略谈 KMP

作者：cms42
原文链接：https://ld246.com/article/1563949376229
来源网站：链滴

注：本文字符串下标从0开始

略谈KMP

啥是KMP?

在计算机科学中，Knuth-Morris-Pratt字符串查找算法（简称为KMP算法）可在一个主文本字符串S查找一个词W的出现位置。此算法通过运用对这个词在不匹配时本身就包含足够的信息来确定下一个配将在哪里开始的发现，从而避免重新检查先前匹配的字符。

简而言之，KMP用来在一个字符串S中查找所有子串W的位置

比如，在字符串S=kmp-is-used-to-ak-ioi 中，子串W=i 出现了三次：

kmp-\color{blue}{i}s-used-to-ak-\color{blue}{i}o\color{blue}{i}

而子串W'=-i 一共出现了两次：

kmp\color{blue}{-i}s-used-to-ak\color{blue}{-i}oi

够简单吧

朴素做法

用两个指针i,j表示“当前正在处理S_i,W_j”，那么每次判断是否满足S_i=W_j，如果是那么i=i+1,j=j+1，处理下一位，否则i=i-j+1，j=0，相当于起点后移一位重复上述过程。

这种做法的时间复杂度如何呢？事实上，在随机数据测试时，朴素做法可以达到接近于O(|S|+|W|)的杂度，但如果碰上出题人恶意构造的数据，可能被卡成O(|S||W|) \color{red}{（设想S='AAA....AAB'} W='AAA...A'，|S|=10^6，|W|=10^5， 朴素算法的循环次数将非常可怕）

优化

考虑当S_i \neq W_j时的情况。此时其实有隐含条件：S_{i-1}=W_{j-1} ， S_{i-2}=W_{j-2} , ... , S_{i-j}W_0 ，在朴素做法中，这些信息完全被浪费了

那么我们该怎么利用这些信息呢？很简单，定义数组Next（部分讲稿也有叫kmp或者f的），表示当S_i \neq W_j（称作“失配”）时，就直接j=Next[j]，相当于模式串W向右移动j-Next[j]

休息一下，马上回来

参考资料

维基百科

阮一峰的网络日志

《算法竞赛入门经典训练指南》