

Chapitre 11 : Algorithmique du texte

Table des matières

1	Recherche dans un texte	2
1.1	Introduction	2
1.1.1	Problème	2
1.1.2	Algorithme naïf	2
1.1.3	Remarque	2
1.2	Algorithme de BOYER-MOORE	3
1.2.1	Introduction	3
1.2.2	Algorithmique de BOYER-MOORE-HORSPPOOL	3
1.2.3	Algorithme de BOYER-MOORE (simplifié)	4
1.2.4	Algorithme de BOYER-MOORE	5
1.3	Algorithme de KARP-RABIN	7
1.3.1	Principe	7
1.3.2	Choix de la fonction de hachage	7
1.3.3	Implémentation en OCaml	8
1.3.4	Étude de la complexité	9
1.3.5	Extension à la recherche de plusieurs motifs	10
2	Algorithmes de compression	10
2.1	Contexte	10
2.1.1	Principe	10
2.1.2	Remarque	10
2.1.3	Algorithmes au programme	10
2.2	Algorithme de HUFFMAN	11
2.2.1	Principe	11
2.2.2	Représentation de l'encodage des symboles	11
2.2.3	Encodage et décodage	12
2.2.4	Arbre de HUFFMAN	13
2.2.5	Optimalité de l'arbre de HUFFMAN	15
2.2.6	Aspects pratiques	17
2.3	Algorithme de LEMPEL-ZIV-WELCH (LZW)	19
2.3.1	Introduction	19
2.3.2	Principe	19
2.3.3	Compression	19
2.3.4	Décompression	20

1 Recherche dans un texte

1.1 Introduction

1.1.1 Problème

Étant donné un texte (un fichier / une chaîne de caractères) t , et un motif (une chaîne de caractères) x , on veut trouver toutes les occurrences de x dans t (*i.e* on veut les positions (indices)).

1.1.2 Algorithme naïf

On teste toutes les positions dans t pour déterminer si ce sont des positions d'occurrences de x :

```

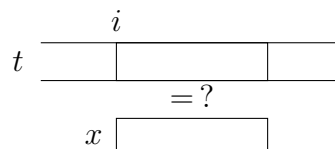
1 | let recherche_naive (x : string) (t : string) : int list =
2 |   let l = ref [] in
3 |   let n = String.length t
4 |   and m = String.length x in
5 |
6 |   for i = 0 to n - m do
7 |     let j = ref 0 in
8 |     while !j < m && x.[!j] = t.[i + !j] do
9 |       incr j
10 |    done;
11 |
12 |    if !j = m then l := i::!l
13 |  done;
14 |  !l

```

Complexité : dans le pire cas, la boucle `while` s'arrête au dernier indice dans x (exemple : $t = a^n = \underbrace{a \cdots a}_{n \text{ fois } a}$, et $x = a^{m-1}b$) : $\mathcal{O}((n - m + 1)m) = \mathcal{O}(nm)$ si m est petit devant n .

1.1.3 Remarque

Cet algorithme fait partie d'une famille d'algorithmes, dits de *fenêtre glissante* : on fait glisser une fenêtre sur le texte en notant toutes les positions auxquelles la fenêtre contient le motif.



Le côté naïf de cet algorithme vient du fait que l'on fait systématiquement glisser la fenêtre d'un rang, quel que soit son contenu. En analysant les raisons de l'échec d'une comparaison, on peut espérer décaler la fenêtre de plusieurs rangs.

1.2 Algorithme de BOYER-MOORE

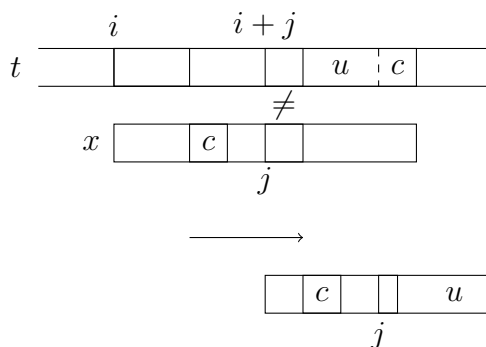
1.2.1 Introduction

L'algorithme de BOYER-MOORE est un algorithme de fenêtre glissante, mais la comparaison entre le contenu de la fenêtre et le motif se fait de la droite vers la gauche, afin de repérer en cas d'échec le caractère le plus à droite qui ne correspond pas au motif. À partir de cette position qui entraîne l'échec de la comparaison, on peut calculer un décalage pour la fenêtre.

En pratique le décalage est précalculé pour toutes les positions du motif et il existe de nombreuses variantes suivant la manière dont le décalage est calculé.

1.2.2 Algorithmique de BOYER-MOORE-HORSPOL

Dans cette variante, si la comparaison de $x_0 \dots x_{m-1}$ et $t_i \dots t_{i+m-1}$ (la fenêtre) échoue à l'indice j , i.e $x_{j+1} \dots x_{m-1} = t_{i+j+1} \dots t_{i+m-1}$ et $x_j \neq t_{i+j}$, on cherche à aligner t_{i+m-1} avec son occurrence la plus à droite dans x (sauf x_{m-1})



Exemple : $x = aababab$ et $t = aabbbababacaabbaba$

$aababab$
 \times
 $aababab$
 \times
 $aababab$
 \times

Algorithme : pour toute lettre a , on note $d(a)$ le décalage à effectuer pour aligner cette lettre avec son occurrence la plus à droite dans x (sauf la dernière lettre) en cas d'échec d'une comparaison avec une fenêtre dont la dernière lettre est a .

$$d(a) = \begin{cases} |u| & \text{si } u \text{ est le plus petit suffixe non vide de } x \text{ tq } au \text{ est suffixe de } x \\ |x| & \text{si } u \text{ n'existe pas, i.e si } x \text{ ne contient pas } a, \text{ ou alors seulement en dernière position} \end{cases}$$

Pseudo-code :

```

i ← 0
Tant que i ≤ n - m :
    j ← m - 1
    Tant que j ≥ 0 et xj = ti+j :
```

```

     $j \leftarrow j - 1$ 
    Si  $j = -1$  :
         $i$  est la position d'une occurrence de  $x$ 
         $i \leftarrow i + 1$ 
    Sinon :
         $i \leftarrow i + d(t_{i+m-1})$ 

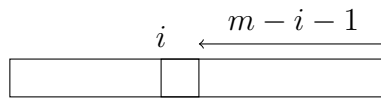
```

Précalcul de d :

```

    Pour toute lettre  $a$ ,  $d(a) \leftarrow m$ 
    Pour  $i$  de 0 à  $m - 2$  :
         $d(x_i) \leftarrow m - 1 - i$ 

```



Complexité :

- Précalcul : $\mathcal{O}(|A| + m)$ où A est l'alphabet, i.e l'ensemble des symboles possibles dans un texte.
 - Algorithme : dans le pire cas, on décale toujours d'un rang (exemple : $t = a^n$ et $x = ba^{m-1}$) : même complexité que l'algorithme naïf.
- En pratique, c'est plus efficace : $\mathcal{O}(n)$ en moyenne (admis).

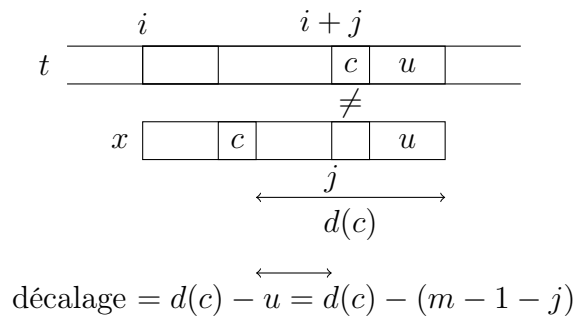
1.2.3 Algorithme de BOYER-MOORE (simplifié)

L'algorithme de BOYER-MOORE-HORSPOOL ne tient pas compte de ce qu'il se passe lors de la lecture de la fenêtre, mais seulement de son dernier caractère.

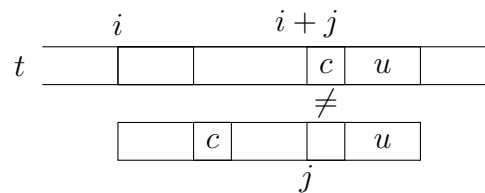
Par exemple : si $x = aababab$ et la fenêtre vaut $aabcbab$, le décalage calculé vaut 2 alors que l'absence de c dans x permet un décalage de 4 rangs.

L'idée de BOYER et MOORE est d'aligner plutôt de caractère qui provoque l'échec de la comparaison avec son occurrence la plus à droite dans x (sauf la dernière lettre).

Plus précisément, si $x_j \neq t_{i+j}$ et $x_{j+1} \dots t_{i+j+1} \dots t_{i+m-1}$, alors on décale la fenêtre de $d(t_{i+j}) - (m - 1 - j)$



Attention : cela ne fait pas toujours progresser la recherche :



On obtient un décalage négatif ! Dans ce cas, on décale seulement d'un rang par sécurité.
 Algorithme :

```

 $i \leftarrow 0$ 
Tant que  $i \leq n - m$  :
   $j \leftarrow m - 1$ 
  Tant que  $j \geq 0$  et  $x_j = t_{i+j}$  :
     $j \leftarrow j - 1$ 
  Si  $j = -1$  :
    Occurrence de  $x$  à la position  $i$ 
     $i \leftarrow i + 1$ 
  Sinon :
     $i \leftarrow i + \max(1, d(t_{i+j}) - (n - 1 - j))$ 
  
```

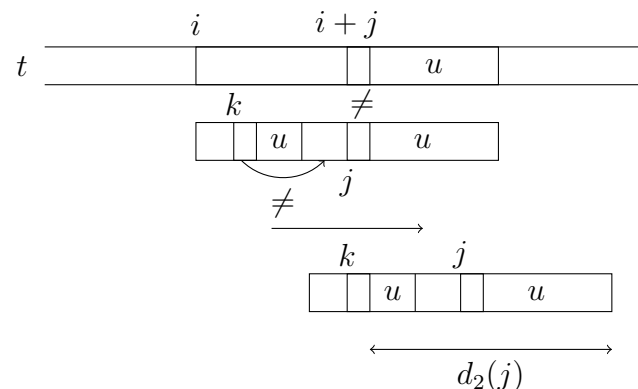
Complexité : dans le pire cas, $\mathcal{O}(nm)$ avec le même exemple qu'en 1.2.2 (page 3).

1.2.4 Algorithme de BOYER-MOORE

La version complète de l'algorithme de BOYER-MOORE utilise une seconde fonction de décalage.

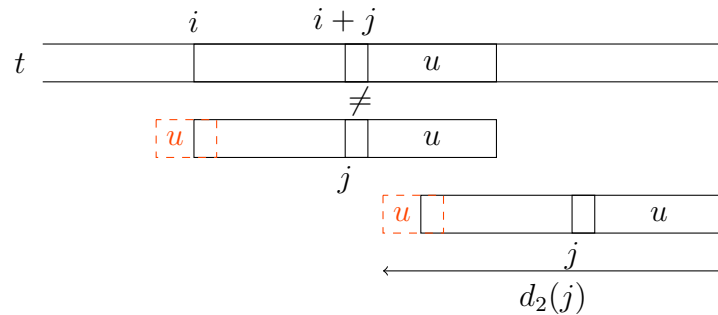
Celle de 1.2.3 (page 4) correspond à la règle du *mauvais caractère* : si un caractère de la fenêtre fait échouer la comparaison avec le motif, on essaie de l'aligner avec sa dernière occurrence dans le motif (sauf la dernière lettre).

Il y a aussi la règle du *bon suffixe* : lorsqu'un caractère fait échouer la comparaison, on a réussi à lire un suffixe de x dans la fenêtre. On peut essayer d'aligner ce suffixe dans le texte avec son occurrence la plus à droite dans x à condition qu'elle soit précédée d'un caractère différent.



$d_2(j)$ = longueur du plus court suffixe de x qui a $x_{j+1} \dots x_{m-1}$ comme suffixe et préfixe et qui n'est pas précédé dans x de x_j .

Si un tel suffixe n'existe pas, on peut chercher le plus long suffixe de u qui est préfixe de x et l'aligner avec le u de la fenêtre.



$d_2(j)$ = longueur de plus court mot w qui a u comme préfixe et x comme suffixe.

Remarque : $|w| \leq |u| + |x|$.

Remarque : $\forall j, d_2(j) \geq 1 + m - j - 1 = m - j$

L'algorithme de BOYER-MOORE utilise le décalage maximal entre ceux calculés à partir de d et d_2 .

Remarque : si on trouve une occurrence de x ($u = x$), alors on tombe dans le deuxième cas du calcul de d_2 : on cherche le plus long suffixe de x qui est aussi préfixe de x , que l'on appelle le *bord* de x .

Algorithme :

```

i ← 0
Tant que i ≤ n + m :
    j ← m - 1
    Tant que j ≥ 0 et x_j = t_{i+j} :
        j ← j - 1
    Si j = -1 :
        Occurrence de x à la position i
        i ← i + d_2(j) - m
    Sinon :
        i ← i + max(d(t_{i+j}) - (m - j - 1), d_2(j) - (m - j - 1))

```

Exemple :

$x = aababab$

j	-1	0	1	2	3	4	5	6
x		a	a	b	a	b	a	b
$d_2(j)$	14	13	12	6	10	6	8	1

$$\begin{array}{cccccccccccccccccccc}
 t = & a & a & b & b & b & a & b & a & b & a & c & a & a & b & b & a & b & a \\
 & a & a & b & a & b & a & b & & & & & & & & & & & \\
 & & & & \times & & & & & & & & & & & & & & \\
 & & & & & & a & a & b & a & b & a & b & & & & & & \\
 & & & & & & & & \times & & & & & & & & & & \\
 & & & & & & & & a & a & b & a & b & a & b & & & & \\
 & & & & & & & & & & \times & & & & & & & &
 \end{array}$$

Complexité : on admet que d_2 peut être précalculée en temps $\mathcal{O}(n)$. Dans le pire des cas, la recherche est encore en $\mathcal{O}(mn)$ (exemple : $x = a^m$ et $t = a^n$)

Remarque : il existe une version plus complexe avec un prétraitement qui permet une complexité $\mathcal{O}(n + m)$ (H.P).

1.3 Algorithme de KARP-RABIN

1.3.1 Principe

L'idée de l'algorithme de KARP-RABIN est d'éviter les comparaisons en temps linéaire entre le motif et le contenu d'une fenêtre en utilisant une fonction de hachage. On parle d'empreinte pour désigner le haché d'une fenêtre d'une chaîne de caractère par la fonction de hachage choisie et on compare les empreintes du motif et du contenu de la fenêtre. Pour contrer le problème des collisions, en cas d'égalité des empreintes, on compare aussi les deux chaînes de caractères.

Algorithme :

Entrées : Texte $t = t_0 \dots t_{n-1}$ (+ fonction de hachage h choisie à l'avance)
 Motif $x = x_0 \dots x_{m-1}$

$h_x \leftarrow h(x_0 \dots x_{m-1})$

$h_t \leftarrow h(t_0 \dots t_{m-1})$

Pour i de 0 à $n - m$:

Si $h_x = h_t$ et (paresseux) $x_0 \dots x_{m-1} = t_i \dots t_{i+m-1}$:
 Occurrence de x à la position i

Si $i < n - m$:

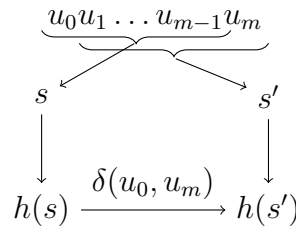
$h_t \leftarrow h(t_{i+1} \dots t_{i+m})$

1.3.2 Choix de la fonction de hachage

Plusieurs contraintes :

- On doit pouvoir comparer des empreintes en temps constant : en pratique, les empreintes sont des entiers machine, donc ce n'est pas un problème.
- Il doit y avoir peu de collisions avec le haché du motif.
- On doit pouvoir calculer l'empreinte de la fenêtre à l'itération suivante sans lire tous ses caractères (sinon on retrouve la complexité de l'algorithme naïf).

On peut par exemple utiliser une fonction de hachage déroulante, *i.e* une fonction h telle qu'il est possible de calculer en temps constant $h(u_1 \dots u_m)$ à partir de $h(u_0 \dots u_{m-1})$



Exemple : on considère que les caractères sont codés sur un octet, *i.e* on les assimile à des entiers compris entre 0 et $r - 1$ (avec $r = 2^8$). On peut voir une chaîne de caractères de longueur m comme l'écriture en base r d'un entier compris entre 0 et $r^m - 1$.

$$P(u_0 \dots u_{m-1}) = \sum_{i=0}^{m-1} u_i r^{m-1-i} = u_0 r^{m-1} + u_1 r^{m-2} + \dots + u_{m-1}$$

On choisit un nombre premier p et on définit $h(u) = P(u) \bmod p$

$$\begin{aligned} h(u_1 \dots u_m) &= \sum_{i=1}^m u_i r^{m-1-i+1} \bmod p \\ &= \left(\sum_{i=1}^{m-1} u_i r^{m-i} + u_m \right) \bmod p \\ &= \left(r \sum_{i=1}^{m-1} u_i r^{m-1-i} + u_m \right) \bmod p \\ &= \left(r \left(\sum_{i=0}^{m-1} u_i r^{m-1-i} - u_0 r^{m-1} \right) + u_m \right) \bmod p \\ &= \left(r \left(h(u_0 \dots u_{m-1}) - u_0 r^{m-1} \right) + u_m \right) \bmod p \end{aligned}$$

Donc

$$\delta(u_0, u_m) : e \mapsto \left(r \left(e - u_0 r^{m-1} \right) + u_m \right) \bmod p$$

Remarque : si on précalcule r^{m-1} alors $\delta(u_0, u_m)$ est bien calculable en temps constant.

Complexité du précalcul : $\mathcal{O}(\log m)$ avec l'exponentiation rapide (modulaire).

1.3.3 Implémentation en OCaml

```

1 | let hash (r : int) (p : int) (s : string) : int =
2 |   let e = ref 0 in
3 |   for i = 0 to String.length s - 1 do
4 |     e := (r * !e + (Char.code s.[i])) mod p
5 |     (*Schema de Horner*)
6 |   done;
7 |   !e;;
8 |
9 | let delta (r : int) (p : int) (rm : int) (u0 : char) (um : char) (
10 |   e : int) : int =
    (r * (e - (Char.code u0) * rm) + Char.code um) mod p;;

```




```

11
12 let harp_rabin (t : string) (x : string) : int list =
13   let n = String.length t
14   and m = String.length x in
15   let l = ref [] in
16   let r = 256
17   and p = 0x7fffffff in (*p = 231 - 1*)
18   let rm = fast_exp_mod r (m - 1) p in (*r(m - 1) mod p*)
19   let hx = hash r p x
20   and e = ref (hash r p (String.sub 0 m t)) in
21
22   for i = 0 to n - m do
23     if hx = !e && x = String.sub i n t then
24       l := i::!l;
25
26     if i < n - m then
27       e := delta r p rm t.[i] t.[i + m] !e
28   done;
29   !l

```

1.3.4 Étude de la complexité

- Complexité de l'initialisation :
 - Complexité de r_m (\mathbf{rm}) : $\mathcal{O}(\log m)$
 - Calcul de h_x et $e = h(t_0 \dots t_{n-1})$: $\mathcal{O}(m)$
 - $\mathcal{O}(m)$ au total pour l'initialisation;
- Complexité de la boucle :
 - À chaque itération : une comparaison et un calcul de delta en $\mathcal{O}(1)$;
 - En cas d'égalité d'empreintes : une extraction et une comparaison de chaîne de taille m : $\mathcal{O}(m)$

Au total : $\mathcal{O}(m) + (n - m)\mathcal{O}(1) + \# \text{collisions} \cdot \mathcal{O}(m) = \mathcal{O}(n + m \cdot \# \text{collisions})$

Avec $t = a^n$ et $x = a^m$, on a égalité à chaque itération et on obtient une complexité $\mathcal{O}((n - m)m)$.

Exemple de cas où le pire cas est atteint sans occurrence de x dans t : $x = aa$, $t = arar \dots ar$, $p = 17$, $r = 26$.

$h(aa) = 0 = h(ar) = h(ra)$.

Remarque : dans la publication de KARP et RABIN, l'idée est de choisir un nombre premier p au hasard parmi un ensemble prédéfini pour limiter le risque de collisions.

Admis : si on choisit deux chaînes de taille m aléatoirement, uniformément, la probabilité de collision avec la fonction de hachage choisie est de l'ordre de $\frac{1}{p}$ (qui est $< 10^{-9}$ si $p = 2^{31} - 1$)

Attention, en pratique les chaînes étudiées ne sont pas du tout choisies uniformément.

1.3.5 Extension à la recherche de plusieurs motifs

Même si cet algorithme est moins efficace que l'algorithme de BOYER-MOORE pour la recherche d'un seul motif, il devient plus intéressant pour la recherche de plusieurs motifs car on peut l'adapter pour éviter de lancer une recherche par motif.

En choisissant une structure d'ensemble adaptée, il est possible de vérifier en temps constant si l'empreinte d'une sous-chaîne est égale à l'empreinte de l'un des motifs (test d'appartenance).

Il suffit de remplacer le test d'égalité par un test d'appartenance à l'ensemble des empreintes des motifs (précalculées) dans l'algorithme de KARP-RABIN

2 Algorithmes de compression

2.1 Contexte

2.1.1 Principe

On dispose d'un texte (en fait, d'un fichier, quel que soit son contenu, puisque c'est une séquence finie d'octets) et on souhaite réduire l'espace mémoire qu'il occupe *via* un changement d'encodage. On s'intéresse ici à la compression sans perte, *i.e* on doit pouvoir retrouver l'intégralité de l'information initiale à partir du texte compressé.

Formellement, on note Σ l'ensemble des symboles autorisés dans le texte, appelé alphabet (en pratique, $\Sigma = \{0, 1\}$), et Σ^* l'ensemble des séquences finies d'éléments de Σ . On cherche alors une fonction de Σ^* dans Σ^* , nommée **compression**, telle qu'il existe **decompression** : $\Sigma^* \longrightarrow \Sigma^* \mid \forall t \in \Sigma^*, \text{decompression}(\text{compression}(t)) = t$

2.1.2 Remarque

On ne peut pas avoir $\forall t \in \Sigma^*, |\text{compression}(t)| < |t|$.

En effet, la condition **decompression** \circ **compression** = id impose que **compression** soit injective. Mais dans ce cas, on aurait une fonction injective de l'ensemble des textes d'une taille donnée vers l'ensemble des textes de taille strictement inférieure : impossible car le cardinal du second ensemble est strictement inférieur à celui du premier.

Il existe donc toujours des textes dont la version compressée est plus grande que la version décompressée.

L'important est alors de trouver des algorithmes qui compressent efficacement les textes qui sont pertinents pour l'utilisateur.

2.1.3 Algorithmes au programme

On va étudier l'algorithme de HUFFMAN, qui exploite des données sur la fréquence d'apparition des caractères dans le texte pour déterminer un encodage, et l'algorithme de LENPEL-ZIV-WELCH, qui ne nécessite pas la connaissance de l'intégrité du texte



pour calculer un encodage car il construit incrémentalement un dictionnaire de codes pour les motifs apparaissant dans le texte.

2.2 Algorithme de HUFFMAN

2.2.1 Principe

L'idée principale de cet algorithme est d'associer un code plus court aux caractères les plus fréquents pour diminuer la taille du texte.

Exemple : `abaabc` nécessite 12 bits avec un code de taille fixe (3 lettres avec 2 bits par caractère) mais seulement 9 bits avec un code de taille variable (ex : `a` = 0, `b` = 10, `c` = 11).

Remarque : pour pouvoir décompresser sans ambiguïté un encodage de taille variable, on ne peut pas encoder un caractère avec un préfixe du code d'un autre caractère. On appelle cela un *code préfixe*.

L'algorithme de HUFFMAN est un algorithme qui, étant donné un texte, produit un code préfixe optimal pour la compression de ce texte, en termes de la taille du texte compressé.

2.2.2 Représentation de l'encodage des symboles

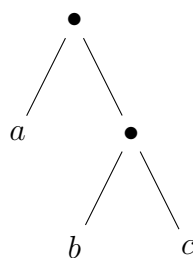
On s'intéresse uniquement à des encodages binaires.

Un code préfixe binaire peut être représenté à l'aide d'un arbre binaire (strict) : dans cet arbre, les caractères sont placés aux feuilles, et le chemin de la racine vers une feuille donne le code du caractère correspondant (descendre à gauche correspond au bit 0, et descendre à droite au bit 1).

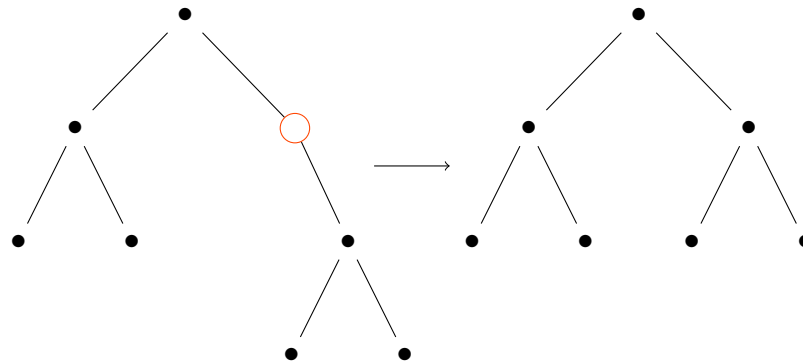
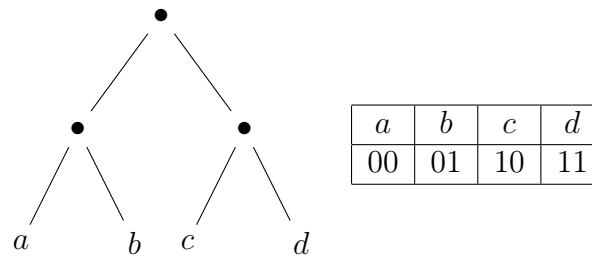
Exemple : le code

<i>a</i>	<i>b</i>	<i>c</i>
0	10	11

 est représenté par :



Autre exemple :



Pour avoir un code optimal, l'arbre doit être binaire strict.

Remarque : la longueur du code d'un caractère est la profondeur de la feuille correspondante dans l'arbre représentant l'encodage. Déterminer un code préfixe optimal revient à chercher un minimum pour

$$\varphi(t) = \sum_{x \in \text{feuilles}(t)} f(x) \cdot p(x, t)$$

où $p(x, t)$ est la profondeur de la feuille x dans l'arbre t , et $f(x)$ est la fréquence du caractère x dans le texte.

2.2.3 Encodage et décodage

On considère un arbre binaire t représentant un code préfixe et on suppose que l'on travaille sur l'alphabet des octets (type `char`).

- Pour décoder une séquence de bits, il suffit de descendre dans l'arbre selon les bits lus et, lorsque l'on atteint une feuille, de produire le caractère correspondant et de poursuivre la lecture en repartant de la racine.

Code (on représente un bit par un booléen (0 : `false`, 1 : `true`)) :

```

1 | type arbre =
2 |   | Feuille of char
3 |   | Noeud of arbre*arbre

```

```

1 | let decode (l : bool list) (t : arbre) : char list =
2 |   let rec aux (l : bool list) (a : arbre) : char list =
3 |     match l, a with
4 |     | _, Feuille c -> c :: aux l t
5 |     | [], _ -> [] (*match order is important here*)
6 |     | b::q, Noeud(g, d) -> if b then aux q d else aux q g
7 |   in aux l t

```

Complexité : \mathcal{O} (nombre de bits dans le texte compressé).

- Pour encoder un texte, étant donné un arbre représentant un encodage pour chacun des caractères du texte, on ne veut pas parcourir l'arbre à chaque caractère pour déterminer son code. On calcule d'abord une table d'encodage associant à chaque caractère son code. On peut la calculer avec un unique parcours de l'arbre :

```

1 | let codes (t : arbre) : bool list array =
2 |   let a = Array.make 256 [] in (*on travaille sur les octets*)
3 |   let rec aux (t : arbre) (acc : bool list) : unit =
4 |     match t with
5 |     | Feuille c -> a.(Char.code c) <- List.rev acc
6 |     | Noeud(g, d) -> aux g (false::acc) ; aux d (true::acc)
7 |   in aux t [];
8 |   a

```

Complexité : chaque nœud interne est traité en $\mathcal{O}(1)$, et chaque feuille x en $\mathcal{O}(p(x, t))$. En notant n le nombre de caractères encodés ($n \leq |\Sigma|$), il y a n feuilles, et $n - 1$ nœuds internes car l'arbre est binaire strict (cf chapitre 6, §1.1.10). La hauteur de l'arbre est comprise entre $\lceil \log_2(n) \rceil$ et n .

Donc la complexité est en $\Omega(n \log n)$ et $\mathcal{O}(n^2)$.

Une fois la table d'encodage calculée, il suffit de remplacer chaque caractère par son code :

```

1 | let encode (l : char list) (t : arbre) : bool list =
2 |   let a = codes t in
3 |   List.flatten (List.map (fun c -> a.(Char.code c)) l)

```

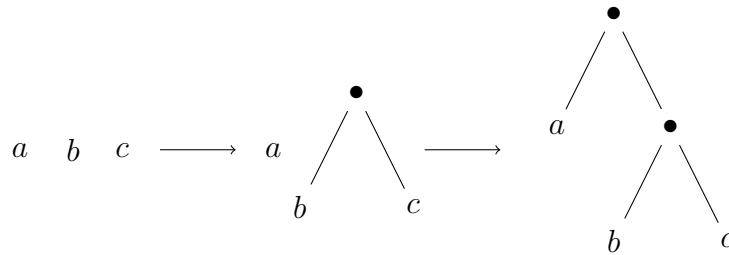
2.2.4 Arbre de HUFFMAN

Il reste à déterminer comment calculer un arbre représentant un code préfixe adapté à la compression d'un texte donné. L'algorithme de HUFFMAN est un algorithme glouton déterminant un arbre optimal.

Cet algorithme part d'une forêt de feuilles et les fusionne deux à deux jusqu'à l'obtention d'un unique arbre. La fusion de deux arbres la construction d'un nouveau nœud dont les deux arbres sont les fils.

Exemple :





L'idée du choix glouton est la suivante : la fusion incrémente la longueur du code des feuilles des deux arbres concernés donc les arbres contenant les caractères les moins fréquents doivent subir le plus de fusion donc être fusionnés en priorité.

On généralise la fonction f de 2.2.2 (page 11) aux arbres en définissant :

$$f(t) = \sum_{x \in \text{feuilles}(t)} f(x)$$

L'algorithme de HUFFMAN s'écrit alors ainsi :

Entrée : texte s

```

occ ← table des occurrences des caractères de s
F ← ensemble des Feuille c où c est tel que occ(c) > 0
Tant que |F| ≥ 2 :
    Extraire t1 de F tel que f(t1) est minimale
    Extraire t2 de F tel que f(t2) est minimale
    F ← F ∪ {fusion(t1, t2)}
Renvoyer l'unique élément de F

```

Implémentation : la forêt F se comporte comme une file de priorité min où la priorité d'un arbre t est $f(t)$.

On suppose donné un type `fp` associé aux primitives suivantes :

- `create` : `unit -> fp` qui crée une file de priorité vide;
- `add` : `arbre -> int -> fp -> unit` qui insère un arbre avec la priorité (entière) donnée dans la file;
- `take_min` : `fp -> arbre*int` qui extrait un arbre de priorité min et qui renvoie aussi sa priorité;
- `size` : `fp -> int` qui renvoie la taille de la file

```

1 | let huffmann (s : char list) : arbre =
2 |   let occ = Array.make 256 0 in (*On travaille sur des octets*)
3 |   List.iter (fun c -> occ.(Char.code c) <- 1 + occ.(Char.code c)) s;
4 |   let f = create () in
5 |   for i = 0 to 255 do
6 |     if occ.(i) <> 0 then
7 |       add (Feuille (Char.chr i)) occ.(i) f
8 |   done;
9 |   while size f > 1 do
10 |     let t1, f1 = take_min f in
11 |     let t2, f2 = take_min f in

```

```

12 ||      add (Noeud(t1, t2)) (f1 + f2) f
13 ||      done;
14 ||      fst (take_min f)

```

Complexité : on note $n(s)$ le nombre de caractères distincts de s

- Le comptage des occurrences : $\mathcal{O}(|s|)$
- Création de la file : avec une implémentation par tas : $\mathcal{O}(n(s) \log(n(s)))$ (insertion en $\mathcal{O}(\log(n(s)))$)
- Une itération de la boucle : $\mathcal{O}(\log(n(s)))$.

Il y a $\mathcal{O}(n(s))$ itérations.

Au total : $\mathcal{O}(|s| + n(s) \log(n(s))) = \mathcal{O}(|s|)$ car $n(s) \leq 256$ car on travaille sur des octets.

2.2.5 Optimalité de l'arbre de HUFFMAN

Rappel : on cherche à trouver un arbre t minimisant

$$\varphi(t) = \sum_{x \in \text{feuille}(t)} f(x)p(x, t)$$

où $f(x)$ est la fréquence d'apparition de x dans le texte et $p(x, t)$ la profondeur de x dans t . On généralise f aux arbres par

$$f(t) = \sum_{x \in \text{feuille}(t)} f(x)$$

et on généralise φ aux forêts par

$$\varphi(F) = \sum_{t \in F} \varphi(t)$$

Lemme : en notant F' la forêt obtenue après un tour de boucle appliquée à F , i.e $F' = (F \setminus \{t_1, t_2\}) \cup \{\text{Noeud}(t_1, t_2)\}$ où t_1 et t_2 sont deux éléments de F qui minimisent f , on a :

$$\varphi(F') = \varphi(F) + f(t_1) + f(t_2)$$

□ Démonstration :

On a :

$$\begin{aligned}
 \varphi(F') &= \sum_{t \in F'} \varphi(t) \\
 &= \underbrace{\sum_{t \in F} \varphi(t)}_{\varphi(F)} - \varphi(t_1) - \varphi(t_2) + \varphi(\text{Noeud}(t_1, t_2))
 \end{aligned}$$

$$\begin{aligned}
& \text{Or} \\
& \varphi(\text{Noeud}(t_1, t_2)) \\
= & \sum_{x \in \text{feuille}(t_1) \cup \text{feuille}(t_2)} f(x)p(x, \text{Noeud}(t_1, t_2)) \\
= & \underbrace{\sum_{x \in \text{feuilles}(t_1)} f(x) \underbrace{p(x, \text{Noeud}(t_1, t_2))}_{1+p(x, t_1)}}_{\underbrace{\sum_{x \in \text{feuilles}(t_1)} f(x)}_{f(t_1)} + \underbrace{\sum_{x \in \text{feuilles}(t_1)} f(x)p(x, t_1)}_{\varphi(t_1)}} + \underbrace{\sum_{x \in \text{feuilles}(t_2)} f(x)p(x, \text{Noeud}(t_1, t_2))}_{f(t_2) + \varphi(t_2)} \\
& \text{d'où } \varphi(F') = \varphi(F) + f(t_1) + f(t_2) \blacksquare
\end{aligned}$$

Proposition : Si les lettres les moins fréquentes sont a et b , alors il existe un arbre optimal tel que les feuilles associées à a et b sont sœurs dans l'arbre et de profondeur maximale.

□ Démonstration :

On considère un arbre optimal t_{opt} et on note c et d deux feuilles sœurs de profondeur maximale dans t_{opt} (possible car t_{opt} est binaire *strict*).

Sans perte de généralité, on suppose $f(a) \leq f(b)$ et $f(c) \leq f(d)$.

Par définition de a et b , on sait que $f(a) \leq f(c)$ et $f(b) \leq f(d)$

En échangeant dans t_{opt} la feuille a et c , on obtient t tel que

$$\begin{aligned}
\varphi(t) &= \varphi(t_{\text{opt}}) - f(a)p(a, t_{\text{opt}}) - f(c)p(c, t_{\text{opt}}) + f(a)p(c, t_{\text{opt}}) + f(c)p(a, t_{\text{opt}}) \\
&= \varphi(t_{\text{opt}}) + \underbrace{(f(a) - f(c))}_{\leq 0} \underbrace{(p(c, t_{\text{opt}}) - p(a, t_{\text{opt}}))}_{\geq 0}
\end{aligned}$$

Donc $\varphi(t) \leq \varphi(t_{\text{opt}})$.

En échangeant de même les feuilles b et d dans t , on obtient t' avec a et b feuilles sœurs de profondeur maximale et telles que $\varphi(t') \leq \varphi(t) \leq \varphi(t_{\text{opt}})$ donc telle que t' est optimal. ■

Théorème : l'algorithme de HUFFMAN renvoie un arbre optimal.

□ Démonstration :

On procède par récurrence sur le nombre de caractères distincts dans le texte :

– Si le texte est écrit à l'aide d'un unique caractère, il n'y a qu'un seul arbre possible donc l'algorithme est optimal.

– Si l'algorithme est optimal pour les textes à n caractères distincts, considérons un texte à $n + 1 \geq 2$ caractères distincts. Notons a et b les deux caractères les moins fréquents dans le texte.

La première étape de l'algorithme de HUFFMAN consiste à fusionner les feuilles associées à a et b . Cela revient à considérer un nouveau texte où a et b sont



remplacés par un nouveau caractère z dont la fréquence d'apparition est donc $f(a) + f(b) = f(\text{Noeud}(a, b))$.

Ce nouveau texte ayant n caractères distincts, on sait que l'algorithme de HUFFMAN produit un arbre t_{opt} optimal pour ce texte par hypothèse de récurrence.

L'arbre obtenu par l'algorithme de HUFFMAN pour le texte initial est t_{opt} où la feuille associée à z est remplacée par le Noeud(a, b).

$$\varphi(t) = \varphi(t_{\text{opt}}) + f(a) + f(b)$$

(même démonstration que le lemme)

On sait qu'il existe une solution optimale t_{opt} telle que a et b sont des feuilles sœurs de profondeur maximale.

On construit t' en remplaçant dans t_{opt} Noeud(a, b) par Feuille(z).

Donc $\varphi(t_{\text{opt}}) = \varphi(t') + f(a) + f(b) \overset{\text{optimalité de } t_{\text{opt}}}{\geq} \varphi(t_{\text{opt}}) + f(a) + f(b) = \varphi(t)$, et ainsi t est optimal ■

2.2.6 Aspects pratiques

- Lectures / écritures : en pratique, on lit et on écrit dans un fichier octet par octet. Or le codage calculé par l'algorithme de HUFFMAN n'est pas de taille fixe, donc on aurait plutôt besoin de lire / écrire bit par bit.

Une solution simple pour l'écriture consiste à accumuler les bits en attendant d'en avoir assez. Il suffit d'un entier pour implémenter cet accumulateur : si $n = \langle b_1 \dots b_k \rangle_2$ avec $k < 8$, on peut ajouter le bit b_{k+1} en calculant $\langle b_1 \dots b_k b_{k+1} \rangle_2 = 2 \langle b_1 \dots b_k \rangle_2 + b_{k+1} = 2n + b_{k+1}$

Problème s'il n'y a pas assez de bits pour construire un octet en fin de fichier.

Solution : on complète avec des 0 à la fin pour finir l'octet (cela s'appelle du *padding*).

Problème : comment ne pas confondre ces 0 avec le code d'un caractère ?

Solution : on ajoute un octet supplémentaire en fin de fichier qui représente le nombre de 0 ajoutés.

Pour la décompression, il faut déterminer quand on arrive à l'avant-dernier octet pour savoir qu'il faut lire l'octet suivant afin de connaître le nombre de 0 à ignorer dans cet avant-dernier octet.

Pour cela, on peut lire le fichier avec 2 octets d'avance pour repérer la fin du fichier, ou alors on peut déterminer à l'avance la taille du fichier et compter le nombre d'octets lus.

- Stockage de l'arbre : la connaissance de l'arbre de HUFFMAN est nécessaire pour pouvoir décompresser le fichier. Il faut donc pouvoir le stocker sous la forme d'un fichier, qui sera transmis avec le fichier compressé.

La transformation de l'arbre en fichier s'appelle *sérialisation de l'arbre*. Comme l'arbre de HUFFMAN est binaire strict, il est aisé de la sérialiser : on définit $\text{repr}(t)$ par induc-

tion structurelle sur t :

$$\text{repr}(\text{Feuillec}) = 0 \quad \underbrace{\text{code}(c)}_{\text{octet représentant } c}$$

$$\text{repr}(\text{Noeud}(g, d)) = 1\text{repr}(g)\text{repr}(d)$$

```

1 | let rec output_tree (f : out_channel) (t : arbre) : unit =
2 |   match t with
3 |   | Feuille c ->
4 |     output_byte f 0;
5 |     output_byte f (Char.code c)
6 |   | Noeud(g, d) ->
7 |     output_byte f 1;
8 |     out_tree f g;
9 |     output_tree f d
10 |
11 | let rec input_tree (f : in_channel) : arbre =
12 |   match input_byte f with
13 |   | 0 -> Feuille (Char.chr (input_byte f))
14 |   | _ ->
15 |     let g = input_tree f in
16 |     let d = input_tree f in
17 |     Noeud(g, f)

```

Remarque : la sérialisation est bien sûr plus générale que son application aux arbres binaires.

Contrainte importante : il faut connaître *a priori* l'encodage utilisé dans la sérialisation pour être capable de désérialiser.

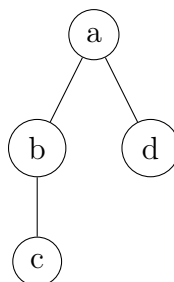
Il existe des formats standards pour la sérialisation, qui peuvent être génériques, par exemple le format JSON, ou bien dédiés à certaines applications, comme le format DIMACS CNF ou le langage DOT pour représenter les graphes.

Exemple pour le langage DOT :

```

1 | graph G {
2 |   a -- b -- c;
3 |   a -- d;
4 | }

```



2.3 Algorithme de LEMPEL-ZIV-WELCH (LZW)

2.3.1 Introduction

L'algorithme de HUFFMAN a deux inconvénients principaux : il est nécessaire de connaître à l'avance l'intégrité du texte pour construire l'encodage, et le texte compressé seul ne suffit pas pour effectuer la décompression : il est nécessaire de l'accompagner par un fichier représentant l'arbre d'encodage. L'algorithme de LZW répond à ces deux problèmes : il est possible de compresser un flux de données *via* la construction incrémentale d'un dictionnaire de codes pour les motifs apparaissant dans le texte, et il n'y a pas besoin d'informations supplémentaires pour décompresser car il est possible de reconstruire le dictionnaire à la volée lors de la lecture du flux compressé.

2.3.2 Principe

On construit incrémentalement le dictionnaire en garantissant que pour chaque motif introduit, tous ses préfixes sont déjà présents dans le dictionnaire. En particulier, le dictionnaire est initialisé en stockant un code pour chaque caractère possible (*via* la table ASCII). Les codes choisis pour les motifs sont consécutifs et ne dépendent que de l'ordre d'apparition des motifs dans le texte. C'est ce qui permet de reconstruire le dictionnaire lors de la décompression en fonction de l'ordre d'apparition des codes.

2.3.3 Compression

On note $t_0 \dots t_{n-1}$ le texte.

```

 $d \leftarrow$  dictionnaire construit à partir de la table ASCII (entier entre 0 et 255 associé à chaque caractère, i.e octet possible)
 $m \leftarrow t_0$ 
Pour  $i$  de 1 à  $n - 1$  :
    Si  $d(mt_i)$  est défini :
         $m \leftarrow mt_i$ 
    Sinon :
        Produire  $d(m)$  en sortie
        Ajouter  $mt_i$  à  $d$  (avec le dernier code dans  $d$  plus 1)
         $m \leftarrow t_i$ 
Produire  $d(m)$  en sortie

```

Exemple :

$t = aababaaab$. Rappel : table ASCII :

a	b	\dots	z
97	98	\dots	122

i	0	1	2	3	4	5	6	7	8
m	a	a	b	a	ab	a	aa	a	ab
Sortie	97	97	98		257		256		257

Dictionnaire : $\{(aa, 256), (ab, 257), (ba, 258), (aba, 259), (aaa, 260)\}$

Attention : ce n'est pas un code préfixe, donc les codes doivent être de taille fixe pour permettre la décompression.

Problème : si tous les codes sont sur k bits, le dictionnaire peut contenir au plus 2^k entrées, donc il y a un risque de saturation.

Pour résoudre ce problème, il y a plusieurs options :

- On efface le dictionnaire lorsqu'il arrive à saturation et on repart de zéro comme si on travaillait sur un nouveau fichier. On produit un code spécial pour signaler l'effacement. Un code peut donc représenter plusieurs motifs selon sa position dans le texte compressé.
- On ajoute un bit aux codes, ce qui revient à doubler la taille du dictionnaire (comme un tableau dynamique). Comme l'algorithme de décompression reconstruit incrémentalement le dictionnaire, la saturation est détectée et on sait quand il faut lire des codes de taille $k + 1$.

2.3.4 Décompression

On reconstruit le dictionnaire en parallèle de la lecture des codes.

Le premier code est nécessairement le code ASCII du premier caractère du fichier. Pour les codes suivants, on a besoin de se souvenir du dernier code lu : notons c_0 le dernier code lu, et c_1 le code en cours de lecture. Le code c_1 correspond à un mot de la forme xs , où x est un caractère, et s un mot potentiellement vide, et c_0 correspond à un certain mot u . Au moment de la production de c_0 , le motif m vaut u et est remplacé par le caractère t_i au cours de la lecture. Comme ce caractère est le premier du motif associé au prochain code produit, *i.e.* c_1 , on sait que $t_i = x$. On sait alors que le motif ux doit être inséré dans le dictionnaire. Ainsi, on effectue les mêmes opérations sur le dictionnaire que l'algorithme de compression, mais avec un temps de retard.

Exemple :

	97	97	98	257	256	257
i	0	1	2	3	4	5
u	a	a	b	ab	aa	ab

Sortie : $aababaaab$

Dictionnaire : $\{(aa, 256), (ab, 257), (ba, 258), (aba, 259), (aaa, 260)\}$

Problème : comme on construit le dictionnaire avec un temps de retard, il est possible que le dernier code lu ne soit pas encore dans le dictionnaire.

Exemple : $t = aaa$

i	0	1	2	
m	a	a	aa	Dictionnaire : $\{(aa, 256)\}$
Sortie	97		256	

Décompression :

	97	256	
i	0	1	Sortie : a ?
u	a		Dictionnaire : \emptyset (la table ASCII)

Remarque : le code qui pose problème vaut nécessairement la taille du dictionnaire, donc correspond au dernier code inséré dans le dictionnaire par l'algorithme de compression à ce stade de la lecture.

Solution : en reprenant les notations précédentes, $c_1 = |d|$ est le code inséré dans le dictionnaire par l'algorithme de compression au moment de la production de c_0 . Donc le motif associé à c_1 s'écrit ux où u est associé à c_0 et x est la lettre lue à ce moment. Or, comme l'algorithme de compression repart du motif $m = x$ au moment de la production de c_0 , x est la première lettre du motif associé à c_1 , *i.e.* ux . Comme u est non vide, x est donc aussi la première lettre de u et on peut reconstruire le motif ux associé à c_1 .

Algorithme (code $c_0 \dots c_{m-1}$) :

```

 $d \leftarrow$  dictionnaire construit à partir de la table ASCII
Produire  $d(c_0)$  en sortie
Pour  $i$  de 1 à  $m - 1$  :
    Si  $c_i < |d|$  :
        Produire  $d(c_i)$  en sortie
         $x \leftarrow$  première lettre de  $d(c_i)$ 
    Sinon (on sait que  $c_i = |d|$ ) :
         $x \leftarrow$  première lettre de  $d(c_{i-1})$ 
        Produire  $d(c_{i-1})x$  en sortie
    Ajouter  $d(c_{i-1})x$  au dictionnaire

```