Datové struktury 1

10. přednáška 9.12.

Pavel Veselý

Plán: řetězcové a geometrické DS

- Konstrukce LCP pole ze sufixového v lineárním čase
- Konstrukce sufixového pole v lineárním čase (jen skeč, nezkouší se)
- FM-index: Ještě lepší textový index
  - Burrows-Wheelerova transformace + efektivní ranky na bitovým vektorem
- Geometrické DS: úvod

Sufixové pole X a spol.

- Udává lexikografické pořadí sufixů S
  - $X[i]$ = pozice začátku i-tého sufixu S v lex. pořadí
- Rankové pole R = inverzní permutace k X
  - $R[i]$ = kolikátý v lexikografickém pořadí je sufix $S[i:]$
- LCP pole L (pole společných prefixů, "longest common prefix")
  - $\mathrm{LCP}(\alpha, \beta)$ = nejdelší společný prefix slov α a β
  - $L[i] = \mathrm{LCP}(S[X[i]:], S[X[i+1]:])$
- Aplikace:
  - Vyhledávání výskytu jehly
  - Počet k-gramů = počet různých podslov délky k
  - Nejdelší opakující se podslovo
  - Nejdelší společné podslovo
  - …

i	$X[i]$	$R[i]$	$L[i]$	*suffix*
0	13	3	0	ε
1	1	1	5	arokoarokoko
2	6	12	0	arokoko
3	0	10	0	barokoarokoko
4	11	5	2	ko
5	4	8	2	koarokoko
6	9	2	0	koko
7	12	13	1	o
8	5	11	1	oarokoko
9	10	6	3	oko
10	3	9	3	okoarokoko
11	8	4	0	okoko
12	2	7	4	rokoarokoko
13	7	0	–	rokoko

Algoritmy pro sufixové pole

1. Výpočet sufixového pole X:
   - Tedy seřazení všech sufixů S
   a) Zdvojováním v čase $O(n \log n)$
      - Seřadíme dle prvních 2^i písmen pro $i = 0, 1, 2, \ldots, \lceil \log_2 n \rceil$
   b) Algoritmus Skew v čase $O(n)$ – pro abecedy, které lze setřídit v $O(n)$
      - [Kärkkäinen&Sanders '03]
2. Výpočet LCP pole ze sufixového pole X
   - Kasaiův algoritmus

i	$X[i]$	$R[i]$	$L[i]$	*suffix*
0	13	3	0	ε
1	1	1	5	arokoarokoko
2	6	12	0	arokoko
3	0	10	0	barokoarokoko
4	11	5	2	ko
5	4	8	2	koarokoko
6	9	2	0	koko
7	12	13	1	o
8	5	11	1	oarokoko
9	10	6	3	oko
10	3	9	3	okoarokoko
11	8	4	0	okoko
12	2	7	4	rokoarokoko
13	7	0	–	rokoko

Alg. Skew – lineární konstrukce sufixového pole

- Seřadíme 2/3 sufixů slova α rekurzivně
- Zbytek dotřídíme v lineárním čase

Doplnění: porovnávání sufixů pro slití $X^{(01)}$ a $X^{(2)}$

$$\begin{aligned}
\alpha_0[i:] &\leq \alpha_2[j:] \\
\alpha[3i:] &\leq \alpha[3j+2:] \\
(\alpha[3i], \alpha[3i+1:]) &\leq (\alpha[3j+2], \alpha[3j+3:]) \\
(\alpha[3i], \alpha_1[i:]) &\leq (\alpha[3j+2], \alpha_0[j+1:]) \\
(\alpha[3i], R_{01}[|\alpha_0|+i]) &\leq (\alpha[3j+2], R_{01}[j+1])
\end{aligned}$$

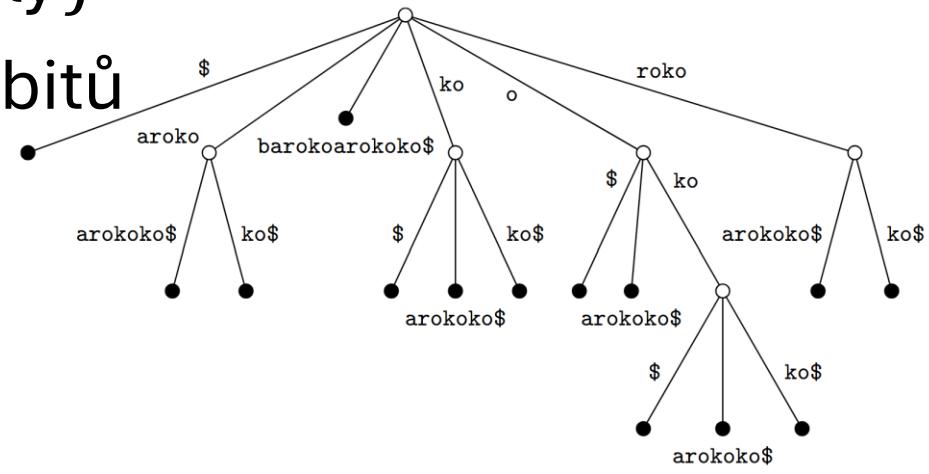
Podobně porovnáme suffix z α_1 se suffixem z α_2:

$$\begin{aligned}
\alpha_1[i:] &\leq \alpha_2[j:] \\
\alpha[3i+1:] &\leq \alpha[3j+2:] \\
(\alpha[3i+1], \alpha[3i+2], \alpha[3i+3:]) &\leq (\alpha[3j+2], \alpha[3j+3], \alpha[3j+4:]) \\
(\alpha[3i+1], \alpha[3i+2], \alpha_0[i+1:]) &\leq (\alpha[3j+2], \alpha[3j+3], \alpha_1[j+1:]) \\
(\alpha[3i+1], \alpha[3i+2], R_{01}[i+1]) &\leq (\alpha[3j+2], \alpha[3j+3], R_{01}[|\alpha_0|+j+1])
\end{aligned}$$

Textové indexy pro podřetězce

= DS pro text α, která umožňuje vyhledání jehly J

- Sufixový strom: $\approx 2 \cdot n \cdot (|\sum| + 2) \cdot \lceil \log_2 n \rceil$ bitů
  - Vyhledávání v $O(|J|)$
- Sufixové+LCP pole: $\approx 2 \cdot n \cdot \lceil \log_2 n \rceil$ bitů
  - Vyhledávání v $O(|J| + \log n)$ za použití LCP
- FM-index: $\approx (1 + o(1)) \cdot n \cdot \lceil \log_2 |\sum| \rceil$ bitů
  - Počet výskytů v $O(|J|)$
  - S pamětí navíc lze najít i pozice výskytů



Obrázek 13.4: Suffixový strom pro slovo barokoarokoko

FM-index (zjednodušený)

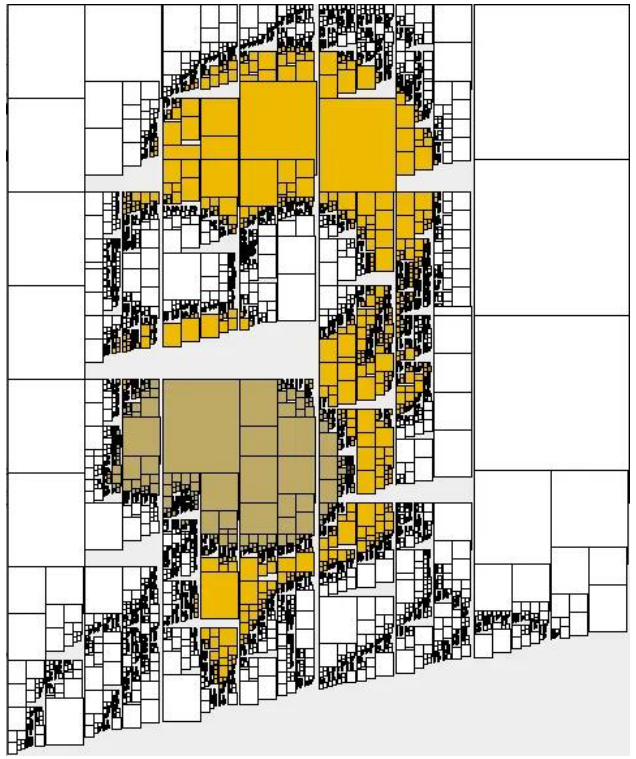
[Ferragina & Manzini ‘05]

- Burrowsova–Wheelerova transformace
  - Permutace $\alpha\$$
  - Seřadíme sufixy $\alpha\$$ a uložíme poslední sloupec
  - Používá se ke kompresi (např. bzip2)
- Occurrence function $\mathrm{occ}(x, i)$
  - Počet výskytů znaku x v $\mathrm{BWT}(\alpha)[:i]$
  - Stačí prostor $o(n)$ navíc k $\mathrm{BWT}(\alpha)$
  - A vyhodnotíme v $O(1)$

I	F	L
1	$abracadabra	
2	a$abracadabr	
3	abra$abracad	
4	abracadabra$	
5	acadabra$abr	
6	adabra$abrac	
7	bra$abracada	
8	bracadabra$a	
9	cadabra$abra	
10	dabra$abraca	
11	ra$abracadab	
12	racadabra$ab	

Příště

- DS pro reprezentaci množiny bodů a intervalové dotazy
  - k-d stromy
  - Range trees



Zdroj: digitalist-alex.blogspot.com/2011/03/quadtree-art.html