Лабораторная работа №9. Методы эффективного кодирования информации

1. Цель работы

Изучить алгоритм Хаффмена для оптимального префиксного кодирования алфавита с минимальной избыточностью.

2. Теоретические сведения

Алгоритм Хаффмена (англ. Huffman) – алгоритм оптимального префиксного кодирования алфавита с минимальной избыточностью был разработан в 1952 году доктором Массачусетского технологического института Дэвидом Хаффменом. Он более практичен и никогда по степени сжатия не уступает методу Шеннона-Фэно, более того, он сжимает максимально плотно.

Этот метод кодирования состоит из двух основных этапов:

Построение оптимального кодового дерева.

Построение отображения код-символ на основе построенного дерева.

Алгоритм кодирования может быть представлен следующим образом:

Символы первичного алфавита m_1 выписывают в порядке убывания вероятностей.

Последние n_0 символов объединяют в новый символ, вероятность которого равна сумме этих символов, удаляют эти символы и вставляют новый символ в список остальных на соответствующее место (по вероятности). n_0 вычисляется из системы:

$$\begin{cases} 2 \le n_0 \le m_2 \\ n_0 = m_1 - a \cdot (m_2 - 1) \end{cases}$$

где a – целое число,

m_1 и m_2 – мощность первичного и вторичного алфавита соответственно.

Последние m_2 символов снова объединяют в один и вставляют его в соответствующей позиции, предварительно удалив символы, вошедшие в объединение.

Предыдущий шаг повторяют до тех пор, пока сумма всех m_2 символов не станет равной 1.

Этот процесс можно представить как построение дерева, корень которого – символ с вероятностью 1, получившийся при объединении символов из последнего шага, его m_2 потомков – символы из предыдущего шага и т.д.

Каждые m_2 элементов, стоящих на одном уровне, нумеруются от 0 до m_{2-1}. Коды получаются из путей (от первого потомка корня и до листка). При декодировании можно использовать то же самое дерево, считывается по одной цифре и делается шаг по дереву, пока не достигается лист – тогда выводится символ, стоящий в листе и производится возврат к корень.

Для двоичного кода описанная методика значительно упрощается. В этом случае $n_0 = m_2 = 2$, а полученное в результате построения дерево называется двоичным.

Рассмотрим применение алгоритма Хаффмена для последовательности *«aa bbb cccc ddddd»* (таблица 12 , таблица 13).

Энтропия исходного сообщения равна:

$$H = \sum_i P_i H_i = -\sum P_i p_i(j) \cdot \log p_i(j) = -\sum p_i \cdot \log p_i = -\frac{5}{17}\log_2\frac{5}{17} - \frac{4}{17}\log_2\frac{4}{17} - 2 \cdot \frac{3}{17}\log_2\frac{3}{17} - \frac{2}{17}\log_2\frac{2}{17} \approx 2.2569$$

Таблица 12 - Процедура оптимального двоичного кодирования

Вероятности					
Оригинальное множество сообщений		Дополнительные множества сообщений			
		Шаг 1	Шаг 2	Шаг 3	Шаг 4
Символ	Вероятность				
				10/17	1
			7/17	7/17	
d	5/17	5/17	5/17		
		5/17	5/17		
c	4/17	4/17			
<space>	3/17	3/17			
b	3/17				
a	2/17				

Таблица 13 - Результат кодирования по методу Хаффмена

Символ	p_i	L_i	$p_i L_i$	Код
d	5/17	2	0.59	00
c	4/17	2	0.47	10
<space>	3/17	2	0.35	11
b	3/17	3	0.53	010
a	2/17	3	0.35	011
			$ML(X)$ = 2.29	

Отметим, что в ряде случаев вместо таблицы, отражающей процедуру кодирования по методу Хаффмена, удобнее работать со списком букв и весами:

$$d_5\, c_4\, <space>_3\, b_3\, a_2,$$

осуществляя с помощью скобок группировку символов. Например, первый шаг может быть представлен следующим образом:

$$d_5\,(b_3 + a_2)_5\, c_4\, <space>_3$$
$$(c_4 + <space>_3)_7\, d_5\,(b_3 + a_2)_5$$
$$[d_5 + (b_3 + a_2)_5]_{10} + (c_4 + <space>_3)_7$$

Полученные соотношения удобно представлять в виде дерева, по которому можно составить код для каждого символа (рисунок 23).

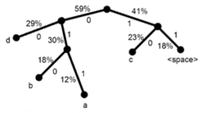

Рисунок 23 - Кодовое дерево

Средняя длина сообщения, кодирующего заданную последовательность, равна:

$$ML(X) = \Sigma L_i \cdot p_i = 2 \cdot \frac{5}{17} + 2 \cdot \frac{4}{17} + 2 \cdot \frac{3}{17} + 3 \cdot \frac{3}{17} + 3 \cdot \frac{2}{17} \approx 2.2941$$

3. Порядок выполнения работы

- Ознакомиться с теоретическими сведениями.
- Получить вариант задания у преподавателя.
- Выполнить задание.
- Продемонстрировать выполнение работы преподавателю.
- Оформить отчет.
- Защитить лабораторную работу.

4. Требования к оформлению отчета

Отчет по лабораторной работе должен содержать следующие разделы:

- титульный лист;
- цель работы:
- задание на лабораторную работу;
- ход работы;
- ответы на контрольные вопросы;
- выводы по проделанной работе.

5. Задание на работу

Построить кодовое дерево и код Хаффмена для последовательности символов в соответствии с вариантом (таблица 14).

Таблица 14 - Варианты заданий на работу

Вариант	Текст
1	one important method of transmitting messages
2	transmit in their place sequences of symbols.
3	there are more messages which might be sent t
4	there are kinds of symbols available, then so
5	the messages must use more than one symbol. I
6	is assumed that each symbol requires the same

Подсчитайте энтропию исходного сообщения и среднюю длину кодирующего сообщения.

6. Контрольные вопросы

1. Какова суммарная вероятность всех символов, участвующих в кодировании по методу Хаффмена?
2. Сколько раз кодеру Хаффмена необходимо просматривать сжимаемый текст для получения окончательного результата?
3. Может ли среднее количество бит на единицу сообщения для кодирования по методу Хаффмена быть меньше энтропии сообщения? Почему?
4. Нужно ли при кодировании по методу Хаффмена кроме сжатого сообщения передавать какую-либо дополнительную информацию? Поясните ответ.
5. Какой вариант сжатия – обратимое или необратимое – реализует алгоритм Хаффмена?
6. Почему кодирование по Хаффмену называется префиксным?

7. Литература

1. Huffman D. A Method for the Construction of Minimum-Redundancy Codes. Proceedings of IRE, vol.40, N9, pp.1098-1101, September 1952.
2. Кормен Т., Лейзерсон Ч., Ривест Р. Алгоритмы: построение и анализ. – М.: МЦНМО, 2001. – 960 с.