Bloomův filtr

1 December, 2021 23:12

- Chci uložit množinu $S \subseteq \mathcal{U}$, potřebuji minimalizovat prostor, ale dovoluji false-positive pro Find/Member,

  tj. $\forall x \in S$ MEMBER(x) = YES

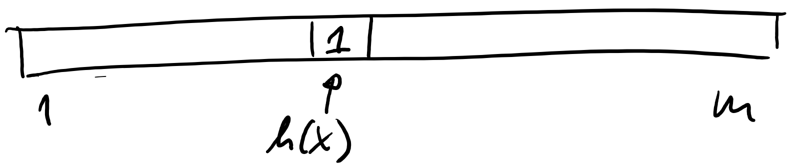
  pro většinu $x \notin S$ MEMBER = NO

  (mohou tedy existovat false-positive)

- lze použít jako předstupeň pro dražší operace

jednoduchý Bloomův filtr

- bitová tabulka velikosti m



hešovací fce $h: \mathcal{U} \to [m]$

INSERT(x) = " $F[h(x)] = 1$ "

MEMBER(x) = " $F[h(x)]$ "

- pro pevnou množinu $S \subseteq \mathcal{U}$, $|S| = n$ a dotaz $z \notin S$

$$\Pr_{h \in \mathcal{H}}[\text{MEMBER}(z) = \text{YES}] \le \frac{n}{m}$$

pokud h je vzato náhodně z univerzálního systému $\mathcal{H}$

Dk: $\Pr_h[\text{MEMBER}(x) = \text{YES}] = \Pr_h[\exists x \in S;\ h(x) = h(z)]$

$$\le \sum_{x \in S} \Pr_h[h(x) = h(z)] \le n \cdot \frac{1}{m} \quad \blacksquare$$

Př. $m = 2n \to \Pr[\text{false-positive}] \le 0.5$

Pr: $m = 2n \to \Pr[\text{false-positive}] \le 0.5$

k-pásmový Bloomův filtr

→ k jednoduchých Bloomových filtrů pro různé, nezávislé zvolené hešovací fce $h_1, \dots h_k$

INSERT(x) = for i=1 to k : $F_i[h_i(x)] = 1$;
MEMBER(x) = for i=1 to k : if $F_i[h_i(x)] == 0$ return NO;
return YES;

- pro pevnou množinu $S \subseteq U$; $|S| = n$ a dotaz $z \notin S$

$$\Pr_h[\text{MEMBER}(z) = \text{YES}] \le \frac{1}{2^k}.$$

→ Pokud chci chybu pro maximálně zlomek ε prvků v U, zvol $k = \lg 1/\varepsilon$ a $m = 2n$
$\to 2n \cdot \lg 1/\varepsilon$ bitů

- Pro h zcela nezávislé, lze použít analýzu

$m = n \cdot \lg e$ $\qquad \Pr[h(z) \text{ je prázdný}] = \left(1 - \frac{1}{m}\right)^n$

$$\le \frac{1}{e^{n/m}} = \frac{1}{e^{1/\lg e}} = \frac{1}{(2^{\lg e})^{1/\lg e}} = \frac{1}{2}$$

$\lg e = 1.44\dots$

$k = \lg 1/\varepsilon$

$\longrightarrow 1.44 n \cdot \lg 1/\varepsilon$ bitů pro chybu ε

DELETE(x)? – nelze bez znalosti celého S.

→ počítací Bloomův filtr:

místo bitu, malý čítač

INSERT(x) = for i=1 to k : $F_i[h_i(x)]$++ ;

DELETE(x) = for i=1 to k : $F_i[h_i(x)]$-- ;

MEMBER(x) = viz výše.

pokud daný čítač dosáhne své maximální hodnoty,
pak se již dále nemění ani při ++ ani při --.
(je "zaseklý")

napŕ: 4-bitový čítač se zasekne po 15 "++".

pro $m = n \cdot \lg e$

$$\Pr[\text{daná přihrádka dostane} \geq t \text{ prvků}] \leq \binom{n}{t} \cdot \left(\frac{1}{m}\right)^t \leq \frac{n^t}{t!} \cdot \frac{1}{(n \lg e)^t} = \frac{1}{t!\,(\lg e)^t}$$

$$\Rightarrow \Pr[\text{přihrádka} \geq 15] \leq 3{,}22 \times 10^{-15}$$

pro 10^9 přihrádek $\Pr[\exists \text{ přihrádka větší} \geq 15] \leq 10^{-5}$.