Hachage parfait

Référence : Cormen, p. 277 de la 3ème édition

2011-2012

On considère que l'ensemble des clefs est fixé.

L'idée est ici d'utiliser des tables de hachage à deux niveaux :
- Le premier niveau est identique au hachage par chainage : on hache n clefs dans m cases grâce à une fonction de hachage h bien choisie dans une famille de fonctions de hachage universelles.
- Plutôt que de créer une liste chaînée des éléments hachés en case j, on utilise une seconde table de hachage S_j associée à une fonction de hachage h_j.

En choisissant bien h_j, on peut assurer qu'il n'y aura pas de collision dans ce second niveau.

Pour cela, on supposera que la taille m_j de la table S_j est n_j^2, où n_j est le nombre de clefs hachées en case j.

On choisit h dans $\mathscr{H}_{p,m}$ (où p est une nombre premier plus grand que n'importe quelle clef), et $h_j \in \mathscr{H}_{p,m_j}$.

Montrons d'abord qu'il n'y a pas de collision au second niveau :

Théorème 1
Supposons qu'on stocke n clefs dans une table de hachage de taille $m = n^2$, par une fonction h choisie aléatoirement dans une classe de fonctions de hachage universelles.

Alors la probabilité qu'il y ait une ou plusieurs collisions est moins de $\frac{1}{2}$.

Démonstration. On a C_n^2 paires de clefs qui peuvent entrer en collision, et chaque paire entre en collision avec une probabilité $\frac{1}{m}$.

Soit X la variable aléatoire qui compte le nombre de collisions. On a :

$$\begin{aligned}\mathbb{E}X &= C_n^2 \cdot \frac{1}{n^2} \\ &= \frac{n^2 - n}{2} \cdot \frac{1}{n^2} \\ &< \frac{1}{2}\end{aligned}$$

On a alors par inégalité de Markov :

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

□

En choisissant au hasard nos fonctions de hachage, on a donc de grandes chances de tomber rapidement sur des fonctions sans collision.

Regardons maintenant l'espace occupé par cette double table.

Théorème 2
Supposons qu'on stocke n clefs dans une table de hachage de taille $m = n$ par une fonction de hachage choisie aléatoirement dans une classe de fonctions de hachage universelles. Alors :

$$\mathbb{E}\left(\sum_{j=0}^{m-1} N_j^2\right) < 2n,$$

où N_j est la variable aléatoire qui compte le nombre de clefs hachées en case j.

Démonstration. On utilise l'identité $a^2 = a + 2C_a^2$.

$$\begin{aligned}
\mathbb{E}\left(\sum_{j=0}^{m-1} N_j^2\right) &= \mathbb{E}\left(\sum_{j=0}^{m-1} N_j + 2C_{N_j}^2\right) \\
&= \mathbb{E}\left(\sum_{j=0}^{m-1} N_j\right) + 2\mathbb{E}\left(\sum_{j=0}^{m-1} C_{N_j}^2\right) \\
&= \mathbb{E}n + 2\mathbb{E}\left(\sum_{j=0}^{m-1} C_{N_j}^2\right) \\
&= n + 2\mathbb{E}\left(\sum_{j=0}^{m-1} C_{N_j}^2\right)
\end{aligned}$$

La quantité $\mathbb{E}\left(\sum_{j=0}^{m-1} C_{N_j}^2\right)$ représente le nombre total de paires de clefs qui entrent en collision. Par propriété du hachage universel, la somme vaut au plus :

$$\begin{aligned}
C_n^2 \frac{1}{m} &= \frac{n(n-1)}{2m} \\
&= \frac{n-1}{2}
\end{aligned}$$

Donc on a

$$\begin{aligned}
\mathbb{E}\left(\sum_{j=0}^{m-1} N_j^2\right) &\leqslant n + 2\frac{n-1}{2} \\
&= 2n - 1 < 2n
\end{aligned}$$

□

Corollaire 3
Supposons qu'on stocke n clefs dans une table de hachage de taille $m = n$ par une fonction de hachage choisie aléatoirement dans une classe de fonctions de hachage universelles.

On choisit les tailles des tables secondaires comme $m_j = n_j^2$.

Alors l'espace mémoire utilisé pour toutes les tables secondaires dans le hachage parfait est en moyenne inférieur à $2n$.

Démonstration. $m_j = n_j^2$, donc c'est immédiat. □

Corollaire 4
Supposons qu'on stocke n clefs dans une table de hachage de taille $m = n$ par une fonction de hachage choisie aléatoirement dans une classe de fonctions de hachage universelles.

On choisit les tailles des tables secondaires comme $m_j = n_j^2$.

Alors, la probabilité que l'espace mémoire utilisé pour les tables secondaires soit supérieur ou égal à $4n$ est moins de $\frac{1}{2}$.

Démonstration. On applique l'inégalité de Markov au corollaire précédent :

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

□

Donc, en testant des fonctions de hachage aléatoirement dans une famille universelle, on en trouvera rapidement une qui utilise un espace de stockage raisonnable.