Datové struktury 1

8. Přednáška 25.11.

Pavel Veselý

Plán: Naposledy hešování

- Bloomův filtr
  - Pravděpodobnostní DS pro přibližnou reprezentaci množiny
- Hešování vektorů a řetězců
  - + jak udělat z univerzální funkce 2-nezávislou

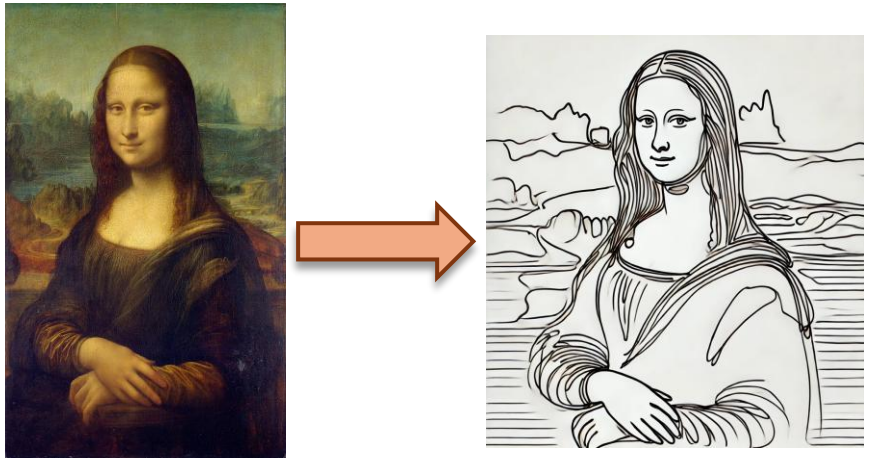
Proudové (streaming) algoritmy a datové skeče

- Jeden průchod přes data
- S malou pamětí
  - menší než data

→ Skeč dat
= obsah paměti algoritmu

- Cena: ztratíme přesnost
  - Povolíme si „malé“ chyby

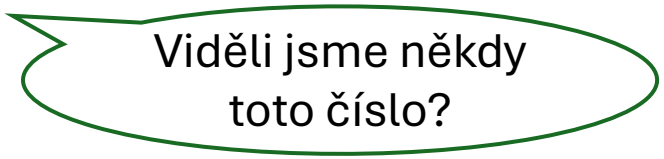
- Triviální: počet prvků, součet čísel
- Zajímavější: počet *různých* prvků
- Často založené na hešování…



Paměť a rychlost vs. chyba

Bloomův filtr pro přibližnou reprezentaci množiny [Bloom '70]

- Pro daný prvek řekne, jestli už byl přidán nebo ne
  - Pokud odpoví NE, tak jsme prvek neviděli
  - Pokud odpoví ANO, tak nevíme…
    - „False positive“ (FP) – prvek jsme neviděli, ale Bloom filtr řekl ANO



- Jak vypadá Bloom filtr?
  - Parametr m = prostor v počtu bitů
  - Pole m bitů
  - k hešovacích funkcí $h_1, \dots, h_k$
    - Např. $k = 5$
- `Insert`(x): Nastavíme buňky $h_1(x), \dots, h_k(x)$ na 1
- `Query`(x): Viděli jsme prvek x?
  - NE, pokud je jedna z buněk $h_1(x), \dots, h_k(x)$ rovna 0
  - Jinak ANO
    - může se splést

Bloomův filtr: varianty

- 1-pásmový: jedno pole m bitů, ale k hešovacích funkcí
  - Stejné garance, ale těžší analýza
- Počítací Bloomův filtr – umí i mazání
  - V buňkách b-bitová počítadla
    - Např. $b = 4$
  - Insert zvýší k počítadel, Delete sníží
  - Query vrátí NE, pokud je jedno z k počítadel nulové
  - Pokud $B_i[j] = 2^b - 1$, pak počítadlo zamrzne

Přehlídka rodin hešovacích funkcí

Pro čísla:

- Lineární kongruence: $\mathcal{L} := \{h_{a,b} | a, b \in \mathbb{Z}_p\}$
  - $h_{a,b}(x) = (a \cdot x + b) \bmod p \bmod m$
  - 2-nezávislé
- Náhodný polynom stupně $k - 1$ nad $\mathbb{Z}_p$
  - $h_{a_0,\ldots,a_{k-1}}(x) = (\sum_{\{i=1\}}^{k} a_i \cdot x^{i-1}) \bmod p \bmod m$
  - k-nezávislé
- Multiply-shift: w-bitová čísla na ℓ-bitová
  - $h_a(x) = \frac{a \cdot x \bmod 2^w}{2^{w-\ell}}$, kde a je liché
  - 2-univerzální (bez důkazu)
- Tabulkové hešování (tabelace)
  - 3-nezávislé a má jiné zajímavé vlastnosti

Pro vektory a řetězce:

- Skalární součin
- Rolling hash – polynom s koeficienty ze vstupního vektoru vyhodnocený v náhodném bodě

Vskuvka: zrychlení operace modulo

- Např. $(a \cdot x + b) \bmod p \bmod m$
- Modulo 10 krát nebo víc pomalejší než násobení
- Co s tím?
  - p známé při kompilaci
  - $\bmod 2^k$ rychlé (AND)
  - Mersennovo prvočíslo $p = 2^q - 1$
    - Př.: 3, 7, 31, 127, $2^{13} - 1$, $2^{17} - 1$, $2^{19} - 1$, $\mathbf{2^{31} - 1}$, $\mathbf{2^{61} - 1}$, $\mathbf{2^{89} - 1}$
    - Pro $x \leq 2^{2q} - 1$:
      - `a = (x & p) + (x >> q)`
      - `If a >= p then return a - p`
      - `Else return a`

Příště

- Už nebude hešování ☺
- DS pro řetězce: sufixové pole a stromy