最优前缀码

二元前缀码

- 二元前缀码

  用0-1字符串作为代码表示字符，要求任何字符的代码都不能作为其它字符代码的前缀

- 非前缀码的例子

  ***a*: 001, *b*: 00, *c*: 010, *d*: 01**

- 解码的歧义，例如字符串 0100001

  解码1: 01, 00, 001 *d, b, a*

  解码2: 010, 00, 01 *c, b, d*

前缀码的二叉树表示

前缀码：

{00000, 00001, 0001, 001, 01,100,101,11}

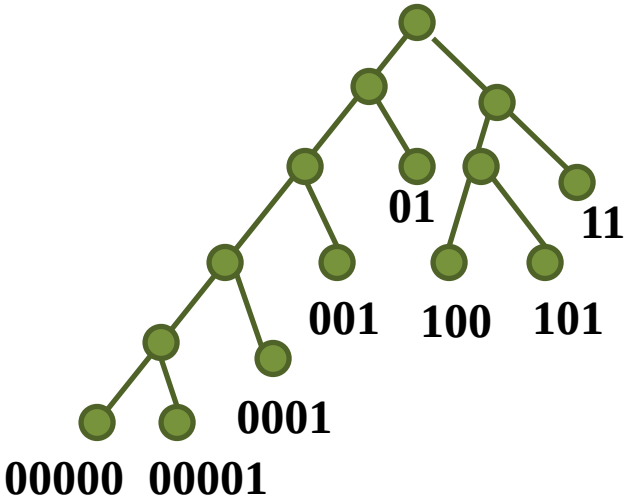
构造树：

0 -左子树

1-右子树

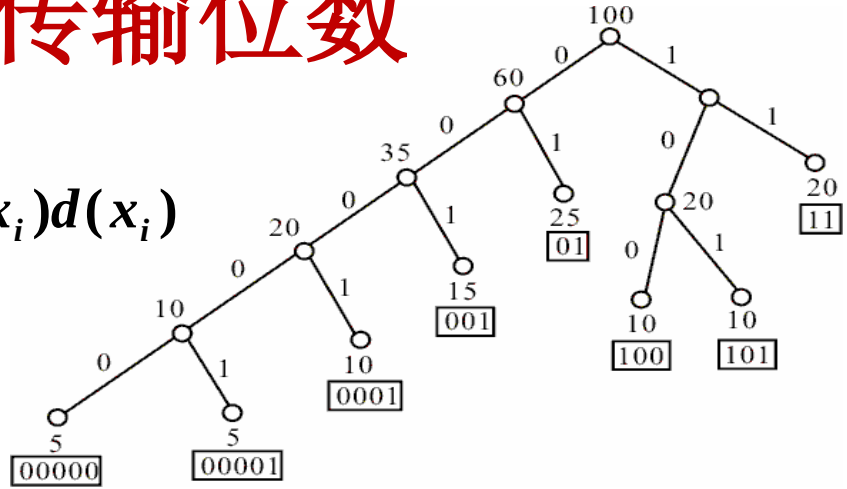
码对应一片树叶

最大位数为树深



平均传输位数

$$B=\sum_{i=1}^{n} f(x_i)d(x_i)$$



B=[(5+5)×5+10×4+(15+10+10)×3

+(25+20)×2]/100=2.85

问题：给定字符集$C=\{x_1,x_2,...,x_n\}$和每个字符的频率$f(x_i)$，$i=1,2,...,n$. 求关于C的一个最优前缀码(平均传输位数最小).

哈夫曼树算法伪码

算法 **Huffman(*C*)**

输入：$C=\{x_1, x_2,\ldots, x_n\}$, $f(x_i)$, $i=1,2,\ldots,n$.

输出：*Q*　　/ /队列

1. $n \leftarrow |C|$
2. $Q \leftarrow C$　　//频率递增队列*Q*
3. for $i \leftarrow 1$ to $n-1$ do
4. 　　*z*←Allocate-Node()　//生成结点 *z*
5. 　　*z.left*←*Q*中最小元　//最小作*z*左儿子
6. 　　*z.right*←*Q*中最小元　//最小作*z*右儿子
7. 　　$f(z) \leftarrow f(x)+f(y)$
8. 　　Insert(*Q*,*z*)　　// 将 *z* 插入*Q*
9. return *Q*

实例

输入 *a*:45; *b*:13; *c*:12; *d*:16; *e*:9; *f*:5

编码：

f--0000,

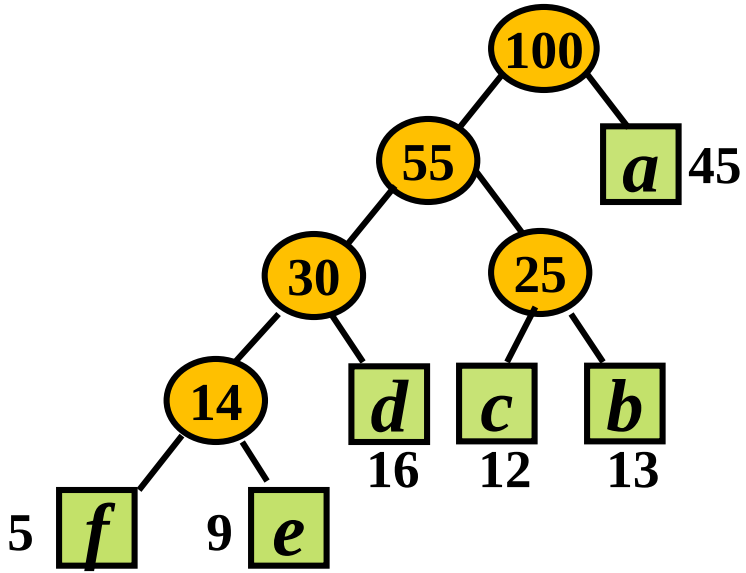
e--0001,

d--001,

c--010,

b—011,

a--1



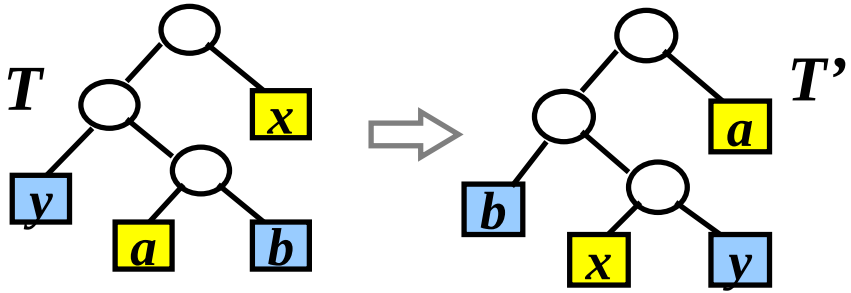
平均位数：

4×(0.05+0.09)+3×(0.16+0.12+0.13)+1×0.45=2.24

最优前缀码性质：引理1

引理1: C是字符集，$\forall c \in C$, $f(c)$为频率, $x,y \in C$, $f(x)$, $f(y)$频率最小, 那么存在最优二元前缀码使得 x, y 码字等长且仅在最后一位不同.

$f(x) \le f(a)$
$f(y) \le f(b)$



$$B(T) - B(T') = \sum_{i \in C} f[i] d_T(i) - \sum_{i \in C} f[i] d_{T'}(i) \ge 0$$

其中$d_T(i)$为i 在T 中的层数(i到根的距离)

引理2

引理 设T是二元前缀码的二叉树,$\forall x,y \in T$, x,y是树叶兄弟，z 是 x,y的父亲, 令

$$T' = T - \{x, y\}$$

且令 z 的频率

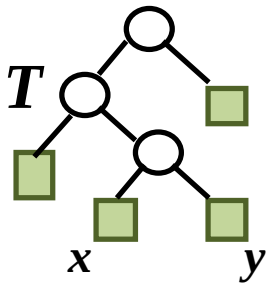
$$f(z) = f(x)+f(y)$$

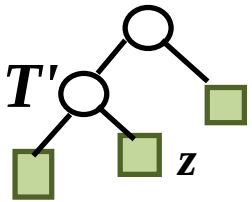
T'是对应二元前缀码

$$C' = (C - \{ x, y \}) \cup \{ z \}$$

的二叉树，那么

$$B(T)=B(T')+f(x)+f(y)$$





引理2证明

证 $\forall c \in C-\{x,y\}$,有

$d_T(c) = d_{T'}(c) \Rightarrow f(c)d_T(c) = f(c)d_{T'}(c)$

$d_T(x) = d_T(y) = d_{T'}(z) + 1$

$$B(T) = \sum_{i \in T} f(i)d_T(i)$$

$$= \sum_{i \in T, i \neq x,y} f(i)d_T(i) + f(x)d_T(x) + f(y)d_T(y)$$

$$= \sum_{i \in T', i \neq z} f(i)d_{T'}(i) + f(z)d_{T'}(z) + (f(x) + f(y))$$

$$= B(T') + f(x) + f(y)$$

小结

- 二元前缀码及其二叉树表示
- 给定频率下的平均传输位数计算公式
- 最优前缀码——平均传输位数最少
- 哈夫曼算法
- 前缀码的性质