通信网络精品图书

吉林大学本科“十二五”规划教材建设项目

信息与编码理论

• 杨晓萍　主编

中国工信出版集团

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

通信网络精品图书

吉林大学本科“十二五”规划教材建设项目

信息与编码理论

杨晓萍　主编

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书系统讲述了信息论及编码的基础理论和方法，主要包括离散信源及熵、离散信道及信道容量、离散信源编码与香农第一定理、离散信道与香农第二定理、连续信源与连续信道、率失真函数、香农第三定理等。采用较多的通信和信息系统相关的背景例题和图示阐述基本概念，注重编码理论、编码方法的实现过程的教学内容编写，给出重要算法的实现流程图，并附有编程算法的实现程序，便于读者对课程的理解和应用。

本书可作为理工科高等院校电子信息类、通信工程及相关专业的研究生、本科生的教材，也可供相关专业的科技人员学习参考。

图书在版编目（CIP）数据

信息与编码理论 / 杨晓萍主编. —北京：电子工业出版社，2016.5
（通信网络精品图书）
ISBN 978-7-121-28892-0

Ⅰ. ①信… Ⅱ. ①杨… Ⅲ. ①信息论－高等学校－教材②信源编码－通信理论－高等学校－教材 Ⅳ. ①TN911.2

中国版本图书馆 CIP 数据核字（2016）第 111638 号

策划编辑：宋 梅
责任编辑：刘真平
印 刷：三河市华成印务有限公司
装 订：三河市华成印务有限公司
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：787×1092 1/16 印张：14 字数：358.4 千字
版 次：2016 年 5 月第 1 版
印 次：2016 年 5 月第 1 次印刷
印 数：2 500 册 定价：43.00 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）88254888，88258888。

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。

本书咨询联系方式：mariams@phei.com.cn。

目　录

第1章 绪 论

信息对于我们来说并不陌生，生活中每天都会接触到大量的信息，各行各业都十分重视信息管理，可以说现在已经进入了信息时代。信息论是由通信技术、概率论、随机过程和数理统计等知识相结合产生的一门学科。它研究信息的基本理论，包括可能性和存在性等问题。信息论涉及的内容不局限于传统的通信范畴，进入了更广阔的信息科学领域。

1.1 信息的概念

信息论的创始人是克劳德·香农，被誉为“信息论之父”。香农于 1948 年发表的论文“A Mathematical Theory of Communication”，被人们认为是现代信息论研究的开端。这篇文章部分基于哈里·奈奎斯特和拉尔夫·哈特莱之前的研究成果，香农在这篇文章中创造性地采用概率论的方法来研究通信中的问题，提出了信息熵的概念。香农的这篇论文及其 1949 年发表的另一篇论文一起奠定了现代信息论的基础。

信息是指各个事物运动的状态及状态变化的方式。信息是对客观事物的反映，从本质上说信息是对社会、自然界的事物特征、现象、本质及规律的描述。人们从对周围世界观察得到的数据中获得信息，信息是看不到摸不着的抽象的意识或知识。在日常生活中，信息常常被认为是消息，信息与消息之间有着密切的关系，但是信息的含义更加深刻，它不能等同于消息。消息是包含信息的语言、文字、图像和声音等。在通信中，消息是指担负着传送信息任务的符号和符号序列。消息是具体的，它承载着信息，但它不是物理性的。以下面的情况为例说明信息和消息的区别。人们接到电话、听到广播、看了电视或者上网浏览了网页以后，就说得到了“信息”，其实这是不准确的。人们在收到消息后，如果消息使我们知道了很多以前不知道的内容，我们就收获了很多信息；但是如果消息的内容我们以前基本都知道，那么我们得到的信息就不多了，甚至是没有得到任何的信息。信息是认识主体接收到的、可以消除对事物认识不确定性的新内容和新知识，信息是可以度量的。

1928 年哈特莱首先研究了通信系统传输信息的能力，提出对数度量信息的概

念，即一个消息所包含的信息量用它的所有可能取值的对数来表示。香农受到了哈特莱工作的启发，他注意到消息的信息量不但和它的可能性数值有关，还和消息本身的不确定性有关。一个消息之所以含有信息，是因为它具有不确定性，一个没有不确定性的消息不能包含任何信息，通信的目的就是要尽量地消除这种不确定性。香农对于信息的定义为：**信息是对事物运动状态或存在方式的不确定性的描述**。

用数学语言来讲，不确定性即为随机性，具有不确定性的事件就是随机事件。概率论和随机过程作为研究随机事件的数学工具，可以用来测度不确定性的大小。信息量是信息论中度量信息多少的一个物理量，它从量上反映具有确定概率的事件发生时所传递的信息。信息的度量与它所代表事件的随机性或者说事件发生的概率有关，当事件发生的概率大，事先容易判断，有关此事件的消息发生的不确定程度小，则包含的信息量就小；反之，当事件发生的概率小，事先不容易发生，则发生后包含的信息量就大。例如天气预报，以长春五月份的天气为例，经常出现的是晴、多云、阴、晴间多云等天气，小雨不常出现，小雪出现的概率极小，大雪出现的可能性微乎其微。在看天气预报前，我们大体可以猜测出气象状况，出现晴、多云、阴、晴间多云等天气的概率较大，我们比较能确定这些天气的出现，所以预报出现晴或阴的时候，我们并不奇怪，和我们预期的情况是一致的，所消除的不确定性就小，获得的信息量也就小。当预报明天有小雪时，我们就会很意外，觉得反常，获得的信息量就很大。如果是预报大雪的话，所获得的信息量就会更大。

在信息论中，将消息用随机事件表示，发出这些消息的信源则用随机变量表示。我们将某个消息 x_i 出现的不确定性的大小定义为**自信息**，用这个消息出现概率的对数的负值来计算自信息量，表示为

$$I(x_i)=-\log p(x_i) \tag{1.1}$$

信息的基本概念强调的是事物状态的不确定性，任何已经确定的事物都不含有信息，信息的特征有以下几点：

① 接收者在收到信息之前，对它的内容是未知的，信息是新知识、新内容；

② 信息可以产生、消失，也可以被传输、储存和处理；

③ 信息是可以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；

④ 信息是可以度量的，并且信息量有多和少的差别。

信息在传输、处理及存储的过程中，难免受到噪声等无用信号的干扰，信息论就是为准确有效地将信息从传递的数据中提取出来提供依据和方法。信息论是建立在信息可以度量的基础上的，对如何有效、可靠地传递信息进行研究。它涉及信息特性、信息度量、信息传输速率、信道容量及干扰对信息传输影响等内容。这是狭义信息论，也称为香农信息论。广义信息论包含通信的全部统计问

题，除了狭义信息论之外，还包括信号设计、噪声理论、信号检测与估值等。本书所描述的信息论是狭义信息论。这种建立在概率模型上的信息概念，排除了生活中“信息”概念中所包含的主观性和主观意义，而是对消息统计特性的定量描述。根据香农的定义，任何一个消息对于任何一个接收者来说，所包含的信息量是一致的。但是，实际上信息有很强的主观性和实用性，对于不同的人同样的一个消息有不同的主观意义和价值，获得的信息量也是不同的。香农信息论的定义和度量是科学的，它能反映出信息的某些本质，但是它有缺陷和局限性，适用范围受到一定的限制。

1.2 信息论的研究对象、目的和内容

1.2.1 研究对象

信息论从诞生到今天，它的发展对人类社会的影响是广泛和深刻的。现在信息论研究的内容不单是通信，还包括与信息有关的自然和社会领域。香农信息论发展成为涉及范围极广的信息科学。信息论的研究对象是广义的通信系统，将各种通信系统模型中具有共同特点的部分抽取出来，概括为一个统一的模型，如图 1-1 所示。

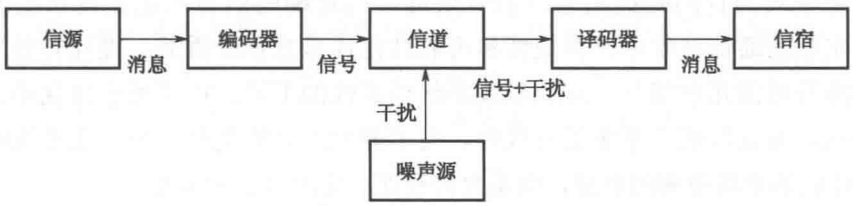

图 1-1 通信系统模型

这个通信系统模型不仅适用于电话、传真、电视、广播、遥感、雷达和导航等狭义的通信系统，还适用于其他的信息流通系统，如生物有机体的遗传系统、神经系统、视觉系统等，甚至是人类社会的管理系统。信息论的研究对象是这种统一的通信系统模型。信息以消息的形式在这个通信系统中传递，人们通过系统中消息的传输和处理来研究信息传输和处理的共同规律，其目的是提高通信的有效性和可靠性。图 1-1 所示的通信系统模型主要分成以下五个部分。

（1）信源

信源是产生消息和消息序列的源，它向通信系统提供消息。信源可以是人、生物、机器或其他事物，它是事物各种运动状态或存在状态的集合。例如，各种

天气状况是信源，通信网中向外发布消息的终端也是信源。信源本身是复杂的，在信息论中我们只对信源的输出进行研究。信源输出的是消息，消息是具体的，但它不是信息本身。消息携带着信息，消息是信息的表达者。

信源可能出现的状态，即信源输出的消息是随机的、不确定的，但又有一定的规律性，因而用随机变量或随机矢量等数学模型表示信源。信源的核心问题是它到底包含多少信息，怎么将信息定量地表示出来，即如何确定信息量。

（2）编码器

编码是把消息变换成信号的措施，而译码就是编码的反变换。编码器输出的是适合信道传输的信号。信号是消息的物理体现，为了在信道上传输消息，就必须将消息加载到具有某种特征的信号上去。信号携带着消息，它是消息的载荷者，是物理性的。

编码器可分为两种，信源编码器和信道编码器。**信源编码**是对信源输出的消息进行适当的变换和处理，目的是为了提高信息传输的效率。信源编码有两个作用，一是把信源发出的消息变换为由二进制或多进制码元组成的代码组，即基带信号；二是通过信源编码来压缩信源的冗余度。**信道编码**是为了提高信息传输的可靠性而有目的地对信源编码器输出的代码添加一些附加码元，使之具有检错、纠错的能力。通常，信道中的干扰会使通信质量下降，对于模拟信号，表现在接收信号信噪比的下降；对于数字信号，表现在误码率的增加。

信源编码的目的是提高系统的有效性，信道编码的目的是提高系统的可靠性。在实际的通信系统中，有效性和可靠性往往是互相矛盾的，提高有效性必须去掉信源符号的冗余部分，这会导致系统可靠性的下降；提高可靠性就必须增加监督码元，这就降低了系统的有效性。为了兼顾二者的关系，不一定要求绝对准确地在接收端重现原来的消息，而是允许存在一定的失真和误差。

（3）信道

信道是指通信系统把载荷消息的信号从甲地传输到乙地的媒介。信道是传递消息的通道，又是传送物理信号的设施。信道除了传播信号外，还可以存储信号。在狭义的通信系统中，实际信道有明线、电缆、波导、光纤、无线电波传播空间等，这些都是属于传输电磁波能量的信道。对于广义的通信系统来说，信道还可以是其他的传输媒介。信道问题主要是它能够传送多少信息，即信道容量的大小。

（4）噪声源

在信道中引入噪声和干扰，这是一种简化的表达方式，将系统其他部分产生

的干扰和噪声都等效地折合成信道干扰，看成是由一个噪声源所产生的，它作用于传输信号。这样，信道输出的是叠加了干扰的信号。**噪声源**是通信系统中各个干扰的集中反映，用来表示消息在信道中传输时遭受干扰的情况。干扰的性质或大小影响着通信系统的性能。由于干扰或噪声往往具有随机性，所以信道的特性也可以用概率空间来描述。而噪声源的统计特性又是划分信道的依据。

（5）译码器

译码就是把信道输出的编码信号（已叠加了干扰）进行反变换，变换成能够理解的消息。一般认为这种变换是可逆的。译码器也可以分为信源译码器和信道译码器。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；信道译码器具有检错、纠错功能，它可以将落在其检错、纠错范围内的误传码元检测出来，并加以纠正，以提高通信系统的可靠性。

（6）信宿

信宿是消息传送的对象，即接收消息的人或机器。根据实际情况，信宿接收的消息形式可以与信源发出的消息相同，也可以不同。当它们形式不同时，信宿所接收的消息是信源发出消息的一个映射。信宿要研究的是能够收到和提取多少信息量。

图 1-1 所示的通信系统模型只适合用于收发两端单向通信的情况，它只有一个信源和一个信宿，信息传输也是单向的。在实际的通信网络中，信源和信宿可能会有若干个，信息传输的方向也可以是双向的。要研究复杂的通信系统，需要对两端单向通信系统模型进行修正，把两端单向通信的信息理论发展为多用户通信信息理论。

1.2.2 研究目的

信息论的研究目的是在通信系统中找到信息传输过程的共同规律，来提高信息传输的可靠性、有效性、保密性和认证性，以达到信息传输系统的最优化。

1.2.3 研究内容

对于信息论研究的具体内容，以往的学者有着争议。目前，对于信息论研究的内容一般有三种理解，分别是：狭义信息论、一般信息论和广义信息论，其中狭义信息论和广义信息论我们已经在 1.1 节中进行了简单的描述。

（1）狭义信息论

狭义信息论又称为香农信息论，主要通过数学描述与定量分析，研究信息的测度、信道容量以及信源和信道编码理论等问题。通过编码和译码使接收和发射两端联合最优化，并以定理的形式证明极限的存在。狭义信息论的内容是信息论的基础理论。

（2）一般信息论

一般信息论又称为工程信息论，主要研究的问题也是信息传输和处理。除了狭义信息论的内容外，还包括噪声理论、信号滤波和预测、统计检测和估计、调制理论、信息处理和保密理论等内容。

（3）广义信息论

广义信息论又称为信息科学，它的研究内容不但包括狭义信息论和一般信息论的内容，而且还包括所有与信息有关的自然和社会科学领域，比如模式识别、机器翻译、遗传学、心理学、神经生理学等，甚至还包括社会学中有关信息的问题。它是新兴的信息科学理论。

本书讲述的信息论的基本内容是与通信学科密切相关的狭义信息论，也就是香农信息论，涉及信息论中的很多基本问题。

第 2 章　信息的测度

本章讨论信息测度的相关概念，主要包括一个事件发生时包含的自信息、事件集合所包含的信息熵、事件之间所能相互给出的互信息等，需要深刻理解相关的定义和计算方法。

2.1　自　信　息

在绪论中已经讲过，信源发出的消息（事件）具有不确定性，而事件发生的不确定性与事件发生的概率大小有关。概率越小，不确定性越大，事件发生后所含有的信息量就越大。小概率事件不确定性大，一旦出现必然使人感到意外，因此产生的信息量就大，特别是几乎不可能出现的事件一旦出现，必然产生极大的信息量；大概率事件因为是意料之中的事件，不确定性小，即使发生也没有多少信息量，特别是概率为 1 的确定事件发生以后，不会给人以任何信息量。因此，随机事件的自信息量 $I(x_i)$是该事件发生概率 $p(x_i)$的函数，并且 $I(x_i)$应该满足以下公理化条件：

（1）$I(x_i)$是 $p(x_i)$ 的严格递减函数。当 $p(x_1) < p(x_2)$ 时，$I(x_1) > I(x_2)$，概率越小，事件发生的不确定性越大，事件发生以后所包含的自信息量越大。

（2）极限情况下，当$p(x_i) = 0$ 时，$I(x_i) \to \infty$；当$p(x_i) = 1$ 时，$I(x_i) = 0$。

（3）两个相对独立的不同消息所提供的信息量应等于它们分别提供的信息量之和，即自信息量满足可加性。

根据上述条件可以从数学上证明事件的自信息量 $I(x_i)$与事件的发生概率 $p(x_i)$之间的函数关系满足对数形式。

定义 2.1　随机事件的自信息量定义为该事件发生概率的对数的负值。设事件 x_i的概率为 $p(x_i)$，则它的自信息量为

$$I(x_i) = -\log p(x_i) = \log \frac{1}{p(x_i)} \tag{2.1}$$

由式（2.1）绘出自信息量 $I(x_i)$与事件的发生概率 $p(x_i)$之间的函数关系，如图 2-1 所示，显然，自信息量的定义满足公理性条件，在定义域 [0, 1]内，自信息量是非负的。

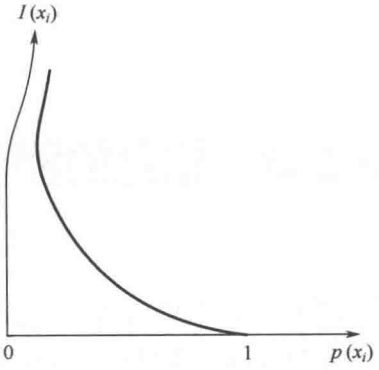

图 2-1　自信息量

$I(x_i)$代表两种含义：在事件 x_i 发生以前，代表事件 x_i 发生的不确定性的大小；在事件 x_i 发生以后，表示事件 x_i 所含有或所能提供的信息量。在无噪信道中，事件 x_i 发生以后，能正确无误地传输到收信者，所以 $I(x_i)$就等于收信者接收到 x_i 后所获得的信息量。这是因为消除了 $I(x_i)$大小的不确定性，才获得如此大小的信息量。

自信息量的单位与对数计算选用的底有关。当分别选用底为 2、e 和 10 时，自信息的单位分别为比特、奈特和哈特，详述如下：

（1）对数的底取 2 时，信息量的单位为比特（bit, binary unit）。当 $p(x_i) = 1/2$ 时，$I(x_i) = 1$ 比特，即概率等于 1/2 的事件具有 1 比特的自信息量。例如，一枚均匀硬币的任何一种抛掷结果均含有 1 比特的信息量。比特是信息论中最常用的信息量单位，为了书写简洁，当取对数的底为 2 时，底数 2 常省略不写。注意：计算机术语中 bit 是位的单位（bit, binary digit），与信息量单位含义不同。

（2）对数的底取自然对数（以 e 为底）时，自信息量的单位为奈特（nat, natural unit）。理论推导中或用于连续信源时用以 e 为底的对数比较方便。

$$1 \text{ 奈特} = \log_2 \mathrm{e} \text{ 比特} \approx 1.443 \text{ 比特}$$

（3）工程上取以 10 为底的对数比较方便，自信息量的单位为哈特（hart），用来纪念哈特莱（Hartley）首先提出用对数来度量信息的贡献。

$$1 \text{ 哈特} = \log_2 10 \text{ 比特} \approx 3.322 \text{ 比特}$$

（4）如果取以 r 为底的对数（$r > 1$），则 $I(x_i) = -\log_r p(x_i)$（r 进制单位）。

$$1\ r \text{ 进制单位} = \log_2 r \text{ 比特}$$

例 2.1　英文字母中“e”的出现概率为 0.105，“a”的出现概率为 0.064，“c”的出现概率为 0.022。求：（1）分别计算它们的自信息量；（2）假定前后字母出现是相互独立的，计算“ac”的自信息量。

【解】（1）“e”的自信息量　　$I(\mathrm{e}) = -\log 0.105 = 3.252$ 比特

“a”的自信息量　　$I(\mathrm{a}) = -\log 0.064 = 3.966$ 比特

“c”的自信息量　　$I(\mathrm{c}) = -\log 0.022 = 5.506$ 比特

（2）由于前后字母出现是相互独立的，“ac”出现的概率为 0.064×0.022，所以

“ac”的自信息量　　$I(\mathrm{ac}) = -\log(0.064\times 0.022) = I(\mathrm{a}) + I(\mathrm{c}) = 9.472$ 比特

由上面的计算可知，出现概率高的字符携带的自信息较小，两个相互独立事件的自信息量满足可加性，也就是由两个相对独立的事件的积事件所提供的信息量等于它们分别提供的信息量之和。

例 2.2　对于 2^n 进制的数字序列，假设每一个符号的出现完全随机且概率相等，求任一符号的自信息量。

【解】设 2^n 进制数字序列任一码元 x_i 的出现概率为 $p(x_i)$，则有

$$p(x_i) = \frac{1}{2^n}$$

$$I(x_i) = \log\frac{1}{p(x_i)} = \log 2^n = n \text{比特}$$

显然，事件的自信息量只与其概率有关，而与它的取值无关。

2.2　平均自信息

2.2.1　平均自信息的概念

自信息量是信源发出某一具体消息所含有的信息量，发出的消息不同它的自信息量就不同，所以自信息量本身为随机变量，不能用来表征整个信源的不确定度。我们**用平均自信息量来表征整个信源的不确定度**。平均自信息量又称为**信息熵、信源熵**，简称**熵**。

因为信源具有不确定性，所以把信源用随机变量来表示，用随机变量的概率分布来描述信源的不确定性。通常把一个随机变量的所有可能的取值和这些取值对应的概率 $[X, P(X)]$ 称为**信源的概率空间**。

假设随机变量 X 有 q 个可能的取值 x_i，$i = 1, 2, \cdots, q$，各种取值出现的概率为 $p(x_i)$，则信源的概率空间表示为

$$\begin{bmatrix} X \\ P(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_i & \cdots & x_q \\ p(x_1) & p(x_2) & \cdots & p(x_i) & \cdots & p(x_q) \end{bmatrix} \qquad \sum_{i=1}^{q} p(x_i) = 1 \quad (2.2)$$

式（2.2）中，$p(x_i)$ 满足概率空间的基本特性，$0 \leqslant p(x_i) \leqslant 1$，即具有非负性和完备性。

定义 2.2　随机变量 X 的每一个可能取值的自信息 $I(x_i)$ 的统计平均值定义为随机变量 X 的平均自信息量，也就是**熵**，表示为