Základní pravděpodobnostní analýza hašování

1 Hašování a cíl jeho analýzy

Základní situace při hašování je popsána následovně:

U je množina, často nazývaná *universum*, které má P prvků;
T je množina (hašovací tabulka), která má M prvků;
$h : U \to T$ je funkce;
$A \subset U$ je množina, která má N prvků.

U je obvykle obrovská, např. 4-bytová celá čísla. Prvky množiny T jsou samy o sobě datové struktury, do kterých je možno ukládat prvky universa. Například to mohou být spojové seznamy, do kterých je možno ukládat celá čísla. O funkci h se předpokládá pouze, že množiny $h^{-1}(t)$ jsou pro všechny prvky $t \in T$ zhruba stejně velké, tedy se do každého t zobrazuje zhruba stejně prvků universa. O množinách A a T předpokládáme, že jsou alespoň řádově stejně velké a přitom jsou malé nebo rozumně velké (např. tak, aby se pole s M prvky vešlo bez problémů do počítače).

A je množina vybraných prvků universa, kterou chceme v hašovacích tabulkách uložit. Udělá se to tak, že pro každý prvek $a \in A$ se určí $h(a)$ a pak se prvek a uloží do $h(a)$ ($h(a)$ je prvek množiny T a tedy, jak bylo řečeno výše, je to datová struktura). Jestliže pak například hledáme, zda prvek $u \in U$ v tabulkách je, stačí se podívat, zda je uložen v datové struktuře $h(u)$, jinde být nemůže.

Pro rychlost zpracování je důležité, aby prvky množiny A byly do datových struktur tabulky T rozloženy rovnoměrně, tedy aby v žádné z nich nebylo příliš mnoho prvků z množiny A.

Formálně si pro $t \in T$ označíme jako A_t množinu $a \in A$ takových, že $h(a) = t$, velikost množiny A_t označíme N_t. Budeme tedy chtít dokázat, že žádná z množin A_t není moc velká.

V ideálním případě když jsou všechny tyto množiny stejně velké, bude jejich velikost N/M (tedy pokud např. $N = M$, mohou být A_t jednoprvkové); lepší už to být nemůže, ale může to být horší a to o hodně horší. Může se dokonce stát, že všechny prvky množiny A se zobrazí funkcí h do stejného prvku v T a pak má jedna z A_t prvků N a ostatní jsou prázdné.

Mým cílem je dokázat, že existuje konstanta C (nezávislá na množinách U, T, A a funkci h) taková, že pokud se prvky A vyberou náhodně z U tak, aby A měla N prvků, pak je velmi nepravděpodobné, že by existovala množina A_t, která by měla více než $C \log N$ prvků, tedy je nepravděpodobné že by neplatil následující vztah: $\max_{t \in T} N_t \leq C \log N$.

To je sice horší než ideální případ, kdy hodnota $\max_{t \in T} N_t$ je blízká N/M (což je konstanta, pokud M a N jsou zhruba stejně velké), ale výrazně lepší než když jedna z množin A_t má N nebo skoro N prvků. Jelikož při hašování je doba na provedení operace shora omezena číslem $\max_t N_t$, vyplyne z toho, že průměrná rychlost hašování je logaritmická (vzhledem k velikosti uchovávané

množiny A) a tedy srovnatelná s rychlostí vyvážených stromových datových struktur s nesrovnatelně větší logickou složitostí.

Zbývá ale vysvětlit, co to znamená "vyberou náhodně" a co znamená "je nepravděpodobné". To uděláme v následujícím oddílu.

2 Formální cíl jeho analýzy

Pro danou množinu A mne budou zajímat velikosti množin $A_t = \{a \in A \mid h(a) = t\}$.

Pro dané k si označím:

$\mathcal{M}_k$ je množina $A \subset U$ takových, že $|A| = N$ a pro některé $t \in T$ je $A_t > k$.
$\mathcal{M}$ je množina všech množin $A \subset U$ takových, že $|A| = N$.
$\mathcal{N}_k$ je počet množin $A \subset U$ takových, že $|A| = N$ a pro některé $t \in T$ je $A_t > k$.
$\mathcal{N}$ je počet *všech* množin $A \subset U$ takových, že $|A| = N$.

V našem označení je tedy $|\mathcal{M}_k| = \mathcal{N}_k$ a $|\mathcal{M}| = \mathcal{N}$.

Podíl $\mathcal{N}_k/\mathcal{N}$ je možno brát jako matematicky přesnou definici vágního pojmu "pravděpodobnost, že pro náhodně zvolenou množinu A prvků universa je některé A_t větší než k".

Je známo, že

$$\mathcal{N} = \binom{P}{N} = \frac{P(P-1)\cdots(P-N+1)}{N(N-1)\cdots 1}.$$

Jestliže jako "příliš veliké" budu chápat "má více než k prvků", pak $\mathcal{N}_k/\mathcal{N}$ je možno brát jako matematicky přesnou definici vágního pojmu "pravděpodobnost, že pro náhodnou A je některé A_t příliš velké".

Mým cílem je dokázat následující odhad:
Věta: Nechť platí
(1) $4eN^2 \leq P$,
(2) pro každé $t \in T$ $|h^{-1}(t)| < 2P/M$,

pak pro každé celé nezáporné číslo D existuje konstanta C (závislá na D, ale nezávislá na U, T, A a h) taková, že pokud $k \geq 2N/M + C \log N$, pak $\mathcal{N}_k/\mathcal{N} \leq N^{-D}$. ♣

Jestliže $N \approx M$, pak N/M je konstantní a věta tvrdí, že relativní počet výběrů náhodné množiny N prvků universa, při kterých počet prvků zobrazených do jedné datové struktury tabulky T překročí logaritmickou hranici (vzhledem k počtu všech výběrů A) klesá k nule rychleji jež jakákoli racionální funkce.

Požadavek (1) je obvykle bez problémů splněný požadavek, znamenající, že U je o hodně větší než A, (2) říká, že do žádného prvku množiny T se nezobrazuje hašovací funkcí h více než dvojnásobek průměrného počtu prvků (citelné oslabení požadavku, aby $h^{-1}(t)$ byly zhruba stejně velké).

3 Úvod důkazu věty

Pro dané číslo k zavedu následující označení:

$\mathcal{M}_{k,t}$ je množina všech množin $A \subset U$ takových, že $|A| = N$ a $A_t \geq k$,
$\mathcal{N}_{k,t}$ je počet všech množin $A \subset U$ takových, že $|A| = N$ a $A_t \geq k$,
$\overline{\mathcal{M}}_{\ell,t}$ je množina všech množin $A \subset U$ takových, že $|A| = N$ a $A_t = \ell$,
$\overline{\mathcal{N}}_{\ell,t}$ je počet všech množin $A \subset U$ takových, že $|A| = N$ a $A_t = \ell$,
a nakonec je-li pro $B \subset A$ takový, že $|B| = \ell$, pak
$\overline{\mathcal{M}}_{\ell,t,B}$ množina všech množin A takových, že $B \subset A \subset U$ $|A| = N$ a pro $a \in A$ je $h(a) = t$ právě když $a \in B$ a
$\overline{\mathcal{N}}_{\ell,t,B}$ počet všech množin A takových, že $B \subset A \subset U$ $|A| = N$ a pro $a \in A$ je $h(a) = t$ právě když $a \in B$.

Je zřejmé, že platí $|\mathcal{M}_{k,t}| = \mathcal{N}_{k,t}$, $|\overline{\mathcal{M}}_{\ell,t}| = \overline{\mathcal{N}}_{\ell,t}$, $|\overline{\mathcal{M}}_{k,t,B}| = \overline{\mathcal{N}}_{k,t,B}$. Kromě toho

$$\mathcal{M}_k = \bigcup_{t \in T} \mathcal{M}_{k,t},$$

$$\mathcal{M}_{k,t} = \bigcup_{\ell=k}^{N} \overline{\mathcal{M}}_{\ell,t},\ a$$

$$\mathcal{M}_{\ell,t,B} = \bigcup_{|B|=\ell} \overline{\mathcal{M}}_{\ell,t,B},$$

což dává

$$\mathcal{N}_k \leq \sum_{t \in T} \mathcal{N}_{k,t},$$

protože jestliže se nějaké s započítává na levé straně, pak se pro nějaké $t \in T$ započítává i ve sčítanci $\mathcal{N}_{k,t}$ na pravé straně, ale může se započítávat ve více sčítancích a proto pravá strana může být větší.

Dále platí

$$\mathcal{N}_{k,t} \leq \sum_{\ell=k}^{N} \overline{\mathcal{N}}_{\ell,t}$$

Kromě toho je zřejmé, že platí

$$\mathcal{N}_{\ell,t} \leq \sum_{|B|=\ell} \mathcal{N}_{\ell,t,B}.$$

Nakonec

$$\mathcal{N} = P(P-1)\cdots(P-N+1)/N!,$$

$$\mathcal{N}_{\ell,t,B} = P_t(P_t-1)\cdots(P_t-\ell+1)(P-P_t)(P-P_t-1)\cdots(P-P_t-(N-\ell)+1)/N!,$$

takže dostáváme

$$\frac{\mathcal{N}_k}{\mathcal{N}} \le \sum_{t\in T}\sum_{\ell=k}^{N}\sum_{|B|=\ell}\frac{\mathcal{N}_{\ell,t,B}}{\mathcal{N}}, \tag{3}$$

kde

$$\frac{\mathcal{N}_{\ell,t,B}}{\mathcal{N}} = \frac{P_t}{P}\frac{(P_t-1)}{P-1}\cdots\frac{(P_t-\ell+1)}{P-\ell+1}\frac{(P-P_t)}{P-\ell}\frac{(P-P_t-1)}{P-\ell-1}\cdots\frac{(P-P_t-(N-\ell)+1)}{P-N+1}. \tag{4}$$

4 Zjednodušení vzorce

Výraz, kterým jsme skončili si teď trochu zjednodušíme. Nejprve uvážíme, že v uvedeném součinu zlomků je prvních ℓ zlomků přibližně P_t/P a zbývajících $N-\ell$ zlomků je přibližně $(P-P_t)/P = 1-P_t/P$. Přesněji

$$\frac{P_t-i}{P-i} \le \frac{P_t}{P-i} = \frac{P_t}{P}\left(1+\frac{i}{P-i}\right) \le \frac{P_t}{P}\left(1+\frac{2i}{P}\right) \le \frac{P_t}{P}\left(1+\frac{2N}{P}\right),$$

$$\frac{P-P_t-i}{P-\ell-i} \le \frac{P-P_t}{P-\ell-i} \le \frac{P-P_t}{P}\left(1+\frac{\ell+i}{P-\ell-i}\right) \le$$
$$\le \frac{P-P_t}{P}\left(1+\frac{2(\ell+i)}{P}\right) \le \frac{P-P_t}{P}\left(1+\frac{4N}{P}\right).$$

Tedy

$$\frac{\mathcal{N}_{\ell,t,B}}{\mathcal{N}} \le \left(\frac{P_t}{P}\right)^{\ell}\left(\frac{P-P_t}{P}\right)^{N-\ell}\left(1+\frac{4N}{P}\right)^{N} \le \left(\frac{P_t}{P}\right)^{\ell}\left(\frac{P-P_t}{P}\right)^{N-\ell}e^{4N^2/P}, \tag{5}$$

takže označíme-li $p_t = P_t/P$, je

$$\frac{\mathcal{N}_{\ell,t,B}}{\mathcal{N}} \le p_t^{\ell}(1-p_t)^{N-\ell}e. \tag{6}$$

Výraz na pravé straně nezávisí na B a proto

$$\sum_{|B|=\ell}\frac{\mathcal{N}_{\ell,t,B}}{\mathcal{N}} \le e\binom{N}{\ell}p_t^{\ell}(1-p_t)^{N-\ell}. \tag{7}$$

To co se objevilo na pravé straně v následujícím oddílu probereme podrobněji.

5 Binomické rozdělení

Výraz $B(N,p;\ell) = \binom{N}{\ell} p^\ell (1-p)^{N-\ell}$ je velmi významný vzorec, udávající ℓ-tý člen *binomickeho rozdělení*, viz teorie pravděpodobnosti, kde se odvozují velmi přesné odhady jeho velikosti i to, že binomické rozložení v limitě přechází v ještě důležitější *normální rozdělení*. Zde si uděláme jen velmi hrubý odhad, který ale bude zcela dostatečný pro naše účely.

Ve zbytku tohoto oddílu zafixujeme N a p a položíme

$$\alpha_\ell = \frac{B(N,p;\ell+1)}{B(n,p;\ell)},$$

tedy α_ℓ je podíl dvou po sobě jdoucích členů binomického rozdělení. Platí

$$\alpha_\ell = \frac{N-\ell}{\ell+1}\frac{p}{1-p}.$$

Z toho je vidět, že α_ℓ s rostoucím ℓ klesá.

Zkusme nyní, kolik je α_{Np} a α_{Np-1}:

$$\alpha_{Np} = \frac{N-Np}{Np+1}\frac{p}{1-p} \le \frac{N-Np}{Np}\frac{p}{1-p} = 1$$

$$\alpha_{Np-1} = \frac{N-Np+1}{Np}\frac{p}{1-p} \ge \frac{N-Np}{Np}\frac{p}{1-p} = 1$$

Tedy až do asi $\ell = Np$ je α_ℓ větší než 1 a pak už je stále menší než 1. Jinými slovy $B(N,p;\ell)$ až asi do $\ell = Np$ roste (každý další člen je větší než předchozí, protože jejich poměr α_ℓ je větší než 1, a pak $B(N,p;\ell)$ stále klesá. Důležité ale je, jak klesá:

$$\text{je-li } \ell \ge 2Np, \text{ pak } \alpha_{2Np} = \frac{N-2Np}{2Np+1}\frac{p}{1-p} \le \frac{N-2Np}{2Np}\frac{p}{1-p} = \frac{1}{2}\frac{1-2p}{1-p} \le \frac{1}{2}$$

Položme nyní $k = 2\lceil Np_t \rceil$ a zvolme celé nezáporné q. Pak

$$B(N,p;k+q) = B(N,p;k)\alpha_k\alpha_{k+1}\cdots\alpha_{k+q-1} \le B(N,p;k)2^{-q},$$

takže pro libovolné celé $\ell \ge k$ je

$$\sum_{i=\ell}^{N} B(N,p;i) \le B(N,p;\ell)\left(1 + 2^{-1} + 2^{-2} + \cdots\right) \le 2B(N,p;\ell).$$

Navíc pro $\ell = k + C\log_2 N$ je

$$\sum_{i=\ell}^{N} B(N,p_t;i) \le 2B(N,p_t;\ell) \le B(N,p_t;k)2^{-C\log_2 N} \le B(N,p_t;k)N^{-C} \le N^{-C}. \tag{8}$$

6 Dokončení důkazu

Z (2) víme, že $p_t = P_t/P \leq (2P/M)/P = 2/M$, takže pokud je $k \geq 4N/M + C \log N$, je také $k \geq 2\lceil Np_t \rceil + C \log N$ a dosadíme-li do (3) nerovnosti z (7) a (8), dostaneme, že

$$\frac{\mathcal{N}_k}{\mathcal{N}} \leq e \sum_{t \in T} N^{-C} = eMN^{-C} \leq N^{2-C} \tag{9}$$

pro dostatečně velké N.