А

Адаптивное сжатие - однопроходный метод сжатия, изменяющий свою работу и/или свои параметры в зависимости от данных, поступающих на вход.

Алфавит - множество всех символов входного потока. При сжатии англоязычных текстов обычно используют множество из 128 ASCII кодов. При сжатии изображений множество значений пиксела может содержать 2,16,256 или другое количество элементов.

Арифметическое кодирование (сжатие) - *статистический метод сжатия*, присваивающий один (обычно длинный) код всему входному потоку, в отличие от методов, присваивающих коды отдельным символам. Метод читает входной поток символ за символом и дополняет код несколькими битами каждый раз когда обрабатывается очередной символ.

Архив - множество из одного или более файлов, собранных в один файл. Отдельные члены архива могут быть сжаты. Архив является удобным способом передачи или хранения группы родственных файлов.

Архиватор - программа для создания и ведения *архивов*.

Б

Биграмма - пара последовательно стоящих *символов*.

Бит/букву - бит на букву. Мера степени сжатия текстовой информации. Также используется в качестве меры *энтропии*.

Бит/символ - бит на символ. Основная мера степени сжатия.

Блочное кодирование - общий термин для методов сжатия, разбивающих входной поток на блоки и кодирующих каждый блок отдельно.

Д

Декодер - программа (или алгоритм) расжатия (восстановления) данных.

К

Код - это *символ*, ставящийся в соответствие другому *символу*.

Коды переменной длинны (неравномерные коды) - коды, у которых длина кодовых слов различна. Используются *статистическими методами сжатия*. Такие коды должны

обладать *свойством префикса* и должны назначаться символам входного потока согласно их вероятностям появления.

Кодер - программа (или алгоритм) сжатия данных.

Контекст - N символов предшествующих данному символу. Контекстные *модели* используют контекст для присваивания вероятностей символам.

Коэффициент сжатия - величина, обратная *отношению сжатия*. Определяется как

$$\text{Коэффициент сжатия} = \frac{\text{размер входного потока}}{\text{размер выходного потока}}$$

Значения >1 обозначают сжатие, а значения <1 - расширение.

Л

Локально адаптивное сжатие - метод сжатия, настраивающийся по локальным свойствам входного потока и изменяющий свои настройки при переходе от одной области входного потока к другой.

М

Модель - метод "предсказания" (присваивания вероятностей символам) данных, поступающих на вход алгоритму сжатия. При использовании статистических алгоритмов, перед тем как начнется сжатие, строится модель данных. Простейшая модель может быть построена при предварительном просмотре входного потока, подсчете числа встречаемых символов и вычисления вероятности появления каждого символа. После этого входной поток просматривается еще раз и сжатие осуществляется основываясь на информации вероятностной модели. Одним из свойств *арифметического кодирования* является легкость разделения модели (таблиц с частотами и вероятностями) и собственно операций кодирования/декодирования. Это позволяет легко кодировать, например, первую половину входного потока при помощи одной модели, а вторую половину - при помощи другой.

О

Отношение сжатия - одна из наиболее часто используемых величин для обозначения эффективности метода сжатия.

$$\text{Отношение сжатия} = \frac{\text{размер выходного потока}}{\text{размер входного потока}}$$

Значение 0.6 означает, что данные занимают 60% от первоначального объема. Значения >1 означают, что выходной поток больше входного (отрицательное сжатие, или расширение). Иногда величина `100 x (1 - Отношение сжатия)` используется для обозначения качества сжатия. Значение 60 означает, что входной поток занимает 40% первоначального размера (или в результате сжатия удалось выиграть 60% объема данных).

П

Полуадаптивное сжатие - метод сжатия, использующий два прохода: во время первого прохода по входным данным собирается статистика, а на втором проходе собственно и осуществляется сжатие.

Предсказание - присваивание вероятностей *символам*, на основе каких-либо правил или свойств сжимаемых данных.

Префикса свойство - код обладает свойством префикса, если никакое кодовое слово не является префиксом другого кодового слова.

С

Символ - наименьшая единица данных, подлежащая сжатию. Обычно, символ - байт, но может быть битом, тритом {0,1,2}, или чем либо еще.

Словарное сжатие - методы сжатия, хранящие фрагменты данных в "словаре" (некоторая структура данных). Если строка новых данных, поступающих на вход, идентична какому-либо фрагменту уже находящемуся в словаре, в выходной поток помещается указатель на этот фрагмент.

Статистические методы - методы сжатия, присваивающие *коды переменной длины* символам входного потока, причем более короткие коды присваиваются символам или группам символам, имеющим большую вероятность появления во входном потоке.

Стопки книг, метод сжатия - основная идея метода заключается в хранении *алфавита* в виде списка, в котором чаще встречаемые символы находятся ближе к голове. *Символ* кодируется числом символов, предшествующих ему в списке. После кодирования *символ* перемещается в начало списка. (Англ. название Move-to-Front Coding.)

Т

Токен - единица данных, записываемая в сжатый поток некторым алгоритмом сжатия. Токен состоит из нескольких полей фиксированной или переменной длины.

Ф

Фраза - фрагмент данных, помещаемый в словарь для дальнейшего использования в сжатии. Эффективность конкретного словарного метода в некоторой степени зависит от того, как этот метод строит фразы, помещаемые в словарь.

Х

Хафмановское кодирование (сжатие) - широко используемый метод сжатия, присвивающий символам *алфавита коды переменной длины* основываясь на вероятностях пояления этих *символов*. Метод кодирования Хафмана схож с *методом Шеннона-Фано*. В общем случае, метод Хафмана строит более оптимальные коды и, так же как и метод Шеннона-Фано, строит оптимальный код в случае, когда вероятности появления символов алфавита являются отрицательными степенями 2. Основное отличие между этими методами состоит в том, что метод *Шеннона-Фано* строит код сверху-вниз (от самого левого бита к самому правому), тогда как метод Хафмана - снизу-вверх (код строится справа налево).

Ш

Шеннона-Фано кодирование (сжатие) - один из ранних методов сжатия, заключающийся в постреении *кода переменной длинны* по заданным вероятностям появления *символов* входного потока. Позже был превзойден *методом Хафмана*.

Э

Энтропия - энтропия *символа* a_i определяется как $-P_i \log_2 P_i$, где P_i - вероятность появления *символа* a_i. Энтропия a_i - наименьшее число бит необходмых, в среднем, для представления *символа* a_i.

В

BWT метод - Burrows-Wheeler Transform, обратимый алгоритм перестановки символов блока входных данных, позволяющий более эффективно сжимать преобразованный блок данных.

С

CM (Context Modeling) - Контекстное моделирование.

D

DC (Distance Coding) - кодирование расстояний.

DCT (Discrete Cosine Transform) - дискетное косинусное преобразование.

DMC (Dynamic Markov Compression) - динамическое марковское сжатие.

DWT (Discrete Wavelet Transform) - дискетное вэйвлетное преобразование.

I

IFS (Iterated Function Systems - Системы итерируемых функций.

L

LPC (Linear Prediction Coding) - линейное предсказывающее кодирование.

LZ Методы - все *словарные методы сжатия*, основанные на работах J. Ziv и A. Lempel, опубликованных в 1977 и 1978 годах. Сегодня эти методы обозначаются LZ77 и LZ78 соответсвенно.

LZ77 - чередуются указатели и символы. Указатели ссылаются на подстроку, расположенную в предыдущих N символах. Авторы: Ziv, Lempel. 1977.

LZ78 - на выходе чередуются указатели и символы, указатели ссылаются на *фразу* из словаря, предварительно выделенную интерактивным анализатором. Авторы: Ziv, Lempel. 1978.

LZB - аналогичен методу *LZSS* кроме различных способов кодирования указателей. Автор: Bell. 1987.

LZC - алгоритм, реализованный программой `compress` в системах UNIX. Модифицированный *LZW*. Автор: Thomas и др. 1985.

LZFG - указатели выбирают узел в дереве поиска. Строки в дереве выбираются из текущего окна. Авторы: Fiala, Green. 1989.

LZH - аналогичен алгоритму *LZSS*, на втором этапе кодирование указателей осуществляется по *методу Хафмана*. Автор: Brent. 1987.

LZJ - выход содержит только указатели; указатели означают подстроку в любом месте предыдущего текста. Автор: Jakobsson. 1985.

LZMW - аналогичен методу *LZT*, но фразы строятся путем конкатенации предыдущих фраз. Авторы: Miller, Wegman. 1984.

LZR - на выходе чередуются указатели и символы. Указатели означают строку, расположенную в любом месте предыдущего текста. Авторы: Rodeh и др. 1981.

LZSS - указатели и символы разделяются битом флага. Указатели ссылаются на подстроки в предыдущих N символах. Автор: Bell. 1986.

LZT - аналогичен *LZC*, но с фразами в LRU-списке. Автор: Tischer. 1987.

LZW - выход содержит только указатели, ссылающиеся на предварительно выделенные подстроки и имеющие фиксированный объем. Автор: Welch. 1984.

M

MNP5, MNP7 - методы сжатия, разработаные фирмой Microcom, Inc. для использования в производимых ею модемах. MNP5 - двустадийный процесс: на первом этапе используется RLE, на втором - адаптичное частотное кодирование. MNP7 сочетает RLE с двумерным вариантом адаптивного *Хафмановского кодирования*.

MTF (Move To front) - английское название метода *Стопки книг*.

P

PBS (Parallel Blocks Sorting) - сортировка параллельных блоков.

PPM (Prediction by Partial Match) - метод сжатия, присваивающий вероятности символам на основе *контекста* (длинного или короткого), в котором они встречаются.

R

RC (Range Coding) - интервальное кодирование (вариант *арифметического кодирования*).

RLE - общее название для методов сжатия, заменяющих последовательность одинаковых символов *токеном*, содержащим *символ* и количество его повторений.

S

SC (Subband Coding) - субполосное кодирование.

SEM (Separate Exponents and Mantissas) - разделение экспонент и мантисс (представление целых чисел).

ST (Sort Transformation) - частичное сортирующее преобразование.

V

V.42bis, Протокол - стандарт, разработанный ITU-T для использования в скоростных модемах. Основывается на протоколе V.32bis и предложен для использования на высоких скоростях передачи, до 57.6К бод. V.42bis определяет два режима работы: прозрачный, когда сжатие не используется, и режим сжатия, использующий вариант *LZW*. Первый используется для потоков, которые не сжимаются. Примером таких потоков является уже сжатый файл.

VQ (Vector Quantization) - векторное квантование.