欢迎各位同学拷贝！

- 此课件挂在mail.google.com网站上

用户名：jpkctxyl@gmail.com

密　码：jpkctxyl123456

以饱满的热情
优良的学风
明显的进步

迎接教育部对我校进行本科教学评估

第2章 无失真信源编码

- 信源编码概述
- 赫夫曼编码
- 游程编码
- 算术编码
- 通用编码

教学目的与要求

1. 深刻理解信源编码原理，明白为什么通过编码能压缩代码长度。

2. 学习信源编码基本概念，了解Shannon定长码与变长码编码定理的内容和意义。

3. 熟练掌握Huffman编码方法。（重点）

4. 掌握游程编码、算术编码（难点）和字典编码原理。

参考文献（见课本182页）

1.周炯磐：**信源编码原理**

人民邮电出版社（1996年10月第一版）

2. 吴乐南：**数据压缩**

电子工业出版社（2001年6月第一版）

3.吴伟陵：**信息处理与编码**

人民邮电出版社（1999年7月第一版）

4.曹雪虹：**信息论与编码**

北京邮电大学出版社 (2001年8月第一版)

第2章 无失真信源编码

2.1 信源编码的目的、原理和方法概述

（第3讲 2007.9.11.）

- 计划学时：2学时
- 要求掌握的主要内容：

  1. 深刻理解信源编码原理和意义。
  2. 熟练掌握编码有关概念：等长码、变长码、唯一可译性、码树、平均码长等。
  3. Shannon编码定理----概率匹配原则。
- 重点难点：

  重点----信源编码原理

  难点----Shannon编码定理

- 外语关键词：

信源编码： *Source Coding*　　码树：*Code Tree*

变长码编码： *Variable Length Coding*

唯一可译性： *Uniquely Decodable*

即时性：*Instantaneously Decodable*

平均码长： *Average code Length*

克拉夫特不等式： *Kraft Inequality*

概率匹配原则： *Principle of Matching with Probability*

香农定理： *Shannon's Theorems*

[温旧引新]

- 等概信源具有最大熵$H_0 = \log m$
- 不等概信源单符号信息熵$H(X)=H_1 < H_0$
- 有记忆信源信息熵随着序列的增长而变小：$H_0 \geqslant H_1 \geqslant H_2 \geqslant H_3 \geqslant \cdots\cdots \geqslant H_N \geqslant H_\infty$
- H_∞是极限熵，代表实际信源的信息熵。
- 这些理论是信源编码的基础。

2.1.1 编码 (coding)

1． 编码的意义：

- 广义地说，编码是用符号(或数字)表达信息的一种方案，是表达信息的符号组合方式。
- 更确切地说编码是不同表达形式信息代码之间的变换。
- 现代通信与计算机技术中，为了更加高效、可靠、安全地对信息进行传输、存储与利用，经常需要把信息符号通过设定的数学关系，用另一套代码来替换原来的代码，因而出现了各种类型的编码。

- 用ASCII码表达字符，用GB2312区位码表达汉字，只是完成了符号文字的数字化，为传输与存储的方便，还需要做进一步处理。

提问：通信技术中有哪三大类编码？

- 进行信源编码：将代码变得更简练（压缩掉代码中的冗余)。
- 进行信道编码：使代码变得更可靠（具有检错与纠错功能)。
- 进行加密编码：使代码变得更安全（具有保密与认证功能)。

2. 编码的基本要求

信息的可恢复性：不论代码形式如何变换，接收端最终应能正确地译出原消息。

提问：编码为什么能传承信息？

信息的可传递性：编码过程虽然改变了“载体”，但由于新旧两套代码之间存在着对应关系，原代码的概率特征被继承了下来，新代码就具有了同样的不确定度，于是信息得以传承。

2.1.2 编码术语

1．约定：

编码是按一定数学规则对信源符号序列进行的一种变换。为了表述方便，设：

- 信源发信的符号集 $A=[a_1，a_2，\cdots\cdots，a_m]$
- 编码符号集 $X=[x_1，x_2，\cdots\cdots，x_r]$
- 信源消息的分组 $S_i=(s_1^i,s_2^i,\cdots\cdots,s_N^i)$；$s_n^i\in A$
- 编码符号的分组 $W_j=(w_1^j,w_2^j,\cdots\cdots,w_L^j)$；$w_l^j\in X$
- 编码对应关系：$S_i \longleftrightarrow W_j$

2. 术语：

（1）码字：变换后的各个新符号串 W_j 被称为码字。

（2）码长：码字 W_j 的长度(符号数) L_j 被称为码长。

（3）码元：组成码字 W_j 的各位代码符号 x_j

（ $j = 1, 2, \ldots\ldots, r$ ），称为码元。

（4）码：所有码字的集合称为“码”。

（5）编码：全部 $S_i \leftrightarrow W_i$ 的映射关系称之为编

3．等长码与变长码 (Fixed-Length Codes and Variable-length Codes)

（1）等长码：

编码中要求所有码字长度都相同，这样的编码叫等长码。

（2）变长码：

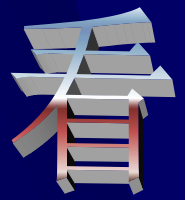
编码并不要求所有码字长度都相同，这样的编码称为不等长码，或曰变长码。

2.1.3 信源编码

1. 信源编码的目的

为压缩代码长度而对信源消息进行的编码。

思考：信源消息为什么能够被压缩？



演示，

问题

2. 信源的相对信息率和冗余度

提问: 实际信源发出消息，为什么会有冗余?

信源每个符号最大可以荷载的信息量是 $H_0=\log m$，但是只有等概信源的信息熵才可达到这个最大值。

实际信源由于非等概和有记忆两个原因，其信息熵H_∞远小于H_0，表明平均每个符号的实际信息荷载量都没有达到最大，存在着信息含量不饱满的现象。

以英语为例，26个字母加上空格，共27个符号，H_0 =log27=4. 76（比特/符号）。然而各个字母在文章中出现概率是不同的，百分比见下表所示：

a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*
6.42	1.27	2.18	3.17	10.31	2.08	1.52	4.67	5.75
j	*k*	*l*	*m*	*n*	*o*	*p*	*q*	*r*
0.08	0.49	0.32	1.98	5.74	6.32	1.52	0.08	4.84
s	*t*	*u*	*v*	*w*	*x*	*y*	*z*	空格
5.14	7.96	2.28	0.83	1.75	0.13	1.64	0.05	18.59

由表不难算出：H_1= $H(X)$= 4.03（比特/符号）。

- 在一阶和二阶马尔科夫信源近似下，有人曾求得：H_2= 3.32(比特/符号)，H_3=3.10(比特/符号)。
- 如果再计及词法、句法、语法、修辞等的制约，计入更高阶的关联，极限熵被估计为H_∞=1. 4（比特/符号）。
- H_0代表平均每个英文符号最多所能承载的信息量。H_∞代表英语文章中平均每个信源符号实际所荷载的信息量。拥有27个符号的英文信源，平均每个符号是有能力荷载4.76比特信息的，但实际上它们每符号却只承载着 1.4比特的信息。

这种信息率不饱满的情况在各种信源中普遍存在，因此定义：

•相对信息率： $\mu = H_\infty / H_0$

•信源冗余度（或剩余度）： $\gamma = 1 - \mu$

来反映信源实际所含信息的饱满程度。对于英文， $\mu = 0.29$， $\gamma = 0.71$，冗余是比较大的。

- 汉语也有类似的情况：

对**10000**个通用汉字统计结果如下：

最常见的只有**140**个，出现概率**50%**；

较常见的**485**个，出现概率**35%**；

一般性的**1775**个，出现概率**14.7%**；

不常见的**7600**个，出现概率**0.3%**。

- 由此算出：$H_0 = log 10000 = 13.29$（*bit* /汉字）

$H_1 = 10.25$（*bit* /汉字）

仅算到H_1：　μ **=0.77**，　γ **=0.23**，

3. 信源编码原理

❖任务：实际信源发出的符号序列，一般总含有一定的冗余。怎样将这些冗余压缩掉？

❖办法：寻找一种更短的代码序列，在不损失信息的前提下，替代原来的符号序列。

❖思路：应当尽量使所找的编码序列各个码元相互独立且等概，就会使单位符号信息含量更多，代码就比原来更短。

- 设有一个1000个英文字符的文件
- 用ASCII码表示，需要8000bit ≈8*Kb*
- 如果只需要表达32个符号(字母与常用标点),用5bit自然码即可满足，则总代码长度为5 *Kb*
- 考虑到各个字母不等概出现而采用下面的编码，则总代码长度为4117 *bit* ≈4*Kb*
- 如果找到了一种最佳编码，将原文所蕴涵的1400bit信息用1400个二元符号表达，则总代码长度只有1.4 *Kb*

仅仅根据英文字母不等概而进行的一个编码例子：

符号	概率	编码	符号	概率	编码
其它	0.1639	000	c	0.0179	001111
空格	0.1524	010	f	0.0170	011100
e	0.0845	0010	m	0.0162	011101
t	0.0652	0110	w	0.0143	011110
r	0.0562	1000	y	0.0134	110000
a	0.0526	1001	g	0.0125	0111111
o	0.0518	1011	p	0.0125	1100010
i	0.0471	1101	b	0.0104	1100011
n	0.0470	1110	v	0.0068	1100100
s	0.0421	1111	k	0.0040	1100101
h	0.0383	00110	x	0.0011	1100110
l	0.0263	10100	j	0.0007	1100111
d	0.0260	10101	q	0.0007	01111100
u	0.0187	001110	z	0.0004	01111101

平均码长是:4.117

4. 信源编码的分类

- **无失真信源编码和限失真信源编码**

  根据能否无失真地恢复信源消息来区分。

- **离散信源编码和连续信源编码**

  根据信源发出消息是否连续来区分。

- **无记忆信源编码和相关信源编码**

  根据信源是否有记忆来区分。

- **分组编码和序列流编码：**

  根据编码结构是否进行分组来区分。

2.1.4 等长码信源编码

1. 等长码信源编码原理：

- 在固定长度的各种信源符号串与长度不变的各个码字之间建立一一对应关系。

❖起码要求：唯一可译性

- 长度为N的信源符号串共有m^N种。
- 长度为L的码字符号串共有r^L种。
- 只要码长L足够大，使$r^L \geqslant m^N$，则每个信源符号串都可以找到一一对应的码字。
- 这时，$L\cdot\log r \geqslant N\cdot\log m$，即：$l=\frac{L}{N}\geq\frac{H_0}{\log r}$

举例：以m=27，r=2 为例来讨论：

- N=1，起码要求：要对长度为1的单符号（27 个）信源符号分组编码，
  - $r^L \geqslant m^N = 27$
  - $\Rightarrow L \geqslant N \cdot \log m / \log r = \log 27$
  - $\Rightarrow L \geqslant 4.76 \approx 5$;
- N=10，起码要求：要对长度为10的全部（$27^{10}=2\times10^{14}$ 个）信源符号分组作编码，
  - $r^L \geqslant m^N = 27^{10} = 2\times10^{14}$
  - $\Rightarrow L \geqslant N \cdot \log m / \log r$
  - $\Rightarrow L \geqslant 10^{\log 27} \approx 48$;
  - 为了都能找到对应码字，至少 L=48，才有足够多（$2^{48}=2.8\times10^{14}$）的码字。

❖最佳要求：

- 仅满足唯一可译未必能使码长最短。
- m^N 种不同的排列的信源符号串中，大多数可能都是杂乱无章的符号堆积，香农把它们称为非典型序列。
- 能出现在实际信源消息中有语义的文字序列只占很少数，香农把它们称为典型序列。
- 如果只对典型序列进行编码，则需要的码字数量就少得多。r不变的情况下，L就可以取得较小，也就是说代码可以变得更短。

- A=[父，母，老，师]
- 序列长为2，N＝2的符号序列S_i是：
- S_i ＝{父母，老师，师父，师母，老父(?)，老母(?)，父老，母老，父父，母母，老老，师师，父师，母师，母父，师老}

m^N种的信源符号串中，大多数可能都是杂乱无章的符号堆积，香农把它们称为非典型序列。

能出现在实际信源消息中有语义的文字序列只占很少数，香农把它们称为典型序列。

- 以N=10，m=27，r=2 为例来讨论：
- 起码要求：要对长度为10的全部（$27^{10}=2\times10^{14}$个）信源符号分组作编码，为了都能找到对应码字，至少L=48，才有足够多（$2^{48}=2.8\times10^{14}$）的码字。
- 最佳要求：假设英文中实际可能出现的长度为10的字串数目不到3万个，只对这些典型序列作编码，则只要取L=15， 就有2^{15}=32768个码字，足够编码使用。
- 压缩比：48/15=3.2。

- 牛津辞典第二版收录了615,000 个词条，the, be, to, of, and, a, in, that, have和I这十个词条在牛津全集一百万个词汇中出现频率是25%。
- 类似的，100个最常用词条占出现频率的50%。1000个常用词条占了75%，而7000个词条却占了牛津全集的90%，5万个词条则占了95%。
- 在100个最常用的词条中，长度为1的仅2条，占2/26 =0.077；长度为2的有22条，占$22/26^2$=0.033；长度为3的有28个，占$28/26^3$=0.0016；长度为4的有35个，占$35/26^4$=0.00077。

- 随着序列长度N的增大，虽然典型序列的数目也会增加，但是它们在序列总数中所占比例却迅速地减少，压缩效果会越来越好。
- 当$N\to\infty$时，长为N的符号串平均具有NH_∞的信息量。
- 最佳的编码应使码元符号独立且等概出现，这时平均每个码字的信息荷载量可达到最大值$L\cdot\log r$。
- 从信息传承的角度讲，应要求$L\cdot\log r\geqslant NH_\infty$，即要求平均每信源符号对应编码的最小长度为：

$$l=\frac{L}{N}\geq\frac{H_\infty}{\log r}$$

- 从起码要求的：$L_1 \cdot \log r \geqslant N \cdot H_0$
- 到最佳要求的：$L_2 \cdot \log r \geqslant N \cdot H_\infty$
- 因为$H_\infty < H_0$，所以码长得到了压缩。
- 如果真达到了$L_2 \cdot \log r = N \cdot H_\infty$，代码中便没有冗余。
- L_1 / L_2是码长压缩比，H_0 / H_∞是相对信息率，
- 由$L_1 / L_2 = H_0 / H_\infty$知，二者相等，确实达到了最佳。

2. Shannon等长码编码定理：

一个信息熵为$H(S)$的离散无记忆信源，若对长度为N的信源序列进行等长码编码，码字从r个码元的符号集中选取L个码元构成。对任意 $\varepsilon>0$，只要满足：

$$L/N=(H_N+\varepsilon)/\log r$$

则当N足够大时，几乎可实现无失真编码，使错误任意小。反之，若

$$L/N=(H_N-2\varepsilon)/\log r$$

则不能实现无失真编码。

定理告诉我们三个结果：

（1）定长码在理论上是能够进行无失真信源编码的；

（2）最短的极限码长是 $l_0=H_N/\log r$；

（3）条件是实际码长 $l=L/N>l_0$，并且信源序列分组N必须足够大。

然而，定理并没有给出具体的编码方案，所以说**Shannon**定理是一个存在性定理。

定理的证明：（可略）

（1）契贝谢夫不等式：

$$P\{\,|I(S)-E[I(S)]|\geq N\varepsilon\,\}\leq\delta=\frac{D[I(S)]}{(N\varepsilon)^2}$$

- 式中$I(S)$是长为N的信源序列S的自信息；
- 随机变量$I(S)$的统计平均值$E[I(S)]=H(S)=N\cdot H_N$；
- $D[I(S)]$是方差，ε是任意小正数；
- 契贝谢夫不等式告诉我们，统计规律的实质在于：当随机样本数目较大时，随机变量的取值偏离它的统计平均值较大的概率是很小的。

（2）典型序列与非典型序列：

❖香农把自信息取值处于$H(S)\mp N\varepsilon$范围内的那些序列叫做典型序列。典型序列的集合记作G；

❖把自信息取值处于$H(S)\mp N\varepsilon$范围外的那些序列叫做非典型序列。

❖契贝谢夫不等式告诉我们，非典型序列出现的概率是很小的，同时也就等于告诉我们典型序列出现的概率是很大的，所以典型序列集合又叫高概率集。

（3）典型序列的渐进等概性质：

典型序列满足：$|I(S)-H(S)|<N\varepsilon$

即：$-N\varepsilon<[I(S)-N\cdot H_N]<N\varepsilon$

或：$N\cdot H_N-N\varepsilon<I(S)<N\cdot H_N+N\varepsilon$

$N\cdot H_N-N\varepsilon<-\log p(S)<N\cdot H_N+N\varepsilon$

可见：$2^{-M[H_N-\varepsilon]}>p(S)>2^{-M[H_N+\varepsilon]}$

表明高概率集的所有典型序列其概率几乎都相等，挤在 $2^{-H(S)\mp N\varepsilon}$ 的区间。

（4）典型序列的数目：

设典型序列数目为M，则典型序列集G出现的概率应满足：$M \cdot 2^{-N(H_N+\varepsilon)} < P(G) < 1$

由此便知：$M < 2^{N(H_N+\varepsilon)}$

M占总序列数的比率：

$$\xi = \frac{M}{m^N} = \frac{2^{N(H_N+\varepsilon)}}{2^{\log m^N}} = 2^{-N(\log m - H_N - \varepsilon)}$$

一般总有$\log m = H_0 > H_N$，使$[\log m - H_N - \varepsilon]$为正数，

故当$N \to \infty$时：$2^{-N[\log m - H_N - \varepsilon]} \to 0$

表明典型序列虽然经常出现，但占总序列数目的比率却很小，并且随着序列变长，比率越来越小。

（5）定长码编码定理的得出：

- 若码长L取得足够大，使：$r^{L} \geqslant 2^{N(H_N+\varepsilon)} > M$，

则所有的典型序列都有唯一的码字可对应。

此时 $L\cdot\log r \geqslant N(H_N+\varepsilon)$；即：

$$l=\frac{L}{N}\geq\frac{H_N+\varepsilon}{\log r}$$

- 反之，若码长L取得不够大，使 $r^{L} \leqslant 2^{N(H_N-2\varepsilon)}$

$$l=\frac{L}{N}\leq\frac{H_N-2\varepsilon}{\log r}$$

即 $L\cdot\log r \leqslant N(H_N-2\varepsilon)$；或：

这时就会有一部分典型序列没有码字可对应。

- 设有码字可对应的典型序列集合为g，它们应当是典型序列G的一部分。g出现的概率为：

$$P(g) \leqslant r^{L} \cdot \max[P(s)] \leqslant 2^{N(H_N - 2\varepsilon)} \cdot 2^{-N(H_N - \varepsilon)} = 2^{-N\varepsilon}$$

- 则当$N \to \infty$时：　$P(g) \to 0$
- 于是，无码字对应的典型序列的概率$P(G-g) \to 1$
- 表明这种情况下，错误必然存在，不可能是无失真编码。
- 定理得到证明。

3. 等长码可行性讨论

- 这个令人鼓舞的理论结果可行性如何呢？
- 香农定理给出任意小的错误概率P_E满足：

$$P_E \le \delta = \frac{D[I(\mathbf{x})]}{N\varepsilon^2}$$

- 这里任意小量 ε 来自实际编码码长 $l=(H_N+\varepsilon)/\log r$ 与理论极限码长 $l_0=H_N/\log r$的接近程度；
- 这里$D[I(X)]=E[I^2(X)]-\{E[I(X)]\}^2$是自信息$I(X)$的方差；
- 若指定了所要求的错误概率和编码效率 $\eta=l_0/l$，由以上关系式能够求出相应的N值。

- 不妨以一个实例来讨论：

设：无记忆二元信源概率为：p={3/4,1/4}，

可求出$H_N=H(X)$= (3/4)log(4/3)+(1/4)log4=0.811，

$D[I(X)]$=3/4(log4/3)^2+1/4(log4)^2−0. 811^2 =0.4715，

若要求 $P_E \leqslant \delta=10^{-5}$且 $\eta=l_0/l=0.96$， 则：

$$\varepsilon=\frac{1-\eta}{\eta}H_N=0.0338;\quad N\leq\frac{D[I(X)]}{\varepsilon^2\delta}=4.13\times10^7$$

- 此例表明，定长码定理所讲的“只要N足够大”，真是太大了，将信源序列以这样大的分组来编码，显然是无法办到的。

- 结论：定长码编码不可行。

2.1.4 变长码编码原理

1. 唯一可译性：

- 首先要解决“断字”问题. 如何将连在一起的编码按原来的分组拆分开呢？
- 例如，当信宿收到的码流为10101011时，按照表中给出的三种编码各应如何翻译呢？

	A	B	C	D	翻译结果
码 I	0	10	11	101	BBBC，DABC，BDAC
码 II	1	10	100	1000	BBBAA
码III	1	01	001	0001	ABBBA

- 按码Ⅰ编码，译为“BBBC”，“DABC”还是“BDAC”都对！译码失去唯一性。
- 按码Ⅱ编码，唯一地只能译为BBBAA。码Ⅱ的特点是1打头，凡是见到1就是新码字头。
- 按码Ⅲ编码，唯一地只能译为ABBBA。码Ⅲ的特点是1结尾，凡是见到1就是码字尾。
- 码Ⅱ与码Ⅲ都是唯一可译码。
- 码Ⅰ不能唯一可译。

2. 即时性：

❖码III在码字一结束就能进行翻译，而码 II 必须等收到下一个码字开头时才能进行翻译。因此码III是即时码，码 II 是非即时码。

❖即时码在结构上有两个典型特征：

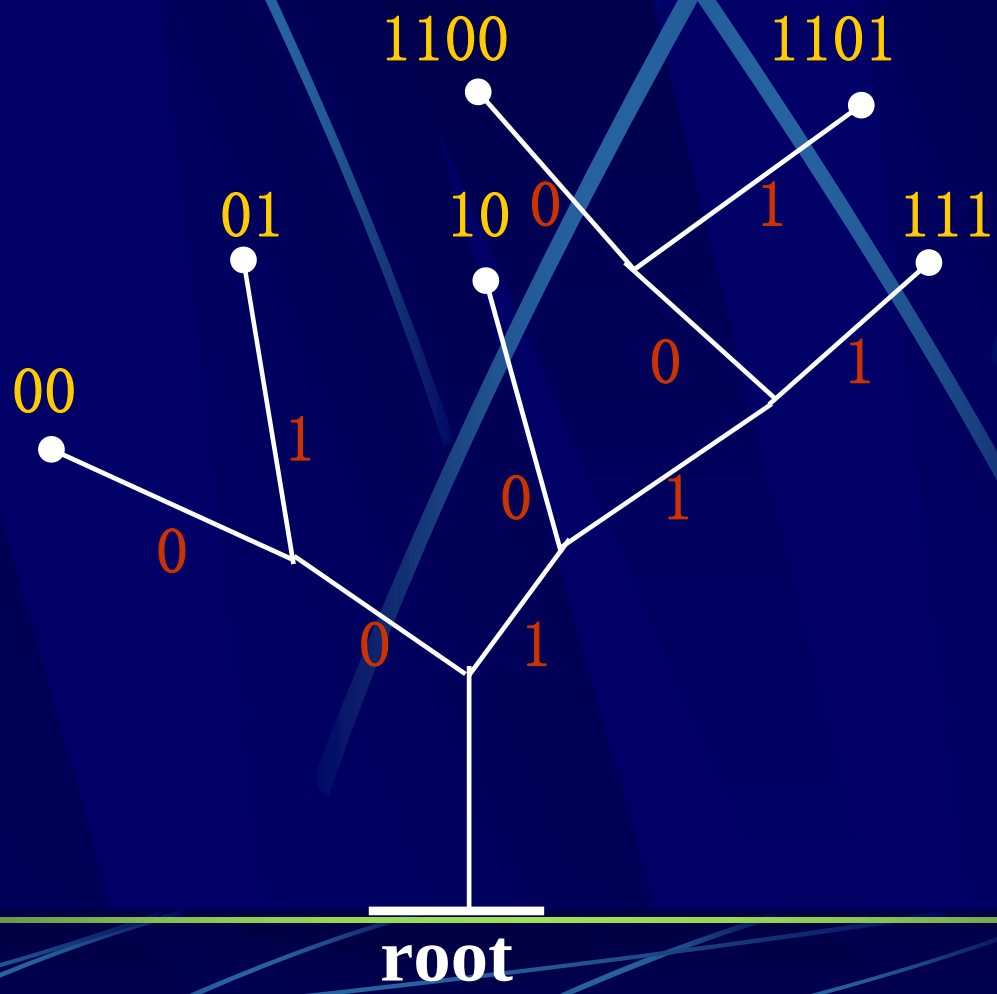
（1）异前缀性：任一码字都不会是另一码字的前缀。因此不会出现一个码字没收完却被判为其它码字的情况。

（2）非延长性：任一码字都不会是另一码字的延长。因此不会出现一个码字已收完却还需要等待其它码字的情况。

❖我们希望的变长码应当是唯一可译的即时码。

3. 码树：

- 码树是设计唯一可译即时码的一个好办法。



码树的构造方法：

- 二元码应分布在一棵二叉树上。每枝分两叉，每个分叉的左右两枝分别标记为0与1。
- 生了叶的枝便不再分叉

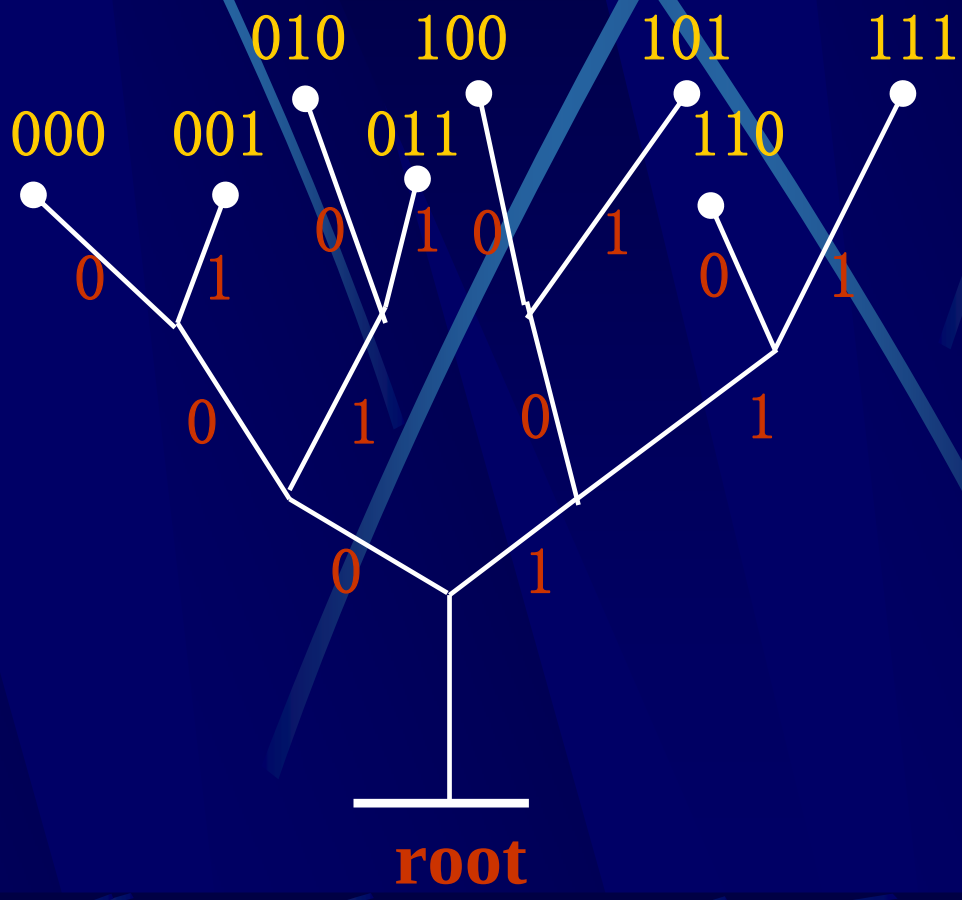
——非延长性。

- 从根到叶的路径只能有唯一的一条，这条路径就给出了树叶所代表的码字

——异前缀性。

- 只要改变每个结点处分叉的个数，就能推广到多元码树。

- 等长码形成一棵“整树”，所有的叶子在离根相同距离n的节点生成，共2^n个.



4. Krafft不等式：（唯一可译的必要条件）

- 我们用“剪枝”的方式从n阶整树上剪出一个最大码长不大于n的变长码的码树来。
- 若在第1阶分叉处张出一片码长为1的叶子，相当于整树被剪掉一半，则因此失去2^{n-1}片叶子。
- 若在第2阶分叉处张出一片码长为2的叶子，相当于整树被剪掉1/4，则因此失去2^{n-2}片叶子。
- 同理，若在第 l 阶分叉的一枝上张出一片码长为l 的叶子，则因此失去2^{n-l}片叶子。

设 m个变长码字的码长分别为$l_1, l_2, \ldots\ldots, l_m$；

它们是从n阶整树上剪出。必须满足：

$$\sum_{i=1}^{m} 2^{n-l_i} \le 2^n$$

才能有足够的枝叶可剪。因此：

$$\sum_{i=1}^{m} 2^{-l_i} \le 1$$

- 叫krafft不等式，是唯一可译码满足的必要条件。
- 克拉夫特不等式不能判断是不是即时码。

[例] 用krafft不等式判断下表给出的三种编码是不是唯一可译码。

	A	B	C	D
码 I	0	10	11	101
码 II	1	10	100	1000
码III	1	01	001	0001

解：码 I：$\because 2^{-1}+2^{-2}+2^{-2}+2^{-3}=\frac{9}{8}>1$

所以码 I 不是唯一可译码。

而码 II 与码III：$\because 2^{-1}+2^{-2}+2^{-3}+2^{-4}=\frac{15}{16}<1$

所以码 II 与码III有可能是唯一可译码。

5. 概率匹配原则(*Principle of match with probability*)

- 平均码长(*Average code length*): $\bar{l} = \sum_{i=1}^{m} p_i l_i$
- 信息传承要求 $\bar{l} \cdot \log(r)$应不小于$H(X)$；并尽量接近。即要求：

$$\bar{l}\log(r) - H(X) = \sum_{i=1}^{m} p_i l_i \log(r) - (-\sum_{i=1}^{m} p_i \log p_i)$$

$$= \sum_{i=1}^{m} p_i [l_i \log(r) + \log p_i] \to 0$$

如果求和中的每一项都为零，或者说如果每一个 i 都满足：$l_i \log(r)+\log(p_i)=0$ （i=1, 2, ……, m ）求和必然为零。它等价于：

$$l_i = -\frac{\log(p_i)}{\log(r)} = \log_r \frac{1}{p_i} = I_r(a_i)$$

- 此式表明，只要每一码字的长度都等于它所对应的信源符号的自信息（以r为底），就能使编码最短。这个原理叫概率匹配原则。

概率匹配原则的解释

- 从信息的荷载能力讲，信息量大的符号用长码，信息量小的符号用短码，是合理的。
- 自信息小的符号必然概率大，经常出现，采用较短的码字表示，必能节省代码长度。
- 自信息大的符号，虽然采用较长的码字表示，但由于它的概率小，不常出现，从总体上讲，不会明显影响平均码长。

1843年莫尔斯根据当地报馆使用铅字的情况编出Morse码

符号	铅字数	Morse码	符号	铅字数	Morse码
e	12000	•	m	3000	- - -
t	9000	-	f	2500	• • - •
a	8000	• -	w	2000	• - -
i	8000	• •	y	2000	- • - -
n	8000	- •	g	1700	- - •
o	8000	- -	p	1700	• - - •
s	8000	• • •	b	1600	- • • •
h	6400	• • • •	v	1200	• • • -
r	6200	• - •	k	800	- • -
d	4400	- • •	q	500	- - • -
l	4000	• - • •	j	400	• - - -
u	3400	• • -	x	400	- • • -
c	3000	- • - •	z	200	- - • •

6. Shannon变长码编码定理：

- 单个信源符号的编码：

考虑到码长只能取整数，概率匹配原则可写为：

$$\log_r(1/p_i) \leqslant l_i \leqslant 1 + \log_r(1/p_i)$$

对各符号取统计平均，得到：

$$\frac{H(X)}{\log r} \leq \bar{l} \leq \frac{H(X)}{\log r} + 1$$

- 它给出最小平均码长的一个范围。

- N个信源符号为分组的编码：

把N个信源符号的序列当作一个符号来编码，其码字平均码长L满足：

$$\frac{H(X_1X_2\cdots X_N)}{\log r} \le \bar{L} \le \frac{H(X_1X_2\cdots X_N)}{\log r} + 1$$

为便于比较，仍然平均到单个信源符号上，就有：

$$\frac{H_N}{\log r} \le \bar{l} \le \frac{H_N}{\log r} + \frac{1}{N}$$

当$N \to \infty$时，就有：$\bar{l} \to (H_\infty / \log r)$；

这就是Shannon给出的极限码长。

编码效率

定义：$$\eta = \frac{H_\infty}{\bar{l} \log r}$$

- 编码效率，代表实际编码的码长与极限码长的逼近程度；
- 有时也被称为**信息率**，代表编码后平均单位码元所荷载的信息量。
  - 式中**log r**是由于采用**r**进制码元引起的信息单位换算。
  - 对于无记忆信源， $H_\infty = H(X)$.

小结：

- **信源编码的原理：**
  - 等长码编码原理-----典型序列与非典型序列
  - 变长码编码原理-----概率匹配原则
- **唯一可译性与即时性：**
  - 码树-----构造唯一可译码的作图法
  - Krafft不等式-----唯一可译码的必要条件
- **Shannon信源编码定理**

课后复习题

- 思考题：

等长码和变长码各自为什么能够压缩代码长度？

- 作业题：

教材第63页习题二第1、3题；

第2章 无失真信源编码

2.2 赫夫曼(Huffman)编码

（第4讲 2007.9.20.）

- 计划学时：2学时
- 要求掌握的主要内容：

  1. 熟练掌握 Huffman编码方法。
  2. 熟练掌握计算平均码长与编码效率的方法。
  3. 深刻理解Huffman编码的适用条件，并掌握用N次扩展信源实现Huffman编码的方法。
- 重点难点：

  重点---- 画码树进行Huffman编码

  难点---- Huffman编码用于相关信源

[温旧引新]

- 码树-----构造唯一可译即时码的作图法。
- 变长码编码原理-----概率匹配原则：

$$\log_r(1/p_i) \leqslant l_i \leqslant 1+\log_r(1/p_i)$$

经常出现的符号采用较短的码字表示，不常出现的符号采用较长的码字表示，使平均码长最短。

- 平均码长：$\bar{l}=\sum_{i=1}^{m} p_i l_i$
- 极限码长：$l_0 = H_\infty / \log r$

2.2.1 香农(Shannon)码

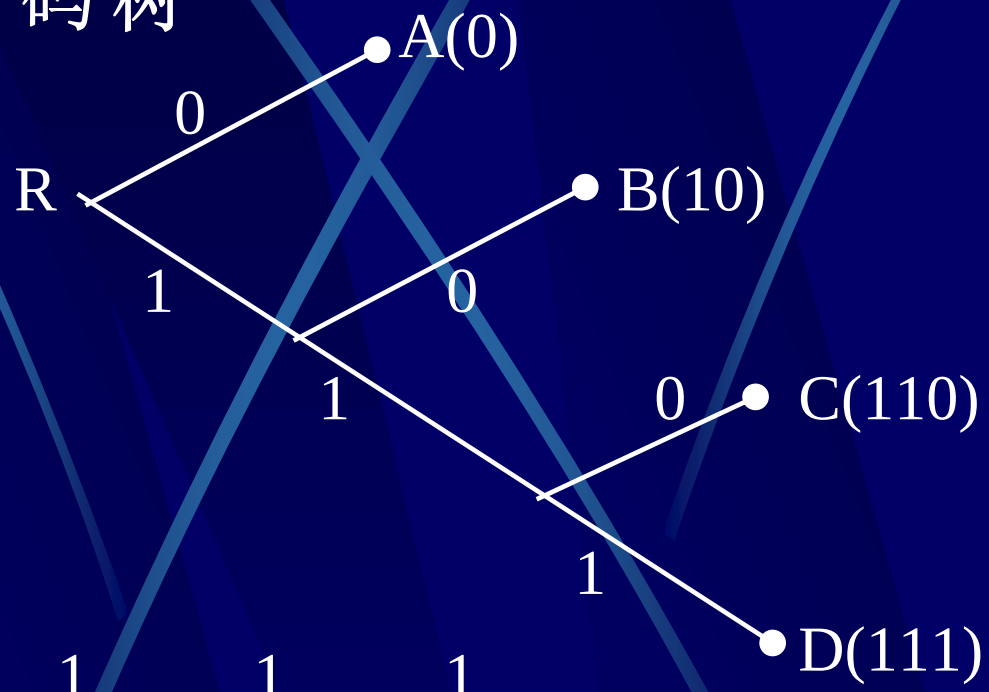
直接利用概率匹配原则进行单符号信源编码，就叫**Shannon**码。

[例1]无记忆信源发出四符号：

	A	B	C	D
已知概率***p***分别为：	1/2	1/4	1/8	1/8
可算出自信息**-log *p***：	1	2	3	3
根据概率匹配原则码长可取为：	1	2	3	3

$$l_i = I_r(a_i) = -\log_r p_i \qquad (a_i = A, B, C, D)$$

根据码长作出码树图：



∵ 平均码长：$\bar{l}=\sum_{i=1}^{m}p_i l_i=\frac{1}{2}\times 1+\frac{1}{4}\times 2+\frac{1}{8}\times 3+\frac{1}{8}\times 3=1.75$

$$H(X)=-\sum_{i=1}^{m}p_i\log p_i=\frac{1}{2}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{4}+\frac{1}{8}\log\frac{1}{8}+\frac{1}{8}\log\frac{1}{8}=1.75$$

∴ 效率 η=100 %

[例2]无记忆信源发出五符号：

	a_1	a_2	a_3	a_4	a_5
已知概率p分别为：	0.4	0.3	0.2	0.05	0.05
可算出自信息$-\log p$：	**1.32**	**1.74**	**2.32**	**4.32**	**4.32**
根据概率匹配原则码长可取为	**2**	**2**	**3**	**5**	**5**

这是因为现在只能取不小于自信息的最小整数：

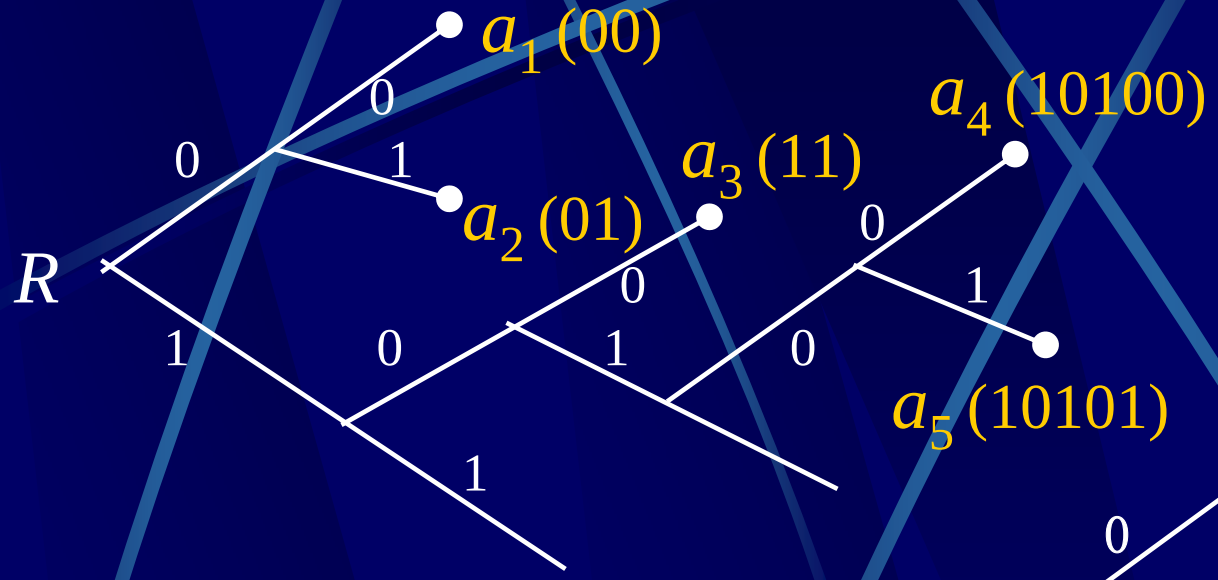
$$l_i \geq I_r(a_i) = -\log_r p_i \qquad (i = 1,2,3,4)$$

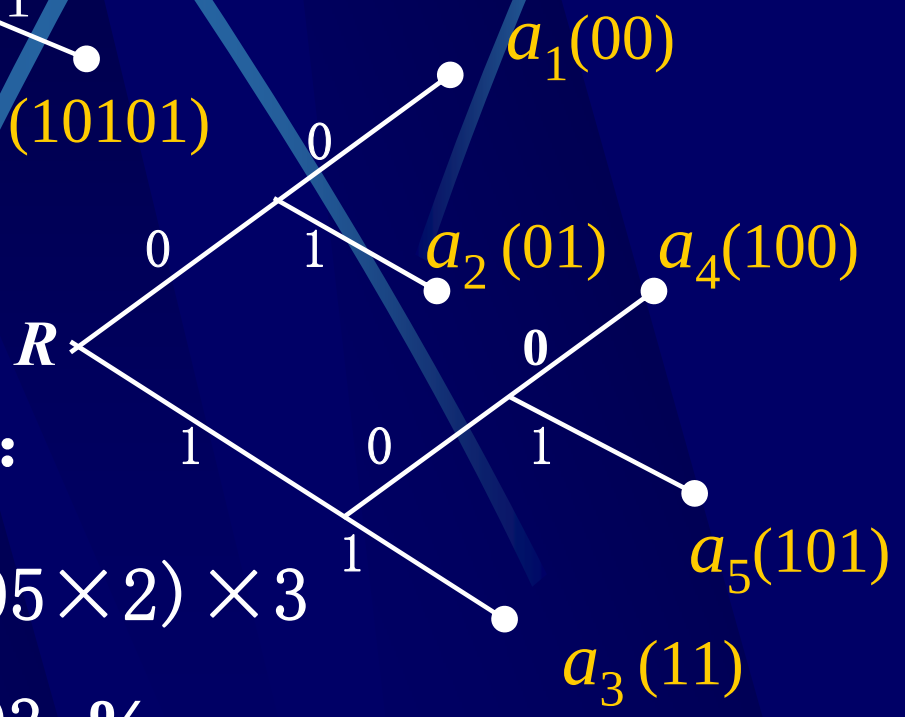
码长与概率不能严格匹配，编码效率也就不高。

$$\bar{l} = \sum_{i=1}^{m} p_i l_i = 0.4\times 2 + 0.3\times 2 + 0.2\times 3 + 0.05\times 5\times 2 = 2.5$$

$$H(X) = -\sum_{i=1}^{m} p_i \log p_i = 1.95 \qquad \therefore\ \eta = \mathbf{78\ \%}$$

码长为：2、2、3、5、5 时的码树(左图)：





调整布局后的码树(右图)：

$\bar{l}=(0.4+0.3+0.2)\times 2+(0.05\times 2)\times 3$

$=2.1$；　提高效率到 $\eta=93\ \%$

上例结果表明，整体的匹配比个体的匹配更重

2.2.2 费诺(Fano)码

费诺提出一种从根到叶的编码法，先把符号集分成两组，使各组总概率大致相等，每组符号再按概率大体相等的原则继续一分为二，直到每组只有一个符号为止。

- **[例3]用Fano编码法再编[例2]。**

[例2]无记忆信源：a_1 a_2 a_3 a_4 a_5

概率p分别为： 0.4 0.3 0.2 0.05 0.05

自信息： 1.32 1.74 2.32 4.32 4.32

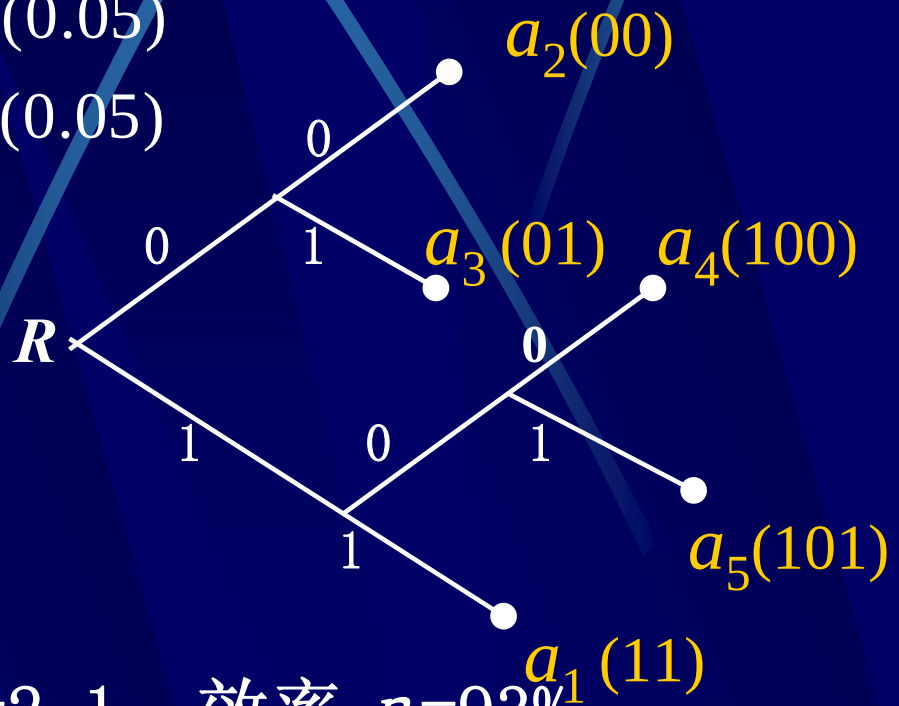
码长可取为 2 2 3 5 5

- 解：分组为：

$$\begin{cases} a_1(0.4) \\ a_1a_4a_5(0.5) \begin{cases} a_4a_5(0.1) \begin{cases} a_4(0.05) \\ a_5(0.05) \end{cases} \end{cases} \\ a_2a_3(0.5) \begin{cases} a_2(0.3) \\ a_3(0.2) \end{cases} \end{cases}$$

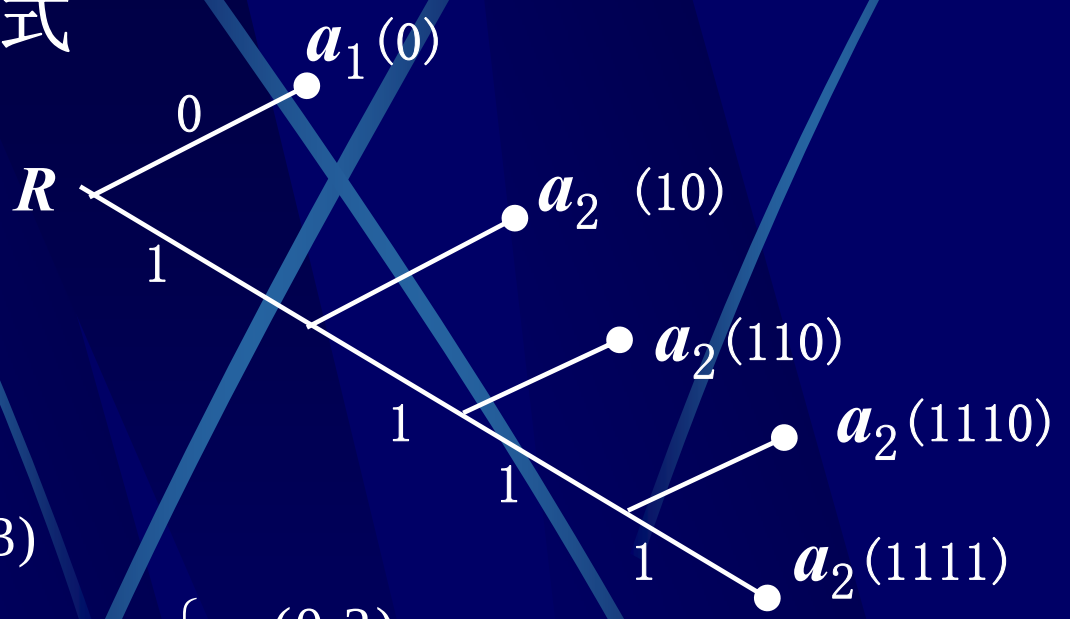
按分组情况构造码树，直接就得出的了与[例2]经调整后形状相同的码树 ：

$$\begin{cases} a_1a_4a_5(0.5)\begin{cases} a_1(0.4) \\ a_4a_5(0.1)\begin{cases} a_4(0.05) \\ a_5(0.05) \end{cases} \end{cases} \\ a_2a_3(0.5)\begin{cases} a_2(0.3) \\ a_3(0.2) \end{cases} \end{cases}$$



平均码长仍为：$\bar{l}$ =2. 1；效率 η=93%；

我们如果按下面方式进行分组和编码，发现结果会更好：



$$\begin{cases} a_1(0.4) \\ a_2a_3a_4a_5(0.6)\begin{cases} a_2(0.3) \\ a_3a_4a_5(0.3)\begin{cases} a_3(0.2) \\ a_4a_5(0.1)\begin{cases} a_4(0.05) \\ a_5(0.05) \end{cases} \end{cases} \end{cases} \end{cases}$$

$\bar{l}$

=0.4×1+0.3×2+0.2×3+(0.05×2)×4=2.0；效率 η=97.5 %；

2.2.3 霍夫曼(Huffman)码

- 先局部后全局的香农码和先全局后局部的费诺码，都难以统筹兼顾全局利益和局部利益。霍夫曼提出从叶到根的编码方法，把单个码字的匹配与整局布局的匹配巧妙地结合起来。
- 在构造码树的过程中，始终保持概率较大的码字离根较近，概率较小的码字离根较远。

霍夫曼编码方法：

(1) 把信源符号集中的所有符号按概率从大到小排队。

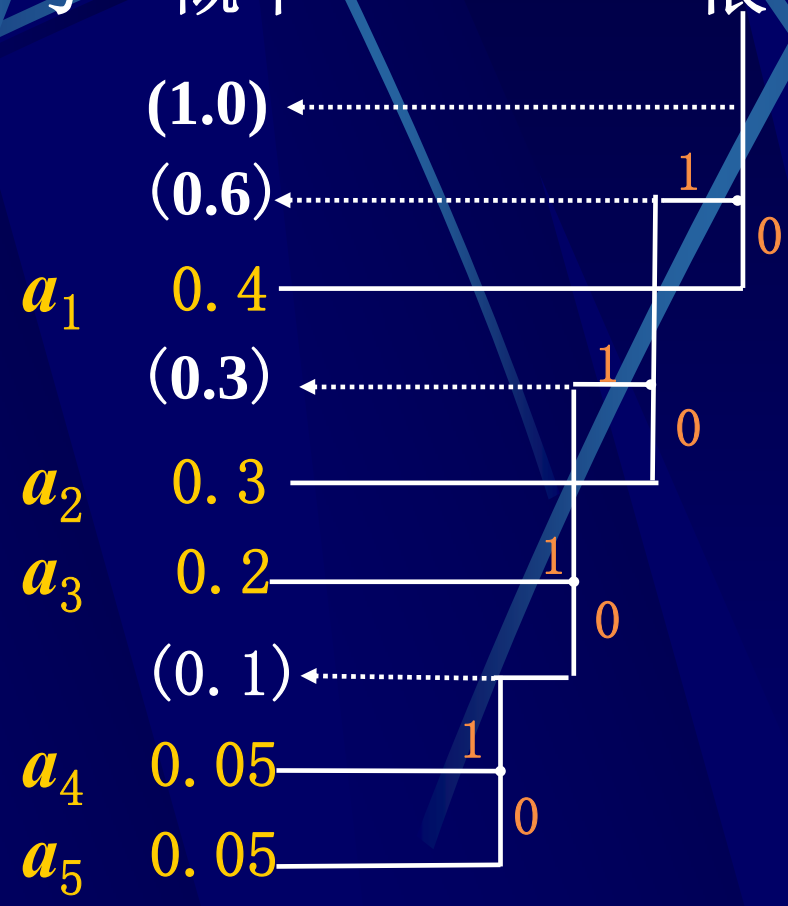
(2) 取概率最小的两个符号作为两片叶子合并到一个节点。

(3) 视此节点为新符号，其概率等于被合并的两个概率之和，参与概率排队。

(4) 重复(2)(3)两步骤，直至全部符号都被合并到根。

(5) 从根出发，对各分枝标记0和1。从根到叶的路径就给出了各个码字的编码和码长。

[例4]用Huffman编码法再编[例2]。



可以证明，霍夫曼编码是单符号信源编码的最佳方案

设$p(a_i) \geqslant p(a_j)$，则必有$l_i \leqq l_j$;

若将任意两符号a_i和a_j的码字W_i和W_j码字对换，即令

$$a_i \longleftrightarrow W_j,\ a_j \longleftrightarrow W_i;$$

而其它符号与码字的对应关系不变;

则互换后平均码长将由原来的$\bar{l}$ 变为：

$$\bar{l}' = \bar{l} - [p(a_i)l_i + p(a_j)l_j] + [p(a_i)l_j + p(a_j)l_i]$$

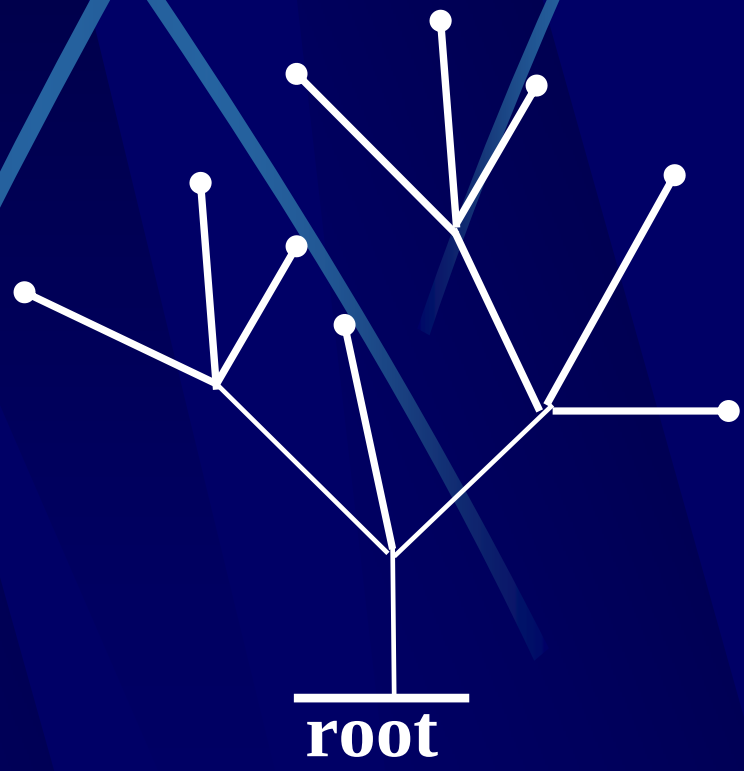
$$= \bar{l} + (l_j - l_i)[p(a_i) - p(a_j)]$$

式中的正、负号总是相同的，所以总有：$\bar{l}' \geq \bar{l}$

2.2.4 霍夫曼编码的推广

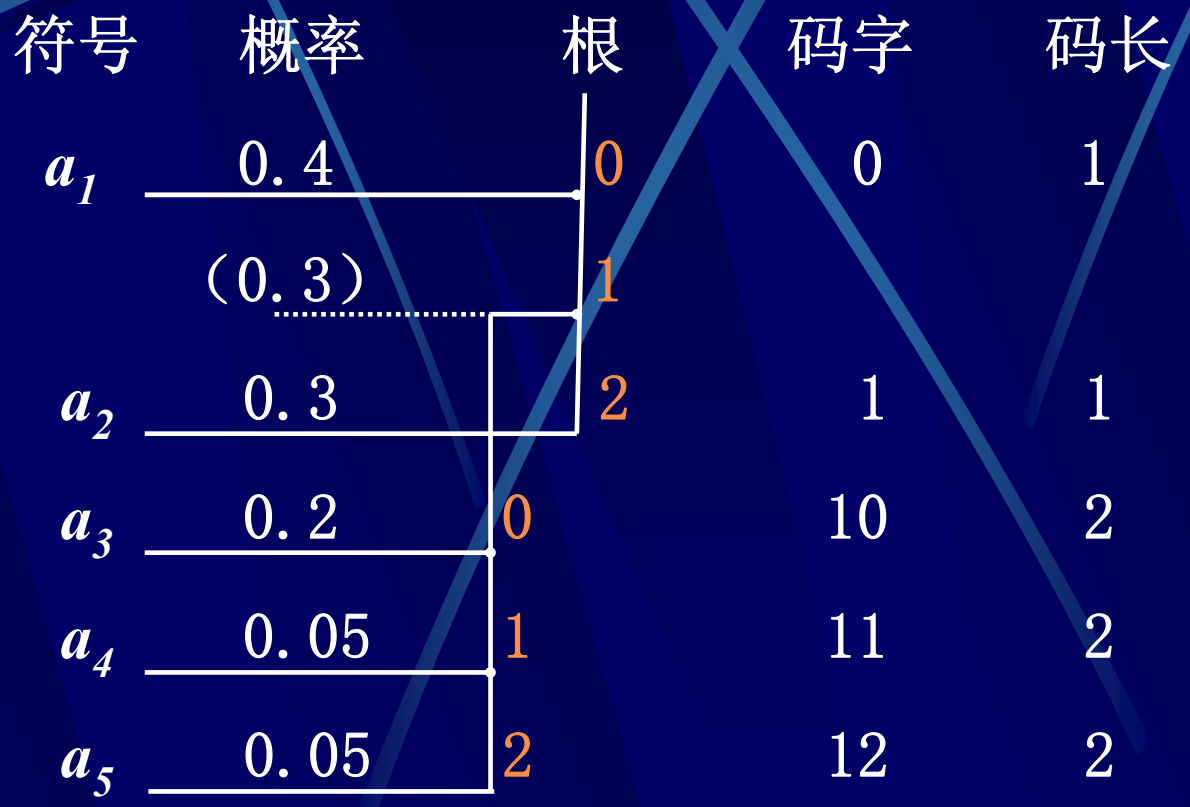
1. 多元霍夫曼编码

多元码可以用多元码树表示。*r*元码树每个节点分为*r*个枝权，编码时，仍然从概率最小的符号开始，每次*r*个符号合并为一个节点。



[例5]仍对例2的5个符号进行三元和四元编码。

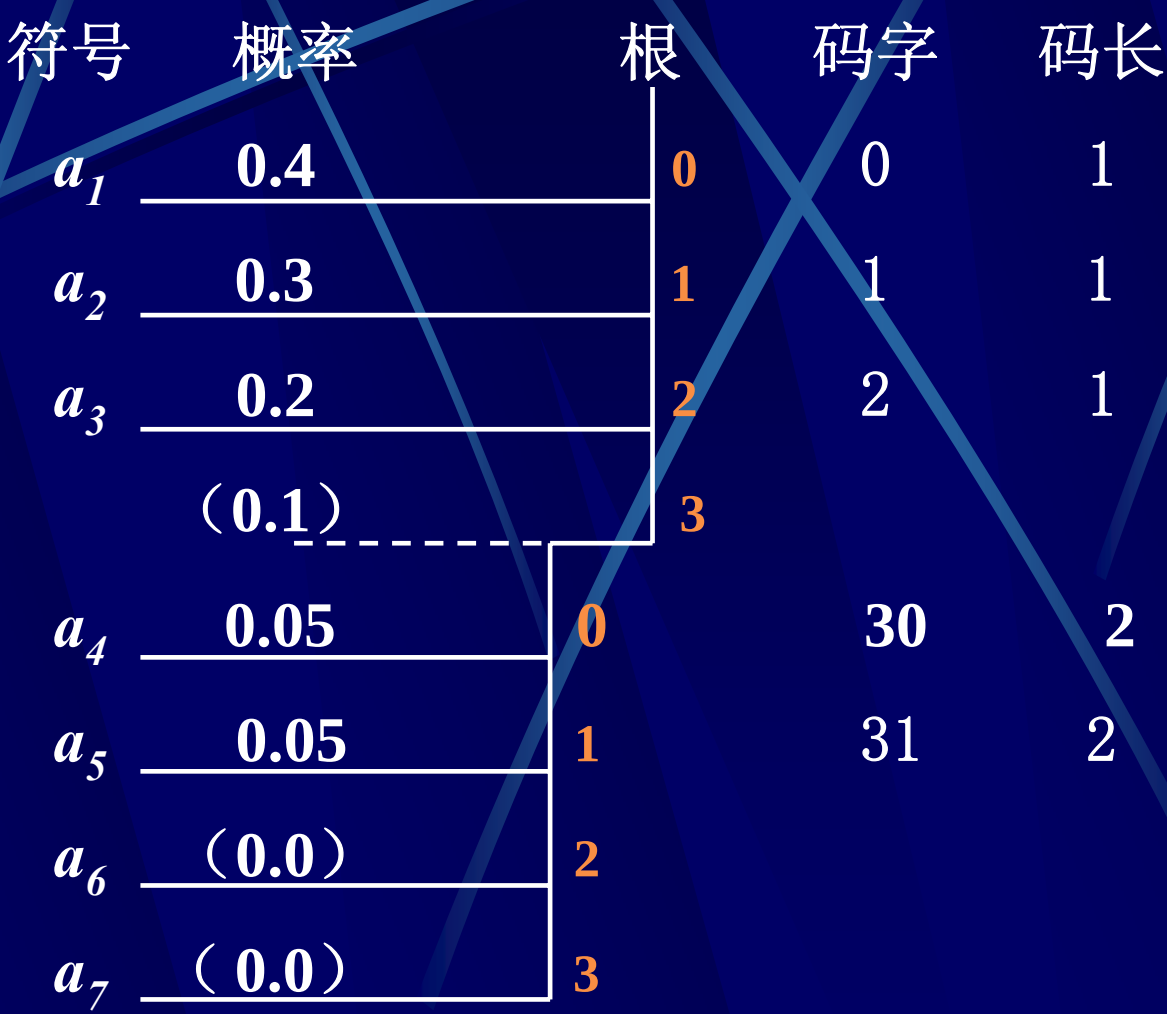
(1)三元编码：



$$\bar{l}=1.3; \qquad \eta=\frac{H(X)}{\bar{l}\cdot\log r}=\frac{1.95}{1.3\log 3}=94.6\%$$

- 三元编码第一次合并后，正好剩下3枝，再合并就到根。
- 而四元编码若第一次将四枝合并，则剩下2枝，没法进行第二次合并。为此，事先添加了两个概率为零的“虚”符号a_6和a_7，凑成4枝，使第二次合并时正好还有4枝。

(2) 四元编码：



$$\bar{l} = 1.1; \qquad \eta = \frac{H(X)}{\bar{l} \cdot \log r} = \frac{1.95}{1.1 \times \log 4} = 88.6\%$$

一般情况下，每分一次枝，增加r片叶，同时原先的叶就变成了节点，总体上只增加了$r-1$片叶。若分枝共S次，则共有叶子（即码字）数目为$r+S(r-1)$；设信源符号数目为m，应添加的虚符号数目为n，则当：

$$m+n=r+S(r-1)$$

才能使符号数与叶子数相同，经S次合并后正好到根。不难验证，三元码$m=5$，$r=3$，$S=1$时$n=0$；四元码$m=5$，$r=4$，$S=1$时$n=2$；都符合此式。

- 另外我们看到，从二元码到三元码、四元码，编码效率越来越低。如果编制五元码，则编码过程不过是换一套表示符号，不起任何压缩作用，等于没有编码。

- 结论是：霍夫曼编码只适用于信源符号数目m比码元符号r数目大很多的情况。

2．二元霍夫曼编码：

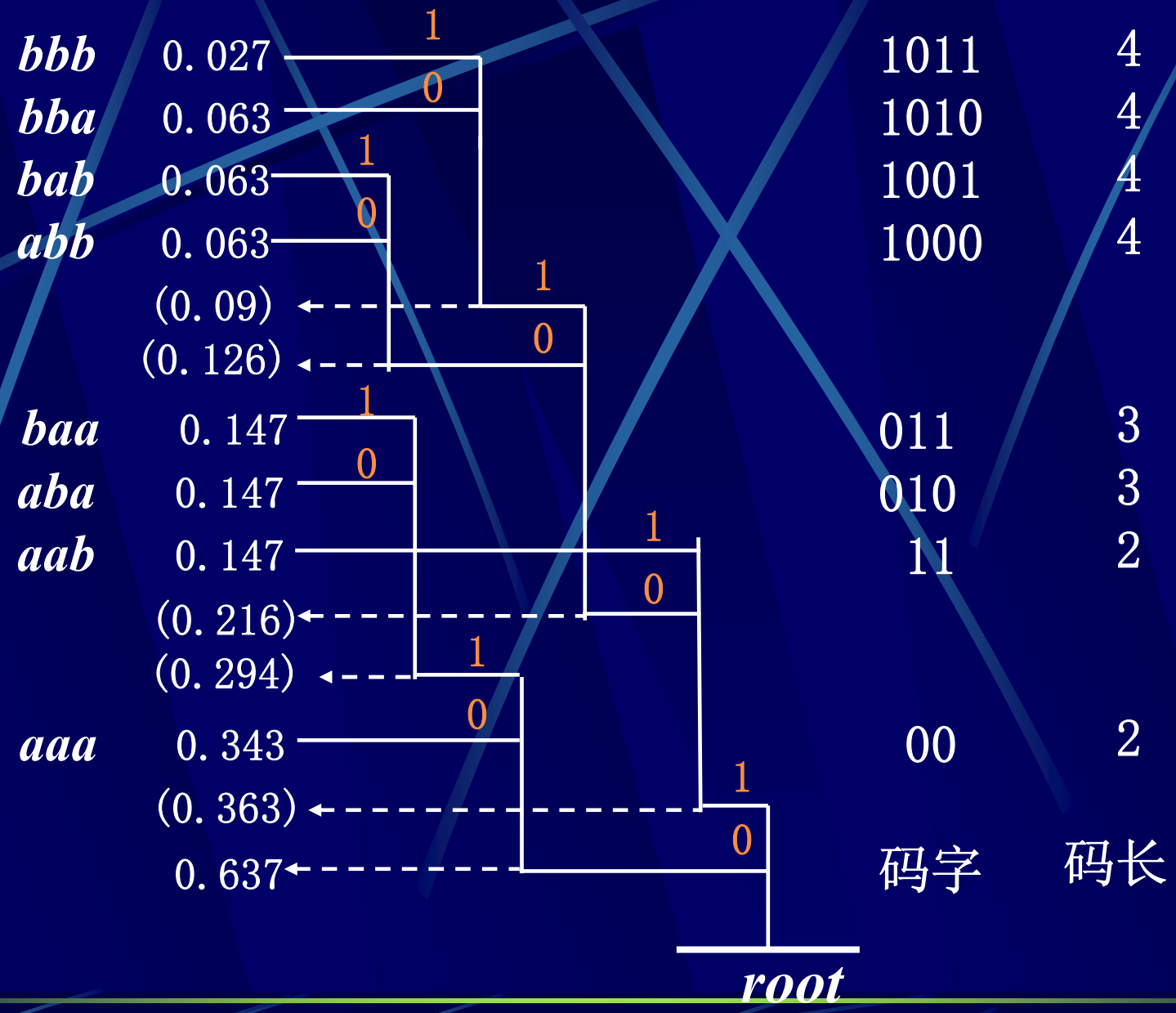
- 既然$m=r=2$时霍夫曼编码不起任何压缩作用，那么二元信源如何编成二元码字呢？
- 可以采用N维扩展信源来进行编码。所谓N维扩展信源指以把N个二元符号的符号串当作一个“符号”看待的信源。
- N维扩展信源扩展信源中共有2^N个不同的“符号”，由于$2^N >> r$，赫夫曼编码就能发挥威力，大大提高编码效率。

[例6]二元无记忆信源发出a、b两个符号，概率分别为0.7和0.3，试用三次扩展信源进行编码。

解：根据 $p(x_1x_2x_3)=p(x_1)p(x_2)p(x_3)$ 不难求出三次扩展信源的概率空间为：

$$\begin{pmatrix} X^3 \\ p(X^3) \end{pmatrix} = \begin{pmatrix} aaa & aab & aba & baa & abb & bab & bba & bbb \\ 0.343 & 0.147 & 0.147 & 0.147 & 0.063 & 0.063 & 0.063 & 0.027 \end{pmatrix}$$

按8个符号的概率排队，进行赫付曼编码


bbb 0.027
bba 0.063
bab 0.063
abb 0.063
(0.09)
(0.126)
baa 0.147
aba 0.147
aab 0.147
(0.216)
(0.294)
aaa 0.343
(0.363)
0.637
root
1011 4
1010 4
1001 4
1000 4
011 3
010 3
11 2
00 2
码字 码长

$$\begin{pmatrix} X^3 \\ p(X^3) \end{pmatrix} = \begin{pmatrix} aaa & aab & aba & baa & abb & bab & bba & bbb \\ 0.343 & 0.147 & 0.147 & 0.147 & 0.063 & 0.063 & 0.063 & 0.027 \end{pmatrix}$$

码字	00	11	010	011	1000	1001	1010	1011
码长	2	2	3	3	4	4	4	4

码字平均长度： L_N=2.726；

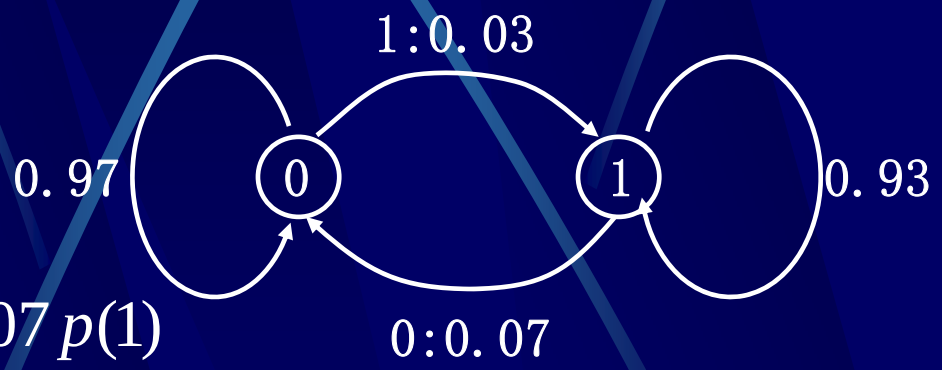
信源符号平均编码长度：$\overline{l}$ =2.726/3=0.909

编码效率： η=0.881/0.909=96.9%

- 采用N次扩展信源的另一个好处是可以用于有记忆信源，因为能够计入N个符号的内部关联。

[例7]一阶马尔科夫信源条件概率分别为$p(0|0)=0.97$和$p(1|1)=0.93$，试用三次扩展信源进行编码。

解：先作状态转移图



并求解稳态方程：

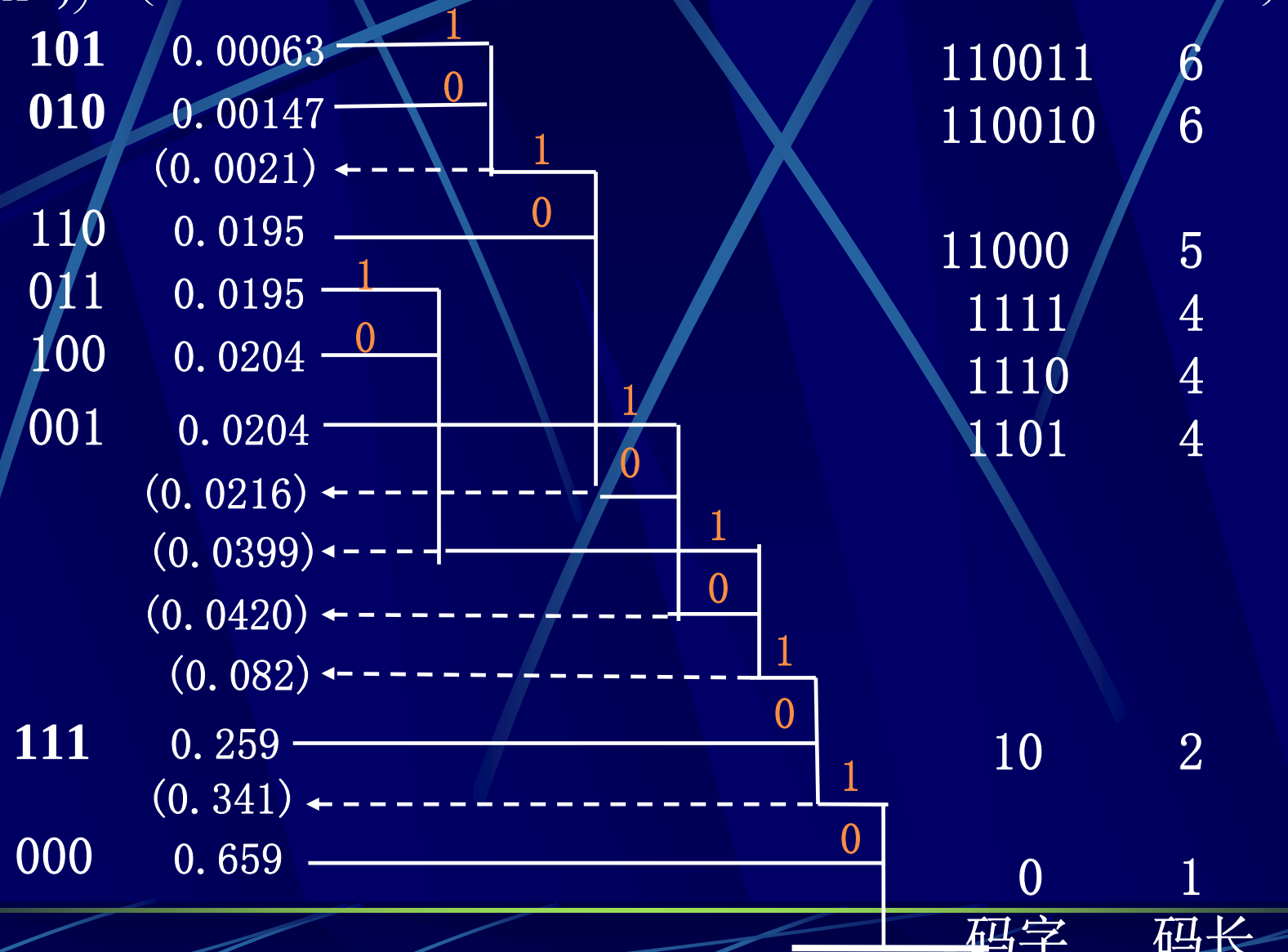
$$\begin{cases} p(0)+p(1)=1 \\ p(0)=0.97\,p(0)+0.07\,p(1) \end{cases}$$

得到：$p(0)=0.7$；$p(1)=0.3$;

根据 $p(x_1x_2x_3)=p(x_1)p(x_2|x_1)p(x_3|x_2)$ 不难求出三次扩展信源的概率空间为：

$$\begin{pmatrix} X^3 \\ p(X^3) \end{pmatrix}=\begin{pmatrix} 000 & 001 & 010 & 100 & 011 & 101 & 110 & 111 \\ 0.659 & 0.0204 & 0.00147 & 0.0204 & 0.0195 & 0.00063 & 0.0195 & 0.259 \end{pmatrix}$$

$$\begin{pmatrix} X^3 \\ p(X^3) \end{pmatrix} = \begin{pmatrix} 000 & 001 & 010 & 100 & 011 & 101 & 110 & 111 \\ 0.659 & 0.0204 & 0.00147 & 0.0204 & 0.0195 & 0.00063 & 0.0195 & 0.259 \end{pmatrix}$$



$$\bar{L} = 0.659\times 1 + 0.259\times 2 + (0.0204 + 0.0204 + 0.0195)\times 4$$
$$+ 0.0195\times 5 + (0.0147 + 0.000063)\times 6 = 1.53$$

$$\bar{l} = \bar{L}/3 = 0.51$$

$$H_\infty = H(X_2 \mid X_1) = -0.7(0.97\log 0.97 + 0.03\log 0.03)$$
$$- 0.3(0.93\log 0.93 + 0.07\log 0.07) = 0.246\ bit/\text{符号}$$

$$\eta = \frac{H_\infty}{\bar{l}} = 0.482$$

尽管码长的数值已经比无记忆信源编码短得多，但仍然还有很大的可压缩空间。

小结：

- **Huffman编码方法**
- **计算平均码长和编码效率：**

$$\bar{l} = \sum_{i=1}^{m} p_i l_i \qquad \eta = \frac{H}{\bar{l} \cdot \log r}$$

- **Huffman编码的扩展**

  多元码的Huffman编码

  扩展信源的二元编码

  相关信源的二元编码

课后复习题

- 实践题：

  试编程实现英文字母的Huffman编码。

（概率表参见课本14页）

- 作业题：

  教材第63页习题二第6、8题；

第2章 无失真信源编码

2.3 游程编码

（第5讲 2007.9.25.）

- 计划学时：1.5学时
- 要求掌握的主要内容：

  1. 深刻理解游程编码原理和意义。
  2. 掌握传真编码有关概念。
  3. 掌握修正的Huffman编码方法。

- 重点难点：

  重点----传真编码原理

  难点----游程编码信息熵理论

- 外语关键词：
  - 游程：**Run**
  - 游程长度：**Run Length**
  - 游程编码：**Run Length Encoding**
  - 传真编码：**Fax Encoding**
  - 激光扫描：**Laser Scan**
  - 黑(白)像素：**Black（White）Pixel**
  - 相关信源：**Correlated Information Source**

[温旧引新]

- **Huffman**编码的原理方法。
- **Huffman**编码的适用条件。
- 信息熵的计算公式：

$$H(X) = -\sum_{i=1}^{m} p(x_i) \log p(x_i)$$

- 平均码长的计算公式：

$$\bar{l} = \sum_{i=1}^{m} p_i l_i$$

2.3.1 什么是游程编码

- 在黑白图像及传真等情况下，信源序列往往呈现连0和连1的分布。
- 把连0的段叫0游程，把连1的段叫1游程。
- 把连0或连1的个数叫做该游程的长度。
- 把各个游程的长度值按原来的顺序组成一个序列，叫做游程序列。
- 二元序列变成游程序列的变换叫做游程编码。

游程编码方法：

如：000101111001100000111.....

游程编码为：31142253.....；

特点：

（1）二元码变成了多元码。

（2）0游程和1游程总是互相穿插的。

（3）一般规定0游程打头。

2.3.2 游程编码的概率特性

设信源为二元无记忆平稳信源，发0发1的先验概率为p_0和p_1

1、游程长度的概率分布：

用$m= L(0)$表示0游程的长度，用$n=L(1)$表示1游程的长度，显然m和n互相穿插，取值范围都是1到∞。

$m=1$的概率为p_1，这是因为既然是0游程，0就已经出现，游程长度为 1要求下一个码元必须发1，而发1 的概率就是p_1，

- 0游程长度m=2的概率等于0后面先发一个0再发一个1的概率，即为：p_0p_1,
- 0游程长度m=3的概率等于0后面先发两个0再发一个1的概率，即为：$p_0^2p_1$,
- 0游程长度m=m的概率等于0后面先发m-1个0再发一个1的概率，即为：$p_0^{m-1}p_1$；
- 结论是：　$p(m)=p_0^{m-1}p_1$；
- 同理：　　$p(n)=p_1^{n-1}p_0$；

- 不难证明，$p(m)$ 和 $p(n)$都是归一化的：

$$\sum_{m=1}^{\infty} p(m) = \sum_{m=1}^{\infty} p_0^{m-1} p_1 = p_1 \frac{1}{1-p_0} = p_1 \cdot \frac{1}{p_1} = 1$$

同理：$\sum_{n=1}^{\infty} p(n) = 1$

2、游程的平均长度：

$$l_0 = E[m] = \sum_{m=1}^{\infty} m p(m) = \sum_{m=1}^{\infty} m p_0^{m-1} p_1 = p_1 \sum_{m=1}^{\infty} \frac{d}{dp_0} p_0^m$$

$$= p_1 \frac{d}{dp_0} \sum_{m=1}^{\infty} p_0^m = p_1 \frac{d}{dp_0} \left(\frac{p_0}{1-p_0}\right) = \frac{p_1}{(1-p_0)^2} = \frac{1}{p_1}$$

同理：$l_1 = \frac{1}{p_0}$

- 不难证明，$p(m)$ 和 $p(n)$都是归一化的：

$$\sum_{m=1}^{\infty} p(m) = \sum_{m=1}^{\infty} p_0^{m-1} p_1 = p_1 \frac{1}{1-p_0} = p_1 \cdot \frac{1}{p_1} = 1$$

同理：$\sum_{n=1}^{\infty} p(n) = 1$

2、游程的平均长度：

$$l_0 = E[m] = \sum_{m=1}^{\infty} m p(m) = \sum_{m=1}^{\infty} m p_0^{m-1} p_1 = p_1 \sum_{m=1}^{\infty} \frac{d}{dp_0} p_0^m$$

$$= p_1 \frac{d}{dp_0} \sum_{m=1}^{\infty} p_0^m = p_1 \frac{d}{dp_0} \left(\frac{p_0}{1-p_0}\right) = \frac{p_1}{(1-p_0)^2} = \frac{1}{p_1}$$

同理：$l_1 = \frac{1}{p_0}$

3、游程码的信息熵

$$H(m)=-\sum_{m=1}^{\infty}p(m)\log p(m)=-\sum_{m=1}^{\infty}p(m)\log\ (p_0^{m-1}p_1)$$

$$=-\sum_{m=1}^{\infty}p_0^{m-1}p_1\log\ p_0^{m-1}-\sum_{m=1}^{\infty}p(m)\log\ p_1$$

$$=-\sum_{m=1}^{\infty}p_0p_1\log p_0\frac{d}{dp_0}p_0^{m-1}-\log p_1$$

$$=-p_1(p_0\log p_0)\frac{d}{dp_0}(\frac{1}{1-p_0})-\log p_1=\frac{1}{p_1}H(p)$$

式中：$H(p)=-p_0\log p_0-p_1\log p_1$

利用：$l_0 = \frac{1}{p_1}$

就有：$H(m)=l_0 H(p)$

同理：$H(n)=l_1 H(p)$

平均一个0游程加一个1游程的总信源量：

$$H(m)+H(n) = (l_0+l_1) \cdot H(p)$$

- 表明游程编码是等熵变换，变换并未改变信息 浓度，也就是 说游程编码并不能压缩代码长度。

- 游程编码的意义在于：将原来的二元码变为多元码，为下一步使用霍夫曼编码创造了条件。

2.3.3 游程编码用于相关信源

- 下面来证明，对于一阶和二阶马尔科夫信源，通过游程编码，原来的相关序列将变换成无关联的多元符号序列；对于三阶和三阶以上的马尔科夫信源，通过游程编码可以变换成弱关联的多元符号序列。

- 如：(.....111)(0X.....0)表示两相邻游程。式中X取1还是取0，将决定0游程的长度等于1还是大于1。对于一阶和二阶马尔科夫信源， X的取值与前面1游程的长度没有关系。

- 对于三阶马尔科夫信源，X与前面1游程中倒数两个码元有关，表明1游程的长度等于1还是大于1对X的影响是不同的。但是，1游程长度等于和大于2的所有情况对于X的影响是相同的。可见，两相邻游程长度之间的关联减弱了，并非任何长度都存在不同的关联，而仅仅是前游程长度等于1时有不同的关联。
- 结论是游程编码具有消除或减弱符号之间关联的作用，这对于相关信源的编码非常有用。

2.3.4 游程编码的应用-----传真编码

1．传真通信原理：

- 传真是借助电话线路把纸面文字或图画传输给对方的技术。
- 页面被分割成m行n列，并且当m和n很大时，每个小方格内画面的细节将变得不重要，可以用一个平均亮度或颜色来代替，称之为像素。（变画面为点阵）
- 黑白传真，其像素只有二色，白色为0，黑色为1
- 用激光对页面一行行地扫描，把各个像素的亮度或颜色，变成串行的电信号，就能加以传输。（变并行为串行）

- 像素划分得越小，图像和文字就越逼真，然而像素的增加会引起传输效率的降低。
- 当每毫米有6个像素时，就可以避免文字斜向笔划的台阶效应；
- 当每毫米达到27—40个像素时，就能显示出画面的灰度层次。
- 我国根据汉字特点采用8像素/*mm* 的行分辨率且8行/*mm* 或4行/*mm*的两种列分辨率标准。

2．传真编码原理：

- 前邮电部对七种样张进行统计，得到结果是白(*W*)、黑(*B*)二色像素出现的概率分别为0.933和0.067，信息熵为$H(X)$=-0.933log0.933-0.67log0.067=0.346，即使不考虑相关，也可达到 1/0.346=2.82的压缩比。
- 考虑到像素间的相关性，北京邮电大学信息工程系对人民日报的抽样统计表明，同色像素之间关联很强，条件概率$p(W|W)=0.97$，$p(B|B)=0.90$，由此算出二阶马尔科夫链的信息熵为H_2=0.2547，无失真编码压缩比约为4；
- 再考虑二维相关性，如大片的背景、规律的行间距及文字笔画的连续性等等，实际压缩比可达10以上。

3．传真编码方法：

传真编码是游程编码与霍夫曼编码相结合并加以改造的产物。

（1）游程编码：

以A4纸黑白传真为例，每行有1728个像素，从左到右，白游程开头，白色为0，黑色为1，将扫描所得的0、1序列变成随机游程长度l_0和l_1相间的队列，l_0和l_1的取值范围均为[0，1727]。

（2）修正的霍夫曼(MH)编码：

对各种白、黑游程长度分别统计出现概率，进行霍夫曼编码，结果存在(MH)编码表中。

- 由于白、黑游程各有1728个值，码字总数为2×1728=3456个，存储和查表都比较繁。
- 修正的霍夫曼编码把游程长度大于63的情况，都写成 $l=64K+R$，K的取值范围是(1,27)，叫做区间码。
- R取值范围是(0,63)，表示所余的零头，叫做结尾码。
- 这样一来，码字总数减少为 2×(27+64)=182个。
- 课本190页给出结尾码，191页给出区间码。

结尾码（部分）					
游程长度	白游程编码	黑游程编码	游程长度	白游程编码	黑游程编码
0	00110101	0000110111	32	000111011	000001101010
1	000111	010	33	00010010	000001101011
2	0111	11	34	00010011	000011010010
3	1000	10	35	00010100	000011010011
4	1011	011	36	00010101	000011010100
5	1100	0011	37	00010110	000011010101
6	1110	0010	38	00010111	000011010110
7	1111	00011	39	00101000	000011010111
8	10011	000101	40	00101001	000001101100
9	10100	000100	41	00101010	000001101101

构造码（部分）					
64	11011	0000001111	960	011010100	0000001110011
128	10010	000011001000	1024	011010101	0000001110100
192	010111	000011001001	1088	011010110	0000001110101
256	0110111	000001011011	1152	011010111	0000001110110
320	00110110	000000110011	1216	011011000	0000001110111
384	00110111	000000110100	1280	011011001	0000001010010
448	01100100	000000110101	1344	011011010	0000001010011
512	01100101	0000001101100	1448	011011011	0000001010100
576	01101000	0000001101101	1472	010011000	0000001010101
640	01100111	0000001001010	1536	010011001	0000001011010

[例1]某行游程序列为：131*W*, 4*B*, 6*W*, 65*B*；求编码。

解：编码为：10010+1000，011，1110，0000001111+010，

000000000001(结束符)；

[例2]某行MH编码压缩的传真信号为：

001101010101101010111101100001100110000000000001；

请恢复黑白像素序列，并计算压缩比。

解：码字为00110101，010，110101，011，11011+000011，0011，000000000001(结束符)；

查表即知原来的信号是：**0**白**1**黑**15**白**4**黑**77**白**5**黑**1626**白；

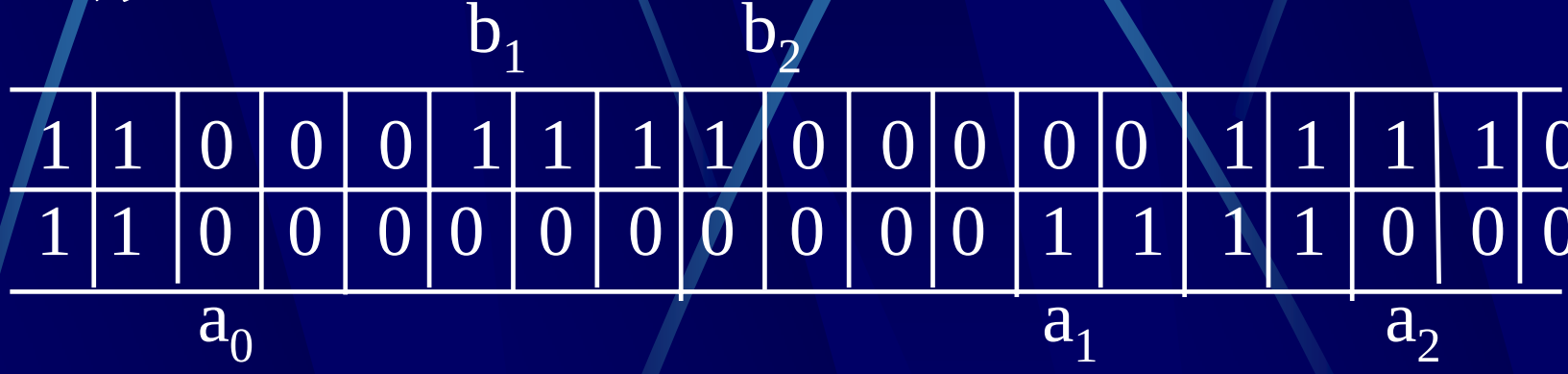
压缩比=1728/47=**36.7**:**1**；

（2）二维READ码：

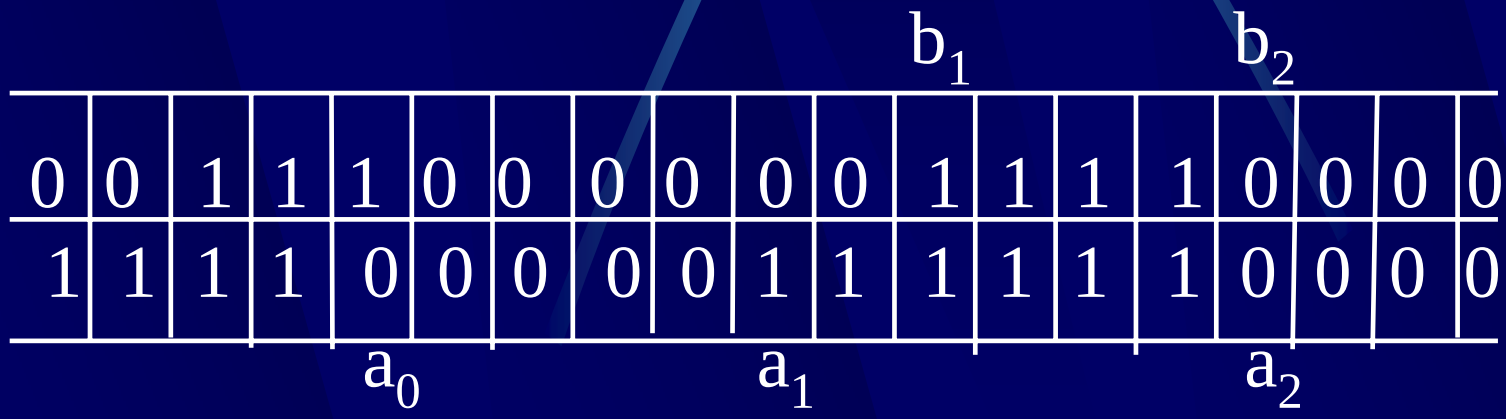
- 鉴于相邻行一般都很相似，所以按修正霍夫曼编码得到一行后，下行就能以上行为参考，用较少的代码来反映变化情况即可。
- 设a_0表示当前正在编码的像素，a_1表示位于待编行a_0右面且不同于a_0的像素（因只有黑白两种像素），a_2 表示位于待编行a_1右面且不同于a_1的像素；b_1表示位于参考行上a_0右面且不同于a_0的像素，b_2 表示位于参考行b_1右面且不同于b_1的像素。

- 分三种情况进行不同的编码处理：

①通过型：b_2位于a_1左面的情况。编码为0001，然后把b_2下面的像素作为新的a_0，继续进行。



②垂直型：a_1位于b_1正下方左、右3个像素之内的情况。 正下方记做V(0),左面一位记做V(-1)，右面一位记做V(1)；编码分别为：V(0)=1，V(-1)010 ，V(1)=011 ，V(-2)=000010 ，V(2)=000011，V(-3)=0000010，V(3)=0000011；此后，把a_2作为新的a_0，继续进行。



③水平型：除了上述两种情况之外，都属于水平型。其编码为：001+*MH*(a_0a_1)+ *MH*(a_1a_2)；式中*MH*(a_0a_1)表示游程(a_0a_1)的修正霍夫曼编码，*MH*(a_1a_2)表示游程(a_1a_2)的修正霍夫曼编码。此后，把a_2作为新的a_0，继续进行。

•由于采用二维READ码，可使压缩比达到12—15。

小结：

❖游程编码的原理：

提出了变二元符号为多元（游程长度）的思路；

证明了游程编码不改变信息熵；

讨论了游程编码有利于减弱信源的记忆；

❖传真编码的原理：

图像的数字化-----游程编码-----修正的Huffman编码------二维READ码

课后复习题

- 思考题：

只用游程编码为什么不能压缩代码长度？

- 作业题：

教材第63页习题二第10、11题；