Práce s řetězci - suffixové pole

9 December, 2021 10:42

Práce s řetězci

$T \in \Sigma^m$ ... text

$P \in \Sigma^n$ ... hledaný slovo

úkol: najdi všechny výskyty P v T.

Klasické řešení: Aho-Corasick – z P se sestrojí konečný automat, který vyhledá výskyty P v T.

čas $O(m+n)$.

typicky: T se mění (fixní databáze) a přijde řada dotazů $P_1, P_2, \ldots$ & $n \ll m$

→ chceme lepší řešení

sufixové pole

$T[1..m], T[2..m], T[3..m], \ldots T[m..m]$

... sufixy T

→ setřídíme lexikograficky

→ $S[1..m]$ $S[i] = j$

$T[j..m]$ je lexikograficky i-tý sufix.

$ koncový symbol, např. $\notin \Sigma$

Př: T = banan$

$T[1..m]$ = banan$ $T[2..m]$ = anan$

$T[3..m]$ = nan$

⋮

$T[m..m]$ = $

$S[1]$ = $ $S[2]$ = an$ $S[3]$ = anan$ $S[4]$ = banan$

$S[5]$ = n$ $S[6]$ = nan$

vyhledání P pomocí sufixového pole

- binárně vyhledej P v sufixovém poli

$O(n \cdot \lg m)$

- lze zlepšit na $O(n + \lg m)$ pomocí fce pro nejdelší společný prefix řetězců $T[1..m]$ a $T'[m..m]$

$lcp(a,b) = \max_k a[1..k] = b[1..k]$

- v čase $O(m \lg m)$ příp. $O(m)$ lze vybudovat strukturu, která zodpovídá dotazy

$LCP(i,j) = lcp(T[i..m], T[j..m])$

v čase $O(1)$.

Vybudování sufixového pole:

- lze vybudovat v čase $O(m \lg m)$ pomocí variant bucket-sortu [Karp-Miller-Rosenberg '72]
- $\lg m$ fází, po fázi i jsou sufixy $T[j..m]$ setříděny podle prvních 2^i znaků.
- fáze $i=0$: setřídím $T[j..j]$, sufix $T[j..m]$ dostane pořadové číslo $H[j]$ prvního výskytu znaku $T[j..j]$. // $T[j]$
- fáze $i+1$: $T[j..m]$ označím dvojicí $(H[j], H[j+2^i])$ setřídím získané dvojice změním pořadové číslo $T[j..m]$ na pořadové číslo dvojice $(H[j], H[j+2^i])$. pozor → nové $H[j]$.
- po fázi $i = \lg m$ mi $H[j]$ udává pořadové číslo $T[j..m]$ mezi sufixy

⇒ rank $T[j...m]$.

$R[j] := H[j]$ ... pořadové číslo,

$T[j..m]$ nějaký sufix

$S[R[j]] = j$ tj. $R[..]$ je inverz $S[...]$.

- z R lze snadno dopočítat S jedním průchodem
  pro j: R: for $j=1..m$: $S[R[j]] := j$.

- i-tou fází algoritmu lze implementovat pomocí
  bucket sortu / radix sortu v čase $O(m)$.
  → celkový čas $O(m \lg m)$ na spočítání $S[..]$.

LCP:
- pole $L[1..m]$; $L[i] = lcp(T[i], T[i+1])$.