Hachage parfait

Manon Ruffini

On considère un univers de clés U et un ensemble de clés figé $K \subset U$, de cardinal n. Par codage, on peut supposer que l'univers U est inclus dans $\mathbb{N}$.

Définition 1

*Un hachage est dit **parfait** lorsque le nombre d'accès mémoire requis pour faire une recherche est, dans le cas le plus défavorable, $O(1)$.*

Pour créer un hachage parfait, on va utiliser deux niveaux de hachage :

— Le premier niveau correspond à un hachage par chaînage : les n clefs sont hachées vers m alvéoles grâce à une fonction de hachage h choisie parmi une famille de fonctions de hachage universelle.

— Pour chaque alvéole j, on utilise une table de hachage secondaire S_j, avec une fonction de hachage h_j. Si on choisit bien h_j, on pourra s'assurer qu'il n'y aura pas de collision dans ce second niveau.

Pour ça, on suppose que la taille m_j de S_j est n_j^2, où n_j est le nombre de clefs hachées dans l'alvéole j.

On commence par décrire une classe de fonctions de hachage universelle.
Soit p un nombre premier tel que toute clef possible soit dans $[|0, p-1|]$. On suppose de $|U| > m$, donc on a $p > m$.

Définition 2

Pour tout $a \in (\mathbb{Z}/p\mathbb{Z})^*$, et pour tout $b \in \mathbb{Z}/p\mathbb{Z}$, on définit la fonction suivante :

$$h_{a,b} : k \mapsto ((ak+b) \mod p) \mod m$$

On définit $\mathcal{H}_{p,m} = \{h_{a,b}, a \in (\mathbb{Z}/p\mathbb{Z})^*, b \in \mathbb{Z}/p\mathbb{Z}\}$

Théorème 1

La classe $\mathcal{H}_{p,m}$ des fonctions de hachage ainsi définie est universelle ; c'est-à-dire que si on prend une fonction h aléatoirement dans $\mathcal{H}_{p,m}$, les chances de collisions sont inférieures ou égales à $1/m$.

Preuve [1]

Théorème 2

Si on stocke n clefs dans une table de taille $m = n^2$, à l'aide d'une fonction h choisie aléatoirement dans une classe universelle de fonctions de hachage, alors la probabilité d'avoir des collisions est inférieure à $\frac{1}{2}$.

1. Soient k,l deux clefs distinctes de $\mathbb{Z}/p\mathbb{Z}$. Pour une fonction de hachage $h_{a,b}$ donnée, notons $r = (ak+b) \mod p$ et $s = (al+b) \mod p$. On remarque que $r - s \equiv a(k-l) \mod p$, avec a et $k-l$ non nuls $\mod p$. Donc, r et s sont distincts ; donc il y a $p(p-1)$ paires (r,s) possibles, ce qui correspond au nombre de choix pour (a,b). De plus, on a : $a = (r-s)((k-l)^{-1} \mod p) \mod p$ et $b = (r-ak) \mod p$. Ainsi, il y a une bijection entre les paires (a,b), avec $a \neq 0$ et les (r,s), avec $r \neq s$. Ainsi, pour $h = h_{a,b}$ choisie aléatoirement, $\mathbb{P}(h(k) = h(l)) = \mathbb{P}(R \mod m = S \mod m)$ où (R,S) est une variable aléatoire uniforme sur $(\mathbb{Z}/p\mathbb{Z})^2 \setminus \{(q,q), q \in \mathbb{Z}/p\mathbb{Z}\}$ Or, pour une valeur de r fixée, $\mathbb{P}(S \equiv r \mod m) \leqslant \lceil p/m \rceil - 1 \leqslant \frac{p-1}{m}$ Donc, $\mathbb{P}(R \equiv S \mod m) \leqslant \frac{p-1}{m(p-1)} = \frac{1}{m}$. Ainsi, $\mathbb{P}(h(k) = h(l)) \leqslant \frac{1}{m}$
Donc $\mathcal{H}_{p,m}$ est universelle.

On compte $\binom{n}{2}$ paires de clefs susceptibles d'entrer en collision. Chaque paire a une probabilité d'entrer en collision inférieure à $1/m$. Soit X une variable aléatoire qui compte le nombre de collisions, alors :

$$\begin{aligned}\mathbb{E}[X] &= \binom{n}{2}\frac{1}{n^2} \\ &= \frac{n^2-n}{2}\frac{1}{n^2} \\ &< \frac{1}{2}\end{aligned}$$

Ensuite, on applique l'inégalité de Markov :

$$\mathbb{P}(X \geqslant 1) \leqslant \frac{\mathbb{E}(X)}{1} < \frac{1}{2}$$

Donc, dans cette situation, une fonction h choisie aléatoirement a plus de chances de ne pas avoir de collisions que d'en avoir. Comme l'ensemble des clefs à hacher est statique, après quelques essais aléatoires, on trouve une fonction h sans collision.

Mais, quand n est grand, on ne veux pas que la table soit de taille n^2. On fait donc un hachage à deux niveaux : on stocke n clefs dans une table de hachage de taille $m = n$ via h choisie aléatoirement dans une classe de fonctions de hachage universelle ; puis on prend pour chaque table de hachage secondaire une taille $m_j = n_j^2$. Cette stratégie permet de faire des consultations en temps constant et sans risque de collision. De plus, on peut montrer que la mémoire utilisée est $O(n)$

Théorème 3

Si on stocke n clefs dans une table de hachage de taille $m = n$ via h choisie aléatoirement dans une classe de fonctions de hachage universelle, alors

$$\mathbb{E}\left[\sum_{j=0}^{n} n_j^2\right] < 2n$$

où n_j est la variable aléatoire qui correspond au nombre de clefs hachées dans l'alvéole j.
Donc, si on prend pour chaque table de hachage secondaire une taille $m_j = n_j^2$, la quantité moyenne de mémoire requise par toutes les tables de hachages secondaires d'une stratégie de hachage parfait est inférieure à $2n$.

Remarquons que $\forall p \in \mathbb{N}^*, p^2 = p + 2\binom{p}{2}$. On a :

$$\begin{aligned}\mathbb{E}\left[\sum_{j=0}^{n} n_j^2\right] &= \mathbb{E}\left[\sum_{j=0}^{n}\left(n_j + 2\binom{n}{2}\right)\right] && \text{(d'après la remarque précédente)} \\ &= \mathbb{E}\left[\sum_{j=0}^{n} n_j\right] + 2\mathbb{E}\left[\sum_{j=0}^{n}\binom{n}{2}\right] && \text{(par linéarité de l'espérance)} \\ &= \mathbb{E}\left[n\right] + 2\mathbb{E}\left[\sum_{j=0}^{n}\binom{n}{2}\right] \\ &= n + 2\mathbb{E}\left[\sum_{j=0}^{n}\binom{n}{2}\right]\end{aligned}$$

Or, $\sum\limits_{j=0}^{n} \binom{n}{2}$ correspond en fait au nombre de paires d'éléments qui entrent en collision. Mais comme on a choisi un hachage universel :

$$\sum_{j=0}^{n} \binom{n}{2} \leqslant \binom{n}{2} \frac{1}{m} = \frac{n(n-1)}{2m} = \frac{n-1}{2}$$

car $m = n$, d'où :

$$\mathbb{E}\left[\sum_{j=0}^{n} n_j^2\right] \leqslant n + 2\frac{n-1}{2} < 2n$$

Puis, puisqu'on prend $m_j = n_j^2$, pour $j \in [|0, m-1|]$, on obtient :

$$\mathbb{E}\left[\sum_{j=0}^{n} m_j\right] = \mathbb{E}\left[\sum_{j=0}^{n} n_j^2\right] < 2n$$

Corollaire 1

Avec la stratégie décrite précédemment, la probabilité que l'espace total consommé par les tables secondaires dépasse $4n$ est inférieure à $1/2$.

On applique ici l'inégalité de Markov :

$$\mathbb{P}\left(\sum_{j=0}^{m-1} m_j \geqslant 4n\right) \leqslant \frac{\mathbb{E}\left[\sum_{j=0}^{m-1} m_j\right]}{4n} < \frac{2n}{4n} = \frac{1}{2}$$

Références

[1] Thomas H. Cormen, *Algorithmique*. Dunod, 3è édition, 2010.