Общая дискуссия о предельной степени сжатия

Автор: Andrew Polar [Пт май 04, 2012 22:45]

Заголовок сообщения: **Общая дискуссия о предельной степени сжатия**

Привет всем, я неожиданно обнаружил дискуссию на *encode.ru* под названием *Modelling... what exactly?* Речь идет о степени сжатия, она начата *Shelwien*, скорее всего это *Шелвин*. Я несколько раз пытался там открыть акаунт но система не пропускает меня по неизвестной причине.

Вопрос достаточно интересный и доказывает что Шеннон придумал энтропию, глядя в потолок и почесав репу. Это видно из его фундаментальной статьи, но однако он попал в яблочко, с одной оговоркой, его энтропия приближенная оценка и довольно близка к точной. Энтропия - это просто лог. бинарной длины числа перестановок деленный на количество символов. Например **100** бит, **20** из которых нули. Количество перестановок **100!/20!/80!**, логарифм от количества перестановок **68.8,** делим на количество символов и получаем **0.69** бит на символ.

Это точно, а Шенон заменил факториал на степень (от балды или непонимания) и получил **100^100/20^20/80^80**. Если взять логарифм и разделить на количество символов то получим **0.72**. Числа близкие по причине аппроксимации Стирлинга лог(м!) равно приближенно лог(м^м)-м, но Шеннон об этом не знал, иначе бы он написал об этом. Разница уменьшается с возрастанием длины. Для тысячи символов она уже доли процента. А мы имеем дело с миллионами символов, так что это неважно. Я об этом писал у себя на сайте *http://ezcodesample.com/reanatomy.html,* но похоже никто не читает. Кроме того никто никак не соглашается, что Шеннон чего-то недопонимал, когда писал свою статью, что довольно странно, он что классик марксизма, который никогда не ошибался. Я даже дал новое определение этропии как характеристика сообщения численно равная (при достаточно большой длине) логарифму числа перестановок выраженному в битах деленному на длину сообщения. Это четко и ясно а не то, нечто путанное и туманное в статье Шеннона. Когда мы делаем контекстное или адаптивное сжатие и достигаем лучшего сжатия, мы не противоречим определению, оно остается в силе просто применяется к контекстным подпоследовательностям.

Автор: Andrew Polar [Пт май 04, 2012 23:21]

Заголовок сообщения: Re: **Общая дискуссия о предельной степени сжатия**

Я извиняюсь за дополнение, но я придерживаюсь мнения, что на момент написания своей статьи Шеннон кое-какие элементарные вещи не понимал. Четкое объяснение энтропии должно звучать примерно так:

При заданном распределении символов в сообщении, вычисляем число перестановок. Это число выраженно десятичными знаками, считаем сколько бит

потребуется, если выразить это число в двоичной системе, для чего берем логариф при основании два и делим результат на количество символов. Но считать таким образом неудобно, поэтому заменяем факториалы на степени, используя формулу Стирлинга, делаем пару алгебраических преобразований, и получаем энтропию приблизительно выраженную через вероятности символов.

Вот собственно и все, а теперь перечитайте статью и увидите, что объяснение Шеннона в несколько раз длинее и невнятно.

Автор: gazlan [Сб май 05, 2012 23:15]

Заголовок сообщения: Re: **Общая дискуссия о предельной степени сжатия**

Цитата:

"*на момент написания своей статьи Шеннон кое-какие элементарные вещи не понимал*"

Андрей, мне кажется, вы излишне очарованы комбинаторной моделью энтропии (взятой у Стратоновича?).

Разумеется, для заданного пространства вы можете задать любую удобную метрику и, разумеется, метрика Шеннона крайне неудобна во многих отношениях (например, для энтропии порядка 0, расстояние между любыми анаграммами будет нулевым), но если вы взглянете в рисунки оригинальной статьи, то увидите, что Шеннон, фактически, использовал "*рычаг Архимеда*" при "уравнивании" двух деревьев. Никакого обоснования этой нетривиальной операции он не дал, но интуитивно, она очевидна.

Дискретное сообщение обладает некоторым спектром (эллипсоидом) состояний (в выбранной системе кодов), два из которых являются экстремальными. В смысле компрессии, это оэначает предельное сжатие и предельное растяжение (короткая и длинная оси эллипсоида), все другие состояния являются промежуточными (возможно и вырождение - кратные собственные числа).

"*Сжатие*" означает некоторый обобщенный сдвиг и поворот в этом пространстве состояний (приведение к собственным осям), причем, на практике интересуются только короткой осью (предельное сжатие).

Все это удобно описывать в терминах преобразования матриц (сдвиг и поворот), комбинаторика и формула Стирлинга здесь ни при чем.

Автор: gazlan [Ср май 09, 2012 19:39]

Заголовок сообщения: Re: **Общая дискуссия о предельной степени сжатия**

Andrew Polar писал(а):

"*Более сложные объяснения найти можно, а вот приведите более простое.*"

Не знаю, покажется ли вам мое объяснение простым, но я постараюсь сделать его развернутым и не прибегать к операциям с матрицами :-)

Коротко:

Прежде всего, я вообще не понимаю этого сакрального интереса к энтропии, на мой взгляд, это довольно малозначимая характеристика сообщения. Во-вторых, просто не вижу предмета для спора - выберете вы тот или иной вариант записи формулы, оба они одинаково плохи.

Подробно:

Сообщение - это многомерный объект, а энтропия - попытка оценки (описания) каких-то важных для вас характеристик этого объекта одним числом (скаляром) - со всеми недостатками такого подхода.

Какие характеристики считать важными и какую функцию от них сконструировать - вопрос вкуса и текущих потребностей. Шеннону требовалась монотонная, аддитивная и зависящая от длины сообщения характеристика. Единственной такой функцией может быть логарифм (коэффициенты, свободный член, соль и сахар добавить по вкусу).

Будет этот логарифм введен прямолинейно, как это сделал Шеннон или с экивоками, через формулу Стирлинга, значения не имеет.

Вопрос выбора той или иной логарифмической меры - это свифтовский спор между остро- и тупоконечниками.

Для практически значимых по размеру сообщений, численные значения шенноновской и комбинаторной энтропии будут близки, а в теоретическом плане обе они одинаково плохи (или одинаково хороши - в зависимости от ваших целей).

Заметьте, я не хочу сказать что Шеннон был в чем-то неправ, ошибался, или чего-то не понимал.

Больше того, я уверен, что Шеннон блестяще решил стоявшую перед ним задачу.

Но, Шеннон сконструировал логарифмическую энтропию для своих (совершенно определенных) целей и использование ее для чего-то другого, скорее всего, окажется безрезультатным.

Классический силлогизм:

Энтропия - это метрика изотропной системы.
Сообщение анизотропно.
Ergo: энтропия не может быть метрикой сообщения.

Простой пример.

В системах связи принято контролировать правильность передачи сигнала с помощью контрольной суммы. В простейшем варианте, это сумма байт сообщения,

скажем, по модулю 256. Такая сумма нечувствительна к порядку букв в сообщении, так что "MAPC" и "CPAM" для нее неразличимы.

Более сложные схемы проверки (например, CRC или ECC) чувствительны к порядку бит, но лишены свойства аддитивности.

Предположим, на вас лежит техническое обеспечение эксперимента по телепатической передаче изображений карт Зенера, и вам нужен метод различения геометрических фигур.

На первый случай, вы решаете оценивать фигуры одним числом - их площадью. До какого-то момента такой подход срабатывает, пока не попадается равновеликая кругу звезда.

Если для ваших целей годится утверждение: "звезда - это круг" или вы готовы поверить, что "MAPC" и "CPAM" - это одно и то же сообщение, то логарифмическая энтропия - ваш метод.

Если же вам требуется отличать "MAPC" от "CPAM", а круг - от звезды - потребуется что-то более изощренное.

MD5, например, замечательно чувствительна к порядку бит, но, разумеется, лишена свойства аддитивности.

К слову, впервые этот подход, вероятно, был использован в *Гематрии*, когда слова с одинаковой числовой суммой букв полагались эквивалентными и в смысловом отношении. Позднейшие нумерологические системы заимствовали тот же самый метод, пока Теория Информации не закрепила его окончательно в виде (логарифмической) суммы, объявленной "*количеством информации*".

На моей домашней страничке можно найти пояснения, почему ни к информации, ни к ее количеству, эта величина отношения не имеет.

Если же вернуться к изначальной трактовке сообщения "по Хартли" - как последовательности выборов (дереву выбора), то делается очевидным, что энтропия по Шеннону - это длина пути по дереву выбора от корня до выбранного терминального узла, а шенноновская избыточность - это "разность хода" между длиной пути по идеальному уравновешенному (минимальной высоты равновероятному дереву) и реальному, сконструированному в данной системе кодов дереву выбора (Шеннон приводит выражение для избыточности в относительной, а не в абсолютной форме).

Если же говорить не о дереве выбора, а об эллипсоиде состояний, то разницу длин пути следует заменить на разницу объемов реального (эксцентричного) эллипсоида и идеальной сферы.

При этом, из выражения для длины пути по дереву выбора, энтропия Шеннона получается автоматически, не требуя ни привлечения комбинаторных формул, ни апелляции к достаточно большой длине сообщения и, IMHO, такой вывод намного проще комбинаторного.

Также очевидно, что такая мера не позволяет различать никакие равноудаленные от

корня терминальные узлы (*анаграммы*).

Автор: gazlan [Сб май 26, 2012 18:02]

Заголовок сообщения: Re: **Общая дискуссия о предельной степени сжатия**

Прежде всего, извиняюсь за задержку с ответом. Я сейчас почти не бываю on-line и может пройти пара недель, прежде чем я увижу ваше сообщение.

Во-вторых, я вполне разделяю ваши взгляды на практичность и полезность. Но если энтропия "по Шеннону" и не принесла пользы вашему компрессору, это еще не делает ее неправильной или неприменимой в иных случаях.

В-третьих, я опять не согласен с вашими замечаниями.

Например, когда вы пишете: "По Шеннону, имеет информацию сообщение из 4 символов, в котором все 4 символа разные", это неверно. Они должны быть статистически независимыми. А будут они одинаковыми или разными - дело случая. (Сходное заблуждение немецких криптографов в период **WWII**, немало помогло Алану Тьюрингу во взломе сообщений, созданных с использованием шифровальной машины "*Enigma*").

Можно, конечно, поставить вопрос о правомерности использования (статистического) понятия "энтропия" для оценки коротких и ультракоротких сообщений. Но, скажем у Яглома ("*Вероятность и информация*"), энтропия успешно применена для решения (ультракоротких) задач о взвешивании монет (определение фальшивой за минимальное число взвешиваний).

К слову, я давно и с интересом просматриваю и ваш сайт и ваши сообщения на форуме и только ваш подход к энтропии (еще в прошлых, трехлетней, кажется, давности сообщениях) - единственное, что вызывает возражения.

Далее, по поводу термина "*количество информации*".

Как показывает история науки, первооткрывателям редко удается ввести удачную терминологию, и она впоследствии заменяется более подходящей.

Некоторые авторы предлагают термин "*информационная емкость*", что вполне согласуется и с природой этой величины и со взглядами Шеннона на информацию.

Если у вас есть, например, пол-литровая коньячная бутыль, в нее можно налить спирт, минеральную воду, крысиный яд... ничего не налить, но налить в нее больше, чем 0.5 литра, гарантированно, не удастся. И если у вас есть литр коньяка, разлить его меньше, чем по двум бутылкам, также не получится.

Шеннону, для решения задачи о транспорте сообщений, нужна была именно характеристика вместимости тары и он ее сконструировал.

Сообщение может содержать информацию, не содержать информации, содержать дезинформацию, но гарантированно не может содержать больше, чем это позволяет

информационная емкость (энтропийный предел).

Теперь, собственно, об определении *энтропии*.

Не поймите меня неправильно...

Когда я писал о том, что это скалярная оценка многомерного объекта, я не хотел сказать, что это непременно плохо или неприемлемо.

Если две вещи не могут быть сравнены непосредственно - наложением, приходится изобретать скалярные оценки - такие как площадь, деньги, энергия итп.

Знтропия Шеннона плоха не именно тем, что это скаляр. Она плоха тем, что в отличие от площади, денег или энергии, неинвариантна к системе отсчета (выбранной системе кодов).

Иными словами, она не является тем, что заявлено - "*количеством*" информации.

Полагаю, было бы правильно не использовать всуе слово "информация" (к которому, как мы убедились, энтропия отношения не имеет) и столь же бессодержательному в данном контексте как, например, "биополе", а ввести конструкционный параметр, аналогичный импедансу Хевисайда в теории электрических цепей (или волновому сопротивлению в теории длинных линий).

Тогда, для согласования канала передачи с сообщением, необходимо, чтобы совпадали (в смысле какой-то подходящей метрики, наименьших квадратов, например) их сопряженные характеристики ("импедансы") - спектр кратностей символов сообщения и спектр битовых размеров кодов. Это выглядит как задача оптимального согласования генератора и нагрузки. (Арифметический кодер [*трансформатор*] в наибольшей степени реализует подобный подход). Для частного случая, когда один из кодов равномерный, а другой нет, получаем два классических метода трансформации - Хаффмана и Зива-Лемпеля. Или, обобщая, все техники *Lossless Data Compression* - это частные случаи *Tearing*-трансформации Крона (разрывное комбинаторное преобразование, сохраняющее площадь фигуры - как в задачах раскроя).

Замечу, что (предельный) коэффициент трансформации определяется только и исключительно конструкцией трансформатора - выбранными системами кодов (эксцентриситетами эллипсоидов кодовых пространств состояний), без всякой нужды в лукавых терминах типа "*энтропия*" и "*избыточность*".

Точнее говоря, в так называемом "моделировании" в техниках сжатия смешаны две различные фазы трансляции - парсинг и конструирование грамматики. Иными словами, это - ортогонализация. Даже идеальный трансформатор ничем не поможет вашему компрессору при неудачной факторизации. А, поскольку, парсинг - NP-проблема, остается простор для эвристики.

Спектральный подход к транспорту сообщений приводит к чисто инженерной задаче конструирования "*рычага Архимеда*" - трансформатора импеданса (согласования передаточных характеристик на границе раздела двух сред), без необходимости привлечения спорного и плохо определенного понятия "информация", использования

вероятностных характеристик, каких-либо логарифмических оценок, устраняет все проблемы, связанные с семантикой сообщений и мог бы поставить точку в нашей дискуссии :-)

В любом случае, на сегодняшний день, энтропия Шеннона - это единственная аддитивная оценка, и в Теории Информации она выполняет ту же роль, что и равенство Парсеваля в теории рядов Фурье.

При этом, энтропия по Шеннону имеет два врожденных порока: базируется на понятии случайности и на понятии непрерывной величины.

Первое приводит к неприменимости понятий и выводов теории Информации ко многим содержательным задачам, второе - к сингулярностям (таким, как "Проблема нулевой частоты" в кодировании Хаффмана).

Если взглянуть на то общее, что есть в альтернативных подходах к определению информации (Эшби, Бриллюэн, Колмогоров), то это "три кита" - A.D.D. - аддитивность, отказ от случайности (детерминированность), отказ от непрерывности (дискретность).

И если два последних пункта: детерминированность и дискретность выглядят прямо напрашивающимися, то с аддитивностью все много хуже.

Если выстроить условную аддитивную шкалу и разместить на ней различные скалярные оценки текста, то самой левой точкой этой шкалы будет энтропия Шеннона (абсолютная аддитивность), где-то посредине - метрики Хэмминга и Левенштейна, а крайней правой - различные криптографические хэш-функции (абсолютная неаддитивность).

"Комбинаторная" энтропия (такая как метрика Бриллюэна или Стратоновича) окажется очень близко к левому краю, но все же, не крайней слева.

Предложение заменить энтропию Шеннона на комбинаторную выглядит как предложение заменить гелиоцентрическую систему мира Коперника (с "гладкими" траекториями) на геоцентрическую систему Птоломея ("корявая" аппроксимация первыми членами ряда Фурье по эпициклам и деферентам).

С практической точки зрения, расхождения в численных расчетах ничтожны, также как неважно для практики, сходятся ли где-то далеко за горизонтом - на бесконечности - параллельные прямые.

Но, в плане идеологии, речь идет о выборе между евклидовой и неевклидовой геометриями.

Если мы говорим о сообщении, как о геометрическом объекте и об энтропии - как его конструкционном параметре (аналоге объема), то очевидно, что шенноновская и комбинаторная энтропии - два представления того же самого (дискретного и детерминированного) объекта (пространства состояний), связанные через нелинейное преобразование (гамма-функцию). Причем одно из них аддитивно на всей области определения, а второе - только "приблизительно аддитивно".

Еще раз: энтропия Шеннона несовершенна (и, возможно, ошибочна). Я убежден, что

дальнейшее развитие научной мысли вытеснит ее куда-то на задворки истории, рядом с месмеризмом и флогистоном. Но, к настоящему времени, все здание теории Информации выстроено на ее свойстве абсолютной аддитивности (возможности "точно разлить коньяк по бутылкам"). Отказ от аддитивности равносилен отказу от пятого постулата Эвклида и требует пересмотра "всей геометрии". Понятие энтропии не может быть просто заменено, без ревизии всех теорем, в которых она участвует.

Это не означает, что в вашем компрессоре нельзя использовать какие угодно удобные вам численные оценки, но пока вы, подобно Лобачевскому, не докажете заново все теоремы, у вас нет права опираться на выводы из них.

Автор: gazlan [Чт мар 21, 2013 22:19]

Заголовок сообщения: Re: **Общая дискуссия о предельной степени сжатия**

sito писал(а):

"Если нет такого - то что будет если появится?"

Нет и не будет.

Задайтесь простым вопросом - до какого предельного размера можно разжать заданный файл?

Для любого заданного алфавита и набора кодов есть ровно два экстремальных значения длины кодированного текста: минимальная и максимальная.

В вырожденном случае (равная кратность всех символов алфавита) эти значения совпадают.

По аналогии - прямоугольный объект можно повернуть длинной стороной, можно - короткой.

Круглый - как ни крути - все того же размера.

Математически, "сжатие" (термин абсолютно неверный, но общепринятый) - это поворот.

Так вот, "круглое" - НЕ сжимается. Как ни крути :-)

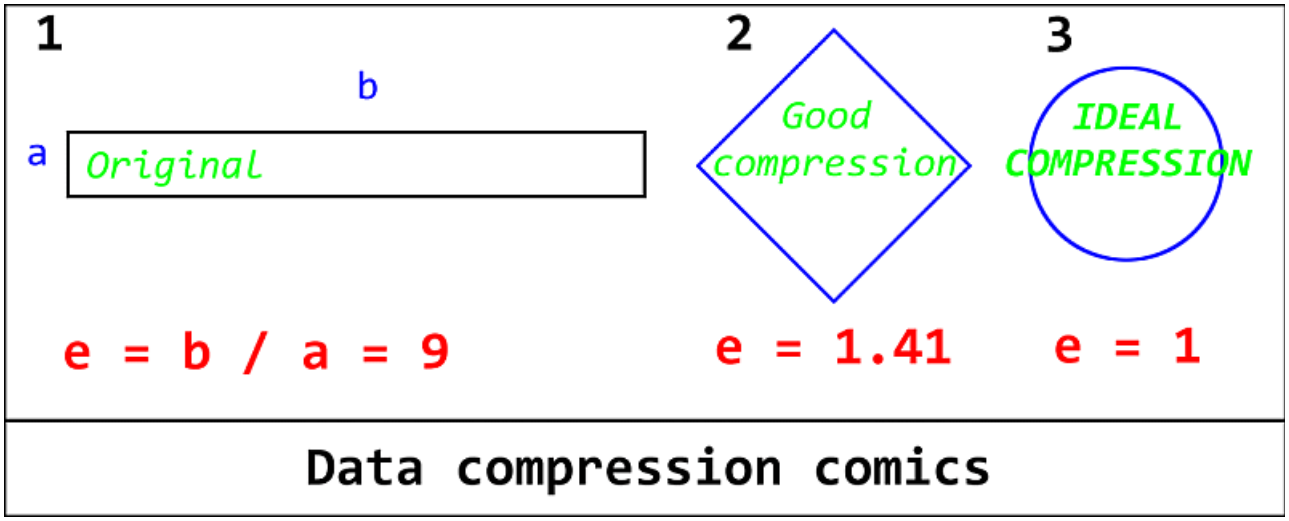
Автор: gazlan [Вс мар 24, 2013 14:10]

Заголовок сообщения: Re: **Общая дискуссия о предельной степени сжатия**

horlon писал(а):

"Банальный уход от ответа"

"На пальцах"



Компрессия основана на том, что в некоторых измерениях предмет скорее длинный, чем короткий. Назовем отношение самой длинной стороны к самой короткой - эксцентриситетом e. Если оригинальный файл (**1**) достаточно "эксцентричен", он может быть сжат "утаптыванием углов".

Хороший компрессор сжимает примерно как (**2**).

Если вам не жалко времени и денег, можете сжать сколь угодно близко к (**3**) с эксцентриситетом **e = 1** (энтропийный предел).

Но нет никакого способа сделать "круглое" еще более круглым.

Дальнейшее уменьшение эксцентриситета приведет только к росту размера сжатого файла.

Автор: gazlan [Пн мар 25, 2013 21:40]

Заголовок сообщения: Re: **Общая дискуссия о предельной степени сжатия**

horlon писал(а):

"*Я просто стараюсь доказать, что предела сжатию нет. Во сколько бы вы раз не сжали архив, есть способ сжать его сильнее...*"

Гм. Можно, я попытаюсь еще раз?

Зайду с другой масти.

Оперируя (общепринятым) термином "избыточность", можно сказать, что сжатие - это устранение избыточности. Условно, - "очистка" продукта от примеси.

Рассмотрим, к примеру, очистку от примесей золота.

Предположим, вы владелец небольшого аффинажного компрессора и принимаете в переработку файлы с содержанием продукта 37.5% (текстовые) и 58.3% (двоичные), получая на выходе сжатый продукт - 90.0% чистого вещества - так же, как и большинство остальных присутствующих на рынке.

Как человека творческого, 90.0% вас не устраивают и вы создаете суперкомпрессор Миллера, поднимающий степень сжатия с жалких 90.0% до вполне приемлемых 99.95%.

Так вот, если вам не жалко времени и денег, то, возможно, вы сможете улучшить степень очистки еще на "пару девяток" - до 99.9995%, но очевидно, что никакими ухищрениями невозможно добиться более чем 100% содержания продукта.

"Энтропийный предел" - это и есть 100% содержание (эксцентриситет e = 1). Никакая дальнейшая "очистка" (сжатие) невозможны.
