Занятие № 3 «Кодирование источника»

Условия существования оптимального неравномерного кода

При передаче сообщения осуществляется его преобразование в сигнал, пригодный для передачи по каналу связи. При этом необходимо согласовывать источник с каналом путём определения правила, по которому каждому элементу сообщения ставится в соответствие некоторый код, преобразуемый далее в сигнал.

В настоящее время существует два основных направления развития теории кодирования. В одном из них рассматриваются вопросы повышения достоверности передачи в каналах с помехами, решаемые применением помехоустойчивых кодов, которые позволяют обнаруживать или исправлять ошибки. Такое кодирование называется помехоустойчивым. При этом избыточность кодовой последовательности выше, чем избыточность источника сообщений. Благодаря этому и оказывается возможным обнаружение и исправление ошибок передачи.

Другое направление теории кодирования связано с вопросами устранения избыточности при передаче сообщений в каналах без помех. Цель кодирования при этом состоит в таком преобразовании сообщения, при котором избыточность кодовой последовательности должна стать меньше, чем избыточность сообщений источника. В результате появляется возможность увеличения скорости передачи информации, либо снижаются требования к пропускной способности канала. Это особенно важно в случае, когда источники сообщений имеют большую избыточность, например, источники речевых сообщений, изображений и т.д.

Процесс кодирования с целью уменьшения избыточности источника сообщений носит название согласования источника с каналом или сжатия источника (экономного кодирования, энтропийного кодирования).

Основными задачами, решаемыми кодированием в процессе передачи сообщений, являются:

согласование источника сообщений с каналом по объемам алфавитов;

повышение скорости передачи информации по каналу за счет устранения избыточности в последовательности сообщений;

повышение помехоустойчивости передачи информации.

Первые две задачи решаются в кодере источника сообщений. Третья задача решается в кодере канала.

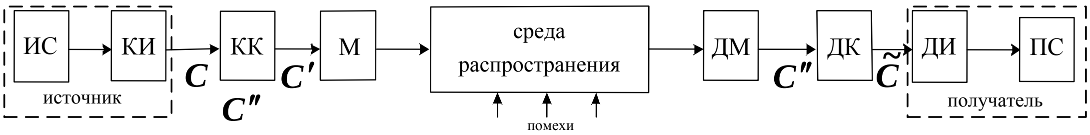

Рис. 1. Структурная схема системы электрической связи

Классификация рассматриваемых методов кодирования приведена на рис. 2. Эта классификация не является исчерпывающей, в нее включены лишь некоторые методы, которые широко используются в современных системах связи.

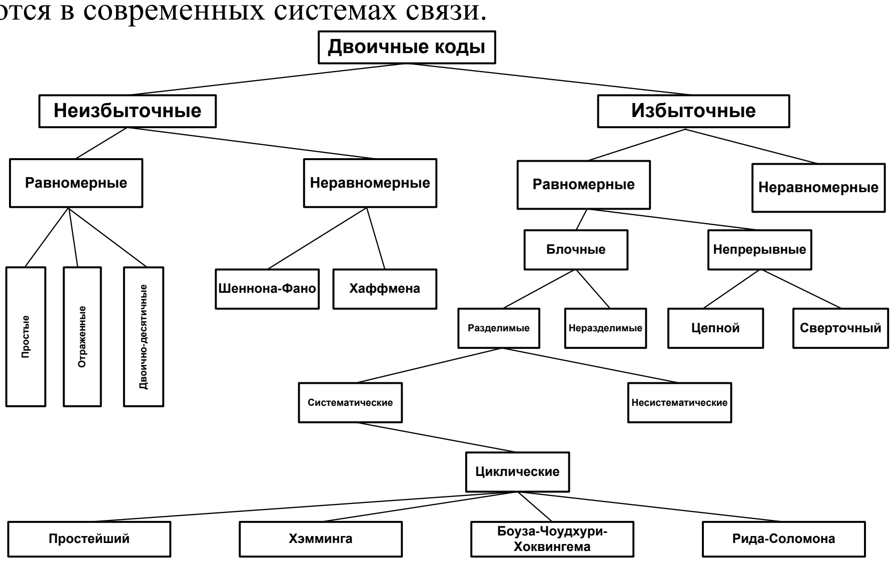

Рис. 2. Классификация кодов

1. Количественное определение информации

В основу измерения количества информации положены вероятностные характеристики передаваемых сообщений, которые не связаны с конкретным содержанием сообщений, а отражают степень их неопределенности. Естественно, что чем меньше вероятность сообщения, тем больше информации оно несет.

Количество информации $I(x_i)$ в отдельно взятом единичном сообщении x_i определяется величиной, обратной вероятности появления сообщения $p(x_i)$ и вычисляется в логарифмических единицах:

$$I(x_i) = log_b\left(\frac{1}{p(x_i)}\right) = -log_b(p(x_i)). \qquad (1)$$

Если источник выдает зависимые сообщения $x_i = x_1,..., x_m$, то они характеризуются условными вероятностями $p\left(\frac{x_i}{x_1,..., x_m}\right)$. И в этом случае количество информации вычисляется по формуле (1) с подстановкой в нее условных вероятностей сообщений.

Единицы измерения количества информации.

Выбор основания логарифмов в формуле (1) определяет единицы измерения количества информации. При использовании десятичного логарифма $(b = 10)$ информация измеряется в десятичных единицах – **дитах**. В случае использования натуральных логарифмов единицей измерения является натуральная единица – **нат**.

Более удобно в системах, работающих с двоичными кодами, использовать основание логарифма $b = 2$ и тогда информация измеряется в двоичных единицах – **дв.ед**. Весьма часто вместо двоичных единиц используется эквивалентное название – **бит**, возникшее как сокращенная запись английских слов binary digit (двоичная цифра). **1 бит** это количество информации, которое передается единичным символом сообщения, вероятность передачи которого $p(x_i) = 0,5$:

$$I(x_i) = -log_2(0,5) = -log_2\left(\frac{1}{2}\right) = -log_2(2^{-1}) = 1 \text{ (бит)}.$$

В настоящее время термин бит в информатике, вычислительной и импульсной технике употребляется не только как единица количества информации, но и для обозначения числа двоичных символов 0 и 1, поскольку они обычно равновероятны и каждый из них несет 1 бит информации.

Количество информации в сообщении, составленном из n символов, определяется по формуле [2]:

$$I(x,n) = -n \cdot \sum_{i=1}^{m} p(x_i) \cdot log_2\ p(x_i),$$

где $i = \{1,..., m\}$ – номер символа x_i из алфавита источника;

$p(x_i)$ – вероятность передачи i-го символа.

Энтропия источника сообщений

Для большинства реальных источников сообщения имеют разные вероятности. Например, в тексте буквы **А, О, Е** встречаются сравнительно часто, а **Щ, Ы** – редко.

При разных вероятностях сообщения несут различное количество информации $I(x_i)$. При решении большинства практических задач необходимо знать ***среднее количество информации***, приходящееся на один элемент сообщения. Это среднее количество информации при общем числе элементов сообщения источника n и числе символов алфавита m равно:

$$H(X) = \frac{I(x,n)}{n} = -\sum_{i=1}^{m} p(x_i) \cdot log_2\ p(x_i) \text{ (бит/сообщение)}. \qquad (2)$$

Величину $H(X)$ называют энтропией источника сообщений. Энтропия всегда положительна и принимает максимальное значение при равновероятных сообщениях:

$$H_{max}(X) = -\sum_{i=1}^{m} \frac{1}{m} \cdot log_2\left(\frac{1}{m}\right) = log_2(m). \qquad (3)$$

Минимальное значение энтропии $H_{min}(X) = 0$ соответствует случаю, когда одна из вероятностей $p(x_i) = 1$, а остальные равны нулю, т.е. имеется полная определенность.

Энтропия двоичного источника будет максимальна, существует полная неопределенность, т.е. $p(x_2) = p(x_1) = 0{,}5$. При этом $H(X) = log_2 m = 1$.

Избыточность источника сообщений

Избыточными в источнике являются сообщения, которые несут малое, иногда нулевое, количество информации. Наличие избыточности означает, что часть сообщений можно и не передавать по каналу связи, а восстановить на приеме по известным статистическим связям.

Количественно избыточность оценивается коэффициентом избыточности:

$$\chi = \frac{H_{\max}(X) - H(X)}{H_{\max}(X)} = 1 - \frac{H(X)}{H_{\max}(X)}, \qquad (4)$$

где $H(X)$ – энтропия источника; $H_{max}(X) = log_2 m$ – максимальная энтропия источника с алфавитом из m сообщений.

Количество информации переданной по дискретному каналу

Основной задачей систем связи является передача информации от источника к получателю. Решение этой задачи сопряжено с определенными трудностями, связанными не только с представлением информации в виде сообщения, пригодного для восприятия и обработки, но и с преобразованием данного сообщения в сигнал, пригодный для передачи по линии связи. Представим графически процесс передачи информации по каналу связи (рис. 3) и учтем при этом возможное влияние помех.

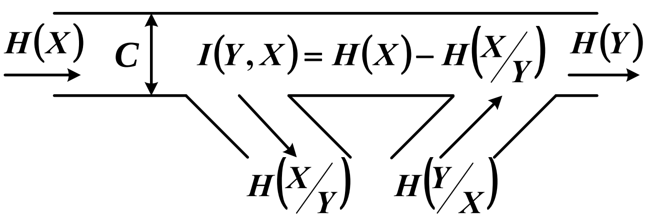

Рис. 3. Иллюстрация передачи информации по каналу связи

Пусть дискретный канал определяется:

$X = \{x_i\}$ – алфавитом источника сообщений;

$Y = \{y_j\}$ – алфавитом получателя сообщений;

$H(X/Y)$ – «потерями» информации;

$H(Y/X)$ – ложной информацией, создаваемой помехами;

$I(Y,X)$ – количеством информации переданной по каналу.

Условная энтропия характеризует среднюю степень неопределенности принимаемых сигналов, обусловленную действием помех.

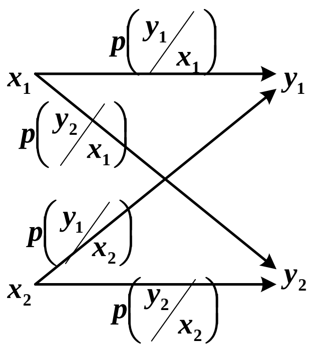

Рис. 4. Вероятностные характеристики передачи информации

При сопряжении входа канала с любым источником двоичной информации на вход могут поступать двоичные символы x_1 и x_2 с вероятностями $p(x_1)$ и $(1 - p(x_1))$ соответственно (рис. 4). На выходе канала появляются двоичные символы y_1 и y_2. Вероятности ошибки при передаче любого символа равна $p_{ош}$. Таким образом, $p(y_1/x_1) = p(y_2/x_2) = 1 - p_{ош}$, а $p(y_1/x_2) = p(y_2/x_1) = p_{ош}$, в общем случае, для m-ичного дискретного канала:

$$p(y_j/x_i) = \begin{cases} \dfrac{p_{ош}}{m-1} & \text{при } i \neq j, \\ 1 - p_{ош} & \text{при } i = j. \end{cases}$$

Тогда «шумовая» энтропия будет определяться выражением:

$$H\left(Y/X\right)=\sum_{i=1}^{m}p(x_i)\sum_{j=1}^{m}p\left(y_j/x_i\right)\cdot log_2\frac{1}{p\left(y_j/x_i\right)}= \tag{5}$$

$$=-\left(1-p_{ош}\right)\cdot log_2\left(1-p_{ош}\right)-p_{ош}\cdot log_2\frac{p_{ош}}{m-1}.$$

Из полученного выражения следует, что энтропия определяемая только помехой не зависит от вероятности $p(x_i)$ появления символов на входе канала.

Если символы на входе канала выбираются независимо от предыдущих символов и равными вероятностями, то энтропия выходных символов $H(Y)$ достигает своего максимального значения, равного $log_2 m$.

Таким образом, **количество информации, переданной по каналу,** это разность между энтропией на выходе и энтропией шума, т.е.:

$$I(Y,X)=H(Y)-H\left(Y/X\right)=log_2 m+\left(1-p_{ош}\right)\cdot log_2\left(1-p_{ош}\right)+p_{ош}\cdot log_2\frac{p_{ош}}{m-1}. \tag{6}$$

Количество информации, переданной по каналу связи, обладает следующими основными свойствами:

$I(Y,X)\geq 0$, причем $I(Y,X)=0$ тогда и только тогда, когда входные и выходные сообщения в канале взаимно независимы;

$I(Y,X)\leq H(X)$, причем $I(Y,X)=H(X)$ тогда и только тогда, когда входная последовательность определяется однозначно по выходной последовательности, например, когда в канале нет помех;

$I(Y,X)=I(X,Y)=H(Y)-H(Y/X)$ следует из того, что количество информации не изменится, если входную и выходную последовательность поменять местами.

Пропускная способность дискретного канала

Пропускной способностью канала, рассчитанной на один входной символ – называется максимальное количество информации, передаваемое по каналу, причем максимум ищется по всем возможным источникам X, имеющим различные (произвольные) вероятностные характеристики:

$$C'=\max_X I(Y,X)\ [\textit{бит/символ}].$$

Часто более удобно пользоваться пропускной способностью канала, рассчитанной не на один входной символ, а на единицу времени:

$$C=\frac{1}{T}\cdot C'.$$

Величину C называют пропускной способностью канала в единицу времени или просто пропускной способностью.

Пропускная способность канала обладает следующими основными свойствами:

$C\geq 0$, $C=0$ тогда и только тогда, когда вход и выход канала статистически независимы;

$C\leq\frac{1}{T}\cdot log_2 m$ для канала без помех.

Пропускная способность дискретного канала, по которому передается m дискретных сигналов с учетом (6), вычисляется по формуле:

$$C=V_И\cdot\left[log_2 m+\left(1-p_{ош}\right)\cdot log_2\left(1-p_{ош}\right)+p_{ош}\cdot log_2\frac{p_{ош}}{m-1}\right], \tag{7}$$

где $V_И=\frac{1}{T}$ – скорость модуляции, бод; T – длительность сигнала; $p_{ош}$ – вероятность ошибки сигналов в канале.

2. Методы сжатия дискретных сообщений

Пусть источник имеет алфавит из четырёх символов *А*, *Б*, *В*, *Г* с вероятностями $p(A) = 0{,}5$; $p(Б) = 0{,}25$; $p(В) = p(Г) = 0{,}125$.

Энтропия такого источника:

$$H(X) = -\sum_{i=1}^{4} p(x_i) \cdot log_2\, p(x_i) = -0{,}5 \cdot log_2\, 0{,}5 - 0{,}25 \cdot log_2\, 0{,}25 -$$

$$-0{,}125 \cdot log_2\, 0{,}125 - 0{,}125 \cdot log_2\, 0{,}125 = 1{,}75.$$

Для передачи по каналу будем использовать равномерное кодирование, например, $A \to 00$, $Б \to 01$, $В \to 10$, $Г \to 11$. Тогда среднее число двоичных символов в сообщении, приходящихся на один символ источника, равно двум. Поскольку это на $12{,}5\%$ больше энтропии источника, то используемый код не является оптимальным.

Рассмотрим теперь неравномерный код: $A \to 0$; $Б \to 10$; $В \to 110$; $Г \to 111$. В этом случае среднее число двоичных символов, приходящихся на один символ источника в сообщении, при этом равно:

$$n_{ср} = \sum_{i=1}^{4} p(x_i) \cdot n_i = 0{,}5 \cdot 1 + 0{,}25 \cdot 2 + 0{,}125 \cdot 3 + 0{,}125 \cdot 3 = 1{,}75.$$

Таким образом среднее число двоичных символов, приходящихся на один символ источника, равно энтропии источника, т.е. для указанного источника неравномерный код оказывается более экономичным, чем равномерный.

Важно отметить, что при кодировании неравномерным кодом должна обеспечиваться возможность однозначного декодирования символов сообщения. Например для рассмотренного источника, нецелесообразно применять код: $A \to 0$; $Б \to 1$; $В \to 10$; $Г \to 11$, поскольку прием последовательности 10 может означать передачу символа $В$, или двух символов $Б$ и A. Неоднозно декодирование и пары символов 11. Для однозначного декодирования неравномерные коды должны удовлетворять условию префиксности: никакое более короткое слово не должно являться началом более длинного слова. Неравномерные коды, удовлетворяющие этому условию, называют префиксными.

Неравномерные коды позволяют затратить в среднем меньшее число двоичных символов на единичное информационное сообщение, однако, им присущ существенный недостаток: при возникновении ошибки она распространяется на все последующие элементы сообщения. Этот недостаток отсутствует в равномерных кодах. При кодировании равномерными кодами используется одно и то же число двоичных символов – блок, поэтому такие коды еще называют блоковыми.

Кодирование методом Шеннона-Фано

Кодирование методом Шеннона – Фано рассмотрим на примере.

Пусть алфавит источника содержит шесть элементов {*А, Б, В, Г, Д, Е*} с вероятностями $p(A) = 0{,}15$, $p(Б) = 0{,}25$, $p(В) = 0{,}1$, $p(Г) = 0{,}13$, $p(Д) = 0{,}25$, $p(Е) = 0{,}12$. Энтропия такого источника:

$$H(X) = -0{,}15 \cdot log_2\, 0{,}15 - 0{,}25 \cdot log_2\, 0{,}25 - 0{,}1 \cdot log_2\, 0{,}1 -$$

$$-0{,}13 \cdot log_2\, 0{,}13 - 0{,}25 \cdot log_2\, 0{,}25 - 0{,}12 \cdot log_2\, 0{,}12 = 2{,}492.$$

Алгоритм построения сжимающего кода Шеннона – Фано заключается в следующем:

1. Все m символов дискретного источника располагаются в порядке убывания вероятностей (табл. 2);

2. Образованный столбец символов делится на две группы таким образом, чтобы суммарные вероятности каждой группы мало отличались друг от друга;

3. Верхняя группа кодируется символом «1», а нижняя – «0»;

4. Каждая группа делится на две подгруппы с близкими суммарными вероятностями; верхняя подгруппа кодируется символом «1», а нижняя – «0»;

5. Процесс деления и кодирования продолжается до тех пор, пока в каждой подгруппе не окажется по одному символу сообщения источника;

6. Записывается код для каждого символа источника; считывание кода осуществляется с лева

на право.

Таблица 2

Построение кода Шеннона-Фано

Элемент сообщения	Вероятность элемента	Деление сообщения на группы и подгруппы	Код
Б	0,25	1	11
Д	0,25	1 0	10
А	0,15	1	011
Г	0,13	1 0	010
Е	0,12	0 1	001
В	0,1	0 0	000

При использовании простейшего равномерного кода для кодирования шести элементов алфавита источника потребуется по три двоичных символа на каждую букву сообщения. Если же используется код Шеннона – Фано, то среднее число символов на одну букву составляет:

$$n_{cp} = \sum_{i=1}^{6} p(x_i) \cdot n_i = 0{,}25 \cdot 2 + 0{,}25 \cdot 2 + 0{,}15 \cdot 3 + 0{,}13 \cdot 3 + 0{,}12 \cdot 3 + 0{,}1 \cdot 3 = 2{,}5,$$

что меньше, чем при простейшем равномерном коде и незначительно отличается от энтропии источника.

Кодирование методом Хаффмена

Рассмотрим кодирование методом Хаффмена (табл. 3)для источника сообщений, заданного выше.

Алгоритм построения сжимающего кода Хаффмена включает в себя следующие действия:

1. Все m символов дискретного источника располагаются в таблице в порядке убывания вероятностей;

2. Два символа, имеющих наименьшие вероятности, объединяются в один блок, а их вероятности суммируются;

3. Ветви скобки, идущей к большей вероятности, присваивается символ «1», а идущей к меньшей – символ «0»;

4. Операции 2 и 3 повторяются до тех пор, пока не сформируется один блок с вероятностью единица;

5. Записывается код для каждого символа источника; при этом считывание кода осуществляется справа налево.

Среднее число символов на одну букву для полученного кода составляет:

$$n_{cp} = \sum_{i=1}^{6} p(x_i) \cdot n_i = 0{,}25 \cdot 2 + 0{,}25 \cdot 2 + 0{,}15 \cdot 3 + 0{,}13 \cdot 3 + 0{,}12 \cdot 3 + 0{,}1 \cdot 3 = 2{,}5,$$

Таблица 3

Построение кода Хаффмена

Элемент сообщения	Вероятность элемента	Деление сообщения на группы и подгруппы	Код
Б	0,25	0	10
Д	0,25	1 (0,53) 1	01
А	0,15	1 (0,28) (1) 1	111
Г	0,13	0 (0,47)	110
Е	0,12	1 (0,22) 0	001
В	0,1	0 0	000

Таким образом для данного примера кодирование методами Хаффмена и Шеннона–Фано приводит к одинаковой эффективности. Однако, опыт кодирования показывает, что код Хаффмена частот оказывается экономичнее кода Шеннона–Фано.