第4章 离散无记忆信源无失真编码

主要内容

1、基本概念

2、码的唯一可译性

3、定长编码定理和定长编码方法

4、变长编码定理

5 变长编码方法

6 几种实用的无失真信源编码

1、基本概念

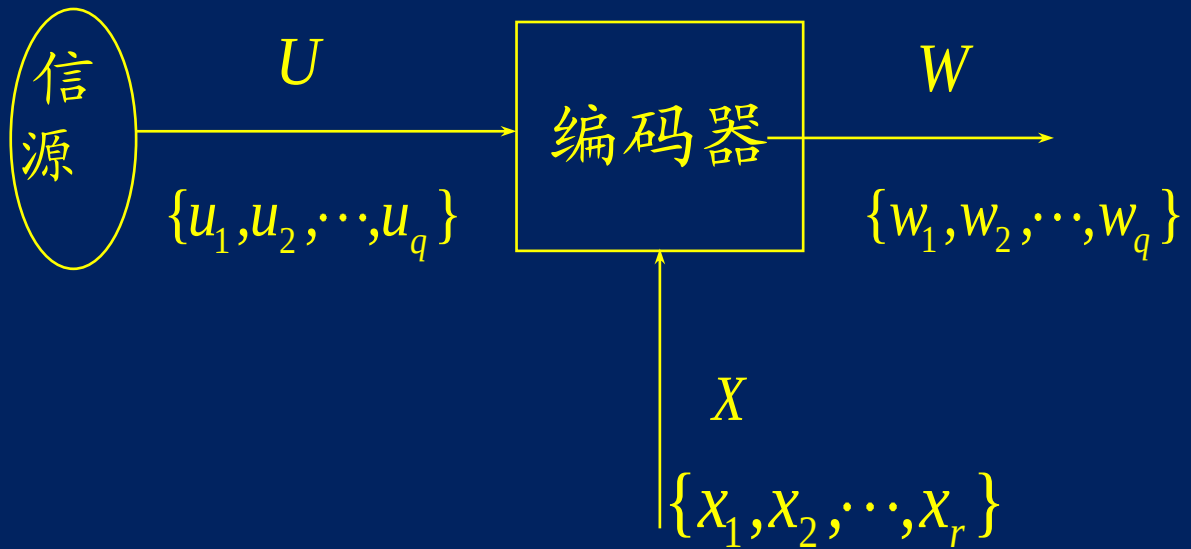
信源发出的消息序列通常不能直接送给信道传输，需要经过信源编码和信道编码。

信道编码的目的是降低差错率，提高传送的可靠性。

信源编码的目的是为了降低冗余度，提高通信的有效性。

编码是一种映射，是将输入符号映射成码字。无失真编码，映射一一对应，可逆。

编码器模型：



码长：码字所含码元的个数

定长编码：所有码字均有相同的码长，对应的码叫做定长码(*FLC*, Fixed Length code)；否则为变长编码。

平均码长：码中所有码字码长的统计平均，即 $\bar{l} = \sum_{i=1}^{q} P(w_i) l_i = \sum_{i=1}^{q} P(u_i) l_i$ 码元/符号

编码效率：编码后的实际信息率与编码后的最大信息率之比

$$\eta_c = \frac{R}{R_{\max}} = \frac{H(X)}{H_{\max}(X)} = \frac{H(U)/\bar{l}}{\log r} = \frac{H(U)}{\bar{l}\log r}$$

冗余度：

$$\gamma_c = 1 - \eta_c$$

2、码的唯一可译性

（1）基本概念

奇异码：一组码中含相同码字。

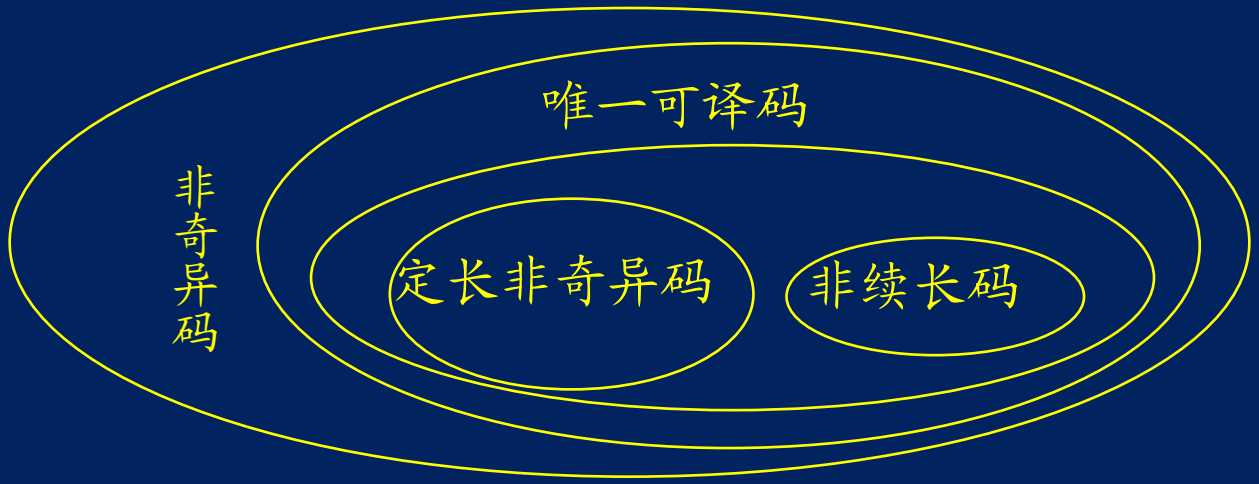
非奇异码：所有的码字都不相同。

唯一可译性：码字组成的任意有限长码字序列都能恢复成唯一的信源序列。

续长码：有些码字是在另一些码字后面添加码元得来的。

及时码：码字的最后一个码元出现时，译码器能立即判断一个码字已经结束，可以立即译码。

非续长码：任一码字都不是其它码字的延长。



5种不同的码

U	$P(u_i)$	W_1	W_2	W_3	W_4	W_5
u_1	½	00	00	1	0	0
u_2	¼	01	00	00	10	01
u_3	⅛	10	10	01	110	011
u_4	⅛	11	11	10	111	111

（2）码树和*Kraft*不等式

从树根开始，生长r个树枝，在节点处再各自生长r个树枝。

节点：树枝与树枝的交点。

1阶节点：经过1根树枝到达的节点。

整树：节点长出的树枝数等于r

定理：对于任一r进制非续长码，各码字的码长必须满足***Kraft***不等式：$\sum_{i=1}^{q} r^{-l_i} \le 1$

反过来，若上式成立，就一定能构造一个r进制非续长码。

定理 对于任一进制唯一可译码（*UDC*），各码字的码长必须满足***Kraft***不等式：

$$\sum_{i=1}^{q} r^{-l_i} \leq 1$$

反过来，若上式成立，就一定能构造一个进制唯一可译码（*UDC*）。

3、定长编码定理和定长编码方法

对信源U的N次扩展信源U^N进行r进制编码

无失真条件：

$$r^{l_N} \geq q^N$$

整理得

$$\frac{l_N}{N} \geq \frac{\log q}{\log r} = \frac{H_{\max}(U)}{\log r} = H_{r\max}(U)$$

定长无失真编码定理

用r元符号表对离散无记忆信源U的N长符号序列进行定长编码，N长符号序列对应的码长为l_N，若对于任意小的正数ε，有不等式

$$\frac{l_N}{N} \geq \frac{H(U)+\varepsilon}{\log r}$$

就几乎能做到无失真编码，且随着序列长度N的增大，译码差错率趋于0。反过来，若

$$\frac{l_N}{N} \leq \frac{H(U)-2\varepsilon}{\log r}$$

就不可能做到无失真编码，且随着N的增大，译码差错率趋于1。

4、变长编码定理

变长编码可以减少平均码长。

无失真变长编码定理（香农第一定理）

对无记忆信源U的N次扩展信源进行变长编码，平均码长为 $\overline{l}_N$，总可以找到一种无失真编码方法，构成惟一可译码，使其平均码长满足

$$H_r(U) \le \frac{\overline{l}_N}{N} < H_r(U) + \frac{1}{N}$$

证明 设信源U的N次扩展信源 U^N 的信源空间为 $[U^N, P_{U^N}] = [\bar{u}_j, P(\bar{u}_j) \mid j = 1,2,\cdots,q^N]$

（1）先证明平均码长的下界，这只须证明 $H(U^N) - \bar{l}_N \log r \le 0$ 即可。

$$H(U^N) - \bar{l}_N \log r = \sum_{j=1}^{q^N} P(\bar{u}_j) \log \frac{1}{P(\bar{u}_j)} - \sum_{j=1}^{q^N} P(\bar{u}_j) l_j \log r$$

$$= \frac{1}{\ln 2} \sum_{j=1}^{q^N} P(\bar{u}_j) \ln \frac{r^{-l_j}}{P(\bar{u}_j)}$$

$$\le \frac{1}{\ln 2} \sum_{j=1}^{q^N} P(\bar{u}_j) \left(\frac{r^{-l_j}}{P(\bar{u}_j)} - 1 \right) = \frac{1}{\ln 2} \left[\sum_{j=1}^{q^N} r^{-l_j} - \sum_{j=1}^{q^N} P(\bar{u}_j) \right]$$

$$\le \frac{1-1}{\ln 2} = 0$$

（2）上界证明

令 $\alpha_j = -\log_r P(\bar{u}_j)$

若α_j是整数，选 $l_j = \alpha_j$

若不是整数，选 $l_j = (\alpha_j)$，表示不小于α_j的整数，$\alpha_j \le l_j < \alpha_j + 1$

即 $\log_r \frac{1}{P(\bar{u}_j)} \le l_j < \log_r \frac{1}{P(\bar{u}_j)} + 1 \ , j = 1, 2, \cdots q^N$

整理左边不等式得 $r^{-l_j} \le P(\bar{u}_j) \ , j = 1, 2, \cdots q^N$

$$\sum_{j=1}^{q^N} r^{-l_j} \le \sum_{j=1}^{q^N} P(\bar{u}_j) = 1$$

根据右边不等式得

$$\bar{l}_N = \sum_{j=1}^{q^N} P(\bar{u}_j) l_j < \sum_{j=1}^{q^N} P(\bar{u}_j) \left(\log_r \frac{1}{P(\bar{u}_j)} + 1 \right)$$

$$= H_r(U^N) + 1$$

$$= N H_r(U) + 1$$

$$\frac{\bar{l}_N}{N} < H_r(U) + \frac{1}{N}$$

5 变长编码方法

（1） 霍夫曼编码

二进制霍夫曼编码

A、将q个信源符号按概率递减次序排列;

B、用0、1符号分别代表概率最小的两个信源符号，将这两个信源合并成一个符号，得到包含q-1个符号的新信源，称缩减信源;

C、继续上述两步，直至信源只剩两个符号为止，将最后两个信源符号分别用0和1示。

D、倒推得码字。

r进制霍夫曼编码

与二进制霍夫曼编码类似，但每次缩减信源时，将r个最小概率符号合并。为了保证缩减到最后正好剩下r个符号，可给信源添加几个零概率符号。

信源符号数q应满足

$$q = (r-1)\theta + r$$

θ 为信源缩减的次数。

序列编码的效率更高。

(2) 费诺编码

费诺编码的步骤如下：

A、将信源符号按概率从大到小的排序；

B、将信源符号分成2组，使2组信源符号的概率之和近似相等，并给2组信源符号分别赋码元“0”和“1”；

C、接下来再把各小组的信源符号细分为2组并赋码元，方法与第一次分组相同，直至每组只含一个信源符号。

D、由此即可构造一个码树，所有终端节点上的码字组成费诺码。

（3）香农编码

编码步骤

A、排序：将信源符号按概率从大到小排列；

B、计算码长：$-\log P(u_i) \le l_i < -\log P(u_i)+1$

C、计算累加概率：

$$\begin{cases} P_1 = 0 \\ P_i = \sum_{k=1}^{i-1} P(u_k) \quad i = 2,3,\cdots,q \end{cases}$$

将累加概率转换成二进制数；

D、决定码字：取累加概率二进制数小数点后l_i个二进制数字作为第i个信源符号的码字。

6 几种实用的无失真信源编码

(1) 游程编码

基本原理

游程编码主要用于黑、白二值文件的传真。表示背景（白色）时像素为码元“0”，表示内容（黑字）时像素为码元“1”。

信源的符号中重复出现的连“0”或连“1”像素序列经游程编码表示后，可表示为统一的编码单元结构：

符号码	标识码	游程长度

MH码

MH码是CCITT提出的文件、传真类一维数据压缩编码的国际标准，是由游程编码及霍夫曼编码集合而成的一种改进型霍夫曼码。

MH码使用固定编码表进行编码，即在信源与信宿两端，利用预先确定的编码表各自独立进行编码和解码。

由于采用固定的编码表，对不同的信源，编码效率各不相同。

编码规则如下：

①每页文件以同步码EOL（000000000001）开始，以6个EOL结束。

②每行必须以白游程开始，以同步码EOL结束，每行游程总和为1728个像素

③游程长度在0-63之间时，码字直接由相应的终止码表示；

②游程长度在64-1728之间时，码字由一个组合码加上一个终止码构成。

（2）算术编码

算术编码是香农编码方法与累积概率分布函数的递推算法的结合，是香农编码的思想在信源序列上的应用。

编码过程：将信源符号序列依累积概率分布函数的大小映射到[0,1)区间，每个符号序列均有一个唯一的小区间与之对应，因而可在小区间内取点来代表该符号序列。将此点的累积概率分布函数值用二进制数表示，取小数点后的前 位，即l_i是信源符号序列的算术码。

(3) 基于字典的编码

LZ编码的基本原理

LZ码是由两位以色列研究人员共同提出的一种基于字典的编码方法。

LZ编码的基本思路与查字典极为相似，即在组成并拥有词典的情况下，通过“单词”的位置信息，间接的表达“单词”的内容。因而传递“单词”的位置信息就是对“单词”内容的传递。

LZW码

LZW算法是LZ算法的一种修正。为了使长短不一的“单词”更便于处理，专门为“单词”建立了一种通用的格式。其格式规定如下:

A、每个“单词”均由前缀字符串和尾字符串两部分组成。

B、前缀字符串为字典中已有的“单词”，尾字符是本“单词”的最后一个字符。

C、对本身已经是单字节的“单词”，没有前缀词时则在前面加上一个空前缀，并规定字典最后一个“单词”为“空”。