Hachage parfait [Cormen p 258 - 262]

Pour le dev : Jouir que le noir

On considère 1 univers U dans lequel les clés prennent leurs valeurs et 1 ensemble figé $K \subset U$ de clés de taille m (= taille de la table d'entrée). Par codage, on a $U \subset \mathbb{N}$

déf On dit qu'1 hachage est parfait qd le nb d'accès mémoire au pire cas pour 1 recherche est en $O(1)$

+ def classe de fonction de hachage universelle

plan Pour créer 1 hachage parfait, on utilise 2 niveaux de hachage :

On commence par hacher les n clés de m alvéoles (à préciser) avec 1 certaine fct° h

Puis pour chaque alvéole j, on la remplit avec 1 table de hachage secondaire S_j avec 1 fonction de hachage h_j. On montre qu'en fixant la taille de S_j égale à $m_j = {n_j}^2$ où n_j = nb de clés hachées de l'alvéole j par h on n'aura pas de collision au 2e niv

déf Soit $p \in \mathbb{P}$ tq $p > |U|$. Pour tout $a \in \mathbb{Z}/p\mathbb{Z}^*$ et tout $b \in \mathbb{Z}/p\mathbb{Z}$

on note $h_{a,b}$: U assimilé à $\mathbb{Z}/p\mathbb{Z} \longrightarrow [\![0, m-1]\!]$, $k \longmapsto (ak + b \bmod p) \bmod m$.

(ça c'est des nb)

et on note $\mathcal{H}_{p,m} = \{h_{a,b} \mid a \in \mathbb{Z}/p\mathbb{Z}^*, b \in \mathbb{Z}/p\mathbb{Z}\}$

th [ADMIS] au moins do la def

$\mathcal{H}_{p,m}$ est 1 classe de fonction de hachage universelle car si h est tiré uniformément aléatoirement de $\mathcal{H}_{p,m}$ et k, ℓ sont 2 clés $\neq$ de $\mathbb{Z}/p\mathbb{Z}$, $\mathbb{P}(h(k) = h(\ell)) \leq \frac{1}{m}$ (chances de collision $\leq \frac{1}{m}$)

th Si on stocke n clés de 1 table de taille $m = n^2$ à l'aide d'1 fonction h choisie uniformément aléatoirement de 1 classe de fonction de hachage universelle alors la proba d'avoir des collisions est $< \frac{1}{2}$

démo Soit X la va qui compte le nb de collisions. Il y a $\binom{n}{2}$ paires de clés susceptibles d'entrer en collision et chaque paire a 1 proba $\leq \frac{1}{m}$ d'entrer en collision (def classe universelle). $\mathbb{E}[X] = \mathbb{E}\left[\sum_{k,\ell \text{ paires de clés}} \mathbb{1}_{\{h(k)=h(\ell)\}}\right] = \sum_{k,\ell} \mathbb{P}(h(k) = h(\ell)) \leq \dots$

Donc $\mathbb{E}[X] \leq \frac{1}{m}\binom{n}{2} = \frac{1}{m^2} \times \frac{n(n-1)}{2} = \frac{n-1}{n} \times \frac{1}{2} < \frac{1}{2}$

On applique ensuite l'$\leq$ de Markov : $\mathbb{P}(X \geq 1) \leq \frac{\mathbb{E}[X]}{1} < \frac{1}{2}$ ce qui conclut

Pour trouver 1 table sans collisions, il suffit d'essayer plusieurs R jusqu'à ce que ça marche : la proba d'avoir 1 collision après k choix de R est $\leq \frac{1}{2^k}$ donc on va y arriver !

Pb : en mémoire, la table est de taille m^2 = quadratique en le nb de clés = trop gros si beaucoup de clés.

On hache alors les n clés de 1 table de taille $m = n$ via R puis pour chaque alvéole j on construit 1 table secondaire S_j de taille $m_j = n_j^2$ où n_j = nb de clés qui arrivent de l'alvéole j via R. On garantit là encore qu'en peu d'essais (linéaire en n) on a 1 table à 2 niveaux sans collision. Montrons qu'on a gagné en mémoire / à la 1ère situation :

(TH) Si on stocke n clés de 1 table de taille $m = n$ via R choisie uniformément de 1 classe de fonctions de hachage universelle alors $E\left[\sum_{j=0}^{m-1} n_j^2\right] < 2m$

où n_j est la va correspondant au nb de clés hachées de l'alvéole j par R

démo On sait $\forall p \in \mathbb{N}^*$, $p^2 = p + 2\binom{p}{2}$

donc $E\left[\sum_{j=0}^{m-1} n_j^2\right] = E\left[\sum_{j=0}^{m-1} \left(n_j + 2\binom{n_j}{2}\right)\right]$

$= E\left[\underbrace{\sum_{j=0}^{m-1} n_j}_{=n}\right] + 2E\left[\underbrace{\sum_{j=0}^{m-1} \binom{n_j}{2}}_{\text{= nb de paires d'éléments qui entrent en collision dans la grande table principale (donnée par R)}}\right]$ par linéarité

$= n$

= 2 x du TH d'avant

Donc comme pr le TH d'avant, $E\left[\sum_{j=0}^{m-1} \binom{n_j}{2}\right] \leq \frac{1}{m}\binom{n}{2} = \frac{1}{m}\frac{n(n-1)}{2} = \frac{n-1}{2}$

Donc $E\left[\sum_{j=0}^{m-1} n_j^2\right] \leq n + n - 1 < 2n$.

On en déduit que, si on prend la table S_j de taille $m_j = n_j^2$, la quantité moyenne de mémoire requise pour toutes les tables secondaires est $< 2n$. Et $\hat{m}$:

(TH) Avec la stratégie décrite précédemment, la proba que l'espace total occupé par 2 tables secondaires dépasse $4m$ est $< \frac{1}{2}$

démo / inégalité Markov : $P\left(\sum_{j=0}^{m-1} n_j^2 \geq 4m\right) \leq \frac{E\left[\sum_{j=0}^{m-1} n_j^2\right]}{4m} < \frac{1}{2}$

Rappel : l'aléa vient du choix des fonctions de hachages, pas des clés qui elles sont fixées

Pour faire 1 hachage sans collision et en $O(m)$ en espace : tant que $\sum n_j^2 \geq 4m$, choisir une nouvelle R (on va trouver 1 R qui va par). Ça garantit 1 espace linéaire. Puis $\forall j \in [0, m-1]$, tant qu'il y a 1 collision de S_j, choisir 1 nouvelle R_j (on va en trouver 1