17. Код Хаффмана

АлгоритмГаффмана — адаптивний жадібний
алгоритм оптимального префіксного кодування алфавіту з мінімальною надмірністю. Був розроблений аспірантом Массачусетського технологічного інституту Девідом Гаффманом при написанні ним курсової роботи та надрукований в статті 1952 року «A Method for the Construction of Minimum-Redundancy Codes».В даний час використовується в багатьох програмах стиснення даних без втрат.

На відміну від алгоритму Шеннона — Фано, алгоритм Гаффмана залишається завжди оптимальним і для вторинних алфавітів m_2 з більш ніж двома символами.

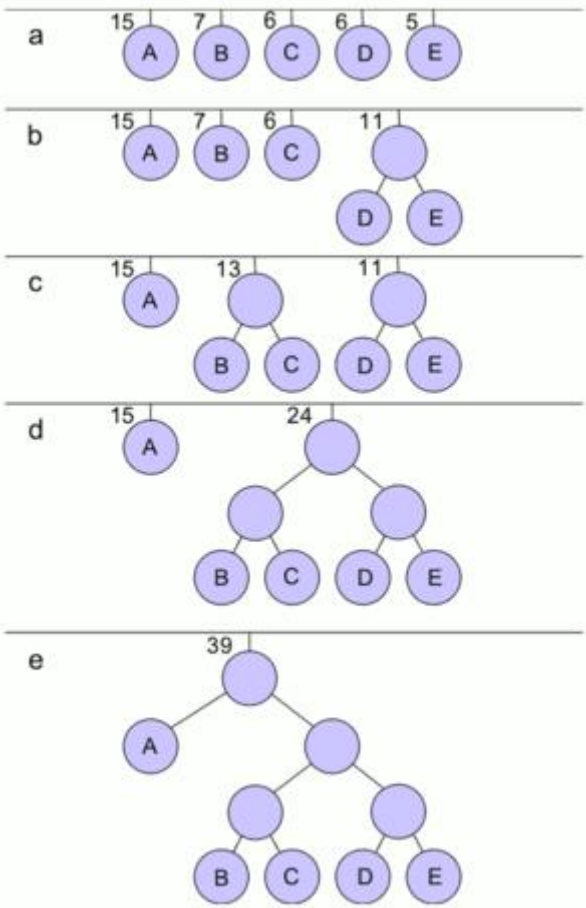
Цей метод кодування складається з двох основних етапів:

1. Побудова оптимального кодового дерева
2. Побудова відображення код-символ на основі побудованого дерева

Один з перших алгоритмів ефективного кодування інформації був запропонований Д. А. Гаффманом в 1952 році. Ідея алгоритму така: знаючи ймовірності появи символів у повідомленні, можна описати процедуру побудови кодів змінної довжини, що складаються з цілої кількості бітів. Символам з більшою ймовірністю ставляться у відповідність коротші коди. Коди Гаффмана володіють властивістю префіксності (тобто жодне кодове слово не є префіксом іншого), що дозволяє однозначно їх декодувати.

Класичний алгоритм Гаффмана на вході отримує таблицю частот з якими зустрічаються символи у повідомленні. Далі на підставі цієї таблиці будується дерево кодування Гаффмана (Н-дерево).

1. Символи вхідного алфавіту утворюють список вільних вузлів. Кожен лист має вагу, яка може бути рівною або ймовірності, або кількості входжень символу у стиснене повідомлення.
2. Вибираються два вільних вузли дерева з найменшими вагами.
3. Створюється їхній батьківський вузол з вагою, рівною їх сумарній вазі.
4. Вузол-батько додається в список вільних вузлів, а два його нащадки видаляються з цього списку.
5. Одній дузі, котра виходить з вузла батька, ставиться у відповідність біт 1, інший — біт 0.
6. Кроки, починаючи з другого, повторюються доти, поки в списку вільних вузлів не залишиться тільки один вільний вузол. Він і буде вважатися коренем дерева.



Припустимо, у нас є наступна таблиця частот:

15	7	6	6	5
A	B	C	D	E

Цей процес можна подати як побудову дерева, корінь якого — символ з сумою ймовірностей об'єднаних символів, отриманий при об'єднанні символів з останнього кроку, його n_0 нащадків — символи з попереднього кроку і т. д.

Щоб визначити код для кожного із символів, що входять у повідомлення, потрібно пройти шлях від листка дерева, який відповідає поточному символу, до його кореня, накопичуючи біти при переміщенні по гілках дерева (перша гілка в шляху відповідає молодшому біту). Отримана таким чином послідовність бітів є кодом даного символу, записаним у зворотному порядку.

Для даної таблиці символів коди Гаффмана будуть виглядати так:

A	B	C	D	E
0	100	101	110	111

Оскільки жоден з отриманих кодів не є префіксом іншого, вони можуть бути однозначно декодовані при читанні їх з потоку. Крім того, найбільш частий символ повідомлення A закодований найменшою кількістю біт, а найбільш рідкісний символ E — найбільшою.

При цьому загальна довжина повідомлення, що складається з наведених у таблиці символів, складе 87 біт (в середньому 2,2308 біта на символ). При використанні рівномірного кодування загальна довжина повідомлення склала б 117 біт (рівно 3 біти на символ). Зауважимо, що ентропія джерела, яке незалежним чином породжує символи із зазначеними частотами, складає ~ 2,1858 біта на символ, тобто надмірність побудованого для такого джерела коду Гаффмана, що розуміється, як відмінність середнього числа біт на символ від ентропії, становить менше 0,05 біта на символ.

Класичний алгоритм Гаффмана має ряд істотних недоліків. По-перше, для відновлення вмісту стиснутого повідомлення декодер повинен знати таблицю частот, якою

користувався кодер. Отже, довжина стиснутого повідомлення збільшується на довжину таблиці частот, яка повинна надсилатися попереду даних, що може звести нанівець всі зусилля щодо стиснення повідомлення. Крім того, необхідність наявності повної частотної статистики перед початком власне кодування вимагає двох проходів по повідомленню: одного для побудови моделі повідомлення (таблиці частот і Н-дерева), іншого для власне кодування. По-друге, надмірність кодування обертається на нуль лише в тих випадках, коли ймовірності кодованих символів є оберненими степеням числа 2. По-третє, для джерела з ентропією, що не перевищує 1 (наприклад, для двійкового джерела), безпосереднє застосування коду Гаффмана позбавлене сенсу.