УДК 004.65

Алгоритм сжатия текста со словарем

Куприянов А.Б.
Белорусский национальный технический университет

Большинство алгоритмов сжатия информации рассматривают текст как простой набор символов не учитывая особенности языка. Если составить словарь различных комбинаций символов текста, закодированного с помощью таблицы ASCII, то количество комбинаций из m символов будет определяться числом размещений из 255 символов по m. Количество таких комбинаций будет определяться формулой

$$A_n^m = \frac{n!}{(n-m)!},$$

где n =255 – количество символов в таблице ASCII.

Число размещений для комбинаций из 5 символов составляет $1{,}03*10^{12}$. Значит полный словарь комбинаций будет иметь размер порядка $S=10^{12}$. Для нумерации комбинаций потребуется число байт равное $N=(\log_2 S)/8=5$. Для замены комбинаций из 5 символов, занимающей 5 байт потребуется число размером 5 байт. Значит никакого сжатия не произойдет.

В русском языке около 500 тысяч слов [1]. «Толковый словарь живого великорусского языка» В. И. Даля насчитывает около 200 тысяч слов. Наиболее употребительными словами, согласно «Частотному словарю русского языка» под редакцией Л. Н. Засориной, являются около 30 тысяч слов, а наибольшую частоту имеют чуть более 6 тысяч слов, покрывающих более 90 % обработанных при составлении этого словаря текстов.

По современным оценкам словарный запас учащегося первого класса школы составляет 2000 слов. Человек с высшим образованием знает порядка 10 тыс. слов, эрудиты – до 50 тыс. слов.

Следовательно, реальный словарь наиболее употребительных слов русского языка содержит примерно 10 тысяч слов

Для исследования возможностей сжатия текста при использовании номеров слов из словаря нужно определить

1. Сколько существует слов из N символов (реальный размер словаря) для различных языков (русский, английский, язык программирования, html-страницы).

2. Какой реальный коэффициент сжатия может обеспечить предложенный метод.

Разработана программа, позволяющая пополнять словарь, сжимать и распаковывать фалы, использование которой даст ответы на оба вышепоставленных вопроса.