Осн. структурная схема системы связи, рассматриваемая в Т. и., приведена на рис. 1. Информацию в виде сообщений создаёт источник сообщений. Сообщения представляют собой слова или наборы слов, записанные буквами нек-рого алфавита. Источниками сообщений могут быть человеческая речь, тексты на любых естеств. или формальных языках, данные систем сбора и обработки информации, а также нек-рые математич. модели — вероятностные процессы, создающие последовательности

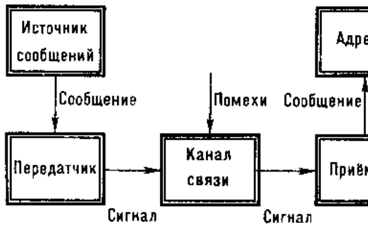

Рис. 1.

букв. Передатчик преобразует передаваемое сообщение в сигнал, соответствующий физ. природе канала связи. Канал связи — это среда для передачи сигнала от передатчика к приёмнику. При прохождении сигнала по каналу на него могут воздействовать помехи, вносящие искажения в значения информац. параметров сигнала. Приёмник восстанавливает по принятому в общем случае с искажениями сигналу исходное сообщение. Восстановленное сообщение поступает адресату — нек-рому лицу или техн. устройству.

Источники сообщений, рассматриваемые в теории информации, имеют статистич. характер, т. е. появление каждого из возможных сообщений (полный набор к-рых предполагается заранее известным) определяется соответствующей априорной вероятностью. Согласно Шеннону [1], считается, что чем больше априорная вероятность данного сообщения, тем меньше неопределённости относительно его действительного появления и, следовательно, тем меньше информации оно несёт. Если вероятность появления сообщения — единица, т. е. его появление достоверно, то неопределённости нет и считается, что сообщение не несёт информации.

Для оценки кол-ва информации в сообщении в Т. и. используется логарифмич. мера, введённая Р. Хартли [2], вероятностная интерпретация к-рой была дана в работах Шеннона [1]. Если вероятность появления сообщения x есть $p(x)$, причем $0<p(x)<1$, то количество информации $I(x)$, содержащееся в сообщении, определяется ф-лой:

$$I(x) = -\log p(x). \qquad (1)$$

Ф-ла (1) определяет кол-во информации с точностью до основания логарифма, т. е. с точностью до пост. множителя. Как правило, в качестве основания логарифма выбирается число 2 и единицей кол-ва информации является бит, что соответствует используемой в вычислит. технике двоичной системе счисления. При любом основании логарифма $I(x)\geqslant 0$, $I(x)=0$ при $p(x)=1$ и $I(x)\to\infty$ при $p(x)\to 0$ (рис. 2).

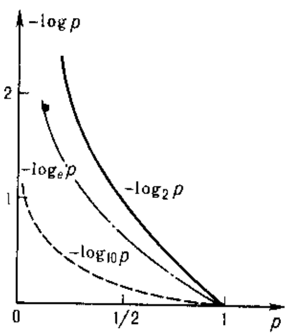

Рис. 2.

Если x_1 и x_2 — сообщения от двух независимых источников с вероятностями появления $p(x_1)$ и $p(x_2)$, то вероятность их совместного появления $p(x_1, x_2) = p(x_1)\cdot p(x_2)$, а соответствующее кол-во информации

$$I(x_1, x_2) = I(x_1) + I(x_2).$$

Это аддитивное свойство логарифмич. меры служит основанием для выбора её в качестве меры кол-ва информации, т. к. соответствует интуитивным представлениям о суммировании кол-ва информации, содержащегося в независимых сообщениях.

Для сообщений $x_1, \dots, x_n$, создаваемых источником с вероятностями $p_1, \dots, p_n$, причём $\sum_{i=1}^{n} p_i = 1$, Шеннон ввёл ср. меру кол-ва информации усреднением по множеству сообщений

$$H = -p_1 \log p_1 - \dots - p_n \log p_n. \qquad (2)$$

Логарифм здесь, как и в ф-ле (1), обычно берётся по основанию 2. Ф-ция H, характеризующая информац. свойства источника сообщений, наз. энтропией, т. к. по форме она совпадает с *энтропией* в статистич. физике, характеризующей априорную неопределённость нахождения статистич. системы в состояниях $x_1, \dots, x_n$, имеющих вероятности $p_1, \dots, p_n$. Очевидна прямая аналогия ф-лы (1) для кол-ва информации в сообщении и ф-лы Больцмана для физ. энтропии S:

$$S = k \ln W + \text{const},$$

где k — постоянная Больцмана, W — *термодинамическая вероятность*. В самой Т. и. и её приложениях эта аналогия с физикой не играет, однако, существенной роли.

Если в ф-ле (2) лишь одна из вероятностей равна единице, а остальные — нули, неопределённости в появлении сообщений нет и $H=0$. Если сообщения равновероятны: $p_1 = \dots = p_n = 1/n$, и неопределённость в том, какое из них появится, максимальна, то $H = \log n$. Это значение энтропии является максимальным и равно в этом случае кол-ву информации, получаемому от каждого отд. сообщения.

Для источника, создающего два сообщения x_1 и x_2, к-рые можно закодировать в двоичном коде как 0 и 1 соответственно, при вероятностях сообщений p и $q = 1 - p$

$$H = -p \log_2 p - q \log_2 q.$$

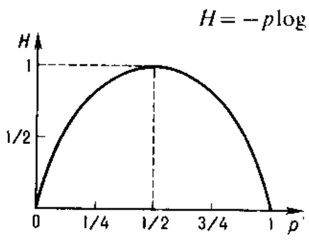

Рис. 3.

График энтропии для этого случая приведён на рис. 3. Энтропия максимальна, когда априорная неопределённость максимальна, т. е. при $p, q = 1/2$. При этом $H = 1$ бит, что соответствует одному двоичному символу (букве), используемому в кодах сообщений.

Если источник создаёт четыре сообщения x_1, x_2, x_3, x_4, то их можно закодировать в двоичном коде так: 00, 01, 10 и 11. При $p_1 = \dots = p_4 = 1/4$ энтропия максимальна, $H = 2$ бит, что соответствует двум двоичным символам, используемым для кодирования сообщений. Вообще для источника, создающего n сообщений, макс. значение энтропии $H = \log_2 n$, что соответствует мин. числу двоичных символов в кодовых словах одинаковой длины, образующих равномерные коды и необходимых для кодирования n равновероятных сообщений.

Если сообщения не являются равновероятными, то для экономии ср. времени на их передачу по каналу связи предпочтительно использование неравномерных кодов, образованных более короткими кодовыми словами для более вероятных сообщений и более длинными — для менее вероятных сообщений. Для n кодовых слов, имеющих $l_1, \dots, l_n$ символов, средняя длина слова (сообщения) определяется ф-лой

$$L = \sum_{i=1}^{n} l_i p_i,$$

где p_i — вероятности появления соответств. слов (сообщений). Энтропия задаёт ниж. границу для L, т. е. $L \geqslant H$. Уменьшение L, т. е. приближение L к H и как следствие уменьшение ср. времени передачи сообщений, возможно за