Лекция №2

Измерение информации.
Кодирование информации.

28/90 Измерение инф-ии

28/83 Кодирование инф-ии

Тема 1

Измерение информации

Методы измерения информации: алфавитный и вероятностный

Единицей измерения количества информации является бит.

1 бит – это количество информации, необходимое для однозначного определения одного из двух равновероятных событий.

В информатике принято рассматривать последовательности длиной 8 бит. Такая последовательность называется **байтом**.

Производные единицы измерения информации:

1 **байт** = 8 бит = 2^3 бит
1 **килобайт** (**Кб**) = 1024 = 2^{10} байт,
1 **мегабайт** (**Мб**) = 1024 килобайт = 2^{20} байт,
1 **гигабайт** (**Гб**) = 1024 мегабайт = 2^{30} байт,
1 **терабайт** (**Тб**) = 1024 гигабайт = 2^{40} байт

АЛФАВИТНЫЙ ПОДХОД К ИЗМЕРЕНИЮ ИНФОРМАЦИИ

При измерении количества информации в тексте, записанном с помощью N-символьного алфавита, используют следующие формулы:

$$I = i \cdot k,$$
$$i = \log_2 N,$$
$$N = 2^i,$$

где ***I*** – количество информации в тексте,
i – количество информации, которое несет один символ (в битах),
k – количество символов в тексте,
N – мощность алфавита.

АЛФАВИТНЫЙ ПОДХОД К ИЗМЕРЕНИЮ ИНФОРМАЦИИ

Задача. Сообщение, записанное с помощью 64-символьного алфавита, занимает 3 страницы, на каждой странице по 240 символов. Найти количество информации в сообщении (в байтах).

Решение:

$$i = \log_2 N = \log_2 64 = \log_2 2^6 = 6 \cdot \log_2 2 = 6 \cdot 1 = 6 \text{ (бит)}$$

$$k = 3 \cdot 240 = 720 \text{ (символов)}$$

$$I = i \cdot k = 6 \cdot 720 = 4320 \text{ (бит)}$$

$$4320 \text{ бит} = 4320 : 8 \text{ байт} = 540 \text{ байт}$$

Измерение информации

СОДЕРЖАТЕЛЬНЫЙ ПОДХОД

Если после получения какого-то сообщения неопределенность знаний уменьшается в 2 раза, то это сообщение несет в себе 1 бит информации. Т.е., если событие имеет 2 исхода, то при наступлении каждого из них неопределенность знаний уменьшается в 2 раза.

Количество информации, полученное из сообщения о том, что наступило одно из N равновозможных событий, можно вычислить по формуле **Хартли**:

$$x = \log_2 N,$$

где x – количество информации в сообщении (в битах),

N – количество равновозможных (равновероятных) событий, только одно из которых наступило.

СОДЕРЖАТЕЛЬНЫЙ ПОДХОД

Задача 1. Бросают игральный кубик. Найти количество информации в сообщении о том, что выпало число 5.

Решение:

$$N = 6$$
$$x = \log_2 N = \log_2 6 \approx 2{,}58 \text{ бит}$$

Задача 2. В корзине лежат 8 шаров, все разного цвета. Найти количество информации в сообщении о том, что наугад вынули красный шар.

Решение:

$$N = 8$$
$$x = \log_2 N = \log_2 8 = 3 \text{ бита}$$

Измерение информации

ВЕРОЯТНОСТНЫЙ ПОДХОД

Формулу для вычисления количества информации в случае различных вероятностей событий предложил **К. Шеннон**:

$$I = -\sum_{i=1}^{N} p_i \log_2 p_i$$

где ***I*** -количество информации;

N - количество возможных событий;

P_i – вероятность i-го события

ВЕРОЯТНОСТНЫЙ ПОДХОД

Пусть в результате испытания наступило некоторое событие. **Вероятность его наступления** можно вычислить по формуле:

$$P = \frac{K}{N},$$

где ***N*** – количество всех возможных исходов испытания,

K – количество исходов испытания, удовлетворяющих данному событию.

Количество информации в сообщении о том, что наступило одно из возможных событий можно вычислить по формуле:

$$x = \log_2 \frac{1}{P},$$

где ***P*** – вероятность наступления события,

x – количество информации в сообщении о том, что наступило данное событие.

ВЕРОЯТНОСТНЫЙ ПОДХОД

Задача. В непрозрачном мешочке хранятся 10 белых, 20 красных, 30 синих и 40 зеленых шариков. Какое количество информации будет содержать зрительное сообщение о цвете вынутого шарика.

Решение:

10 + 20 + 30 + 40 = 100 - шариков всего

$p_б = 10/100;\ p_к = 20/100;\ p_с = 30/100;\ p_з = 40/100$

$p_б = 0,1;\ p_к = 0,2;\ p_с = 0,3;\ p_з = 0,4$

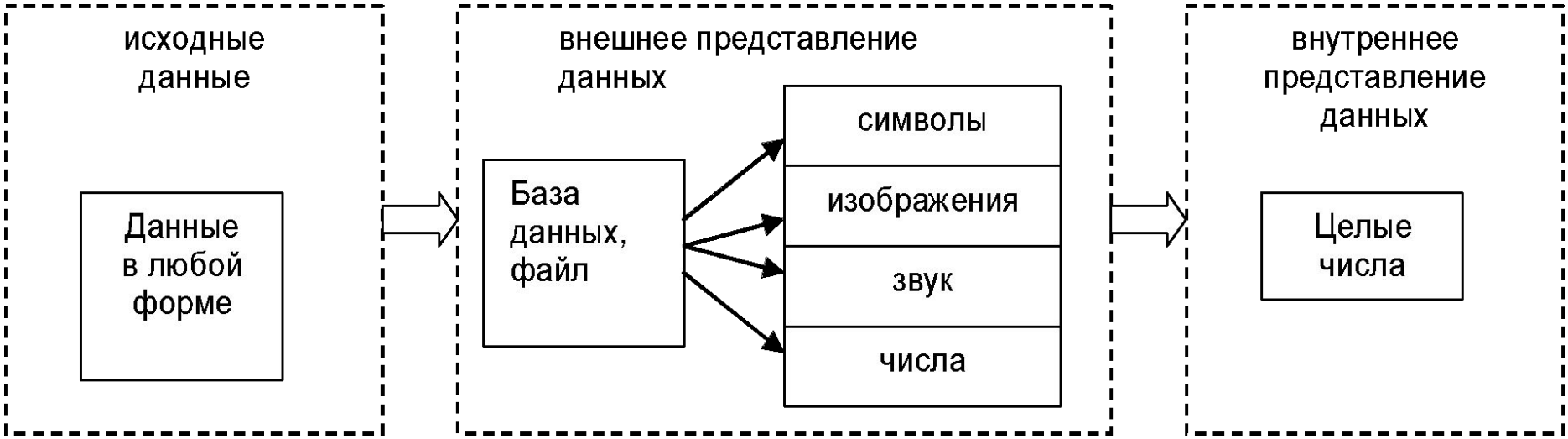
$$I = -(0,1 \cdot \log_2 0,1 + 0,2 \cdot \log_2 0,2 + 0,3 \cdot \log_2 0,3 + 0,4 \cdot \log_2 0,4) \text{ бит}$$

Таким образом, **I ≈ 1,85** бит.

Тема 2

Представление (кодирование) информации в компьютерах

Представление информации в компьютерах



Представление числовой информации в компьютерах

Целые числа без знака - 1 байт

Целые числа со знаком - 2 байта

Представление числовой информации в компьютерах

Вещественные числа - 4 байта

$$R = mP^n$$

m – мантисса числа;
P – основание системы счисления;
n – порядок

$5{,}14 = 0{,}514 \cdot 10^{1} = 51{,}4 \cdot 10^{-1}$

Кодирование текстовой информации

Кодирование символа – это присвоение символу конкретного числового кода.

При вводе в компьютер текстовой информации происходит ее двоичное кодирование.

Код символа хранится в оперативной памяти компьютера. В процессе вывода символа на экран производится обратная операция – **декодирование**, т.е. преобразование кода символа в его изображение.

Как правило, для хранения кода символа используется 1 байт (8 бит), поэтому коды символов могут принимать значение от 0 до 255. Такие кодировки называют **однобайтными**. Они позволяют использовать 256 символов ($N = 2^I = 2^8 = 256$).

Кодирование текстовой информации

Таблица однобайтных кодов символов называется **ASCII** (**American Standard Code for Information Interchange** – Американский стандартный код для обмена информацией).

Первая часть таблицы ASCII-кодов (от 0 до 127) одинакова для всех IBM-PC-совместимых компьютеров и содержит:

- коды управляющих символов;
- коды цифр, арифметических операций, знаков препинания;
- некоторые специальные символы;
- коды больших и маленьких латинских букв.

Вторая часть таблицы ASCII (коды от 128 до 255) бывает различной в разных компьютерах. Она содержит коды букв национального алфавита, коды некоторых математических символов, коды символов псевдографики. Для русских букв в настоящее время имеется пять различных кодовых таблиц: КОИ-8, CP1251, CP866, Mac, ISO.

Кодирование текстовой информации

ASCII (American Standard Code of Information Interchange)

код	символ	код	символ	код	символ	код	символ	код	символ	код	символ
32	Пробел	48	.	64	@	80	P	96	'	112	p
33	!	49	0	65	A	81	Q	97	a	113	q
34	"	50	1	66	B	82	R	98	b	114	r
35	#	51	2	67	C	83	S	99	c	115	s
36	$	52	3	68	D	84	T	100	d	116	t
37	%	53	4	69	E	85	U	101	e	117	u
38	&	54	5	70	F	86	V	102	f	118	v
39	'	55	6	71	G	87	W	103	g	119	w
40	(	56	7	72	H	88	X	104	h	120	x
41	)	57	8	73	I	89	Y	105	i	121	y
42	*	58	9	74	J	90	Z	106	j	122	z
43	+	59	:	75	K	91	[	107	k	123	{
44	,	60	;	76	L	92	\	108	l	124	|
45	-	61	<	77	M	93	]	109	m	125	}
46	.	62	>	78	N	94	^	110	n	126	~
47	/	63	?	79	O	95	_	111	o	127	DEL

Кодирование текстовой информации

Широкое распространение в последнее время получил новый международный стандарт **Unicode**. В нем отводится по два байта (16 бит) для кодирования каждого символа, поэтому с его помощью можно закодировать 65536 различных символов ($N = 2^{16} = 65536$). Коды символов могут принимать значения от 0 до 65536.

Пример. С помощью кодировки Unicode закодирована фраза: *Я хочу поступить в университет*.
Нужно определить информационный объем этой фразы.

Решение. В данной фразе содержится 31 символ (*включая пробелы и знак препинания*). Поскольку в кодировке Unicode каждому символу отводится 2 байта памяти, для всей фразы понадобится $31 \cdot 2 = 62$ байта, или $31 \cdot 2 \cdot 8 = 496$ бит.
Ответ: 62 байта, или 496 бит.

Кодирование текстовой информации

Двоичный код	Десятичный код	КОИ8	CP1251	CP866	Mac	ISO
11000010	194	б	В	-	-	Т

Unicode 16 бит

$N=2^{16}=65\ 536$ символов

Для символов кириллицы в Unicode выделено два диапазона кодов:

Cyrillic (#0400 — #04FF)

Cyrillic Supplement (#0500 — #052F)

Кодирование текстовой информации

Два текста содержат одинаковое количество символов. Первый текст записан на русском языке, а второй на языке племени нагури, алфавит которого состоит из 16 символов.

Чей текст несет большее количество информации?

$$I = K * i$$

Т.к. оба текста имеют одинаковое число символов (*K*), то разница зависит от информативности одного символа алфавита (*i*).

$$2^{i1} = 32, \quad i1 = 5 \text{ бит},$$

$$2^{i2} = 16, \quad i2 = 4 \text{ бит}.$$

$$I1 = K * 5 \text{ бит}, \quad I2 = K * 4 \text{ бит}.$$

Значит, текст, записанный на русском языке в 5/4 раза несет больше информации.

Кодирование текстовой информации

Объем сообщения, содержащего 2048 символов, составил 1/512 часть Мбайта. Определить мощность алфавита

I = 1/512 * 1024 * 1024 * 8 = 16384 бит. - перевели в биты информационный объем сообщения

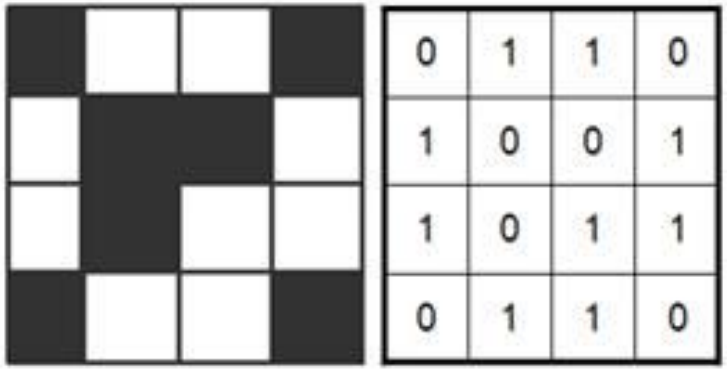
$i = I / K$ = 16384 / 2048 = 8 бит - приходится на один символ алфавита

2^8 = 256 символов - мощность использованного алфавита

Кодирование графической информации

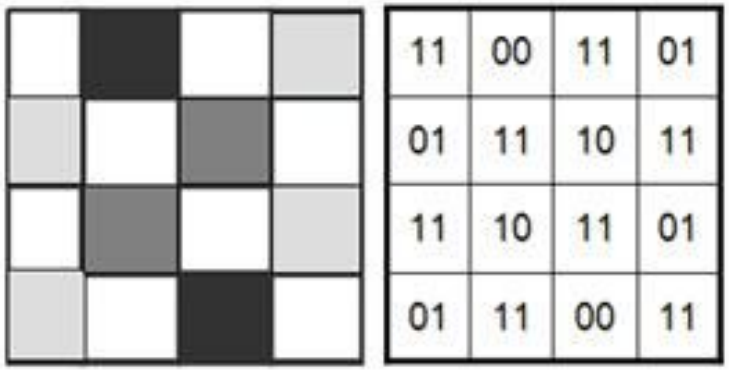
Растровое изображение

Кодирование графической информации



0	1	1	0
1	0	0	1
1	0	1	1
0	1	1	0

1 бит на пиксель – **2 цвета**



11	00	11	01
01	11	10	11
11	10	11	01
01	11	00	11

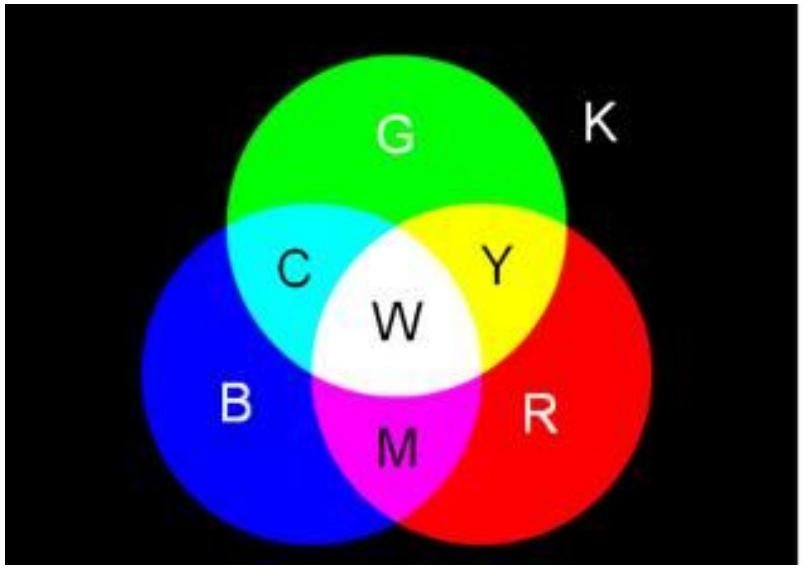
2 бита на пиксель – **4 цвета**

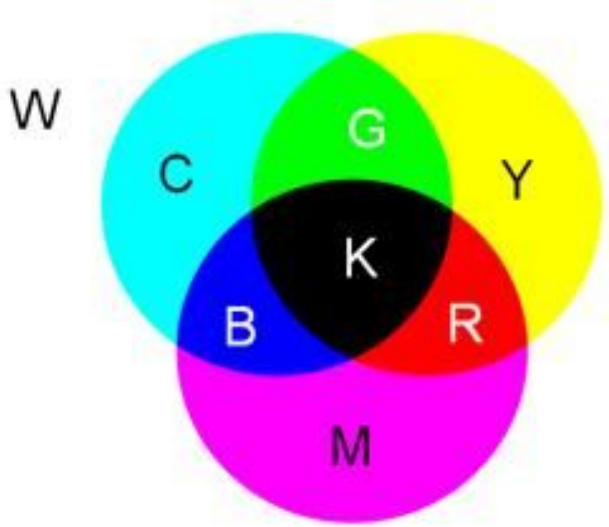
256 градаций серого цвета (от черного (0) до белого (255))

$\log_2(256) = 8$ бит

Цветовые модели

1. **HSB** - оттенок цвета(**H**ue), насыщенность цвета (**S**aturation) и яркость цвета (**B**rightness)
2. **RGB** - красный (Red, **R**), зеленый (Green, **G**), синий (Blue,**B**)
3. **CMYK**





Излучающий объект **RGB** Отражающий объект **CMYK**

Режимы представления цветной графики

1. полноцветный (True Color)
2. High Color
3. индексный

$K = 2^i$

i	*K*	**Достаточно для…**
8	$2^8 = 256$	Рисованных изображений типа мультфильмов, но недостаточно для изображений живой природы
16 (High Color)	$2^{16} = 65536$	Изображений в журналах и на фотографиях
24 (True Color)	$2^{24} = 16\ 777\ 216$	Обработки и передачи изображений, не уступающих по качеству наблюдаемым в живой природе

Кодирование графической информации

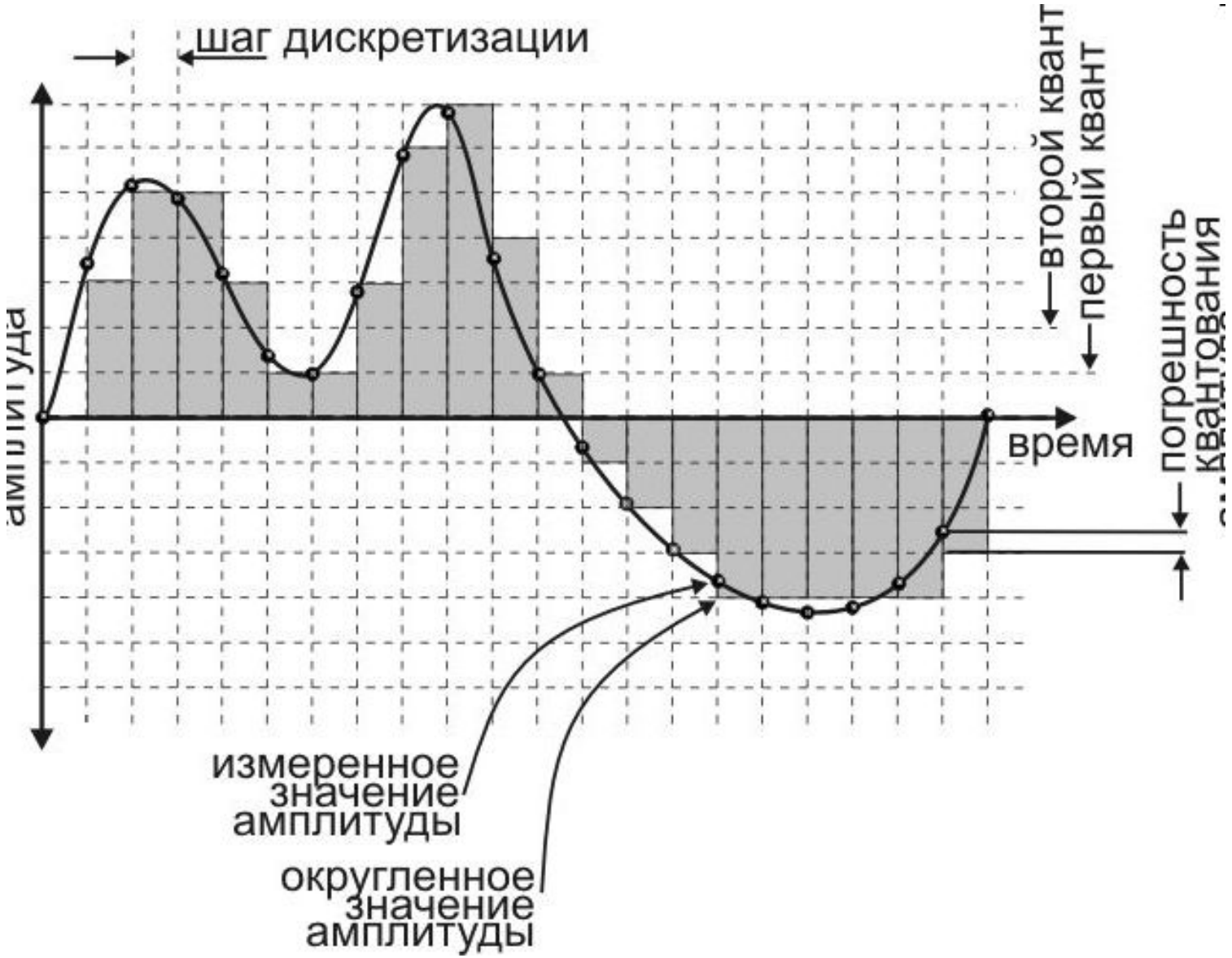
Сколько бит требуется, чтобы закодировать информацию о 130 оттенках?

8 бит (то есть 1 байт), поскольку при помощи 7 бит можно сохранить номер оттенка от 0 до 127, а 8 бит хранят от 0 до 255.

Объем изображения, размером 40x50 пикселей, составляет 2000 байт. Изображение использует:

A - 8 цветов;
B - 256 цветов;
C - 16777216 цветов.

Кодирование звуковой информации



Кодирование звуковой информации

$$K = 2^a$$

a	K	Применение
8	256	Недостаточно для достоверного восстановления исходного сигнала, так как будут большие нелинейные искажения. Применяют в основном в мультимедийных приложениях, где не требуется высокое качество звука
16	65 536	Используется при записи компакт-дисков, так как нелинейные искажения сводятся к минимуму.
20	1 048 576	Где требуется высококачественная оцифровка звука.

Кодирование звуковой информации

Рассчитайте объем стереоаудиофайла длительностью 20 секунд при 20-битном кодировании и частоте дискретизации 44.1 кГц.

Решение:

20 бит * 20 с * 44100 Гц * 2 = 35 280 000 бит =

= 4 410 000 байт = 4,2 Мб