Современные проблемы информатики

Источники информации. Энтропия

Фионов Андрей Николаевич

СибГУТИ

2020

Возникновение теории информации

Клод Шеннон (Claude Shannon) 1948

Возникновение теории информации

Reprinted with corrections from *The Bell System Technical Journal*,
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

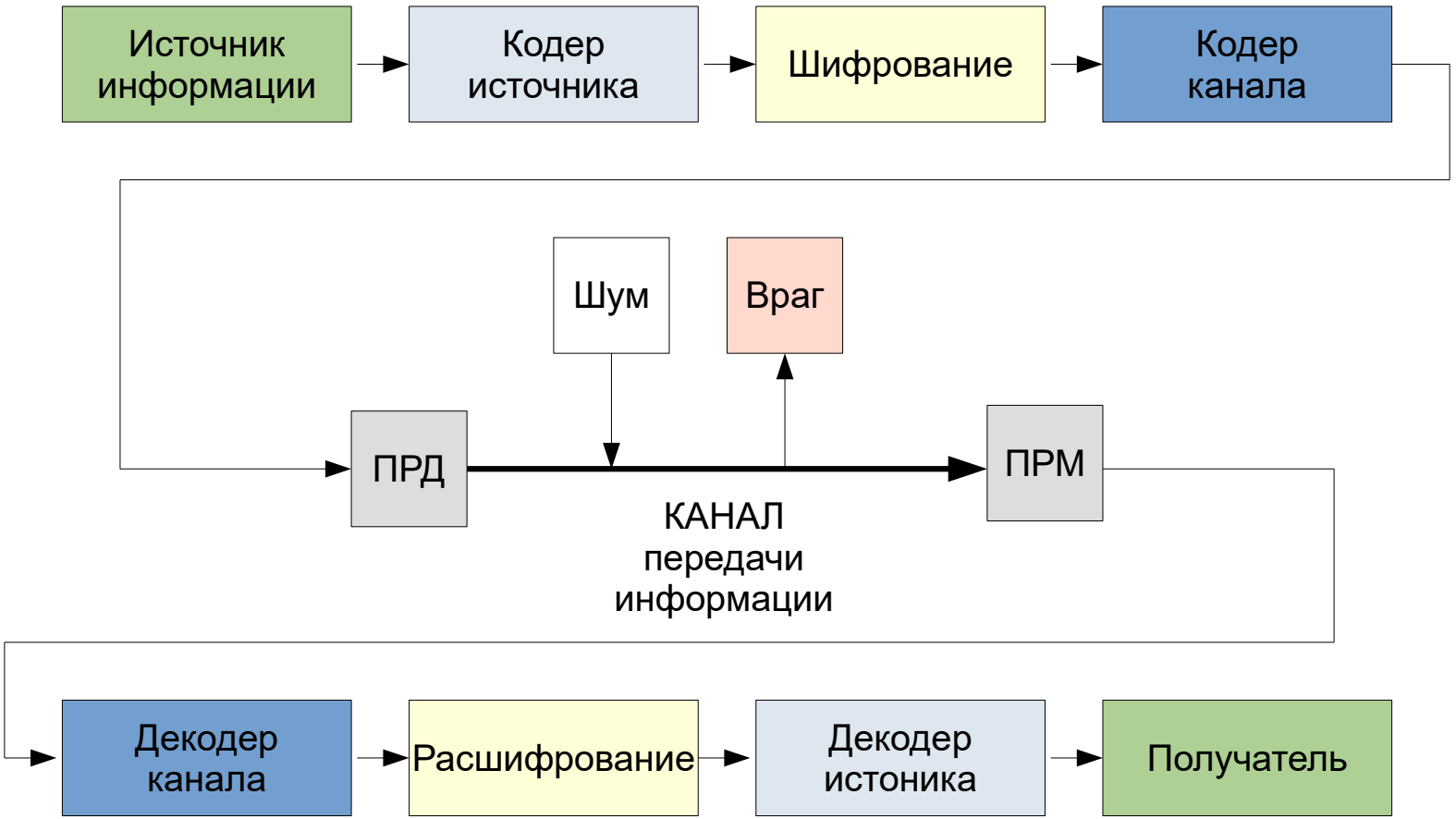
The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all

Литература

- Кудряшов Б. Д. Теория информации
  https://books.ifmo.ru/file/pdf/723.pdf
- Ватолин Д., Ратушняк А., Смирнов М., Юкин В. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. М.: ДИАЛОГ-МИФИ, 2002. 384 с.
  http://www.compression.ru/book/
- Сайт "Всё о сжатии данных, изображений и видео"
  http://www.compression.ru/
- A Mathematical Theory of Communication By C. E. SHANNON
- Thomas M. Cover, Joy A. Thomas, "Elements of Information Theory"

Система передачи данных (СПД)



Источник информации

Источник информации (вероятностный источник) — дискретный случайный процесс, характеризуемый конечным алфавитом $A = \{a_1, a_2, \ldots, a_N\}$ и распределением вероятностей последовательностей букв (символов) этого алфавита $\mathbb{P} = P(X_1, X_2, \ldots, X_n)$, $X_i \in A$, $n = 1, 2, 3, \ldots$.

Обозначения:
X – случайная величина, x – её значение
$P(X)$ – распределение, $p(x)$ – вероятность

$P(X = a_1) = p(a_1) = p_1$
$P(X = x) = p(x) = p_x$

$x_1 x_2 \ldots x_n$ – сообщение источника

Источник информации

Стационарность (неформально).
Источник называется *стационарным*, если его вероятностное описание не меняется со временем.

Эргодичность (очень неформально).
Источник называется *эргодическим*, если все порождаемые им сообщения имеют одинаковое вероятностное описание.

Модели источников

1. Источник без памяти — все порождаемые символы независимы

Описание: алфавит + вероятности букв

Пример.
$A = \{a, b, c\}$, $\mathbb{P} = (p(a), p(b), p(c))$

Модели источников

2. Марковский источник — очередной символ зависит от μ предыдущих символов

μ – порядок источника (длина контекста)

Описание: алфавит, набор условных вероятностей букв

Пример.
$A = \{a, b, c\}$, $\mu = 1$,

$$\mathbb{P} = \begin{pmatrix} p(a|a), & p(b|a), & p(c|a) \\ p(a|b), & p(b|b), & p(c|b) \\ p(a|c), & p(b|c), & p(c|c) \end{pmatrix}$$

Модели источников

3. Древовидный источник (tree source) — символы зависят от контекстов *переменной* длины

Описание: алфавит, набор условных вероятностей букв

Пример.
$A = \{a, b\}$,

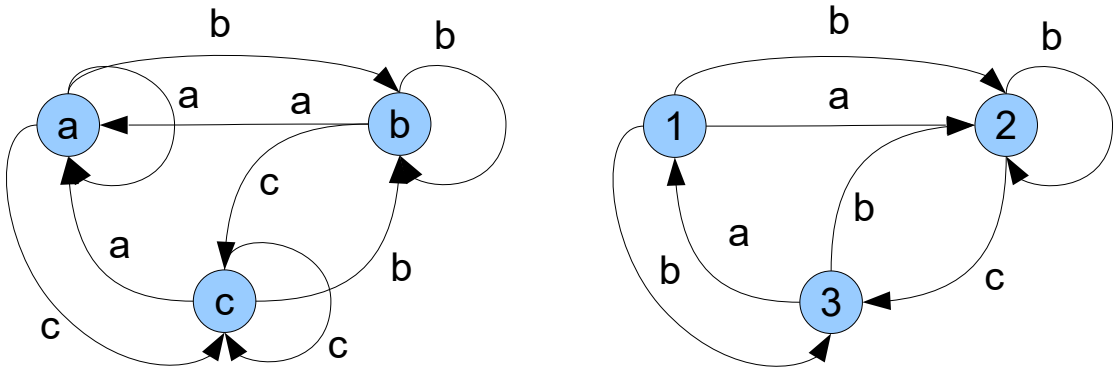
$$\mathbb{P} = \begin{pmatrix} p(a|a), & p(b|a) \\ p(a|ba), & p(b|ba) \\ p(a|bb), & p(b|bb) \end{pmatrix}$$

Модели источников

4. Автоматный источник (FSM source) — последовательность генерируемых символов определяется работой автомата с конечным числом состояний

Описание: алфавит, граф автомата с набором переходных вероятностей

Пример.



Модели источников

5. Грамматический источник — последовательность генерируемых символов определяется правилами некоторой грамматики

Описание: алфавит, грамматика, набор переходных вероятностей

Пример.

$X = 01001100011100001111 0\ldots$ Описание: $0^k 1^k,\ k = 1, 2, 3, \ldots$

$X = 01011011101111011111 0\ldots$ Описание: $01^k,\ k = 1, 2, 3, \ldots$

Марковские источники

Определение стационарных вероятностей (на примере)

$A = \{a, b\}$, $\mu = 1$, $\mathbb{P} = (p(a|a), p(b|a), p(a|b), p(b|b))$

$$\begin{cases} p(a) = p(a)p(a|a) + p(b)p(a|b) \\ p(b) = p(a)p(b|a) + p(b)p(b|b) \\ p(a) + p(b) = 1 \end{cases}$$

Марковские источники

Задание источника через вероятности пар, троек, ... (на примере)

$A = \{a, b\}$, $\mu = 1$, $\mathbb{P} = (p(aa), p(ab), p(ba), p(bb))$

$p(a) = p(aa) + p(ab)$
$p(b) = p(ba) + p(bb)$

$p(aa) = p(a)p(a|a) \quad \Rightarrow \quad p(a|a) = p(aa)/p(a)$
$p(ab) = p(a)p(b|a) \quad \ldots$
$p(ba) = p(b)p(a|b)$
$p(bb) = p(b)p(b|b) \quad \Rightarrow \quad p(b|b) = p(bb)/p(b)$

Аппроксимация англоязычного текста

Буквы независимы и равновероятны

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD

Аппроксимация англоязычного текста

Буквы независимы, частота встречаемости как в англоязычном тексте

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

Аппроксимация англоязычного текста

Частота встречаемости пар букв как в англоязычном тексте

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

Аппроксимация англоязычного текста

Частота встречаемости троек букв как в англоязычном тексте

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

Аппроксимация англоязычного текста

Частота встречаемости слов как в англоязычном тексте

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TOOF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE

Аппроксимация англоязычного текста

Частота встречаемости пар слов как в англоязычном тексте

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

Аппроксимация англоязычного текста

Частота встречаемости пар слов как в англоязычном тексте

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

Информация

Определение. Информация (собственная информация, количество информации)

$$I(x) = \log \frac{1}{p(x)} = -\log p(x).$$

$\log p \equiv \log_2 p$, единица измерения – бит

вычисление: $\log_2 p = \ln p / \ln 2$

Информация

$p(x)$	$I(x)$
1	0
1/2	1
1/4	2
1/8	3
1/16	4

Энтропия

$A = \{a_1, a_2, \ldots, a_N\}$, $\mathbb{P} = (p_1, p_2, \ldots, p_N)$

Определение. Энтропия случайной величины $X \in A$ (энтропия распределения $\mathbb{P}$)

$$H(X) = H(\mathbb{P}) = -\sum_{i=1}^{N} p_i \log p_i \qquad (0 \log 0 = 0)$$

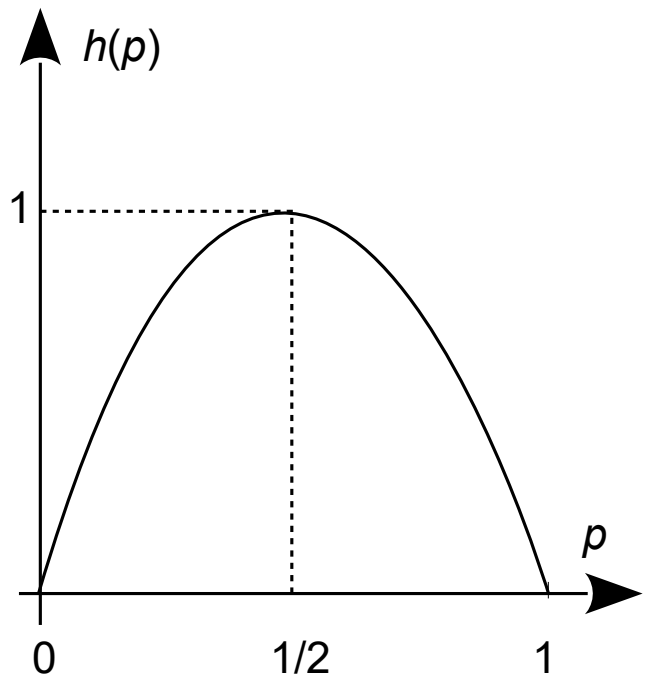
Пример.
$A = \{a, b, c\}$, $\mathbb{P} = (1/2, 1/4, 1/4)$: $H(\mathbb{P}) = 1.50$

$A = \{a, b, c\}$, $\mathbb{P} = (1/3, 1/3, 1/3)$: $H(\mathbb{P}) = 1.58$

Двоичная энтропия

$A = \{a, b\}, \quad p(a) = p, p(b) = 1 - p$

$$h(p) = -p \log p - (1 - p) \log(1 - p)$$



Свойства энтропии

1. $H(X) \geq 0 \quad (= 0, \text{ если } p_j = 1)$
2. $H(X) \leq \log N \quad (= \log N, \text{ если все } p_i = 1/N)$

Энтропия сообщения

Пусть источник порождает сообщение из двух букв $x, y \in A$ с распределениями $\mathbb{P}_x, \mathbb{P}_y$, $p(x, y) = P(X = x, Y = y)$.

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$$

Условная энтропия

Первая буква – x, вторая буква – y: $p(x)$, $p(y|x)$, $p(x,y) = p(x)p(y|x)$

$$H(Y|X=x) = -\sum_{y} p(y|x) \log p(y|x)$$

Условная энтропия

$$H(Y|X) = \sum_x p(x) H(Y|X = x)$$

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x)$$

Свойства условной энтропии

1. $H(Y|X) \leq H(Y)$
2. $H(X,Y) = H(X) + H(Y|X)$
3. $H(X,Y) \leq H(X) + H(Y)$

Обобщение на произвольную длину сообщения

$$X^n = X_1, X_2, \ldots, X_n$$

$$H(X^n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots$$
$$+H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1)$$

Удельная энтропия (entropy rate)

$$h_n^+ = \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

$$h_n^- = H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1)$$

Предельная энтропия

Теорема. Для стационарного эргодического источника существует предельная энтропия при длине сообщения $n \to \infty$

$$h_\infty = h_\infty^+ = h_\infty^-.$$

Энтропия английского языка (Шеннон)

Алфавит = 26 букв + пробел

Символы равновероятны и независимы

$$h_0 = -\log(1/27) = 4.76$$

Символы генерируются с характерными частотами

$$h_1 = 4.03$$

Пары символов генерируются с характерными частотами

$$h_2 = 3.32$$

Тройки символов генерируются с характерными частотами

$$h_3 = 2.8$$

$$0.6 < h_\infty < 1.3$$

К О Н Е Ц