高等院校
通信与信息专业规划教材

信息论与编码基础

戴善荣　编著

机械工业出版社
CHINA MACHINE PRESS

本书以 Shannon 信息理论为依据，分基础篇、信道编码篇、信源编码篇、网络篇共 12 章讲述了信息与编码理论的基本概念、基本原理和在通信及信息工程等领域的应用。内容包括：信息的定义及度量、信源及其信息量、信道及其容量、分组码、卷积码、TCM 与 Turbo 码、离散信源无失真编码、限失真信源编码理论、信源编码实践、网络信息论初步和信息安全中的密码技术等。

本书可作为通信、计算机、信息工程等专业的教材或参考书，也可供信息领域科技工作者、工程技术人员参考。

图书在版编目（CIP）数据

信息论与编码基础/戴善荣编著．—北京：机械工业出版社，2004.10
高等院校通信与信息专业规划教材
ISBN 7-111-14887-8

Ⅰ.信… Ⅱ.戴… Ⅲ.①信息论-高等学校-教材②信源编码-编码理论-高等学校-教材③信道编码-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（2004）第 067921 号

机械工业出版社（北京市百万庄大街 22 号　邮政编码 100037）
责任编辑：陈振虹　版式设计：张世琴
责任校对：李秋荣　责任印制：李　妍
北京机工印刷厂印刷·新华书店北京发行所发行
2005 年 1 月第 1 版第 1 次印刷
787mm×1092mm 1/16·17 印张·415 千字
0 001—5 000 册
定价：24.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
本社购书热线电话（010）68993821、88379646
68326294、68320718

高等院校通信与信息专业规划教材
编委会名单

（按姓氏笔画排序）

出 版 说 明

为了培养21世纪国家和社会急需的通信与信息领域的高级科技人才，为了配合高等院校通信与信息专业的教学改革和教材建设，机械工业出版社会同全国在通信与信息领域具有雄厚师资和技术力量的高等院校，组成阵容强大的编委会，组织长期从事教学的骨干教师编写了这套面向普通高等院校的“高等院校通信与信息专业规划教材”，并且将陆续出版。

这套教材力求做到：专业基础课教材概念清晰、理论准确、深度合理，并注意与专业课教学的衔接；专业课教材覆盖面广、深度适中，不仅体现相关领域的最新进展，而且注重理论联系实际。

这套教材的选题是开放式的。随着现代通信与信息技术日新月异地发展，我们将不断更新和补充选题，使这套教材及时反映通信与信息领域的新发展和新技术。我们也欢迎在教学第一线有丰富教学经验的教师及通信与信息领域的科技人员积极参与这项工作。

由于通信与信息技术发展迅速，而且涉及领域非常宽，在这套教材的选题和编审过程中难免会有缺点和不足之处，诚恳希望各位老师和同学提出宝贵意见，以利于今后不断地改进。

机械工业出版社

高等院校通信与信息专业规划教材编委会

前　言

本书是机械工业出版社组织编写的“高等院校通信与信息专业规划教材”中的《信息论与编码基础》教材。本书内容的组织没有像专著那样分为信息理论与编码理论两个系统。而是综合成一个以形成高效、可靠、安全的信息传输（存储）码为目标，以 Shannon 关于信息的定义与质量为基础的 Shannon 信息论（或称狭义信息论）框架结构。该框架结构由 1 个共同的基础理论即 Shannon 信息理论，和相对独立的 3 个编码理论即信源编码理论、信道编码理论和密码理论所组成。这个框架基本上涵盖了以 Shannon 信息理论为基础的所有信息编码领域。在内容选材上还包括了新近正发展的、虽不成熟但已形成研究热点的课题，如网络信息论、Turbo 码等，以及为大众所关心的音视频信源编码的应用标准介绍。

信息理论与编码理论的发展与数学密不可分，而且曾是数学的一个专门化领域，其数学基础并非为一般本科大学生所具有。因此，虽有许多优秀的专著，但并不适合作教材。从已出版的教材中可以看到不同程度的普及化努力，这些给予了作者很好的借鉴。然而作者在教学实践中仍然觉得这是一个很棘手的问题，是本课程的一个特殊性。作为教材，要求必须有一定的数学理论高度，事实上数学是一个具有严谨性和推理功能的理论工具，用它可以准确、简洁、明了地表达出定理与科学结论。但如果使用太深的数学工具，就会使读者觉得抽象、不知所云与不可接受，而起反作用。本书所要求的数学知识基础为：普通高等数学、基础概率论（古典概率论）和一般的线性代数知识。没有引用数论与近世代数等知识，也没有引入典型序列概念。对一些定理与结论的证明会因此而发生困难，只好采取应用例证与物理概念解释相结合的办法。对于像公钥密码体制那样非应用数论知识不可的，则归纳出 14 条结论予以例证，然后应用这些结论推导公钥密码算法。

与框架结构相对应，本书由基础篇、信道编码篇、信源编码篇和网络篇共 12 章组成。其中基础篇 4 章，介绍 Shannon 信息论的形成、发展和框架结构；信息的度量；信源及其信息量；信道及其容量等关于信息的基本理论。这是 Shannon 信息理论的基本知识，也是本书其余各章的共同基础。

信道编码篇由分组码（第 5 章）、卷积码（第 6 章）和 TCM 与 Turbo 码（第 7 章）组成。分组码和卷积码在理论与技术上都比较成熟，而 TCM 与 Turbo 码是正在研究发展中的两种编码技术，理论上并未完善，技术上也未成熟，但却显示出各自的优异性能，被认为是 20 世纪末在信道编码领域中具有里程碑意义的两大成就。将它们编成一章，除了两者都应用了有反馈的卷积码这个共同点外，并无更多的理论与技术上的原因，两者都是信道编码领域中正在发展的颇受关注的新技术，这是更主要的考虑。

信源编码篇除了第 8 章离散信源无失真编码，第 9 章限失真信源编码理论外，还包括第 10 章信源编码实践。这是因为信源编码的理论与技术在微电子与计算技术的催化下，使音、视频系统数码技术的发展、应用成了 20 世纪 90 年代信息领域的一大亮点，引发了人们对信息技术的关注与兴趣。在学习信源编码基本理论的同时，了解一下这些理论的应用与标准对理解抽象的理论是有益的。

将第 11 章网络信息论初步和第 12 章信息安全中的密码技术合编成网络篇，唯一的理由由它们都是网络通信中有特殊意义的课题。应该说密码理论与技术并非以网络为依托的，但网络通信离开密码技术就会失去其实用意义。至于网络信息论，显然应该是网络发展的基础，但由于理论上的困难，至今研究成果甚少，与网络的高速发展极不相称。本章对已报导的一些成果作了初步的介绍，其意义不在于应用这些结果，更在于从中可了解、发现网络信息论研究中的困难与问题。

本书每章均附有习题，以加深对基本内容的理解，难度不高，只是书本内容的直接应用。在教学中，针对不同的专业需要，应增加一些结合专业的综合练习题。

作者要感谢在编写过程中参阅过的相关著作的作者，他们的著作给作者以很大的启发与借鉴，恕不能一一举名致谢。特别要感谢王育民教授、王新梅教授、郑志航教授以及美国的 R. B. Wells 教授。作者在编写本书的过程中曾参阅过他们的著作并应用了其中的某些资料。作者还要感谢徐澄圻教授对编写大纲及书稿所提出的宝贵建议与改进意见。

尽管作者在准确性与合理性方面作了努力，但疏漏之处终难避免，祈盼指正，以便不断改进。

作　　者

目　录

第 3 篇 信源编码篇

第4篇 网 络 篇

第1篇 基 础 篇

Shannon 信息论是以通信系统为基础，以概率统计为数学工具来研究信息及其交换的一门学科。Shannon 综合了前人及同时代的许多实践和理论研究成果，创立了信息研究的基础理论。本篇介绍的是信息理论中的基础内容，即关于信息的定义与度量，关于信源和信道的定性分类和定量表述。

本篇共分 4 章，第 1 章概述。介绍 Shannon 信息论的形成和发展，讨论信息论的组成与系统架构并给出本书的结构安排。第 2 章讲述信息的度量，给出自信息量、熵函数、互信息量和微分熵等的概念与测度。第 3 章与第 4 章分别讨论信源与信道的定性分类和定量表述。这些内容是本书其他各篇的共同基础，也是对信息科学具有奠基性作用的基础知识。

第1章 概 论

1.1 信息论的形成和发展

信息及其交换，几乎是伴随人类社会的产生、发展而同时进行的。然而，作为一门学科的“信息论”的建立，却是在 20 世纪 40 年代末。1948 年，C. E. Shannon 在 Bell 系统技术杂志（B. S. T. J.）上发表的著名论文“通信的数学原理”（A Mathematical Theory of Communication）是信息科学发展的起点。这篇关于现代信息论的开创性的权威著作，是在对通信系统研究的基础上论述信源和信道特性，给出了信息的定义和度量，从而为信息论的创立作出了独特的贡献。这篇论文的发表标志着信息论作为一门学科的诞生，Shannon 也因此被公认为是信息论的奠基人。

信息理论的创建与发展是以通信实践与理论为基础的。抛开对 19 世纪以前关于通信实践与理论成就的追溯，在 20 世纪 20 ~ 30 年代 Nyquist、Hartlay 和 Armstrong 的研究工作，可以说是为 Shannon 的论文作了前期工作，给予很大影响。Nyquist 在 1924 年 B. S. T. J. 杂志上发表的论文“影响电报速率的某些因素”（Certain Factors Affecting Telegraph Speed）阐明的信号带宽与信息速率之间的数量关系，至今仍然是通信系统设计的指导准则；1928 年 Hartlay 发表在 B. S. T. J. 上的论文“信息传输”（The Transmission of Information），最早研究了通信系统传输信息的能力，并引入了按等概事件定义信息量的概念；Amstrong 则于 1936 年对调频实验的研究中，得出“增大带宽可以提高抗干扰能力”的结论。20 世纪 40 年代，Shannon 在前人工作的基础上，用概率统计的方法研究通信系统，从而揭示了通信系统中传送的对象是信息，并对信息给以科学的定量描述，提出了信息熵的概念；揭示出通信系统设计的中心问题是在干扰噪声中如何有效而可靠地传送信息，指出可以用编码方法实现这一目

标；还从理论上证明了通信系统可以达到的最佳性能极限。

许多与 Shannon 同时期的学者，也采用概率统计方法研究信息，其中最著名的是被誉为控制论的奠基人 N. Wiener。他从控制论和从噪声中提取信息的最佳滤波器设计角度研究信息，他提出的最佳滤波理论也因此成为通信理论的一个重要分支。由此可见，Shannon 信息论的产生，是通信理论发展的必然产物，也是同时期学者的共同成果。

在 1948 年 Shannon 论文发表之后，也曾有过信息论价值的“泡沫”现象。当时，相关论文一哄而起，大学里也开设相关课程，对信息论能使通信系统达到最佳性能寄予厚望。然而，由于信息论只给出可能达到的性能极限，却没有给出系统设计的具体方法，厚望渐渐地变成失望。在经过初期的热望和随后的失望之后，人们对信息论的认识也逐渐趋向恰当。特别是被 Shannon 指出为达到有效与可靠通信目标的编码技术的研究进展，使信息论走向实用化的发展阶段。

从逻辑上讲是编码理论导致信息理论，而信息理论给出了编码所能达到的性能极限，又促进了编码技术的发展。这就使这两个原来独立发展的理论，变得密不可分，以致融合成同一学科，本书就反映了这一融合关系。

1.2 通信系统模型

信息论源于对通信系统的研究，又服务于通信系统。Shannon 给出的通信系统模型如图 1-1 所示。这个通信系统实际上包含了信息的传输与存储两大功能。从信源发出消息，经过编码器将消息变换成适合于传输或存储的信号形式，可靠而有效地送入信道。这里的信道实际上是一种媒介，如果系统是作为通信传输功能，则信道就可能是常见的电话线，光缆，无线电磁波…的一种；如果系统是用作存储功能，则信道就可能是磁盘，光盘，磁带，…的一种，它是将消息在时间域内从现在（存储时刻）传输到将来某时刻（回放时刻）的一种通信。译码器接收信道输出的信号并将信号反变换成原来的消息送到称为信宿的信息接受器。

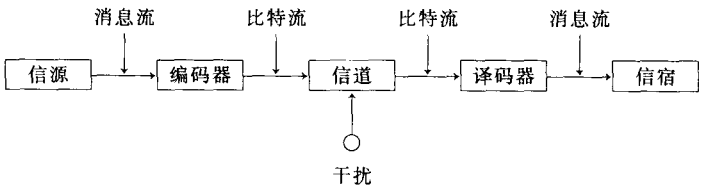

图 1-1 通信系统的基本模型

整个通信系统的要求就是希望恢复消息与原始发送消息一致，或控制失真在要求的范围之内，这就是可靠性的要求。此外，还希望作为信道的媒介质的利用率高一些。具体到传输媒介就是提高频带利用率；对存储媒介则希望能提高单位存储容量（信息容量）。这就是有效性的要求。信息论研究的主要问题是在通信系统设计中如何实现可靠性和有效性。而研究的对象就是图 1-1 中的编码器和译码器，编码器是信源编码器、纠错编码器和调制器的总称，它们的功能是将信源发出的消息变换成适合于信道传送的信号。译码器原则上是编码器的逆操作，但由于信号受到干扰而使正确译码变得很困难，并因此成为通信系统设计中的主

攻点。中心问题是研究各种可实现的解调和译码方法，并且要求有较高的性能价格比。因此，与干扰作斗争就成了通信工程师的基本使命。

信道研究的中心课题是研究信道的统计特性和它的传信能力。所谓统计特性其实就是干扰的统计特性，干扰的形式与特性有各种各样，如衰落、多径、码间串扰、非线性失真、…等等。它对信号的影响方式可分为两大类。一类是由外界引入的随机干扰，如天电干扰，设备内部噪声等，他们与信号统计无关，是一种客观存在的噪声。对传输信号的影响就是把干扰加入到信号中，称之为加性干扰；另一类干扰是由于信号在传播过程中，信道的物理条件发生了变化，如温度，电离层密度等的变化，引起信号参数的随机变化，如频率色散，幅度衰减，相位偏移等。这种影响是某些随机变量与信号相乘的结果，故称之为乘性干扰。加性干扰与乘性干扰的存在是影响信道传输能力的决定因素。

信宿是信息的接收者，可以是人或设备。应该说，信息的最终接收者大部分是人。因此，在信源编码技术中，人的生理、心理特性常常被用来作为压缩数据的条件。然而在Shannon信息论中，并没有计及人的主观因素对信息的作用。这也反映出Shannon信息论的局限性。尽管它对数字通信的发展起了巨大的推动与指导作用，至今尚无更完善的理论可以取代。然而在信息时代，信息的涵义与应用已远远超出了Shannon信息论所讨论的范围，更不是本节的通信系统模型所能描述的了。对信息概念和信息理论的进一步认识和发展也是历史发展的必然。本书的讨论，将仍以图1-1的基本模型为基础来理解信息和信息理论，从而认识通信科学中已获得成功应用的技术成就。

1.3 Shannon信息论的框架与本书的编排

1.3.1 Shannon信息论的框架结构

Shannon信息论的框架结构可用图1-2表示。

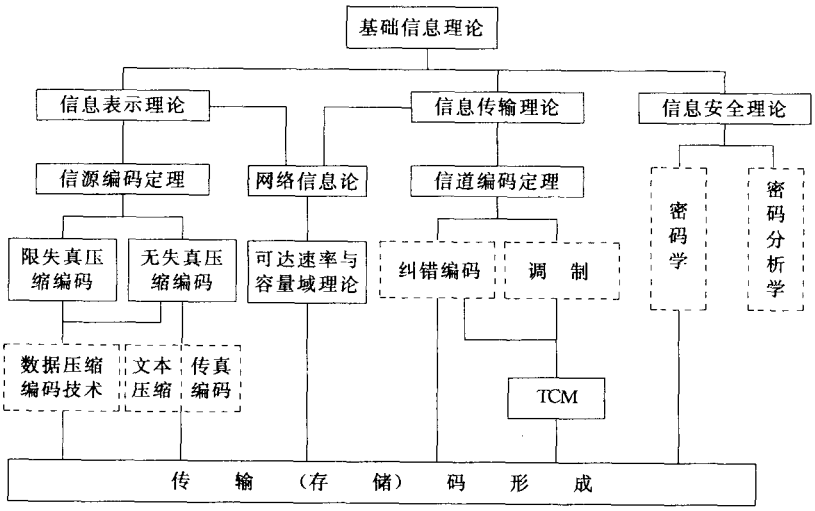

图1-2 Shannon信息论框架结构

图中在基础信息理论框之下，列出了信息表示理论、信息传输理论和信息安全理论三大块。这三大块都是 Shannon 信息论框架的支柱，各自都有独立的理论体系，相互间没有依赖关系。图中的虚线框表示该框所列的内容已发展成一门自成体系的学科。它们在理论上与技术上的成就都已超出基础信息论的范畴，本书只介绍其属于基础信息论的基本内容。框架结构表明，应用信息论技术的最终目标是形成传输或存储码，该码字将以高效、可靠的性能输入至信道传输或存储介质存储。这个过程，在当今的信息社会里，每时每刻都在大量地进行着。其范围之广几乎涉及到人类社会的各个方面，从家庭日常生活到宇宙航行，都离不开信息。同时，这个过程又是不被察觉地被人们所操作，因为它已被精心设计成各种简单易操作的产品。然而，作为一个科技工作者，则应该作为一种不可缺少的常识来学习、理解它。

信息表示理论是以 Nyquist 定理和信源编码定理为基础，将实际信源发出的信息表示成数字信号的各种理论与技术方法。其内容包括两大类：一类是以率-失真函数理论为基础的对模拟信源作数字化表示，又称为限失真压缩编码，因为该编码技术要在满足一定的保真度要求（即限失真）下，达到最低的数码率。在 20 世纪末，由于计算技术与微电子技术的高速发展，使各种复杂的算法的实时应用成为可能，数据压缩编码技术也就成了一门独立发展的学科，是各种数码技术与产品的基础。另一类是基于保熵编码定理对数字信源信息的无失真压缩编码。该技术根据数据符号发生的概率统计特性或符号的序列特性，用比较少的数据符号来表示原来的信源信息而不产生任何失真。在计算机数据的文本压缩与数字传真及图像编码中已广泛地采用了无失真压缩编码技术。

信息传输理论所关注的是信息传输的可靠性和信道频带的高效利用。这两方面都已发展为相对独立的理论体系，即纠错编码与调制理论。此两者联合起来统称为信道编码，其理论与技术都是基于信道编码定理而发展的。本书不包括调制理论，只在 TCM 介绍中略有涉及。

21 世纪的通信技术已发展到网络通信时代，基于点对点通信的单信道模型的基础信息理论，已满足不了多用户、多信道的通信技术要求，网络信息论或称多用户信息论的发展已形成强烈的需求。网络信息论仍以信息表示理论与信息传输理论为基础，所研究的基本问题也仍为信源编码速率与信道容量。然而由于信源与信道的多样性及相关性，使理论研究变得非常复杂困难。信源编码输出速率与信道容量的 Shannon 界不再是一个数，而是一个区域，分别称为可达速率域与容量域。Shannon 在 1961 年提出了两端双路通信系统（Two-Way Communication Channels）的论文，涉及了多用户信息理论中的许多基本概念，可以认为是网络信息论的起点。半个多世纪以来，虽然针对具体的信源与信道，比如多点接入与广播信道，得到了一些研究结果。但是，关于网络的一般信息理论远未建立起来，现有的一些结果也称不上成熟。因此，网络信息论是正在发展中的一个信息论分支，本书介绍的一些结果与思路，只能说是为读者提供一些启发性的材料。

网络时代所引发的另一个需求，就是信息安全。由于网络技术的发展和应用的普及，尤其是 Internet 在全球联通，使人类的活动，包括政治、经济、文化生活…等都被网络联系在一起。信息安全问题日益为人们所关注。Shannon 于 1949 年所建立的“保密系统的通信理论”，已发展成“密码学”与“密码分析学”这两门既相关又互相独立的学科。密码学中的以 DES（Data Encryption Standard）为代表的对称密码体制和以 RSA 算法为代表的公钥（非对称）密码体制在传输（存储）码形成中得到了切实有效的应用，成为信息安全的核心技术。由于密码学理论的数学工具要求较高，本书将放松其数学严谨性，而着重介绍原理及实

际应用。密码分析将不在本书介绍。

以上对 Shannon 信息论框架结构的基本内容与相互关系的概括性介绍，对初次接触信息论的读者来说，这种总体轮廓性的了解对具体内容的学习将是有益的。

1.3.2 本书的编排

基于图 1-2 所示的框架结构，本书分基础篇（4 章），信道编码篇（3 章），信源编码篇（3 章）和网络篇（2 章），共 12 章。其中基础篇的内容是其他各篇的理论基础和必备知识，其他 3 篇则相对地比较独立，相互之间没有结构上的联系，这也是采用分篇结构的原因。信道编码篇与信源编码篇的内容比较完整，各章之间联系也比较紧密，自成体系。然而网络篇的两章只是从应用特点出发放在同一篇中。就内容与理论基础而论，它们并无太多关联，编为一篇，似乎有点勉强，也算是结构形式上的需要吧。

采用分篇结构的另一个考虑是便于作选择性的阅读或教学实施。同是一门“信息论与编码基础”课，对不同专业、不同要求和不同水平的读者来说，内容会有不同的选择。除基础篇为必读外，其他 3 篇的次序与取舍都可自由选择而不会影响学习理解。因此，对篇目的排序号，只是一种编排形式，并无先后次序的含义。全书统一的章节序号，其前后次序只在同一篇目内有意义。比如不在同一篇目内的第 8 章与第 5 章，哪一章先读都不会影响理解。但在同一编码篇内的第 5，第 6，第 7 章还是按顺序阅读为好。

随着信息时代的到来，在计算技术与微电子技术高度发达的催化作用下，信息科学与技术已转化为庞大的信息产业，推动着整个社会的进步。与此同时，人们对信息科学与技术的兴趣与关注也日益增涨。为此，作者在选材与叙述方法上，更多地采用物理概念或基础数学工具来阐明抽象的关于信息的基础理论。希望本书既能满足大学本科的教学要求，又能满足信息时代带来的社会需求，为信息科学知识的普及出一点力。

第2章　信息的度量

2.1　离散变量的自信息量

2.1.1　消息、信息与概率空间

消息是人们很熟悉的词，但信息是什么呢？Shannon 等人把消息中含有不确定性的成份叫信息。让我们来分析一则有奖竞猜的新闻报导："第 17 届世界杯足球赛即将在日本和韩国举行，中国国家足球队首次参加。本报特组织"中国队能否进入十六强？冠军属谁家？"的有奖竞猜活动……。无疑这是一则消息，有信息吗？按 Shannon 关于信息的定义，要看是否有不确定性。在日本与韩国举行第 17 届世界杯，中国队首次参加，是早已确定了的事，不含有信息。中国队能否进入 16 强？冠军是谁？这完全是未知数，具有很大的不确定性，因此蕴涵很大的信息。要举办有奖竞猜活动，在见报之前，也是个未知因素，具有不确定性，故也有信息含量，但自见报开始它就是一个确知事件了。按 Shannon 关于信息的定义就不再有信息含量。应该强调的是以上关于信息的讨论是"对事不对人的"。举行世界杯、中国队首次参加、要举办有奖竞猜，这几件事因确定无疑而不再含有信息。当它们刚刚确定的时刻，因解除不确定性而失去信息的同时给人们带来了信息。然而信息的传播是需要媒体和时间的，因而每个人接受信息的方式与时间各不相同，可以听广播、看电视、读报纸……，也可以用语言交流。传播时间，可以是即时（实时通信）的，也可以是非实时的，甚至很久远。因此对某个人来说，这些经过时间域的传输已消除了不确定性的"过时消息"，只要他是从未知到已知，在消除不确定性的同时，仍然获得信息。这种个人感知信息，不属于 Shannon 信息所讨论的内容。

现在从人类感知特性来讨论消息与信息的关系。

消息是由媒体所表示的客观存在的一种事件或一种状态。这里的媒体是通常所遇到的文字、符号、声音、图像……等的通称；事件则是事物的存在、发生、变化……等的总称。消息是以媒体为载体的，总是可以为人们所感知。但消息不一定含有信息。只有当消息本身包含不确定性（如上例中中国能否进 16 强？冠军属谁家？）或该消息能消除某些不确定性（比如举办有奖竞猜，消除了是否举办有奖竞猜的不确定性）时，该消息才有信息。含有信息的消息是信息的载体，即信息是蕴涵于消息之中的一种数学抽象，不被人们所直接感知，人们感知的乃是含有信息的消息。这也是人们常常混淆消息与信息的缘故。另一方面，信息又是客观世界存在四大要素（物质、能量、时间和信息）之一，没有信息就没有世界，人类的生活时时都离不开信息。

可以用概率大小来描写含有不确定性的事件，比如中国队进入 16 强这个事件发生的可能性极小（即概率小），而概率小即不确定性大。一旦中国队进入 16 强这件事发生，会使中国乃至世界球迷觉得惊奇，也就产生很大信息。可见小概率事件的发生，产生大的信息量。于是，可以用事件及其发生概率联合，并用概率空间的数学术语来表述某事件发生所具

有的信息情况。

将所有含有信息的消息所代表的多种可能状态排列成一个集合 S 称为状态空间。每一种可能状态就是状态空间中的一个元素，用 E_i 表示，即

$$S = \{E_1, E_2, \cdots\cdots E_K\}$$

对每一个可能状态 E_i，赋以一个概率值 p_i 以表示可能性的大小，称为概率测度。状态空间与其对应的概率测度联合起来，构成一个概率空间，即

$$\begin{Bmatrix} S \\ P \end{Bmatrix} = \begin{Bmatrix} E_1, E_2, \cdots\cdots E_K \\ P_1, P_2, \cdots\cdots P_K \end{Bmatrix} \tag{2-1}$$

式(2-1)中的 p_i 应具有概率特性，即

$$p_i \geqslant 0 \quad i = 1, 2, \cdots, K$$

$$\sum_{i=1}^{K} p_i = 1 \tag{2-2}$$

概率空间也可以记作 $\{S, P\}$。用概率空间来定义和度量信息称为 Shannon 信息或概率信息。

2.1.2 离散变量的自信息量

用概率空间模型来表示信息源，就是一种离散信源模型。此时状态空间就是信源，其符号表用 $A = \{a_1, a_2, \cdots\cdots, a_k\}$ 表示，它是一个离散集合。各元素的概率测度集合为 $P = \{P_1, P_2, \cdots\cdots, P_K\}$，也称概率矢量。比如英文句子的符号表 $A = \{a, b, c, \cdots\cdots, z, \sqcup\}$，它含有 26 个字母及一个空格共 27 个字符，每个字符的出现是随机的，视句子而定，即具有不确定性。其统计概率记作 $p(\mathrm{a}), p(\mathrm{b}), \cdots\cdots$，用 X 记具有随机性的概率空间，则离散信源可记作

$$X = \begin{Bmatrix} a_1, a_2, \cdots\cdots, a_K \\ p_1, p_2, \cdots\cdots, p_K \end{Bmatrix} = \{A, P\}$$

其中

$$p_i \geqslant 0$$

$$\sum_{i=1}^{K} p_i = 1$$

如果随机变量 a_i 又是互相独立的，则 X 又叫离散无记忆信源，记作 d. m. s(discrete memoryless source)。

定义 2-1 用 $\log \dfrac{1}{p(a)}$ 来度量随机变量 a_i 自身具有的信息量，称作自信息量，记作

$$I(a_i) = \log \frac{1}{p(a_i)} = -\log p(a_i) \tag{2-3}$$

这是根据以上关于信息性质的讨论而作出的一种数学定义，它能定量地描述信息的一些特性，比如确定事件($p(a_i) = 1$)不含有信息；出现概率小的事件含有大的信息量；同时出现的两个事件的信息量可以相加等，正是式(2-3)所代表的数学特性。即 $p(a_i) = 1$ 时 $I(a_i) = 0$；$p(a_i)$ 减小则 $I(a_i)$ 增大；当 a_1, a_2 互相独立，又同时出现时 $I(a_1, a_2) = \log \dfrac{1}{p(a_1, a_2)} =$

$\log\frac{1}{p(a_1)p(a_2)}=\log\frac{1}{p(a_1)}+\log\frac{1}{p(a_2)}$。由于$0\leqslant p(a_i)\leqslant 1$，式(2-3)还反映出自信息量要么有，要么无，但不可能为负值的特性。这也是符合以上关于信息量的定性讨论结论的。

2.1.3 信息量单位

信息量的单位由对数的底来决定。最常用的关于信息量的度量是以 2 为底的。此时用 bit(比特)作单位，取二进制单位(binary unit)之意。二进制符号只有两种可能，平等地取其中之一的概率为$\frac{1}{2}$，任一符号的出现就具有 $\log_2 2=1\text{bit}$ 的信息量。

【例 2-1】 统计独立事件的联合事件自信息。

设 64 个完全一样的球，排成 8×8 矩阵，问随机摸到第 i 行第 j 列的球的自信息量为多少 bit?

解：可以用两种方法来解这个问题。

方法 1：视在 64 个点中摸取其中之一完全是独立随机事件，共有 64 种可能。摸取第 i 行第 j 列的概率为 $p_{ij}=\frac{1}{64}$，根据定义式(2-3)，可计算出该事件的自信息为

$$-\log_2\frac{1}{64}\text{bit}=6\text{bit}$$

方法 2：把行和列分别视作彼此独立事件，即在 8 行中刚好摸到第 i 个球的概率为 $p_i=\frac{1}{8}$，而 8 列中摸到第 j 列球的概率也为 $p_j=\frac{1}{8}$。该两件事同时发生，称为联合事件，根据概率论，独立事件的联合概率可用概率乘表示即 $p_{ij}=p_i\cdot p_j$，从而联合事件的自信息量为

$$-\log_2(p_{ij})=-\log_2(p_i,p_j)=-\log_2 p_i-\log_2 p_j=(3+3)\text{bit}=6\text{bit}$$

这就是前面讨论以对数函数来定义的信息量具有可加性的结果，即独立事件的联合事件自信息等于各事件自信息之和。

若以 e 为底，用 Nat(奈特)作单位，表示取自然对数。

若以 10 为底，用 Hart(哈特)作单位，这是由 Hartly 首先采用的。

应用换底公式 $\log_2 x=\log_r x/\log_r 2$(令 r 为 e 或 10)，可以得出与 bit 单位的关系为：1Nat = 1.443bit；1Hart = 3.322bit。比较常用的为 $\log_2 x=1.443\ln x$，即用自然对数来计算。

应该指出 bit 还用作二进制数的数位或简称位，它不同于信息单位“比特”。由于二元信息最为常见，比如开关的开与关；回答问题的是与否等等，用二进制数 0 与 1 来作相应的代表最为方便。即使多元信息，比如四元信息也常用两位二进制数 00，01，10，11 来分别代表四种不同状态。于是在编码与传输理论中常常应用 m 位二进制序列，并不表示该序列所代表的事件的信息量为 $m$bit。

【例 2-2】 m 位二进制序列的自信息量。

解：因为 m 位二进制数的可能状态为 2^m 种，比如 $m=3$，$2^3=8$ 即共有 000，001，010，…，111，8 种状态。随机地取其中一种的概率为 $P=\frac{1}{2^m}$，于是其中任一组出现都带来 $-\log_2 2^{-m}=m$bit 的信息量，正好等于序列长度。所以这个结果容易产生误会。应该强调，这里的 $m$bit 自