TP 9 : Tables de hachage

Soient K et E deux ensembles, les éléments de K étant appelés *clefs*. Une *table d'association* (ou dictionnaire) m de K vers E est une partie de $K \times E$ telle que pour toute clef $k \in K$ il existe au plus un $e \in E$ tel que le couple (k, e) soit dans m. Une implémentation simple d'une telle table d'association peut naturellement être effectuée à l'aide d'une liste d'association formée de couples de $K \times E$.

Les tables de hachage (*hash tables* en anglais) sont une implémentation souvent plus efficace d'une telle structure de données. L'idée est la suivante : pour chaque clef k, on calcule un *entier de hachage* $h_w(k)$ compris entre 0 et $w-1$ (w est appelé la *largeur* de la table). On utilise ensuite un tableau de w listes pour stocker les enregistrements : la liste numéro i contenant les couples (k, e) de la table d'association tels que $h_w(k) = i$.

1 Fonctions de hachage

Pour implémenter une table de hachage, il est nécessaire de disposer d'une fonction de hachage h_w de K vers $[\![0, w-1]\!]$. Pour que le hachage soit efficace, il est utile que cette fonction assure une bonne répartition des clefs dans les différentes cases de la table, c'est à dire, de manière informelle, que étant donné un entier i de $[\![0, w-1]\!]$, la probabilité que $h_w(k) = i$ soit voisine de $1/w$. Dans cette section, nous donnons quelques exemples de telles fonctions.

1.1 Entiers naturels

Pour le cas où les clefs sont des entiers, le *hachage par division* de largeur w consiste à hacher la clef entière k en le reste de la division de k par w ($k \bmod w$) qui appartient bien à l'intervalle $[\![0, w-1]\!]$.

QUESTION 1 Écrivez une fonction `hash_int : int -> int -> int` telle que `hash_int w k` retourne le haché de la clef entière k en utilisant un hachage par division de largeur `w`.

Il arrive que la répartition des clefs ne permette pas au hachage par division de donner de bons résultats. Dans ce cas, on peut utiliser un *hachage par multiplication*. Pour cela, on choisit un entier a fixé et on définit h_w par

$$h_w(k) = \left\lfloor \frac{w(ak \bmod \mathrm{card}(K))}{\mathrm{card}(K)} \right\rfloor$$

1.2 Chaînes de caractères

Dans le cas où les clefs sont des chaînes de caractères, on peut se ramener au cas des entiers en utilisant le code ASCII des caractères (le code ASCII d'un caractère est un entier compris entre 0 et 255 identifiant le caractère de manière unique). Une chaîne $s = s_0 \dots s_{n-1}$ peut alors être vue comme la représentation en base 256 de l'entier

$$\sum_{k=0}^{n-1} \text{code}(s_k) \times 256^k$$

Le code d'un caractère est retourné en Caml par la fonction `int_of_char` (de type `char -> int`).

QUESTION 2 Écrivez une fonction `convert_string : string -> int` qui retourne l'entier représenté par une chaîne de caractères comme expliqué ci-dessus.

QUESTION 3 Déduisez-en une fonction de hachage (par division) des chaînes de caractères `hash_string: int -> string -> int`.

2 Tables de hachage de taille fixe

Nous représentons en Caml une table de hachage de clefs de type `'a` vers des valeurs de type `'b` par un enregistrement du type suivant :

```
type ('a, 'b) hashtbl =
    { hash: 'a -> int;
      data: ('a * 'b) list vect
    }
  ;;
```

Soit `m` une table de hachage de largeur w. `m.hash` est la fonction de hachage utilisée pour hacher les clefs stockées dans la table (notée dans l'énoncée h_w) et `m.data` est un tableau de longueur w. La case numéro i de ce tableau contient la liste des entrées (k, e) de la table tels que $h_w(k) = i$.

QUESTION 4 Écrivez une fonction `create : ('a -> int) -> int -> ('a, 'b) hashtbl` telle que `create h w` retourne une nouvelle table de hachage vide de largeur `w` utilisant la fonction de hachage `h`.

Nous allons maintenant écrire des fonctions permettant d'effectuer les opérations de base sur ces tables : recherche, ajout et suppression.

QUESTION 5 Écrivez une fonction `mem : ('a, 'b) hashtbl -> 'a -> bool` telle que `mem m k` rende un booléen indiquant si la clef `k` est présente dans la table `m`.

QUESTION 6 Écrivez une fonction `find : ('a, 'b) hashtbl -> 'a -> 'b` telle que `find m k` retourne la valeur `e` associé à la clef `k` dans la table `m`. Si la clef `k` n'est pas présente dans la table `m`, votre fonction lèvera l'exception `Not_found`.

QUESTION 7 Écrivez une fonction `add : ('a, 'b) hashtbl -> 'a -> 'b -> unit` telle que `add m k e` ajoute l'entrée (k, e) à la table de hachage m. (Vous préciserez le comportement de votre fonction si la clef k est déjà présente dans la table.)

QUESTION 8 Écrivez enfin une fonction `remove` telle que `remove m k : ('a, 'b) hashtbl -> 'a -> unit` supprime l'entrée de la clef `k` dans la table `m`. (Vous préciserez le comportement de votre fonction si la clef k n'est pas présente dans la table.)

3 Tables de hachage de taille dynamique

On constate en général que la partie coûteuse de la recherche d'une entrée dans la table est le parcours de la liste des enregistrements correspondant à la valeur de hachage de la clef considérée. Dans les tables que nous avons considéré jusqu'à présent, la largeur est fixée une fois pour toutes au moment de la création de la table. Ainsi, au fur et à mesure que l'on ajoute des entrées, la longueur des listes est susceptible d'augmenter et, par conséquent, le coût des recherches.

Dans cette section, on se propose d'améliorer ce point en utilisant des tables de hachage de taille dynamique : l'idée est d'augmenter la largeur de la table dès lors qu'il y a trop d'éléments. Ainsi, au fur et à mesure de l'ajout d'entrées, on garde (en moyenne) des listes courtes dans lesquelles la recherche est rapide.

Pour cela, on définit une nouvelle représentation des tables de hachage :

```
type ('a, 'b) dyn_hashtbl =
    { hash: int -> 'a -> int;
      mutable size: int;
      mutable data: ('a * 'b) list vect
    }
  ;;
```

Dans cette nouvelle représentation, on dispose d'un champ supplémentaire, `size`, qui permet de stocker le nombre d'entrées de la table. Ce champ est déclaré *mutable* de manière à pouvoir être mis à jour à chaque ajout ou suppression. Vous noterez également que la fonction de hachage d'une table de hachage dynamique prend un argument supplémentaire : la largeur de hachage.

QUESTION 9 Écrivez une fonction `dyn_create : (int -> 'a -> int) -> int -> ('a, 'b) dyn_hashtbl` permettant de créer une table de hachage dynamique. Les arguments de cette fonction seront la fonction de hachage et la largeur initiale. Écrivez ensuite deux fonctions `dyn_mem : ('a, 'b) dyn_hashtbl -> 'a -> bool` et `dyn_find : ('a, 'b) dyn_hashtbl -> 'a -> 'b`.

Nous utiliserons le principe de redimensionnement suivant : lors de l'ajout d'une entrée, si la taille de la table (i.e. le nombre d'entrées) dépasse le double de la largeur courant w, alors on réarrange la table sur une largeur $2w$ avant de procéder à l'ajout.

QUESTION 10 Écrivez une fonction `dyn_rearrange : ('a, 'b) dyn_hashtbl -> unit` qui réarrange une table en doublant sa largeur.

QUESTION 11 Déduisez-en une fonction `add_dyn : ('a, 'b) dyn_hashtbl -> 'a -> 'b -> unit` telle que `add_dyn m k e` ajoute l'entrée (k, e) à la table `m` en effectuant un réarrangement si nécessaire.

On s'intéresse maintenant à la suppression d'une entrée.

QUESTION 12 Que pensez-vous de la stratégie consistant à réarranger une table lors d'une suppression si sa taille devient inférieure à la moitié de sa largeur. Quelle autre solution proposez-vous ? Implémentez ainsi une fonction `dyn_remove : ('a, 'b) dyn_hashtbl -> 'a -> unit` supprimant une entrée d'une table dynamique.