ЛАБОРАТОРНАЯ РАБОТА № 11

МЕТОДЫ СЖАТИЯ ПО ШЕННОНУ И ХАФФМЕНУ

Цель работы: ознакомление со статистическими принципами сжатия информации с использованием методов Шеннона — Фано и Хаффмена.

Примечание. Для выполнения лабораторной работы на компьютере необходимо установить файл **Shannon-Huffman.exe**, который находится в архиве Методы сжатия по Шеннону и Хаффмену.rar.

Описание лабораторной работы. Работа выполняется на персональном компьютере с использованием программы **Shannon-Huffman.exe**. Программа предназначена для демонстрации методов сжатия информации по алгоритмам Шеннона — Фано и Хаффмена (рис. 3.8).

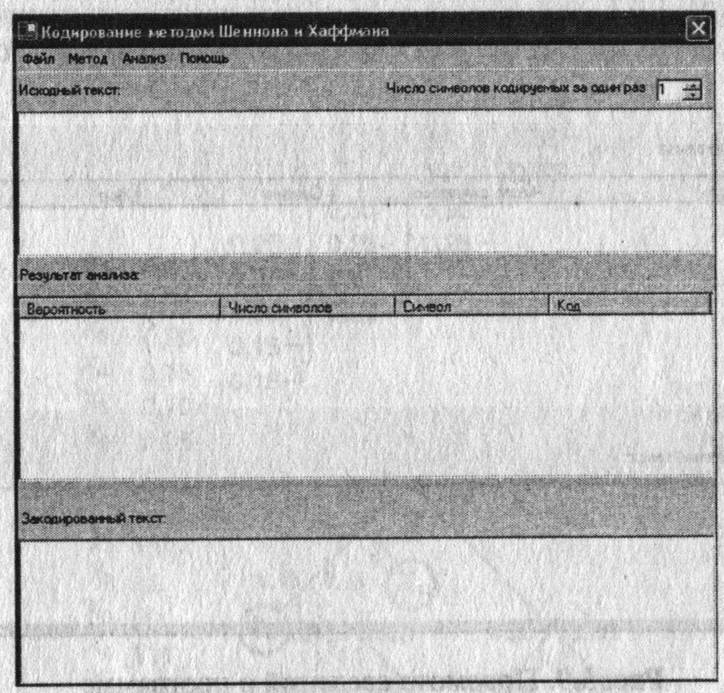


Рис. 3.8. Главное окно программы

Для работы с программой пользователь выбирает требуемый метод сжатия (см. рис. 3.8).

В окне программы ИСХОДНЫЙ ТЕКСТ записывается сообщение произвольной длины (или загружается сообщение из заранее подготовленного файла в формате .txt). Затем необходимо указать число символов, кодируемых за один раз.

Для подсчета вероятности появления букв во введенном сообщении P, определения энтропии источника сообщений H, среднего числа символов при кодировании одной буквы сообщения L необходимо выбрать закладку АНАЛИЗ.

Для определения эффективности кодирования рассчитывается избыточность кода ($L - H$).

Кратко ознакомиться со сведениями о программе вы можете, выбрав закладку ПОМОЩЬ (рис. 3.9).

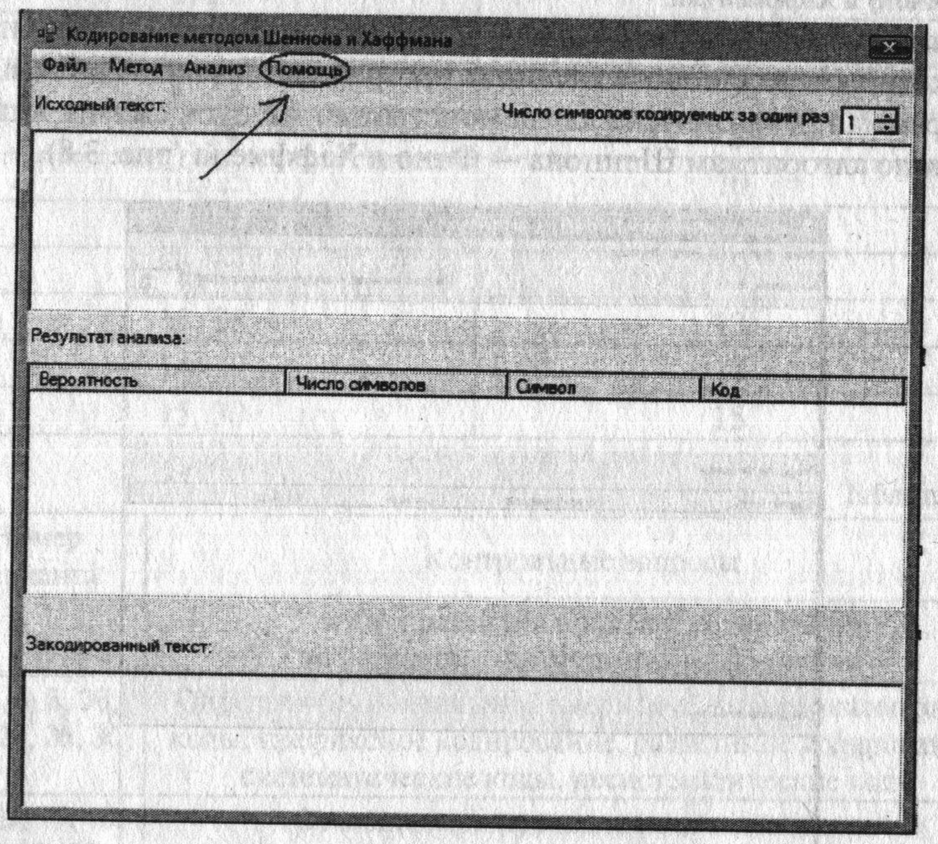


Рис. 3.9. Просмотр сведений о программе

Чтобы выйти из программы, достаточно выбрать закладку ФАЙЛ и далее ВЫХОД.

Пример 3.12. Исходное сообщение *аббввггггдддддеееее*, состоящее из символов шестибуквенного алфавита $A = \{а, б, в, г, д, е\}$, закодируем методом Хаффмена.

Определяем вероятность появления символа в исходном сообщении:

$$P = n/N,$$

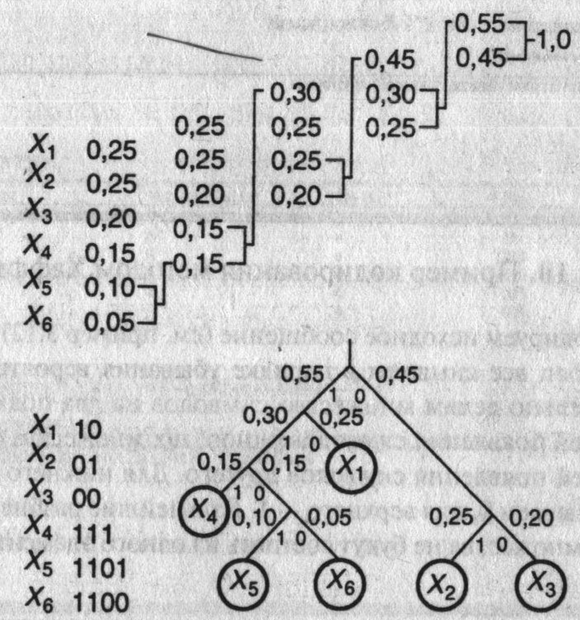
где n — число повторов символа в сообщении; N — длина сообщения.

Выпишем в столбец все символы алфавита в порядке убывания вероятности их появления в тексте (табл. 3.15).

Таблица 3.15

Символ	Число повторений символов в данном сообщении	Вероятность появления символов
д	5	0,25
е	5	0,25
г	4	0,20
в	3	0,15
б	2	0,10
а	1	0,05

Последовательно объединяя два символа с наименьшими вероятностями появления символов новый составной символ, вероятность появления которого полагается равной сумме вероятностей составляющих его символов, построим дерево, каждый узел которого имеет суммарную вероятность всех узлов, находящихся ниже него. Проследим путь к каждому листу дерева, помечая направление к каждому узлу (например, направо — 0, налево — 1):



Среднее количество символов на букву сообщения:

$$L = \sum_{i=1}^{n} P(i)n(i) = 2 \cdot 2 \cdot 0{,}25 + 2 \cdot 0{,}20 + 3 \cdot 0{,}15 + 4 \cdot (0{,}10 + 0{,}05) = 2{,}45,$$

где $n(i)$ — количество знаков в кодовой комбинации i-го символа алфавита;

$P(i)$ — вероятность появления i-го символа алфавита.

Энтропия:

$$H = \sum_{i=1}^{n} P(i) \log \frac{1}{P(i)} = 2{,}425.$$

Значение избыточности кода:

$$(L - H) = 0{,}025.$$

Сравниваем полученные данные с результатами работы программы (рис. 3.10).

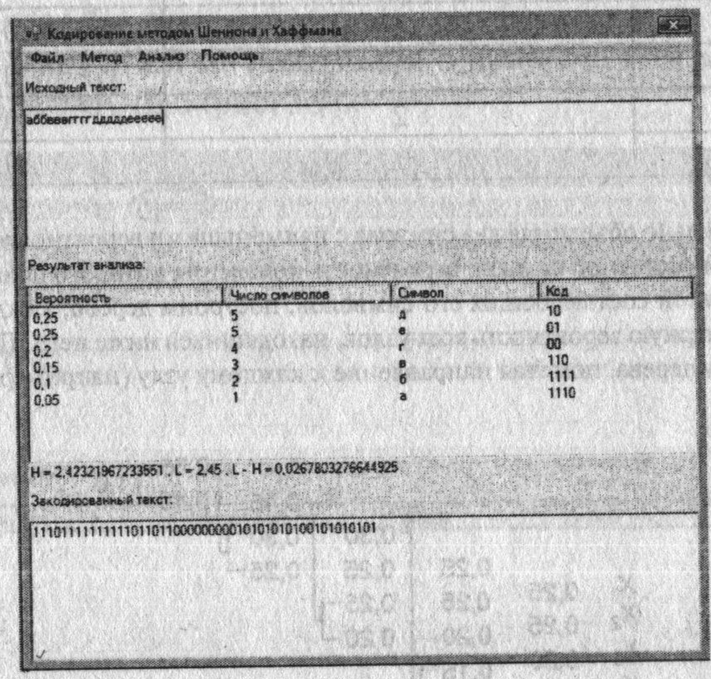


Рис. 3.10. Пример кодирования методом Хаффмена

Пример 3.13. Закодируем исходное сообщение (см. пример 3.12) методом Шеннона.

Выпишем в столбец все символы в порядке убывания вероятности появления их в тексте. Последовательно делим множество символов на два подмножества так, чтобы сумма вероятностей появления символов одного подмножества была примерно равна сумме вероятностей появления символов другого. Для нижнего подмножества каждому символу приписываем 0, для верхнего — 1. Дальнейшие разбиения повторяются до тех пор, пока все подмножества не будут состоять из одного элемента (табл. 3.16).

Таблица 3.16

Символ	Число повторений символов в данном сообщении	Вероятность появления символов	Код Шеннона		
д	5	0,25	1	1	
е	5	0,25	1	0	
г	4	0,20	0	1	1
в	3	0,15	0	1	0
б	2	0,10	0	0	1
а	1	0,05	0	0	0

Среднее количество символов на букву сообщения:

$$L = \sum_{i=1}^{n} n(i)\,P(i) = 0{,}25 \cdot 2 + 0{,}25 \cdot 2 + 0{,}2 \cdot 3 + 0{,}15 \cdot 3 + 0{,}1 \cdot 3 + 0{,}25 \cdot 3 = 2{,}5.$$

Энтропия:

$$H = \sum_{i=1}^{n} P(i) \log \frac{1}{P(i)} = 2{,}425.$$

Значение избыточности кода:

$$(L - H) = 0{,}075.$$

Сравниваем полученные данные с результатами работы программы (рис. 3.11).

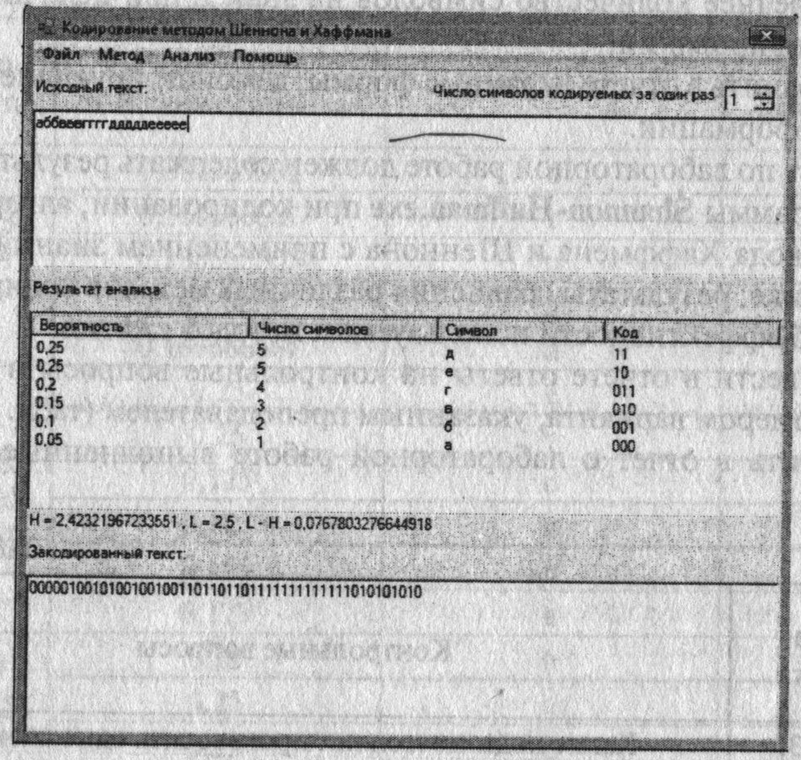


Рис. 3.11. Пример кодирования методом Шеннона

Задание

1. Ознакомиться со сведениями о программе **Shannon-Huffman.exe** и рассмотреть примеры эффективного кодировании информации.

2. Запустить программу **Shannon-Huffman.exe**, предназначенную для демонстрации методов сжатия информации с использованием алгоритмов Шеннона — Фано и Хаффмена.

3. Выполнить следующие задания:

1) число символов алфавита $k = m$ (m — номер варианта). Составить такое исходное сообщение, чтобы:
   а) символы алфавита встречались в сообщении с равными вероятностями,
   б) символы алфавита встречались в сообщении с разными вероятностями;
2) ввести произвольный связный текст на русском языке. Это может быть пословица, стихотворение или произвольный текст. Используя результаты работы программы, следует проанализировать алфавит введенного сообщения: подсчитать количество символов алфавита, значение энтропии H, среднее количество символов на знак L при целочисленном кодировании.

4. Сохранить в отчете экранные формы, демонстрирующие процесс сжатия информации.

5. Отчет по лабораторной работе должен содержать результаты анализа программы **Shannon-Huffman.exe** при кодировании; алгоритм построения кода Хаффмена и Шеннона с применением знаний по данной тематике; результаты сравнения различных методов кодирования; выводы об эффективности используемых методов сжатия.

6. Привести в отчете ответы на контрольные вопросы в соответствии с номером варианта, указанным преподавателем (табл. 3.17).

Включить в отчет о лабораторной работе выполненные задания № 1—3.

Таблица 3.17

Номер варианта	Контрольные вопросы
1, 5, 7, 3, 9, 18, 28	Какие коды позволяют производить однозначное декодирование даже без использования разделительного символа? Приведите примеры таких кодов
2, 4, 6, 8, 20, 22, 24, 26, 30	Назовите условия построения оптимальных кодов
11, 13, 15, 10, 17, 19, 27	С какой целью используются эффективные коды и какие из них вам известны?
12, 14, 16, 21, 23, 25, 29	Перечислите основные методы сжатия информации без потерь

Варианты заданий

Задание 1. Сообщение состоит из последовательности двух букв A и B, вероятности появления каждой из которых не зависят от того, какая была передана раньше, и равны 0,8 и 0,2 соответственно. Произведите кодирование по методу Шеннона: а) отдельных букв; б) блоков, состоящих из двухбуквенных сочетаний; в) блоков, состоящих из трехбуквенных сочетаний. Сравните полученные коды по их эффективности.

Задание 2. Составьте текст, который бы соответствовал данным, приведенным в табл. 3.18. Используя программу **Shannon-Huffman.exe**, закодируйте текст методом Хаффмена.

Таблица 3.18

Номер варианта	Вероятность появления символов	Символ	Число символов
1, 5, 7, 3, 9, 12, 14, 18, 28	0,333333333	о	2
	0,166666667	г	1
	0,166666667	р	1
	0,166666667	д	1
	0,166666667	а	1
2, 4, 6, 8, 16, 21, 20, 22, 24, 30	0,25	е	2
	0,25	т	2
	0,125	о	1
	0,125	п	1
	0,125	р	1
	0,125	а	1
11, 13, 15, 10, 17, 19, 23, 25	0,25	р	2
	0,25	а	2
	0,25	с	2
	0,125	т	1
	0,125	е	1

Задание 3. Для вариантов *а*, *б*, *в*, приведенных на рис. 3.12 и в табл. 3.19, составьте код Хаффмена. Рассчитайте среднее количество символов на знак L; избыточности $(L - H)$ и относительной избыточности полученного кода $(L - H)/L$. Сравните полученные значения с L, H, $(L - H)$ для кода Шеннона, сделайте выводы.

Результат анализа:

Вероятность	Число символов	Символ	Код
0,4375	7	а	11
0,1875	3	б	10
0,1875	3	г	01
0,125	2	в	001
0,0625	1	д	000

H = 2,05242128362936 , L = 2,1875 , L - H = 0,135078716170635

а

Результат анализа:

Вероятность	Число символов	Символ	Код
0,357142857142857	5	к	11
0,285714285714286	4	е	10
0,142857142857143	2	ж	011
0,142857142857143	2	з	010
0,0714285714285714	1	и	00

H = 2,1209520310264 , L = 2,28571428571429 , L - H = 0,164762254687883

б

Результат анализа:

Вероятность	Число символов	Символ	Код
0,363636363636364	4	п	11
0,272727272727273	3	о	10
0,181818181818182	2	м	01
0,0909090909090909	1	н	001
0,0909090909090909	1	т	000

H = 2,11807820934971 , L = 2,18181818181818 , L - H = 0,0637399724684737

в

Рис. 3.12. Варианты заданий

Таблица 3.19

Номер варианта	Задание
1, 5, 7, 3, 9, 12, 14, 18, 28	Задание 3. Вариант *а*
2, 4, 6, 8, 20, 22, 23, 24, 25, 30	Задание 3. Вариант *б*
11, 13, 15, 16, 21, 10, 17, 19	Задание 3. Вариант *в*