Министерство образования и науки РФ

Севастопольский государственный университет

Институт информационных технологий и управления в технических системах

Лабораторная работа №1

ИССЛЕДОВАНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ КОДИРОВАНИЯ, СЖАТИЯ И ШИФРОВАНИЯ ДАННЫХ

Выполнил:

ст.гр. ИСб-22д

Воронин И.Ю.

Проверил:

Маслова М.А

Севастополь

2015

1.Цель работы

Изучение особенностей методов статического сжатия данных и алгоритмов простого безизбыточного и оптимального кодирования (Шеннона- Фано, Хаффмана, арифметического), а также приобретение практических навыков информационных расчетов эффективности работы кодеров/декодеров систем передачи и хранения данных.

2.Вариант задания

Вариант 99

Вышел месяц из тумана, вынул пышку из кармана: — буду деточек кормить, а тебе, дружок, водить!

3.Краткая теория

Кол-во информации: I = n\*∑pi\*log21/pi

Энтропия: H = ∑pi\*log21/pi

Избыточность: D= 1- H/log2m

4.Ход работы

Проанализировав данное сообщение, получим следующие количества символов в тексте:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| а- 5 | ё- 0 | л- 2 | с-1 | ш-2 | я-0 |
| б- 1 | ж-1 | м-2 | т- 4 | щ-0 | :-1 |
| в-3 | з-0 | н-3 | у-5 | ъ-0 | - -1 |
| г-0 | и-2 | о-4 | ф-0 | ы-3 | ,-3 |
| д-4 | й-0 | п-1 | ч-1 | ь-2 | !-1 |
| е -4 | к-5 | р- 3 | ц-1 | ю-0 | Пробел-15 |

Всего используется 27 символов. Вычислим частоту их встречаемости.

Таблица частоты встречаемости символов



Рассчитаем информационные параметры источника и сообщения: количество информации, энтропию, избыточность:

Кол-во информации: m=27, n=80, значит I = 8,19201\*80.

Энтропия: H = 8,19201.

Избыточность: D= 0,72.

Закодируем сообщение двоичными безизбыточным кодом.

Всего у нас 27 символов, вычислим, сколько необходимо разрядов двоичного числа, чтобы их закодировать.

2m>= 27 значит m = 5;

Значит все символы будут закодированы комбинациями от 00000 до 110011. И в таком случае сообщение будет иметь вес: 80 \*5= 400 бит. В кодировке ASCCI каждый символ имеет 1 байт (8 бит). Следовательно, объём данных был бы равен 8\*80 = 640 бит. Значит было использовано на 37,5% меньше бит.

|  |  |  |
| --- | --- | --- |
| пробел | 15 | 00000 |
| а | 5 | 00001 |
| к | 5 | 00010 |
| у | 5 | 00011 |
| д | 4 | 00100 |
| е | 4 | 00101 |
| о | 4 | 00110 |
| Т | 4 | 00111 |
| в | 3 | 01000 |

…………………………….

Закодируем сообщение методом Шеннона-Фано.

Подсчёты проводятся в программе Microsoft Excel. Запишем частоту каждого символа. В сумме все частоты дают 1, что есть проверкой на правильность подсчёта букв в предложении. Делим на 2 группы данные ячейки, по принципы примерного равенства сумм их содержимого. Верхняя ячейка деления будет принимать значение ноль, а нижняя-единица. Продолжаем деление, пока для каждой из цифр можно будет представить двоичный код.

Количество цифр в записи двоичного кода равно количеству битов, которым будет закодирован сигнал. Суммарное число битов равно в данном случае 285. Если бы мы просто кодировали символы 5-разрядным двоичным кодом, то в таком случае объём памяти был бы равен 400 бит. Это значит, что на 125 бит(29%) сигнал был сжат.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| пр. | 15 | 0,1875 | 0 | 0 | 0 | |  |  |  |  | 0 |
| а | 5 | 0,0625 | 1 | 0 |  |  |  |  | 10 |
| к | 5 | 0,0625 | 1 |  |  |  |  | 11 |
| у | 5 | 0,0625 | 1 | 0 | 0 |  |  |  |  | 100 |
| д | 4 | 0,05 | 1 |  |  |  |  | 101 |
| е | 4 | 0,05 | 1 | 0 |  |  |  |  | 110 |
| о | 4 | 0,05 | 1 |  |  |  |  | 111 |
| т | 4 | 0,05 | 1 | 0 | 0 | 0 |  |  |  |  | 1000 |
| в | 3 | 0,0375 | 1 |  |  |  |  | 1001 |
| н | 3 | 0,0375 | 1 | 0 | 0 |  |  |  | 10100 |
| р | 3 | 0,0375 | 1 |  |  |  | 10100 |
| ы | 3 | 0,0375 | 1 |  |  |  |  | 1011 |
| , | 3 | 0,0375 | 1 | 0 | 0 | 0 |  |  |  | 11000 |
| и | 2 | 0,025 | 1 |  |  |  | 11001 |
| л | 2 | 0,025 | 1 | 0 | 0 |  |  | 110100 |
| м | 2 | 0,025 | 1 |  |  | 110101 |
| ш | 2 | 0,025 | 1 |  |  |  | 11011 |
| ь | 2 | 0,025 | 1 | 0 | 0 |  |  |  | 11100 |
| б | 1 | 0,0125 | 1 | 0 | 0 |  | 1110100 |
| ж | 1 | 0,0125 | 1 |  | 1110101 |
| п | 1 | 0,0125 | 1 |  |  | 111011 |
| с | 1 | 0,0125 | 1 | 0 | 0 | 0 |  | 1111000 |
| ч | 1 | 0,0125 | 1 |  | 1111001 |
| ц | 1 | 0,0125 | 1 |  |  | 111101 |
| : | 1 | 0,0125 | 1 | 0 | 0 |  | 1111100 |
| - | 1 | 0,0125 | 1 |  | 1111100 |
| ! | 1 | 0,0125 | 1 |  |  | 111111 |

Таблица 4.1.-Применение метода Шеннона-Фано.

Закодируем сообщение методом Хаффмана.

В данном разборе метода будут приведены лишь основные данные, так как из-за длины данного сообщения существует несколько вариантов, а так же это крайне трудоёмкий процесс при реализации вручную.

В данном методе мы последовательно объединяем два равных значения(параллельно возможно несколько таких операций). Делаем это до момента, пока не будет только одно множество.

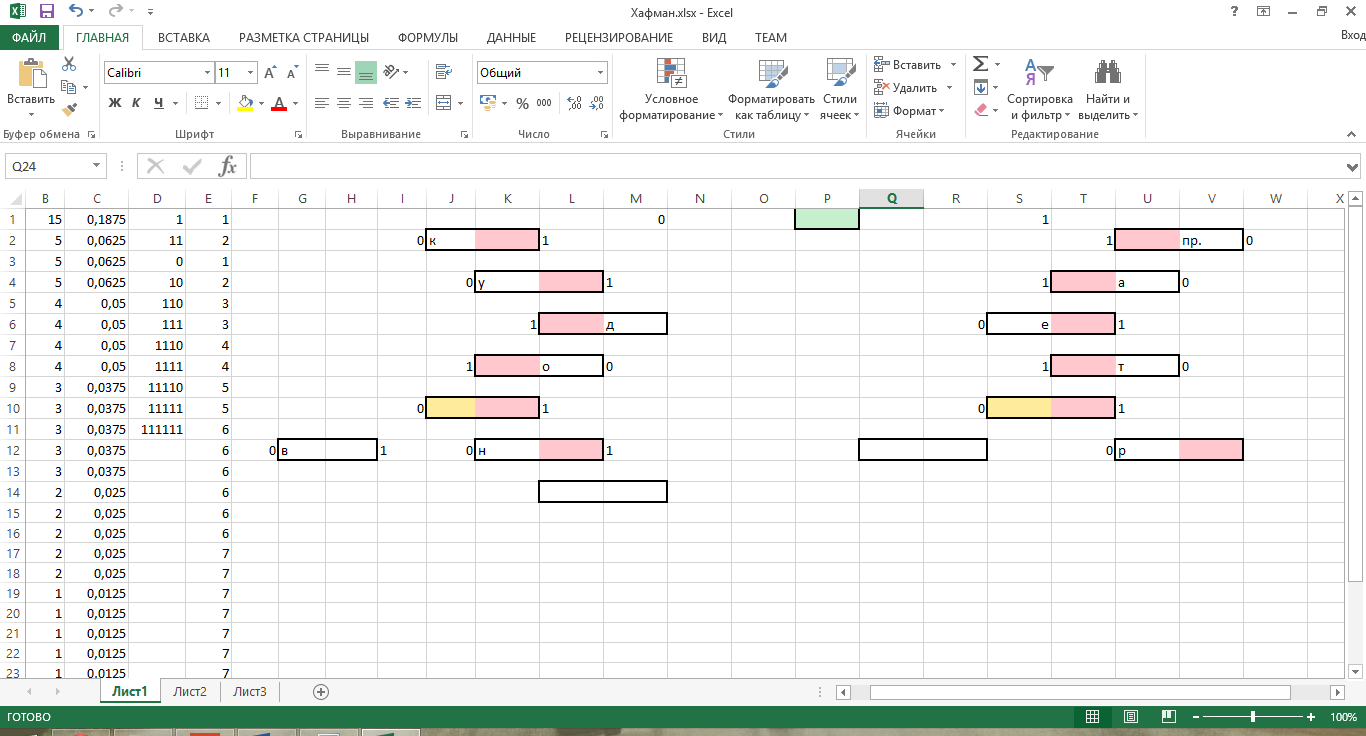


Таблица 4.2- Метод Хаффмана.

После этого начинается обратный ход. Код символа образуется путём слияния в один двоичный код всех цифр от листа до корня. Следовательно, приблизительная сумма байт будет равна: 307-320 байт. Это меньше, чем при безизбыточным двоичном кодировании, однако больше, чем при методе Шеннона-Фано.

Арифметический код.

Составим таблицу значений по вероятности, для реализации арифметического кода.



Проведём анализ по первым 6 символам.



Таким образов анализируется всё предложение, а после необходимо взять логарифм от числа из последнего диапазона. Допустим, что для диапазона №6 возьмём число 0,015968293. Результат 1,79 дит. Тогда примерно для предложения получим 24 дита.

ВЫВОДЫ

Были изучены различные методы кодирования информации, которые в разной степени уменьшали передаваемый объём данных или упрощали их передачу. Изучены особенности методов статического сжатия данных и алгоритмов простого безизбыточного и оптимального кодирования.