УДК 004.42

ДВУХУРОВНЕВОЕ СЖАТИЕ ТЕКСТОВЫХ ДАННЫХ В БАЗЕ ДАННЫХ

Куприянов А.Б., Усович В.А.
Белорусский национальный технический
университет

Базы данных широко применяются в системах автоматизированного формирования и обработки документов для хранения готовых документов или их фрагментов. При этом основная часть информации хранится в текстовом виде и имеет достаточно большой объем. Для уменьшения объема хранимой информации и ускорения ее передачи по сети целесообразно использовать словарный алгоритм сжатия, позволяющий получить коэффициент сжатия $K_{сж}$=10. Словарь языка служебных документов содержит примерно 3-4 тысячи слов, поэтому для кодирования слова достаточно 12 бит, что позволяет использовать двухбайтное кодирование слов с выделением 4 бит на служебную информацию, определяющую особенности написания слова (с заглавной буквы, полностью заглавными буквами, со знаками препинания в конце слова). В служебной документации часто используются одни и те же фразы, поэтому представляется целесообразным в дополнение к словарю слов сформировать словарь фраз. В этом словаре каждой фразе, состоящей из закодированных слов можно поставить в соответствие некоторый код, длина которого будет определяться количеством используемых фраз. Считая, что количество фраз не превосходит 1-2 тысяч можно организовать их двухбайтное кодирование с выделением нескольких бит для служебной информации.

Алгоритм двухуровневого сжатия текста можно сформулировать следующим образом.

1. Чтение текста, выделение слов и внесение новых слов в словарь слов.
2. Замена слов текста их кодами из словаря и дополнение кода до двух байт служебной информацией.
3. После формирование нескольких образцов текста в базе данных чтение сжатого текста, выделение в нем повторяющихся фраз и внесение их в словарь фраз.
4. Замена фраз в сжатом тексте кодами фраз.

Считая, что в среднем каждая фраза будет содержать 4-5 слов, можно при таком двухуровневом сжатии получить коэффициент сжатия, доходящий до 40-50.