字符串哈希

kiri

2024 年 2 月 1 日

1 用途

快速判断两个子串是否相同，也可以判断一个字符串是否是另一个字符串的子串

2 基本思想

1. 先考虑字符串前缀哈希，因为知道前缀哈希可以推出任意子段的哈希。可以预处理出字符串所有前缀的哈希, 比如对于串 $'ABCABCDEFGHIJKLMN'$,$h[0] = 0$,$h[1] = hash('A')$ ($hash('A')$ 代表 $'A'$ 的哈希值) $h[2] = hash('AB')$,$h[3] = hash('ABC')\dots$

2. 哈希值怎么求? 比如:$'ABCD'$ 可以看做 p 进制下的一个数 $(ABCD)_p$ 转换成十进制就是 $(A \times p^3 + B \times p^2 + C \times p + D \times p^0)_{10}$ 又因为这个数可能有点大，所以我们要 mod 一个较小的数 Q 即:$(A \times p^3 + B \times p^2 + C \times p + D \times p^0)_{10} \mod Q$，一般我们取 p=131 或者 13331;Q 一般取 2^{64} 那么 mod Q 操作我们可以用 $unsignedlonglong$ 代替, 因为 $unsignedlonglong$ 爆了之后就相当于对 2^{64} 取模

   注意:

   - 字符串不能被映射为 0
   - 一般是不会存在冲突

3. 求出前缀哈希值后就可以求每个子串的哈希值了:

   - 假设我们要求 L 到 R 段的哈希值 (包含 L 和 R) 我们已知了整个子串 (1 到 n) 的前缀哈希值那么我们可以把左侧 (字符串开始的一端) 视作低位，右侧 (字符串结束的一端) 视作高位
   - 那么 $h[R]$ 就相当于一个 R 位的数 1 是第 $R-1$ 位，R 是第 0 位即 $h[R] = p^{R-1}\dots p^0$; 同理，$h[L-1] = p^{l-2}\dots p^0$
   - 那么我们算 L 到 R 的哈希值，先将 $h[L-1]$ 最高位移到和 $h[R]$ 最高位相同即 $h[L-1] \times p^{R-L+1}$
   - 所以 L 到 R 的哈希值就是 $h[R] - h[L-1] \times p^{R-L+1}$