Эффективное кодирование

Эффективное кодирование

- Эффективное кодирование решает задачу более компактной записи сообщений, вырабатываемых источником за счет их перекодировки.
- Применяется практически во всех архиваторах типа Rar, Zip и др.
- Позволяют сжать информацию в в 2-3, max в 4 раза, но при этом происходит полное восстановление сжатой информации «бит в бит»

Сжатие в большее число раз

- Применяется, если же не требуется восстановление информации «бит в бит»

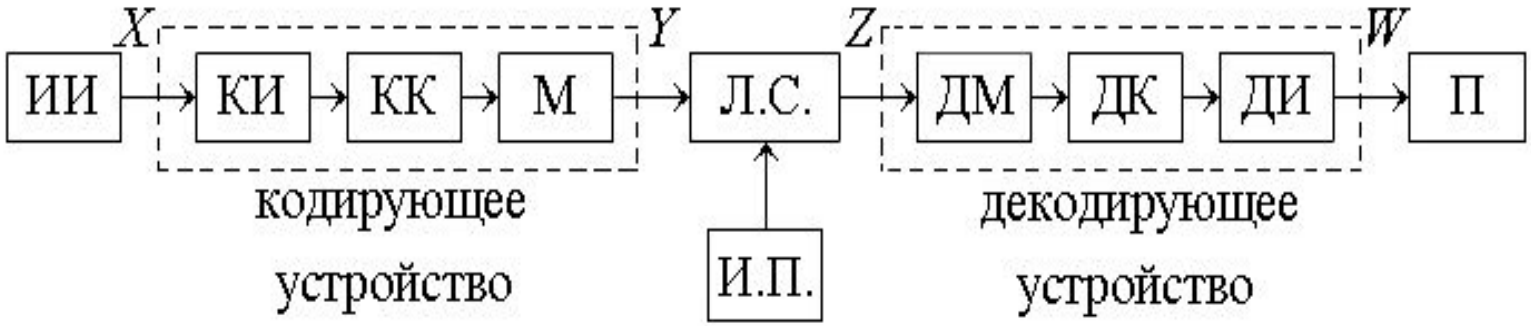
- Например, при передаче речи можно допускать искажения, которые получатель голосового сообщения не заметит из-за нечувствительности слухового аппарата человека к этим изменениям.

Общее определение кодирования и кода

- Кодирование – в широком смысле слова – это представление сообщений в форме, удобной для передачи по данному каналу.

- Операция, обратная кодированию, называется декодирование.

Схема системы передачи информации



- ИИ – источник информации
- КИ – кодер источника
- КК –кодер канала
- М – модулятор
- ЛС –линия связи
- ДМ - демодулятор
- ДК- декодер канала
- ДИ – декодер источника
- П - приемник

- Сообщению *X* на выходе источника информации (ИИ) необходимо поставить в соответствие определенный сигнал.

- Дискретные сообщения складываются из букв,
а непрерывные сообщения можно представить последовательностью цифр в каждый момент отсчета,
то *можно использовать конечное число образцовых сигналов, соответствующих отдельным буквам алфавита источника*.

- При большом объеме алфавита прибегают к представлению букв в другом алфавите с меньшим числом букв, которые будем называть символами. Т. е. выполняется КОДИРОВАНИЕ

- Поскольку алфавит символов меньше алфавита букв, то каждой букве соответствует некоторая последовательность символов, называемая кодовой комбинацией.

- Число символов в кодовой комбинации называется ее значностью.

В процессе преобразования букв в символы нужно:

1. Преобразовать информацию в такую систему символов (код), чтобы он обеспечивал.:
   - простоту аппаратуры различения отдельных символов;
   - минимальное время передачи;
   - минимальный объем запоминающего устройства при хранении;
   - простоту выполнения в принятой системе арифметических и логических действий.

Статистические свойства источника сообщений и помех в канале связи при этом не принимаются во внимание.

- Техническая реализация процесса кодирования в таком простейшем виде при непрерывном входном сигнале осуществляется аналого-кодовыми (цифровыми) преобразователями.

В процессе преобразования букв в символы нужно:

2. Второй целью кодирования является на основании теорем Шеннона – согласование свойств источника сообщений со свойствами канала связи.

- При отсутствии помех это непосредственно дает выигрыш во времени передачи или в объеме запоминающего устройства.
- Такое кодирование получило название эффективное кодирование.

При наличии помех

- в канале эффективное кодирование позволяет преобразовать входную информацию в последовательность символов, наилучшим образом подготовленную для дальнейшего преобразования (максимально сжатую)

КК - кодер канала

имеет целью обеспечить заданную достоверность при передаче или хранении информации путем дополнительного внесения избыточности, но уже по простым алгоритмам и с учетом статистических закономерностей помехи в канале связи.

Такое кодирование получило название помехоустойчивого.

Выбор кодирующих и декодирующих устройств зависит от **статистических свойств источника**, а так же **уровня и характера помех** в канале связи

- Если избыточность источника сообщений (ИС) и помехи в канале связи практически отсутствуют, то введение как КИ, так и КК нецелесообразно.

- Если избыточность ИС высока, а помехи малы, целесообразно введение только КИ.
- Когда избыточность источника мала, а помехи велики, целесообразно введение КК.

- При большой избыточности и высоком уровне помех целесообразно введение обоих дополнительных кодирующих и декодирующих устройств.

После кодера канала КК

- кодированный сигнал поступает в устройство кодирования символов сигналами – модулятор М.
- Получаемый на выходе модулятора сигнал *Y* подготовлен к передаче по конкретной линии связи ЛС.
- В устройство декодирования сигналов в символы (демодулятор ДМ) из линии связи приходит сигнал, искаженный шумом, который обозначен на схеме – *Z*.

Устройства декодирования

- Устройство декодирования помехоустойчивого кода декодер канала ДК и
- устройство декодирования сообщений (декодер источника ДИ)
- выдают декодированное сообщение *И* получателю П (человеку или машине).

Итак

- При кодировании каждая буква исходного алфавита представляется различными последовательностями, состоящими из кодовых букв (цифр).

- Если исходный алфавит содержит m букв, то для построения равномерного кода с
- использованием k кодовых букв необходимо удовлетворить соотношение

$m \le k_q$,

где q - количество элементов в кодовой последовательности.

$$q \ge \frac{\log m}{\log k} = \log_k m$$

Для построения равномерного кода

достаточно пронумеровать буквы исходного алфавита и записать их коды как q - разрядные числа в k-ичной системе счисления

- Например,

при двоичном кодировании 32 букв русского алфавита используется

$q = \log_2 32 = 5$ разрядов,

на чем и основывается телетайпный код.

- Кроме двоичных кодов, наибольшее распространение получили восьмеричные коды.

- Например, необходимо закодировать алфавит, состоящий из 64 букв.

Для этого потребуется $q = \log_2 64 = 6$ двоичных разрядов или $q = \log_8 64 = 2$ восьмеричных

- разрядов.
- При этом буква с номером 13 при двоичном кодировании получает код 001101, а при
- восьмеричном кодировании 15.

Основная теорема Шеннона о эффективном кодировании

- В любом реальном сигнале всегда присутствуют помехи.
- Однако, если их уровень настолько мал, что вероятность искажения практически равна нулю, можно условно считать, что все сигналы передаются неискаженными.

теорема Шеннона

- В этом случае среднее количество информации, переносимое одним символом, можно считать:
- $J(Z;\ Y) = H_{апр}(Z) - H_{апост}(Z) = H_{апр}(Y)$,

- так как $H(Y) = H(Z)$ и $H(Y/Z) = 0$, а $\max\{J(Z;\ Y)\} = H_{max}(Y)$ – max энтропия источника сигнала, получающаяся при равномерном распределении вероятностей символов алфавита Y.
- $p(\ y_1) = p(\ y_2) = \ldots = p(\ y_m) = 1/M_y$,
- т.е. $H_{max}(Y) = \log_a M_y$

Теорема Шеннона

- следовательно, пропускная способность дискретного канала без помех в единицах информации за единицу времени равна:

- $C_y = V_y \cdot \max\{J(Z; Y)\} = V_y \cdot H_{\max}(Y) = V_y \cdot \log_a M_y$
- или $C_k = V_k \cdot \log_a M_y$,
- V_k – скорость передачи определяется частотными переключательными способностями канала:

$$V_k = \frac{1}{\Delta t_k} = \frac{1}{1/2 \cdot f_{\max}} = 2 \cdot f_{\max}.$$

Теорема Шеннона

- Если источник информации создает поток информации
- ,

$$H'(x) = \overline{H}(x) = V_x \cdot H(x) = \frac{H(x)}{T_x}$$

- Такой что, производительность источника информации равна пропускной способности канала , т.е. равна энтропии источника, приходящаяся на единицу времени(бит,сек=бод)

- а канал связи обладает пропускной способностью *C* ед. информации в единицу времени, то при $H(x) \le C$:

1. Сообщения, вырабатываемые источником, всегда можно закодировать так, чтобы скорость их передачи была сколь угодно близкой к

$$V_{x\max} = \frac{C_k}{H(x)}$$

2. Не существует способа кодирования, позволяющего сделать эту скорость большей, чем $V_{x\max}$.

Теорема Шеннона

- Согласно сформулированной теореме существует метод кодирования, позволяющий при:
- $H(x) \leq C$ – передавать всю информацию, вырабатываемую источником при ограниченном объеме буфера;
- $H(x) > C$ – такого метода кодирования не существует, так как требуется буфер, объем которого определяется превышением производительности источника над пропускной способностью канала, умноженной на время передачи.

Следствия из теоремы Шеннона

- если источник информации имеет энтропию $H(X)$,
- то сообщения можно закодировать так, чтобы средняя длина кода $l_{ср}$ (количество символов сигнала на одну букву сообщения) была сколь угодно близкой к величине:

$$l_{ср} \rightarrow \frac{H(x)}{\log_a M_y}$$

Следствия из теоремы Шеннона

$$l_{ср} \rightarrow \frac{H(x)}{\log_a M_y}$$

- то есть при $a = 2$ (бит) и $M_y = 2$ {0; 1} имеем:

$$l_{ср} = \sum_{i=1}^{M} p_i \cdot l_i = H(x) + \varepsilon$$

- где p_i – вероятность встречи данного элемента алфавита;
- l_i – количество символов в i-ой кодовой комбинации;
- ε – бесконечно малая величина ≥ 0, т.е. $\lim l_{ср} = H(x)$.

- Это следует из равенства:
- .
- Таким образом, $l_{ср}$ выступает критерием эффективности кодирования. Чем ближе $l_{ср}$ к $H(x)$, тем лучше мы закодировали. В инженерной практике это различие можно считать допустимым 3÷5% (до 10%).

Следствия из теоремы Шеннона

- Это следует из равенства:

$$\overline{H}(x) = V_{x\max} \cdot H(x) = C_k = V_{y\max} = V_{x\max} \cdot l_{ср}$$

(Производительность ист.=попускн.сп .к=макс.ск. Канала без помех=произ макс. скорости передачи сообщения на ср длину)

- Таким образом, $l_{ср}$ выступает критерием эффективности кодирования. Чем ближе $l_{ср}$ к $H(x)$, тем лучше мы закодировали.

- В инженерной практике это различие можно считать допустимым 3÷5% (до 10%).

Следствия из теоремы Шеннона

Из этого же критерия следует, что если буквы имеют равномерное распределение вероятностей их употребления, то

- $H(x) = \log_2 M$,
- а $l_{ср} = \log_2 M$.

Пределы эффективного кодирования:

- $H(x) \le l_{ср} \le \log_2 M$.

Методики эффективного кодирования

- В большинстве случаев буквы сообщений преобразуются в последовательности двоичных символов.
- Учитывая статистические свойства источника сообщений, можно минимизировать среднее число двоичных символов, требующихся для выражения одной буквы сообщения, что при отсутствии шума позволяет уменьшить время передачи или объем запоминающего устройства.

Методики эффективного кодирования

Шеннон доказал, 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

Из этого следует, что при выборе каждого символа кодовой комбинации необходимо стараться, чтобы он нес максимальную информацию.

Каждый символ должен принимать значения 0 и 1 по возможности с равными вероятностями и каждый выбор должен быть независим от значений предыдущих символов.

Методика Шеннона-Фэно

Алгоритм использует коды переменной длины: часто встречающийся символ кодируется кодом меньшей длины, редко встречающийся — кодом большей длины.

Коды Шеннона — Фано префиксные, то есть никакое кодовое слово не является префиксом любого другого.

Это свойство позволяет однозначно декодировать любую последовательность кодовых слов.

Методика Шеннона-Фэно

1. Символы первичного алфавита m_1 выписывают по убыванию вероятностей.
2. Символы полученного алфавита делят на две части, суммарные вероятности символов которых максимально близки друг другу.
3. В префиксном коде для первой части алфавита присваивается двоичная цифра «0», второй части — «1».
4. Полученные части рекурсивно делятся и их частям назначаются соответствующие двоичные цифры в префиксном коде.

Методика Шеннона-Фэно

- Наибольший эффект сжатия получается в случае, когда вероятности букв представляют собой целочисленные отрицательные степени двойки. Среднее число символов на букву в этом случае точно равно энтропии источника.

- На шаге деления алфавита существует неоднозначность, так как разность суммарных вероятностей может быть одинакова для двух вариантов разделения (учитывая, что все символы первичного алфавита имеют вероятность больше нуля).

Методика Шеннона-Фэно Пример:

Символ источника	Вероятность появления $P(a_i)$	Разбиение на подгруппы	Кодовые слова длиной L_i
a_1	0,4	I	1
a_2	0,2	I	01
a_3	0,2	I	001
a_4	0,1	II I	0001
a_5	0,05	II II II I	00001
a_6	0,05	II	00000

Рис. 1. Кодирование методом Шеннона-Фано

В равномерном коде длина сообщения n=3

Среднее число двоичных символов кода Шеннона-Фано, приходящихся на один символ алфавита источника, равно:

$n_{cp} = \Sigma L_i * P(a_i) = 1*0,4 + 2*0,2 + 3*0,2 + 4*0,1 + 5(0,05+0,05) = 2,3$ бит

Методика Шеннона-Фэно

- В равномерном коде длина сообщения n=3
- $2^2<6<2^3$,

- Среднее число двоичных символов кода Шеннона-Фано, приходящихся на один символ алфавита источника, равно:
- ncp= Σ Li*P(ai) =1*0,4+ 2*0,2+3*0,2+4*0,1+5(0,05+0,05)=2,3 бит
- Энтропия источника равна:
-

$$H(A) = -\sum_{i=1}^{6} P(a_i) * \log_2 P(a_i) = 2{,}22 \qquad бит$$

6

Методика Шеннона-Фэно

Условие эффективного кодирования:

$\max H(Z)$: $\log_2 m \geq l_{ср} \geq H(Z) + \varepsilon$,

- но $H(Z) < l_{ср.}$

- Следовательно, некоторая избыточность в последовательностях символов осталась.

- Из теоремы Шеннона следует, что эту избыточность можно устранить, если перейти к кодированию достаточно большими блоками.

Методика Шеннона-Фэно кодирование блоками

- Рассмотрим сообщения, образованные с помощью алфавита, состоящего из 2-х букв Z_1 и Z_2
- с вероятностями появления $P(Z_1) = 0.7$ и $P(Z_2) = 0.3$.
- Поскольку $P(Z_1)$ не равно $P(Z_2)$, то последовательность из таких букв будет обладать избыточностью. Однако при буквенном кодировании мы никакого эффекта не получим.
- Действительно, на передачу каждой буквы требуется либо 1, либо 0, то есть
- $I_{ср} = 1 \cdot 0.9 + 1 \cdot 0.1 = 1$,
- в то время как
- $H(Z) = -0.7\log_2 0.7 - 0.3\log_2 0.3 = 0.88$бит.
- Избыточность составляет R(A)=0,12 бит/символ.

Методика Шеннона-Фэно кодирование блоками

Комбинация символов	$P(a_i)$	Разбиение на подгруппы	Кодовое слово
AA	0,49	I	1
AB	0,21	I	01
BA	0,21	II I	001
BB	0,09	II II	000

В этом случае средняя длина кодового слова составляет:

с =1,81 бит.

На один символ алфавита источника приходится в среднем 1,81/2=0,905 бит/символ.

Избыточность составляет R(A)=0,905 – 0,88=0,025 бит/символ.

Методика Хаффмена

От указанного недостатка свободна методика Хаффмена.

МХ гарантирует однозначное построение кода с наименьшей для данного распределения вероятностей средним числом символов на букву.

Метод широко применяется в факсимильных устройствах.

Построение кода Хаффмана основывается на преобразовании, которое называется сжатием алфавита.

Методика Хаффмена
Алгоритм:

1. Расположить символы исходного алфавита А в порядке убывания вероятности.
2. Два наименее вероятных символа алфавита А будем считать одним символом нового сжатого алфавита А1.
3. Располагаем символы алфавита А1 в порядке убывания вероятности. И снова подвергаем его сжатию.
4. Повторяем процедуру, пока не придем к алфавиту, содержащему всего два символа.

Методика Хаффмена

5. Припишем символам последнего алфавита кодовые обозначения 0 (например - верхнему) и 1 (в нашем примере – нижнему). Это старшие символы будущих кодовых слов.
6. В предпоследнем алфавите кодовые обозначения получаются следующим образом:
   - Символ, который сохранился в последнем алфавите, имеет то же кодовое обозначение.
   - Символам, которые слились в последнем алфавите, приписывают справа 0 (в нашем примере – верхнему символу) и 1 (нижнему символу).
7. Повторяем процедуру, последовательно возвращаясь к исходному алфавиту.

Методика Хаффмена пример 1

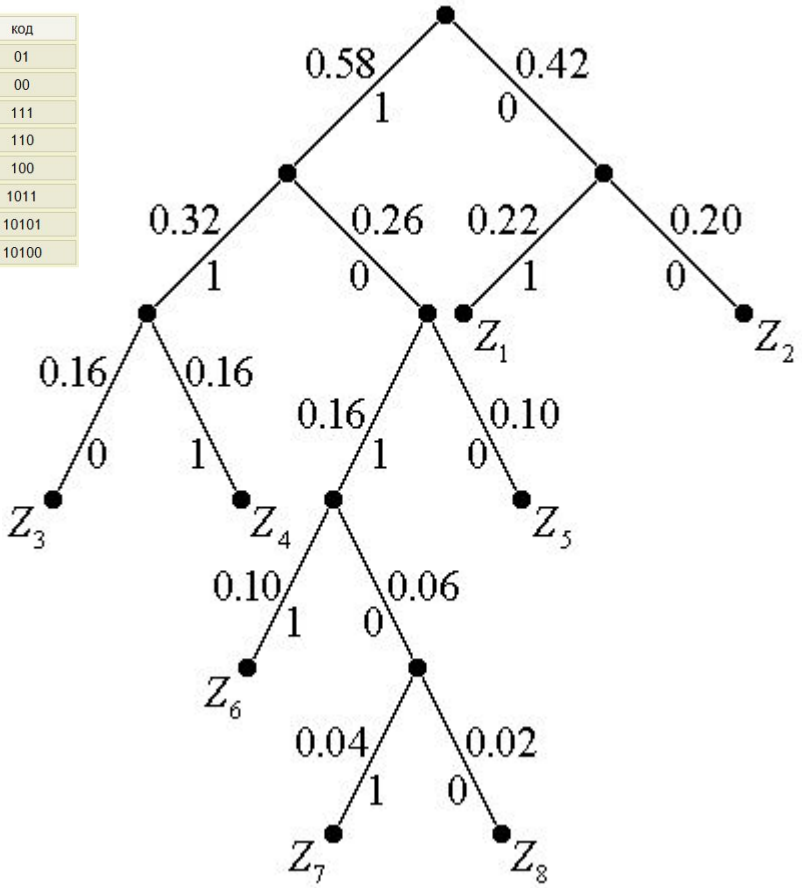
Вероятности и кодовые обозначения						
A	$P(a_i)$	Кодовые слова	A_1	A_2	A_3	A_4
						0,6 **0**
a_1	0,4	**1**	0,4 **1**	0,4 **1**	0,4 **1**	0,4 **1**
					0,4	**00**
a_2	0,2	**01**	0,2 **01**	0,2 **01**	0,2	**01**
a_3	0,2	**000**	0,2 **000**	0,2	**000**	
a_4	0,1	**0010**	0,1 **0010**			
				0,2		
a_5	0,05	**00110**				
			0,1 **0011**			
a_6	0,05	**00111**				

Методика Хаффмена пример2

		1	2	3	4	5	6	7	код
Z_1	0,22	0,22	0,22	0,26	0,32	0,42	0,58	1	01
Z_2	0,20	0,20	0,20	0,22	0,26	0,32	0,42		00
Z_3	0,16	0,16	0,16	0,20	0,22	0,26			111
Z_4	0,16	0,16	0,16	0,16	0,20				110
Z_5	0,10	0,10	0,16	0,16					100
Z_6	0,10	0,10	0,10						1011
Z_7	0,04	0,06							10101
Z_8	0,02								10100

Методика Хаффмена пример 2

		1	2	3	4	5	6	7	код
Z_1	0,22	0,22	0,22	0,26	0,32	0,42	0,58	1	01
Z_2	0,20	0,20	0,20	0,22	0,26	0,32	0,42		00
Z_3	0,16	0,16	0,16	0,20	0,22	0,26			111
Z_4	0,16	0,16	0,16	0,16	0,20				110
Z_5	0,10	0,10	0,16	0,16					100
Z_6	0,10	0,10	0,10						1011
Z_7	0,04	0,06							10101
Z_8	0,02								10100



Методика Хаффмена

Доказано, что код Хаффмана является самым экономным в том смысле, что никакой другой метод кодирования алфавита не позволяет получить среднее число

- двоичных символов на один символ алфавита меньше, чем в случае кода Хаффмана.