

Algorithme de Huffman

Groupe : Poupelin Bastien, Cherruau Anthony, Thebaudin Corentin

Table des matières

1	Introduction	3
2	Compter la fréquence des caractères	4
2.1	Structure de liste chaînée	4
2.2	Problème des caractères de l'ASCII étendu	4
2.3	Insertion et tri dans la liste chaînée	4
3	Arbre	5
3.1	Principe	5
3.2	Structure pour l'arbre	5
3.3	Création de l'arbre	5
3.4	Réalisation d'un dictionnaire listant le codage de chaque caractère	6
4	Compression	6
4.1	Principe	6
5	Décompression	7
5.1	Principe	7
6	Comparaison différente compression	7
7	Amélioration	7

1 Introduction

David albert Huffman

Qui est Huffman ?

Il est un professeur né le 9 août 1925 et mort le 7 octobre 1999. Il est un pionnier dans le domaine informatique. Durant toute sa vie, Huffman apporta des contributions importantes à l'étude des machines à états finis. Mais 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 numériques comme les fax, les modems, les réseaux informatiques et la télévision à haute définition. Huffman contribua beaucoup au développement de différents domaines, notamment dans la théorie de l'information et du codage, où il a été un pionnier dont les découvertes sont à la base des systèmes de compression de fichiers informatiques dans toutes les machines de nos jours.

Principe du codage de huffman

Le principe du codage de huffman repose sur la création d'un arbre composé de nœuds. Dans une phrase à coder on compte d'abord le nombre d'occurrence de chaque lettre. Chaque caractère constitue donc une feuille de l'arbre auquel on y associe sa valeur d'occurrence. Puis on crée l'arbre. On associe les deux nœuds ou feuilles avec les occurrences les plus faibles pour créer un nouveau nœud dont son occurrence est la somme des deux occurrences des feuilles. On fait cela jusqu'à qu'il en reste un qui devient la racine. Ensuite on associe le code 0 pour la branche gauche et le code 1 pour la branche droite. Ainsi pour obtenir le code binaire de chaque caractère on parcourt l'arbre à partir de la racine jusqu'au feuilles en rajoutant un 0 ou un 1 en fonction de la branche suivie.

Déroulement du projet

- Pour réaliser le projet, nous nous sommes réparti le travail en 3 parties :
- La fréquence des caractères du document source réalisée par Bastien
 - La création de l'arbre binaire réalisée par Anthony
 - Réalisation d'un dictionnaire listant le codage de chaque caractère réalisé par Corentin

2 Compter la fréquence des caractères

Le principe de l'algorithme est de remplacer les lettres du document par des codes. Afin de gagner de la place, plus la lettre apparaît de fois et plus le code lui correspondant doit être court. Avant d'attribuer le code il est donc nécessaire de calculer la fréquence des caractères dans le document.

2.1 Structure de liste chaînée

Pour compter les caractères j'ai utilisé une structure de liste chaînée contenant le caractère et l'occurrence de son apparition dans le document. Pour implémenter cette liste, on parcourt les caractères du fichier un à un pour les ajouter dans la liste.

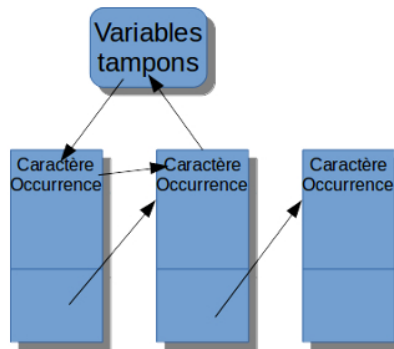
2.2 Problème des caractères de l'ASCII étendu

Malheureusement, un caractère UTF-8 ou ISO 8859-1 est composé de plusieurs octets pour pouvoir utiliser l'ASCII étendu (>127). Mais le langage C ne gère pas nativement ce codage car dans ce langage, 1 caractère = 1 octet. Ainsi seul les caractères de l'ASCII simples étaient correctement comptés donc les caractères accentués de la langue française, par exemple, ne l'étaient pas. Après une recherche de solutions improductives, nous avons en fin de projet, trouvé une solution. Avec la librairie WCHAR, on définit tous les caractères en type `wchar_t` et toutes les fonctions telle que `fprint` sont transformées en `wfprintf`. Le problème des caractères de l'ASCII étendu sont donc désormais correctement dénombrés.

2.3 Insertion et tri dans la liste chaînée

Pour revenir à la liste chaînée, elle est au début vide mais on ajoute à chaque fois le caractère lu à la première position si il n'existe pas déjà dans la liste. Sinon, si il est déjà présent, on incrémente l'occurrence correspondant à ce caractère. Afin d'avoir un algorithme performant la liste doit être triée de l'élément le moins présent à celui le plus présent. Ainsi lors de l'incrémement, il faut parfois trier la liste chaînée.

J'ai réalisé le tri non pas en échangeant les pointeurs de 2 maillons mais simplement en inversant les données (caractère, occurrence) entre maillons jusqu'à être à la bonne place. C'est à dire que l'occurrence contenue dans la liste à droite de celle qu'on déplace doit être supérieure ou égale à l'occurrence de cette dernière.



3 Arbre

3.1 Principe

La création de l'arbre se fait en parcourant la liste trié. Elle prend les deux premières valeurs de la liste on va créer les feuilles puis on va créer un nœud et donc un arbre. Le nœud est crée en prenant les deux feuilles qui sont le fils gauche et le fils droit du nœud. Le nœud a pour valeur la somme des fréquences de ses fils. Ensuite on va réitérer cela a toute les valeurs de la liste.

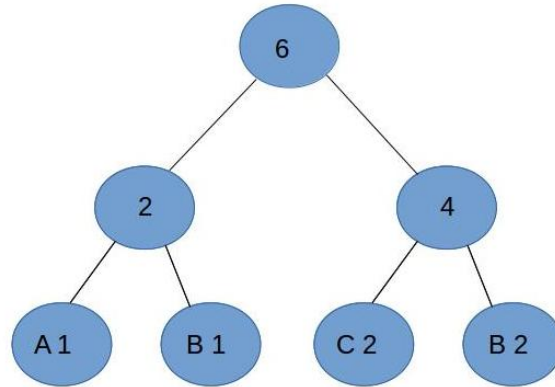
3.2 Structure pour l'arbre

Description

L'arbre contient la fréquence des lettres compté précédement et les lettres qui vont nous permettre de créer les feuilles et les noeuds. Enfin on a une structure fils gauche et fils droit pour les noeuds ou le fils gauche et le fils droit vont pointer vers un noeud ou une feuille.

3.3 Création de l'arbre

La création de l'arbre se fait en parcourant la liste triée. Si l'élément suivant de la liste est différent de nulle alors on regarde si l'arbre est nulle ou non. Si c'est le cas alors on va créer le premier noeud avec les deux premières valeur de la liste triée. Si ce n'est pas le cas alors on va regarder les fréquences dans la liste. Si la valeur de la fréquence est identique alors on va créer un noeud et on le met dans un autre arbre. Sinon on crée un noeud avec le premier noeud puis on assemble les deux arbres. A la fin il reste une valeur et donc on va créer la racine de l'arbre.



exemple d'arbre obtenu

3.4 Réalisation d'un dictionnaire listant le codage de chaque caractère

Ce bloc permettra de coder chaque caractère du fichier à compresser en binaire. C'est cette opération que l'on appellera compression du document.

Pour réaliser ce bloc, il faudra utiliser l'arbre créé auparavant dans le bloc 2. Grâce à cet arbre, on pourra coder chaque caractère en binaire suivant le chemin emprunté dans l'arbre. Par exemple ici, on établira que pour chaque chemin vers la gauche emprunté, un "0" sera codé pour le caractère trouvé. Lorsque le chemin emprunté est celui du fils droit, le codage sera un "1".

C'est pourquoi la réalisation de ce bloc se basera sur une boucle de lecture de l'arbre. En premier temps la réalisation de cette lecture serait faite de manière préfixe, avec une variable binaire qui contiendrait le codage du caractère, s'incrémentant par un décalage de bit à chaque mouvement dans l'arbre jusqu'à attendre une feuille. Par soucis de temps et de solutions, un second choix a été réalisé : L'arbre sera lu de haut en bas jusqu'à atteindre une feuille, le codage et le caractère sera écrit dans le fichier puis la feuille sera supprimée. La lecture sera relancée, jusqu'à ce que l'arbre soit vide.

4 Compression

4.1 Principe

Après avoir réalisé ces 3 blocs, nous avons à disposition tout les outils pour réaliser la compression et la décompression d'un fichier.

La compression consiste à lire le fichier à compresser et remplacer chaque caractère par son codage listé dans le dictionnaire.

5 Décompression

5.1 Principe

La décompression va lire le fichier compressé qui contient une en-tête et le fichier texte compressé. L'en-tête contient l'arbre créé précédemment. Pour décompresser le fichier on lit le fichier qui contient des 1 et des 0 puis on parcourt l'arbres. Dès que l'on arrive a une feuille dans l'arbre alors on mets la lettre trouvé dans le fichier décompressé et on continue ainsi jusqu'à la fin du fichier.

(A voir) La décompression peut se faire de 2 manières : l'en-tête peut contenir le dictionnaire et donc on lira le fichier compressé et on retrouvera le caractère correspondant au codage. Cette méthode est plus simple à réaliser pour l'en-tête, mais un peu plus complexe pour la décompression. La seconde méthode consiste à intégrer l'arbre dans l'en-tête, et on retrouvera le caractère en suivant le chemin suivi avec le codage du fichier compressé. Cette méthode est plus complexe pour intégrer l'arbre à l'en-tête, mais plus simple à la suite pour décompresser le fichier.

6 Comparaison différente compression

7 Amélioration

Notre code est minimaliste et de nombreux point peut être amélioré.

encodage : L'encodage actuelle ne prend pas les accents et donc dans notre langue le français cela est problématique. Notre programme prend donc que des caractères non accentués.

arbre : Quand on a un fichier vide ou un fichier contenant deux caractères ce cas n'est pas géré pour la création de l'arbre.