УДК 621.391

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ СТАТИСТИЧЕСКИХ И СТРУКТУРНЫХ МЕТОДОВ СЖАТИЯ

Ю.А. Зубань, м.н.с., В.Н. Гапич, инж., А.А. Маринченко, студ., СумГУ

На сегодняшний день для сжатия данных широкое распространение получили методы сжатия, устраняющие статистическую избыточность сообщений, вызванную различной вероятностью их появления. К таким методам относят методы Хаффмана, Шеннона-Фано и арифметическое кодирование.

Также для многих видов данных оказываются удобными структурные методы сжатия, такие как нумерационное кодирование и метод локальных сдвигов. Так для двоичных последовательностей длины n и содержащих k единиц можно получить множество всех возможных комбинаций и присвоить каждой номер. По номеру можно однозначно определить двоичную комбинацию. Сжатие происходит благодаря тому, что для представления номера необходимо меньше разрядов, чем для исходной комбинации. Получение номера — сложная задача, требующая больших аппаратно-программных затрат, особенно для последовательностей большой длины. Однако существует возможность получения номера комбинации существенно меньшими затратами, например, с помощью метода локальных сдвигов.

Вероятности символов сообщения можно определить исходя из свойств комбинаторных комбинаций. Сжимаемые последовательности являются, как правило,

бинарными, и вероятности «1» и «0» символов определяются из выражений:

$$p(1)=\frac{k}{n},\ p(0)=1-\frac{k}{n}$$

Энтропия такого источника вычисляется по формуле Шеннона

$$H=-\frac{k}{n}\log_2\frac{k}{n}-\left(1-\frac{k}{n}\right)\log_2\left(1-\frac{k}{n}\right)$$

Дальнейшее сжатие возможно, арифметическим кодером. Для алфавитов с мощностью более 2 возможно эффективное применение методов Хаффмана или Шеннона-Фано.

Выражение для энтропии определяет количество информации в одном символе кодируемой последовательности.

Исходя из вероятностного подхода к описанию комбинаторных последовательностей, количество информации в каждой из них определяется как произведение количества символов на энтропию источника:

$$I=n\cdot H$$

С точки зрения структурного подхода к вычислению количества информации, информационная емкость каждой комбинаторной комбинации определяется как C_n^k.

$$I\geq C_n^k$$

Вышеприведенные выражения равны только при $k=0$. Это свидетельствует о том, что описание сообщений статистическими методами избыточно, а значит сжатие последовательностей структурными методами дает лучший результат.