

9. cvičení

Datové struktury I, 4. 12. 2025

<https://iuuk.mff.cuni.cz/~chmel/2526/ds1/>

Úloha 1 (Rolling hash je d -univerzální)

Pro prvočíslo p a délku vektoru d definujeme třídu hashovacích funkcí $\mathcal{R} = \{h_a: a \in \mathbb{Z}_p\}$, kde $h_a(x) = \sum_{i=0}^{d-1} x_{i+1}a^i$, a všechno počítáme modulo p . (Bereme $x \in \mathbb{Z}_p^d$, indexujeme od jedničky.)

Dokažte, že tato třída hashovacích funkcí je $(d-1)$ -univerzální.

Hint: *„If you fix a few more elements – you can’t find a polynomial of fixed degree that vanishes at all of them“*

Úloha 2 (Vyhledávání jehly v textu)

Vymyslete algoritmus na nalezení všech výskytů podřetězce x délky n v textu T délky m pomocí hashování, který běží v průměrném čase (tj. ve střední hodnotě) $\mathcal{O}(n + m + k \cdot n)$, kde k je počet výskytů x v T .

Úloha 3 (Přetečení v počítačím Bloomově filtru)

Uvažme počítačím Bloomův filtr s maximální hodnotou jednoho počítadla ℓ . Tento Bloomův filtr bude mít m políček s počítadly, a budeme používat k zcela náhodných hashovacích funkcí. Určete pravděpodobnost toho, že nám tento Bloomův filtr přeteče (tedy budeme mít aspoň jedno počítadlo, které se dostalo na hodnotu ℓ). Pro jednoduchost můžete začít nejprve s $k = 1$, a potom zobecnit pro libovolné k .

Úloha 4 (Bitová krize)

Máte Bloomův filtr s bitovým polem délky $m = 2^b$. Bohužel ale přišla krize, RAMky skokově zdražily¹ a vaše RAMka zrovna vyhořela (ale naštěstí předtím zvládla dumpnout všechno, co v ní bylo, na disk). Koupíte si tedy novou, ale protože je drahá, můžete si dovolit jenom 2^{b-1} bitů. Zároveň ale nechcete přijít o váš filtr – umíte ho nějak upravit, aby fungoval i s poloviční pamětí?

Úloha 5 (Operace na Bloomových filtrech)

Uvažme dva jednopásové Bloomovy filtry s poli F_1, F_2 , které mají stejnou délku, používají stejné hashovací funkce (potenciálně i více) a reprezentují množiny A_1, A_2 . Uvažme Bloomův filtr s polem F_3 , které vznikne tak, že $F_3[i] := F_1[i] \wedge F_2[i]$.

- Je F_3 funkční Bloomův filtr pro množinu $A_1 \cap A_2$? (Tedy chceme, aby F_3 nikdy chybně neodpověděl „ne“.)
- Je F_3 identický s Bloomovým filtrem, který by vzniknul postupným přidáváním prvků z množiny $A_1 \cap A_2$?
- Umíte tento postup upravit i pro počítačím filtry?
- A co pro sjednocení? (A počítačím filtry?)

Bonusové úlohy

Úloha 6 (Dosáhneme?)

Sestrojte pravděpodobnostní algoritmus² pro rozhodnutí dosažitelnosti v neorientovaném grafu (tedy odpověď na otázku „existuje cesta z u do v ?“), který selže s pravděpodobností nejvýše $1/4$, a potřebuje jenom logaritmický prostor.

Technická poznámka k logaritmickému prostoru: když máme algoritmus běžící v sublineárním prostoru, počítáme jej tak, že máme vstup na „read-only“ disku, kde k němu můžeme přistupovat, ale nemůžeme jej upravovat (ale podle potřeby si jej můžeme bez problému kopírovat).³

Pravděpodobně se vám bude hodit následující tvrzení: když v n -vrcholovém souvislém neorientovaném grafu vyjdeme z vrcholu u , a budeme v každém kroku uniformně náhodně vybírat ze všech sousedů vrchol, kam půjdeme v dalším kroku, střední hodnota doby než navštívíme jiný vrchol v , je nejvýše $2n^3$.

Taky se může hodit připomenout si Markovovu nerovnost: Buď X nezáporná náhodná veličina. Pak $\forall \varepsilon > 0$ platí $P[X \geq \varepsilon] \leq \frac{\mathbb{E}[X]}{\varepsilon}$. Ekvivalentně pro jakékoliv $d \geq 1$, $P[X \geq d \cdot \mathbb{E}[X]] \leq \frac{1}{d}$.

Fun fact. Octomliky mají vyvinutou jakousi verzi Bloomova filtru se zapomínáním pro pachy. Více viz <https://www.pnas.org/doi/10.1073/pnas.1814448115> – Dasgupta, Sheehan, Stevens, Navlakha: A neural data structure for novelty detection, *Proc. Natl. Acad. Sci. U.S.A.* 115 (51) 13093-13098.

¹Podobnost se skutečnými událostmi je čisté náhodná.

²Pravděpodobnostní algoritmy jste asi už potkali, ale zkráceně: bude se jednat o algoritmus, který se bude moct v každém kroku rozhodnout náhodně z několika možností

³Technicky tuto třídu definujeme na Turingových strojích, kde máme read-only vstupní pásku, a druhou pracovní pásku. Do prostorové složitosti se nám počítá jenom pracovní pásku.