УДК 621.311

ОЦЕНКА ВОЗМОЖНОСТЕЙ СЛОВАРНОГО СЖАТИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ

Хамраев А.Ш.,
Научный руководитель – Куприянов А.Б. к.т.н., доцент

Текстовая информация, хранимая в базах данных сайтов, в виде отдельных статей имеет, как правило, небольшой объем и относится к узкой предметной области, определяемой тематикой сайта. Небольшой объем отдельных статей сайта не позволяет эффективно сжимать их с помощью известных архиваторов (winzip,winraги др.). С помощью созданной программы формирования словаря сайта были сформированы словари нескольких сайтов – интернет магазинов по продаже различных товаров.

Анализ словарей показал следующие особенности:

1. Объем словаря сайта не превышает 2-3 тысячи слов.
2. Количество слов в отдельной статье составляет от нескольких сотен до нескольких тысяч.
3. Средняя длина слова составляет 8 символов.

Словарный алгоритм сжатия предполагает замену каждого слова в статье его номером в словаре в двоичной системе счисления. ,Количество бит для кодирования слов можно определить по формуле $N_{бит} = \log_2 V_{сл}$, где $V_{сл}$– количество слов в словаре.Считая, что словарь имеет объем не более 3 тысяч слов, получим $N_{бит} \leq 11$. При использовании 11-битового кодирования средний коэффициент словарного сжатия статей исследованных сайтов составил Ксж = 6,5, при этом коэффициент сжатия тех же статей известными архиваторами не превысил значение 1,2.

Основным недостатком словарного сжатия является необходимость хранения словаря. Словарь может передаваться вместе со статьями и в последующем храниться на компьютере клиента. В этом случае при обнаружении в сжатом тексте слов, отсутствующих в словаре клиент может запрашивать на сервере только обновления словаря, что существенно сократит трафик, исключив передачу словаря в каждом сеансе. сети.