Информационная система
(система передачи инф-и)

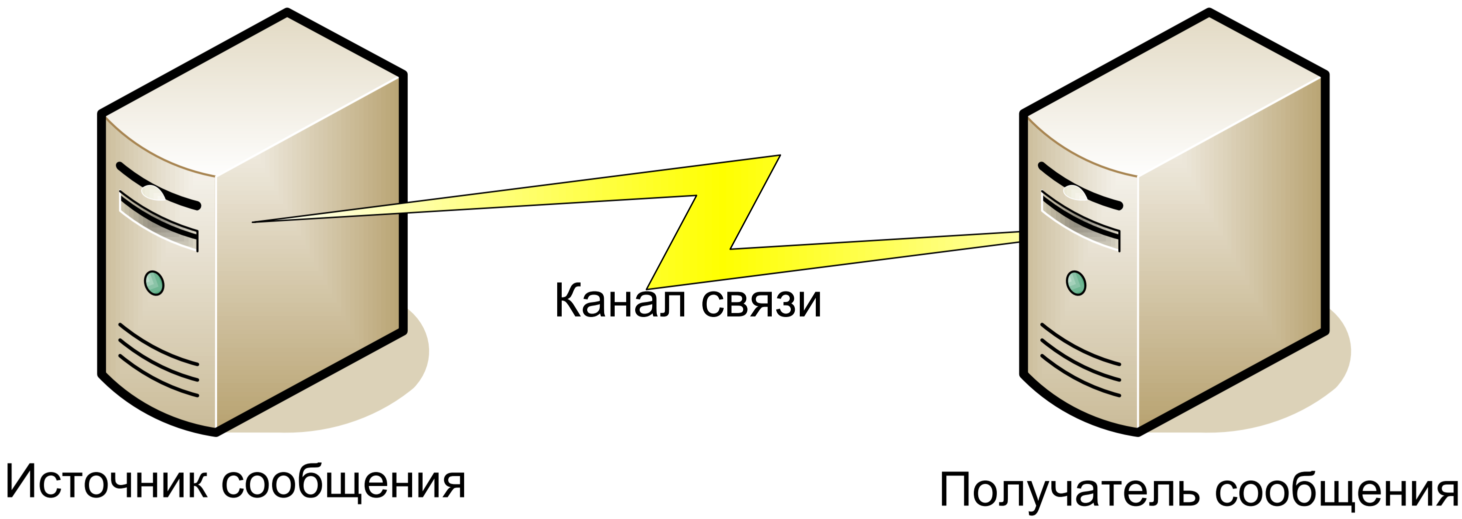
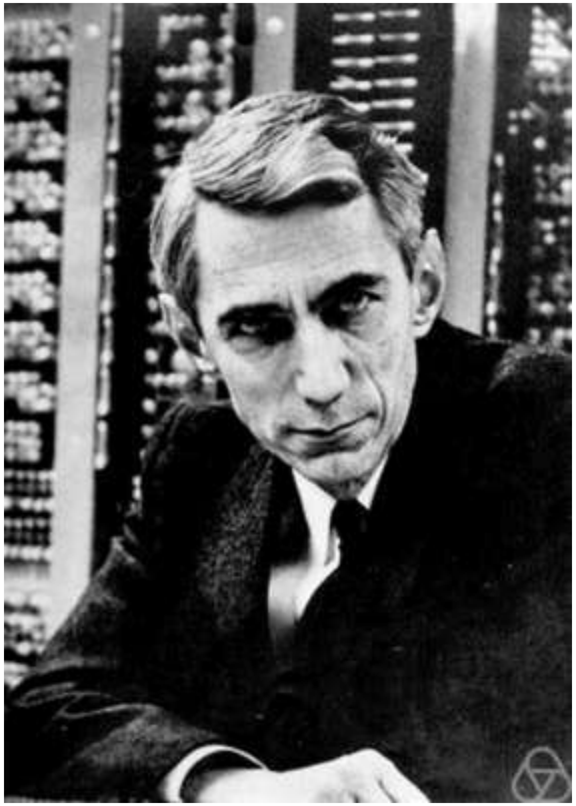


Рис1.



Claude Elwood Shannon (April 30, 1916 – February 24, 2001) was an American mathematician, electrical engineer, and cryptographer known as "the father of information theory".

Shannon, C. E. "A Mathematical Theory of Communication." *The Bell System Technical J.* **27**, 379-423 and 623-656, July and Oct. 1948:
http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf
http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

Основные характеристики элементов ИС

Определение 1. **Алфавит** – конечная совокупность символов (знаков), с помощью которых можно представить любое сообщение в ИС: **$A\{a_i\}$**

a_i – i- ый символ алфавита

Определение 2. **Мощность алфавита** – количество символов, составляющих алфавит: **N(A)**

Вероятность того, что произвольный символ **ξ** произвольного документа (текст, база данных, текст программы) будет буквой «$\mathbf{a_i}$»: **$P(\xi = a_i) = p(a_i)$**

$$\sum_{i=1}^{N} P(a_i) = 1 \qquad (1)$$

Пример 1.

$$\mathrm{P}(\xi = e) = 0.13;$$
$$\mathrm{P}(\xi = q) = 0.0011; \mathrm{P}(\xi = z) = 0.0007.$$

Определение 3. Информационной характеристикой алфавита (источника сообщений на основе этого алфавита) является ***энтропия***.

$$\mathrm{H}_S(A) = -\sum_{i=1}^{N} P(a_i) * \log_2 P(a_i), \qquad (2)$$

где $i = \overline{1, N}$, a_i – элемент алфавита, $P(a_i)$ – вероятность $P(\xi = a_i)$.

С физической точки зрения *энтропия показывает, какое количество информации (бит) приходится в среднем на один символ алфавита*.

Пример 2. **N=10 ; p(ai)= ….**

Энтропия двоичного алфавита

A{0, 1}, N=2

$P(\xi = 0) = p(0); P(\xi = 1) = p(1)$

Энтропия двоичного алфавита на основе (2):

$$H(A_2) = -p(0)*\log_2(p(0)) - p(1)*\log_2(p(1)) \qquad (3)$$

$p(0) + p(1) =$

Обозначим **$p(0) = 1 – p(1)$**

С учетом этого (3) имеет вид:

$$H(A_2) = -(1-p(1))*\log_2(1-p(1)) - p(1)*\log_2(p(1)) \quad (4)$$

Если $p(1) = 0$, то $H(A_2) =$

Если $p(0) = 0$, то $H(A_2) =$ (см (3))

$$\frac{d\,H(A_2)}{d\;p(1)} = 0 \longrightarrow p(1) = 0.5 \longrightarrow p(0) = 0.5$$

Подставим последние значения в (3):

$H(A2) = -\,0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$ **(бит)**

График **H(**A2**)** = **f** (p(1), p(0))

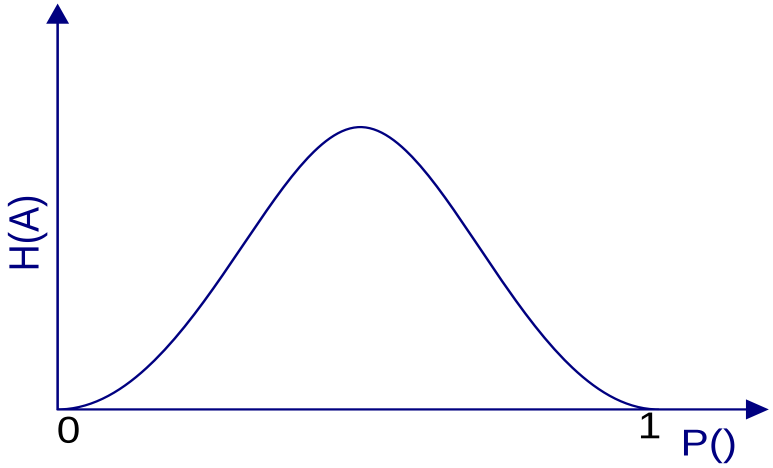


Рис.2

Энтропия Хартли

Имеет место при ***const* p(a$_i$)**

Тогда **p(a$_i$)** при известном **N** равно **…..**

Подставим **p(a$_i$)** в (2):

$$H_c\,(A) = -\sum_{1}^{N} (1/N) * \log_2 (1/N) = \log_2 N \qquad (5)$$

Количество информации

Определение 4. Количество информации **I** в произвольном сообщении X_k, где **k** – число символов в сообщении определяется соотношением

$$I(X_k) = H(A) * k \qquad (6)$$

Пример 3. X_k **= 'I love you'**

Подсчитать $I(X_k$ **= 'I love you')** при **H(A) = 4.7** бит

Энтропийная оценка информации при ее передаче

Пусть в ИС сообщение $X_k = x_1, x_2, \dots x_i, \dots . x_k$ на входе канала формируется на основе $A = \{a_i\}, \; i = 1 \dots N$

Сообщение на выходе канала ($Y_k = y_1, y_2, \dots y_j, \dots . y_k$) формируется на основе того же алфавита : A

При передаче сообщения по каналу могут появляться ошибки.

Пример 4. Двоичный симметричный канал (ДСК)

N=2

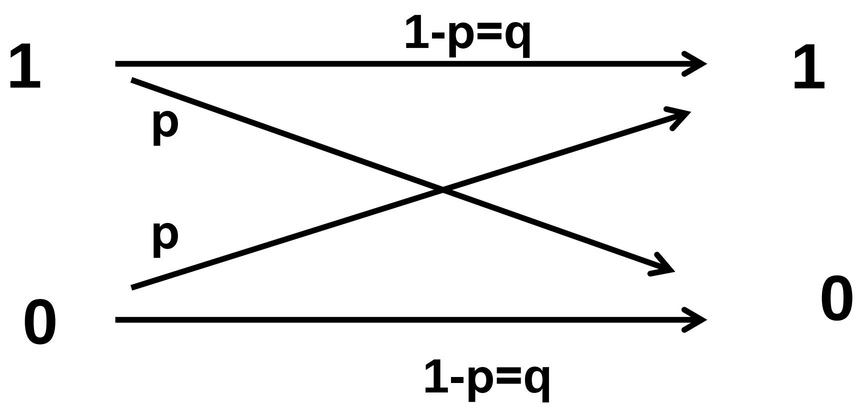


Рис.3

Обозначим условную вероятность события:

$P(x_i \mid y_j)$: $P(0|0) = P(1|1) = q$; $P(1|0) = P(0|1) = p$

Стоит задача: определить количественно потери информации, вызванные несовершенством ИС (канала), т.е. при $p > 0$

Задача относится к области **проверки гипотез и принятия статистических решений.**

Математич. основа – теорема Байеса:

Совместная вероятность случайных событий А и В:

$$P(A,B) = P(A|B)\,P(B) = P(B|A)\,P(A) \quad (7)$$

Или

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (8)$$

В соответствии с (8) для ДСК можно записать (используя дискретную форму теоремы Байеса):

$$P(x_i \mid y_j) = \frac{P(y_j|x_i)\, P(x_i)}{P(y_j)} \qquad (9)$$

где

$$P(y_j) = \sum_{i,j=1}^{N} P(y_j|x_i)\, P(x_i) \qquad (10)$$

В общем случае i и j могут принимать различные значения.

В соответствии с (9) и (10) для ДСК:

$P(x=0|y=0) = [P(y=0|x=0) * P(x=0)] / [P(y=0|x=0) * P(x=0) +$
$+ P(y=0|x=1) * P(x=1)]$

$P(x=1|y=0) = [P(y=0|x=1) * P(x=1)] / [P(y=0|x=0) * P(x=0) +$
$+ P(y=1|x=1) * P(x=1)]$

$P(x=0|y=1) = [P(y=1|x=0) * P(x=0)] / [P(y=1|x=0) * P(x=0) +$
$+ P(y=1|x=1) * P(x=1)]$

$P(x=1|y=1) = [P(y=1|x=1) * P(x=1)] / [P(y=1|x=0) * P(x=0) +$
$+P(y=1|x=1) * P(x=1)]$

Если $p > 0$, то это можно трактовать как неоднозначность (по Шеннону – *equivocation*) между переданным и принятым сообщениями.

Эта неоднозначность определяется как **условная энтропия (частная)** сообщения x_i, обусловленная полученным сообщением Y:

$$H(Y|x_i) = - \sum_{J=1} P(y_j|x_i) * \log P(y_j|x_i) \qquad \textbf{(11)}$$

В соответствии с (11):

$$H(Y|x=0)= - P(y=0|x=0) * \log P(y=0|x=0) –$$
$$- P(y=1|y=1) * \log P(y=1|y=1) = - q \log q - p \log p$$

аналогично

$$H(Y|x=1)= - P(y=0|x=1) * \log P(y=0|x=1) –$$
$$- P(y=1|x=1) * \log P(y=1|x=1) = - p \log p - q \log q \qquad \textbf{(12)}$$

или в ином виде:

$$H(X|y_j) = - \sum_{i=1} P(x_i|y_j) * \log P(x_i|y_j) \qquad \textbf{(12)}$$

В соотв. с (12) можно определить, какому количеству инф соот-т один символ сообщения X , если на вых получен 0:

$$H(X|y=0)= - P(x=0|y=0) * \log P(x=0|y=0) -$$
$$- P(x=1|y=0) * \log P(x=1|y=0) = - q \log q - p \log p$$

То же, если на вых получена 1: **(13)**

$$H(X|y=1)= - P(x=0|y=1) * \log P(x=0|y=1) -$$
$$- P(x=1|y=1) * \log P(x=1|y=1) = - p \log p - q \log q$$

<u>**Определение**</u>. *Условной энтропией <u>Источника дискретного сообщения</u>* **X** в ДСК называем величину

$$H(Y|X) = - \Sigma\ P(x_i)\ H(Y|x_i) = \Sigma_i\ P(x_i)\ \Sigma_{i,\ j}\ P(y_i|x_j) * \log P(y_i|x_j) =$$

$$= P(x=0)*(- p\ log_2\ p - q\ \ log_2\ q) + P(x=1)*\ (- p\ log_2\ p - q\ \ log_2\ q)$$
$$= \ - p\ log_2\ p - q\ log_2\ q\ , \qquad (14)$$

так как $P(x=0) + P(x=1) = 1$

H(Y|X) – <u>энтропия со стороны источника сообщения</u>

H(Y|X) означает <u>**потерю информации на каждый символ переданного сообщения**</u>

Определение. **Условная энтропия H(X|Y) источника дискретного сообщения** в ДСК рассчитывается на основе формулы

$$H(X|Y) = P(y=0)H(X|y=0) + P(y=1)H(X|y=1) \qquad (15)$$

H(X|Y) соответствует энтропии со стороны получателя сообщения

H(X|Y) означает средний объем информации, соответствующей одному из символов сообщения X, относительно принятого сообщения Y

или потерю информации на каждом символе отправленного сообщения

Пример 5. Пусть известно, что P(*X=0*) = P(*X=1*) =0.5 и p=0.01.

Из (14) определим

H(X|Y) = - p log p - q log q = -0.01 * log 0.01 – 0.99 * log 0.99 = = 0.081 *бит*

Шеннон показал, что эффективная информация на выходе канала относительно входной в расчете на 1 символ (Эфф энтропия алфавита) составляет:

$$\mathbf{H_e = H(X) - H(X|Y)} \qquad (15)$$

Для случая из примера 5 $\mathbf{H_e}$ = 0.919 *бит*

Пример 6. Какое количество информации будет передано по каналу связи за 1 час при скорости передачи 1 Мбит/с, вероятность ошибки равна 0.5?

Пример 7. Какое количество информации будет передано по каналу связи за 1 час при скорости передачи 1 Мбит/с, вероятность ошибки равна 1.0?