Профессор
Игорь Н. Бекман

КОМПЬЮТЕРНЫЕ НАУКИ

Курс лекций

Лекция 3. ИНФОРМАЦИЯ

Содержание

Информатика – молодая научная дисциплина, изучающая вопросы, связанные с поиском, сбором, хранением, преобразованием и использованием информации в самых различных сферах человеческой деятельности. Генетически информатика связана с вычислительной техникой, компьютерными системами и сетями, так как именно компьютеры позволяют порождать, хранить и автоматически перерабатывать информацию в таких количествах, что научный подход к информационным процессам становится одновременно необходимым и возможным.

В данной лекции мы рассмотрим свойства некоторых видов информации, введём единицы информации и способы измерения количества информации. Заключительная часть лекции будет посвящена проблеме передачи сообщения от источника информации к приёмнику.

1. ИНФОРМАЦИЯ

Слово «информация» происходит от латинского – разъяснение, изложение, осведомленность. В течение многих веков понятие информации не раз претерпевало изменения, то расширяя, то предельно сужая свои границы. Сначала под этим словом понимали «представление», «понятие», затем – «сведения», «передачу сообщений». В XX в. бурное развитие получили всевозможные средства связи (телефон, телеграф, радио), назначение которых заключалось в передаче сообщений. Эксплуатация их выдвинула ряд проблем: как обеспечить надежность связи при наличии помех, какой способ кодирования сообщения применять в том или ином случае, как закодировать сообщение, чтобы при минимальной его длине обеспечить передачу смысла с определенной степенью надежности. Эти проблемы требовали разработки теории передачи сообщений, т.е. теории информации. Одним из основных вопросов этой теории был вопрос о возможности измерения количества информации.

Попытки количественного измерения информации предпринимались неоднократно. Первые отчетливые предложения об общих способах измерения количества информации были сделаны **Р.Фишером** (1921) в процессе решения вопросов математической статистики. Проблемами хранения информации, передачи её по каналам связи и задачами определения количества информации занимались **Р.Хартли** (1928) и **Х.Найквист** (1924). Р. Хартли заложил основы теории информации, определив меру количества информации для некоторых задач. Наиболее убедительно эти вопросы были разработаны и обобщены американским инженером **Клодом Шенноном** в 1948. С этого времени началось интенсивное развитие теории информации вообще и углубленное исследование вопроса об измерении ее количества в частности.

Что такое информация – науке не известно. К настоящему времени дано более 400 определений этого термина, что лишний раз доказывает беспомощность современных учёных и философов. Информации нет, но наука Информатика есть. Как показано в нашем курсе лекций ИНФОРМАТИКА (http://profbeckman.narod/ru - лекции), из информационного моря можно выделить пять видов информации, которые худо-бедно удалось как-то определить и даже ввести некоторую меру. Это - физическая (в двух ипостасях – термодинамическая и статистическая), техническая, смысловая, алгоритмическая и квантовая информации.

Есть ли между ними что-то общее, или это – совершенно разные сущности – узнаем лет через двести.

В данном курсе лекций нас будут интересовать два вида информации: техническая (в её компьютерном приложении) и смысловая (в плане баз данных, баз знаний и поиска в Интернет).

Наилучшее определение информации дали философы: понятие «информация» - первичное и ***неопределяемое*** понятие (как, например «точка» в геометрии, «множество» в математике).

Хорошо сказал по этому поводу и основатель кибернетики Норберт Винер: «*Информация – это информация, не вещество и не энергия, и этого достаточно!*»

Сводная энциклопедия Википедия выдаёт по этому поводу следующее:

Информация (от лат. *informatio* - осведомление, разъяснение, изложение)
- сведения (сообщения, данные) независимо от формы их представления.
- абстрактное понятие, имеющее множество значений, в зависимости от контекста.

(Что правда, то правда! Именно что в зависимости от контекста и именно абстрактное).

Согласно официальным источникам:

По российскому ГОСТу 7.0-99:

Информация - сведения, воспринимаемые человеком и (или) специальными устройствами как отражение фактов материального или духовного мира в процессе коммуникации.

По российскому федеральному закону от 27 июля 2006 года № 149-ФЗ «Об информации, информационных технологиях и о защите информации» (Статья 2):

Информация - сведения (сообщения, данные) независимо от формы их представления.

Главный научный центр СССР в области научной информации - ВИНИТИ в своё время дал такое определение:

Информация - объективное содержание связи между взаимодействующими материальными объектами, проявляющееся в изменении состояний этих объектов.

Информация - сведения, передаваемые людьми устным, письменным или другим способом (с помощью условных сигналов, технических средств и т. д.); 20 в. общенаучное понятие, включающее обмен сведениями между людьми, человеком и автоматом, автоматом и автоматом; обмен сигналами в животном и растительном мире; передачу признаков от клетки к клетке, от организма к организму.

Техническая (кибернетическая, компьютерная) **информация** – характеристика управляющего сигнала, передаваемого по линии связи. Это та часть знаний, которая используется для ориентирования, активного действия, управления, т.е. в целях сохранения, совершенствования, развития системы. Информация уменьшает общую неопределённость (неизвестность); она - мера устранения неопределенности в системе.

Информация - величина уменьшаемой неопределенности в результате получения сообщения, она же - мера разнообразия в объектах и процессах и т. д.

В данном курсе лекций мы под технической информацией будем понимать сообщения, передаваемые в форме знаков или сигналов. Применительно ***к компьютерной обработке данных*** под информацией понимаем некоторую последовательность символических обозначений (букв, цифр, закодированных графических образов и звуков и т.п.), несущую смысловую нагрузку и представленную в понятном компьютеру виде. Каждый новый символ в такой последовательности символов увеличивает информационный объём сообщения.

Смысловая информация – знания, которые получает человек из различных источников с помощью органов чувств или технических средств. Это - сведения, содержащиеся в данном сообщении и рассматриваемые как объект передачи, хранения и обработки; причём это данные, организованные таким образом, что имеют смысл для имеющего с ними дело человека».

Более-менее обобщающее сказанное определение информации имеет вид:

Информация *– 1) что-то сказанное, новости; знание, полученное любым способом; 2) в информационной теории и теории компьютеров: это точная мера информации, измеренная в битах и охватывающая диапазон от нуля (это когда все известно заранее) и до какого-то максимального значения, когда ничего заранее о содержании сообщения не известно; 3) любые данные, хранящиеся в компьютере.*

Таким образом:

Техническая информация (она же кибернетическая или компьютерная) – информация, передаваемая азбукой Морзе, по радио- или телеканалу, информация в компьютерах и прочих технических машинах. Мы рассмотрим вопросы передачи информации по линии связи, вопросы кодирования-декодирования информации, и способы переработки информации компьютерами. А истинна эта информация или ложна, ценна или бесполезна, нас ни с какого бока интересовать не будет. Не интересно нам здесь также материальна она или нет. Важно сколько раз я должен ударить по телеграфному ключу, чтобы передать азбукой Морзе ваше сообщение жене о полярной зимовке, и насколько разнообразен этот текст (можно вспотеть, сутками передавая одну букву, но информации в этом будет немного).

Смысловая (семантическая и прагматическая) **информация**: информация, которую воспринимает человек (и которая, к примеру, передаётся средствами массовой информации). Это то, что можно осмыслить, оценить, купить-продать-подарить, накапливать, хранить, охранять, терять; она способна исчезать и появляться, может быть полезной и вредной, истинной и ложной, переходя в дезинформацию. Мы рассмотрим методы её поиска в Интернете, способы создания банка данных, базы данных и банка знаний.

2. ЕДИНИЦЫ ИНФОРМАЦИИ

В отличие от термодинамической, техническая информация безразмерна, тем не менее, для её количественного описания существуют специальные единицы.

В компьютерной технике измерение информации касается объёма компьютерной памяти и объёма данных, передаваемых по цифровым каналам связи. При этом, если ячейка памяти способна, в зависимости от внешнего воздействия, принимать одно из двух состояний, которые условно обозначаются обычно как «0» и «1», она обладает минимальной информационной ёмкостью.

За единицу количества информации принимают такое количество информации, при котором неопределённость уменьшается в два раза. Такая единица названа **бит**.

Информационный объём сообщения - количество двоичных символов, используемое для кодирования этого сообщения.

Информационная ёмкость одной ячейки памяти компьютера, способной находиться в двух различных состояниях, принята за единицу измерения количества информации - 1 **бит**.

В технике под количеством информации понимают количество кодируемых, передаваемых или хранимых символов.

Бит (*binary digit – двоичная цифра, двоичное число; также игра слов: bit - немного*) (один двоичный разряд в двоичной системе счисления) - одна из самых известных единиц измерения информации.

Бит (bit) - двоичный знак двоичного алфавита {0, 1}.

Бит – термин, обозначающий наименьшую единицу информации, с которой может оперировать вычислительная машина

Бит - минимальная единица измерения информации, минимальная передаваемая единица информации. Сочетания битов могут указывать букву, число, передавать сигнал, выполнять переключение или другие функции.

Бит - единица измерения информационной ёмкости и количества информации, а также информационной энтропии.

Бит - информация, содержащаяся в одном дискретном сообщении источника равновероятных сообщений с объемом алфавита равного двум.

Бит – двоичный логарифм вероятности равновероятных событий или сумма произведений вероятности на двоичный логарифм вероятности при равновероятных событиях.

Бит – единица информации, представляющая собой такое количество, которое необходимо, чтобы сократить количество альтернатив в ситуации выбора на половину. Например, если вы имеете четыре альтернативы и получаете информацию, устраняющую две из них, вы получили один бит информации.

Бит – простое двоичное число (цифра или символ), принимающее значения 1 или 0 и служащее для записи и хранения данных в ЭВМ. Бит является минимальной двоичной единицей измерения энтропии и количества информации в ЭВМ, соответствующей одному двоичному разряду.

Бит – единица измерения количества информации, равная количеству информации, содержащемуся в опыте, имеющем два равновероятных исхода. Это тождественно количеству информации в ответе на вопрос, допускающий ответы «да» либо «нет» и никакого другого (то есть такое количество информации, которое позволяет однозначно ответить на поставленный вопрос).

Если подбросить монету и проследить, какой стороной она упадёт, то мы получим определенную информацию. Обе стороны монеты «равноправны», поэтому одинаково вероятно, что выпадет как одна, так и другая сторона. В таких случаях говорят, что событие несёт информацию в 1 бит. Если положить в мешок два шарика разного цвета, то, вытащив вслепую один шар, мы также получим информацию о цвете шара в 1 бит. Бит одна из самых безусловных единиц измерения. Если единицу измерения длины можно было положить произвольной: локоть, фут, метр, то единица измерения информации не могла быть по сути никакой другой. На физическом уровне бит является ячейкой памяти, которая в каждый момент времени находится в одном из двух состояний: «0» или «1».

Если каждая точка некоторого изображения может быть только либо чёрной, либо белой, такое изображение называют битовым, потому что каждая точка представляет собой ячейку памяти емкостью 1 бит. Количество информации равное 1 биту можно получить в ответе на вопрос типа «да»/ «нет». Если изначально вариантов ответов было больше двух, количество получаемой в конкретном ответе информации будет больше, чем 1 бит, если вариантов ответов меньше двух, т.е. один, то это не вопрос, а утверждение, следовательно, получения информации не требуется, раз неопределенности нет. Информационная ёмкость

ячейки памяти, способной воспринимать информацию, не может быть меньше 1 бита, но количество получаемой информации может быть и меньше, чем 1 бит. Это происходит тогда, когда варианты ответов «да» и «нет» неравновероятны. Неравновероятность в свою очередь является следствием того, что некоторая предварительная (априорная) информация по этому вопросу уже имеется, полученная, допустим, на основании предыдущего жизненного опыта.

В зависимости от точек зрения, бит может определяться следующими способами:

1. Один разряд двоичного кода (двоичная цифра). Может принимать только два взаимоисключающих значения: да/нет, 1/0, включено/выключено, и т.п. В электронике 1 биту соответствует 1 двоичный триггер.
2. 2. Двоичный логарифм вероятности равновероятных событий или сумма произведений вероятности на двоичный логарифм вероятности при равновероятных событиях.
3. Базовая единица измерения количества информации, равная количеству информации, содержащемуся в опыте, имеющем два равновероятных исхода. Это тождественно количеству информации в ответе на вопрос, допускающий ответы «да» либо «нет» и никакого другого (т. е. такое количество информации, которое позволяет однозначно ответить на поставленный вопрос).

Бит как единица информации не привязан конкретно к виду информационного пакета. Поэтому можно содержание информации определять не абсолютно, а относительно каких либо конкретных информационных пакетов. Например, 10 коров и 10 домов могут содержать одинаковое количество информации с точки зрения субъекта. Хотя внутри этих информационных пакетов скрыто от наблюдателя огромное количество пока не нужной информации.

В вычислительной технике данных обычно значения 0 и 1 передаются различными уровнями напряжения либо тока. В вычислительной технике, особенно в документации и стандартах, слово «бит» часто применяется в значении «двоичный разряд». В компьютерной технике бит соответствует физическому состоянию носителя информации: намагничено - не намагничено, есть отверстие - нет отверстия. При этом одно состояние принято обозначать цифрой 0, а другое - цифрой 1. Выбор одного из двух возможных вариантов позволяет также различать логические истину и ложь. Последовательностью битов можно закодировать текст, изображение, звук или какую-либо другую информацию. Такой метод представления информации называется **двоичным кодированием** (*binary encoding*).

Бит - слишком мелкая единица измерения. Существует более крупная единица – **байт** (*byte*), вообще говоря равная произвольному числу битов, но в компьютерной практике обычно равная восьми битам, т.к. именно восемь битов требуется для того, чтобы закодировать любой из 256 символов алфавита клавиатуры компьютера ($256=2^8$). При кодировании каждому символу соответствует своя последовательность из восьми нулей и единиц, т. е. **байт.** Соответствие байтов и символов задается с помощью таблицы, в которой для каждого кода указывается свой символ. Если бит позволяет выбрать один вариант из двух возможных, то байт = 1 из 256 (2^8).

Определенное количество бит составляет размер других единиц – двоичных слов, в том числе, байта, килобайта, мегабайта и т.д.

Байт *(byte)* – это двоичное слово, способное записывать и хранить в памяти ЭВМ один буквенно-цифровой или другой символ данных. Каждый символ записывается в виде набора двоичных цифр (битов) при помощи определенного кода, например *ASC*II.

Байт – единица количества информации, обычно состоящая из 8 бит и используемая как одно целое при передаче, хранении и переработки информации компьютером. Байт служит для представления букв или специальных символов (занимающих обычно весь байт). Информация в компьютере обрабатывается отдельными байтами, либо группами байтов (полями, словами).

Байт - в запоминающих устройствах - наименьшая адресуемая единица данных в памяти компьютера, обрабатываемая как единое целое. По умолчанию байт считается равным 8 битам. Обычно в системах кодирования данных байт представляет собой код одного печатного или управляющего символа.

Байт - в измерении информации - единица измерения количества информации, объема памяти и ёмкости запоминающего устройства.

Байт - единица количества информации. Для конкретного компьютера байт - минимальный шаг адресации памяти. В стандартном виде байт равен восьми битам (может принимать 256 (2^8) различных значений). Байт в современных компьютерах - минимально адресуемая последовательность фиксированного числа битов. При хранении данных в памяти существует также бит чтения-записи, а для цифровых микросхем - бит синхронизации. Иногда байтом называют последовательность битов, которые составляют подполе машинного слова, используемое для кодирования одного текстового символа (хотя правильней это называть символом, а не байтом).

Байт определяется как минимальный шаг адресации памяти, который на старых машинах не обязательно был равен 8 битам (не у всех компьютеров память состоит из битов, пример: троичный

компьютер). Сейчас байт считают равным восьми битам. В таких обозначениях как Кбайт (русское) или *KB* (английское) под байт (*B*) подразумевается именно 8 бит, хотя сам термин «байт» не вполне корректен с точки зрения теории.

Далее в лекциях:

1 байт = 8 битам

В компьютере информация представляется в виде последовательности из нулей и единиц (двоичное кодирование). Цифры 0 и 1 можно рассматривать как два равновероятных события, а один двоичный разряд содержит количество информации, равное 1 биту. Два двоичных разряда несут соответственно 2 бита информации. Информационный объём сообщения - количество двоичных символов, используемое для кодирования этого сообщения. Каждому символу в компьютере соответствует последовательность из 8 нулей и единиц, называемая байтом: 1 байт = 8 битам. Например, слово МИР в компьютере выглядит следующим образом: *{М}*11101101 *{И}*11101001 *{Р}*11110010. Последовательностью нулей и единиц можно закодировать и графическую информацию, разбив изображение на точки. Если только чёрные и белые точки, то каждую можно закодировать 1 битом.

Количество бит в байте определяет его разрядность, которая может составлять 8, 16, 32 и т.д. тогда байт называют 8-разрядным, 16-разрядным и т.д. Один 8-разрядный байт может определять 256 разных значений, например десятичных чисел от 0 до 256. Увеличение разрядности ведёт к соответствующему увеличению числа возможных вариантов комбинаций, кодируемых одним байтом. Например, 16-разрядным - до 65536 или 216, 32-разрядным - до 232 и т.д.

Широко используются ещё более крупные производные единицы информации:

- 1 Килобайт (Кбайт) = 1024 байт = 2^{10} байт,
- 1 Мегабайт (Мбайт) = 1024 Кбайт = 2^{20} байт,
- 1 Гигабайт (Гбайт) = 1024 Мбайт = 2^{30} байт.
- 1 Терабайт (Тбайт) = 1024 Гбайт = 2^{40} байт,
- 1 Петабайт (Пбайт) = 1024 Тбайт = 2^{50} байт.

Табл. 1. Единицы информации.

Измерения в байтах						
Десятичная приставка			Двоичная приставка			
Название	Символ	Степень	Название	Символ		Степень
				МЭК	ГОСТ	
байт	B	10^0	байт	B	байт	2^0
килобайт	kB	10^3	кибибайт	KiB	Кбайт	2^{10}
мегабайт	MB	10^6	мебибайт	MiB	Мбайт	2^{20}
гигабайт	GB	10^9	гибибайт	GiB	Гбайт	2^{30}
терабайт	TB	10^{12}	тебибайт	TiB	Тбайт	2^{40}
петабайт	PB	10^{15}	пебибайт	PiB	Пбайт	2^{50}
эксабайт	EB	10^{18}	эксбибайт	EiB	Эбайт	260
зеттабайт	ZB	10^{21}	зебибайт	ZiB	Збайт	2^{70}
йоттабайт	YB	10^{24}	йобибайт	YiB	Йбайт	2^{80}

Килобайт, Кбайт (kilobyte) – единица измерения ёмкости памяти или длины записи, равная 1024 байтам. Часто под килобайтом понимается также величина, равная 10^3 байт.

Мегабайт, Мбайт (megabyte) – единица измерения ёмкости памяти или длины записи, равная 1024 Кбайт. Часто под мегабайтом понимается также величина, равная 10^3 килобайт или 106 байт.

Гигабайт, Гбайт (gigabyte) – единица измерения ёмкости памяти или длины записи, равная 1024 Мбайт. Часто под гигабайтом понимается также величина, равная 10^3 мегабайт, 10^6 килобайт или 10^9 байт.

Терабайт, Тбайт (terabyte) - единица измерения ёмкости памяти или длины записи, равная 1024 Гбайт. Часто под терабайтом понимается также величина, равная 10^3 гигабайт, 10^6 мегабайт, 10^9 килобайт или 1012 байт.

Кратные приставки для образования производных единиц для байта применяются не как обычно: 1) уменьшительные приставки не используются совсем, а единицы измерения информации меньшие, чем байт, называются специальными словами (нибл и бит); 2) увеличительные приставки означают за каждую тысячу $1024=2^{10}$ (килобайт равен 1024 байтам, мегабайт равен 1024 килобайтам, или 1048576 байтам; и т.д. с гига-, тера- и пентабайтами).

Целые количества бит отвечают количеству состояний, равному степеням двойки. Особое название имеет 4 бита - нибл (полубайт, тетрада, четыре двоичных разряда), который вмещают в себя количество информации, содержащейся в одной шестнадцатеричной цифре.

Именно к байту (а не к биту) непосредственно приводятся все большие объёмы информации, исчисляемые в компьютерных технологиях. Для измерения больших количеств байтов служат единицы «килобайт» = 1000 байт и «Кбайт» (кибибайт) = 1024 байт. Единицы «мегабайт» = 1000 килобайт = 1000000 байт и «Мбайт» (мебибайт) = 1024 Кбайт = 1048576 байт применяются для измерения объёмов носителей информации. Единицы «гигабайт» = 1000 мегабайт = 1000000000 байт и «Гбайт» (гибибайт) = 1024 Мбайт = 230 байт измеряют объём больших носителей информации, например жёстких дисков. Для исчисления ещё больших объёмов информации имеются единицы терабайт-тебибайт (10^{12} и 2^{40} соответственно), петабайт-пебибайт (10^{15} и 2^{50} соответственно) и т. д.

Смысл единицы информации: если задавать вопросы «да» или «нет», то число вопросов будет точно соответствовать неопределённости задачи в битах. Можно закодировать эти вопросы нулём («нет») и единицей («да»), тогда любая последовательность содержательных ответов будет представлена кодовой последовательностью определённой длины. Длина этой последовательности равна количеству информации, содержащейся в ответах. Особенно это полезно, когда последовательность вопросов стандартна, а отвечают на них многие. Например, ответ на референдум «да, да, нет, да» прдставляется кодом 1101.

Российский ГОСТ 8.417-2002 («Единицы величин») для обозначения байта регламентирует использование русской заглавной буквы «Б». Кроме того, констатируется традиция использования приставок СИ вместе с наименованием «байт» для указания двоичных множителей (1 Кбайт = 1024 байт, 1 Мбайт = 1024 Кбайт, 1 Гбайт = 1024 Мбайт и т. д.), причём используется прописная «К» вместо строчной «к», обозначающей множитель 10^3. Использование прописной буквы «Б» для обозначения байта соответствует требованиям ГОСТ и позволяет избежать путаницы между сокращениями от байт и бит. Важно, что в стандарте нет сокращения для «бит», поэтому использование записи вроде «Гб» как синонима для «Гбит» недопустимо.

Долгое время разнице между множителями 1000 и 1024 не придавали большого значения. Во избежание недоразумений следует чётко понимать различие между: двоичными кратными единицами, обозначаемыми согласно ГОСТ 8.417-2002 как «Кбайт», «Мбайт», «Гбайт» и т. д. (два в степенях кратных десяти); единицами килобайт, мегабайт, гигабайт и т. д., понимаемыми как научные термины (десять в степенях кратных трём). Последние равны соответственно 10^3, 10^6, 10^9 байт.

Единицы измерения информации зависят от основания логарифма. В случае логарифма с основанием 2 единицей измерения является **бит**, если используется натуральный логарифм - то **нат**, если десятичный - то **хартли**.

Основание логарифма	Единица измерения	Количество информации о падении монеты «орлом» вверх
2	бит	$-\log_2(1/2) = \log_2 2 = 1$ бит
e	нат	$-\ln(1/2)=\ln 2 \approx 0{,}69$ ната
10	хартли	$-\log_{10}(1/2)=\log_{10} 2 \approx 0{,}30$ хартли

100 Мб это много или мало? 100 Мб могут вместить:

страниц текста	50 000 или 150 романов
цветных слайдов высочайшего качества	150
аудиозапись речи видного политического деятеля	1.5 часа
музыкальный фрагмент качества CD -стерео	10 минут
фильм высокого качества записи	15 секунд
протоколы операций с банковским счётом	за 1000 лет

3. КОЛИЧЕСТВО ИНФОРМАЦИИ

Информационный объём сообщения (информационная ёмкость сообщения) - количество информации в сообщении, измеренное в битах, байтах или производных единицах (Кбайтах, Мбайтах и т.д.).

Количество информации - мера уменьшения неопределенности.

Количество технической информации - числовая характеристика сигнала, которая не зависит от его формы и содержания и характеризует неопределенность, которая исчезает после получения сообщения в виде данного сигнала. Оно зависит от вероятности получения сообщения о том или ином событии. Для абсолютно достоверного события (событие обязательно произойдет, поэтому его вероятность равна 1) количество информации в сообщении о нём равно 0. Чем невероятнее событие, тем большее количество информации несёт сообщение о нём. Лишь при равновероятных ответах ответ «да» или «нет» несёт один бит информации.

Понятие количества информации возникает в следующих типовых случаях:

1. Равенство вещественных переменных $a=b$, заключает в себе информацию о том, что а равно b. Про равенство $a^2=b^2$ можно сказать, что оно несёт меньшую информацию, чем первое, т.к. из первого следует второе, но не наоборот. Равенство $a^3=b^3$ несёт в себе информацию по объему такую же, как и первое.
2. Пусть происходят некоторые измерения с некоторой погрешностью. Тогда чем больше будет проведено измерений, тем больше информации об измеряемой сущности будет получено.
3. Математическое ожидание некоторой случайной величины, содержит в себе информацию о самой случайной величине. Для случайной величины, распределенной по нормальному закону, с известной дисперсией знание математического ожидания даёт полную информацию о случайной величине.
4. Рассмотрим схему передачи информации. Пусть передатчик описывается случайной величиной, X, тогда из-за помех в канале связи на приёмник будет приходить случайная величина, $Y=X+Z$, где Z - это случайная величина, описывающая помехи. В этой схеме можно говорить о количестве информации, содержащейся в случайной величине, Y, относительно X. Чем ниже уровень помех (дисперсия Z мала), тем больше информации можно получить из Y. При отсутствии помех Y содержит в себе всю информацию об X.

Мерой количества информации, связанной с тем или иным объектом или явлением, может служить редкость его встречаемости или сложность его структуры.

В компьютерной технике измерению обычно подвергается информация, представленная дискретным сигналом. При этом различают следующие подходы:

1. **Структурный (алфавитный, объёмный).** Измеряет количество информации простым подсчётом информационных элементов, составляющих сообщение. Применяется для оценки возможностей запоминающих устройств, объёмов передаваемых сообщений, инструментов кодирования без учета статистических характеристик их эксплуатации. Алфавитный подход к измерению информации не связывает количество информации с содержанием сообщения. Это - объективный подход к измерению информации. Количество информации зависит от объёма текста и мощности алфавита. Ограничений на максимальную мощность алфавита нет, но есть достаточный алфавит мощностью 256 символов. Этот алфавит используется для представления текстов в компьютере. Поскольку $256=2^8$, то один символ несёт в тексте 8 бит информации.
2. **Статистический (вероятностный).** Учитывает вероятность появления сообщений: более информативным считается то сообщение, которое менее вероятно, т.е. менее всего ожидалось. Применяется при оценке значимости получаемой информации. Все события происходят с различной вероятностью, но зависимость между вероятностью событий и количеством информации, полученной при совершении того или иного события можно выразить формулой Шеннона.
3. **Семантический (содержательный).** Учитывает целесообразность и полезность информации. Применяется при оценке эффективности получаемой информации и её соответствия реальности. Сообщение – информативный поток, который в процессе передачи информации поступает к приемнику. Сообщение несёт информацию для человека, если содержащиеся в нем сведения являются для него новыми и понятными. Информация - знания человека - сообщение должно быть информативно. Если сообщение не информативно, то количество информации с точки зрения человека = 0. (Пример: вузовский учебник по высшей математике содержит знания, но они не доступны первокласснику).

В этой лекции мы более подробно рассмотрим два первых подхода к измерению информации.

3.1 Алфавитный подход

В рамках алфавитного (структурного) подхода выделяют три меры информации:

- **Геометрическая.** Определяет максимально возможное количество информации в заданных объемах. Мера может быть использована для определения информационной ёмкости памяти компьютера;
- **Комбинаторная**. Оценивает возможность представления информации при помощи различных комбинаций информационных элементов в заданном объёме. Комбинаторная мера может использоваться для оценки информационных возможностей некоторого системы кодирования.
- **Аддитивная** (мера Хартли).

Геометрическая мера определяет максимально возможное количество информации в заданных объёмах. Единица измерения – информационный элемент. Мера может быть использована для определения информационной ёмкости памяти компьютера. В этом случае в качестве информационного элемента выступает минимальная единица хранения – бит.

Пример 1. Пусть сообщение, состоящее из 14 символов) 5555 6666 888888 закодировано методом кодирования повторений – и имеет вид: 5(4) 6(4) 8(6). Требуется измерить информацию в исходном и закодированном сообщениях геометрической мерой и оценить эффективность кодирования. В качестве информационного элемента зададимся символом сообщения. Тогда: $I^{(\text{исх.})} = n^{(\text{исх.})} = 14$ символов; $I^{(\text{закод.})} = n^{(\text{закод.})} = 12$ символов, где $I^{(\text{исх.})}$, $I^{(\text{закод.})}$ – количества информации, соответственно, в исходном и закодированном сообщениях; $n^{(\text{исх.})}$, $n^{(\text{закод.})}$ – длины (объёмы) тех же сообщений, соответственно. Эффект кодирования определяется как разница между $I^{(\text{исх.})}$ и $I^{(\text{закод.})}$ и составляет 2 символа.

Очевидно, геометрическая мера не учитывает, какими символами заполнено сообщение. Так, одинаковыми по количеству информации, измеренной геометрической мерой, являются, например, сообщения «компьютер» и «программа»; а также 346 и 10*B*.

Комбинаторная мера оценивает возможность представления информации при помощи различных комбинаций информационных элементов в заданном объёме. Использует типы комбинаций элементов и соответствующие математические соотношения, которые приводятся в одном из разделов дискретной математики – комбинаторике.

Комбинаторная мера может использоваться для оценки информационных возможностей некоторого автомата, который способен генерировать дискретные сигналы (сообщения) в соответствии с определенным правилом комбинаторики. Пусть, например, есть автомат, формирующий двузначные десятичные целые положительные числа (исходное множество информационных элементов {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}). В соответствии с положениями комбинаторики, данный автомат генерирует размещения (различаются числа, например, 34 и 43) из 10 элементов (используются 10 цифр) по 2 (по условию задачи, формируются двузначные числа) с повторениями (очевидно, возможны числа, состоящие из одинаковых цифр, например, 33). Тогда можно оценить, сколько различных сообщений (двузначных чисел) может сформировать автомат, иначе говоря, можно оценить информационную ёмкость данного устройства равна $10^2 = 100$.

Комбинаторная мера используется для определения возможностей кодирующих систем.

Пример 2. Определить емкость *ASC*II-кода, представленного в двоичной или шестнадцатеричной системе счисления. *ASC*II-код – это сообщение, которое формируется как размещение с повторениями: 1) для двоичного представления – из информационных элементов {0, 1}, сообщение длиной (объемом) 8 символов; 2) для шестнадцатеричного представления – из информационных элементов {0, 1, 2, …., *A*, *B*, *C*, …. *F*}, сообщение длиной (объемом) 2 символа. Тогда в соответствии с положениями комбинаторики: $I^{(\text{двоичное})} = 2^8 = 256$; $I^{(\text{шестнадцатеричное})} = 16^2 = 256$, где $I^{(\text{двоичное})}$,

$I^{(\text{шестнадцатеричное})}$ – количества информации, соответственно, для двоичного и шестнадцатеричного представления *ASC*II-кода. Таким образом, емкость *ASC*II-кода для двоичного и шестнадцатеричного представления одинакова и равна 256.

Комбинаторная мера является развитием геометрической меры, т. к. помимо длины сообщения учитывает объём исходного алфавита и правила, по которым из его символов строятся сообщения. Ею измеряется информация не конкретного сообщения, а всего множества сообщений, которые могут быть получены.

Единицей измерения информации в комбинаторной мере является число комбинаций информационных элементов.

Аддитивная мера называется мерой Хартли. Хартли рассмотрел кодировку сообщения с помощью некоторого набора знаков (если для данного набора установлен порядок следования знаков, то он называется алфавитом). Самой сложной частью работы оказалось определение количества информации, содержащейся в каждом отдельном символе: остальная часть процедуры весьма проста.

Алфавит – вся совокупность символов, используемых в некотором языке для представления информации.

С целью анализа некоторого текста, введём обозначения: N - мощность алфавита, использованного для написания текста, т.е. число символов в алфавите (размер алфавита); n – число символов в сообщении (длина текста, например, одного слова); I –количество информации в сообщении; $i=I/n$ – информационный вес символа (количество информации в одном символе).

Если речь идёт не об анализе текста, передаваемого в сообщении, а об описании каких-либо событий, то приводимый ниже математический аппарат полностью сохраняется, если под используемыми обозначениями понимать: N – общее число возможных исходов; n - число возможных исходов интересующего нас события (здесь все возможные события считаются равновероятными), I – количество информации о всех случившихся событиях; i – количество информации в сообщении о том, что произошло одно из N событий.

Предположим, что какое-то событие имеет N_a равновероятных исходов. Пусть такие события наступают в передаваемом по каналам связи сообщении, которое представлено в виде набора некоторых смысловых элементов или символов (например, букв какого-то алфавита). Здесь событием является появление любого символа из алфавита. Пусть далее в языке, использованном для написания текста этого сообщения, общее количество (объём алфавита) смысловых символов (букв) равно N_a, а одно сообщение составлено из n элементов (если сообщение состоит из одного слова, то n – количество букв в слове, т.е. длина слова; если сообщение состоит из многих слов, то n – число знаков во всём тексте).

Вопрос: как измерить количество информации, которое может быть передано при помощи такого алфавита?

Необходимо посчитать число N возможных сообщений, которые могут быть переданы при помощи этого алфавита. Очевидно, что если сообщение формируется из одного символа, то $N=N_a$, если из двух, то $N=N_a \cdot N_a=N_a^2$. Если сообщение содержит n символов, то число возможных сообщений

$$N=N_a^n. \qquad (4)$$

Если сообщение – это текст, состоящий только из отдельных букв алфавита, причём вероятности появления конкретной буквы в тексте одинаковы, n=1 (текст – случайный набор букв), то $N=N_a$. Если события передаются последовательностью цифр некоторой разрядности, то, естественно, $N \neq N_a$, а n-разрядность числа.

Так, например, с помощью двухразрядного десятичного числа (n= 2, N_a = 10) можно записать $N = 10^2$ = 100 различных чисел от 0 до 99. В частности, при средней длине русского слов» n = 5 букв и алфавите в N_a = 32 буквы можно составить 33,5 млн. различных слов.

Казалось бы, искомая мера количества информации найдена. Её можно понимать как меру неопределенности исхода опыта, если под опытом подразумевать случайный выбор какого-либо сообщения из некоторого числа возможных. Мера эта, однако, не удобна, так как:

1) Не выполняется условие пропорциональности между длиной слова (длительностью сигнала) и количеством содержащейся в нём информации. Между тем удвоение времени передачи сообщений должно приводить к удвоению количества передаваемой информации. Для двух независимых источников сообщений (или алфавита) с N_1 и N_2 числом возможных сообщений, общее число возможных сообщений $N = N_1 \cdot N_2$, в то время как логичнее было бы считать, что количество информации, получаемое от двух независимых источников, должно быть не произведением, а суммой составляющих величин

2) При наличии алфавита, состоящего из одного символа, т.е. когда N_a = 1, возможно появление только этого символа. Следовательно, неопределенности в этом случае не существует, и появление этого символа не несёт никакой информации. Между тем, значение N при N_a = 1 не обращается в нуль.

Р. Хартли предложил в качестве меры количества информации использовать логарифм числа возможных сообщений

Выход из положения был найден Р. Хартли, который предложил информацию I, приходящуюся на одно сообщение, определять логарифмом по некоторому основанию a от общего числа возможных сообщений N:

$$I(N) = \log_a N \qquad (2)$$

Если же всё множество возможных сообщений состоит из одного ($N = N_a = 1$), то $I(N) = \log 1 = 0$, что соответствует отсутствию информации в этом случае. При наличии независимых источников информации с N_1 и N_2 числом возможных сообщений

$$I(N) = \log N = \log N_1 N_2 = \log N_1 + \log N_2, \qquad (3)$$

т.е. количество информации, приходящееся на одно сообщение, равно сумме количеств информации, которые были бы получены от двух независимых источников, взятых порознь. Формула, предложенная Хартли, удовлетворяет предъявленным требованиям. Поэтому её можно использовать для измерения количества информации.

Полная информация, содержащаяся в сообщении, определяется как количество сведений (при их длине *n)* пропорциональное числу смысловых символов *N* ***формулой Хартли****:*

$$I = n\log_a N \qquad (4)$$

Согласно этому соотношению, количество информации в передаваемом сообщении пропорционально его длительности (числу символов). Выбор основания логарифма a влияет только на размерность, т.е. на единицу измерения количества информации. Наиболее удобным оказалось основание логарифма $a = 2$.

Хартли впервые ввёл специальное обозначение для количества информации – I и предложил следующую логарифмическую зависимость между количеством информации и мощностью исходного алфавита (для равновероятных событий, а у Хартли они именно равновероятные, $N=N_a$):

$$I = n\log_a N = n \cdot i \qquad (5)$$

т.е.

$$i = \frac{I}{n} = \log_a N \qquad (5а)$$

Количество информации, содержащееся в одном элементе сигнала, называют *удельной информативностью* или *энтропией* сигнала:

$$i = H = \frac{I}{n} = \log_a N \qquad (6)$$

По существу энтропия есть мера неопределенности или мера недостающей информации исследуемого процесса (сообщения). В частности, энтропия русского алфавита (32 знака) равна $H = 5$ бит/символ.

$$N = a^{\frac{I}{n}} = a^i \qquad (7)$$

где a – основание логарифма. Если алфавит построен на двоичной системе {0,1}, т.е. a=2, то

$$N = 2^i, \qquad (8)$$

и **формула Хартли** приобретает вид

$$I = n\log_2 N \qquad (9)$$

При $n = 1$; $N = 2$ и основании логарифма, равном a=2, имеем $I = 1 \cdot \log_2 2 = 1$ - аналитическое определение бита по Хартли: это количество информации, которое содержится в двоичной цифре. Единицей измерения информации в аддитивной мере является бит.

Если хотят отразить подход Хартли на языке тории вероятности, то вводят понятие вероятности реализации одного из N событий, (по Хартли они имеют равновероятный исход, $N_a=N$), $p = \frac{1}{N}$. Тогда $N = \frac{1}{p}$ и $p = \frac{n}{N}$, где n – число возможных исходов интересующего нас события и

$$i = \log_2 N = log_2 N_a = \log_2(1/p) = -\log_2 p. \qquad (10)$$

т.е. количество информации на каждый равновероятный сигнал равно отрицательному логарифму от вероятности отдельного сигнала.

Из формулы

$$i = \log_2 \frac{1}{p} \qquad (11)$$

очевидно, что чем вероятнее событие, тем меньше информации оно несёт.

Полученная формула позволяет для некоторых случаев определить количество информации. Однако для практических целей необходимо задаться единицей его измерения. Для этого предположим, что информация – это устраненная неопределенность. Тогда в простейшем случае неопределенности выбор будет производиться между двумя взаимоисключающими друг друга равновероятными сообщениями, например между двумя качественными признаками: положительным и отрицательным импульсами, импульсом и паузой и т.п. Количество информации, переданное в этом простейшем случае, наиболее удобно принять за единицу количества информации. Именно такое количество информации может быть получено, если в формуле Хартли брать логарифм по основанию $a = 2$. Тогда

$$I = -\log_2 p = -\log_2(1/2) = \log_2 2 = 1. \qquad (12)$$

Полученная единица количества информации, представляющая собой выбор из двух равновероятных событий, получила название двоичной единицы, или бита.

За единицу количества информации в системах передачи дискретных и цифровых сообщений был принят один *бит* (*binary digit* - двоичная цифра) - двоичный разряд - символ, принимающий значение 0 или 1. Так, символы 101 есть 3-битовое число. Бит является не только единицей количества информации, но и единицей измерения степени неопределенности. При этом имеется в виду неопределенность, которая содержится в одном опыте, имеющем два равновероятных исхода.

Данная мера представления является универсальной и позволяет сравнить различные сообщения и количественно определить ценность различных источников информации, оценить величину её потерь при передаче, приёме, обработке, хранении, использовании и т. д. Как уже упоминалось, в цифровой технике (компьютерах) в качестве единицы представления данных используется *байт* (*byte* - слог) - слово (набор) из восьми двоичных разрядов (битов). Легко посчитать, что байтом можно передать одно из $2^8 = 256$ различных сообщений.

На количество информации, получаемой из сообщения, влияет фактор неожиданности его для получателя, который зависит от вероятности получения того или иного сообщения. Чем меньше эта вероятность, тем сообщение более неожиданно и, следовательно, более информативно. Сообщение, вероятность которого высока и, соответственно, низка степень неожиданности, несет немного информации.

Р.Хартли понимал, что сообщения имеют различную вероятность и, следовательно, неожиданность их появления для получателя неодинакова. Но, определяя количество информации, он пытался полностью исключить фактор «неожиданности». Поэтому формула Хартли позволяет определить количество информации в сообщении только для случая, когда появление символов равновероятно и они статистически независимы. На практике эти условия выполняются редко. При определении количества информации необходимо учитывать не только количество разнообразных сообщений, которые можно получить от источника, но и вероятность их получения. Для событий, вероятность наступления которых не одинакова, формулу для расчёта вероятности, предложил К.Шеннон уже не как аддитивную, а как вероятностную меру.

При выводе своей формулы Хартли предполагал, что буквы в тексте появляются с одинаковой вероятностью. Это - грубая модель, но зато очень простая. Если применять формулу Хартли не к тексту, а к событиям, то она применима исключительно для анализа равновероятных событий. Для событий, реализующихся с разной вероятностью, вычисление информации следует проводить по формуле Шеннона.

Пример 3. Рассчитать количество информации, которое содержится в шестнадцатеричном и двоичном представлении *ASC*II-кода для числа 1. В соответствии с таблицей *ASC*II-кодов имеем: шестнадцатеричное представление числа 1 – 31, двоичное представление числа 1 – 00110001. Тогда по формуле Хартли получаем: для шестнадцатеричного представления $I = 2\log_2 16 = 8$ бит; для двоичного представления $I = 8\ \log_2 2 = 8$ бит. Таким образом, разные представления *ASC*II-кода для одного символа содержат одинаковое количество информации, измеренной аддитивной мерой.

В целом алфавитный подход основан на определении количества информации в каждом из знаков дискретного сообщения с последующим подсчётом количества этих знаков в сообщении. В простейшем варианте он заключается *подсчёте числа символов в сообщении,* т. е. связан только с длиной сообщения и не учитывает его содержания. Длина сообщения зависит от числа знаков, употребляемых для записи сообщения. Например, слово «мир» в русском алфавите записывается тремя знаками, в английском - пятью *(peace),* а в КОИ -8 - двадцатью четырьмя битами (111011011110100111110010).

Пример 4.

Исходное сообщение		Количество информации		
на языке	в машинном представлении (КОИ - 8)	в символах	в битах	в байтах
рим	11110010 11101001 11101101	3	24	3
мир	11101101 11101001 11110010	3	24	3

миру мир!	11101101 11101001 11110010 11110101 00100000 11101101 1110101 11110010 00100001	9	72	9

Количество информации в техническом сообщении совпадает с количеством символов (нулей и единиц) в нём. Так, в слове «Рим» содержится 24 бита (3 байта) информации, а в «Миру мир!» – 72 бита (9 байтов).

***Количество информации**, которое содержит сообщение, закодированное с помощью знаков системы, равно количеству информации, которое несёт один знак, умноженному на число знаков в сообщении.*

Поскольку все символы «равноправны», естественно, что объём информации в каждом из них одинаков. Следовательно, остаётся полученное значение *I* умножить на количество символов в сообщении, и мы получим общий объём информации в нём. Осмысленность сообщения в описанной процедуре нигде не требуется, напротив, именно при отсутствии смысла предположение о равновероятном появлении всех символов выполняется лучше всего!

Описанный простой способ кодирования, когда коды всех символов имеют одинаковую длину, не является единственным. Часто при передаче или архивации информации по соображениям экономичности тем символам, которые встречаются чаще, ставятся в соответствие более короткие коды и наоборот. Можно показать, что при любом варианте кодирования (чем экономичнее способ кодирования, тем меньше разница между этими величинами).

Проиллюстрируем алфавитный подход на примере анализа русского текста.

Рассмотрим метод двоичного поиска на примере игры, использующей этот поиск.

Пусть требуется отгадать задуманное число из данного диапазона целых чисел. Игрок, отгадывающий число, задаёт вопросы, на которые можно ответить только «да» или «нет». Если каждый ответ отсекает половину вариантов (уменьшает выбор в два раза), то он несёт 1 бит информации. Тогда общее количество информации в (битах) полученной при угадывании числа, равно количеству заданных вопросов.

Пример 5. Требуется отгадать задуманное число из диапазона от 1 до 8.

1. Число меньше 5? Нет 1 бит
2. Число меньше 7? Да 1 бит
3. Это число 5? Нет 1 бит

8 возможных варианта – 3 вопроса – 3 бита информации

Всё множество используемых в языке символов будем традиционно называть алфавитом. Обычно под алфавитом понимают только буквы, но поскольку в тексте могут встречаться знаки препинания, цифры, скобки, то мы их тоже включим в алфавит. В алфавит также следует включить и пробел, т.е. пропуск между словами. Полное количество символов алфавита принято называть мощностью алфавита, *N*. Например, мощность алфавита из русских букв и отмеченных дополнительных символов равна 54. Представьте себе, что текст к вам поступает последовательно, по одному знаку, словно бумажная ленточка, выползающая из телеграфного аппарата. Предположим, что каждый появляющийся на ленте символ с одинаковой вероятностью может быть любым символом алфавита. В действительности это не так, но для упрощения примем такое предположение. В каждой очередной позиции текста может появиться любой из *N* символов. Тогда каждый такой символ несёт *I* бит информации, причём *I* - решение уравнения: $2^I = 54$. Получаем: $I = 5.755$ бит - столько информации несёт один символ в русском тексте! А теперь для того, чтобы найти количество информации во всем тексте, нужно посчитать число символов в нем и умножить на *I*. Посчитаем количество информации на одной странице книги. Пусть страница содержит 50 строк. В каждой строке - 60 символов. Значит, на странице умещается 50x60=3000 знаков. Тогда объем информации равен: $5{,}755 \cdot 3000 = 17265$ бит.

При алфавитном подходе к измерению информации количество информации зависит не от содержания, а от размера текста и мощности алфавита.

Применение алфавитного подхода удобно при использовании технических средств работы с информацией. Алфавитный подход является объективным способом измерения информации в отличие от субъективного содержательного подхода. Удобнее всего измерять информацию, когда размер алфавита *N* равен целой степени двойки. Например, если *N*=16, то каждый символ несёт 4 бита информации потому, что $2^4 = 16$. А если *N* =32, то один символ «весит» 5 бит. Ограничения на максимальный размер алфавита теоретически не существует. Однако есть алфавит, который можно назвать достаточным. При работе с компьютером - это алфавит мощностью 256 символов. В алфавит такого размера можно поместить все практически необходимые символы: латинские и русские буквы, цифры, знаки арифметических операций, всевозможные скобки, знаки препинания.... Поскольку $256 = 2^8$, то один символ этого алфавита «весит» 8 бит. Причем 8 бит информации - это настолько характерная величина, что ей даже присвоили свое название - *байт*.

Пример 6. При угадывании целого числа в диапазоне от 1 до *N* было получено 6 бит информации. Чему равно *N*? Здесь *i*=6, получаем $N=2^6=64$.

Пример 7. В корзине лежат 16 шаров разного цвета. Сколько информации несёт сообщение о том, что из корзины достали красный шар? Вытаскивание любого из 16 шаров – события равновероятные, поэтому *i*=4 бита.

Сегодня при подготовке писем, документов, статей, книг и пр. широко используются компьютерные текстовые редакторы. Если 1 символ алфавита несёт 1 байт информации, то следует сосчитать количество символов; полученное число даст информационный объем текста в байтах. Пусть небольшая книжка, сделанная с помощью компьютера, содержит 150 страниц; на каждой странице - 40 строк, в каждой строке - 60 символов. Значит, страница содержит 40·60=2400 байт информации. Объём всей информации в книге: 2400·150 = 360 000 байт.

Сообщение, уменьшающее неопределенность знаний в два раза, несет 1 бит информации.

Примеры 8:

1. Определить информацию, которую несёт в себе 1-й символ в кодировках *ASC*II и *Unicode*.

В алфавите *ASC*II предусмотрено 256 различных символов, т.е. *M* = 256, а

$$I = \log_2 256 = 8 \text{ бит} = 1 \text{ байт}$$

В современной кодировке *Unicode* заложено гораздо большее количество символов. В ней определено 256 алфавитных страниц по 256 символов в каждой. Предполагая для простоты, что все символы используются, получим, что

$$I = \log_2 (256 \cdot 256) = 8 + 8 = 16 \text{ бит} = 2 \text{ байта}$$

2. Текст, сохраненный в коде *ASC*II, состоит исключительно из арифметических примеров, которые записаны с помощью 10 цифр от 0 до 9, 4 знаков арифметических операций, знака равенства и некоторого служебного кода, разделяющего примеры между собой. Сравните количество информации, которое несет один символ такого текста, применяя вероятностный и алфавитный подходы. Легко подсчитать, что всего рассматриваемый в задаче текст состоит из *N* = 16 различных символов. Следовательно, по формуле Хартли $I_{\text{вероятностная}} = \log_2 16 = 4$ бита. В то же время для символа *ASC*II

$$I_{\text{алфавитная}} = 8 \text{ бит}$$

3. Пусть некто вынимает одну карту из колоды. Нас интересует, какую именно из 36 карт он вынул. Изначальная неопределенность, рассчитываемая по формуле $I=\log_2 N$, составляет $I=log_2(36)\approx 5{,}17$ *бит*. Вытянувший карту сообщает нам часть информации. Используя формулу $I = \log_2(\frac{N_{до}}{N_{после}})$, определим, какое количество информации мы получаем из этих сообщений:

Вариант A. "Это карта красной масти".

$I=\log_2(36/18)=\log_2(2)=1$ бит (красных карт в колоде половина, неопределенность уменьшилась в 2 раза).

Вариант B. "Это карта пиковой масти".

$I=\log_2(36/9)=\log_2(4)=2$ бита (пиковые карты составляют четверть колоды, неопределенность уменьшилась в 4 раза).

Вариант C. "Это одна из старших карт: валет, дама, король или туз".

$I=\log_2(36)-\log_2(16)=5{,}17-4=1{,}17$ бита (неопределенность уменьшилась больше чем в два раза, поэтому полученное количество информации больше одного бита).

Вариант D. "Это одна карта из колоды".

$I=\log_2(36/36)=\log_2(1)=0$ бит (неопределенность не уменьшилась - сообщение не информативно).

Вариант D. "Это дама пик".

$I=\log_2(36/1)=\log_2(36)=5{,}17$ бит (неопределенность полностью снята).

Табл. 2. Количество информации в сообщении об одном из *N* равновероятных событий:

N	*i*	*N*	*i*	*N*	*i*	*N*	*i*
1	0,00000	17	4,08746	33	5,04439	49	5,61471
2	1,00000	18	4,16993	34	5,08746	50	5,64386
3	1,58496	19	4,24793	35	5,12928	51	5,67243
4	2,00000	20	4,32193	36	5,16993	52	5,70044
5	2,32193	21	4,39232	37	5,20945	53	5,72792
6	2,58496	22	4,45943	38	5,24793	54	5,75489
7	2,80735	23	4,52356	39	5,28540	55	5,78136
8	3,00000	24	4,58496	40	5,32193	56	5,80735
9	3,16993	25	4,64386	41	5,35755	57	5,83289
10	3,32193	26	4,70044	42	5,39232	58	5,85798
11	3,45943	27	4,75489	43	5,42626	59	5,88264
12	3,58496	28	4.80735	44	5,45943	60	5.90689

13	3,70044	29	4,85798	45	5,49185	61	5,93074
14	3,80735	30	4,90689	46	5,52356	62	5,95420
15	3,90689	31	4,95420	47	5,55459	63	5,97728
16	4,00000	32	5,00000	48	5,58496	64	6,00000

Процесс получения информации - уменьшение неопределенности в результате приёма сигнала, а количество информации - количественная мера степени снятия неопределённости.

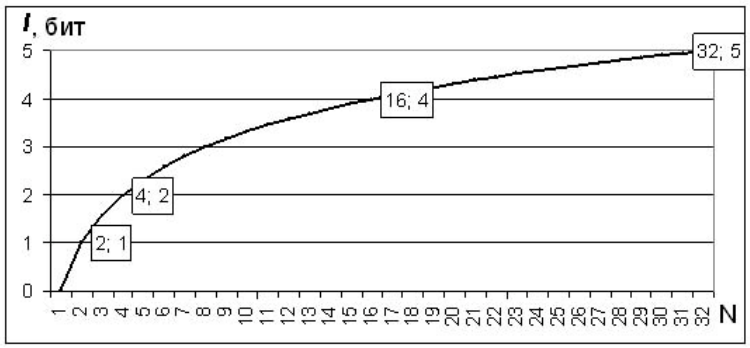


Рис. 1. Зависимость количества информации он числа равновероятных событий, *N* (по данным **Табл.2**).

Количество информации в сообщении об одном из *N* равновероятных событий

В связи с методологической важностью осуждаемого здесь простейшего понятия об информации, рассмотрим более подробно подход Хартли.

Будем считать, что если существует множество элементов и осуществляется выбор одного из них, то этим самым сообщается или генерируется определенное количество информации. Эта информация состоит в том, что если до выбора не было известно, какой элемент будет выбран, то после выбора это становится известным. Найдём вид функции, связывающей количество информации, получаемой при выборе некоторого элемента из множества, с количеством элементов в этом множестве, т.е. с его мощностью.

Если множество элементов, из которых осуществляется выбор, состоит из одного-единственного элемента, то ясно, что его выбор предопределен, т.е. никакой неопределенности выбора нет. Таким образом, если мы узнаем, что выбран этот единственный элемент, то, очевидно, при этом мы не получаем никакой новой информации, т.е. получаем нулевое количество информации. Если множество состоит из двух элементов, то неопределенность выбора минимальна. В этом случае минимально и количество информации, которое мы получаем, узнав, что совершен выбор одного из элементов. Минимальное количество информации получается при выборе одного из двух равновероятных вариантов. Это количество информации принято за единицу измерения и называется «бит».

Чем больше элементов в множестве, тем больше неопределенность выбора, тем больше информации мы получаем, узнав о том, какой выбран элемент. Рассмотрим множество, состоящее из чисел в двоичной системе счисления длиной *i* двоичных разрядов. При этом каждый из разрядов может принимать значения только 0 и 1 (**Табл. 3**).

Табл. 3. – К эвристическому выводу формулы количества информации по Хартли.

Количество двоичных разрядов *i* и количество состояний *N*, которое можно пронумеровать *i*-разрядными двоичными числами				Основание системы счисления и номера разрядов						
				10		**16**	**2**			
4	**3**	**2**	**1**	**2**	**1**	**1**	**4**	**3**	**2**	**1**
				0	0	0	0	0	0	0
			2	0	1	1	0	0	0	1
				0	2	2	0	0	1	0
		4		0	3	3	0	0	1	1
				0	4	4	0	1	0	0
				0	5	5	0	1	0	1
				0	6	6	0	1	1	0
	8			0	7	7	0	1	1	1
				0	8	8	1	0	0	0
				0	9	9	1	0	0	1
				1	0	A	1	0	1	0
				1	1	B	1	0	1	1
				1	2	C	1	1	0	0
				1	3	D	1	1	0	1
				1	4	E	1	1	1	0
16				1	5	F	1	1	1	1

Из **Табл. 3** видно, что при увеличении количества разрядов в двоичных числах на **один** количество состояний, которые можно пронумеровать с помощью этих чисел возрастает в **два** раза:

Количеств двоичных разрядов (i)	Количество состояний N, которое можно пронумеровать i-разрядными двоичными числами
1	2
2	4
3	8
4	16
***	***
i	$N=2^i$

Это верно для чисел в любой системе счисления: при увеличении количества разрядов в числах в системе счисления с основанием E на один количество состояний, которые можно пронумеровать с помощью их чисел возрастает в E раз. Например, при дописывании нуля к десятичному числу справа оно увеличивается в 10 раз, к двоичному – в два раза, к шестнадцатеричному – в 16 раз.

Из **Табл. 3** видно, что количество чисел (элементов) в множестве равно: $N=2^i$

Рассмотрим процесс выбора чисел из рассмотренного множества. До выбора вероятность выбрать любое число одинакова. Существует объективная неопределенность в вопросе о том, какое число будет выбрано. Эта неопределенность тем больше, чем больше N – количество чисел в множестве, а чисел тем больше – чем больше разрядность i этих чисел.

Выбор одного числа даёт количество информации: $i=\log_2 N$

Количество информации, содержащейся в двоичном числе, равно количеству двоичных разрядов в этом числе. Это количество информации i мы получаем, когда случайным равновероятным образом выпадает одно из двоичных чисел, записанных i разрядами, или из некоторого множества выбирается объект произвольной природы, пронумерованный этим числом (предполагается, что остальные объекты этого множества пронумерованы остальными числами и этим они и отличаются).

Формула Хартли полностью совпадает с выражением для энтропии (по Больцману и Эшби), которая рассматривалась ими как количественная мера степени неопределенности состояния системы.

Информация связана с выбором и принятием решений. Поэтому для принятия решений нужна информация, без информации принятие решений невозможно, значение информации для принятия решений является определяющим, процесс принятия решений генерирует информацию.

3.2 Статистический подход к измерению информации

Другой подход к измерению информации **- статистический (вероятностный) –** рассматривает информацию как снятую неопределенность. К. Шеннон предложил связать количество информации, которое несёт в себе некоторое сообщение, с вероятностью получения этого сообщения.

Шеннон измерял количество информации как меру достоверности передаваемого сигнала в битах. Информация уничтожает неопределенность. Степень неопределенности принято характеризовать с помощью понятия «вероятность».

Вероятность - величина, которая может принимать значения в диапазоне от 0 до 1. Она может рассматриваться как мера возможности наступления какого-либо события, которое может иметь место в одних случаях и не иметь места в других.

Вероятность p – количественная априорная (т.е. известная до проведения опыта) характеристика одного из исходов (событий) некоторого опыта. Измеряется в пределах от 0 до 1. Если заранее известны все исходы опыта, сумма их вероятностей равна 1, а сами исходы составляют **полную группу событий**. Если все исходы могут свершиться с одинаковой долей вероятности, они называются **равновероятными**.

Вероятность - численная мера достоверности случайного события, которая при большом числе испытаний близка к отношению числа случаев, когда событие осуществилось с положительным исходом, к общему числу случаев. Два события называют равновероятными, если их вероятности совпадают.

Для определения информации в одном символе алфавита можно также использовать вероятностные методы, поскольку появление конкретного знака в конкретном месте текста есть явление случайное.

Примеры равновероятных событий

1. при бросании монеты: «выпала решка», «выпал орёл»;
2. на странице книги: «количество букв чётное», «количество букв нечётное»;

3. при бросании игральной кости: «выпала цифра 1»,«выпала цифра 2»,«выпала цифра 3»,«выпала цифра 4»,«выпала цифра 5»,«выпала цифра 6».

Неравновероятные события

Определим, являются ли равновероятными сообщения «первой из дверей здания выйдет женщина» и «первым из дверей здания выйдет мужчина». Однозначно ответить на этот вопрос нельзя. Во-первых, как известно количество мужчин и женщин неодинаково. Во-вторых, всё зависит от того, о каком именно здании идет речь. Если это военная казарма, то для мужчины эта вероятность значительно выше, чем для женщины.

Пусть опыт состоит в сдаче студентом экзамена по информатике. Очевидно, у этого опыта всего 4 исхода (по количеству возможных оценок, которые студент может получить на экзамене). Тогда эти исходы составляют полную группу событий, т.е. сумма их вероятностей равна 1. Если студент учился хорошо в течение семестра, значения вероятностей всех исходов могут быть такими: $p(5) = 0.5$; $p(4) = 0.3$; $p(3) = 0.1$; $p(2) = 0.1$, где запись $p(\nu)$ означает вероятность исхода, когда получена оценка ν ($\nu = \{2, 3, 4, 5\}$). Если студент учился плохо, можно заранее оценить возможные исходы сдачи экзамена, т.е. задать вероятности исходов, например, следующим образом: $p(5) = 0.1$; $p(4) = 0.2$; $p(3) = 0.4$; $p(2) = 0.3$.

Рассмотрим источник информации, передающий элементарные сигналы n различных типов. Проследим за достаточно длинным отрезком сообщения. Пусть в нём имеется N_1 сигналов первого типа, N_2 сигналов второго типа, ..., N_n сигналов n-го типа, причем $N_1 + N_2 + ... + N_n = N$ – общее число сигналов в наблюдаемом отрезке, f_1, f_2, ..., f_n – частоты соответствующих сигналов. При возрастании длины отрезка сообщения каждая из частот стремится к фиксированному пределу, т.е.

$$\lim f_i = p_j, (j = 1, 2, ..., n), \qquad (13)$$

где p_j можно считать вероятностью сигнала. Предположим, получен сигнал j-го типа с вероятностью pj, содержащий – $\log p_j$ единиц информации. В рассматриваемом отрезке j-й сигнал встретится примерно N_{pj} раз (будем считать, что N достаточно велико), и общая информация, доставленная сигналами этого типа, будет равна произведению $N_{pj} \log p_j$. То же относится к сигналам любого другого типа, поэтому полное количество информации, доставленное отрезком из N сигналов, будет примерно равно

$$-N\sum_{i=1}^{n} p_i \log p_i. \qquad (14)$$

Чтобы определить среднее количество информации, приходящееся на один сигнал, т.е. удельную информативность источника, нужно это число разделить на N. При неограниченном росте приблизительное равенство переходит в точное. В результате возникает асимптотическое соотношение для среднего количества информации (количество бит в сообщении, что любое из n событий произошло), получаемой со всеми n сообщениями – **формула Шеннона**

$$I = -\sum_{j=1}^{n} p_j \log p_j = \sum_{j=1}^{n} p_j \log_2 \frac{1}{p_j}, \qquad (15)$$

$$0<p_j<1;\ \sum_{j=1}^{n} p_j = 1 \qquad (16)$$

где I - символ количества информации, воплощенной в некоторой системе, j - индекс состояний системы, n - число состояний, p_j - вероятность состояния j.

Формула Шеннона предназначена для измерения количества информации в системах, которым присуще конечное количество дискретных состояний, различающихся по распространенности внутри соответствующих систем. Доказано, что эта формула выражает единственно возможную меру количества информации с системах указанного в ней типа (с точностью до постоянного множителя, который служит для выбора единицы информации). Величина I в формуле Шеннона представляет собой математическое ожидание информации, воплощенной в некоторой системе, имеющей различные состояния j. Единичная информация, воплощенная в состоянии j, определяется по формуле:

$$i = -\log_2 p_j = \log_2(1/p_j), \qquad (17)$$

где p_j – вероятность j-го сообщения (исхода).

Если система имеет только одно состояние (n=1), то для него p_j=1, j=0. Если $n \geq 2$, то все вероятности $p<1$, все $j>0$. Вклад каждого из состояний в общую (среднюю) информацию, содержащуюся в ней, определяется величиной

$$I_j = p_j \cdot i \qquad (18)$$

Было бы неверно полагать, что этот вклад тем больше, чем больше i, т.е. чем меньше p_j. I_j^{max} достигается при $p_j=e^{-1}\sim0.37$ ($I_j^{max}=0.5307$); I_j убывает вплоть до нуля как при уменьшении, так и при увеличении вероятности по отношению к этому значению. Общее количество информации в системе есть сумма ее величин, воплощенных в состояниях системы:

$$I = \sum_{j=1}^{n} I_j \qquad (19)$$

Информация случайной величины точно равна логарифму количества состояний лишь при равномерном распределении. Во всех прочих случаях количество информации будет меньше.

Формула Хартли - частный случай более общей формулы Шеннона. Если в формуле Шеннона принять, что $p_1 = p_2 = ... = p_j = ... = p_N = 1/N$, то $I = -\sum_{j=1}^{N} \frac{1}{N} \log \frac{1}{N} = -\log \frac{1}{N} = \log N$ или $N=2^I$, где N - количество равновероятных событий.

Пример 9. Определить количество информации, содержащейся в сообщении о результате сдачи экзамена для студента-хорошиста. Пусть $I(j)$ – количество информации в сообщении о получении оценки j. В соответствии с формулой Шеннона имеем: $I(5) = -\log_2 0{,}5 = 1$, $I(4) = -\log_2 0{,}3 = 1{,}74$, $I(3) = -\log_2 0{,}1 = 3{,}32$, $I(2) = -\log_2 0{,}1 = 3{,}32$.

Пример 10. Определить количество информации, содержащейся в сообщении о результате сдачи экзамена для нерадивого студента: $I(5) = -\log_2 0{,}1 = 3{,}32$, $I(4) = -\log_2 0{,}2 = 2{,}32$, $I(3) = -\log_2 0{,}4 = 1{,}32$, $I(2) = -\log_2 0{,}3 = 1{,}74$.

Таким образом, количество получаемой с сообщением информации тем больше, чем неожиданнее данное сообщение. Этот тезис использован при эффективном кодировании кодами переменной длины (т.е. имеющими разную геометрическую меру): исходные символы, имеющие большую частоту (или вероятность), имеют код меньшей длины, т.е. несут меньше информации в геометрической мере, и наоборот. Формула Шеннона позволяет определять также размер двоичного эффективного кода, требуемого для представления того или иного сообщения, имеющего определенную вероятность появления.

Пример 11. Есть 4 сообщения: *a, b, c, d* с вероятностями, соответственно, $p(a) = 0{,}5$; $p(b) = 0{,}25$; $p(c) = 0{,}125$; $p(d) = 0{,}125$. Определить число двоичных разрядов, требуемых для кодирования каждого их четырех сообщений. В соответствии с формулой Шеннона имеем: $I(a) = -\log_2 0{,}5 = 2$, $I(b) = -\log_2 0{,}25 = 2$, $I(c) = -\log_2 0{,}125 = 3$, $I(d) = -\log_2 0{,}125 = 3$.

Пример 12. Определить размеры кодовых комбинаций для эффективного кодирования сообщений из примера 1. Для вещественных значений объемов информации (что произошло в примере 1) в целях определения требуемого числа двоичных разрядов полученные значения округляются до целых по традиционным правилам арифметики. Тогда имеем требуемое число двоичных разрядов: для сообщения об оценке 5 – 1, для сообщения об оценке 4 – 2, об оценке 3 – 3, об оценке 2 – 3.

Пример 5. Определить среднее количество информации, получаемое студентом-хорошистом, по всем результатам сдачи экзамена. В соответствии с приведенной формулой имеем: $I = -(0{,}5\cdot\log_2 0{,}5 + 0{,}3\cdot\log_2 0{,}3 + 0{,}1\cdot\log_2 0{,}1 + 0{,}1\cdot\log_2 0{,}1) = 1{,}67$.

Пример 6. Определить среднее количество информации, получаемое нерадивым студентом, по всем результатам сдачи экзамена. В соответствии с приведенной формулой имеем: $I = -(0{,}1\cdot\log_2 0{,}1 + 0{,}2\cdot\log_2 0{,}2 + 0{,}4\cdot\log_2 0{,}4 + 0{,}3\cdot\log_2 0{,}3) = 1{,}73$. Большее количество информации, получаемое во втором случае, объясняется большей непредсказуемостью результатов: в самом деле, у хорошиста два исхода равновероятны.

Пример 7. Рассчитать количество информации, получаемое при бросании несимметричной четырехгранной пирамидки, если вероятности отдельных событий: 1/2, 1/4, 1/8, 1/8. Количество информации, получаемое после реализации одного из них, рассчитывается по формуле Шеннона:
I=-(1/2*log1/2+1/4*log1/4+1/8*log1/8+1/8*log1/8) бит = 1,75 бит.

Пусть у опыта два равновероятных исхода, составляющих полную группу событий, т.е. $p_1 = p_2 = 0{,}5$. Тогда имеем в соответствии с формулой для расчета I:

$$I = -(0{,}5\cdot\log_2 0{,}5 + 0{,}5\cdot\log_2 0{,}5) = 1. \qquad (20)$$

Эта формула есть аналитическое определение бита по Шеннону, как среднего количества информации содержащегося в двух равновероятных исходах некоторого опыта, составляющих полную группу событий. Единица измерения информации при статистическом подходе – бит.

На практике часто вместо вероятностей используются частоты исходов. Это возможно, если опыты проводились ранее и существует определенная статистика их исходов. Например, в построении эффективных кодов участвуют не частоты символов, а их вероятности.

Пусть имеется строка текста, содержащая тысячу букв. Буква «о» в тексте встречается 90 раз, буква «р» - 40 раз, буква «ф» - 2 раза, буква «а» - 200 раз. Поделив 200 на 1000, получим величину 0.2 - средняя частота, с которой в рассматриваемом тексте встречается буква «а». Вероятность появления буквы «а» в тексте ($p_а$) можем считать равной 0.2. Аналогично, $p_р = 0.04$, $p_ф = 0.002$, $p_о = 0.09$. Далее берём двоичный

логарифм от величины 0.2 и получаем количество информации, которую переносит одна-единственная буква «а» в рассматриваемом тексте. Точно такую же операцию проделаем для каждой буквы. Тогда количество собственной информации, переносимой одной буквой равно

$$i_j = \log_2 1/p_j = -\log_2 p_j, \qquad (21)$$

где p_j - вероятность появления в сообщении j-го символа алфавита.

Можно показать, что функция $I = -\sum_{j=1}^{N} p_j \log_2 p_j$ принимает максимальное значение в точке $\left(\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}\right)$ и оно равно $\log_2 N$. Это позволяет оценить величину информации о распределении вероятностей $p_1, p_2, ..., p_N$, или подробно

$$I(p_1, p_2, ..., p_N) = I\left(\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}\right) - I(p_1, p_2, ..., p_N) \qquad (22)$$

$$I(p_1, p_2, ..., p_N) = \log_2 N + \sum_{j=1}^{N} p_j \log_2 p_j \qquad (23)$$

Информация по Шеннону характеризует уменьшение неопределённости, т.е. снижение трудности решения задачи.

Информация оценивает уменьшение трудности решения задачи, показывает, в какой мере полученное сообщение облегчает решение задачи. Информацию можно использовать, как удобный способ измерения реальной ценности сообщений, независимо от того, насколько оно длинно и многословно. Чем больше вероятность события, тем выше уверенность в том, что оно произойдёт, и тем меньше информации содержит сообщение об этом событии. Когда же вероятность события мала, сообщение о том, что оно случилось, очень информативно. Исходя из понятия «информационный пакет», возникает возможность измерять атрибутивную информацию количеством и размерами информационных пакетов в единице объёма.

Полное количество информации в некотором объекте измерить не возможно. Можно измерить различие в содержании информации двух разных объектов, причём нулевое количество информации выбирается условно. Количество информации в объекте можно характеризовать количеством информационных пакетов выбранного произвольного уровня, входящих в объект. Один пакет - один бит.

Количество информации *I,* характеризующей состояние, в котором пребывает объект, можно определить, используя формулу Шеннона:

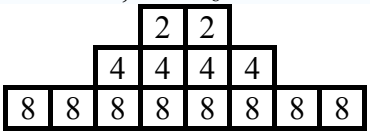
$$I = -[p_1 \cdot \log_2(p_1) + p_2 \cdot \log_2(p_2) + ... + p_n \cdot \log_2(p_N)], \qquad (24)$$

здесь N - число возможных состояний; $p_1, ... p_N$ - вероятности отдельных состояний. Знак минус перед суммой позволяет получить положительное значение для I, поскольку значение $\log_2(p_j)$ всегда не положительно.

Итак, количество информации в сообщении зависит от числа разнообразий, присущих источнику информации и их вероятностей.

Единицы измерения информации служат для измерения объёма информации - величины, исчисляемой логарифмически. Это означает, что когда несколько объектов рассматриваются как один, количество возможных состояний перемножается, а количество информации - складывается. Не важно, идёт речь о случайных величинах в математике, регистрах цифровой памяти в технике или в квантовых системах в физике. Чаще всего измерение информации касается объёма компьютерной памяти и объёма данных, передаваемых по цифровым каналам связи.

В связи с важностью формулы Шеннона, как обобщения меры Хартли для неравновероятных событий, поясним её вывод. Представим себе, что имеются объекты различных видов, причем: всего имеется n видов объектов; объектов каждого i-го вида имеется N_i. Пусть, например, в мешке имеются бильярдные шары, на которых написаны числа 2, 4, 8 и имеется: 4 шара с надписью «2», т.е. N_2=2; 4 шара с надписью «4», т.е. N_4=4; 8 шаров с надписью «8», т.е. N_8=8:



всего, 14 шаров.

Тогда **по Хартли**, если мы извлекаем один из объектов j-го вида, то получаем i_j бит информации $i_j = \log_2 N_j$. В частности:

– при извлечении шара с надписью «2» мы получаем 1=$\log_2 2$ бит информации;
– при извлечении шара с надписью «4» мы получаем 2=$\log_2 4$ бит информации;
– при извлечении шара с надписью «8» мы получаем 3=$\log_2 8$ бит информации.

В среднем по $\bar{i}_j$, т.е. на один объект *j*-го вида: $\bar{i}_j = \frac{i_j}{N_j} = \frac{\log_2 N_j}{N_j}$. Или количественно в нашем примере:

$\bar{i}_2 = \frac{1}{2}$; $\bar{i}_4 = \frac{2}{4}$; $\bar{i}_8 = \frac{3}{8}$. Сумма этих средних будет равна:

$$\bar{I} = \sum_{j=1}^{n} \frac{\log_2 N_j}{N_j} = -\sum_{j=1}^{n} \frac{1}{N_j} \log_2 \frac{1}{N_j} = -\sum_{j=1}^{n} p_j \log_2 p_j \qquad (25)$$

*Формула **Шеннона** позволяет рассчитать **средневзвешенное** количество информации, приходящееся на один объект, получаемое при предъявлении объектов различных видов, причём внутри объектов каждого вида выбор равновероятен, а количество объектов разного вида вообще говоря различно.*

При переходе от частот к вероятностям в формуле Шеннона использованы элементарные свойства логарифмов и введено обозначение: $p_i=1/N_i$, которое **традиционно** интерпретируется как вероятность встречи объектов i-го вида.

Это не совсем точно, т.к. из элементарной теории вероятностей известно, что **в общем случае** вероятность определяется иначе, а именно как отношение количества событий определенного вида к общему количеству всех возможных событий всех видов. Если использовать приведенные выше обозначения, то эта вероятность должна выражаться следующей формулой: $p_j = \frac{N_j}{\sum_{j=1}^{n} N_j}$. Из сравнения этих выражений видно, что общее выражение переходит в $p_j=1/N_j$ в случае, когда все $N_j=1$ и сумма всех $N_j=N$.

Количественно: $\bar{I} = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} = 2{,}375$ бит.

Если количество объектов каждого вида одинаково, то формула Шеннона преобразуется в формулу Хартли.

Идентификация объектов, как относящихся к тому или иному виду (*i*-му виду) осуществляется на основе ***признаков*** этих объектов. В простейшем варианте это может быть и один признак, например номер вида на бильярдном шаре, но в реальных случаях признаков может быть очень много и их различные наборы сложным и неоднозначным образом могут быть связаны с принадлежностью объектов к тем или иным классам.

Но главный вывод от этого не изменяется: ***формула Шеннона даёт средневзвешенное количество информации, приходящееся на один объект, получаемое при предъявлении объектов различных видов (классов), отличающихся своими наборами признаков. Мера Шеннона является обобщением меры Хартли для неравновероятных событий.***

Формула Шеннона предназначена для измерения количества информации в системах, которым присуще конечное количество дискретных состояний, различающихся по распространенности внутри соответствующих систем. Доказано, что формула Шеннона выражает единственно возможную меру количества информации с системах указанного в ней типа (с точностью до постоянного множителя, который служит для выбора единицы информации). Величина *I* в формуле Шеннона представляет собой математическое ожидание информации, воплощенной в некоторой системе, имеющей различный состояния *j*. Единичная информация, воплощённая в состоянии *i*, определяется по формуле:

$$i_j = \log_2(1/p_j) \qquad (26)$$

Если система имеет только одно состояние (n=1), то для него p_j= 1, i_j=0, соответственно I=0. Если $n\geq 2$, то все вероятности $p<1$, все $i_j>0$, соответственно, $i>0$. Вклад каждого из состояний в общую (среднюю) информацию, содержащуюся в ней, определяется величиной

$$I_j = p_j i_j \qquad (27)$$

Было бы неверно полагать, что этот вклад тем больше, чем больше I_i, т.е. чем меньше p_i. I_i^{max} достигается при $p_j=e^{-1}\sim 0.37$ (I_j^{max}=0.5307); I_j убывает вплоть до нуля как при уменьшении, так и при увеличении вероятности по отношению к этому значению. Имеем: $I = \sum_{j=1}^{n} I_j$, т.е. общее количество информации в системе есть сумма ее величин, воплощенных в состояниях системы.

В порядке подведения итогов сравним вероятностный и алфавитный подходы. Первый подход позволяет вычислить предельное (минимально возможное) теоретическое значение количества информации, которое несёт сообщение о данном исходе события. Второй - каково количество информации на практике с учетом конкретной выбранной кодировки. Очевидно, что первая величина есть однозначная характеристика рассматриваемого события, тогда как вторая зависит еще и от способа кодирования: в «идеальном» случае обе величины совпадают, однако на практике используемый метод кодирования может иметь ту или иную степень избыточности. С рассмотренной точки зрения вероятностный подход имеет преимущество. Но, с другой стороны, алфавитный способ заметно проще и с некоторых позиций (например, для подсчета требуемого количества памяти) полезнее.

4. ИНФОРМАЦИЯ ПРИ ПЕРЕДАЧЕ СООБЩЕНИЙ

Применим подходы Хартли-Шеннона к передаче сообщений по линии связи.

Информация, подлежащая передаче и выраженная в определенной форме, называется *сообщением.* Сообщение может быть представлено в форме текста телеграммы, некоторых сведений, передаваемых по телефону или телеграфу, телевизионного изображения, данных на выходе компьютера и т.д.

Ценность любых сведений, содержащихся в переданном получателю сообщении, характеризует количество заключенной в нём информации. Данная величина может определяться степенью изменения поведения получателя под воздействием принятого сообщения. В теории связи количественная оценка информации основывается на концепции выбора наиболее важного сообщения из всей совокупности возможных сообщений. При этом, чем менее вероятен выбор данного сообщения, т.е., чем более оно неожиданно для получателя, тем большее количество информации в нём содержится. Совершенно очевидно обратное: достоверное (заранее известное) сообщение нет смысла передавать, поскольку оно не является неожиданным и, следовательно, не содержит никакой информации. Поэтому любые реальные сообщения следует рассматривать как *случайные события (случайные процессы).*

В практике передачи сообщений по линии связи важно не только количество информации, но и скорость её передачи. Численно *скорость передачи информации* определяется её количеством, переданным за секунду. Предельные возможности скорости передачи информации оцениваются *пропускной способностью* (часто используется термин *ёмкость) канала связи.* Пропускная способность канала численно равна *максимальному количеству информации,* которое можно передать по каналу за 1 с.

Сообщение (а, следовательно, и информация) может быть передано на какое-либо расстояние с помощью определенного материального носителя. Например, при передаче сообщения по почте материальным носителем служит бумага. В радиотехнике в качестве носителей сообщений используются различные сигналы. *Сигнал* (*signum* - знак) - физический процесс (или явление), несущий информацию о состоянии какого-либо объекта наблюдения. По своей физической природе сигналы могут быть электрическими, световыми, звуковыми и др. В радиотехнике в основном используются электрические сигналы.

Для передачи информации с помощью свободного пространства используются специальные электрические сигналы *(переносчики сообщений),* которыми являются хорошо излучающиеся и распространяющиеся в свободном пространстве мощные высокочастотные гармонические электромаганитные колебания *(несущие колебания).* Сами несущие колебания не содержат никакой информации (можно сказать, что *передают с нулевой скоростью),* а только её переносят. Передаваемая информация закладывается в один или ряд параметров несущего колебания.

В современной радиотехнике используются электромагнитные колебания, расположенные в диапазоне частот (радиодиапазоне) от 10 до 10^{13} Гц. Электромагнитные колебания с такими частотами принято называть *радиоволнами* (часто просто *волнами).*

Для обеспечения устойчивой и надежной радиосвязи очень важна длина волны несущего колебания. На выбор того или другого диапазона радиоволн для конкретной системы передачи информации влияет ряд факторов, связанных с особенностью излучения и распространения электромагнитных волн, характером имеющихся в заданном диапазоне помех, параметрами сообщения, характеристиками и габаритными размерами передающих и приемных антенн.

Рассмотрим вопрос о количественном измерении информации, доставляемой сигналом. Как уже говорилось, реальный (случайный) сигнал можно заменить его упрощенной дискретной моделью, согласно которой существенными (информативными) считаются лишь те его значения, которые соответствуют ближайшим узлам решетки (сетки), полученной в результате дискретизации сигнала по времени и уровню. Если сигнал имеет конечную длительность *T,* то число его дискретных отсчетов во времени можно приближенно оценить с помощью теоремы Котельникова

$$n = \frac{T}{\Delta t} = 2F_{\max}T, \quad (28)$$

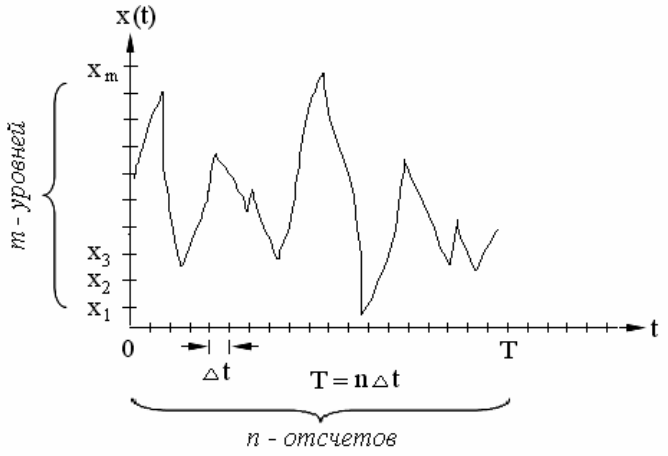
здесь F_{max} - максимальная частота в спектре сигнала x (t). Число уровней сигнала $x\{t)$ определяют соотношением (28), где шаг квантования Δx определяется требуемой точностью обработки информации.

Полагая, что количество информации, которое можно перенести сигналом, будет тем больше, чем больше число возможных сообщений (комбинаций сигнала), дадим оценку числу таких сообщений в рассматриваемом случае. Так как в каждый дискретный момент времени сигнал может принимать одно из N значений, то с помощью двух соседних отсчётов сигнала можно передать уже N^2 различных сообщений, за три отсчета - N^3 сообщений и т. д. В общем случае число различных комбинаций сигнала за время $T=n\Delta t$ составляет $M = N^n$. Полученное таким образом число N дает комбинаторную оценку информации, содержащейся в произвольном дискретном сообщении (слове) из n элементов (букв), каждая из которых принимает одно из N возможных значений, составляющих вместе некоторый алфавит. (Так, с помощью двухразрядного десятичного числа можно записать 100 различных чисел от 0 до 99. При средней длине слова в русском языке $n = 5$ и алфавите с $N = 32$ буквами можно составить 33,5 миллиона различных слов.) Вместе с тем использование N в качестве меры информации неудобно, так как в данном случае не выполняется условие аддитивности, т. е. пропорциональности между длиной слова (длительностью сигнала) и количеством содержащейся в ней информации. Между тем удвоение времени передачи должно приводить к удвоению количества передаваемой информации.

Р. Хартли предложил в качестве меры количества информации использовать логарифм числа возможных сообщений:

$$I = \log_a N_a \quad (29)$$

Согласно Хартли, количество информации в сигнале пропорционально длительности сигнала (числу отсчетов n) и логарифму от мощности использованного алфавита. Выбор основания логарифма a влияет лишь на размерность, т. е. на единицу измерения количества информации. Наиболее часто принимается $a = 2$, при этом значение I измеряется в битах. 1 бит - количество информации, соответствующее одному из двух равновозможных сообщений (да - нет, включить - выключить, исправно - неисправно). В вычислительной технике 1 бит обозначает 1 двоичный разряд - символ, принимающий значение 0 или 1. В качестве единицы представления данных в ЭВМ используется байт - слово (набор) из восьми двоичных разрядов (битов). Легко видеть, что байтом можно передать одно из $2^8 = 256$ различных сообщений.



Подход Хартли, не учитывает, что различные значения дискретного сигнала могут приниматься им с различными вероятностями. Пусть p_j (j = 1, 2,..., N) - априорные вероятности появления i-го значения (уровня) сигнала $\bar{x}_i$.

Рис. 2. К определению количества информации в сигнале.

Пусть n_1 отсчетов сигнала принимают значение x_1, n_2 отсчетов - значение x_2 и т. д. Вероятность подобного события

$$p = p_1^{n_1} p_2^{n_2},..., p_N^{n_N} = \prod_{i=1}^{N} p_i^{n_i} \quad (30)$$

Если общее число отсчетов сигнала $n = \sum_{j=1}^{N} n_j$ достаточно велико, то можно положить $n_1=p_1{}^n$; $n_2=p_2{}^n$; …; $n_N=p_N{}^n$. Эта формула показывает, что вероятность события есть количество благоприятных исходов к общему числу событий.

Тогда $p = \prod_{j=1}^{N} p_j^{p_j n}$. При достаточно большом n можно считать возникающие при этом возможные сочетания равновероятными, т. е. $p = 1/N_a$, откуда получим число возможных сочетаний

$$M = \frac{1}{p} = -\frac{1}{\prod_{j=1}^{N} p_j^{p_j n}} = \prod_{j=1}^{N} p_i^{p_i n} \quad (31)$$

Логарифмируя, находим количество информации в сигнале для этого случая:

$$I = \log_2 N_a = -n \log_2 p_j. \qquad (32)$$

Формула К. Шеннона даёт статистическую оценку количества информации, содержащейся в *n* дискретных отсчётах случайного сигнала (дискретного сообщения).

Количество информации, приходящееся на один отсчёт (элемент) сигнала, называют удельной информативностью или энтропией сигнала:

$$H = \frac{I}{n} = -\sum_{j=1}^{N} p_j \log_2 p_j. \qquad (33)$$

Количество информации в сообщении равно нулю, если это сообщение известно заранее. Если рассматриваемый сигнал принимает какое-либо значение x_k с вероятностью, равной единице ($p_k = 1$), то $p_j = 0$ и $I = 0$. Для того чтобы сигнал содержал информацию, он должен быть принципиально случайным.

Количество информации в сигнале максимально, если все его значения равновероятны, т. е. $p_j = 1/N$ ($j = 1, 2, ..., N$). Если вероятности всех значений сигнала одинаковы, то предсказать поведение сигнала практически невозможно. Неопределенность ситуации велика, поэтому каждый следующий отсчёт, снимая неопределенность, несёт большую информацию. Если же какое-то из значений сигнала существенно преобладает, можно с большим успехом предсказать дальнейшее поведение сигнала. Вероятность ошибки такого предсказания мала. Если все значения сигнала равновероятны, то $p_j = 1/N$ и

$$I = -n \sum_{j=1}^{N} \left(\frac{1}{N} \log \frac{1}{N} \right) = n \log_2 N \qquad (34)$$

Это означает, что формула Хартли даёт верхнюю оценку для количества информации, содержащейся в сигнале с неравномерным распределением значений.

Количество информации, получаемой при измерении непрерывного сигнала *x(t):*

$$I = -n \int_{-\infty}^{\infty} f(x) \log_2 f(x) dx - n \lim_{\Delta x \to 0} \log_2 \Delta x \qquad (35)$$

где $f(x)$ - плотность вероятности случайного (стационарного) процесса x (t); n - число измерений (отсчетов); Δx - погрешность измерения.

Шенноном предложена абстрактная схема связи, состоящая из пяти элементов (источника информации, передатчика, линии связи, приемника и адресата), и сформулированы теоремы о пропускной способности, помехоустойчивости, кодировании и т.д.

Сильной и в то же время слабой стороной классической теории информации, обеспечивающей её универсальность, стало абстрагирование от содержания и природы передаваемых данных. Такую теорию интересуют лишь два аспекта: количество передаваемой информации и качество передачи. Названные характеристики связаны обратной зависимостью: чем точнее передаётся сообщение при наличии помех в канале связи, тем более замедляется передача. Особое внимание в теории информации уделяется оптимальным характеристикам, таким как пропускная способность канала, т.е. максимально возможная скорость передачи при использовании кодирования-декодирования, обеспечивающего исправление ошибок, вызванных помехами.