第9卷 第6期 电子科学学刊 Vol. 9, No. 6
1987 年 11 月 JOURNAL OF ELECTRONICS Nov., 1987

数据压缩器的输出格式及其性能*

隋 厚 棠
（中国科学院空间科学技术中心）

提 要

本文从发送和接收两地之间的同步、编译码的容易程度、传输中产生错误的扩散范围和效率等方面分析讨论了多种遥测数据压缩器的输出格式. 并对几种典型格式的编码效率进行了计算比较和计算机模拟.

一、引 言

通常，固定时分多路遥测数据的输出格式在发送端是容易产生的；数据在接收端也是容易恢复的. 但是，采用了数据压缩技术之后，原来的数据关系就被破坏了. 为了在接收端能正确地识别出每个非多余字，必须随同非多余字一起发送通道和时间识别信息. 为了有效地利用通道，同时又不过分降低系统的比特压缩比，希望得到一种权衡了各种因素的高效的输出格式.

迄今为止，人们对数据压缩的研究主要限于源编码，而由此带来的对传输的要求所产生的逆作用却很少讨论. 本文针对这个问题，从效率、同步等方面分析讨论了几种典型的输出格式.

二、几种编码格式的分析讨论

压缩后的多路遥测数据可采用如下三种识别非多余字的格式编码方法[1]：（1）与非多余字一起发送一个通道地址码.（2）与非多余字一起发送一个能给出相邻非多余字之间的多余字个数的码字.（3）随同每一帧非多余字序列发送一个能识别出多余和非多余字序列状态的码字，这种方式可称为比特映射[2].

为了有效地表示非多余字的相对位置，文献[1,3]提出了两种游程编码. 一种为字溢出的游程编码，另一种为位溢出的游程编码. 位溢出编码比字溢出编码具有更高的效率.

文献[4]在文献[1]的基础上，从提高效率和克服同步困难出发，提出了一种输出格式，但性能如何，正如作者本人指出的，需要作理论分析. 该文讨论的格式采用了文献[1]提出的位溢出游程编码. 帧识别采用了新帧标志位[5]. 不难看出，这种格式有如下缺点：

* 1985 年 6 月 11 日收到，1987 年 7 月 1 日修改定稿.

(1) 同步图样 S 只给出了一个字的起始位置，所带信息太少；(2) 由于 F 是分散插入的，因此，只要一个 F 出错，就会使以后的帧时间关系遭到很大破坏；(3) 游程码本身和格式结构使编译码变得复杂化. 图 1 给出了这种格式的结构图. 设第 i 个组合字长为 B_{0i}，则有 $B_{0i} = F + L + K_i + C + W$，$(1 \leqslant i \leqslant A_0)$，式中 W 表示字码长度，A_0 表示一个同步块中的组合字数. 当 $F = 1, C = 1$，得位码长度为 $D_{0i} = 2 + L + K_i$，$(1 \leqslant i \leqslant A_0)$. 一个同步块的总长度为

$$M_0 = \sum_{i=1}^{A_0} (W + D_{0i}) + b_0 = A_0(W + 2 + L) + \sum_{i=1}^{A_0} K_i + b_0, \tag{1}$$

式中 b_0 为总开销(要保证 M_0 是常量). 总位码长度

$$M_D = A_0(2 + L) + \sum_{i=1}^{A_0} K_i. \tag{2}$$

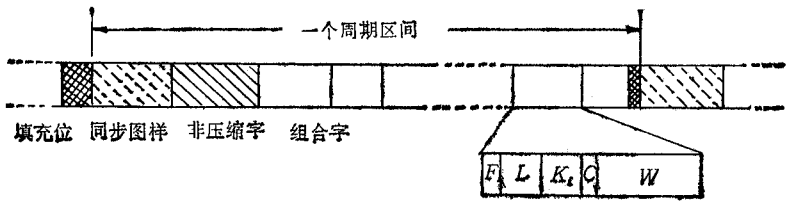


图 1 游程编码的一种输出格式

一个组合字的结构成份：F 为帧识别标志，1bit； L 为游程码字，L bit；K_i 为溢出标志，K_i bit； C 为结束识别，1bit； W 为数据字，W bit.

设 K_i 的均值为 $\bar{K}$，则 $\bar{K} = \sum_{i=1}^{A_0} (K_i/A_0)$，则平均位码长度为

$$\bar{D}_0 = 2 + L + \bar{K}. \tag{3}$$

平均组合字码长度为

$$\bar{B}_0 = \bar{D}_0 + W = 2 + L + \bar{K} + W. \tag{4}$$

一个同步块的长度可表示成

$$M_0 = A_0(\bar{D}_0 + W) + b_0. \tag{5}$$

总位码长度可表示成

$$\bar{M}_D = A_0\bar{D}_0 = A_0(2 + L + \bar{K}). \tag{6}$$

为了确定一帧的起始点，必须在插入的同步图样 S 确定的信息块中，先对 F 进行探测，F 本身的错误和 L, K, C 的错误都会破坏对 F 的探测，导致同步失败.

设比特错误率为 p_e，则 $\bar{M}_D$ 码不出错的概率为 $(1 - p_e)^{\bar{M}_D}$，至少出一个错的概率为

$$p_F = 1 - (1 - p_e)^{\bar{M}_D}. \tag{7}$$

表 1 给出了 $A_0 = 64$，$\bar{D}_0 = 5, 6$ 时，由 (7) 式计算的结果. 不难看出，为了使这种格式能正常工作，除非 p_e 非常小 $(p_e \ll 1)$. 但这样高的通道质量通常是无法保证的.

根据文献 [6] 给出的另一种游程编码方法，我们可组合成一种效率更高些，同步捕获和编译码都更容易些的输出格式.

将一帧中所有的通道状态按时间顺序映射成一个由“0”和“1”组成的序列，设“0”

表 1　$A_0 = 64$

$\bar{D}_0$ / p_F / p_e	5	6
0.01	0.960	0.979
0.001	0.274	0.319
0.0001	0.031	0.038

表示采样值是多余的，出现的概率为 q，“1”表示是非多余的，出现的概率为 p，则 $p+q=1$，$q>p$. 用一个“0”表示 2^L 个“0”(L 是正整数)，余下的用 L 比特码来表示. 表 2 给出了 $A_1=32$ 的编码过程举例，↓所指的“1”对应非多余字，△表示非多余字应插入的位置.

表 2　$A_1 = 32$

原序列 / L	编码序列	长度	压缩比
	0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1	32	
0	0 0 0 0 0 0 0 0 ↓1 0 0 0 ↓1 0 0 0 0 0 0 0 0 0 0 0 0 ↓1 0 0 0 0 0 0 ↓1 △ △ △ △	32	1
1	0 0 0 0 ↓1 0 0 ↓1 1 0 0 0 0 0 ↓1 1 0 0 0 ↓1 0 △ △ △ △ △	21	1.5238
2	0 0 ↓1 0 0 ↓1 1 1 0 0 ↓1 1 1 0 ↓1 1 0 △ △ △ △ △	17	1.8824
3	0 ↓1 0 0 0 ↓1 0 1 1 0 ↓1 0 1 1 ↓1 1 1 0 △ △ △ △ △	18	1.7778
4	↓1 1 0 0 0 ↓1 0 0 ↓1 1 1 ↓1 0 1 1 ↓1 0 1 1 0 △ △ △ △	20	1.6000

将位码和字码组合起来可构成一个组合字，它的平均位码长度为

$$\bar{D}_1 = \bar{K} + 1 + L. \tag{8}$$

平均组合字长度为

$$\bar{B}_1 = \bar{D}_1 + W = \bar{K} + 1 + L + W. \tag{9}$$

由 (3) 式和 (8) 式可知，$\bar{D}_0 > \bar{D}_1$，可见前者的位码效率不如后者. 图 2 (a) 给出了这种组合字的结构图. 设一帧中有 A_1 个可压缩字，由贝努里分布可知，非多余字出现的均值为 $\bar{A}_1 = E(A_1) = A_1 p$. 将一帧中的非多余字和 S 连结起来，L 码等均可方便地插在 S 的后面，这样就构成一帧信息，如图 2(b) 所示. 平均帧长度为

$$\bar{M}_1 = pA_1(W + L + 1) + \sum_{i=1}^{\bar{A}_1} K_i + b_1, \quad (0 \leqslant \bar{A}_1 \leqslant A_1), \tag{10}$$

式中 b_1 表示填充位和其它开销之和，它是个变量. 由 K_i 的均值 $\bar{K} = \sum_{i=1}^{\bar{A}_1} (K_i/\bar{A}_1)$，可得 $\bar{K}$ 表示的 $\bar{M}_1$ 为

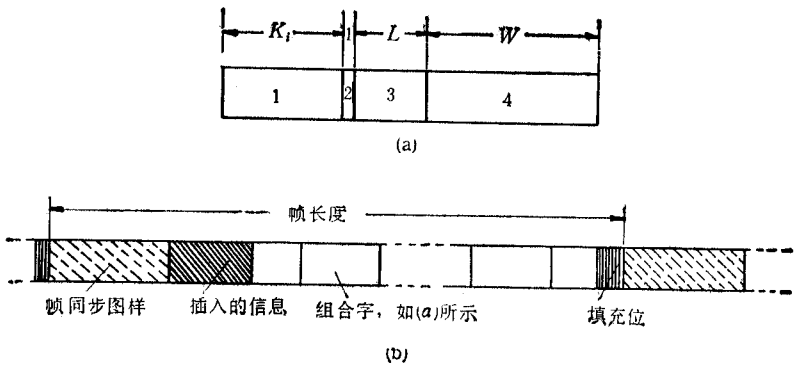


图 2 帧结构和组成成份

(a) 一个组合字的组成成份

1 表示有 K_i 个 "0"，每个代表 2^L 个 "0"；2 为非多余字的标志位；3 为 L bit 二进制码；4 为 W bit 采样值.

(b) 帧格式及其成份

$$\bar{M}_1 = pA_1(W + 1 + L + \bar{K}) + b_1. \tag{11}$$

比较 (5) 式和 (11) 式，当 $E(A_1) = A_0$，$b_1 = b_0$ 时，前者的帧效率低于后者.

当 $p_e \ll 1$ 时，这种格式的效率较高，帧识别也不成问题. 除工作开始阶段，接收机需对输入数据流按位扫描之外，一旦探测到一个 S，便可进入以组合字长为单位的跨越方式探测，这样可大大减少逐位探测时出现的假帧同步指示概率. 但由于组合字长是随机变化的，一旦位码出错，就会丢失大量的数据. 从表 1 ($A_0 = 64$，$\bar{D}_0 = 5$，相当于 $E(A_1) = 64$，$\bar{D}_1 = 5$) 可知，当 p_e 较大，如 $p_e > 0.0001$ 时，这种格式是不能正常工作的. 因此可用填充的办法，使每一帧位长度为某个整数倍. 如为 10 的整数倍，则填充长度 g 满足 $0 \leqslant g \leqslant 9$. 这样一旦找到一个 S，便可进入以 10 bit 为单位的跨越方式探测. 表 3 给出了两种帧同步捕获概率[7]；M 表示帧长度；n 表示 S 的长度；E 表示容错个数；p_1 表示逐位探测时的捕获概率. 不难看出，p_0 比 p_1 有明显改善，尤其是当 E 比较大时，改善更明显.

表 3 $p_e = 0.01$，$n = 16$，$M = 600$

E	0	1	2	3
p_1	0.8437	0.8432	0.2757	0
p_0	0.8506	0.9738	0.8817	0.5283

设跨越探测的间距为 G，填充长度 g 为均匀分布，其密度函数 $f(g) = \frac{1}{G}$，$g \in [0, G)$；$f(g) = 0$，$g \bar{\in} [0, G)$，$g_{\max} = G - 1$；平均填充长度 $\bar{g}$ 为

$$\bar{g} = \sum_{g=0}^{G-1} gf(g) = \sum_{g=0}^{G-1} \frac{g}{G}, \quad g \in [0, G). \tag{12}$$

由 (12) 式可算得，当 $G = 5$ 时，$\bar{g} = 2$，当 $G = 10$ 时，$\bar{g} = 4.5$. 通常压缩后的帧长度应在 $M = 1000$ 左右，因此，加入填充位不会对效率有明显影响.

因为 S 是在以 G 为等间隔的一些可预测的点上，所以可以先实现帧识别，然后实现字

识别，其优点：(1) 不会因位码出错而丢失帧同步，(2) 位码出错，其错误的扩散范围也会被限制在一帧之内．下面给出 L 和 q 之间的表达式．

从表 2 可以发现 $L_{\min}=2$ 为最佳值．如何选择 $L_{\min}$，有两种方法：(1) 用不同的 L 多次编码，选择 $L_{\min}$ 的发送．但 $L_{\min}$ 是随机变化的，所以要经常发送．(2) 用已知的 q 算出 $L_{\min}$．设“0”游程为几何分布，则“0”游程长度为 R 的出现概率为 $q(R)=q^{R-1}p$，$(R=1,2,\cdots)$．则 R 的期望值为

$$E(R)=\sum_{R=1}^{\infty}Rq^{R-1}p=\frac{1}{p},\quad \left(q>\frac{1}{2}\right). \tag{13}$$

设 $2^L=m$，编码后游程长度 R_c 的出现概率为 $q(R_c)$，则 R_c 的期望值为

$$E(R_c)=\sum_{K=0}^{\infty}\sum_{R=Km}^{(K+1)m}q^{R-1}p(L+K+1). \tag{14}$$

经简化处理后得

$$E(R_c)=L+\frac{1}{1-q^m}. \tag{15}$$

对 (15) 式经差分运算，求出极值点处的 L 和 q 之间的关系为

$$q^{2^{L_{\min}}-1}=\frac{1}{2}(\sqrt{5}-1)=0.618. \tag{16}$$

从 (16) 式解出 $L_{\min}$（将 $L_{\min}-1$ 看成 $L_{\min}$）有

$$L_{\min}=\left[3.322\log\left(\frac{\log 0.618}{\log q}\right)\right], \tag{17}$$

$[\cdot]$表示取最大整数．给定 q 之后，便可从 (17) 式求得 $L_{\min}$，以满足系统的最佳设计要求．为了识别一帧是否结束，映射序列中的最末一位应是“1”，可用强迫发送或用一个非压缩字来实现．下面给出 $\bar{K}$ 和 q 之间的函数关系．

同样设 $q>p$，$m=2^L$，则连续出现 m 个“0”的概率为 $U=q^m$，而“0”游程长度 $R_j=(j-1)m+(t-1)$，$j=1,2,\cdots$，$t=1,2,\cdots,2^L$ 的出现概率为

$$q_t(R_j)=U_1U_2\cdots U_{j-1}q_1q_2\cdots q_{t-1}p,\quad j=1,2,\cdots,\ t=1,2,\cdots,2^L. \tag{18}$$

用随机变量 X 表示 j，则“1”出现在第 j 块中的概率为

$$p_X(j)=p\{X=j\}=U^{j-1}\sum_{t=1}^{m}q^{t-1}p,\quad j=1,2,\cdots, \tag{19}$$

其中

$$p\sum_{t=1}^{m}q^{t-1}=1-q^m=1-U,\quad (q<1). \tag{20}$$

因此得

$$p_X(j)=(q^m)^{j-1}(1-q^m)=U^{j-1}(1-U). \tag{21}$$

由 (21) 式可知，压缩后的“0”游程同样满足几何分布，因此得 j 的期望值 $E(j)$ 为

$$E(j)=\sum_{j=1}^{\infty}jU^{j-1}(1-U)=\frac{1}{1-U}. \tag{22}$$

由 (22) 式可知，对于不同的 L，可求得不同的 $E(j)$．将 (17) 式求得的 $L_{\min}$ 代入 (22)

式可得

$$E(j_0)=\frac{1}{1-U_0}=\frac{1}{1-q^{2L_{\min}}}. \tag{23}$$

由游程编码的定义可知，K_i 的均值 $\bar{K}=E(j)-1$. 设满足最佳关系的 $\bar{K}$ 为 $\bar{K}_0$，这时有

$$\bar{K}_0=E(j_0)-1=\frac{1}{1-q^{2L_{\min}}}-1. \tag{24}$$

由(8)式可得最佳位码长度为

$$\bar{D}_{1\min}=\frac{1}{1-U_0}+L_{\min}. \tag{25}$$

考虑到最后一个字为强迫发送字，则压缩后的映射序列长度为

$$M'_{\min}=\left(\frac{1}{1-U_0}+L_{\min}\right)[(A_1-1)p+1]. \tag{26}$$

将(25)式代入(9)式得

$$\bar{B}_{1\min}=\frac{1}{1-U_0}+L_{\min}+W. \tag{27}$$

一帧中平均出现非多余字的个数为 $E(A_1)=p(A_1-1)$，由(11)式得

$$\begin{aligned}\bar{M}_{1\min}&=\bar{B}_{1\min}(A_1-1)p+\bar{B}_{1\min}+\bar{g}+b_1\\&=\left(\frac{1}{1-U_0}+L_{\min}+W\right)(pA_1-p+1)+\bar{g}+b_1,\end{aligned} \tag{28}$$

式中 $\bar{g}$ 可由(12)式求得. b_1 中包括 S 和其它信息长度之和. S 的长度 n 可由

$$n=e\log_2 E(A_1) \tag{29}$$

估算，e 可在 3-5 之间选取，也可由文献[7,8]给出的公式准确求得. 这样可从(28)式算出压缩后的帧长度. 当不采用压缩技术时，$p=1$，$\bar{g}=0$，$\bar{B}_{1\min}=W$，代入(28)式得未压缩的帧长度为

$$M_1=A_1W+b_1. \tag{30}$$

压缩比可定义为未压缩的帧长度与压缩后的帧长度之比

$$CR_1=\frac{M_1}{\bar{M}_{1\min}}=\frac{A_1W+b_1}{\bar{B}_{1\min}(pA_1-p+1)+\bar{g}+b_1}. \tag{31}$$

膨胀比可定义为压缩比的倒数

$$ER_1=\frac{\bar{M}_{1\min}}{M_1}. \tag{32}$$

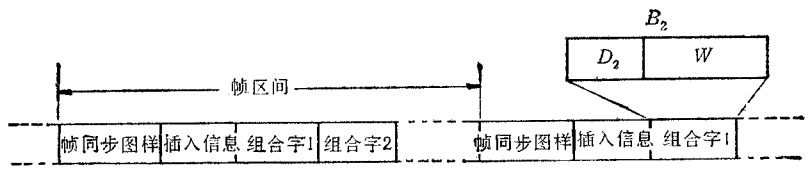
当 $p=1$ 时，得最大膨胀比

$$ER_{1\max}=\frac{(1+L_{\min}+W)A_1+\bar{g}+b_1}{A_1W+b_1}=1+\frac{(1+L_{\min})A_1+\bar{g}}{A_1W+b_1}. \tag{33}$$

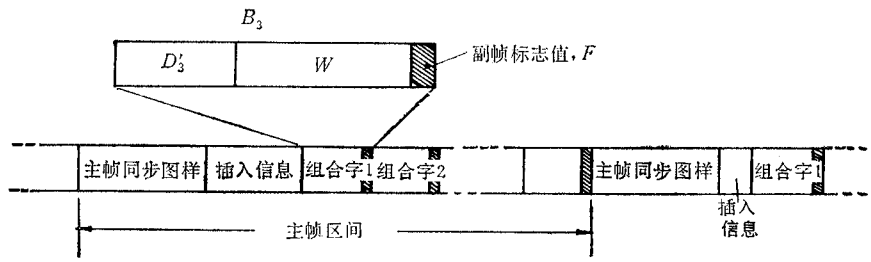
当 $p=0$ 时，得最大压缩比为

$$CR_{1\max}=\frac{A_1W+b_1}{\bar{B}_{1\min}+\bar{g}+b_1}. \tag{34}$$

由(33)和(34)式可知，只有当 p 较小时才能采用数据压缩技术，否则会出现膨胀.



(a) 一般帧结构的一种输出格式



(b) 一种主副帧结构的输出格式

图 3　带绝对地址字的两种输出格式

为了减少组合字的相互依赖性，以便减少错误扩散的程度．可用数据字带绝对地址的方式．这可分为两种，见图 3 (a),(b)．对于图 3 (a)，它的组合字长为

$$B_2 = W + D_2, \tag{35}$$

D_2 表示位码长度．　如果一帧中有 A_2 个压缩字，则 $D_2 = [\log_2 A_2]$，[·] 表示取最大整数．若一帧中平均非多余字个数为 $E(A_2)$，则一帧的平均长度为 $\bar{M}_2 = E(A_2)B_2 + b_2$．因 $E(A_2) = A_2 p$，因此得

$$\bar{M}_2 = A_2 p B_2 + b_2. \tag{36}$$

对于未采用压缩技术的系统，$D_2 = 0$，$B_2 = W$，$p = 1$，这时帧长度（设附加信息的长度是相等的）

$$M_2 = A_2 W + b_2. \tag{37}$$

压缩比为

$$CR_2 = \frac{M_2}{\bar{M}_2} = \frac{A_2 W + b_2}{A_2 p B_2 + b_2}. \tag{38}$$

由 (38) 式可看出，当 p 较大时，帧长度同样会出现膨胀．当 $p = 1$ 时，膨胀比为

$$ER_{2\max} = 1 + \frac{A_2 D_2}{A_2 W + b_2}. \tag{39}$$

当 $p \ll 1$ 时，这种格式的效率很高，极限情况，$p = 0$（相当于空帧），这时的压缩比为

$$CR_2 = 1 + \frac{A_2 W}{b_2}. \tag{40}$$

对于表 2 给出的例子，$A_1 = 32$，按绝对地址编码，总位码长度为 $4 \times 5 = 20$，可见其效率比 $L = 1$ 时的要好，但比 $L = 2$ 或 3 时的要差．

这种绝对地址码格式，抗干扰能力较强，出现错误影响的范围小．如果 D_2 码错到相

邻非多余字之外，可用 D_2 码序号将其排除．而错在相邻非多余字之内，常可用野值甄别技术将其排除．

图 3 (b) 给出主副帧结构．当 $p_e \ll 1$，每副帧中平均非多余字数 $E(A_3)$ 比较少 (A_3 表示一副帧中可压缩字个数)，这时可采用主副帧结构[5,9]．如位码长度为 $D_3 = D_3' + F$，组合字长度为

$$B_3 = D_3 + W = D_3' + F + W, \tag{41}$$

则一副帧平均组合字长为 $E(A_3)B_3$．如果用 N 个副帧组成一个主帧，且 $F = 1$，$B_3 = D_3' + 1 + W$，则平均主帧长度为

$$\bar{M}_3' = NE(A_3)(W + D_3' + 1) + b_3', \tag{42}$$

式中 b_3' 表示主帧附加信息．当 $D_3' = D_2$ 时，有 $B_3 > B_2$，可见后者的字码效率不如前者．但是如果 $E(A_3) < E(A_2)$，合理的选用 n 和 N，其效率就可能比前者高．由 (42) 式可得一主帧长度为

$$\bar{M}_3' = NA_3p(W + D_3' + 1) + b_3'. \tag{43}$$

平均副帧长度为

$$\bar{M}_3 = \frac{\bar{M}_3'}{N} = A_3p(W + D_3' + 1) + b_3, \tag{44}$$

式中 $b_3 = b_3'/N$．无压缩时的副帧长度为

$$M_3 = A_3W + b_3. \tag{45}$$

得压缩比为

$$CR_3 = \frac{M_3}{\bar{M}_3} = \frac{N(A_3W + b_3)}{NA_3pB_3 + b_3'}. \tag{46}$$

当 $p = 0$ 时，得最大压缩比

$$CR_{3\max} = \frac{N(A_3W + b_3)}{b_3'}. \tag{47}$$

当 $p = 1$ 时，得最大膨胀比为

$$ER_{3\max} = \frac{NA_3B_3 + b_3'}{N(A_3W + b_3)}. \tag{48}$$

图 4 给出了由 (28)，(36) 和 (44) 式计算的三种格式帧长度 M 与 q 之间的函数关系．为了验证理论的正确性，我们在 IBM-PC 机上进行了模拟，所得结果和理论值很一致．图中仅给出了游程法的模拟结果．

图 5 给出了由 (31)，(38) 和 (46) 式计算的 CR 和 q 之间的关系曲线．可以看出，当 q 较大，约大于 0.9 时，采用游程法不如采用绝对地址法．当 q 进一步增大，则用主副帧结构更加有效．当 q 小到 0.5 左右时，采用绝对地址法，CR 已降到 1 左右，起不到压缩作用．再进一步减少 q，就出现膨胀．而游程法直到 $q = 0.4$ 时，仍有压缩作用，即仍有 $CR > 1$．

本文给出的分析方法同样适用于文献 [10] 给出的变帧长和变字长格式．此外，对于定时不稳的系统，同样可采用多比特窗口技术加以解决[8]．

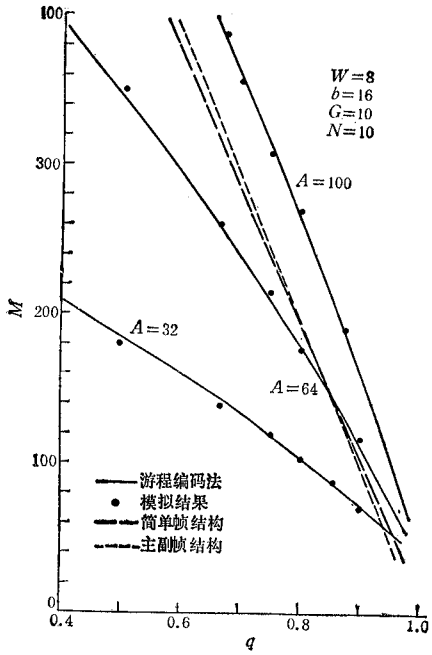


图 4　三种格式的帧长度M与q之间的函数关系

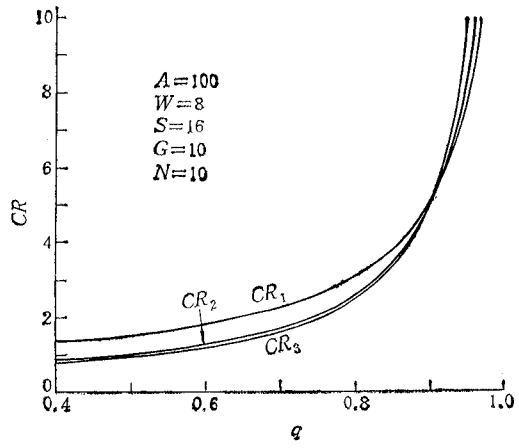


图 5　三种格式的压缩比 CR 与q之间的函数关系

三、结　束　语

一种输出格式是否最佳，与应用的条件有关．本文在分析讨论了多种格式的基础上，对三种最好格式的性能作了计算比较．一般来说，在 p 较小，$E(A)$ 不大，而 p_e 较大的场合，可用绝对地址法，相反用游程法．

在计算机技术得到广泛应用的今天，收发两地采用计算机来实现同步捕获是非常有利的．这样不仅能容易地产生各种复杂的输出格式，而且译码也相当容易．同时也便于采用最优化的数据压缩技术和缓冲器的自动控制．

参　考　文　献

[1] S. A. Sheldahl, Channel identification coding for data compressors, EASCON, Washington D. C., Sept. 1968, p.319.

[2] G. Held, Data compression, Chichester, Wiley, 1933, p.18.

[3] M. F. Lynch, Compression of biliographic files using an adaptation of run-length coding, Information Storage and Retrieval, 9(1973), 207.

[4] W. M. Rathbone, Output format coding data compressors, Proc. National Telemetering Conference, Los Angeles, CA, pp. 117—121, 1970.

[5] J. R. Hulme and R. A. Schomburg, A data bandwidth compressor for space vehicle telemetry, Proc. National Telemetering Conference, Washington D. C., May 1962, paper 3—2.

[6] A. G. Carlton, H. D. Zink, R. L. Appel, and G. P. Gafke, A real-time data compression system for Apollo PCM telemetry, AD-707333, 1969.

[7] 隋厚棠，电子科学学刊，**6**(1984)，p. 1.

[8] 隋厚棠，黄明亚，陈小敏，电子学报，1985 年，第 6 期，第 46 页．

[9] H. N. Massey, An experimental data compressor, Proc. National Telemetering Conference, Houston, Texas, April 1965, pp. 25—28.

[10] B. Maglaris and M. Schwartz, *IEEE Trans. on COM*, **COM-39**(1981), 800.

DATA COMPRESSOR OUTPUT FORMATS AND THEIR PERFORMANCES

Sui Houtang

(*Space Science and Technology Center, Academia Sinica*)

In order to get better output formats for data compressors, the synchronization, efficiency and complexity of encoding and decoding, and the spreading range of error generated in transmission are analysed and discussed. Then theoretical analyses of the preformances of several typical output formats are carried out. Finally, some simulation results are given.