信息论与编码

XINXILUN YU BIANMA

曹雪虹　张宗橙　编

北京邮电大学出版社
www.buptpress.com

信息论与编码

曹雪虹　张宗橙　编

北京邮电大学出版社
·北　京·

内 容 简 介

本书重点介绍信息论的基本理论以及编码的理论和实现原理。在介绍了有关信息度量的基础上，重点讨论了无失真信源编码、限失真信源编码、信道编码和密码学中的理论知识及其实现原理。本书注重概念，采用通俗的文字，联系目前通信系统，用较多的例题和图示阐述基本概念、基本理论及实现原理，尽量减少繁杂的公式定理证明；在各章的最后还附有大量习题，便于加深理解。本书可作为理工科高等院校信息工程、通信工程及相关专业的本科学生教材，亦可供信息、通信、电子等有关专业的科技人员作为参考书。

图书在版编目(CIP)数据

信息论与编码/曹雪虹，张宗橙编 .—北京：北京邮电大学出版社，2001.8
ISBN 7-5635-0524-5

Ⅰ.信 … Ⅱ.①曹 …②张 … Ⅲ.①信息论②信息—编码 Ⅳ.TN911.2

中国版本图书馆 CIP 数据核字(2001)第 048272 号

出版发行：北京邮电大学出版社
网　　址：www.buptpress.com
社　　址：北京市海淀区西土城路 10 号(100876)
电话传真：010-62282185(发行部)/010-62283578(FAX)
E - mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：北京忠信诚胶印厂印刷
开　　本：787mm×1 092mm　1/16
印　　张：12
字　　数：308 千字
印　　数：3 001—6 000 册
版　　次：2001 年 8 月第 1 版　2002 年 1 月第 2 次印刷

ISBN　7-5635-0524-5/TP·53　　**定　价**：22.00 元

前　言

当前信息产业发展很快，需要大量从事信息、通信、电子工程类专业的人才，而《信息论与编码》是这些专业的基础课，必须掌握，它可以指导理论研究和工程应用。

由于《信息论与编码》这门课本身理论性很强，介绍的内容是信息论基础和编码理论，现有的一些教材除了介绍理论和公式外，都用了大量篇幅来证明这些理论和公式，这些用作研究生教材是比较适合的。

而作为电子、信息、通信工程的本科生及相关专业的工程技术人员，由于他们理论基础的不足以及实际应用的需要，不可能花很多精力去研读那些在他们看来是非常难懂而枯燥乏味的证明，迫切需要一本介绍有关信息理论的基本知识且与实际应用紧密联系的书籍，本书就是出于这样的目的而写。

本书注重基本概念，用较通俗的文字解释其物理意义，辅以一定的例题和图示说明，不再用繁杂的公式来证明这些早已是非常成熟的公理。联系当前实际通信技术来讲述，使读者研读本书后概念清晰，有目标地应用在实际工作中。

本书共6章，由曹雪虹主编。第5章由张宗橙编写，其余各章由曹雪虹编写。在编写过程中，得到了徐澄圻教授和胡建彰教授的大力帮助，在此表示衷心感谢。

限于编者的水平，书中不妥或谬误之处难免，殷切希望读者指正。

编　者

2001年6月

目　　录

第1章　绪 论 …… 1
1.1　信息论的形成和发展 …… 1
1.2　通信系统的模型 …… 3
第2章　信源及信源熵 …… 6
2.1　信源的描述和分类 …… 6
2.2　离散信源熵和互信息 …… 7
2.2.1　自信息量 …… 7
2.2.2　离散信源熵 …… 8
2.2.3　互信息 …… 10
2.2.4　数据处理中信息的变化 …… 12
2.2.5　熵的性质 …… 13
2.3　连续信源的熵和互信息 …… 15
2.3.1　连续信源熵 …… 15
2.3.2　最大熵定理 …… 17
2.4　离散序列信源的熵 …… 18
2.4.1　离散无记忆信源的序列熵 …… 18
2.4.2　离散有记忆信源的序列熵 …… 18
2.5　冗余度 …… 27
习 题 …… 29
第3章　无失真信源编码 …… 34
3.1　编码的定义 …… 34
3.2　定长编码定理 …… 37
3.3　变长编码定理 …… 40
3.4　最佳编码 …… 42
3.4.1　香农编码方法 …… 42
3.4.2　费诺编码方法 …… 43
3.4.3　哈夫曼编码方法 …… 44
习 题 …… 48
第4章　限失真信源编码 …… 51
4.1　平均失真和信息率失真函数 …… 51
4.1.1　失真函数 …… 51
4.1.2　平均失真 …… 52

4.1.3 信息率失真函数 $R(D)$ …… 53
4.1.4 信息率失真函数的性质 …… 55
4.2 离散信源和连续信源的 $R(D)$ 计算 …… 58
4.3 限失真信源编码定理 …… 60
4.4 常用信源编码方法简介 …… 61
4.4.1 游程编码 …… 61
4.4.2 算术编码 …… 62
4.4.3 矢量量化 …… 66
4.4.4 预测编码 …… 68
4.4.5 变换编码 …… 70
习 题 …… 73
第 5 章 信道编码 …… 75
5.1 信道模型和信道容量 …… 75
5.1.1 信道模型 …… 75
5.1.2 信道容量 …… 77
5.2 有扰离散信道的编码定理 …… 83
5.2.1 随机编码 …… 83
5.2.2 编码定理 …… 85
5.3 差错控制与信道编译码的基本原理 …… 88
5.3.1 差错控制的途径 …… 88
5.3.2 码距与纠、检错能力 …… 91
5.3.3 最优译码与最大似然译码 …… 92
5.4 线性分组码 …… 94
5.4.1 线性分组码基本概念 …… 94
5.4.2 生成矩阵和校验矩阵 …… 97
5.4.3 伴随式与译码 …… 101
5.4.4 循环码 …… 106
5.5 卷积码 …… 114
5.5.1 卷积码的基本概念和描述方法 …… 114
5.5.2 卷积码的最大似然译码——维特比算法 …… 120
5.5.3 卷积码的性能限与距离特点 …… 127
5.6 网格编码调制与级联码简介 …… 130
5.6.1 网格编码调制 …… 130
5.6.2 级联码简介 …… 136
习 题 …… 141
第 6 章 密码学 …… 145
6.1 密码学的基础知识 …… 145
6.1.1 密码学的基本概念 …… 145
6.1.2 密码学中的熵概念 …… 148

6.2 数据加密标准 DES …… 150
6.2.1 换位和替代密码 …… 150
6.2.2 DES 密码算法 …… 151
6.2.3 DES 密码的安全性 …… 155
6.2.4 DES 密码的改进 …… 157
6.3 国际数据加密算法 …… 158
6.3.1 算法原理 …… 158
6.3.2 加密解密过程 …… 159
6.3.3 算法的安全性 …… 161
6.4 公开密钥加密法 …… 161
6.4.1 公开密钥密码体制 …… 161
6.4.2 RSA 密码体制 …… 162
6.4.3 报文摘要 …… 164
6.5 模拟信号加密 …… 168
6.6 通信网络中的加密 …… 168
6.7 信息安全和确认技术 …… 169
6.7.1 信息安全的基本概念 …… 170
6.7.2 数字签名 …… 170
6.7.3 防火墙 …… 173
6.7.4 密码学在电子支付系统中的应用 …… 174
6.7.5 密码学在电子数据交换中的应用 …… 175
习 题 …… 175
附录:符号及含义 …… 176
部分习题参考答案 …… 179
参考文献 …… 183

第1章 绪 论

“信息”这个词相信大家不陌生，几乎每时每刻都会接触到。不仅在通信、电子行业，其他各个行业也都十分重视信息，所谓进入了“信息时代”。信息不是静止的，它会产生也会消亡，人们需要获取它，并完成它的传输、交换、处理、检测、识别、存储、显示等功能。研究这方面的科学就是信息科学，信息论是信息科学的主要理论基础之一。它研究信息的基本理论(Information Theory)，主要研究可能性和存在性问题，为具体实现提供理论依据。与之对应的是信息技术(Information Technology)，主要研究如何实现、怎样实现的问题。

通过本章的学习，可以了解下列问题：信息论的形成和发展；信息论研究的内容及信息的基本概念；并结合通信系统模型介绍模型中各部分的作用及编码的种类和研究内容。

1.1 信息论的形成和发展

信息论理论基础的建立，一般来说开始于香农(C. E. Shannon)研究通信系统时所发表的论文。随着研究的深入与发展，信息论具有了较为宽广的内容。

信息在早些时期的定义是由奈奎斯特(Nyquist, H.)和哈特莱(Hartley, L. V. R.)在20世纪20年代提出来的。1924年奈奎斯特解释了信号带宽和信息速率之间的关系；1928年哈特莱最早研究了通信系统传输信息的能力，给出了信息度量方法；1936年阿姆斯特朗(Armstrong)提出了增大带宽可以使抗干扰能力加强。这些工作都给香农很大的影响，他在1941～1944年对通信和密码进行深入研究，用概率论的方法研究通信系统，揭示了通信系统传递的对象就是信息，并对信息给以科学的定量描述，提出了信息熵的概念。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送信息以及实现这一目标的主要方法是编码等。这一成果于1948年以《通信的数学理论》(A mathematical theory of communication)为题公开发表。这是一篇关于现代信息论的开创性的权威论文，为信息论的创立作出了独特的贡献。香农因此成为信息论的奠基人。

50年代信息论在学术界引起了巨大的反响。1951年美国IRE成立了信息论组，并于1955年正式出版了信息论汇刊。60年代信道编码技术有较大进展，使它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，使分组码技术发展到了高峰，找到了大量可纠正多个错误的码，而且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Viterbi译码方法。

信源编码的研究落后于信道编码。香农1959年的文章(Coding theorems for a discrete source with a fidelity criterion)系统地提出了信息率失真理论，它是数据压缩的数学基础，为各种信源编码的研究奠定了基础。

到70年代，有关信息论的研究，从点与点间的单用户通信推广到多用户系统的研究。1972年盖弗(Cover)发表了有关广播信道的研究，以后陆续有关于多接入信道和广播信道模型的研究，但由于这些问题比较难，到目前为止，多用户信息论研究得不多，还有许多尚待

解决的课题。

信息论主要应用在通信领域，在含噪信道中传输信息的最优方法到今天还不十分清楚。特别是当数据的信息量大于信道容量的情况，更是毫无所知，这是经常遇到的情况。因为从信源提取的信息常常是连续的，也就是信号的信息含量为无限大。在一般信道中传输这样的信号，是不可能不产生误差的。引入信道容量和信息量的概念以后，这类问题就可以得到满意的解释，并可给出一个通信系统的最佳效果，这样就为设计通信系统提供了理论依据。

信息论是在信息可以量度的基础上，研究有效地和可靠地传递信息的科学，它涉及信息量度、信息特性、信息传输速率、信道容量、干扰对信息传输的影响等方面的知识。通常把上述范围的信息论称为狭义信息论，又因为它的创始人是香农，故又称为香农信息论。广义信息论则包含通信的全部统计问题的研究，除了香农信息论之外，还包括信号设计、噪声理论、信号的检测与估值等。当信息在传输、存储和处理的过程中，不可避免地要受到噪声或其它无用信号的干扰，信息理论就是为能可靠地有效地从数据中提取信息，提供必要的根据和方法。这就必须研究噪声和干扰的性质以及它们与信息本质上的差别，噪声与干扰往往具有按某种统计规律的随机特性，信息则具有一定的概率特性，如度量信息量的熵值就是概率性质的。因此，信息论、概率论、随机过程和数理统计学是信息论应用的基础和工具。

本书讲述的信息理论的基本内容是与通信科学密切相关的狭义信息论，涉及到信息理论中很多基本问题。例如：

(1) 什么是信息？如何度量信息？

(2) 在信息传输中，基本的极限条件是什么？

(3) 信息的压缩和恢复的极限条件是什么？

(4) 从环境中抽取信息的极限条件是什么？

(5) 设计什么样的设备才能达到这些极限？

(6) 实际上接近极限的设备是否存在？

在信息论和通信理论中经常会遇到信息、消息和信号这三个既有联系又有区别的名词。下面将它们的定义比较如下：

信息：信息是指各个事物运动的状态及状态变化的方式。人们从来自对周围世界的观察得到的数据中获得信息。信息是抽象的意识或知识，它是看不见、摸不到的。人脑的思维活动产生的一种想法，当它仍储存在脑子中的时候它就是一种信息。

消息：消息是指包含有信息的语言、文字和图像等，例如我们每天从广播节目、报纸和电视节目中获得各种新闻及其他消息。在通信中，消息是指担负着传送信息任务的单个符号或符号序列。这些符号包括字母、文字、数字和语言等。单个符号消息的情况，例如用 x_1 表示晴天，x_2 表示阴天，x_3 表示雨天。符号序列消息的情况，例如“今天是晴天”这一消息由 5 个汉字构成。可见消息是具体的，它载荷信息，但它不是物理性的。

信号：信号是消息的物理体现，为了在信道上传输消息，就必须把消息加载(调制)到具有某种物理特征的信号上去。信号是信息的载荷子或载体，是物理性的。如电信号、光信号等。

按照信息论或控制论的观点，在通信和控制系统中传送的本质内容是信息，系统中实际传输的则是测量的信号，信息包含在信号之中，信号是信息的载体。信号到了接收端(信息论里称为信宿)经过处理变成文字、语声或图像，人们再从中得到有用的信息。在接收端将

含有噪声的信号经过各种处理和变换，从而取得有用信息的过程就是信息提取，提取有用信息的过程或方法主要有检测和估计两类。载有信息的可观测、可传输、可存储及可处理的信号均称为**数据**。

信息的基本概念在于它的不确定性，任何已确定的事物都不含有信息。其特征有：

- 接收者在收到信息之前，对它的内容是不知道的，所以信息是新知识、新内容；
- 信息是能使认识主体对某一事物的未知性或不确定性减少的有用知识；
- 信息可以产生，也可以消失，同时信息可以被携带、贮存及处理；
- 信息是可以量度的，信息量有多少的差别。

1.2 通信系统的模型

图 1-2-1 是目前较常用的、也是较完整的通信系统模型，下面介绍模型中各个部分的作用及需要研究的核心问题。

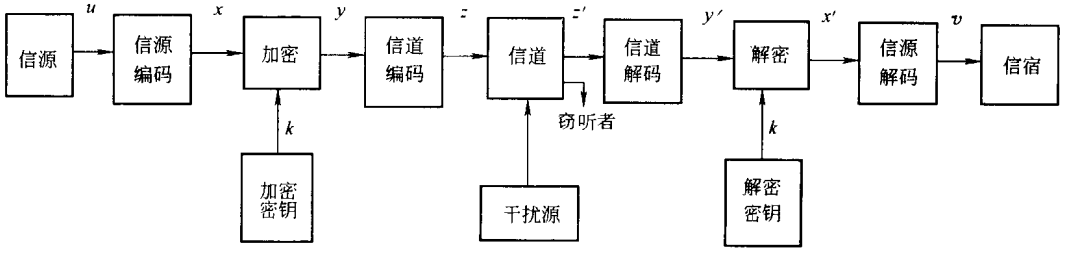

图 1-2-1 通信系统的物理模型

信源是向通信系统提供消息(u)的人和机器。信源本身是十分复杂的，在信息论中我们仅研究信源的输出。信源输出的是以符号形式出现的具体消息，它载荷信息。信源输出的消息可以有多种形式，但可归纳成两类：①离散消息，例如由字母、文字、数字等符号组成的符号序列或者单个符号。②连续消息，例如话音、图像、在时间上连续变化的电参数等。因为通信系统的接收者(信宿)在收到消息之前并不知道信源所发出消息的内容，所以一般地说信源发出的是随机性的消息。但因信源发出的消息都携带着信息，可见消息的变化是具有一定规律性的，因此严格地说信源发出消息并不是完全随机性的。信源的核心问题是它包含的信息到底有多少，怎样将信息定量地表示出来，即如何确定信息量。

信宿是消息传递的对象，即接收消息的人或机器。根据实际需要，信宿接收的消息(v)其形式可以与信源发出的消息(u)相同，也可以不相同，当两者形式不相同时，v 是 u 的一个映射。信宿需要研究的问题是能收到或提取多少信息。

信道是传递消息的通道，又是传送物理信号的设施。信道可以是一对导线、一条同轴电缆、传输电磁波的空间、一条光导纤维等传输信号的媒质。信道的问题主要是它能够传送多少信息的问题，即信道容量的大小。

干扰源是整个通信系统中各个干扰的集中反映，用以表示消息在信道中传输时遭受干扰的情况。对于任何通信系统而言，干扰的性质、大小是影响系统性能的重要因素。

密钥源是产生密钥 k 的源。信源编码器输出信号 x 经过 k 的加密运算后，就把明文 x

变换为密文y。若窃听者未掌握发端采用的密钥k,则他就很难从窃听到的信号z'解出明文。而收端的信宿则因知道事先已约定好的密钥k,因此能从收到的信号z'解出明文。对于二进制的代码而言,加密相当于$y=x\oplus p$运算(其中序列p通常是受密钥控制的伪随机序列),而解密则相当于$x'=y'\oplus p$运算。这里x',y',z'之所以不同于发端的x,y,z,是考虑到信号z在信道中传输时所受到的干扰影响。但在正常通信条件下,总会有$x'\approx x$,$y'\approx y$,$z'\approx z$的结果。

一般地说,通信系统的性能指标主要是有效性、可靠性、安全性和经济性。通信系统优化就是使这些指标达到最佳。除了经济性外,这些指标正是信息论的研究对象。根据信息论的各种编码定理和上述通信系统的指标,编码问题可分解为三类:信源编码、信道编码和密码。

信源编码器的作用是把信源发出的消息变换成由二进制码元(或多进制码元)组成的代码组,这种代码组就是基带信号。同时通过信源编码可以压缩信源的冗余度(即多余度),以提高通信系统传输消息的效率。信源编码可分为无失真信源编码和限失真信源编码。前者适用于离散信源或数字信号,后者主要用于连续信源或模拟信号,如语音、图像等信号的数字处理。从提高通信系统的有效性意义上说,信源编码器的主要指标是它的编码效率,即理论上能达到的码率与实际达到的码率之比。一般来说,效率越高,编译码器的代价也将越大。信源译码器的作用是把信道译码器输出的代码组变换成信宿所需要的消息形式,它的作用相当于信源编码器的逆过程。

信道编码器的作用是在信源编码器输出的代码组上有目的地增加一些监督码元,使之具有检错或纠错的能力。信道译码器具有检错或纠错的功能,它能将落在其检错或纠错范围内的错传码元检出或纠正,以提高传输消息的可靠性。信道编码包括调制解调和纠错检错编译码。信道中的干扰常使通信质量下降,对于模拟信号,表现在收到的信号的信噪比下降;对于数字信号,就是误码率增大。信道编码的主要方法是增大码率或频带,即增大所需的信道容量。这恰与信源编码相反。

密码学是研究如何隐蔽消息中的信息内容,使它在传输过程中不被窃听,提高通信系统的安全性。将明文变换成密文,通常不需要增大信道容量,例如在二进码信息流上叠加一密钥流;但也有些密码要求占用较大的信道容量。

在实际问题中,上述三类编码应统一考虑来提高通信系统的性能。这些编码的目标往往是相互矛盾的。提高有效性必须去掉信源符号中的冗余部分,此时信道误码会使接收端不能恢复原来的信息,也就是必须相应提高传送的可靠性,不然会使通信质量下降;反之,为了可靠而采用信道编码,往往需扩大码率,也就降低了有效性。安全性也有类似情况。编成密码,有时需扩展码位,这样就降低有效性;有时也会因失真而使授权用户无法获得信息,必须重发而降低有效性,或丢失信息而降低可靠性。从理论方面来说,若能把三种码合并成一种码来编译,即同时考虑有效、可靠和安全,可使编译码器更理想化,在经济上可能也更优越。这种三码合一的设想是当前众所关心的课题,但因理论上和技术上的复杂性,要取得有用的结果,还是相当困难。值得注意的是信息论分析的问题是存在性问题,即符合条件的编码是存在的,但并没有给出如何去寻找。

本书用了大量篇幅讨论编码问题,着重介绍信源和信道的编码定理,主要从概念上解释了这些定理的结论,并没有从严格意义上加以证明。顺便指出,不是所有的通信系统都采用

如图 1-2-1 所示的那样全面的技术。例如，点对点的有线电话，只要有一对电话机和一条电话线路（铜线）就够了，话音基带信号通过电话机变成相应的电信号（模拟信号），就能在电话线上传送，收端的电话机再把电信号恢复成人耳能听得清的话音。如果是点对点的无线电话，则在发端需要一台发信机，把模拟信号调制到射频上，再用大功率发射机经天线发射出去，然后在无线信道中传输；收端则应使用收信机把收到的调制射频信号解调恢复为发端的原始话音。若在这样的系统中增加加密和解密装置，就构成无线保密通信系统。在干扰大、信道容量有限的通信系统中，就需要采用信源编码和信道编码技术，以提高传输消息的有效性和可靠性。

这里首先举几个例子来说明编码的应用，如电报常用的莫尔斯码就是按信息论的基本编码原则设计出来的；在一些商品上面有一张由粗细条纹组成的标签，从这张标签可以得知该商品的生产厂家、生产日期和价格等信息，这些标签是利用条形码设计出来的，非常方便，非常有用，应用越来越普遍；计算机的运算速度很高，要保证它几乎不出差错，相当于要求有 100 年的时间内不得有一秒钟的误差，这就需要利用纠错码来自动地及时地纠正所发生的错误；每出版一本书，都给定一个国际标准书号（ISBN），大大方便图书的销售、编目和收藏工作。可以说，人们在日常生活和生产实践中，正在越来越多地使用编码技术。

本书的内容安排如下：

第 2 章介绍信息论的一些基本概念，包括自信息量、条件自信息量、互信息量、条件互信息量、平均互信息量、单符号熵、熵的性质以及连续信源熵、最大熵定理和随机序列的熵等，并解释了冗余度的由来及作用。

第 3 章介绍无失真信源编码定理，包括定长编码定理和变长编码定理，并详细阐述了最佳编码中的香农码、费诺码和霍夫曼码的编码方法及其性能比较。

第 4 章主要介绍了失真函数和信息率失真函数的定义及性质，简述了限失真信源编码定理。最后还简单提及了常用的几种信源编码方法。

第 5 章介绍信道及信道编码，其中包括信道、信道容量等基本概念，以及信道编码定理，还介绍了差错控制与信道编译码的基本原理及线性分组码、卷积码、级联码的基本原理。

第 6 章在介绍密码体制的基础知识及其熵的概念后，简述了具有代表性的秘密密钥加密算法 DES，IDEA 和公开密钥加密算法 RSA 和 MD5 等。还引入了信息安全性概念以及数字签名、防火墙等技术。

第 2 章　信源及信源熵

2.1　信源的描述和分类

在信息论中，信源是发出消息的源，信源输出以符号形式出现的具体消息。如果符号是确定的而且预先是知道的，那么该消息就无信息可言。只有当符号的出现是随机的，预先无法确定，一旦出现某个符号就给观察者提供了信息。因此可用随机变量或随机矢量来表示信源，运用概率论和随机过程的理论来研究信息，这就是香农信息论的基本点。

实际应用中分析信源所采用的方法往往依信源特性而定。按照信源发出的消息在时间上和幅度上的分布情况可将信源分成**离散信源**和**连续信源**两大类。**离散信源**是指发出在时间和幅度上都是离散分布的离散消息的信源，如文字、数字、数据等符号都是离散消息。**连续信源**是指发出在时间和幅度上都是连续分布的连续消息（模拟消息）的信源，如语言、图像、图形等都是连续消息。

下面来分析离散情况。离散信源可进一步分类：

离散信源
- 离散无记忆信源
  - 发出单个符号的无记忆信源
  - 发出符号序列的无记忆信源
- 离散有记忆信源
  - 发出符号序列的有记忆信源
  - 发出符号序列的马尔可夫信源

发出单个符号的信源是指信源每次只发出一个符号代表一个消息；**发出符号序列的信源**是指信源每次发出一组含二个以上符号的符号序列代表一个消息。**离散无记忆信源**所发出的各个符号是相互独立的，发出的符号序列中的各个符号之间没有统计关联性，各个符号的出现概率是它自身的先验概率。**离散有记忆信源**所发出的各个符号的概率是有关联的。这种概率关联性可用两种方式表示，一种是用信源发出的一个符号序列的整体概率（即联合概率）反映有记忆信源的特征，这就是上图中**发出符号序列的有记忆信源**。一般情况下，表述有记忆信源要比表述无记忆信源困难得多，尤其当记忆长度很长甚至无限长时。在实际问题中，我们往往试图限制记忆长度，即某一个符号出现的概率只与前面一个或有限个符号有关，而不依赖更前面的那些符号，这样的信源可以用信源发出符号序列内各个符号之间的条件概率来反映记忆特征，这就是**发出符号序列的马尔可夫信源**。

例如一个离散信源发出的各个符号消息的集合为 $X=\{x_1,x_2,\cdots,x_n\}$，它们的概率分别为 $P=\{p(x_1),p(x_2),\cdots,p(x_n)\}$，$p(x_i)$称为符号 x_i 的**先验概率**。通常把它们写到一起，称为概率空间：

$$\begin{bmatrix} X \\ P \end{bmatrix}=\begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ p(x_1) & p(x_2) & \cdots & p(x_n) \end{bmatrix}$$

显然有 $p(x_i)\geqslant 0$，$\sum_{i=1}^{n}p(x_i)=1$。

最简单的有记忆信源是 $N=2$ 的情况，此时信源 $X=X_1X_2$，其信源的概率空间为

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} a_1a_1 & a_1a_2 & \cdots & a_qa_q \\ p(a_1a_1) & p(a_1a_2) & \cdots & p(a_qa_q) \end{bmatrix}$$

对于无记忆信源联合概率为 $p(x_1,x_2,\cdots,x_n)=p(x_1)p(x_2)\cdots p(x_n)$，当进一步满足平稳性时，$p(x_1,x_2,\cdots,x_n)=p(x_1)p(x_2)\cdots p(x_n)=p^n$。

在分析有记忆信源时，有时也可将多个符号合并成一个符号来处理。例如有 L 个符号，每个符号取值于 A 空间，有 n 种可能性。将这 L 个符号组成一个 L 维随机矢量，则该随机矢量取值于 A^L 空间，共 n^L 个可能的取值，这样就把有记忆的 L 个符号的信源转化成单符号问题。

2.2 离散信源熵和互信息

2.2.1 自信息量

一个随机事件的**自信息量**定义为其出现概率对数的负值。即

$$I(x_i)=-\log p(x_i) \tag{2-2-1}$$

因为概率 $p(x_i)$ 越小，x_i 的出现就越稀罕，一旦出现，所获得的信息量应是较大的。由于 x_i 是随机出现的，它是 X 的一个样值，所以是一个随机量。而 $I(x_i)$ 是 x_i 的函数，它必须也是一个随机量。

自信息量的单位与所用的对数底有关。在信息论中常用的对数底是 2，信息量的单位为比特(bit)；若取自然对数，则信息量的单位为奈特(nat)；若以 10 为对数底，则信息量的单位为笛特(det)。这三个信息量单位之间的转换关系如下：

$$1\ \text{nat} = \log_2 e \approx 1.433\ \text{bit},\quad 1\ \text{det} = \log_2 10 \approx 3.322\ \text{bit}$$

如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：

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

若是一个 m 位的二进制数，因为该数的每一位可从 0,1 两个数字中任取一个，因此有 2^m 个等概率的可能组合。所以 $I=-\log_2\frac{1}{2^m}=m\,\text{bit}$，就是需要 m 比特的信息来指明这样的二进制数。

这里要引入随机事件的**不确定度**概念。根据日常知识，各个出现概率不同的随机事件所包含的不确定度是有差别的。一个出现概率接近于 1 的随机事件，发生的可能性很大，所以它包含的不确定度就很小。反之，一个出现概率很小的随机事件，很难猜测在某个时刻它能否发生，所以它包含的不确定度就很大。若是确定性事件，出现概率为 1，则它包含的不确定度为 0。注意：**随机事件的不确定度在数量上等于它的自信息量，两者的单位相同，但含义却不相同**。具有某种概率分布的随机事件不管发生与否，都存在不确定度，不确定度表征了该事件的特性，而自信息量是在该事件发生后给予观察者的信息量。

若有两个消息 x_i,y_j 同时出现，可用联合概率 $p(x_iy_j)$ 来表示，这时的联合自信息量定义为

$$I(x_iy_j)=-\log p(x_iy_j) \tag{2-2-2}$$

当 x_i 和 y_j 相互独立时，有 $p(x_iy_i)=p(x_i)p(y_j)$，那么就有 $I(x_iy_j)=I(x_i)+I(y_j)$。x_iy_j

所包含的不确定度在数值上也等于它们的自信息量。

若两个消息出现不是独立的，而是有相互联系的，这时可用条件概率 $p(x_i/y_j)$ 来表示，即在事件 y_j 出现的条件下，随机事件 x_i 发生的条件概率，则它的条件自信息量定义为条件概率对数的负值：

$$I(x_i/y_j) = -\log p(x_i/y_j) \tag{2-2-3}$$

在给定 y_j 条件下，随机事件 x_i 所包含的不确定度在数值上与条件自信息量相同，但两者含义不同。

由于一个随机事件的概率和条件概率总是在闭区间[0,1]内，所以自信息量和条件自信息量均为非负值。

例 2-2-1 英文字母中“e”的出现概率为 0.105，“c”的出现概率为 0.023，“o”的出现概率为 0.001。分别计算它们的自信息量。

解： “e”的自信息量 $I(\mathrm{e}) = -\log_2 0.105 = 3.25$ bit

“c”的自信息量 $I(\mathrm{c}) = -\log_2 0.023 = 5.44$ bit

“o”的自信息量 $I(\mathrm{o}) = -\log_2 0.001 = 9.97$ bit

2.2.2 离散信源熵

例 2-2-2 一个布袋内放 100 个球，其中 80 个球是红色的，20 个球是白色的，若随机摸取一个球，猜测其颜色，求平均摸取一次所能获得的自信息量。

这一随机事件的概率空间为

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ 0.8 & 0.2 \end{bmatrix}$$

其中 x_1 表示摸出的球为红球事件，x_2 表示摸出的球是白球事件。

这是一个随机事件试验。试验结果，当被告知摸出的是红球，则获得的信息量是

$$I(x_1) = -\log_2 p(x_1) = -\log_2 0.8 \text{ bit}$$

当被告知摸出的是白球，那么获得的信息量是

$$I(x_2) = -\log_2 p(x_2) = -\log_2 0.2 \text{ bit}$$

如果每次摸出一个球后又放回袋中，再进行下一次摸取。那么如此摸取 n 次，红球出现的次数为 $np(x_1)$ 次，白球出现的次数为 $np(x_2)$ 次。随机摸取 n 次后总共所获得的信息量为

$$np(x_1)I(x_1) + np(x_2)I(x_2)$$

而平均随机摸取一次所获得的信息量则为

$$\begin{aligned} H(X) &= \frac{1}{n}[np(x_1)I(x_1) + np(x_2)I(x_2)] \\ &= -[p(x_1)\log_2 p(x_1) + p(x_2)\log_2 p(x_2)] \\ &= -\sum_{i=1}^{2} p(x_i)\log_2 p(x_i) = 0.72 \text{ 比特/次} \end{aligned}$$

从此例可以看出，自信息量 $I(x_1)$ 和 $I(x_2)$ 只是表征信源中各个符号的不确定度，一个信源总是包含着多个符号消息，各个符号消息又按概率空间的先验概率分布，因而各个符号的自信息量就不同。所以自信息量不能作为信源总体的信息量，而上面求出的平均自信息

量，也即信息熵 $H(X)$是从平均意义上来表征信源的总体特征，可以表征信源的平均不确定性。我们定义信源的**平均不确定度** $H(X)$为信源中各个符号不确定度的数学期望。即

$$H(X) = E[I(X)] = \sum_i p(x_i)I(x_i) = -\sum_i p(x_i)\log p(x_i) \qquad (2\text{-}2\text{-}4)$$

单位为比特/符号或比特/符号序列。

因为 X 中各符号 x_i 的不确定度 $I(x_i)$为非负值，$p(x_i)$也是非负值，且 $0 \leqslant p(x_i) \leqslant 1$，故信源的平均不确定度 $H(X)$也是非负量。平均不确定度 $H(X)$的定义公式与热力学中熵的表示形式相同，所以又把 $H(X)$称为信源 X 的熵。熵是在平均意义上来表征信源的总体特性的。正如不确定度与自信息量的关系那样，**信源熵**是表征信源的平均不确定度，**平均自信息量**是消除信源不确定度时所需要的信息的量度，即收到一个信源符号，全部解除了这个符号的不确定度。或者说获得这样大的信息量后，信源不确定度就被消除了。两者在数值上相等，但含义不同。某一信源，不管它是否输出符号，只要这些符号具有某些概率特性，必有信源的熵值；这熵值是在总体平均上才有意义，因而是一个确定值，一般写成 $H(X)$，X 是指随机变量的整体(包括概率分布)。而另一方面，信息量则只有当信源输出符号而被接收者收到后，才有意义，这就是给予接收者的信息度量，这值本身也可以是随机量，也可以与接收者的情况有关，如考虑信息的有用性时就是如此。

在(2-2-4)式中，当某一符号 x_i 的概率 p_i 为零时，$p_i\log p_i$ 在熵公式中无意义，为此规定这时的 $p_i\log p_i$ 也为零。当信源 X 中只含一个符号 x 时，必定有 $p(x) = 1$，此时信源熵 $H(X)$为零。

例 2-2-3 电视屏上约有 $500 \times 600 = 3 \times 10^5$ 个格点，按每点有 10 个不同的灰度等级考虑，则共能组成 $n = 10^{3\times 10^5}$个不同的画面。按等概计算，平均每个画面可提供的信息量为

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) = -\log_2 10^{-3\times 10^5}$$

$$= 3 \times 10^5 \times 3.32 \approx 10^6 \text{ 比特/画面}$$

另外，有一篇千字文章，假定每字可从万字表中任选，则共有不同的千字文

$$N = 10000^{1000} = 10^{4000}\text{篇}$$

仍按等概计算，平均每篇千字文可提供的信息量为

$$H(X) = \log_2 N = 4 \times 10^3 \times 3.32 \approx 1.3 \times 10^4 \text{ 比特/千字文}$$

可见，“一个电视画面”平均提供的信息量要丰富得多，远远超过“一篇千字文”提供的信息量。

例 2-2-4 设信源符号集 $X = \{x_1, x_2, x_3\}$，每个符号发生的概率分别为 $p(x_1) = 1/2$，$p(x_2) = 1/4$，$p(x_3) = 1/4$，则信源熵为

$$H(X) = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 = 1.5 \text{ 比特/符号}$$

例 2-2-5 二元信源是离散信源的一个特例。该信源 X 输出符号只有两个，设为 0 和 1。输出符号发生的概率分别为 p 和 q，$p + q = 1$。即信源的概率空间为

$$\begin{pmatrix} X \\ P \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ p & q \end{pmatrix}$$

根据(2-2-4)式可得二元信源熵为

$$H(X) = -p\log p - q\log q$$

$$= -p\log p - (1-p)\log(1-p) = H(p)$$

信源信息熵 $H(X)$是概率 p 的函数，通常用 $H(p)$表示。p 取值于[0,1]区间。$H(p)$函数曲线如图 2-2-1 所示。从图中看出，如果二元信源的输出符号是确定的，即 $p=1$ 或 $q=1$，则该信源不提供任何信息。反之，当二元信源符号 0 和 1 以等概率发生时，信源熵达到极大值，等于 1 比特信息量。

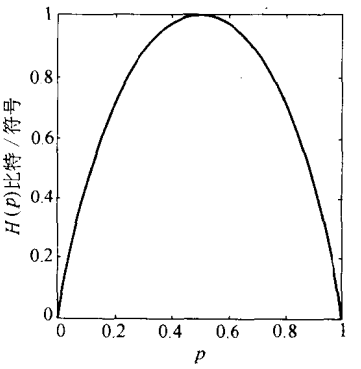

图 2-2-1　熵函数 $H(p)$

在给定 y_j 条件下，x_i 的条件自信息量为 $I(x_i/y_j)$，X 集合的条件熵 $H(X/y_j)$为

$$H(X/y_j) = \sum_i p(x_i/y_j) I(x_i/y_j)$$

进一步在给定 Y(即各个 y_j)条件下，X 集合的条件熵 $H(X/Y)$定义为

$$H(X/Y) = \sum_j p(y_j) H(X/y_j) = \sum_{i,j} p(y_j)\, p(x_i/y_j) I(x_i/y_j)$$
$$= \sum_{i,j} p(x_i y_j) I(x_i/y_j) \qquad (2\text{-}2\text{-}5)$$

即**条件熵**是在联合符号集合 XY 上的条件自信息量的联合概率加权统计平均值。条件熵 $H(X/Y)$表示已知 Y 后，X 的不确定度。

相应地，在给定 X(即各个 x_i)条件下，Y 集合的条件熵 $H(Y/X)$定义为

$$H(Y/X) = \sum_{i,j} p(x_i y_j) I(y_j/x_i) = -\sum_{i,j} p(x_i y_j)\log p(y_j/x_i) \qquad (2\text{-}2\text{-}6)$$

联合熵是联合符号集合 XY 上的每个元素对 $x_i y_j$ 的自信息量的概率加权统计平均值，定义为

$$H(XY) = \sum_{i,j} p(x_i y_j) I(x_i y_j) = -\sum_{i,j} p(x_i y_j)\log p(x_i y_j) \qquad (2\text{-}2\text{-}7)$$

联合熵 $H(XY)$表示 X 和 Y 同时发生的不确定度。联合熵 $H(XY)$与熵 $H(X)$及条件熵 $H(Y/X)$之间存在下列关系：

$$\left.\begin{aligned} H(XY) &= H(X) + H(Y/X) \\ H(XY) &= H(Y) + H(X/Y) \end{aligned}\right\} \qquad (2\text{-}2\text{-}8)$$

2.2.3　互信息

最简单的通信系统模型，如图 2-2-2 所示，X 是信源发出的离散符号集合，Y 是信宿收到的符号集合。由于信宿事先不知道信源在某一时刻发出的是哪一个符号，所以每个符号消息是一个随机事件。信源发出符号通过有干扰的信道传递给信宿。通常信宿可以预先知道信息 X 发出的各个符号消息的集合，以及它们的概率分布，即预知信源 X 的先验概率 $p(x_i)$。当信宿收到一个符号消息 y_j 后，信宿可以计算信源各消息的条件概率 $p(x_i/y_j)$，$i=1,2,\cdots,N$，这种条件概率称为**后验概率**。互信息量定义为后验概率与先验概率比值的对数，即

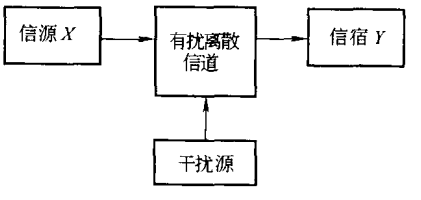

图 2-2-2　简单的通信系统模型