1. ЕДИНИЦЫ ИНФОРМАЦИИ

Информация и техническая энтропия безразмерны, тем не менее, для их количественного описания существуют специальные единицы, одинаковые для информации и энтропии.

Единицы измерения информации служат для измерения объёма информации - величины, исчисляемой линейно или логарифмически. Это означает, что когда несколько объектов рассматриваются как один, количество возможных состояний перемножаются, 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.

Разнообразие необходимо при передаче информации. Нельзя нарисовать белым по белому, одного состояния недостаточно. Если ячейка памяти способна находиться только в одном (исходном) состоянии и не способна изменять свое состояние под внешним воздействием, это значит, что она не способна воспринимать и запоминать информацию. Информационная емкость такой ячейки равна 0. Минимальное разнообразие обеспечивается наличием двух состояний. 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.

Понятно, что для измерения количества информации (без учёта её содержания) необходима единица информации. За единицу количества информации приняли такое количество информации, при котором неопределённость уменьшается в два раза. Такая единица названа **бит**.

Информационный объем 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 (теория кодирования и передачи сообщений) под количеством информации понимают количество кодируемых, передаваемых или хранимых символов.

Бит (*англ binary digit – двоичная цифра, двоичное число; также игра слов: англ. 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 - минимальная единица измерения количества передаваемой или хранимой информации, соответствующая одному двоичному разряду, способному принимать значений 0 или 1.

1 бит информации - количество информации, посредством которого выделяется одно из двух равновероятных состояний объекта.

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. Следовательно, 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. Энтропия сообщения, выраженная в битах, определяется средним числом символов, необходимых для записи этого сообщения.

Бит – 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 (то есть такое количество информации, которое позволяет однозначно ответить на поставленный вопрос).

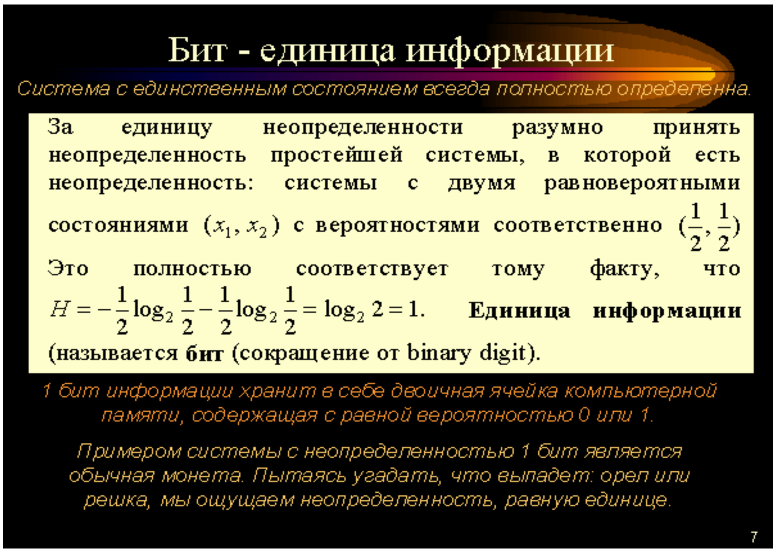
Замечание. В литературе часто встречается утверждение, что единицу «бит» предложил К.Шеннон. Это, естественно, не верно. Этот термин предложен знаменитым американским статистиком, профессором Джоном Туки в 1946. (Туки - пионер современного эмпирического спектрального анализа. В 1949 году он открыл основания для спектрального анализа, используя анализ корреляций, ограниченных временных последовательностей. В 1965 году он в соавторстве с Джимом Кули описал алгоритм для вычисления в цифровой форме преобразования Фурье. Туки – член американской делегации по переговорам с СССР о запрещении подземных испытаний ядерного оружия: предложил методику выявления ядерных взрывов на фоне природных сейсмических явлений, например, землетрясений).

К.Шеннон воспользовался предложением Д.Туки и в 1948 публикацией в “Bell Systems Technical Journal” своей статьи ввёл термин «бит» в практическую информатику.

Если подбросить монету и проследить, какой стороной она упадет, 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 несет 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. Классический пример, иллюстрирующий 1 бит информации – количество информации, получаемое в результате подбрасывания монеты – «орел» или «решка». 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» не равновероятны. Неравновероятность в свою очередь является следствием того, что некоторая предварительная (априорная) информация по этому вопросу уже имеется, полученная, допустим, на основании предыдущего жизненного опыта.



Замечание. Сказанное справедливо только для равновероятного случая.

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. По Шеннону бит - это двоичный логарифм вероятности равновероятных событий или сумма произведений вероятности на двоичный логарифм вероятности при равновероятных событиях.

3. По Шеннону, 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 (то есть такое количество информации, которое позволяет однозначно ответить на поставленный вопрос).

В вычислительной технике и сетях передачи 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». Аналогом бита в квантовых компьютерах является q-бит.

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. На практике чаще применяется 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 ($256=2^8$). В большинстве современных ЭВМ 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, соответственно, 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 ASCII.

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).

Замечание. Термин байт (byte), обозначающий последовательность битов, необходимых для компьютерного представления одного символа (как правило, 8 бит), гораздо моложе термина бит. Его предложил в 1964 доктор Вернер Бухгольц из в IBM. (Первое упоминание байта встречается в одной из статей в "IBM Systems Journal" за 1964). Что же касается происхождения самого термина, то тут существует несколько гипотез. По одной из них, термин байт (byte) произошел от слов BInary digiT Eight (двоичное число восемь) путем замены в образовавшемся слове BITE буквы I на Y. Последнее было сделано для того, чтобы не путать в произношении и написании новый термин с уже существовавшим «битом». Сторонники другой гипотезы утверждают, что «байт» произошел от сокращения слов BinarY TErm (двоичный термин) без всякой возни с заменой одной буквы на другую. Наконец, есть и третьи, утверждаютщие, что «байт» просто был переделан из «бита» для того, чтобы термины для обозначения однородных величин и в звучании были похожи друг на друга.

В принципе, байт определяется для конкретного компьютера 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). В современной традиции, байт часто 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. Во французском языке используются обозначения o, Ko, Mo и т. д. (от слова octet) дабы подчеркнуть, что речь идёт именно о 8 битах.

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: *{M}*11101101 *{И}*11101001 *{P}*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. Соответственно 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 ведет 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 также 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.

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) – это единица измерения емкости памяти или длины записи, равная 1024 байтам. Часто под килобайтом понимается также величина, равная 103 байт.
Мегабайт, Мбайт (megabyte) – это единица измерения емкости памяти или длины записи, равная 1024 Кбайт. Часто под мегабайтом понимается также величина, равная 103 килобайт или 106 байт.
Гигабайт, Гбайт (gigabyte) – единица измерения емкости памяти или длины записи, равная 1024 Мбайт. Часто под гигабайтом понимается также величина, равная 103 мегабайт, 106 килобайт или 109 байт.
Терабайт, Тбайт (terabyte) - это единица измерения емкости памяти или длины записи, равная 1024 Гбайт. Часто под терабайтом понимается также величина, равная 103 гигабайт, 106 мегабайт, 109 килобайт или 1012 байт.
Кубит (quantum bit, qubit) – это «Квантовый бит» – мера и измерения объема памяти в теоретически возможном виде компьютера, использующем квантовые носители, например - спины электронов. Кубит может принимать не два различных значения ("0" и "1"), а несколько, соответствующих нормированным комбинациям двух основных состояний спина, что дает большое число возможных сочетаний. Так, 32 кубита могут образовать около 4 млрд состояний.

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.

Российский ГОСТ 8.417-2002 («Единицы величин») в «Приложении А» 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. Однако следует учитывать, что в стандарте нет сокращения для «бит», поэтому использование записи вроде «Гб» как синонима для «Гбит» недопустимо.

Замечание. 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.

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

В теории информации количеством информации называют числовую характеристику сигнала, которая не зависит от его формы и содержания и характеризует неопределенность, которая исчезает после получения сообщения в виде данного сигнала. В этом случае количество информации 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 нем равно 0. Чем невероятнее событие, тем большее количество информации несет сообщение о нем. Лишь при равновероятных ответах ответ «да» или «нет» несёт один бит информации.

Более строго единицу информации вводят, основываясь на определении информации:

$$I = K \ln P_0, \qquad (1)$$

где K – постоянная, ln – натуральный логарифм, P_0 – число возможных равновероятных событий.

Замечание. Не важно, какой именно логарифм выбрать, поскольку численные величины логарифмов по разным основаниям пропорциональны. Таким образом, вопрос выбора единицы измерения информации фактически равнозначен выбору основания для логарифма количества состояний. Следует помнить, что информация случайной величины точно равна логарифму количества состояний лишь при равномерном распределении. Во всех прочих случаях количество информации будет меньше.

Информация I – безразмерная величина (отвлечённое число), поэтому постоянная K – отвлечённое число. Наиболее удобная система основана на двоичных единицах:

$$K = \frac{1}{\ln 2} = \log_2 e. \qquad (2)$$

Другая система единиц может быть введена, если мы сравним информацию с термодинамической энтропией и будем измерять обе величины в одних и тех же единицах. Энтропия имеет размерность энергии, деленной на температуру. Для энтропии имеется формула Больцмана, очень сходная с (1) и содержащая коэффициент

$$k=1{,}38\cdot 10^{-16} \text{ эрг/градус} \qquad (3)$$

Эта константа известна как постоянная Больцмана. Если мы введем эту постоянную k вместо K в (1), то мы будем измерять информацию в единицах энтропии. Мы можем сделать еще один шаг и выбрать единицы гак, чтобы как энтропия, так и информация были безразмерными величинами и представлялись отвлеченными числами. Для этого нужно измерять температуру в единицах энергии. Обычная стоградусная шкала применима, если k имеет численное значение (3) и рассматривается как отвлеченное число. При этом отношение единиц наших двух систем также есть отвлеченное число, равное

$$\frac{k}{K} = k \ln 2 \approx 10^{-16}.$$

Это численное значение играет важную роль во всех применениях теории.

Соотношение между битом и энтропийной единицей

1 бит = $2{,}3\cdot 10^{-24}$ э.е.

1 э.е. = $4{,}3\cdot 10^{23}$ бит

Полученная Шенноном формула позволила вывести единицы измерения энтропии и количества информации. Для этого приравняем выражение для энтропии системы к единице:

$$H = -\sum_{1}^{n} p_i \log p_i = 1 \qquad (4)$$

где n - число возможных состояний системы n, log - основание логарифма a, p_i - распределение вероятностей.

Для решения уравнения необходимо задаться из каких-либо соображений двумя переменными и вычислить третью.

Рассмотрим теперь единицы информации несколько подробнее.

БИТ

Рассмотрим физическую систему с двумя равновероятными состояниями. Количество информации равное единице может быть получено, если в формуле Шеннона и взять логарифм по основанию 2. Пусть n=2, тогда

$$H = -\frac{1}{2}\log p_1 - \frac{1}{2}\log p_2 = -\log\frac{1}{2} = \log_2 2 = 1 \qquad (5)$$

Следовательно, в данном случае единицей энтропии служит энтропия системы с двумя равновероятными состояниями, вычисленная с помощью логарифма с основанием два. Полученная единица количества информации, представляющая собой выбор из двух равновероятных событий, получила название двоичной единицы, или бита.

Другими словами, если мы приняли, что информация – это устраненная неопределенность, тогда в случае неопределенности физической системы с двумя равновероятными состояниями 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, принято за единицу количества информации - бит. Бит является единицей количества информации и представляет собой информацию, содержащуюся в одном дискретном сообщении источника равновероятных сообщений с объемом алфавита равного двум.

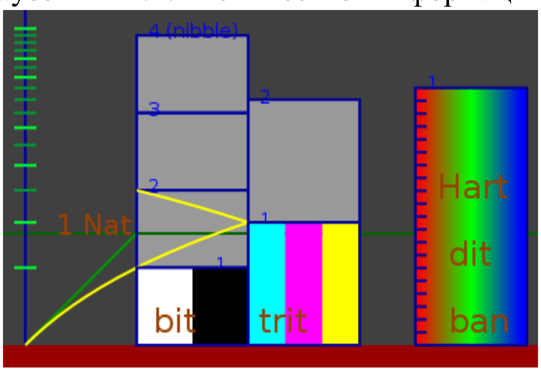


Рис. 1. Сравнение разных единиц измерения информации. Дискретные величины представлены прямоугольниками, единица «нат» - горизонтальным уровнем. Риски слева - логарифмы натуральных чисел.

За единицу информации можно было бы выбрать количество информации, необходимое для различения, например, десяти равновероятных сообщений. Это будет не двоичная (бит), а десятичная (дит) единица информации. Объёмы информации можно представлять как логарифм количества состояний. Бит (бело-чёрный) - одна из самых известных используемых единиц информации. Наименьшее целое число, логарифм которого положителен - 2. Соответствующая ему единица – бит - является основой исчисления информации в цифровой технике. Замена основания логарифма 2 на e, 3 или 10 приводит соответственно к редко употребляемым единицам нат, трит и хартли=дит, равным соответственно: $\mathrm{Log}_2 e \approx 10/7$, $\log_2 3 \approx 8/5 \approx 1{,}585$, $\log_2 10 \approx 19/3 \approx 3{,}322$ бита. Единица как нат (nat, e-бит), соответствующая натуральному логарифму применяется в вычислительной технике в инженерных и научных расчётах.

Не важно, какой именно логарифм выбрать, поскольку численные величины логарифмов по разным основаниям пропорциональны. Таким образом, вопрос выбора единицы измерения информации фактически равнозначен выбору основания для логарифма количества состояний. Следует также заметить, что информация случайной величины точно равна логарифму количества состояний лишь при равномерном распределении. Во всех прочих случаях количество информации будет меньше.

ДИТ (ХАРТЛИ)

За единицу информации можно выбрать количество информации, необходимое для различения, например, десяти равновероятных сообщений. Это будет не двоичная (**бит**), а десятичная (**дит**) единица информации.

Возьмем основание логарифма равным 10-ти и рассмотрим физическую систему с числом n равновероятных состояний. Определим, чему должна быть равна переменная n, чтобы количество информации в формуле Шеннона было равно единице, когда взят логарифм по основанию 10:

$$p_i = \frac{1}{n}, \quad i = 1, \ldots, n$$

Пусть a=10 и

$$H = -\sum_{1}^{n} \log_{10} \frac{1}{n} = 1$$

Отсюда следует, что тогда число состояний n=10. Итак, дит – это энтропия системы с десятью равновероятными состояниями, вычисленная с помощью логарифма с основанием десять.

$$\log_{10} n = 1$$

Можно заметить, что основание логарифма равно числу состояний. Это – важное замечание! Перейдем теперь к рассмотрению виртуальной системы с количеством состояний, равным натуральной единице e. Дит - единица количества информации, содержащейся в одном дискретном сообщении источника равновероятных сообщений с объемом алфавита, равного десяти.

НАТ

Если взять физическую систему с e состояниями, получим натуральную единицу количества информации, называемую натом, при этом основание логарифма в формуле Шеннона равно e=2,7.

Взаимосвязь между единицами количества информации:

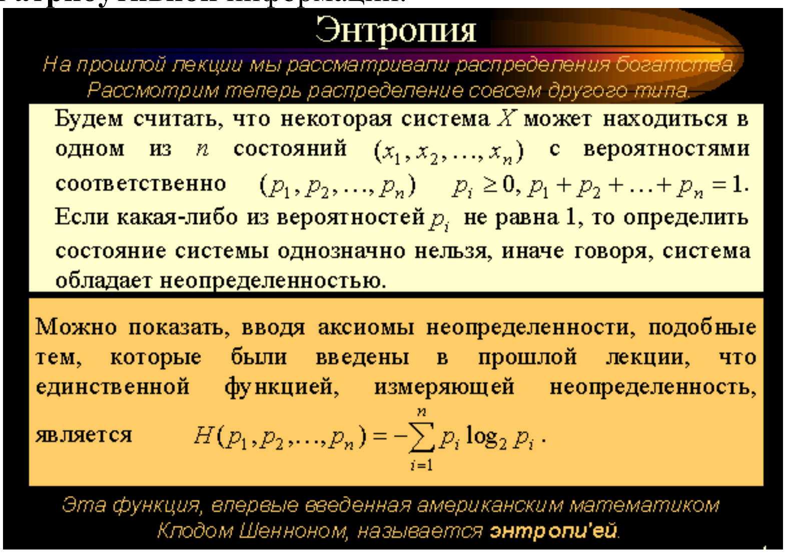
$$1 \quad бит = \frac{1}{1{,}51} нат = \frac{1}{3{,}32} дит$$

Таким образом, 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

2. КОЛИЧЕСТВЕННОЕ ОПРЕДЕЛЕНИЕ ИНФОРМАЦИИ

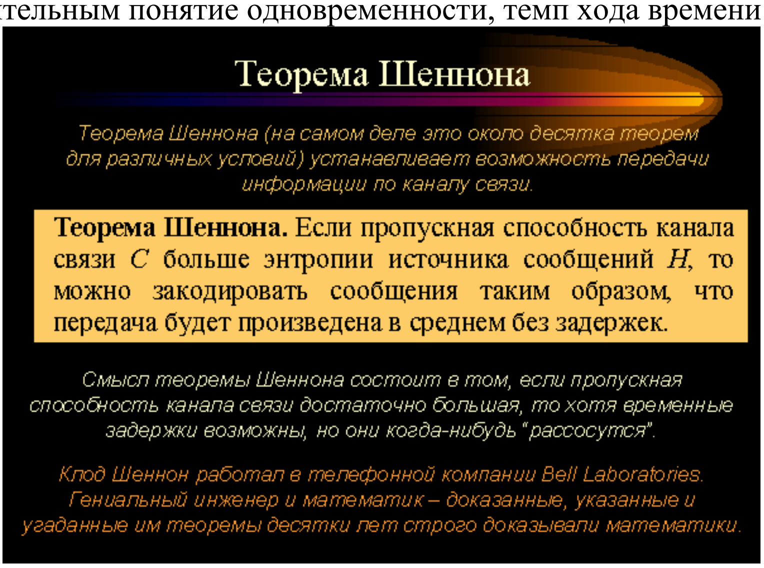
Шеннон измерял количество информации как меру достоверности передаваемого сигнала в битах. Для этого он использовал функцию, отдалённо напоминающую функцию энтропии Л.Больцмана. Н.П. Рашевский В 1955 году предложил топологический подход для измерения количества информации. В 1965 году академик А.Н. Колмогоров предложил алгоритмическое определение количества информации. Количество информации по Колмогорову определяется как минимальная длина программы, позволяющая однозначно преобразовать один объект (множество) в другой объект (множество). Чем больше различаются объекты, тем длиннее оказывается переход от одного к другому, тем больше разность количества информации между этими объектами. Метод Колмогорова не позволяет определять абсолютное количество информации, содержащейся в объекте, но может определять приращение информации. Этот метод универсален. Он может быть реализован, как для оценки **функциональной** информации, так и для оценки **атрибутивной** информации.



Ситуация с определением количества информации в чём-то напоминает ситуацию с определением внутренней энергии, содержащейся в веществе. Обычно полная внутренняя энергия неизвестна. Она может быть очень велика. Определяют только приращение энергии, происходящее при изменении состояния объекта. Химики ограничиваются рассмотрением только энергии химических связей. Физики к этому могут приплюсовать энергию связей между нуклонами атомных ядер. Если идти дальше в глубь микромира, то каждый нуклон имеет свою внутреннюю энергию (связь между кварками, образующими нуклон). А далее возникает энергия вакуума и ещё неизвестно чего. А сколько же энергии в грамме вещества? Нет ответа.

Но проблема с измерением количества энергии это всё же пустяк, по сравнению с проблемой измерения количества информации по Колмогорову. Все виды энергии можно измерять известными единицами (джоуль, ватт, электрон-вольт). А какими единицами измерять информацию, как перейти от одного разнообразия к другому? Как, например, составить алгоритм перевода слона в мышь? Квадрат перевести в треугольник ещё можно. А как перевести один цвет в другой или один запах в другой? Кто станет возражать против того, что цвет и запах это тоже вид информации. Все эти трудности привели в ходе эволюции к особым приёмам переработки информации в человеческом мозге, которые можно назвать моделированием и системным взглядом на мир. Измерение информации носит признаки релятивизма (относительности). Отсутствие нуля (абсолютной точки отсчета) делает все сравнения относительными.

Очевидно, что среда совершенно однородная может быть принята за отсутствие разнообразия. Совершенно однородная среда (процесс) несёт очень мало информации для наблюдателя, но все-таки информация есть, т.к. сообщается сам факт присутствия однородного объекта. Как видно, проблема нуля информации и бесконечности в общем виде не решена. Решения имеются лишь в частных случаях (у Шеннона). Нечто похожее имеет место в специальной теории относительности, где для измерений и наблюдений нет начальной, неподвижной точки отсчета. Выбор произвольной точки отсчета делает относительным понятие одновременности, темп хода времени может изменяться и т.д.



Разработаны различные способы оценки количества информации. В технике чаще всего используется способ оценки, предложенный К. Шенноном.

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.

Если событие никогда не может произойти, его вероятность считается равной 0. Так, вероятность события *"Завтра будет 5 августа 1832 года"* равна нулю в любой день, кроме 4 августа 1832 года. Если событие происходит всегда, его вероятность равна 1.

Чем больше вероятность события, тем выше уверенность в том, что оно произойдет, 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 объема.

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.

Информационный пакет может быть принят за условную единицу количества атрибутивной информации. Один пакет – одна единица независимо от сложности пакета. Нуклоны содержат три единицы информации (3 кварка). Ядра атомов от одного до сотни единиц (нуклоны). Клетка – сотни миллионов молекул и т.д. Подход явно субъективный, так как точка отсчета количества информации принимается произвольно. Но там, где приходиться оперировать с бесконечностями, другого выхода нет. Именно поэтому, согласно теории систем, элемент системы выбирается произвольно. «Элементом сложной системы считается некоторая ее часть, достаточная для понимания функционирования системы». Например, элемент автомобиля это карбюратор, колесо, коленчатый вал, но не атом железа, который является материальной основой всех перечисленных узлов. Элементом человека для анатома являются внутренние органы. Для цитолога – клетки, для химика – белки. Любой объект можно дробить на более мелкие части, но слишком большое дробление не внесет ясности в понимание функционирования объекта, поэтому деление всегда ограничивается некоторым элементом. Как видно, <u>элемент системы и информационный пакет – синонимы.</u>

Существует ещё проблема идентичности объектов. Если два объекта (события) неотличимы, имеющимися у наблюдателя средствами, то из этого не следует, что они идентичны. Порог различимости состояний зависит от очень многих условий. Неразличимость двух объектов по Колмогорову указывает на то, что в них содержится одинаковое количество информации. Однако усовершенствование методов наблюдения может позволить выявить отличия и тогда неизвестно, откуда появится новая информация. Помехи и шумы также маскируют различия между объектами (сигналами). Полное устранение шумов невозможно, поэтому идентичность объектов является условной, субъективной. В связи с выше изложенным, становится понятным, например, почему мы считаем, что все атомы водорода одинаковы, а молекулы воды и др. соединений неотличимы друг от друга. Это искусственный приём упрощения осуществляется с целью возможности описания. В противном случае из-за безграничного количества информации наши интеллектуальные системы не смогли бы перерабатывать её и соответственно принимать решения. Фактически молекулы содержат разное количество движения и этим отличаются друг от друга. Мы замещаем объект его упрощенным образом – моделью и только это позволяет строить картину мира, хотя и в упрощенном варианте. Моделирование означает признание того, что в каждом объекте содержится если не бесконечное, то хотя бы неизмеримое количество информации.

Таким образом, 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:

$$I = -[p_1 * \log_2(p_1) + p_2 * \log_2(p_2) + ... + p_n * \log_2(p_n)], \qquad (6)$$

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

Формула Шеннона, в принципе, может быть использована и для оценки количества информации в непрерывных величинах.

При оценке количества дискретной информации часто используется также **формула Хартли:**

$$I = \log_2(n), \qquad (7)$$

где $\boldsymbol{n}$ - число возможных равновероятных состояний.

Формула Хартли применяется в случае, когда вероятности состояний, в которых может находиться объект, одинаковые.

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.

Понятие количества информации естественно возникает, например, в следующих типовых случаях:

1. Равенство вещественных переменных $a=b$, заключает в себе информацию о том, что a равно b. Про равенство $a^2=b^2$ можно сказать, что оно несет меньшую информацию, чем первое, т.к. из первого следует второе, но не наоборот. Равенство $a^3=b^3$ несет 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 дает 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.

3. ИЗМЕРЕНИЕ ИНФОРМАЦИИ

Информация является важнейшим понятием и основным объектом изучения в информатике. Неудивительно поэтому, что проблема измерения информации имеет фундаментальное значение.

В информатике, как правило, измерению подвергается информация, представленная дискретным сигналом. При этом различают следующие подходы:

1. **Структурный (алфавитный, объёмный).** Измеряет количество информации простым подсчетом информационных элементов, составляющих сообщение. Применяется для оценки возможностей запоминающих устройств, объемов 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. Он удобен при использовании технических средств работы с информацией, т.к. не зависит от содержания сообщения. 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=28, то один символ несет 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 которую в 1948 году предложил Шеннон.
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 1-класснику)

Количество информации - мера уменьшения неопределенности

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

3.1 Структурный подход

В рамках ***структурного подхода*** выделяют три меры информации:

- **геометрическая.** Определяет максимально возможное количество информации в заданных объемах. Мера может быть использована для определения информационной емкости памяти компьютера;
- **комбинаторная**. Оценивает возможность представления информации при помощи различных комбинаций информационных элементов в заданном объеме. Комбинаторная мера может использоваться для оценки информационных возможностей некоторого системы кодирования;
- **аддитивная**, или мера Хартли.

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

Пример 1. Пусть сообщение 5555 6666 888888 закодировано одним из специальных методов эффективного кодирования – кодирование 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: $\mathbf{I}^{(\text{исх.})} = \boldsymbol{l}^{(\text{исх.})} = 14$ символов; $\mathbf{I}^{(\text{закод.})} = \boldsymbol{l}^{(\text{закод.})} = 12$ символов, где $\mathbf{I}^{(\text{исх.})}$, $\mathbf{I}^{(\text{закод.})}$ – количества информации, соответственно, в исходном и закодированном сообщениях; $\boldsymbol{l}^{(\text{исх.})}$, $\boldsymbol{l}^{(\text{закод.})}$ – длины (объемы) тех же сообщений, соответственно. Эффект кодирования определяется как разница между $\mathbf{I}^{(\text{исх.})}$ и $\mathbf{I}^{(\text{закод.})}$ 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 10B.

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 емкость данного устройства: $P^n({}_{10}{}^2) = 10^2 = 100$.

Комбинаторная мера используется для определения возможностей кодирующих систем, которые широко используются в информационной технике.

Пример 1. Определить емкость ASCII-кода, представленного в двоичной или шестнадцатеричной системе счисления. ASCII-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: $I^{(\text{двоичное})} = P^{П}({}_2{}^8) = 2^8 = 256$; $I^{(\text{шестнадцатеричное})} = P^{П}({}_{16}{}^2) = 16^2 = 256$, где $I^{(\text{двоичное})}$, $I^{(\text{шестнадцатеричное})}$ – количества информации, соответственно, для двоичного и шестнадцатеричного представления ASCII-кода. Таким образом, емкость ASCII-кода для двоичного и шестнадцатеричного представления одинакова и равна 256.

Следует отметить, что все коды постоянной длины формируются по правилам комбинаторики или их комбинациям. В случае, когда сообщения формируются как размещения с повторениями из элементов алфавита мощности h и известно количество сообщений M, можно определить требуемый объем сообщения (т.е. его длину l) для того, чтобы в этом объеме представить все сообщения: $l = \log_h M$. Например, есть 4 сообщения – a, b, c, d. Выполняется двоичное кодирование этих сообщений кодом постоянной длины. Для этого требуются 2 двоичных разряда. В самом деле: $l = \log_2 4 = 2$.

Очевидно, комбинаторная мера является развитием геометрической меры, так как помимо длины сообщения учитывает объем исходного алфавита и правила, по которым из его символов строятся сообщения. Особенностью комбинаторной меры является то, что 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 предложена – мера Хартли. Хартли впервые ввел специальное обозначение для количества информации – I и предложил следующую логарифмическую зависимость между количеством информации и мощностью исходного алфавита: $I = l \log h$, где I – количество информации, содержащейся в сообщении; l – длина сообщения; h – мощность исходного алфавита. При исходном алфавите {0,1}; $l = 1$; h = 2 и основании логарифма, равном 2, имеем $I = 1*\log_2 2 = 1$. Данная формула даёт аналитическое определение бита по Хартли: это количество информации, которое содержится в двоичной цифре. Единицей измерения информации в аддитивной мере является бит.

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

Структурный (объёмный или алфавитный) подход основан на определении количества информации в каждом из знаков дискретного сообщения с последующим подсчетом количества этих знаков в сообщении. В простейшем варианте он заключается *подсчете 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

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
(** */	00101000 00101010 00101010 00100000 00101010 00101111	6	48	6

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).

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

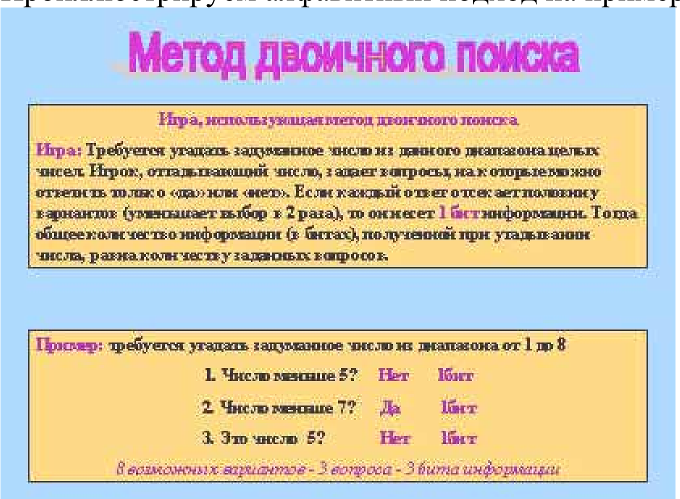
страниц текста	50 000 или 150 романов
цветных слайдов высочайшего качества	150
аудиозапись речи видного политического деятеля	1.5 часа
музыкальный фрагмент качества CD -стерео	10 минут
фильм высокого качества записи	15 секунд
протоколы операций с банковским счётом	за 1000 лет

Пусть сообщение кодируется с помощью некоторого набора знаков. Заметим, что если для данного набора установлен порядок следования знаков, то он называется алфавитом. Наиболее сложной частью работы при объемном измерении информации является определение количества информации, содержащейся в каждом отдельном символе: остальная часть процедуры весьма проста. Для определения информации в одном символе алфавита можно также использовать вероятностные методы, поскольку появление конкретного знака в конкретном месте текста есть явление случайное.

Самый простой метод подсчета заключается в следующем. Пусть алфавит, с помощью которого записываются все сообщения, состоит из M символов. Для простоты предположим, что все они появляются в тексте с одинаковой вероятностью (конечно, это грубая модель, но зато очень простая). Тогда в рассматриваемой постановке применима формула Хартли для вычисления информации об одном из исходов события (о появлении любого символа алфавита): $I = log_2 M$. Поскольку все символы «равноправны», естественно, что объем информации в каждом из них одинаков. Следовательно, остается полученное значение I умножить на количество символов в сообщении, и мы получим общий объем информации в нем. 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.

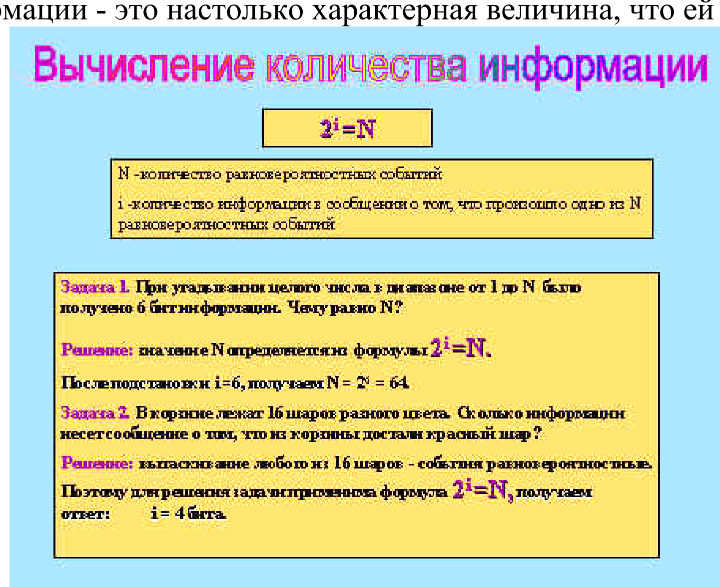


Все множество используемых в языке символов будем традиционно называть алфавитом. Обычно под алфавитом понимают только буквы, но поскольку в тексте могут встречаться знаки препинания, цифры, скобки, то мы их тоже включим в алфавит. В алфавит также следует включить и пробел, т.е. пропуск между словами. Полное количество символов алфавита принято называть мощностью алфавита. Будем обозначать эту величину буквой 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 совсем так, но для упрощения примем такое предположение. В каждой очередной позиции текста может появиться любой из N символов. Тогда, согласно известной нам формуле, каждый такой символ несет I бит информации, которое можно определить из решения уравнения: $2^I = 54$. Получаем: $I = 5.755$ бит. Вот сколько информации несет 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 будет равен: 5,755*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. В этом случае теряют смысл понятия «новые — старые», «понятные — непонятные» сведения. 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 несет 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. С ним мы скоро встретимся 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*.



Сегодня очень многие люди для подготовки писем, документов, статей, книг и пр. используют компьютерные текстовые редакторы. Компьютерные редакторы, в основном, работают с алфавитом размером 256 символов. В этом случае легко подсчитать объем информации в тексте. Если 1 символ алфавита несет 1 байт информации, то надо просто 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. Объем всей информации в книге: 2400*150 = 360 000 байт.

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

Примеры:

1. Определить информацию, которую несет в себе 1-й символ в кодировках ASCII и Unicode.

В алфавите ASCII 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

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

2. Текст, сохраненный в коде ASCII, 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, согласно вычислениям примера 3, для символа ASCII

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

Табл. 1. 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

3.2 Статистический подход к измерению информации

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

Понятие количества информации естественно возникает, например, в следующих типовых случаях:

1. Равенство вещественных переменных $a = b$, заключает в себе информацию о том, что a равно b. Про равенство $a^2 = b^2$ можно сказать, что оно несет 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 дает полную информацию о случайной величине;
4. Рассмотрим схему передачи информации. Пусть передатчик описывается случайной величиной, X, тогда из-за помех в канале связи на приемник 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.

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.

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, все 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 $p(\mathrm{j})$ означает вероятность исхода, когда получена оценка j (j = {2, 3, 4, 5}). Если студент учился плохо, можно заранее оценить возможные исходы сдачи экзамена, т.е. задать вероятности исходов, например, следующим образом: $p(5) = 0.1$; $p(4) = 0.2$; $p(3) = 0.4$; $p(2) = 0.3$.

В обоих случаях выполняется условие:

$$\sum_{i=1}^{n} p_i = 1$$

где n – число исходов опыта, i – номер одного из исходов.

Пусть можно получить n сообщений по результатам некоторого опыта (т.е. у опыта есть n исходов), причем известны вероятности получения каждого сообщения (исхода) - p_{i}. Тогда в соответствии с идеей Шеннона, количество информации I в сообщении i определяется по формуле:

$$I = -\log_2 p_{\mathrm{i}},$$

где p_{i} – вероятность i-го сообщения (исхода).

Пример 1. Определить количество информации, содержащейся в сообщении о результате сдачи экзамена для студента-хорошиста. Пусть $I(\mathrm{j})$ – 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 2. 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 3. 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: $\boldsymbol{I}(a) = -\log_2 0{,}5 = 2$,

$$\boldsymbol{I}(b) = -\log_2 0{,}25 = 2,\ \boldsymbol{I}(c) = -\log_2 0{,}125 = 3,\ \boldsymbol{I}(d) = -\log_2 0{,}125 = 3.$$

Пример 4. 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, для сообщения об оценке 3 – 3, для сообщения об оценке 2 – 3.

Помимо информационной оценки одного сообщения, Шеннон предложил количественную информационную оценку всех сообщений, которые можно получить по результатам проведения некоторого опыта. Так, среднее количество информации $I_{\text{ср}}$, получаемой со всеми n сообщениями, определяется по формуле:

$$l_{cp} = -\sum_{i=1}^{n} p_i \log p_i \qquad (8)$$

где p_i – вероятность i-го сообщения.

Пример 5. Определить среднее количество информации, получаемое студентом-хорошистом, по всем результатам сдачи экзамена. В соответствии с приведенной формулой имеем: I_{cp} = - (0,5*$\log_2$0,5 + 0,3*$\log_2$0,3 + 0,1*$\log_2$0,1 + 0,1*$\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_{cp} = - (0,1*$\log_2$0,1 + 0,2*$\log_2$0,2 + 0,4*$\log_2$0,4 + 0,3*$\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.

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_{cp}:

$$I_{cp} = -(0{,}5*\log_2 0{,}5 + 0{,}5*\log_2 0{,}5) = 1.$$

Эта формула есть аналитическое определение бита по Шеннону: это среднее количество информации, которое содержится в двух равновероятных исходах некоторого опыта, составляющих полную группу событий.

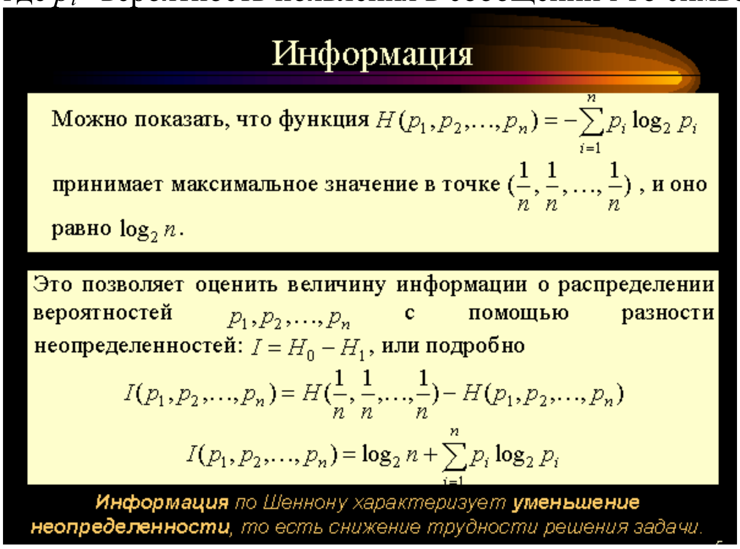
Единица измерения информации при статистическом подходе – бит.

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

Пусть имеется строка текста, содержащая тысячу букв. Буква «о» в тексте встречается примерно 90 раз, буква «р» ~ 40 раз, буква «ф» ~ 2 раза, буква «а» ~ 200 раз. Поделив 200 на 1000, мы получим величину 0.2, которая представляет собой среднюю частоту, с которой в рассматриваемом тексте встречается буква «а». Вероятность появления буквы «а» в тексте ($p_а$) можем считать приблизительно 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 называем то, что получилось количеством 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

$$h_i = \log_2 1/p_i = -\log_2 p_i, \qquad (9)$$

где p_i - вероятность появления в сообщении i-го символа алфавита.



Удобнее в качестве меры количества информации пользоваться не значением h_i, а средним значением количества информации, приходящейся на один символ алфавита

$$H = S\, p_i\, h_i = -S\, p_i \log_2 p_i \qquad (10)$$

Значение H достигает максимума при равновероятных событиях, то есть при равенстве всех p_i

$$p_i = 1/N. \qquad (11)$$

В этом случае формула Шеннона превращается в формулу Хартли.

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.

Смысловым подходом мы в данной лекции заниматься не будем.

5. ЗАДАЧИ И ПРИМЕРЫ

Для решения задач по информатике требуется знание формул Шеннона и Хартли.

Формула К.Шеннона:

Если I - количество информации, N - количество возможных событий, p_i - вероятности отдельных событий, то количество информации для событий с различными вероятностями можно определить по формуле:

$$I = - \mathrm{Sum}\, p_i \log_2 p_i, \qquad (12)$$

где i принимает значения от 1 до N.

Формула Хартли - частный случай формулы Шеннона:

$$I = -\mathrm{Sum} 1/N \log_2(1/N) = I = \log_2 N \text{ или } N=2^I$$

где N - количество равновероятных событий; I - количество бит в сообщении, такое, что любое из N событий произошло.

Иногда формулу Хартли записывают так: $I=\log_2 N=\log_2(1/p) =- \log_2 p$, т. к. каждое из N событий имеет равновероятный исход $p=1/N$, то $N = 1/ p$. $p=K/N$, р - вероятность события; N - общее число возможных исходов; K - число возможных исходов интересующего нас события; I – количество информации

$$N/K= 2^I$$

При равновероятных событиях получаемое количество информации максимально.

При анализе текстов:

$$N=2^i\ ;\ i=\log_2 N;\ I=K{\cdot}I, \qquad (13)$$

где N - полное количество символов в алфавите; i - количество информации, которое несёт каждый символ; K - размер текста; I - размер информации, содержащейся в тексте.

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

1. После экзамена по информатике, объявляются оценки ("2", "3", "4", или "5"). Какое количество информации будет нести сообщение об оценке учащегося *A*, который выучил лишь половину билетов, и сообщение об оценке учащегося *B,* который выучил все билеты? Для учащегося *A* все четыре оценки равновероятны, тогда количество информации вычисляется по формуле: $I=\log_2 4 = 2$ бита. Для учащегося *B* наиболее вероятной оценкой является "5" ($p=1/2$), вероятность оценки "4" в два раза меньше ($p_2=1/4$), а вероятности оценок "2" и "3" еще в два раза меньше ($p_3=p_4 =1/8$). Так как события не равновероятны, воспользуемся для подсчета количества информации формулой Шеннона:

$I=-(1/2 \log_2 1/2 + 1/4 \log_2 1/4 + 1/8 \log_2 1/8 + 1/8 \log_2 1/8) = 1.75$ бит

Сообщение об оценке учащегося А равно 2 бита, а учащегося В – 1,75 бит.

2. Сколько бит информации получит человек при бросании монеты? При бросании монеты сообщение о результате жребия (например, выпал орел) несёт 1 бит информации, поскольку количество возможных вариантов результата равно 2 («орел» или «решка»). Оба эти варианта равновероятны. Ответ может быть получен из решения уравнения: $2^i= 2$, откуда, очевидно, следует: $i = 1$ бит. Вывод: в любом случае *сообщение об одном событии из двух равновероятных несет 1 бит информации.*

3. В барабане для розыгрыша лотереи находится 32 шара. Сколько информации содержит сообщение о первом выпавшем номере (например, выпал номер 15)? Поскольку вытаскивание любого из 32 шаров равновероятно, то количество информации об одном выпавшем номере находится из уравнения: $2^i = 32$**.** Но $32 = 2^5$. Следовательно, $i = 5$ бит. Очевидно, ответ не зависит от того, какой именно выпал номер. *5 бит содержит сообщение о первом выпавшем номере.*

4. Оперативная память компьютера содержит 163840 машинных слов (наибольшую последовательность бит, которую процессор может обрабатывать как единое целое), что составляет 0,625 Мбайт. Сколько бит содержит каждое машинное слово? Переведем 0,625 Мбайт в биты: 0,625 (Мбайт) = $0{,}625 \cdot 2^{10}$ (Кбайт)= $0{,}625 \cdot 2^{10} \cdot 2^{10}$ (байт)=$0{,}625 \cdot 2^{10} \cdot 2^{10} \cdot 2^3$ (бит)=$0{,}625 \cdot 2^{23}$ (бит)=0,625*8388608(бит)=5242880(бит). Разделим

объем оперативной памяти на количество машинных слов: 5242880 (бит)/163 840 (машинных слов) = 32 (бит).
5. Получить внутреннее представление целого числа 1607_{10} в 2-х байтовой ячейке. Воспользуемся правилом представления беззнакового целого числа. Переведём число в двоичную систему: $1607_{10}=11001000111_2$. Внутреннее представление этого числа в 2-х байтовой (16 бит) ячейке будет следующим: 0000 0110 0100 0111.
6. Получить внутреннее представление целого отрицательного числа -1607_{10} в 2-х байтовой ячейке. Воспользуемся правилом представления знакового целого числа. Используя результат предыдущего примера, запишем внутреннее представление положительного числа 1607: 0000 0110 0100 0111. Инвертированием получим обратный код: 1111 1001 1011 1000. Добавим единицу: 1111 1001 1011 1001 - это и есть внутреннее двоичное представление числа -1607.
7. Представить число 0,005089 в нормализованной форме с плавающей точкой в десятичной системе счисления. Чтобы мантисса была меньше единицы и её первая цифра была отлична от нуля, возьмем её равной 0,5089. Тогда порядок будет равен -2. Наше число 0,005089 в нормализованном виде будет равно 0,5089 x 10^{-2}.
8. Сколько различных чисел можно закодировать с помощью 8 бит? I=8 бит, $K=2^I=2^8=256$ различных чисел.
9. Какое количество информации необходимо, чтобы отгадать одно число из набора чисел от единицы до ста. По формуле Хартли можно вычислить, какое количество информации для этого требуется: $I=\log_2 100=6{,}644$. Таким образом, сообщение о верно угаданном числе содержит количество информации, приблизительно равное 6,644 единицы информации.
10. Какое минимальное количество двоичных разрядов потребуется для того, чтобы закодировать алфавит языка племени Мумба-Юмба, состоящий их 16 символов? $2^x = 16$; $2^x = 24$; x = 4. Ответ: 4 разряда
11. 256 символов на клавиатуре кодируются последовательностью из 0 и 1. Сколько же потребуется таких 0 и 1 (то есть разрядов)? $2^x = 256$; $2^x = 28$; x = 8. Ответ: 8 бит или 1 байт
12. Какое минимальное количество двоичных разрядов потребуется для того, чтобы закодировать цифры десятичной системы счисления? $2^x = 10$; $2^x = 23 = 8$ (мало); $2^x = 24 =16$; x = 4. Ответ: 4 разряда.
13. При угадывании целого числа в некотором диапазоне было получено 8 бит информации. Сколько чисел содержал этот диапазон? i=8 бит; K=1 (угадано одно число); $N/K=2^i$; $N/1=2^8$; N=256. Ответ: 256
14. Текст состоит из 20-ти символов (каждый символ кодируется 8 битами). Сколько места в памяти он занимает? Раз каждый символ занимает 8 бит, то 20 символов будут занимать 20*8 = 160 бит = 20 байт. Ответ: текст содержит 20 байт
15. Книга содержит 100 страниц; на каждой странице - 35 строк, в каждой строке - 50 символов. Рассчитать объем информации, содержащийся в книге. Т. к. один символ - 1 байт, то страница содержит 35 x 50 = 1 750 байт информации. Объем всей информации в книге (в разных единицах): 1750x100=175000байт. 175000/1024=170,8984 Кбайт. 170,8984/1024 =0,166893 Мбайт.
16. Чему равен в байтах объём текстовой информации в книге из 258 страниц, если на одной странице размещается в среднем 45 строк по 60 символов (включая пробелы)? Один символ в двоичной форме содержит 1 байт. Строка будет содержать 61 байт, учитывая и служебный символ окончания строки. Тогда 61 байт*45 строк = 2745 байт. Так как в книге 258 страниц текста и на каждой странице в среднем по 2745 байт информации, то объём алфавитно-цифровой информации в книге 2745 байт*258 страниц = 708210 байт = 692 Кбайт. Ответ: текст книги имеет объём 692 Кбайт.
17. Сосчитайте энтропию двухмерного кристалла, показанного на **рис. 2.**

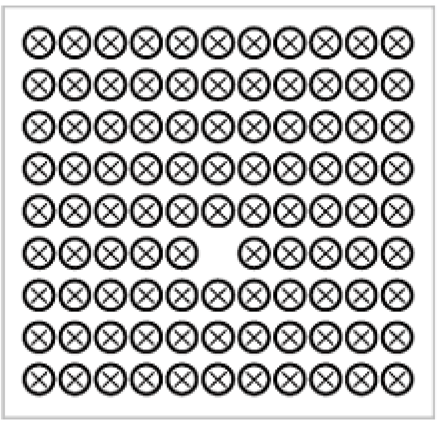


Рис. 2. К задаче 14: Двухмерная модель кристалла с одной вакансией

В этом кристалле 9*11–1=98 атомов. Вакансия может быть в любом из 99 узлов решетки, следовательно, W=99. Энтропия $S=k\ \ln W = 1{,}38\cdot10^{-23}\cdot\ln 99 = 1{,}38\cdot10^{-23}\cdot4{,}60 = 6{,}35\cdot10^{-23}$ Дж/К.
15. Увеличивается или уменьшается энтропия системы, если в ней происходит испарение воды? Если система открытая, то водяной пар покидает её и энтропия оставшейся воды уменьшается, т.к. уменьшается её количество. Если система закрытая и количество вещества в ней не меняется, то ее энтропия
18. Какое количество вопросов достаточно задать вашему собеседнику, чтобы наверняка определить месяц, в котором он родился? Будем рассматривать 12 месяцев как 12 возможных событий. Если спрашивать о конкретном месяце рождения, то, возможно, придется задать 11

вопросов (если на 11 первых вопросов был получен отрицательный ответ, то 12-й задавать не обязательно, так как он и будет правильным). Однако, такой способ слишком длинный. Удобнее задавать «двоичные» вопросы, т.е. вопросы, на которые можно ответить только «Да» или «Нет». Например, «Вы родились во второй половине года?». Каждый такой вопрос разбивает множество вариантов на два подмножества: одно соответствует ответу «Да», а другое - ответу «Нет». Правильная стратегия состоит в том, что вопросы нужно задавать так, чтобы количество возможных вариантов каждый раз уменьшалось вдвое. Тогда количество возможных событий в каждом из полученных подмножеств будет одинаково и их отгадывание равновероятно. В этом случае на каждом шаге ответ («Да» или «Нет») будет нести информацию равную 1 биту. Т.к. события равновероятны и количество возможных событий N=12, то используя формулу Хартли находим количество информации: $I=\log_2 12=3{,}6$ бит. Количество полученных бит информации соответствует количеству заданных вопросов, однако количество вопросов не может быть нецелым числом. Округляем до большего целого числа и получаем 4. Ответ**:** при правильной стратегии необходимо задать не более 4 вопросов.

19. Какое количество информации будет содержать зрительное сообщение о цвете вынутого шарика, если в непрозрачном мешочке находится 50 белых, 25красных, 25 синих шариков
1) всего шаров 50+25+25=100
2) вероятности шаров 50/100=1/2, 25/100=1/4, 25/100=1/4
3) $I= -(1/2 \log_2 1/2 + 1/4 \log_2 1/4 + 1/4 \log_2 1/4) = -(1/2(0-1) +1/4(0-2) +1/4(0-2))$ = <u>1,5 бит</u>

20. В корзине лежит 16 шаров разного цвета. Сколько информации несет сообщение, что достали белый шар? т.к. N = 16 шаров, то $I = \log_2 N = \log_2 16$ = <u>4 бит.</u>

21. Шарик находится в одной из трех урн: А, В или С. Определить сколько бит информации содержит сообщение о том, что он находится в урне В. Такое сообщение содержит $I = \log_2 3 = 1{,}585$ бита информации.

22. В коробке 5 синих и 15 красных шариков. Какое количество информации несет сообщение, что из коробки достали синий шарик? N=15+5=20 всего шариков; K=5 – синих (его достали); N/K=20/5=4; $2^i=4$; i=2 бита. Ответ: 2 бита.

23. В коробке находятся кубики трех цветов: красного, желтого и зеленого. Причем желтых в два раза больше красных, а зеленых на 6 больше чем желтых. Сообщение о том, что из коробки случайно вытащили желтый кубик, содержало 2 бита информации. Сколько было зеленых кубиков? Ж – желтые, К – красные, З – зеленые; Ж=2К; З=Ж+6=2К+6; N=К+Ж+З=К+2К+2К+6= 5К+6; К=Ж=2К; i=2 бита; $N/K=2^i$; $(5K+6)/2K=2^2$; 5К+6=4*2К; К=2; З=2*2+6=10. Ответ: было 10 зеленых кубиков

24. В коробке имеется 50 шаров. Из них 40 белых и 10 черных. Обозначим $p_ч$ - вероятность попадания при вытаскивании черного шара, $p_б$ - вероятность попадания белого шара. Тогда: $p_ч$=10/50=0,2; $p_б$40/50=0,8, т.е. вероятность попадания белого шара в 4 раза больше, чем черного. Количественная зависимость между вероятностью события *p* и количеством информации в сообщении о нем x выражается формулой: $x=\log_2(1/p)$. Количество информации в сообщении о попадании белого шара и черного шара получится: $x_б$=log2(1/0,8)=$\log_2$(1,25)=0,321928; $x_ч$= $\log_2$(1/0,2)=log25=2,321928

25. В колоде содержится 32 карты. Из колоды случайным образом вытянули туза, потом его положили обратно и перетасовали колоду. После этого из колоды опять вытянули этого же туза. Какое количество бит информации в сумме содержат эти два сообщения? N=32; K_1=4 (4 туза в колоде); K_2=1(в колоде один туз определенной масти, который был вытянут в первый раз); N/ K_1=32/4=8; 2^{i1}=8; i_1=3 бита; N/ K_2=32/1=324; 2^{i2}=32; i_2=5 бита; i_{1+} i_2=3+5=8 бит. Ответ: 8 бит

26. В колоде содержится 32 карты. Из нее наугад взяли 2 карты. Какое количество информации несет сообщение о том, что выбраны туз и король одной масти? N_1=32; K_1=4 (4 туза в колоде); N_1/ K_1=32/4=8; $i_1=\log_2 8$=3 бита. После этого в колоде остается 31 карта. N_2=31; K_2=1 (только один король той же масти, что и туз, вытянутый в первый раз); N_2/ K_2=31/1=31; $i_1=\log_2 31$ бит; i_{1+} $i_{2=}$3+ $\log_2 31$ бит. Ответ: 3+ $\log_2 31$ бит

27. Пусть имеется колода из 32 карт (в колоде отсутствуют шестерки). Задумана одна карта (например, туз пик). Сколько двоичных вопросов нужно задать, чтобы отгадать задуманную карту. Двоичным называют вопрос, который предполагает только два взаимоисключающих ответа «да» или «нет». Сколько бит информации несет сообщение о задуманной карте? Давайте рассуждать. Сначала, все карты нам нужно разбить на два подмножества так, чтобы вероятности принадлежности задуманной карты к ним, были равными. Разобьем карты по цветам масти: красная и черная. Первый вопрос: Задумана карта черной масти? Сколько бит информации мы получили? 1 бит информации. Почему? Потому что уменьшили неопределенность ровно вдвое.

Далее, описанную процедуру следует повторить. Второй вопрос: Масть карты пики? (1 бит информации). Третий вопрос: Задумана карта-картинка? (1 бит информации). Четвёртый вопрос: Задумана карта старше дамы? (1 бит информации). Пятый вопрос: Это король? (1 бит информации). Ответ: задуманная карта несет 5 бит информации, поэтому нужно задать пять двоичных вопросов ($\log_2 32 = 5$).

28. Из колоды выбрали 16 карт (все «картинки» и тузы) и положили на стол рисунком вниз. Верхнюю карту перевернули. Сколько информации будет заключено в сообщении о том, какая именно карта оказалась сверху? Все карты одинаковы, поэтому любая из них могла быть перевернута с одинаковой вероятностью. В таких условиях применима формула Хартли. Событие, заключающееся в открытии верхней карты, для нашего случая могло иметь 16 возможных исходов. Следовательно, информация о реализации одного из них равняется $I = \log_2 16 = 4$ бита

29. Пусть теперь верхняя перевернутая карта оказалась чёрной дамой. Отличие данной задачи от предыдущей заключается в том, что в результате сообщения об исходе случайного события не наступает полной определенности: выбранная карта может иметь одну из двух чёрных мастей. В этом случае, прежде чем воспользоваться формулой Хартли, необходимо вспомнить, что информация есть уменьшение неопределенности знаний: $I = H_1 - H_2$. До переворота карты неопределенность (энтропия) составляла $H_1 = \log_2 N_1$ после него – $H_2 = \log_2 N_2$ (причем для нашей задачи $N_1 = 16$, а $N_2 = 2$). В итоге информация вычисляется следующим образом:

$$I = H_1 – H_2 = \log_2 N_1 – \log_2 N_2 = \log_2 N_1/N_2 = \log_2 16/2 = 3 \text{ бита}$$

Заметим, что в случае, когда нам называют карту точно (пример 1), неопределенность результата исчезает, $N_2 = 1$, и мы получаем "традиционную" формулу Хартли. И еще одно полезное наблюдение. Полная информация о результате рассматриваемого опыта составляет 4 бита. В данном же случае, мы получили 3 бита информации, а оставшийся четвёртый описывает сохранившуюся неопределенность выбора между двумя дамами черной масти.

30. Определить количество информации, получаемое при реализации одного из событий, если бросают а) несимметричную четырехгранную пирамидку; б) симметричную и однородную четырехгранную пирамидку.

а) Будем бросать несимметричную четырехгранную пирамидку. Вероятность отдельных событий будет такова: $p_1 = 1/2$, $p_2 = 1/4$, $p_3 = 1/8$, $p_4 = 1/8$, тогда количество информации, получаемой после реализации одного из этих событий, рассчитывается по формуле: $I=-(1/2\log_2 1/2+1/4 \log_2 1/4+1/8\log_2 1/8+1/8\log_2 1/8) = 1/2 + 2/4 + 3/8 + 3/8 = 14/8 = 1{,}75$ (бит).

б) Теперь рассчитаем количество информации, которое получится при бросании симметричной и однородной четырехгранной пирамидки: $I = \log_2 4 = 2$ (бит). Заметим, что для симметричной пирамидки количество информации оказалось больше, чем для несимметричной пирамидки. Максимальное значение количества информации достигается для равновероятных событий.

31. Молекулы ДНК (дезоксирибонуклеиновой кислоты) состоят из четырех различных составляющих (нуклеотидов), которые образуют генетический алфавит. Информационная емкость знака этого алфавита составляет: $4 = 2^I$, т.е. $I = 2$ бит.

32. Какую информацию несёт каждая буква русского алфавита (если считать, что е=ё)? Ответ: 5 бит ($32 = 2^I$).

33. Какова мощность алфавита, с помощью которого записано сообщение, содержащее 2048 символов, если его объём составляет 1,25 Кбайт. Переведём информационный объем сообщения в биты: I = 1,25 Кбайт * 1024 * 8 = 10240 бит. Определяем количество бит, приходящиеся на один символ: 10240 бит : 2048 = 5 бит. Определяем количество символов в алфавите: $N = 2^I = 2^5 = 32$.

Ответ. *32 символа в алфавите.*

34. Пусть имеется два объекта. С каждого из них в определенные моменты времени диспетчеру передаётся одно из двух сообщений: включён или выключён объект. Диспетчеру известны типы сообщений, но неизвестно, когда и какое сообщение поступит. Пусть также, объект А работает почти без перерыва, т.е. вероятность того, что он включен, очень велика (например, $pА_{вкл} = 0{,}99$ и $pА_{выкл}=0{,}01$, а объект Б работает иначе и для него $pБ_{вкл} =pБ_{выкл}=0{,}5$). Тогда, если диспетчер получает сообщение том, что А включен, он получает очень мало информации. С объектом Б дела обстоят иначе. Подсчитаем для этого примера среднее количество информации для указанных объектов, которое получает диспетчер: Объект А: $I=-[0{,}99*\log_2(0{,}99)+0{,}01*\log_2(0{,}01)] =0{,}0808$. Объект Б: $I=-[0{,}50*\log_2(0{,}50)+0{,}50*\log_2(0{,}50)]=1$. Итак, каждое сообщение объекта Б несет 1 бит информации.

35. Студенты группы изучают один из трех языков: английский, немецкий или французский. Причем 12 студентов не учат английский. Сообщение, что случайно выбранный студент Петров изучает английский, несет $\log_2 3$ бит информации, а что Иванов изучает французский – 1 бит. Сколько студентов изучают

немецкий язык? Y – учат французский; X – учат английский, 12 не учат английский. Всего 12+X студентов $N=12+X$; $K_1=X$; $i_1=\log_2 3$ бит; $(12+X)/X=2^{\log_2 3}$; $12+X=3X$; $X=6$; $N=12+6=18$; $i_2=1$ бит; $K_2=Y$; $18/Y=2^1$; $Y=9$; 18-X-Y=18-6-9=3 студента изучают немецкий.
Ответ: 3
36. Ученики класса, состоящего из 21 человека, изучают немецкий или французский языки. Сообщение о том, что ученик *А* изучает немецкий язык, несет $\log_2 3$ бит информации. Сколько человек изучают французский язык? N=21; $\mathrm{Log}_2 3=(21/K)$; $21/K=2^{\mathrm{Log}_2 3}$; 21/K=3; K=7; 21-7=14 учеников изучают французский язык. Ответ: 14
37. В составе 16 вагонов, среди которых К – купейные, П – плацкартные и СВ – спальные. Сообщение о том, что ваш друг приезжает в СВ несет 3 бита информации. Определите, сколько в поезде вагонов СВ. N=16; K=CB; i=3 бита; $i=\log_2(N/K)$; $3=\log_2(16/CB)$; 16/CB=8; CB=2. Ответ: 2
38. Для записи письма был использован алфавит мощностью в 16 символов. Письмо состояло из 25 строк. В каждой строке вместе с пробелами было 64 символа. Сколько байт информации содержало письмо? N=16, $i=\log_2 16=4$ бит, K=25*64=1600, *I*=K*i=1600*4 бит=6400 бит=800 байт. Ответ: 800 байт
39. Письмо состояло из 30 строк. В каждой строке вместе с пробелами по 48 символов. Письмо содержало 900 байт информации. Какова мощность алфавита (количество символов), которым было написано письмо? K=30*48=1440; I=900 байт=7200 бит; i =I/K=5 бит; $N=2^5=32$ символа
40. Для шифрования информации был использован код, состоящий из 64 различных знаков. Какое количество байт содержит шифровка, состоящая из 110 групп по 12 знаков в каждой группе? N=64; K=110*12=1320; $i=\log_2 64=6$ бит; I=6*1320 бит=7920 бит=990 байт. Ответ: 990
41. Шифровка состояла из 36 групп символов по 6 символов в группе и содержала 81 байт информации. С помощью скольких различных знаков была закодирована шифровка? K=36*6= 216; I=81 байт = 648 бит; i=648/216=3 бита; $N=2^3=8$ символов. Ответ: 8
42. В доме 160 квартир. Сколько бит должно содержать двоичное слово, чтобы закодировать в это доме двоичным кодом все квартиры? Пусть слово содержит i=7 бит, тогда всего можно закодировать $2^7=128$ квартир. Нельзя закодировать все 160. Значит, пусть слово содержит i=8 бит, тогда всего можно закодировать $2^8=256$ квартир. Этого достаточно чтобы закодировать 160 квартир. Ответ: 8
43. Даны два текста, содержащих одинаковое количество символов. Первый текст состоит из алфавита мощностью 16 символов, а второй текст – из 256 символов. Во сколько раз информации во втором тексте больше, чем в первом? $K_1=K_2$; $N_1=16$; N_2+256; $i_1=\log_2 16=4$; $i_2=\log_2 256=8$; $I_1=K_1*i_1$; $I_2=K_2*i_2$; $I_2/I_1=(K_2*i_2)/(K_1*i_1)=(K_2*8)/(K_2*4)=8/4=2$. Ответ: в 2 раза
44. Телеграфистка в течение пяти минут передавала информационное сообщение со скоростью 20 байт в секунду. Сколько символов содержало данное сообщение, если она использовала алфавит из 32 символов? 5минут=5*60 с=300 с; 20 байт/сек=160 бит/с; I=160 бит/с * 300с = 48000 бит; N=32; $i=\log_2 32=5$ бит; K=48000/5=9600 символов. Ответ: 9600
45. Текст занимает 3 страницы по 25 строк. В каждой строке записано по 60 символов. Сколько символов в используемом алфавите, если все сообщение содержит 1125 байт? K=3*25*60=4500; I=1125 байт = 9000 бит; i=I/K; i=9000 бит/4500=2 бит; $N=2^i$; $N=2^2=4$ символа в алфавите. Ответ: 4
46. Пусть имеется два объекта. С каждого из них в определенные моменты времени диспетчеру передается одно из двух сообщений: включен или выключен объект. Диспетчеру известны типы сообщений, но неизвестно, когда и какое сообщение поступит. Пусть также, объект А работает почти без перерыва, т.е. вероятность того, что он включен, очень велика (например, $pA_{вкл} = 0,99$ и $pA_{выкл} = 0,01$, а объект Б работает иначе и для него $pБ_{вкл} =pБ_{выкл}=0,5$). Тогда, если диспетчер получает сообщение том, что А включен, он получает очень мало информации. С объектом Б дела обстоят иначе. Подсчитаем по формуле Шеннона для этого примера среднее количество информации для указанных объектов, которое получает диспетчер: Объект А: $I = - [0,99*\log_2(0,99)+0,01* \log_2(0,01)] = 0,0808$. Объект Б : $I = -[0,50*\log_2(0,50)+0,50*\log_2(0,50)] = 1$.
Ответ. Каждое сообщение объекта Б несет 1 бит информации.
47. Пусть объект может находиться в одном из восьми равновероятных состояний. Тогда по формуле Хартли количество информации, поступающей в сообщении о том, в каком именно он находится, будет равно $I = \log_2(8) = 3$ [бита].
48. Книга, набранная с помощью компьютера, содержит 150 страниц; на каждой странице - 40 строк, b в каждой строке - 60 символов. Каков объем информации в книге? Мощность компьютерного алфавита равна 256. Один символ несет 1 байт информации. Значит, страница содержит 40 х 60 = 2400 байт информации. Объем всей информации в книге (в разных единицах):

2400 x 150 == 360 000 байт. 360000/1024 = 351,5625 Кбайт. 351,5625/1024 = 0,34332275 Мбайт.

49. Подсчитайте объем информации, содержащейся в романе А. Дюма «Три мушкетёра», и определите, сколько близких по объему произведений можно разместить на одном лазерном диске? (590 стр., 48 строк на одной странице, 53 символа в строке). 90*48*53=1500960 (символов). 1500960байт» 1466Кбайт» 1,4Мбайт. На одном лазерном диске ёмкостью 600 Мбайт можно разместить около 428 произведений, близких по объёму к роману А. Дюма «Три мушкетера».

50. На диске объемом 100 Мбайт подготовлена к выдаче на экран дисплея информация: 24 строчки по 80 символов, эта информация заполняет экран целиком. Какую часть диска она занимает? Код одного символа занимает 1 байт. 24*80=1920 (байт). Объем диска 100*1024*1024 байт = 104857600 байт. 1920/104857600=0,000018 (часть диска).

51. В барабане для розыгрыша лотереи находится 32 шара. Сколько информации содержит сообщение о первом выпавшем номере (например, выпал номер 15)? Поскольку вытаскивание любого из 32 шаров равновероятно, то количество информации об одном выпавшем номере находится из уравнения: $2^i = 32$. Но $32 = 2^5$. Следовательно, i = 5 бит. Очевидно, ответ не зависит от того, какой именно выпал номер.

52. При игре в кости используется кубик с шестью гранями. Сколько бит информации получает игрок при каждом бросании кубика? Выпадение каждой грани кубика равновероятно. Поэтому количество информации от одного результата бросания находится из уравнения: $2^i = 6$. Решение этого уравнения: $i = \log_2 6$. i = 2,585 бит. Этот ответ точен, но можно воспользоваться таблицей логарифмов и с точностью до 3-х знаков после запятой вычислить, что i = 2,585 бит. Ответ. *Количество информации от одного бросания кубика равно* $Log_2 6$.

53. Определить информационный объём полноэкранного графического изображения на экране, имеющем разрешающую способность 640*320 и 16 возможных цветов. $x=\log_2 16=4$ (количество бит, необходимых для кодирования цвета) 640*320*4 = 819200 бит = 102400 байт = 100 Кбайт.

54. Для хранения области экрана монитора размером 256*128 точек выделено 32 Кбайта оперативной памяти. Сколько максимально цветов допустимо использовать для раскраски точек?
256*128 = 32768 бит; 32 Кбайт = 32*1024 байт = 32768 байт = 32768*8 бит; Для цвета остается 8 бит; 28 = 256 (количество цветов). Ответ: 256 цветов

55. Черно-белое изображение имеет 8 градаций яркости. Размер изображения 10*15 см. Разрешение 300 точек на дюйм (1 дюйм = 2,5 см). Сколько Кбайт памяти требуется для хранения изображения в несжатом виде? N=8; $i=\log_2 8=3$ бит (на каждую точку). Размер изображения = 10*15 см = 4*6 дюйм = 24 дюйм^2 На дюйм – 300 точек, на дюйм^2=300^{2}точек=90000точек. K=90000точек*24 дюйм^2=2160000 точек. *I*=K*i=2160000*3бит=6480000 бит = 810000 байт = 810 Кбайт. Ответ: 810

56. Цветное изображение имеет 256 цветов. Размер изображения 7,5*12,5 см. Для хранения изображения требуется $432 \cdot 10^5$ бит памяти. Каково разрешение изображения в точках на дюйм? (1 дюйм=2,5 см). N=256; $i=\log_2 256=8$ бит; I=43200000 бит; 7,5 см*12,5 см = 3 дюйм*5 дюйм=15 дюйм^2; X - точек на дюйм; X^2 - точек на дюйм^2; $K=X^2*15$; $I=15X^2*8$ бит; 43200000 бит=$15X^2*8$ бит; X^2=360000; X=600 точек на дюйм. Ответ: 600

57. Графическое 16 цветное изображение имеет размер 256 пикселей на 200 пикселей. Какое место в памяти оно занимает? Для представления 16 цветного изображения требуется $\log_2 16=4$ бита, следовательно, цвет пикселя кодируется 4 битами. Размер изображения 256 на 200, значит количество информации в картинке 256*200*4=204 800 байт=200 Кбайт. Ответ: изображение занимает в памяти 200 килобайт

58. Какое максимально возможное количество цветов в изображении, если цвет кодируется 24 битами? Для того чтобы вычислить максимальное количество цветов, достаточно возвести два в степень 24. 224=16777216. Ответ: 24 битами можно представить 16777216 цветов.

59. Глаз человека способен различать порядка 4 тысяч цветов, сколько бит достаточно для представления такого количества? Для ответа на вопрос задачи нужно решить уравнение $\log_2 4000 = x$; или эквивалентное ему 2x = 4000. Поскольку 212 = 4096, то достаточно 12 бит по 4 бита на составляющие красного, зеленого и синего цвета. Ответ: для представления 4 тысяч цветов достаточно 12 бит.