МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«МОРДОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
им. Н. П. ОГАРЁВА»

ОСНОВЫ ЦИФРОВОЙ РАДИОСВЯЗИ.

ЭНТРОПИЯ И ИЗБЫТОЧНОСТЬ ИСТОЧНИКА ИНФОРМАЦИИ

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

САРАНСК
ИЗДАТЕЛЬСТВО МОРДОВСКОГО УНИВЕРСИТЕТА
2013

УДК 621.391

Составители: *Д. В. Пьянзин, В. И. Королев*

Рецензент: *С. М. Мурюмин*, кандидат физико-математических наук, доцент

Основы цифровой радиосвязи. Энтропия и избыточность источника информации : метод. указания / сост.: Д. В. Пьянзин, В.И. Королев. – Саранск : Изд-во Мордов. ун-та, 2013. – 20 с.

Приведены сведения, поясняющие основные характеристики источников информации, а также алгоритмы энтропийного кодирования. Текстовый материал дополняется примерами кодирования сообщений алгоритмами Шеннона – Фано и Хаффмана, а также заданиями для практического выполнения.

Предназначено для студентов специальностей «Радиотехника» и «Радиоэлектронные системы и комплексы» очной и вечерней форм обучения, может быть полезно студентам других специальностей, связанных с электронной техникой.

Учебное издание

ОСНОВЫ ЦИФРОВОЙ РАДИОСВЯЗИ.

ЭНТРОПИЯ И ИЗБЫТОЧНОСТЬ ИСТОЧНИКА ИНФОРМАЦИИ

Методические указания

Составители: **Пьянзин Денис Васильевич**
Королев Валерий Иванович

Печатается в авторской редакции
в соответствии с представленным оригинал-макетом

Подписано в печать .09.13. Формат 60×84 1/16. Усл. печ. л. 1,4.
Тираж 100 экз. Заказ № .
Издательство Мордовского университета
Типография Издательства Мордовского университета
430005, г. Саранск, ул. Советская, 24

ПРЕДИСЛОВИЕ

В современном виде теория информации – это научная дисциплина, изучающая способы передачи и хранения информации наиболее надежным и экономным методом.

Под информацией (от латинского *information* – разъяснение, изложение) первоначально подразумевались сведения о процессах в обществе и природе, передаваемые устно, письменно или другим способом, а также сам процесс передачи или получения этих сведений. В середине XX века в связи с бурным развитием науки и техники произошли изменения в трактовке понятия информации. Оно было расширено путем включения в него обмена сведениями не только между людьми, но и между человеком и автоматом, между автоматом и автоматом, а также обмена сигналами в животном и растительном мире.

Сама по себе информация не является материей. Материальной формой представления информации являются сообщения и сигналы. Сообщения служат для обработки, преобразования, хранения и непосредственного использования информации, а сигналы – для ее передачи по каналу связи от источника к получателю (при этом информация должна быть соответствующим образом закодирована). Примеры сообщений: человеческая речь, изображение и т. д.

Для сравнения между собой различных источников сообщений необходимо было ввести некоторую количественную меру, которая дала бы возможность объективно оценивать информацию, содержащуюся в сообщении. Такая мера впервые была введена К. Шенноном – это энтропия источника.

Данные методические указания предназначены для выполнения практических и лабораторных работ по дисциплинам «Основы цифровой радиосвязи» и «Основы цифрового телевидения».

1. КОЛИЧЕСТВЕННАЯ МЕРА ИНФОРМАЦИИ

Как известно, любая информация может быть получена только в результате опыта. Под опытом будем понимать, например, просмотр телепередач, визуальное наблюдения какого-либо события, измерение некоторого параметра экспериментального процесса тем или иным прибором и т. п. При этом до проведения опыта должна существовать некоторая неопределенность в том или ином его исходе, так как если получателю заранее известно, какое сообщение он получит, то, получив его, он не приобретет никакого количества информации. Таким образом, до опыта всегда имеется большая или меньшая неопределенность в интересующей нас ситуации. После опыта (после получения информации) ситуация становится более определенной и на поставленный вопрос можно ответить однозначно, либо число возможных ответов уменьшится и, следовательно, уменьшится существовавшая ранее неопределенность. Количество уменьшенной неопределенности после опыта, очевидно, можно отождествить с количеством полученной информации.

Таким образом, чтобы установить формулу для вычисления количества информации *I,* необходимо уметь вычислять неопределенность некоторой ситуации до и после опыта.

Предположим вначале, что после проведения опыта неопределенность отсутствует. Примером такой ситуации может служить бросание монеты, когда возможны две ситуации: выпадет «орел» или «решка». Пусть после проведения опыта неопределенности исхода нет – выпал «орел». В этой ситуации к количеству информации (или, что-то же самое, к количеству неопределенности до опыта) можно предъявить следующие требования:

1. Количество получаемой информации больше в том опыте, у которого большее число возможных исходов *n*:

$$I(n_1) \geq I(n_2), \ \textit{если} \ n_1 \geq n_2. \qquad (1)$$

2. Опыт с единственным исходом несет количество информации, равное нулю:

$$I(n=1)=0. \qquad (2)$$

3. Количество информации от двух независимых опытов должно равняться сумме количеств информации от каждого из них:

$$I(n_1 n_2) = I(n_1) + I(n_2), \qquad (3)$$

так как опыт, объединяющий два опыта с исходами n_1 и n_2 имеет $(n_1 n_2)$ исходов.

Очевидно, что единственной функцией аргумента *n*, удовлетворяющей трем поставленным условиям, является логарифмическая функция (логарифм единицы равен нулю; логарифм произведения равен сумме логарифмов сомножителей). Тогда выражение количества информации от опыта с *n* исходами при условии, что после опыта неопределенность отсутствует, имеет следующий вид:

$$I = c \log_a n, \qquad (4)$$

где *c* и *a* – некоторые постоянные.

При получении формулы (4) не различали исходы опыта по степени их возможности наступить, т. е. маловероятный исход опыта не отличали от исхода опыта, имеющего большую вероятность. Поэтому исходы опыта в (4) следует считать рав-

новероятными, т. е. вероятность любого исхода $p = \frac{1}{n}$ (равномерное распределение вероятностей). Учитывая это, представим формулу (4) в следующем виде:

$$I = -c \log_a p. \qquad (5)$$

Выбор постоянной c и основания логарифма здесь несущественен, так как переход от одной системы логарифмов к другой сводится лишь к простому изменению масштаба I. Поэтому для простоты полагают $c = 1$, а с учетом того, что на практике удобнее всего пользоваться логарифмами с основанием $a = 2$ (это хорошо согласуется с двоичной системой счисления), выражение (5) приобретает следующий вид [7]:

$$I = -\log_2 p. \qquad (6)$$

В данном случае единицу количества информации называют бит. Как видно из (6), бит есть количество информации, получаемое в результате опыта с двумя равновероятными исходами. В дальнейшем по тексту, если не оговорено, под символом log будем понимать двоичный логарифм.

Формула (6) изменится, если исходы опыта не обладают равной вероятностью. Пусть у опыта X имеется n исходов x_i с соответствующими вероятностями наступления p_i (i = 1, 2,..., n) и $\sum_{i=1}^{n} p_i = 1$. В этом случае количество информации о том, что наступил исход x_i, является уже случайной величиной:

$$I(x_i) = -\log_2 p(x_i) \qquad (7)$$

Пример 1. В городе 25 % населения составляют студенты. Среди студентов 50 % юношей. Всего же юношей в городе 35 %.

Сколько дополнительной информации содержится в сообщении, что встречный юноша – студент?

Решение. Обозначим событие «встречен юноша» через x_1, а событие «встречен студент» через x_2. По теореме умножения вероятность совместного наступления событий x_1 и x_2 (вероятность совместного наступления двух событий равна произведению одного из них на условную вероятность другого, вычисленную при условии, что первое имело место; при этом безразлично, какое событие произошло первым) равна:

$$P(x_1)P(x_2 \mid x_1) = P(x_2)P(x_1 \mid x_2).$$

По условию задачи $P(x_1) = 0{,}35$; $P(x_2) = 0{,}25$; $P(x_1 \mid x_2) = 0{,}5$. Подставив эти значения в предыдущую формулу, найдем вероятность того, что встреченный юноша – студент:

$$P(x_2 \mid x_1) = \frac{0{,}25 \cdot 0{,}5}{0{,}35} = 0{,}357.$$

Искомое количество информации

$$I = -\log_2 P(x_2 \mid x_1) = -\log_2 0{,}357 = 1{,}49 \text{ бит}.$$

2. ЭНТРОПИЯ И ИЗБЫТОЧНОСТЬ ИСТОЧНИКА СООБЩЕНИЙ

Рассмотрим источник информации, который выдает последовательность независимых дискретных сообщений x_i. Каждое сообщение случайным образом выбирается из алфавита источника $X = x_1,\ldots,x_n$ (n – размер алфавита источника). Такой источник информации называется источником без памяти с конечным дискретным алфавитом, а сообщения, вырабатываемые им, называются простыми. В дальнейшем для упрощения расчетов будем работать именно с такими источниками.

Количество информации, содержащейся в одном элементарном сообщении источника (см. формулу 7), еще никак его не характеризует, так как одни элементарные сообщения могут нести в себе много информации, но при этом передаваться редко, а другие сообщения могут нести мало информации, но передаваться часто. Поэтому источник может быть охарактеризован средним количеством информации, приходящимся на одно элементарное сообщение – энтропия источника [5, 6, 7]:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \text{ [бит]}, \tag{8}$$

где X – алфавит сообщений источника информации; n – размер алфавита источника.

Энтропия обладает рядом свойств.

1. Во-первых, $H(X) \geq 0$. Положительность $H(X)$ видна из (8), так как вероятности положительны и заключены между нулем и единицей; логарифмы таких чисел отрицательны, а равенство нулю возможно только для такого случая, когда вероятность появления одного из сообщений источника равна единице, а для остальных – нулю.

2. Во-вторых, при заданном размере алфавита источника n энтропия максимальна и равна $\log_2 n$, когда вероятности появления сообщений источника равны, т.е. сообщения равновероятны.

3. В-третьих, энтропия обладает свойством аддитивности:

$$H(X,Y) = H(X) + H(Y), \tag{9}$$

где $H(X)$ – энтропия первого источника информации; $H(Y)$ – энтропия второго источника информации.

Пример 2. Представим источник сообщений в виде корзины, в которой находятся шары трех цветов: красный, зеленый и синий. Данные шары (сообщения) определяют размер алфавита источника.

Рассчитаем энтропию источника сообщений, если:

1) красных шаров – 7 шт., зеленых шаров – 5 шт., синих шаров – 2 шт.

2) красных, зеленых и синих шаров – 2 шт.

Решение: В корзине находятся шары трех цветов, следовательно, размер алфавита источника $n = 3$.

1) Вероятность появления красного шара $p_1 = \frac{7}{14} = \frac{1}{2}$; зеленого шара $p_2 = \frac{5}{14}$; синего шара $p_1 = \frac{2}{14} = \frac{1}{7}$;

Рассчитаем энтропию источника:

$$H(X) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{5}{14}\log_2\frac{5}{14} + \frac{1}{7}\log_2\frac{1}{7}\right) = 1{,}43 \text{ [бит].}$$

2) Вероятность появления красного шара $p_1 = \frac{2}{6} = \frac{1}{3}$; зеленого шара $p_2 = \frac{2}{6} = \frac{1}{3}$; синего шара $p_1 = \frac{2}{6} = \frac{1}{3}$.

Рассчитаем энтропию источника:

$$H(X) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 1{,}59 \text{ [бит].}$$

Как видно из вышеизложенного, энтропия определяет среднее число двоичных знаков, необходимых для кодирования исходных символов источника информации. Она максимальна, когда символы вырабатываются источником с равной вероятностью. Если же некоторые символы появляются чаще других, энтропия уменьшается. Чем меньше энтропия источника отличается от максимальной, тем рациональнее он работает, тем большее количество информации несут его символы.

Для сравнения источников по их информативности вводится параметр, называемый избыточностью и равный [5 – 7]:

$$R = \frac{H_{\max}(X) - H(X)}{H_{\max}(X)} = 1 - \frac{H(X)}{H_{\max}(X)}, \qquad (10)$$

где $H_{\max}(X)$ – максимальная энтропия источника.

Источник, избыточность которого $R = 0$, называют оптимальным. Все реальные источники имеют избыточность $R \neq 0$.

Предположим, что мы получили одинаковое количество информации I_0 от реального и оптимального источников. Тогда число символов k, затраченных на передачу этого количества информации реальным источником, будет больше числа символов $k_{\min}$, затраченных оптимальным источником. Зная число символов k и $k_{\min}$ можно также рассчитать избыточность:

$$R = 1 - \frac{k_{\min}}{k}. \qquad (11)$$

Избыточность увеличивает время передачи информации, поэтому она нежелательна. Однако при передаче сообщений при наличии помех в канале связи избыточность используется для увеличения помехозащищенности передаваемых сообщений (помехоустойчивое кодирование).

Пример 3. Пусть источник информации передает русский текст. Если не различать буквы «е» и «ё», а также мягкий и твердый знаки, то в русском алфавите 31 буква; добавим пробел между словами и получим 32 символа.

Покажем, что пятиразрядный двоичный код (код Боде) не является оптимальным для передачи русского текста.

Решение. В данном коде на представление каждой буквы тратятся пять элементарных символов. Максимальная энтропия источника, использующего для передачи русского алфавита пятизначный код Боде равна $H_{\max}(X) = \log_2 32 = 5$ (бит). В данном случае считается, что все буквы русского алфавита имеют одинаковую вероятность и статически независимы.

С учетом различной вероятности появления букв в тексте энтропия равна:

$$H(X) = 4{,}42 \text{ [бит]}.$$

С учетом корреляции между двумя и тремя соседними буквами энтропия равна:

$$H(X) = 3{,}52 \text{ [бит]}.$$

С учетом корреляции между восемью и более символами энтропия равна:

$$H(X) = 2 \text{ [бит]}.$$

Далее все остается без изменений.

Рассчитаем избыточность представленного источника информации при кодировании символов пятиразрядным двоичным кодом Боде:

$$R = 1 - \frac{H(X)}{H_{\max}(X)} = 1 - \frac{2}{5} = 0{,}6.$$

Таким образом, можно сделать вывод, что каждые 6 букв из десяти являются избыточными и могут просто не передаваться, т.е. избыточность русского текста составляет 60 %.

Такой же и более высокой избыточностью обладают и другие источники информации – речь, музыка, ТВ - изображения и т.д.

Зная энтропию $H(X)$ и время $T_{\text{ср}}$, которое занимает в среднем каждое элементарное сообщение, можно рассчитать одну из важнейших характеристик источника– производительность (среднее количество информации в единицу времени) [5 – 7]:

$$\overline{H(X)} = \frac{H(X)}{T_{\text{ср}}}, \left[\frac{\text{бит}}{\text{с}}\right]. \quad (12)$$

Время $T_{\text{ср}}$ рассчитывается следующим образом:

$$T_{\text{ср}} = \sum_{i=1}^{n} T_i p(x_i), \quad (13)$$

где T_i – длительность i-го сообщения; $p(x_i)$ – вероятность появления i-го сообщения.

3. КОДЫ ШЕННОНА – ФАНО

Одно и то же сообщение можно закодировать различными способами. Наиболее выгодным является такой код, при использование которого на передачу сообщений затрачивается минимальное время. Если на передачу каждого элемента символа (например, 0 или 1) тратится одно и то же время, то оптимальным будет такой код, при использование которого на передачу сообщения заданной длины будет затрачено минимальное количество элементарных символов. Коды Шеннона – Фано являются префиксными, т.е. никакое кодовое слово не является префиксом любого другого. Данное свойство позволяет однозначно декодировать любую последовательность кодовых слов.

Рассмотрим принцип построения одного из первых алгоритмов сжатия, который сформулировали американские ученые Шеннон и Фано, на примере букв русского алфавита. Алгоритм использует коды переменной длины, т.е. часто встречающийся символ кодируется кодом меньшей длины, редко встречающийся – кодом большей длины [3, 7].

Чтобы составить такой код, очевидно, нужно знать частоты появления букв в русском тексте. Эти частоты приведены в таблице 1 [3]. Буквы в таблице расположены в порядке убывания частот.

Таблица 1

Частота появления букв русского алфавита

Буква	Частота	Буква	Частота	Буква	Частота	Буква	Частота
«−»	0,145	р	0,041	я	0,019	х	0,009
о	0,095	в	0,039	ы	0,016	ж	0,008
е	0,074	л	0,036	з	0,015	ю	0,007
а	0,064	к	0,029	ъ, ь	0,015	ш	0,006
и	0,064	м	0,026	б	0,015	ц	0,004
т	0,056	д	0,026	г	0,014	щ	0,003
н	0,056	п	0,024	ч	0,013	э	0,003
с	0,047	у	0,021	й	0,010	ф	0,002

Пользуясь таблицей, можно составить наиболее экономичный код на основе соображений, связанных с количеством информации. Очевидно, код будет самым экономичным, когда каждый элементарный символ будет передавать максимальную информацию. Рассмотрим элементарный символ (т. е. изображающий его сигнал) как физическую систему с двумя возможными состояниями: 0 и 1. Информация, которую дает этот символ, равна энтропии этой системы и максимальна в случае, когда оба состояния равновероятны; в этом случае элементарный символ передает информацию 1 (двоичная единица). Поэтому основой оптимального кодирования будет требование, чтобы элементарные символы в закодированном тексте встречались в среднем одинаково часто.

Идея кодирования состоит в том, что кодируемые символы (буквы или комбинации букв) разделяются на две приблизительно равновероятные группы: для первой группы символов на первом месте комбинации ставится 0 (первый знак двоичного числа, изображающего символ); для второй группы − 1. Далее каждая группа снова делится на две приблизительно равновероятные подгруппы; для символов первой подгруппы на втором месте ставится 0; для второй подгруппы − единица и т. д.

Продемонстрируем принцип построения кода Шеннона − Фано на примере материала русского алфавита (см. табл. 1). Отсчитаем первые шесть букв (от «−» до «т»); суммируя их вероятности (частоты), получим 0,498; на все остальные буквы от «н» до «ф» придется приблизительно такая же вероятность – 0,502. Первые шесть букв (от «−» до «т») будут иметь на первом месте двоичный знак 0. Остальные буквы (от «н» до «ф») будут иметь на первом месте единицу. Далее снова разделим первую группу на две приблизительно равновероятные подгруппы: от «−» до «о» и от «е» до «т»; для всех букв первой подгруппы на втором месте поставим нуль, а второй подгруппы − единицу. Процесс будем продолжать до тех пор, пока в каждом подразделении не останется ровно одна буква, которая будет закодирована определенном двоичным числом. Механизм построения показан на таблице 2, а сам код приведен в таблице 3.

Таблица 2

Механизм построения кода Шеннона – Фано на примере русского алфавита

	Двоичные знаки								
Буквы	1й	2й	3й	4й	5й	6й	7й	8й	9й
«–»	0	0	0						
о			1						
е		1	0	0					
а				1					
и			1	0					
т				1					
н	1	0	0	0					
с				1					
р			1	0	0				
в					1				
л				1	0				
к					1				
м		1	0	0	0				
д					1	0			
п						1			
у			0	1	0				
я					1	0			
ы						1			
з			1	0	0	0			
ъ, ь						1			
б					1	0			
г						1			
ч				1	0	0			
й						1	0		
х							1		
ж					1	0	0		
ю							1		
ш						1	0	0	
ц								1	
щ							1	0	
э									0
ф								1	1

Таблица 3

Результат кодирования букв русского алфавита кодом Шеннона - Фано

Буква	Двоичное число	Буква	Двоичное число	Буква	Двоичное число
«–»	000	к	10111	ч	111100
о	001	м	11000	й	1111010
е	0100	д	110010	х	1111011
а	0101	п	110011	ж	1111100
и	0110	у	110100	ю	1111101
т	0111	я	110110	ш	11111100
н	1000	ы	110111	ц	11111101
с	1001	з	111000	щ	11111110

р	10100	ъ, ь	111001	э	111111110
в	10101	б	111010	ф	111111111
л	10110	г	111011		

Пример 4. Запишем фразу «способ кодирования», используя код Шеннона – Фано.

Решение. Воспользуемся таблицей 3 и получим следующий результат:

(1001)с (110011)п (001)о (1001)с (001)о (111010)б (000)пробел
(10111)к (001)о (110010)д (0110)и (10100)р (001)о (10101)в
(0101)а (1000)н (0110)и (110110)я

Заметим, что здесь нет необходимости отделять друг от друга буквы специальным знаком, так как и без этого декодирование выполняется однозначно благодаря свойству префиксности: ни одна более короткая кодовая комбинация не является началом более длинной кодовой комбинации. Действительно, из таблицы 3 видно, что самыми короткими являются коды для символов «пробел» и «о». При этом ни один другой более длинный код не имеет в начале последовательности 000 («пробел») и 001 («о»). То же самое можно наблюдать и для всех других двоичных последовательностей кода Шеннона – Фано, которые приведены в таблице 3.

Необходимо отметить, что любая ошибка при кодирование (случайное перепутывание знаков 0 или 1) при таком коде губительна, так как декодирование всего следующего за ошибкой текста становится невозможным.

Пример 5. Определим, является ли рассмотренный нами код оптимальным при отсутствие ошибок.

Решение. Найдем среднюю информацию, приходящуюся на один элементарный символ (0 или 1), и сравним ее с максимально возможной информацией, которая равна единице. Для этого найдем сначала среднюю информацию, содержащуюся в одной букве передаваемого текста, т. е. энтропию на одну букву (см. формулу 8):

$$H(\textit{Буква}) = -\sum_{i=1}^{32} p(x_i)\cdot \log_2 p(x_i) = -(0{,}145\log_2 0{,}145 + 0{,}095\log_2 0{,}095 + \ldots + 0{,}003\log_2 0{,}003 + 0{,}002\log_2 0{,}002)$$

$\approx 4{,}42$ [бит].

По таблице 1 определяем среднее число элементарных символов на букву:

$$k_{\text{ср}} = 3\cdot 0{,}145 + 3\cdot 0{,}095 + 4\cdot 0{,}074 + \ldots + 9\cdot 0{,}003 + 9\cdot 0{,}002 = 4{,}45 \text{ [бит]}.$$

Далее рассчитаем информацию на один элементарный символ:

$$I_c = \frac{H(\textit{Буква})}{k_{\text{ср}}} = \frac{4{,}42}{4{,}45} \approx 0{,}994 \text{ [бит]}.$$

Таким образом, информация на один символ весьма близка к своему верхнему пределу – единице, а данный код весьма близок к оптимальному.

В случае использования пятиразрядного двоичного кода информация на один символ

$$I_c = \frac{4{,}42}{5} \approx 0{,}884 \text{ [бит]}.$$

Пример 6. Пусть по каналу связи получено сообщение (слово на русском языке), закодированное кодом Шеннона – Фано: 10111001110010010010100.

Необходимо декодировать данную последовательность.

Решение. Процесс декодирования основывается на свойстве префиксности кода и выполняется слева направо. Из таблицы 3 видно, что минимальная длина кода составляет три бита. Отсчитаем три бита от начала принятой кодовой комбинации, получим 101. В таблице такой код отсутствует, поэтому добавляем еще один бит, получим 1011. Данного кода также нет в таблице, следовательно, необходимо добавить еще один бит, получим комбинацию 10111, которой соответствует буква «к». Кодовая комбинация 10111 исключается из принятой кодовой комбинации и заменяется исходным символом (буква «к»). Процесс декодирования остальных букв принятого сообщения выполняется аналогично.

Полный процесс декодирования приведен в таблице 4. Знак «–» в таблице означает, что в таблице 3 отсутствует выбранный код.

Таблица 4

Процесс декодирования сообщения

Принятая кодовая последовательность																					
1	0	1	1	1	0	0	1	1	1	0	0	1	0	0	1	0	0	1	0	1	0	0
–			1	1	0	0	1	1	1	0	0	1	0	0	1	0	0	1	0	1	0	0
–				1	0	0	1	1	1	0	0	1	0	0	1	0	0	1	0	1	0	0
к					0	0	1	1	1	0	0	1	0	0	1	0	0	1	0	1	0	0
к					о			1	1	0	0	1	0	0	1	0	0	1	0	1	0	0
к					о			–			0	1	0	0	1	0	0	1	0	1	0	0
к					о			–				1	0	0	1	0	0	1	0	1	0	0
к					о			–					0	0	1	0	0	1	0	1	0	0
к					о			д						0	1	0	0	1	0	1	0	0
к					о			д						–			0	1	0	1	0	0
к					о			д						е				1	0	1	0	0
к					о			д						е				–			0	0
к					о			д						е				–				0
к					о			д						е				р				

Итак, слово, полученное в результате декодирования принятой кодовой комбинации, – «кодер».

4. КОДЫ ХАФФМАНА

Один из первых алгоритмов эффективного кодирования информации был предложен Д. А. Хаффманом в 1952 году. Идея алгоритма состоит в следующем: зная вероятности символов в сообщении, можно описать процедуру построения кодов переменной длины, состоящих из целого количества битов. Символам с большей вероятностью ставятся в соответствие более короткие коды. Коды Хаффмана обладают свойством префиксности, что позволяет однозначно их декодировать. В отличие от кодов Шеннона – Фано, которые в настоящее время практически не применяются, коды Хаффмана находят широкое применение при кодирование сообщений [1, 2, 4, 6, 7].

Классический алгоритм Хаффмана на входе получает таблицу частот встречаемости символов в сообщении. Далее на основании этой таблицы строится дерево кодирования Хаффмана.

Рассмотрим алгоритм построения кодов Хаффмана на следующем примере: пусть имеется источник с алфавитом X, включающий шесть сообщений $x_1,\ldots, x_6$. У каждого сообщения имеется своя вероятность появления, приведенная в таблице 5.

Таблица 5

Вероятность появления сообщений источника информации

Сообщения источника	x_1	x_2	x_3	x_4	x_5	x_6
Вероятность появления сообщения	0,3	0,2	0,15	0,15	0,1	0,1

1. Выбираем два сообщения с наименьшими вероятностями и заменяем их одним с вероятностью равной сумме вероятности данных сообщения (выбранные сообщения выделяем темным цветом) (табл. 6).

Таблица 6

Формирование промежуточных алфавитов

Действие	x_1	x_2	x_3	x_4	x_5	x_6
Объединяем сообщения	0,3	0,2	0,15	0,15	0,1	0,1
Упорядочиваем по вероятности появления	0,3	0,2	0,2	0,15	0,15	
Объединяем сообщения	0,3	0,2	0,2	0,15	0,15	
Упорядочиваем по вероятности появления	0,3	0,3	0,2	0,2		
Объединяем сообщения	0,3	0,3	0,2	0,2		
Упорядочиваем по вероятности появления	0,4	0,3	0,3			
Объединяем сообщения	0,4	0,3	0,3			
Упорядочиваем по вероятности появления	0,6	0,4				
Объединяем сообщения	0,6	0,4				
	1					

2. В результате проведенных операций мы получили 4 промежуточных алфавита ($X_1 – X_4$). Результат перепишем в следующем виде (табл. 7).

Таблица 7

Промежуточные алфавиты

Вероятности				
Исходный алфавит X	Промежуточные алфавиты			
	X_1	X_2	X_3	X_4
0,3	0,3	0,3	0,4	0,6
0,2	0,2	0,3	0,3	0,4
0,15	0,2	0,2	0,3	
0,15	0,15	0,2		
0,1	0,15			
0,1				

3. Далее проведем процедуру кодирования (табл. 8). Кодирование выполняется в обратном порядке от алфавита X_4 к исходному алфавиту X.

Таблица 8

Процедура кодирования

Вероятности				
Исходный алфавит X	Промежуточные алфавиты			
	X_1	X_2	X_3	X_4
0,3 (00)	0,3 (00)	0,3 (00)	0,4 (1)	0,6 (0)
0,2 (10)	0,2 (10)	0,3 (01)	0,3 (00)	0,4 (1)
0,15 (010)	0,2 (11)	0,2 (10)	0,3 (01)	
0,15 (011)	0,15 (010)	0,2 (11)		
0,1 (110)	0,15 (011)			
0,1 (111)				

Двум знакам последнего алфавита X_4 присваиваем коды 0 (сообщение с вероятностью 0,6) и 1 (сообщение с вероятностью 0,4). Условимся в дальнейшем, что верхний знак будет кодироваться символом 0, а нижний – 1.

Сообщение с вероятностью 0,6 алфавита X_4 было получено как сумма вероятностей двух сообщений алфавита X_3 с вероятностями 0,3 и 0,3. В данном случае эти сообщения кодируются уже двухразрядным кодом. Старшим разрядам обоих сообщений присваивается 0, так как нулем было закодировано сообщение с вероятностью 0,6 алфавита X_4. Младшему разряду верхнего сообщения в таблице присваивается 0, а нижнему – 1 (было определено выше). Сообщение с вероятностью 0,4 алфавита X_3 будет кодироваться так же, как и сообщение с этой же вероятностью в алфавите X_4.

По аналогии кодируются остальные сообщения алфавитов, в результате получаем исходный алфавит, закодированный кодом Хаффмана, в котором сообщения, имеющие большую вероятность, кодируются кодом меньшей длины, а с меньшей вероятностью – кодами большей длины.

Пример 7. Закодируем русский алфавит с помощью описанного выше алгоритма Хаффмана, используя таблицу 1.

Решение. Результат кодирования приведен в таблице 9.

Таблица 9

Результат кодирования букв русского алфавита кодом Хаффмана

Буква	Двоичное число	Буква	Двоичное число	Буква	Двоичное число
«–»	001	к	10001	ч	101111
о	111	м	11000	й	0001001
е	0100	д	11001	х	0101100
а	0110	п	000000	ж	0101101
и	0111	у	000001	ю	1011100
т	1001	я	000101	ш	00010000
н	1010	ы	010111	ц	10111010
с	1101	з	100000	щ	10111011
р	00001	ъ, ь	100001	э	000100010
в	00011	б	101100	ф	000100011
л	01010	г	101101		

Пример 8. Пусть по каналу связи получено сообщение (слово на русском языке), закодированное кодом Хаффмана: 10001101001111011010110.

Необходимо декодировать данную последовательность, используя таблицу 9.

Решение: Процесс декодирования основывается также на свойстве префиксности кода и выполняется слева направо (табл. 10).

Таблица 10

Процесс декодирования сообщения

Принятая кодовая последовательность																					
1	0	0	0	1	1	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	1	0
–			0	1	1	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	1	0
–				1	1	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	1	0
к					1	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	1	0
к					–			0	0	1	1	1	1	0	1	1	0	1	0	1	1	0
к					н				0	1	1	1	1	0	1	1	0	1	0	1	1	0
к					н				–			1	1	0	1	1	0	1	0	1	1	0
к					н				и				1	0	1	1	0	1	0	1	1	0
к					н				и				–			1	0	1	0	1	1	0
к					н				и				–				0	1	0	1	1	0
к					н				и				–					1	0	1	1	0
к					н				и				г						0	1	1	0
к					н				и				г						–			0
к					н				и				г						а			

Можно повысить эффективность кодирования, если строить код не для символа, как в рассмотренных выше примерах, а для блоков из n символов. В этом случае частота блока рассчитывается как произведение частот символов, которые входят в блок.

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ВЫПОЛНЕНИЮ ПРАКТИЧЕСКОЙ РАБОТЫ

1. Изучение кодов Шеннона – Фано.

1.1. Закодируйте согласно своему варианту задания словосочетания, приведенные в таблице 11, используя алгоритм Шеннона – Фано, полученный в п. 3 (см. табл. 3).

Таблица 11

Варианты заданий

Номер варианта	Словосочетания для кодирования
1	Теорема Шеннона
2	Канал связи
3	Цифровое телевидение
4	Энтропия и избыточность
5	Цифровой сигнал
6	Цифровая модуляция
7	Циклический код
8	Код Хэмминга
9	Скорость передачи
10	Преобразование Фурье
11	Вейвлет-преобразование
12	Сжатие изображений
13	Межкадровая корреляция
14	Внутрикадровое кодирование
15	Квантование сигнала
16	Нелинейное квантование
17	Дискретный сигнал
18	Теорема Котельникова
19	Квадратурная модуляция
20	Цифровой фильтр
21	Корреляция и свертка
22	Аддитивная помеха
23	Решающее устройство
24	Уплотнение каналов
25	Цифровая радиосвязь

1.2. Выполните декодирование сообщений, закодированных с помощью кода Шеннона – Фано, согласно варианту заданий (табл. 12). При декодировании используйте таблицу 3. Результат представьте по аналогии с примером 6.

Таблица 12

Варианты заданий

Номер варианта	Закодированные сообщения
1	0101100001011011001101110000001010001011110100010111110111
2	1011001101000010011110101000010111011000011111101010011011111001
3	0011110101010001010111100001011101100001001101011101101110000111001
4	1010111110110011100101000001111101000010010110111011110000010110110
5	0101101101110110011010001100111110000001010001011110100010111110111
6	1100001000111001110010000101000101111010001011111011
7	1010101101100100100001000101110011100100100100100
8	1100001011110111000011001111000001010000011001100110110100100
9	0011100110011010010001101111111010000100101101110111000010110110
10	1000010111110001011011011100110000101110110000111111111010111100000101
11	1100100100101010110010111111101011011011000011110001011001011100101111110111
12	1010001001111100011011000000111000010111110111010101010110101
13	1011100110110010011101001010111010010110111001100011011111111010
14	1100001011110111000011001111000110111111101000011001100101011100110111
15	0110100101110011111001000011010111000110011011001110101100000110110110
16	1100110110011101011111101111111100101110110000100101000111111001
17	1100110110110010000101000100111011110100101101101100111001100100
18	1100000111111110100000110010111111001000110011001110000100111101101110
19	110011011010110001011100010010110111011100001011011O
20	1001101010100011100110101001111101000011001100101110011011O111
21	0011110101111111001011101100000111001111100101110101
22	0101101110111011010101100001011101100000101100001110100100010000101
23	1000011011111001000011001000001111000101100101110010111110111
24	1111111110101111000001101011101111111010000110010010001110100101110111001101O0
25	1100111010001100100110001000011010111000110011001110000100111101

1.3. Введите в закодированное сообщение (табл. 12) ошибку в любой из разрядов кода и выполните его декодирование. Сделайте вывод по полученным результатам.

1.4. Закодируйте буквы английского алфавита кодом Шеннона – Фано. Частота появления букв приведена в таблице 13.

Таблица 13

Частота появления букв английского алфавита

Буква	Частота	Буква	Частота	Буква	Частота	Буква	Частота
e	0,127	h	0,0609	w	0,0236	k	0,0077
t	0,0906	r	0,0599	f	0,023	x	0,0015
a	0,0817	d	0,0425	g	0,0202	j	0,0015
o	0,0751	l	0,0403	y	0,0197	q	0,001
i	0,0697	c	0,0278	p	0,0193	z	0,0007
n	0,0675	u	0,0276	b	0,0149		
s	0,0633	m	0,0241	v	0,0098		

2. Изучение кодов Хаффмана.

2.1. Закодируйте согласно своему варианту задания словосочетания, приведенные в таблице 11, используя алгоритм Хаффмана, полученный в п. 4 (см. табл. 3).

2.2. Введите в закодированное сообщение, полученное в пункте 2.1, ошибку в любой из разрядов кода и выполните его декодирование. Сделайте вывод по полученным результатам.

2.3. Распишите механизм кодирования букв русского алфавита алгоритмом Хаффмана.

Контрольные вопросы

1. Как рассчитать количество информации, содержащееся в сообщении?
2. Что такое энтропия источника сообщений?
3. Свойства энтропии.
4. Избыточность различных источников информации.
5. Производительность источника сообщений.
6. Сущность энтропийного кодирования.
7. Алгоритм построения кода Шеннона – Фано.
8. Алгоритм построения кода Хаффмана.
9. Декодирование кода Шеннона – Фано и Хаффмана.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Гонсалес Р. Цифровая обработка изображений / Р. Гонсалес, Р. Вудс. – М. : Техносфера, 2005. – 1072 с.

2. Гонсалес Р. Цифровая обработка изображений в среде MATLAB / Р. Гонсалес, Р. Вудс, С. Эддинс. – М. : Техносфера, 2006. – 616 с.

3. Кошкин Г. М. Энтропия и информация / Г. М. Кошкин // Сорос. образоват. журн. 2001 – том 7, № 11. – С. 122-127.

4. Смирнов А. В. Цифровое телевидение : от теории к практике / А. В. Смирнов, А. Е. Пескин – 2-е изд. – М. : Горячая линия – Телеком, 2012. – 352 с.

5. Хэмминг Р. В. Теория кодирования и теория информации : пер. с англ./ Р. В. Хэмминг. – М. : Радио и связь, 1983. – 176 с.

6. Скляр Б. Цифровая связь. Теоретические основы и практическое применение : пер. с англ. / Б. Скляр. – Изд. 2-е, испр.: – М. : Изд. дом «Вильямс», 2003. – 1104 с.

7. Шульгин В. И. Основы теории передачи информации учеб. пособие: в 2 ч. Ч.1. Экономное кодирование / В. И. Шульгин. – Харьков : Нац. аэрокосм. ун-т «Харьковский авиационный институт», 2003. – 102 с.

СОДЕРЖАНИЕ