AC 自动机入门

郭炼

哈尔滨工业大学
计算学部

2020 年 8 月 10 号

哈爾濱工業大學

目录

AC 自动机

- ~~AC 自动机就是自动 AC 的机器。~~
- AC 自动机是一种基于 Trie 树，结合 KMP 思想的自动机。
- KMP 算法的模式串只能有一个，但是实际中我们可能需要进行多模式匹配。
- AC 自动机在 Trie 树的基础上为每个结点添加了一个 fail 指针，其定义与 KMP 中的 fail 指针类似。
- fail[x] 是满足 string (fail[x]) 和 string (x) 的后缀相等且长度最长的点。

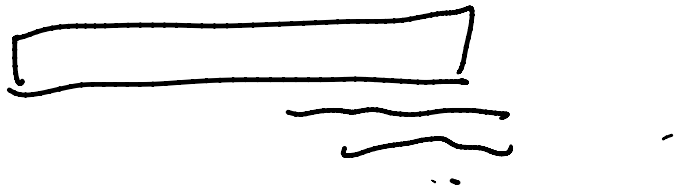
fail 指针构建思想

fail 指针的构建思想与 KMP 十分类似。考虑字典树中当前的结点 u，父结点是 p，p 通过字符 c 的边指向 u。假设深度小于 u 的所有结点的 fail 指针都已求得。

1. 若 next[fail[p], c] 存在，则让 u 的 fail 指针指向 next[fail[p], c]。
2. 若对应结点不存在，则沿着 fail 指针一路向上爬，直到存在。
3. 若到根结点都不存在，那么 fail 指针指向根节点。

时间复杂度为 $O(\sum |s_i|)$

fail 树



- AC 自动机的另一个重要的应用是关于其 fail 树。
- 由于 fail 指针总是指向深度更浅的结点，所以 fail 指针其实构成了一棵有向树。
- 沿着 fail 指针向上走，经过的点都是出发结点的后缀。
- 所以 u 的子树大小就是 string (u) 在模式串中出现的次数。
- 利用 fail 树，我们可以干更多的事情，比如 DP，计数，~~AC 自动机 fail 树 dfs 序建可持久化线段树~~等。

Trie 图

Trie 图是 AC 自动机的确定化形式，即把每个结点不存在字符的 next 指针使用 fail 指针补全了。这样匹配时就可以直接转移而不用沿着 fail 指针爬，而是直接转移了。

Trie 图和 AC 自动机本质相同。

单词

TJOI 2013 单词

给定一个由 n 个字符串构成的文章，询问每个单词在文章中共出现了多少次。

$n \leq 200, \sum |s_i| \leq 10^6$

单词

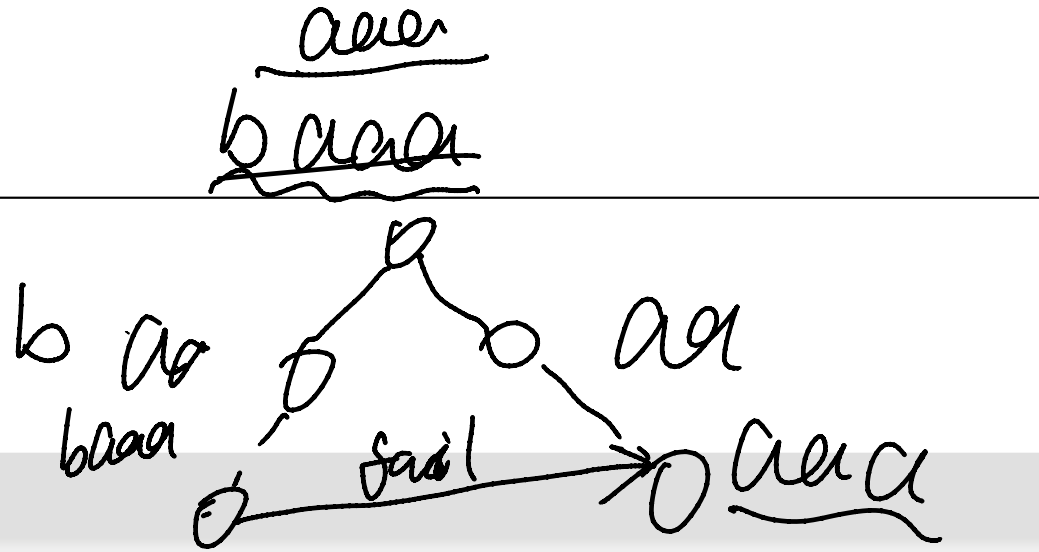
TJOI 2013 单词

给定一个由 n 个字符串构成的文章，询问每个单词在文章中共出现了多少次。

$n \leq 200, \sum |s_i| \leq 10^6$

- 将字符串插入 Trie 树是为每个结点自加一。

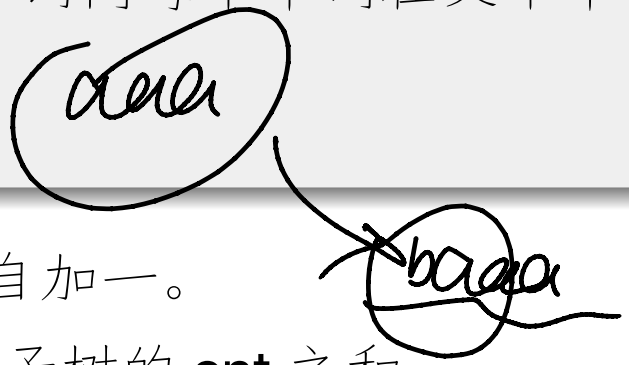
单词



TJOI 2013 单词

给定一个由 n 个字符串构成的文章，询问每个单词在文章中共出现了多少次。

$n \le 200, \sum |s_i| \le 10^6$



- 将字符串插入 Trie 树是为每个结点自加一。
- 每个字符串的答案，就是 fail 树对应子树的 cnt 之和。

病毒

POI 2000 病毒

n 个 01 病毒串，总长不超过 30000。问是否存在无限长的不包含病毒串的 01 串。

病毒

POI 2000 病毒

n 个 01 病毒串，总长不超过 30000。问是否存在无限长的不包含病毒串的 01 串。

- 如果不包含病毒串而且无限长也就是我们可以一直沿着 Trie 图跑。

病毒

POI 2000 病毒

n 个 01 病毒串，总长不超过 30000。问是否存在无限长的不包含病毒串的 01 串。

- 如果不包含病毒串而且无限长也就是我们可以一直沿着 Trie 图跑。
- 即存在一个环，使得环上不含病毒串（即终止结点）。

文本生成器

JSOI 2007 文本生成器

给定 n 个串，字符集为大写字母。问长度为 m 的字符串中，有多少个串包含至少一个给定串。答案对 10007 取模。

$n \le 60, m, |s_i| \le 100$

文本生成器

JSOI 2007 文本生成器

给定 n 个串，字符集为大写字母。问长度为 m 的字符串中，有多少个串包含至少一个给定串。答案对 10007 取模。

$n \le 60, m, |s_i| \le 100$

- 考虑没出现给定串的字符串。

文本生成器

JSOI 2007 文本生成器

给定 n 个串，字符集为大写字母。问长度为 m 的字符串中，有多少个串包含至少一个给定串。答案对 10007 取模。

$n \le 60, m, |s_i| \le 100$

- 考虑没出现给定串的字符串。
- 定义 $dp[i][j]$ 为前 i 个字符，其中最后一个字符落在自动机的 j 号节点上的非法串数目。

文本生成器

JSOI 2007 文本生成器

给定 n 个串，字符集为大写字母。问长度为 m 的字符串中，有多少个串包含至少一个给定串。答案对 10007 取模。

$n \le 60, m, |s_i| \le 100$

- 考虑没出现给定串的字符串。
- 定义 $dp[i][j]$ 为前 i 个字符，其中最后一个字符落在自动机的 j 号节点上的非法串数目。
- $dp[i][j] = \sum dp[i-1][k], (k, j) \in$ Trie 图。

String

2017 Multi-University Training Contest String

有 n 个由小写字母构成的串 W_i，现在有 q 次询问，每次给一个前缀 P_i 和后缀 S_i，求这 n 个串中有多少个串满足给的前缀和后缀（前缀和后缀不能在这个字符串中重叠）?

$$0 < n, q \leq 100000, \sum |S_i| + |P_i| \leq 500000, \sum |W_i| \leq 500000$$

String

2017 Multi-University Training Contest String

有 n 个由小写字母构成的串 W_i，现在有 q 次询问，每次给一个前缀 P_i 和后缀 S_i，求这 n 个串中有多少个串满足给的前缀和后缀（前缀和后缀不能在这个字符串中重叠）?

$$0 < n, q \leq 100000, \sum |S_i| + |P_i| \leq 500000, \sum |W_i| \leq 500000$$

- 离线处理。

String

2017 Multi-University Training Contest String

有 n 个由小写字母构成的串 W_i，现在有 q 次询问，每次给一个前缀 P_i 和后缀 S_i，求这 n 个串中有多少个串满足给的前缀和后缀（前缀和后缀不能在这个字符串中重叠）？

$$0 < n, q \leq 100000, \sum |S_i| + |P_i| \leq 500000, \sum |W_i| \leq 500000$$

- 离线处理。
- 将询问做成 s + # + p 的模式串插入到 AC 自动机里。
- 将每一个原串 w 做成 w + # + w 的串，用 AC 自动机进行匹配，沿着 fail 指针更新答案。

String

2017 Multi-University Training Contest String

有 n 个由小写字母构成的串 W_i，现在有 q 次询问，每次给一个前缀 P_i 和后缀 S_i，求这 n 个串中有多少个串满足给的前缀和后缀（前缀和后缀不能在这个字符串中重叠）？

$$0 < n, q \leq 100000, \sum |S_i| + |P_i| \leq 500000, \sum |W_i| \leq 500000$$

- 离线处理。
- 将询问做成 s + # + p 的模式串插入到 AC 自动机里。
- 将每一个原串 w 做成 w + # + w 的串，用 AC 自动机进行匹配，沿着 fail 指针更新答案。
- 为了防止 aaa 匹配上 aa + aa，还需要判断长度。