Compression de données

La compression est un processus qui permet de compresser des données informatiques afin de libérer de la place dans la mémoire d'un ordinateur, dans les unités de stockage ou lors des transferts de fichiers.

On sait que tout fichier informatique (en particulier les fichier textes, images ou musiques) est formé d'une suite de caractères. Chacun de ces caractères est codé par une suite de 0 et de 1.

Codage de Huffman

Le codage de Huffman est un **algorithme de compression de données sans perte**.

Le principe du codage de Huffman est de repérer les caractères les plus fréquents et de leur attribuer des codes courts (c'est-a-dire nécessitant moins de 0 et de 1) alors que les caractères les moins fréquents auront des codes longs.

David Albert Huffman

> Le professeur David Albert Huffman (9 août 1925 - 7 octobre 1999, Etats-Unis) fut un pionnier dans le domaine de l'informatique.
>
> Huffman est principalement connu pour l'invention du codage de Huffman utilisé dans presque toutes les applications qui impliquent la compression et la transmission de données digitales comme les fax, les modems, les réseaux informatiques et la télévision à haute définition.
>
> Source : wikipedia – janvier 2012

Découverte de l'algorithme

Nous allons coder le mot : « ELECTRONICIEN »

Dans un premier temps, il faut affecter à chaque caractère son nombre d'occurrences (son « poids »).

(le fichier fréquence-lettre.xls téléchargeable sur http://www.info-isn.fr/ permet de réaliser cette tâche)

ELECTRONICIEN													
Lettre	A	B	C	D	E	F	G	H	I	J	K	L	M
Poids	0	0	2	0	3	0	0	0	2	0	0	1	0
Lettre	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Poids	2	1	0	0	1	0	1	0	0	0	0	0	0

Les caractères « L », « O », « R », « T » sont représentés chacun une fois.

Les caractères « C », « I », « N » sont représentés chacun deux fois.

Le caractère « E » est représenté trois fois.

Dans un deuxième temps, il faut créer un arbre composé de nœuds et dont chaque feuille représente un caractère affecté de son poids.

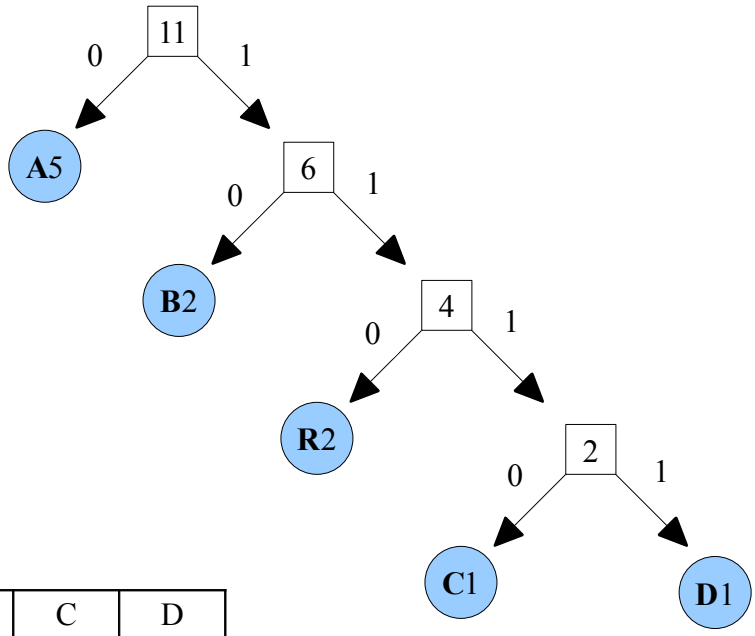
Dans notre exemple, les caractères sont : $\boxed{C\,2}$ $\boxed{E\,3}$ $\boxed{I\,2}$ $\boxed{L\,1}$ $\boxed{N\,2}$ $\boxed{O\,1}$ $\boxed{R\,1}$ $\boxed{T\,1}$

Deux principes de base régissent la construction de l'arbre :

1. les données rares sont codées sur une longueur binaire supérieure à la moyenne, et les données très fréquentes sur une longueur binaire très courte.
2. une séquence binaire ne peut jamais être à la fois représentative d'un élément codé et constituer le début du code d'un autre élément. Par exemple, si un caractère est représenté par la combinaison binaire 100 alors la combinaison 10001 ne peut être le code d'aucune autre information.

Exemple : codage du mot « ABRACADABRA »

Lettre	A	B	R	C	D
Poids	5	2	2	1	1



Lettre	A	B	R	C	D
Code	0	10	110	1110	1111

Au total, on utilise $(1\times 5+2\times 2+2\times 3+1\times 4+1\times 4)=23$ bits

au lieu de 11 caractères de 8 bits, soit 88 bits

Le taux de compression est $\frac{23}{88}\approx 0{,}26$

Exercice

Proposer un codage du mot « ELECTRONICIEN » et calculer le taux de compression.