ISGA
Institut supérieur
d'ingénierie et des affaires

Algorithme de compression de Huffman

M.BAIDADA

Qu’est ce que le codage de Huffman ?

- Il s’agit d’un type de codage utilisé pour la compression de données sans perte
- Il est de type « codage de préfixe » : qui exige qu’il n’y ait pas de mode de code entier ne soit un préfixe d’un autre mot de code dans le système
- Exemple :
  - {101, 0 , 111 , 1100 , 1101} est un code préfixe
  - {**10**1, 0, 111 , **10** , **11** , **11**00, **11**01} n’est pas un code préfixe

Historique

- Proposé, en 1951, par David Albert Huffman, étudiant à l'époque au MIT
- Il devait faire une recherche, dans un cours de théorie de l'information, pour trouver le code binaire le plus efficace
- Huffman, incapable de trouver les code optimaux, a eu l'idée d'utiliser un arbre binaire, et a trouvé que cette méthode est la plus efficace
- Ainsi, Huffman a surpassé son professeur Fano, inventeur de la théorie de l'information, qui avait développé avec Shannon, la méthode de codage de Shannon-Fano

Principe

- Consiste à représenter le texte à coder sus forme d’un arbre binaire
- Un parcours particulier de cet arbre permet d’associer à chaque lettre, selon sa fréquence dans le texte initial, un code binaire unique
- La technique assure que les lettre les plus fréquentes sont représentées par les codes les plus courts.

Etapes

1. Calculer le nombre d'occurrences de chaque caractère dans le texte initial
2. Ces caractères seront placés comme feuilles d'un arbre, en associant à chaque nœud un poids qui vaut son nombre d'occurrence
3. L'arbre est ensuite construit du bas en haut de la manière suivante :
   - On associe à chaque fois les deux nœuds ayant le poids le plus faible pour créer un nœud père dont le poids est la somme des poids de ces deux nœuds
   - On continue jusqu'à obtenir un seul nœud (la racine de l'arbre)

Etapes

4. On associe ensuite, par exemple, le code 0 aux branches gauches et 1 aux branches droites
5. Le code binaire de chaque caractère (placés dans les feuilles), est obtenu en remontant l'arbre de la racine jusqu'aux feuilles, et en rajoutant à chaque fois un 1 ou un 0 selon la brache suivie.
6. L'arbre est ensuite construit du bas en haut de la manière suivante :
   - On associe à chaque fois les deux nœuds ayant le poids le plus faible pour créer un nœud père dont le poids est la somme des poids de ces deux nœuds
   - On continue jusqu'à obtenir un seul nœud (la racine de l'arbre)

Exemple

- Considérons le texte suivant :

this is an example of a huffman tree

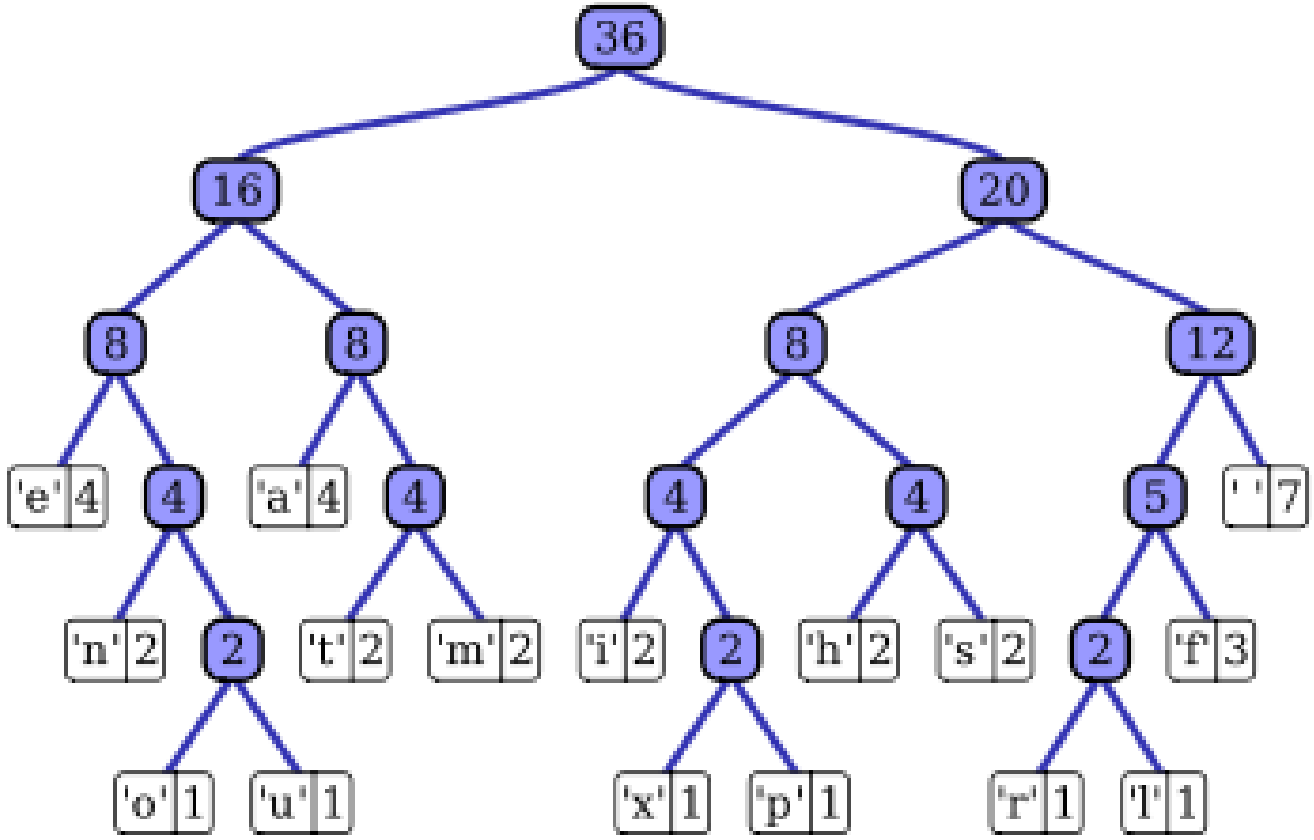
Exemple tiré de wikipédia.org

Classement par fréquence

Caractère	Fréquence
X	1
P	1
L	1
O	1
U	1
R	1
T	2
H	2
I	2
S	2
N	2
M	2
F	3
A	4
E	4
Espace	7

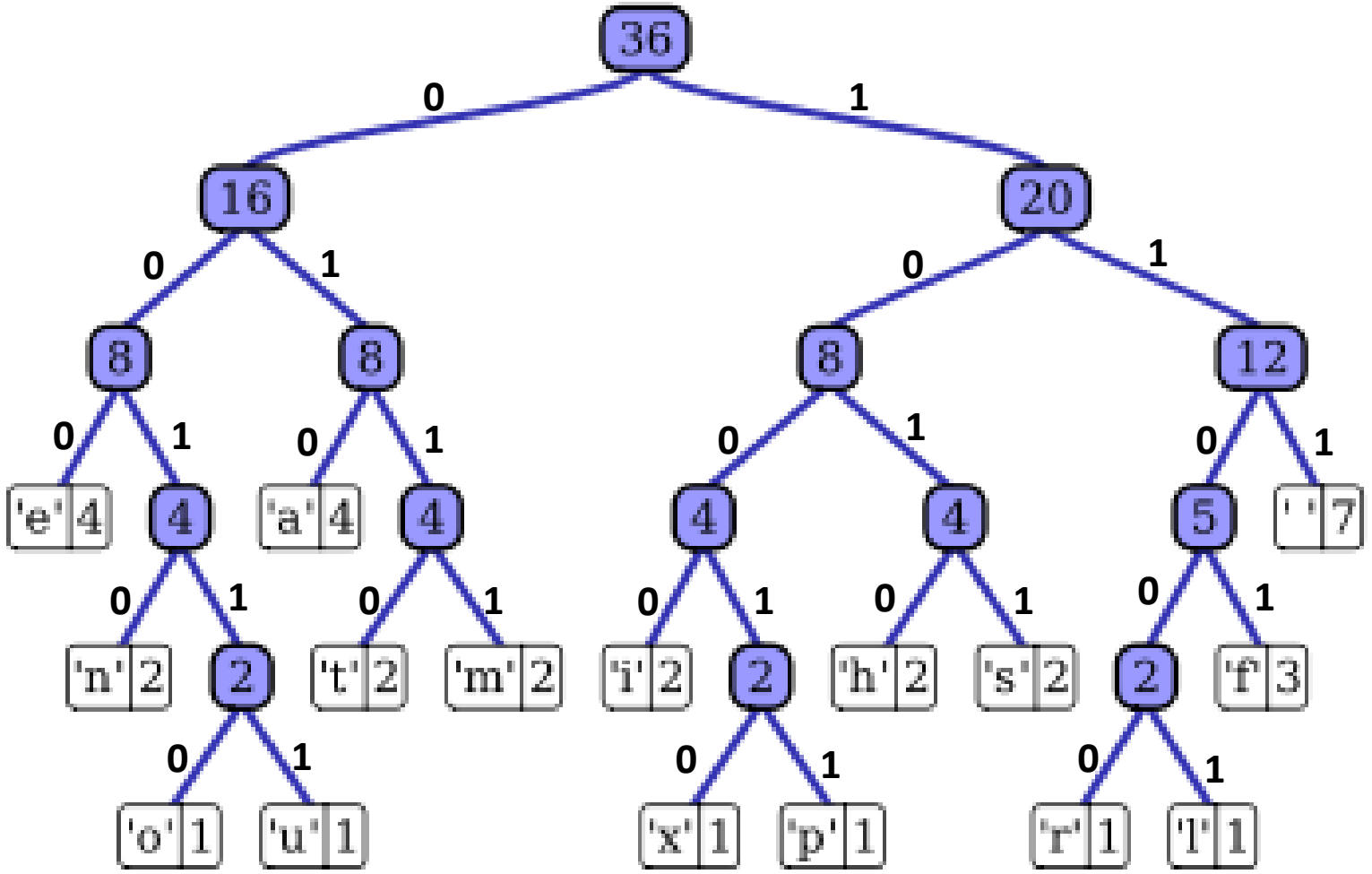
Représentation dans un arbre binaire

Caractère	Fréquence
X	1
P	1
L	1
O	1
U	1
R	1
T	2
H	2
I	2
S	2
N	2
M	2
F	3
A	4
E	4
Espace	7



Exemple tiré de wikipédia.org

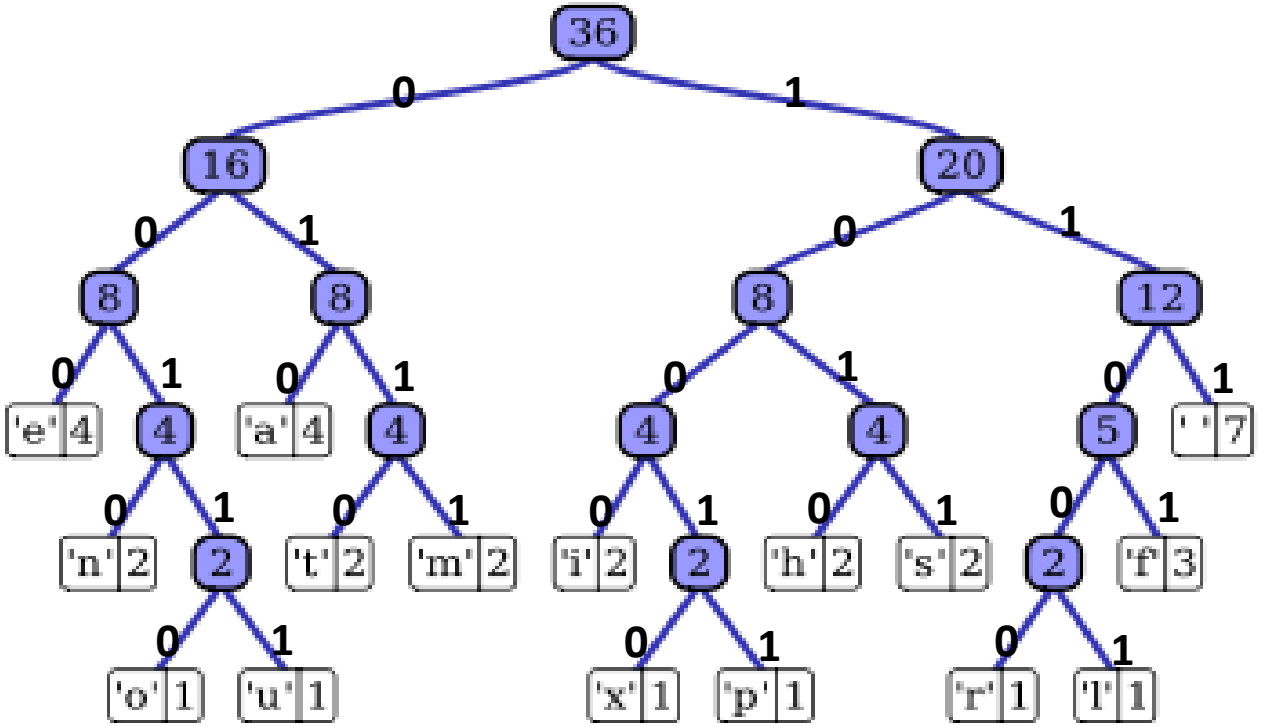
Codage



On associe des 1 aux branches droites et des 0 aux branches gauches

Table du codage

Le codage de chaque caractère est obtenu en parcourant l(arbre de la racine aux feuilles, et en regroupant les 0 et les 1 relatifs aux branches parcourues



Caractère	Fréquence	Code
X	1	1 0 0 1 0
P	1	1 0 0 1 1
L	1	1 1 0 0 1
O	1	0 0 1 1 0
U	1	0 0 1 1 1
R	1	1 1 0 0 0
T	2	0 1 1 0
H	2	1 0 1 0
I	2	1 0 0 0
S	2	1 0 1 1
N	2	0 0 1 0
M	2	0 1 1 1
F	3	1 1 0 1
A	4	0 1 0
E	4	0 0 0
Espace	7	1 1 1

Décodage

- Le texte compressé peut être décodé en le parcourant les bits le constituant et suivre les chemins droite/gauche selon la valeur rencontrée si c’est 1 ou0, jusqu’à arriver à une feuille, et afficher le caractère qu’elle contient.
- Notons que le message compressé doit être toujours accompagné de la table du codage ou bien de l’arbre correspondant, pour qu’il puisse être décodé.