

数据压缩

许织新 编著

国防工业出版社

数 据 压 缩

许 织 新 编 著

國 防 工 業 出 版 社

内 容 简 介

本书重点阐述了数据压缩理论和技术三项主要内容,即预测编码、变换编码和统计编码。为了深入地掌握这些原理和方法,还介绍了数据压缩的理论极限和信源统计特性分析的有关问题。作为提高性的内容,又将数据压缩的基本理论——码率失真理论作为一章,供有条件的读者赏阅。

本书适合于数字通信、信号处理、信息工程等专业的师生作教材用。也可从事于上述领域的科研、工程技术人员作参考书籍。

数 据 压 缩

许织新 编著

*

国防工业出版社出版、发行

新华书店经售

国防工业出版社印刷厂印刷

*

787×1092 1/16 印张 21 482千字

1990年1月第一版 1990年1月第一次印刷 印数: 0,001—2,000册

ISBN 7-118-00268-2/TN51 定价: 10.70元



笔者寄语

本书的原型,是笔者在国防科技大学任教期间,为本科高年级学生及研究生编写的专业课教材。经过多年来在校内外教学和科研实践中的风风雨雨,不少质疑和嘉许,都为本书提供了珍贵的鉴益。如果说那本原型还只能是在搜集资料基础上,简单编集和加工而成的粗材的话,那末经过不断的补充和完善,今天所敬献给读者的,无妨说已是焕然一新之卷了。

就目前而言,以“数据压缩”这一主题出现的,较为完整和系统的专著,在国内外似不多见。因此,立志于填补这一尚未开发的空隙,便是笔者本人多年来的宿愿和奢望。也正由于这样,所以本书在构思、体系、内容和写法上,颇有无从借鉴之苦。自然也当有值得推敲斟酌之处。加之笔者的水平有限,故书中的错漏自当难免,伏望前辈和各界人士不吝赐教!

本书在编写过程中,承蒙国防科技大学刘景伊、周祖同教授、汪漱玉、王绍霖副教授,以及杨述明讲师等同志审阅了文稿,并在实践过程中,还得益于许多同志的指点和教诲。另外,李晓军、周梁两同志也仔细校阅过修改稿,提出过不少中肯的意见。在出版过程中,曾得到航天工业部李邦复高级工程师和国防科技大学孙仲康教授的帮助和支持。在此则一并致以由衷的谢意!

许织新 谨 识 1984. 10.

目 录

第一章 结论	1
§ 1.1 数据压缩的内容、意义和用途	1
§ 1.2 数据压缩技术发展的沿革	2
§ 1.3 数据压缩技术的分类	2
§ 1.4 本书的宗旨、体系和建议	3
第二章 数据压缩的理论极限	5
§ 2.1 概述	5
2.1.1 数据压缩与信息论	5
2.1.2 数字传输系统的模型	6
§ 2.2 离散无记忆信源的熵	7
2.2.1 信息的度量和熵	7
2.2.2 熵与概率分布的依从关系——数据压缩的途径之一	10
2.2.3 熵与信源相关性的依从关系——数据压缩的途径之二	12
§ 2.3 互信息量	15
2.3.1 有扰通道模型	15
2.3.2 互信息量的定义	16
2.3.3 平均互信息量	16
2.3.4 互熵的性质	17
2.3.5 互熵与数据压缩的极限	19
§ 2.4 连续信源的互信息量	20
2.4.1 连续信源熵的引出	20
2.4.2 连续互熵的定义和性质	23
§ 2.5 随机序列的信息量	25
2.5.1 平稳过程模型	25
2.5.2 定义和性质	25
2.5.3 举例和讨论	27
§ 2.6 模糊信息与广义熵	28
2.6.1 模糊集合与模糊信息	28
2.6.2 广义信息与广义熵	30
2.6.3 模糊信息论对数据压缩的指导意义	35
习题	36
第三章 信源统计特性分析	40
§ 3.1 信源的数学模型	40
3.1.1 引言	40
3.1.2 信源的确定性模型	41
3.1.3 信源的参数性统计模型	41
3.1.4 信源的非参数性统计模型	43
3.1.5 马尔科夫过程模型	47
3.1.6 二维信源模型的特殊性	53
3.1.7 小结	56
§ 3.2 图象信号的统计特性	57

3.2.1	图象的数学描述方法	57
3.2.2	图象信号的自相关函数	61
3.2.3	图象信号的功谱密度函数	66
3.2.4	图象信号的幅度分布	68
§ 3.3	语音信号的统计特性	69
3.3.1	语音产生的物理机理与数学模型	69
3.3.2	语音信号的自相关函数	70
3.3.3	语音信号的功谱密度函数	72
3.3.4	语音信号的幅度分布	73
§ 3.4	遥测信号的统计特性	74
3.4.1	遥测信号形式举例和分类	74
3.4.2	遥测信号的自相关函数	75
3.4.3	遥测信号幅度的差分布	76
3.4.4	遥测信号的条件概率	77
3.4.5	遥测信号的熵	77
§ 3.5	计算机数据的统计特性	78
§ 3.6	结语	79
	习题	79
第四章	预测编码理论基础	82
§ 4.1	引言	82
4.1.1	简要的回顾	82
4.1.2	预测编码的基本型式	83
4.1.3	线性预测模型的导出	85
§ 4.2	最小均方误差准则	88
4.2.1	确定信号	88
4.2.2	随机信号	90
4.2.3	预测参数的计算	91
4.2.4	预测滤波器的稳定性	97
4.2.5	响应函数和 G 的计算	98
§ 4.3	功谱包络匹配准则	100
4.3.1	频域公式	100
4.3.2	功谱包络匹配	101
4.3.3	预测误差分析	104
4.3.4	最小均方误差准则和功谱包络匹配准则的比较和其它	108
§ 4.4	预测系数不变性准则	108
4.4.1	公式的导出	109
4.4.2	性能讨论	111
§ 4.5	量化特性	111
4.5.1	通用最优量化	111
4.5.2	预测误差 e_n 的量化	113
4.5.3	预测参数的量化	115
4.5.4	几种编码型式的量化性能分析	117
4.5.5	向量量化	121
§ 4.6	线性预测性能分析	124
4.6.1	线性预测的信噪比	124
4.6.2	信道误码的影响	126
§ 4.7	最大误差准则	128
4.7.1	多项式预测模型	128

4.7.2	最大误差准则下预测系数的推导	130
4.7.3	压缩效率与性能分析	131
4.7.4	信道误码对多项式预测性能的影响	137
4.7.5	孔径 K , 内插法及其它	138
§ 4.8	结语	139
	习题	139
第五章	变换编码理论基础	143
§ 5.1	引言	143
5.1.1	数据压缩与变换编码	143
5.1.2	变换编码的数学模型	145
5.1.3	变换编码的系统构成	146
5.1.4	变换编码的分类及发展沿革	147
§ 5.2	最优变换 ($K-L$ 变换)	147
5.2.1	公式的导出	148
5.2.2	方法及讨论	152
§ 5.3	离散傅氏变换	153
5.3.1	次优变换	153
5.3.2	DFT 的定义及功能分析	154
5.3.3	DFT 的快速算法	157
§ 5.4	离散余弦变换	159
5.4.1	公式的引出及定义	159
5.4.2	EDCT 矩阵的写法和性能分析	161
5.4.3	DCT 的算法与反 DCT (IDCT)	162
§ 5.5	离散正弦变换	163
5.5.1	定义及公式	163
5.5.2	DST 的正交性	163
5.5.3	DST 对角化矩阵的条件	165
§ 5.6	沃尔什-阿达玛变换	166
5.6.1	非正弦正交函数的引入	166
5.6.2	沃尔什-阿达玛变换	170
5.6.3	WHT 的性能估计	174
5.6.4	快速 WHT	177
§ 5.7	其它变换	179
5.7.1	哈尔变换	180
5.7.2	斜变换	182
5.7.3	二维变换略述	186
5.7.4	关于各种变换方法的评述	188
§ 5.8	变换域的统计分析	188
5.8.1	输出的协方差及功谱	188
5.8.2	输出的幅度分布函数	189
§ 5.9	变换采样准则	191
5.9.1	带状采样	191
5.9.2	门限采样	193
§ 5.10	变换样本的编码	195
5.10.1	分组量化法	195
5.10.2	主观评定因素	198
§ 5.11	参数选择和性能比较	198
5.11.1	变换区间 N 的选择	199

5.11.2 信道误码的影响	199
5.11.3 变换编码与预测编码	201
§ 5.12 结语	203
习题	203
第六章 统计编码理论基础	210
§ 6.1 引言	210
6.1.1 统计编码的基本思想	210
6.1.2 统计编码的分类	213
§ 6.2 变长码的基本原理	213
6.2.1 问题的提出	213
6.2.2 树码及非续长性	214
6.2.3 单义可译码	218
6.2.4 变长信源编码定理	219
6.2.5 变长码的扩展	221
6.2.6 缩短分组码	222
§ 6.3 最优变长码——哈夫曼码	223
6.3.1 最优变长码的理论基础	223
6.3.2 最优变长码的编码方法	227
6.3.3 特定情况下的哈夫曼码	231
6.3.4 信源概率起伏的影响	232
6.3.5 哈夫曼码的冗余度	234
6.3.6 哈夫曼码的发展和前景	235
§ 6.4 其它变长码	236
6.4.1 香农-费诺码	236
6.4.2 哈夫曼-香农-费诺码	240
6.4.3 变移哈夫曼码	242
6.4.4 无失真通用伴随式信源编码	244
§ 6.5 定长编码	246
6.5.1 定长编码的基本定理	247
6.5.2 多径搜索编码的基本方法	251
6.5.3 码树/栅格编码	253
§ 6.6 结语	259
习题	260
第七章 码率失真理论基础	265
§ 7.1 引言	265
7.1.1 码率失真理论的目的和意义	265
7.1.2 码率失真理论研究的范围	266
§ 7.2 离散无记忆信源的 $R(D)$ 函数	267
7.2.1 失真的度量	267
7.2.2 保真度标准	268
7.2.3 试验信道的互熵和码率失真函数	268
7.2.4 无记忆随机序列的失真度量	269
7.2.5 无记忆随机序列的 $R(D)$ 函数	270
§ 7.3 $R(D)$ 函数的性质	271
7.3.1 定义区间 $D_{\min}$ 的确定	271
7.3.2 定义区间 $D_{\max}$ 的确定	273
7.3.3 $R(D)$ 是凸函数	275
7.3.4 $R(D)$ 是严密递减函数	275

VIII

7.3.5	$R(D)$ 函数的基本性质	279
§ 7.4	$R(D)$ 函数的计算	279
7.4.1	$R(D_{\min})$ 的计算	279
7.4.2	离散信源 $R(D)$ 函数的参量表述法	280
7.4.3	离散信源 $R(D)$ 的计算例	285
7.4.4	连续无记忆信源 $R(D)$ 函数的参量表述法	289
7.4.5	连续信源 $R(D)$ 的计算例	293
§ 7.5	有记忆信源的 $R(D)$ 函数	296
7.5.1	有记忆信源模型	296
7.5.2	有记忆信源的 $R(D)$ 函数	297
7.5.3	$R(D)$ 函数的频域计算	299
§ 7.6	用 $R(D)$ 理论证明信源编码定理	301
§ 7.7	码率失真函数与信道容量代价函数的对偶关系	304
7.7.1	容量代价函数简述	304
7.7.2	$R(D)$ 与 $C(\beta)$ 的对偶关系	304
§ 7.8	结语	305
附录 A	z 变换和数字滤波	306
§ A.1	引言	306
§ A.2	z 变换	307
A.2.1	z 变换的定义和性质	307
A.2.2	z 变换与拉氏变换	308
A.2.3	反 z 变换	308
§ A.3	数字滤波器的构成	309
A.3.1	一次非递归形滤波器	309
A.3.2	一次递归形滤波器	310
A.3.3	双二次形滤波器	311
§ A.4	数字滤波器的原理设计方法	311
A.4.1	标准 z 变换	311
A.4.2	双线性 z 变换	312
A.4.3	匹配 z 变换	313
§ A.5	二维数字滤波器	314
A.5.1	二维采样定理的证明	314
A.5.2	二维线性系统	317
A.5.3	二维 z 变换	318
A.5.4	二维差分方程	318
附录 B	熵 $H(p)$ 函数表	319
参考文献		320
主要符号一览表		324

第一章 绪 论

§ 1.1 数据压缩的内容、意义和用途

数据压缩 (Data Compression) 是这样一门科学和/或技术, 它以一定的质量损失为容限, 按照某种方法从给定的信源中推拿 (Mas-Saging) 出已简化的或“被压缩”的数据表述。大家清楚, 任何物理信源泛泛的自然表述, 并不总是为反映其本质所必须的。古谚语说: “只有精纯, 才是本质的伟大朋友”。所以, 探求任何信源的这种精纯的表述, 而又不致带来重大的失真, 这就是数据压缩所要解决的命题。因此也可以说, 数据压缩是一门信息处理的科学和/或技术。

我们知道, 信息科学的一个重要分支, 就是**信源编码**。它除了必要的模数转换以外, 数据压缩就成为信源编码的基本内容。因而可以说, 数据压缩多少是面向信源的一种信息科学和/或技术。

显然, 这样一种减少数据的努力, 对社会经济的发展来说, 是有积极意义的。这也许是由于存储容量的限制, 也许是信道带宽或容量上的不足, 也可能是希望精选出有属性数据的某种要求等等。因此, 这是一种节省社会财富, 最大限度地利用信息和物质资源的积极措施。所以对于我国的社会主义现代化建设来讲, 不无重大意义。从目前来看, 数据压缩的应用领域颇为广泛, 诸如通信、语音和图象处理、模式识别、信息恢复、信息存储和保密学等等。为了更具体地阐明它的用途, 不妨借鉴一下国外有重要意义的应用场合, 作为我们的参考。

美国国家海洋气象局 (NOAA) 从卫星上接收到的地面图象, 从指令数据录取站向数据处理和分析中心传输时, 用了一条 300km 的微波信道。在 1969 年以前一直没有使用数据压缩技术。但后来由于要接收和传输更高分辨率的气象图片, 因此由南加州大学为其研制了一整套压缩算法供其使用。1972 年后已正式投入运行。其方法是预测编码再加上统计编码的复合技术, 压缩比是 2:1。硬件是用计算机来实现。据报道, 效果是良好的。

在美国著名的**阿波罗**宇航计划中, 从载人空间飞行网络站到哥达特空间飞行中心之间, 有一条传输阿波罗数字遥测数据的线路, 其容量为 $2 \times 2400\text{bit/s}$ 。但实际上从飞船向地面传来的遥测数据是 51.2kbit/s 。因此由约翰霍普金斯大学为其研制了一套数据压缩算法, 主要也是预测编码与统计编码的复合运用。这项技术已得到我国遥测界人士的赞赏, 并已研究付诸使用。

众所周知, 在最新的航天飞机实施计划中, 语音的传输已采用某种自适应的预测编码技术, 把一路数字语音从 64kbit/s 压缩到 32 或 24kbit/s 。并准备在后续型号中, 对电视图象采用二维预测编码技术, 可将 64Mbit/s 的码率压缩到 13.1Mbit/s 。对上行传真图片, 则采用统计编码, 压缩比可达 5:1。

在目前出现的大型计算中心那里, 常常有浩大的数据库存在。为了减少这种日益增

长的存储要求的压力，加拿大贝尔公司就运用统计编码方法进行数据压缩，其压缩比为30:7。

除了上面提到的几种压缩方法以外，还有一种变换编码法也正被应用到实际中去。比如，美国在无人驾驶飞机上已经应用这种能够压缩图象的实用技术等等。

凡此种种，确实是不胜枚举。有兴趣的读者可参阅有关资料。

§ 1.2 数据压缩技术发展的沿革

可以指出，最早的数据压缩的理论和实践，应该追溯到十八世纪末期。这在1898年，W. F. Sheppards首先研究了任意实数如何舍入成固定的十进位数的问题，并发表了相应的论文：“《On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equidistant Divisions of a Scale》Proc. London Math. Soc. 24, Pt2, 353~380”。显然可看出，这个问题实际上就是最基本最古老的数据压缩命题。从此以后，数据压缩就沿着两条不同的途径发展起来。一方面，是许多数学工作者所遵循着的。他们试图去建立信源和数据压缩系统的数学模型，并从中找出一些质量指标及最优性能界限，以便去比较这些系统的优劣。而另一方面，则是由更多的实用工程师们所从事着的。他们更关心于如何设计一个精巧的系统，来实现预定的数据压缩功能，或者去分析或模拟这些特殊的系统，以确定它们的性能指标。后来，由于经典信息论和估值理论不断发展，也由于对实际的数据压缩系统的深入剖析，就导致了新的抽象数学理论的应用。于是这两个方面都出现了殊途同归、互相结合的趋势。

必须指出，和信息理论的其它分支一样，在六十年代初期，美国和西方世界对这一理论的研究出现了低潮。但是事隔不久，由于空间计划的推进，计算技术的发展，特别是模式识别和图象处理科学的兴起，数据压缩理论和实际工作的发展，又再次崛起。特别是作为它的基础理论——码率失真理论，也在六十年代后期得到较大的进展，并且在七十年代里，数据压缩就开始接近于工程实际应用阶段了。到现在来看，可以说这门学科已经日趋成熟。

§ 1.3 数据压缩技术的分类

大家知道，数据压缩还有许多同义词被学术界所采用。比如，Data Compaction, Bandwidth Compression, Redundancy Removal, Redundancy Reduction, Adaptive Sampling, 等等。但是，这些词语所反映的内容，除了细节上有所差异外，从本质上来讲都同出一辙。因而在本书里就不再予以区别了。

至于数据压缩技术的分类，则更是五花八门。据有人统计，仔细分来可达三四十种。但是如按大的特征来区分，一般可以有如下几种分法。

有人认为，数据压缩或信源编码基本上可分为分组编码 (Block Codes) 和非分组编码 (Nonblock Codes) 两大类。前者是将连续的非重叠的数据组，处理成已编码的非重叠的数据组。这样分组的长度既可是固定的，也可以是可变的。实质上它对每一组新的数据，都是重新开始处理。而后者则不然，它的编码输出既可能依赖于过去，也可能依赖于将来的数据值。当然，它也可以处理连续的数据组，但它们是互相重叠的。因

此既受过去值影响,也受未来值影响。所以它实质上是一个具有记忆和延迟的非线性滤波器。至于这两者的具体体现,将在后面讨论时细述。

也有人认为,数据压缩技术可分为**熵压缩法**(Entropy Compression)和**无失真编码法**(Noiseless Coding)两种。前者是一种不可逆的处理,也就是说在压缩过程中要损失一部分熵,因而在再现时就不能完整无误地复演出来。而后者又可称为**准确编码**、**熵编码**、**信息保持编码**或**无失真冗余度压减**等。这是一种可逆的处理,由于信源的熵没有受损失,因而在再现时就能准确无误地复演出来。

最后一种分法多半更能说明它们的本质区别,且也是我们所将依循的分类法。不妨认为,从目前实用的压缩技术来讲,似以分为**预测编码**、**变换编码**、**统计编码**这三大类为好。因为它们分别以不同的理论准则为指导,且可把一些具体的压缩方法概括在一起。所以这是目前普遍采用和受欢迎的分类法。

§ 1.4 本书的宗旨、体系和建议

1. 根据多年来的实践和调研,确定本书的宗旨为:

(1) 由于数据压缩在理论研究和工程实践方面,都有丰富的内容和材料,所以本书既不拘泥于某些工程技术方法的罗列和比较,也不致力于纯理论的推导和分析。确切地说,是从一般工程技术方法中抽出本质的规律来,并以工程技术人员所能接受的理论深度加以提高。因而是一种理工结合的产物。

(2) 由于数据压缩与信息论的关系是如此紧密,而且信息论知识的普及又势在必行,所以本书注意到把数据压缩与香农的信息理论尽可能地联系起来。为尔后深入学习信息论而创造有利的条件。

(3) 为了照顾到不同读者的需要和可能,本书可以为基础性的进度,比如大学本科学生的程度作出安排。也可以为提高性的进度,比如研究生或在职科技人员的程度作出安排。

2. 根据上述宗旨和它的具体内容,本书的体系为:

(1) 三条主线

以**预测编码**(第四章),**变换编码**(第五章),**统计编码**(第六章)为三条主线,详细阐述数据压缩的原理和方法。

(2) 二项基础理论

所有数据压缩技术,都以信源为其研究对象。因此有必要在讲述前面三条主线之前,对**信源统计特性**作一基础分析,故写为第三章。但就此而言,还需要有统计理论和信息理论作指导,来找到数据压缩的理论极限。同时,这些信息论知识还直接指导着三条主线的展开。因此在这之前,还安排有**信息论基础**作为第二章。

本来,**码率失真理论**是数据压缩的基础理论,但考虑到它的内容较深,牵涉到许多抽象的概念和冗长的推导,所以只宜作为提高性进度的内容,故列为第七章。

(3) 二个附录

在学习预测编码理论及其它内容时,要有一定的 **z 变换**和**数字滤波**的知识。因此在最后列为附录A,以供不具有此方面知识的读者作准备。另外,在附录B里,列出了有关计算熵值的表,以供查阅。

这样就构成了本书完整的体系和安排。请见图 1-1。此外,本书在每一章的末尾,都选辑有若干练习题,可供读者加深和巩固正文内容之用。

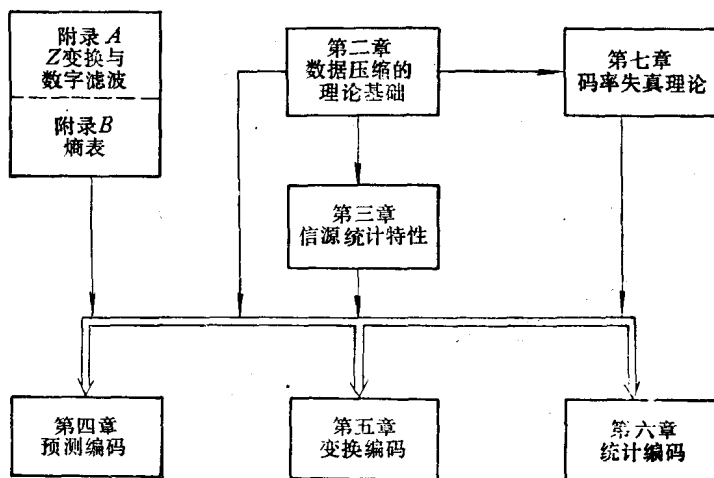


图1-1 全书的体系和安排

3. 对于大专院校本科学生或相当程度的读者来说,可以不学码率失真理论,而其余部分是比较接近于工程实际的。对于研究生和专门从事于本专业的工程技术人员来讲,建议在学好码率失真理论的基础上,把握住整个数据压缩的全局。

对于一些想要掌握信息论中有关信源编码内容的读者,建议学好第二章和第六章的有关章节。因为它们已把有关信源编码定理包含在内了。另外第三章也应有所了解。

对于一些急切想掌握工程设计原理的读者,也可以单刀直入地学好第四、五、六章,以一睹方法为快。如在必要和条件许可时,回过头来再钻研第二、三、七章的内容,也不失为一条行之有效的捷径。

第二章 数据压缩的理论极限

§ 2.1 概 述

2.1.1 数据压缩与信息论

要论述数据压缩与信息论这两者关系，我们可以先从一个传说讲起⁽¹⁾。

据说在欧洲的一个小城镇里，有一位年青学者，名叫阿尔伯特 (Albert)。他在地区专利局里得到了一个从事于专利审核工作的职位。他自信是不难胜任的。可是就在他到职的头一天，上司就吩咐他去审查一项有关通信和信息处理的专利。在这份专利中，申请人自称已经制成了一种装置，它能将任意一份计算机存储器中的英文文件，以 10 个拉丁字母只需 1bit 的压缩率存储起来。他特别留意到是任意一份。

阿尔伯特一开始就觉得十分沮丧，因为他有一种束手无策的预感。于是他只好移樽就教，找他的老朋友克劳德去了。克劳德是一位受人尊敬的通信工程师，他在听取了阿尔伯特有关该项专利的阐述以后，很快就告诉阿尔伯特说：“这个所谓的发明，已经破坏了信息论中的基本极限，因而也是不可能实现的。”他同时还建议阿尔伯特去学一点信息论知识，只有这样才能据理以斥。

经过几个月的苦心钻研以后，阿尔伯特懂得了这样一个道理，就是在 1948 年由香农 (Shannon) 所创立的信息论，对数据压缩这门学科来说，有着极重要的指导意义。它一方面给出了数据压缩的理论极限，而另一方面又指明了数据压缩的技术途径。

就基本极限而言，香农信息论认为信源所含有的平均信息量（或称为熵），就是进行无失真压缩的理论极限。举例来说，如将由 26 个字母和一个间隔符号所组成的英文文件看成是某个信源的话，那么该信源的熵可算出为

$$H(X) = 1 \sim 2 \text{ bit/字符}$$

由于计算机内的文件是不允许有任何失真的，所以对这种信源进行压缩后的平均字长 $\bar{n}$ ，就不能低于 $H(X)$ 。换句话说，低于此极限 $H(X)$ 的无失真编码方法是找不到的。而只要不低于此极限，那就总能找到某种适宜的编码方法，使 $\bar{n}$ 任意地逼近于 $H(X)$ 。这样，就使阿尔伯特最终看清了该份专利的荒谬所在。用大大低于 $H(X)$ 的 0.1bit/字符的平均字长，去无失真地压缩任意一份英文文件，就好比在热力学中去搞永动机那样天真和可笑。

就压缩途径而言，香农信息论认为信源中含有或多或少的自然冗余度。这些冗余度既来自于信源本身的相关性中，又来自于信源概率分布的不均匀性中。于是只要找到去除相关性，或改变概率分布不均匀性的方法和手段，就可以实现有效的压缩。大家知道，目前计算机中英文文件的编码，普遍采用了 ASCII 码，每 1 字符需用 7 个比特。这就使该信源含有较多的自然冗余度。其原因就在于文件中字符之间，常常具有较大的关联性，而并不是相互独立的。同时，26 个字母所出现的机会也远非均等，一般来说，字母

“e”出现的概率最大，而“q”的概率最小。这样看来，该份专利虽然并不可取，但恰到好处的压缩却又是可能的。这就使阿尔伯特多少有些沾沾自喜，因为这是他在审核这份专利后的意外收获。

我们所讲的传说就此结束。发人深思的不只在于该传说本身，而还在于启示我们为什么要去研究现代科学领域中的一个重要学科——信息论。仅仅以本书的主题——数据压缩而言，其目的就是要

(1) 掌握数据压缩的理论极限。

(2) 找到数据压缩的基本途径。

这就是本章所阐述内容的宗旨所在。

2.1.2 数字传输系统的模型

在现代信息社会里，信息传输将是信息科学中最基本问题之一。而且，如从更广义的角度上去理解，它就不仅限于经典的通信问题。信息如从“这里”传输到“那里”，这当然是典型的“通信”的概念。信息如从“现在”传输到“将来”，那就是所谓“存储”的概念。于是，不妨将这两种物理过程，用统一的数字传输系统的模型加以概括，以俾能用更抽象的方法予以研究。此模型示于图 2-1 中。

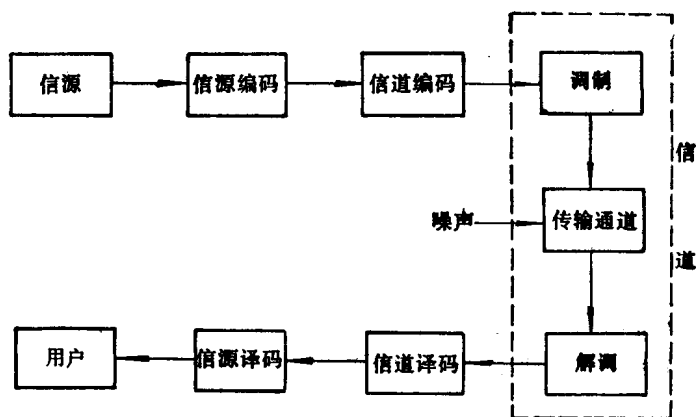


图2-1 数字传输系统模型

应该指出，香农正是在该模型基础上创建了著名的信息论，而且形成了信源编码 (Source Coding) 和信道编码 (Channel Coding) 两大分支。信源编码是面向信源的一种信息处理手段，而信道编码是面向信道的另一种处理手段。两者既有区别，但又有联系。当然，信道有较广泛的内在含义，比如既可表示电缆、微波等传输介质，又可表示磁盘、磁带等存储器件。它们在传输信息过程中，均不同程度地遭受着噪声的干扰。因而在信道的输出端，就会出现差错现象。

如果能够根据信道的特性，来妥善地设计信道编译码器，从而使得这种差错现象全部或绝大部分能纠正过来，那么在信源编码器和信源译码器之间，呈现出一条无噪声信道 (Noiseless Channel)。于是整个数字传输系统又可简化成如图 2-2 所示那样的模型，称为信源编码模型。这就意味着可把信源编码器和信源译码器直接连通起来。

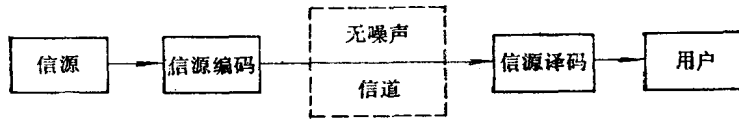


图2-2 信源编码模型

我们通常把信源看成是一个随机数据发生器，整个数字传输系统的职能，就是有效而又可靠地将此随机数据传送到用户那里。因而对信源编码器来说，它的任务首先就是将信源的输出，变换成相应的数字序列。假如信源本身能产生数字输出的话，那么信源编码器就一一对应地将其变换成另一种适宜的数字序列，而信源译码器就一一对应地反变换成原始的数字数据。

但是，客观世界的许多物理量，往往是渐变和有连续性的。所以，它的输出就不再是数字序列，而是在时间上和幅度上，都为连续变化的模拟信号。或者说，它的取值是一个不可数的无限集合。因此这时的信源编码器就做不到一一对应的变换，而只能在容忍一定失真的条件下，变换成数字序列。当然信源译码器也只能做到近似的反变换。在这种情况下，信源编码器要完成模-数变换功能，而信源译码器可能要完成数-模变换功能。

不过，信源编码器有着更为重要的主题内容，就是数据压缩。这意味着如何能以最少的比特数，来表述所给定的信源。并在信源译码器那里，能准确地或尽可能准确地再现出信源数据。因此，信源编码作为信息论的一个分支来讲，其核心内容仍是个数据压缩问题。况且正如在绪论里已提过的那样，从广义上来认识模数转换这一功能，实质上它也是个数据压缩问题。所以从这个意义上说，就可以把信源编码和数据压缩这两个概念等同起来。

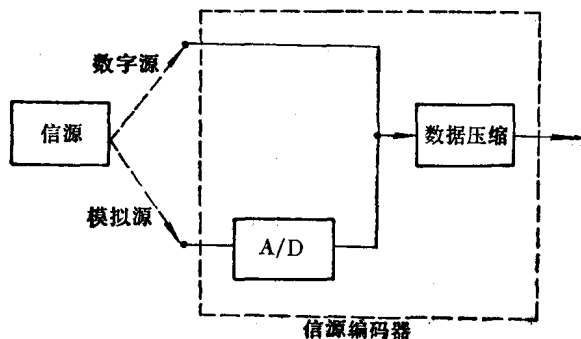


图2-3 信源编码器构成图

综上所述，就可得到一个完整的信源编码器的构成框图，如图 2-3 所示。于是，运用香农信息论有关信源编码这一部分的理论和概念，来研究数据压缩的基本极限和技术途径，就显得十分有力而贴切。

§ 2.2 离散无记忆信源的熵

2.2.1 信息的度量和熵⁽²⁾

1. 离散无记忆信源的定义

为了便于引入信息的有关概念和定理，通常总是从最基本而又最简单的离散无记忆信源出发，因此有必要对这种信源作一明确的定义。

定义2-1 若信源具备下述特性：

(1) 它的输出是一随机变量 x ，而此变量是从某个有限字母集合 $A = \{a_1, a_2, \dots, a_n\}$ 中选取。

(2) 某个字符 a_i 出现的概率 $P(a_i) = p_i$ 也已确定。

(3) 信源送出的随机变量，与前一时刻和后一时刻的输出互相独立。

则称此信源为离散无记忆信源 (Discrete Memoryless Source, DMS)。本节所有的讨论，都将限于这种信源而不再附加说明。

2. 消息与信息

在人们日常生活中，如果收到一封电报，接到一个电话，或收到一组数据，从信息论的观点来看，这些都只能被叫做收到“消息”，而不能说收到“信息”。消息和信息，其英文分别为 *Message* 和 *Information*，其日文分别为“通报”和“情报”，可见是有区别的。不过，信息寓于消息之中，而且同样一个消息会带有不相等的信息量。这是因为每个消息送到用户那里，对用户所提供的收益和价值，将因事因地而异。比方同样收到一封仪器试验成功的电报，如果原先已经预计到成功的可能性是 99%，那么它所带来的信息量是微少的，好象有“不出所料”之感。但是如果预计成功的可能性不大，比如只有 1% 的可能性，而这时收到成功的电报，就会给人们提供极大的信息量，似乎有“十分意外”的味道。此时对人们心理上的鼓励作用，在技术上敢于肯定的积极态度，都表明该消息所含有的信息量是多的。

那么用什么样的科学语言来定义信息的度量呢？由上述例子可见，在未收到电报之前，接收者对是否会出现成功这一事件，是处于某种疑惑的或不肯定的状态中。而且对前一情况而言，它的不肯定性要小。后一情况的不肯定性则要大得多。而在收到成功电报以后，接收者就解除了这些疑义度或不肯定性。显然，前一情况所解除的不肯定性要少，后一情况则多。因此，可用解除不肯定性的多少来描述从消息中所获得信息量的多少。这样一来，关于信息度量的定义就可写为

定义 2-2 用户在收到消息之前，是处于某种不肯定状态之中，而在收到消息之后，就解除或部分解除了此不肯定性，从而获得了信息。因此解除不肯定性的多少，就可作为信息的度量。

3. 不肯定性的数学描述

用什么数学模型来描述不肯定性呢？香农信息论则是应用了“概率”这一概念。也就是说，事件出现的概率小，则相当于不肯定性多。反之则少。如在上例中，可能出现的事件是二个，即成功与失败。如设成功事件为 a_1 ，其概率为 $P(a_1) = p_1$ ；失败事件为 a_2 ，其概率为 $P(a_2) = p_2$ 。则可记不肯定性是 p_i 的函数，为 $\mathcal{H}(p_i)$ 。但问题还在于要找到一个确切的函数关系。

为了科学地定义不肯定性函数 $\mathcal{H}(p_i)$ ，必须考虑到如下几个基本要求，

(1) 当事件为肯定事件时，即 $p_i = 1$ ，则要有

$$\mathcal{H}(p_i) = 0 \quad (2.2.1)$$

此时，即毫无不肯定性可言。

(2) 当 p_i 愈小时，不肯定性 $\mathcal{H}(p_i)$ 愈大。反之则愈小。

(3) 如有二个事件集同时存在，则有联合不肯定性 $\mathcal{H}(p_{i,k}) = \mathcal{H}[P(a_i, b_k)]$ 存在。其中 $P(a_i, b_k)$ 是二个事件的联合概率。但当二个事件集互相独立时，能有