05-J1

5.二叉树

Huffman编码树

PFC编码

句读之不知，惑之不解，或师焉，或不焉，小学而大遗，吾未见其明也

邓俊辉

deng@tsinghua.edu.cn

应用

❖ 通讯 / 编码 / 译码

❖ 二进制编码

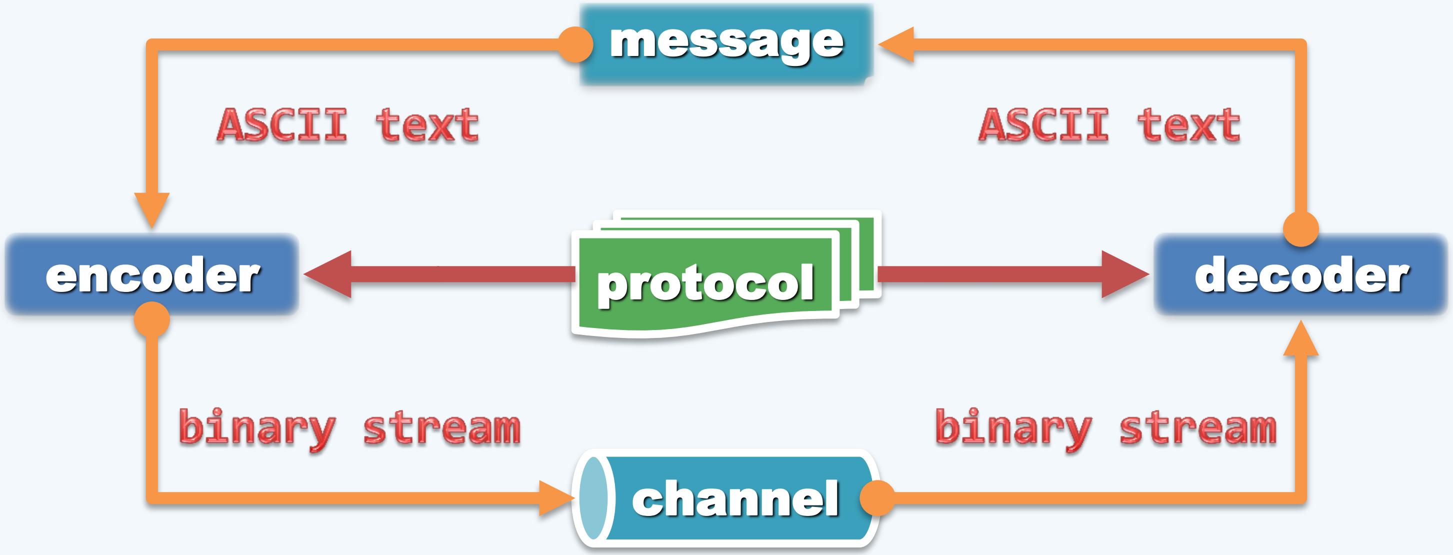
- 组成数据文件的字符来自字符集Σ
- 字符被赋予互异的二进制串

❖ 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？



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

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 $\boxed{\text{左}}$ / $\boxed{\text{右}}$ 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

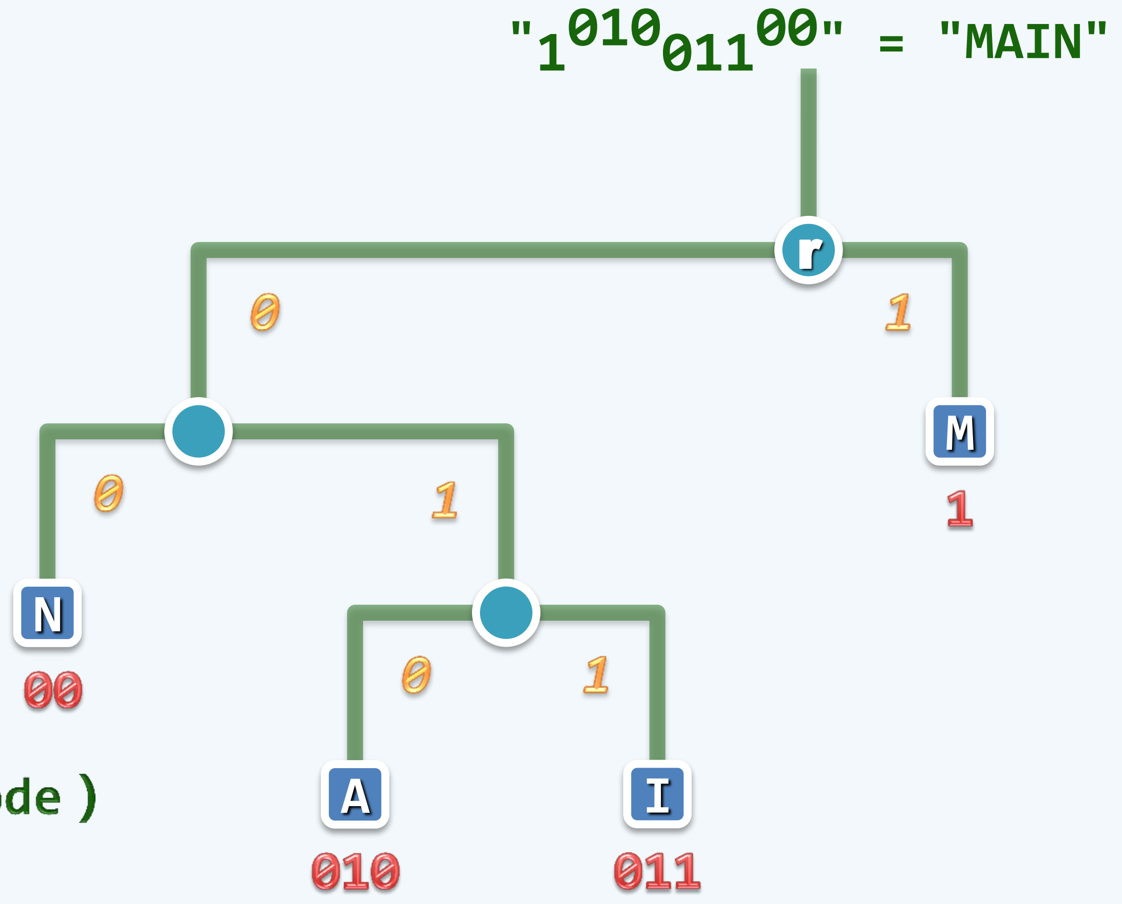
❖ 优点：字符编码不必等长，且

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 $\boxed{\text{互不为前缀}}$ ，故不致歧义

❖ 缺点：你能发现吗？



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))

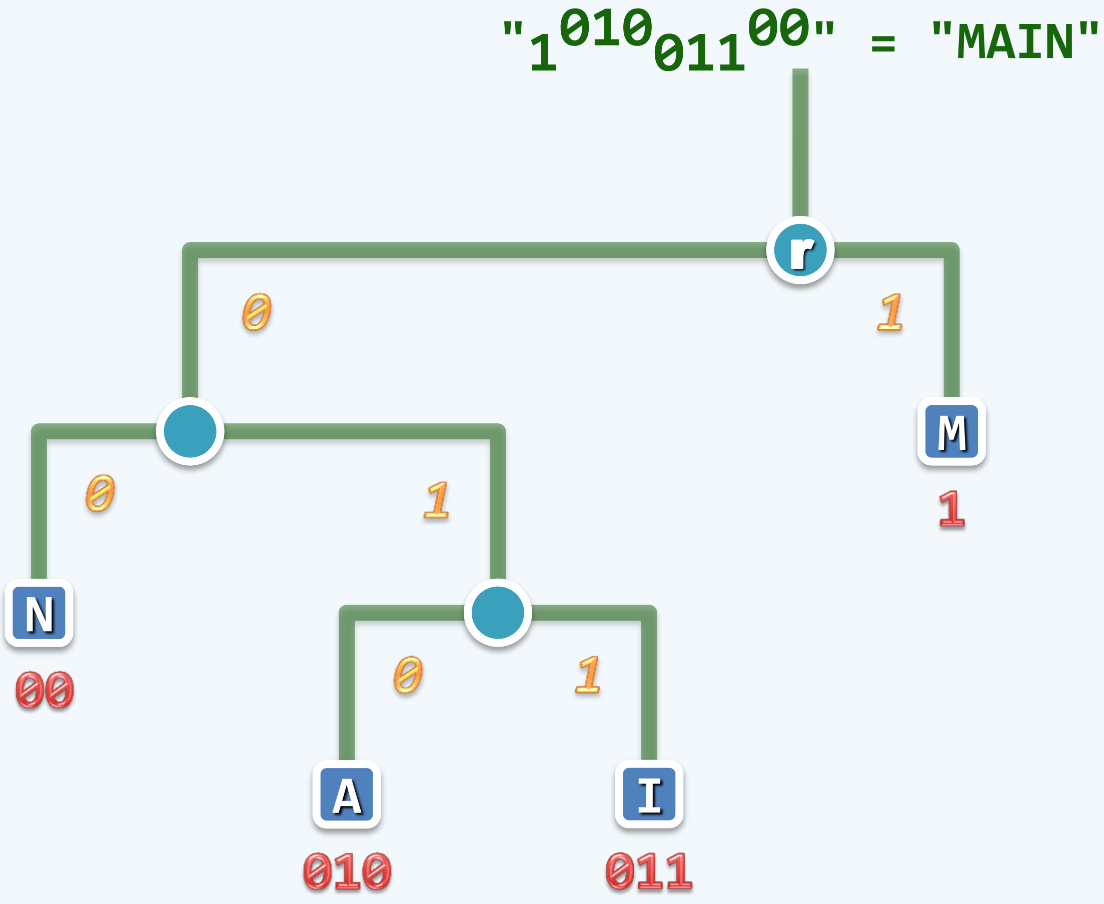
  平均编码长度

  = $\Sigma_x$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？



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...

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

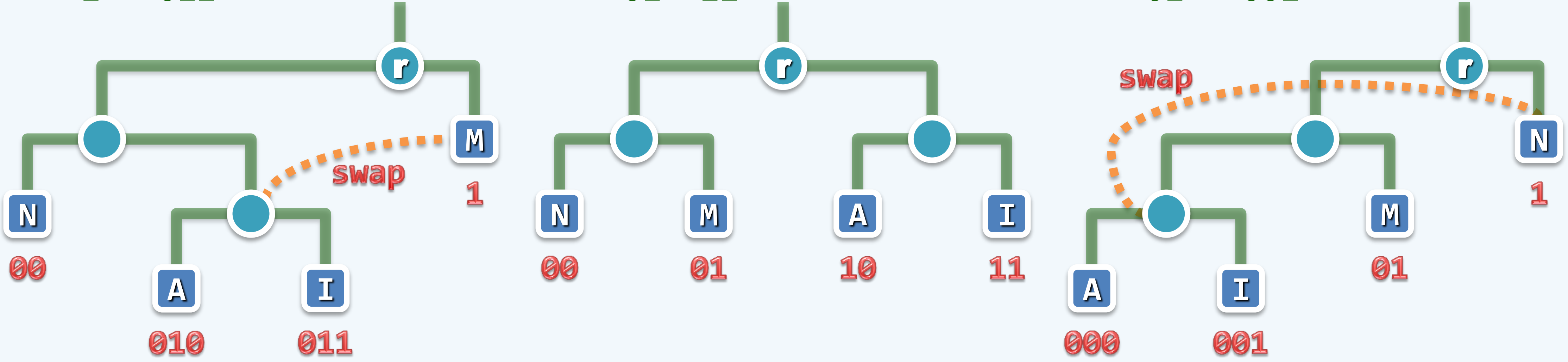
"$1^{010}011^{00}$" = "MAIN"

ald(T)*4 = 2+2+2+2 = 8

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

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

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



❖ 特别地，真完全树即是最优编码树

出现频率

❖ 实际上，字符的出现频率不尽相同，且往往相差极大...

序	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
字	不	人	山	无	风	一	日	云	有	何	来	天	中	时	花	上	水	春	月	相	年	为	生	君	长	心	自	如	知	白	归	秋
次	26482	20995	16080	15823	15718	15272	14984	13421	12701	12379	12315	12202	11459	11377	11367	11148	11097	11096	11004	9666	9562	9506	9428	9325	9185	9078	9025	8751	8677	8530	8364	8121

❖ 已知各字符的期望频率，如何构造最优编码树？

带权编码长度 vs. 叶节点平均带权深度

❖ **文件长度 $\propto$ 平均带权深度**

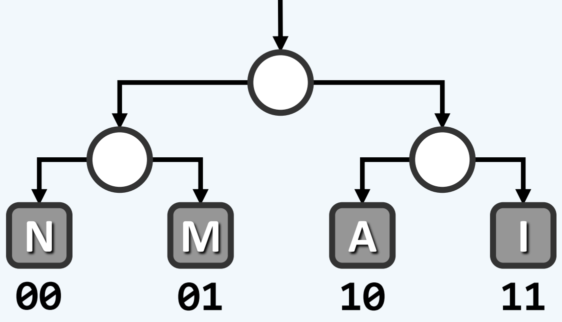
$= \mathrm{wald}(T) = \Sigma_x \, \mathrm{rps}(x) \times w(x)$

❖ **此时，完全树$\boxed{\text{未必}}$就是最优编码树**

比如，考查"mamani"和"mammamia"...

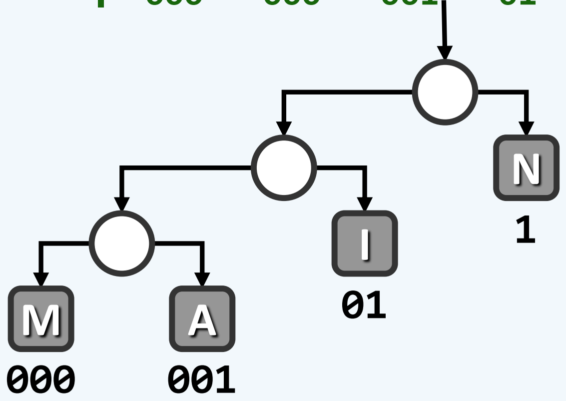
$|\text{"}{}_{01}{}^{10}{}_{01}{}^{10}{}_{00}{}^{11}\text{"}| = 12$

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



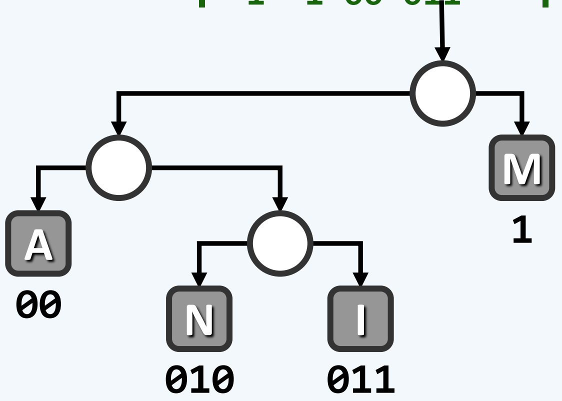
$|\text{"}{}_{000}{}^{001}{}_{000}{}^{001}{}_{1}{}^{01}\text{"}| = 15$

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



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

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



最优带权编码树

❖ 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)