Трансформатор Хаффмана. Мысью по древу

Однажды Ходжу Насреддина спросили: - Как строятся самые высокие минареты - Очень просто, - ответил тот. - Роют глубокий колодец, а потом выворачивают его наизнанку.

Алгоритм Хаффмана в точности следует этой идеологии: вначале строится дерево максимальной высоты, а затем выворачивается наизнанку - инвертируется в дерево минимальной высоты. Можно считать историческим курьезом, что на протяжении семидесяти лет алгоритм инверсии взвешенного дерева излагается исключительно как метод компресии данных.

Если надавить сверху на пластилиновый кубик, то он станет меньше в высоту и больше - в ширину, сохранив при этом общее количество пластилина.

Параметры системы, которые не изменяются при ее деформации, называются инвариантами.

Если записать фразу и проиграть ее на повышенной скорости, тон звука повысится, а время исполнения уменьшится. Этим часто пользуются при озвучке мультфильмов, создавая "писклявые" голоса медийных персонажей. Наоборот, если проиграть эту же фразу при пониженной скорости, тон ее станет ниже, а время исполнения больше.

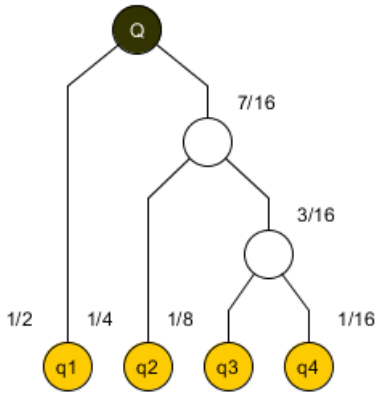
В 1928 году Ральф Хартли ввел понятие "энергетического инварианта" при передаче аналоговых сообщений (инвариант Хартли). Вот как он писал об этом:

"Это приводит нас к важному выводу, что максимальная скорость передачи информации, возможная в системе, частотный диапазон которой ограничен некоторой областью, пропорциональна ширине этой полосы частот. Отсюда и следует, что общее количество информации, которое может быть передано посредством такой системы, пропорционально произведению передаваемой полосы частот на время, в течение которого система используется для передачи. Произведение передаваемой полосы частот на время и есть упомянутый мною в начале статьи количественный критерий сравнения передающих систем"

Для коммерческих текстов часто характерна частотная неравномерность: некоторые символы встречаются чаще других. При кодировании текста, метод Хаффмана реализует простую идею, первоначально заложенную в кодах Морзе: частым символам ставятся в соответствие короткие коды, а редким - длинные. Это позволяет сократить время передачи и сэкономить электроэнергию в электрическом телеграфе.

Для построения дерева максимальной высоты список узлов сортируется по весу. Два узла минимального веса изымаются из списка и заменяются узлом суммарного веса, возвращаемого в сортированный список. Эти два узла (в нашем случае, **1/8** и **1/16**) образуют затравку дерева. К ним, в качестве родительского, присоединяется узел

суммарного веса (**1/8 + 1/16 = 3/16**). Далее, процедура повторяется с обновленным списком узлов до его исчерпания и дерево надстраивается аналогичным образом.



В результате, узлу наименьшего веса соответствует самый длинный путь от корня дерева, и узлу наибольшего веса отвечает самый короткий (единичный) путь от корня дерева.

Исторически, алгоритм (трансформатор) Хаффмана - один из первых алгоритмов сжатия данных. Он был предложен в 1952 году Давидом Хаффманом (в то время, аспирантом Роберта Фано). Как оптимизационный алгоритм на дереве, метод Хаффана может быть применен где угодно в задачах, отображаемых на деревья, но традиционно рассматривается на примерах сжатия текстов.

Рассмотрим распределительную систему из четырех потоков **q[i]**, отображаемых на дерево **T**. Входной поток **Q** равен сумме всех потоков терминальных узлов. Пусть распределение потоков организовано так, что полный поток делится в геометрической прогрессии, так что каждый правый поток вдвое меньше соседнего с ним левого потока. Примем полный поток **Q** равным **1**, тогда **q[1] = 1/2**, **q[2] = 1/4**, **q[3] = 1/8** и **q[4] = 1/16**. Геометрическое распределение потоков гарантирует уравновешенность в узлах и максимальную высоту дерева.

Природа потока **Q** не имеет значения, важно только соблюдение принципа неразрывности потока (первого закона Кирхгофа): алгебраическая сумма всех входных и выходных потоков равна нулю.

Припишем каждой их ветвей единичный вес, тогда вес пути до терминального узла **w[1] = 1**, **w[2] = 2**, **w[3] = 3** и **w[4] = 3**.

Введем понятия Момента ветви, как произведение веса пути на вес инцидентного ему терминального узла **M[i] = w[i] * q[i]** и Периметра дерева, как суммы всех моментов ветвей: **L = sum(M[i]) = sum(w[i] * q[i])**.

В случае, когда на дерево отображен текст с кратностями символов **q[i]** и битовыми размерами кодов **w[i]**, периметр дерева имеет смысл битовой длины сообщения.

Структура дерева определяется только распределением потоков и не зависит от их абсолютной величины. В задачах обработки текста это позволяет вместо абсолютных значений кратностей символов использовать их частоту в тексте (в терминологии Шеннона - вероятность).

В рассматриваемом примере, инверсия дерева максимальной высоты в дерево минимальной высоты дает максимальный коэффициент трансформации, то есть построенное дерево Хаффмана обладает минимальным периметром. Иными словами, для текста, например, невозможно получить лучшее сжатие (в целочисленных кодах).

Если выписать спектры битовых размеров кодов **{ 1, 2. 3, 3 }** и величин потоков **{ q/2, q/4, q/8, q/16 }**, то очевидно, что данное дерево реализует логарифмическое отображение величин потоков на размеры кодов. Иными словами **w[i] <- -Log(q[i])**. (Так как при геометрическом распределении потоков **q[i] = Q * 1/2^i**, то, если использован двоичный логарифм, **Log(q[i]) <- -(Log[1] - i*Log[2]) = i**, **i = 1..N** - число потоков).

Поскольку для дерева максимальной высоты размеры кодов пропорциональны логарифмам величин потоков, несложно выписать формулу величины периметра

L = -sum (q[i] * Log(q[i])).

Сравнивая эту формулу с энтропией Шеннона, мы видим, что так называемая "информационная энтропия" является относительным (на единицу потока) периметром дерева, а энтропийный предел - относительным минимальным периметром (периметром дерева максимальной высоты).

Введение понятия "энтропия" не требует обращения к случайности (или использования таких заклинаний как "информация" и "неопределенность"), а связано только с комбинаторным ростом числа состояний распределительных систем (в частности, текстов). При этом энтропия, как логарифмическая оценка, является естественной характеристикой процессов с экспоненциальным ростом.

Энтропия вычисляется не как абсолютная оценка (от величин потоков), а как относительная (от соотношения вероятностей), что дает нормированную (на символ текста) оценку.

Информационная энтропия **H** была введена Шенноном как функция состояния системы (третья аксиома). Это означает, что полученная оценка не зависит от внутренней структуры распределительной системы, а только от соотношения величин потоков в терминальных узлах. По историческим причинам, будем называть эту оценку "количеством информации в сообщении" **S** (хотя никакого отношения к "информации" она не имеет), **S = H * N**, где **N** - число символов сообщения, а **H** - информационная энтропия на один символ.

При "сжатии" сообщения (фактически, перекодировании) распределение потоков не изменяется и, следовательно, не изменяется вычисляемая от них функция. Однако, при этом изменяются битовые размеры кодов, размер сообщения и, тем самым, его энтропия.

Сравнивая выражение **S = H * N** с "энергетическим инвариантом" Хартли, нетрудно заметить, что при сжатии текста "количество информации" является инвариантом трансформации, размер сообщения является аналогом длительности сигнала, а энтропия - аналогом ширины спектра. При этом сама трансформация Хаффмана может быть описана как гиперболический поворот ("лоренц-сжатие") в координатах "размер сообщения - энтропия".