Лабораторная работа № 2
Неравномерные префиксные коды

1. Цель работы

Изучить алгоритмы сжатия информации методами Шеннона-Фано и Хаффмана, получить практические навыки в формировании архивов этими методами.

2. Общие сведения

Задачу кодирования (сжатия) сообщений источника, имеющего отличное от равномерного распределение вероятностей появления символов его алфавита, позволяют решить неравномерные префиксные коды.

Рассмотрим принцип построения таких кодов методами Шеннона-Фано и Хаффмана [1, 2]. Оба этих кода основываются на статистических свойствах источника сообщений и ставят в соответствие часто встречающимся символам алфавита короткие кодовые комбинации.

На рисунке показана гистограмма абсолютных частот появления букв русского алфавита в книге [3]

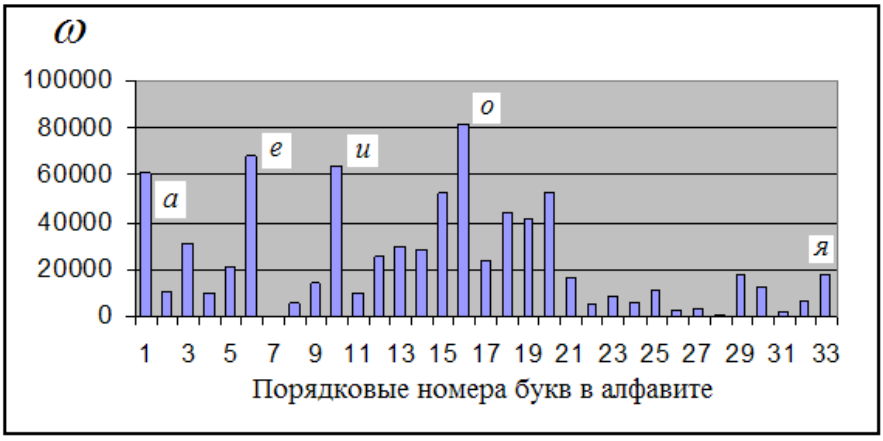


Рисунок показывает существенную неравномерность в использовании букв русского алфавита.

Идея, основанная на учёте частот появления символов в тексте, была разработана **Клодом Шенноном** и независимо от него **Робертом Фано**, а затем в 1952 г. развита **Дэвидом Хаффманом** - аспирантом Массачусетского

технологического института при написании им курсовой работы. Идея базируется на том факте, что в обычном тексте частоты появления различных символов неодинаковые. Для сжатия предложено использовать кодовые комбинации различной длины, по этой причине такие коды называют **неравномерными.**

В неравномерных методах архивации стандартные кодовые таблицы не используются, а при каждом архивировании создаются собственные кодовые таблицы (**словари**). Вид словаря каждый раз изменяется и зависит от содержимого архивируемого документа.

Как и код Шеннона-Фано, код Хаффмана требует получения априорных сведений о статистических свойствах источника сообщения, то есть необходима таблица абсолютных частот символов данного сообщения. На основе этих данных строится кодовое дерево, также называемое деревом Хаффмана или Н-деревом. В отличие от кода Шеннона-Фано, дерево Хаффмана строится в направлении от листьев к корню (в обратном направлении).

Коды Шеннона-Фано и Хаффмана обладают свойством **префиксности**. Это означает, что ни одна кодовая комбинация не является началом другой, что позволяет обеспечить однозначное декодирование и нет необходимости двоичные символы разделять пробелами.

Как уже отмечалось ранее, для кодирования и декодирования сообщений, сжатых по методам Шеннона-Фано и Хаффмана, кодек должен обладать априорной информацией о статистике сообщения. Поэтому кроме самого архива на приёмную сторону необходимо передать таблицу частот символов данного сообщения, что увеличивает длину передаваемых данных и снижает фактическую эффективность сжатия. Тем не менее, этот недостаток нивелируется при сжатии больших объёмов данных, например, при сохранении изображений в формате JPEG.

3. Задания на выполнение лабораторной работы

3.1. Задание 1. Выполнить сжатие информации методом Шеннона-Фано

Используя фразу из табл. 3.1.1, построить кодовое дерево и определить коэффициент сжатия методом Шеннона-Фано. Рассчитать энтропию и избыточность.

Табл. 3.1.1

Вар	Текст	Вар	Текст
1	Заработали 522211112	9	Кккккттттттто тттттам?
2	До дембеля 60440000 с	10	Длинношеее животное
3	Кредитка 235556999922	11	Ураааааааааааа в атаку
4	ИНН 8825577777488856	12	Долг 3255566667444444
5	Шифр 159222666644444	13	Телефон 8904222211111
6	Улов 98544477778555 кг	14	Ауууууууу заблудились
7	Пароль RRWQQQQ6666	15	Свидетельство 22263333
8	Пароль 778SSЫЫIzzzzN	16	Возраст 1000000000 лет

3.2. Задание 2. Выполнить сжатие информации методом Хаффмана

Используя фразу из табл. 3.1.1, построить кодовое дерево и определить коэффициент сжатия методом Хаффмана. Рассчитать энтропию и избыточность.

4. Порядок выполнения лабораторной работы

4.1. Методические указания к заданию 3.1

В качестве примера использования кода Шеннона – Фано рассмотрим порядок сжатия сообщения: «ИНН 637322757237».

На первом этапе построения кода Шеннона – Фано формируется таблица абсолютных частот символов.

Таблица 4.1.1

Символ	Абсолютная частота ω_i	Символ	Абсолютная частота ω_i
7	4	5	1
2	3	6	1
3	3	И	1
Н	2	Пробел	1

Заданный текст содержит избыточность, которая определяется по формуле:

$$L = \left(1 - \frac{H}{n}\right) \cdot 100\%,$$

где H - энтропия сообщения;

n - длина кодовой комбинации при равномерном кодировании.

Энтропия сообщения вычисляется по формуле:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i ,$$

где N - объем алфавита источника (для русского алфавита N =33);

p_i - относительная частота (вероятность) появления символа в сообщении.

Относительная частота встречаемости символа определяется как отношение абсолютной частоты появления символа в сообщении к общей длине сообщения (числу символов в сообщении):

$$p_i = \frac{\omega_i}{m} ,$$

где ω_i - абсолютная частота (частость) встречаемости i-ого символа алфавита источника; m – число символов в сообщении.

В данном случае энтропия сообщения равна:

$$H = -\left(\frac{4}{16}\log_2\frac{4}{16} + 2\cdot\frac{3}{16}\log_2\frac{3}{16} + \frac{2}{16}\log_2\frac{2}{16} + 4\cdot\frac{1}{16}\log_2\frac{1}{16}\right) = 2{,}781$$

бит/символ,

где $p_1 = \frac{4}{16} = \frac{1}{4}$ - относительная частота появления символа «7»;

$p_2 = p_3 = \frac{3}{16}$ - относительная частота появления символов «2» и «3»;

$p_4 = \frac{2}{16} = \frac{1}{8}$ - относительная частота появления символа «Н»;

$p_5 = p_6 = p_7 = p_8 = \frac{1}{16}$ - относительная частота появления символов «5», «6», «И», Пробел.

При необходимости расчёта логарифма по основанию два через логарифм по основанию десять можно воспользоваться соотношением:

$$\log_2 x = \frac{\lg x}{\lg 2}.$$

При использовании равномерного кода (например, СР-1251) длина кодовой комбинации определяется так:

$$n = \lceil \log_2 N \rceil,$$

$\lceil x \rceil$ - функция округления аргумента до ближайшего целого значения, не меньшего, чем x.

В данном примере $n = \lceil \log_2 8 \rceil = 3$ бита.

Избыточность сообщения при кодировании равномерным кодом равна:

$$L = \left(1 - \frac{2,781}{3}\right) \cdot 100\% = 7,3\%.$$

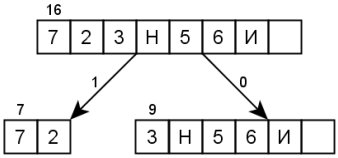
Для получения кодовых комбинаций строится кодовое дерево. При построении кода Шеннона-Фано дерево строится от корня к листьям (в отличие от настоящего дерева здесь корень располагается вверху, а листья – внизу). В качестве корня используется множество всех символов алфавита сообщения (см. рис.), упорядоченное по частоте встречаемости символов. Число сверху таблицы равно суммарной частоте символов в исходном сообщении (суммарное число символов в сообщении).

16

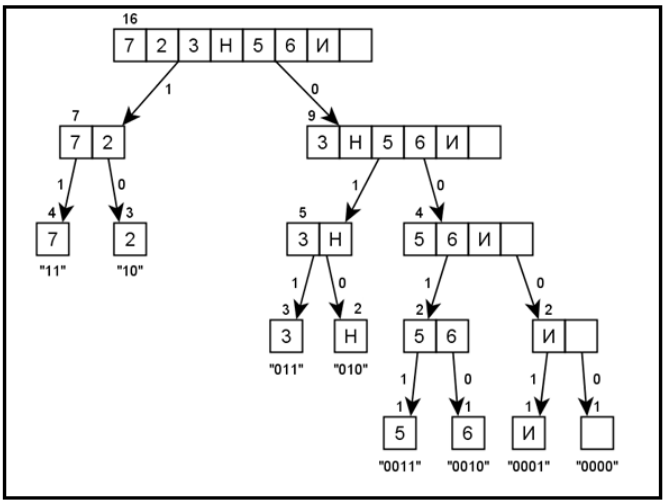
7	2	3	Н	5	6	И	

Затем множество символов делят на два подмножества так, чтобы новые множества имели равные, насколько это возможно, суммарные частоты встречаемости входящих в них символов. (Например, вес 11 желательно разделить на два подмножества 5 и 6. Если есть возможность разделения на 5 и 6, то деление на 4 и 7 будет ошибочным). Эти подмножества соединяются с корнем дерева ветвями, становясь потомками. Левая ветвь дерева обознача-

ется символом 1, а правая ветвь – символом 0 (см. следующий рисунок).



Полученные подмножества также рекурсивно делятся до тех пор, пока не будут сформированы листья дерева – отдельные символы сообщения.



Кодовые комбинации (на предыдущем рисунке они указаны в кавычках под соответствующими листьями) получаются при движении от корня дерева к кодируемому символу-листу путём сбора бит, присвоенных пройдённым ветвям дерева. Запись кодовой комбинации ведут в направлении от старших разрядов к младшим. Например, при кодировании символа «3» сначала следует пройти по правой ветви к множеству {3, Н, 5, 6, И, Пробел} (к кодовой комбинации добавляется бит 0). Затем нужно пройти по левой ветви к множеству {3, Н} (к кодовой комбинации добавляется бит 1). Наконец, нужно пройти по левой ветви, чтобы достичь листа «3». Таким образом, получена кодовая комбинация «011»

При декодировании биты считываются из входного потока и используются, как указатели направления движения по кодовому дереву от корня к искомому листу. При достижении листа найденный символ записывается в выходной поток, а движение по кодовому дереву снова начинают от корня.

Например, декодирование комбинации «010» происходит следующим образом. Из потока считывается бит 0, следовательно, нужно пройти по правой ветви от корня дерева к узлу {З, Н, 5, 6, И, Пробел}. Следующий бит единичный, что требует пройти по левой ветви к множеству {З, Н}. Наконец, следующий бит 0 приводит декодер по правой ветви к листу «Н».

В следующей таблице приведены все разрешённые комбинации полученного кода Шеннона-Фано. Это так называемый **словарь** сообщения. Он передаётся на приёмную сторону вместе с архивом.

Таблица 4.1.2

Символ	Кодовая комбинация	Символ	Кодовая комбинация
7	11	5	0011
2	10	6	0010
З	011	И	0001
Н	010	Пробел	0000

Закодированное сообщение (архив) будет иметь вид:
000101001000000010011110111010110011111001111
Общая длина закодированного сообщения составляет 45 бит.

Средняя длина кодовой комбинации равна (напомним, что число символов в сообщении – 16):

$$n = \frac{45}{16} = 2,813 \text{ бит/символ.}$$

Избыточность сообщения после применения кода Шеннона-Фано снизилась до значения:

$$L = \left(1 - \frac{2,781}{2,813}\right) \cdot 100\% = 1,13\% .$$

Несложно убедиться, что применение кода Шеннона-Фано позволило существенно уменьшить избыточность сообщения. При равномерном кодировании рассмотренного сообщения с помощью кодовой таблицы CP-1251 пришлось бы передать 128 бит.

4.2. Методические указания к заданию 3.2

Построение кодового дерева по методу Хаффмана начинают с того, что формируют набор листьев, имеющих веса, равные частотам появления символов в исходном (сжимаемом) сообщении. Листья ранжируют в соответствии с их весами (записывают веса справа налево в порядке их возрастания). Запись листьев желательно вести в одну строчку. Затем выбирают пару узлов (листьев), имеющих наименьший вес, которые соединяют дугами с новым узлом, вес которого равен сумме весов присоединённых к нему потомков. Образовавшийся узел называется родителем. Новый узел (родитель) участвует в дальнейших построениях дерева. Процедура объединения свободных узлов ведётся до тех пор, пока не останется единственный узел (корень дерева).

В отличие от метода Шеннона-Фано построение дерева ведётся не сверху-вниз, а снизу-вверх.

Предположим, что требуется выполнить сжатие фразы «**Проездной 7977977**». Ниже приведена таблица абсолютных частот использованных символов в заданном сообщении.

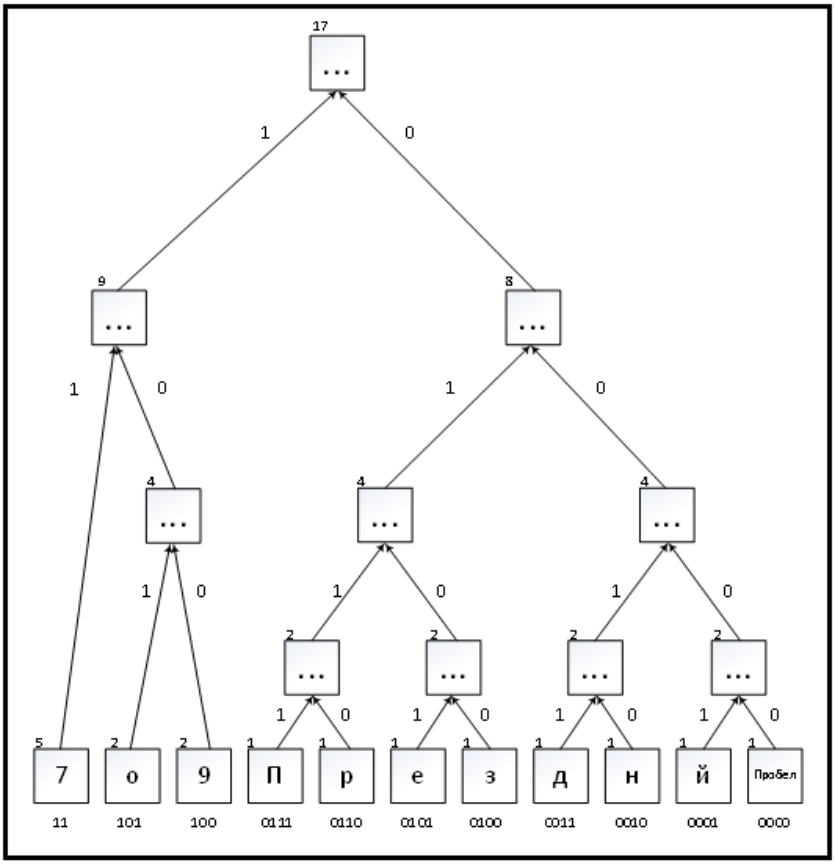
Таблица 4.2.1

Символ	Абсолютная частота ω_i	Символ	Абсолютная частота ω_i
7	5	з	1
о	3	д	1
9	2	н	1
П	1	й	1
р	1	Пробел	1
е	1		

В соответствии с построенным деревом составим словарь замен.

Таблица 4.2.2

Символ	Кодовая комбинация	Символ	Кодовая комбинация
7	11	з	0100
о	101	д	0011
9	100	н	0010
П	0111	й	0001
р	0110	Пробел	0000
е	0101		



Закодированное сообщение **Проездной 7977977** будет иметь вид:
01110110101010101000011001010100010000111001111100111 1.
Общая длина закодированного сообщения составляет 54 бит.
Энтропия сообщения имеет значение:

$$H = -\left(\frac{5}{17}\log_2\frac{5}{17} + 2\cdot\frac{2}{17}\log_2\frac{2}{17} + 8\cdot\frac{1}{17}\log_2\frac{1}{17}\right) = 3{,}17 \text{ бит/символ.}$$

Средняя длина кодовой комбинации составляет n = 3, 177 бит/символ.

5. Требования к отчёту

Отчёт подготавливается в электронном виде. Он должен содержать результаты сжатия информации с помощью двух неравномерных префиксных кодов (кодовые деревья, словари, расчёты избыточности и энтропии). В отчёте следует привести не только конечные результаты расчётов, но и описать порядок расчётов.

6. Контрольные вопросы

6.1. Перечислите известные Вам методы сжатия информации без потерь.

6.2. В чём состоит отличие методов сжатия с потерями и без потерь?

6.3. Сколько бит в управляющем байте отводят для указания числа повторяющихся байтов при сжатии методом кодирования длин серий?

6.4. О чём говорит равенство единице старшего бита в управляющем байте при сжатии методом кодирования длин серий?

6.5. Перечислите известные Вам архиваторы.

6.6. Целесообразно ли выполнять сжатие файлов формата JPEG, MP3, MPEG?

6.7. Рисунок какого формата будет сжат сильнее BMP или JPEG?

6.8. Какой код является неравномерным: RLE или Хаффмана?

6.9. Что называется кодом?

6.10. Чем отличаются алгоритмы построении кодов Шеннона-Фано и Хаффмана?

6.11. Перечислите коды, которые обладают свойством префиксности.

6.12. Что называется входным алфавитом?

7. Список литературы

1. Shannon C. E. «A mathematical theory of communication», Bell Sys. Tech. Jour., vol. 27, pp. 379-423; July, 1948.

2. Huffman D. A., «A method for the construction of minimum-redundancy codes», Proc. Inst. Radio Engineers, vol. 40, no. 9, pp. 1098-1101, Sep. 1952.

3. Алексеев А.П. Информатика 2007. – СОЛОН-ПРЕСС, 2007. – 608 с.