Algav – Projet Huffman Dynamique

Principe Huffman classique

La compression par l'algorithme de Huffman peut être décomposé en 4 étapes distinctes :

(1) Lister symbole, compter les occurrences et trier en fonction des occurrences.

(2) Prendre les deux nœuds de plus faible occurrence, les combiner en faire un nœud dont l'occurrence est égale à la somme des deux occurrences et marquer les nœuds fusionnés par 0 pour le premier et 1 pour le second (par ordre de probabilités décroissantes). Le nouveau nœud crée comportera les adresses les adresses des nœuds fils.

(2bis) L'étape (2) est répétée jusqu'à ce qu'il ne reste plus qu'un seul nœud qui deviendra la racine de l'arbre.

(3) Parcourir l'arbre depuis la racine jusqu'aux feuilles afin de déterminer les codes de Huffman pour chaque symbole.

(4) Lire le fichier à compresser en remplaçant chaque symbole par son codage de Huffman.

Principe Huffman Dynamique

- Ne parcourir qu'une seule fois le fichier à compresser.
- Respecter les propriétés suivantes à chaque incrément ou ajout dans l'arbre :

Propriété 1. Soit H un AHA avec n feuilles et $n-1$ nœuds internes ; dans la numérotation hiérarchique GDBH *GaucheDroiteBasHaut* $x_1, x_2, \ldots, x_{2n-1}$ des nœuds, on a :

(1) $W(x_1) \leq W(x_2) \leq \ldots \leq W(x_{2n-1})$, où $W(x_i)$ est le poids du nœud x_i.

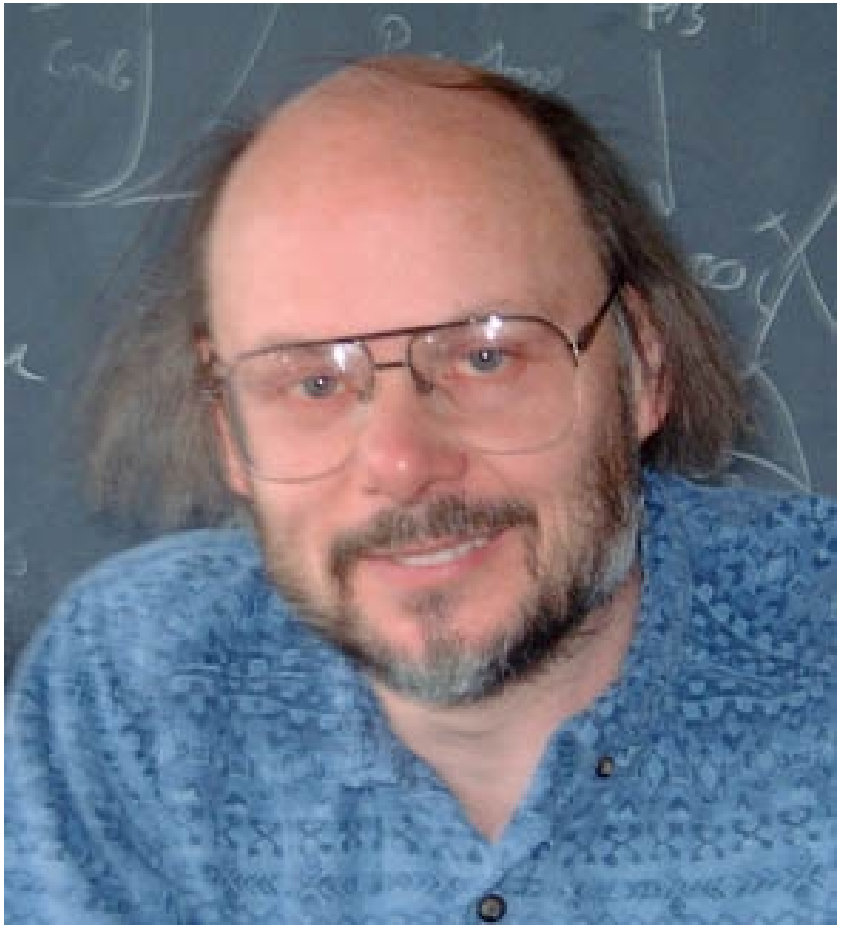
(2) Pour $1 \leq i \leq n-1$, les nœuds x_{2i-1} et x_{2i} sont frères (i.e. ont le même père dans l'arbre).

Propriété 2. Etant donné un AHA et une feuille f, de numéro x_{i_0}, dont le chemin à la racine est $\Gamma_f = x_{i_0}, x_{i_1}, \ldots, x_{i_k}$ $(i_k = 2n-1)$, alors l'arbre résultant de l'incrémentation de $W(f)$ sera encore un AHA ssi $W(x_{i_j}) < W(x_{i_j+1})$, pour $0 \leq j \leq k-1$. On dira dans ce cas que tous les nœuds du chemin Γ_f sont *incrémentables*.

Compression par Huffman Dynamique

Tant que l'on peut lire dans le fichier :
    On lit un caractère dans le fichier.
        Si c'est un nouveau caractère :
            On l'ajoute à l'arbre.
            On transmet le chemin de la feuille spéciale + le codage du nouveau symbole.
        Sinon si le caractère existe déjà :
            On incrémente son poids dans l'arbre.
            On transmet le chemin jusqu'à cette feuille.
    On effectue les éventuelles modifications nécessaires sur l'arbre.
FinTantQue

Choix du langage



Structure du programme

Fichier : main.cpp
Rôle : Appelle les fonctions de compression ou décompression de Huffman.cpp.

Fichiers : Huffman.hpp, Huffman.cpp
Rôle : Contient les deux fonctions principales de compression et décompression, ainsi que des méthodes « debug » permettant d'afficher le déroulement des algos.

Fichiers : Aha.hpp, Aha.cpp
Classes : Aha et Node
Rôle : Contient la classe Aha représentant un arbre de Huffman adaptatif ainsi que la classe Node représentant un nœud composant d'un arbre.

Fichiers : Binarytools.hpp, Binarytools.cpp
Classe : NBuffer
Rôle : Contient la classe NBuffer qui permet d'écrire et de lire les transmissions d'huffman dynamique dans des fichiers.

Hall of Fame (en debug)

Fichier	Taille initiale	Taille compressé	Taux de compression	Taille décompressé	Caractères	Temps (en seconde)
"abracadabra"	11	9	18,18	11	5	0,001
"carambarbcm"	11	10	9,09	11	5	0,001
Data\Test1\test0	3 000	1 893	36,90	3 000	30	0,02
Data\Test1\test1	2 046	525	74,34	2 046	10	0,009
Data\Test1\test2	65 534	16 047	74,96	65 534	15	1,335
Data\Test1\test2rev	65 534	16 394	74,98	65 534	15	1,65
Data\Test1\test3	80 000	54 291	32,14	80 000	40	2,349
Data\Test1\test4	1 048 574	262 178	75,00	1 048 574	19	28,139
Data\Test1\upmc.eps	585 794	101 507	82,67	585 794	63	15,705
Data\Experimental Data\urns.m12.n4000.s100000	1 088 958	483 286	55,62	1 088 958	27	30,873
Data\Experimental Data\exp-rnd-33.txt	174 333	73 098	58,07	174 333	45	5,074
Data\Wikipedia\Wikipedia-20091130153946.xml	578 837	380 264	34,31	578 837	193	16,585
Data\Wikipedia\Wikipedia-20091130154119.xml	132 841	91 661	31,00	132 841	195	3,678
Data\Textual Data\phdthesis.ps	3 457 327	2 192 860	36,57	3 457 327	97	98,63
Data\Textual Data\eng-dict-1913.txt	8 025 699	4 613 720	42,51	8 025 699	87	222,569
Data\Textual Data\Bible\BDS\40026000	9 677	4 880	49,57	9 677	26	0,274
Data\Textual Data\Bible\ITA\40026000	8 318	4 176	49,80	8 318	22	0,241
Data\Textual Data\Bible\KJV\40026000	8 067	4 133	48,77	8 067	26	0,263
Data\Textual Data\Gutenberg\2505.txt	1 220 550	628 962	48,47	1 220 550	27	35,977
Data\Textual Data\Gutenberg-S3\2505-s3.txt	3 904 440	2 056 800	47,32	3 904 440	30	111,963

Temps d’exécution en fonction de la taille initiale