Belarusian State Technological University
Department of Information Systems and Technology

Pavel Urbanovich

INFORMATION PROTECTION

Part 4: INFORMATION SYSTEM. ENTROPY. AMOUNT OF INFORMATION.

pav.urb@yandex.by, p.urbanovich@belstu.by

Information System (IS)

Definition of Information Systems

"Information Systems(IS) is the study of complementary networks of hardware and software that people and organizations use to collect, filter, process, create, and distribute data."

"Information systems are combinations of hardware, software, and telecommunications networks that people build and use to collect, create, and distribute useful data, typically in organizational settings."

"**Information systems are interrelated components working together to** collect, process, store, and disseminate information to support decision making, coordination, control, analysis, and vizualization in an organization."

•These definitions focus on two different ways of describing information systems: the ***components*** that make up an information system and the ***role*** that those components play in an organization.

•There are many definitions, but most include the idea that information systems is at the intersection of technology, people and processes within an organization.

•That is, it combines the components of people, technology and process to achieve a goal.

The Components of IS

Hardware:
Hardware is the physical piece of technology: computer, support equipment (input and output devices, storage devices and communications devices and networks).

Software:
Computer programs (produce useful information from data).

Data:
Like programs, data are generally stored in machine-readable form on disk or tape until the computer needs them.

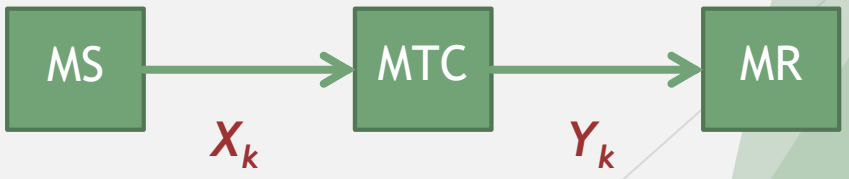
Procedures (technology):
"Procedures are to people what software is to hardware" is a common analogy that is used to illustrate the role of procedures in a system.

People (**users**; often: senders and recipients of the message (information)):
include "not only the users, but those who operate and service the computers, those who maintain the data, and those who support the network of computers".

Technical Means of IS

• **Message Source** (**MS**) - generates (creates) a message X_k **(signal)** consisting of k characters.
• **Message Transmission Channel** (**MTC**) or **Data Line**.
• **Message Recipient** (**MR**) -receives a message (Y_k which may be different from the sent one, for example, as a result of interference), consisting of k characters.
• **Analog** and **digital signals** are used to transmit information, usually through electric **signals**. In both these technologies, the information, such as any audio or video, is transformed into electric **signals**.

Digital sources – examples:
• Keyboard,
• Telegraph

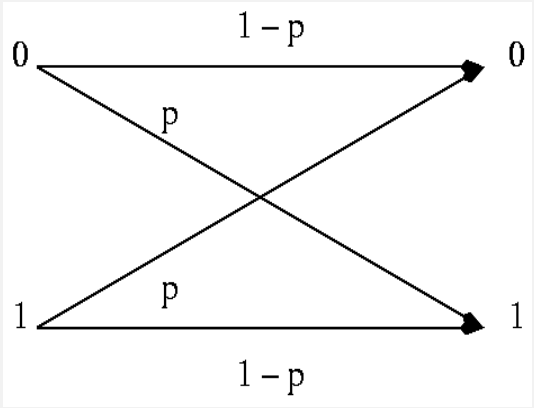


Discrete Digital Channel

- at the input and output of channel - discrete (digital) signals;
- a communications path that handles only digital signals.

A Binary Symmetric Channel (or BSC) is a common communications **channel** model used in coding theory and information theory.
In this model, a transmitter wishes to send a bit (a zero or a one), and the receiver receives a bit.
BSC is a channel with binary input and binary output and probability of error *p*; that is, if *X* is the transmitted random variable and *Y* the received variable, then the channel is characterized by the conditional probabilities:

$$P(y = 0 \mid x = 0) = 1 - p,$$
$$P(y = 0 \mid x = 1) = p,$$
$$P(y = 1 \mid x = 0) = p,$$
$$P(y = 1 \mid x = 1) = 1 - p,$$
$$1 - p = q,$$



$X=\{0,1\}$, $Y=\{0,1\}$

MS (X) **MR (Y)**

An **Analog signal** is any continuous signal for which the time varying feature (variable) of the signal is a representation of some other time varying quantity, i.e., analogous to another time varying signal. It differs from a digital signal in terms of small fluctuations in the signal which are meaningful.

Analog (Continuous) channel - at the input and output of channel - continuous (not digital, but for example, analog) signals.

Analog sources – examples:

- Thermometer, barometer,
- The vocal cords,
- Musical instruments,
- Electronic and electrical devices based on analog signals.

- An **analog signal** is a continuous function, with an unlimited number of values at different times.

The frequency (F) and period (T) of the signal are the most important system's parameters.

Alphabet of MS and MR

• All communication (and data processing) is achieved through the use of symbols.

• **Alphabet** $A=\{a_i\}$ of the MS (or MR) - the number of characters (symbols) from which the message is formed.

• In computer science, an alphabet is a finite non-empty set.

• **Finite Alphabet** - is a alphabet whose number of elements is finite; there exists a nonnegative integer ***N*** equal to the number of elements of this alphabet.

•**Capacity of Alphabet** is number of characters in the alphabet.

•The alphabet which used in computer science $A=\{a_i\}=\{0,1\}$ is called the **binary alphabet** because it contains of two symbols.

•**Number** (numeral) **systems** (notation): (decimal, binary (dual), hexadecimal, etc.) – is a system of writing numbers using a specific set of digits.

How to represent numbers in different systems?

Entropy and Amount of Information

Entropy of discrete source of information and Amount of information

- **Units of Information**

bit *(b) - basic, smallest and indivisible unit of digital information that can be processed by a computer (**BI**nary digi**T**)*

2^{10} b = 1024 b = 10^3 = 1Kb,
2^{20} b = = 10^6 = 1Mb,
2^{30} b = = 10^9 = 1Gb,
……

The **concept of entropy** has been borrowed from *thermodynamics* to determine the average amount of information from a message source.

• The **concept of information entropy** was introduced by **Claude Shannon** in his 1948 paper *„A Mathematical Theory of Communication”*.

• **Information entropy** is defined as the average amount of information produced by a probabilistic stochastic source of data.

• **The mathematical meaning of information entropy** is the logarithm of the number of available states of the system.

• The base (n) of the logarithm ($\log_n$) can be different, it defines the unit of **measurement of entropy**:
n = 2 – **bit**, n = e – **nat**, n = 10 – **dit** — ***decimal digit***.

•In 1928, the American engineer **R. Hartley** proposed a scientific approach to the evaluation of message:

Let the states of the system (alfabet) be equiprobable and have the probability ***p***, then the number of states ***N = 1/p*** .

The formula he proposed for evaluation was as follows:

$$H = \log (1/p) = \log N, \text{ bit.} \qquad (1)$$

The logarithm of the probability distribution is useful as a measure of entropy, because it is *additive for independent sources*.

Additivity (Latin: additivus-additive) is a property of quantities consisting in the fact that the value of a quantity corresponding to an entire object is equal to the sum of the values of the quantities corresponding to its parts:

$$F(a+b) = F(a) + F(b).$$

This formula (1) determines the **Hartley's entropy** under the condition of the same probability (***p=1/N***) of occurrence in an arbitrary message of an arbitrary symbol of the alphabet.

Task. 1

The ball is in one of three urns: A, B or C.
Determine how many bits of information contain a message that it is in the urn B.

Decision.

According to (1) such message contains $H = \log_2 3 = 1{,}585$ bit of information, because N=3.

Task. 2

How much information will be obtained when playing (one attempt) in a roulette with 32 sectors?

Task. 3

How much information is contained in the event, if I ask one of the students (anyone) a question?

C. Shannon's Entropy and Amount of Information in Message

C. Shannon proposed the formula for calculating the amount of information in the event of **various probabilities of events** in 1948.

- **Information binary entropy for independent random events** ***x*** *with* ***N*** possible states, distributed with probabilities p_i (i= 1, ..., N) is calculated by the formula:

$$H(x) = -\Sigma p_i(x) \log_2 p_i(x) \qquad (2)$$

If ***x*** is an alphabet consisting of ***N*** symbols, then (2) defines entropy of alphabet with different probabilities of the appearance of symbols in the message:

$$H(A) = -\Sigma p_i(a_i)\ log_2\ p_i(a_i), \qquad (3)$$

$A=\{a_i\}$, $i=1,2,\ldots, N$

The entropy of the alphabet corresponds to the average amount of information contained in one alphabet symbol (message).

- **Amount of information *I* in the message X_k**, *consisting of* ***k*** *characters with entropy of alphabet* ***H(A)*****:**

$$I(X_k) = k \cdot H(A)\ . \qquad (4)$$

Binary Alphabet Entropy

$A = \{0, 1\}$, $N = 2$
$P(\xi = 0) = p(0)$; $P(\xi = 1) = p(1)$

$$H(A_2) = -p(0)\log_2(p(0)) - p(1)\log_2(p(1)) \quad (5)$$

If $p(0) = 1 - p(1)$
then:

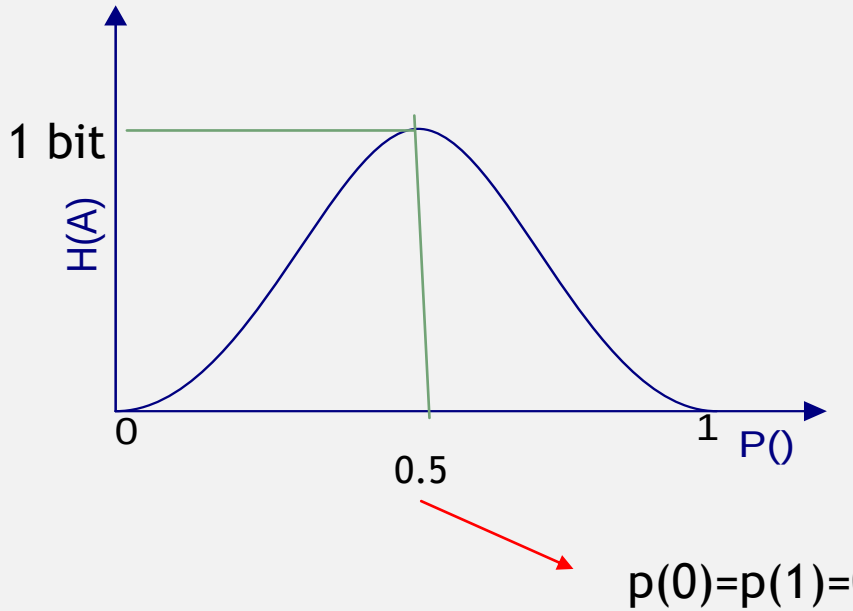
$$H(A_2) = -(1-p(1))\log_2(1-p(1)) - p(1)\log_2(p(1)) \quad (6)$$

$dH(A_2)/dp(1) = 0 \longrightarrow p(1) = 0.5$ and $p(0) = 0.5$

$H(A_2)|$with $p(1) = 0.5$, $p(0) = 0.5|$ = 1 bit

$H(A_2)|$with $p(1) \neq 0.5$, $p(0) \neq 0.5|$ = ?

Qualitative characteristic of the binary alphabet entropy at different values of the probabilities of the appearance of characters in the message



Conditional Entropy

In information theory, the **conditional entropy** (or **equivocation**) quantifies the remaining entropy (i.e. uncertainty) of a random variable Y given that the value of a second random variable X is known.

It is referred to as *the entropy of Y conditional on X*, and is written $H(Y|X)$.

Like other entropies, the conditional entropy is measured usually in bits.

On the other hand, for BSC the conditional entropy of the message X (X_k) due to the received message Y (Y_k):

$$H(X|Y) = - \sum P(x_i|y_j) \log_2 P(x_i|y_j) =$$
$$= - \sum P(y_j) \sum P(x_i|y_j) \log_2 P(x_i|y_j) \quad (7)$$

In accordance with (7) you can determine, how much information corresponds to one character of the message X (X_k) , if **0** is received at the output of channel:

$$H(X|y=0) = -P(x=0|y=0)\log_2 P(x=0|y=0) -$$
$$- P(x=1|y=0) \log_2 P(x=1|y=0) = - q \log_2 q - p \log_2 p \quad (8)$$

The same, if the output of the channel received **1**:

$$H(X|y=1) = -P(x=0|y=1)\log_2 P(x=0|y=1) - P(x=1|y=1)\log_2 P(x=1|y=1) = -p\log_2 p - q\log_2 q \quad (9)$$

Thus, the conditional entropy of the source of a discrete (binary) message ***X*** we called the value

$$H(X|Y) = P(y=0)H(X|y=0) + P(y=1)H(X|y=1) = -p\log p - q\log q \quad (10)$$

H (X|Y) means the average amount of information for the input symbol relative to the received message ***Y*** or the loss of information for each character of the transmitted message ***X***

Task. 1

Let it be known that $P(x=0) = P(x=1) = 0.5$ and $p = 0.01$.
Find how much information in this case will be lost for each bit of the message ($X=X_k$= 'Hello' is sent in ASCII codes)

Decision.

From (10) we define

$H(X|Y) = - p \log_2 p - q \log_2 q = -0.01 \log_2 0.01 - 0.99 \log_2 0.99 = 0.081$ bits

C.Shannon showed that the **effective information** at the channel output relative to the input per 1 symbol is:

$$H_e = H(X) - H(X|Y)$$

H_e is **Effective entropy**

Task. 2 Calculate how much information will be transmitted over a communication channel in 1 hour with transmission rate of 1 Mb/s, the error probability is 0.5?

Task. 3 Calculate how much information will be transmitted over a communication channel in 1 hour with transmission rate of 1 Mb/s, the error probability is 1.0?

References:

1. Dave Bourgeois and David T. Bourgeois, Information Systems for Business and Beyond, (URL: https://bus206.pressbooks.com/chapter/chapter-1/)

2.Shannon, Claude E. A Mathematical Theory of Communication, Bell System Technical Journal, July – October 1948, 27 (3), p. 379–423

3. Урбанович, П. П. Информационная безопасность и надежность систем : учебно-методическое пособие по одноименному курсу для студентов специальности 1-40 01 02-03 "Информационные системы и технологии" / П. П. Урбанович, Д. М. Романенко, Е. В. Романцевич. - Минск : БГТУ, 2007. - 87 с. (URL: http://elib.belstu.by/handle/123456789/2937)

4. Урбанович, П. П. Защита информации и надежность информационных систем: пос. для студ. вузов спец. 1-40 05 01-03 «Информационные системы и технологии (издательско-полиграфический комплекс)» / П. П. Урбанович, Д. В. Шиман.- Минск: БГТУ, 2014. - 91 с. (URL: https://elib.belstu.by/handle/123456789/23761)

5. Урбанович,П.П. Лабораторный практикум по дисциплинам «Защита информации и надежность информационных систем» и «Криптографические методы защиты информации». Ч.1: Кодирование информации: учебно-метод.пос./П.П. Урбанович, Д.В.Шиман, Н.П. Шутько. – Минск: БГТУ,2019. – 95 с.