4.5 变长编码方法

变长编码采用非续长码；

- 力求平均码长最小，此时编码效率最高，信源的冗余得到最大程度的压缩；
- 对给定的信源，使平均码长达到最小的编码方法称为最佳编码，编出的码称为最佳码；
- 三种变长编码方法：霍夫曼编码、费诺编码以及香农编码；
- 霍夫曼编码是真正意义下的最佳编码。

4.5.1 霍夫曼编码

二进制霍夫曼编码过程如下：

(1)将信源符号按概率大小排序；

(2)对概率最小的两个符号求其概率之和，同时给两符号分别赋予码元“0”和“1”；

(3)将“概率之和”当作一个新符号的概率，与剩下符号的概率一起，形成一个缩减信源，再重复上述步骤，直到“概率之和”为1为止；

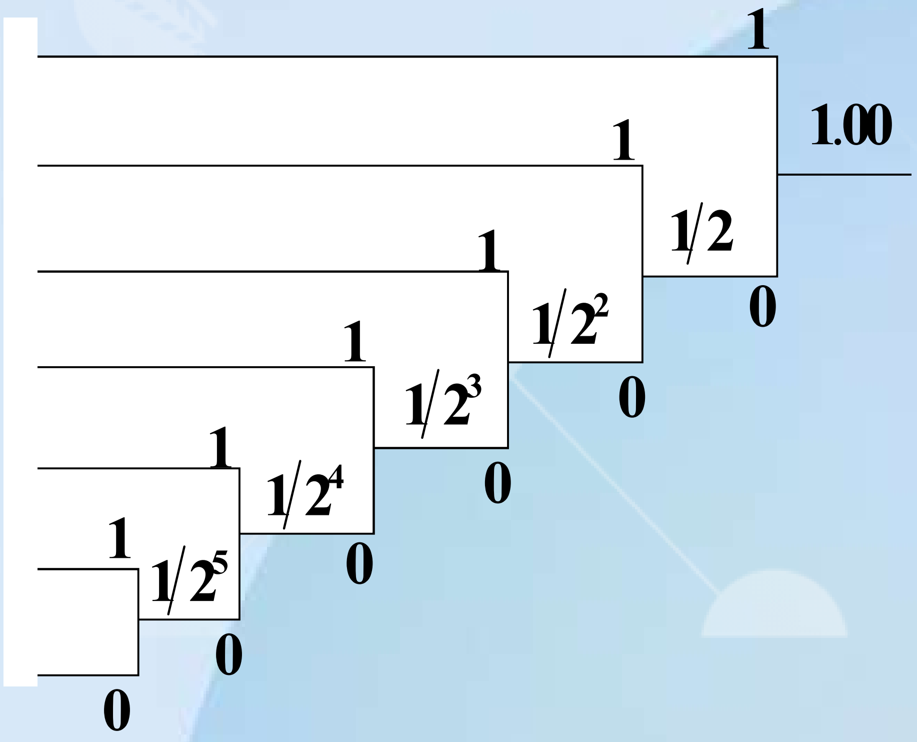
(4)按上述步骤实际上构造了一个码树，从树根到端点经过的树枝即为码字。

$$\begin{bmatrix} U \\ P_U \end{bmatrix} = \begin{bmatrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 \\ \frac{1}{2} & \frac{1}{2^2} & \frac{1}{2^3} & \frac{1}{2^4} & \frac{1}{2^5} & \frac{1}{2^6} & \frac{1}{2^6} \end{bmatrix}$$

2进制霍夫曼编码。

码元集：X={0, 1}

符号 u_i	概率 $P(u_i)$	码字 W_i	码长 l_i
u_1	$1/2$		
u_2	$1/2^2$		
u_3	$1/2^3$		
u_4	$1/2^4$		
u_5	$1/2^5$		
u_6	$1/2^6$		
u_7	$1/2^6$		



$$\bar{l} = \frac{1}{2}\times 1 + \frac{1}{2^2}\times 2 + \frac{1}{2^3}\times 3 + \frac{1}{2^4}\times 4 + \frac{1}{2^5}\times 5 + \frac{1}{2^6}\times 6 + \frac{1}{2^6}\times 6 = \frac{63}{32}$$

码元/符号

$$\eta_c = \frac{H(U)}{\bar{l}\log r} = \frac{\frac{63}{32}}{\frac{63}{32}\times\log 2} = 100\%$$

霍夫曼编码的基本特点

- 编出的码是非续长码：霍夫曼编码实际上构造了一个码树，码树从最上层的端点开始构造，直到树根结束，最后得到一个横放的码树，而且码字在终端节点上。
- 平均码长最小：霍夫曼编码采用概率匹配方法来决定各码字的码长，概率大的符号对应于短码，概率小的符号对应于长码。
- 码字不唯一：每次对概率最小的两个符号求概率之和形成缩减信源时，就构造出两个树枝，由于给两个树枝赋码元是任意的，码字不唯一。

定长编码与变长编码冗余压缩效果比较

$$\begin{bmatrix} U \\ P_U \end{bmatrix} = \begin{bmatrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 \\ \frac{1}{2} & \frac{1}{2^2} & \frac{1}{2^3} & \frac{1}{2^4} & \frac{1}{2^5} & \frac{1}{2^6} & \frac{1}{2^6} \end{bmatrix}$$

$$H(U) = 63/32 \text{ bit/符号}$$

$$\gamma = 1 - \frac{H(U)}{H_{\max}(U)} = 1 - \frac{63/32}{\log 7} \approx 0.3$$

定长编码：{001，010，011，100，101，110，111}

变长编码：{1，01，001，0001，00001，000001，000000}

定长编码	变长编码
$\bar{l} = l = 3$ 码元/符号	$\bar{l} = 63/32$ 码元/符号
$H(X) = \frac{H(U)}{\bar{l}} = \frac{63/32}{3} = 0.65625$ bit/码元	$H(X) = \frac{H(U)}{\bar{l}} = \frac{63/32}{63/32} = 1$ bit/码元
$\eta_c = \frac{H(X)}{H_{\max}(X)} = \frac{0.65625}{\log 2} = 65.625\%$	$\eta_c = \frac{H(X)}{H_{\max}(X)} = \frac{1}{\log 2} = 100\%$
$\gamma_c = 1 - \eta_c = 0.34375$	$\gamma_c = 1 - \eta_c = 0$

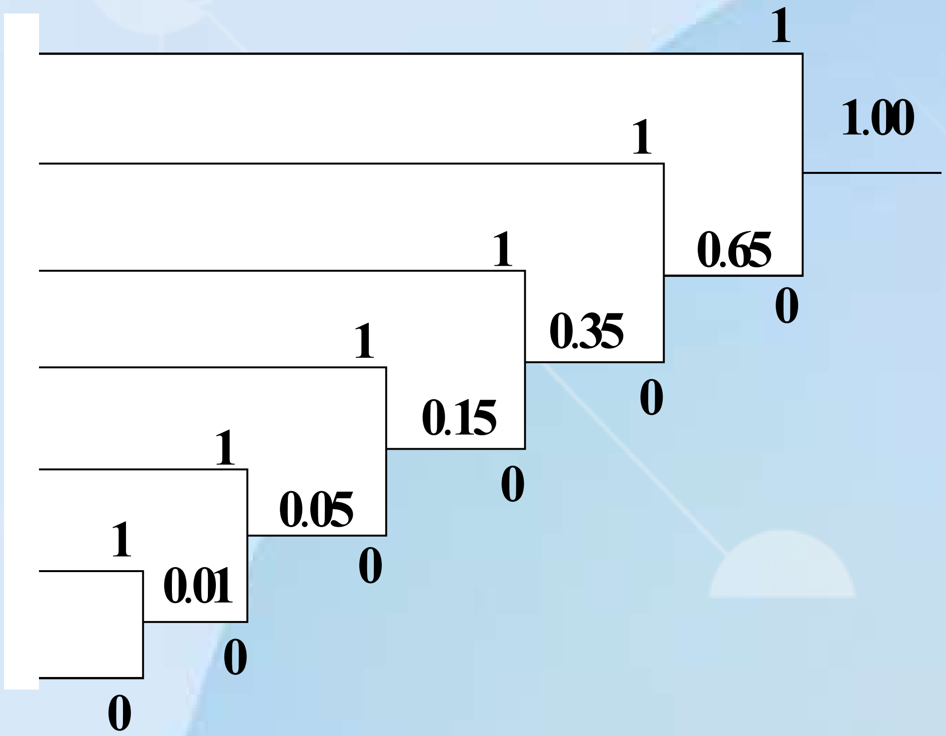
码子不唯一（1）

$$\begin{bmatrix} U \\ P_U \end{bmatrix} = \begin{bmatrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 \\ 0.35 & 0.30 & 0.20 & 0.10 & 0.04 & 0.005 & 0.005 \end{bmatrix}$$

2进制霍夫曼编码。

码元集：$X=\{0, 1\}$

符号 u_i	概率 $P(u_i)$	码字 W_i	码长 l_i
u_1	**0.35**		
u_2	**0.30**		
u_3	**0.20**		
u_4	**0.10**		
u_5	**0.04**		
u_6	**0.005**		
u_7	**0.005**		



$$\bar{l} = 0.35\times1 + 0.30\times2 + 0.20\times3 + 0.10\times4 + 0.04\times5 + 0.005\times6 + 0.005\times6$$
$$= 2.21 \text{ 码元/符号}$$

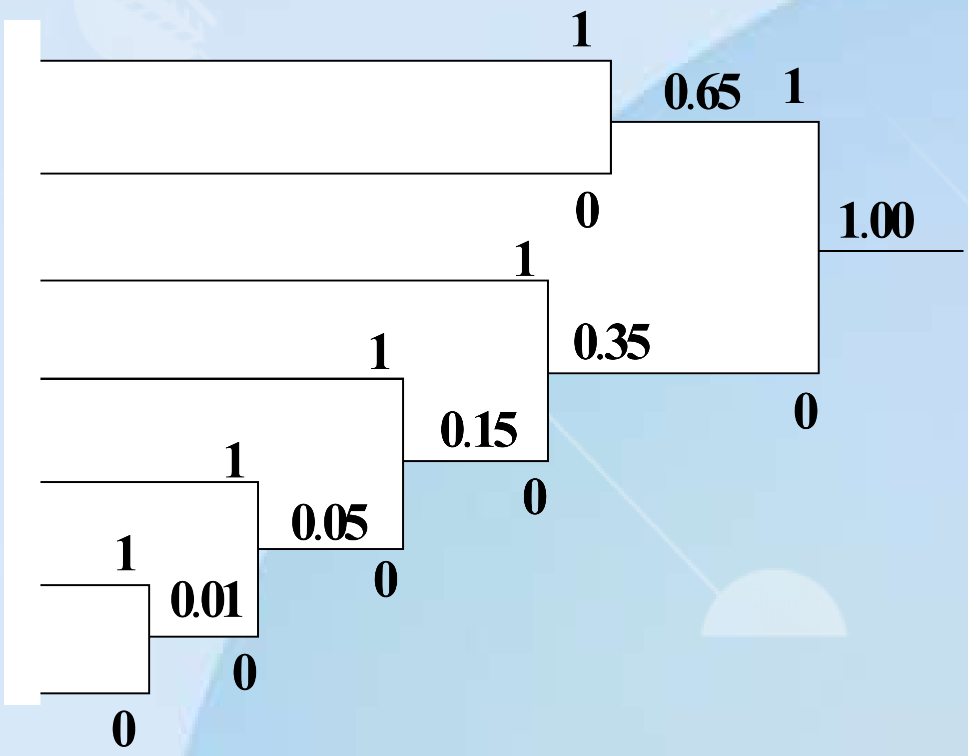
码子不唯一（2）

2进制霍夫曼编码。

码元集：$X=\{0, 1\}$

$$\begin{bmatrix} U \\ P_U \end{bmatrix} = \begin{bmatrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 \\ 0.35 & 0.30 & 0.20 & 0.10 & 0.04 & 0.005 & 0.005 \end{bmatrix}$$

符号 u_i	概率 $P(u_i)$	码字 W_i	码长 l_i
u_1	0.35		
u_2	0.30		
u_3	0.20		
u_4	0.10		
u_5	0.04		
u_6	0.005		
u_7	0.005		



$$\bar{l} = 0.35\times2+0.30\times2+0.20\times2+0.10\times3+0.04\times4+0.005\times5+0.005\times5$$
$$=2.21 \text{ 码元/符号}$$

码字 W_i	码长 l_i
1	1
01	2
001	3
0001	4
00001	5
000001	6
000000	6

码字 W_i	码长 l_i
11	2
10	2
01	2
001	3
0001	4
00001	5
00000	5

码字不同，码长也不同，但平均码长相同，因此编码效率相同。

码方差：

$$\sigma^2(l) = E[(l_i - \bar{l})^2] = \sum_{i=1}^{q} P(u_i)(l_i - \bar{l})^2$$

$$\sigma_1^2(l) = 1.4259$$

$$\sigma_2^2(l) = 0.3059$$