Единицы количества информации: вероятностный и объемный подходы

Определить понятие «количество информации» довольно сложно. В решении этой проблемы существуют два основных подхода. Исторически они возникли почти одновременно. В конце 40-х годов XX века один из основоположников кибернетики, американский математик Клод Шеннон, развил вероятностный подход к измерению количества информации, а работы по созданию ЭВМ привели к «объемному» подходу.

Вероятностный подход

Рассмотрим в качестве примера опыт, связанный с бросанием правильной игральной кости, имеющей N граней. Результаты данного опыта могут быть следующие: выпадение грани с одним из следующих знаков: 1, 2, ..., N.

Введем в рассмотрение численную величину, измеряющую неопределенность — **энтропию** (обозначим ее H). Согласно развитой теории, в случае равновероятного выпадания каждой из граней величины N и H связаны между собой **формулой Хартли** $H = \log_2 N$.

Единицы количества информации: вероятностный и объемный подходы

Важным при введении какой-либо величины является вопрос о том, что принимать за единицу ее измерения. Очевидно, H будет равно единице при $N = 2$. Иначе говоря, в качестве единицы принимается количество информации, связанное с проведением опыта, состоящего в получении одного из двух равновероятных исходов (примером такого опыта может служить бросание монеты, при котором возможны два исхода: «орел», «решка»). Такая единица количества информации называется «бит».

В случае, когда вероятности P_i результатов опыта (в примере, приведенном выше, — бросания игральной кости) неодинаковы, имеет место **формула Шеннона** $H = -\sum_{i-1}^{N} P_i \times \log_2 P_i$. В случае равновероятности событий $P_i = \frac{1}{N}$, и формула Шеннона переходит в формулу Хартли.

В качестве примера определим количество информации, связанное с появлением каждого символа в сообщениях, записанных на русском языке. Будем считать, что русский алфавит состоит из 33 букв и знака «пробел» для разделения слов. По формуле Хартли $H = \log_2 34 \approx 5{,}09$ бит.

Единицы количества информации: вероятностный и объемный подходы

Однако в словах русского языка (равно как и в словах других языков) различные буквы встречаются неодинаково часто. Ниже приведена табл. 1.1 вероятностей частоты употребления различных знаков русского алфавита, полученная на основе анализа очень больших по объему текстов.

Воспользуемся для подсчета H формулой Шеннона: $H \approx 4{,}72$ бит. Полученное значение H, как и можно было предположить, меньше вычисленного ранее. Величина H, вычисляемая по формуле Хартли, является максимальным количеством информации, которое могло бы приходиться на один знак.

Аналогичные подсчеты H можно провести и для других языков, например, использующих латинский алфавит — английского, немецкого, французского и др. (26 различных букв и «пробел»). По формуле Хартли получим $H = \log_2 27 \approx 4{,}76$ бит.

Единицы количества информации: вероятностный и объемный подходы

Таблица 1.1

Частотность букв русского языка

i	Символ	$P(i)$	i	Символ	$P(i)$	i	Символ	$P(i)$
1	_	0,175	12	Л	0,035	23	Б	0,014
2	О	0,090	13	К	0,028	24	Г	0,012
3	Е	0,072	14	М	0,026	25	Ч	0,012
4	Ё	0,072	15	Д	0,025	26	Й	0,010
5	А	0,062	16	П	0,023	27	Х	0,009
6	И	0,062	17	У	0,021	28	Ж	0,007
7	Т	0,053	18	Я	0,018	29	Ю	0,006
8	Н	0,053	19	Ы	0,016	30	Ш	0,006
9	С	0,045	20	З	0,016	31	Ц	0,004
10	Р	0,040	21	Ь	0,014	32	Щ	0,003
11	В	0,038	22	Ъ	0,014	33	Э	0,003
						34	Ф	0,002

Единицы количества информации: вероятностный и объемный подходы

Рассмотрим алфавит, состоящий из двух знаков 0 и 1. Если считать, что со знаками 0 и 1 в двоичном алфавите связаны одинаковые вероятности их появления ($P(0) = P(1) = 0{,}5$), то количество информации на один знак при двоичном кодировании будет равно $H = \log_2 2 = 1$ бит.

Таким образом, количество информации (в битах), заключенное в двоичном слове, равно числу двоичных знаков в нем.

Объемный подход

В двоичной системе счисления знаки 0 и 1 называют **битами** (*bit* — от английского выражения *Binary digiTs* — двоичные цифры). В компьютере бит является наименьшей возможной единицей информации. Объем информации, записанной двоичными знаками в памяти компьютера или на внешнем носителе информации, подсчитывается просто по числу требуемых для такой записи двоичных символов. При этом, в частности, невозможно нецелое число битов (в отличие от вероятностного подхода).

Для удобства использования введены и более крупные, чем бит, единицы количества информации. Так, двоичное слово из восьми знаков содержит один **байт**

Единицы количества информации: вероятностный и объемный подходы

информации. 1024 байта образуют килобайт (Кбайт), 1024 килобайта — мегабайт (Мбайт), а 1024 мегабайта — гигабайт (Гбайт).

Между вероятностным и объемным количеством информации соотношение неоднозначное. Далеко не всякий текст, записанный двоичными символами, допускает измерение объема информации в вероятностном (кибернетическом) смысле, но заведомо допускает его в объемном. Далее, если некоторое сообщение допускает измеримость количества информации в обоих смыслах, то это количество не обязательно совпадает, при этом кибернетическое количество информации не может быть больше объемного.

В прикладной информатике практически всегда количество информации понимается в объемном смысле.