УДК 621.311

СЛОВАРНЫЙ АЛГОРИТМ СЖАТИЯ ИНФОРМАЦИИ

магистрант Усович В.А.,
Научный руководитель – Куприянов А.Б., к.т.н., доцент

Текстовая информация в узкой предметной области характеризуется словарем небольшого объема. В результате анализа текстов на сайтах посвященных вопросам программирования и информационных технологий получена зависимость объема словаря (количество слов) от объема анализируемого текста (в байтах), показанная на рисунке 1.

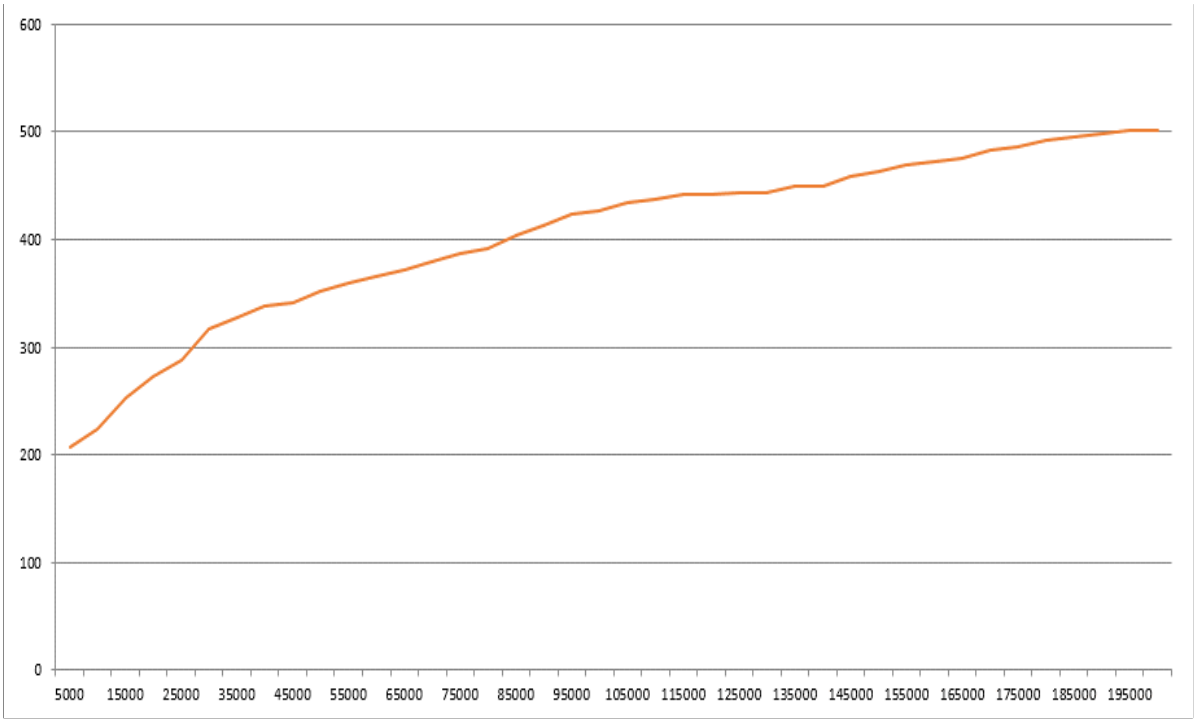

Рисунок 1. Зависимость количества слов в словаре от объема текста.

Из полученной зависимости следует, что объем словаря сайтов узкой предметной области не превышает 2-3 тысяч слов. При этом под словом понимается любой набор символов от начала абзаца до пробела или между двумя пробелами. Следовательно, при кодировании слова его номером в словаре понадобится двоичный код размером $N=ceil(\log_2 V)\approx 12$ бит, где V – объем словаря. Средняя длина слова в исследуемой области составила 8 символов. Значит, средний коэффициент сжатия текста может составлять 8*8/12 = 5,3 при ASCII-кодировке текста и 10,6 при Unicode-кодировке. Сжатие статей на сайтах известными архиваторами получены коэффициенты сжатия от 1,5 до 2.

Хранение статей сайта в базе данных в сжатом виде позволит существенно уменьшить объем базы данных и трафик при передаче

данных клиенту. Для декодирования текста на стороне клиента можно размещать словарь в сети и на компьютере клиента. При обнаружении кода, которого нет в словаре клиента, будет происходить обращение к сетевому словарю и обновление словаря клиента. В этом случае клиенту передается только сжатый текст и обновления словаря в случае необходимости.

Литература

1. В. Пекелис «Кибернетическая смесь», М., «Знание», 1991, стр. 323—324; «IEEE Proc.», 1985, Vol.68, No.7
2. Ватолин Д., Ратушняк А., Смирнов М., Юкин В. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. - М.: ДИАЛОГ-МИФИ, 2003. - 384 с.