Двухпроходное кодирование кодом Хаффмана

5 января 2023 г. 1:24

Код Хаффмана предполагает, что мы знаем распределение вероятностей символов источника.
Т.к. в реальных задачах данное распределение вероятностей неизвестно:
● Можно оценить распределение вероятностей и передать его декодеру (двухпроходное кодирование)
● Можно оценивать распределение вероятностей адаптивно одинаковым образом используя уже закодированные/декодированные символы (адаптивное кодирование)

- Проход 1.
  1. Оценить $\boldsymbol{\theta}$ (обозначим оценку через $\hat{\boldsymbol{\theta}}$).
  2. Кодировать $\hat{\boldsymbol{\theta}}$. На выходе получим кодовые слова $\boldsymbol{c}_1$.
- Проход 2.
  1. Кодировать символы источника $\boldsymbol{x}$ используя $\hat{\boldsymbol{\theta}}$. На выходе получим кодовые слова $\boldsymbol{c}_2$.
  2. Сформировать кодовое слово из двух частей $\boldsymbol{c} = (\boldsymbol{c}_1, \boldsymbol{c}_2)$.

Предположим, мы реализовываем код Хаффмана, оценили вероятности и хотим передать их. Осуществлять передачу самих вероятностей не рационально, т.к. они займут много бит итогового файла, файл может даже получиться больше исходного. Очевидно, тогда выгоды от сжатия никакой - одни убытки.
Рационально будет передать само дерево Хаффмана.

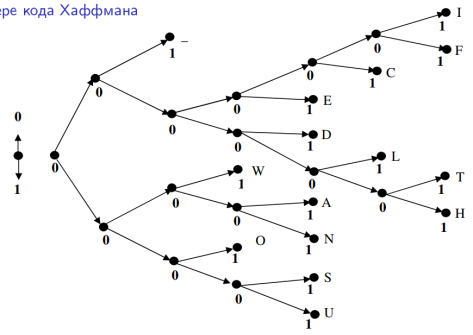
```
IF_WE_CANNOT_DO_AS_WE_WOULD_WE_SHOULD_DO_AS_WE_CAN
```

$$l(\boldsymbol{x}) = l_1(\boldsymbol{x}) + l_2(\boldsymbol{x})$$

При равномерном кодировании получим 50×8=400 бит.

x	Число появлений x в $\boldsymbol{x}$, $\tau(x)$	Длина кодового слова, $l(x)$	Кодовое слово	$\tau(x) \times l(x)$
I	1	6	010000	6
F	1	6	010001	6
_	12	2	00	24
W	5	3	100	15
E	4	4	0101	16
C	2	5	01001	10
A	4	4	1010	16
N	3	4	1011	12
O	5	3	110	15
T	1	6	011110	6
D	4	4	0110	16
S	3	4	1110	12
U	2	4	1111	8
L	2	5	01110	10
H	1	6	011111	6
Всего $l_2(\boldsymbol{x})$				178 бит

На примере кода Хаффмана



$$\boldsymbol{c}_1 = (0\ 00\ 1000\ 001010\ 01101111\ 0110\ 1111,\ \text{ASCII(x)}, \ldots)$$

$$l_1 = 29 + 8 \times 15 = 149 \text{ бит}, l = l_1 + l_2 = 149 + 178 = 327 \text{ бит}.$$

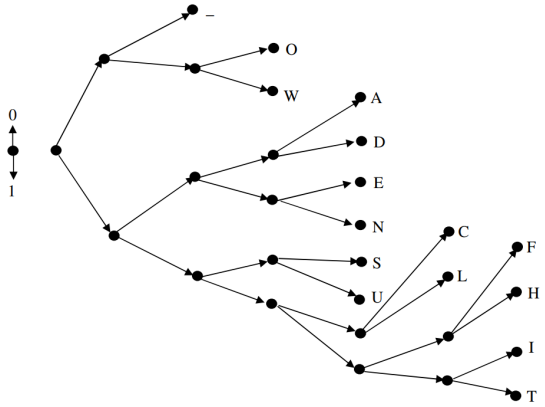
На примере канонического кода Хаффмана

- Для заданного распределения вероятностей можно построить несколько одинаково эффективных кодов Хаффмана.
- Код Хаффмана называется ***каноническим***, если его короткие кодовые слова лексикографически предшествуют более длинным.

x	Длина кодового слова $l(x)$	Кодовое слово
_	2	00
O	3	010
W	3	011
A	4	1000
D	4	1001
E	4	1010
N	4	1011
S	4	1100
U	4	1101
C	5	11100
L	5	11101
F	6	111100
H	6	111101
I	6	111110

T	6	111111

На примере канонического кода Хаффмана



- Достаточно указать количество концевых вершин для ярусов с номерами $0, ..., l_{max}$, где l_{max} – максимальная длина кодового слова.

Ярус	Число вершин	Число концевых вершин n_i	Диапазон значений n_i	Затраты в битах
0	1	0	0...1	1
1	2	0	0...2	2
2	4	1	0...4	3
3	6	2	0...6	3
4	8	6	0...8	4
5	4	2	0...4	3
6	4	4	0...4	3
Total				19

$\boldsymbol{c_1}$ = (0 00 001 010 0110 010 100 ASCII(x), ...)

$l_1 = 19 + 8 \times 15 = 139$ бит, $l = l_1 + l_2 = 139 + 178 = 317$ бит.

Скорость кодирования кодом Хаффмана

Theorem

Полное кодовое дерево, имеющее M концевых вершин, имеет $M - 1$ промежуточных вершин. Поэтому, $M + M - 1 = 2M - 1$ бит достаточно для описания полного описания дерева.