

Алфавитный (объёмный) подход к измерению информации

Алфавитный (объёмный) подход к измерению информации позволяет определить количество информации, заключенной в тексте, записанном с помощью некоторого алфавита.

Алфавит - множество используемых символов в языке.

Обычно под алфавитом понимают не только буквы, но и цифры, знаки препинания и пробел.

Мощность алфавита (N) - количество символов, используемых в алфавите.

Например, мощность алфавита из русских букв равна **32** (буква ё обычно не используется).

Если допустить, что все символы алфавита встречаются в тексте с одинаковой частотой (равновероятно), то количество информации, которое несет каждый символ, вычисляется по формуле Хартли:

$$i = \log_2 N,$$

где N - мощность алфавита.

Формула Хартли задает связь между количеством возможных событий N и количеством информации i :

$$N = 2^i$$

Из базового курса информатики известно, что в компьютерах используется двоичное кодирование информации. Для двоичного представления текстов в компьютере чаще всего используется равномерный восьмиразрядный код. С его помощью можно закодировать алфавит из **256** символов, поскольку $256 = 2^8$.

В стандартную кодовую таблицу (например, ASCII) помещаются все необходимые символы: английские и русские прописные и строчные буквы, цифры, знаки препинания, знаки арифметических операций, всевозможные скобки и пр.

В двоичном коде один двоичный разряд несет одну единицу информации, которая называется **1 бит**.

Например, в **2**-символьном алфавите каждый символ «весит» **1 бит** ($\log_2 2 = 1$); в **4**-символьном алфавите каждый символ несет **2 бита** информации ($\log_2 4 = 2$); в **8**-символьном - **3 бита** ($\log_2 8 = 3$) и т. д.

Один символ из алфавита мощностью 256 (2^8) несет в тексте 8 битов информации. Такое количество информации называется байтом.

1 байт = 8 битов

Информационный объем текста в памяти компьютера измеряется в **байтах**. Он равен количеству знаков в записи текста.

Для измерения информации используются и более крупные единицы:

Название единицы измерения	Численная величина в байтах	Точное количество байтов
Килобайт (Кбайт)	2^{10}	1024 байт
Мегабайт (Мбайт)	2^{20}	1024 килобайт 1 048 576 байт
Гигабайт (Гбайт)	2^{30}	1024 мегабайт 1 073 741 824 байт
Терабайт (Тбайт)	2^{40}	1024 гигабайт 1 099 511 627 776 байт
Петабайт (Пбайт)	2^{50}	1024 терабайт 1 125 899 906 842 624 байт
Эксабайт (Эбайт)	2^{60}	1024 петабайт 1 152 921 504 606 846 976 байт
Зеттабайт (Збайт)	2^{70}	1024 эксабайт 1 180 591 620 717 411 303 424 байт
Йоттабайт (Йбайт)	2^{80}	1024 зеттабайт 1208925819614629174706176 байт

Единицы измерения количества информации, в названии которых есть приставки «кило», «мега» и т. д., с точки зрения теории измерений не являются корректными, поскольку эти приставки используются в метрической системе мер, в которой в качестве множителей кратных единиц используется коэффициент

10, где $n=3,6,9$ и т. д.

Для устранения этой некорректности *Международная электротехническая комиссия*, занимающаяся созданием стандартов для отрасли электронных технологий, утвердила ряд новых приставок для единиц измерения количества информации: **киби** (kibi), **меби** (mebi), **гиби** (gibi), **теби** (tebi), **пети** (peti), **эксби** (exbi). Однако пока используются старые обозначения единиц измерения количества информации, и требуется время, чтобы новые названия начали широко применяться.

Обрати внимание!

Последовательность действий при переводе одних единиц измерения информации в другие приведена на следующей схеме:



Если весь текст состоит из K символов, то при алфавитном подходе объём V содержащейся в нем информации равен:

$$V = K \cdot i$$

где i - информационный вес одного символа в используемом алфавите.

Зная, что $i = \log_2 N$, данную выше формулу можно представить в другом виде:

если количество символов алфавита равно N , а количество символов в записи сообщения - K , то информационный объём V данного сообщения вычисляется по формуле:

$$V = K \cdot \log_2 N$$

При алфавитном подходе к измерению информации информационный объём текста зависит только от размера текста и от мощности алфавита, а не от содержания. Поэтому нельзя сравнивать информационные объёмы текстов, написанных на разных языках, по размеру текста.

Пример:

1. Считая, что каждый символ кодируется одним байтом, оцените информационный объем следующего предложения: *Белеет Парус Одинокий В Тумане Моря Голубом!*

Решение.

Так как в предложении **44** символа (считая знаки препинания и пробелы), то информационный объем вычисляется по формуле:

$$V = 44 \cdot 1 \text{ байт} = 44 \text{ байта} = 44 \cdot 8 \text{ бит} = 352 \text{ бита}$$

2. Объем сообщения равен **11** Кбайт. Сообщение содержит **11264** символа. Какова мощность алфавита?

Решение.

Выясним, какое количество бит выделено на **1** символ. Для этого переведем объем сообщения в биты:

11 Кбайт = $11 \cdot 2^{10}$ байт = $11 \cdot 2^{10} \cdot 2^3$ бит = $11 \cdot 2^{13}$ бит и разделим его на число символов.

$$\text{На } 1 \text{ символ приходится: } \frac{11 \cdot 2^{13}}{11264} = \frac{11 \cdot 2^{13}}{11 \cdot 2^{10}} = 2^3 = 8 \text{ бит.}$$

Мощность алфавита определяем из формулы Хартли: $N = 2^8 = 256$ символов.

Содержательный подход к измерению информации

В **содержательном подходе** количество информации, заключенное в сообщении, определяется объемом знаний, который это сообщение несет получающему его человеку.

Вспомним, что с «человеческой» точки зрения **информация** - это знания, которые мы получаем из внешнего мира. Количество информации, заключенное в сообщении, должно быть тем больше, чем больше оно пополняет наши знания.

Вы уже знаете, что за единицу измерения информации принимается **1** бит.

1 бит - минимальная единица измерения количества информации.

Проблема измерения информации исследована в теории информации, основатель которой - *Клод Шеннон*.

В теории информации для бита дается следующее определение:

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

Что такое неопределенность знания, поясним на примерах. Допустим, вы бросаете монету, загадывая, что выпадет: орел или решка. Есть всего два возможных результата бросания монеты. Причем ни один из этих результатов не имеет преимуществ перед другим. В таком случае говорят, что они **равновероятны**.

В случае с монетой перед ее подбрасыванием неопределенность знания о результате равна двум.

Игральный же кубик с шестью гранями может с равной вероятностью упасть на любую из них. Значит, неопределенность знания о результате бросания кубика равна шести.

Еще пример: спортсмены-лыжники перед забегом путем жеребьевки определяют свои порядковые номера на старте. Допустим, что имеется **100** участников соревнований, тогда неопределенность знания спортсмена о своем номере до жеребьевки равна **100**.

Следовательно, можно сказать так:

Неопределенность знания о результате некоторого события (бросание монеты или игрального кубика, вытаскивание жребия и др.) - это количество возможных результатов.

Вернемся к примеру с монетой. После того как вы бросили монету и посмотрели на нее, вы получили зрительное сообщение, что выпал, например, орел. Определился один из двух возможных результатов. Неопределенность знания уменьшилась в два раза: было два варианта, остался один. Значит, узнав результат бросания монеты, вы получили 1 бит информации.

Сообщение об одном из двух равновероятных результатов некоторого события несет **1 бит** информации.

Пусть в некотором сообщении содержатся сведения о том, что произошло одно из **N** равновероятных событий.

Тогда количество информации **i**, содержащееся в сообщении о том, что произошло одно из **N** равновероятных событий, можно определить из **формулы Хартли**:
 $N=2^i$.

Данная формула является **показательным уравнением** относительно неизвестного **i**.

Из математики известно, что **решение такого уравнения** имеет вид:

$i=\log_2 N$ - логарифм **N** по основанию **2**.

Если **N** равно целой степени двойки (**2,4,8,16** и т. д.), то такое уравнение можно решить «в уме».

Пример:

Шахматная доска состоит из 64 полей: 8 столбцов на 8 строк.

Какое количество бит несет сообщение о выборе одного шахматного поля?

Решение.

Поскольку выбор любой из 64 клеток равновероятен, то количество бит находится из формулы:

$$2^i = 64,$$

$$i = \log_2 64 = 6, \text{ так как } 2^6 = 64.$$

Следовательно, $i = 6$ бит.

В противном случае количество информации становится нецелой величиной, и для решения задачи придется воспользоваться таблицей двоичных логарифмов.

Также, если N не является целой степенью 2, то можно выполнить округление i в большую сторону.

При решении задач в таком случае i можно найти как $\log_2 K$, где K - ближайшая к N степень двойки, такая, что $K > N$.

Пример:

При игре в кости используется кубик с шестью гранями.

Сколько битов информации получает игрок при каждом бросании кубика?

Решение.

Выпадение каждой грани кубика равновероятно. Поэтому количество информации от одного результата бросания находится из уравнения: $2^i = 6$.

Решение этого уравнения: $i = \log_2 6$

Из таблицы двоичных логарифмов следует (с точностью до 3-х знаков после запятой):

$$i = 2,585 \text{ бита.}$$

Данную задачу также можно решить округлением i в большую сторону: $2^i = 6 < 8 = 2^3$,

$$i = 3 \text{ бита.}$$

Вероятностный подход к измерению информации

В реальной жизни существует множество ситуаций с различными вероятностями. Например, если у монеты одна сторона тяжелее другой, то при ее бросании вероятность выпадения «орла» и «решки» будет различной.

Сначала разберемся с понятием «**вероятность**». Введем следующие понятия:

испытание - любой эксперимент;

единичное испытание - испытание, в котором совершается одно действие с одним предметом (например, подбрасывается монетка, или из корзины извлекается шар);

исходы испытаний - результаты испытания (например, при подбрасывании монеты выпал «орел», или из корзины извлекли белый шар);

множество исходов испытания - множество всех возможных исходов испытания;

случайное событие - событие, которое может произойти или не произойти (например, выигрыш билета в лотерее, извлечение карты определенной масти из колоды карт).

Вероятностью случайного события (p) называется отношение числа благоприятствующих событию исходов (m) к общему числу исходов (n):

$$p = m/n.$$

Заметим, что вероятность случайного события может изменяться от **0** до **1**.

Пример:

*В беспроигрышной лотерее разыгрывается **3** книги, **2** альбома, **10** наборов маркеров, **10** блокнотов. Какова вероятность выиграть книгу?*

Решение.

*Общее число исходов $2 + 3 + 10 + 10 = 25$; число благоприятствующих исходу событий равно **3**.*

Вероятность выигрыша книги вычисляется по формуле: $p = \frac{3}{25} = 0,12$.

Заметим, что во многих случаях события происходят с разной вероятностью, а значит формула $N = 2^i$ не всегда применима.

Вероятностный подход предполагает, что возможные события имеют различные вероятности реализации.

В этом случае, зная **вероятность (p)** событий, можно определить **количество информации (i)** в сообщении о каждом из них из формулы:

$$2^i = \frac{1}{p}.$$

Количество информации будет определяться по формуле Шеннона, предложенной им в 1948 г. для различных вероятностных событий:

$$I = - \sum_{i=1}^N p_i \log_2 p_i$$

или

$$I = -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \dots + p_N \log_2 p_N),$$

где I - количество информации;

N - количество возможных событий;

p_i - вероятность i -го события.

Качественная связь между вероятностью события и количеством информации в сообщении состоит в следующем: чем меньше вероятность некоторого события, тем больше информации содержит сообщение об этом событии.

Пример:

В корзине лежат 8 черных шаров и 24 белых. Сколько бит информации несет сообщение о том, что достали черный шар?

Решение. Общее число исходов: $8 + 24 = 32$, число благоприятствующих исходу событий равно 8.

Вероятность выбора черного шара определяется как $p = \frac{8}{32} = \frac{1}{4} = 0,25$

Количество информации вычисляем из соотношения $2^i = \frac{1}{0,25} = \frac{1}{1/4} = 4$,

значит, $i = 2$ бита.

Пример:

Пусть при бросании несимметричной четырехгранной пирамидки вероятности отдельных событий равны:

$$p_1 = \frac{1}{2}; p_2 = \frac{1}{4}; p_3 = \frac{1}{8}; p_4 = \frac{1}{8}.$$

Тогда количество информации, которое будет получено после реализации одного из событий, можно вычислить по формуле Шеннона:

$$I = - \left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{8} \log_2 \frac{1}{8} + \frac{1}{8} \log_2 \frac{1}{8} \right) = \left(\frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} \right) = \frac{14}{8} \text{ (бит)} = 1,75 \text{ (бита)}.$$