Реплика. Компрессия данных

Компрессия - это (очень) частный случай трансляции. И то, что на жаргоне компрессорщиков называется "modelling", на самом деле есть парсинг и конструирование грамматики. Будет ли набор правил записан в нотации Хомского или в виде цепей Маркова, по большому счету, значения не имеет - это вопрос традиций вашей научной школы.

Но, на мой взгляд, использование псевдохарактеристик, таких как "вероятность" и "энтропия" затемняет суть дела - необходимо оперировать наблюдаемыми величинами. Для сообщения таковыми являются набор (система) кодов, их кратности (а не вероятности), и их битовая ширина.

В заданной системе кодов сообщению соответствует некоторое пространство состояний (обычно, меньшее, чем полное кодовое пространство), которое, в теории, должно бы быть инвариантом любого обратимого преобразования.

Однако, часто не все состояния являются независимыми, так что сообщение может быть вложено в пространство меньшей размерности (ортогонализация).

С другой стороны, кодер, неизбежно, добавляет к сообщению собственную информацию для декодера, так что пространство состояний сжатого сообщения (его избыточность) может оказаться больше исходного.

Над сообщением возможны две недеструктивные операции: перестановка кодов (символов) и изменение их битовой ширины.

Соответственно, все техники обратимого сжатия могут использовать только эти две операции, которые будем называть Обобщенный поворот (пермутации) и Обобщенный сдвиг (манипуляции битовой шириной).

Понятно, что пермутации не изменяют размера сообшения.

Манипуляция битовой шириной позволяет использовать две дуальные техники, которые будем называть Метод Хаффмана (H): замена системы моноширинных (равномерных) кодов на систему пропорциональных (неравномерных) кодов (арифметическое кодирование - это вариант метода Хаффмана) и (зеркальный к нему) Метод Зива-Лемпеля (LZ) - замена неравномерных кодов на равномерные.

В первом случае решается задача построения кодов с минимальной средней длиной, во втором - обратная к ней - построения словаря с максимальной средней длиной слов.

На практике, оба сводятся к построению дерева кодов, обладающего данным экстремальным свойством.

Хаффман (1952) предложил процедуру создания дерева, по которому конструируются префиксные коды, зеркально, Зив и Лемпель (1978), создают суффиксное дерево, включающее все префиксы каждого внесенного в словарь слова.

В статическом варианте оба метода двухфазны - парсинг и кодирование. Парсинг равномерных кодов тривиален и метод Хаффмана просто трансформирует один префиксный код в другой префиксный код. Построение суффиксного дерева и поиск в словаре более трудоемки, но позволяют оперировать со словами произвольно большой длины, что затруднительно для произвольно длинного равномерного кода.

Фактически, оба метода не обязательно используются в "чистом" виде. Кодирование Хаффмана может быть применено к неравномерным входным кодам, а для суффиксного дерева может быть сгенерирован неравномерный выходной код.

Все существующие техники сжатия являются либо оберткой над одной из этих двух техник, либо оберткой над их комбинацией.

Окончательно, все техники, манипулирующие битовой шириной (H, LZ, RLE...) будем относить к классу Обобщенных сдвигов (S), и все техники пермутаций (FT, MTF, BWT, ST...) к классу Обобщенных поворотов (R).

Используя геометрическое представление, легко увидеть, что собственно "сжатие" заключается в минимизации эксцентриситета эллипсоида пространства состояний (приведении его к "почти" сфере) путем:

1. Ортогонализации (устранение "нулевых" осей) и

2. Масштабирования по осям (гиперболический поворот).