实验三、实验四
预习材料

一、 信源编码

1.1 通信系统模型

一般的通信系统模型如图 1 所示：

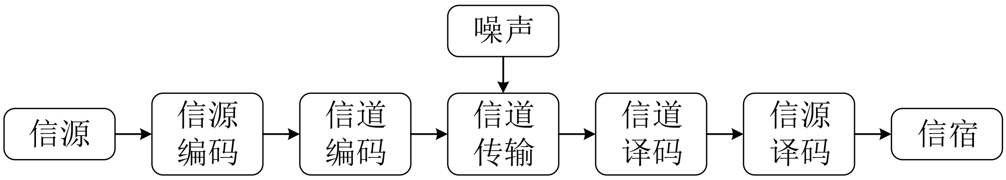

图 1 通信系统模型

可以发现，信源编码处于通信系统模型的最前端，是信息进行传递时的第一个步骤，对信息传输具有重要意义。

1.2 信源编码的定义

在计算机科学和信息论中，信源编码是按照特定的编码机制用比未经编码少的数据比特（或者其它信息相关的单位）表示信息的过程。信源编码是一种以提高通信有效性为目的而对信源符号进行的变换，或者说为了减少或消除信源冗余度而进行的信源符号变换。具体说，就是针对信源输出符号序列的统计特性来寻找某种方法，把信源输出符号序列变换为最短的码字序列，使后者的各码元所载荷的平均信息量最大，同时又能保证无失真地恢复原来的符号序列。

1.3 信源编码的作用

信源编码的作用之一是，即通常所说的数据压缩；作用之二是将信源的模拟信号转化成数字信号，以实现模拟信号的数字化传输。

1.4 信源编码的分类

根据信源的性质进行分类，则有信源统计特性已知或未知、无失真或限定失真、无记忆或有记忆信源的编码；按编码方法进行分类可分为分组码或非分组码、等长码或变长码等。然而最常见的是讨论统计特性已知条件下，离散、平稳、无失真信源的编码，消除这类信源剩余度的主要方法有统计匹配编码和解除相关性

编码。比如香农码、哈夫曼码，它们属于不等长度分组码，算术编码属于非分组码；预测编码和变换编码是以解除相关性为主的编码。对限定失真的信源编码则是以信息率失真函数$R(D)$为基础，最典型的是矢量量化编码。对统计特性未知的信源编码称为通用编码。

二、 香农编码

2.1 香农编码简介

香农第一定理指出了平均码长与离散信源概率之间的关系，同时也指出了可以通过编码使平均码长达到极限值，这是一个很重要的极限定理。根据香农第一定理，当选择每个码字的长度l_i为满足下式的一个整数时，这种编码方法就是香农编码。

$$I(x_i) \le l_i \le I(x_i) + 1$$

其中$I(x_i)$表示信源符号x_i的信息量，即，$I(x_i) = -log_2 p(x_i)$。

经过香农编码之后，码字的平均码长$\bar{L}_S$为：

$$\bar{L}_S = \sum_{i=1}^{n} p(x_i) * l_i$$

其中，$p(x_i)$表示信源符号x_i的统计概率，l_i表示信源符号x_i经过香农编码得到的码字的长度。

香农编码的效率为：

$$\eta = \frac{\text{信源熵}}{\text{平均码长}} = \frac{H(x)}{\bar{L}_S}$$

2.2 香农编码流程

香农编码严格意义上来说不是最佳码，它是采用信源符号的累计概率分布函数来分配码字。

其编码具体步骤如下：

(1) 将信源符号按概率从大到小顺序进行排列；

(2) 根据不等式$-\log_2\big(p(x_i)\big) \le l_i \le -\log_2\big(p(x_i)\big) + 1$，计算信源符号$x_i$对应的码字的码长（$l_i$取在此范围内的整数）；

(3) 计算排序后信源符号x_i的累加概率P_i；

(4) 将累加概率P_i变换成二进制小数，取小数点后的l_i位二进制数作为信源x_i的码字。

香农编码的效率不高，实用性不大，但对其他编码方法有很好的理论指导意义。一般情况下，按照香农编码方法编出来的码，其平均码长不是最短的，即不是紧致码（最佳码）。只有当信源符号的概率分布使(2)中不等式左边的等号成立时，编码效率才能达到最高。

2.3 香农编码示例分析

有一离散信源 X，其概率分布如下：

$$\begin{bmatrix} X \\ p(X) \end{bmatrix} = \begin{Bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ 0.25 & 0.15 & 0.2 & 0.05 & 0.1 & 0.25 \end{Bmatrix}$$

则对该信源进行二进制香农码的编码过程如下表所示：

表 1 香农编码流程

信源符号x_i	符号概率$p(x_i)$	累加概率P_i	$-log\,p\,(x_i)$	码长l_i	码字
x_1	0.25	0	2	2	00
x_6	0.25	0.25	2	2	01
x_3	0.2	0.5	2.32	3	100
x_2	0.15	0.7	2.74	3	101
x_5	0.1	0.85	3.32	4	1101
x_4	0.05	0.95	4.32	5	11110

以$\mathrm{i}=2$为例，$-log_2 0.15 \le l_2 < -log_2 0.15 + 1$，即$2.74 \le l_2 < 3.74$，因此，$l_2 = 3$，累加概率$P_2$为 0.7，变成二进制数前三位为 0.101。编程时可采用如下方法得到二进制数：用排序后信源符号x_i的累加概率P_i乘以 2 后再次赋值给P_i，如果整数部分有进位，则P_i转化为二进制后的小数点后第一位为 1，否则为 0，将新的P_i的小数部分再次乘以 2，同样，如果整数部分有进位，则转化为二进制后的小数点后第二位为 1，否则为 0，从而得到小数点后的第二位，依此类推，直到得到了满足要求的位数，或者没有小数部分了为止，最终即可获得所有信源符号对应的码字。

同时，我们可以通过相应的公式计算得出如下信息：

信源信息熵：$\mathrm{H}(x) = -\sum_{i=1}^{6} p(x_i)\log_2 p(x_i) = 2.4232$ bits

平均码长：$\bar{L}_S = \sum_{i=1}^{6} p(x_i) l_i = 2.7$ bits

编码效率： $\eta = H(x)/\bar{L}_S = 0.8975$

三、 哈夫曼编码

3.1 哈夫曼编码简介

变字长编码的最佳编码定理：在变字长码中，对于概率大的信息符号编为短字长的码；对于概率小的信息符号编为长字长的码。如果码字长度严格按照符号概率的大小顺序排列，则平均码字长度一定小于以任何顺序排列方式得到的码字长度。

哈夫曼编码就是利用了这个定理，根据信源符号的概率分布，采用不等长编码。概率大的符号，使用短的码字编码；概率小的符号，使用长的码字编码。哈夫曼编码把信源符号按概率大小顺序排列，并设法按逆次序分配码字的长度。在分配码字的长度时，首先将出现概率最小的两个符号相加，合成一个概率；第二步把这个合成的概率看成是一个新组合符号的概率，重复上述做法，直到最后只剩下两个符号的概率为止。完成以上概率相加顺序排列后，再反过来逐步向前进行编码。每一步有两个分支，各赋予一个二进制码，可以对概率大的编为 0 码，概率小的编为 1 码，反之亦然。

经过哈夫曼编码之后，码字的平均码长$\bar{L}_H$为：

$$\bar{L}_H = \sum_{i=1}^{n} p(x_i) * l_i$$

其中，l_i表示信源符号x_i经过哈夫曼编码得到的码字的长度。

哈夫曼编码的效率为：

$$\eta = \frac{\text{信源熵}}{\text{平均码长}} = \frac{H(x)}{\bar{L}_H}$$

3.2 哈夫曼编码流程

哈夫曼编码的具体步骤归纳如下：

(1) 统计 n 个信源符号，得到 n 个不同概率的信息符号；

(2) 将这 n 个信源符号按其概率从小到大依次排序；

(3) 取两个概率最小的信息符号分别配以 1 和 0 两个码元，并将这两个概率相加作为一个新的信息符号概率，和未分配的信息符号构成新的信息符

号序列；

(4) 将剩余的信息符号，按概率从小到大重新进行排序；

(5) 重复步骤(3)，将排序后的最小的两个概率相加，相加和与其他概率再排序；

(6) 如此反复重复 n-2 次，最后只剩下两个概率值；

(7) 从最后一级开始，向前返回得到各个信源符号所对应的码元序列，即相应的码字，构成霍夫曼编码字，编码结束。

3.3 哈夫曼编码示例分析

设信源共有 7 个离散符号消息，其概率如下表所示:

表 2 离散信源分布概率

信源符号 x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7
符号概率 $p(x_i)$	0.15	0.19	0.10	0.17	0.01	0.18	0.20

对信源符号进行哈夫曼编码的过程如下图所示：

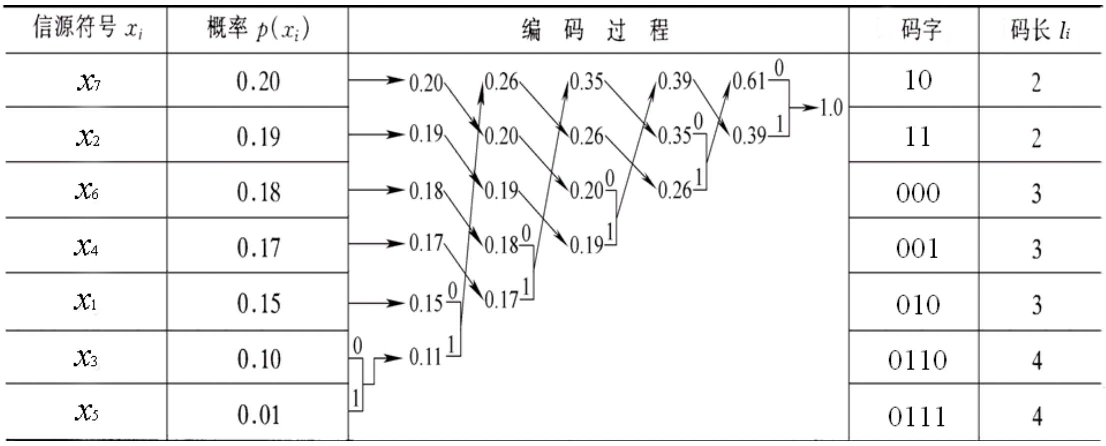

图 2 哈夫曼编码过程

信源符号的信息熵为：$\mathrm{H(x)} = -\sum_{i=1}^{7} p(x_i) \log_2 p(x_i) = 2.6087$ bits

该哈夫曼码的平均码长为：$\bar{L}_H = \sum_{i=1}^{7} p(x_i) l_i = 2.72$ bits

编码效率为：$\eta = H(x)/\bar{L}_H = 2.6087/2.72 = 0.95907$

四、 香农编码与哈夫曼编码的比较

哈夫曼编码的平均码长小于或等于香农编码的平均码长，它们的关系可以用下面的不等式表示：

$$\mathrm{H(X)} \le \bar{L}_H \le \bar{L}_S \le H(X) + 1$$

其中，H(X)表示离散信源符号的信息熵，$\bar{L}_H$表示使用哈夫曼编码得到的码字的平均码长，$\bar{L}_S$表示使用香农编码得到的码字的平均码长。

下面通过一个具体的实示例对二者进行比较比较，有一离散信源分布为$\mathrm{p}(x_i) = [0.36, 0.34, 0.25, 0.05]$。

可以求出，信源的信息熵为：$\mathrm{H(X)} = 1.78\ \ \mathrm{bits}$

香农编码：

$$-\log_2\left(p(x_i)\right) = [1.47, 1.56, 2.00, 4.32]$$

$$l_i = \lceil -\log_2(p(x_i)) \rceil = [2; 2; 2; 5]$$

$$\bar{L}_S = 2.15\ \ \mathrm{bits}$$

其中，运算符“$\lceil x \rceil$”表示对于x进行向上取整。

哈夫曼编码：

$$l_i = [1; 2; 3; 3]$$

$$\bar{L}_H = 1.94\ \ \mathrm{bits}$$

可以发现，哈夫曼码的平均码长要小于香农码的平均码长。