# Hešovací funkce

**(slabá) univerzalita**

Systém $\mathcal{H}$ hešovacách funkcí $\mathcal{U} \to [m]$ je (slabě) $c$-univerzální ($c \geq 1$) $\Leftrightarrow$

$$(\forall x \neq y \in \mathcal{U}) \ \Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq c/m$$

"Pravděpodobnost kolize dvou prvků je úměrná velikosti cílového univerza"

Pokud hešujeme (různé) prvky množiny $N$ ($|N| = n$) pomocí $c$-univerzálního sytému do $[m]$, celkový očekávaný očekávaný počet kolizí hešu se určí následovně:

$$\mathop{\mathbb{E}}_{h \in \mathcal{H}}[\#colisions] = \sum_{x,y \in M} \Pr_{h \in \mathcal{H}}[h(x) = h(y)] = \binom{n}{2} \frac{c}{m} = \mathcal{O}_{\mathbb{E}}(n^2/m)$$

Kde první rovnost je využití linearity střední hodnoty. Formálně počet kolizí je součet indikátorových náhodných proměnných indukujících zda nastaly kolize mezi jednotlivými páry. Střední hodnota indikátorové proměnné je pravděpodobnost indikovaného jevu.

Obdobně můžeme spočítat očekávaný počet kolizí fixního prvku $x$ s ostatními prvky (opět předpoládáme, že všechny prvky jsou různé).

$$\mathop{\mathbb{E}}_{h \in \mathcal{H}}[\#x - colisions] = \sum_{y \in M; y \neq x} \Pr_{h \in \mathcal{H}}[h(x) = h(y)] = (n-1)\frac{c}{m} = \mathcal{O}_{\mathbb{E}}(n/m)$$

Z univerzality tedy umíme odhadnout počet kolizí, ale nemáme žádné garance jejich distribuce. Existují příklady $c$-univerzálních systémů, kde pravděpodnost kolize všech prvků do stejné hodnoty je právě $c/m$.

**silná univerzalita**

Systém $\mathcal{H}$ hešovacách funkcí $\mathcal{U} \to [m]$ je silne $c$-univerzální ($c \geq 1$) $\Leftrightarrow$

$$(\forall x, y \in \mathcal{U}, \forall a, b \in [m]) \ \Pr_{h \in \mathcal{H}}[h(x) = a \wedge h(y) = b] \leq c/m^2$$

"Korelace mezi hodnotami dvojic prvků nejsou moc velké"

Fakt: Silná $c$-univerzalita implikuje (slabou) $c$-univerzalitu.

Fakt: Silná $c$-univerzaliza implikuje následující rovnoměrnost hodnot

$$(\forall x \in \mathcal{U}, \forall a \in [m]) \ \Pr_{h \in \mathcal{H}}[h(x) = a] \leq c/m$$

(tato vlastnost ale neimplikuje ani slabou univerzalitu!)

**Střední časová složitost**

Pokud používáme hešování pro identifikaci stejných prvků (nebo podobným způsobem), složitost algoritmu se určí jako složitost algoritmu za předpokladu neexistence kolizí (perfektní volba hešovací funkce) plus cena časové složitosti každé falešné kolize přenásobená očekávaným počtem kolizí.

Př. pro $n$ prvků chceme ověřit, zda jsou všechny různé a vypsat všechny shodné páry. Vytvoříme hešovací tabulku velikosti $m$, prvky zahešujeme do tabulky a porovnáme prvky se stejným hešem. Vypíšeme všechny stejné dvojice.

Časová složitost bude asymptoticky součet ceny inicializace tabulky, zahešování všech prvků, zpracování falešných kolizí (použijeme očekávaný početkolizí) a výpis skutečně shodných prvků.

Pro $c$-univerzální systém a čísla bude celková složitost $\mathcal{O}_{\mathbb{E}}(m + n + \frac{n^2}{m} + \#pairs)$, pro volbu $m = n$ tedy $\mathcal{O}_{\mathbb{E}}(n + \#pairs)$. Pro $c$-univerzální systém a řetězce délky $k$ dostaneme $\mathcal{O}_{\mathbb{E}}(m + kn + k\frac{n^2}{m} + k * \#pairs)$. Celková složitost bude $\mathcal{O}_{\mathbb{E}}(kn + k * \#pairs)$ pro volbu $m = kn$ (optimalizce času na kolizích) nebo $m = n$ (optimalizace času inicializace). V komplexnějších případech může být třeba volit parametry chytřeji podle toto kterou část alg. je třeba optimalizovat.

**Speciální hešovací funkce**

Často se pro návrh algoritmu hodí hešovací funkce produkující heše, v určitým vztahem.

Např. se hodí "pseudo-lineární" systém hešovacích funkcí t.ž. pro libovolné dva prvky $a, b$ (a volbu $h$) platí $h(a + b) = g_h(h(a), h(b))$ pro nějakou jednoduchou funkci $g_h$ (závislou na volbě $h$). Tedy umíme najít heše součtů bez znalosti původních prvků. Např. pro konstrukci hešovacích systémů ze skalárního součinu lze tuto vlastnost snadno nahlédnout.

Analogicky existují hešovací funkce, které umožňují různé další operace (např. rozšíření/zkrácení řetězce o nějaké znaky, viz. ADS2 a vyhledávání v textu). Dokonce existují systémy, které umožňují kostrukci celé algebry, a tak provádět výpočty na "zašifrovaných" datech (užitečné třeba pro konstrukci bezpečných a anonymních volebních systémů, čená magie).

**Varování**

- Hešovací funkce nejsou totéž co náhodné funkce. Nelze je analyzovat jako náhodné funkce. Hešovací funkce mají masivní korelace mezi hodnotami.
- Hešovací funkce sama o sobě žádné vlastnosti nemá, pouze celý systém hešovacích funkcí má příslušné vlastnosti. Nelze tedy nikdy používat fixní hešovací funkci.
- Nevytvářejte vlastní návrhy funkcí s tím, že se budou chovat jako hešovací, nebudou. Typický prohřešek bývá součet cifer čísla nebo hodnot znaků v řetezci. Obzvláště součet má velmi ne-hešové vlastnosti (třeba vůbec nezávisí na pořadí prvků).