Кодирование информации

Кодирование информации

Сигналы, используемые для передачи информации:

- **световые;**
- **звуковые;**
- **тепловые;**
- **электрические;**
- **в виде жеста;**
- **в виде движения;**
- **в виде слова и т. д.**



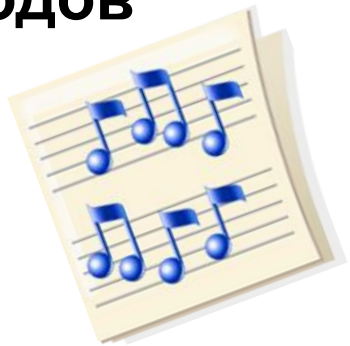


Для того чтобы передача информации была успешной, приёмник должен не только получить сигнал, но и расшифровать его. Необходимо заранее договариваться, как понимать те или иные сигналы, т.е. требуется разработка **кода.**

Кодирование информации

Разнообразие используемых нами кодов

Нотные знаки кодируют музыкальные произведения:



Правила дорожного движения кодируются специальными знаками:

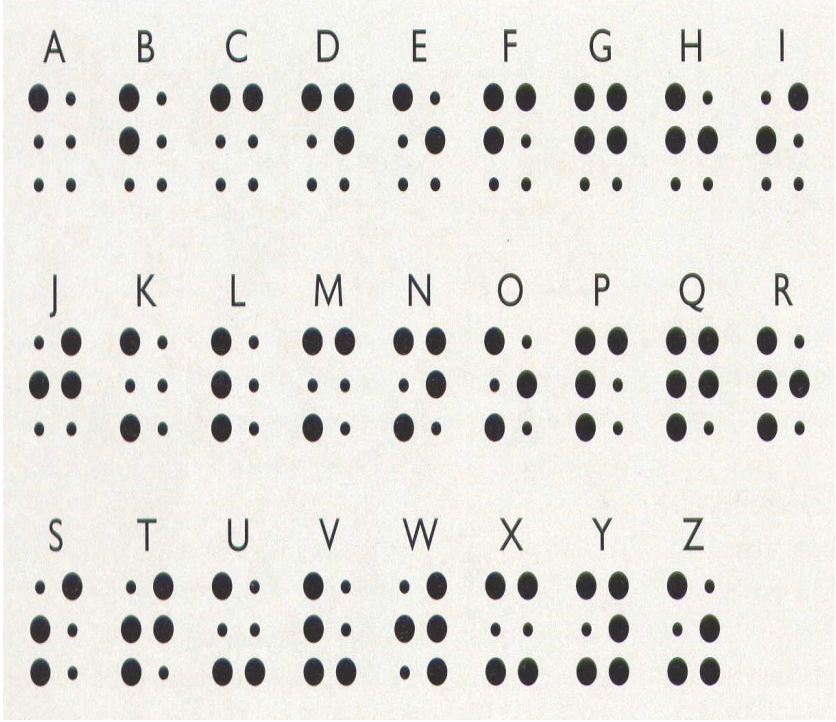
Свой код из шести цифр (индекс) имеет каждый населенный пункт:

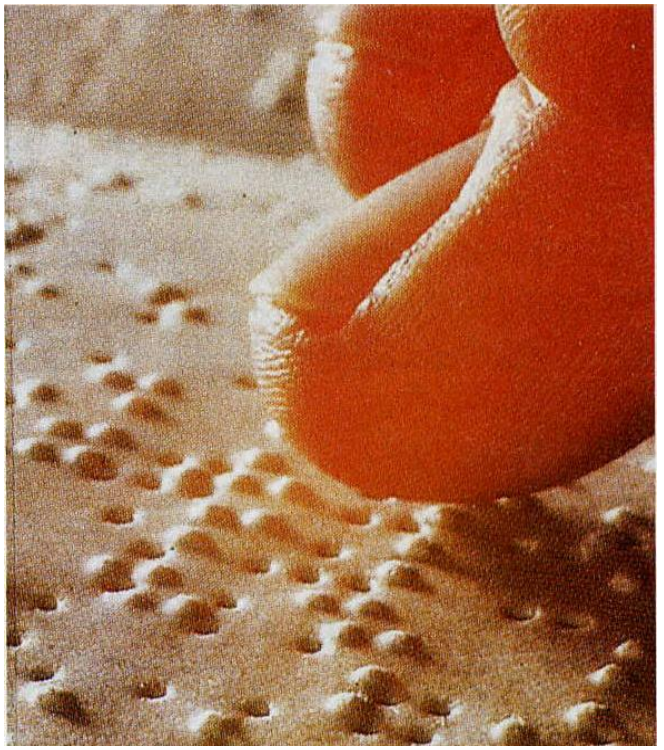
Товары маркируются специальными кодами:

Кодирование информации

Разнообразие используемых нами кодов

В середине XIX века французский педагог Луи Брайль придумал специальный способ представления информации для незрячих людей.



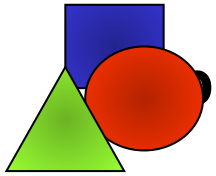


Кодирование информации

Способы кодирования

Одна и та же информация может быть представлена разными кодами.

Существуют **три** основных способа кодирования информации:



Графический – с помощью рисунков и значков;

Числовой – с помощью чисел;

Символьный – с помощью символов того же алфавита, что и исходный текст.

Кодирование информации

Числовое кодирование

В алфавите любого разговорного языка буквы следуют друг за другом в определенном порядке. Это дает возможность присвоить каждой букве алфавита ее порядковый номер.

Например, числовое сообщение

01112001030918

соответствует слову

АЛФАВИТ

Кодирование информации

Символьное кодирование

Смысл этого способа заключается в том, что символы алфавита (буквы) заменяются символами (буквами) того же алфавита по определенному правилу.

Например, а →б, б → в, в → г и т.д. Тогда слово **АЛФАВИТ** будет закодировано последовательностью **БМХБГЙУ**.

Кодирование информации

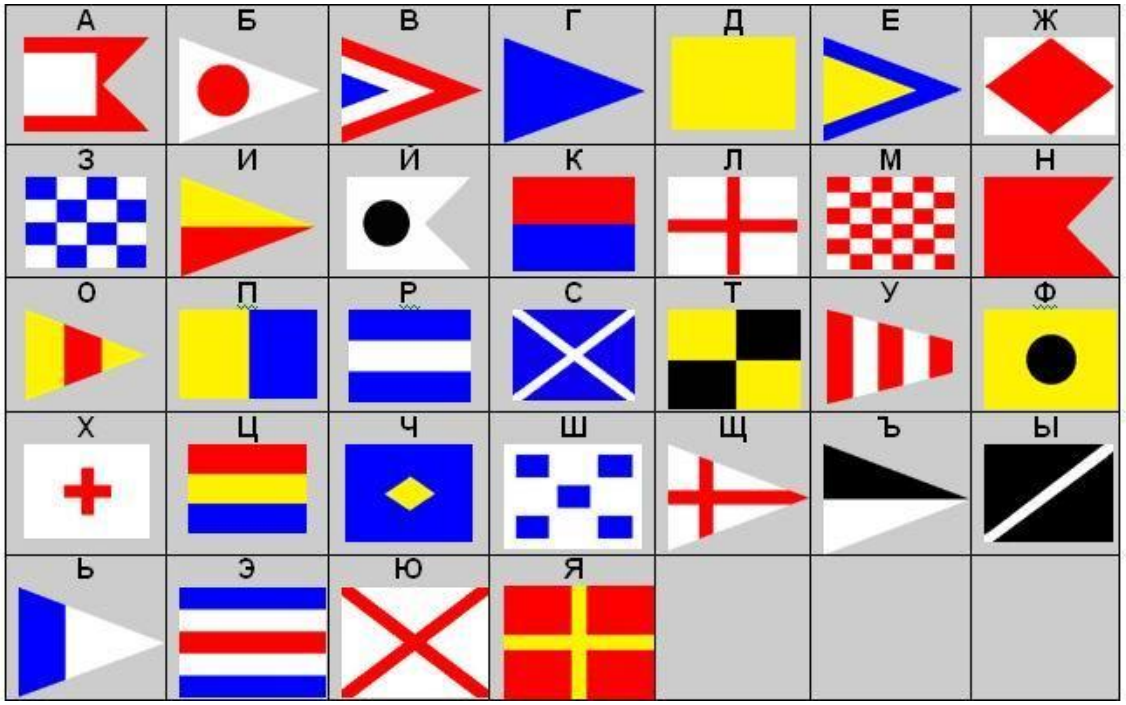
Графическое кодирование

Это кодирование информации при помощи разнообразных рисунков или значков:

Кодирование информации

Графическое кодирование ВМФ России

Специальные сигнальные флаги появились в России ещё в 1696 г. В СССР существовали 32 буквенных, 10 цифровых флагов, 4 дополнительных и 13 специальных флагов. Эта же система с незначительными изменениями используется в ВМФ России.



Общие вопросы кодирования информации

Задача кодирования информации представляется как некоторое преобразование числовых данных в заданной системе счисления.

Двоично-десятичный код				Десятичный код
0	0	0	0	0
0	0	0	1	1
0	0	1	0	2
0	0	1	1	3
0	1	0	0	4
0	1	0	1	5
0	1	1	0	6
0	1	1	1	7
1	0	0	0	8
1	0	0	1	9

Так как любая позиционная система не несет в себе избыточности информации и все кодовые комбинации являются разрешенными, **использовать такие системы для контроля правильности передачи не представляется возможным**.

Общие вопросы кодирования информации

Систематический код - код, содержащий в себе кроме информационных еще и контрольные разряды.

ВАСЯ ⇉ 0101 00 1001 00 1011 10 0001 01

В контрольные разряды записывается некоторая информация об исходном числе. Поэтому можно говорить, что систематический код обладает ***избыточностью***. При этом абсолютная избыточность будет выражаться количеством контрольных разрядов *k*, а относительная избыточность - отношением *k*/*n* , где *n* = *m* + *k* - общее количество разрядов в кодовом слове (*m* - количество информационных разрядов).

Абсолютная избыточность = 2, относительная избыточность 2/6

Общие вопросы кодирования информации

Понятие **корректирующей способности кода** обычно связывают с возможностью обнаружения и исправления ошибки. Количественно корректирующая способность кода определяется вероятностью обнаружения или исправления ошибки. Если имеем ***n***-разрядный код и вероятность искажения одного символа ***p***, то вероятность того, что искажены ***k*** символов, а остальные ***n - k*** символов не искажены, по теореме умножения вероятностей будет

$$w = p^k(1-p)^{n-k}.$$

Число кодовых комбинаций, каждая из которых содержит ***k*** искаженных элементов, равна числу сочетаний из ***n*** по ***k***:

$$C_n^k = \frac{n!}{k!(n-k)!}$$

Тогда полная вероятность искажения информации

$$P_\Sigma = \sum_{i=1}^{k} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}$$

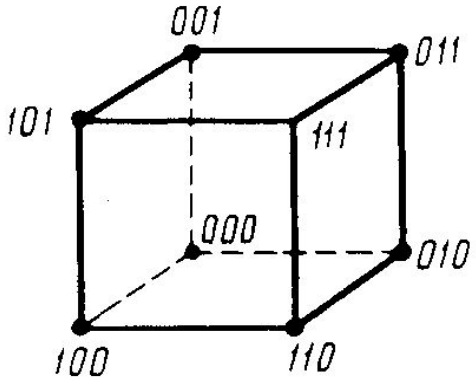
Общие вопросы кодирования

информации

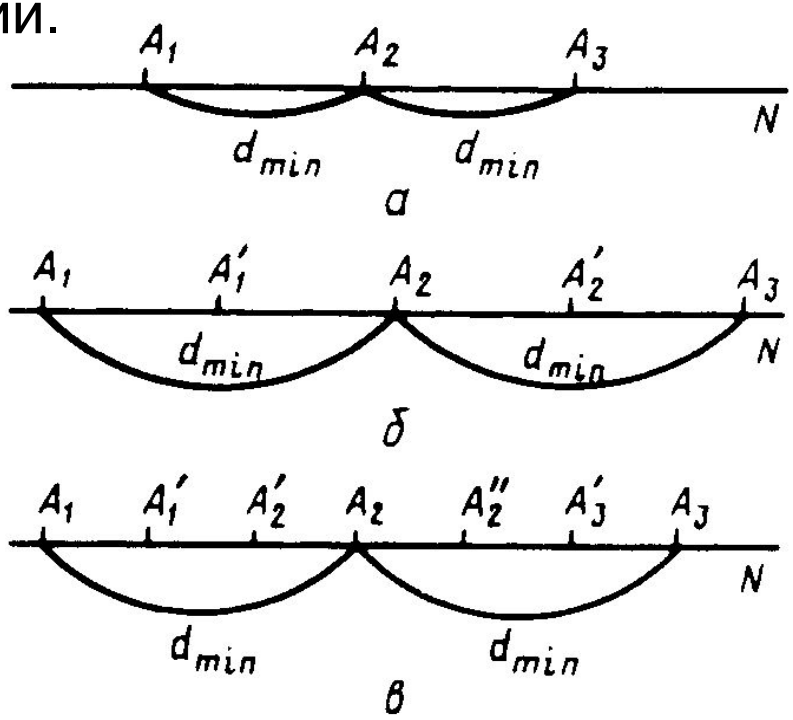
Кодовое расстояние $d(A, B)$ для кодовых комбинаций A и B определяется как вес третьей кодовой комбинации, которая получается поразрядным сложением исходных комбинаций по модулю 2 (операция XOR) или количество различающихся разрядов в двух кодовых комбинациях.

Вес кодовой комбинации $V(A)$ - количество единиц, содержащихся в кодовой комбинации.

Кодовое расстояние = сумма длин ребер между соответствующими вершинами куба



Геометрическое представление кодов



Кодовые расстояния

Общие вопросы кодирования информации

В тех случаях, когда необходимо **не только обнаружить** ошибку, но и **исправить** ее (т. е. указать место ошибки), минимальное кодовое расстояние должно быть

$$d_{min} \geq 2t+1.$$

В теории кодирования показано, что систематический код способен обнаружить ошибки **только когда минимальное кодовое расстояние для него больше или равно 2*t***, т.е.

$$d_{min} \geq 2t,$$

где t - кратность обнаруживаемых ошибок (в случае одиночных ошибок $t = 1$ и т. д.).

Это означает, что между соседними разрешенными кодовыми словами должно существовать по крайней мере одно кодовое слово

Общие вопросы кодирования информации

Допустим, имеется следующий набор кодовых комбинаций:

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

Геометрическая модель этого кода – куб. Для рассматриваемого кода d_{min}= 1. Учитывая, что рассматриваемый код по построению является неизбыточным, можно утверждать, что **любой неизбыточный код имеет dmin= 1 и наоборот, если dmin= 1, код является неизбыточным**.

Общие вопросы кодирования информации

Для построения простейшего избыточного кода, который может обнаруживать одну ошибку, нужно отобрать рабочие комбинации на расстоянии $d(A, B) \geq 2$.

В рассматриваемом коде можно выбрать следующие комбинации:

$$\left.\begin{array}{l} 0\,1\,0 \\ 1\,0\,0 \\ 1\,1\,1 \\ 0\,0\,1 \end{array}\right\} M_p = 4$$

где M_p – число разрешенных (или рабочих) комбинаций.

Избыточность полученного кода

$$R=k/n=(n-m)/n=1/3\approx 33\%.$$

Общие вопросы кодирования информации

Если требуется обнаруживать две ошибки, то рабочих комбинаций будет только две, например

$$\left.\begin{matrix}0\,0\,0\\1\,1\,1\end{matrix}\right\} M_p = 2$$

$$\left.\begin{matrix}0\,1\,0\\1\,0\,1\end{matrix}\right\} M_p = 2$$

минимальное кодовое расстояние в этом случае $d_{min} = 3$, избыточность

$$R=k/n=(n-m)/n=2/3\approx 67\%.$$

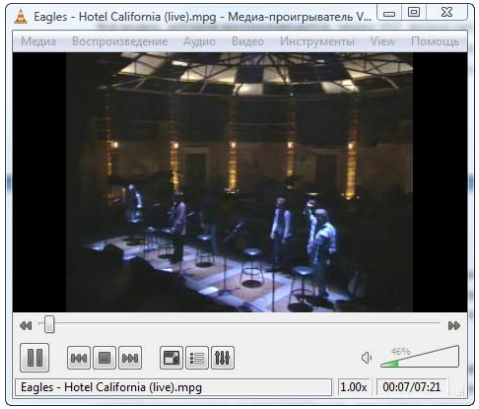
Если требуется обнаруживать три ошибки, $d_{min} \geq 4$, что невозможно обеспечить в рассматриваемом коде, так как кодовое расстояние $d(A, B) \leq 3$.

Общие вопросы кодирования информации

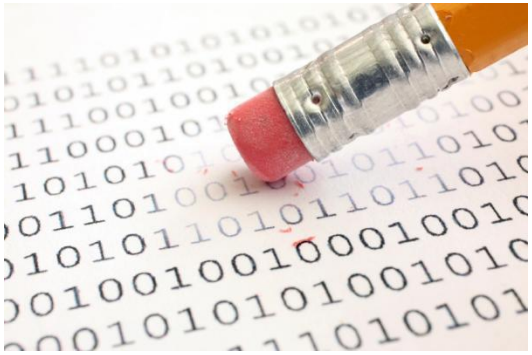
Возможны несколько стратегий борьбы с ошибками:

обнаружение ошибок в блоках данных и автоматический запрос повторной передачи повреждённых блоков

обнаружение ошибок в блоках данных и отбрасывание повреждённых блоков (потоковые мультимедиа-системы)

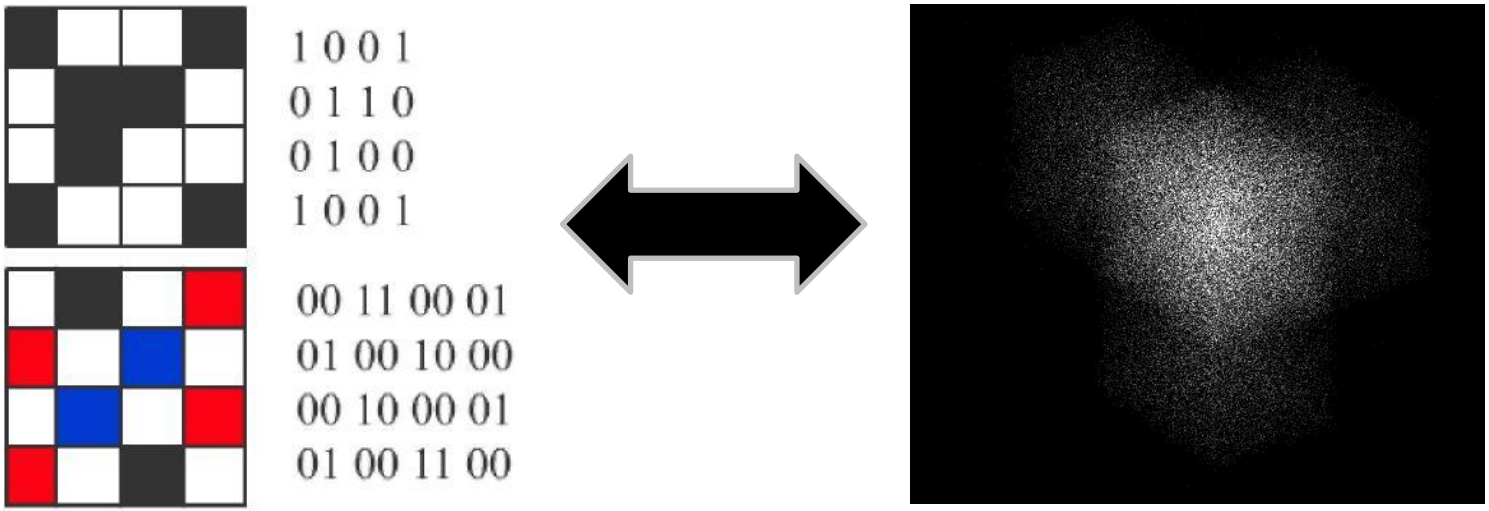


исправление ошибок



Общие вопросы кодирования информации

Эффективное кодирование базируется на основной теореме Шеннона для каналов без шума, в которой доказано, что сообщения, составленные из букв некоторого алфавита, можно закодировать так, что среднее число двоичных символов на букву будет сколь угодно близко к энтропии источника этих сообщений, но не меньше этой величины.



Эффективное кодирование информации

Методика кодирования Шеннона-Фано

Код строится следующим образом: буквы алфавита сообщений выписываются в таблицу в порядке убывания вероятностей. Затем они разделяются на две группы так, чтобы суммы вероятностей в каждой из групп были по возможности одинаковы. Всем буквам верхней половины в качестве первого символа приписывается 1, а всем нижним - 0. Каждая из полученных групп, в свою очередь, разбивается на две подгруппы с одинаковыми суммарными вероятностями и т. д. Процесс повторяется до тех пор, пока в каждой подгруппе останется по одной букве.

Буквы	Вероятности	Кодовые комбинации
z_1	0,22	11
z_2	0,20	101
z_3	0,16	100
z_4	0,16	01
z_5	0,10	001
z_6	0,10	0001
z_7	0,04	00001
z_8	0,02	00000

$$(z) = -\sum_{i=1}^{8} p(z_i)\log p(z_i) \approx 2{,}76$$

Среднее число символов на букву

$$l_{\text{ср}} = -\sum_{i=1}^{8} p(z_i)\, n(z_i) \approx 2{,}84,$$

Эффективное кодирование информации

Методика кодирования Шеннона-Фано

Множество вероятностей в предыдущей таблице можно было разбить иным образом:

Буквы	Вероятности	Кодовые комбинации
z_1	0,22	11
z_2	0,20	10
z_3	0,16	011
z_4	0,16	010
z_5	0,10	001
z_6	0,10	0001
z_7	0,04	00001
z_8	0,02	00000

Буквы	Вероятности	Кодовые комбинации
z_1	0,22	11
z_2	0,20	101
z_3	0,16	100
z_4	0,16	01
z_5	0,10	001
z_6	0,10	0001
z_7	0,04	00001
z_8	0,02	00000

Предыдущий вариант

$$(z) = -\sum_{i=1}^{8} p(z_i)\log p(z_i) \approx 2{,}76$$

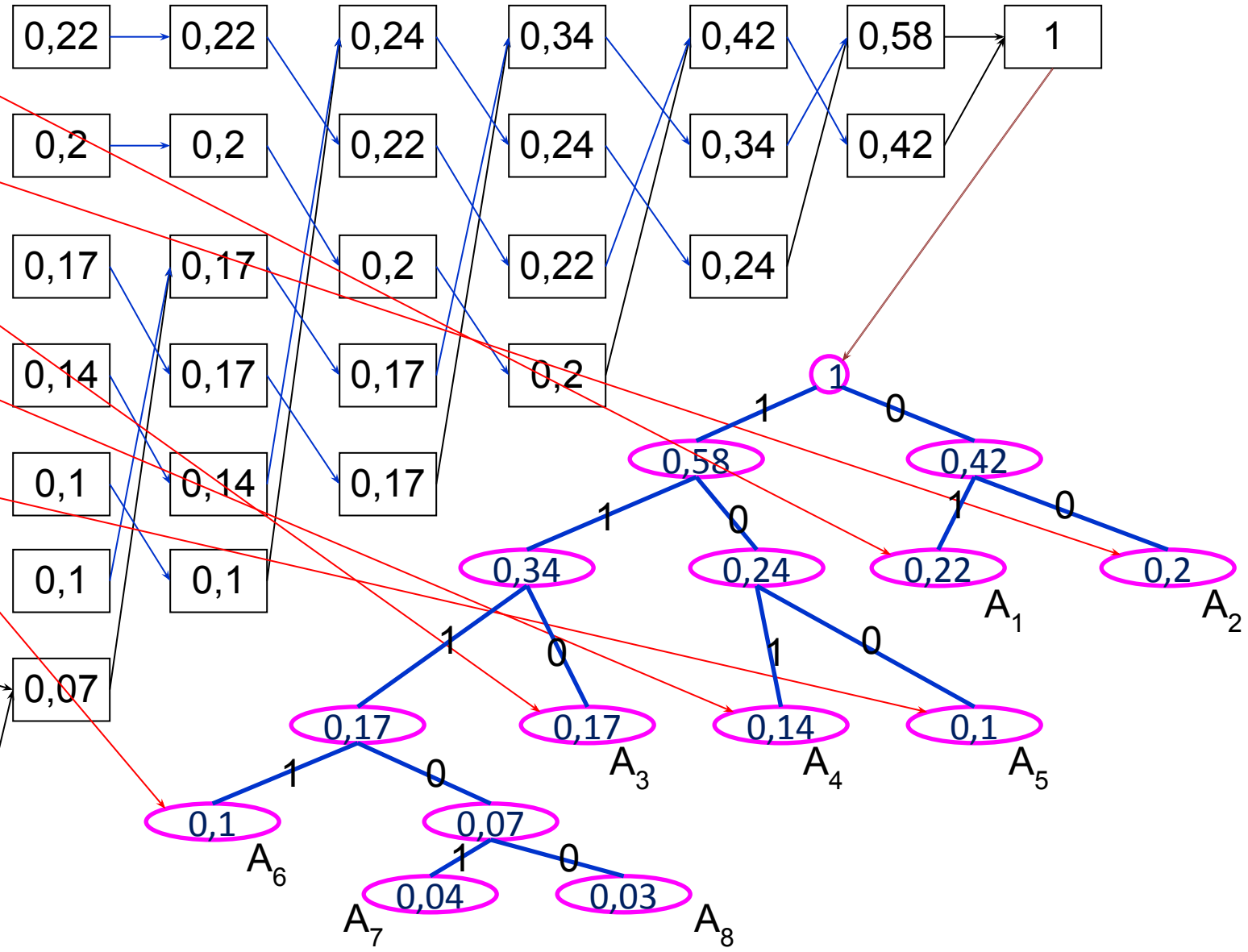
Среднее число символов на букву

$$l_{\text{ср}} = -\sum_{i=1}^{8} p(z_i)\, n(z_i) \approx 2{,}8$$

Эффективное кодирование информации

Методика кодирования Хаффмена (Huffman)

A	P
A_1	0,22 **01**
A_2	0,2 **00**
A_3	0,17 **110**
A_4	0,14 **101**
A_5	0,1 **100**
A_6	0,1 **1111**
A_7	0,04 **11101**
A_8	0,03 **11100**



Эффективное кодирование информации

Методика кодирования по четности-нечетности

Если в математическом коде выделен один контрольный разряд (k = 1), то к каждому двоичному числу добавляется один избыточный разряд и в него записывается 1 или 0 с таким условием, чтобы сумма цифр в каждом числе была по модулю 2 равна 0 для случая четности или 1 для случая нечетности. Появление ошибки в кодировании обнаружится по нарушению четности (нечетности). При таком кодировании допускается, что может возникнуть только одна ошибка.

Число	Контрольный разряд	Проверка
10101011	1	0
11001010	0	0
10010001	1	0
11001011	0	1-нарушение

Такое кодирование имеет минимальное кодовое расстояние, равное 2.

Эффективное кодирование информации

Методика кодирования по четности-нечетности

Можно представить и несколько видоизмененный способ контроля по методу четности - нечетности. Длинное число разбивается на группы, каждая из которых содержит *l* разрядов. Контрольные разряды выделяются всем группам по строкам и по столбцам согласно следующей схеме:

a_1	a_2	a_3	a_4	a_5	k_1
a_6	a_7	a_8	a_9	a_{10}	k_2
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	k_3
a_{16}	a_{17}	a_{18}	a_{19}	a_{20}	k_4
a_{21}	a_{22}	a_{23}	a_{244}	a_{25}	k_5
k_6	k_7	k_8	k_9	k_{10}	

Проверка

$\Sigma_i(a_i+k_i)$

Эффективное кодирование информации

Методика кодирования Хэмминга

Предположим, что имеется код, содержащий m информационных разрядов и k контрольных разрядов. Запись на k позиций определяется при проверке на четность каждой из проверяемых k групп информационных символов. Пусть было проведено k проверок. Если результат проверки свидетельствует об отсутствии ошибки, то запишем 0, если есть ошибка, то запишем 1 . Запись полученной последовательности символов образует двоичное, контрольное число, указывающее номер позиции, где произошла ошибка. При отсутствии ошибки в данной позиции последовательность будет содержать только нули. Полученное таким образом число описывает $n = (m + k + 1)$ событий. Следовательно, справедливо неравенство

$$2^k \geq (m + k + 1).$$

Пример кода: 1001011001 0100 $2^k \geq (m + k + 1) \Rightarrow 16 \geq 15$

m k

Эффективное кодирование информации

Методика кодирования Хэмминга

Полученное таким образом число описывает $n = (m + k + 1)$ событий. Следовательно, справедливо неравенство

$$2^k \geq (m + k + 1).$$

Определить максимальное значение m для данного k можно из следующего:

n… 1 2 3 4… 8…15 16…31 32…63 64
m…0 0 1 1… 4…11 11…26 26…57 57
k… 1 2 2 3… 4…4 5…5 6…6 7

Пример кода: 1001011001 0100 $2^k \geq (m + k + 1) \Rightarrow 16 \geq 15$
m k

Эффективное кодирование информации

Методика кодирования Хэмминга

Определим теперь позиции, которые надлежит проверить в каждой из *k* проверок. Если в кодовой комбинации ошибок нет, то контрольное число содержит только нули. Если в первом разряде контрольного числа стоит 1, то, значит, в результате первой проверки обнаружена ошибка. Имея таблицу двоичных эквивалентов для десятичных чисел, можно сказать, что, например, первая проверка охватывает позиции 1, 3, 5, 7, 9 и т. д., вторая проверка — позиции 2, 3, 6, 7, 10.

Проверка Проверяемые разряды

1... **1**,3,5,7,9,11,13,15...
2... **2**,3,6,7, 10, 11, 14, 15, 18, 19,22,23...
3... **4**, 5, 6, 7, 12, 13, 14, 15, 20, 21, 22, 23...
4... **8**,9,10,11,12,13,14,15,24...

Эффективное кодирование информации

Методика кодирования Хэмминга

Кодирование информации по методу Хэмминга для 7-миразрядного кода

n=7, *m*=4, *k*=3 и контрольными будут разряды 1, 2, 4

Разряды двоичного кода							Кодируемая десятичная информация
1 k_1	2 k_2	3 m_1	4 k_3	5 m_2	6 m_3	7 m_4	
0	0	0	0	0	0	0	0
1	1	0	1	0	0	1	1
0	1	0	1	0	1	0	2
1	0	0	0	0	1	1	3
1	0	0	1	1	0	0	4
0	1	0	0	1	0	1	5
1	1	0	0	1	1	0	6
0	0	0	1	1	1	1	7
1	1	1	0	0	0	0	8
0	0	1	1	1	0	1	9
1	0	1	1	0	1	0	10
0	1	1	0	0	1	1	11
0	1	1	1	1	0	0	12
1	0	1	0	1	0	1	13
0	0	1	0	1	1	0	14
1	1	1	1	1	1	1	15

Проверка	Проверяемые разряды
1…	1,3,5,7,9…
2…	2,3,6,7, 10…
3…	4, 5, 6, 7, 12…
4…	8,9,10,11…

000

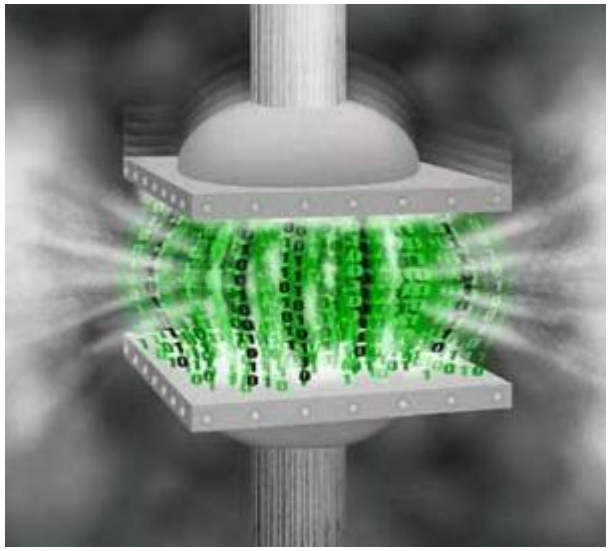
$101_2=5_{10}$

Эффективное кодирование информации

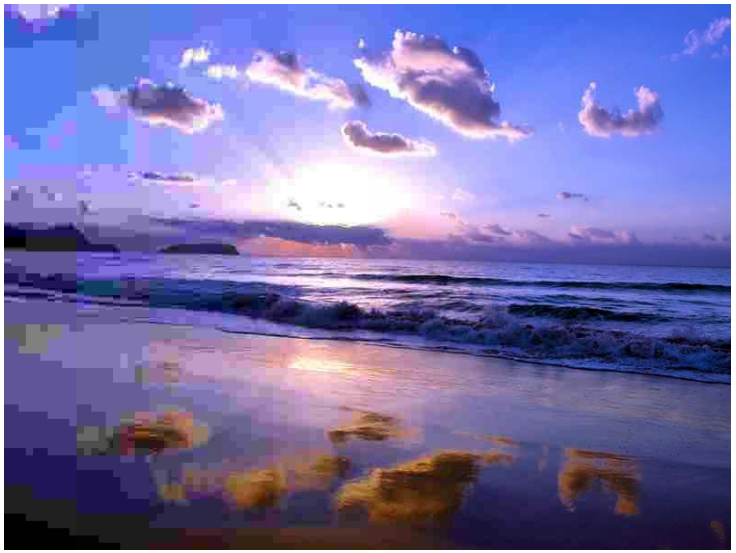
Алгоритмы сжатия информации

Обычно выделяют два класса алгоритмов сжатия

Сжатие без



Сжатие с



Эффективное кодирование информации

Алгоритмы сжатия информации

Классический алгоритм Лемпела-Зива – LZ77, названный так по году своего опубликования, предельно прост. Он формулируется следующим образом : **"если в прошедшем ранее выходном потоке уже встречалась подобная последовательность байт, причем запись о ее длине и смещении от текущей позиции короче чем сама эта последовательность, то в выходной файл записывается ссылка (смещение, длина), а не сама последовательность"**.

"КОЛОКОЛ_ОКОЛО_КОЛОКОЛЬНИ"

"КОЛО(-4,3)_(-5,4)О_(-14,7)ЬНИ"

Алгоритм RLE (англ. Run Length Encoding)

"AAAAAAA"

"(A,7)"