ALGO1 – Table de hachage

François Schwarzentruber

September 19, 2020

1 Motivation

Rêve : implémenter efficacement les types abstraits ensemble et tableau associatif

1.1 Algorithmique

Somme de nombres On veut par exemple savoir s'il existe deux nombres a et b dans un tableau T tel que $a + b = 0$. Voici un algorithme en $O(n)$:

- On met tous les éléments dans une structure de données avec un accès en $O(1)$;
- On parcourt les éléments a du tableau et on teste si $-a$ se trouve dans la structure de données.

Relation symétrique Tester si une relation donnée sous la forme d'une liste de paires (a, b) est symétrique.

1.2 Texte : mots les plus fréquents

Étant donné un roman, quels sont les 100 mots les plus fréquents ?

1.3 Base de données

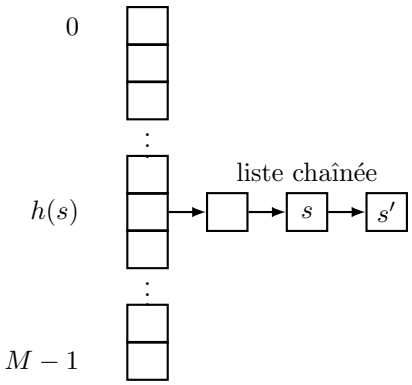
On a une liste de triplets $(prixapayer, client)$. On souhaite en temps linéaire calculer pour chaque client le prix total qu'il doit payer.

1.4 Dans un compilateur/interpréteur

Pour stocker l'association entre nom de variables et données.

2 Définition

Définition 1 *Une table de hachage par chaînage est une structure de données composée d'un tableau avec M cases, chacune contenant une liste chaînée.*



fonction contient(s)
  retourner liste_contient($T[h(s)], s$)

fonction ajouter(s)
  retourner liste_ajouter($T[h(s)], s$)

fonction supprimer(s)
  retourner liste_supprimer($T[h(s)], s$)

Définition 2 (alvéole) *Une case du tableau s'appelle une alvéole.*

Définition 3 (collision) *Une collision est un ensemble de 2 clefs logées dans la même alvéole.*

Notations

Ω	= l'univers des mondes possibles (pour parler de probabilités)
$\mathbb{K}$	= l'ensemble des clefs possibles (pas forcément dans la table)
n	= nombre de clefs présentes dans la table
M	= nombre d'alvéoles dans la table
h	= fonction de hachage
a_j	= nombre de clefs dans l'alvéole numéro j
$\alpha = \frac{n}{M}$	= facteur de remplissage

3 Hachage uniforme

Définition 4 (Hypothèse du hachage uniforme simple) *Soit $K : \Omega \to \mathbb{K}$ est une variable aléatoire représentant une clef. La clef K vérifie l'hypothèse de hachage uniforme simple si pour tout alvéole $j \in \{0, \dots, M-1\}$,*

$$\mathbb{P}(h(K) = j) = \frac{1}{M}.$$

demo: Hachage uniforme

Dans ce cas là, on a la proposition suivante, avec $\leq$ car les éléments qu'on ajoute peuvent être égaux et il n'y a pas de doublons :

Proposition 5 *Supposons que l'on a ajouté n éléments : $K_1, \dots, K_n : \Omega \to \mathbb{K}$ sous l'hypothèse de hachage uniforme simple. On a alors pour tout $j \in \{0, \dots, M-1\}$,*

$$\mathbb{E}(a_j) \leq \alpha = \frac{n}{M}.$$

Démonstration. Soit $\mathcal{P}$ une propriété. On note $1_{\mathcal{P}}$ la fonction de $\Omega \to \{0, 1\}$ qui vaut 1 si la propriété $\mathcal{P}$ est vraie et 0 sinon.

Le nombre de tentatives d'ajout d'une clef dans l'alvéole numéro j vaut $\sum_{i=1}^{n} 1_{h(K_i)=j}$. On a donc :

$$a_j \leq \sum_{i=1}^{n} 1_{h(K_i)=j}.$$

Le $\leq$ provient du fait qu'il peut y avoir des clefs égales. Par linéarité de l'espérance :

$$\begin{array}{ll} \mathbb{E}(a_j) & \leq \mathbb{E}(\sum_{i=1}^{n} 1_{h(K_i)=j}) \\ & = \sum_{i=1}^{n} \mathbb{E}(1_{h(K_i)=j}) \\ & = \sum_{i=1}^{n} \mathbb{P}(h(K_i) = j) \\ & = \sum_{i=1}^{n} \frac{1}{M} \\ & = \frac{n}{M} \end{array}$$

■

Corollaire 6 *Les opérations d'ajout, suppression et recherche sont $O(1+\alpha)$ en moyenne.*

4 Hachage universel

Problème : Un intrus qui connaît la fonction de hachage peut attaquer la table de hachage.

demo: Hachage uniforme : attaque

Solution : choisir aléatoirement une fonction de hachage H.

Exemple 7 *Soit p un nombre premier et supposons $\mathbb{K} = \mathbb{Z}/_{p\mathbb{Z}}$. Soit $\mathcal{D} = \mathbb{Z}/^{*}_{p\mathbb{Z}} \times \mathbb{Z}/_{p\mathbb{Z}}$.*

Tirer aléatoirement uniformément (a, b) dans $\mathcal{D}$ et choisir comme fonction de hachage

$$\begin{array}{rccl} h_{ab} : & \mathbb{Z}/_{p\mathbb{Z}} & \to & \{0, \dots, M-1\} \\ & k & \mapsto & ((ak+b)[p])[M]. \end{array}$$

demo: Hachage universel

4.1 Fonction de hachage aléatoire universelle

Définition 8 *Soit $H : \Omega \to (\mathbb{K} \to \{1, \ldots, M\})$ une fonction de hachage aléatoire est universelle si pour toute paire de clef $(k, \ell) \in \mathbb{K}$ tel que $k \neq \ell$, on a*

$$\mathbb{P}(H(k) = H(\ell)) \leq \frac{1}{M}.$$

Théorème 9 *Supposons que H est une fonction de hachage aléatoire universelle. Alors pour tout éléments $x_1, \ldots, x_n$ distincts deux à deux, pour tout x, après l'insertion de $x_1, \ldots, x_n$ on a :*

$$\mathbb{E}(a_{H(x)}) \leq \frac{n}{M} + 1.$$

Démonstration.

$$\begin{array}{rl} \mathbb{E}(a_{H(x)}) & = \mathbb{E}(\sum_{i=1}^{n} 1_{H(x_i)=H(x)}) \\ & = \sum_{i=1}^{n} \mathbb{E}(1_{H(x_i)=H(x)}) \\ & = \sum_{i=1}^{n} \mathbb{P}(H(x_i) = H(x)) \end{array}$$

Si $x_i = x$, $\mathbb{P}(H(x_i) = H(x)) = 1$. Si $x_i \neq x$, $\mathbb{P}(H(x_i) = H(x)) \leq \frac{1}{M}$.
D'où le résultat. ■

4.2 L'exemple est une classe universelle (*)

Théorème 10 *(admis) Soit $\mathcal{D} = \mathbb{Z}/_{p\mathbb{Z}}^{*} \times \mathbb{Z}/_{p\mathbb{Z}}$. Si on tire aléatoirement uniformément (a, b) dans $\mathcal{D}$, alors la variable aléatoire h_{ab} de l'exemple 7 est universelle.*

Démonstration. Rappel :

$$\begin{array}{rccl} h_{ab} : & \mathbb{Z}/_{p\mathbb{Z}} & \to & \{0, \ldots, M-1\} \\ & k & \mapsto & (\underbrace{(ak+b)[p]}_{f_{ab}(k)})[M]. \end{array}$$

Par définition, montrer que h_{ab} est universelle revient à montrer que pour toute paire de clef (k, ℓ) tel que $k \neq \ell$, on a

$$|\{(a, b) \in \mathcal{D} \mid h_{ab}(k) = h_{ab}(\ell)\}| \leq \frac{|\mathcal{D}|}{M}.$$

Il faut montrer que pour toute paire de clef (k, ℓ) tel que $k \neq \ell$, on a

$$|\{(a, b) \in \mathcal{D} \mid f_{ab}(k) = f_{ab}(\ell)[M]\}| \leq \frac{|\mathcal{D}|}{M}.$$

Soit $\mathcal{S} = \{(r, s) \in \mathbb{Z}/_{p\mathbb{Z}}^{2} \mid r \neq s\}$. Soit $k, \ell \in U$ telles que $k \neq \ell$. Posons

$$\begin{array}{rccl} \theta : & \mathcal{D} & \to & \mathcal{S} \\ & (a, b) & \mapsto & (f_{ab}(k), f_{ab}(\ell)) \end{array}$$

Lemme 11 *θ est une bijection.*

Démonstration. Bien défini θ est bien à valeurs dans $\mathcal{S}$. En effet :

- D'une part, $f_{ab}(k), f_{ab}(\ell) \in \mathbb{Z}/_{p\mathbb{Z}}$ par définition de f_{ab}.
- D'autre part $f_{ab}(k) \neq f_{ab}(\ell)$. En effet, par définition de $f_{ab}(..)$, il faut montrer que $ak + b \neq a\ell + b[p]$. Autrement dit, il faut montrer que $a(k - \ell) \neq 0[p]$. Mais :
  - $a \neq 0$ (par définition de $\mathcal{D}$).
  - Et comme p est plus grand que les valeurs des clefs (ie. $k < p$ et $\ell < p$), et $k - \ell \neq 0$ implique que $(k - \ell) \neq 0[p]$.

  Comme $\mathbb{Z}/_{p\mathbb{Z}}$ est intègre, on a bien $a(k - \ell) \neq 0[p]$. CQFD.

Surjection
Pour tout $(r, s) \in \mathcal{S}$, montrons qu'il existe $(a, b) \in \mathcal{D}$ tel que $\theta(a, b) = (r, s)$. Le candidat (a, b) est :

- $a = (r-s)(k-\ell)^{-1}[p]$;
- $b = (r - ak)[p]$.

On remarque que comme $k \neq \ell$ et comme $r \neq s$, alors par intégrité de $\mathbb{Z}/_{p\mathbb{Z}}$, $a \neq 0[p]$. Donc $(a,b) \in \mathcal{D}$. D'autre part,

$$\begin{aligned}
\theta(a,b) &= (ak+b, a\ell+b) \\
&= (ak + \underbrace{(r-ak)}_{b}, \quad \underbrace{(r-s)(k-\ell)^{-1}}_{a}\ell + r - \underbrace{\underbrace{(r-s)(k-\ell)^{-1}}_{a}k}_{b}) \\
&= (r, \quad (r-s)(k-\ell)^{-1}(\ell-k) + r) \\
&= (r, \quad s - r + r) \\
&= (r,s).
\end{aligned}$$

Conclusion
Comme $|\mathcal{D}| = |\mathcal{S}|$, on en conclut que θ est une bijection.
■
Vu que θ est une bijection, on a :

$$|\{(a,b) \in \mathcal{D} \mid f_{ab}(k) = f_{ab}(\ell)[M]\}| = |\{(r,s) \in \mathcal{S} \mid r = s[M]\}|.$$

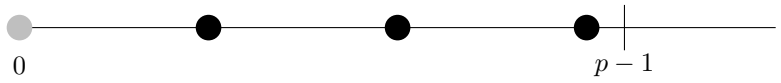
Ainsi, l'objectif est de montrer que :

$$|\{(r,s) \in \mathcal{S} \mid r = s[M]\}| \leq \frac{|\mathcal{S}|}{M}.$$

Pour un $r \in \mathbb{Z}/_{p\mathbb{Z}}$ fixé,
$|\{s \in \mathbb{Z}/_{p\mathbb{Z}} \mid (r,s) \in \mathcal{S} \text{ et } r = s[M]\}|$

$$\begin{aligned}
&= |\{s \in \mathbb{Z}/_{p\mathbb{Z}} \mid (r,s) \in \mathbb{Z}/^2_{p\mathbb{Z}} \text{ et } r \neq s[p] \text{ et } r = s[M]\}| \\
&\leq \left\lceil \tfrac{p}{M} \right\rceil - 1 \\
&\leq \tfrac{p+M-1}{M} - 1 = \tfrac{p-1}{M}
\end{aligned}$$

Exemple 12 *Par exemple pour $r = 0$, voici les candidats (en noir) espacés de M :*



Ainsi,

$$|\{(r,s) \in \mathcal{S} \mid r = s[M]\}| \leq p\frac{p-1}{M} = \frac{|\mathcal{S}|}{M}$$

■

4.3 Nombre de collisions

Lemme 13 *Soit H une fonction de hachage aléatoire avec l'hypothèse d'universalité. On a :*

$$\mathbb{E}(nb \text{ de collisions obtenues avec } H) \leq \binom{n}{2}\frac{1}{M}.$$

DÉMONSTRATION. Etant donné deux clefs k et ℓ, on définit

$$X_{k,\ell} = 1_{k \text{ et } \ell \text{ donne une collision avec } H} = 1_{H(k)=H(\ell)}.$$

On a, en notant X le nombre de collisions obtenues avec H :

$$X = \sum_{\{k,\ell\} \text{ où } k,\ell \text{ dans la table } |k\neq\ell} X_{k,\ell}.$$

Par linéarité de l'espérance on a :

$$\mathbb{E}(X) = \sum_{\{k,\ell\} \text{ où } k,\ell \text{ dans la table } |k\neq\ell} \mathbb{E}(X_{k,\ell})$$

Mais

$$\begin{aligned} \mathbb{E}(X_{k,\ell}) &= \mathbb{P}(X_{k,\ell} = 1) \\ &= \mathbb{P}(H(k) = H(\ell)) \\ &\leq \tfrac{1}{M} && \text{car } H \text{ est universelle.} \end{aligned}$$

On a :

$$\mathbb{E}(X) = \sum_{\{k,\ell\} \text{ où } k,\ell \text{ dans la table } |k\neq\ell} \mathbb{E}(X_{k,\ell}) \leq \sum_{\{k,\ell\} \text{ où } k,\ell \text{ dans la table } |k\neq\ell} \frac{1}{M} \leq \binom{n}{2}\frac{1}{M}.$$

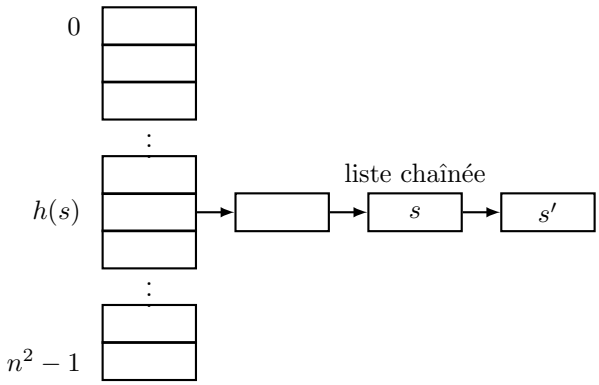
■

5 Hachage parfait

Problème : les complexités sont en moyenne mais pas dans le pire cas.
Solution : étant donné n clés **fixées et connues**, *compiler* une structure de données pour avoir un accès pire cas en $O(1)$ et un espace mémoire en $O(n)$

5.1 Etape 1 : table avec accès $O(1)$ et taille $O(n^2)$

demo: Hachage parfait



Théorème 14 *Posons $M = n^2$. Soit H une fonction de hachage choisie aléatoire l'hypothèse d'universalité. Alors :*

$$\mathbb{P}(\textit{pas de collision avec } H) > \frac{1}{2}.$$

DÉMONSTRATION. Soit X le nombre de collisions que l'on a avec H. Via le lemme 11, on a :

$$\mathbb{E}(X) \leq \frac{n(n-1)}{2}\frac{1}{n^2} < \frac{1}{2}.$$

Rappel de l'inégalité de Markov : $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$.

Ici, pour $t = 1$, on obtient $\mathbb{P}(X \geq 1) \leq \frac{1}{2}.$, et donc $\mathbb{P}(X = 0) > \frac{1}{2}$. ■
Solution : tirer aléatoirement h dans $\mathcal{H}$ et s'arrêter quand on n'a pas de collision.

Théorème 15 *Il faut en moyenne $\frac{1-p}{p}$ essais pour avoir une fonction de hachage qui ne donne pas de collisions, où p est la probabilité de ne pas avoir de collisions.*

DÉMONSTRATION. Soit T le nombre d'essais ratés jusqu'à ne pas avoir de collision.
$\mathbb{P}(T = 0) = p$
$\mathbb{P}(T = 1) = (1-p)p$
$\vdots$
$\mathbb{P}(T = k) = (1-p)^k p$
$\vdots$

Rappel que : $\sum_0^\infty kx^k := \frac{x}{(1-x)^2}$

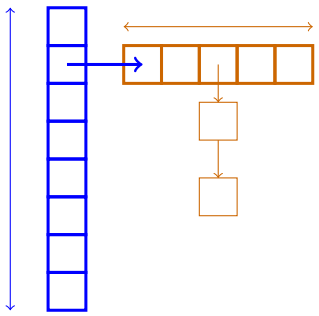
Ainsi, $\mathbb{E}(T) = \sum_0^\infty k\mathbb{P}(T = k) = \sum_0^\infty k(1-p)^k p = p\frac{1-p}{(1-(1-p))^2} = \frac{1-p}{p}$.
La fonction $p \mapsto \frac{1-p}{p}$ est décroissante en p et vaut 1 en $\frac{1}{2}$. Il faut donc en moyenne "moins de 1 tirage" pour ne pas avoir de collision. ■

5.2 Etape 2 : obtenir une taille en $O(n)$

On construit une structure à deux niveaux :

1. Premier niveau avec $M = n$ alvéoles, avec une fonction de hachage h. Il peut y avoir des collisions.
2. Dans l'alvéole numéro j du premier niveau, on met une table de hachage sans collision T_j. On utilise le principe de la sous-section précédente, i.e. on met ${a_j}^2$ sous-alvéoles dans T_j où a_j = le nombre d'éléments dans l'alvéole numéro j.

   On montre que l'on peut choisir h du premier niveau tel que l'espace du deuxième niveau $\sum_{j=0}^{M-1} {a_j}^2 = O(n)$.



Théorème 16 *Posons $M = n$. Soit H une fonction de hachage aléatoire universelle. Soit ${a_j}^H$ la variable aléatoire qui représente le nombre d'éléments dans l'alvéole numéro j au premier niveau avec H. Alors :*

$$\mathbb{E}(\sum_{j=0}^{M-1} \left({a_j}^H\right)^2) < 2n.$$

DÉMONSTRATION. On va calculer $\sum_{j=0}^{M-1} \left({a_j}^H\right)^2$.

Rappel : on a l'égalité pour tout $a \in \mathbb{N}$: $a^2 = a + 2\binom{a}{2}$.

On a :

$$\sum_{j=0}^{M-1} \left({a_j}^H\right)^2 = \underbrace{\sum_{j=0}^{M-1} {a_j}^H}_{n} + 2 \underbrace{\sum_{j=0}^{M-1} \binom{{a_j}^H}{2}}_{X}$$

En fait, X est le nombre de collisions dans la table de hachage principale. En effet, $\binom{{a_j}^H}{2}$ est le nombre de couples de clefs $\{k, \ell\}$ avec $k \neq \ell$ où $H(k) = H(\ell) = j$. En faisant la somme, on compte toutes les collisions.

Ainsi, on a :

$$\begin{array}{lll} \mathbb{E}(\sum_{j=0}^{M-1} {a_j}^{H\,2}) & \leq n + 2\mathbb{E}(X) & \text{par la linéairité de l'espérance} \\ & \leq n + 2\binom{n}{2}\frac{1}{n} & \text{par le lemme 11} \\ & \leq n + 2\frac{n-1}{2} & \\ & < 2n. & \end{array}$$

■

5.3 Construction de la structure

1. On tire au hasard h pour le premier niveau jusqu'à avoir $\sum_{j=0}^{M-1} {a_j}^2 < 4n$.

   La probabilité d'avoir $\sum_{j=0}^{M-1} {a_j}^2 < 4n$ est minoré par

$$\begin{array}{ll} \mathbb{P}(\sum_{j=0}^{M-1} {a_j}^2 < 4n) & = 1 - \mathbb{P}(\sum_{j=0}^{M-1} {a_j}^2 \geq 4n) \\ & \geq 1 - \frac{\mathbb{E}(\sum_{j=0}^{M-1} {a_j}^{H\,2})}{4n} \text{inégalité de Markov} \\ & > 1 - \frac{2n}{4n} \\ & > \frac{1}{2}. \end{array}$$

2. Pour tout j, pour la table de hachage secondaire dans l'alvéole numéro j, on choisit une fonction de hachage h_j qui ne donne pas de collision (cf. sous-section 4.1).

6 Notes bibliographiques

Cette note de cours résume le chapitre sur les tables de hachage de [CLRS09]. A la fin du chapitre, ils indiquent que Donald Knuth attribue l'inention des tables de hachage à Hans Peter Luhn. Le hachage parfait avec des clefs statiques a été proposé dans [FKS84]. Le cas dynamique, avec des ajouts et suppressions en $O(1)$ en cout amorti est proposé dans [DKM+94].

References

[CLRS09] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2009.

[DKM+94] Martin Dietzfelbinger, Anna R. Karlin, Kurt Mehlhorn, Friedhelm Meyer auf der Heide, Hans Rohnert, and Robert Endre Tarjan. Dynamic perfect hashing: Upper and lower bounds. *SIAM J. Comput.*, 23(4):738–761, 1994.

[FKS84] Michael L. Fredman, János Komlós, and Endre Szemerédi. Storing a sparse table with 0(1) worst case access time. *J. ACM*, 31(3):538–544, 1984.