字符串相关问题

西安交通大学少年班 王袤广

- 模式串匹配算法 KMP
- 待匹配串T(长度为m)，模式串S(长度为n)
- 普通算法复杂度$O(mn)$
- 在随机数据情况下表现好过KMP算法
- KMP算法通过$next$数组将复杂度降至读入复杂度$O(m + n)$
- $next[i]$表示前i个字符组成的子串中最长的满足后缀与前缀相等的后缀

- NOI2014 D2T1
- 对于串T，$num[i]$表示$T[0\sim i]$有多少个后缀与前缀相等且后缀与前缀不覆盖.
- 即对于$T[0\sim i]$，有多少个j使得$T[0\sim j]$与$T[i-j+1\sim i]$相等且$j\le\frac{i}{2}$
- $n\le 10^6$

- 先忽略$j \le \frac{i}{2}$这个条件
- 显然，对于长度为i的前缀，$next[i]$是最长的满足条件的子串，$next[next[i]]$是次长的，$next[next[next[i]]]$是第三长的，以此类推
- $num[i]$即为$next[next \cdots next[i] \cdots] = 0$ （经过$num[i]+1$）次迭代
- 直接递推即可$O(n)$

- 处理$j \le \frac{i}{2}$
- 直接倍增即可$O(n \log n)$ 需注意常数
- $next[i] -> i$建树，$num[i]$即为i到根节点路径上编号$\le \frac{i}{2}$的结点个数
- DFS用树状数组维护即可$O(n \log n)$比起倍增常数小
- $cnt[i]$长度为i的前缀最多经过的迭代次数，再次匹配一次利用cnt数组求解，时间复杂度$O(n)$

- 对于多个待匹配串，需在线性时间内求出模式串是否在待匹配串中出现
- 引入Trie树（字母树）
- 线性时间复杂度，线性空间复杂度

- 线性时间多模式串匹配：AC自动机
- 1.建立Trie树，所有模式串建立Trie树
- 2.构造Fail指针，类似KMP的方法构造失配指针
- 3.匹配主串，类似KMP的方法匹配
- 整体与KMP算法很像，相当于KMP是链结构，AC自动机是树结构

- HDU 2222
- 给定多个模式串，求有多少个模式串在主串中出现
- $N \leq 10000, |S| \leq 50, |T| \leq 1000000$

- 裸AC自动机
- 时间复杂度$O(N|S| + |T|)$

- HDU 3065
- 对于多个模式串，求出每个模式串在主串中出现的次数
- $N \leq 1000, |S| \leq 50, |T| \leq 2000000$

- 与上一题区别不大，加个数组统计即可
- 时间复杂度$O(N|S| + |T|)$

- POJ 2778
- 对于m个模式串，求长度为L的主串有多少个不包含任何模式串
- 模式串和主串均只包含A C G T四个字母，答案对10^5取模
- $m \leq 10, |S| \leq 10, L \leq 2 * 10^9$

- 将所有模式串建立AC自动机
- 题目即使求有多少个长度为n的主串在AC自动机上无法匹配成功
- 即，从根节点走n步，不经过加标记的点(经过代表匹配成功，注意如果一个节点的Fail指针指向的点加标记则这个点加标记)的路径条数

- 子问题
- 对于n个点的有向图，求i到j恰好走m步的路径条数
- $n \leq 100, m \leq 2 * 10^9$

- 建立邻接矩阵F
- 则$F^m[i][j]$即为答案
- 时间复杂度$O(n^3 \log m)$

- 回到原问题
- AC自动机相当于有向图，将所有加标记的点的入边出边删掉
- 建立邻接矩阵，原问题变为子问题
- $result = \sum F^L[root]$
- 时间复杂度$O((m|S|)^3 \log L)$

- HDU 2243
- 求有多少个长度小于等于L的主串，满足至少包含一个模式串
- 主串和模式串均只包含小写字母，答案对2^{64}取模
- $0 < N < 6, |S| \leq 5, 0 < L < 2^{31}$

- 设转移矩阵为F（即AC自动机）
- 则答案为$\sum 26^k - \sum\sum F^k[root]$
- $\sum 26^k$ 可用二分的方法求出
- $\sum\sum F^k[root]$即为$\sum\ (\sum F^k)\,[root]$可用二分的方法求出
- 时间复杂度$O((N|S|)^3 \log^2 L)$

- 根据矩阵的性质
- $\begin{vmatrix} F & 1 \\ 0 & 1 \end{vmatrix}^{L} = \begin{vmatrix} F^{L} & \sum_{i=1}^{L-1} F^{i} + 1 \\ 0 & 1 \end{vmatrix}$
- 其中1为和F同阶的单位矩阵，0为和F同阶的零矩阵
- 时间复杂度$O((N|S|)^3 \log L)$

- POJ 1625
- 求有多少种长度为M的主串满足不包含任何模式串，模式串和主串中仅可能出现N种字符
- $N \leq 50, M \leq 50, P \leq 10, P$为模式串个数

- 需要高精度
- DP转移
- $F[i][j]$表示长度为i的串在AC自动机上的j节点时的答案
- $F[i][j] = \sum F[i-1][k], flag[k]\&\&next[k][son] = i$
- 即k未标记且i为k的儿子节点

- HDU 2825
- 求长度为n的串至少包含k个模式串的方案数
- 模式串和主串均只包含小写字母，答案对20090717取模
- $n \leq 25, k \leq m \leq 10, m$为模式串个数

- 和上题类似，至少包含k个模式串这个条件我们可以将状态多加一维来解决
- 状态压缩动态规划
- $F[i][j][\text{state}]$表示长度为i的串在AC自动机上的j节点且包含模式串状态为state时的答案
- $F[i][j][state] = \sum F[i-1][k][state'], next[k][son] = i$
- 即i为k的儿子节点
- $O(nm^2 2^m)$

- ZOJ 3494
- 求在A和B之间有多少个数满足BCD编码不包含非法01串
- BCD编码即为十进制中每一位分别转换为二进制
- 答案对10^9+9取模
- $0 \le N \le 100, 0 < A \le B < 10^{200}, |S| \le 20, N$为非法01串个数