

Minulý týždeň: Plnotextové vyhľadávanie kľúčových slov

Definícia problému

Dokument: Postupnosť slov

Cieľ: Vytvoriť index pre statickú množinu dokumentov, aby sme vedeli rýchlo odpovedať na dotazy.

Dotaz: Dané slovo w , nájdite všetky dokumenty obsahujúce w .

Príklad:

Dokument 0: Ema ma mamu.

Dokument 1: Mama ma Emu.

Dokument 2: Mama sa ma. Ema sa ma.

Dotaz: Mama vráti dokumenty 1,2.

Invertovaný index

Pre každé slovo zoznam výskytov (ID dokumentov)

ema: 0,2

emu: 1

ma: 0,1,2

mama: 1,2

mamu: 0

sa: 2

Rôzne implementácie: utriedené polia, bin. vyhľadávacie stromy,
hashovanie, lexikografické stromy, ...

Vyhľadávanie viacerých kľúčových slov

Nájdí prienik dvoch utriedených postupností dĺžok m a n

- Merge $O(m + n)$
- m -krát binárne vyhľadávanie $O(m \log n)$
- Doubling search $O(m \log \frac{n}{m})$

String matching (vyhl'adávanie vzorky v texte)

Vstup: vzorka (pattern) P dĺžky m , text T dĺžky n .

Úloha: Nájdí všetky indexy $\{i_1, i_2, \dots\}$ t.ž. $T[i_j..i_j + m - 1] = P$.

Príklad:

Vstup: $P = \text{ma}$, $T = \text{Ema ma mamu}$

Výstup: 1, 4, 7

Vstup: $P = \text{"a ma"}$, $T = \text{Ema ma mamu}$

Výstup: 2, 5

Triviálny algoritmus $O(mn)$

```
1  for ( i=0; i<=n-m; i++) {  
2      j=0;  
3      while ( j<m && P[j]==T[i+j] ) { // (*)  
4          j++;  
5      }  
6      if ( j==m ) {  
7          print( i );  
8      }  
9  }
```

Najhorší prípad: $P = a^m, T = a^n$

Nepovinná DÚ

Koľko najviac spraví triviálny algoritmus porovnaní,
ak sa $P[0]$ vyskytuje v P najviac k krát?

Funkcia parametrov k, m, n , pričom $1 \leq k \leq m \leq n$

Plán na dnes

- Dokončíme analýzu priemerného prípadu
- Použitie konečných automatov

Priemerný počet porovnaní triv. alg. pre $\Sigma = \{0, 1\}$

Priemer cez všetky možné vstupy:

$$\begin{aligned} & \frac{1}{2^{n+m}} \sum_{P \in \{0,1\}^m, T \in \{0,1\}^n} \sum_{i=0}^{n-m} c_i(P, T) \\ &= (n - m + 1) \left(\sum_{j=1}^{m-1} j/2^j + m/2^{m-1} \right) \\ &= (n - m + 1) \left(2 - 1/2^{m-1} \right) \\ &\leq 2(n - m + 1) \end{aligned}$$

Súčet geometrickej postupnosti: $\sum_{i=0}^n q^i = (1 - q^{n+1})/(1 - q)$

Pre $q = 1/2$ máme $2 - 1/2^n$

Priemerný počet porovnaní triv. alg. pre $\Sigma = \{0, 1\}$

Stredná hodnota počtu porovnaní na náhodne vygenerovanom vstupe

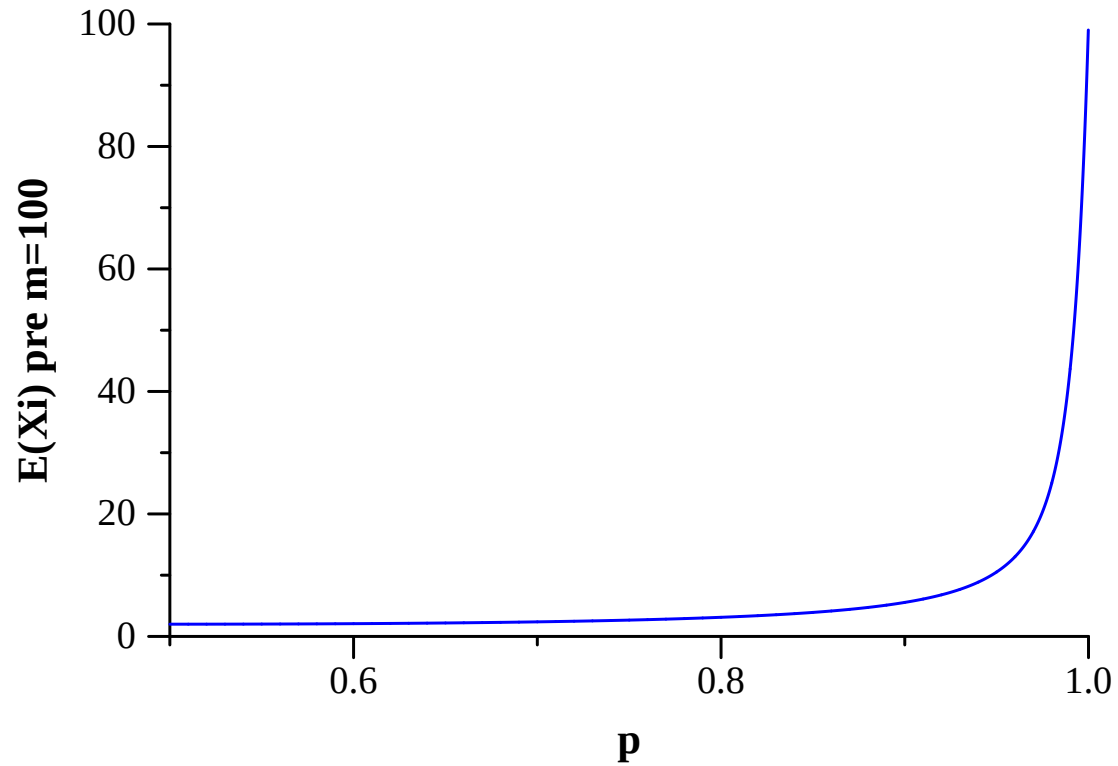
X – celkový počet porovnaní

X_i – počet porovnaní v iterácii i (t.j. keď $P[0]$ zarovnáme s $T[i]$)

Y_i – dĺžka najdlhšieho spoločného prefixu P a $T[i..n]$.

Hľadáme $E(X)$

**Středná hodnota počtu porovnaní,
ak $P(T[i] = 1) = P(P[i] = 1) = p$**



Simulácia nedeterministického konečného automatu na texte T

```
1  S = {0};
2  for(i=0; i<n; i++){
3      S1 = empty_set;
4      foreach state j in S {
5          add delta(j, T[i]) to S1;
6      }
7      if (m in S1){
8          print i-m+1;
9      }
10     S = S1;
11 }
```

Simulácia deterministického konečného automatu na texte T

```
1  state = 0;
2  for(i=0; i<n; i++){
3      state= delta(state ,T[i]);
4      if (state==m){
5          print i-m+1;
6      }
7  }
```