Développement : Hachage parfait

Pierron Théo-Lacoste Cyril

17 mars 2014

Référence : Cormen, p.258

Si l'on dispose d'un ensemble de n clés statiques (représentées par des entiers), on peut construire une table de hachage de taille $O(n)$ et telle que la recherche nécessite $O(1)$ opérations dans le pire des cas (on ne s'intéresse ni à l'insertion ni à la suppression car l'ensemble des clés est statique).

Nous allons pour cela procéder en deux temps :

1. Complexité temporelle : si on utilise une table de taille $m = n^2$ on peut trouver une fonction de hachage telle qu'il n'y ait aucune collision, donc une recherche en $O(1)$.
2. Complexité spatiale : au lieu de considérer une seule table de taille n^2, on prendra une table primaire de taille n, puis pour chaque alvéole j, une table secondaire de taille $m_j = n_j^2$ où n_j est le nombre de clés hachées vers l'alvéole j, et on s'assurera qu'il n'y a pas de collision dans les tables secondaires (possible par le point 1). On peut trouver dans ce cas des fonctions de hachage telles que l'espace mémoire total utilisé soit en $O(n)$, et la recherche sera toujours en $O(1)$.

Nous choisirons nos fonctions de hachage aléatoirement dans des classes universelles de fonctions de hachage, c'est-à-dire que pour tout couple de clés distinctes, la probabilité d'avoir une collision entre ces clés est de $\frac{1}{m}$ où m est la taille de la table. On peut par exemple considérer un nombre premier p strictement supérieur à toutes les clés, et la classe $\mathcal{H}_{p,m}$ constituée des fonctions $h_{a,b}$ définies par :

$$h_{a,b}(k) = ((ak + b)[p])[m]$$

pour $a \in \{1, \ldots, p - 1\}$ et $b \in \{0, \ldots, p - 1\}$.

Proposition 0.1 Si on stocke n clés dans une table de taille $m = n^2$ via une fonction h choisie aléatoirement dans une classe universelle, alors la probabilité d'avoir une collision est inférieure à $\frac{1}{2}$.

Démonstration. Il y a $\binom{n}{2}$ paires de clés susceptibles d'entrer en collision, chaque paire ayant une probabilité de collision de $\frac{1}{m}$ si la fonction de hachage

h est aléatoire. Ainsi si X est la variable aléatoire qui compte le nombre de collisions :

$$E[X] = \binom{n}{2}.\frac{1}{m} = \frac{n(n-1)}{2m} = \frac{n-1}{2n} < \frac{1}{2}$$

Donc par l'inégalité de Markov :

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

□

Ainsi en testant plusieurs fonctions de hachage pour chaque table secondaire, on tombera rapidement sur des tables sans collision.

Proposition 0.2 En procédant comme indiqué en introduction, la mémoire totale utilisée est $O(n)$ avec une probabilité supérieure à $\frac{1}{2}$.

Démonstration. Il suffit de montrer qu'il existe une constante K telle que, en espérance, $\sum_{j=1}^{n} n_j^2 \leqslant K.n$. On va voir que $K = 2$ convient. En utilisant l'identité : $n_j^2 = n_j + 2\binom{n_j}{2}$ on a par linéarité de l'espérance :

$$\begin{aligned} E\left[\sum_{j=1}^{n} n_j^2\right] &= E\left[\sum_{j=1}^{n} n_j\right] + 2E\left[\sum_{j=1}^{n} \binom{n_j}{2}\right] \\ &= n + 2\frac{n-1}{2} \\ &= 2n - 1 \\ &< 2n \end{aligned}$$

La deuxième égalité provient du fait que le terme de droite est le nombre de collisions dans la table primaire et l'on réutilise le résultat de la proposition précédente (avec dans ce cas $m = n$). Donc à nouveau par l'inégalité de Markov :

$$P\left(\sum_{j=1}^{n} n_j^2 \geqslant 4n\right) < \frac{2n}{4n} = \frac{1}{2}$$

□