Лабораторная работа № 5

Тема: Сжатие текстовой информации

Цель работы: Освоить методику сжатия текстовой информации с использованием метода Хаффмана и арифметического кодирования

Работа состоит в выполнении двух заданий: построения кода Хаффмана и арифметического кода.

Краткие теоретические сведения. Описание метода Хаффмана

Метод был предложен Хаффменом в 1952 году. Этот алгоритм стал базой для большого количества программ сжатия информации. Например, кодирование по Хаффмену используется в программах сжатия **ARJ**, **ZIP**, **RAR**, в алгоритме сжатия графических изображений с потерями **JPEG**, а также встроено в современные факс-аппараты.

Алгоритм представлен в виде последовательности шагов:

1. Буквы входного алфавита образуют список свободных узлов будущего дерева кодирования. Каждый узел в этом списке имеет вес, равный вероятности появления соответствующей буквы в сообщении.
2. Выбираются два свободных узла дерева с наименьшими весами. Если имеется более двух свободных узлов с наименьшими весами, то можно брать любую пару.
3. Создается их родитель с весом, равным их суммарному весу.
4. Родитель добавляется в список свободных узлов, а двое его детей удаляются из этого списка.
5. Одной дуге, выходящей из узла-родителя, ставится в соответствие бит 1, другой -- 0.
6. Пункты 2, 3, 4, 5 повторяются до тех пор, пока в списке свободных узлов не останется только один узел. Этот узел будет являться корнем дерева. Его вес получается равным единице -- суммарной вероятности всех букв сообщения.

Пример построения кодов Хаффмана приведен ниже.

Таблица

X_i	Вероят-ности	Шаговая процедура (кодовое дерево)							Кодовые слова
	$p(X_i)$	1	2	3	4	5	6	7	
X_1	0,729	0,729	0,729	0,729	0,729	0,729	0,729	1	1
X_2	0,081	0,081	0,081	0,081	0,109	0,162	0,271	0	011
X_3	0,081	0,081	0,081	0,081	0,081	0,109	0		010
X_4	0,081	0,081	0,081	0,081	0,081	0			001
X_5	0,009	0,010	0,018	0,028	0				00011
X_6	0,009	0,009	0,010	0					00010
X_7	0,009	0,009	0						00001
X_8	0,001	0							00000

Содержание первой части задания

1. Для выполнения работы с использованием метода Хаффмана необходимо подготовить следующие таблицы:

Таблица 1

Алфавит источника	Обозначения кодовых слов	Вероятность $p(S_i)$	Код Хафмена	Число символов
x_1	X1			
x_2	X2			

Группировка по 2

Таблица 2

Алфавит источника	Обозначения кодовых слов	Вероятность $p(S_i)$	Код Хафмена	Число символов
$x_1\ x_1$	Y1			
$x_1\ x_2$	Y2			
$x_2\ x_1$	Y3			
$x_2\ x_2$	Y4			

Группировка по 3 Таблица 3

Алфавит источника	Обозначения кодовых слов	Вероятность $p(S_i)$	Код Хафмена	Число символов
$x_1\ x_1\ x_1$	Z1			
$x_1\ x_1 x_2$	Z2			
$x_1 x_2\ x_1$	Z3			
.....	.....			
$x_2\ x_2\ x_2$	Z8			

Группировка по 4 Таблица 4

Алфавит источника	Обозначения кодовых слов	Вероятность $p(S_i)$	Код Хафмена	Число символов
$x_1\ x_1\ x_1\ x_1$	Q1			
$x_1 x_1\ x_1 x_2$	Q2			
$x_1 x_1 x_2\ x_1$	Q3			
.....	.....			
$x_2\ x_2\ x_2\ x_2$	Q16			

Таблица 5

ℓ	$p(x_k)$	$H(S)$	$H_1(S)$	n_{minc}	$\chi_и$	$\bar{n}$	$\bar{n}_c$	$n_{minc}/\ \bar{n}_c$	$\chi_к$
1									
2									
3									
4									

Где

- среднюю длину кодового слова $\bar{n}$,
- среднюю длину на символ $\bar{n}_c$,
- энтропию источника $H(S)$,
- удельную энтропию $H_1(S)$,
- избыточность источника $\chi_и$,
- избыточность кода $\chi_к$,

2. Задать произвольным образом вероятности появления символов.
3. Заполнить все таблицы.
4. Построить на одном графике зависимости $H_{max}(S) = f(\ell)$ и $H1_{max}(S) = f(\ell)$, где ℓ - количество объединяемых в блок исходных символов x1 и x2, ($\ell = 1, 2, 3, 4$).
   На основании данных таблицы 5 построить следующие зависимости: $\bar{n} = f(\ell)$ и $n_c = f(\ell)$ на одном графике; $\chi_и = f(\ell)$ и $\chi_к = f(\ell)$ на одном графике;
5. Сделать выводы

Содержание второй части задания. Построение арифметического кода

Краткие теоретические сведения о методе.

Арифметическое кодирование является методом, позволяющим упаковывать символы входного алфавита без потерь при условии, что известно распределение частот этих символов. Арифметическое кодирование является оптимальным, достигая теоретической границы степени сжатия.

Текст, сжатый арифметическим кодером, рассматривается как некоторая двоичная дробь из интервала [0, 1). Результат сжатия можно представить как последовательность двоичных цифр из записи этой дроби. Каждый символ исходного текста представляется отрезком на числовой оси с длиной, равной вероятности его появления и началом, совпадающим с концом отрезка символа, предшествующего ему в алфавите. Сумма всех отрезков, очевидно должна равняться единице. Если рассматривать на каждом шаге текущий интервал как целое, то каждый вновь поступивший входной символ “вырезает” из него подинтервал пропорционально своей длине и положению.

Построение интервала для сообщения "АБВГ...":

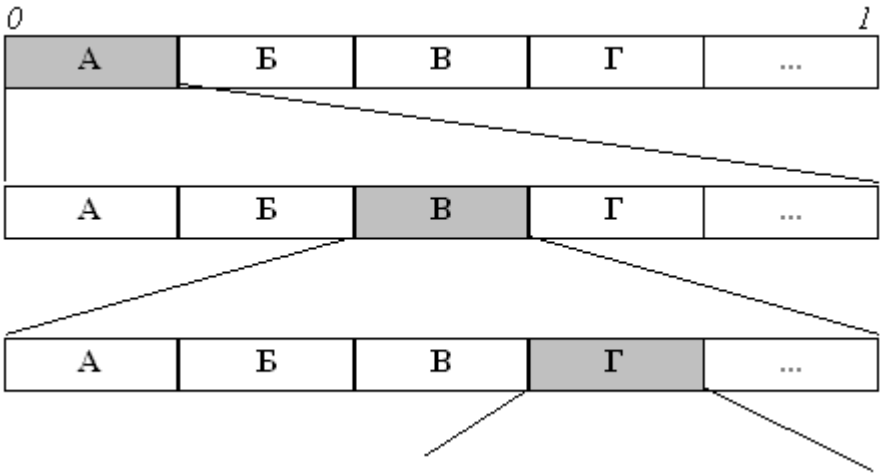

Рисунок 1. Графическая иллюстрация арифметического кодирования

Поясним работу метода на примере:

Пусть алфавит состоит из двух символов: “a” и “b” с вероятностями соответственно 3/4 и 1/4.

Рассмотрим (открытый справа) интервал [0, 1). Разобьем его на части, длина которых пропорциональна вероятностям символов. В нашем случае это [0, 3/4) и [3/4, 1). Суть алгоритма в следующем: каждому слову во входном алфавите соответствует некоторый подинтервал из [0, 1). Пустому слову соответствует весь интервал [0, 1). После получения каждого очередного символа арифметический кодер уменьшает интервал, выбирая ту его часть, которая соответствует вновь поступившему символу. Кодом сообщения является интервал, выделенный после обработки всех его символов, точнее, число минимальной длины, входящее в этот интервал. Длина полученного интервала пропорциональна вероятности появления кодируемого текста.

Выполним алгоритм для цепочки “aaba”:

Шаг	Просмотренная цепочка	Интервал
0	“”	[0, 1) = [0, 1)
1	“a”	[0, 3/4) = [0, 0.11)
2	“aa”	[0, 9/16) = [0, 0.1001)
3	“aab”	[27/64, 36/64) = [0.011011, 0.100100)
4	“aaba”	[108/256, 135/256) = [0.01101100, 0.10000111)

На первом шаге мы берем первые 3/4 интервала, соответствующие символу "a", затем оставляем от него еще только 3/4. После третьего шага от предыдущего интервала останется его правая четверть в соответствии с положением и вероятностью символа "b". И, наконец, на четвертом шаге мы оставляем лишь первые 3/4 от результата. Это и есть интервал, которому принадлежит исходное сообщение.

Содержание второй части задания

1. Закодировать арифметическим кодом текст на русском языке. Размер текста определен в соответствии с вариантом. Использовать усредненные частоты появления символов.

Вариант	Среднее число символов текста
1.	1000
2.	1500
3.	2000
4.	1600
5.	1000
6.	1200
7.	800
8.	900
9.	1300
10.	1400
11.	1100
12.	950
13.	1150
14.	1250
15.	1350
16.	1450
17.	1650

18.	1550
19.	1350
20.	870

Для выполнения обоих заданий должен быть разработан проект.

Проект должен обеспечивать вывод результатов для таблиц 1-5 (задание 1). Для арифметического кода проект должен обеспечивать ввод исходного текста и вывод соответствующего ему кода в виде дробного числа и двоичного кода.

Письменный отчет по лабораторной работе должен содержать:

1. Титульный лист. (Название лабораторной работы. Фамилия, имя, отчество, номер группы исполнителя, дата сдачи.)
2. Все требуемые математические соотношения для вычислений.
3. Распечатку текстов подпрограмм.
4. Выводы по лабораторной работе. Выводы содержат сравнительный анализ методов Хаффмана и арифметического кодирования.