M2201 TP 1

Codage de Huffman

23 février 2019 – B. Colombel

De quoi s'agit-il ?

Pour coder les caractères sur un clavier, on les représente par un mot binaire d'une longueur donnée. C'est le cas avec la code ASCII où les caractères sont représentés par des mots binaires de longueur 7.

Plus on codera de caractères, plus on aura de bits à manipuler, stocker, transmettre, etc.
On cherche alors à réduire la quantité de bits produits sans perdre d'information. C'est la *compression des données*.

Le codage de Huffman est un algorithme de compression de données sans perte élaboré par David Albert Huffman, lors de sa thèse de doctorat au MIT. Le principe est le suivant :
- on utilise un mot binaire court pour les caractères les plus utilisés
- on réserve les mots binaires long aux caractères les plus rares.

Le codage de Huffman peut-être représenté par un *arbre binaire*.

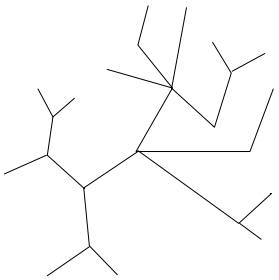
1 Arbres binaires

1.1 Arbres et arbres enracinés

Définition 1 (Arbres)
Un graphe non orienté est un arbre si :
- il est connexe ;
- il ne contient pas de cycle (et donc ni boucle ni arcs parallèles).

Vocabulaire.
- Les arcs d'un arbres s'appellent des *branches* ;
- Les sommets au bout des branches s'appellent des *feuilles* ;
  Plus précisément, une feuille est un sommet de degré 1.
- Les sommets qui ne sont pas des feuilles s'appellent des *nœuds*.

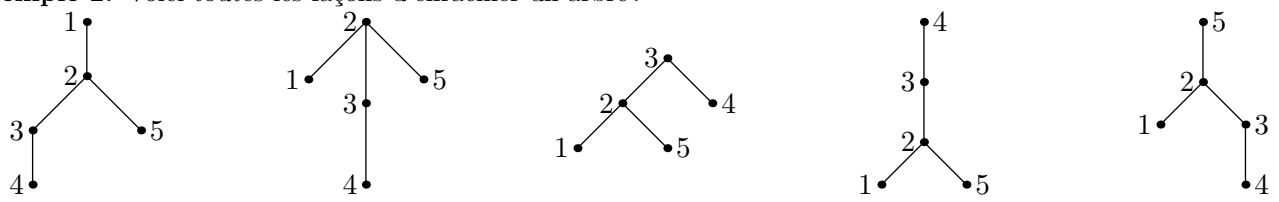
Exemple 1. Le graphe ci-contre est un arbre.



Concrètement, on utilise le plus souvent des arbres dans lequel un sommet joue un rôle particulier : on appelle ce sommet *racine* et un arbre muni d'une racine est un *arbre enraciné*.

Représentation. Lorsqu'on représente un arbre enraciné,on le dessine toujours comme si il avait poussé vers le bas. La racine est alors le point le plus haut du dessin.

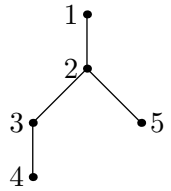
Exemple 2. Voici toutes les façons d'enraciner un arbre !



Les arbres enracinés permettent de représenter les enchaînements de choix successifs. Ils sont donc muni d'une orientation naturelle (du haut vers le bas).

Définition 2
Le *niveau* d'un sommet est le distance qui le sépare de la racine.
La *hauteur* d'un arbre est le plus haut niveau atteint.

Exemple 3. Sur l'arbre ci-contre, la racine est de niveau 0, le sommet 2 de niveau 1, les sommets 3 et 5 sont de niveau 2 et le sommet 4 de niveau 3.



1.2 Arbres binaires

Les arbres enracinés les plus importants sont les arbres binaires :

Définition 3
Un arbre enraciné est un *arbre binaire* si :
- Chaque nœud a au plus 2 successeurs : ce sont ses fils.
- Si un nœud possède deux fils, ils sont ordonnés : il y a un fils ainé et un fils cadet.

Par convention, on représente toujours le fils ainé à gauche te le fils cadet à droite.

Exercice 1 : On considère un arbre enraciné qui vérifie :
- il a au moins trois sommets
- le degré de la racine est 2
- le degré de tous les autres nœuds est 3

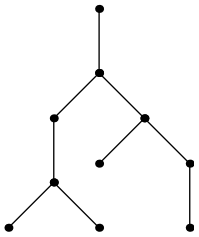
1. Démontrer qu'il s'agit d'un arbre binaire.
2. Comment pourrait-on définir cette sorte d'arbre binaire.
3. Démontrer qu'un tel arbre a un nombre impair de sommets.

Numéro binaire d'un sommet

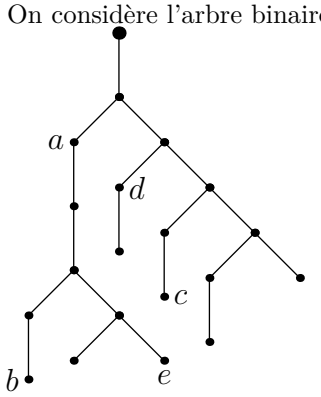
On peut représenter chaque nœud par un mot binaire qui indique le position du sommet dans l'arbre. Pour numéroter les arbres binaires, on procède comme suit :

Méthode. On part de la racine avec le mot sans lettre et on descend jusqu'au sommet dont on cherche le numéro binaire. À chaque fois que l'on quitte un sommet, on ajoute un bit à droite du mot :
- si on descend vers un fils ainé, le bit vaut 0 ;
- si on descend vers un fils cadet, le bit vaut 1 ;
- si le sommet n'a qu'un seul fils, le bit vaut 0.

Exercice 2 : Écrire, sur le graphe ci-dessous, à coté de chaque sommet son numéro binaire.



Exercice 3 : On considère l'arbre binaire ci-dessous.



1. Quel est le numéro binaire des sommets a, b, c, d et e ?
2. Quel sommet a pour numéro binaire 000010 ?

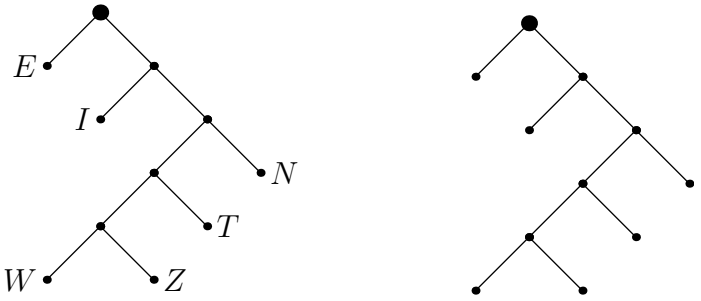
2 Codage de Huffman

Dans un code de Huffman, les caractères sont associés aux feuilles d'un certain arbre binaire et le mot binaire qui code ce caractère est le numéro binaire de sa feuille.

Avant de voir la méthode de construction de l'arbre, regardons comment on l'utilise.

2.1 Décoder un mot binaire

Exercice 4 : On considère l'arbre binaire construit pour le codage de Huffman de la figure 1a. Sur la figure 1b écrire les mots binaires correspondants.



(a) Arbre des lettres (b) Mots binaires

FIGURE 1 – Utilisation de l'arbre d'Huffman

Exercice 5 : On peut aussi coder les 6 caractères $\{E,I,N,T,W,Z\}$ en n'utilisant que des mots binaires de longueur 1 ou 2. Par exemple :

E	I	N	T	W	Z
0	1	00	01	10	11

1. Coder les mots ETIENNE, ETIEEEEEE, EEIIEEEEEE et EEZEENN.
2. Pourquoi ce codage n'est-il pas satisfaisant ?

Propriétés des codes de Huffman

1. La méthode que nous allons voir permet de décoder un message sans ambiguïté
2. Plus un caractère est utilisé, plus le mot qui le code est court

Méthode pratique pour décoder un mot binaire

- ➢ On part de la racine de l'arbre et on se déplace de sommet en sommet lorsque l'on lit les bit du mot
- ➢ Quand on lit le bit 0, on descend vers la gauche (fils ainé) et :
  - — si c'est un nœud, on pour lire le bit suivant
  - — si c'est une feuille, on note le caractère associé à cette feuille et on retourne à la racine
- ➢ Quand on lit le bit 1, on descend vers la droite (fils cadet) et :
  - — si c'est un nœud, on s'arrête pour lire le bit suivant
  - — si c'est une feuille, on note le caractère associé à cette feuille et on retourne à la racine

Exercice 6 : Décoder avec l'arbre de la figure 1 le mot : 1101011110011001.

2.2 Construction de l'arbre de Huffman

Pour construire l'arbre d'Huffman, il faut :
- — décider de l'ensemble des caractères à coder ;
- — connaitre leur fréquences

Remarque. Bien évidemment, les fréquences changent suivant que la langue employée est le français, le polonais, le tahitien, le C ou n'importe quoi d'autre.

1re étape : travail sur les fréquences

- ➢ On écrit en ligne et en ordre croissant les fréquences de caractères.
  On départage arbitrairement les ex æquo.
- ➢ On additionne les deux fréquences les plus faibles
- ➢ On réordonne la nouvelle liste
- ➢ On recommence jusqu'à ce qu'il ne reste plus que 2 fréquences.

Exercice 7 : Dans la langue française les fréquences des caractères $\{E,I,N,T,W,Z\}$ sont approximativement :

E	I	N	T	W	Z
17,66	7,387	7,24	7,08	0,02	0,13

Faire ce travail sur les fréquences.

2e étape : Construction de l'arbre

- ➢ On dessine un arbre binaire à 2 feuilles, de hauteur 1.
- ➢ On inscrit les deux dernières fréquences à côté des deux feuilles, la plus petite à gauche.
- ➢ Pour chaque feuille :
  — si c'est la fréquence d'un caractère, on écrit ce caractère à côté de la feuille
  — si c'est la somme de deux fréquences intermédiaires, on dessine deux branches et on écrit les deux fréquences intermédiaires en face de chaque feuille (la plus petite à gauche)
- ➢ On continue jusqu'à ce que chaque caractère corresponde à une feuille de l'arbre.

Exercice 8 :

1. Construire l'arbre de Huffman de l'exercice précédent.
2. Décoder les mots suivants : 110101011111010 0110101011111010 01101110101111011110

Exercice 9 : Voici une liste de caractères avec leurs fréquences (exprimées dans une certaine unité) :

A	C	E	G	H	L	M	N	R	T	$*$
8	2	10	3	4	9	1	6	5	7	1000

1. Construire l'arbre de Huffman pour ces caractères.
   En cas d'ex æquo, on placera en premier la somme des fréquences qui vient d'être calculé.
2. Donner le codage des caractères.
3. Coder ATTACHER.
4. Décoder 011100010010010110001010100100100001010011001010001011100010110101010101001

Exercice 10 : On considère un texte source formé à partir de 5 symboles distincts (a, b, c, d, r) avec les fréquences d'apparition suivantes :

a	b	c	d	r
0,43	0,20	0,10	0,09	0,18

1. Générer un arbre de Huffman binaire et proposer le codage correspondant.
2. Coder le texte suivant et calculer le gain de compression par rapport à un code binaire de longueur fixe et minimale :

   abracadabrabracadabra

Exercice 11 : Supposons que l'on ait à coder les caractères a, b, c, d, e et f. les lettres a, b, c, d et e ont des probabilités d'apparition respectivement égales à 0,07 ; 0,09 ; 0,12 ; 0,22 ; 0,23.

1. Quelle est la probabilité d'apparition du caractère f ?
2. Trouver le codage de Huffman pour ces 6 lettres en dessinant l'arbre binaire correspondant.
3. Quelle est la longueur moyenne du codage ?
4. Il faut au minimum 3 bits pour coder les 6 lettres. Pour quoi ne pas avoir choisi le code : code(a) = 0 ; code(b) = 1 ; code(c) = 00 ; code(d) = 01 ; code(e) = 10 ; code(f) = 11.