5.二叉树

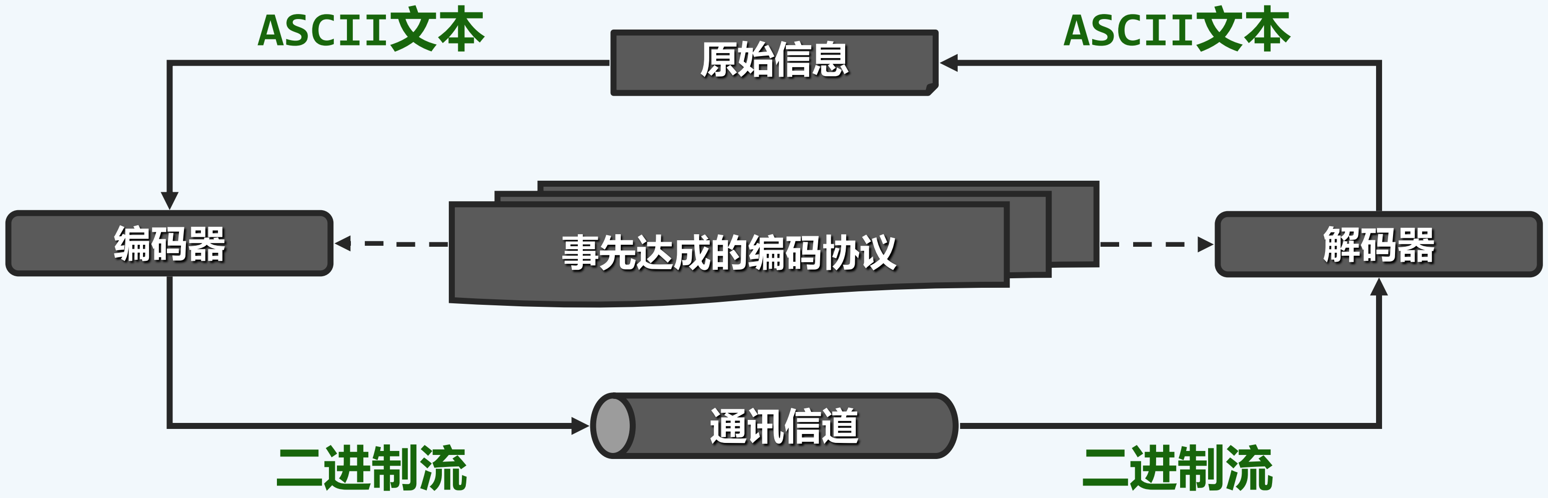
(f) PFC编码

邓俊辉

deng@tsinghua.edu.cn

问题

❖ 通讯 / 编码 / 译码



❖ 二进制编码

组成数据文件的字符来自字符集Σ

字符被赋予互异的二进制串

"1 010 011 00" = "MAIN"

M	A	I	N
1	010	011	00

❖ 文件的大小取决于

字符的数量 × 各字符编码的长短

❖ 通讯带宽有限时

如何对各字符编码，使文件最小？

二叉编码树

❖ 将Σ中的字符组织成一棵二叉树

以$\boxed{0}$/$\boxed{1}$表示$\boxed{左}$/$\boxed{右}$孩子

各字符$\boxed{x}$分别存放于对应的叶子$\boxed{v(x)}$中

❖ 字符x的编码串rps(v(x)) = rps(x)

由根到v(x)的通路（root path）确定

❖ 优点：字符编码不必等长，且

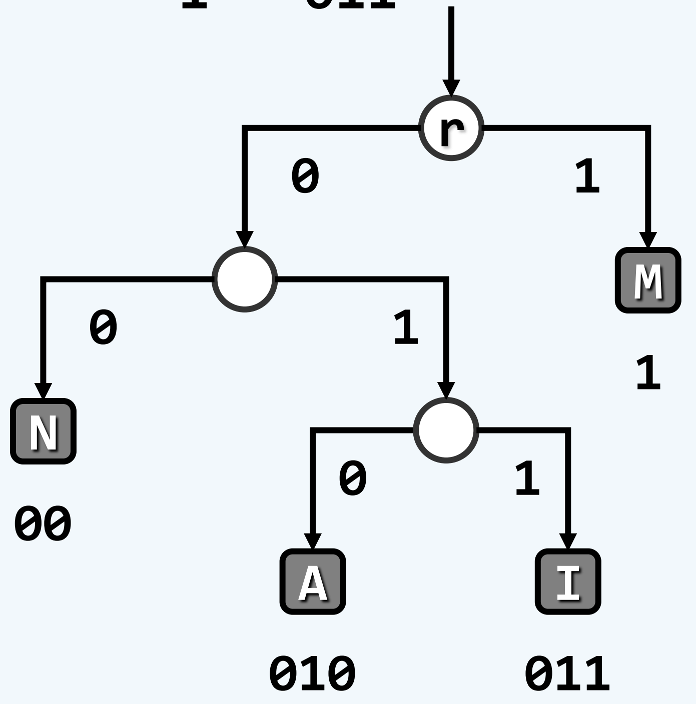
不致出现解码歧义

❖ 这属于“前缀无歧义”编码（prefix-free code）

不同字符的编码$\boxed{互不为前缀}$，故不致歧义

❖ 缺点：你能发现吗？

"$_{1}{}^{010}{}_{011}{}^{00}$" = "MAIN"



编码长度 vs. 叶节点平均深度

- $|rps(x)| = depth(v(x))$
- 编码总长 $= \Sigma_x depth(v(x))$

  平均编码长度

  $= \Sigma_x depth(v(x))/|\Sigma|$

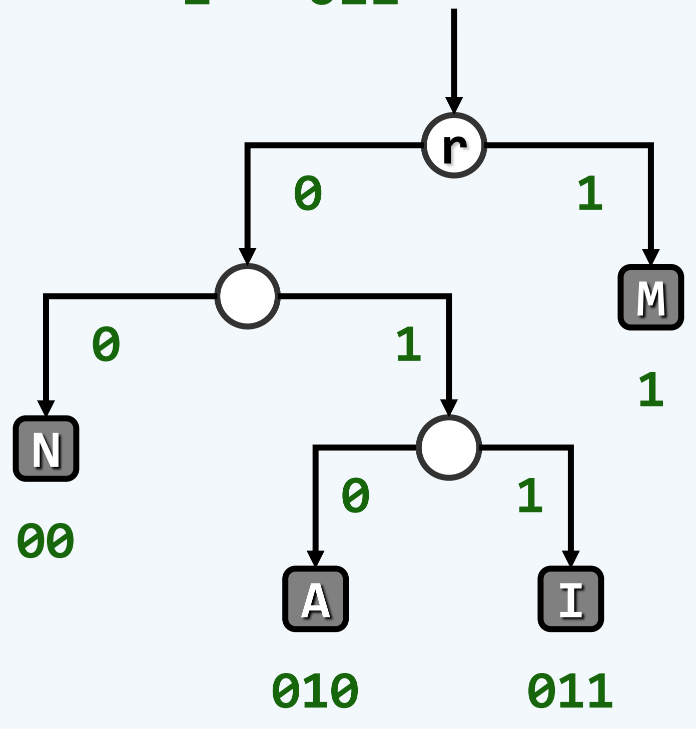
  $=$ 叶节点平均深度ald(T)
- 对于特定的Σ

  ald()最小者即为最优编码树T_{opt}
- 最优编码树必然存在，但不见得唯一

  它们具有哪些特征？

ald(T)*4 = 2+3+3+1 = 9

"$1^{010}{}_{011}{}^{00}$" = "MAIN"



最优编码树

- $\forall\ v \in T_{opt},\ deg(v) = 0$ $\boxed{\text{only if}}$ $depth(v) \geq depth(T_{opt})-1$

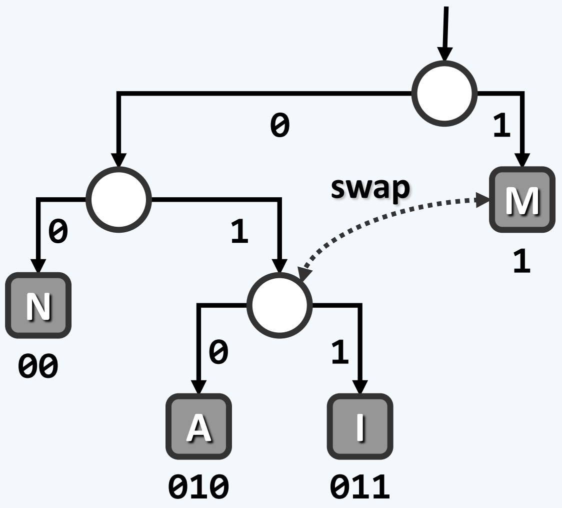
  亦即，叶子只能出现在倒数两层内——否则，通过节点交换即可...

- 特别地，真完全树即是最优编码树

- 实际上，字符的出现频率不尽相同，例如w('E') >> w('Z')

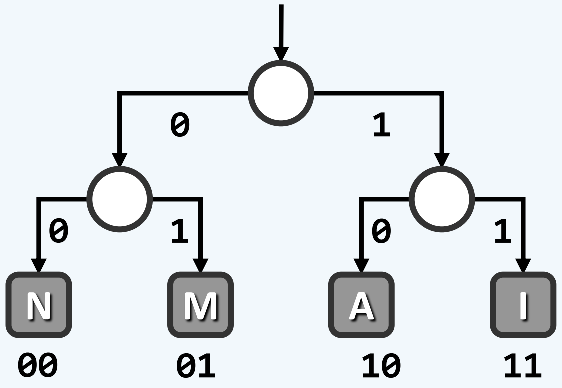
ald*4 = 2+3+3+1 = 9

$"1^{010}011^{00}"$ = "MAIN"



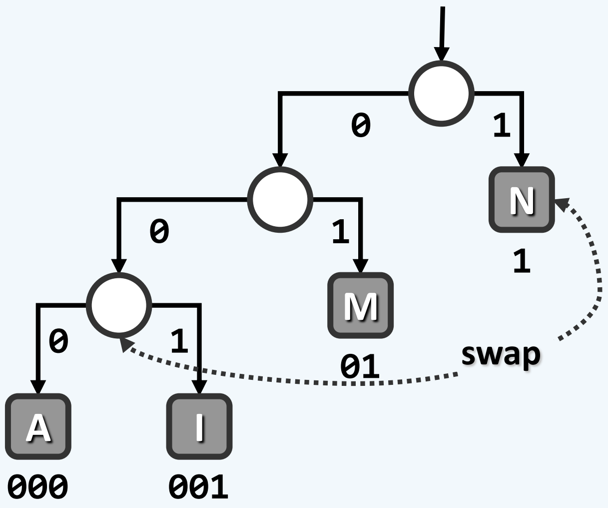
ald*4 = 2+2+2+2 = 8

$"01^{10}11^{00}"$ = "MAIN"



ald*4 = 2+3+3+1 = 9

$"01^{000}001^{1}"$ = "MAIN"



带权编码长度 vs. 叶节点平均带权深度

❖ 已知各字符的 期望频率 ，如何构造最优编码树？

❖ 文件长度 $\propto$ 平均带权深度

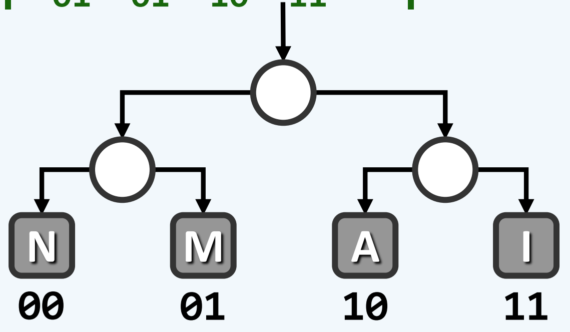
$= \mathrm{wald}(T) = \Sigma_x\, \mathrm{rps}(x) \times w(x)$

❖ 此时，完全树 未必 就是最优编码树

比如，考查"mamani"和"mammamia"...

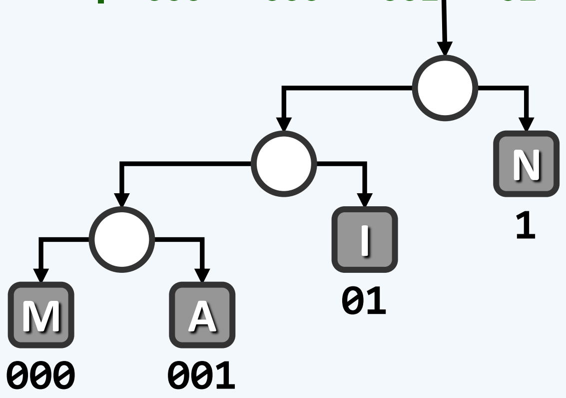
$|\text{"}01^{10}01^{10}00^{11}\text{"}| = 12$

$|\text{"}01^{10}01^{01}10^{01}11^{10}\text{"}| = 16$



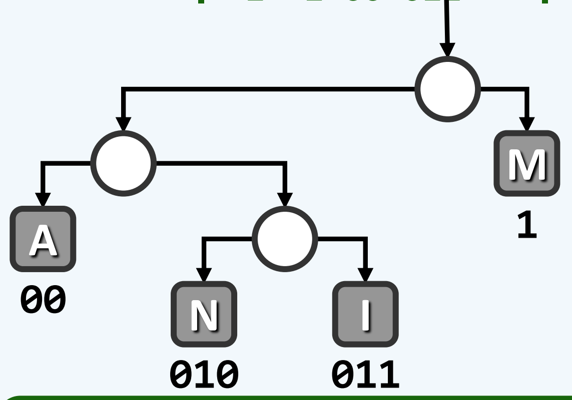
$|\text{"}000^{001}000^{001}1^{01}\text{"}| = 15$

$|\text{"}000^{001}000^{000}001^{000}01^{001}\text{"}| = 23$



$|\text{"}1^{00}1^{00}010^{011}\text{"}| = 12$

$|\text{"}1^{00}1^{1}00^{1}011^{00}\text{"}| = 13$



最优编码树

- ❖ 同样，频率$\boxed{高}$/$\boxed{低}$的（超）字符，应尽可能放在$\boxed{高}$/$\boxed{低}$处
- ❖ 故此，通过$\boxed{交换}$，同样可以缩短wald(T)