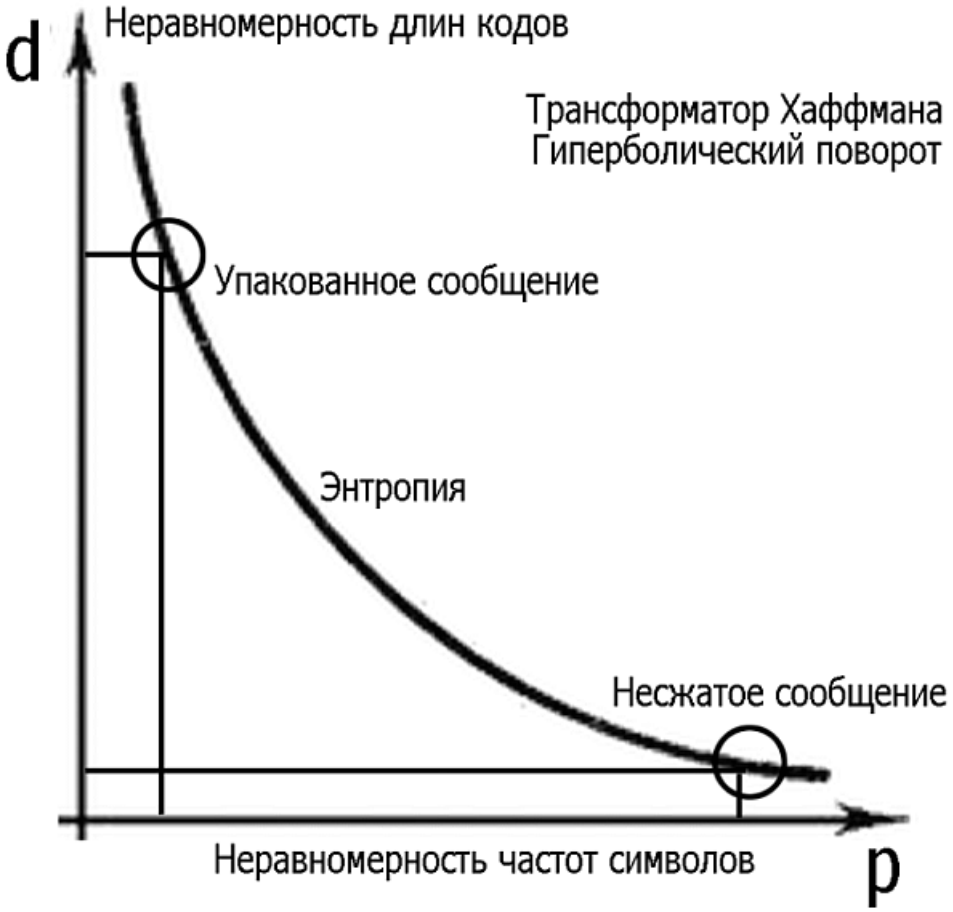
Размышления. Трансформатор Хаффмана



Трансформатор Хаффмана все еще остается важным классическим алгоритмом в техниках сжатия данных без потерь и достаточно поучителен для того, чтобы остановиться на нем подробнее.

Все методы энтропийного неравномерного кодирования используют асимметрию кратностей символов текста. В самом деле, если все символы текста встречаются одинаково часто, не имеет никакого значения, какой длины код назначен каждому из них - результирующий размер не изменится. Иными словами, полностью симметрированный текст (часто ошибочно трактуемый как "случайный") несжимаем.

Как и в методах Шеннона-Фано, ключевая идея заключается в построении префиксного дерева кодов (ни один код не является началом другого), но в отличие от дерева Шеннона-Фано, которое строится сверху-вниз, от корня, дерево Хаффмана строится снизу-вверх от самого низкочастотного символа, который получает в результате самый длинный код. Соответственно, самый короткий код назначается самому высокочастотному символу.

Таким образом, если дерево равномерных кодов является сбалансированным (деревом минимальной высоты), то дерево неравномерных кодов (дерево

Хаффмана), напротив, асимметрично и, в предельном случае, является деревом максимальной высоты. Процесс кодирования (ремэппинга равномерных кодов в неравномерные), традиционно называемый "сжатием", фактически, является инверсией.

Для заданной системы неравномерных кодов и выбранного асимметричного сообщения возможны ровно два экстремальных (двойственных) состояния: сообщение минимальной длины (энтропийный предел) и сообщение максимальной длины.

Введем понятие Идеального текста. Идеальным текстом будем называть алфавитно упорядоченную цепочку, содержащую все символы данного алфавита. Размер такой цепочки в точности равен мощности алфавита. Заметим, что идеальный текст и любая его пермутация принципиально несжимаемы энтропийным кодером.

Для неидеального текста того же самого размера (совпадающего с мощностью алфавита), будем оценивать асимметрию (отклонение от идеала) дисперсией кратностей - суммой квадратов уклонений кратностей символов от единицы. Очевидно, для идеального текста дисперсия равна нулю и эта оценка растет с ростом асимметрии текста.

Аналогично, для системы неравномерных кодов, будем оценивать асимметрию (отклонение от равномерности) дисперсией битовых размеров кода - суммой квадратов уклонений размеров кодов от битового размера равномощного равномерного кода.

Для исходного случая асимметричного сообщения, записанного равномерным кодом дисперсия кратностей отлична от нуля и дисперсия размеров тождественно равна нулю. Результатом работы трансформатор Хаффмана является инверсия: в идеале, дисперсия кратностей стремится к нулю, дисперсия размеров отлична от нуля.

По аналогии с качаниями маятника, изменение битового размера кода будем называть качанием системы кодов. Крайним положениям маятника соответствуют сбалансированное дерево кодов (равномерный код) и дерево максимальной высоты (неравномерный код).

Поскольку дисперсия (квадратичная форма) имеет физический смысл Энергии, эксплуатируя аналогию с маятником, припишем дисперсии кратностей смысл Кинетической энергии сообщения и дисперсии размеров - смысл Потенциальной энергии сообщения.

Разумеется, термин Энергия сообщения в данном контексте достаточно условен и не имеет отношения к реальной энергии физического процесса. С другой стороны, он удобен, так как отражает взаимный переход дисперсий. По аналогии с взаимным переходом энергии при качании маятника, можно уравнять предельные значения кинетической и потенциальной энергии сообщения, получив коэффициент, зависящий только от конструктивных особенностей выбранной системы кодов.

С другой стороны, для заданной системы кодов несложно прямо вычислить число и размер кодов дерева максимальной высоты и высоту сбалансированного дерева, что позволяет найти предельный коэффициент сжатия (динамический диапазон).

Переход от сбалансированного дерева кодов к дереву максимальной высоты и обратно соответствует предельному циклу компрессии-декомпрессии.

Рассматривая текст как аналог идеального газа, можно увидеть аналогию предельного цикла с циклом Карно. Информационная энтропия, очевидно, является аналогом термодинамической, а длина сообщения - аналогом абсолютной температуры.