Министерство образования и науки РФ
ФГБОУ ВО «Томский государственный университет
систем управления и радиоэлектроники»
Кафедра комплексной информационной безопасности
электронно-вычислительных систем (КИБЭВС)

И.А. Ходашинский, М.Б. Бардамова

ТЕОРИЯ ИНФОРМАЦИИ

Методические указания для выполнения практических и самостоятельных работ
для студентов специальности и направления
10.03.01 – «Информационная безопасность»,
10.05.02 – «Информационная безопасность телекоммуникационных систем»,
10.05.03 – «Информационная безопасность автоматизированных систем»,
10.05.04 – «Информационно-аналитические системы безопасности»

В-Спектр
Томск, 2018

УДК 519.72
ББК 32.811
Х 69

Х 69 **Ходашинский И.А., Бардамова М.Б.** Теория информации: методические указания для выполнения практических и самостоятельных работ. – Томск, В-Спектр, 2018. – 64 с.
ISBN 978-5-91191-394-6

Практикум содержит описания практических и самостоятельных работ по дисциплине «Теория информации» для специальностей 10.05.02 – «Информационная безопасность телекоммуникационных систем», 10.05.03 – «Информационная безопасность автоматизированных систем», 10.05.04 – «Информационно-аналитические системы безопасности» и направления 10.03.01 – «Информационная безопасность», теоретический материал, практические задания, методические указания по выполнению заданий, вопросы для самоконтроля.

УДК 519.72
ББК 32.811

ISBN 978-5-91191-394-6

Содержание

ВВЕДЕНИЕ

Понятие «информация» наряду с понятиями материи и энергии является базисом в естественных науках.

Под *информацией* понимают совокупность сведений о каких-либо событиях, процессах, явлениях и т.п., рассматриваемых в разных аспектах. Термин широко используется в лингвистике, психологии, биологии и других науках. Однако в разных областях знаний в него вкладывают разный смысл. Разнообразие информационных процессов и широкий интерес к ним в разных областях знаний породили много толкований определений понятия «информация».

Аспекты информации, связанные со смыслом высказываний, называются ее *семантической частью*, а аспекты, связанные с эффектом, производимым информацией, называются ее *прагматической частью*. Семантические аспекты информации являются в своей основе предметом изучения в философии и ее составной части – логике, а прагматические аспекты – в психологии, медицине и биологии, поскольку, например, одни и те же сведения могут вызвать у разных людей различные, иногда противоположные, состояния.

Но для того, чтобы проанализировать смысл высказываний или оценить эффект, ими произведенный, необходима некоторая система знаков, например, естественный или искусственный язык. Данная система знаков имеет свои внутренние свойства, составляющие *синтаксическую часть* информации. Именно эта часть информации, без учета смысла или эффекта, изучается в математической дисциплине, называемой *теорией информации*.

Теория информации рассматривает две задачи:

– передачу информации, т.е. пересылку из одного места в другое;

– хранение информации – т.е. пересылку от одного момента времени до другого.

Теория информации, ориентированная на формализованное описание сообщений, процессов их формирования, передачи и приема, развивалась особенно активно во второй половине XX в., начиная с исследований Р. Хартли («Передача информации», 1928 г.) и К. Шеннона («Математическая теория связи», 1948 г., «Теория связи в секретных системах», 1949 г.). Создание теории информации позволило решить основные теоретические проблемы и создать эффективные системы передачи сигналов и хранения информации.

ТЕМА 1

ПРЕДВАРИТЕЛЬНЫЕ МАТЕМАТИЧЕСКИЕ СВЕДЕНИЯ

Цель занятия – повторение необходимых сведений теории вероятностей и комбинаторики.

1.1. Основные понятия теории вероятностей

Вероятностью события A называют отношение числа благоприятствующих этому событию исходов к общему числу всех равновозможных несовместных элементарных исходов, образующих полную группу [1]. Вероятность события A определяется формулой

$$P(A) = m/n,$$

где m – число элементарных исходов, благоприятствующих A, n – число всех элементарных исходов испытания.

Из определения вероятности вытекают следующие ее свойства:

1) вероятность достоверного события равна 1;

2) вероятность невозможного события равна 0;

3) вероятность случайного события есть положительное число из интервала [0, 1].

Вероятность появления одного из нескольких попарно несовместимых событий, безразлично какого, равна сумме вероятностей этих событий:

$$P(A_1+A_2+\ldots+A_n) = P(A_1) + P(A_2) +\ldots+ P(A_n).$$

Сумма вероятностей событий A_1, A_2, …, A_n, образующих полную группу, равна единице:

$$P(A_1) + P(A_2) +\ldots+ P(A_n) = 1.$$

Сумма вероятностей противоположных событий равна единице:

$$P(A) + P(\bar{A}) = 1.$$

Произведением двух событий A и B называют событие AB, состоящее в совместном появлении этих событий.

Условной вероятностью $P(B|A)$ называют вероятность события B, вычисленную в предположении, что событие A уже наступило.

Пример. В урне 3 белых и 3 черных шара. Из урны дважды вынимают по одному шару, не возвращая их обратно. Найти вероятность появления белого шара при втором испытании (B), если при первом испытании был извлечен черный шар (A).

Решение. После первого испытания в урне осталось 5 шаров, из них 3 белых. Условная вероятность появления белого шара во втором испытании:

$$P(B|A) = 3/5.$$

Условная вероятность события B при условии, что событие A уже наступило и $P(A) > 0$, равна:

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

Вероятность совместного появления двух событий равна произведению вероятности одного из них на условную вероятность другого, вычисленную в предположении, что первое событие уже наступило:

$$P(AB) = P(A)\,P(B|A).$$

Вероятность появления хотя бы одного из событий $A_1, A_2, \ldots, A_n$, независимых в совокупности, равна разности между единицей и произведением вероятностей противоположных событий $\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_n$.

1.2. Основные понятия комбинаторики

Число различных подмножеств n-элементного множества равно 2^n.

Перестановками называют комбинации, состоящие из одних и тех же различных элементов и отличающиеся только порядком их расположения. Число всех возможных перестановок n-элементного множества

$$P_n = n! = 1\cdot 2\cdot 3\cdot \ldots \cdot n.$$

Пример. Сколькими разными способами можно расположить четыре книжки на книжной полке?

Решение. Искомое число является числом способов упорядочения множества из четырех элементов, $P_4 = 4! = 1\cdot 2\cdot 3\cdot 4 = 24$.

Размещением называют комбинации, составленные из n различных элементов по m элементов, которые отличаются либо составом элементов, либо их порядком. Число всех возможных размещений

$$A_n^m = n(n-1)...(n-m+1) = \frac{n!}{(n-m)!}.$$

Пример. Студенту необходимо сдать два экзамена на протяжении четырех дней. Сколькими способами это можно сделать?

Решение.

$$A_4^2 = \frac{4!}{2!} = 3\cdot 4 = 12.$$

Сочетаниями называют комбинации, составленные из n различных элементов по m элементов, которые отличаются хотя бы одним элементом. Число сочетаний

$$C_n^m = \frac{n!}{m!(n-m)!} = \frac{n(n-1)(n-2)...(n-m+1)}{m!}.$$

Пример. Сборная команда по биатлону состоит из 9 человек.

Сколько разных вариантов должен рассмотреть тренер перед формированием эстафетной гонки (четыре человека)?

Решение.

$$C_9^4 = \frac{9!}{4!5!} = \frac{9\cdot 8\cdot 7\cdot 6}{4\cdot 3\cdot 2\cdot 1} = 126.$$

Пусть $k_1, k_2, \ldots, k_m$ – натуральные числа, такие что $k_1 + k_2 + \ldots + k_m = n$. Количество способов построения разбиения n-элементного множества на классы $A_1, A_2, \ldots, A_m$, число элементов которых соответственно $k_1, k_2, \ldots, k_m$, равно

$$C_n(k_1, k_2, \ldots, k_m) = \frac{n!}{k_1! k_2! \ldots k_m!}.$$

Пример. Сколько различных слов можно построить перестановкой букв в слове «банан»?

Решение. В слово «банан» входит одна буква «б» и по две буквы «а» и «н». Общее число букв в слове равно пяти. Ответ следующий:

$$C_5(1, 2, 2) = \frac{5!}{1!2!\,2!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{1 \cdot 2 \cdot 2} = 30.$$

Количество различных упорядоченных k-элементных подмножеств n-элементного множества, все элементы которого не обязательно различны, равно n^k [2].

1.3. Аудиторные задания

1. Сколько различных слов длиною в восемь букв можно составить из алфавита {0, 1}?

2. Сколько различных слов длиною в восемь букв можно построить из букв «0» и «1», таких, что количество букв «0» не превышает четырех?

3. Внутрь круга наудачу брошена точка. Найти вероятность того, что точка окажется внутри вписанного в круг квадрата.

4. Вероятность того, что при одном выстреле стрелок попадает в цель, равна 0,6. Сколько выстрелов должен сделать стрелок, чтобы с вероятностью не менее 0,95 он попал в цель хотя бы один раз?

5. Куб, все грани которого окрашены, распилен на 64 кубика одинакового размера, которые затем перемешаны. Найти вероятность того, что наудачу извлеченный кубик будет иметь окрашенных граней: а) одну, б) две, в) три; не будет иметь окрашенных граней.

6. На каждой из шести одинаковых карточек напечатана одна из следующих букв: б, и, о, р, т. Карточки перемешаны. Найти: а) вероятность того, что на трех, вынутых по одной и расположенных по порядку карточках можно будет прочесть слово «рот»; б) вероятность того, что на четырех, вынутых по одной и расположенных по порядку карточках можно будет прочесть слово «борт»; в) вероятность того, что на пяти, вынутых по одной и расположенных по порядку карточках можно будет прочесть слово «орбит».

7. Алфавит сообщения состоит из двух букв «0» и «1», появляющихся с вероятностями $P(0)$ и $P(1)$.

Сколько разных последовательностей длиной K можно составить в данном алфавите?

Какова вероятность появления последовательности, в которой подряд идут N_1 букв «0» и N_2 букв «1» ($N_1 + N_2 = K$)?

Какова вероятность появления последовательности, в которой произвольно встречаются N_1 букв «0» и N_2 букв «1» ($N_1 + N_2 = K$)?

1.4. Самостоятельная работа

1. Чему равна вероятность того, что при бросании трех игральных костей 6 очков появится хотя бы на одной из костей?

2. В турнире принимали участие 10 команд. Каждые две команды сыграли между собой только один раз. Сколько игр было сыграно в турнире?

3. Пусть подбрасываются N монет. Определить вероятность результата, состоящего в выпадении «N_G гербов и N_R решеток».

4. Пусть подбрасываются N игральных костей. Определить вероятность результата, состоящего в выпадении «N_1 единиц, N_2 двоек, N_3 троек, N_4 четверок, N_5 пятерок, N_4 шестерок».

5. Алфавит сообщения состоит из двух букв 0 и 1, появляющихся с вероятностями $P(0)$ и $P(1)$. Какова вероятность появления последовательности, в которой произвольно встречаются N_1 букв «0» и N_2 букв «1» ($N_1 + N_2 < K$)?

1.5. Контрольные вопросы

1. Приведите примеры зависимых и независимых случайных событий.
2. Что называется:
   а) перестановкой n-элементного множества,
   б) сочетанием из n элементов по m элементов?
3. Каково основное свойство полной группы событий?
4. Дайте определение условной вероятности.
5. Как определяется вероятность появления хотя бы одного из множества событий?
6. Приведите пример геометрической интерпретации вероятности.

ТЕМА 2

МЕРА КОЛИЧЕСТВА ИНФОРМАЦИИ

Цель занятия – систематизация, углубление, закрепление полученных теоретических знаний о мере информации.

2.1. Постулаты теории информации

Исследования теории информации основаны на четко сформулированных постулатах.

1. Источник сообщения осуществляет выбор сообщения из некоторого множества с определенной вероятностью.

2. Сообщения могут передаваться по каналу связи в закодированном виде. Кодированные сообщения образуют множество, являющееся взаимно однозначным отображением множества сообщений. Правило декодирования известно декодеру.

3. Сообщения следуют друг за другом, причем число сообщений может быть сколь угодно большим.

4. Сообщение считается принятым верно, если в результате декодирования оно может быть в точности восстановлено. При этом не учитывается, сколько времени прошло с момента передачи сообщения до момента окончания декодирования, и какова сложность операций кодирования и декодирования.

5. Количество информации не зависит от смыслового содержания сообщения, от его эмоционального воздействия, полезности и даже от его отношения к реальной действительности.

2.2. Энтропия как мера неопределенности выбора

Количество информации в сообщении о некотором событии существенно зависит от вероятности этого события. В основу определения меры количества информации положен вероятностный подход.

Количество информации в отдельно взятом сообщении определяется следующим образом:

$$I(a) = \log_k \frac{1}{P(a)} = -\log_k P(a) .$$

Логарифмическая мера обладает естественным в данном случае свойством аддитивности, согласно которому количество информации, содержащееся в нескольких независимых сообщениях, равно сумме количества информации в каждом из них. Так как совместная вероятность n независимых сообщений

$$P(a_1, a_2, \ldots, a_n) = P(a_1)P(a_2)\ldots P(a_n),$$

то количество информации в этих сообщениях равно

$$I(a_1, a_2, \ldots, a_n) = -\log_k P(a_1, a_2, \ldots, a_n) = -\sum_{i=1}^{n} \log_k P(a_i) = \sum_{i=1}^{n} I(a_i).$$

Основание логарифма k может быть любым.

Если все сообщения равновероятны:

$$P(a_1) = P(a_2) = \ldots = P(a_m) = P(a) = \frac{1}{m},$$

то количество информации в каждом из них определяется величиной

$$I(a) = -\log P(a) = \log m.$$

Среднее количество информации, приходящееся на одно сообщение, в общем случае определяется как математическое ожидание $I(a_i)$:

$$H(a) = \overline{I(a_i)} = \sum_{i=1}^{m} P(a_i) \cdot I(a_i) = -\sum_{i=1}^{m} P(a_i) \cdot \log P(a_i)$$

Величина $H(a)$ называется энтропией.

В теории информации энтропия $H(a)$ характеризует неопределенность ситуации до передачи сообщения, поскольку заранее неизвестно, какое из сообщений будет передано. Важно понимать, что чем больше энтропия, тем сильнее неопределенность и тем большую информацию в среднем несет одно сообщение источника.

2.3. Аксиомы Хинчина. Аксиомы Фаддеева

А.Я. Хинчин установил, что энтропия конечной вероятностной схемы однозначно определяется перечисленной ниже системой аксиом

1) $H(p_1, p_2, \ldots, p_m)$ непрерывна относительно $p_1, p_2, \ldots, p_m$ в области

$$0 \le p_i \le 1, \ \sum_{i=1}^{m} p_i = 1.$$

2) $H(p_1, p_2, \ldots, p_m)$ симметрична относительно $p_1, p_2, \ldots, p_m$.

3) $H(p_1, p_2, \ldots, p_m, 0) = H(p_1, p_2, \ldots, p_m)$. Это означает, что добавление к множеству состояний невозможного состояния не изменяет неопределенность.

4) $H(p_{11}, p_{12}, \ldots, p_{1n}; p_{21}, p_{22}, \ldots, p_{2n}; p_{m1}, p_{m2}, \ldots, p_{mn}) = H(p_1, p_2, \ldots, p_m) + \sum_{i=1}^{m} p_i H\left(\frac{p_{i1}}{p_i} + \ldots + \frac{p_{in}}{p_i}\right)$,

где p_{ij} – нормированное совместное распределение статистически зависимых объектов с

$$\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} = 1,\ p_i = \sum_{j=1}^{n} p_{ij}\ (p_{ij} \geq 0)\ ;$$

5) $H(p_1, p_2, \ldots, p_m) \leq H(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m})$.

Это означает, что энтропия имеет наибольшее значение при $p_1 = p_2 = \ldots = p_m = 1/m$.

Д.К. Фаддеев упростил систему аксиом Хинчина и предложил следующие аксиомы:

1. $H(p_1, p_2) = H(1 - p, p)$ непрерывна при $0 \leq p \leq 1$ и положительна хотя бы в одной точке.

2. $H(p_1, p_2, \ldots, p_m)$ симметрична относительно $p_1, p_2, \ldots, p_m$.

3. При $m \geq 2$

$$H(p_1, p_2, \ldots, p_{m-1}, p_m, p_{m+1}) = H(p_1, p_2, \ldots, p_{m-1}, p_m + p_{m+1}) +$$

$$+ (p_m + p_{m+1}) H(\frac{p_m}{p_m + p_{m+1}}, \frac{p_{m+1}}{p_m + p_{m+1}}),\ p_m + p_{m+1} > 0.$$

Переход к общему случаю с $m > 2$ осуществляется методом математической индукции на основании аксиомы 3.

Разница в этих двух системах аксиом заключается в следующем. Во-первых, аксиома 5 (экстремальность) в системе Хинчина заменяется требованием положительности энтропии в одной точке и, во-вторых, аксиомы 3 и 4 заменяются одной аксиомой 3 системы Фаддеева, очень естественной, если рассмотреть энтропию как меру неопределенности состояний случайного объекта.

2.4. Аудиторные задания

1. Определить энтропию сообщения из шести букв, если общее число букв в алфавите равно 32 и а) появление всех букв равновероятно, б) три буквы появляются в два раза чаще трех остальных.

2. Опыт $X = \{x_1, x_2, x_3\}$ имеет три исхода со следующими вероятностями: $p(x_1) = 0{,}25$, $p(x_2) = 0{,}35$, $p(x_3) = 0{,}4$. Определить количество информации, которое несут каждый из исходов по отдельности и среднее количество.

3. Вычислить энтропии следующих систем.

$$X_1 = \begin{vmatrix} x_1 & x_2 \\ 1/256 & 255/256 \end{vmatrix},\ X_2 = \begin{vmatrix} x_1 & x_2 \\ 1/2 & 1/2 \end{vmatrix},\ X_3 = \begin{vmatrix} x_1 & x_2 \\ 1/4 & 3/4 \end{vmatrix},\ Y = \begin{vmatrix} y_1 & y_2 & y_3 \\ 5/16 & 4/16 & 7/16 \end{vmatrix}.$$

Провести сравнительный анализ.

Для вероятностной схемы Y проверить выполнимость третьей аксиомы Фаддеева.

4. Проверить свойство аддитивности энтропии на следующей системе

$$Y = \begin{vmatrix} y_1 & y_2 & y_3 & y_4 & y_5 \\ 1/2 & 1/4 & 1/8 & 1/16 & 1/16 \end{vmatrix}.$$

5. Имеются три урны, содержащие шары белого, синего и красного цветов. Состав шаров в этих урнах следующий:

– 23 белых шара, 0 синих и 0 красных шаров;

– 11 белых, 11 синих и 1 красный шар;

– 3 белых, 3 синих и 1 красный шар.

Сравнить значения энтропии соответствующих вероятностных схем, реализующихся при извлечении одного шара из каждой урны.

6. Имеются две системы:

$$X = \begin{vmatrix} x_1 & x_2 \\ p_1 & p_2 \end{vmatrix}, Y = \begin{vmatrix} y_1 & y_2 & y_3 \\ q_1 & q_2 & q_3 \end{vmatrix}.$$

Определить, какая система обладает большей неопределенностью в случае, если

a) $p_1 = p_2$, $q_1 = q_2 = q_3$,

b) $p_1 = q_1$, $p_2 = q_2 + q_3$, $q_1 = q_2 = q_3$.

2.5. Самостоятельная работа

Определить значение вероятностей, при которых энтропия достигает максимального значения для вероятностных схем с двумя, тремя и четырьмя элементами.

2.6. Контрольные вопросы

1. Что называется энтропией?
2. Что характеризует информационная энтропия?
3. В каких единицах измеряется энтропия?
4. Назовите свойства энтропии.

ТЕМА 3

ЭНТРОПИЯ И ИНФОРМАЦИЯ СЛОЖНЫХ СИСТЕМ

Цель занятия – систематизация, углубление, закрепление полученных теоретических знаний об объединенной вероятностной системе и условной энтропии.

3.1. Условная энтропия

Рассмотрим две вероятностные схемы X и Y:

$$X = \begin{pmatrix} x_1, & x_2, & ..., & x_r \\ P(x_1), & P(x_2), & ..., & P(x_r) \end{pmatrix},$$

$$Y = \begin{pmatrix} y_1, & y_2, & ..., & y_s \\ P(y_1), & P(y_2), & ..., & P(y_s) \end{pmatrix},$$

которые определяются не только собственными вероятностями $P(x_i)$ и $P(y_j)$, но и условными вероятностями $P(y_j|x_i)$, $P(x_i|y_j)$, где $i = 1, 2, ..., r$; $j = 1, 2, ..., s$.

Условным распределением составляющей X при $Y = y_j$ (y_j сохраняет одно и то же значение при всех возможных значениях X) называют совокупность условных вероятностей $P(x_1|y_j)$, $P(x_2|y_j)$, ..., $P(x_r|y_j)$.

Аналогично определяется условное распределение составляющей Y.

Условные вероятности составляющих X и Y вычисляют соответственно по формулам:

$$P(y_j \mid x_i) = \frac{P(x_i, y_j)}{P(x_i)}, \; P(x_i \mid y_j) = \frac{P(x_i, y_j)}{P(y_j)}.$$

Так как условная вероятность события y_j при условии выполнения события x_i принимается по определению

$$P(y_j \mid x_i) = \frac{P(x_i, y_j)}{P(x_i)},$$

то вероятность совместного появления совокупности событий

$$P(x_i, y_j) = P(x_i) P(y_j \mid x_i).$$

Аналогично, совместная вероятность события x_i при условии выполнения события y_j:

$$P(x_i, y_j) = P(y_j) P(x_i \mid y_j).$$

Тогда энтропию объединенных вероятностных схем X и Y или совместную энтропию определяют по формуле Шеннона:

$$H(X,Y)=-\sum_{i=1}^{r}\sum_{j=1}^{s}P(x_i,y_j)\log P(x_i,y_j)=$$

$$=-\sum_{i=1}^{r}P(x_i)\log P(x_i)-\sum_{i=1}^{r}P(x_i)\sum_{j=1}^{s}P(y_j|x_i)\log P(y_j|x_i)=H(X)+H(Y|X),$$

где $H(X)$ – энтропия вероятностной схемы X; $H(Y|X)$ – условная энтропия схемы Y при условии, что сообщение схемы X известны:

$$H(Y|X)=-\sum_{i=1}^{r}P(x_i)\sum_{j=1}^{s}P(y_j|x_i)\log P(y_j|x_i).$$

Выражая $P(x_i,y_j)=P(y_j)P(x_i|y_j)$ через другую условную вероятность, получим

$$H(X,Y)=H(Y)+H(X|Y),$$

где $H(Y)$ энтропия вероятностной схемы Y; $H(X|Y)$ – условная энтропия схемы X при условии, что сообщение схемы Y известны:

$$H(X|Y)=-\sum_{j=1}^{s}P(y_j)\sum_{i=1}^{r}P(x_i|y_j)\log P(x_i|y_j).$$

Таким образом, совместная энтропия двух сообщений равна сумме безусловной энтропии одного сообщения и условной энтропии второго сообщения.

Обобщая полученные выше результаты, совместную энтропию n вероятностных схем можно вычислить следующим образом:

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_{n-1}, \ldots, X_2, X_1).$$

Для *независимых* событий X и Y:

$$P(y_j|x_i) = P(y_j),$$

$$H(X,Y)=-\sum_{i=1}^{r}\sum_{j=1}^{s}P(x_i,y_j)\log P(x_i,y_j)=$$

$$=-\sum_{i=1}^{r}P(x_i)\log P(x_i)-\sum_{j=1}^{s}P(y_j)\log P(y_j)=H(X)+H(Y)$$

или для n вероятностных схем:

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2) + \ldots + H(X_n).$$

Если X и Y полностью зависимы, т.е. при появлении x_i неизбежно следует y_j, то $P(x_i,y_j)$ равна единице при $i = j$ и нулю при $i \neq j$. Поэтому $H(Y|X) = 0$, и, следовательно, $H(X,Y) = H(X)$, т.е. при полной зависимости двух ансамблей один из них не вносит никакой информации.

Между энтропией объединенной вероятностной схемы и энтропиями составляющих схем существует соотношение:

$$H(X, Y) \le H(Y) + H(X),$$

равенство имеет место тогда и только тогда, когда X и Y независимы.

Условная энтропия всегда меньше или равна безусловной:

$$H(X|Y) \le H(X),$$
$$H(Y|X) \le H(Y).$$

3.2. Взаимная информация

Рассмотрим две вероятностные схемы X и Y:

$$X = \begin{pmatrix} x_1, & x_2, & ..., & x_r \\ P(x_1), & P(x_2), & ..., & P(x_r) \end{pmatrix},$$

$$Y = \begin{pmatrix} y_1, & y_2, & ..., & y_s \\ P(y_1), & P(y_2), & ..., & P(y_s) \end{pmatrix}.$$

Исследуем вопрос об измерении количества информации, содержащейся в исходе x_i при условии, что исход y_j, принадлежащий Y, реализовался. Естественно, что появление исхода y_j изменяет безусловную априорную вероятность появления исхода x_i до апостериорной условной вероятности $P(x_i|y_j)$. В качестве количественной меры такого изменения примем величину

$$I(x_i; y_j) = \log \frac{P(x_i \mid y_j)}{P(x_i)}.$$

Если реализация исхода y_j не влияет на вероятность появления исхода x_i, т.е. если эти исходы независимы, то $P(x_i|y_j) = P(x_i)$, и введённое количество информации, как следует из приведенной выше формулы, будет равно нулю.

Можно рассматривать количество информации, получаемое при реализации исхода y_j, если известно, что исход x_i имеет место. Аналогично получаем:

$$I(y_j; x_i) = \log \frac{P(y_j \mid x_i)}{P(y_j)}.$$

Умножая на $P(y_j)$ числитель и знаменатель под логарифмом в формуле

$$I(x_i; y_j) = \log \frac{P(x_i \mid y_j)}{P(x_i)},$$

и учитывая равенства

$$P(x_i \mid y_j) P(y_j) = P(x_i, y_j) = P(y_j \mid x_i) P(x_i),$$

получим:

$$I(x_i;y_j)=\log\frac{P(x_i\mid y_j)}{P(x_i)}=\log\frac{P(x_i\mid y_j)P(y_j)}{P(x_i)P(y_j)}=\log\frac{P(x_i,y_j)}{P(x_i)P(y_j)}=$$

$$\log\frac{P(y_j\mid x_i)P(x_i)}{P(x_i)P(y_j)}=\log\frac{P(y_j\mid x_i)}{P(y_j)}=I(y_j;x_i).$$

Отсюда следует, что y_j несет об x_i такое же количество информации, какое x_i несет об y_j (свойство симметрии). Поэтому величина $I(x_i; y_j)$ называется взаимной информацией исходов x_i и y_j. Взаимная информация $I(x_i; y_j)$ может принимать положительные значения ($P(x_i|y_j) > P(x_i)$), отрицательные ($P(x_i|y_j) < P(x_i)$) и быть равной нулю ($P(x_i|y_j) = P(x_i)$).

Поскольку сложная система случайным образом приходит в то или иное состояние, определяемое парой чисел (x_i, y_j), то $I(x_i; y_j)$ будет случайной величиной, которую можно усреднить по всему множеству состояний. Для среднего этой случайной величины введено обозначение:

$$I(X;Y)=\overline{I(x_i;y_j)}=\sum_{i=1}^{r}\sum_{j=1}^{s}P(x_i,y_j)\log\frac{P(x_i\mid y_j)}{P(x_i)}.$$

Величина $I(X; Y)$ называется средней взаимной информацией вероятностных схем X и Y.

Важным свойством средней взаимной информации является свойство выпуклости: средняя взаимная информация не имеет локальных минимумов или седловых точек.

Другим свойством взаимной информации является симметричность. Взаимная информация симметрична относительно пары вероятностных схем: $I(A; B) = I(B; A)$.

Если сообщение A и B – независимы, т.е. не совместны, то взаимная информация $I(A; B) = 0$.

3.3. Аудиторные задания

1. Даны $H(X)$ и $H(Y)$, $H(Y|X) = 1$, найти $H(X|Y)$.

2. Элементы алфавита X и Y статистически связаны. Известно, что $H(X)$ =8 бит, $H(Y) = 12$ бит. В каких пределах меняется условная энтропия $H(Y|X)$ при изменении $H(X|Y)$ в максимально возможных пределах.

3. Ансамбли событий X (x_1, x_2, x_3) и Y (y_1, y_2) объединены. Вероятности совместных событий приведены ниже.

y_j	x_i		
	x_1	x_2	x_3
y_1	0,1	0,2	0,3
y_2	0,25	0	0,15

Определить:

а) энтропию ансамблей X и Y;

б) энтропию объединенного ансамбля X,Y; в) условные энтропии ансамблей;

г) существование статистической связи между двумя ансамблями.

4. Система описывается следующей матрицей

$$P(X,Y)=\begin{vmatrix} 1/8 & 1/8 & 1/8 \\ 1/8 & 0 & 1/8 \\ 1/8 & 1/8 & 1/8 \end{vmatrix}.$$

Определить энтропии $H(X)$, $H(Y)$, $H(X|Y)$, $H(Y|X)$, $H(X|y_3)$, $H(Y|x_2)$, $H(X, Y)$.

5. Ракеты двух пусковых установок используются для поражения двух целей. Ракета, пущенная с первой установки, поражает цель номер 1 с вероятностью 0,5; цель номер 2 – с вероятностью 0,3 и дает промах с вероятностью 0,2. Ракета второй установки поражает первую цель с вероятностью 0,3; вторую – вероятностью 0,5 и вероятность промаха равна 0,2. Вероятность выбора первой установки равна 0,4.

Чему равна неопределенность выбора установки, если известно, что а) поражена вторая цель; б) произошел промах.

Какова неопределенность исхода, если пущена любая ракета?

6. Ансамбли событий X (x_1, x_2, x_3) и Y (y_1, y_2, y_3) объединены. Вероятности совместных событий приведены ниже. Определить среднюю взаимную информацию вероятностных схем X и Y.

$P(x_i, y_j)$	x_1	x_2	x_3
y_1	0,06	0,1	0,04
y_2	0,24	0,01	0,35
y_3	0,02	0,15	0,03

3.4. Самостоятельная работа

1. Задана матрица вероятностей состояний системы, объединяющей источники X и Y.

y_j	x_i		
	x_1	x_2	x_3
y_1	0,4	0,1	0
y_2	0	0,2	0,1
y_3	0	0	0,2

Определить:

а) энтропию ансамблей X и Y;

б) энтропию объединенного ансамбля X,Y;

в) условные энтропии ансамблей;

г) существование статистической связи между двумя ансамблями.

2. Определить среднюю взаимную информацию вероятностных схем X и Y.

$P(x_i, y_j)$	x_1	x_2
y_1	0,12	0,08
y_2	0,48	0,32

3.5. Контрольные вопросы

1. Что называется условной энтропией?
2. Что характеризует условная энтропия?
3. В каких единицах измеряется условная энтропия?
4. Назовите свойства условной энтропии.
5. Какие значения может принимать взаимная информация?

ТЕМА 4

ИСТОЧНИКИ ДИСКРЕТНЫХ СООБЩЕНИЙ

Цель занятия – систематизация, углубление, закрепление полученных теоретических знаний о марковских источниках сообщений и информационных характеристиках источника дискретных сообщений.

4.1. Марковские источники

В классе дискретных источников с памятью особое место занимают марковские источники. Их описания базируются на математическом аппарате марковских цепей, с помощью которых можно составить математическую модель многих процессов, наиболее близких к практике. Значение марковских цепей в теории информации заключается в том, что реальные источники информации вырабатывают сообщения при наличии *статистической зависимости* между отдельными событиями, символами.

В реальных источниках вероятность выбора какого-либо очередного символа зависит от предшествующих символов.

Марковским процессом называется случайный процесс, который можно полностью задать с помощью двух величин: вероятности $P(x, t)$ того, что случайная величина $x(t)$ в момент времени t равна x, и вероятности $P(x_2, t_2|x_1, t_1)$ того, что если x при $t = t_1$ равен x_1, то при $t = t_2$ он равен x_2. Вторая из этих величин называется вероятностью перехода из состояния x_1 при $t = t_1$ в состояние x_2 при $t = t_2$.

Дискретный по времени и состояниям марковский процесс называется *марковской цепью*. Цепь Маркова порядка n характеризует последовательность событий, вероятности которых зависят от того, какие n событий предшествовали данному. Эти n конкретных событий определяют состояние источника, в котором он находится при выдаче очередного знака. При объеме алфавита знаков l число R различных состояний источника не превышает l^n. Обозначим эти состояния через
$S_1, S_2,\ldots, S_q,\ldots, S_R$, а вероятности выбора в состоянии S_q, знака x_i – через $P_q(x_i)$. При определении вероятности $P_q(x_i)$ естественно предположить, что к моменту выдачи источником очередного знака известны все знаки, созданные им ранее, а, следовательно, и то, в каком состоянии находится источник.

Если источник находится в состоянии S_q, его частная энтропия $H(S_q)$ определяется следующим соотношением:

$$H(S_q)=\sum_{i=1}^{l} P_q(x_i)\log P_q(x_i).$$

Усредняя случайную величину $H(S_q)$ по всем возможным состояниям $q = 1, \ldots, R$, получаем энтропию марковского источника сообщений:

$$H(X) = \sum_{q=1}^{R} P(S_q) \sum_{i=1}^{l} P_q(x_i) \log P_q(x_i).$$

где $P(S_q)$ – вероятность того, что источник сообщений находится в состоянии S_q.

Величина $H(X)$ характеризует неопределенность, приходящуюся в среднем на один знак, выдаваемый источником сообщений.

Определим энтропию источника сообщений для нескольких частных случаев. Если статистические связи между знаками полностью отсутствуют, то после выбора источником знака x_i его состояние не меняется ($R = 1$). Следовательно, $P(S_1) = 1$, и для энтропии источника сообщений справедливо выражение:

$$H(X) = \sum_{i=1}^{l} P(x_i) \log P(x_i).$$

Когда корреляционные связи наблюдаются только между двумя знаками (простая цепь Маркова), максимальное число различных состояний источника равно объему алфавита. Следовательно, $R = l$ и $P_q(x_i) = P(x_i|x_q)$, где $q = 1, \ldots, l$. При этом

$$H(X) = \sum_{q=1}^{R} P(x_q) \sum_{i=1}^{l} P(x_i \,|\, x_q) \log P(x_i \,|\, x_q).$$

При наличии корреляционной связи между тремя знаками состояния источника определяются двумя предшествующими знаками. Поэтому для произвольного состояния источника S_q удобно дать обозначение с двумя индексами $S_{k,h}$, где $k = 1, \ldots, l$ и $h = 1, \ldots, l$. Тогда

$$P(S_q) = P(S_{k,h}) = P(x_k, x_h) \text{ и } P_q(x_i) = P(x_i \,|\, x_k x_h)$$

Энтропия такого источника сообщений будет вычислена следующим образом:

$$H(X) = -\sum_{k=1}^{l} \sum_{h=1}^{l} P(x_k, x_h) \sum_{i=1}^{l} P(x_i \,|\, x_k x_h) \log P(x_i \,|\, x_k x_h).$$

Аналогично можно получить выражения для энтропии источника сообщений и при более протяженной корреляционной связи между знаками.

4.2. Информационные характеристики источника дискретных сообщений

Пусть энтропии двух источников сообщений $H_1 < H_2$, а количество информации, получаемое от них одинаковое, т.е. $I = n_1 H_1 = n_2 H_2$, где n_1 и n_2 – длины сообщения от первого и второго источников. Обозначим

$$\mu = \frac{n_2}{n_1} = \frac{H_1}{H_2}.$$

При передаче одинакового количества информации сообщение тем длиннее, чем меньше его энтропия. Величина μ, называемая *коэффициентом сжатия*, характеризует степень укорочения сообщения при переходе к кодированию состояний элементов, характеризующихся большей энтропией.

При этом доля излишних элементов оценивается *коэффициентом избыточности*:

$$r = \frac{H_0 - H_1}{H_0} = 1 - \frac{H_1}{H_0} = 1 - \mu.$$

В русском алфавите, содержащем 32 элемента, при одинаковых вероятностях появления всех элементов алфавита, неопределенность, приходящаяся на один элемент, составляет $H_0 = \log 32 = 5$ бит.

Анализ показывает, что с учетом неравномерного появления различных букв алфавита $H_1 = 4{,}39$ бит, а с учетом зависимости двухбуквенных сочетаний $H_2 = 3{,}52$ бит, с учетом корреляции между тремя символами $H_3 = 3{,}05$ бит, а с учетом корреляции между восемью символами $H_8 = 2$ бит и дальше остается неизменной. Следовательно, избыточность русского языка составляет

$$r = \frac{5-2}{5} = 0{,}6.$$

Под *производительностью источника сообщений* подразумевают количество информации, вырабатываемое источником в единицу времени. Эту характеристику источника называют также скоростью создания сообщений или потоком входной информации. Поскольку возможное воздействие помех на источник сообщений принято учитывать эквивалентным изменению характеристик модели канала связи, то производительность источника сообщений равна энтропии источника, приходящейся на единицу времени.

Длительность выдачи знаков источником в каждом из состояний в общем случае может быть различной. Обозначим длительность выдачи знака x_i, формируемого источником в состоянии S_q, через τ_{qxi}. Тогда средняя длительность выдачи источником одного знака

$$\tau_u = \sum_{q=1}^{R} P(S_q) \sum_{i=1}^{l} P_q(x_i) \tau_{qxi},$$

Производительность источника $\bar{I}(X)$ теперь можно выразить формулой

$$\bar{I}(X) = H(X) / \tau_u.$$

Как следует из приведенных выше формул, повышение производительности источника возможно не только за счет увеличения энтропии, но и за счет снижения средней длительности формирования знака. Длительности выдачи знаков желательно выбирать обратно пропорциональными вероятностям их появления.

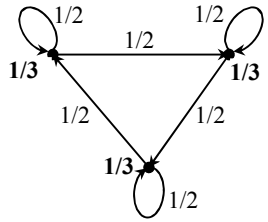
Если длительность выдачи знака не зависит от состояния источника, для всех знаков одинакова и равна τ, то $\tau_u = \tau$. Выражение для $\bar{I}(X)$ принимает вид

$$\bar{I}(X) = H(X) / \tau.$$

Наибольшая производительность источника в этом случае достигается при максимальной энтропии.

4.3. Аудиторные задания

1. Найти энтропию источника, описываемого графом вероятностей перехода.



2. Алфавит источника сообщений состоит из трех букв {1, 2, 3}, вероятности $P(1) = 0{,}4$, $P(2) = 0{,}5$, $P(1) = 0{,}1$. Вероятности появления пар букв приведены в таблице.

x_i, x_j	x_1, x_1	x_1, x_2	x_1, x_3	x_2, x_1	x_2, x_2	x_2, x_3	x_3, x_1	x_3, x_2	x_3, x_3
$P(x_i, x_j)$	0,1	0,2	0,1	0,2	0,3	0	0,1	0	0

Определить энтропию а) заданного источника, б) источника, у которого отсутствуют связи между буквами, в) источника с независимыми появлениями равновероятных символов. Провести сравнительный анализ.

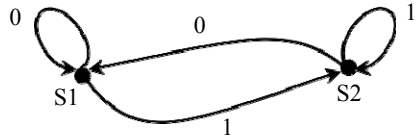
3. Найти энтропию источника сообщений, если вероятности появления сообщений x_1, x_2, x_3, x_4 равны $P(x_1) = 1/2$, $P(x_2) = 1/4$, $P(x_3) = P(x_4) = 1/8$. Между сообщениями имеются корреляционные связи, описанные в таблице.

x_i, x_j	$P(x_i \mid x_j)$	x_i, x_j	$P(x_i \mid x_j)$	x_i, x_j	$P(x_i \mid x_j)$	x_i, x_j	$P(x_i \mid x_j)$
x_1, x_1	0,8125	x_1, x_2	0,125	x_1, x_3	0	x_1, x_4	0,5
x_2, x_1	0,1875	x_2, x_2	0,5	x_2, x_3	0	x_2, x_4	0,25
x_3, x_1	0	x_3, x_2	0,375	x_3, x_3	0	x_3, x_4	0,25
x_4, x_1	0	x_4, x_2	0	x_4, x_3	1	x_4, x_4	0

4. Избыточность ряда европейских языков лежит в пределах 5065%. Определить энтропию их алфавитов.

4.4. Самостоятельная работа

1. Марковский источник с начальным распределение $P(S1) = P(S2) = 1/2$ задан следующей диаграммой:



Вероятности порождения символов 0, 1 следующие: $P(0|S1) = 0,7$, $P(0|S2) = 0,3$, $P(1|S1) = 0,4$ $P(1|S2) = 0,6$. Найти энтропию данного источника.

2. Лектор произносит в среднем около 46 шестибуквенных слов в минуту. Рассматривая его как источник дискретных сообщений определить его производительность, считая, что все буквы алфавита равновероятны и статистически независимы.

4.5. Контрольные вопросы

1. Что есть цепь Маркова?
2. Как определяется энтропия марковского источника сообщений?
3. Что есть избыточность7
4. Каковы последствия от наличия избыточности сообщений?
5. Что есть производительность источника дискретных сообщений7
6. При каких условиях достигается наибольшая производительность источника?

ТЕМА 5

ДИСКРЕТНЫЕ КАНАЛЫ СВЯЗИ

Цель занятия – систематизация, углубление, закрепление полученных теоретических знаний о моделях дискретных каналов связи.

5.1. Общее описание дискретного канала

Дискретным каналом называют совокупность средств, предназначенных для передачи дискретных сигналов. Такие каналы широко используются, например, при передаче данных, в телеграфии, радиолокации.

Дискретные сообщения, состоящие из последовательности знаков алфавита источника сообщений (первичного алфавита) $\mathbf{X} = \{x_1, x_2, \ldots, x_l\}$, преобразуются в кодирующем устройстве в последовательности символов. Объем m алфавита символов (вторичного алфавита) $\mathbf{U} = \{u_1, u_2, \ldots, u_m\}$, как правило, меньше объема l алфавита знаков, но они могут и совпадать.

Информационная модель канала с помехами задается множеством символов на его входе и выходе и описанием вероятностных свойств передачи отдельных символов. В общем случае канал может иметь множество состояний и переходить из одного состояния в другое как с течением времени, так и в зависимости от последовательности передаваемых символов.

В каждом состоянии канал характеризуется матрицей условных вероятностей $p(v_j|u_i)$ того, что переданный символ u_i будет воспринят на выходе как символ v_j. Значения вероятностей в реальных каналах зависят от многих различных факторов: свойств сигналов, являющихся физическими носителями символов (энергия, вид модуляции и т.д.), характера и интенсивности воздействующих на канал помех, способа определения сигнала на приемной стороне.

При наличии зависимости переходных вероятностей канала от времени, что характерно практически для всех реальных каналов, он называется *нестационарным каналом связи*. Если эта зависимость несущественна, используется модель в виде стационарного канала, переходные вероятности которого не зависят от времени. Нестационарный канал может быть представлен рядом стационарных каналов, соответствующих различным интервалам времени.

Канал называется с «памятью» (с последействием), если переходные вероятности в данном состоянии канала зависят от его предыдущих состояний. Если переходные вероятности постоянны, т.е. канал имеет только одно состояние, он называется стационарным каналом без памяти. Под k-ичным каналом подразумевается канал связи, у которого число различных символов на входе и выходе одинаково и равно k.

5.2. Модели дискретного канала

Стационарный дискретный двоичный канал без памяти однозначно определяется четырьмя условными вероятностями: $P(0|0)$, $P(1|0)$, $P(0|1)$, $P(1|1)$. Такую модель канала принято изображать в виде графа, представленного на рис. 5.1.

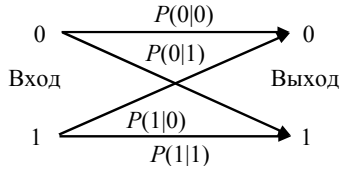


Рис. 5.1. Дискретный двоичный канал без памяти

Здесь $P(0|0)$, $P(1|1)$ – вероятности неискаженной передачи символов, а $P(1|0)$, $P(0|1)$ – вероятности искажения (трансформация) символов 1 и 0, соответственно.

Если вероятности искажения символов можно принять равными, т.е. $P(1|0) = P(0|1) = q$, то такой канал называют *двоичным симметричным каналом* (при $P(1|0) =/= P(0|1)$ канал называется *несимметричным*). Символы на его выходе правильно принимают с вероятностью p и неправильно – с вероятностью 1 $p = q$. Математическая модель упрощается. Именно этот канал исследовался наиболее интенсивно не столько в силу своей практической значимости (многие реальные каналы описываются им весьма приближенно), сколько в силу простоты математического описания.

Следует отметить еще одну модель канала, которая в последнее время приобретает все большее значение. Это дискретный канал со стиранием. Для него характерно, что алфавит выходных символов отличается от алфавита входных символов. На входе, как и ранее, символы 0 и 1, а на выходе канала фиксируются состояния, при которых сигнал с равным основанием может быть отнесен как к единице, так и к нулю. На месте такого символа не ставится ни нуль, ни единица: состояние отмечается дополнительным символом стирания *S*. При декодировании значительно легче исправить такие символы, чем ошибочно определенные.

На рис. 5.2 ниже приведены модели стирающего канала при отсутствии (*а*) и при наличии (*б*) трансформации символов.

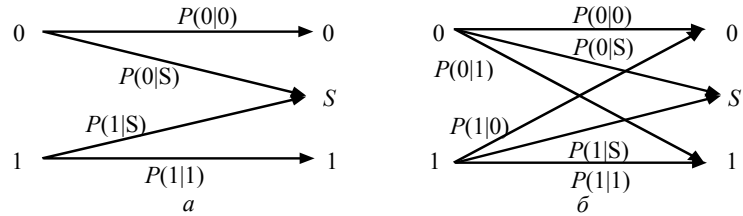


Рис. 5.2. Модели стирающего канала: *а*) с отсутствием трансформации символов; *б*) с трансформацией символов

5.3. Информационных характеристики дискретных каналов

Характеризуя дискретный канал связи, используют два понятия скорости передачи: технической и информационной.

Под технической скоростью передачи $V_т$, называемой также скоростью манипуляции, подразумевают число элементарных сигналов (символов), передаваемых по каналу в единицу времени. Она зависит от свойств линии связи и быстродействия аппаратуры канала.

С учетом возможных различий в длительностях символов скорость

$$V_т = 1/\tau_{ср},$$

где $\tau_{ср}$ – среднее значение длительности символа.

При одинаковой продолжительности τ всех передаваемых символов $\tau_{ср} = \tau$.

Единицей измерения технической скорости служит бод – скорость, при которой за одну секунду передается один символ.

Информационная скорость, или *скорость передачи информации*, определяется средним количеством информации, которое передается по каналу в единицу времени. Она зависит как от характеристик данного канала связи, таких, как объем алфавита используемых символов, техническая скорость их передачи, статистические свойства помех в линии, так и от вероятностей поступающих на вход символов и их статистической взаимосвязи.

При известной скорости манипуляции $V_т$ скорость передачи информации по каналу $\bar{I}(V, U)$ задается соотношением

$$\bar{I}(V, U) = V_т I(V, U),$$

где $I(V, U)$ – среднее количество информации, переносимое одним символом.

Для теории и практики важно выяснить, до какого предела и каким путем можно повысить скорость передачи информации по конкретному каналу связи. Предельные возможности канала по передаче информации характеризуются его пропускной способностью.

Пропускная способность канала $C_д$ равна той максимальной скорости передачи информации по данному каналу, которой можно достигнуть при самых совершенных способах передачи и приема:

$$C_д = \max \bar{I}(V, U) = \max V_т I(V, U).$$

При заданном алфавите символов и фиксированных основных характеристиках канала (например, полосе частот, средней и пиковой мощности передатчика) остальные характеристики должны быть выбраны такими, чтобы обеспечить наибольшую скорость передачи по нему элементарных сигналов, т.е. обеспечить максимальное значение $V_т$. Максимум среднего количества информации, приходящейся на один символ принятого сигнала $I(V, U)$, определяется на множестве распределений вероятностей между символами $u_1 \ldots u_i \ldots u_m$.

Пропускная способность канала, как и скорость передачи информации по каналу, измеряется числом двоичных единиц информации в секунду (дв. ед./с).

Так как в отсутствие помех имеет место взаимно-однозначное соответствие между множеством символов $\{v\}$ на выходе канала и $\{u\}$ на его входе, то

$$I(V, U) = I(U, V) = H(U).$$

Максимум возможного количества информации на символ равен $\log m$, где m – объем алфавита символов, откуда пропускная способность дискретного канала без помех:

$$C_{\text{д}} = V_{\text{т}} \log m.$$

Следовательно, для увеличения скорости передачи информации по дискретному каналу без помех и приближения ее к пропускной способности канала последовательность букв сообщения должна подвергнуться такому преобразованию в кодере, при котором различные символы в его выходной последовательности появлялись бы по возможности равновероятно, а статистические связи между ними отсутствовали бы.

Расширение объема алфавита символов m приводит к повышению пропускной способности канала, однако возрастает и сложность технической реализации.

При наличии помех соответствие между множествами символов на входе и выходе канала связи перестает быть однозначным. Среднее количество информации $I(V, U)$, передаваемое по каналу одним символом, определяется в этом случае соотношением

$$I(V, U) = H(V) - H(V|U) = H(U) - H(U|V).$$

Если статистические связи между символами отсутствуют, энтропия сигнала на выходе линии связи равна

$$H(V) = -\sum_{i=1}^{m} P(v_i) \log P(v_i).$$

При наличии статистической связи энтропию определяют с использованием цепей Маркова.

Апостериорная энтропия характеризует уменьшение количества переданной информации вследствие возникновения ошибок. Она зависит как от статистических свойств последовательностей символов, поступающих на вход канала связи, так и от совокупности переходных вероятностей, отражающих вредное действие помехи.

Если объем алфавита входных символов u равен m_1, а выходных символов $v - m_2$, то

$$H(U|V) = -\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} P(v_j, u_i) \log P(v_j, u_i).$$

Тогда среднее количество информации, передаваемое по каналу одним символом

$$I(V,U)=\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}P(v_j,\,u_i)\log\frac{P(v_j,\,u_i)}{P(v_j)P(u_i)}.$$

Скорость передачи информации по каналу с помехами

$$I(V,U)=V_m\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}P(v_j,\,u_i)\log\frac{P(v_j,\,u_i)}{P(v_j)P(u_i)}.$$

Считая скорость манипуляции $V_т$ предельно допустимой при заданных технических характеристиках канала, величину $I(V,\,U)$ можно максимизировать, изменяя статистические свойства последовательностей символов на входе канала посредством преобразователя (кодера канала). Получаемое при этом предельное значение $C_д$ скорости передачи информации по каналу называют *пропускной способностью* дискретного канала связи с помехами:

$$C_д=\max_{p\{u\}}V_т I(V,U)\,,$$

где $p\{u\}$ – множество возможных распределений вероятностей входных сигналов.

Важно подчеркнуть, что при наличии помех пропускная способность канала определяет наибольшее количество информации в единицу времени, которое может быть передано со сколь угодно малой вероятностью ошибки.

Предельные возможности канала никогда не используются полностью. Степень его загрузки характеризуется коэффициентом использования канала

$$\lambda = I(Z) / C_д,$$

где $\bar{I}(Z)$ – производительность источника сообщений; $C_д$ – пропускная способность канала связи.

Поскольку нормальное функционирование канала возможно при изменении производительности источника в пределах

$$0 \le \bar{I}(Z) \le C_д,$$

λ теоретически может изменяться в пределах от 0 до 1.

5.4. Аудиторные задания

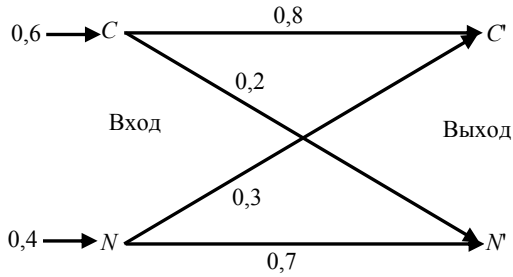
1. Определить среднее количество информации в системе передачи информации, описываемой матрицей

$$P(X,Y)=\begin{vmatrix} 1/8 & 1/8 & 1/8 \\ 1/8 & 0 & 1/8 \\ 1/8 & 1/8 & 1/8 \end{vmatrix}.$$

2. Вычислить среднее количество информации $I(X, y_2)$ о переданных сообщениях $X = \{x_1, x_2, x_3\}$, доставляемое принятым сообщением y_2 ансамбля $Y = \{y_1, y_2, y_3\}$, если система передачи описывается матрицей

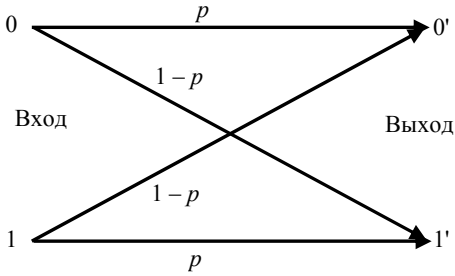
$$P(X,Y) = \begin{vmatrix} 0{,}05 & 0{,}2 & 0 \\ 0 & 0{,}3 & 0{,}1 \\ 0{,}05 & 0 & 0{,}3 \end{vmatrix}.$$

3. Сигнал C подается на вход канала с вероятностью 0,6 и отсутствует на входе с вероятностью 0,4. Поступивший сигнал воспроизводится на выходе канала с вероятностью 0,8 и теряется с вероятностью 0,2. При отсутствии сигнала на выходе возможен ложный сигнал C' с вероятностью 0,3. Диаграмма канала представлена ниже.



Найти среднее количество информации о входном сигнале по фиксируемому выходному.

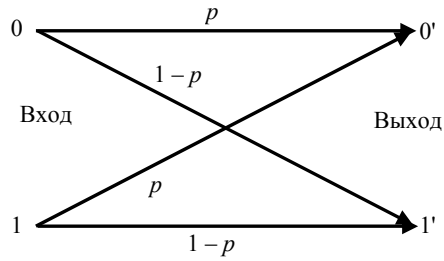
4. Диаграмма двоичного симметричного канала без памяти представлена ниже.



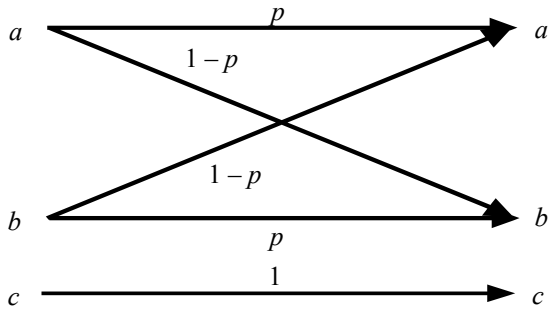
Определить пропускную способность канала.

Построить график $C_д = f(p)$. Дать интерпретацию поведения функции при $p = 1/2$.

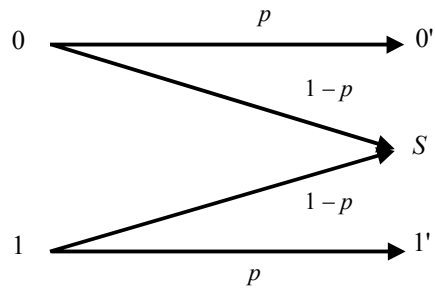
5. Определить пропускную способность двоичного канала с независимым входом и выходом.



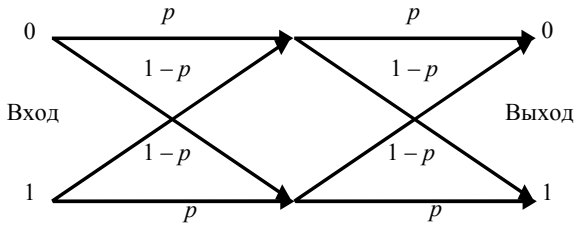
6. Определить пропускную способность канала. Диаграмма канала представлена ниже.



7. Определить пропускную способность двоичного симметричного канала со стиранием при отсутствии трансформации символов.



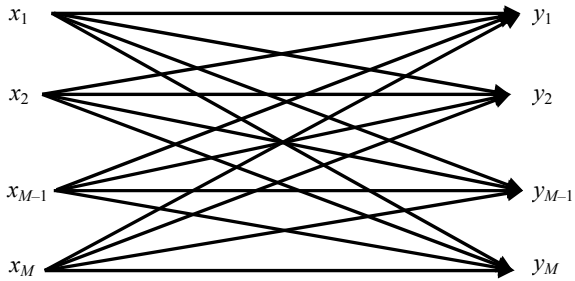
8. Вычислить пропускную способность канала, образованного последовательным соединением двух двоичных симметричных каналов. Диаграмма канала представлена ниже.



5.5. Самостоятельная работа

1. Определить количество информации, которое содержится в сообщении о том, что сумма выпавших очков на двух игральных костях равна семи.

2. Канал связи



задан вероятностной схемой

$$X = \begin{vmatrix} x_1 & x_2 & \dots & x_M \\ p(x_1) & p(x_2) & \dots & p(x_M) \end{vmatrix}.$$

Вероятности переходов равны

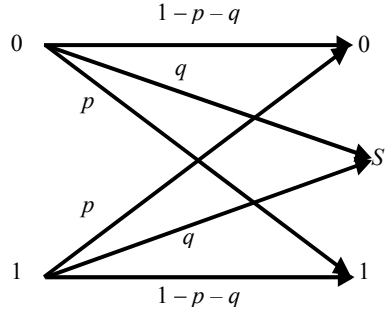
$$P(y_j \mid x_i) = \begin{cases} 1-p, \text{ если } j = i, \\ \dfrac{p}{M-1}, \text{ если } j \neq i. \end{cases}$$

Определить пропускную способность канала.

3. Определить пропускную способность троичного симметричного канала, задаваемого следующей матрицей вероятностей:

$$P(Y \mid X) = \begin{pmatrix} p & q & 0 \\ 0 & p & q \\ q & 0 & p \end{pmatrix}.$$

4. Определить пропускную способность двоичного симметричного канала со стиранием при наличии трансформации символов.



5. Вычислить пропускную способность канала, образованного последовательным соединением трех двоичных симметричных каналов.

5.6. Контрольные вопросы

1. В чем отличия между стационарным и нестационарным каналом связи?
2. Как задается информационная модель канала с помехами?
3. Какие каналы называют каналами с памятью?
4. В чем особенность канала со стиранием?
5. Как определяется скорость передачи информации?
6. От каких величин зависит пропускная способность дискретного канала?
7. Какой величиной характеризуется степень загрузки канала?

ТЕМА 6

ОПТИМАЛЬНОЕ КОДИРОВАНИЕ

Цель занятия – обучение построению оптимальных кодов.

6.1. Основные понятия и определения

Под кодированием понимают преобразование алфавита источника сообщений $A = \{a_i\}$, ($i = 1, 2, \ldots, K$) в алфавит некоторых кодовых символов $R = \{x_j\}$, ($j = 1, 2, \ldots, N$). Обычно размер алфавита кодовых символов существенно меньше размера алфавита источника. Последовательность кодовых символов, описывающих сообщение источника, называется *кодовым словом*.

Совокупность правил, в соответствии с которыми производятся операции преобразование алфавита источника сообщений в кодовые слова, называют *кодом*.

Необходимо иметь в виду, что кодирование, обеспечивающее изменение структуры сигнала, не должно изменять количество информации, заключенной в первоначальном сообщении.

По условиям построения кодовых слов коды делятся на равномерные и неравномерные. Если кодовые слова имеют разную длину, то код называется *неравномерным*. Типичным представителем этой группы является код Морзе.

В *равномерных кодах* все сообщения передаются кодовыми словами с одинаковым числом элементов. Примером является телеграфный код с длиной кодового слова равной 5.

По размеру алфавита кодовых символов различают следующие виды кодов: единичные, двоичные, многопозиционные.

Если исходный алфавит содержит K символов, то для построения равномерного кода с использованием N кодовых символов необходимо обеспечить выполнение следующего условия:

$$K \le N^q \text{ или } q \ge \frac{\log K}{\log N},$$

где q – количество элементов кодовой последовательности.

Заметим, что поиск равномерного кода означает, что каждая буква исходного алфавита кодируется кодовой последовательностью длиной q. Очевидно, что при различной вероятности появления букв исходного алфавита равномерный код является избыточным, так как энтропия, характеризующая информационную емкость сообщения максимальна при равновероятных буквах исходного текста. Например, для телетайпного кода $H_0 = 5$ бит, а с учетом неравномерности появления различных букв исход-

ного алфавита $H \approx 4{,}35$ бит. Устранение избыточности достигается применением неравномерных кодов, в которых буквы, имеющие наибольшую вероятность, кодируются наиболее короткими кодовыми последовательностями, а более длинные комбинации присваиваются редким буквам. Если i-я буква, вероятность которой P_i, получает кодовую комбинацию длины q_i, то средняя длина комбинации

$$q_{\text{ср}} = \sum_{i=1}^{K} P_i q_i .$$

Считая кодовые буквы равномерными, определяем наибольшую энтропию закодированного алфавита как $q_{\text{ср}} \log K$, которая не может быть меньше энтропии исходного алфавита H, т.е.

$$q_{\text{ср}} \log K \geq \text{H}.$$

Отсюда имеем $q_{\text{ср}} \geq (H / \log K)$.

Чем ближе значение $q_{\text{ср}}$ к энтропии H, тем более эффективно кодирование. В идеальном случае, когда $q_{\text{ср}} \approx H$, код называют *оптимальным*.

При построении неравномерных кодов необходимо обеспечить возможность их однозначной расшифровки. В равномерных кодах такая проблема не возникает, т.к. при расшифровке достаточно кодовую последовательность разделить на группы, каждая из которых состоит из q элементов. В неравномерных кодах можно использовать разделительный символ между буквами алфавита (так поступают, например, при передаче сообщений с помощью азбуки Морзе).

Если же отказаться от разделительных символов, то следует запретить такие кодовые комбинации, начальные части которых уже использованы в качестве самостоятельной комбинации. Например, если 101 означает код какой-то буквы, то нельзя использовать комбинации 1, 10 или 10101. Такие неравномерные коды называются *префиксными*. Рассмотрим пример кодирования сообщений a_i из алфавита объемом $K = 8$ с помощью произвольного q-разрядного двоичного кода.

Пусть источник сообщений выдает некоторый текст с алфавитом от А до З и одинаковой вероятностью букв $P(a_i) = 1/8$.

Кодирующее устройство кодирует эти буквы равномерным трехразрядным кодом.

Определим основные информационные характеристики источника с таким алфавитом:

– энтропия источника: $H = -\sum_{i=1}^{K} P_i \cdot l \log P_i$, $H_0 = \log K = 3$;

– избыточность источника: $r_H = \dfrac{H_0 - H}{H_0} = 0$;

– среднее число символов в коде: $q_{\text{ср}} = \sum_{i=1}^{K} q_i \cdot P_i = \sum_{i=1}^{8} 3 \cdot \frac{1}{8} = 3$;

– избыточность кода: $r_k = \frac{q_{\text{ср}} - H}{q_{\text{ср}}} = 0.$

Таким образом, при кодировании сообщений с равновероятными буквами избыточность выбранного (равномерного) кода оказалась равной нулю.

Пусть теперь вероятности появления в тексте различных букв будут разными (табл. 6.1).

Таблица 6.1

Вероятности букв

А	Б	В	Г	Д	Е	Ж	З
$P_{\text{а}} = 0{,}6$	$P_{\text{б}} = 0{,}2$	$P_{\text{в}} = 0{,}1$	$P_{\text{г}} = 0{,}04$	$P_{\text{д}} = 0{,}025$	$P_{\text{е}} = 0{,}015$	$P_{\text{ж}} = 0{,}01$	$P_{\text{з}} = 0{,}01$

Энтропия источника

$$H = -\sum_{i=1}^{K} P_i \cdot l\log P_i$$

в этом случае, естественно, будет меньшей и составит H =1,781.

Среднее число символов на одно сообщение при использовании того же равномерного трехразрядного кода

$$q_{\text{ср}} = \sum_{i=1}^{K} q_i \cdot P_i = \sum_{i=1}^{8} 3 \cdot \frac{1}{8} = 3\,.$$

Избыточность кода в этом случае будет

$$r_k = 1 - \frac{1{,}781}{3} \approx 0{,}41,$$

или довольно значительной величиной (в среднем 4 символа из 10 не несут никакой информации).

Ниже рассмотрим методы построения двоичных префиксных кодов.

6.2. Метод Фано

Метод сводится к последовательному выполнению следующих шагов.

1. Буквы алфавита **A** упорядочиваем по убыванию вероятностей:

$$P(a_1) \geq P(a_2) \geq \ldots \geq P(a_K).$$

2. Множество упорядоченных букв разбивается на два подмножества $\mathbf{A}^{(0)}$, $\mathbf{A}^{(1)}$ с помощью некоторого порогового целого числа $1 \leq k^{(1)} \leq K - 1$ так, чтобы величина

$$J^{(1)} = \left| \sum_{i=1}^{k^{(1)}} P(a_i) \; - \sum_{i=k^{(1)}+1}^{K} P(a_i) \right|$$

достигала наименьшего возможного значения. Буквам из подмножества $\mathbf{A}^{(0)}$ приписываем 0, буквам подмножества $\mathbf{A}^{(1)}$ приписываем 1.

3. Если подгруппы $\mathbf{A}^{(0)}$, $\mathbf{A}^{(1)}$ состоят более чем из двух букв, то разбиваем множество букв каждой из подгрупп на две подгруппы $\mathbf{A}^{(00)}$, $\mathbf{A}^{(01)}$ и $\mathbf{A}^{(10)}$, $\mathbf{A}^{(11)}$, соответственно, с помощью пороговых целых чисел

$$1 \le k^{(11)} \le k^{(1)} - 1,\ k^{(1)} \le k^{(12)} \le K - 1$$

так, чтобы величины

$$J^{(21)} = \left| \sum_{i=1}^{k^{(11)}} P(a_i) \ - \sum_{i=k^{(11)}+1}^{k^1} P(a_i) \right|,$$

$$J^{(22)} = \left| \sum_{i=k^{(1)}+1}^{k^{(12)}} P(a_i) \ - \sum_{i=k^{(12)}+1}^{K} P(a_i) \right|$$

достигали наименьших возможных значений. Буквам из подгрупп $\mathbf{A}^{(00)}$, $\mathbf{A}^{(10)}$ с нулевыми последними индексами приписываем 0, буквам из подгрупп $\mathbf{A}^{(10)}$, $\mathbf{A}^{(11)}$ приписываем 1.

Если подгруппа $\mathbf{A}^{(ij)}$ i,j = 0,1 состоит более чем из одной буквы, то переходим к шагу 4. Если все подгруппы состоят из одной буквы, то переходим к шагу 5.

4. Если есть подгруппы, состоящие более чем из одной буквы, то разбиваем каждую из них на две подгруппы, исходя из соотношения, введенного на шаге 2. Буквам подгрупп с нулевыми последними индексами приписываем нуль, остальным приписываем единицу.

Если все образовавшиеся подгруппы состоят из одной буквы, то переходим к шагу 5. Если есть подгруппы, состоящие более чем из одной буквы, то повторяем шаг 4.

5. Если образовавшиеся подгруппы состоят из одной буквы, то последовательно, начиная с первой метки, выписываем нули и единицы, относящиеся к каждой букве алфавита **A**.

В итоге получается двоичный префиксный код для заданного источника.

Демонстрационный пример приведен в табл. 6.2.

Таблица 6.2

Пример построения кода методом Фано

Буква	$P(a_i)$	I		II		III		IV		Код	$q_i \cdot P_i$
a_1	0,36	$A^{(0)}$	0	$A^{(00)}$	0					00	0,72
a_2	0,18		0	$A^{(01)}$	1					01	0,36
a_3	0,18	$A^{(1)}$	1	$A^{(10)}$	0					10	0,36
a_4	0,12		1	$A^{(11)}$	1	$A^{(110)}$	0			110	0,36
a_5	0,09		1		1	$A^{(111)}$	1	$A^{(1110)}$	0	1110	0,36
a_6	0,07		1		1		1	$A^{(1111)}$	1	1111	0,28

Средняя длина для построенного кода $q_{ср} = 2{,}44$. Соответствующее кодовое дерево имеет вид, представленный на рис. 6.1.

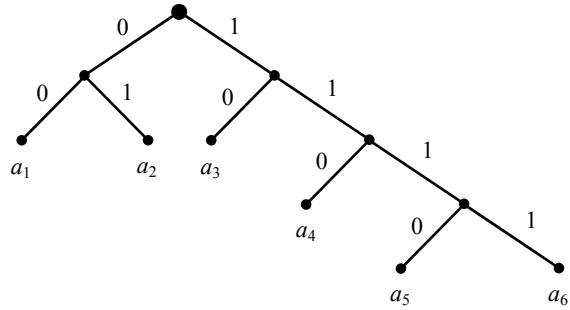


Рис. 6.1. Кодовое дерево построенного кода

6.3. Вектор Крафта

Полученный в предыдущем примере код называют еще префиксным множеством. В нашем случае это множество следующее:

$$\mathbf{S} = \{00, 01, 10, 110, 1110, 1111\}.$$

Если $\mathbf{S} = \{w_1, w_2, ..., w_K\}$ – префиксное множество, то можно определить некоторый вектор $\mathbf{v}(\mathbf{S}) = (q_1, q_2, ..., q_K)$, состоящий из чисел, являющихся длинами соответствующих префиксных последовательностей т.е q_i – это длина w_i.

Вектор $(q_1, q_2, ..., q_k)$, состоящий из неуменьшающихся положительных целых чисел, называется ***вектором Крафта***.

$$2^{-q_1} + 2^{-q_2} + \ldots + 2^{-q_K} \le 1.$$

Для него справедливо следующее утверждение: если **S** – какое-либо префиксное множество, то **v(S)** – вектор Крафта.

Иными словами, длины двоичных последовательностей в префиксном множестве удовлетворяют неравенству Крафта.

Неравенство Крафта формулируется в виде следующей теоремы:

Теорема. Пусть $q_1, q_2, ..., q_K$ набор положительных целых чисел. Для того чтобы существовал префиксный код с длинами кодовых слов $q_1, q_2, ..., q_K$, необходимо и достаточно, чтобы выполнялось неравенство

$$2^{-q_1} + 2^{-q_2} + \ldots + 2^{-q_K} \le 1.$$

Заметим, что теорема не утверждает, что любой код с длинами кодовых слов, удовлетворяющими неравенству Крафта, является префиксным. Например, множество двоичных кодовых слов {0, 01, 11} удовлетворяет неравенству, но не обладает свойством префикса.

Примеры простейших префиксных множеств и соответствующие им векторы Крафта:

$$\mathbf{S}_1 = \{0, 10, 11\} \text{ и } \mathbf{v}(\mathbf{S}_1) = (1, 2, 2);$$
$$\mathbf{S}_2 = \{0, 10, 110, 111\} \text{ и } \mathbf{v}(\mathbf{S}_2) = (1, 2, 3, 3);$$
$$\mathbf{S}_3 = \{0, 10, 110, 1110, 1111\} \text{ и } \mathbf{v}(\mathbf{S}_3) = (1, 2, 3, 4, 4);$$
$$\mathbf{S}_4 = \{0, 10, 1100, 1101, 1110, 1111\} \text{ и } \mathbf{v}(\mathbf{S}_4) = (1, 2, 4, 4, 4, 4).$$

6.4. Метод Шеннона

1. Буквы алфавита **A** упорядочиваем по убыванию вероятностей:

$$P(a_1) \geq P(a_2) \geq \ldots \geq P(a_K).$$

2. Находим числа q_i, $i = 1, \ldots, K$, исходя из условия:

$$\frac{1}{2^{q_i}} \leq P(a_i) \leq \frac{1}{2^{q_i - 1}}.$$

3. Подсчитываем накопленные суммы:

$$P_1 = 0,\ P_2 = P(a_1),\ P_3 = P(a_1) + P(a_2),\ \ldots,\ P_n = \sum_{i=1}^{K} P(a_i).$$

4. Находим первые после запятой q_i знаков в разложении числа P_i в двоичную дробь: $i = 1, \ldots, K$. Цифры этого разложения, стоящие после запятой, являются кодовым словом, соответствующим букве a_i.

5. Если необходимо, производим операцию усечения.

Рассмотрим кодовое дерево, максимальный порядок концевых вершин которого равен n. Предположим, что вершине a n-го порядка предшествует вершина b (n–l)-го порядка, из которой выходит единственное ребро, ведущее в a. Других рёбер из вершины b не выходит. Операцию удаления этого ребра и превращения промежуточной вершины b (n–l)-го порядка в концевую вершину будем называть *операцией усечения*. Средняя кодовая длина после операции усечения становится меньше. Пример приведен на рис. 6.2.

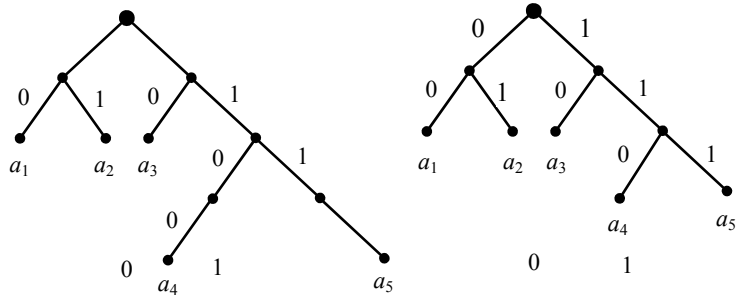


Рис. 6.2. Пример проведения процедуры усечения

Для сравнения методов рассмотрим тот же самый источник сообщений и проделаем последовательно шаги алгоритма (табл. 6.3).

Таблица 6.3

Кодирование методом Шеннона

	1	2	3	4	5	
Буква	$P(a_i)$	q_i	Накопленная P_i	Код до усечения	Код после усечения	$q_i*P(a_i)$
a_1	0,36	2	0 = 0,00	00	00	0,72
a_2	0,18	3	0,36 = 0,01011…	010	01	0,36
a_3	0,18	3	0,54 = 0,10001…	100	100	0,54
a_4	0,12	4	0,72 = 0,10111…	1011	101	0,36
a_5	0,09	4	0,84 = 0,11010…	1101	110	0,27
a_6	0,07	4	0,93 = 0,11101…	1110	111	0,21

Средняя длина для построенного кода $q_{ср} = 2{,}46$. Соответствующее кодовое дерево до и после усечения имеет вид, приведенный на рис. 6.3.

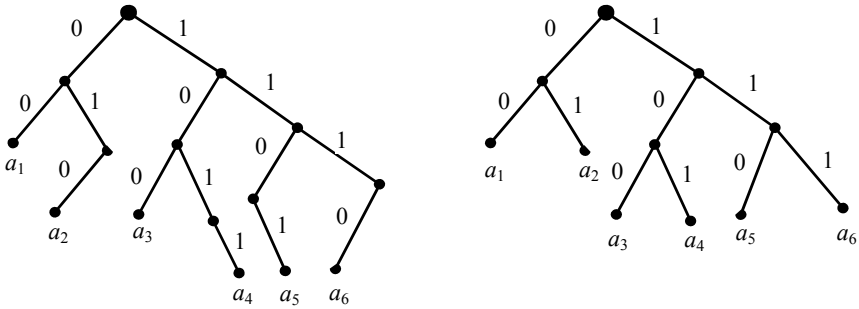


Рис. 6.3. Усечение кода, построенного методом Шеннона

Алгоритм Шеннона-Фано применим и при основании кода больше двух ($K > 2$). В этом случае алфавит разбивается на K частей примерно одинаковой суммарной вероятности.

6.5. Метод Хаффмана

Рассмотренные выше алгоритмы кодирования не всегда приводят к хорошему результату, вследствие отсутствия четких рекомендаций относительно того, как делить множество кодируемых знаков на подгруппы. Рассмотрим методику кодирования Хаффмана, которая свободна от этого недостатка.

Кодируемые знаки располагают в порядке убывания их вероятностей. Далее на каждом этапе две последние позиции списка заменяются одной и ей приписывают вероятность, равную сумме вероятностей заме-

няемых позиций. После этого производится пересортировка списка по убыванию вероятностей, с сохранением информации о том, какие именно знаки объединялись на каждом этапе. Процесс продолжается до тех пор, пока не останется единственная позиция с вероятностью, равной 1 (табл. 6.4).

Таблица 6.4

Процесс группировки вероятностей по методу Хаффмана

Буква	$P(a_i)$	Вспомогательные группировки				
		1	2	3	4	5
a_1	0,36	0,36	0,36	0,36	**0,64**	1
a_2	0,18	0,18	0,28	**0,36**	**0,36**	
a_3	0,18	0,18	**0,18**	**0,28**		
a_4	0,12	**0,16**	**0,18**			
a_5	**0,09**	**0,12**				
a_6	**0,07**					

Жирным шрифтом в таблице выделены объединяемые позиции, подчеркиванием – полученные при объединении позиции.

После этого строится кодовое дерево. Корню дерева ставится в соответствие узел с вероятностью, равной 1. Далее каждому узлу приписываются два потомка с вероятностями, которые участвовали в формировании значения вероятности обрабатываемого узла. Так продолжают до достижения узлов, соответствующих вероятностям исходных знаков (рис. 6.4).

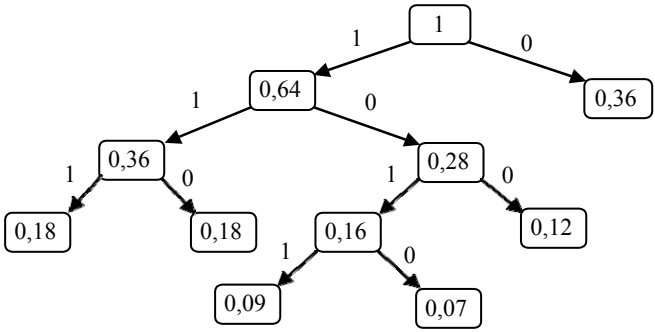


Рис. 6.4. Построение кода методом Хаффмана

Процесс кодирования по кодовому дереву осуществляется следующим образом. Одной из ветвей, выходящей из каждого узла, например, с более высокой вероятностью, ставится в соответствие символ 1, а с меньшей – 0. Спуск от корня к нужному знаку дает код этого знака. Правило кодирования в случае равных вероятностей оговаривается особо.

Средняя длина для построенного кода $q_{ср} = 2{,}44$, $H = 2{,}3595$.

6.6. Аудиторные задания

1. Символы источника без памяти

$$A=\left|\begin{matrix}a_1 & a_2 & a_3 & a_4\\ 0{,}4 & 0{,}3 & 0{,}2 & 0{,}1\end{matrix}\right|$$

кодируются двоичными последовательностями:

a) 1, 01, 001, 0001;

b) 1, 10, 100, 1000.

Являются ли данные коды префиксными, оптимальными? Какова средняя длина кодового слова?

2. Определить среднюю длину кодового слова и избыточность кода при кодировании по методу Фано, Шеннона, Хаффмана знаков ансамбля, приведенного ниже

$$X=\left|\begin{matrix}x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8\\ 0{,}25 & 0{,}25 & 0{,}15 & 0{,}15 & 0{,}05 & 0{,}05 & 0{,}05 & 0{,}05\end{matrix}\right|.$$

3. Источник информации задан вероятностной схемой

$$X=\left|\begin{matrix}x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8\\ 0{,}125 & 0{,}125 & 0{,}125 & 0{,}125 & 0{,}125 & 0{,}125 & 0{,}125 & 0{,}125\end{matrix}\right|.$$

Закодировать двоичными последовательностями ансамбль сообщений: 1) кодом Фано; 2) кодом Шеннона; 3) кодом Хаффмана; 4) равномерным двоичным кодом. Определить среднюю длину кодового слова и избыточность кода.

4. Источник информации задан вероятностной схемой

$$X=\left|\begin{matrix}x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9\\ 0{,}2 & 0{,}15 & 0{,}15 & 0{,}12 & 0{,}1 & 0{,}1 & 0{,}08 & 0{,}06 & 0{,}04\end{matrix}\right|.$$

Закодировать двоичными последовательностями ансамбль сообщений: 1) кодом Фано; 2) кодом Шеннона; 3) кодом Хаффмана; 4) равномерным двоичным кодом. Определить среднюю длину кодового слова и избыточность кода.

5. Сообщение кодируется при помощи двух групп символов, причем в первой группе имеется k символов, встречающихся с вероятностями

$$P_{11}, P_{12}, \ldots, P_{1k}, \sum_{i=1}^{k} P_{1i} = \alpha,$$

а во второй группе имеется m символов, встречающихся с вероятностями

$$P_{21}, P_{22}, \ldots, P_{2m}, \sum_{j=1}^{m} P_{2j} = 1-\alpha.$$

Считая, что значение α задано, определить, при каких вероятностях P_{1i} и P_{2j} энтропия максимальна. Для $k = m = 4$ и $\alpha = 0{,}4$ закодировать двоичными последовательностями ансамбль сообщений: 1) кодом Фано; 2) кодом Шеннона; 3) кодом Хаффмана. Определить среднюю длину кодового слова и избыточность кода.

6. Символы источника без памяти

$$A=\begin{vmatrix} a_1 & a_2 & \dots & a_5 \\ P(a_1) & P(a_2) & \dots & P(a_5) \end{vmatrix}, P(a_i)=p_i; \sum_{i=1}^{5} p_i = 1$$

кодируются двоичными последовательностями. Определить, можно ли построить двоичный префиксный код, длины l_i кодовых слов которых будут равны:

a) $l_1 = 1$, $l_2 = 1$, $l_3 = l_4 = l_5 = 3$;
b) $l_1 = 1$, $l_2 = l_3 = l_4 = l_5 = 2$;
c) $l_1 = 2$, $l_2 = 2$, $l_3 = l_4 = l_5 = 3$;
d) $l_1 = 2$, $l_2 = 2$, $l_3 = 2$, $l_4 = l_5 = 3$;
e) $l_1 = 1$, $l_2 = l_3 = l_4 = l_5 = 3$.

7. В сообщениях используются символы алфавита A_1, A_2, A_3, A_4 с вероятностями 0,45; 0,1; 0,15; 0,3. Для передачи сообщения по каналу связи могут быть применены два кода. В первом символам алфавита соответствуют символы кода a, b, c, d, во втором – a, d, b, c. Длительность элементов кода в условных единицах равны: $t_a = 8$, $t_b = 6$, $t_c = 5$, $t_d = 3$. Определить количество информации, передаваемое каждым кодом в единицу времени. Построить коды Фано, Шеннона, Хаффмана. Определить среднюю длину кодового слова и избыточность кода.

8. Закодировать знаки ансамблей сообщений, приведенных в задачах 1, 2, 3 и 4, троичным кодом. Определить среднюю длину кодового слова и избыточность кода.

9. Ансамбль из девяти сообщений представлен в виде четырех одноэлементных и пяти двухэлементных кодовых комбинаций. Определить минимальное кодовое основание, при котором код получается префиксным.

6.7. Самостоятельная работа

1. Определить, можно ли построить двоичное кодовое дерево с концевыми вершинами порядков

a) 1, 2, 3, 3, 4;
b) 2, 2, 3, 3, 3, 3;
c) 1, 3, 3, 3, 4, 4;

2. Из приведенных ниже кодов выделить однозначно декодируемые: a) {0,01}; b) {0,10,11}; c) {0,01,11}; d) {0,01,10}; e) {10,11,110}; f) {00, 10,01,11}; g) {00,10,11,100,110}.

3. Источник информации задан вероятностной схемой

$$X = \left| \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{matrix} \right|.$$

Закодировать двоичными и троичными последовательностями ансамбль сообщений: 1) кодом Фано; 2) кодом Шеннона; 3) кодом Хаффмана; 4) равномерным двоичным кодом. Определить среднюю длину кодового слова и избыточность кода.

4. Символы источника без памяти

$$A = \left| \begin{matrix} a_1 & a_2 & \dots & a_n \\ P(a_1) & P(a_2) & \dots & P(a_n) \end{matrix} \right|,$$

$$P(a_i) = \frac{1}{2^i}; i = 1, \dots, n-1; P(a_{n-1}) = \frac{1}{2^{n-1}},$$

кодируются двоичными последовательностями. Определить минимальное среднее число символов на сообщение.

5. Предложить пошаговую схему декодирования префиксного кода при известной таблице кодирования.

6.8. Контрольные вопросы

1. С какой целью кодируют сообщения?
2. Что такое код?
3. Назовите достоинства и недостатки равномерных и неравномерных кодов.
4. Для каких кодов требуется разделяющий символ?
5. Какие коды называют префиксными?
6. Чем отличаются кодовые деревья префиксных и непрефиксных кодов?
7. Каковы основные концепции, лежащие в основе оптимального кодирования?
8. Назовите недостатки оптимального кодирования.

ТЕМА 7

БЛОЧНОЕ КОДИРОВАНИЕ

Цель занятия – обучение построению блочных кодов.

7.1. Основные понятия

Для построения блочного кода вектор данных делят на блоки заданной длины и заменяют каждый блок кодовым словом из префиксного множества двоичных слов. Полученную последовательность кодовых слов объединяют в результирующую двоичную строку на выходе кодера. О блочном коде говорят, что он – блочный код k-го порядка, если все блоки имеют длину, равную k.

Рассмотрим процедуру эффективного кодирования сообщений, образованных с помощью алфавита, состоящего всего из двух знаков x_1 и x_2 с вероятностями появления соответственно $P(x_1) = 0{,}9$ и $P(x_2) = 0{,}1$.

Так как вероятности не равны, то последовательность из таких букв будет обладать избыточностью. Однако при побуквенном кодировании мы никакого эффекта не получим. Действительно, на передачу каждой буквы требуется символ либо 1, либо 0, т.е. $q_{ср} = 1$, в то время как энтропия равна $H(Z) = 0{,}47$.

При кодировании блоков, содержащих по две буквы, получим коды, показанные в табл. 7.1. Так как знаки статистически не связаны, вероятности блоков определяются как произведение вероятностей составляющих знаков.

Таблица 7.1

Построение блочного кода второго порядка

Блоки	Вероятности	Кодовые комбинации	$q_i \cdot P_i$
x_1x_1	0,81	0	0,81
x_1x_2	0,09	10	0,18
x_2x_1	0,09	110	0,27
x_2x_2	0,01	111	0,03

Среднее число символов на блок получается равным 1,29, а на букву – 0,645.

Кодирование блоков, содержащих по три знака, дает еще больший эффект. Соответствующий ансамбль и коды приведены в табл. 7.2.

Среднее число символов на блок равно 1,598, а на знак – 0,533, что всего на 13 % больше энтропии.

Теоретический минимум $H(Z) = 0{,}47$ может быть достигнут при кодировании блоков, включающих бесконечное число знаков:

$$\lim_{n\to\infty} q_{\text{ср}} = H(Z).$$

Таблица 7.2

Построение блочного кода третьего порядка

Блоки	Вероятности	1	2	3	4	5	Код	$q_i * P_i$
$x_1x_1x_1$	0,729	0					0	0,729
$x_1x_1x_2$	0,081			0			100	0,243
$x_1x_2x_1$	0,081		0	1			101	0,243
$x_2x_1x_1$	0,081			0			110	0,243
$x_1x_2x_2$	0,009					0	11100	0,045
$x_2x_1x_2$	0,009	1			0	1	11101	0,045
$x_2x_2x_1$	0,009		1	1		0	11110	0,045
$x_2x_2x_2$	0,001				1	1	11111	0,005

Следует подчеркнуть, что увеличение эффективности кодирования при укрупнении блоков не связано с учетом статистических связей, так как нами рассматривались алфавиты с некоррелированными знаками. Повышение эффективности определяется лишь тем, что набор вероятностей, получающихся при укрупнении блоков, можно делить на более близкие по суммарным вероятностям подгруппы.

7.2. Аудиторные задания

1. Сообщение составлено из неравновероятных независимых элементов x_1 с вероятностью 0,89 и x_2 с вероятностью 0,11. Построить эффективный двоичный код указанного сообщения так, чтобы избыточность была меньше 0,1.

Закодировать последовательность $x_1\ x_1\ x_1\ x_2\ x_1\ x_1\ x_1\ x_1\ x_2\ x_1\ x_1\ x_1$.

2. Сообщение составлено из равновероятных зависимых элементов x_1 ($P(x_1)$=0,5) и x_2 ($P(x_2)$=0,5). Заданы условные вероятности:

$$P(i|j)=\begin{vmatrix}0,9 & 0,1\\ 0,1 & 0,9\end{vmatrix}$$

Построить эффективный двоичный код указанного сообщения так, чтобы избыточность была меньше 0,15.

Закодировать последовательность $x_1\ x_2\ x_1\ x_2\ x_2\ x_1\ x_1\ x_2\ x_2\ x_2\ x_1\ x_1$

3. Сообщения составляются из трех независимых букв x_1, x_2, x_3, вероятности появления которых равны соответственно 0,7; 0,2; 0,1. Произвести кодирование оптимальным кодом: a) отдельных сообщений; b) блоков по два сообщения; c) блоков по три сообщения. Сравнить по избыточности полученные коды.

4. Сообщение составлено из неравновероятных зависимых элементов x_1 ($P(x_1)$=0,6) и x_2 ($P(x_2)$=0,4). Заданы условные вероятности:

$$P(i|j)=\begin{vmatrix}0,8 & 0,2\\ 0,2 & 0,8\end{vmatrix}.$$

Построить эффективный двоичный код указанного сообщения так, чтобы избыточность была меньше 0,1.

7.3. Самостоятельная работа

1. Источник вырабатывает независимо друг от друга два разных сообщения: x_1 (P (x_1)=0,9) и x_2 ($P(x_2)$=0,1). Закодировать a) отдельные сообщения; b) блоки по два сообщения; c) блоки по три сообщения. Сравнить по избыточности полученные коды.

2. Сообщения составляются из трех независимых букв x_1, x_2, x_3, вероятности появления которых равны соответственно 0,6; 0,3; 0,1. Произвести кодирование оптимальным кодом: a) отдельных сообщений; b) блоков по два сообщения; c) блоков по три сообщения. Сравнить по избыточности полученные коды.

7.4. Контрольные вопросы

1. Что такое блочный код?
2. Как строится блочный код?
3. Какие цели преследует построение блочного кода?

ТЕМА 8

ПОМЕХОУСТОЙЧИВОЕ КОДИРОВАНИЕ

Цель занятия – обучение построению помехоустойчивых кодов.

8.1. Общие принципы

Кодирование в канале, или помехоустойчивое кодирование информации, используется для уменьшения количества ошибок, возникающих при передаче по каналу с помехами.

Коды, которые обеспечивают возможность обнаружения и исправления ошибки, называют помехоустойчивыми.

Эти коды используют для 1) исправления ошибок; 2) обнаружения ошибок.

Способность кода обнаруживать и исправлять ошибки обусловлена наличием в нем избыточных символов.

На вход кодирующего устройства поступает последовательность из k информационных двоичных символов. На выходе ей соответствует последовательность из n двоичных символов, причем $n > k$.

Всего может быть 2^k различных входных и 2^n различных выходных последовательностей. Из общего числа 2^n выходных последовательностей только 2^k последовательностей соответствуют входным. Их называют *разрешенными кодовыми комбинациями*. Остальные $2^n - 2^k$ возможных выходных последовательностей для передачи не используются. Их называют *запрещенными кодовыми комбинациями*.

Искажения информации в процессе передачи сводятся к тому, что некоторые из передаваемых символов заменяются другими – неверными.

Так как каждая из 2^k разрешенных комбинаций в результате действия помех может трансформироваться в любую другую, то всегда имеется $2^k * 2^n$ возможных случаев передачи. В это число входят:

1) 2^k случаев безошибочной передачи;

2) $2^k (2^k - 1)$ случаев перехода в другие разрешенные комбинации, что соответствует необнаруженным ошибкам;

3) $2^k (2^n - 2^k)$ случаев перехода в неразрешенные комбинации, которые могут быть обнаружены.

Следовательно, часть обнаруживаемых ошибочных кодовых комбинаций от общего числа возможных случаев передачи составляет

$$2^k (2^n - 2^k) / (2^k \cdot 2^n) = 1 - \frac{2^k}{2^n}.$$

Любой метод декодирования можно рассматривать как правило разбиения всего множества запрещенных кодовых комбинаций на 2^k непересе-

кающихся подмножеств M_i, каждое из которых ставится в соответствие одной из разрешенных комбинаций. При получении запрещенной комбинации, принадлежащей подмножеству M_i, принимают решение, что передавалась запрещенная комбинация A_i. Ошибка будет исправлена в тех случаях, когда полученная комбинация действительно образовалась из A_i, т.е. в $2^n - 2^k$ случаях.

Всего случаев перехода в неразрешенные комбинации $2^k(2^n - 2^k)$. Таким образом, при наличии избыточности любой код способен исправлять ошибки.

Отношение числа исправляемых кодом ошибочных кодовых комбинаций к числу обнаруживаемых ошибочных комбинаций равно

$$(2^n - 2^k) / (2^k (2^n - 2^k)) = \frac{1}{2^k}.$$

Способ разбиения на подмножества зависит от того, какие ошибки должны направляться конкретным кодом.

Большинство разработанных кодов предназначено для корректирования взаимно независимых ошибок определенной кратности и пачек (пакетов) ошибок.

Взаимно независимыми ошибками называют такие искажения в передаваемой последовательности символов, при которых вероятность появления любой комбинации искаженных символов зависит только от числа искаженных символов r и вероятности искажения обычного символа p.

Кратностью ошибки называют количество искаженных символов в кодовой комбинации.

При взаимно независимых ошибках вероятность искажения любых r символов в n-разрядной кодовой комбинации:

$$p_r = C_n^r p^r (1-p)^{n-r},$$

где p – вероятность искажения одного символа; r – число искаженных символов; n – число двоичных символов на входе кодирующего устройства; C_n^r – число ошибок порядка r.

Если учесть, что $p << 1$, то в этом случае наиболее вероятны ошибки низшей кратности. Их следует обнаруживать и исправлять в первую очередь.

8.2. Связь корректирующей способности кода с кодовым расстоянием

Степень отличия любых двух кодовых комбинаций характеризуется *расстоянием* между ними в *смысле Хэмминга* или просто *кодовым расстоянием*.

Кодовое расстояние выражается числом символов, в которых комбинации отличаются одна от другой, и обозначается через d.

Чтобы получить кодовое расстояние между двумя комбинациями двоичного кода, достаточно подсчитать число единиц в сумме этих комбинаций по модулю 2. Например,

$$\begin{array}{l} 1001111101 \\ \oplus \\ \underline{1100001010} \\ 0101110111 \quad d = 7. \end{array}$$

Сложение «по модулю 2»: $y = x_1 \oplus x_2$, сумма равна 1 тогда и только тогда, когда x_1 и x_2 не совпадают.

Минимальное расстояние, взятое по всем парам кодовых разрешенных комбинаций кода, называют *минимальным кодовым расстоянием*.

Декодирование после приема производится таким образом, что принятая кодовая комбинация отождествляется с той разрешенной, которая находится от нее на наименьшем кодовом расстоянии.

Такое декодирование называется декодированием по *методу максимального правдоподобия*.

Очевидно, что при кодовом расстоянии $d = 1$ все кодовые комбинации являются разрешенными. Например, при $n = 3$ разрешенные комбинации образуют следующее множество:

000, 001, 010, 011, 100, 101, 110, 111.

Любая одиночная ошибка трансформирует данную комбинацию в другую разрешенную комбинацию. Это случай безызбыточного кода, не обладающего корректирующей способностью.

Если $d = 2$, то ни одна из разрешенных кодовых комбинаций при одиночной ошибке не переходит в другую разрешенную комбинацию.

Например, подмножество разрешенных кодовых комбинаций может быть образовано по принципу четности в нем числа единиц. Например, для $n = 3$:

000, 011, 101, 110 – разрешенные комбинации;

001, 010, 100, 111 – запрещенные комбинации.

Код обнаруживает одиночные ошибки, а также другие ошибки нечетной кратности (при $n = 3$ тройные).

В общем случае при необходимости обнаруживать ошибки кратности до r включительно минимальное хэммингово расстояние между разрешенными кодовыми комбинациями должно быть по крайней мере на единицу больше r, т.е

$$d_{0\min} \geq r + 1.$$

Действительно, в этом случае ошибка, кратность которой не превышает r, не в состоянии перевести одну разрешенную кодовую комбинацию в другую.

Для *исправления* одиночной ошибки кодовой комбинации необходимо сопоставить подмножество запрещенных кодовых комбинаций. Чтобы эти подмножества не пересекались, хэммингово расстояние между разрешенными кодовыми комбинациями должно быть не менее трех. При $n = 3$ за разрешенные кодовые комбинации можно, например, принять 000 и 111. Тогда разрешенной комбинации 000 необходимо приписать подмножество запрещенных кодовых комбинаций 001, 010, 100, образующихся в результате единичной ошибки в комбинации 000.

Подобным же образом разрешенной комбинации 111 необходимо приписать подмножество запрещенных кодовых комбинаций: 110, 011, 101, образующихся в результате возникновения единичной ошибки в комбинации 111.

В общем случае для обеспечения возможности исправления всех ошибок кратности до s включительно при декодировании по методу максимального правдоподобия, каждая из ошибок должна приводить к запрещенной комбинации, относящейся к подмножеству исходной разрешенной кодовой комбинации.

Минимальное хэммингово расстояние между разрешенными кодовыми комбинациями должно удовлетворять соотношению

$$d_{\text{и\_min}} \geq 2s + 1.$$

Для исправления всех ошибок кратности s и одновременного обнаружения всех ошибок кратности r минимальное хэммингово расстояние нужно выбирать из условия

$$d_{\text{и\_min}} \geq r + s + 1.$$

8.3. Понятие качества корректирующего кода

Одной из основных характеристик корректирующего кода является избыточность кода, указывающая степень удлинения кодовой комбинации для достижения определенной корректирующей способности.

Если на каждые n символов выходной последовательности кодера канала приходится k информационных и $(n - k)$ проверочных, то относительная избыточность кода может быть выражена соотношением:

$$R_k = (n - k) / k.$$

Оценим возможное наибольшее число Q разрешенных комбинаций n-значного двоичного кода, обладающего способностью исправлять взаимно независимые ошибки кратности до s включительно. Это равносильно отысканию числа комбинаций, кодовое расстояние между которыми не менее $d = 2s + 1$.

Общее число различных исправляемых ошибок для каждой разрешающей комбинации составляет $\sum_{i=1}^{s} C_n^i$, где C_n^i – число ошибок кратности i.

Каждая из таких ошибок должна приводить к запрещенной комбинации, относящейся к подмножеству данной разрешенной комбинации. Совместно с этой комбинацией подмножество включает $1+\sum_{i=1}^{s} C_n^i$ комбинаций.

Однозначное декодирование возможно только в том случае, когда названные подмножества не пересекаются. Так как общее число различных комбинаций n-значного двоичного кода составляет 2^n, число разрешенных комбинаций не может превышать

$$2^n / (1+\sum_{i=1}^{s} C_n^i) \text{ или } Q \le 2^n / (1+\sum_{i=0}^{s} C_n^i).$$

Эта верхняя оценка найдена Хэммингом. Для некоторых конкретных значений кодового расстояния d, соответствующие Q укажем в табл. 8.1.

Таблица 8.1

Количество передаваемых комбинаций в зависимости от кодового расстояния

d	Q	d	Q
1	2^n	5	$\le \frac{2^{n+1}}{n^2+n+2}$
2	$\le 2^{n-1}$	…	
3	$\le \frac{2^n}{n+1}$	…	
4	$\le \frac{2^{n-1}}{n}$	$2k+1$	$\le \frac{2^n}{1+C_n^1+C_n^2+...+C_n^k}$

Коды, для которых в приведенном соотношении достигается равенство, называют также плотноупакованными.

8.4. Построение двоичного группового кода

Построение конкретного корректирующего кода производят, исходя из требуемого объема кода Q, т. е. необходимого числа передаваемых команд или дискретных значений измеряемой величины и статистических данных о наиболее вероятных векторах ошибок в используемом канале связи.

Вектором ошибки называют n-разрядную двоичную последовательность, имеющую единицы в разрядах, подвергшихся искажению, и нули во всех остальных разрядах. Любую искаженную кодовую комбинацию можно рассматривать теперь как сумму (или разность) по модулю 2 исходной разрешенной кодовой комбинации и вектора ошибки.

Опираясь на неравенство $2^k - 1 \geq Q$ (нулевая комбинация часто не используется, так как не меняет состояния канала связи), определяем число информационных разрядов k, необходимое для передачи заданного числа команд обычным двоичным кодом.

Каждой из $2^k - 1$ ненулевых комбинаций k -разрядного безызбыточного кода нам необходимо поставить в соответствие комбинацию из n символов. Значения символов в $n - k$ проверочных разрядах такой комбинации устанавливаются в результате суммирования по модулю 2 значений символов в определенных информационных разрядах.

Поскольку множество 2^k комбинаций информационных символов (включая нулевую) образует подгруппу группы всех n-разрядных комбинаций, то и множество 2^k n-разрядных комбинаций, полученных по указанному правилу, тоже является подгруппой группы n-разрядных кодовых комбинаций. Это множество разрешенных кодовых комбинаций и будет *групповым кодом*.

Нам надлежит определить число проверочных разрядов и номера информационных разрядов, входящих в каждое из равенств для определения символов в проверочных разрядах.

Разложим группу 2^n всех n-разрядных комбинаций на смежные классы по подгруппе 2^k разрешенных n-разрядных кодовых комбинаций, проверочные разряды в которых еще не заполнены. Помимо самой подгруппы кода в разложении насчитывается $2^{n-k} - 1$ смежных классов. Элементы каждого класса представляют собой суммы по модулю 2 комбинаций кода и образующих элементов данного класса. Если за образующие элементы каждого класса принять те наиболее вероятные для заданного канала связи вектора ошибок, которые должны быть исправлены, то в каждом смежном классе сгруппируются кодовые комбинации, получающиеся в результате воздействия на все разрешенные комбинации определенного вектора ошибки. Для исправления любой полученной на выходе канала связи кодовой комбинации теперь достаточно определить, к какому классу смежности она относится. Складывая ее затем (по модулю 2) с образующим элементом этого смежного класса, получаем истинную комбинацию кода.

Ясно, что из общего числа 2^{n-1} возможных ошибок групповой код может исправить всего $2^{n-k} - 1$ разновидностей ошибок по числу смежных классов.

Чтобы иметь возможность получить информацию о том, к какому смежному классу относится полученная комбинация, каждому смежному классу должна быть поставлена в соответствие некоторая контрольная последовательность символов, называемая *опознавателем* (*синдромом*).

Каждый символ опознавателя определяют в результате проверки на приемной стороне справедливости одного из равенств, которые мы составим для определения значений проверочных символов при кодировании.

Ранее указывалось, что в двоичном линейном коде значения проверочных символов подбирают так, чтобы сумма по модулю 2 всех символов (включая проверочный), входящих в каждое из равенств, равнялась нулю. В таком случае число единиц среди этих символов четное. Поэтому операции определения символов опознавателя называют проверками на четность. При отсутствии ошибок в результате всех проверок на четность образуется опознаватель, состоящий из одних нулей. Если проверочное равенство не удовлетворяется, то в соответствующем разряде опознавателя появляется единица. Исправление ошибок возможно лишь при наличии взаимно однозначного соответствия между множеством опознавателей и множеством смежных классов, а, следовательно, и множеством подлежащих исправлению векторов ошибок.

Таким образом, количество подлежащих исправлению ошибок является определяющим для выбора числа избыточных символов $n - k$. Их должно быть достаточно для того, чтобы обеспечить необходимое число опознавателей. Если, например, необходимо исправить все одиночные независимые ошибки, то исправлению подлежат n ошибок:

000...01

000...10

.....

010...00

100...00.

Различных ненулевых опознавателей должно быть не менее n. Необходимое число проверочных разрядов, следовательно, должно определяться из соотношения:

$$2^{n-k} - 1 \geq n \text{ или } 2^{n-k} - 1 \geq C_n^1.$$

Если необходимо исправить не только все единичные, но и все двойные независимые ошибки, соответствующее неравенство принимает вид:

$$2^{n-k} - 1 \geq C_n^1 + C_n^2.$$

В общем случае для исправления всех независимых ошибок кратности до s включительно получаем:

$$2^{n-k} - 1 \geq C_n^1 + C_n^2 + \ \ldots + C_n^s.$$

В приведенных соотношениях указывается теоретический предел минимально возможного числа проверочных символов, который не всегда можно реализовать практически. Часто проверочных символов требуется больше, чем следует из соответствующего равенства.

8.5. Составление таблицы опознавателей

Начнем для простоты с установления опознавателей для случая исправления одиночных ошибок ($s = 1$). Допустим, что необходимо закодировать 15 команд ($Q = 15$). Так как используется двоичный код, то требуемое число информационных разрядов k равно четырем ($2^k - 1 = 15$). Для исправления одиночной ошибки минимальное кодовое расстояние d должно равняться трем ($2s + 1 = 3$). Пользуясь неравенством для $d = 3$ из табл. 8.1, найдем, что общее число разрядов кода должно равняться семи ($15 \le 2^n / (n+1)$).

Три избыточных разряда позволяют использовать в качестве опознавателей трехразрядные двоичные последовательности. В данном случае ненулевые последовательности в принципе могут быть сопоставлены с подлежащими исправлению ошибками в любом порядке. Однако более целесообразно сопоставлять их с ошибками в разрядах, начиная с младшего, в порядке возрастания двоичных чисел, как указано в табл. 8.2.

Таблица 8.2

Векторы ошибок и соответствующие им опознаватели для исправления одиночной ошибки

Векторы ошибок	Опознаватели
0000001	001
0000010	010
0000100	011
0001000	100
0010000	101
0100000	110
1000000	111

При таком сопоставлении каждый опознаватель представляет собой двоичное число, указывающее номер разряда, в котором произошла ошибка.

Коды, в которых опознаватели устанавливаются по указанному принципу, известны как коды Хэмминга.

Возьмем теперь более сложный случай исправления одиночных и двойных независимых ошибок ($s = 2$). В качестве опознавателей одиночных ошибок в первом и втором разрядах можно принять, как и ранее, комбинации 0...001 и 0...010. Однако в качестве опознавателя одиночной ошибки в третьем разряде комбинацию 0...011 взять нельзя. Такая комбинация соответствует ошибке одновременно в первом и во втором разрядах, а она также подлежит исправлению и, следовательно, ей должен соответствовать свой опознаватель 0...011.

В качестве опознавателя одиночной ошибки в третьем разряде можно взять только трехразрядную комбинацию 0...0100, так как множество двухразрядных комбинаций уже исчерпано. Подлежащий исправлению

вектор ошибки 0...0101 также можно рассматривать как результат суммарного воздействия двух векторов ошибок 0...0100 и 0...001 и, следовательно, ему должен быть поставлен в соответствие опознаватель, представляющий собой сумму по модулю 2 опознавателей этих ошибок, т.е. 0...0101. Аналогично находим, что опознавателем вектора ошибки 0...0110 является комбинация 0...0110.

Определяя опознаватель для одиночной ошибки в четвертом разряде, замечаем, что еще не использована одна из трехразрядных комбинаций, а именно 0...0111. Однако, выбирая в качестве опознавателя единичной ошибки в i-м разряде комбинацию с числом разрядов, меньшим i, необходимо убедиться в том, что для всех остальных подлежащих исправлению векторов ошибок, имеющих единицы в i-м и более младших разрядах, получатся опознаватели, отличные от уже использованных. В нашем случае подлежащими исправлению векторами ошибок с единицами в четвертом и более младших разрядах являются:

0...01001, 0...01010, 0...01100.

Если одиночной ошибке в четвертом разряде поставить в соответствие опознаватель 0...0111, то для указанных векторов опознавателями должны были бы быть соответственно

$$\begin{array}{lll} 0...0111 & 0...0111 & 0...0111 \\ \oplus & \oplus & \oplus \\ \underline{0...0001} & \underline{0...0010} & \underline{0...0100} \\ 0...0110 & 0...0101 & 0...0011 \end{array}.$$

Однако эти комбинации уже использованы в качестве опознавателей других векторов ошибок, а именно:

0...0110, 0...0101, 0...0011.

Во избежание неоднозначности при декодировании в качестве опознавателя одиночной ошибки в четвертом разряде следует взять четырехразрядную комбинацию 1000. Тогда для векторов ошибок

0...01001, 0...01010, 0...01100

опознавателями соответственно будут:

0...01001, 0...01010, 0...01100.

Аналогично можно установить, что в качестве опознавателя одиночной ошибки в пятом разряде может быть выбрана не использованная ранее четырехразрядная комбинация 01111.

Действительно, для всех остальных подлежащих исправлению векторов ошибок с единицей в пятом и более младших разрядах получаем опознаватели, отличающиеся от ранее установленных (табл. 8.3).

Продолжая сопоставление, можно получить таблицу опознавателей для векторов ошибок данного типа с любым числом разрядов.

В табл. 8.4 и 8.5 указаны проверочные последовательности для кодов, исправляющих двойные и тройные ошибки.

Таблица 8.3

Векторы ошибок и опознаватели для $s = 2$

Векторы ошибок	Опознаватели
0…010001	0…01110
0…010010	0…01101
0…010100	0…01011
0…011000	0…00111

Таблица 8.4

Код для исправления двойной ошибки

Позиция ошибки	Проверочная последовательность	Позиция ошибки	Проверочная последовательность
1	0000000001	16	0011101101
2	0000000010	17	0011110111
3	0000000100	18	0100000000
4	0000001000	19	0100010111
5	0000001111	20	0100101001
6	0000010000	21	0110111101
7	0000100000	22	1000000000
8	0000110011	23	1000011001
9	0001000000	24	1 000101101
10	0001010101	25	1001010010
11	0001101010	26	1010000011
12	0010000000	27	1100100011
13	0010010110	28	1101011111
14	0010110101	29	1111100110
15	0011011011		

Таблица 8.5

Код для исправления тройной ошибки

Позиция ошибки	Проверочная последовательность
1	0000000001
2	0000000010
3	0000000100
4	0000001000
5	0000010000
6	0000100000
7	0000111111
8	0001000000
9	0010000000
10	0100000000
11	0110111101
12	1000000000
13	1011011001
14	1101101010
15	1110110100

8.6. Определение проверочных равенств

Для любого кода, имеющего целью исправлять наиболее вероятные векторы ошибок заданного канала связи (взаимно независимые ошибки или пачки ошибок), можно составить таблицу опознавателей одиночных ошибок в каждом из разрядов. Пользуясь этой таблицей, нетрудно определить, символы каких разрядов должны входить в каждую из проверок на четность.

Рассмотрим в качестве примера опознаватели для кодов, предназначенных исправлять единичные ошибки (табл. 8.6).

Таблица 8.6

Опознаватели для исправления единичной ошибки

Номер разряда	Опознаватель	Номер разряда	Опознаватель	Номер разряда	Опознаватель
1	0001	7	0111	12	1100
2	0010	8	1000	13	1101
3	0011	9	1001	14	1110
4	0100	10	1010	15	1111
5	0101	11	1011	16	10000
6	0110				

В принципе можно построить код, усекая эту таблицу на любом уровне. Однако из таблицы видно, что оптимальными будут коды (7, 4), (15, 11), где первое число равно n, а второе k, и другие, которые среди кодов, имеющих одно и то же число проверочных символов, допускают наибольшее число информационных символов.

Усечем эту таблицу на седьмом разряде и найдем номера разрядов, символы которых должны войти в каждое из проверочных равенств.

Предположим, что в результате первой проверки на четность для младшего разряда опознавателя будет получена единица. Очевидно, это может быть следствием ошибки в одном из разрядов, опознаватели которых в младшем разряде имеют единицу. Следовательно, первое проверочное равенство должно включать символы 1, 3, 5 и 7-го разрядов:

$$a_1 \oplus a_3 \oplus a_5 \oplus a_7 = 0.$$

Единица во втором разряде опознавателя может быть следствием ошибки в разрядах, опознаватели которых имеют единицу во втором разряде. Отсюда второе проверочное равенство должно иметь вид

$$a_2 \oplus a_3 \oplus a_6 \oplus a_7 = 0.$$

Аналогично находим и третье равенство:

$$a_4 \oplus a_5 \oplus a_6 \oplus a_7 = 0.$$

Чтобы эти равенства при отсутствии ошибок удовлетворялись для любых значений информационных символов в кодовой комбинации, в нашем распоряжении имеется три проверочных разряда. Необходимо так вы-

брать номера этих разрядов, чтобы каждый из них входил только в одно из равенств. Это обеспечит однозначное определение значений символов в проверочных разрядах при кодировании. Этому условию удовлетворяют разряды, опознаватели которых имеют одну единицу. В нашем случае это первый, второй и четвертый разряды.

Таким образом, для кода (7, 4), исправляющего одиночные ошибки, искомые правила построения кода, т. е. соотношения, реализуемые в процессе кодирования, принимают вид:

$$a_1 = a_3 \oplus a_5 \oplus a_7,$$
$$a_2 = a_3 \oplus a_6 \oplus a_7,$$
$$a_4 = a_5 \oplus a_6 \oplus a_7.$$

Поскольку построенный код имеет минимальное хэммингово расстояние $d_{\min} = 3$, он может использоваться с целью обнаружения единичных и двойных ошибок. Обращаясь к предыдущей таблице, легко убедиться, что сумма любых двух опознавателей единичных ошибок дает ненулевой опознаватель, что и является признаком наличия ошибки.

Пример. Построим групповой код объемом 15 слов, способный исправлять единичные и обнаруживать двойные ошибки.

В соответствии с $d_{\text{иомin}} \geq r+s+1$ код должен обладать минимальным хэмминговым расстоянием, равным 4. Такой код можно построить в два этапа.

Сначала строим код заданного объема, способный исправлять единичные ошибки ($d_{\text{имin}} = 3$). Это код Хэмминга (7, 4), в котором всего семь разрядов, из которых четыре информационных и три – контрольных. Таблица опознавателей для такого кода приведена в табл. 8.2.

Затем добавляем еще один проверочный разряд, который обеспечивает четность числа единиц в разрешенных комбинациях. Получаем код Хэмминга (8, 4), способный обнаруживать двойные ошибки и исправлять единичные. В процессе кодирования реализуются соотношения:

$$a_1 = a_3 \oplus a_5 \oplus a_7,$$
$$a_2 = a_3 \oplus a_6 \oplus a_7,$$
$$a_4 = a_5 \oplus a_6 \oplus a_7,$$
$$a_8 = a_1 \oplus a_2 \oplus a_3 \oplus a_4 \oplus a_5 \oplus a_6 \oplus a_7.$$

Для декодирования уравнения следующие:

$$a_1 \oplus a_3 \oplus a_5 \oplus a_7 = 0,$$
$$a_2 \oplus a_3 \oplus a_6 \oplus a_7 = 0,$$
$$a_4 \oplus a_5 \oplus a_6 \oplus a_7 = 0,$$
$$a_1 \oplus a_2 \oplus a_3 \oplus a_4 \oplus a_5 \oplus a_6 \oplus a_7 \oplus a_8 = 0.$$

Обозначив синдром кода (7, 4), проверяющийся по первым трем уравнениям, через S_1, а результат общей проверки на четность (четвертое

уравнение) через S_2, и пренебрегая возможностью возникновения ошибок кратности выше 2, запишем алгоритм декодирования:

– при $S_1 = 0$ и $S_2 = 0$ – ошибок нет;

– при $S_1 = 0$ и $S_2 = 1$ – ошибка в восьмом разряде;

– при $S_1 \neq 0$ и $S_2 = 0$ – двойная ошибка (коррекция блокируется, посылается запрос повторной передачи);

– при $S_1 \neq 0$ и $S_2 = 1$ – одиночная ошибка (осуществляется ее исправление).

Разберем алгоритм на примере. Допустим, источнику необходимо передать последовательность 1011.

Результат кодирования последовательности 1011 кодом Хэмминга (8, 4) представлен в табл. 8.7.

Таблица 8.7

Закодированная последовательность 1011

Разряд	$1^к$	$2^к$	3	$4^к$	5	6	7	$8^к$
Код	0	1	1	0	0	1	1	0

В первой строке показан номер разряда, во второй – значение разряда, то есть сам код. Разряды под номерами 3, 5, 6 и 7 – информационные, они заполняются кодируемой последовательностью. Разряды 1, 2, 4 и 8 являются контрольными, их значения находятся по уравнениям кодирования:

$$a_1 = a_3 \oplus a_5 \oplus a_7 \Rightarrow a_1 = 1 \oplus 0 \oplus 1 = 0,$$
$$a_2 = a_3 \oplus a_6 \oplus a_7 \Rightarrow a_2 = 1 \oplus 1 \oplus 1 = 1,$$
$$a_4 = a_5 \oplus a_6 \oplus a_7 \Rightarrow a_4 = 0 \oplus 1 \oplus 1 = 0,$$
$$a_8 = a_1 \oplus a_2 \oplus a_3 \oplus a_4 \oplus a_5 \oplus a_6 \oplus a_7 \Rightarrow a_8 = 0 \oplus 1 \oplus 1 \oplus 0 \oplus 0 \oplus 1 \oplus 1 = 0.$$

Закодированная последовательность 01100110 передается по каналу связи. Если помехи не оказали влияния на код, то приемник, получив сигнал 01100110, воспользуется уравнениями декодирования для расчета синдромов:

$$a_1 \oplus a_3 \oplus a_5 \oplus a_7 = 0 \oplus 1 \oplus 0 \oplus 1 = 0,$$
$$a_2 \oplus a_3 \oplus a_6 \oplus a_7 = 1 \oplus 1 \oplus 1 \oplus 1 = 0,$$
$$a_4 \oplus a_5 \oplus a_6 \oplus a_7 = 0 \oplus 0 \oplus 1 \oplus 1 = 0,$$
$$a_1 \oplus a_2 \oplus a_3 \oplus a_4 \oplus a_5 \oplus a_6 \oplus a_7 \oplus a_8 = 0 \oplus 1 \oplus 1 \oplus 0 \oplus 0 \oplus 1 \oplus 1 \oplus 0 = 0.$$

Так как $S_1 = 000$ и $S_2 = 0$, то приемник убедится, что передача была безошибочной.

Если же помехи привели к искажению последовательности, то возможны три варианта.

Вариант первый.

При передаче произошла ошибка в восьмом символе. Приемник получает последовательность 01100111 и проверяет её:

$$a_1 \oplus a_3 \oplus a_5 \oplus a_7 = 0 \oplus 1 \oplus 0 \oplus 1 = 0,$$
$$a_2 \oplus a_3 \oplus a_6 \oplus a_7 = 1 \oplus 1 \oplus 1 \oplus 1 = 0,$$
$$a_4 \oplus a_5 \oplus a_6 \oplus a_7 = 0 \oplus 0 \oplus 1 \oplus 1 = 0,$$
$$a_1 \oplus a_2 \oplus a_3 \oplus a_4 \oplus a_5 \oplus a_6 \oplus a_7 \oplus a_8 = 0 \oplus 1 \oplus 1 \oplus 0 \oplus 0 \oplus 1 \oplus 1 \oplus 1 = 1.$$

По значению синдромов $S_1 = 000$ и $S_2 = 1$ приемник делает вывод, что произошла ошибка в восьмом разряде, исправляет последнюю 1 на 0 и получает правильную последовательность 01100110.

Вариант второй.

При передаче произошла двойная ошибка.

Сделаем для примера любые два изменения в исходной последовательности 01100110. Пусть ошибка будет в контрольном разряде 2 и информационном разряде 6, тогда последовательность, полученная приемником, будет выглядеть следующим образом: 00100010. Приемник использует уравнения декодирования на полученном коде:

$$a_1 \oplus a_3 \oplus a_5 \oplus a_7 = 0 \oplus 1 \oplus 0 \oplus 1 = 0,$$
$$a_2 \oplus a_3 \oplus a_6 \oplus a_7 = 0 \oplus 1 \oplus 0 \oplus 1 = 0,$$
$$a_4 \oplus a_5 \oplus a_6 \oplus a_7 = 0 \oplus 0 \oplus 0 \oplus 1 = 1,$$
$$a_1 \oplus a_2 \oplus a_3 \oplus a_4 \oplus a_5 \oplus a_6 \oplus a_7 \oplus a_8 = 0 \oplus 0 \oplus 1 \oplus 0 \oplus 0 \oplus 0 \oplus 1 \oplus 0 = 0.$$

Приемник получает синдромы $S_1 = 100$ и $S_2 = 0$. Так как первый синдром не равен нулю, а второй равен нулю, приемник понимает, что сделана двойная ошибка, однако установить, в каких именно разрядах, он не может. Поэтому принятая последовательность удаляется, а источнику отправляется уведомление об ошибке и запрос на повторную отправку.

Вариант третий.

Помехи в канале привели к возникновению одиночной ошибки. Для примера сделаем ошибку в разряде 3, тогда код, полученный приемником, будет выглядеть следующим образом: 01000110. Проверяем:

$$a_1 \oplus a_3 \oplus a_5 \oplus a_7 = 0 \oplus 0 \oplus 0 \oplus 1 = 1,$$
$$a_2 \oplus a_3 \oplus a_6 \oplus a_7 = 1 \oplus 0 \oplus 1 \oplus 1 = 1,$$
$$a_4 \oplus a_5 \oplus a_6 \oplus a_7 = 0 \oplus 0 \oplus 1 \oplus 1 = 0,$$
$$a_1 \oplus a_2 \oplus a_3 \oplus a_4 \oplus a_5 \oplus a_6 \oplus a_7 \oplus a_8 = 0 \oplus 1 \oplus 0 \oplus 0 \oplus 0 \oplus 1 \oplus 1 \oplus 0 = 1.$$

Получаем синдромы $S_1 = 011$ (синдром S_1 составляется, начиная с последнего уравнения, не считая уравнения для S_2) и $S_2 = 1$. При обнаружении одиночной ошибки приемник может осуществить её исправление, сравнив полученный в S_1 синдром с опознавателями из таблицы 8.2. Опознаватель 011 соответствует ошибке в третьем разряде. Исправив в третьем разряде 0 на 1, приемник получит исходную последовательность 01100110.

8.7. Аудиторные задания

1. Даны кодовые слова: 00001, 11100, 10110, 01110. Оцените корректирующие способности данного кода.

2. Постройте таблицу декодирования по методу максимального правдоподобия для двоичного кода, состоящего из четырех кодовых слов 0000, 0011, 1100, 1111:

а) для двоичного симметричного канала;

б) для двоичного стирающего канала при отсутствии трансформации символов.

3. Какое количество символов первичного алфавита можно передать с помощью пятизначного двоичного кода, исправляющего одиночную ошибку? Определить количество информационных двоичных символов в коде. Построить код и таблицу декодирования.

4. Построить код Хэмминга с исправлением одиночной ошибки для передачи 15 информационных сообщений. Закодировать четыре информационных сообщения: 0001, 0010, 1000, 1011. Проверить правильность принятых сообщений: 1111001, 0101010, 1100011.

5. Построить код Хэмминга с исправлением одиночной ошибки и обнаружением двойной ошибки для передачи 15 информационных сообщений. Закодировать четыре информационных сообщения: 0010, 0100, 1100, 1011. Проверить правильность построенного кода на следующих кодовых комбинациях: а) переданных без искажений; б) искажен один разряд; в) искажено два разряда;

6. Построить код Хэмминга с исправлением двойной ошибки для передачи 3 информационных сообщений. Закодировать четыре информационных сообщения: 0100, 0010, 1010, 1001. Проверить правильность принятых сообщений: 11110010, 00101010, 10100011.

8.8. Самостоятельная работа

1. Даны кодовые слова: 000010, 111000, 010110, 011101. Оцените корректирующие способности данного кода.

2. Какое количество символов первичного алфавита можно передать с помощью шестизначного двоичного кода, исправляющего одиночную ошибку? Определить количество информационных двоичных символов в коде. Построить код и таблицу декодирования.

3. Построить код Хэмминга с исправлением двойной ошибки для передачи 7 информационных сообщений. Закодировать четыре информационных сообщения: 0100, 0010, 1010, 1001. Проверить правильность принятых сообщений: 11110010, 00101010, 10100011.

4. Предложить пошаговую схему декодирования кода Хэмминга.

8.9. Контрольные вопросы

1. Какая основная концепция лежит в основе построения помехоустойчивых кодов?

2. Какова вероятность искажения любых r символов в n-разрядной кодовой комбинации при взаимно независимых ошибках?

3. Как определяется расстояние Хэмминга?

4. Поясните связь минимального кодового расстояния с обнаружением и исправлением ошибок.

5. Как ведется декодирование по методу максимального правдоподобия?

6. Что есть опознаватель или синдром? Какова его функция?

7. Как строится код Хэмминга?

8. Что представляет собой синдром при обнаружении и исправлении ошибок?

ЛИТЕРАТУРА

1. Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшее образование, 2008. 479 с.

2. Капитонова Ю.В. Лекции по дискретной математике / Ю.В. Капитонова, С.Л. Кривой, А.А. Летичевский, Г.М. Луцкий / СПб.: БХВ-Петербург, 2004. 624 с.

3. Дмитриев В.И. Прикладная теория информации. М.: Высш. шк., 1989. 320 с.

4. Теория кодирования. Под ред. Э.Л. Блоха. М.: Мир, 1964. 258 с.

Учебное издание

Илья Александрович Ходашинский
Марина Борисовна Бардамова

Теория информации

Методические указания для выполнения практических и самостоятельных работ
для студентов специальности и направления
10.03.01 – «Информационная безопасность»,
10.05.02 – «Информационная безопасность телекоммуникационных систем»,
10.05.03 – «Информационная безопасность автоматизированных систем»,
10.05.04 – «Информационно-аналитические системы безопасности»

Верстка – В.М. Бочкаревой
Текст дан в авторской редакции, без корректуры

Издательство «В-Спектр»
Подписано к печати 25.08.2018.
Формат $60\times84^{1}/_{16}$. Печать трафаретная.
Печ. л. 4. Тираж 100 экз. Заказ 222.

Тираж отпечатан ИП Бочкаревой В.М.
ИНН 701701817754
634055, г. Томск, пр. Академический, 13-24, тел. 49-09-91.
E-mail: bvm@sibmail.com