



普通高等教育“十一五”国家级规划教材

21世纪

高等院校电子信息类规划教材

Xinxilun
Jichu

信息论 基础

◎ 田宝玉 杨洁 贺志强 王晓湘 编著

◎ 吴伟陵 审

 人民邮电出版社
POSTS & TELECOM PRESS

普通高等教育“十一五”国家级规划教材

21 世纪高等院校电子信息类规划教材

信息论基础

田宝玉 杨 洁 贺志强 王晓湘 编著
吴伟陵 审

人民邮电出版社
北京

图书在版编目 (CIP) 数据

信息论基础/田宝玉等编著. —北京: 人民邮电出版社,
2008.8
21 世纪高等院校电子信息类规划教材,
ISBN 978-7-115-17790-2

I. 信… II. 田… III. 信息论—高等学校—教材
IV. G201

中国版本图书馆 CIP 数据核字 (2008) 第 032205 号

内 容 提 要

本书是作者多年教学和科研实践的积累, 是在吸收国内外优秀教材优点的基础上, 进一步优化整合教学内容, 并进行改进和补充而成的。全书共分为 12 章, 内容包括: 信息的基本概念、香农信息论研究的内容与进展, 离散信息的度量, 离散信源, 连续信息与连续信源, 无失真信源编码, 离散信道及其容量, 有噪信道编码, 波形信道, 信息率失真函数, 有约束信道及其编码, 网络信息论初步, 信息理论与方法与应用等。

本书强调定理中物理概念和结论的理解与掌握, 简化烦琐的数学推导, 注重使用明确、直观的物理概念, 增加实例, 力求让讲述的内容更适合工科专业学生的学习。本书有*号的部分可以作为选学内容, 有利于教师根据教学计划进行调整。每章的思考题和习题可以作为课后的练习。

本书主要用做高等院校信息与通信及相关专业的教材, 还可作为相关专业研究生和工程技术人员的参考书。

普通高等教育“十一五”国家级规划教材

21 世纪高等院校电子信息类规划教材

信息论基础

-
- ◆ 编 著 田宝玉 杨 洁 贺志强 王晓湘
审 吴伟陵
责任编辑 张孟伟
 - ◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街 14 号
邮编 100061 电子函件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
北京通州大中印刷厂印刷
 - ◆ 开本: 787×1092 1/16
印张: 17.75
字数: 432 千字 2008 年 8 月第 1 版
印数: 1—3 000 册 2008 年 8 月北京第 1 次印刷

ISBN 978-7-115-17790-2/TN

定价: 29.80 元

读者服务热线: (010) 67170985 印装质量热线: (010) 67129223

反盗版热线: (010) 67171154

前 言

信息论基础即香农 (C. E. Shannon) 信息论, 是用概率论与随机过程的方法研究通信系统传输有效性和可靠性极限性能的理论, 是现代通信与信息处理技术的理论基础, 也是通信与电子信息类专业的重要基础课程。“信息论基础”课程以信息熵为基本概念, 以香农 3 个编码定理 (无失真信源编码定理、有噪信道编码定理和限失真信源编码定理) 为核心内容, 研究通信系统中信源信息的度量、信源的压缩, 以及信息通过信道有效和可靠传输等问题。

学习本课程的主要目的是, 理解香农信息论的基本原理, 掌握信息与熵的基本运算方法, 培养利用信息论的基本原理分析和解决实际问题的能力, 为今后进行更深入的研究奠定良好的理论基础。

从 20 世纪 60 年代开始, 国内出版了很多关于香农信息论的教材与专著, 推动了信息论的理论研究与基本知识的普及。本书是作者多年教学和科研实践的积累, 是在吸收国内外优秀教材优点的基础上, 进一步优化整合教学内容, 并进行改进和补充而成的。

本书分为 12 章, 内容包括: 第 1 章介绍信息的基本概念、通信系统模型、香农信息论研究的内容与进展; 第 2 章离散信息的度量, 主要包括信息熵和互信息的概念与性质, 这是学习本课程的基础知识; 第 3 章离散信源, 主要包括离散无记忆信源和马尔可夫信源熵的计算; 第 4 章连续信息与连续信源, 介绍连续信息的度量、连续信源的熵与互信息的性质, 主要包括离散时间高斯信源的熵以及最大熵定理; 第 5 章无失真信源编码, 主要包括无失真信源编码定理, 异前置码的性质以及 Huffman 编码方法; 第 6 章离散信道及其容量, 主要包括某些重要的离散信道的性质及其信道容量的计算; 第 7 章有噪信道编码, 主要包括常用译码准则、有噪信道编码定理以及简单的纠错编码技术; 第 8 章波形信道, 主要包括离散时间连续信道容量和 AWGN 信道容量; 第 9 章信息率失真函数, 主要包括离散与连续信源 $R(D)$ 函数的计算, 重点是高斯信源的 $R(D)$ 函数; 第 10 章有约束信道及其编码, 主要包括有约束信道容量、有约束信道编码定理和常用的有约束信道编码实例; 第 11 章网络信息论初步, 主要包括多接入信道、广播信道和相关信源编码; 第 12 章信息理论方法及其应用, 主要包括信源熵估计、最大熵原理与最小交叉熵原理与应用等。

本书遵循与时俱进的原则。在内容的选择上既考虑基础性又考虑先进性和时代性。在保证基础理论讲授的前提下, 加入信息领域的最新研究成果作为补充。本书的另一个特点是注重理论与技术的应用, 例如第 12 章所介绍的信息理论方法与技术已经应用到很多领域, 并获得了成功。

虽然“信息论基础”课程是一门理论性较强的课程, 通常采用“以理论教学为主、以实验教学为辅”的教学方式, 但是在理论教学中还要结合工科学生的特点。本书强调定理中物理概念和结论的理解与掌握, 简化而不陷入烦琐的数学推导, 注重使用明确或直观的物理概念, 增加实例, 力求让讲述的内容更适合工科专业学生的学习。

本书将内容分成层次, 注 * 号的可以作为选学内容, 这不但有利于教师根据教学计划进行调整, 也便于学生自学。本书每章均附有思考题和习题, 思考题侧重于基本概念的理解,

而习题侧重于基本运算和解决实际问题能力的训练。为便于读者对某些问题进行深入研究，书的末尾列出了主要的参考文献。

学习本课程需要具备较扎实的概率论与随机过程的知识，如果有通信原理的基础则更好一些，但本课程可以与“通信原理”课程同步学习。本书主要用作高等院校信息与通信及相关专业的教材，还可作为相关专业研究生和工程技术人员的参考书。

本书由田宝玉担任主编，负责全书内容编写协调并撰写第1、7、10、12章；杨洁撰写第2、5、6章，贺志强撰写第8、9章，王晓湘撰写第3、4、11章。吴伟陵教授审阅了全书，并提出了很多宝贵意见。

在本书的写作过程中，作者得到人民邮电出版社出版基金的资助，同时也得到北京邮电大学张英海副校长和北京邮电大学信息工程学院郭军院长、张琳副院长的关心与支持，得到北京邮电大学信息工程学院信息理论与技术教研中心张惠民教授和林家儒教授以及广大教师的支持与帮助；并始终得到了我国信息论奠基人之一——北京邮电大学周炯槃院士，信息论著名专家吴伟陵教授，智能信息著名专家钟义信教授，信号与信息处理著名专家南京邮电大学郑宝玉教授的指导；很多教师和同学也为本书的完成做出了贡献，作者在此一并表示感谢。

因编写水平所限，错误与疏漏之处在所难免，敬请读者批评指正。

编 者

2008年2月于北京邮电大学

目 录

第1章 绪论	1	3.1.3 离散有记忆信源的数学模型	39
1.1 信息的基本概念	1	3.2 离散无记忆信源的熵	39
1.1.1 信息论的产生	1	3.2.1 单符号离散无记忆信源的熵	39
1.1.2 信息的基本概念	2	3.2.2 离散无记忆信源 N 次扩展源的熵	40
1.2 香农信息论研究的内容	3	3.3 离散平稳信源的熵	40
1.2.1 通信系统模型	4	3.3.1 离散平稳信源	40
1.2.2 香农信息论的主要内容	6	3.3.2 离散平稳有记忆信源的熵	41
1.3 香农信息论研究的进展与应用	8	3.4 有限状态马尔可夫链	42
1.3.1 香农信息论创立的背景	8	3.4.1 马氏链基本概念	43
1.3.2 香农的主要贡献	9	3.4.2 齐次马氏链	43
1.3.3 香农信息论的研究进展	9	3.4.3 马氏链状态分类	46
1.3.4 香农信息论的应用	12	3.4.4 马氏链的平稳分布	47
思考题	12	3.5 马尔可夫信源	48
第2章 离散信息的度量	14	3.5.1 马氏源的基本概念	48
2.1 自信息和互信息	14	3.5.2 马氏源的产生模型	50
2.1.1 自信息	14	3.5.3 马氏链 N 次扩展源的熵的计算	51
2.1.2 互信息	17	3.5.4 马氏源符号熵的计算	53
2.2 信息熵	18	3.6 信源的相关性与剩余度	55
2.2.1 信息熵的定义与计算	18	3.6.1 信源的相关性	55
2.2.2 条件熵与联合熵	21	3.6.2 信源剩余度(冗余度)	55
2.2.3 熵的基本性质	22	3.6.3 自然语言的相关性和剩余度	56
2.3 平均互信息	27	本章小结	59
2.3.1 平均互信息的定义	27	思考题	59
2.3.2 平均互信息的性质	28	习题	60
2.3.3 平均条件互信息	30	第4章 连续信息与连续信源	64
本章小结	33	4.1 连续随机变量集合的熵	64
思考题	34	4.1.1 连续随机变量的离散化	65
习题	34	4.1.2 连续随机变量集的熵	65
第3章 离散信源	37	4.1.3 连续随机变量集的条件熵	65
3.1 离散信源分类与数学模型	37	4.1.4 连续随机变量集的联合熵	66
3.1.1 离散信源分类	37	4.1.5 连续随机变量集差熵的性质	66
3.1.2 离散无记忆信源的数学模型	38	4.1.6 连续随机变量集的信息散度	68
		4.2 离散时间高斯信源的熵	69

4.2.1 一维高斯随机变量集的熵	69	5.5.3 L-Z 编码	101
4.2.2 多维独立高斯随机变量集的熵	69	本章小结	101
4.2.3 多维相关高斯随机变量集的熵	69	思考题	102
4.3 连续最大熵定理	70	习题	102
4.3.1 限峰值最大熵定理	71	第 6 章 离散信道及其容量	105
4.3.2 限功率最大熵定理	71	6.1 概述	105
4.3.3 熵功率和剩余度	72	6.1.1 信道的分类	105
4.4 连续随机变量集的平均互信息	72	6.1.2 离散信道的数学模型	106
4.4.1 连续随机变量集的平均互信息	72	6.1.3 信道容量的定义	109
4.4.2 连续随机变量集平均互信息的性质	73	6.2 单符号离散信道及其容量	109
4.5 离散集与连续集之间的互信息	75	6.2.1 离散无噪信道的容量	109
4.5.1 离散事件与连续事件之间的互信息	76	6.2.2 离散对称信道的容量	110
4.5.2 离散集合与连续集合的平均互信息	76	6.2.3 一般离散信道的容量	112
本章小结	77	6.3 级联信道及其容量	116
思考题	77	6.4 多维矢量信道及其容量	118
习题	77	6.4.1 多维矢量信道输入与输出的性质	118
第 5 章 无失真信源编码	80	6.4.2 离散无记忆扩展信道及其容量	120
5.1 概述	80	6.4.3 并联信道及其容量	122
5.1.1 信源编码器	80	6.4.4 和信道及其容量	122
5.1.2 信源编码的分类	81	6.5 信道容量的迭代计算	123
5.1.3 分组码	82	本章小结	125
5.2 定长码	83	思考题	126
5.2.1 无失真编码条件	83	习题	126
5.2.2 信源序列分组定理	84	第 7 章 有噪信道编码	129
5.2.3 定长码信源编码定理	86	7.1 概述	129
5.3 变长码	88	7.1.1 信道编码的基本概念	129
5.3.1 异前置码的性质	88	7.1.2 判决与译码规则	130
5.3.2 变长码信源编码定理	90	7.1.3 译码错误概率	131
5.4 哈夫曼编码	93	7.2 最佳判决与译码准则	132
5.4.1 二元哈夫曼编码	93	7.2.1 最大后验概率准则	132
5.4.2 多元哈夫曼编码	96	7.2.2 最大似然准则	133
5.4.3 马氏源的编码	97	7.3 信道编码与最佳译码	134
*5.5 几种实用的编码方法	99	7.3.1 线性分组码	134
5.5.1 算术编码	99	7.3.2 序列最大似然译码	135
5.5.2 游程编码	100	7.3.3 几种简单的分组码	136

7.4 费诺(Fano)不等式	137	8.4.1 有色高斯噪声信道容量	173
7.4.1 信道疑义度	137	8.4.2 AWGN 信道容量的进一步讨论	175
7.4.2 费诺(Fano)不等式	138	* 8.5 数字调制系统的信道容量	176
7.5 有噪信道编码定理	139	本章小结	179
7.5.1 联合典型序列	140	思考题	180
7.5.2 有噪信道编码定理	141	习题	180
7.5.3 无失真信源信道编码定理	143	第9章 信息率失真函数	183
7.6 纠错编码技术简介	144	9.1 概述	183
7.6.1 线性分组码的编译码	144	9.1.1 系统模型	184
7.6.2 几种重要的分组码	148	9.1.2 失真测度	184
7.6.3 卷积码简介	149	9.2 离散信源信息率失真函数	185
* 7.7 信道编码性能界限	150	9.2.1 信息率失真函数	185
7.7.1 汉明球包界	150	9.2.2 $R(D)$ 函数的性质	185
7.7.2 Varsharmov-Gilbert 界	151	9.3 限失真信源编码定理	188
7.7.3 Plotkin 界	152	9.3.1 码率的压缩	188
本章小结	153	9.3.2 限失真信源编码定理	189
思考题	153	9.3.3 限失真信源信道编码定理	190
习题	154	9.4 离散信源信息率失真函数的 计算	190
第8章 波形信道	159	9.4.1 $R(D)$ 参量表示法求解	191
8.1 离散时间连续信道	159	9.4.2 $R(D)$ 求解过程归纳	192
8.1.1 时间离散连续信道模型	159	9.4.3 参量 s 的意义	193
8.1.2 平稳无记忆连续信道	160	9.5 连续信源信息率失真函数	195
8.1.3 多维矢量连续信道的性质	160	9.5.1 信息率失真函数与性质	195
8.1.4 离散时间连续信道的容量	160	9.5.2 $R(D)$ 函数的计算	195
8.2 加性噪声信道与容量	161	9.5.3 差值失真测度	196
8.2.1 加性噪声信道的容量	161	9.6 高斯信源的 $R(D)$ 函数	197
8.2.2 加性高斯噪声信道的容量	162	9.6.1 离散时间无记忆高斯信源	197
8.2.3 一般加性噪声信道容量界	163	9.6.2 独立并联高斯信源	199
8.2.4 并联加性高斯噪声信道的 容量	164	9.7 一般连续信源 $R(D)$ 函数	201
8.3 AWGN 信道的容量	166	* 9.8 有损数据压缩技术简介	201
8.3.1 加性高斯噪声波形信道	166	9.8.1 量化	202
8.3.2 波形信道的互信息与容量	167	9.8.2 预测编码	202
8.3.3 AWGN 信道的容量	168	9.8.3 子带编码	203
8.3.4 高斯噪声信道编码定理	170	9.8.4 变换编码	203
8.3.5 功率利用率和频谱利用率的 关系	171	本章小结	204
8.4 有色高斯噪声信道	173	思考题	205
		习题	205

第 10 章 有约束信道及其编码	208	11.2.3 多址接入信道的容量	238
10.1 标号图的性质	208	11.3 广播信道	238
10.1.1 标号图的基本概念	208	11.3.1 退化广播信道	239
10.1.2 标号图的变换	210	11.3.2 退化广播信道的容量区域	240
10.2 有约束信道容量	211	11.4 相关信源编码	242
10.2.1 有约束信道容量的定义	211	11.4.1 典型的相关信源编码模型	243
10.2.2 等时长符号有约束信道的 容量	212	11.4.2 Slepian-Wolf 相关信源编码 定理	244
10.2.3 不等时长符号无约束信道的 容量	213	本章小结	247
10.2.4 不等时长符号有约束信道的 容量	214	思考题	248
10.3 有约束序列的性质	215	习题	249
10.3.1 信道对传输序列的约束	215	* 第 12 章 信息理论方法及其应用	250
10.3.2 游程长度受限序列(RLL)	215	12.1 信源熵的估计	250
10.3.3 部分响应最大似然(PRML) 序列	217	12.1.1 离散信源序列熵的估计	251
10.3.4 直流平衡序列	218	12.1.2 连续信源熵的估计	254
10.3.5 其他频域受限序列	220	12.2 最大熵原理	255
10.4 有约束信道编码定理	220	12.2.1 最大熵原理的描述	255
10.4.1 编码器的描述	220	12.2.2 熵集中定理	258
10.4.2 有约束信道编码定理	221	12.2.3 几种重要的最大熵分布	259
10.4.3 有限状态编码定理	221	12.3 最小交叉熵原理	261
10.4.4 编码器性能指标	222	12.3.1 最小交叉熵原理	261
* 10.5 有约束序列编码与应用	222	12.3.2 交叉熵的性质	263
10.5.1 块编码器	222	12.3.3 最小交叉熵推断的性质	264
10.5.2 实用直流平衡序列	223	12.3.4 交叉熵法	265
10.5.3 常用有约束序列编码及应用	225	12.4 信息理论方法的应用	265
本章小结	227	12.4.1 DNA 序列的熵估计和压缩	265
思考题	227	12.4.2 最大熵谱估计和最小交叉熵谱 估计	267
习题	227	12.4.3 最大熵建模及其在自然语言处理中的 应用	269
第 11 章 网络信息论初步	230	12.4.4 最大熵原理在经济学中的 应用	271
11.1 概述	230	12.4.5 信息理论方法应用展望	273
11.2 多址接入信道	231	本章小结	273
11.2.1 二址接入信道的容量	232	思考题	274
11.2.2 不同多址方式下的接入信道容量 分析	236	习题	274
		参考文献	276

第 1 章 绪 论

信息论即香农 (Claude Edwood Shannon, 1916—2001) 信息论, 也称经典信息论, 是研究通信系统极限性能的理论。从信息论产生到现在几十年来, 随着人们对信息的认识不断深化, 对信息论的研究日益广泛和深入, 信息论的基本思想和方法已经渗透到许多学科。在人类社会已经进入信息时代的今天, 信息理论在自然科学和社会科学研究领域还会发挥更大的作用。

本章简要地介绍了香农信息论的概况, 内容安排如下: 首先介绍信息的基本概念, 说明香农信息属于语法信息中的概率信息; 然后以通信系统模型为基础, 简单介绍香农信息论所研究的内容; 最后介绍香农信息论产生的背景、主要的研究进展及其应用。

1.1 信息的基本概念

1.1.1 信息论的产生

我们知道, 组成客观世界三大基本要素是: 物质、能量和信息。人类社会从农业时代经过工业时代发展到信息时代, 特别是在今天的信息时代, 社会的发展都离不开物质 (材料)、能量 (能源) 和信息资源。美国学者欧廷格说: “没有物质什么都不存在, 没有能量什么都不发生, 没有信息什么都没意义。” (“Without materials nothing exists. Without energy nothing happens. Without information nothing makes sense.”) 可见, 关于信息的课程本应该像物理、化学、生物等课程一样, 是基础课。但是, 由于信息的抽象性以及当前人们对信息的认识并不完全清楚, 所以在当前只能是专业基础课。

人们普遍认为, 1948 年美国工程师和数学家香农发表的《通信的数学理论》 (A Mathematical Theory of Communication, BSTJ, 1948) 这篇里程碑性的文章标志着信息论的产生, 而香农本人也成为信息论的奠基人。

香农指出, 通信的基本问题是在一点精确地或近似地恢复另一点所选择的消息。人们从这个基本问题出发, 对通信系统制定了三项性能指标: 传输的有效性、传输的可靠性、传输的安全性。

有效性是指: ①对于离散信源, 信源符号平均代码长度应尽量短; ②信息传输应尽可能快, 即高的传信率 (单位时间传送信息的速率), 实际上是有效利用时间资源; ③信息传送应该有高的频谱利用率, 实际上是有效利用频率资源。

可靠性是指：传输差错要尽量少，对于数字传输就是要求低的误码率。

安全性是指：传输的信息不能泄露给未授权人。

三项性能指标所对应的三项基本技术是：数据压缩、数据纠错和数据加密。

香农信息论解决了前两项技术的理论问题：提高有效性可通过信源编码（即信源压缩编码）来实现，并给出了压缩编码码率的极限；提高可靠性可通过信道编码来实现，并给出实现可靠传输的信息传输速率的极限。可以说，香农揭示了数据压缩和传输的基本定律。

实际上，传输安全性的理论问题也是香农首先解决的。不过关于传输安全性的问题往往被认为属于信息安全或密码学领域。早在“二战”期间，香农就对密码学感兴趣。他认识到，密码学中的基本问题与通信中的问题密切相关。1945年，他写了《密码学的数学理论》(A Mathematical Theory of Cryptography)，1949年改名为《保密系统的通信理论》(Communication Theory of Secrecy System)公开发表。这篇文章建立了保密系统的数学理论，对密码学产生了很大的影响。人们认为，是香农的工作才把密码学从技术变成科学。

因此，我们说，香农建立了通信中的三项基本技术的理论基础，信息论是前两项技术的理论基础。

1.1.2 信息的基本概念

信息论的产生引起了专家学者对信息研究的兴趣，他们从不同的角度和侧面研究和定义信息。据说到目前为止已有上百种信息的定义或说法。例如，“信息是事物之间的差异”，“信息是物质与能量在时间与空间分布的不均匀性”，“信息是收信者事先不知道的东西”等。

正因为信息的定义种类繁多，所以当前还没有一个公认的关于信息的定义，但这并不影响我们对信息的基本特征的认识。信息有许多与物质、能量相同的特征，例如信息可以产生、消失、携带、处理和量度；信息也有与物质、能量不同的特征，例如，信息可以共享，可以无限制地复制等。

实际上，信息可以划分为两个大的层次：本体论层次和认识论层次。从本体论层次上看，信息是客观的，即它是独立于人或其他有感知的事物而存在的，这就是说，在人类出现以前信息就存在了。从认识论层次上看，信息是通过认识主体的感受而体现出来的。现在我们所说的信息实际上是指认识论层次的信息。

1949年，Weaver在《通信的数学》中解释香农的工作时，把通信问题分成以下3个层次：

第1层：通信符号如何精确传输？（技术问题）

第2层：传输的符号如何精确携带所需要的含义？（语义问题）

第3层：所接收的含义如何以所需要的方式有效地影响行为？（效用问题）

Weaver认为香农的工作属于第一层，但他又指出香农的工作是交叉的，对第二、三层也有意义。

当前一种比较普遍的描述信息的说法是：信息是认识主体（人、生物、机器）所感受的和所表达的事物运动的状态和运动状态变化的方式。以这种定义为基础，可以把信息分成3个基本层次，即语法（Syntactic）信息、语义（Semantic）信息和语用（Pragmatic）信息，分别反映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。

可以看到，现在这种比较普遍认同的对信息的描述与Weaver对信息的解释基本一致。

语法信息是事物运动的状态和变化方式的外在形式，不涉及状态的含义和效用。像语言学领域的“词与词的结合方式”，不考虑词的含义与效用，在语言学中称为语法学。语法信息还可细分为概率信息、偶发信息、确定信息、模糊信息等。

语义信息是事物运动的状态和变化方式的含义。在语言学里，研究“词与词结合方式的含义”的学科称为语义学。

语用信息是事物运动状态及其状态改变方式的效用。

下面举例说明信息3个层次的含义。有一个情报部门，其主要任务是对经济情报进行收集、整理与分析以提供给决策机构。该部门设3个组：信息收集组、信息处理组和信息分析组。信息收集组的任务是将收集到的资料按中文、英文或其他文字、明文、密文进行分类，不管这些资料的含义如何都交到信息处理组。信息处理组根据资料的性质进行翻译或破译得到这些资料的含义，然后交到信息分析组。信息分析组从这些资料中挑选出有价值的情报提交给决策机构。可见，信息收集组是根据所得到的消息提取出语法信息，信息处理组是根据所得到的语法信息提取出语义信息，而信息分析组是根据所得到的语义信息提取出语用信息。

可以看到，研究语义信息要以语法信息为基础，研究语用信息要以语义信息和语法信息为基础。三者之间，语法信息是最简单、最基本的层次，语用信息则是最复杂、最实用的层次。

让我们再完整地引用香农在1948年的经典论断：“通信的基本问题是在一点精确地或近似地恢复另一点所选择的消息。通常，这些消息是有含义的，即它对于某系统指的是某些实在的或抽象的实体。这些通信的语义方面与通信问题无关，而重要的方面是实际消息是从一个可能消息集合中选择出的一条消息。”

可见，香农在研究信息理论时，排除了语义信息与语用信息的因素，先从语法信息入手，解决当时最重要的通信工程一类的信息传递问题。同时他还把信源视为具有输出的随机过程，所研究的事物运动状态和变化方式的外在形式遵循某种概率分布。显然香农信息论或经典信息论所研究的信息是语法信息中的概率信息。

不过，随着信息论研究的深入，香农信息论的方法已经渗透到语义信息领域，例如最大熵建模方法用于机器翻译等自然语言处理问题。

有人还提出通信信息（应属于语法信息）也有3个层次，即信号、消息与信息。其中信号为最低的层次，信息是最高层次。消息是信息的携带者，信息包含于消息之中，信号是消息的载体，消息是信号的具体内容。

信息各层次之间的关系如图1.1所示。

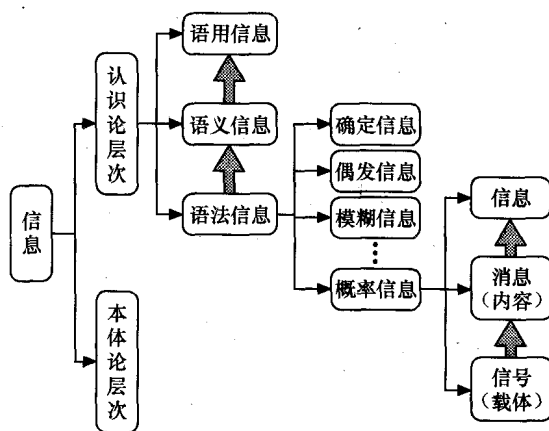


图 1.1 信息各层次之间的关系

1.2 香农信息论研究的内容

本节在介绍通信系统模型中各主要模块功能的基础上，简单讲述香农信息论研究的主要问题。

1.2.1 通信系统模型

如前面所述, 香农创立信息论是从研究通信系统开始的, 并首先建立了通信系统模型。由于技术发展水平的限制, 当时的通信基本限制在点对点的通信, 所以这种通信系统模型是指“从一个地方向另一个地方传送信息的系统”, 例如, 电话、电报、电视、无线通信、光通信等。而存储系统在某种意义上也可视为现在向将来发送信息的通信系统, 例如, 磁盘或光盘驱动器、磁带记录器、视频播放器等。可以说, 一般的通信系统是从空间的一点到另一点传送信息的, 而存储系统是从时间的一点到另一点传送信息的。

随着通信与信息网(其中包括电信网、互联网、移动通信网、广播电视网、光通信网等)的飞速发展, 需要将传统的通信系统模型进行扩展以适应新的研究需要。实际上, 多个点对点的通信系统通过一个公用信道, 就构成了多点对多点的通信系统模型。所以关于传统的点对点通信模型的知识是最基本的。

一般的通信系统模型框图如图 1.2 所示。下面对模型的主要组成部分进行简单描述。

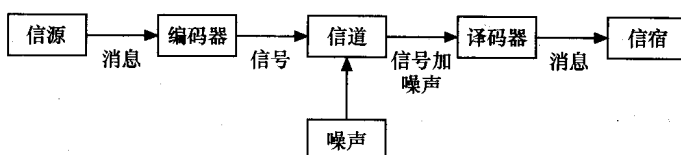


图 1.2 通信系统模型

1. 信源 (information source)

信源是信息的来源, 其功能是直接产生可能包含信息的信息。信源按输出符号的取值, 分为离散和连续信源两大类。在离散时间发出取值离散符号的信源为离散信源, 例如, 字符序列, 包括文件、信件、书报、杂志、电报、电传等都是离散信源。而连续信源又分为两种, 一种是在离散时间发出取值连续符号的信源, 称为离散时间连续信源; 另一种是输出为连续时间波形(连续时间, 符号取值连续)的信源, 称为波形信源或模拟信源。无线广播信号、电视信号、话音、图像信号以及多媒体信号等都是模拟信源, 而模拟信源在时间域、频率域的抽样或通过其他变换方式得到的等价的离散时间序列都是离散时间连续信源。

离散信源和离散时间连续信源也有共性, 就是它们的输出都是序列, 只不过是其符号的取值范围不同, 前者取自可数符号集, 而后者取自实数集。

信源按输出符号之间的依赖关系分类, 可分为无记忆和有记忆信源。如果信源输出符号的概率与以前输出的符号无关, 就称为无记忆信源; 否则就称为有记忆信源。离散信源和离散时间连续信源可以是无记忆的, 也可以是有记忆的, 而模拟信源大多是有记忆的。

信源研究的核心问题是信源消息所包含信息的量度。

2. 编码器 (encoder)

编码器的功能是将消息变成适合于信道传输的信号。在通信系统中称做发信机, 而在存储系统中称做记录器或写入器。编码器包括: 信源编码器 (source encoder)、信道编码器 (channel encoder)、调制器 (modulator), 如图 1.3 所示。应该指出, 在模拟通信系统中的编码器仅包含调制器。编码器中主要部分的功能介绍如下。

信源编码器的功能是将信源消息变成符号, 目的是提高传输有效性, 也就是压缩每个信

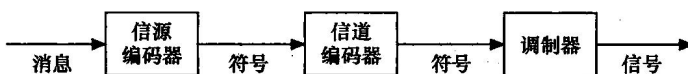


图 1.3 编码器的组成

源符号传输所需代码（通常为二进制代码）的数目（对二进制代码称比特数）。例如，一个信源含 4 个符号 {a, b, c, d}，概率分别为 $1/2$, $1/4$, $1/8$, $1/8$ 。如果不采用信源编码，每个信源符号至少需要用两个二进制代码传输。如果采用信源编码，分别将 a, b, c, d 编码成为：0, 10, 110, 111，那么平均每个信源符号只需 1.75 个二进制代码传输。可见，采用合适的信源编码确实能通过压缩码率提高传输有效性，信源编码也称为信源压缩编码。

信道编码器给信源编码符号增加冗余符号，目的是提高传输可靠性。信源编码输出直接传送，不能保证传输可靠性。利用信道编码对信源编码器的输出符号增加一些冗余符号，并让这些符号满足一定的数学规律，就可使传输具有纠错或检错能力。因为出现传输错误就会破坏这种数学规律，在接收端就会发现错误。例如，最简单的奇偶纠错，将信源编码输出的每个码组的尾端补一个 1 或 0，使得整个码组“1”的个数为奇或偶（或模二加为 1 或 0）。当传输发生奇数差错，打乱了“1”数目的奇偶性，就可以检测出错误。这是最简单的检错方式，而实际信道编码技术要复杂得多。

图 1.4 说明增加冗余符号可以提高传输可靠性。其中，图 (a)：4 个消息用 4 个二维矢量传送，没有冗余符号。如果出现任何差错都会使传送的码字变成另一个码字，所以无检错能力；图 (b)：在图 (a) 的基础上每个码字增加一个校验符号，构成奇校验，4 个消息用 4 个三维矢量传送。如果出现任何奇数差错都会使传送的码字变成不是码字的三维矢量，这样就能检测出错误，但不能纠正错误；图 (c)：用两个汉明距离为 3 的三维矢量传送两个消息。如果出现一个错误，可以根据接收矢量和码字汉明距离的大小判定是哪个消息被传输，可以纠正 1 个错误。

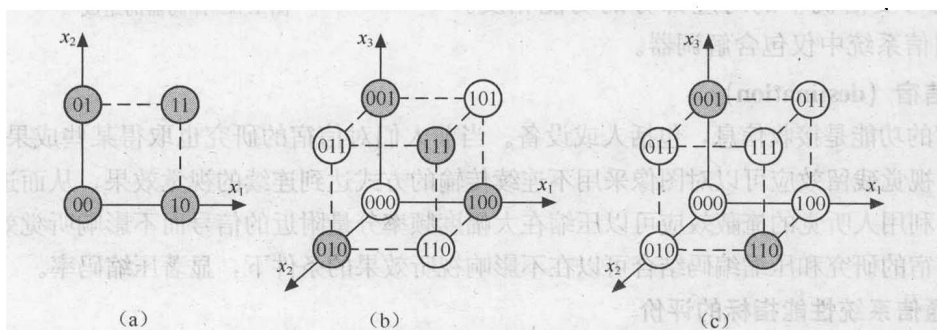


图 1.4 增加冗余符号可以提高传输可靠性

(图中，有阴影的点表示码字)

调制器的功能是，将编码器的输出符号变成适合信道传输的信号，目的是提高传输效率（使远距离传输成为可能）。信道编码符号不能直接通过信道输出，要将编码器的输出符号变成适合信道传输的信号，例如，0, 1 符号变成两个电平，为远距离传输，还需载波调制，例如，ASK, FSK, PSK 等。

3. 信道 (channel)

信道是信号从编码器传输到译码器的中间媒介。信道可以分为狭义信道和广义信道。狭义信道是某些物理通信信道，也可以是物理的存储介质。例如有线、无线、光纤、磁盘、光盘等。广义信道是一种逻辑信道，它和信息所通过的介质无关，只反映信源与信宿的连接关

系。信息论中主要研究广义信道。

信道还分为无噪声信道和有噪声信道。通常，系统中其他部分的噪声和干扰都等效成信道噪声。通信系统中主要有两种噪声：加性噪声和乘性噪声。一般地讲，背景噪声为加性，而衰落为乘性。这里主要研究加性噪声。在信息论中研究最多的是理想加性高斯白噪声(AWGN)信道。研究高斯噪声的主要原因是它的普遍性和易于处理的特性。

高斯分布的普遍性主要基于两种原因：

(1) 根据中心极限定理，无数独立随机变量的和的分布趋近高斯分布，显然高斯噪声普遍存在；

(2) 在限平均功率条件下产生最大熵的信源分布为高斯分布，而最大熵分布是最容易被观察到的分布。

与信源的分类类似，信道也分为离散信道、离散时间连续信道和波形信道（或模拟信道），其中，离散信道和离散时间连续信道输入与输出都是符号序列，只不过符号取值不同，前者取离散值，而后者取连续值；而波形信道的输入与输出均为时间连续波形。

信道也有无记忆和有记忆的区别，离散信道和离散时间连续信道可以是无记忆的，也可以是有记忆的；而波形信道通常是有记忆的。

4. 译码器 (decoder)

译码器实现与编码器相反的功能，即从信号中恢复消息。在通信系统中称做接收机，而在存储系统中称做回放系统或读出器。译码器包括：解调器、信道译码器、信源译码器，如图 1.5 所示。解调器的功能是，将信道输出信号恢复成符号；信道译码器的功能是，去掉解调器输出符号中的冗余符号；信源译码器的功能是，将信道译码器输出符号变成消息。总之，功能与发信机中的对应部分的功能相反。

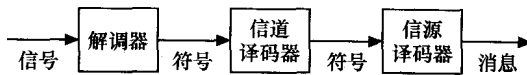


图 1.5 译码器的组成

在模拟通信系统中仅包含解调器。

5. 信宿 (destination)

信宿的功能是接收信息，包括人或设备。当前人们对信宿的研究也取得某些成果。例如，利用人的视觉残留效应可以对图像采用不连续传输的方式达到连续的视觉效果，从而进一步压缩码率。利用人听觉的掩蔽效应可以压缩在大幅度频率分量附近的信号而不影响听觉效果。总之，对信宿的研究和压缩编码结合可以在不影响视听效果的条件下，显著压缩码率。

6. 通信系统性能指标的评价

有效性用频谱复用程度（模拟系统）或频谱利用率（数字系统）来衡量。提高有效性的措施是，采用性能好的信源编码以压缩码率，采用频谱利用率高的调制减小传输带宽。

可靠性用输出信噪比（模拟系统）和传输错误率（数字系统）来衡量。提高数字传输可靠性的措施是，采用高性能的信道编码以降低错误率。

安全性用信息加密强度来衡量，提高安全性的措施是，采用强度高的密码与信息隐藏或伪装方法。

1.2.2 香农信息论的主要内容

信息论是在概率论、随机过程和通信技术相结合的基础上发展起来的学科，可分为狭义

信息论和广义信息论。狭义信息论即香农信息论或经典信息论，所研究的基本问题是：信源、信道及编码问题，其核心是3个编码定理。广义信息论包括，香农或经典信息论、信源编码、信道编码、近代信息论、统计通信理论、通信网理论、信号与信息处理、保密通信等。

香农信息论的内容可用一句话概括为：“一个概念，三个定理”，就是信息熵的概念和3个编码定理。

1. 关于信源信息的度量

如前所述，香农将信源限制为具有某一先验概率的随机过程，对信息进行量度是信息论的首要问题。香农定义了信息熵作为信源所含信息的量度，是信息论中最重要的概念。信息熵的含义与应用以及不同于香农定义的信息量度方法是研究的主要内容。

2. 关于无失真信源编码

无失真信源编码定理（香农第一定理）是信源无损压缩编码的理论基础，其内容是：如果信源编码码率（编码后传送信源符号所需比特数）不小于信源的熵，就存在无失真编码；反之，不存在无失真编码。可以简述为：

$$R \geq H \Leftrightarrow \text{存在无失真信源编码}$$

其中， R 为信源编码码率， H 为信源的熵。例如，英文字母加空共27符号，用代码传送，每个符号需5bit。但通过对大量的英文文本统计得知，实际的英文字母信源熵大约为1.4bit/符号，所以根据香农第一定理，存在某种信源编码方式，使得平均每字母仅用1.4个二进制符号就能无失真传送英文文本，这样可以显著提高传输效率。

由于定理解决的是信源无损压缩极限的理论问题，并未给出普遍的信源编码的方法，所以寻找接近或达到信源熵的压缩编码技术、分析其性能并有效实现是研究的主要问题。

3. 关于信道容量与信息的可靠传输

有噪信道编码定理（香农第二定理）是信道编码的理论基础，其内容是：如果信息传输速率小于信道容量，则总可找到一种编码方式使得当编码序列足够长时传输差错任意小，反之不存在使差错任意小的编码。可以简述为：

$$R \leq C \Leftrightarrow \text{存在译码差错任意小的信道编码}$$

其中， R 为信息传输速率，也称为信道编码码率， C 为信道容量。例如，一个带宽为3kHz的系统，用二进制符号传输，如果不进行信道编码，无码间干扰的最大数据传输速率为6kbit/s，按照香农AWGN信道容量公式可知，通过合适的信道编码，在信噪比为26dB的条件下，数据实现可靠传输的最大速率可达到25.6kbit/s。

实际上信道类型是复杂的，所以根据实际研究的问题建立信道模型，分析特性，计算容量以及在这些信道条件下各类通信系统的性能是研究的主要问题。此外，定理解决的是信息传输极限的理论问题，并未给出普遍的信道编码的方法，为此寻找接近或达到信源信道容量的信道编码技术、分析其性能并有效实现也是研究的主要问题。不过这些内容主要归入信道编码理论与技术的范畴。

4. 信息率失真理论（有损数据压缩的理论基础）

在很多情况下，我们并不需要信息精确的传输，而是容许有一定限度差错的传输。这样在保证获取足够信息的前提下，可以提高传输效率，降低通信成本。香农指出：“实际上，当我们有一个连续信源时，我们感兴趣的不是精确的传输，而只是在一个给定容限内的传

输。问题就是，当我们仅需要一定的以合适方式度量的恢复保真度时，我们能不能给连续信源分配一个确定的速率。”为实现限失真传输的有效性，我们总是希望在满足一定的失真要求条件下，使编码器的码率最小，这是一个最优有损数据压缩编码的理论问题，其理论基础就是信息率失真理论，核心是限失真信源编码定理。

限失真信源编码定理（香农第三定理）是有损压缩编码的理论基础，其内容是：对任何失真测度 $D \geq 0$ ，只要码字足够长，总可找到一种编码，使得当信源编码的码率 $\geq R(D)$ 时，码的平均失真 $\leq D$ ；反之，如果信源编码的码率 $< R(D)$ ，就不存在平均失真 $\leq D$ 的编码。可以简述为：

$$R \geq R(D) \Leftrightarrow \text{存在平均失真} \leq D \text{ 信源编码}$$

其中， R 为信源编码码率， $R(D)$ 称为信息率失真函数，是满足失真度准则（平均失真 $\leq D$ ）下，每信源符号所需的最小编码比特数。

例如，一个等概率 16 个符号的离散信源，如果要求无失真信源编码，那么每个信源符号需要 4 个二进制符号，即平均码长为 4bit。但如果允许失真，就可以减小平均码长。如果要求平均失真不超过 1/2，即对编码序列进行译码时，最多允许 1/2 的错误。很明显，我们可以将信源的前 7 个符号分别编码，而信源的后 9 个符号都编成一个相同的（第 8 个）码字传输，在接收端译码时，接收到前 7 个码字与接收端符号表的前 7 符号一一对应，无译码差错。当接收到第 8 个码字时，随机地译成接收端符号表中第 8 到第 16 个符号中的任意一个。总译码正确率为 $7/16 + 1/16 = 1/2$ ，错误率也为 1/2，即编码失真为 1/2。如果用最优化二元编码（例如，Huffman 编码），编码的平均码长可达 2.25bit，小于 4bit。但是，在理论上，同样的信源，为实现 1/2 差错率的信源编码最小的码长由 $R(D)$ 函数来决定，为 1.05bit。

与无失真信源编码类似，限失真信源编码定理解决了有损压缩极限的理论问题，也未给出实际的有损信源编码的方法。在各种失真度准则下对信源 $R(D)$ 函数的研究，利用率失真理论对实际有损压缩编码性能的分析，寻找接近或达到理想有损压缩的信源编码方法是该领域研究的主要问题。

5. 信息论的特点

现将香农信息论的特点总结如下：

- (1) 以概率论、随机过程为基本研究工具；
- (2) 研究通信系统的整个过程，而不是单个环节，并以编、译码器为重点；
- (3) 关心的是最优系统的性能和怎样到达这个性能，并不具体设计系统；
- (4) 语法信息中的概率信息，要求信源为随机过程。

1.3 香农信息论研究的进展与应用

1.3.1 香农信息论创立的背景

应该指出，香农信息论的创立主要是由于香农的杰出贡献，但也与当时的技术发展背景和前人的工作密不可分。