Лабораторная работа №1. Сжатие без потерь

Метод оптимального побуквенного кодирования был разработан в 1952 г. Д. Хаффманом. Оптимальный код Хаффмана обладает минимальной средней длиной кодового слова среди всех побуквенных кодов для данного источника с алфавитом $A=\{a_1,\ldots,a_n\}$ и вероятностями $p_i=P(a_i)$.

Рассмотрим алгоритм построения оптимального кода Хаффмана, который основывается на утверждениях лемм предыдущего параграфа.

1. Упорядочим символы исходного алфавита $A=\{a_1,\ldots,a_n\}$ по убыванию их вероятностей $p_1 \geq p_2 \geq \cdots \geq p_n$.

2. Если $A=\{a_1,a_2\}$, то $a_1 \to 0$, $a_2 \to 1$.

3. Если $A=\{a_1,\ldots,a_j,\ldots,a_n\}$ и известны коды $<a_j \to b_j>$, $j = 1,\ldots,n$, то для алфавита $\{a_1,\ldots a_j', a_j''\ldots,a_n\}$ с новыми символами a_j' и a_j'' вместо a_j, и вероятностями $p(a_j)=p(a_j')+ p(a_j'')$, код символа a_j заменяется на коды $a_j' \to b_j0$, $a_j'' \to b_j1$.

Пример. Пусть дан алфавит A=$\{a_1, a_2, a_3, a_4, a_5, a_6\}$ с вероятностями

$$p_1=0.36,\ p_2=0.18,\ p_3=0.18,\ p_4=0.12,\ p_5=0.09,\ p_6=0.07.$$

Здесь символы источника уже упорядочены в соответствии с их вероятностями. Будем складывать две наименьшие вероятности и включать суммарную вероятность на соответствующее место в упорядоченном списке вероятностей до тех пор, пока в списке не останется два символа. Тогда закодируем эти два символа 0 и 1. Далее кодовые слова достраиваются, как показано на рисунке 4.

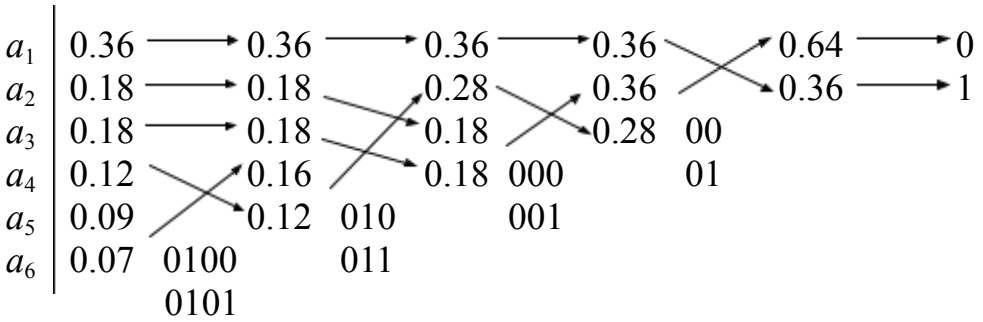

Рисунок 4 Процесс построения кода Хаффмана

Таблица 5 Код Хаффмана

a_i	p_i	L_i	кодовое слово
a_1	0.36	1	1
a_2	0.18	3	000
a_3	0.18	3	001
a_4	0.12	3	011
a_5	0.09	4	0100
a_6	0.07	4	0101

Посчитаем среднюю длину, построенного кода Хаффмана

$$L_{cp}(P)=1\cdot0.36 + 3\cdot0.18 + 3\cdot0.18 + 3\cdot0.12 + 4\cdot0.09 + 4\cdot0.07 =2.44,$$

при этом энтропия данного источника

$$H(p_1,\ldots,p_6) = -\ 0.36\cdot\log0.36 - 2\cdot0.18\cdot\log0.18 -$$
$$- 0.12\cdot\log0.12 - 0.09\cdot\log0.09 -$$
$$0.07\log0.07=2.37$$

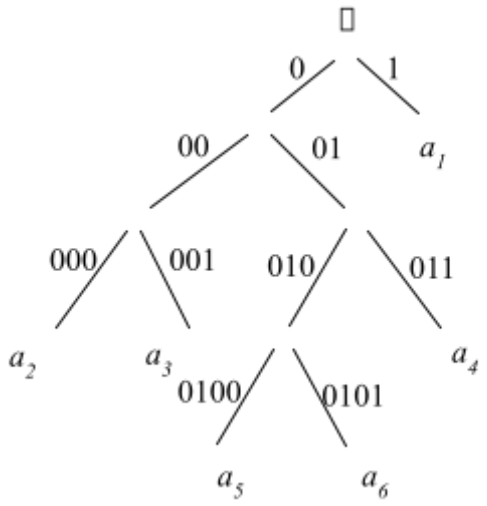

Рисунок 5 Кодовое дерево для кода Хаффмана

Код Хаффмана обычно строится и хранится в виде двоичного дерева, в листьях которого находятся символы алфавита, а на «ветвях» – 0 или 1. Тогда уникальным кодом символа является путь от корня дерева к этому символу, по которому все 0 и 1 собираются в одну уникальную последовательность (рис. 5).

Алгоритм на псевдокоде

Построение оптимального кода Хаффмана (n,P)

Обозначим
n – количество символов исходного алфавита
P – массив вероятностей, упорядоченных по убыванию
C – матрица элементарных кодов
L – массив длин кодовых слов

```
Huffman (n,P)
IF (n=2) C [1,1]:= 0, L [1]:= 1
        C [2,1]:=1, L [2]:=1
ELSE q:= P [n-1]+P [n]
        j:= Up (n,q)                (поиск и вставка суммы)
        Huffman (n-1,P)
        Down (n,j)                  (достраивание кодов)
FI
```

Функция Up (n,q) находит в массиве P место, куда вставить число q, и вставляет его, сдвигая вниз остальные элементы.

```
DO (i=n-1, n-2,…,2)
        IF (P [i−1]≤q) P [i]:=P [i−1]
        ELSE j:=i
 OD
        FI
OD
P [j]:= q
```

Процедура Down (n,j) формирует кодовые слова.

```
S:= C [j,*]  (запоминание j-той строки матрицы элем. кодов в массив S)
L:= L[j]
DO (i=j,…,n-2)
        C [i,*]:= C[i+1,*]  (сдвиг вверх строк матрицы C)
        L [i]:=L [i+1]
OD
C [n-1,*]:= S, C [n,*]:= S  (восстановление префикса кодовых слов из м-ва S)
C [n-1,L+1]:=0
C [n,L+1]:=1
L [n-1]:=L+1
L [n]:=L+1
```

Для зачета по лабораторной работе студенту необходимо представить

- Исходные тексты программ с подробными комментариями;
- Исполняемые файлы;
- Отчет по лабораторной работе.

Отчет должен включать в себя следующие разделы

- Формулировку задания
- Описание основных методов, используемых в лабораторной работе;
- Результаты работы программы (в виде файла или в виде скриншота);
- Анализ результатов.

Порядок выполнения работы

1. Изучить теоретический материал
2. Реализовать процедуру построения оптимального кода Хаффмана.
3. Построить код Хаффмана для текста на языке, обозначенном преподавателем, использовать файл не менее 1 Кб. Распечатать полученную кодовую таблицу в виде:

Символ	Вероятность	Кодовое слово	Длина кодового слова

4. Вычислить энтропию исходного файла и сравнить со средней длиной кодового слова.
5. Восстановить исходный текст

Контрольные вопросы

1. Какой код называется разделимым? Префиксным?
2. Что такое энтропия дискретного вероятностного источника?
3. Какова основная характеристика неравномерного кода?
4. Что такое избыточность кода?
5. Почему код Хаффмана называется оптимальным?