Сто постов про избыточность

Архиваторы: Избыточности не существует

---

gazlan #1 06-02-2009 12:12:13

Обсуждались архиваторы. Почему-то гуру компрессии не отзываются. А послушать критику хочется. Мой пост пока последний в ветке: Пн Фев 02, 2009 09:03 *
http://forum.compression.ru/viewtopic.php?t=2152

Stiver #2 06-02-2009 12:34:01

Потому что ответ тривиален: избыточность внес разархиватор.

---

gazlan #3 06-02-2009 12:40:22

> избыточность внес разархиватор

Иными словами, для каждого возможного сообщения декомпрессор "знает" какую избыточность и куда внести? Откуда взялось это знание?

Stiver #4 06-02-2009 13:02:43

> Откуда взялось это знание?

Из алгоритма естественно. Представь себе например программу, которая удваивает каждую переданную ей строку. Передаем A, получаем AA - избыточность явно растет. Откуда программа взяла "знание"?

---

gazlan #5 06-02-2009 13:11:16

> Из алгоритма

То есть, каждое возможное сообщение также содержит и свой "алгоритм" увеличения избыточности? Или каждое возможное сообщение содержит "универсальный алгоритм"? Или декомпрессор содержит "универсальный алгоритм" для каждого возможного сообщения?

Stiver #6 06-02-2009 13:26:38

> Или декомпрессор содержит "универсальный алгоритм" для каждого возможного сообщения?

Несомненно, он содержит алгоритм обработки каждого возможного сообщения. Это вроде бы очевидно :)

---

asmfan #7 06-02-2009 13:31:51

Алгоритмом создаётся/убирается избыточность. Суть кодирования.

---

gazlan #9 06-02-2009 14:59:13

> содержит алгоритм обработки каждого возможного сообщения

Угу. Тогда я хочу конкретизации. Возьмем, например, static Huffman. В этом случае все сводится к прямому и обратному (биективному) преобразованию, фактически - к мэппингу. Скажем, при (обычном) символьном кодировании это просто словарь из 256 кодов. Где здесь избыточность или устранение/введение избыточности?

4apa #15 06-02-2009 18:56:02

GoldFinch написал:

> информация меряется в битах, тритах, натах, и других подобных единицах

Допустим ты прав. Скажи тогда, товарищ, сколько информации находится в ... ну например в одной молекуле воздуха?

Как и чем ты ее собрался замерять если не секрет ??? Квантовой черпалкой информации что ли ? :)
---
vover #18 08-02-2009 00:44:55

> * битах, тритах, натах, * - это теория информации, которая ничего общего с информацией не имеет. Эта теория оперирует данными.

А по сабжу, информацию просто передали по разному закодированными данными, ничего ниоткуда не появилось и никуда не ушло.
---
gazlan #19 08-02-2009 17:41:19

> Информации не существует

Не существует (конструктивного) определения информации. Но есть определение количества информации - вполне полезное на практике. Сил инерции тоже не существует, но для расчетов они очень удобны :-)

А про избыточность так никто и не хочет рассказать?
---
gazlan #20 08-02-2009 17:49:11

> А по сабжу, информацию просто передали по разному закодированными данными, ничего ниоткуда не появилось и никуда не ушло

+1

В книжках все врут про избыточность :-)

Но вот интересно: нигде, кажется, не говорится явно, что оба - приемник и передатчик - должны разделять ту же самую (идентичную) модель данных, ибо "информация" суть знание этой модели.
---

leo #30 09-02-2009 11:09:21

Избыточность - это когда для кодирования символа \ сообщения \ файла используется большее количество бит, чем это реально необходимо с учетом статистики реального сообщения.

Например, кодировка русского текста в юникоде (ucs-2 = utf-16) является явно избыточной, т.к. из всего огромного набора символов используется лишь одна кодовая страница, поэтому достаточно указать номер этой кодовой страницы и закодировать текст в 8-битным ANSI. Но для осмысленного текста и 8-битная кодировка является избыточной.

Наглядным примером устранения этой избыточности в "железе" является стандартная клавиатура - вместо того, чтобы юзать ~256 кнопок на каждый символ, используется разбивка всего алфавита на поддиапазоны с переключением регистра (верхний \ нижний) и раскладки (рус \ лат).

Аналогичный прием разбивки на поддиапазоны используется и в телеграфных кодах - в итоге на каждый символ из одного поддиапазона используется только 5-6 бит, а при смене поддиапазона вставляется соответствующий управляющий префикс и поскольку эти переключения происходят сравнительно редко, то в среднем получается существенная экономия.

Другими словами, если символы в тексте группируются в серии по диапазонам, то кодировка каждого символа 8 битами является избыточной, т.к. можно указывать номер диапазона только в начале каждой серии, а не в каждом символе - явное устранение избыточности.

Поскольку алгоритм кодирования однозначный, то принимающая сторона может произвести обратное преобразование, добавив недостающие "избыточные" биты в каждый символ.

Что касается кода Хаффмана, то в нем для устранения избыточности используется кодировка символов разным количеством бит в зависимости от вероятности (частоты) их появления в тексте - часто встречающиеся символы кодируются меньшим числом бит, редко встречающиеся - бОльшим (аналогичный подход используется и в коде \ азбуке Морзе).

Говорить о том, что здесь все сводится к "мэппингу" и "словарю из 256 кодов" не совсем корректно, т.к. во-первых, за счет ранжировки символов по вероятности, вообще не встречающиеся символы оказыаются в конце таблицы (дерева) и по сути вообще выпадают из преобразования, т.е. происходит сокращение объема алфавита (длинные комбинации просто отсутствуют).

Во-вторых, за счет кодирования часто встречающихся символов меньшим числом бит (начиная с 2-х) средняя длина закодированного сообщения оказывается меньше исходной. Для декодирования такого кода достаточно знать порядок ранжировки символов - она м.б. или предопределена заранее или же передаваться в начале самого сообщения

---

leo #31 09-02-2009 11:56:53

> используется большее количество бит, чем это реально необходимо

Спасибо за подробное изложение вашего видения "избыточности". Но я предпочел бы ограничиться классическим определением Шеннона - как меры уклонения сообщения от чисто случайного. Иначе мы будем обсуждать различные вещи. Скажем, мультипликатор Stiver (#4 с предыдущей страницы) НЕ изменяет "избыточность", так как нет изменения частотностей. Юникод, коды Бодо и неиспользуемые символы к делу не относятся - очевидно подразумевается, что сообщение реализует всю мощность заданного алфавита. Работа Хаффмановского кодера описана вами правильно, но хотелось бы услышать, куда прячется "избыточность" в канале связи.
---
gazlan #37 09-02-2009 15:50:47

GoldFinch

Признаться, ничего не понял из вашего поста. Оставим в стороне путаницу с разрядностью.

> количество информации в сообщении не зависит от способа кодирования

Это постулат? Или результат экспериментов?

Если я кодирую 100-страничное сообщение как "План Z", как мне подсчитать количество информации в этой строке и убедиться, что оно не изменилось?

Или я постулирую, что оно не изменилось и всю невязку списываю на метод кодирования и "избыточность"?

GoldFinch #38 09-02-2009 16:01:41

> это подтверждается как теоретически так и практически, причем примерно одинаково количество информации в сообщении в битах - это двоичный логарифм числа всех N вариантов сообщения при условии что все варианты сообщения равновероятны (если не равновероятны то все тоже доказывается)

если кодирование обратимо, т.е. сообщение потом можно раскодировать, и при этом информация не потеряется то все эти N вариантов сообщения преобразуются в N кодированных сообщений, при этом количество информации равное log2(N) сохранится после декодирования N вариантов сообщения мы получим N исходных сообщений

например мы передаем в сообщении один из цветов радуги у нас 7 вариантов сообщения: в сообщении log2(7) ~ 2.807 бита информации как его не кодируй, словом "...", номером цвета, речью, непосредственно цветом (воспринимаемым визуально) в нем всеравно будет 2.807 бита информации
---
gazlan #39 09-02-2009 16:38:05

> после декодирования N вариантов сообщения мы получим N исходных сообщений

Обратимость не обсуждается. Я спрашивал о другом. Если вам не нравится фраза

"План Z", обратимся, например, к числу PI (3.1415926...).

Пусть исходное сообщение - это первая 1000 знаков числа PI, закодированное сообщение - фраза (символьный литерал) "первая 1000 знаков числа PI", декодированное сообщение идентично исходному - первая 1000 знаков числа PI (вычисленная декомпрессором).

От чего мы берем двоичный логарифм?

Если такой метод кодирования кажется вам непривычным, назовите его Static Huffman (формально, это так и есть - really!)

В качестве бонуса, поясните, plz, где здесь информация и где - избыточность.

GoldFinch #40 09-02-2009 17:27:09

число Пи известно, единственная информация о которой тут можно говорить, это число его знаков, вы собираетесь ЭТО передать?

другое дело если речь идет о передаче числа из 1000 знаков, тогда все просто, как и в посте выше

когда передается заранее известное сообщение, говорится что передается НОЛЬ информации поэтому если известно что передается число ПИ то информации в любом сообщении которое его передает 0, а избыточность 100%, потомучто но стороне приема число Пи известно

когда говорят о передаче сообщений, не говорят о конкретном сообщении.
---
leo #45 09-02-2009 21:46:21

> Спасибо за подробное изложение вашего видения "избыточности". Но я предпочел бы ограничиться классическим определением Шеннона - как меры уклонения сообщения от чисто случайного

"Мое видение" не отличается от шенноновского. Избыточность кодирования по определению это относительная разность между максимальной энтропией, которую мог бы обеспечить заданный алфавит при чисто случайном распределении символов, и реальной энтропией передаваемого сообщения.

Если юникодом или анси передавать случайную последовательность символов (например, закодированный текст), то его энтропия будет максимальной (16 или 8 бит/симв.) и соотв-но никакой избыточности в таком сообщении не будет. Если же передавать русский текст с явновыраженной разночастотностью и зависимостью появления символов, то его энтропия упадет до 3-4 бит/симв. и соотв-но 8-ми и тем более 16-ти битная кодировка станут избыточными.

> Иначе мы будем обсуждать различные вещи. Скажем, мультипликатор Stiver (#4 с предыдущей страницы) НЕ изменяет "избыточность", так как нет изменения частотностей

Если не учитывать зависимость между символами, то формально не изменяет. Но в

статистической теории в общем случае рассматривают не отдельные символы, а множество (ансамбль) всех возможных сообщений. Если сообщение по определению (или предположению) состоит из независимых символов, то расчет энтропии по множеству сообщений можно свести к расчету по вероятности появления отдельных символов.

Можно также учесть корреляцию двух и более последовательных символов (при наличии корреляции энтропия ес-но будет уменьшаться). Но для экзотических или специальных случаев типа "мультипликатора" Stiver, или цветов радуги GoldFinch нужно рассматривать все возможное мн-во сообщений, тогда станет ясно, что "мультипликатор" дважды гонит одну и ту же строку, а для передачи цветов радуги достаточно не более 3-х бит.

А вот ты как раз со своим "планом Z" или "1000 знаками числа Pi" как раз и уходишь от шенноновских определений фиг знает куда, т.к. по одному сообщению и соответсвующему ему коду ничего сказать нельзя - нужно рассматривать все множество допустимых сообщений. Если в твоей системе заложено всего 2 "плана", известые на обоих сторонах, то независимо от того, по сколько страниц они занимают, выбор одного из двух требует всего 1 бита информации.
---
gazlan #46 10-02-2009 07:32:56

> уходишь от шенноновских определений

Не ухожу. Просто иллюстрирую, что кодирование - суть переименование.

Вернемся к static Huffman, словарю в 256 кодов и набору символов ASCII.

Очевидно, что "сжатое" сообщение - это просто переименованное исходное.

Что у нас с информацией и с избыточностью? К слову, размер произвольного "сжатого" сообщения в таком эксперименте необязательно будет меньше исходного - может быть и больше. Fifty-fifty.
---
gazlan #47 10-02-2009 07:45:13

> число Пи известно, единственная информация о которой тут можно говорить, это число его знаков

Я собираюсь передать 1000 знаков, чудесным образом совпавших с первой тысячей знаков числа PI. "Фрактальный" компрессор сумел уловить это совпадение (детали реализации), а декомпрессор на приемной стороне сгенерировал их по формуле. Акцент - на модели, разделяемой абонентами на обоих концах канала связи. То, что мы привыкли называть "информацией" (и подсчитывать ее количество) имеет смысл только в рамках строго определенной _разделяемой_ модели.
---
aa_dav #48 10-02-2009 10:52:37

Ну я тебе на пальцах покажу где избыточность и куда она девается. =)

Строки "1001(2)" и "9(10)" обозначают одно и то же число, но первая запись занимает на 2 байта больше места в накопителе информации. Вот эти "лишние" 2 байта, на

которых можно сэкономить, которыми можно не захламлять устройства хранения и связи - и есть та самая избыточность, которую ты до сих пор никак не можешь нащупать ни в воздухе ни в радиоволнах, а вот они - тута, щупай сколько влезет и наслаждайся результатом. Наконец то ты увидел и потрогал избыточность в 2 байта, и я рад. =)
---
gazlan #49 10-02-2009 11:14:00

> обозначают одно и то же число

Только в рамках разделяемой модели

Где здесь избыточность по Шеннону?
---
gazlan #50 10-02-2009 11:49:53

Не всем, наверное, ясно чего ради я затеял это обсуждение, поэтому попробую кратко изложить свое понимание информационного обмена:

Обмен информацией происходит между источником и приемником через канал связи (Посылка сообщения).

Материальные аспекты переноса (масса, энергия) игнорируются.

Оба - источник и приемник разделяют идентичную модель системы. (Скажем, атомный цезиевый эталон времени и ходики с кукушкой разделяют модель "часы, способные отображать текущее время с точностью до одной минуты").

Информация, передаваемая по каналу связи используется для синхронизации моделей (Скажем, сигналы точного времени).

Информация из п.4 (синхросигнал) является относительной величиной, не имеющей никакого смысла вне контекста разделяемой модели (невязка между текущим и эталонным состояниями).

Рассматривая Сообщение как многомерный вектор (объект с i степенями свободы), можно выбрать такую систему координат, в которой его представление будет выглядеть простейшим образом (Нормализация, Приведение к Главным осям).

"Сжатие" сообщения есть ни что иное, как нормализация (многомерный сдвиг и поворот).

Шифрование также является многомерным сдвигом и поворотом.

Следствие:

Информация нематериальна, неуничтожима и неперемещаема. (Иными словами, информацию можно только клонировать).
---
black_hole #51 10-02-2009 14:57:46

Ничего не понял, но какая практическая польза?

---
aa_dav #52 10-02-2009 14:58:20

> обозначают одно и то же число

> 1. Только в рамках разделяемой модели

Я не очень понял что есть такое эта разделяемая модель. Знания о том как записывать числа текстом? Если так, то да, конечно, обозначение чисел имеют смысл только в рамках правил/знаний как их записывать. Это как-то даже странно обсуждать...

> 2. Где здесь изыточность по Шеннону?

В первом случае используется почти только 2 состояния буквы алфавита, из всех возможных! А во втором - 10. Что это как не мера уклонения от случайности, которая в первом случае выходит больше, больше и Шенноновская избыточность.
---
aa_dav #53 10-02-2009 15:04:53

> Следствие:

> Информация нематериальна, неуничтожима и неперемещаема. (Иными словами, информацию можно только клонировать).

Эээ... Непонятно куда делся топикстартеровый тезис "избыточности не существует"... =)

Но и собственно непонятно откуда из изложенного "понимания информационного обмена" вытек вдруг вот это вот следствие. Не то чтобы оно не верно - мир Эйдосов (нематериальный мир идей, образов, информации, где отсутствует понятие движения или даже просто местоположения) еще знаете ли до нашей эры придумывали философы. Нет. Просто совершенно непонятно как сам по себе этот "вывод" связан с предшествующими ему 8 пунктами. =)
---
gazlan #54 10-02-2009 16:11:35

> какая практическая польза

На сегодняшний день, понятие Информации - это предмет для всевозможных спекуляций (a la "Инфодинамика"). Предлагаемая концепция, возможно, позволит вернуться на реальную почву и воспользоваться результатами Теории Синхронизации Динамических Систем (Информация - как слабая связь).
---
gazlan #55 10-02-2009 16:21:00

> что есть такое эта разделяемая модель

Неявно, всегда подразумевается некоторый общий базис для источника и приемника. Например, знание Алфавита, который используется в сообщении. Скажем, если передается литера "А", то количество полученной информации будет различным для латинского и русского алфавитов. Разделяемая Модель - это явное

подчеркнутое указание на общий для обоих корреспондентов тезаурус.
---
gazlan #56 10-02-2009 16:30:57

> куда делся топикстартеровый тезис

Никуда не делся. Это следует из самого процесса компрессии, как преобразования системы координат. Собственно информация при этом должна быть инвариантом преобразования (Сравните, например, с каноническим представлением плоских кривых второго порядка). Понятие "избыточности" при этом вряд ли вообще имеет смысл.

> вдруг вот это вот следствие

Это следствие самого понимания информации как синхросигнала. То есть, информация - это некоторая условная величина (невязка), в отличие от объектов реального мира, обладающих свойствами Находиться и Перемещаться. Не являясь физическим объектом (мы абстрагируемся от реальной энергетики процесса переноса), информация подобными свойствами не обладает по определению, она - результат расчета и не более реальна, чем "мнимая единица".
---
aa_dav #57 10-02-2009 16:39:34

ага. ну всё верно. вы же сами говорите, что вектор информации может быть представлен в некоем "минимальном" виде. вот все остальные виды в этом смысле и избыточны. в принципе, я не вижу тут что еще можно обсуждать на тему избыточности - она есть и это факт.
---
gazlan #58 10-02-2009 16:49:39

> не вижу тут что еще можно обсуждать

Мне кажется неправильным само разделение на компоненты. Это тоже самое сообщение, только записанное другими буквами. Взгляд из другой системы координат. Сравните, например, с коническими сечениями. Является ли парабола более "избыточной", чем эллипс?
---
aa_dav #59 10-02-2009 17:18:07

Мы уже выяснили что "1001(2)" более избыточно чем "9(10)", при всём при том что вы тут говорите "что это взгляд из другой системы координат (способ записи)", что "это тоже самое сообщение, только записанное другими буквами", всё так и есть и тем не менее 1001(2) более избыточно, чем 9(10) на 2 знака, в чём легко убедится прямо сейчас сосчитав пальцем эти самые знаки - раз, два, три... и так далее. =) О чём еще речь может быть? Всё вашими же терминами и подтверждается.

Про эллипс и параболу - просто fail - мы же не геометрию тут обсуждаем.
---
black_hole #60 10-02-2009 21:10:19

Избыточность вносится возможностью процессинга :) Говоря вашим языком, возможностью совершать преобразования координат...В чистом виде информации не

существует :) Вы ищите философский камень...пока это имеет практическую выгоду, "информация" удобна, ну а когда начинаются спекуляции, то лучше читать историю и классику:)

GoldFinch #61 11-02-2009 00:54:52

gazlan, ваша "теория" - это не теория а просто набор пустых фраз. Хотите выдвинуть теорию - сначала узнайте что означает термин "теория" а потом уже пишите свои домыслы.

> 1. Обмен информацией происходит между источником и приемником через канал связи (Посылка сообщения).

> 2. Материальные аспекты переноса (масса, энергия) игнорируются.

> 3. Оба - источник и приемник разделяют идентичную модель системы. (Скажем, атомный цезиевый эталон времени и ходики с кукушкой разделяют модель "часы, способные отображать текущее время с точностью до одной минуты").

> 4. Информация, передаваемая по каналу связи используется для синхронизации моделей (Скажем, сигналы точного времени).

> 5. ...

1. что такое информация в контексте вашей теории? что такое источник информации? что такое приемник информации? что такое канал связи?

2. почему игнорируются?

3. какой системы? какой критерий идентичности?

4. как информация может использоваться для синхронизации? зачем нужна синхронизация?

...

> "Рассматривая Сообщение как многомерный вектор (объект с i степенями свободы)"

какому множеству принадлежит i?

сдвиг и поворот - это аффинные преобразования, у вас вектор в каком пространстве?

> "Шифрование также является"

это называется не шифрование а кодирование и только для цифрового представления сообщений.

DEN #62 11-02-2009 18:24:34

Хы) Берем гигабайтный архив rar, создаем свой архиватор и добавляем этот архив в

словарь. Имеем архиватор, пакующий гигабайтный pap в несколько байт :-)
---
gazlan #63 11-02-2009 21:34:23

> 1. что такое информация в контексте вашей теории? что такое источник информации? что такое приемник информации? что такое канал связи?

> 2. почему игнорируются?

> 3. какой системы? какой критерий идентичности?

> 4. как информация может использоваться для синхронизации? зачем нужна синхронизация?

...

>> "Рассматривая Сообщение как многомерный вектор (объект с i степенями свободы)"

> какому множеству принадлежит i?

> сдвиг и поворот - это аффинные преобразования, у вас вектор в каком пространстве?

>> "Шифрование также является"

> это называется не шифрование а кодирование и только для цифрового представления сообщений.

Начну с "системы". Объект характеризуемый набором состояний (т.е. их количеством и возможностью перехода из одного в другое) и набором связей (ограничений). Скажем, правила русского языка запрещают некоторые буквы, как начальные буквы слов.

Совокупность всех используемых правил будем называть моделью системы.

Две системы могут быть (слабо) связаны между собой. Изменение состояния одной из них влекут изменение состояния другой.

Если такая связь существует, мы говорим об информационном обмене. Термин "информационный" в данном контексте означает, что связь настолько слаба, что энергетикой переноса можно пренебречь.

Для простоты, рассматриваем процесс влияния как однонаправленный (довольно частый случай на практике), при этом задающая система является источником сигнала, а ведомая - приемником.

Каналом связи будем называть все что угодно, что служит целям такого обмена (электрическая линия, мировой эфир, компрессор-декомпрессор) - любую "проводящую прокладку" между двумя взаимодействующими системами. Для простоты, считаем канал идеальным (нет шумов, нет потерь).

Реальные процессы переноса всегда сводятся к переносу массы/энергии. Однако, при информационном обмене этот перенос (поток управления) настолько незначителен по сравнению с общей энергетикой управляемой системы (для управления используется другой источник энергии - сравните, например, с управлением автомобилем), что энергетикой обмена можно пренебречь (как это всегда делалось в Теории Информации).

Речь идет о модели из п.1 Идентичность означает, что оба - источник и приемник - используют ту же самую модель. На практике, подмена модели означает, по сути, дезинформацию: корректные сообщения некорректно интерпретируются. Множество литературных произведений построено именно на этом.

Вообще говоря, (авто)синхронизация фундаментальное свойство реальных физических систем (резонанс планет; аплодисменты, переходящие в овации итд.), вероятно отвечающее критерию энергетической устойчивости (потенциальной яме). Практически, имеется бесчисленное количество примеров, где требуется синхронизация двух моделей - от сверки часов до управления армиями. По сути, информирование - это просто синоним для синхронизации моделей.

Считайте, для простоты, что речь идет о счетном множестве. Все реальные объекты конечны.

Термины "сдвиг и поворот" употребляются только как аналогия. Не принимайте их слишком всерьез. Реальные преобразования (та же компресиия) - нелинейны.

Разумеется, есть разница между шифрованием и кодированием, но в данном случае я хочу подчеркнуть общие аспекты преобразования.
---
leo #66 12-02-2009 10:28:42

Может и не бред, а философские измышлизмы, не имеющие отношения к классической шенноновской теории. Но главная неувязочка в том, что об избыточности можно говорить только в случае определения некой меры информации и некой меры затраты неких ресурсов, на передачу единицы информации. У Шеннона это все четко определено, а у философа gazlan-а нет ни того, ни другого - поэтому и понятия избыточности нет :D
---
gazlan #68 12-02-2009 14:14:57

> к теории как к таковой вообще

Спасибо за высокую оценку :-)

На самом деле, понятие модели введено в теорию Компрессии еще в 80-х годах прошлого века и давно стало общепризнанным. Представление дискретных нелинейных преобразований как движения в многомерном криволинейном пространстве (я использовал метафору "Сдвиг и поворот") восходит к 20-м годам прошлого века (работы Г.Крона), Теория Синхронизации - вполне разработанный раздел Прикладной Математики (интересные результаты можно найти в работах И. Блехмана).

Что еще осталось? Нематериальность информации - следует из самого факта

пренебрежения материальными аспектами переноса.

Указание на то, что информационный обмен суть синхронизация? С удовольствием ознакомлюсь с вашей точкой зрения. Желательно кроме эмоций, привести и какие-либо доводы.
---
cppasm #69 12-02-2009 17:54:54

gazlan, какой смысл парить себе и другим мозги?

Для начала: static Huffman - это НЕ переименование. За счёт переименования ты избыточность не уберёш. Основная суть в том, что кодирование исходного сообщения рассчитано на случайные равновероятные данные.

А кодирование "сжатого" - это кодирование с учётом специфики конкретного сообщения. Т.е. если кодировать в ANSI любой текст, получаем 256 вариантов, или 8 бит/символ. Но если кодировать конктретное сообщение, в котором символы не равновероятны - ты получиш для часто повторяющихся символов код короче 8 бит, для редко встречающихся - возможно больше 8 бит. Длина кода будет разная.

Так что кодирование по Хаффману это не простое переименование. Простое переименование - это табличная замена, и никого сжатия она не даст.
---
gazlan #70 13-02-2009 13:55:03

> static Huffman - это НЕ переименование

Ознакомьтесь, для начала с алгоритмом. Это именно табличная замена. Как именно инициализирована таблица, для данного обсуждения неважно.
---
cppasm #71 13-02-2009 14:57:31

С алгоритмом я знаком, а вот ты похоже нет.

У Хаффмана переменная длина кода, а у табличной замены постоянная.
---
gazlan #72 14-02-2009 21:10:55

> а у табличной замены постоянная

??

У static Huffman ни коды ни их длины НЕ изменяются после инициализации таблицы и, в общем случае, не связаны с заданным входным потоком (таблица строится заранее для некоторого класса сообщений).

Я даже больше скажу: Любой алгоритм lossless компресии суть табличная подстановка. Всего возможны три варианта:

1. замена блоков равной длины на блоки неравной длины (ex. Huffman).

2. замена блоков неравной длины на блоки равной длины (ex. LZ).

3. комбинация двух первых (ex. LZH).

В статических методах таблица (S-box) инициализирована статически, в адаптивных - обновляется. При шифровании, в инициализации и обновлении участвует также секретный ключ. Оставляя в стороне детали реализации, компрессия - это шифрование с "прозрачным" ключом.
---
gazlan #73 17-02-2009 07:39:42

Как-то тихо ... подолью масла в огонь ... шарик под левым наперстком - следите за руками :-)

На самом деле, я совершенно серьезен в отношении избыточности. Напомню исходное положение:

Простой мысленный эксперимент: Алиса посылает Бобу текстовое сообщение, сжав его архиватором, удалившим всю избыточность. Боб получает сообщение и разархировав его, получает копию оригинального сообщения со всей его избыточностью. По каналу связи "избыточность" не передавалась. В сжатом сообщении ее не было по определению. Вопрос: откуда она взялась в восстановленном сообщении?

Stiver пишет:

> Потому что ответ тривиален: избыточность внес разархиватор.

Легко показать абсурдность этого утверждения.

Пусть сжатое сообщение на сколько-то бит короче исходного. С точки зрения декодера (декомпрессора) на вход ему поступило "испорченное" исходное сообщение, требующее восстановления: часть бит искажена и часть - утеряна. Используя (остаточную) избыточность сжатого сообщения, декодер восстанавливает исходное: корректирует присутствующие биты и дописывает отсутствующие. Иными словами, из этого следует, что избыточность сжатого сообщения больше избыточности оригинального.
---
asd #74 17-02-2009 08:03:27

gazlan

Фиговый из тебя провокатор за 11 дней всего на 3 страницы тема.
---
gazlan #75 17-02-2009 08:27:23

> Фиговый из тебя провокатор

:-)

На самом деле - никакой провокации. Просто побочный результат размышлений на темы компрессии, шифрования и, если так выразиться, "физической" сущности информации.

Одно только (со)существование нескольких дюжин "теорий информации" - симптом неблагополучия. Количественные оценки не имеют ничего общего с качественными (приходится вводить понятие "новой" информации итп), а вероятностные методы (как указывал Колмогоров) большей частью совершенно неприменимы на практике.

Имей я готовые ответы на все вопросы - не было бы нужды заводить топик. Однако, возникло некоторое иное понимание привычных ранее вещей - и открытая дискуссия кажется мне лучшим методом оценки его правильности.
---
leo #77 17-02-2009 10:24:13

> По каналу связи "избыточность" не передавалась. В сжатом сообщении ее не было по определению. Вопрос: откуда она взялась в восстановленном сообщении?

Или ты туп как дерево или прикидываешься

Избыточность незачем передавать по каналу связи, т.к. ее можно заложить в "модель системы", т.е. перевести в разряд априорного "знания", заранее известного архиватору (передатчику) и разархиватору (приемнику).

Примеров - масса. В дельта-кодировании избыточность устраняется \ восстанавливается за счет априорного знания коррелированности передаваемых числовых значений, что позволяет передавать не каждое число, а только их приращения (аналогичная ситуация и с синхронизацией часов - если часы достаточно стабильны, то можно периодически "подкручивать" только секунды \ миллисекунды \ и т.д. и не передавать каждый раз всю эпоху).

В RLE юзается априорное знание наличия в сообщении длинных серий повторяющихся символов \ чисел.

В классическом коде Хаффмана юзается различие вероятностей и соотв-но разное кол-во бит представления символов.

В твоей интерпретации - юзается особое "преобразование координат многомерноо вектора" и соотв-но избыточность переходит в знание алгоритма этого преобразования. Если декодер \ разархиватор обладает этим знанием, то он может восстановить исходное "сырое" сообщение, добавив в него избыточность по известному алгоритму. А ежели не знает, то для него закодированное без-избыточное сообщение будет выглядеть набором случайных символов

> Иными словами, из этого следует, что избыточность сжатого сообщения больше избыточности оригинального

С какой стати ?! Из этого следует, что избыточность исходного текста "настолько велика", что ее урезание до некого предела позволяет тем не менее устранять ошибки при передаче. Просто чем больше избыточность, тем больше случайных ошибок может быть устранено и наоборот - не более того.
---
gazlan #78 17-02-2009 11:22:50

> в виде формул, основанных на математике

Формулы - это просто вид стенографической записи. Не вижу нужды в греческих буквах там, где достаточно связного русского текста.

> избыточность переходит в знание алгоритма этого преобразования

Выше мы установили, что "преобразование" является простым переименованием. В таком случае, "избыточность" - некая фиктивная характеристика, зависящая от текущей системы координат (способа наименования). Или, лучше - не является инвариантной характеристикой сообщения. То же для случая, когда избыточность "прячется" в кодере/декодере.

> Из этого следует, что избыточность исходного текста "настолько велика", что ее урезание до некого предела позволяет тем не менее устранять ошибки при передаче

Избыточность исходного текста не имеет никакого значения. Мы говорим о восстановлении исходного текста из сжатого. Исключая тривиальный случай простого копирования, избыточность сжатого текста в этом случае _обязана_ быть больше избыточности исходного.
---
gazlan #81 17-02-2009 11:52:22

> Избыточность незачем передавать по каналу связи, т.к. ее можно заложить в "модель системы", т.е. перевести в разряд априорного "знания", заранее известного архиватору (передатчику) и разархиватору (приемнику).

Забыл сказать: да, это НЕ вызывает возражений. Но это и есть то, что я называю разделяемой моделью и синхросигналом.

Вот еще пример: в языке иврит гласные буквы на письме опускаются (трудно было писать на камнях). Зачастую слово может быть восстановлено только в контексте (т.к. возможны различные варианты огласовок). Очевидно, что пишущий (кодер) и читающий (декодер) разделяют сложную модель языка - не только словарь, но и значительный набор правил. С точки зрения классической Теории Информации, ивритский текст обладает очень низкой избыточностью.
---
aa_dav #82 17-02-2009 18:34:39

> С точки зрения декодера (декомпрессора) на вход ему поступило "испорченное" исходное сообщение...

Нет. В этом постулате ошибка.
---
gazlan #83 18-02-2009 08:00:28

Признаться, не вижу в чем, можно подробнее?

Well. Изменим формулировку, уберем "испорченное". Пусть теперь, декодер - обычный транслятор (текст-2-текст), обрабатывающий входной поток в соответствии с некоторой (бесконфликтной) грамматикой. Входной поток считаем синтаксически правильным. Все ограничения (связи) сосредоточены в разделяемой модели (сжатый текст псевдослучаен).

Вопросы:

Где происходит восстановление избыточности?

Является ли эта (семантическая) избыточность избыточностью по Шеннону?
---
aa_dav #84 18-02-2009 08:54:07

> Признаться, не вижу в чем, можно подробнее?

В том что сжатое сообщение не является испорченным ни в коей мере. Если реально испортить хоть один битик в том же хаффмане - всё оригинальное сообщение после этого битика пойдет лесом. Если же битик поменять в оригинальном сообщение испортится только один этот битик (символ, если речь идет о тексте). Т.е. действительно избыточность несжатого сообщения выше, как и обещал прогноз.

> 1. Где происходит восстановление избыточности?

> 2. Является ли эта (семантическая) избыточность избыточностью по Шеннону?

Вам уже ответили несколько раз на эти вопросы. Восстановление происходит в декодере, да, является.
---
gazlan #85 18-02-2009 09:16:20

"Порчу" я понимаю как отличие от оригинального сообщения, а не как нарушение структуры сжатого.

Пример же с испорченным битом некорректен. Например, если изменить (на один бит) поле размера записи в несжатом DBF-файле, то программа чтения (DBF Reader) точно также не сможет обработать (некорректно выдаст) весь оставшийся поток. Избыточность здесь ни при чем. И даже больше - если сжатый поток по размеру меньше оригинального, то восстановление при ошибке (ECC) обойдется дешевле - просто в силу меньшего размера. Опора на избыточность исходного текста (скажем, пропуск гласных букв при письме) требует совсем иной модели восстановления.
---
aa_dav #86 18-02-2009 10:07:31

> "Порчу" я понимаю как отличие от оригинального сообщения, а не как нарушение структуры сжатого.

И совершенно неправильно понимаете. Порча - это изменение сообщения, приводящее к неправильному декодированию этого сообщения. Поэтому оно и называется "порча", а не как-то иначе.

> Пример же с испорченным битом некорректен. Например, если изменить (на один бит) поле размера записи в несжатом DBF-файле, то программа чтения (DBF Reader) точно также не сможет обработать (некорректно выдаст) весь оставшийся поток. Избыточность здесь ни при чем.

Для DBF-а существует repair-утилиты, если вы не знали, как раз пытающиеся на основании избыточности файла восстановить из него хоть что-то - фиксированный размер записей тому способствует. А вот если бы бит был замещен в сжатом в RAR DBF файле, то никакая repair-утилита уже бы не помогла. Что опять таки доказывает вашу неправоту по отношению к избыточности.
---
gazlan #87 18-02-2009 10:31:51

> И совершенно неправильно понимаете

Кажется, мы о разном.

Я не обсуждаю семантику слова "порча" в контексте декодирования - я ее постулирую. Если это слово кажется вам неподходящим - найдем другое. Важно, что имеются два связанных сообщения и процесс получения одного из другого рассматривается как "восстановление", что априори задает большую избыточность первого по отношению ко второму (выводимому из первого).

Я готов согласиться, что избыточность (в виде набора правил) может быть спрятана в самом декодере (как разделяемая модель), но тогда какой смысл приписывать ее сообщению? (Пример правил: русский текст не может содержать более двух одинаковых согласных подряд или более четырех согласных подряд).

Я знаю и про repair утилиты для DBF и про ключ 'r' (Repair archive) в RAR, но не понимаю как это соотносится с приведенным вами примером, где никакая коррекция ошибок НЕ используется
---
aa_dav #88 18-02-2009 10:48:43

> рассматривается как "восстановление", что априори задает большую избыточность первого по отношению ко второму (выводимому из первого).

А кто сказал что это так? =) Вы как раз аппелируете к смыслу "порчи", т.к. действительно, чтобы восстановить испорченное сообщение в первозданном (неиспорченном) виде, требуется чтобы кодированное сообщение было априори более избыточно чем восстановленное. Как пример - CD-диски, на которых хранится примерно на 30% больше бит, чем заявлено на обложке (600-700Мб), за счёт кодов коррекции Рида-Соломона, позволяет читать даже зацарапанный диск.

Это всё известно, но тут как раз семантика слова "порча" заключается именно в смысле "порча", как я его описал.

То о чём вы говорите, порчей не является и не требует "большую избыточность первого по отношению ко второму" никоим образом.
---
gazlan #89 18-02-2009 10:58:06

> То о чём вы говорите, порчей не является и не требует "большую избыточность первого по отношению ко второму" никоим образом

ОК. Забудем про "порчу".

Вы согласны, что если второе сообщение выводится из первого, то это означает, что первое содержит всю информацию из второго + правила вывода? Если же считать правила вывода "упрятанными" в декодер, то какое основание приписывать эту информацию выводимому сообщению?
---
aa_dav #90 18-02-2009 11:06:47

> Вы согласны, что если второе сообщение выводится из первого, то это означает, что первое содержит всю информацию из второго + правила вывода?

Правила вывода содержаться в декодере.

> Если же считать правила вывода "упрятанными" в декодер, то какое основание приписывать эту информацию выводимому сообщению?

Никаких. В декодированном сообщении тем более нет никаких правил вывода.

Но о чём это вы тут? Я не понял к чему эти выводы.
---
gazlan #91 18-02-2009 11:28:29

> Но о чём это вы тут?

Да все о том же: пытаюсь понять "физический" смысл понятия "избыточность". Пока она мне видится фиктивной величиной, зависящей от "системы отсчета" (метода кодирования).
---
aa_dav #92 18-02-2009 11:36:24

А. Ну нет. Пока оно занимает место на накопителях информации и напрягает каналы передачи данных - фиктивным оно не будет. =)
---
leo #93 18-02-2009 11:45:17

> Я готов согласиться, что избыточность (в виде набора правил) может быть спрятана в самом декодере (как разделяемая модель), но тогда какой смысл приписывать ее сообщению?

Никто ее сообщению и не приписывает. В #78 ты наконец-то сделал правильный вывод, о том, что избыточность это характеристика не самого сообщения, а его исходной кодировки \ модели \ _системы_координат_.

Если в твоей модели заложены некие знания, позволяющие представить сообщение (точнее некий класс \ вид сообщений) в более "компактном виде", то можно говорить о том, что исходная кодировка является "избыточной" для данного вида сообщений. В итоге эту "избыточность" можно не хранить для каждого сообщения и не передавать по каналу связи, а заложить ее в алгоритм кодирования \ декодирования. Т.е. одно и то же сообщение может быть представлено в разных кодировках (моделях) с разной степенью избыточности.

Например, возвращаясь к нашим баранам ;), кодировка русского текста в формате юникод является очень избыточной, т.к. "модель" юникода предполагает любое \

случайное чередование символов из множества различных кодовых страниц. Если мы знаем \ определяем, что весь текст принадлежит одной странице, то можем использовать менее избыточную кодировку анси. Но для связного текста она ес-но тоже является избыточной, т.к. преполагает случайное чередование симолов, поэтому можно заюзать другие модели, учитывающие реальную (или среднюю) статистику \ корреляцию символов в сообщении (или классе сообщений).

В итоге, чем более "хитрую" модель мы используем и чем более "компактным" (в общем случае в среднем для всего класса сообщений) в итоге получается закодированное сообщение по сравнению с исходным, тем больше "избыточности" было в исходном представлении сообщения по сравнению с нашей "хитрой" кодировкой/моделью.

Использование избыточности для коррекция ошибок никак не влияет на определение самой избыточности как меры соотношения "компактностей" представления сообщений в разных кодировках \ моделях. Поэтому приплетать сюда возможности коррекции совершенно незачем
---
gazlan #94 18-02-2009 12:03:37

> фиктивным оно не будет

Попытаюсь пояснить, почему я говорю о "фиктивности".

Пусть имеется некое достаточно длинное и достаточно сжимаемое сообщение.

Пусть оно хорошо сжимается как с использование Хаффмановского кодера, так и с использованием словарной схемы LZW. Для простоты, примем обе схемы статическими: первый проход - составление словаря, второй - кодирование. Предположим, что длины всех трех сообщений (Src, Huf and LZW) различаются и, соответственно, имеются три оценки избыточности.

А теперь вспомним, что и Huf и LZW являются просто способами переименования. Для Huf равномерные коды кодируются неравномерными, для LZW - в точности наоборот.

Иными словами, три разных способа разбиения/именования на битовые блоки дают три разных оценки.

Следовательно, эти оценки относятся к способу разбиения (структуре и содержанию словаря), но не к самому объекту.
---
gazlan #95 18-02-2009 12:09:38

> чем более "хитрую" модель мы используем и чем более "компактным" (в общем случае в среднем для всего класса сообщений) в итоге получается закодированное сообщение по сравнению с исходным, тем больше "избыточности" было в исходном представлении сообщения по сравнению с нашей "хитрой" кодировкой

С этим я согласен, но такое представление не является общепринятым. По сути, вы вводите "относительную" избыточность, тогда как классическая теория оперирует "абсолютной".

GoldFinch #96 18-02-2009 12:13:04

> вы вводите "относительную" избыточность, тогда как классическая теория оперирует "абсолютной".

в классической теории избыточность определяется в зависимости от "модели"
---
leo #97 18-02-2009 12:30:40

> По сути, вы вводите "относительную" избыточность, тогда как классическая теория оперирует "абсолютной".

Угу, иди для начала книжки почитай ;) Я уже приводил классическое определение избыточности - в ней фигурирует максимальная энтропия источника сообщений при заданном алфавите или длине исх.сообщения, т.е. если у тебя есть исходное сообщение длиной 100 бит, то соотв-но максимальная энтропия и есть 100 бит в преположении их случайного чередования. Если есть сообщения произвольной длины, использующие "алфавит" анси, то соотв-но макс.энтропия будет 8 бит/символ, а в "алфавите" юникод - 16 бит/символ.

А энтропия русского текста как ни крути ~3-4 бит/символ - вот тебе и разная избыточность в разных кодировках
---
gazlan #98 18-02-2009 12:33:03

> в классической теории избыточность определяется в зависимости от "модели"

Статистической. Семантика остается за бортом.
---
aa_dav #99 18-02-2009 15:22:39

> Следовательно, эти оценки относятся к способу разбиения (структуре и содержанию словаря), но не к самому объекту.

Ну разумеется. Объект - сообщение. А предмет оценки избыточности - его представление. В чём ты тут увидел подвох?
---
gazlan #100 18-02-2009 15:43:52

Мне кажется, что это разделение: Объект (как инвариант) и его представление (как набор чисел в данной системе координат) нигде явно не проводится.
---
leo #105 18-02-2009 22:05:13

Повторю, что в шенноновской теории и в ее современной интерпретации речь идет не об избыточности конкретного сообщения ("объекта"), а об избыточности источника сообщений (или языка), генерирующего сообщения из некоторого набора символов (алфавита). Соотв-но мат.определение избыточности это - единица минус отношение (_реальной_) энтропии источника к максимальной (_воображаемой_) энтропии, которую он мог бы обеспечить, используя тот же набор символов (т.е. при их случайном равновероятном чередовании).

Это определение есть и у самого Шеннона в "Мат.теории связи", и во всех учебниках по теории информации, и в интернете кучу ссылок можно найти, в т.ч. и в википедии (по кр.мере в англ.варианте - redundancy)
---
gazlan #107 19-02-2009 07:28:54

Все, что вы написали, совершенно правильно, но мне не кажется, что из этого следует явное разделение объекта и представления.

Y_Mur писал:

> определить количество "чистой" информации

IMHO, никак. У Колмогорова есть интересные рассуждения на эту тему (на примере "Войны и мира"). Шеннон вводит "абсолютный (термодинамический) нуль отсчета" как энтропию белого шума и уже от него отсчитывается "количество информации".

P.S.

Еще на тему разделяемой модели :-)

Вчера, в блоге Садальского:

"- Король сказал, что двери его сокровищницы открыты передо мной! - буркнул рыцарь" (http://gazlan.narod.ru/library/knight.html).
---
leo #108 19-02-2009 10:18:45

Y_Mur

Не надо подливать масла в огонь ;)

Во-первых, "на пальцах" здесь уже уже с десяток раз пытались объяснить понятие избыточности, но ТС это не устравивает и ему подавай "классику".

Во-вторых, классическое определение как раз ничего не затуманивает, а объясняет другими более абстрактными словами \ понятиями. Да, русский текст можно записать в разных по объему алфавитах - ограниченном алфавите кода Бодо, в расширенном алфавите ANSI или в сверхрасширенном UNICODE. Но энтропия (или "чистая информация") этого текста будет одинаковой во всех этих алфавитах (при этом не важно, как конкретно ее расчитывать, т.е. насколько глубоко учитывать межсимвольные связи, главное чтобы было единообразно для любых алфавитов).

А вот основная составляющая избыточности - максимальная энтропия источника - разумеется будет разной для разных кодировок, т.к. неиспользуемые (или крайне редко используемые) символы алфавита будут давать нулевой вклад в реальную энтропию источника, и "весомый" вклад в его максимальную энтропию (в предположении, что все символы равновероятны). Поэтому, если источник выдает сырой (несжатый) текст, то его максимальная удельная энтропия в расчете на 1 символ всегда равна логарифму из объема используемого алфавита. В общем случае максимальная удельная энтропия (в т.ч. и для сжатых сообщений) = средней

длине сообщения в битах, деленной на число символов в сообщении.

Вывод: не важно, можем мы или не можем расчитать реальную энтропию ("чистую информацию") источника или конкретного сообщения, главное в классической формуле избыточности то, что она показывает основные тенденции:

классический вывод: при заданном алфавите источника (Hmax = const) его избыточность тем больше, чем больше статистических связей \ ограничений в выходных сообщениях (т.е. чем меньше реальное значение H)

очевидный вывод, на котором в класс.теории не заостряют внимание: если взять сообщения с фикс.реальной энтропией H, то избыточность источника будет тем выше, чем больше (в среднем) бит он использует для кодирования \ представления одного символа. Например (в десятый раз повторяю ;), если взять простой русский текст без учета регистра, то Hmax для юникода = 16 бит/симв, для анси - 8, для Бодо - ~5, для Хаффмана - 4.4 бит/симв. Соотв-но не важно как мы считаем реальное значение H (посимвольно или более сложно) - очевидно, что чем больше Hmax, тем выше избыточность источника
---
leo #109 19-02-2009 10:42:11

> мне не кажется, что из этого следует явное разделение объекта и представления

А мне "кажется", т.к. за понятием "источник сообщения" скрывается именно не конкретный "объект", а "представление" этого объекта. Например, для системы связи (или для архиватора) источником может быть телеграфный аппарат, выдающий символы в кодах Бодо, а может быть виндовый блокнот, сохраняющий файлы либо в анси, либо в юникоде, или Adobe Acrobat и т.д. и т.п. Текст м.б. одним и тем же, а его "представление" и соотв-но избыточность - разная.

А твое заявление в #31

> Юникод, коды Бодо и неиспользуемые символы к делу не относятся - очевидно подразумевается, что сообщение реализует всю мощность заданного алфавита

означает страусиный отход от реальности, т.к. ты пытаешься "вынести за скобки" реальную избыточность разных источников сообщений. Но реальные архиваторы и системы связи должны работать с любыми данными, которые им "подсовывают", и не капризничать, требуя "реализации всей мощности заданного алфавита" :D
---
gazlan #110 19-02-2009 11:10:05

> за понятием "источник сообщения" скрывается именно не конкретный "объект", а "представление" этого объекта

Угу. Потому вам и "кажется". Реально, информация почти целиком определяется моделью. Само сообщение (его биты и их логарифмы) это маленькая часть пазла.

Помните историю, как четверо слепых ощупывали слона? Невозможно восприятие вне парадигмы. Если вы прошли по линку в моем предыдущем (#107) ответе, то могли видеть, что сообщение "поразил дракона" совершенно по-разному интерпретировано в двух разных, хотя и очень близких моделях. И ориентироваться

на "логарифм объема используемого алфавита" попросту неверно.

> пытаешься "вынести за скобки" реальную избыточность разных источников сообщений

Нет, мы уже договорились, что избыточность спрятана именно в модели.
---
leo #111 19-02-2009 12:12:31

Мда, ты похоже безнадежно болен...

Источник существует сам по себе (например, txt-файл на диске), а система передачи или архиватор сами по себе. Пока архиватор не проанализирует файл, для него существует только Hmax = размеру файла в битах. Может оказаться, что в этом txt содержится хорошо зашифрованный текст или вообще случайный набор символов и соотв-но его избыточность будет близка к 0 и следовательно его не удасться существенно сжать и проще оставить как есть.

А может, наоборот он содержит кучу повторяющихся символов и его можно значительно сжать элементарным RLE, или еще лучше LZW (который, к слову сказать, юникодный текст сожмет больше ансишного -> зависимость не только от "содержания", но и от "кодировки"). Т.е. конкретный текст (или источник, который его выдал) может содержать некоторую "теоретическую избыточность" по Шеннону, но упрощенный алгоритм RLE может устранить только часть этой избыточности, заложенную в его модели, алгоритм LZW - свою часть, бОльшую чем RLE, а некий идеальный \ абстрактный алгоритм - практически всю "теоретическую" избыточность.

Другими словами есть теоретическая \ потенциальная избыточность, присущая данному источнику (расчитывается по Шеннону), а есть избыточность, реально устраняемая конкретным архиватором, заложенная в его модели и расчитывается она в шенновских терминах вообще элементарно - как единица минус отношение длины сжатого сообщения к длине исходного в битах. Причем если используются динамические алгоритмы сжатия, то архиватор устраняя часть избыточности сообщения, вынужден добавить в него свою "избыточность" = "информацию" для декодера.

Т.е. с точки зрения декодера это служебная информация (заранее неизвестная дельта к его модели), но с точки зрения передачи сообщения по каналу или его хранения на диске - это избыточность, т.к. теоретически возможны другие, статические методы сжатия, которые не будут добавлять к сообщению этих "лишних" битов
---
gazlan #112 19-02-2009 12:24:18

> Мда, ты похоже безнадежно болен

Диагноз по фотографии?

Со всем, изложенным ниже я согласен - и не могу вспомнить, чтобы когда-либо возражал против этого. В использованных мною ранее определениях это выглядит как "модель + синхросигнал". Поясните, пожалуйста: в чем вы видите несогласованность в наших взглядах?

Y_Mur #113 19-02-2009 16:43:40

> Реально, информация почти целиком определяется моделью. Само сообщение (его биты и их логарифмы) это маленькая часть пазла. Помните историю, как четверо слепых ощупывали слона? Невозможно восприятие вне парадигмы.

Развивая эту мысль можно сказать - один человек прочитав книгу из #102 поймёт её всю, другой только 25%, третий 5%, четвёртый не поймёт в ней ничего - вывод - книгу вообще нельзя считать самостоятельным сообщением и пытаться как-то измерить количество информации в ней, соответсвенно рассуждения об избыточности в ней из #107, #108 совершенно беспочвенны...

Не могу согласиться с такой постановкой вопроса - имхо всё таки содержимое книги должно иметь более менее объективную оценку количества заключённой в неё информации.
---
gazlan #114 20-02-2009 07:30:46

> Не могу согласиться с такой постановкой вопроса

Представьте, что эта книга даже не на незнакомом вам языке, а просто в нечитаемой кодировке. Ее "объективная" информативность не изменилась ни на бит, но в вашей модели это просто шум. В шифровании это еще прозрачнее. Есть ключ и алгоритм (модель) - сообщение может быть прочитано. Нет - это практически чистый шум (почти случайный текст).
---
gazlan #115 20-02-2009 08:17:59

Трансформаторы

Не раз уже упоминалось, что lossless сжатие суть переименование - биективное отображение одного множества кодов в другое. Для целей компрессии данных, разумеется, наиболее интересны искажающие отображения - равномерных кодов в неравномерные (и наоборот).

Отображения равномерных кодов в равномерные тривиально (Ex: KOI8R-Win1251), а неравномерных в неравномерные можно рассматривать просто как двойное (каскадное) преобразование. Останавливаться на очевидных (кольцо) групповых свойствах подстановок я не буду.

Желательна некоторая компактная (алгебраическая) форма записи такого отображения (из одного алфавита в другой).

Пусть имеются (равномощные) алфавиты A1 и A2 и некоторое отображение (которое будем называть Трансформацией T), взаимооднозначно связывающее элементы (Литеры) из A1 и A2, так что любому a1(i) сопоставлен a2(j) и наоборот.

Оператор, переводящий текст x записанный в алфавите A1 в текст y, записанный в алфавите A2 будем называть Трансформатором T(1,2).

y = Tx (*)

В силу естественной упорядоченности алфавитов, каждой литере можно сопоставить ее номер (последовательный индекс).

Предлагаю записывать Трансформатор T(1,2) как диагональную матрицу n*n (n - мощность алфавита), где на (i,i)-ом месте находится j - индекс подставляемой литеры второго алфавита.

В силу постулированной биективности трансформации, матрица T невырождена и имеет обратную.

Уравнение (*) будем называть Законом Ома для участка информационной цепи, а T - информационным импедансом.

Последовательному (каскадному) применению трансформаций соответствует перемножение импедансов (слева). Так, например, импеданс LZH может быть получен произведением импедансов LZ и H.
