出某种模糊的回答. 如“家信中得到了更多的信息”. 这个结论可靠吗？ 就算这个结论不错，如进一步问：“家信中含有的信息比同学来信中含有的信息多了多少？”一般来说，很难回答这个问题.

那么，为什么在把“消息”和“信息”不加区分地混为一谈的情况下，信息度量就会显得十分困难呢？ 要回答这个问题，我们首先要来对“消息”作一番剖析. 众所周知，“消息”是用文字、符号、数据、语言、图片、音符、图像等能被人们的感觉器官所感知的形式，对客观物质运动和主观思维活动状态的一种表述，“消息”由形式、语义和语用三因素组成. 不同的“消息”，不仅有不同的形式，而且含有不同的语义和不同的语用效果. 例如“中国获得二〇〇八年第二十九届奥运会主办权”这条消息，它的“形式”，可看作是从汉字表中挑选 20 个字的一种选择，是 20 个汉字的一个时间序列. 在语义上，这条消息含有多个语义层次，构成复杂的语义层次结构：是“中国”，而不是“法国”、“加拿大”、“日本”和“土耳其”等其它国家；是“获得”，而不是“丢失”；是“二〇〇八年”，而不是“二〇〇〇”、“二〇〇四”和“二〇一二年”等其它时间；是“第二十九届”，而不是“第二十八届”、“第三十届”等其它届数；是“奥运会”，而不是“世界杯”、“世界锦标赛”和“公开赛”等其它赛事；是“主办权”，而不是“参赛权”、“电视转播权”等其它权利. 从语用效果上来看，它不仅与消息的语义内容有关，而且与消息的接收者的主观因素有关. 从电视实况转播中我们看到，当“萨马兰奇”主席宣布“第二十九届奥运会的主办城市是北京”这一消息时，在场的中国代表团成员情不自禁地跳跃欢呼，互相拥抱，热泪盈眶. 而在场的其它国家的代表团，绝大多数仍然坐在自己的座位上以鼓掌表示祝贺. 当然很可能有少数代表团会感到失落甚至沮丧. 你看，同一条消息，对不同的接收者来说，作出的反应的反差是如

此的巨大！由此可见，一条“消息”既有形式，又有语义，又有语用效果. 而且消息的形式、语义和语用这三个因素是捆绑在一起的，“消息”是其形式、语义和语用三因素互相交织在一起的一个混合体.

另一方面，要解决信息的度量问题，必然要用数学工具，进行数量运算. 我们知道，数学是刻画物质运动形式的工具，用数学对“消息”的形式进行刻画，不存在法则上的困难，但如何运用数学工具刻画“消息”所含有的语义乃至语用效果，至今仍然是一个巨大的难题.

由此可见，如按照经验性的思路，习惯地把“信息”和“消息”不加区分地混为一谈，把“消息”的形式、语义和语用三因素捆绑交织在一起，综合地解决信息度量问题，必然面临头绪纷繁、无从下手的僵局. 这就是前面提到的例子中，为什么很难说出“家信”和“同学来信”各自含有多少信息量，以及“家信”到底比“同学来信”多了多少信息量的根本原因之所在.

美国科学家香农以敏锐的观察力和新颖的方法、技巧，针对人类通信活动的特点，精辟地提出了“形式化假说”、“非决定论”、“不确定性”等三个观点，明确了“信息”的特定含意，阐明了“信息”和“消息”的联系和区别，提出了“信息”度量方法. 就此打破了僵局，跨出了用数学方法定量描述“信息”的关键一步. 开创了通信领域信息理论的新局面.

（一）形式化假说

通过对通信活动的基本功能的观察分析，香农指出，“通信的基本问题，是在消息的接收端精确地或近似地复制发送端所挑选的消息. 通常消息是有意义的，即是说，它按某种关系与某些物质或概念的实体联系着. 通信的语义方面的问题与工程问题是没有关系的.”这就是说，通信系统的任务只是在接收端把发送端发出的消息从形式上精确地或近似地复制出来，通信工程并不需要对复制出来的消息的语义作任何处理和判断. 对消息的语义内容的解读、处理和判断，是接收者自己的事，不是通信工程本身的任务，与通信工程无关. 至于消息的效用问题，那更应该是接收者自己的主观感受问题，与传送消息的通信系统无关. 正如邮递员一样，邮递员的职责只是把信送到收信者手中，至于对信的内容的解读乃至看了信后收信者是高兴还是愤怒，那是收信者自己的事，与邮递员无关. 又如电视实况转播一场精彩的 NBA 篮球比赛，电视转播系统的职责只是把画面和声音尽量精确地传播到每家的电视接收机的屏幕上，至于这场球赛中各运动员的表现，裁判的判罚是否公正等对这场球赛的解读和判断，那是收看电视节目的观众自己的事，与电视转播系统无关. 当然，看完球赛后观众的情绪是高兴还是愤怒，是热烈欢呼，还是气得把电视机从楼上摔下去等球赛的效用问题，更应该是观众自己的事，与电视转播系统无关. 这就是香农对通信活动的“形式化假说”.

这种通信工程的“形式化假说”，大胆地去掉了消息的语义和语用因素，巧妙地保留了能用数学描述的形式这一因素，这使应用数学工具定量描述信息成为可能，打开了信息理论进入科学殿堂的大门.

（二）非决定论

通过对通信活动的对象的分析研究，香农指出，“重要的是，一个实际的消息，总是从可能发生的消息的集合中选择出来的. 因此，系统必须设计得对每一种选择都能工作，而不是只适合工作于某一种选择. 因为各种消息的选择都是随机的，设计者事先无法知道什么时候会选择什么消息来传

送.”这就是说,一切有通信意义的消息的发生都是随机的,是事先无法预料的.消息传递过程中遇到的噪声干扰也是随机的,通信系统的工程设计者也是无法事先预料的.例如,面对公众的公用电话系统,不是针对某一特定对象设计的.什么人,什么时候来使用公用电话,以及通话人声音的最高频率、频带宽度、峰值功率、平均功率、持续时间等技术参数都是随机的,工程设计者都是无法事先预料的.显然,根据通信系统工程的这一特点,在设计工程时,不可能把某一特定的对象的技术参数作为设计的依据,而是要用概率论、随机过程、数理统计等数学工具,从大量的不可预料的随机消息(包括噪声)中,寻求其统计规律,作为通信工程设计的依据,用非决定论的观点和方法,来观察、描述信息.这就是香农的“非决定论”观点.

这种“非决定论”观点,是对通信活动的总的认识观.它从原则上回答了应采用什么类型的数学工具来解决信息度量问题.

(三) 不确定性

通过对通信活动的机制和作用的剖析研究,香农一针见血地指出“人们只在两种情况下有通信的需要.其一,是自己有某种形式的消息要告知对方,而估计对方‘不知道’这个消息;其二,是自己有某种‘疑问’要询问对方,而估计对方能作出一定的解答”.这里的所谓“不知道”、“疑问”,就是通信前对某事件可能发生的若干种结果不能作出明确的判断,存在某种知识上的“不确定性”.通信后,通过消息的传递,由原先的“不知道”到“知道”,或由“知之不多”到“知之甚多”;原先的“疑问”得到了解答,或部分解答,由原先的“疑问”到“明白”,或部分“明白”.这就是说,通信后,消除或部分消除了通信前存在的“不确定性”.所以,通信的作用就是通过消息的传递,使接收者从收到的消息中获取了一样“东西”,因而消除或部分消除了通信前存在的“不确定性”.这种“东西”,就是“信息”.这样,我们就有理由给“信息”下一个明确的定义:“信息就是用来消除不确定性的东西”,进而,可合理地推断:通信后接收者从收到的消息中获取的“信息”,在数量上等于通信前、后“不确定性”的消除量.这就是香农从“不确定性”观点出发,给“信息”下的明确的定义.

我们知道,“可能性”的大小在数学上可以用概率的大小来表示:概率大即表示出现的“可能性”大;概率小即表示出现的“可能性”小.我们同样知道,“不确定性”与“可能性”是有联系的:“可能性”大就意味着“不确定性”小;“可能性”小就意味着“不确定性”大.这样,“不确定性”就可与消息发生的概率联系起来.例如,“中国女子乒乓球队夺取2008年奥运会冠军”这条消息,根据中国女子乒乓球队历来的表现,夺取奥运会冠军的概率很大,即“可能性”很大,也就意味着“不确定性”很小.这个消息一旦发生,消除的“不确定性”也很小,收信者从这条消息中获取的信息量也很小.相反,“中国男子足球队夺取世界杯赛冠军”这条消息,根据中国男子足球队历来的表现,夺取世界杯赛冠军的概率很小,即“可能性”很小,也就意味着“不确定性”很大.若有朝一日这个消息真的发生了,消除的“不确定性”很大,收信者从这条消息中获取的信息量也很大,甚至惊喜万分、欢呼跳跃.由此可见,“不确定性”与消息发生的概率有内在联系,它应该是消息发生概率的某一函数.

根据香农关于信息的定义,通信后收信者从消息中获取的“信息”,从数量上等于通信前、后“不确定性”的消除.既然“不确定性”一定是消息发生概率的某一函数,那么,“不确定性”的“消除量”也一定是消息发生概率的某一函数.当然,通信后获取的信息量也应该是消息发生概率的某一函数.

对于随机消息来说，我们虽然不能精确预料它能否发生，但表示随机消息的可能性大小的“概率”一定是一个精准的数量. 所以，香农对“信息”的定义，从理论原则上完全解决了信息的度量问题.

香农从“不确定性”观点出发对“信息”的明确定义告诉我们，“信息”与“消息”两者之间既有联系，又有区别，两者不能混为一谈. “消息”是表达“信息”的形式，是载荷“信息”的客体；“信息”是“消息”统计特性的函数，是“消息”的抽象本质. 不同形式的“消息”，可能有相同数量的“信息”；相同形式的“消息”，可能有不同数量的“信息”. 信息论的研究对象不是具体的“消息”，而是抽象于各种不同形式的“消息”的“信息”. 专门讨论“信息”的产生、传输和处理规律的信息论，是一门具有高度抽象性和概括性的学科，是一门具有完整的数学理论演绎体系的学科.

对通信领域中的“信息”的特定含意的界定，以及由此导致的“信息”度量的思路和方法，是《信息论》完整理论体系的起点和基石. 树立《信息论》的“信息观”，明确“信息”与“消息”之间的联系和区别，是打开《信息论》这门学科的“大门”，并达到其“顶峰”的一把必备的钥匙. 这就是这段“引言”的结论，希望它能真正起到“引路入门”的作用.

信息论起源于通信领域，经过半个多世纪的充实、完善和提高，它已成为通信领域中一门完善、系统、成熟的重要基础理论学科. 同时，由于它的独到的思路和新颖的技巧，使它在与其它学科的相互渗透和结合上，也出现了不少可喜的成果. 随着科学技术的迅猛发展和人类文明的不断提高，必将显示出它的勃勃生机和光明的前景.

第一章　单符号离散信源

通信系统一般由信源、信道和信宿三部分组成(如图 1.1).“信源”就是信息的源泉.信息不是消息本身,但它又包含在消息之中.信源是由含有信息的消息组成的集合.若信源是由有限或无限可列个取值离散的符号(如文字、字母、数字等)组成的离散集合,则这种信源称为离散信源.又若一个符号就代表一个完整的消息,则这种离散信源又称为单符号离散信源.单符号离散信源是最简单的离散信源.

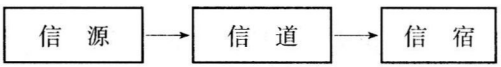

图 1.1

第一节　信源的数学模型

单符号离散信源中的某一符号要含有一定的信息,信源发这一符号必须具有随机性,必须以一定的概率分布发这一符号.单符号离散信源是具有一定概率分布的离散符号的集合.基于对信源的这种认识,可用一个离散随机变量的可能取值,表示信源可能发出的不同符号;用离散随机变量的概率分布,表示信源发出不同符号可能性的大小.总之,可用一个离散随机变量来代表和描述一个单符号离散信源.

例如,掷一个六面质地均匀的骰子,每次出现朝上一面的点数是随机的.如把出现朝上一面的点数作为这个随机试验的结果,并把试验的结果看作信源的输出消息,这个随机试验可看作是一个信源.这个信源输出有限种离散数字,其组成的集合为 A:{1,2,3,4,5,6},而且每一个数字代表一个完整的消息,这个信源是单符号离散信源.可用离散随机变量 X 来描述这个单符号离散信源:X 的可能取值表示信源可能发出的各种不同符号,其状态空间就是信源可能发出的各种不同符号组成的集合 A:{1,2,3,4,5,6};X 的概率分布表示信源发出各种不同符号的先验概率,其概率空间就是信源发出各种不同符号的先验概率组成的概率空间 P:$\left\{P(X=1)=\frac{1}{6},P(X=2)=\frac{1}{6},\cdots,P(X=6)=\frac{1}{6}\right\}$.所以,这个单符号离散信源的数学模型可完整地表示为

$$[X\cdot P]:\begin{cases}X: & 1 & 2 & 3 & 4 & 5 & 6\\ P(X): & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6}\end{cases}$$

因为[$X\cdot P$]完整地描述了信源 X 的信息特性,所以把[$X\cdot P$]称为信源 X 的“信源空间”.信源 X

输出的符号只可能是集合 $A:\{1,2,3,4,5,6\}$ 中的任何一种，不可能是集合 A 以外的其它任何符号. 信源 X 的概率空间是一个完备集，即有

$$P(X=1)+P(X=2)+\cdots+P(X=6)=1$$

在这个典型实例的启发下，可构建一般单符号离散信源的数学模型. 若某信源可能发出 r 种不同的符号 $a_1,a_2,\cdots,a_r$，相应的先验概率分别是 $p(a_1),p(a_2),\cdots,p(a_r)$. 用随机变量 X 表示这个信源，其信源空间可表示为

$$[X\cdot P]:\begin{cases} X: & a_1 & a_2 & \cdots & a_r \\ P(X): & p(a_1) & p(a_2) & \cdots & p(a_r) \end{cases} \tag{1.1}$$

其中

$$0\leqslant p(a_i)\leqslant 1 \quad (i=1,2,\cdots,r)$$

$$\sum_{i=1}^{r} p(a_i)=1 \tag{1.2}$$

不同信源对应不同的信源空间. 如信源给定，这就意味着相应的信源空间已经确定. 反之，如信源空间已经确定，这就意味着相应的信源已经给定. 用信源空间表示信源的数学模型的必要前提，就是信源可能发出的各种不同符号的概率先验可知，或事先可测定. 测定信源的概率空间是构建信源空间的关键. 例如，在一个箱子中，有红、黄、蓝、白四种不同颜色的彩球，它们的大小、质量和重量完全一样. 若从这个箱子中任意摸取出一个球，并把球的颜色当作试验的结果. 显然，这个随机试验就可看作是一个单符号离散信源，信源的输出符号集就是四种不同的颜色组成的集合 $A:\{$红，黄，蓝，白$\}$. 构建这个信源的信源空间的关键，在于测定出现各种不同颜色的概率. 在这个问题中，可把各种不同颜色的彩球的出现频率，近似地看作其出现的概率. 假如，箱子中共有 32 个球，其中：红球 16 个；黄球 8 个；蓝球和白球各 4 个. 则可得各种彩球出现频率，即各种彩球出现的先验概率分别为

$$\text{出现红球的概率}\quad P(\text{红})=\frac{16}{32}=\frac{1}{2}$$

$$\text{出现黄球的概率}\quad P(\text{黄})=\frac{8}{32}=\frac{1}{4}$$

$$\text{出现蓝球的概率}\quad P(\text{蓝})=\frac{4}{32}=\frac{1}{8}$$

$$\text{出现白球的概率}\quad P(\text{白})=\frac{4}{32}=\frac{1}{8}$$

若用随机变量 X 表示这个信源，其信源空间为

$$[X\cdot P]:\begin{cases} X: & \text{红} & \text{黄} & \text{蓝} & \text{白} \\ P(X): & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{cases}$$

用一个离散随机变量 X 代表一个单符号离散信源，这就是用数学描述信源，建立信源的数学模型的基本原则. 随机变量 X 的状态空间和概率空间，是信源空间$[X\cdot P]$的两个基本要素，而概率空间又是决定性要素. 概率可测是香农信息论的基本前提.