2.8 信息冗余度与自然语言的熵

2.8.1 信源的相关性

1. 含义：信源符号间的依赖程度

2. $H_0 \geq H_1 \geq H_2 \geq H_{m+1} \geq \cdots \geq H_\infty$

   ~~~ 称为有限信源

   符号相关性越大，信源熵越小

2.8.2 信源冗余度

对一般离散平稳信源，H_∞就是实际信源熵。理论上只要有传送H_∞的手段，就能把信源包含的信息全部发送出去。但实际上确定H_∞非常困难，只好用H_{m+1}来代替。$H_{m+1} > H_\infty$，所以在传输手段上必然富裕，这样做实际上就是一种浪费，特别是有时只能得到H_1，甚至H_0，非常不经济。

这种浪费是由于实际信源符号具有无限记忆长度的相关性，而又没有求得H_∞的手段。

1. 信源冗余度：$R = 1 - \frac{H_\infty}{H_0}$
2. 熵的相对率：$\eta = \frac{H_\infty}{H_0}$
3. 自然语言的熵：

英文：$H_0 = \log 27 = 4.76$ bit/符号

认为独立：$H_1 = -\sum_{i=1}^{27} p(x_i) \log p(x_i) = 4.03$ bit/sym

视为马尔可夫信源：$H_2 = 3.32$

$H_3 = 3.1$

...

$H_\infty = 1.4$

$R = 1 - \frac{1.4}{4.76} = 71\%$

- 写英语文章时，71%是由语言结构定好的，只有29%是写文字的人可以自由选择的。100页的书，大约只传输29页就可以了，其余71页可以压缩掉。信息的冗余度表示信源可压缩的程度。
- 从提高传输效率的观点出发，总是希望减少或去掉冗余度。但冗余度大的消息抗干扰能力强。能通过前后符号间的关联关系纠正错误。
   ~~~