Данные и информация

Информация – сведения об окружающем мире, которые повышают уровень осведомленности человека, уменьшают меру неопределенности его знаний.

Данные – это результат наблюдений, зарегистрированные сигналы, которые не используются, а только хранятся. Как только данные начинают использоваться в практических целях, появляется информация.

Р. Хартли первым ввел в теорию передачи информации методологию «измерения количества информации». При этом Хартли считал, что информация, которую он собирался измерять, это «…группа физических символов – слов, точек, тире и т. п., имеющих по общему соглашению известный смысл для корреспондирующих сторон».

Если передаётся последовательность из *n* символов a_1, a_2, a_3,…, a_n, каждый из которых принадлежит алфавиту A_m, состоящему из *m* символов, то число различных вариантов таких последовательностей *K* для *n* = 1 (передаётся один символ) - $K = m$, а для *n* = 2 (передаётся 2 символа), то $K = m^2$? в общем случае для последовательности из *n* символов - $K = m^n$.

Количество информации, содержащееся в такой последовательности, Хартли предложил вычислять как логарифм числа *K* по основанию 2:

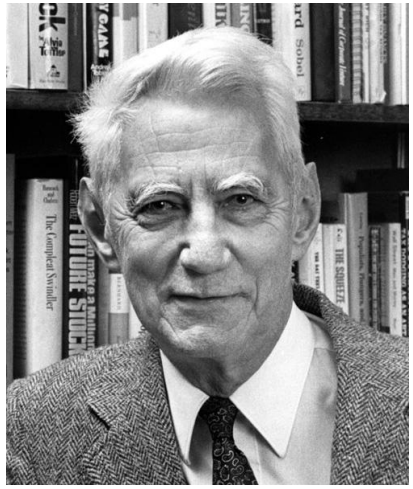
$$I = \log_2 K, \text{ где } K = m^n,$$

а количество информации, содержащееся в последовательности из *n* символов из алфавита A_m, в соответствии с формулой Хартли равно

$$I = \log_2(m^n) = n \log_2 m \ .$$

Замечание 1. Хартли предполагал, что все символы алфавита A_m могут с равной частотой встретиться в любом сообщении.

Замечание 2. Любое сообщение длины *n* в алфавите A_m будет содержать одинаковое количество информации. Это означает, что при вычислении количества информации, содержащегося в сообщении, в расчет не берется его смысловое содержание.



В своих работах К. Шеннон определял количество информации через энтропию.

Им было введено понятие **информационная энтропия** – мера неопределённости состояния некоторой физической системы с конечным числом возможных состояний.

Если *X* – некоторая физическая система, которая может принимать с одинаковой частотой *n* различных состояний x_1, x_2, … x_N, то ее энтропия вычисляется как:

$$H(X) = \log_2 N.$$

Замечание 1. Если система может находиться только в одном состоянии (*N*=1), то её энтропия равна 0, так как её состояние предопределено.

Замечание 2. При оценке энтропии используется логарифм по основанию два. Это означает, что за единицу измерения степени неопределенности принимается неопределенность, содержащаяся в опыте, имеющем два равновероятных исхода, как при подбрасывания монеты.

Такая единица измерения неопределенности принято называть бит.

Шеннон учитывал, что в окружающем мире при наступлении некоторого события, его результаты могут возникать с разной частотой, поэтому использовал понятия теории вероятности: *случайное событие* и *вероятность события*.

Если обозначать события заглавными буквами ***A***, ***B***, ***C*** и т.д, то количественная мера возможности наступления некоторого события ***A*** называется его ***вероятностью***. Вероятность наступления события ***A*** обозначается как $p(A)$ и определяется как отношение количества наступления события ***A*** в опыте к общему числу возможных исходов.

Достоверное событие – событие, которое обязательно наступит, его вероятность равна 1. Достоверное событие информации не несет.

Невозможным называют событие, которое никогда не произойдёт и его вероятность равна 0.

Чем более возможно наступление случайного события, тем больше его вероятность: если ***A*** более возможно чем ***B***, то $p(A) > p(B)$. Для события ***A*** вероятность ее наступления колеблется в диапазоне $0 < p(A) < 1$.

События A_1, A_2, …, A_n образуют ***полную группу***, если в результате опыта обязательно наступит хотя бы одно из них при этом сумма их вероятностей

$$p_1 + p_2 + \ldots + p_n = 1.$$

К. Шеннон, используя подход Р. Хартли, обратил внимание на то, что при передаче словесных сообщений вероятность использования различных букв алфавитов естественных языков не одинакова: некоторые буквы используются часто, другие – редко.

Обозначив через p_i вероятность появления *i*-ого символа в любой позиции передаваемого сообщения, состоящего из *n* символов, то общее количество информации, содержащееся в сообщении из *n* символов:

$$I = -n\sum_{i=1}^{m} p_i \log_2(p_i)$$

Если все символы алфавита A_m появляются с равной вероятностью, то учитывая, что

$$\sum_{i=1}^{m} p_i = 1 \quad \text{è} \quad p = \frac{1}{m}$$

получаем формулу Хартли.

$$I = -n\sum_{i=1}^{m} p_i \log_2(p_i) = -n\sum_{i=1}^{m} \frac{1}{m}\log_2\left(\frac{1}{m}\right) = -n\log_2\left(\frac{1}{m}\right) = n\log_2 m$$

Единицы измерения количества информации

Кроме наименьшей единицы измерения количества информации (***Бит***) используются и более крупные :

1 байт = 8 бит;

1 Кбайт (килобайт) = 1024 байта;

1 Мбайт (мегабайт) = 1024 Кбайта;

1 Гбайт (гигабайт) = 1024 Мбайта.

Представление числовой информации

Система счисления – это способ представления чисел и правила действия над ними. Существуют системы счисления непозиционные и позиционные.

В ***непозиционных*** системах от положения цифры в записи числа не зависит величина, которую она обозначает. Примером может служить римская система. Так CCXXXII складывается из 2-х сотен, 3-х десятков и 2-х единиц и равно 232.

В ***позиционных*** системах величина, обозначаемая цифрой, зависит от ее позиции. Количество используемых цифр называется ***основанием*** системы счисления.

Основание	**Название**	**Алфавит**
n=2	двоичная	01
n=8	восьмеричная	01234567
n=10	десятичная	0123456789
n=16	шестнадцатеричная	0123456789ABCDEF

Развернутой формой записи числа в позиционной системе называется запись в виде:

$$A_q = \pm(a_n q^n + a_{n-1} q^{n-1} + ... + a_1 q^1 + a_0 q^0)$$

Например развернутая форма десятичного числа 589 имеет вид:

$$589_{10} = 5 \cdot 10^2 + 8 \cdot 10^1 + 9 \cdot 10^0 = 5 \cdot 100 + 8 \cdot 10 + 9 \cdot 1$$

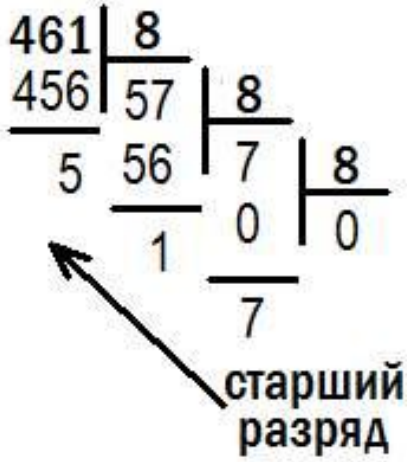
Если все слагаемые в развернутой форме недесятичного числа представить в десятичной системе и вычислить, то получится равное число в десятичной системе. Например:

$$101101_2 = 1 \cdot 2^5 + 0 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 45_{10}$$

Кодирование первых чисел в разных системах счисления

10сс	2сс	8сс	16сс
1	1	1	1
2	10	2	2
3	11	3	3
4	100	4	4
5	101	5	5
6	110	6	6
7	111	7	7
8	1000	10	8
9	1001	11	9
10	1010	12	A
11	1011	13	B
12	1100	14	C
13	1101	15	D
14	1110	16	E
15	1111	17	F

Перевод чисел из десятичной системы счисления в другие системы

В двоичную	В шестнадцатеричную	В восьмеричную
		
Ответ: $12_{10} = 1100_2$	Ответ: $331_{10} = 14B_{16}$	Ответ: $461_{10} = 715_8$

Перевод чисел из двоичной системы счисления в восьмеричную и шестнадцатеричную

$10011010_2 = 10\ \ 011\ \ 010 = 232_8$
$10011010_2 = 1001\ \ 1010 = 9A_{16}$

Кодирование текстовых данных

Для кодирования одного символа используется 8 бит – один байт информации. Такой 8-разрядный код позволяет закодировать 256 различных символов. Институт стандартизации США ввел в действие систему кодирования ***ASCII*** (American Standard Code for Information Interchange – стандартный код информационного обмена США).

В системе ASCII закреплены две таблицы – базовая и расширенная. Первые 128 кодов (с 0 до 127) стандартные и обязательные для всех стран. Вторые – используется для национальных стандартов.

ASCII

	32	5	53	J	74	_	95	t	116	Й	137	Ю	158	│	179	╚	200	▌	221	Є	242
!	33	6	54	K	75	`	96	u	117	К	138	Я	159	┤	180	╔	201	▐	222	є	243
"	34	7	55	L	76	a	97	v	118	Л	139	а	160	╡	181	╩	202	▀	223	Ї	244
#	35	8	56	M	77	b	98	w	119	М	140	б	161	╢	182	╦	203	р	224	ї	245
$	36	9	57	N	78	c	99	x	120	Н	141	в	162	╖	183	╠	204	с	225	Ў	246
%	37	:	58	O	79	d	100	y	121	О	142	г	163	╕	184	═	205	т	226	ў	247
&	38	;	59	P	80	e	101	z	122	П	143	д	164	╣	185	╬	206	у	227	°	248
'	39	<	60	Q	81	f	102	{	123	Р	144	е	165	║	186	╧	207	ф	228	∙	249
(	40	=	61	R	82	g	103	|	124	С	145	ж	166	╗	187	╨	208	х	229	·	250
)	41	>	62	S	83	h	104	}	125	Т	146	з	167	╝	188	╤	209	ц	230	√	251
*	42	?	63	T	84	i	105	~	126	У	147	и	168	╜	189	╥	210	ч	231	№	252
+	43	@	64	U	85	j	106	⌂	127	Ф	148	й	169	╛	190	╙	211	ш	232	¤	253
,	44	A	65	V	86	k	107	А	128	Х	149	к	170	┐	191	╘	212	щ	233	■	254
-	45	B	66	W	87	l	108	Б	129	Ц	150	л	171	└	192	╒	213	ъ	234		255
.	46	C	67	X	88	m	109	В	130	Ч	151	м	172	┴	193	╓	214	ы	235		
/	47	D	68	Y	89	n	110	Г	131	Ш	152	н	173	┬	194	╫	215	ь	236		
0	48	E	69	Z	90	o	111	Д	132	Щ	153	о	174	├	195	╪	216	э	237		
1	49	F	70	[	91	p	112	Е	133	Ъ	154	п	175	─	196	┘	217	ю	238		
2	50	G	71	\	92	q	113	Ж	134	Ы	155	▒	176	┼	197	┌	218	я	239		
3	51	H	72	]	93	r	114	З	135	Ь	156	▓	177	╞	198	▌	219	Ё	240		
4	52	I	73	^	94	s	115	И	136	Э	157	▓	178	╟	199	▄	220	ё	241		

В 1990 г на базе кодировок, использовавшихся в ранних «самопальных» русификаторах Windows, совместно представителями «Параграфа», «Диалога» и российского отделения Microsoft была создана 8-битная кодировка **Windows-1251**, являющаяся стандартной кодировкой для всех русских версий Microsoft Windows вплоть до 10-й версии.

128	Ђ	144	ђ	160		176	°	192	А	208	Р	224	а	240	р
129	Ѓ	145	‘	161	Ў	177	±	193	Б	209	С	225	б	241	с
130	‚	146	’	162	ў	178	І	194	В	210	Т	226	в	242	т
131	ѓ	147	“	163	Ј	179	і	195	Г	211	У	227	г	243	у
132	„	148	”	164	¤	180	г	196	Д	212	Ф	228	д	244	ф
133	…	149	•	165	Ґ	181	µ	197	Е	213	Х	229	е	245	х
134	†	150	–	166	¦	182	¶	198	Ж	214	Ц	230	ж	246	ц
135	‡	151	—	167	§	183	·	199	З	215	Ч	231	з	247	ч
136		152		168	Ё	184	ё	200	И	216	Ш	232	и	248	ш
137	‰	153	™	169	©	185	№	201	Й	217	Щ	233	й	249	щ
138	Љ	154	љ	170	Є	186	Є	202	К	218	Ъ	234	к	250	ъ
139	‹	155	›	171	«	187	»	203	Л	219	Ы	235	л	251	ы
140	Њ	156	њ	172	¬	188	ј	204	М	220	Ь	236	м	252	ь
141	Ќ	157	ќ	173	-	189	S	205	Н	221	Э	237	н	253	э
142	Ћ	158	ћ	174	®	190	s	206	О	222	Ю	238	о	254	ю
143	Џ	159	џ	175	Ї	191	ї	207	П	223	Я	239	п	255	я

В тоже время создается и новый международный стандарт, использующий два байта, получивший название универсальный – ***Unicode***.

Полная спецификация этого стандарта включает в себя существующие, вымершие и искусственные алфавиты, а также математические, музыкальные, химические и прочие символы.

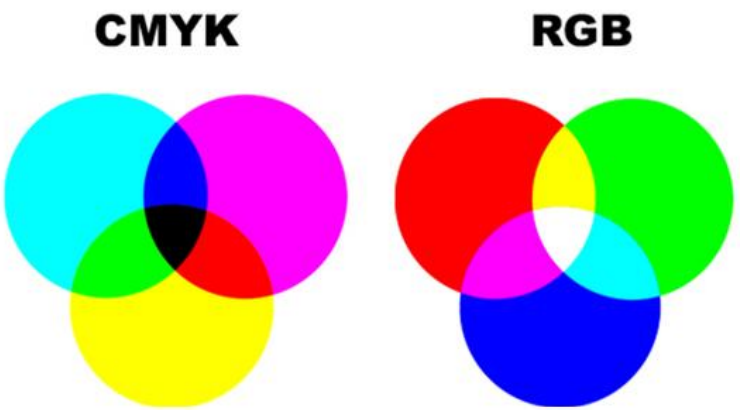
Кодирование графических данных

Графические изображения могут быть представлены в цифровом виде путем их сканирования. Полученный массив прямоугольников называется ***растром***, а сами прямоугольники элементами растра, или ***пикселами*** (picture's element). Качество растрового изображения определяется его разрешением (количеством точек по горизонтали и вертикали) и используемой палитрой цветов.

При кодировании цветных графических изображений один байт может закодировать 256 различных цветов. Если использовать два байта, то 256*256 = 65 536 цветов. При использовании трех байтов можно получить 16,5 миллионов цветов. Этот режим близок к восприятию человеческого глаза красок живой природы.

Физиологические особенности цветового зрения таковы, что глаз воспринимает любой цвет как сумму трех цветов: ***красного***, ***зеленого*** и ***синего***. Система кодирования цвета по трем цветам: красный (Red), зеленый (Green) и синий (Blue) называется системой ***RGB***.

При печати на бумаге действуют другие законы (краски не испускают, а поглощают цвета). Поэтому на печатающих устройствах обычно используется голубой, пурпурный, желтый и черный цвета в качестве основных (такой метод кодировки называется ***CMYK***).



Кодирование звуковой информации

Звук представляет собой аналоговую волну с меняющейся амплитудой и частотой. При преобразование звука в цифровой вид используют два основных метода:

Метод FM (*Frequency Modulation*) предусматривает разложение сложного звукового сигнала на последовательность простейших гармонических сигналов разных частот с последующим квантованием непрерывной волны. Эту работу выполняют специальное устройство – аналого-цифровой преобразователь, расположенный на звуковой плате компьютера. Качество кодирования звука зависит от частоты дискретизации. При таком преобразовании сигналов неизбежны потери информации, поэтому качество звучания имеет оттенок электронной музыки.

Метод таблично-волнового синтеза (*Wave-Table*). При этом методе в памяти компьютера хранятся образы звуков различной природы (***сэмплы)***. Синтез звука основан на последовательном воспроизведении ограниченных по длительности циклических волновых форм, расположенных в памяти в виде матрицы. Сама последовательность вызова той или иной волны, динамическое изменение воспроизводимых волн (синтез и фильтрация) различные способы модуляции и наложение спецэффектов. Все эти изменения могут производиться с помощью математических функций, описывающих степень влияния того или иного параметра на генерируемый сигнал в каждый конкретный момент времени.