Измерение количества информации

Содержательный подход

Содержательный подход к измерению информации применяется для измерения **количества информации** в сообщении, получаемом человеком.

При этом ситуация рассматривается так:

Содержательный подход

1) Перед тем, как событие произойдет, имеется *неопределенность знания* человека об ожидаемом событии. Это неопределенность может быть выражена числом возможных вариантов события или вероятностью ожидаемых вариантов события.

2) После того, как событие произошло, *неопределенность знания* снимается: из некоторого возможного количества вариантов оказался выбранным один.

3) По формуле вычисляется *количество информации* в полученном сообщении, выраженное в битах.

Содержательный подход

Формула, используемая для вычисления количества информации, зависит от ситуаций, которых может быть две:

1. Все возможные варианты события равновероятны. Их число конечно и равно *N*.
2. Вероятности *p* возможных вариантов события разные и они заранее известны:

$$p_i,\ i = 1..N$$

Здесь по-прежнему N — число возможных вариантов события.

Содержательный подход: равновероятные события

Если обозначить за ***i*** количество информации в сообщении о том, что произошло одно из равновероятных событий, то величины ***i*** и ***N*** связаны между собой формулой Хартли:

$$2^i = N \ (1)$$

Величина ***i*** измеряется в битах.

1 бит — это количество информации в сообщении об одном из двух равновероятных событий.

Если i — неизвестная величина, то решение уравнения (1):

$$i = \log_2 N \ (2)$$

Формулы (1) и (2) тождественны друг другу. Иногда в литературе формулой Хартли называют (2).

Пример

Пример 1. Сколько информации содержит сообщение о том, что из колоды карт достали даму пик?

Решение. В колоде 32 карты. В перемешанной колоде выпадение любой карты — равновероятные события. Если i — количество информации в сообщении о том, что выпала конкретная карта (например, дама пик), то из уравнения Хартли:

$$2^i = 32 = 2^5$$

Ответ: $i = 5$ бит.

Пример

Пример 2. Сколько информации содержит сообщение о выпадении грани с числом 3 на шестигранном игральном кубике?

Решение.

Считая выпадение любой грани событием равновероятным, запишем формулу Хартли:

$$2^i = 6$$

$$i = \log_2 6 = 2{,}58496$$

Ответ: 2,58496 бит.

Содержательный подход: неравновероятные события

Если вероятность некоторого события равна *p*, а *i* (бит) — это количество информации в сообщении о том, что произошло это событие, то данные величины связаны между собой формулой:

$$2^i = 1/p \quad (3)$$

Решая показательное уравнение (3) относительно *i*, получаем:

$$i = \log_2(1/p) \quad (4)$$

Формула (4) - формула К.Шеннона.

Пример

Пример 3. На автобусной остановке останавливаются два маршрута автобусов: № 5 и № 7. Определить, сколько информации содержит сообщение о том, что к остановке подошел автобус № 5, и сколько информации в сообщении о том, что подошел автобус № 7. Известно, что в течение 1 рабочего дня к остановке автобусы подходили 100 раз. Из них — 25 раз подходил автобус № 5 и 75 раз подходил автобус № 7.

Решение. Сделав предположение, что с такой же частотой автобусы ходят и в другие дни, вероятность появления на остановке автобуса № 5:

$$p_5 = 25/100 = ¼$$

Вероятность появления автобуса № 7:

$$p_7 = 75/100 = ¾$$

Количество информации в сообщении об автобусе № 5:

$$i_5 = \log_2(1/ ¼) = \log_2 4 = 2 \text{ бита.}$$

Количество информации в сообщении об автобусе № 7 равно:

$$i_7 = log_2(4/3) = log_2 4 – log_2 3 = 2 – 1{,}58496 = 0{,}41504 \text{ бита.}$$

Содержательный подход: неравновероятные события

Формула Хартли (1) – частный случай формулы Шеннона(3). Если имеется N равновероятных событий (результат бросания монеты, игрального кубика и т.п.), то вероятность каждого возможного варианта:

$$p = 1/N.$$

Подставив в (3), получим формулу Хартли: $2^i = N$.

Для примера: если бы автобусы № 5 и № 7 приходили к остановке из 100 раз каждый по 50, то вероятность появления каждого из них была бы равна 1/2.

Тогда количество информации в сообщении о приходе каждого автобуса равно $i = \log_2 2 = 1$ биту.

Содержательный подход: неравновероятные события

Пример 4. На остановке останавливаются автобусы № 5 и № 7. Сообщение о том, что к остановке подошел автобус № 5, несет 4 бита информации. Вероятность появления на остановке автобуса с № 7 в два раза меньше, чем вероятность появления автобуса № 5. Сколько бит информации несет сообщение о появлении на остановке автобуса № 7?

Решение. Запишем условие задачи в следующем виде:

$$i_5 = 4 \text{ бита}, \qquad p_5 = 2 \cdot p_7$$

Вспомним связь между вероятностью и количеством информации и найдем p_5.

Пример

Находим p_5:

$$i_5 = \log_2 \frac{1}{p_5} \qquad 2^{i_5} = \frac{1}{p_5} \qquad 2^4 = \frac{1}{p_5} \qquad 16 = \frac{1}{p_5} \qquad p_5 = \frac{1}{16}$$

Теперь вспомним соотношение $p_5 = 2 \cdot p_7$:

$$2p_7 = \frac{1}{16} \qquad p_7 = \frac{1}{32} \qquad i_7 = \log_2 \frac{1}{1/32} \qquad i_7 = \log_2 32 \qquad i_7 = 5$$

Ответ: $i_7 = 5$ *бит*.

Содержательный подход: неравновероятные события

Из полученного результата следует **вывод**:
уменьшение вероятности события в 2 раза увеличивает информативность сообщения о нем на 1 бит.

Очевидно и **обратное правило**:
увеличение вероятности события в 2 раза уменьшает информативность сообщения о нем на 1 бит.

Зная эти правила, предыдущую задачу можно было решить “в уме”.

Измерение информации: алфавитный подход

Алфавитный подход используется для измерения **количества информации** в тексте, представленном в виде последовательности символов некоторого алфавита.

Такой подход не связан с содержанием текста.

Количество информации при этом называется **информационным объемом текста**, который пропорционален размеру текста — количеству символов, составляющих текст.

Данный подход также называют объемным подходом.

Измерение информации: алфавитный подход

Каждый символ текста несет определенное количество информации – **информационный вес символа**.

Информационный объем текста равен сумме информационных весов всех символов данного текста.

$$I = i_1 + i_2 + i_3 + \ldots + i_K = \sum_{j=1}^{K} i_j \quad (1)$$

Здесь:

i_1 обозначает информационный вес первого символа текста,

i_2 — информационный вес второго символа текста и т.д.;

K — размер текста, т.е. полное число символов в тексте.

Измерение информации: алфавитный подход

Все множество различных символов, используемых для записи текстов, называется **алфавитом**.

Размер алфавита — целое число, которое называется **мощностью алфавита**.

Причем в алфавит входят не только буквы определенного языка, но все другие символы, которые могут использоваться в тексте: цифры, знаки препинания, различные скобки, пробел и пр.

Измерение информации: алфавитный подход

Определение информационных весов символов может происходить в двух приближениях:

1) в предположении равной вероятности (одинаковой частоты встречаемости) любого символа в тексте;

2) с учетом разной вероятности (разной частоты встречаемости) различных символов в тексте.

Измерение информации: алфавитный подход

Если допустить, что все символы алфавита в любом тексте появляются с одинаковой частотой, то информационный вес всех символов будет одинаковым.

Пусть N — мощность алфавита. Тогда доля любого символа в тексте составляет $1/N$-ю часть текста. По определению вероятности эта величина равна вероятности появления символа в каждой позиции текста:

$$p = 1/N$$

Измерение информации: алфавитный подход

Согласно формуле К.Шеннона, количество информации, которое несет символ, вычисляется следующим образом:

$$i = \log_2(1/p) = \log_2 N \quad \text{бит} \quad (2)$$

Следовательно, информационный вес символа (i) и мощность алфавита (N) связаны между собой по формуле Хартли:

$$2^i = N.$$

Измерение информации: алфавитный подход

Зная информационный вес одного символа (i) и размер текста, выраженный количеством символов (K), можно вычислить информационный объем текста по формуле:

$$I = K \cdot i \quad (3)$$

Эта формула есть частный вариант формулы (1) в случае, когда все символы имеют одинаковый информационный вес.

Измерение информации: алфавитный подход

Из формулы (2) следует, что при $N = 2$ (двоичный алфавит) информационный вес одного символа равен 1 биту.

С позиции алфавитного подхода к измерению информации ***1 бит*** *— это информационный вес символа из двоичного алфавита.*

Более крупной единицей измерения информации является ***байт****.*

1 байт — это информационный вес символа из алфавита мощностью 256.

Поскольку $256 = 2^8$*, то из формулы Хартли следует связь между битом и байтом:*

$$2^i = 256 = 2^8$$

Отсюда: $i = 8$ *бит* $= 1$ *байт*

Измерение информации: алфавитный подход

Для представления текстов, хранимых и обрабатываемых в компьютере, часто используется алфавит мощностью 256 символов. Следовательно, 1 символ такого текста “весит” 1 байт.

Более крупные единицы:

1 Кб (килобайт) = 2^{10} байт = 1024 байта,

1 Мб (мегабайт) = 2^{10} Кб = 1024 Кб,

1 Гб (гигабайт) = 2^{10} Мб = 1024 Мб.

Измерение информации: алфавитный подход

Приближение разной вероятности встречаемости символов в тексте

В реальном тексте разные символы встречаются с разной частотой. Поэтому вероятности появления разных символов в тексте различны и, следовательно, различаются их информационные веса.

Измерение информации: алфавитный подход

Статистический анализ русских текстов показывает, что частота появления буквы “о” составляет 0,09. Это значит, что на каждые 100 символов буква “о” в среднем встречается 9 раз.

Это же число обозначает вероятность появления буквы “о” в определенной позиции текста: $p_o = 0{,}09$. Отсюда следует, что информационный вес буквы “о” в русском тексте равен:

$$i_o = \log_2(1/0{,}09) = \log_2(100/9) =$$
$$= \log_2(11{,}11111) = 3{,}47393 \text{ бита}$$

Измерение информации: алфавитный подход

Самой редкой в текстах буквой является буква “ф”. Ее частота равна 0,002. Отсюда:

$$i_Ф = \log_2(1/0,002) = \log_2(1000/2) =$$
$$= \log_2(500) = 8,96578 \text{ бит}$$

Отсюда следует качественный вывод: информационный вес редких букв больше, чем вес часто встречающихся букв.

Измерение информации: алфавитный подход

Информационный объем текста с учетом разных информационных весов символов алфавита вычисляется по следующей формуле:

$$I = \sum_{j=1}^{N} n_j i_j = n_1 i_1 + n_2 i_2 + \ldots + n_N i_N \quad (4)$$

Здесь N — размер (мощность) алфавита;

j – номер символа в тексте

n_j — число повторений символа номер j в тексте;

i_j — информационный вес символа номер j.

Измерение информации: алфавитный подход

Пример 1. Для записи текста используются только строчные буквы русского алфавита и “пробел” для разделения слов. Какой информационный объем имеет текст, состоящий из 2000 символов (одна печатная страница)?

Пример

Решение. В русском алфавите 33 буквы. Сократив его на две буквы (например, “ё” и “й”) и введя символ пробела, получаем очень удобное число символов — 32. Используя приближение равной вероятности символов, запишем формулу Хартли:

$$2^i = 32 = 2^5$$

Отсюда: $i = 5$ бит — информационный вес каждого символа русского алфавита. Тогда информационный объем всего текста равен:

$$I = 2000 \cdot 5 = 10\,000 \text{ бит}$$

Пример

Пример 2. Вычислить информационный объем текста размером в 2000 символов, в записи которого использован алфавит компьютерного представления текстов мощностью 256.

Решение. В данном алфавите информационный вес каждого символа равен 1 байту (8 бит). Следовательно, информационный объем текста равен 2000 байт.

Если пересчитать информационный объем текста из примера 2 в килобайты, то получим:

$$2000 \text{ байт} = 2000/1024 \approx 1{,}9531 \text{ Кб}$$

Пример

Пример 3. Объем сообщения, содержащего 2048 символов, составил 1/512 часть Мегабайта.

Каков размер алфавита, с помощью которого записано сообщение?

Пример

Решение. Переведем информационный объем сообщения из мегабайтов в биты. Для этого данную величину умножим дважды на 1024 (получим байты) и один раз — на 8:

$$\frac{1 \cdot 1024 \cdot 1024 \cdot 8}{512} = \frac{2^{10} \cdot 2^{10} \cdot 2^{3}}{2^{9}} = 2^{14} \quad \text{áèò}$$

Поскольку такой объем информации несут 2048 символа (К), то на один символ приходится:

$$i = \frac{I}{K} = \frac{2^{14}}{2048} = \frac{2^{14}}{2^{11}} = 2^{3} = 8 \ \text{áèò}$$

Отсюда следует, что размер (мощность) использованного алфавита равен $2^8 = 256$ символов.

Пример

Пример 4. В алфавите племени МУМУ всего 4 буквы (А, У, М, К), один знак препинания (точка) и для разделения слов используется пробел. Подсчитали, что в популярном романе содержится всего 10 000 знаков, из них:

букв А — 4000,

букв У — 1000,

букв М — 2000,

букв К — 1500,

точек — 500,

пробелов — 1000.

Какой объем информации содержит книга?

Пример

Решение.

Поскольку объем книги достаточно большой, то можно допустить, что вычисленная по ней частота встречаемости в тексте каждого из символов алфавита характерна для любого текста на языке МУМУ.

Подсчитаем частоту встречаемости каждого символа во всем тексте книги (т.е. вероятность) и информационные веса символов:

Пример

Буква *А*: 4000/10 000 = 0,4;

$i_A = \log_2(1/0,4) = 1,321928$ бит

Буква У: 1000/10 000 = 0,1;

$i_У = \log_2(1/0,1) = 3,1928$ бит

буква М: 2000/10 000 = 0,2;

$i_М = \log_2(1/0,2) = 2,321928$ бит

буква К: 1500/10 000 = 0,15;

$i_К = \log_2(1/0,15) = 2,736966$ бит

точка: 500/10 000 = 0,05;

$i_{точка} = \log_2(1/0,05) = 4,321928$ бит

пробел: 1000/10 000 = 0,1

$i_{пробел} = \log_2(1/0,1) = 3,321928$ бит

Задачи

Общий объем информации в книге вычислим как сумму произведений информационного веса каждого символа на число повторений этого символа в книге:

$$I = i_{А} \cdot n_{А} + i_{У} \cdot n_{У} + i_{М} \cdot n_{М} + i_{К} \cdot n_{К} + i_{точка} \cdot n_{точка} +$$
$$+ i_{пробел} \cdot n_{пробел} = 1.321928 \cdot 4000 +$$
$$+ 3.1928 \cdot 1000 + 2.321928 \cdot 2000 +$$
$$+ 2.736966 \cdot 1500 + 4.321928 \cdot 500 +$$
$$+ 3.321928 \cdot 100 = 22\ 841.84 \text{ бита}$$