УДК 004.42

Двухэтапное сжатие текстовых данных

Куприянов А. Б.
Белорусский национальный технический университет

Словарный метод сжатия текстовой информации, предполагающий замену слова в тексте его номером в словаре, может обеспечить коэффициент сжатия, определяемый формулой

$$K_{сж} = \frac{L_{ср} \cdot k \cdot 8}{ceil(\log_2 V)}, \quad (1)$$

где $L_{ср}$ – средняя длина слова в исходном тексте; k – количество байт на символ в исходном тексте (k = 1 для ASCII кодировки и k = 2 для Unicode кодировки); V – объем словаря (количество слов в словаре).

Исследования нескольких десятков сайтов различной тематики показали, что объемы отдельных статей обычно не превышают 5 кБ, средняя длина слова $L_{ср}$ = 8, а объем словаря V = 2-3 тысячи слов. Следовательно, размер кода слова может составлять 11–12 бит, а коэффициент сжатия может достигать значения 5k.

Для увеличения коэффициента сжатия предлагается использовать второй этап сжатия, заключающийся в создании словаря фраз и использовании в сжатом тексте номеров фраз из словаря. Повторяющиеся фразы из текста могут быть выделены алгоритмом LZW, в котором вместо символов используются 12 битовые коды слов, полученные на первом этапе сжатия.

При таком двухэтапном сжатии текста коэффициент сжатия можно оценить по формуле

$$K_{сж} = \frac{N_{исх} \cdot k \cdot 64}{12 \cdot N_{сж}}, \quad (2)$$

где $N_{исх}$ – количество слов в исходном тексте; $N_{сж}$ – количество 12-битовых кодов в сжатом тексте, определяемое формулой

$$N_{сж} = N_{исх} - M_2 - 2M_3 - ... - (p-1) \cdot M_p \quad (3)$$

где M_i – количество комбинаций из i слов в исходном тексте.

Исследования сайтов в области компьютерной техники и программирования показали, что статьи размером до 500 слов могут содержать до 20 двухсловных фраз и до 10 трехсловных фраз. Следовательно, для таких статей возможно обеспечить коэффициент сжатия $К_{сж}$ = 11. С увеличением размера статьи и количества статей в базе данных коэффициент сжатия будет увеличиваться за счет появления фраз с большим количеством слов и увеличения количества фраз в тексте.