Министерство образования Республики Беларусь
Учреждение образования «Витебский государственный
университет имени П.М. Машерова»
Кафедра инженерной физики

ТЕОРИЯ ИНФОРМАЦИИ

Методические рекомендации

Витебск
ВГУ имени П.М. Машерова
2022

УДК 159.17(076.5)
ББК 22.184я73
Т33

Печатается по решению научно-методического совета учреждения образования «Витебский государственный университет имени П.М. Машерова». Протокол № 3 от 03.03.2022.

Составитель: старший преподаватель кафедры инженерной физики ВГУ имени П.М. Машерова **Т.И. Сапелко**

.

Р е ц е н з е н т :
заведующий кафедрой прикладного и системного программирования
ВГУ имени П.М. Машерова,
кандидат физико-математических наук, доцент *С.А. Ермоченко*

Теория информации : методические рекомендации / сост.
Т33 Т.И. Сапелко. – Витебск : ВГУ имени П.М. Машерова, 2022. – 48 с.

Методические рекомендации к выполнению лабораторных работ написаны в соответствии с учебной программой для специальностей 1-98 01 01-02 Компьютерная безопасность (радиофизические методы и программно-технические средства), 26 03 01 Управление информационными ресурсами. Содержат теорию, методику решения типовых задач, контрольные вопросы для управляемой самостоятельной работы и список литературы. Издание предназначено для освоения теоретического материала и отработки практических навыков при изучении дисциплины «Теория информации».

УДК 159.17(076.5)
ББК 22.184я73

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время теория информации является динамично развивающейся областью математики, что обусловлено как математической красотой результатов, так и множеством приложений в информатике, физике и генетике. Понятие информации, в частности квантовой и генетической, стало в последние десятилетия одним из важнейших при изучении как неживой так и живой природы.

Работой, заложившей основу математической теории информации, является статья К. Шеннона «Математическая теория связи», опубликованная в 1948 году. В этой статье впервые было введено понятие энтропии (количества информации) и были указаны основные задачи и направления развития теории. В настоящее время теория информации (в широком смысле) включает теорию распознавания статистических свойств источников сообщений, теорию кодирования источников сообщений (сжатия данных), теорию передачи сообщений по каналам связи, теорию помехоустойчивого кодирования, теорию поиска информации и криптографию.

Данные методические рекомендации составлены как сборник практических занятий, каждое из которых посвящено одной теме, что по замыслу автора должно облегчить самостоятельную работу над курсом. Дается краткое конспективное изложение основных вопросов. Вместе с тем, автор стремился к тому, чтобы в издании нашли отражение ключевые вопросы теории вероятности, математического описания сигналов, теории информации и кодирования.

Раздел 1
ОСНОВЫ ТЕОРИИ ИНФОРМАЦИИ

1.1 Элементы теории вероятностей в задачах теории информации

СЛУЧАЙНЫЕ СОБЫТИЯ. ВЕРОЯТНОСТЬ

Случайным событием называется всякий факт, который при осуществлении некоторых условий может произойти или не произойти.

События можно классифицировать следующим образом:

по возможности появления:

– достоверные (события, которые в результате опыта непременно должны произойти);

– невозможные (события, которые в результате опыта никогда не произойдут);

– равновозможные (если ни одно из этих событий не является объективно более возможным, чем другое);

– случайные (все другие события – возможные, но не достоверные);

по совместности появления:

– совместные (происходят одновременно);

– несовместные (происходят не одновременно);

по взаимозависимости:

– зависимые (события, при которых вероятность появления одного из них изменяет вероятность появления другого);

– независимые (события, при которых вероятность появления одного из них не изменяет вероятность появления другого);

по сложности:

– элементарные (события – возможные, исключающие друг друга в результате одного испытания);

– сложные (события, состоящие из других событий).

Полной группой называется совокупность единственно возможных событий испытания.

Противоположными называются два единственно возможных события, образующих полную группу.

Вероятностью события А называется число равное отношению числа исходов m, благоприятствующих появлению события, к числу всех равновозможных исходов n:

$$P(A) = \frac{m}{n} \ .$$

Свойства вероятности события:

1. $0 \leq P(A) \leq 1$.

2. Если A – событие невозможное, то $P(A) = 0$.

3. Если B – событие достоверное, то $P(B) = 1$.

В теории вероятностей часто приходится встречаться с элементами комбинаторики.

Комбинаторика изучает способы подсчета числа элементов в различных множествах.

Основными понятиями являются: перестановки, сочетания, размещения.

Перестановками называют комбинации, состоящие из одних и тех же n различных элементов и различающиеся между собой только порядком расположения элементов.

Число перестановок из n элементов находится по формуле

$$P_n = n!\,,$$

где $n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot (n-1) \cdot n$.

Размещениями из n элементов по k ($n \geq k$) называют комбинации, каждая из которых состоит из k элементов, взятых из n данных элементов, и отличающиеся между собой либо самими элементами, либо порядком их расположения.

Число размещений вычисляется по формуле

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

Сочетаниями из n элементов по k ($n \geq k$) называют комбинации, каждая из которых состоит из k элементов, взятых из n данных элементов и отличающиеся между собой хотя бы одним элементом.

Число сочетаний вычисляется по формуле

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

Существуют два основных правила применяемых при решении комбинаторных задач.

Правило суммы. Если некоторый объект A может быть выбран из некоторой совокупности объектов m способами, а объект B может быть выбран из некоторой совокупности объектов n способами, то выбор «либо A, либо B» можно сделать $n + m$ способами.

Правило произведения. Если объект A можно выбрать из некоторой совокупности m способами и после каждого такого выбора объект B можно выбрать n способами, то пара объектов A и B в указанном порядке может быть выбрана $m \cdot n$ способами.

Расчеты вероятности сложного события A через вероятности более простых событий A_1, A_2, … A_n базируются на использовании основных теорем теории вероятностей.

Теорема сложения вероятностей. Вероятность наступления одного из двух несовместных событий равна сумме их вероятностей, то есть:

$$P(A+B) = P(A) + P(B)\ .$$

Следствие 1. Если события A, B, C образуют полную группу, то сумма их вероятностей равна 1.

Следствие 2. Сумма вероятностей двух противоположных событий A и $\bar{A}$ равна 1.

Условной вероятностью события *B* называется вероятность наступления события *B* при условии, что событие *A* уже наступило. Обозначается: $P(B/A)$ или $P_A(B)$.

Теорема умножения вероятностей. Вероятность совместного наступления событий *A* и *B,* равна произведению вероятности одного из них на условную вероятность другого:

$$P(AB) = P(A)P(B / A) \quad \text{или} \quad P(AB) = P(B)P(A / B),$$

Следствие. Вероятность произведения двух независимых событий равна произведению их вероятностей:

$$P(AB) = P(A)P(B).$$

Теорема сложения вероятностей для случая, когда события совместны. Вероятность наступления хотя бы одного из двух совместных событий равна сумме вероятностей этих событий, минус вероятность их совместного появления, то есть

$$P(A+B) = P(A)+P(B) - P(AB).$$

Объединение теорем сложения и умножения выражается в ***формуле полной вероятности***

Теорема. *Вероятность события A, которое может произойти при осуществлении одного из несовместных событий* $B_1, B_2, B_3, \ldots, B_n$, *образующих полную группу, определяется формулой*:

$$P(A) = P(B_1)P(A/B_1) + P(B_2)P(A/B_2) + \cdots + P(B_n)P(A/B_n) = \\ = \sum_{i=1}^{n} P(B_i)P(A/B_i).$$

Замечание. События $B_1, B_2, B_3, \ldots, B_n$ называются ***гипотезами***.

При решении практических задач, когда событие *A,* появляющееся совместно с каким-либо из несовместных событий $B_1, B_2, B_3, \ldots, B_n$, образующих полную группу, произошло и требуется произвести количественную переоценку вероятностей событий $B_1, B_2, B_3, \ldots, B_n$ применяются ***формулы Бейеса*:**

$$P(B_i/A) = \frac{P(B_i)\cdot P(A/B_i)}{\sum_{i=1}^{n} P(B_i)P(A/B_i)}.$$

Пусть производится n последовательных независимых испытаний, в каждом из которых может появиться или не появиться событие *A*.

Вероятность появления события *A* в каждом испытании постоянна и равна p, а вероятность не появления обозначим через q: $P(\bar{A}) = 1 - p = q$. В случае небольшого числа испытаний вероятность того, что в n испытаниях это событие наступит ровно k раз рассчитывается по ***формуле Бернулли:***

$$P_n(k) = C_n^k \cdot p^k \cdot q^{n-k},$$

где C_n^k – число сочетаний из n по k.

СЛУЧАЙНЫЕ ВЕЛИЧИНЫ И ИХ ХАРАКТЕРИСТИКИ

Случайной величиной называется такая переменная величина, которая в результате опыта может принимать одно, заранее неизвестное значение ξ из известного множества значений *X*.

Случайная величина называется ***дискретной*** (прерывной), если множество ее возможных значений конечно или счётно.

Случайная величина называется ***непрерывной,*** если существует неотрицательная функция $W(x)$ такая, что $F(x) = \int_{-\infty}^{+\infty} W(t)dt$.

Например, число студентов на лекции – дискретная случайная величина, а продолжительность лекции – непрерывная случайная величина**.**

Случайные величины обозначаются заглавными буквами латинского алфавита X, Y, Z, а их возможные значения – соответствующими прописными буквами x_1, x_2,..., x_n.

Законом распределения случайной величины называется соответствие, устанавливающее связь между возможными значениями случайных величин и их вероятностями.

Закон распределения может быть задан:

-аналитически;

-таблично;

-графически.

Закон распределения дискретной случайной величины задаётся ***рядом распределения***, т.е. таблицей.

X	x_1	x_2	…	x_n
P	p_1	p_2	…	p_n

в которой x_1, x_2,..., x_n – расположенные по возрастанию значения дискретной случайной величины X, а p_1, p_2, ..., p_n – соответствующие этим значениям вероятности.

Функция распределения – это вероятность того, что случайная величина X примет значение, меньшее заданного значения x:

$$F(x) = P(X < x) .$$

Свойства функции распределения:

1. $0 \leq F(x) \leq 1$;

2. $F(x)$ – неубывающая функция, т.е. если $x_1 < x_2$, то $F(x_1) \leq F(x_2)$.

3. Вероятность попадания случайной величины в интервал (a, b) равна $P(a \leq X < b) = F(b) - F(a)$.

4. Если все возможные значения случайной величины X находятся на интервале (a, b), то $F(x) = 0$ при $x \leq a$ и $F(x) = 1$ при $x \geq b$;

5. $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$.

Функция распределения дискретной случайной величины представляет собой ступенчатую функцию со скачками в точках x_1, x_2,..., x_n

Непрерывная случайная величина задается в виде ***функции плотности вероятностей,*** которая является производной от функции распределения $W(x)=F'(x)$ в точках непрерывности.

Функция распределения непрерывной случайной величины представляет собой непрерывную функцию.

Плотность вероятностей обладает следующими свойствами:

1. $W(x) \geq 0$;

2. Вероятность попадания непрерывной случайной величины в интервал (x_1,x_2) равна интегралу от плотности вероятностей в этих пределах:

$$P(x_2 \leq X < x_1) = \int_{x_1}^{x_2} W(x)dx = F(x_2) - F(x_1);$$

3. $\int_{-\infty}^{+\infty} W(x)dx = 1$.

Во многих случаях определить закон распределения случайной величины невозможно. В таких ситуациях охарактеризовать случайную величину можно с помощью некоторых постоянных величин (числовых характеристик) этого закона.

Математическое ожидание – это число, которое выражает ***среднее значение*** случайной величины с учетом распределения.

Для дискретных величин она вычисляется по формуле

$$M(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n = \sum_{i=1}^{n} x_i \cdot p_i ,$$

где $x_1, x_2,..., x_n$ – возможные значения случайной величины,

$p_1, p_2, ..., p_n$. – их вероятности.

Для непрерывных случайных величин математическое ожидание – это число, которое определяется формулой

$$M(X) = \int_{-\infty}^{+\infty} xW(x)dx \ \text{(сходится)}$$

Свойства математического ожидания:

1. $M(C) = C \quad (C = const)$;

2. $M(CX) = CM(X) \ (C = const)$;

3. $(X \pm Y) = M(X) \pm M(Y)$;

4. $M(XY) = M(X)M(Y)$ (для независимых случайных величин).

Характеристиками рассеивания возможных значений случайной величины вокруг математического ожидания являются ***дисперсия и среднеквадратичное отклонение.***

Дисперсией случайной величины *X* называется величина равная математическому ожиданию квадрата отклонения

$$D(X) = M(X - M(X))^2 .$$

При практических вычислениях используют формулу:

$$D(X) = M(X^2) - (M(X))^2$$

Для непрерывных случайных величин дисперсия вычисляется формулами:

$$D(X) = \int_{-\infty}^{+\infty} (x - M(X))^2 W(x)dx ,$$

$$D(X) = \int_{-\infty}^{+\infty} (x)^2 W(x)dx - (M(X))^2 .$$

Дисперсия характеризует ***меру рассеивания*** возможных значений случайной величины около ее математического ожидания. Из двух величин с равными математическими ожиданиями та считается «лучшей», которая имеет меньший разброс.

Свойства дисперсии:

1. $D(C) = 0 \quad (C = const)$;
2. $D(CX) = C^2 D(X) \ \ (C = const)$;
3. $D(X \pm Y) = D(X) + D(Y)$ (для независимых случайных величин).

Арифметический корень из дисперсии случайной величины называется ***среднеквадратическим отклонением:***

$$\sigma_x = \sqrt{D(X)} \ .$$

Основными законами распределения непрерывных случайных величин являются: нормальный, показательный, равномерный.

Нормальный закон распределения случайной величины задается плотностью распределения по формуле

$$W(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} \quad , \ -\infty < x < \infty$$

Числа $a \in R$ и $\sigma > 0$ называются параметрами нормального закона. Нормальный закон с такими параметрами обозначается $N(a, \sigma)$. В общем случае $M(X) = a$, $D(X) = \sigma^2$.

Показательный закон распределения случайной величины задается плотностью распределения по формуле

$$W(x) = \begin{cases} 0, & x < 0 \\ \alpha \cdot e^{-\alpha x}, & x \geq 0 \end{cases}$$

Числовые характеристики $M(X) = \frac{1}{\alpha}$, $D(X) = \frac{1}{\alpha^2}$.

Равномерный закон распределения случайной величины задается плотностью распределения по формуле

$$W(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases}$$

Числовые характеристики $M(X) = \frac{a+b}{2}$, $D(X) = \frac{(b-a)^2}{12}$

Пример 1.

В последовательности из 6 двоичных символов имеется 3 единицы. При передаче данной последовательности сохраняется 3 символа, остальные теряются. Какова вероятность того, что среди сохранившихся будет не более 2 –х единиц?

Решение.

Пусть А – событие, состоящее в том, что среди двоичных символов будет не более 2-х единиц, т.е. 2 или 1, или ни одной. Тогда вероятность события А определяется как сумма:

$$P(A) = P(X \leq 2) = P(X = 2) + P(X = 1) + P(X = 0)$$

Вероятность каждого слагаемого можно рассчитать, используя гипергеометрическое распределение дискретной случайной величины

$$P(X = m) = \frac{C_M^m \cdot C_{N-M}^{n-m}}{C_N^n} \ .$$

Общее число возможных комбинаций выбора символов равно числу сочетаний 3 из 6, т.е. C_6^3.

Число благоприятных исходов для X=2 определяется как произведение $C_3^2 C_3^1$, где первый сомножитель это число комбинаций выбора 2-х «единиц» из общего числа «единиц» в последовательности. Но с каждой такой комбинацией могут встретиться символы, не являющиеся «единицами». Число таких комбинаций будет C_3^1. Поэтому искомая вероятность запишется в виде

$$P(X = 2) = \frac{C_3^2 C_3^1}{C_6^3} = 0{,}45$$

Аналогично для $P(X = 1) = \frac{C_3^1 C_3^2}{C_6^3} = 0{,}45$ и $P(X = 0) = \frac{C_3^0 C_3^3}{C_6^3} = 0{,}05$.

Таким образом, $P(A) = 0{,}95$.

Пример 2.

По каналу связи с помехами передается одна из двух команд управления в виде 11111 и 00000, вероятности передачи этих команд соответственно равны 0,7 и 0,3. Вероятность правильного приема каждого из символов 0 и 1 равна 0,6. Символы искажаются помехами независимо друг от друга. На выходе канала имеем кодовую комбинацию 10110. Определить какая комбинация была передана.

Решение.

Пусть событие А состоит в приеме комбинации 10110. Это событие может произойти в совокупности с событиями B_1 (передавалась комбинация 11111) и B_2 (передавалась комбинация 00000). При этом $P(B_1) = 0{,}7$, $P(B_2) = 0{,}3$.

Условная вероятность приема комбинации 10110 при условии, что передавалась команда 11111 равна

$$P(A/B_1) = P(1/1) \cdot P(0/1) \cdot P(1/1) \cdot P(1/1) \cdot P(0/1),$$

где $P(1/1) = 0{,}6$, $P(0/1) = 1 - P(1/1) = 0{,}4$.

$$P(A/B_1) = 0{,}6 \cdot 0{,}4 \cdot 0{,}6 \cdot 0{,}6 \cdot 0{,}4 = 0{,}035.$$

Условная вероятность приема комбинации 10110 при условии, что передавалась команда 00000 равна

$$P(A/B_2) = P(1/0) \cdot P(0/0) \cdot P(1/0) \cdot P(1/0) \cdot P(0/0),$$

где $P(0/0) = 0{,}6$, $P(1/0) = 1 - P(0/0) = 0{,}4$.

$$P(A/B_2) = 0{,}4 \cdot 0{,}6 \cdot 0{,}4 \cdot 0{,}4 \cdot 0{,}6 = 0{,}023.$$

По формуле полной вероятности $P(A) = P(B_1)P(A/B_1) + P(B_2)P(A/B_2) = 0{,}7 \cdot 0{,}035 + 0{,}3 \cdot 0{,}023 = 0{,}0314$.

По формуле Байеса $P(B_1/A) = \frac{P(B_1) \cdot P(A/B_1)}{P(A)} = \frac{0{,}7 \cdot 0{,}035}{0{,}0314} = 0{,}78$,

$$P(B_2/A) = \frac{P(B_2) \cdot P(A/B_2)}{P(A)} = \frac{0{,}3 \cdot 0{,}023}{0{,}0314} = 0{,}22.$$

Сравнивая найденные результаты, заключаем, что более вероятна передача команды 11111.

Пример 3.

По двоичному каналу связи с помехами передаются цифры 1 и 0 с вероятностями p_1=p_2=0,5. Вероятность перехода единицы в единицу и нуля в нуль соответственно равны $p(1/1)=p$, $p(0/0)=q$. Определить закон распределения вероятностей случайной величины X – однозначного числа, получаемого на приемной стороне.

Решение.

X=0 на приемной стороне можно получить при передаче нуля или единицы.

$P(B_1) = 0{,}5$ – вероятность передать ноль, $P(B_2) = 0{,}5$ – вероятность передать единицу.

Используя формулу полной вероятности, получим вероятность события А

$$P(A) = P(X = 0) = P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2) = P(0) \cdot P(0/0) +$$
$$+P(1) \cdot P(0/1) = 0{,}5 \cdot q + 0{,}5 \cdot (1 - p) = 0{,}5(q + 1 - p).$$

где $P(0/1) = 1 - P(1/1) = 1 - p$.

Аналогично X=1 на приемной стороне можно получить при передаче нуля или единицы.

Используя формулу полной вероятности, получим вероятность события С

$$P(C) = P(X = 1) = P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2) = P(1) \cdot P(1/1) +$$
$$+P(0) \cdot P(1/0) = 0{,}5 \cdot p + 0{,}5 \cdot (1 - q) = 0{,}5(p + 1 - q)\ ,$$

где $P(1/0) = 1 - P(0/0) = 1 - q$.

Распределение вероятностей удобно представить в виде таблицы

X	0	1
p_i	$0{,}5(q + 1 - p)$	$0{,}5(p + 1 - q)$

Проверка:

$$P(X = 0) + P(X = 1) = 0{,}5(q + 1 - p) + 0{,}5(p + 1 - q) = 1\ .$$

Пример 4.

Производится прием символов 0 и 1 до первого появления символа 1. Вероятность появления 1 при приеме p=0,4. Принимается не более четырех символов. Вычислить $M(X)$, $D(X)$, $\sigma(X)$ величины числа принятых символов.

Решение.

Вероятность появления 0 при приеме p=0,6.

Распределение вероятностей можно рассчитать следующим образом:

$P(X = 1) = p(1) = 0{,}4$ – вероятность получить 1 при первом приеме,

$P(X = 2) = p(0) \cdot p(1) = 0{,}6 \cdot 0{,}4 = 0{,}24$ – вероятность получить 1 при втором приеме,

$P(X = 3) = p(0) \cdot p(0) \cdot p(1) = 0{,}6 \cdot 0{,}6 \cdot 0{,}4 = 0{,}144$ – вероятность получить 1 при третьем приеме,

$P(X = 4) = p(0) \cdot p(0) \cdot p(0) \cdot p(1) + p(0) \cdot p(0) \cdot p(0) \cdot p(0) =$

$= 0{,}6 \cdot 0{,}6 \cdot 0{,}6 \cdot 0{,}4 + 0{,}6 \cdot 0{,}6 \cdot 0{,}6 \cdot 0{,}6 = 0{,}216$ – вероятность получить 1 при четвертом приеме или вероятность получить четыре раза 0.

$$M(X) = \sum_{i=1}^{n} x_i \cdot p_i = 1 \cdot 0{,}4 + 2 \cdot 0{,}24 + 3 \cdot 0{,}144 + 4 \cdot 0{,}216 = 2{,}176.$$

$$D(X) = M(X^2) - \left(M(X)\right)^2 = 1 \cdot 0{,}4 + 4 \cdot 0{,}24 + 9 \cdot 0{,}144 + 16 \cdot 0{,}216 - $$
$$-2{,}176^2 = 1{,}382\ .$$

$$\sigma(X) = \sqrt{D(X)} = 1{,}17\ .$$

Распределение вероятностей удобно представить в виде таблицы

X	1	2	3	4
p_i	0,4	0,24	0,144	0,216

Проверка: $0{,}4 + 0{,}24 + 0{,}144 + 0{,}216 = 1$

Пример 5.

Функция распределения $F(X)$ случайной величины X задана графиком (рисунок 1.1). Найти:

1) аналитическое выражение для функции распределения,

2) построить график плотности вероятностей W(x),

3) определить вероятность попадания случайной величины X в интервал (3,5;4,5).

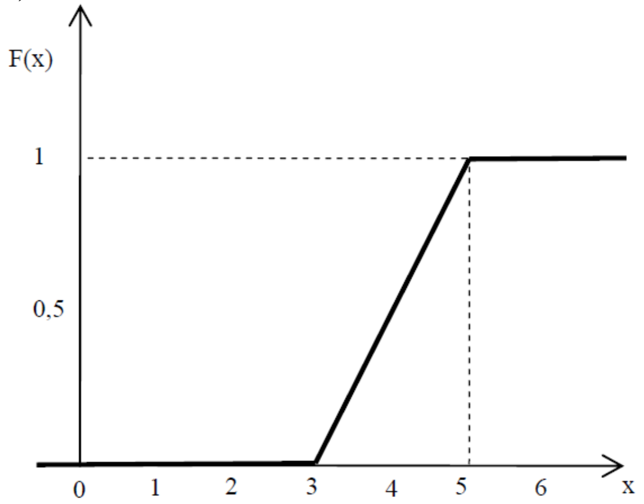


Рисунок 1.1 – Функция распределения $F(X)$ случайной величины X

Решение.

1) Из графика видно, что при $X \in [3;\ 5]$ функция распределения $F(X)$ представляет собой отрезок прямой, проходящей через две точки с координатами (3,0) и (5,1). Используя уравнение прямой, проходящей через две точки $\frac{x-x_1}{x_2-x_1} = \frac{y-y_1}{y_2-y_1}$, получим $\frac{x-3}{5-3} = \frac{y-0}{1-0}$. Откуда $F(X) = \frac{x-3}{2}$.

Следовательно, $F(X) = \begin{cases} 0 & x \le 3, \\ (x-3)/2 \text{ при } 3 < x \le 5, \\ 1 & x > 5. \end{cases}$

2) $W(X) = F'(X)$, поэтому

$$W(X) = \begin{cases} 0, & x \le 3, \\ 0{,}5, & 3 < x \le 5, \\ 0\,, & x > 5. \end{cases}$$

$P(3{,}5 < X < 4{,}5) = F(4{,}5) - F(3{,}5) = \frac{4{,}5-3}{2} - \frac{3{,}5-3}{2} = 0{,}5.$

Или геометрически найденная вероятность – это площадь заштрихованной фигуры (рисунок 1.2).

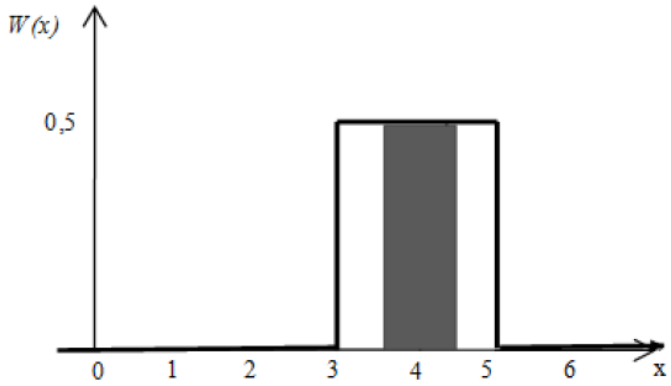


Рисунок 1.2 – График плотности вероятностей W(x)

Контрольные вопросы

1. Что называется событием?

2. Какие события называются противоположными, достоверными, невозможными?

3. Какие события составляют полную группу?

4. Какие события называются элементарными?

5. Сформулируйте классическое определение вероятности.

6. Что называется перестановками? Как они определяются?

7. Что называется сочетаниями? Как вычисляются сочетания?

8. Что называется размещениями? Как вычисляются размещения?

9. Что называется условной вероятностью?

10. Сформулируйте теорему умножения вероятностей.

11. Сформулируйте теорему сложения вероятностей.

12. Сформулируйте формулу Бернулли, когда он а используется?

13. Как формулируется теорема о полной вероятности?

14. Как формулируется теорема Байеса?

15. Что называется случайной величиной?

16. Какие случайные величины называются дискретными, непрерывными?

17. Что называется законом распределения случайной величины?

18. Что называется функцией распределения случайных величин.

19. Перечислите свойства функции распределения.

20. Как определяется функция плотности вероятностей непрерывных случайных величин?

21. Перечислите свойства функции плотности распределения вероятностей.

22. Как вычисляется математическое ожидание для дискретных и непрерывных случайных величин?

23. Перечислите свойства математического ожидания.

24. Как вычисляется дисперсия и среднеквадратическое отклонение для дискретных и непрерывных случайных величин?

25. Перечислите свойства дисперсии.

26. Основные законы распределения.

1.2 Информационная мера Шеннона

КОЛИЧЕСТВО ИНФОРМАЦИИ И ИЗБЫТОЧНОСТЬ

Дискретные системы связи – системы, в которых как реализации сообщения, так и реализации сигнала представляют собой последовательности символов алфавита, содержащего конечное число элементарных символов.

Пусть ξ и η – случайные величины с множествами возможных значений $X = \{x_1, x_2, \cdots, x_n\}$, $Y = \{y_1, y_2, \cdots, y_n\}$.

Количество информации $H(\xi)$ при наблюдении случайной величины $\xi \in X = \{x_1, x_2, \cdots, x_n\}$ с распределением вероятностей $P = \{p_1, p_2, \cdots, p_n\}$ задается формулой Шеннона:

$$H(\xi) = \sum_{i=1}^{n} p_i \log 1/p_i.$$

Единицей измерения количества информации является бит, который представляет собой количество информации, получаемое при наблюдении случайной величины, имеющей два равновероятных значения.

При равномерном распределении $p_1 = p_2 = \cdots = p_n$ количество информации задается формулой Хартли:

$$H(\xi) = \log_2 N.$$

Справедливы следующие соотношения:

1) $0 \le H(\xi) \le \log_2 N$;

2) $N = 2,\ p_1 = p_2 = 0{,}5,\ H(\xi) = 1$;

3) $H(\xi, \eta) = H(\xi) + H(\eta)$, если ξ и η – независимы.

Избыточностью называется $p = 1 - H(\xi)/maxH(\xi) = 1 - -H(\xi)/\log_2 N$.

ЭНТРОПИЯ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

Непрерывные системы передачи информации – системы, в которых как реализации сообщения, так и реализации сигнала на конечном временном интервале $(0, T)$ представляют собой некоторые непрерывные функции времени.

Пусть *x(t)* – реализации непрерывного сообщения на входе какого-либо блока схемы связи, *y(t)* – реализация выходного сообщения (сигнала), *W(x)* – плотность вероятностей ансамбля входных сообщений, *W(y)* – плотность вероятностей ансамбля выходных сообщений.

Формулы для энтропии H непрерывных сообщений получаются путем обобщения формул для энтропии дискретных сообщений. Если Δx – интервал квантования (точность измерения), то при достаточно малом Δx ***энтропия непрерывных сообщений.***

$$H(X) = -\int_{-\infty}^{+\infty} W(x)\log W(x)\,dx - \log\Delta x\int_{-\infty}^{+\infty} W(x)dx = H^*(X) - \log\Delta x =$$
$$= \int_{-\infty}^{+\infty} W(x)\log[W(x)\Delta x]\,dx,$$

где $H^*(X) = -\int_{-\infty}^{+\infty} W(x)\log W(x)\,dx$ – приведенная энтропия.

$H(x)$ можно представить в компактной записи:

$H(x) = M[-\log W(x)\,\Delta x]$ и $H^*(X) = M[-\log W(x)]$.

По аналогии $H(Y) = -\int_{-\infty}^{+\infty} W(y)\log W(y)\,dy$.

Пример 1.

Источник сообщений выдает символы из алфавита $A = \{a_i\}$, $i = \overline{1{,}4}$ с вероятностями $p_1 = 0{,}2$; $p_2 = 0{,}3$; $p_3 = 0{,}4$; $p_4 = 0{,}1$. Найти количество информации и избыточность.

Решение.

По формуле Шеннона $H(\xi) = \sum_{i=1}^{n} p_i \log 1/p_i$.

$$H(A) = -(0{,}2\cdot\log 0{,}2 + 0{,}3\cdot\log 0{,}3 + 0{,}4\cdot\log 0{,}4 + 0{,}1 *$$
$$*\log 0{,}1) = 1{,}86 \text{ бит.}$$

Избыточность $p = 1 - \frac{H(A)}{\log_2 4} = 1 - \frac{1{,}86}{2\log_2 2} = 0{,}07$.

Пример 2.

Имеются два источника информации, алфавиты и распределения вероятностей которых заданы матрицами:

$$\left\|\begin{matrix} X \\ P \end{matrix}\right\| = \left\|\begin{matrix} x_1 & x_2 \\ p_1 & p_2 \end{matrix}\right\|, \quad \left\|\begin{matrix} Y \\ Q \end{matrix}\right\| = \left\|\begin{matrix} y_1 & y_2 & y_3 \\ q_1 & q_2 & q_3 \end{matrix}\right\|.$$

Определить, какой источник дает большее количество информации, если

1) $p_1 = p_2$; $q_1 = q_2 = q_3$;

2) $p_1 = q_1$; $p_2 = q_2 + q_3$;

Решение.

1) Для первого источника при равновероятном распределении воспользуемся формулой Хартли. Для $\xi \in X$ и $\eta \in Y$ имеем

$$\left.\begin{matrix} H(\xi) = \log_2 2 = 1 \\ H(\eta) = \log_2 3 > 1 \end{matrix}\right\} H(\eta) > H(\xi).$$

Следовательно, источник с тремя символами дает большее количество информации.

2) Воспользуемся формулой Шеннона:

$$H(\xi) = \sum_{i=1}^{n} p_i \log 1/p_i \,.$$

с учетом условия задачи имеем $p_1 = q_1;\quad p_2 = q_2 + q_3;$

$$H(\xi) = q_1 \log_2 1/q_1 + q_2 \log_2 1/(q_2 + q_3) + q_3 \log_2 1/(q_2 + q_3).$$

С другой стороны,

$$H(\eta) = q_1 \log_2 1/q_1 + q_2 \log_2 1/q_2 + q_3 \log_2 1/q_3.$$

Сравним слагаемые $H(\xi)$ и $H(\eta)$.

Поскольку

$$\frac{1}{(q_2+q_3)} < \frac{1}{q_2} \;,\; \frac{1}{(q_2+q_3)} < \frac{1}{q_3} \;, \text{ то } H(\xi) < H(\eta).$$

Пример 3.

Определить количество информации и энтропию сообщения из 5 букв, если число букв в алфавите равно 32 и все сообщения равновероятны.

Решение.

Общее число пятибуквенных сообщений равно $N = m^n$, $N = 32^5$.

Энтропия для равновероятных сообщений по формуле Хартли

$$H(\xi) = \log_2 N \,.$$

$$H = \log_2 32^5 = 5 \log_2 2^5 = 25 \text{ бит}$$

Пример 4.

По линии связи передаются непрерывные амплитудно-модулированные сигналы *x(t)*, распределенные по нормальному закону с математическим ожиданием $m_x = 0$ и дисперсией $\sigma_x^2 = \sigma^2 = 8B^2$.

Определить энтропию $H(X)$ сигнала при точности его измерения $\Delta x = 0{,}2B$.

Решение.

По условию плотность вероятностей сигнала *x(t)*

$$W(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x)^2}{2\sigma^2}} \;,$$

$$H(X) = -\int_{-\infty}^{+\infty} W(x) \log W(x)\, dx - \log \Delta x =$$

$$= -\frac{1}{\ln 2}\int_{-\infty}^{+\infty} W(x) \ln W(x)\, dx - \log \Delta x =$$

$$= -\frac{1}{\ln 2}\int_{-\infty}^{+\infty} W(x) \left[\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x)^2}{2\sigma^2}\right] dx - \log \Delta x =$$

$$= -\frac{1}{\ln 2}\left(\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2}\right) - \log \Delta x = \frac{1}{\ln 2} \ln \sqrt{2\pi e \sigma^2} - \log \Delta x = \log \frac{\sqrt{2\pi e \sigma^2}}{\Delta x} \;.$$

Подставляя числовые значения, получаем

$$H(X) = \log \frac{\sqrt{2\pi e 8}}{0{,}2} \cong 5{,}87 \text{ дв.ед.}$$

Пример 5.

Определить полную энтропию системы X, состояние которой имеет экспоненциальное распределение.

Решение.

Полная энтропия системы X

$$H(X) = H^*(X) - \log_2 \Delta x \,,$$

$$H^*(X) = M[-\log_2 W(x)] = M[-\log_2 \alpha \cdot e^{-\alpha x}] =$$
$$= M[-\log_2 \alpha - \log_2 e^{-\alpha x}] = -\log_2 \alpha + M[\alpha x \cdot \log_2 e] =$$
$$= -\log_2 \alpha + \alpha \cdot \log_2 e \cdot M[x] = -\log_2 \alpha + \alpha \cdot \log_2 e \cdot \frac{1}{\alpha} = \log_2 \frac{e}{\alpha} \,,$$
$$H(X) = H^*(X) - \log_2 \Delta x = \log_2 \frac{e}{\alpha} - \log_2 \Delta x = \log_2 \frac{e}{\alpha \Delta x} \,.$$

Контрольные вопросы

1.Дать определение энтропии.
2.Запишите формулу Шеннона.
3.Запишите формулу Хартли.
4.Перечислите основные свойства энтропии.
5. Что является единицей измерения энтропии?
6. В каких случаях энтропия равна нулю?
7. При каких условиях энтропия принимает максимальное значение?
8. В чем состоит правило сложения энтропий для независимых источников?
9. Как определяется количество информации непрерывных сообщений?
10. Запишите формулу избыточности.

1.3 Условная энтропия и взаимная информация

ДИСКРЕТНЫЕ СИСТЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Условной энтропией величины Y при наблюдении величины X называется $H(Y/X) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i y_j) \log_2 \frac{1}{p(y_j/x_i)}$.

Справедливы соотношения:

$$H(X,Y) = H(X) + H(Y)\,, \quad H(X,Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i\,, y_j) \log_2 \frac{1}{p(x_i\,, y_j)}\,,$$

Взаимная информация величин X и Y определяется из рисунка 1.3

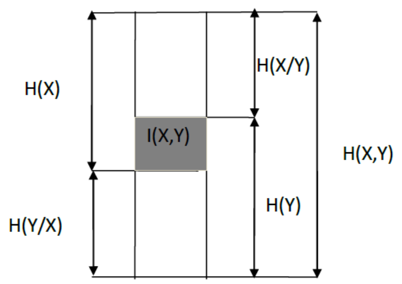


Рисунок 1.3 – Взаимная информация источников сообщений

Это штриховая часть $I(X,Y)$, показывающая какое (в среднем) количество информации содержит сообщение X о сообщении Y или наоборот сообщение Y о сообщении X.

$$I(X,Y) = H(X) - H(X/Y)\ ,$$
$$I(X,Y) = I(Y,X) = H(Y) - H(Y/X),$$
$$I(X,Y) = H(X) - H(X,Y).$$

Если X и Y независимы, то $I(X,Y) = 0$,

Если X и Y полностью зависимы (содержат одну и ту же информацию), то $I(X,Y) = H(X) = H(Y)$.

Справедливы следующие соотношения

$$Y(X,Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i\,, y_j) \log_2 \frac{p(x_i\,,y_j)}{p(x_i)\cdot p(y_j)},$$
$$0 \le I(X,Y) \le H(X), \qquad 0 \le I(X,Y) \le H(Y).$$

Понятие взаимной информации широко используется в теории передачи информации. Требования к взаимной информации различны в зависимости от того, с какой информацией работает потребитель. Например, если X и Y – это сообщения, публикуемые различными газетами, то для получения возможно большей суммарной (совместной) информации, взаимная (то есть одинаковая в данном случае) информация должна быть минимальной. Если же X и Y – это сообщения на входе и на выходе канала связи с помехами, то для получения возможно большей информации ее получателем необходимо, чтобы взаимная информация была наибольшей. Тогда условная энтропия $H(X/Y)$ – это потери информации в канале связи (ненадежность канала). Условная энтропия $H(Y/X)$ – это информация о помехах (энтропия источника помех $H(n)$), поступающая в канал извне или создаваемая внутренними помехами в канале (рисунок 1.4).

При расчетах условной энтропии и взаимной информации удобно пользоваться следующими соотношениями теории вероятностей:

1) теорема умножения вероятностей

$$p(x_i\, y_j) = p(x_i)p(y_j/x_i) = p(y_j)p(x_i/y_j)\ ;$$

2) формула полной вероятности $p(x_i) = \sum_{j=1}^{M} p(x_i\,, y_j)$; $p(y_j) = \sum_{i=1}^{n} p(x_i\,, y_j)$;

3) формула Байеса $p(x_i/y_j) = \frac{p(x_i)\cdot p(y_j/x_i)}{p(y_j)} = \frac{p(x_i)\cdot p(y_j/x_i)}{\sum_{i=1}^{N} p(x_i)\cdot p(y_j/x_i)}$

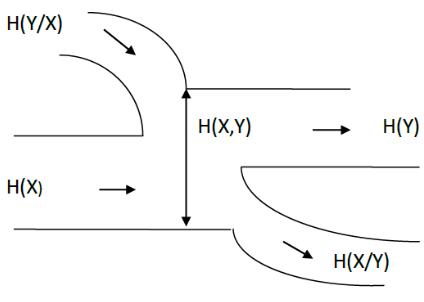


Рисунок 1.4 – Энтропия объединения

НЕПРЕРЫВНЫЕ СИСТЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Пусть *x(t)* – реализации непрерывного сообщения на входе какого-либо блока схемы связи, *y(t)* – реализация выходного сообщения (сигнала), $f_1(x)$ – одномерная плотность вероятностей ансамбля входных сообщений, $f_2(x)$ – одномерная плотность вероятностей ансамбля выходных сообщений, $f(x,y)$ – совместная плотность вероятностей, $f(y/x)$ – условная плотность вероятностей при известном Тогда для количества информации *I* справедливы следующие соотношения:

$$I(x,y) = \log_2 \frac{f(x,y)}{f_1(x)f_2(y)} = \log_2 \frac{f(x/y)}{f_1(x)} = \log_2 \frac{f(y/x)}{f_2(y)} = I(x,y),$$

$$I(X,y) = \int_X I(x,y)f(x/y)dx\,, \qquad I(x,Y) = \int_Y I(x,y)f(y/x)dy,$$

Выражение для ***полной взаимной информации***, содержащейся в двух непрерывных системах *X* и *Y* будет определяться

$$I(X,Y) = \int_X f_2(y)I(X,y)dy = \iint_{XY} f(x,y)I(x,y)dxdy =$$

$$= \iint_{XY} f(x,y)\log_2 \frac{f(x,y)}{f_1(x)f_2(y)}dxdy = I(Y,X),$$

или, применяя знак математического ожидания

$$I_{Y\leftrightarrow X} = M\left[\log_2 \frac{f(X,Y)}{f_1(X)f_2(Y)}\right]$$

$$I(X,y) \geq 0\,, \quad I(x,Y) \geq 0, \quad I(X,Y) \geq 0.$$

Здесь $I(x,y)$ – взаимная информация между каким-либо значением x входного и значением y выходного сообщений, $I(X,y)$ и $I(x,Y)$ – средние значения условной информации, $I(X,Y)$ – полная средняя взаимная информация.

Условная энтропия определяется по формуле:

$$H(X/Y) = -\iint_{-\infty}^{\infty} f(x,y)\log f(x/y)\,dxdy,$$

$$H(Y/X) = -\iint_{-\infty}^{\infty} f(x,y)\log f(y/x)\,dxdy,$$

Можно представить в компактной записи

$$H(X/Y) = M[-\log f(Y/X)] - \log \Delta x.$$

Когда *X* и *Y* статистически связаны между собой, то

$$H(X,Y) = H(X) + H(Y)\,.$$

Полная средняя взаимная информация определяется формулой:

$$I(X,Y) = H(X) - H(X/Y) = H(Y) - H(Y/X).$$

Пример 1.

Дана матрица

$$P(X,Y) = \left\|\begin{matrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & 0 & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{matrix}\right\|, \qquad x \in X\,, y \in Y\,.$$

Решение.

По формуле полной вероятности имеем:

$$P(x_1) = \sum_{j=1}^{3} p(x_1\,, y_i) = \frac{3}{8}\,, \quad P(x_2) = \frac{1}{4}\,, \quad P(x_3) = \frac{3}{8},$$

$$P(y_1) = \sum_{i=1}^{3} p(x_i\,, y_1) = \frac{3}{8}\,, \quad P(y_2) = \frac{1}{4}\,, \quad P(y_3) = \frac{3}{8}.$$

Следовательно,

$$H(X) = \sum_{i=1}^{3} p(x_i) \log_2 \frac{1}{p(x_i)} = 1{,}57;$$

$$H(Y) = \sum_{i=1}^{3} p(y_i) \log_2 \frac{1}{p(y_i)} = 1{,}57.$$

По теореме умножения $p(x_i/y_i) = \frac{p(x_i\,,y_i)}{p(y_i)}$.

$$p(x_1/y_1) = \frac{1}{3}\,,\ \ p(x_1/y_2) = \frac{1}{2}\,,\ \ p(x_1/y_3) = \frac{1}{3},$$

$$p(x_2/y_1) = \frac{1}{3}\,,\ \ p(x_2/y_2) = 0\,,\ \ p(x_2/y_3) = \frac{1}{3},$$

$$p(x_3/y_1) = \frac{1}{3}\,,\ \ p(x_3/y_2) = \frac{1}{2}\,,\ \ p(x_3/y_3) = \frac{1}{2}.$$

Следовательно,

$$H(X/Y) = \sum_{l=1}^{3} \sum_{j=1}^{3} p(x_i\ \ y_i) \log_2 \frac{1}{p(x_i\,/\,y_j)} = 1{,}43.$$

Аналогично $H(Y/X) = 1{,}43$;

Энтропия объединения:

$$H(X,Y) = H(X) + H(Y/X) = 1{,}57 + 1{,}43 = 3;$$

Взаимная информация величин X и Y:

$$I(X,Y) = H(X) - H(X/Y) = 0{,}14.$$

Пример 2.

Канал связи описан следующей канальной матрицей

$$P(Y/X) = \begin{pmatrix} 0{,}98 & 0{,}1 & 0{,}2 \\ 0{,}01 & 0{,}75 & 0{,}3 \\ 0{,}01 & 0{,}15 & 0{,}5 \end{pmatrix}.$$

Найти:

1) Среднее количество информации, которое переносится одним символом сообщения, если вероятности появления символов источника сообщений равны $p(x_1) = 0{,}7$, $p(x_2) = 0{,}2$, $p(x_3) = 0{,}1$.

2) Чему равны информационные потери при передаче сообщения из 1000 символов алфавита $x_1\,, x_2\,, x_3$?

3) Чему равно количество принятой информации?

Решение.

1) Энтропия источника сообщений

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i = -(0{,}7 \cdot \log 0{,}7 + 0{,}2 \cdot \log 0{,}2 + 0{,}1 \cdot \log 0{,}1) =$$
$$= 1{,}16 \text{ бит}$$

2) Общая условная энтропия

$$H(Y/X) = \sum_{I=1}^{N}\sum_{J=1}^{M} p(x_i\,, y_j)\log_2\frac{1}{p(\,y_j\,/x_i\,)} =$$

$$= \sum_{I=1}^{N}\sum_{J=1}^{M} p(x_i)p(\,y_j\,/x_i\,)\log_2\big(y_j\,/x_i\big)$$

$$H(Y/X) = -0{,}7[(0{,}98\cdot\log_2 0{,}98 + 0{,}01\cdot\log_2 0{,}01 + 0{,}01\cdot\log_2 0{,}01) +$$
$$+0{,}2(0{,}1\cdot\log_2 0{,}1 + 0{,}75\cdot\log_2 0{,}75 + 0{,}15\cdot\log_2 0{,}15) +$$
$$+0{,}1(0{,}2\cdot\log_2 0{,}2 + 0{,}3\cdot\log_2 0{,}3 + 0{,}5\cdot\log_2 0{,}5)] = 0{,}473 \text{ бит.}$$

Потери в канале связи $\Delta I = kH(Y/X) = 1000\cdot 0{,}473 = 473$ бит.

3) Энтропия приемника $H(Y) = -\sum_{j=1}^{M} p(y_i)\log_2 p(y_i)$.

Используя формулу умножения $p(x,y) = p(x)\cdot p(y/x)$, осуществим переход к матрице

$$P(X,Y) = \begin{pmatrix} 0{,}7\cdot 0{,}98 = 0{,}686 & 0{,}2\cdot 0{,}1 = 0{,}02 & 0{,}1\cdot 0{,}2 = 0{,}02 \\ 0{,}7\cdot 0{,}01 = 0{,}007 & 0{,}2\cdot 0{,}75 = 0{,}15 & 0{,}1\cdot 0{,}3 = 0{,}03 \\ 0{,}7\cdot 0{,}01 = 0{,}007 & 0{,}2\cdot 0{,}15 = 0{,}03 & 0{,}1\cdot 0{,}5 = 0{,}05 \end{pmatrix}$$

$$P(X,Y) = \begin{pmatrix} 0{,}686 & 0{,}02 & 0{,}02 \\ 0{,}007 & 0{,}15 & 0{,}03 \\ 0{,}007 & 0{,}03 & 0{,}05 \end{pmatrix}$$

Суммируя элементы строк, получим безусловные вероятности $p(y_i)$.

Суммируя элементы столбцов, получим безусловные вероятности $p(x_i)$.

$$p(y_1) = 0{,}726\,,\quad p(y_2) = 0{,}187\,,\quad p(y_3) = 0{,}087.$$

$$H(Y) = -\sum_{j=1}^{M} p\big(y_j\big)\log_2 p\big(y_j\big) = 1{,}094 \text{ бит.}$$

Среднее количество полученной информации

$$I = k(H(Y) - H(Y/X)) = 1000\cdot(1{,}094 - 0{,}473) = 621 \text{ бит.}$$

Пример 3.

Найти энтропию шума $H(Y/X)$ в двоично-симметричном канале без памяти, если энтропия источника на входе канала $H(X) = 3400$ бит, энтропия ансамбля на выходе канала $H(Y) = 6800$ бит, ненадежность канала $H(X/Y) = 700$ бит.

Решение.

Взаимная информация источника сообщений

$$I(X,Y) = H(X) - H(X/Y) = H(Y) - H(Y/X)\ .$$

$$I(X,Y) = H(X) - H(X/Y) = 3400 - 700 = 2700 \text{ бит.}$$

Энтропия шума:

$$H(Y/X) = H(Y) - I(X,Y) = 6800 - 2700 = 4100 \text{ бит.}$$

Пример 4.

Принимаемый сигнал может иметь амплитуду A_1 (событие x_1) или A_2 (событие x_2), а также сдвиг фаз φ_1 (событие y_1) или φ_2 (событие y_2). Вероятности совместных событий имеют следующие значения: $p(x_1, y_1) = 0{,}73$, $p(x_1, y_2) = 0{,}21$, $p(x_2, y_1) = 0{,}02$, $p(x_2, y_2) = 0{,}04$. Вычислить количество информации, получаемой о фазовом сдвиге сигнала, если станет известной его амплитуда.

Решение.

Среднее количество информации о фазовом сдвиге при известной амплитуде $I(X,Y) = H(Y) - H(Y/X)$,

$$p(y_1) = p(x_1, y_1) + p(x_2, y_1) = 0{,}73 + 0{,}02 = 0{,}75.$$

$$p(y_2) = p(x_1, y_2) + p(x_2, y_2) = 0{,}21 + 0{,}04 = 0{,}25,$$

$$H(Y) = -\sum_{j=1}^{M} p(y_j)\log_2 p(y_j) = -(0{,}75\log_2 0{,}75 + 0{,}25\log_2 0{,}25) =$$

$$= 0{,}81 \text{ бит}.$$

$$p(x_1) = p(x_1, y_1) + p(x_2, y_2) = 0{,}73 + 0{,}21 = 0{,}94,$$

$$p(x_2) = p(x_2, y_1) + p(x_2, y_2) = 0{,}02 + 0{,}04 = 0{,}06,$$

По теореме умножения $p(y_j/x_i) = \frac{p(x_i, y_j)}{p(x_i)}$.

$$p(y_1/x_1) = \frac{0{,}73}{0{,}94} = 0{,}78\ ,\ p(y_2/x_1) = \frac{0{,}2}{0{,}94} = 0{,}22,$$

$$p(y_1/x_2) = \frac{0{,}02}{0{,}06} = 0{,}33\ ,\ p(y_2/x_2) = \frac{0{,}04}{0{,}06} = 0{,}67,$$

$$H(Y/X) = \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j)\log_2\frac{1}{p(y_j/x_i)} = 0{,}73\log_2\frac{1}{0{,}78} + 0{,}21\log_2\frac{1}{0{,}22} +$$

$$+0{,}02\log_2\frac{1}{0{,}33} + 0{,}04\log_2\frac{1}{0{,}67} = 0{,}77 \text{ бит}.$$

$$I(Y,X) = H(X) - H(Y/X) = 0{,}81 - 0{,}77 = 0{,}04 \text{ бит}.$$

Пример 5.

На вход приемного устройства воздействует колебание $y(t) = x(t) + n(t)$, где сигнал $x(t)$ и помеха $n(t)$ – независимые гауссовские случайные процессы с нулевыми математическими ожиданиями и дисперсиями, равными соответственно σ_x^2 и σ_n^2.

Определить:

1) количество взаимной информации $I(x,y)$, которое содержится в каком-либо значении принятого колебания $y(t)$ о значении сигнала $x(t)$;

2) полную среднюю взаимную информацию $I(X,Y)$.

Решение.

По условию задачи $y(t)$ представляет собой сумму независимых колебаний $x(t)$ и $n(t)$, которые имеют нормальные плотности вероятностей

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}}exp\left(-\frac{x^2}{2\sigma_x^2}\right),\ f(y) = \frac{1}{\sigma_y\sqrt{2\pi}}exp\left(-\frac{y^2}{2\sigma_y^2}\right),$$

$$\sigma_y^2 = \sigma_x^2 + \sigma_n^2.$$

$$f(y/x) = f(y-x) = \frac{1}{\sigma_n\sqrt{2\pi}}exp\left(-\frac{(y-x)^2}{2\sigma_n^2}\right) = \frac{1}{\sigma_n\sqrt{2\pi}}exp\left(-\frac{n^2}{2\sigma_n^2}\right) = f(n).$$

1. Количество информации определяется по формуле

$$I(x,y) = \log_2 \frac{f(y/x)}{f(y)} = \frac{1}{ln2} ln \frac{f(y/x)}{f(y)} = \frac{1}{ln2} ln \left[\frac{1}{\sigma_n \sqrt{2\pi}} exp\left(-\frac{n^2}{2\sigma_n^2}\right) - \right.$$

$$\left. - \frac{1}{\sigma_y \sqrt{2\pi}} exp\left(-\frac{y^2}{2\sigma_y^2}\right) \right] = \frac{1}{ln2} ln \frac{\sqrt{\sigma_x^2+\sigma_n^2}}{\sigma_n} + \frac{1}{ln2}\left[-\frac{n^2}{2\sigma_n^2} + \frac{y^2}{2(\sigma_x^2+\sigma_n^2)}\right].$$

1. Полная средняя взаимная информация

$$I(X,Y) = \int_X \int_Y f(x,y) I(x,y) dxdy = \int_X \int_Y f(y/x) f(x) I(x,y) dxdy .$$

Или $I(X,Y) = M[I(x,y)] = M\left[\frac{1}{ln2} ln\sqrt{1+\frac{\sigma_x^2}{\sigma_n^2}} + \frac{1}{ln2}\left[\frac{y^2}{2(\sigma_x^2+\sigma_n^2)} - \frac{n^2}{2\sigma_n^2}\right]\right] =$

$$= M\left[\log_2 \frac{\sqrt{\sigma_x^2+\sigma_n^2}}{\sigma_n}\right] + \frac{1}{ln2} M\left[\frac{y^2}{2(\sigma_x^2+\sigma_n^2)} - \frac{n^2}{2\sigma_n^2}\right] = \log_2 \frac{\sqrt{\sigma_x^2+\sigma_n^2}}{\sigma_n} +$$

$$+ \frac{1}{2ln2(\sigma_x^2+\sigma_y^2)} M[y^2] - \frac{1}{2ln2\sigma_n^2} M[n^2] = \log_2 \frac{\sqrt{\sigma_x^2+\sigma_n^2}}{\sigma_n} + \frac{\sigma_y^2}{2ln2(\sigma_x^2+\sigma_n^2)} - \frac{\sigma_n^2}{2ln2(\sigma_n^2)} =$$

$$= \log_2 \frac{\sqrt{\sigma_x^2+\sigma_n^2}}{\sigma_n} + \frac{1}{ln2}\left[\frac{\sigma_x^2+\sigma_n^2}{2(\sigma_x^2+\sigma_n^2)} - \frac{\sigma_n^2}{2(\sigma_n^2)}\right] = \frac{1}{2}\log_2\left(1+\frac{\sigma_x^2}{\sigma_n^2}\right) \text{ дв.ед.}$$

Контрольные вопросы

1. Дать определение условной энтропии.
2. Сформулировать закон аддитивности энтропии в общем случае.
3. Какие формулы используются для расчета условной энтропии?
4. Какие формулы используются для расчета взаимной информации?
5. Как определяется полная средняя взаимная информация?
6. Что понимают под дискретными системами передачи информации?
7. Что понимают под непрерывными системами передачи информации?
8. Как определяется условная энтропия в непрерывной системе передачи информации?

1.4 Передача информации по каналу связи

Понятие энтропии, скорости выдачи информации источником, избыточность позволяют характеризовать свойства информационных систем. Однако для сравнения информационных систем такого описания недостаточно. Потребителя интересует не только передача данного количества информации, но и передача его в более короткий срок, не только хранение определенного количества, но и хранение с помощью минимального объема аппаратуры и т.д.

Пусть количество информации, которое передается по каналу связи за время T равно

$$I_T = H_T - H_T(Y/X) .$$

Если передача длится T единиц времени, то ***скорость передачи информации*** составит

$$V = \frac{I_T}{T} = \frac{1}{T}(H_T - H_T(Y/X)) = H(X) - H(X/Y).$$

Это количество информации, приходящееся в среднем на одно сообщение.

Если в секунду передается n сообщений, то скорость передачи равна

$$V = n(H(X) - H(X/Y)).$$

Пропускная способность канала есть максимально достижимая для данного канала скорость передачи информации

$$c = maxV = n(H(X) - H(X/Y))_{max} = n \cdot I(X,Y)_{max}$$

Скорость передачи может быть технической или информационной.

Под ***технической*** V_T (скорость манипуляции) подразумевается число элементарных сигналов (символов), передаваемых в единицу времени

$$V_T = \frac{1}{\tau}\ \text{бод}$$

Информационная скорость или скорость передачи информации определяется средним количеством информации, которое передается в единицу времени

$$V = nH\ \text{бит/сек.}$$

Для равновероятных сообщений, составленных из равновероятных взаимно независимых символов

$$V = \frac{1}{\tau}\log m.$$

Если символы *не равновероятны*

$$V = -\frac{1}{\tau}\sum_i p_i \log_2 p_i.$$

Если символы имеют *разную длительность*

$$V = -\frac{\sum_i p_i \log_2 p_i}{\sum_i p_i \tau_i}.$$

Выражение для пропускной способности характеризуется *максимальной энтропией*

$$C_{max} = \frac{H_{max}}{\tau}\ \text{бит/сек.}$$

Для двоичного кода

$$C_{max} = \frac{\log_2 2}{\tau} = \frac{1}{\tau}\ \text{бит/сек.}$$

Пропускная способность является важной характеристикой каналов связи. Возникает вопрос: какова должна быть пропускная способность канала, чтобы информация от источника X к приемнику Y поступала без задержек? Ответ дает теорема Шеннона.

1 Теорема Шеннона

Если имеется источник информации с энтропией $H(X)$ и канал связи с пропускной способностью c, то если $c > H(X)$, то всегда можно закодировать достаточно длинное сообщение таким образом, что оно будет передано без задержек. Если $c < H(X)$, то передача информации без задержек невозможна.

В любом реальном канале всегда присутствуют помехи. Однако, если их уровень мал, то вероятность искажения равна нулю и можно считать, что все сигналы передаются неискаженными. В этом случае среднее количество информации, переносимое одним символом

$$I(X,Y) = I(Y,X) = H(X)\ , \qquad H_{max} = \log_2 m\ .$$

Следовательно, *пропускная способность канала без помех* за единицу времени

$$c = n \log_2 m,$$

Реальные каналы характеризуются тем, что в них всегда есть помехи. *Пропускная способность дискретного канала с помехами* вычисляется

$$c = n(H(Y) - H(Y/X))_{max}.$$

где $H(Y) = \log_2 m$.

Для дискретного канала с помехами Шеннон дал вторую теорему.

2 Теорема Шеннона

Пусть имеется источник информации *X*, энтропия которого в единицу времени равна $H(X)$, и канал с пропускной способностью *c*. Если $H(X) > c$, то при любом кодировании передача сообщений без задержек и искажений невозможна. Если $H(X) < c$, то любое достаточно длинное сообщение можно всегда закодировать так, что оно будет предано без задержек и искажений с вероятностью сколь угодно близкой к единице.

Пример 1.

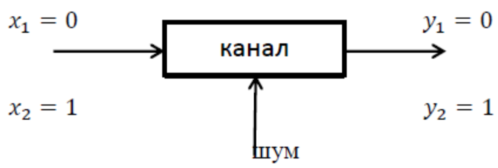
На вход дискретного симметричного канала без памяти поступают двоичные символы $x_1 = 0$ и $x_2 = 1$ с априорными вероятностями $p(x_1) =$ 0,85 и $p(x_2) = 0{,}15$. Переходные вероятности $p(y_j/x_i)$ в таком канале задаются соотношением

$$p(y_j/x_i) = \begin{cases} p, & j \neq i \\ 1 - p, & j = i \end{cases},$$

где $p = 0{,}05$ – вероятность ошибки. Определить все апостериорные вероятности.

Решение.

Ситуация в канале характеризуется схемой



Так как p=0,05, то вероятность правильного приема q=1-0,05.

В таком канале каждый кодовый символ может быть принят с ошибочной вероятностью

$$p(y_1/x_2) = p(y_2/x_1) = p = 0{,}05.$$

Правильно переданная информация описывается

$$p(y_1/x_1) = p(y_2/x_2) = q = 0{,}95.$$

По формуле Байеса определим апостериорные вероятности

$$p(x_i/y_j) = \frac{p(x_i) \cdot p(y_j/x_i)}{p(y_j)} = \frac{p(x_i) \cdot p(y_j/x_i)}{\sum_{i=1}^{N} p(x_i) \cdot p(y_j/x_i)}.$$

$$p(x_1/y_1) = \frac{p(x_1) \cdot p(y_1/x_1)}{p(x_1) \cdot p(y_1/x_1) + p(x_2) \cdot p(y_1/x_2)} = \frac{0{,}85 \cdot 0{,}95}{0{,}85 \cdot 0{,}95 + 0{,}15 \cdot 0{,}05} = 0{,}991,$$

$$p(x_1/y_2) = \frac{p(x_1)\cdot p(y_2/x_1)}{p(x_1)\cdot p(y_2/x_1)+p(x_2)\cdot p(y_2/x_2)} = \frac{0,85\cdot 0,05}{0,85\cdot 0,05+0,15\cdot 0,95} = 0,23,$$
$$p(x_2/y_1) = \frac{p(x_2)\cdot p(y_1/x_2)}{p(x_2)\cdot p(y_1/x_2)+p(x_1)\cdot p(y_1/x_1)} = \frac{0,15\cdot 0,05}{0,15\cdot 0,05+0,85\cdot 0,95} = 0,009,$$
$$p(x_2/y_2) = \frac{p(x_2)\cdot p(y_2/x_2)}{p(x_2)\cdot p(y_2/x_2)+p(x_1)\cdot p(y_2/x_1)} = \frac{0,15\cdot 0,95}{0,15\cdot 0,95+0,85\cdot 0,05} = 0,77.$$

Пример 2.

По каналу связи передается сообщение из ансамбля
$$X = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\ 0,09 & 0,1 & 0,22 & 0,07 & 0,15 & 0,17 & 0,02 & 0,18 \end{pmatrix}.$$

Средняя длительность передачи одного элемента сообщения в канале $\tau = 0,44$ мс. Шум в канале отсутствует. Определить пропускную способность канала и скорость передачи информации.

Решение.

Когда шум в канале отсутствует пропускная способность канала
$$c = V_T \log_2 m \text{ , } V_T = \frac{1}{\tau} = \frac{1}{0,44\cdot 10^{-3}} = 2273 \text{ с}^{-1}.$$

Объем алфавита данного сообщения m=8
$$c = \frac{1}{0,44\cdot 10^{-3}} \log_2 8 = 6819 \text{ бит/сек.}$$

Информационная скорость $V = I(X,Y)\cdot V_T$
$$I(X,Y) = H(X) - H(X/Y)$$

Так как шум в канале отсутствует, то $H(X/Y) = 0$, $I(X,Y) = H(X)$.

Определим энтропию заданного распределения
$$H(X) = \sum_{i=1}^{8} p(x_i)\log_2\frac{1}{p(x_i)} = 0,09\log_2\frac{1}{0,09} + 0,1\log_2\frac{1}{0,1} + 0,22\log_2\frac{1}{0,22} +$$
$$+0,07\log_2\frac{1}{0,07} + 0,15\log_2\frac{1}{0,15} + 0,17\log_2\frac{1}{0,17} + 0,02\log_2\frac{1}{0,02} +$$
$$+0,18\log_2\frac{1}{0,18} = 2,44 \text{ бит.}$$
$$V = 2,44\cdot 2273 = 5546,12 \text{ бит/сек.}$$

Пример 3.

Источник вырабатывает три сообщения с вероятностями: $p_1 = 0,1$; $p_2 = 0,2$; $p_3 = 0,7$. Сообщения независимы и передаются равномерным двоичным кодом (m=2) с длительностью символов равной 1 мс. Определить скорость передачи информации по каналу связи без помех.

Решение.
$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) = -(0,1\log_2 0,1 + 0,2\log_2 0,2 +$$
$$+0,7\log_2 0,7) = 1,16 \text{ бит}$$

Для передачи трех сообщений равномерным кодом необходимо два разряда, при этом длительность кодовой комбинации равна $2t$.

Средняя скорость передачи сигнала
$$V_T = \frac{1}{2t} = 500 \text{ бод}$$

Скорость передачи информации
$$V = V_T H(X) = 500\cdot 1,16 = 580 \text{ бит/сек.}$$

Пример 4.

По каналу связи передаются сообщения, вероятности которых равны: $p(x_1) = 0{,}1; p(x_2) = 0{,}2;\ \ p(x_3) = 0{,}3;\ \ p(x_4) = 0{,}4$. Канальная матрица, определяющая потери информации в канале связи

$$P(Y/X) = \begin{pmatrix} 0{,}99 & 0{,}01 & 0 & 0 \\ 0{,}01 & 0{,}97 & 0{,}02 & 0 \\ 0 & 0{,}01 & 0{,}98 & 0{,}01 \\ 0 & 0 & 0{,}01 & 0{,}99 \end{pmatrix}$$

Определить:

1)энтропию источника информации $H(X)$;

2) безусловную энтропию приемника информации $H(Y)$;

3) общую условную энтропию $H(Y/X)$;

4) скорость передачи информации, если время передачи одного символа первичного алфавита t=0,1 мс;

5) потери информации в канале связи при передаче 500 символов алфавита;

6) среднее количество принятой информации;

7) пропускную способность канала связи.

Решение.

1) Энтропия источника сообщений

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) = -(0{,}1\log_2 0{,}1 + 0{,}2\log_2 0{,}2 + 0{,}3\log_2 0{,}3 + 0{,}4\log_2 0{,}4) = 1{,}8465 \text{ бит}$$

2) Вероятности появления символов на входе приемника рассчитываются по формуле полной вероятности $p(y_i) = \sum_{i=1}^{N} p(x_i, y_i)$;

$$p(y_1) = 0{,}1\cdot 0{,}99 + 0{,}2\cdot 0{,}01 = 0{,}101;$$
$$p(y_2) = 0{,}1\cdot 0{,}01 + 0{,}2\cdot 0{,}97 + 0{,}3\cdot 0{,}01 = 0{,}198;$$
$$p(y_3) = 0{,}2\cdot 0{,}02 + 0{,}3\cdot 0{,}98 + 0{,}4\cdot 0{,}01 = 0{,}302;$$
$$p(y_4) = 0{,}3\cdot 0{,}01 + 0{,}4\cdot 0{,}99 = 0{,}399.$$

Проверка: $0{,}101 + 0{,}198 + 0{,}302 + 0{,}399 = 1$.

Энтропия приемника

$$H(Y) = -\sum_{i=1}^{n} p(y_i)\log_2 p(y_i) = -(0{,}101\log_2 0{,}101 + 0{,}198\log_2 0{,}198 + 0{,}302\log_2 0{,}302 + 0{,}399\log_2 0{,}399) = 1{,}85 \text{ бит}$$

3) Общая условная энтропия

$$H(Y/X) = \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j)\log_2\frac{1}{p(y_j/x_i)}.$$

$$H(Y/X) = -\big(0{,}1(0{,}99\log_2 0{,}99 + 0{,}01\log_2 0{,}01) + 0{,}2(0{,}01\log_2 0{,}01 + 0{,}97\log_2 0{,}97 + 0{,}02\log_2 0{,}02) + 0{,}3(0{,}01\log_2 0{,}01 + 0{,}98\log_2 0{,}98 + 0{,}01\log_2 0{,}01) + 0{,}4(0{,}01\log_2 0{,}01 + 0{,}99\log_2 0{,}99)\big) = 0{,}132.$$

4) Скорость передачи информации

$$V = V_T(H(X) - H(X/Y)) = V_T(H(Y) - H(Y/X)) = \frac{1{,}85 - 0{,}132}{0{,}0001} = 17{,}18 \text{ кбит/с};$$

5) Потери информации в канале связи

$$\Delta I = kH(Y/X) = 500\cdot 0{,}132 = 66 \text{ бит};$$

6) Среднее количество принятой информации

$$I = k(H(Y) - H(Y/X)) = 500 \cdot (1{,}85 - 0{,}132) = 859 \text{ бит};$$

7) Пропускная способность канала связи

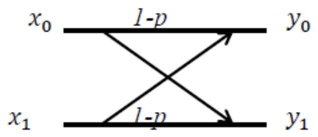
$$c = V(\log_2 m - H(Y/X)) = (2 - 0{,}132)/0{,}0001 = 17{,}18 \text{ кбит}.$$

Пример 5.

Определить скорость передачи по двоичному симметричному каналу связи при $\tau = 0{,}001$ с, если шумы в канале вносят ошибки таким образом, что в среднем четыре символа из 100 принимаются неверно (т.е. 1 вместо 0 и наоборот).

Решение.

Двоично симметричный канал можно представить схемой



Вероятность ошибки $p = \frac{4}{100}$.

Составим таблицу вероятностей

$$p(x_0) = 0{,}5, \quad p(y_0/x_0) = 0{,}96;$$
$$p(x_1) = 0{,}5, \quad p(y_1/x_0) = 0{,}04;$$
$$p(y_0) = 0{,}5, \quad p(y_0/x_1) = 0{,}04;$$
$$p(y_1) = 0{,}5, \quad p(y_1/x_1) = 0{,}96.$$

Пропускная способность двоично симметричного канала

$$c = \frac{1}{\tau}(1 + (1-p)\log_2(1-p) + p\log_2 p) = 10^3(1 + 0{,}96\log_2 0{,}96 + 0{,}04\log_2 0{,}04) = 0{,}76 \text{ кбит/с}.$$

Контрольные вопросы

1. Что называется технической скоростью?
2. Что называется информационной скоростью?
3. Как определяется информационная скорость для равновероятных сообщений?
4. Как определяется пропускная способность канала без помех?
5. Как определяется пропускная способность канала с помехами?
6. Сформулируйте 1-ю теорему Шеннона.
7. Сформулируйте 2-ю теорему Шеннона.

Раздел 2
ОСНОВЫ ТЕОРИИ КОДИРОВАНИЯ СООБЩЕНИЙ

ПРЕДСТАВЛЕНИЕ КОДОВ

Преобразование информации из одной формы представления (знаковой системы) в другую называется ***кодированием.***

Все виды информации в компьютере кодируются на машинном языке в виде логических последовательностей 0 и 1.

Цифры двоичного кода можно рассматривать как два равновероятных состояния (события). Каждая двоичная цифра машинного кода несет информацию 1 бит. Две цифры несут информацию в 2 бита, три цифры – в 3 бита и т.д. Количество информации в битах равно количеству цифр двоичного машинного кода. Восемь последовательных бит составляют байт. В одном байте можно закодировать значение одного символа из 256 возможных ($256=2^8$). Более крупной единицей информации является килобайт (Кбайт), равный 1024 байтам ($1024=2^{10}$) Еще крупные единицы измерения данных: мегабайт, гигабайт, терабайт (1Мбайт=1024 Кбайт, 1 Гбайт=1024 Мбайт, 1Тбайт=1024 Гбайт).

Целые числа кодируются двоичным путём деления числа на два. Для кодирования нечисловой информации используется алгоритм: все возможные значения кодируемой информации нумеруются и эти номера кодируются с помощью двоичного кода.

Понятие кодирование означает преобразование информации в форму, удобную для передачи по определённому каналу связи.

Декодирование – восстановление принятого сообщения из кодированного вида в вид доступный для потребителя.

Одно и тоже сообщение можно закодировать различными способами. Необходимо найти *оптимальный способ* кодирования, при котором на передачу сообщения тратится минимальное время.

Если на передачу каждого элементарного символа (0 или 1) тратится одно и тоже время, то оптимальным будет такой код, при котором на передачу сообщения заданной длины будет затрачено минимальное количество символов.

Например, пусть имеются буквы русского алфавита а, б, в, г,...+ промежуток между словами (-). Если не различать ь и ъ (как принято в телеграфии), то получим 32 буквы. Требуется закодировать двоичным кодом буквы так, чтобы каждой букве соответствовала определенная комбинация символов 0 и 1 и, чтобы среднее число этих символов на букву текста было минимальным.

1-й вариант. Не меняя порядка букв, пронумеровав их от 0 до 31 и перевести их в двоичную систему счисления, получим следующий код:

а ~ 00000

б ~ 00001

в ~ 00010
г ~ 00011
....
я ~ 11110
- ~ 11111

В этом коде на каждую букву тратится ровно пять элементарных символов. Является ли этот код оптимальным? Можно ли составить другой код, при котором на одну букву в среднем приходится меньше элементарных символов?

2 вариант. Так как одни буквы встречаются часто (а,о,е), а другие (щ,э,ф) редко, то часто встречающиеся буквы целесообразно закодировать меньшим числом символов, а реже встречающиеся – большим. Чтобы составить такой код нужно знать частоты букв русского алфавита (табл. 1).

Таблица 1 – Статистические данные русского алфавита

буква	частота	буква	частота	буква	частота
о	0,095	к	0,029	г	0,014
е	0,074	м	0,026	ч	0,013
а	0,064	д	0,026	й	0,01
и	0,064	п	0,024	х	0,009
т	0,056	у	0,021	ж	0,008
н	0,056	я	0,019	ю	0,007
с	0,047	ы	0,016	ш	0,006
р	0,041	з	0,015	щ	0,003
в	0,039	ъ,ь	0,015	э	0,003
л	0,036	б	0,015	ф	0,002

(-) – 0,145

Пользуясь такой таблицей, можно составить наиболее экономичный код на основе соображений, связанных с количеством информации. Код будет самым экономичным, когда каждый 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 системы и максимальна в случае, когда оба состояния равновероятны.

Основой оптимального кодирования будет требование: элементарные символы в закодированном тексте встречались в среднем одинаково часто.

2.1 Метод Шеннона-Фано

Метод Шеннона-Фано соответствует требованию оптимального кодирования. Алгоритм построения кода Шеннона-Фано состоит в том, что кодируемые символы (буквы) разделяются на две равновероятные подгруппы: для символов 1-й подгруппы на втором месте ставится 0, а для 2-й подгруппы – 1 и т.д.

Берутся первые шесть букв (от – до т). Сумма их вероятностей равна 0,498, на все остальные (от н до ф) приходится 0,502. Первые шесть букв будут иметь на первом месте 0, остальные 1.

Далее снова первая группа делится на две приблизительные равновероятные подгруппы: (от – до щ) и (от е до т) и т.д. Для всех букв первой подгруппы на втором месте ставится 0, а второй подгруппы – 1.

Процесс продолжается до тех пор, пока в каждом подразделении не окажется ровно одна буква, которая и будет закодирована определенным двоичным кодом.

Пример 1.

Построить таблицу кодов алфавита методом Шеннона-Фано.

Записать двоичным кодом фразу «теория информации».

Решение.

Составим таблицу 2 кодов алфавита описанным ранее методом.

Таблица 2 – Коды русского алфавита по методу Шеннона-Фано

буквы	вероятности	символы кода									
		1-й	2-й	3-й	4-й	5-й	6-й	7-й	8-й	9-й	код
-	0,145		0	0							000
о	0,095			1							001
е	0,074			0	0						0100
а	0,064	0			1						0101
и	0,064		1	1	0						0110
т	0,056				1						0111
н	0,056	1	0	0	0						1000
с	0,047				1						1001
р	0,041			1	0	0					10100
в	0,039					1					10101
л	0,036				1	0					10110
к	0,029					1					10111
м	0,026		1	0	0	0					11000
д	0,026					1					11001
п	0,024				1	0					11010
у	0,021					1	0				110110
я	0,019						1				110111
ы	0,016			1	0	0	0				111000
з	0,015						1				111001
ъ,ь	0,015					1	0				111010
б	0,015						1				111011
г	0,014				1	0	0				111100
ч	0,013						1				111101
й	0,01					1	0	0			1111100
х	0,009							1	0		11111010
ж	0,008								1		11111011
ю	0,007						1	0	0		11111100
ш	0,006								1		11111101
ц	0,003							1	0	0	111111100
щ	0,003									1	111111101
э	0,003								1	0	111111110
ф	0,002									1	111111111

Фраза «теория информации» будет выглядеть следующим образом:

0111	0100	001	10100	0110	110111	000	0110	1000	111111111	001	10100	11000	0101	111111100	0110	0110
т	е	о	р	и	я	пробел	и	н	ф	о	р	м	а	ц	и	и

Здесь нет необходимости отделять буквы специальным знаком.

Пример 2.

Декодировать сообщение методом Шеннона-Фано, используя таблицу кодов из примера 1:

1001110100011001001111011000101110011100101101010000110101010110000110110111

Решение.

«способ кодирования»

Случайные перестановки 0 и 1 недопустимы.

Для того, чтобы выяснить, является ли построенный код оптимальным, необходимо найти среднюю информацию, приходящуюся на один элементарный символ (0 или 1) и сравнить ее с максимально возможной информацией.

Определим среднюю информацию, содержащуюся в одной букве передаваемого текста, т.е. энтропия на одну букву

$$H(б) = -(0{,}145 \cdot \log_2 0{,}145 + 0{,}95 \cdot \log_2 0{,}95 + \dots + 0{,}002 \cdot \log_2 0{,}002) = \\ = 4{,}42.$$

Определим ***среднее число элементарных символов на букву*** как произведение количества символов кода на вероятность появления данной буквы

$$n_{ср} = 3 \cdot 0{,}145 + 3 \cdot 0{,}095 + 4 \cdot 0{,}074 + \dots + 9 \cdot 0{,}002 = 4{,}45.$$

Информация на один элементарный символ

$$I_{ср} = \frac{H(б)}{n_{ср}} = \frac{4{,}42}{4{,}45} = 0{,}994 \quad \text{дв.ед.}$$

Таким образом, информация на один символ близка к своему верхнему пределу 1. Следовательно, построенный код в целом отвечает принципу оптимальности.

В случае кодирования простым двоичным кодом каждая буква изображается пятью двоичными знаками и информация на один символ

$$I_{ср} = \frac{H(б)}{n_{ср}} = \frac{4{,}42}{5} = 0{,}884 \quad \text{дв.ед.}$$

Это меньше, чем при кодировании методом Шеннона-Фано.

Однако, кодирование по «буквам» не является экономичным, так как между соседними буквами любого текста всегда есть зависимость. Например, после гласной буквы не может быть ь или ъ, после шипящих не может быть я или ю, после нескольких согласных следует гласная

Известно, что при объединении зависимых систем суммарная энтропия меньше суммы энтропий отдельных систем. Следовательно, информация, передаваемая отрезком связанного текста, всегда меньше, чем информация на один символ, умноженная на число символов. Поэтому более экономичный код можно построить, если кодировать не каждую букву в отдельности, а целые «блоки» из букв. Например, «тся», «ает», «ние» и т.д.

Кодируемые блоки располагаются в порядке убывания частот и двоичное кодирование осуществляется по тому же принципу.

В ряде случаев разумно кодировать не на блоки букв, а целые куски текста. Например, «поздравляю новым годом желаю здоровья успехов в работе».

Пример 3.

Имеется алфавит символов и их вероятности, с которыми они встречаются в тексте. Построить таблицу кодов символов методом Шеннона-Фано. Закодировать сообщение «вилка» и раскодировать заданную последовательность кодов.

а	в	л	и	е	с	к
0,3	0,2	0,15	0,1	0,1	0,08	0,07

Решение.

Составим таблицу кодов для символов алфавита

буква	вероятность	символ кода				код
а	0,3	0	0			00
в	0,2		1			01
л	0,15	1	0	0		100
и	0,1			1		101
е	0,1		1	0		110
с	0,08			1	0	1110
к	0,07				1	1111

Сообщению «вилка» соответствует выходная последовательность кодов 01101100111100.

Выходной последовательности кодов 100101111000 соответствует сообщение «лиса».

Пример 4.

Провести кодирование по методу Шеннона-Фано двухбуквенных комбинаций, когда алфавит состоит из двух букв А и В, имеющих вероятности Р(A)=0,8 и Р(B)=0,2. Каково среднее число символов на знак?

Решение.

комбинации букв	вероятность	разбиение на подгруппы			код
АА	0,64	0			0
АБ	0,16	1	0		10
БА	0,16		1	0	110
ББ	0,04			1	111

$$n_{\text{ср}} = 1 \cdot 0{,}64 + 2 \cdot 0{,}16 + 3 \cdot 0{,}16 + 3 \cdot 0{,}04 = 1{,}56.$$

Пример 5.

Закодировать сообщение методом Шеннона-Фано «РоссийскаяСистемаВысшегоТехническогоОбразования».

Решение.

Составляется алфавит кода и для каждого символа этого алфавита и определяется вес [количество повторений символа в сообщении].

Р-2, О-5, С-7, И-4, Й-1, К-2, А-4, Я-2, Т-2, Е-4, М-1, В-2, Ы-1, Ш-1, Г-2, Х-1, Н-2, Ч-1, Б-1, З-1.

Алфавит ранжируется по убыванию веса символа.

С-7, О-5, И-4, А-4, Е-4, Р-2, К-2, Я-2, Т-2, В-2, Г-2, Н-2, Й-1, М-1, Ы-1, Ш-1, Х-1, Ч-1, Б-1, З-1.

Далее создается таблица №3 символов. Она делится на две группы таким образом, чтобы каждая из групп имела приблизительно одинаковую частоту по сумме символов. Первой группе устанавливается начало кода в 0, второй в 1.

Таблица 3 – Коды символов по методу Шенно-Фано

<table>
<tr><th rowspan="2">буква</th><th rowspan="2">частота</th><th colspan="6">символ кода</th><th rowspan="2">код</th></tr>
<tr><th>1-й</th><th>2-й</th><th>3-й</th><th>4-й</th><th>5-й</th><th>6-й</th></tr>
<tr><td>С</td><td>7</td><td rowspan="5">0</td><td rowspan="2">0</td><td>0</td><td rowspan="3"></td><td rowspan="8"></td><td rowspan="15"></td><td>000</td></tr>
<tr><td>О</td><td>5</td><td>1</td><td>001</td></tr>
<tr><td>И</td><td>4</td><td rowspan="3">1</td><td>0</td><td>010</td></tr>
<tr><td>А</td><td>4</td><td rowspan="2">1</td><td>0</td><td>0110</td></tr>
<tr><td>Е</td><td>4</td><td>1</td><td>0111</td></tr>
<tr><td>Р</td><td>2</td><td rowspan="15">1</td><td rowspan="5">0</td><td rowspan="2">0</td><td>0</td><td>1000</td></tr>
<tr><td>К</td><td>2</td><td>1</td><td>1001</td></tr>
<tr><td>Я</td><td>2</td><td rowspan="3">1</td><td>0</td><td>1010</td></tr>
<tr><td>Т</td><td>2</td><td rowspan="2">1</td><td>0</td><td>10110</td></tr>
<tr><td>В</td><td>2</td><td>1</td><td>10111</td></tr>
<tr><td>Г</td><td>2</td><td rowspan="10">1</td><td rowspan="4">0</td><td rowspan="2">0</td><td>0</td><td>11000</td></tr>
<tr><td>Н</td><td>2</td><td>1</td><td>11001</td></tr>
<tr><td>Й</td><td>1</td><td rowspan="2">1</td><td>0</td><td>11010</td></tr>
<tr><td>М</td><td>1</td><td>1</td><td>11011</td></tr>
<tr><td>Ы</td><td>1</td><td rowspan="6">1</td><td rowspan="3">0</td><td>0</td><td>11100</td></tr>
<tr><td>Ш</td><td>1</td><td rowspan="2">1</td><td>0</td><td>111010</td></tr>
<tr><td>Х</td><td>1</td><td>1</td><td>111011</td></tr>
<tr><td>Ч</td><td>1</td><td rowspan="3">1</td><td>0</td><td></td><td>11110</td></tr>
<tr><td>Б</td><td>1</td><td rowspan="2">1</td><td>0</td><td>111110</td></tr>
<tr><td>З</td><td>1</td><td>1</td><td>111111</td></tr>
</table>

Получаем код к каждой букве:

С	000	Р	1000	Г	11000	Ш	111010
О	001	К	1001	Н	11001	Х	111011
И	010	Я	1010	Й	11010	Ч	11110
А	0110	Т	10110	М	11011	Б	111110
Е	0111	В	10111	Ы	11100	З	111111

Используя полученную таблицу кодов, кодируется входной поток – каждый символ заменяется соответствующим кодом.

1000 001 000 000 010 11010 000 1001 0110 1010 000 010 000 10110 0111 11011 0110 10111 11100 000 111010 0111 11000 001 10110 0111 111011 11001 010 11110 0111 000 1001 001 11000 001 111110 1000 0110 11111 001 10111 0110 11001 010 1010

ЭФФЕКТИВНОЕ КОДИРОВАНИЕ

Эффективное или экономичное кодирование используется для уменьшения объемов информации на носителе-сигнале таким образом, чтобы устранить избыточность.

Для кодирования символов исходного алфавита используются двоичные коды переменной длины: чем больше частота символа, тем короче его код. Эффективность кода определяется средним числом двоичных разрядов для кодирования одного символа.

При эффективном кодировании существует предел сжатия, ниже которого не «спускается» ни один метод эффективного кодирования – иначе будет потеряна информация. Этот параметр определяется предельным значением двоичных разрядов возможного эффективного кода

$$l_{\text{пр}} = -\sum_{i=1}^{n} f_i \log_2 f_i \ ,$$

где n – мощность кодируемого алфавита,

f_i – частота i-го символа кодируемого алфавита.

2.2 Метод Хаффмана

Это метод эффективного кодирования. Пусть имеются сообщения входного алфавита $A = \{x_1, x_2, \dots x_n\}$ с соответствующими вероятностями их появления $p_1, p_2, \dots p_n$. Тогда алгоритм кодирования Хаффмана состоит в следующем:

1. Сообщения располагаются в столбец в порядке убывания вероятностей их появления.

2. Два самых маловероятных сообщения объединяются в одно y, которое имеет вероятность равную сумме вероятностей события x_{n-1}, x_n, т.е. $p_{n-1} + p_n$. В результате получаются сообщения $x_1, x_2, \dots x_{n-2}$, y, вероятности которых $p_1, p_2, \dots p_{n-2}, p_{n-1} + p_n$.

3. Повторить шаги 1 и 2 до тех пор, пока не получится единственное сообщение, вероятность которого равна 1.

4. Проводя линии, объединяющие сообщения и образующие последовательности подмножества, получится дерево, в котором отдельные сообщения являются концевыми узлами. Соответствующие им кодовые слова можно определить, приписывая правым ветвям объединения символ 1, а левым – 0 (или наоборот).

Пример 1.

Источник генерирует знак x_1 с вероятностью 0,8 и x_2 с вероятностью 0,2. Построить эффективные коды Шеннона-Фано и Хаффмана для последовательностей из трех знаков $x_i x_j x_k$. Каково среднее число символов на знак? Сравнить с энтропией источника.

Решение.

Энтропия источника $H(X) = -(0{,}8 \cdot \log_2 0{,}8 + 0{,}2 \cdot \log_2 0{,}2) = 0{,}722$. Для наиболее наглядного сравнения энтропии источника со средним значением длины кодового обозначения рассмотрим всевозможные комбинации знаков.

Применим метод Шеннона-Фано

знак	вероятность	код
x_1	0,8	1
x_2	0,2	0

Среднее количество символов в коде (или длина)

$$n_{\text{ср}} = 1 \cdot 0{,}8 + 1 \cdot 0{,}2 = 1$$

Среднее число символов на знак $\frac{n_{\text{ср}}}{1} = 1$, что на 28% больше энтропии.

Построим код Шеннона-Фано для всевозможных двух знаковых комбинаций.

<table>
<tr><th rowspan="2">комбинации знаков</th><th rowspan="2">вероятность</th><th colspan="3">символы кода</th><th rowspan="2">код</th></tr>
<tr><th>1-й</th><th>2-й</th><th>3-й</th></tr>
<tr><td>$x_1 x_1$</td><td>0,64</td><td>1</td><td></td><td rowspan="2"></td><td>1</td></tr>
<tr><td>$x_1 x_2$</td><td>0,16</td><td rowspan="3">0</td><td>1</td><td>01</td></tr>
<tr><td>$x_2 x_1$</td><td>0,16</td><td rowspan="2">0</td><td>1</td><td>001</td></tr>
<tr><td>$x_2 x_2$</td><td>0,04</td><td>0</td><td>000</td></tr>
</table>

Среднее количество символов в коде (или длина)

$$n_{\text{ср}} = 1 \cdot 0{,}64 + 2 \cdot 0{,}16 + 3 \cdot 0{,}16 + 3 \cdot 0{,}04 = 1{,}56,$$

Среднее число символов на знак $\frac{n_{\text{ср}}}{2} = 0{,}78$ (знаменатель означает из скольких знаков состоит одна комбинация), что на 8% больше энтропии.

Построим код Шеннона-Фано для всевозможных трех знаковых комбинаций.

<table>
<tr><th rowspan="2">комбинации знаков</th><th rowspan="2">вероятность</th><th colspan="5">символы кода</th><th rowspan="2">код</th></tr>
<tr><th>1-й</th><th>2-й</th><th>3-й</th><th>4-й</th><th>5-й</th></tr>
<tr><td>$x_1 x_1 x_1$</td><td>0,512</td><td>1</td><td></td><td></td><td rowspan="3"></td><td rowspan="3"></td><td>1</td></tr>
<tr><td>$x_1 x_1 x_2$</td><td>0,128</td><td rowspan="2">0</td><td rowspan="2">1</td><td>1</td><td>011</td></tr>
<tr><td>$x_1 x_2 x_1$</td><td>0,128</td><td>0</td><td>010</td></tr>
</table>

$x_2x_1x_1$	0,128			1			001
$x_1x_2x_2$	0,032				1	1	00011
$x_2x_1x_2$	0,032		0	0		0	00010
$x_2x_2x_1$	0,032				0	1	00001
$x_2x_2x_2$	0,008					0	00000

Среднее количество символов в коде (или длина)

$$n_{\text{ср}} = 1 \cdot 0{,}512 + 3 \cdot 0{,}384 + 3 \cdot 0{,}16 + 5 \cdot 0{,}104 = 2{,}184,$$

Среднее число символов на знак $\frac{n_{\text{ср}}}{3} = 0{,}728$ (знаменатель означает из скольких знаков состоит одна комбинация), что на 0,8% больше энтропии.

Построим код Хаффмана для всевозможных трех знаковых комбинаций (рисунок 2.1). Затем строим кодовое дерево (рисунок 2.2) и составляем код комбинаций:

$x_1x_1x_1$	1
$x_1x_1x_2$	011
$x_1x_2x_1$	010
$x_2x_1x_1$	001
$x_1x_2x_2$	00011
$x_2x_1x_2$	00010
$x_2x_2x_1$	00001
$x_2x_2x_2$	00000

Коды Шенно-Фано и Хаффмана совпали.

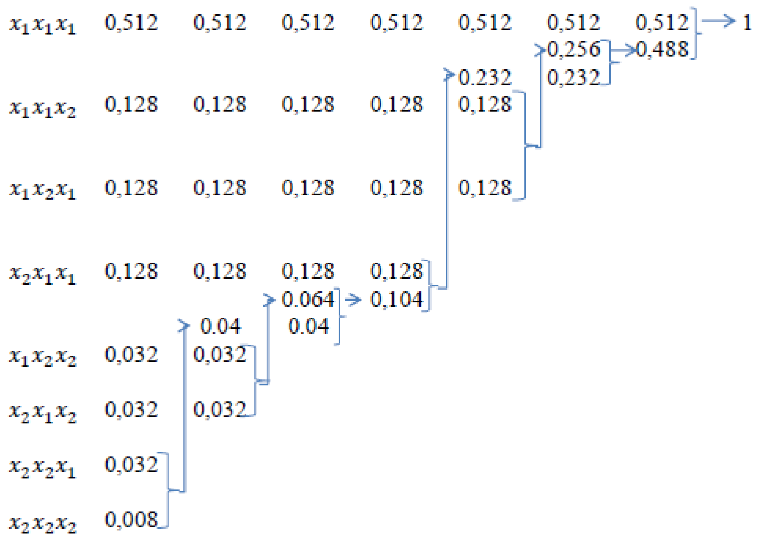


Рисунок 2.1 – Код Хаффмана

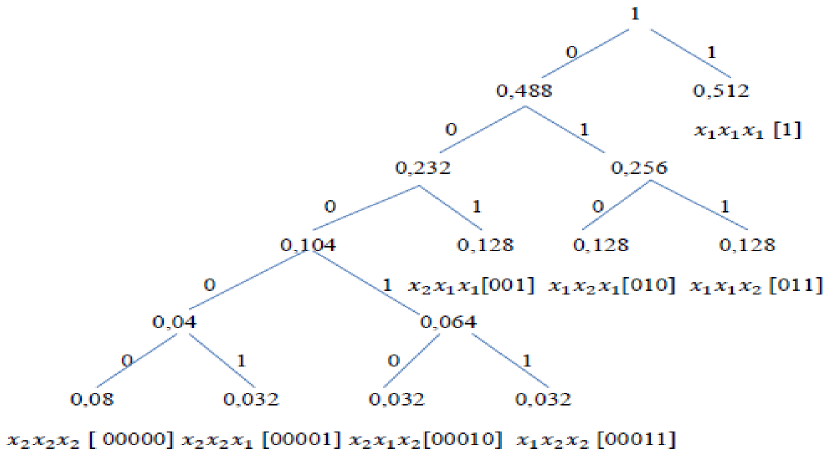


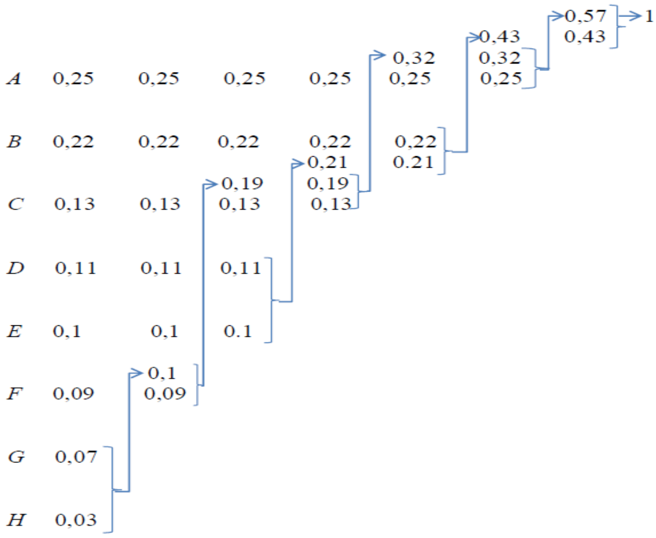
Рисунок 2.2 – Кодовое дерево

Пример 2.

Построить код Хаффмана для следующих данных:

A	B	C	D	E	F	G	H
0,25	0,22	0,13	0,11	0,1	0,09	0,07	0,03

Решение.



На основании полученной таблицы можно построить кодовое дерево (рисунок 2.3).

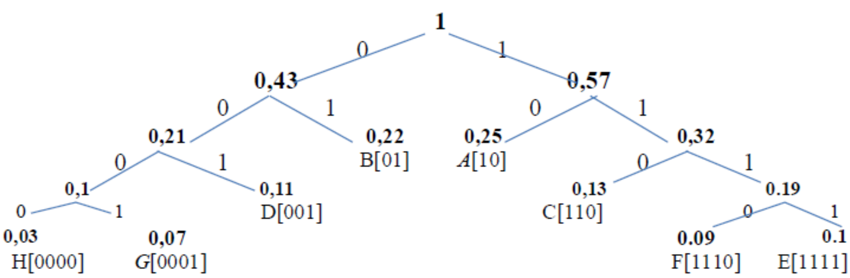


Рисунок 2.3 – Кодовое дерево

Пример 3.

Пусть передалась следующая последовательность 1001110001. Декодировать сообщение.

Решение.

A B C D.

Пример 4.

Сообщение X ссимволами x_1, x_2, x_3, x_4, x_5 передается по дискретному двоичному каналу с вероятностями $p(x_1) = 0{,}1$, $p(x_2) = 0{,}1$, $p(x_3) = 0{,}4$, $p(x_4) = 0{,}3$, $p(x_5) = 0{,}1$. Полоса пропускания канала обеспечивает возможность передачи двоичных символов длительностью $\tau = 10^{-4}$ сек.

Требуется выбрать наилучший способ кодирования.

Решение.

Пропускная способность двоичного дискретного канала $c = \frac{1}{\tau} = 10000$ бит/сек.

1) Сообщение закодировано равномерным двоичным кодом.

X	*P(X)*	код
x_3	0,4	000
x_4	0,3	001
x_1	0,1	010
x_2	0,1	011
x_5	0,1	111

Среднее количество символов в коде (или длина)

$$n_{\text{ср}} = 3 \cdot 0{,}4 + 3 \cdot 0{,}3 + 3 \cdot 0{,}1 + 3 \cdot 0{,}1 + 3 \cdot 0{,}1 = 3\ ,$$

$$\tau_{\text{ср}} = n_{\text{ср}} \cdot \tau = 3 \cdot 0{,}0001 = 0{,}0003\ ,$$

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot \log_2 p(x_i) = -(0{,}4 \cdot \log_2 0{,}4 + 0{,}3 \cdot \log_2 0{,}3 + $$
$$+ 3 \cdot 0{,}1 \cdot \log_2 0{,}1) = 2{,}046 \text{ бит.}$$

$$V = \frac{H(X)}{\tau_{\text{ср}}} = \frac{2{,}046}{0{,}0003} = 6820\ .$$

2) Сообщение закодировано кодом Шенно-Фано.

X	*P(X)*	код
x_3	0,4	0
x_4	0,3	10
x_1	0,1	110
x_2	0,1	1110
x_5	0,1	1111

Среднее количество символов в коде (или длина)

$$n_{\text{ср}} = 1 \cdot 0{,}4 + 2 \cdot 0{,}3 + 3 \cdot 0{,}1 + 4 \cdot 0{,}1 + 4 \cdot 0{,}1 = 2{,}1,$$

$$\tau_{\text{ср}} = n_{\text{ср}} \cdot \tau = 2{,}1 \cdot 0{,}0001 = 0{,}00021,$$

$$V = \frac{H(X)}{\tau_{\text{ср}}} = \frac{2{,}046}{0{,}00021} = 9742{,}86.$$

3) Сообщение закодировано кодом Хаффмана.

X	$P(X)$	код
x_3	0,4	0
x_4	0,3	11
x_1	0,1	100
x_2	0,1	1011
x_5	0,1	1010

Среднее количество символов в коде (или длина)

$$n_{\text{ср}} = 1 \cdot 0{,}4 + 2 \cdot 0{,}3 + 3 \cdot 0{,}1 + 4 \cdot 0{,}1 + 4 \cdot 0{,}1 = 2{,}1,$$

$$\tau_{\text{ср}} = n_{\text{ср}} \cdot \tau = 2{,}1 \cdot 0{,}0001 = 0{,}00021,$$

$$V = \frac{H(X)}{\tau_{\text{ср}}} = \frac{2{,}046}{0{,}00021} = 9742{,}86.$$

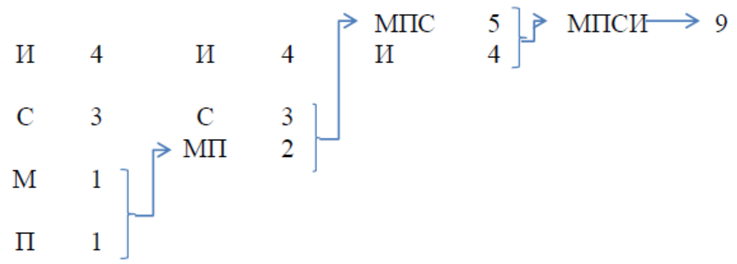
Самым медленным кодом оказался равномерный код $V = 6820$ сим/сек, что в 1,43 раза меньше, чем при кодировании методом Шеннона-Фано и Хаффмана.

Пример 5.

Закодировать методом Хаффмана слово «миссисипи».

Решение.

Рассчитаем частоту появления символов (вес) и составим таблицу.



Построим кодовое дерево (рисунок 2.4).

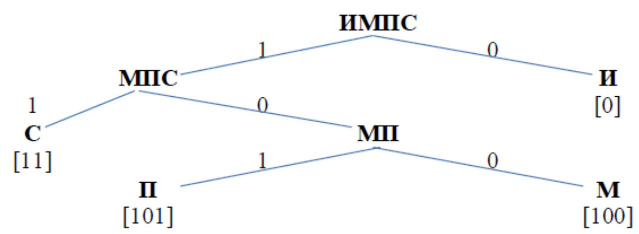


Рисунок 2.4 – Кодовое дерево

Таким образом закодированное слово «миссисипи» выглядит так: 1000111101101010.

Длина закодированного слова – 16 бит.

Контрольные вопросы

1. Назначение и цели эффективного кодирования.
2. Поясните за счет чего, обеспечивается сжатие информации при применении эффективного кодирования.
3. Чем определяется минимальная длина кодовой комбинации при применении эффективного кодирования?
4. Какие проблемы возникают при разделении неравномерных кодовых комбинаций?
5. Объяснить принцип построения кода Хаффмана.
6. Какой код является самым выгодным?
7. За счет чего при эффективном кодировании уменьшается средняя длина кодовой комбинации?
8. До какого предела может уменьшится длина кодовой комбинации при эффективном кодировании?
9. При каком распределении букв первичного алфавита оптимальный неравномерный код оказывается самым эффективным?

2.3 Помехоустойчивое кодирование

Высокие требования к достоверности передачи, обработки и хранения информации диктуют необходимость такого кодирования информации, при котором обеспечивалось бы возможность обнаружения и исправления ошибок. Это достигается путем введения избыточности, которая позволяет выбрать передаваемые последовательности символов, чтобы они удовлетворяли дополнительным условиям, проверка которых на приемной стороне дает возможность обнаружить и исправить ошибки.

Коды, обладающие таким свойством, получили название *помехоустойчивые*. Они используются для обнаружения ошибок и для исправления ошибок (корректирующие коды).

Платой за помехоустойчивость является необходимость увеличения длины слов по сравнению с обычным кодом. Естественно, что введение дополнительных контрольных разрядов увеличивает затраты на хранение и передачу кодированной информации. При этом объем полезной информации остается неизменным.

Контрольные разряды не передают информацию и в этом смысле бесполезны. Относительное число контрольных разрядов называется ***избыточностью помехоустойчивого кода***: $Q = \frac{n_k}{n} \cdot 100\%$.

Корректирующие коды образуются путем добавления к исходной кодовой комбинации n_k контрольных (избыточных) символов. В итоге влинию передаются $n = n_и + n_k$ символов.

При передаче кода может быть искажен любой информационный символ или ни один.

В передаче информации контрольный разряд не участвует, так как является линейно зависимым от информационных. Код с контролем по четности позволяет обнаружить одиночные ошибки в блоках данных при передаче данных. Однако он не может обнаружить двукратные ошибки потому, что двукратная ошибка переводит кодовое слово через промежуток между допустимыми словами и превращает его в другое допустимое слово.

Если всего n символов, то с помощью контрольных символов, входящих в это число должно быть создано 2^{n_k} число комбинаций, чтобы свободно различать $n+1$ вариант. Поэтому n_k должно удовлетворять неравенству:

$$2^{n_k} \geq n+1\ ,\ 2^n = 2^{n_k+n_и} = 2^{n_k} \cdot 2^{n_и}\ ,\ 2^n \geq (n+1) \cdot 2^{n_и},$$

где 2^n – полное число комбинаций кода.

Число информационных символов кода, обнаруживающего и корректирующего одиночную ошибку равно:

$$2^{n_и} \leq \frac{2^n}{n+1}.$$

Для вычисления основных параметров кода Хэмминга задается либо количество информационных символов, либо информационных комбинаций $N = 2^{n_и}$. Соотношения $n, n_k, n_и$ представлены в таблице.

Таблица – Соотношения $n, n_k, n_и$

n	$n_и$	n_k
1	0	1
2	0	2
3	1	2
4	1	3
5	2	3
6	3	3
7	4	3
8	4	4
9	5	4
10	6	4
11	7	4
12	8	4
13	9	4
14	10	4
15	11	4
16	11	5

Для выбора проверочных позиций составляется таблица для ряда натуральных чисел в двоичном коде. Число ее строк $n = n_и + n_k$. Первой строке соответствует проверочный коэффициент a_1, второй a_2 и т.д. Затем выявляют проверочные позиции, выписывая коэффициенты по следующему принципу: в первую проверку входят коэффициенты, которые содержат единицу в младшем разряде $(a_1, a_3, a_5, a_7, a_{11}, \dots)$, во вторую – во втором разряде $(a_2, a_3, a_6, a_7, a_{10}, a_{11}, \dots)$, в третью – в третьем разряде и т.д.

0001	a_1
0010	a_2
0011	a_3
0100	a_4
0101	a_5
0110	a_6
0111	a_7
1000	a_8
1001	a_9
1010	a_{10}
1011	a_{11}

Номера проверочных коэффициентов соответствуют номерам проверочных позиций. Составляется общая таблица 5 проверочных позиций кода Хэмминга.

Таблица 5 – Проверочные позиции кода Хэмминга

N проверки	проверочные позиции	N контрол. символа
1	1,3,5,7,9,11	1
2	2,3,6,7,10,11,14,15,1819,22,24…	2
3	4,5,6,7,12,13,14,15,20,21,22,23…	4
4	8,9,10,11,12,13,14,15,24,25,26,27,28,29,30,31,40,41,42…	8

Номера контрольных символов выбирают по закону $N = 2^i$, i=0,1,2...n.

Затем определяются значения контрольных коэффициентов (0 или 1) по правилу четности: *сумма единиц на проверочных позициях должна быть четной. Если она четная, то значение контрольного коэффициента 0, в противном случае – 1.*

Пример 1.

Определить кодовое расстояние между комбинациями 100101100 и 110110101, используя правило четности.

Решение.

Просуммируем 100101100 $\oplus$110110101=010011001

Вес полученной суммы (количество единиц) равно 4, следовательно, кодовое расстояние d=4.

Пример 2.

Построить макет кода Хэмминга и определить значения корректирующих разрядов для кодовой комбинации ($n_и = 4$) 0101.

Решение.

По таблице №4 минимальное число контрольных символов $n_к = 3$, а общее число символов $n = n_и + n_k = 7$. По закону $N = 2^i$, i=0,1,2...n

определяются позиции проверочных символов: 1,2 и 4, остальные позиции занимают информационные символы.

Кодовая комбинация будет иметь вид:

C=_ _ 0_ 101.

Для определения значения корректирующих разрядов составляются уравнения:

$$a_1 = a_3 \oplus a_5 \oplus a_7 = 0 \oplus 1 \oplus 1 = 0$$
$$a_2 = a_3 \oplus a_6 \oplus a_7 = 0 \oplus 0 \oplus 1 = 1$$
$$a_4 = a_5 \oplus a_6 \oplus a_7 = 1 \oplus 0 \oplus 1 = 0$$

(знак $\oplus$ обозначает сложение по правилу четности).

Следовательно, C=0100101.

Пример 3.

Пусть дана информационная последовательность 11001001. Преобразовать заданное информационное слово в код Хэмминга.

Решение.

Количество информационных символов $n_{\text{и}}$=8. По таблице №4 $n_{\text{к}} = 4$, а общее число символов $n = n_{\text{и}} + n_k = 12$. По закону $N = 2^i$, i=0,1,2...n определяются позиции проверочных символов: 1,2,4 и 8, остальные позиции занимают информационные символы.

Следовательно, кодовая комбинация будет иметь вид:

C=_ _ 1_ 100_1001.

Для определения значения проверочных символов составляются уравнения:

$$a_1 = a_3 \oplus a_5 \oplus a_7 \oplus a_9 \oplus a_{11} = 1 \oplus 1 \oplus 0 \oplus 1 \oplus 0 = 1$$
$$a_2 = a_3 \oplus a_6 \oplus a_7 \oplus a_{10} \oplus a_{11} = 1 \oplus 0 \oplus 0 \oplus 0 \oplus 0 = 1$$
$$a_4 = a_5 \oplus a_6 \oplus a_7 \oplus a_{12} = 1 \oplus 0 \oplus 0 \oplus 1 = 0$$
$$a_8 = a_9 \oplus a_{10} \oplus a_{11} \oplus a_{12} = 1 \oplus 0 \oplus 0 \oplus 1 = 0$$

Следовательно, C=111010001001.

Пример 4.

В результате передачи кодовой комбинации из предыдущего примера 3 принята кодовая комбинация C*=110010001001, т.е. произошло искажение 3-го разряда. Обнаружить ошибку.

Решение.

Значения проверок равны:

$$\mathrm{E}_1 = a_1 \oplus a_3 \oplus a_5 \oplus a_7 \oplus a_9 \oplus a_{11} = 1 \oplus 0 \oplus 1 \oplus 0 \oplus 1 \oplus 0 = 1$$
$$\mathrm{E}_2 = a_2 \oplus a_3 \oplus a_6 \oplus a_7 \oplus a_{10} \oplus a_{11} = 1 \oplus 0 \oplus 0 \oplus 0 \oplus 0 \oplus 0 = 1$$
$$\mathrm{E}_3 = a_4 \oplus a_5 \oplus a_6 \oplus a_7 \oplus a_{12} = 0 \oplus 1 \oplus 0 \oplus 0 \oplus 1 = 0$$
$$\mathrm{E}_4 = a_8 \oplus a_9 \oplus a_{10} \oplus a_{11} \oplus a_{12} = 0 \oplus 1 \oplus 0 \oplus 0 \oplus 1 = 0$$

Тогда контрольное число (синдром) ошибки:

$$S = \mathrm{E}_4\mathrm{E}_3\mathrm{E}_2\mathrm{E}_1 = 0011_{(2)} = 3_{(10)}$$

Переход из двоичной системы счисления в десятичную:

$$1 \cdot 2^0 + 1 \cdot 2^1 + 0 \cdot 2^2 + 0 \cdot 2^3 = 3.$$

Значение синдрома указывает, что в результате однократной ошибки искажен 3-й разряд принятой кодовой комбинации. Для исправления ошибки достаточно инвертировать искаженный разряд.

Пример 5.

В результате передачи кодовой комбинации из предыдущего примера 3 принята кодовая комбинация С*=101000001001, т.е. произошло искажение 2-го и 5-го разрядов. Обнаружить ошибки.

Решение.

Значения проверок равны:

$$\text{Е}_1 = a_1 \oplus a_3 \oplus a_5 \oplus a_7 \oplus a_9 \oplus a_{11} = 1 \oplus 1 \oplus 0 \oplus 0 \oplus 1 \oplus 0 = 1$$

$$\text{Е}_2 = a_2 \oplus a_3 \oplus a_6 \oplus a_7 \oplus a_{10} \oplus a_{11} = 0 \oplus 1 \oplus 0 \oplus 0 \oplus 0 \oplus 0 = 1$$

$$\text{Е}_3 = a_4 \oplus a_5 \oplus a_6 \oplus a_7 \oplus a_{12} = 0 \oplus 0 \oplus 0 \oplus 0 \oplus 1 = 1$$

$$\text{Е}_4 = a_8 \oplus a_9 \oplus a_{10} \oplus a_{11} \oplus a_{12} = 0 \oplus 1 \oplus 0 \oplus 0 \oplus 1 = 0$$

Тогда контрольное число (синдром) ошибки:

$$S = \text{Е}_4\text{Е}_3\text{Е}_2\text{Е}_1 = 0111_{(2)} = 7_{(10)}$$

Переход из двоичной системы счисления в десятичную:

$$1 \cdot 2^0 + 1 \cdot 2^1 + 1 \cdot 2^2 + 0 \cdot 2^3 = 7.$$

Таким образом, при наличии двукратной ошибки декодирование дает номер разряда с ошибкой в позиции 7, в то время как ошибки произошли во 2-м и 5-м разрядах. В этом случае составляется расширенный код Хэмминга, путем добавления одного проверочного символа.

Контрольные вопросы

1. Какие коды называются помехоустойчивыми?
2. Что называется избыточностью?
3. Как образуются корректирующие коды?
4. Объяснить методику построения кода Хэмминга.
5. Назовите основные параметры кода Хэмминга?
6. Как определить общее число элементов кодовых комбинаций кодов Хэмминга?
7. Как определить число проверочных и информационных элементов кода Хэмминга?
8. Как выбираются номера проверочных позиций кода Хэмминга?
9. По какому закону рассчитывают номера контрольных символов?
10. Объяснить правило четности.
11. Как происходит переход из двоичной системы счисления в десятичную?
12. Объяснить особенности кода Хэмминга.

Список использованных источников

1. Митюхин, А.И. Прикладная теория информации: учебно-методическое пособие / А.И. Митюхин. – Минск: БГУИР, 2018. – 168 с.
2. Фурсов, В.А. Лекции по теории информации: учебное пособие под редакцией Н.А. Кузнецова – Самара: Изд-во Самар. гос. аэрокосм. ун-та, 2006. – 148 с.:
3. Свирид Ю.В. Основы теории информации: курс лекций / Ю.В. Свирид. – Мн.: БГУ, 2003. – 139 с.

Учебное издание

ТЕОРИЯ ИНФОРМАЦИИ

Методические рекомендации

Составитель
САПЕЛКО Татьяна Ивановна

Технический редактор *Г.В. Разбоева*
Компьютерный дизайн *В.Л. Пугач*

Подписано в печать .2022. Формат 60x84 $^{1}/_{16}$. Бумага офсетная.
Усл. печ. л. 2,79. Уч.-изд. л. 1,64. Тираж экз. Заказ .

Издатель и полиграфическое исполнение – учреждение образования
«Витебский государственный университет имени П.М. Машерова».

Свидетельство о государственной регистрации в качестве издателя,
изготовителя, распространителя печатных изданий
№ 1/255 от 31.03.2014.

Отпечатано на ризографе учреждения образования
«Витебский государственный университет имени П.М. Машерова».
210038, г. Витебск, Московский проспект, 33.