МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Омский государственный технический университет»

**А. С. Гуменюк, Н. Н. Поздниченко**

# ТЕОРИЯ ИНФОРМАЦИИ И КОДИРОВАНИЯ

*Учебное текстовое электронное издание*
*локального распространения*

Омск
Издательство ОмГТУ
2015



Сведения об издании: 1, 2



ISBN 978-5-8149-2111-6

УДК 007(075)
ББК 32.811я73
Г94

Рецензенты:

*А. А. Колоколов*, д-р ф.-м. н., проф., зав. лабораторией дискретной оптимизации Омского филиала Института математики им. С.Л. Соболева СО РАН;

*Е. М. Раскин*, к. т. н., директор ООО «Автоматика-Э»

**Гуменюк, А. С.**

Г94 Теория информации и кодирования: Учебное пособие / А.С. Гуменюк, Н.Н. Поздниченко ; Минобрнауки России, ОмГТУ. – Омск: Изд-во ОмГТУ, 2015.

ISBN 978-5-8149-2111-6

Приведены разработки авторов, развивающие теорию М. Мазура и представляющие формальные средства описания и исследования упорядоченных массивов данных – анализ строя информационных цепей. Представлены традиционные разделы – элементы математической теории связи К. Шеннона и теории кодирования. Описаны формальные средства измерения числа информаций и показана фундаментальная связь формул К. Шеннона, М. Мазура и характеристик строя цепи. Теоретические положения учебного пособия сопровождаются большим числом практических примеров, контрольных вопросов и заданий.

Предназначено для студентов бакалавриата и магистрантов, обучающихся по направлениям 09.03.01, 09.04.01 – «Информатика и вычислительная техника» и изучающих дисциплины «Теория информации и кодирования», «Прикладная теория информации» и «Теоретические основы информационных процессов».

УДК 007(075)
ББК 32.811я73

*Рекомендовано редакционно-издательским советом Омского государственного технического университета*

ISBN 978-5-8149-2111-6

1 электронный оптический диск

Оригинал-макет издания выполнен в Microsoft Office Word 2007
с использованием возможностей Adobe Acrobat X.

**Минимальные системные требования:**

- процессор Intel Pentium 1,3 ГГц и выше;
- оперативная память 256 Мб;
- свободное место на жестком диске 260 Мб;
- операционная система Microsoft Windows XP/Vista/7;
- разрешение экрана 1024×576 и выше;
- акустическая система не требуется;
- дополнительные программные средства Adobe Acrobat Reader 5.0 и выше.

Редактор *В. А. Маркалева*
Компьютерная верстка *О. Н. Савостеевой*

Сводный темплан 2015 г.
Подписано к использованию 24.11.15.
Объем #,# Мб.

---

Издательство ОмГТУ.
644050, г. Омск, пр. Мира, 11; т. 23-02-12
Эл. почта: info@omgtu.ru

# ПРЕДИСЛОВИЕ

Данное учебное пособие содержательно связано с учебным пособием «Прикладная теория информации» и представляет вторую часть материала [1].

В разделе 1 данного учебного пособия представлены разработки авторов на основе идей К. Шеннона, Н. Винера, М. Мазура, Б. Мандельброта, А. Реньи, Ю. Орлова и Ю. Шрейдера, посвящённые формальному анализу расположения компонентов (строя) в информационных цепях (любых упорядоченных массивах данных). Ранее приведено понятие нового абстрактного объекта, названного строем информационной цепи или расположением компонентов в ней, сформулированы выражения числовых характеристик строя [2, 3].

В учебном пособии представлены также традиционные разделы – элементы математической теории связи К. Шеннона; элементы теории кодирования.

В разделе 2 рассматриваются основные понятия и некоторые теоремы математической теории связи [4, 5, 6], включая понятия условной энтропии источника сообщений и энтропии объединения нескольких источников.

В разделе 3 рассматриваются элементы теории кодирования: суть и цели кодирования в технических системах, классификация кодов и кодирующих систем [6, 7]. Подробно рассмотрены вопросы эффективного кодирования: теоремы Крафта – Макмиллена, Шеннона, процедуры построения эффективных кодов. Помехоустойчивое кодирование представлено построением алгебраических групповых кодов. Процессы, рассматриваемые в разделах 2 и 3, представляют собой кодирование в определениях теории Мазура.

Раздел 4 посвящён измерению информаций, используемых при управлении не только для идентификации, но и при описании сообщений. Современной фундаментальной основой для этого являются формулы Хартли, Шеннона, Мазура. Для исследования и обработки упорядоченных массивов данных (информационных цепей) предложены информационные характеристики, основанные на аппарате анализа строя. Они оказываются эффективнее статистических, так как учитывают не только мощности состава последовательностей, но и порядок следования (расположение, строй) их компонентов, что позволяет оценивать

специфичные свойства этих последовательностей. Кроме того, показана связь всех представленных характеристик друг с другом.

Отметим, что традиционный набор разделов теории информационных процессов включает в себя теорию сигналов, теорию информации (К. Шеннона) и теорию кодирования [6–8]. В рамках данного пособия теория сигналов не рассматривается, так как она ориентирована на технических специалистов по связи и обработке сигналов и основана на использовании математического аппарата, требующего отдельного изучения.

Значительная часть излагаемого в пособии материала (разделы 1 и 4) до настоящего времени не представлена в англоязычной и отечественной учебной и даже научной литературе. Новый материал, который дается в пособии апробирован и использовался более 30 лет для студентов 1–4 курсов и магистрантов направления «Информатика и вычислительная техника» при чтении лекций, на практических занятиях, в лабораторных работах и при выполнении курсовых проектов, а также публиковался в соответствующих учебных и научных изданиях [9–12].

# 1. АНАЛИЗ СТРОЯ ИНФОРМАЦИОННЫХ ЦЕПЕЙ

## 1.1. О предмете исследования

В рамках своей теории М. Мазур рассматривает факт упорядоченности сообщений в отдельной информационной цепи для определения условия правильного информирования. При этом условии (см. в [1] п. 1.10) правильное информирование приемника источником возможно, когда соблюдается один и тот же порядок следования сообщений в цепях: начиная от цепи оригиналов, в цепях промежуточных сообщений и заканчивая цепью образов. Данное утверждение поясняется следующими примерами: ***преобразование расположения*** участков местности в аналогичное ***расположение*** тех же участков на карте этой местности, преобразование расположения букв и слов в тексте отображается в расположение во времени соответствующих им звуков говорящего человека.

Вместе с тем очевидна необходимость реализовывать ***определённый порядок следования сообщений в информационных цепях*** (то есть в любых массивах данных) для организации управления (определённую последовательность управляющих воздействий и определённая очередность сбора информации). Примером может служить любой алгоритм, который по определению является последовательностью действий (порядком действия), направленных на достижение определённой цели (например, алгоритм умножения двух чисел в позиционной системе счисления). При участии человека в управлении техническими системами ключевое значение имеют инструкции по эксплуатации, в которых практически всегда действия описываются в виде нумерованных списков, каждый пункт которых является сообщением информационной цепи (примером могут служить рецепты в любой поваренной книге). При сознательном взаимодействии людей (общении) порядок следования слов в текстах и звуков в речи также имеет ключевое значение. Обратным примером, когда непринятие во внимание порядка следования сообщений приводит к нежелательным последствиям, может служить «эффект гонок» как в аппаратных, так и в программных системах, более частным примером являются dead lock’и (взаимные блокировки) в многопоточных приложениях. Примером случая, когда порядок следования сообщений восстанавливается приемником, могут служить некоторые сетевые протоколы, в которых порядок передачи пакетов одного сообщения по сети может быть произвольным, а приемник упорядочивает их по номерам, которые были присвоены пакетам по порядку перед отправлением.

В данном разделе представлены разработки авторов на основе идей К. Шеннона, Н. Винера, М. Мазура, Б. Мандельброта, А. Реньи, Ю. Орлова, Ю. Шрейдера, направленные на исследование порядка следования сообщений в отдельных

информационных цепях [2, 3]. ***Известные средства и модели математики*** (статистический, корреляционный, математический, спектральный и фрактальный анализ, марковские цепи, потоки заявок и теория очередей, методы математической лингвистики, взвешенные графы) ***непосредственно не учитывают порядок следования компонентов исследуемых объектов***. Такое положение в некоторой степени объясняется ***отсутствием формализма для абстрактного объекта, представляющего порядок следования сообщений в информационной цепи*** и называемого «строем», «построением цепи» или «расположением компонентов». Следует отметить, что разные по природе последовательности событий с одинаковыми статистическими распределениями (в дальнейшем – с равномощными составами) могут иметь один и тот же оригинальный строй. Очевидно также, что множество, которое содержит повторяющиеся элементы (мультимножество), может быть основой для построения различных комбинаций (последовательностей, кортежей) типа «перестановки с повторениями». При этом многие из них будут иметь разное расположение компонентов.

В данном разделе рассматривается подход, который предназначен для формального анализа расположения сообщений в информационной цепи произвольной природы. Такой цепью может быть текст, нуклеотидная последовательность, нотная запись, массив данных измерений и массив данных любой размерности, если его можно преобразовать в одномерный. В последнем случае это могут быть картины, изображения, карты местности, видеоизображения и рельефы различных поверхностей.

## 1.2. Понятие и формализмы строя цепи

Рассмотрим на рис. 1.1 несколько примеров разных по природе упорядоченных множеств символьных последовательностей, сигналов сложной формы и диаграмм.

На первый взгляд ничего общего в построении рассмотренных символьных последовательностей нет. Однако вначале даже поверхностное рассмотрение позволяет определить одинаковую длину этих кортежей (это $n$-ки, где $n = 17$). Дальнейшее исследование показывает одинаковую ***мощность состава этих кортежей – мощности алфавитов и наборов чисел вхождений***, представленных в форме ряда распределения и графиков на рис. 1.2. В первой строке ряда расположены номера символов по мере их встречи в последовательности при чтении «слева направо» или «сверху вниз», эти же номера используются в качестве значений на оси абсцисс гистограммы. Заметим, что сами символы и любые другие события являются неупорядоченными множествами и построить для них графическое изображение распределений невозможно.

$$j_{max} = m = 7,$$

| $j$-е компоненты кортежей | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Числа вхождений $n_j$ | 2 | 1 | 4 | 3 | 1 | 1 | 5 |

$$\sum_{j=1}^{m} n_j = 17.$$

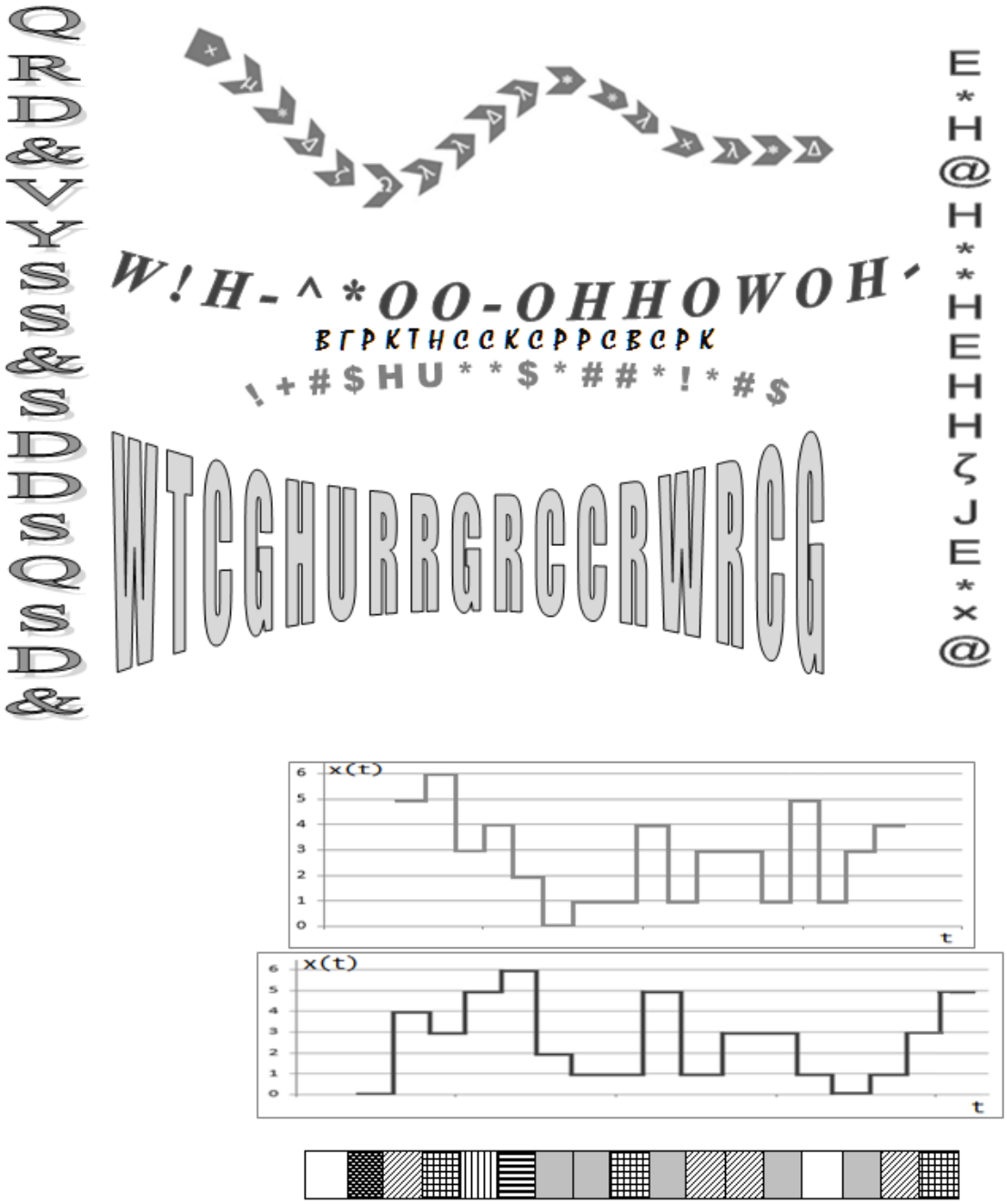



*Рис. 1.1.* Примеры символьных последовательностей, сигналов и диаграмм

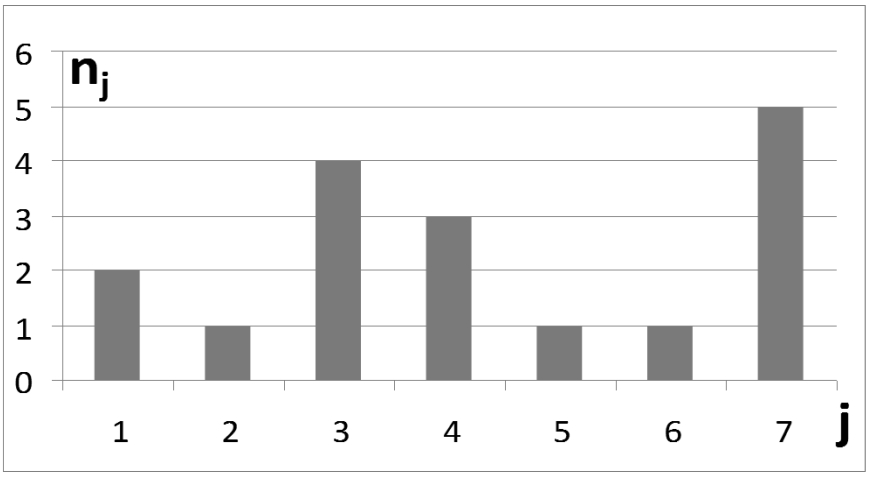



*Рис. 1.2.* График ряда распределения
*j*-х компонентов кортежей

Более глубокое изучение примеров, приведенных на рис. 1.1, позволяет обнаружить одинаковое расположение компонентов этих знаковых цепей, сигналов и диаграмм с помощью представленного на рис. 1.3 графического преобразования.

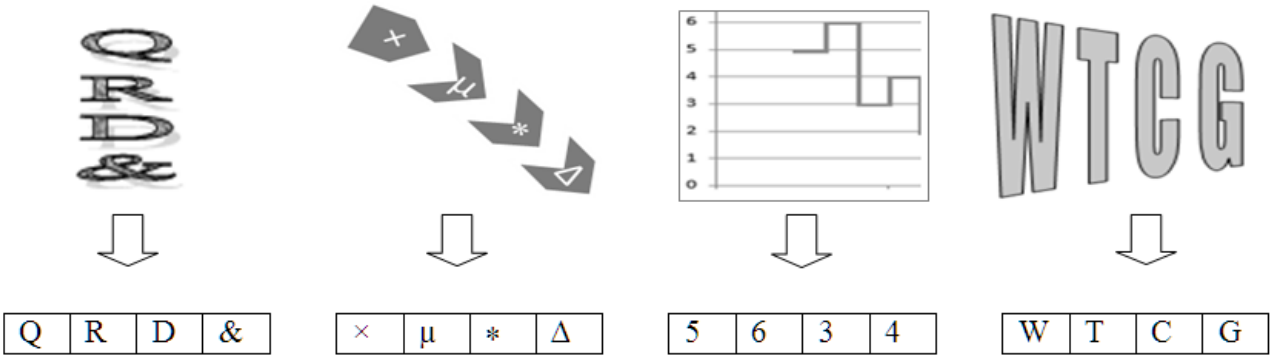



*Рис. 1.3.* Преобразование исходных объектов разной природы в строковый вид

***Строй цепи*** сообщений (событий, знаков и т. п.) – это кортеж (упорядоченное множество), в котором каждому компоненту цепи поставлено в соответствие натуральное число, причем идентичные по выбранному признаку компоненты отображены одним и тем же числом. Первый компонент кортежа – единица, каждый следующий компонент цепи, отличный от всех предыдущих, обозначается натуральным числом, которое на единицу больше максимального из расположенных ранее в кортеже.

В соответствии с определением для формирования строя необходимо учитывать следующие ограничения:

1. ***Алфавит строя*** – это множество всех натуральных чисел из диапазона от 1 до $m$ $\{1, 2, 3, 4, 5, \ldots, m\}$.

2. Мощность алфавита $m$ всегда не больше длины строя $m \leq n$ (предельный случай, когда длина строя равна размеру алфавита ($m = n$) и все элементы (числа) встречаются в строе один раз).

3. Первые вхождения элементов алфавита располагаются на позиции строя по возрастанию, начиная с единицы в первой позиции, возможно с пропусками некоторых мест:

⟨1 2 – 3 – –4 – – – –5 – – – – – – –6 7⟩.

4. Не занятые первыми вхождениями элементов алфавита места на позиции строя заполняются натуральными числами, по значению не превышающими максимального натурального числа среди всех лежащих слева чисел:

⟨1 2 1 3 2 3 4 4 4 1 5 3 4 5 1 1 1 6 7⟩.

***Мощность алфавита строя*** – это количество различных компонентов в цепи.

Примеры разных последовательностей (кортежей) символов с одинаковым строем приведены на рис. 1.4 и 1.5.

Для сравнения по строю нескольких кортежей реальных сообщений, необходимо правильно выполнить однозначное ***прямое преобразование*** для каждого из них, а затем сравнить полученные строи.

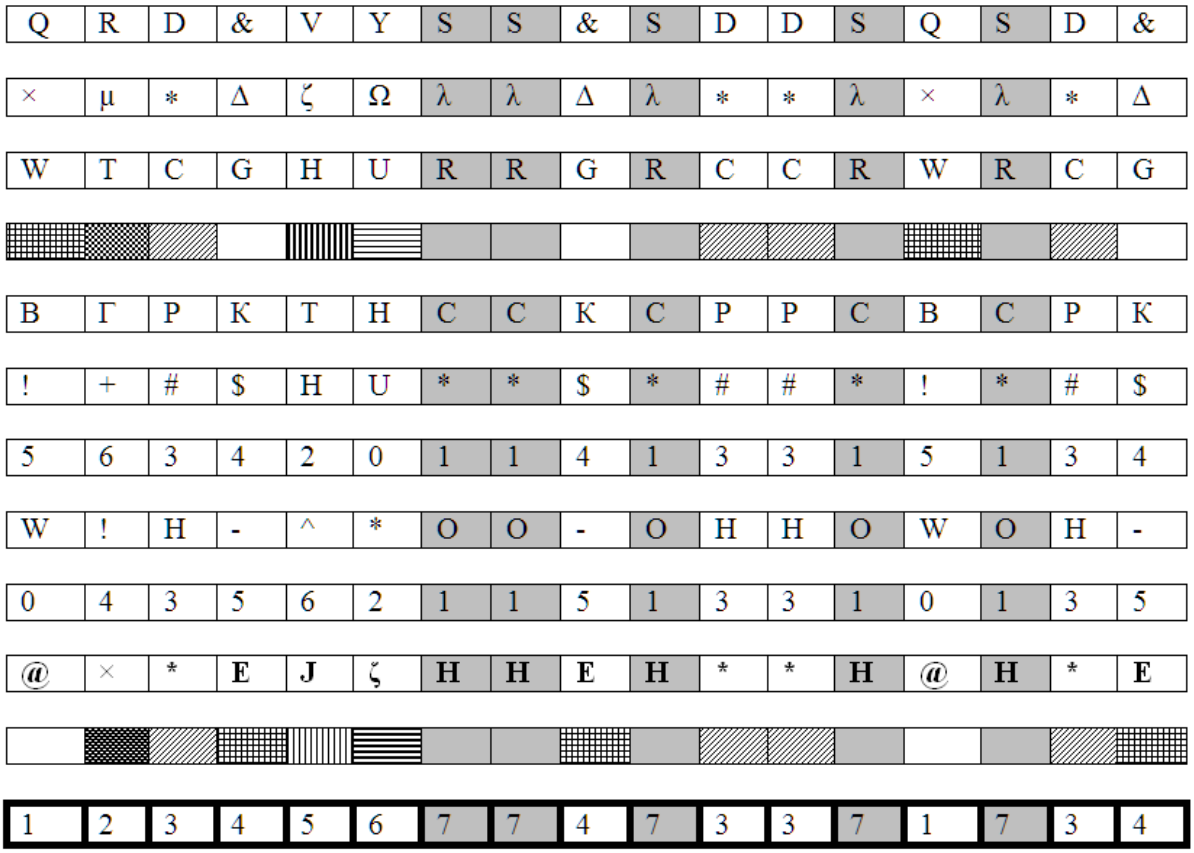



*Рис. 1.4.* Пример прямого преобразования 11 разных знаковых цепей, цифровых последовательностей и диаграмм в строй цепи

| A | C | C | T | G | A | C | T | G | C | T | A | T | C | G | G | A | T | T | G | A | T | A | C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | G | G | A | C | T | G | A | C | G | A | T | A | G | C | C | T | A | A | C | T | A | T | G |
| 1 | 2 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 2 | 3 | 1 | 3 | 2 | 4 | 4 | 1 | 3 | 3 | 4 | 1 | 3 | 1 | 2 |

*Рис. 1.5.* Фрагмент двух комплементарных цепочек РНК
бактерии Candidatus nitrosopumilus maritimus с одинаковым строем

Для кодирования разных знаков при прямом преобразовании цепи сообщений в строй цепи возможно, кроме натуральных чисел, использовать любой (упорядоченный) алфавит символов достаточно большой мощности. Соответствие между исходной и закодированной таким образом последовательностями называется «совпадением с точностью до переименования». Однако такой алфавит необходимо выбрать или специально построить и самое трудное – сделать его общепринятым. Кроме того, все реальные алфавиты и словари неявно упорядочены натуральными числами для удобства запоминания и использования.

Несколько отвлекаясь от темы данного раздела, заметим, что 12 кортежей, изображённых на рис. 1.4, можно рассматривать как искусственный пример, графически представляющий воздействие источника на приёмник как конечное множество сообщений мощностью $12 \cdot 17 = 204$. Это множество можно разбить на два вида подмножеств: поперечные (их 12) и продольные (их 17). Первое сверху – это подмножество оригиналов, если полагать, что эти сообщения формируются на выходе источника. Будучи упорядоченным это же подмножество является информационной цепью оригиналов. Последнее снизу – подмножество образов, если полагать, что эти сообщения формируются на входе приёмника. Будучи упорядоченным это же подмножество является информационной цепью образов. Остальные (сверху вниз) – это поперечные подмножества промежуточных сообщений. Продольные подмножества представлены в столбцах; если рассматривать их как упорядоченные множества, то они являются кодовыми цепями. Ещё раз заметим, что данное множество сообщений не представляет никакую реальную цепь управления.

В теоретико-множественном представлении вектором называется кортеж, компонентами которого являются числа. В соответствии с таким определением вектора, назовем специфически сформированный (организованный) кортеж «***вектором строя***».

Таким образом, строй цепи и вектор строя – это синонимы одного и того же абстрактного объекта. Однако на практике следует различать «вектор строя» данной цепи или некоторого их множества и «вектор строя» как элемент множества разных векторов строя.

Заметим, что при несоблюдении ограничений на порядок расположения натуральных чисел мы получим кортеж, точнее вектор, который не представляет собой строй. Для примера на рис. 1.6 представлен такой вектор.

| 1 | 2 | 3 | 5 | 4 | 6 | 7 | 7 | 5 | 7 | 3 | 3 | 7 | 1 | 7 | 3 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

*Рис. 1.6.* Вектор натуральных чисел,
который не представляет строй цепи

Рассмотрим отличный от представленного на рис. 1.4 строй цепи. Очевидно неоднозначное преобразование уникального строя в знаковые последовательности. Для наглядности, пусть они имеют мощность состава элементов такую же, что и в примере (см. рис. 1.4). Условимся называть преобразование строя в знаковую цепь «***обратным преобразованием строя***» (рис. 1.7).

| 1 | 2 | 3 | 4 | 5 | 1 | 5 | 6 | 7 | 7 | 1 | 5 | 2 | 7 | 5 | 5 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| & | Q | R | V | S | & | S | Y | D | D | & | S | Q | D | S | S | D |
| $ | ! | U | H | * | $ | * | + | # | # | $ | * | ! | # | * | * | # |

*Рис. 1.7.* Обратное преобразование данного строя
в 12- и 13-знаковые последовательности

Из примеров и здравого смысла очевидно, что ***при одинаковой мощности составов знаковых цепей их частотные распределения одинаковы***, т.е. ***инвариантны относительно расположения элементов в цепях.***

***Строй цепи – это идея или план построения*** некоторого множества кортежей, цепей реальных сообщений, сигналов или событий. ***Строй цепи*** в определённом смысле соответствует введенному Гёте понятию «***архетип***». Этот термин предложил использовать Юлий Шрейдер для обозначения описания структуры таксона (класса объектов) [13]. Другими словами, ***строй цепи – это ее архетип***.

В данном разделе рассматриваются только информационные цепи, т.е. одномерные последовательности данных. Само получение строя для многомерных массивов данных не представляет проблем, однако, по причине неоднозначности выбора «траекторий чтения», дальнейший анализ и сравнение разных «построений цепей-траекторий» является нетривиальной задачей. Кроме того, следует заметить, что современные инструментальные исследования движения глаз показали, что изображение (картинка или текст) обрабатывается непараллельно. Оно сканируется по определённой сложной траектории, форма которой зависит от композиции компонентов изображения. При этом отдельные и связанные элементы траектории сканирования (информационной цепи) воспринимаются очень компактной группой светочувствительных элементов,

расположенных в центре сетчатки глаза [14,15]. Альтернативой произвольному чтению (сканированию) многомерного массива человеком является его формальное отображение в одномерный, однако и здесь возможно применение множества различных способов построения проекции. На практике отображение многомерных данных в одномерный массив осуществляется общепринятыми методами [16, 6], которые закрепляются с помощью образовательных систем. При этом они, как правило, не имеют под собой специального обоснования среди множества других способов, кроме технических ограничений и ограничений, отмеченных Мазуром (соображениями удобства). Наконец, отметим, что кроме регулярного сканирования может применяться также случайное чтение (поиск).

Очевидно, возможно осуществлять прямое преобразование текстов и других знаковых последовательностей по траектории чтения кортежа «справа налево», как это принято в арабских языках и текстах на иврите, или «сверху вниз» – для китайских и японских текстов. Для примера прочтем строй (см. рис. 1.4) по направлению «справа налево» и в результате получим кортеж, а точнее вектор, который по определению не является строем цепи (при чтении «слева направо») (рис. 1.8). Расположим рядом вектор строя этой цепи. Из примера видно, что при различных направлениях чтения расположение компонентов в одной и той же последовательности воспринимается по-разному.

| 4 | 3 | 7 | 1 | 7 | 3 | 3 | 7 | 4 | 7 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

| 1 | 2 | 3 | 4 | 3 | 2 | 2 | 3 | 1 | 3 | 3 | 5 | 6 | 1 | 2 | 7 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

*Рис. 1.8.* Вектор строя при чтении «справа налево»
и вектор строя при чтении «слева направо»
(вектор строя и «обратный» к нему)

Разобьем множество символьных последовательностей с равномощными составами $S_k$ на непересекающиеся подмножества $\{S_u\}$ (где $S_u \subset S_k$) по принципу соответствия тому или иному строю. В результате все равномощные по составу знаковые последовательности $S_u$, имеющие одинаковое расположение элементов (композицию), будут отображены одним и тем же строем, который представляет уникальный для данного подмножества цепей порядок следования (построение символов, слов, сообщений). В случае неограниченного набора символов алфавитов подмножество $S_u$ и множество $S_k$ являются счётными; в случае, когда набор символов алфавитов фиксирован, $S_u$ и $S_k$ будут конечными множествами.

Строй знаковой последовательности или информационной цепи сообщений характеризуется мощностью ее состава (но не только алфавита) и оригинальной композицией его компонентов. Все кортежи некоторого счётного или конечного (в случае фиксированного перечня символов алфавита) подмножества $S_u$ отмеченного разбиения ***инвариантны или изоморфны относительно своего строя***, то есть совпадают с точностью до переименований компонентов. Заметим, что данному подмножеству информационных цепей принадлежит и их собственный строй; однако только он представляет в отмеченном смысле строение цепи – порядок следования разных сообщений. Наконец отметим, что знаковые кортежи с разными по мощности составами (но, возможно, одинаковыми или равномощными алфавитами или словарями) по определению отличаются строем цепей.

В более широком, неформальном смысле, конкретный строй фиксирует определённый порядок следования событий в цепи. Это могут быть последовательности акустических элементов речи человека, хроники исторических событий, тексты, последовательности данных о субъективном состоянии человека, цепи сообщений, массивы значений измеряемых величин.

Таким образом, для исследования построения реального массива данных ***вводится особое формальное понятие – строй*** знаковой последовательности, который представляет только определенный порядок следования, расположение различимых и одинаковых его компонентов без учета их обозначений и содержимого.

### 1.2.1. Однородные последовательности и интервально-матричная модель строя цепи

Операция выявления в разных по природе информационных цепях одинаковых построений расширяет возможности междисциплинарных исследований. Однако результат такой операции ограничен описанием строя в форме обычного числового кортежа, хотя и имеющего определение «вектор строя». Рассмотрим более удобное для анализа формальное описание строя, которое позволяет получать компактные числовые характеристики (подобные используемым для описания случайных величин), полезные, в частности, при опознавании строёв цепей и определении степени их различия. Для определения такого формализма строя отдельной цепи при обычном (естественном) способе её чтения «поэлементно» (подряд) введем две нумерации:

- первая нумерует элементы собственного алфавита (словаря) данной знаковой последовательности по мере их первой встречи;
- вторая дает сквозную нумерацию всех компонентов кортежа от начала до конца.

Разложим ***полную неоднородную*** (без пустых мест на позиции) символьную последовательность на $m$ ***неполных однородных кортежей***, на позициях которых заняты одинаковыми знаками только некоторые места (рис. 1.9). Такое разложение цепи называют ***декомпозицией.*** Аналогом однородной последовательности является поток однородных событий (заявок), определенный в теории массового обслуживания. Очевидно, что ***композиция*** всех однородных строёв данного полного строя даёт полный неоднородный строй, аналогом которого в теории очередей является неоднородный поток событий [17]. Вообще разложение цепи может осуществляться по разным правилам. Декомпозиция строя полной неоднородной знаковой цепи на неполные однородные представлена на рис. 1.9, ***разнородные (неполные неоднородные)*** цепи – на рис. 1.10. В последнем случае те места позиции, которые заняты разными знаками, заполняются по следующему правилу: при просмотре цепи от ее начала, в состав первой разнородной цепи выбираются все первые вхождения каждого элемента алфавита, при втором – все вторые вхождения и т. д.


*Рис. 1.9.* Декомпозиция строя неоднородной знаковой цепи на неполные однородные цепи и матрица их интервалов


*Рис. 1.10.* Декомпозиция строя неоднородной знаковой цепи на разнородные цепи и матрица их интервалов

В приведенных на рис. 1.9, 1.10 примерах используется привязка к концу последовательности, то есть последний интервал считается от последнего вхождения компонента до конца последовательности. Кроме данной привязки могут также использоваться следующие варианты: к началу, к началу и к концу, циклическая, либо отсутствие привязки.

Определим ***«интервал»*** как расстояние от выделенного в цепи компонента, до другого ближайшего, отмеченного в направлении просмотра (см. рис. 1.9); ***величина интервала*** – это натуральное число, определённое как модуль разности номеров мест двух выделенных компонентов на позиции кортежа. В дальнейшем для краткости будем называть эти понятия одинаково – интервал.

Назовем направление чтения текста или знаковой цепи «подряд слева направо» «обычным способом считывания». Пусть первое считывание текста осуществляется отличным от обычного способом с самого начала до конца таким образом, что выбираются только элементы строя с номером «1»; при этом последний интервал определяется до знака «финиш» (возможен и другой вариант – определение первого интервала от начала текста – «старта»). Интервалы данной однородной последовательности разместим в соответствии с номерами считываемых элементов в первой строке матрицы. Далее, при втором просмотре строя текста аналогично выберем элементы с номером «2» и разместим вектор интервалов, соответствующий другой однородной последовательности, во второй строке матрицы. В каждой следующей строке помещается вектор интервалов «новой» при очередном просмотре однородной последовательности. Одиночные знаки, слова или сообщения будут представлены всего одним интервалом (до финиша), который размещается в крайнем столбце соответствующей строки матрицы. Число столбцов $n_{j\,max}$ в ***«матрице интервалов» однородных цепей*** равно числу вхождений самого частого знака (или слова) текста. Незанятые интервалами элементы матрицы заполним нулями. Число строк $m$ равно мощности собственного алфавита или словаря текста. Результаты описанных действий представлены на рис. 1.10.

Пусть считывание текста осуществляется вторым (описанным выше) способом по разнородным цепям. В результате получим ***матрицу интервалов разнородных цепей***, в которой число столбцов равно $m$, а число строк – $n_{j\,max}$. В случае правильного выполнения декомпозиций, полученные множества однородных (разнородных) последовательностей (рис. 1.11, 1.12) будут несовместными (так как не содержат занятых мест с одинаковыми номерами на их позициях). ***Композиция*** или «совмещение» всех неполных однородных (неоднородных) кортежей дает исходную полную неоднородную знаковую последовательность.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **T** | **T** | **G** | **G** | **G** | **T** | **T** | **C** | **C** | **G** | **G** | **G** | **G** | **G** | *Cricetulus griseus* |
| **G** | **G** | **A** | **A** | **A** | **G** | **G** | **T** | **T** | **A** | **A** | **A** | **A** | **A** | *Homo sapiens* |
| **1** | **1** | **2** | **2** | **2** | **1** | **1** | **3** | **3** | **2** | **2** | **2** | **2** | **2** | строй, общий для обоих фрагментов |

*Рис. 1.11.* Фрагменты нуклеотидных цепей *Cricetulus griseus* и *Homo sapiens* с одинаковым строем (длина фрагментов 15). Выделены путём просмотра рибосомальной РНК общей длиной 1871и 1559. Совпадение строя фрагментов начинается с позиций: 1157 для первой цепочки и 778 – для второй цепочки

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | - | - | - | - | G | G | - | - | - | - | - | - | - | G | G | - | - | G |
| A | - | - | - | - | A | A | - | - | - | - | - | - | - | A | A | - | - | A |
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

*Рис. 1.12.* Однородные фрагменты нуклеотидных цепей *Cricetulus griseus* и *Homo sapiens* с одинаковым строем (длина фрагментов 19). Совпадение строя начинается с позиций: 108 для первой цепочки и 1847 – для второй цепочки

Следует отметить, также описание знаковых цепей и текстов графами. Обычно это взвешенный граф типа «дерево», узлы которого представляют выделенные по определенным признакам символы или слова, а ребра – это интервалы между ними.

Ещё раз подчеркнём, что ***для исследования построения реальной информационной цепи вводится формальное понятие – строй знаковой последовательности***, который представляет только определенный порядок следования, взаимное расположение различимых и одинаковых его компонентов без учета их обозначений и содержимого. Заметим, что цели и методы исследования строя, если его рассматривать как обычный кортеж, не отличаются от исследований реальных текстов и знаковых цепей и т. п. Если рассматривать строй как новый абстрактный объект, отображающий информационную цепь, то открывается возможность исследовать и использовать его особые свойства, в том числе применять новые формулы для подсчёта информации в массиве данных.

## 1.3. Числовые характеристики строя информационных цепей

### 1.3.1. Интегральные характеристики строя

Используем понятие однородной знаковой последовательности и векторное отображение ее строя в виде соответствующей строки матрицы интервалов для определения числовых характеристик строя текста.

Перемножением всех интервалов выделенной $j$-й однородной последовательности (элементов соответствующей ей строки матрицы, кроме нулевых) определим ***абсолютный объем строя $j$-й однородной цепи*** в виде

$$V_j = \prod_{i=1}^{n_j} \Delta_{ij}, \tag{1.1}$$

где $\Delta_{ij}$ – интервал от $i$-го до $(i+1)$-го вхождения $j$-го символа; $n_j$ – число вхождений $j$-го символа.

***Средний геометрический интервал*** между занятыми местами на позиции строя однородной цепи определяется в виде

$$\Delta_{gj} = \sqrt[n_j]{V_j}. \tag{1.2}$$

***Абсолютный объем строя текста*** определим как произведение абсолютных объемов всех строев $j$-ых однородных последовательностей в виде

$$V = \prod_{j=1}^{m} V_j, \tag{1.3}$$

при подстановке (1.1) в (1.3) получим

$$V = \prod_{j=1}^{m} \prod_{i=1}^{n_j} \Delta_{ij}, \tag{1.4}$$

где $m$ – мощность алфавита (собственного словаря текста).

***Средний геометрический интервал строя цепи*** на множестве всех однородных цепей текста определяется в виде

$$\Delta_g = \sqrt[n]{V}, \tag{1.5}$$

где $n$ – длина текста, равная числу мест для размещения всех слов (знаков) на его позиции.

***Средний арифметический интервал*** строя $j$-й однородной цепи определяется в виде

$$\Delta_{aj} = \frac{n}{n_j} = \frac{1}{P_j}, \tag{1.6}$$

где $P_j$ – частота вхождения или статистическая вероятность вхождения $j$-го элемента в цепи.

***Периодичность*** (следования одинаковых элементов) строя *j*-й однородной цепи $\tau_j$ определим из (1.2) и (1.6) отношением среднего геометрического и среднего арифметического интервалов:

$$\tau_j = \frac{\Delta_{gj}}{\Delta_{aj}}. \tag{1.7}$$

***Регулярность*** (следования одинаковых элементов (4.20) строя полной неоднородной цепи определим из (1.5) и формулы (4.20) для числа описательных информаций *D* (по М. Мазуру [16]) отношением вида

$$r = \frac{\Delta_g}{D}, \tag{1.8}$$

где $\Delta_g \leq D$, как будет показано далее в формулах (4.40), (4.44) и (4.50).

Логарифмирование представленных величин дает набор удобных для практики компактных аддитивных информационных характеристик строя цепи. При этом интервал соответствует ***удаленности определенного j-го символа i-го вхождения*** относительно его $(i+1)$-го вхождения в виде

$$g_{ij} = \log_2 \Delta_{ij}; \tag{1.9}$$

объем строя однородной последовательности выделенного *j*-го символа – ***глубине расположения строя j-й однородной цепи*** в виде

$$G_j = \log_2 V_j, \tag{1.10}$$

при подстановке (1.1) получим

$$G_j = \sum_{i=1}^{n_j} \log_2 \Delta_{ij}. \tag{1.11}$$

Объем строя текста соответствует ***глубине расположения строя всей цепи*** в виде

$$G = \log_2 V; \tag{1.12}$$

при подстановке (1.4) в (1.12) получим

$$G = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta_{ij}. \tag{1.13}$$

При сравнении строёв разных знаковых цепей и текстов могут быть полезны оценки ***относительных глубин расположения*** в виде

$$\delta G_j = G_j / G, \tag{1.14}$$

а также оценки средних удаленностей соседних одинаковых элементов в строе $j$-й однородной цепи, которые отображают средние геометрические интервалы. Из выражений (1.2) и (1.11) ***средняя удаленность выделенного j-го символа*** в строе однородной последовательности определяется в виде

$$g_j = \log_2 \Delta_{gj} = \frac{1}{n_j} \sum_{i=1}^{n_j} \log_2 \Delta_{ij}\,. \tag{1.15), (1.16}$$

Из выражений (1.5) и (1.13) ***средняя удаленность отдельного символа*** в строе данной цепи или текста определяется в виде

$$g = \log_2 \Delta_g = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta_{ij} = \frac{1}{n} \sum_{j=1}^{m} n_j \cdot g_j. \tag{1.17), (1.18}$$

Отношение средней удаленности выделенного $j$-го символа к средней удаленности отдельного символа в виде

$$\delta g_j = g_i / g, \tag{1.19}$$

дает характеристику строя, которая дополняет частоту вхождения $P_j = n_j / n$ и названа ***относительной удаленностью j-го символа***.

Из (1.11), (1.15) и (1.16) следует, что

$$G_j = n_j \cdot g_j, \tag{1.20}$$

а из (1.13), (1.17) и (1.18), что

$$G = n \cdot g. \tag{1.21}$$

Заметим, что с учетом выражений (1.20) и (1.21) для строя любой полной знаковой последовательности (без пустых мест на позиции) ***сумма произведений частот вхождения выделенных элементов на их относительные удаленности равна единице***, так как

$$\sum_{j=1}^{m} \frac{G_j}{G} = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \frac{g_j}{g} = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \frac{g_j}{g} = \sum_{j=1}^{m} P_j \cdot \delta g_j = 1\,. \tag{1.22}$$

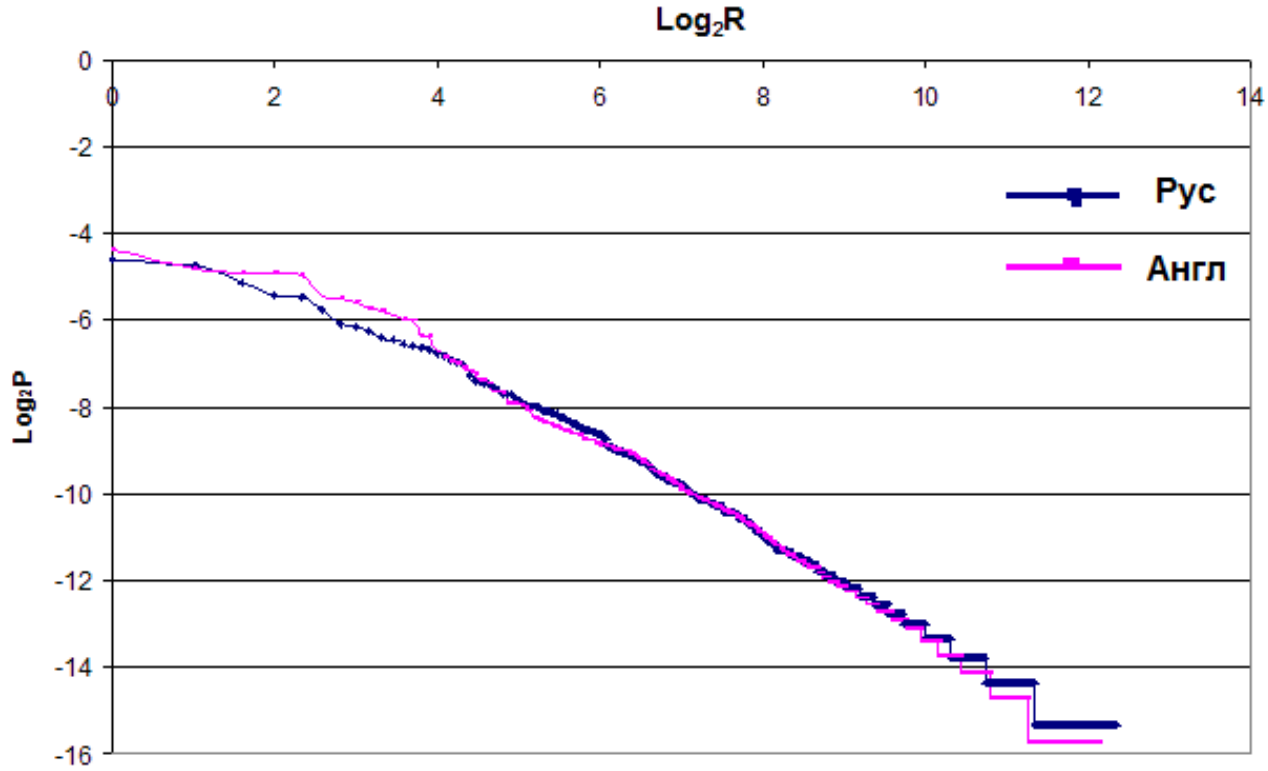
### 1.3.2. Распределения характеристик строя

Множество величин $\{G_j\}$ всех однородных последовательностей является ***распределением глубин*** однородных последовательностей для строя отдельной цепи и представляет, в некотором смысле, его «полное» описание. Множество величин $\{g_j\}$ является ***распределением средних удаленностей*** однородных последовательностей. При анализе и описании строя данной цепи распределение средних удаленностей дополняет обычное частотное распределение знаков $\{n_j\}$ ***комплексным распределением строя*** вида $\{\langle n_j, g_j \rangle\}$.

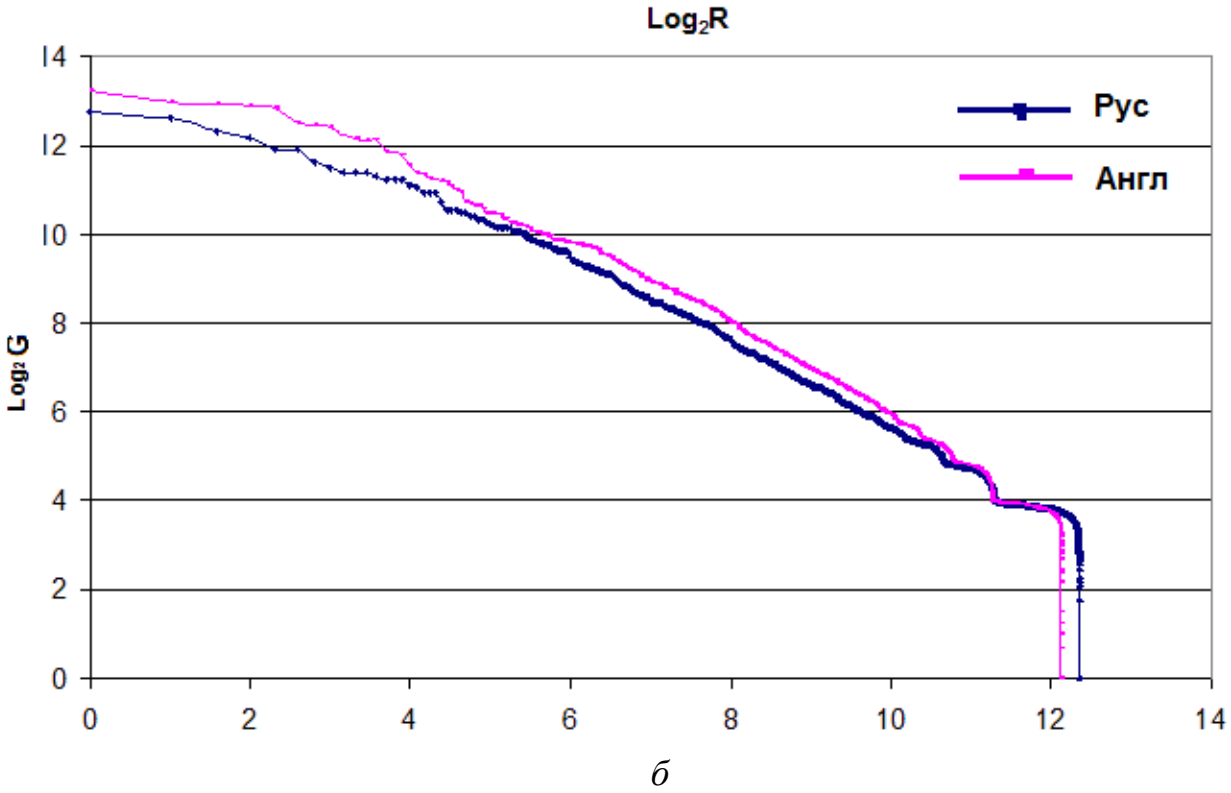
Для примера на рис. 1.13 приведены графики частотно-ранговых распределений (рис. 1.13, *а*) и глубинно-ранговых распределений (рис. 1.13, *б*) нормализованных текстов романа Ф.М. Достоевского «Бедные люди» и его перевода на английский язык.

Графики представлены в двойном логарифмическом масштабе. По оси абсцисс располагается двоичный логарифм ранга слова (рис. 1.13, *а*) и соответствующей однородной цепи (см. рис. 1.13, *б*); на рис. 1.13, *а* по оси ординат – двоичный логарифм частоты слова $P_j$, на рис. 1.13, *б* – двоичный логарифм глубины $G_j$ однородной цепи выделенного слова.

Легко определяются аналогичные по форме, но отличные по содержанию, числовые характеристики строя данной цепи с использованием векторов интервалов, представленных строками матрицы для разнородных цепей. В таком случае формулы (1.1)–(1.22) определяются при выделении знаков $j$ в $i$-ых разнородных цепях и представляют ***числовые характеристики строя на основе разнородных цепей,*** а также его «полное» описание с помощью аналогичных распределений.




*а*




*б*

*Рис. 1.13.* Примеры ранговых распределений оригинала и его перевода

### 1.3.3. Функции характеристик строя

Общепринятым методом исследования больших массивов данных измерений, лингвистических текстов, нуклеотидных последовательностей и длинных цепей другой природы является «просмотр окном». Здесь представлены средства для анализа локальной структуры целостных полноразмерных последовательностей на основе характеристик строя отдельных, но связных фрагментов ($L$-грамм).

Для формального определения функции характеристики строя используются следующие понятия:

- ***место*** – элементарная ячейка, предназначенная для хранения одного компонента цепи;
- ***позиция*** – это упорядоченное множество мест;
- ***фрагмент*** – участок полной цепи;
- ***окно*** – позиция фрагмента (участок позиции полной цепи);
- ***размер окна*** – количество мест на позиции окна;
- ***шаг*** – это смещение окна на позиции полной цепи, позволяющее выделить следующий фрагмент цепи;
- ***размер шага*** – размер смещения окна, измеряемый числом мест;
- ***функция характеристики строя цепи*** – это упорядоченное множество значений характеристик строя, вычисленных для всех фрагментов, последовательно взятых на позиции полной цепи;
- ***отсчётное значение функции характеристики строя*** – это значение функции характеристики строя, вычисленное для отдельного фрагмента, задаваемого его номером, длиной, и размером шага.

Ниже приведены формулы для вычисления отсчётных значений некоторых функций характеристик строя.

$$\Delta_{ij} = x_{i+1j} - x_{ij};\ x_{i+1j}, x_{ij} \in [s * k, s * k + l]; \tag{1.23}$$

$$f_G(k,l,s) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} log_2 \Delta_{ij}\,; \tag{1.24}$$

$$f_g(k,l,s) = f_G(k,l,s)/l\,; \tag{1.25}$$

$$f_{\Delta_g}(k,l,s) = \sqrt[l]{\prod_{j=1}^{m} \prod_{i=1}^{n_j} \Delta_{ij}}\,; \tag{1.26}$$

$$f_D(k,l,s) = \sqrt[l]{\prod_{j=1}^{m}\left(\frac{l}{n_j}\right)^{n_j}}; \quad (1.27)$$

$$f_r(k,l,s) = f_{\Delta_g}(k,l,s)/f_D(k,l,s), \quad (1.28)$$

где $x_{ij}$ – номер места $i$-го вхождения $j$-го элемента алфавита на позиции данного фрагмента; $k$ – номер фрагмента; $s$ – размер шага (при $s = 1$ фрагменты являются $L$-граммами); $l$ – размер окна; $f_G(k,l,s)$ – функция глубины; $f_g(k,l,s)$ – функция средней удалённости; $f_{\Delta_g}(k,l,s)$ – функция среднего геометрического интервала; $f_r(k,l,s)$ – функция регулярности; $f_D(k,l,s)$ – функция числа описательных информаций.

Общее количество фрагментов при заданных длине цепи, шаге и размере фрагмента определяется в виде

$$k_{max} = ]^n/_s - l + s[. \quad (1.29)$$

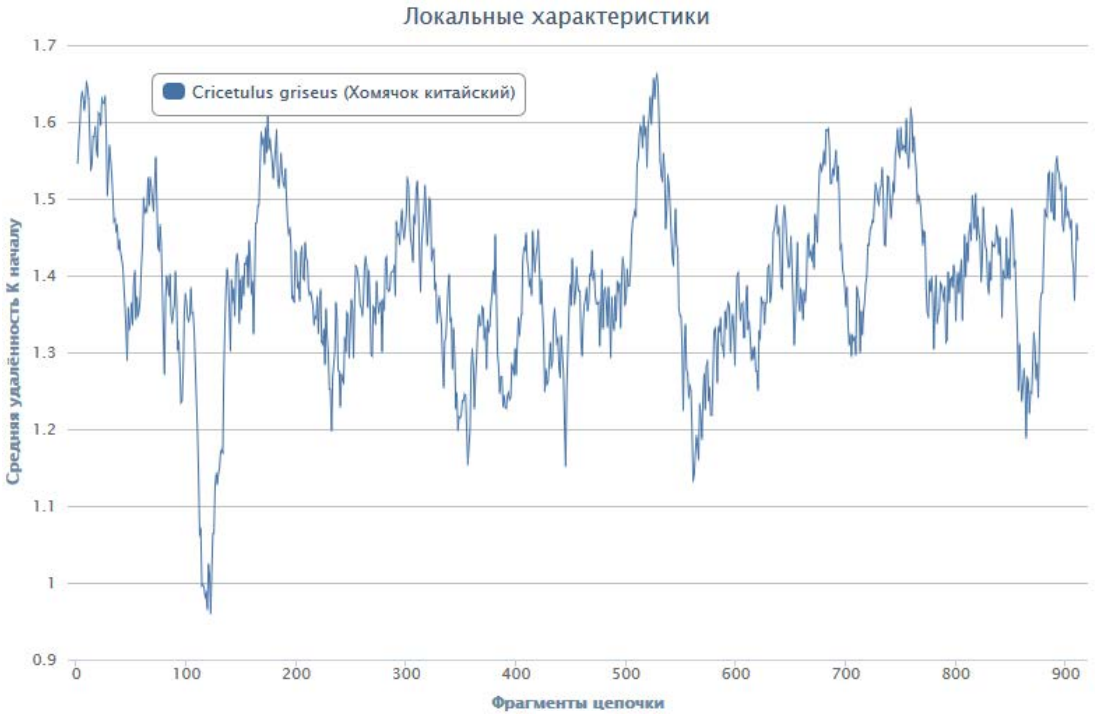
Заметим, что мощность $m$ алфавита данного фрагмента может быть меньше мощности алфавита всей цепи (минимум – 1, если фрагмент полностью заполнен одинаковыми компонентами). Аргументы функций $(k,l,s)$ являются натуральными числами. Таким образом, данные функции являются функциями дискретных аргументов. Зная все три параметра, можно вычислить отдельное значение такой функции. Также возможно вычислить многомерную функцию, изменяя не только номер фрагмента, но и два других параметра.

***Отображение строя последовательностей данных функциями, кроме отмеченных средств, позволяет применять также классические методы математики, а именно: математический анализ, спектральный анализ, корреляционный анализ, и т.п., что было бы невозможно при непосредственном анализе самих последовательностей.***

Графики функции удалённости (рис. 1.14, *а*) и функции регулярности (рис. 1.14, *б*) нуклеотидной последовательности, вычисленные с размером окна 50 и шагом 2 представлены на рис. 1.14 . Из рисунка видно, что данные функции независимы и могут дополнять друг друга при комплексном описании локальной структуры массивов данных.

Следует учитывать, что при увеличении длины шага уменьшается количество вычисляемых отсчётных значений, и для выявления совпадающих цепочек и фрагментов при таком разбиении требуется более сложная поисковая проце-

дура, которая может потребовать дополнительных вычислительных ресурсов. Кроме того, при экспертном анализе сокращение количества отсчётных значений упрощает восприятие графического представления функций, вычисленных для длинных последовательностей.




*а*


*б*

*Рис. 1.14.* Графики функций характеристик строя:
*а* – функция удалённости; *б* – функция регулярности


*Рис. 1.15.* Графики функции глубины 18S рибосомальных РНК двух особей одного вида

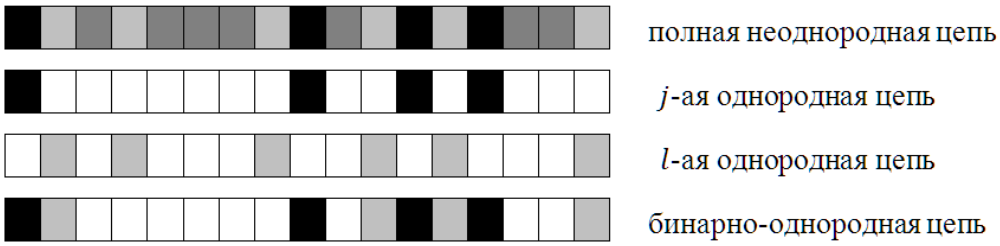

*Рис. 1.16.* Графики функции глубины 18S рибосомальных РНК двух видов, принадлежащих к разным классам одного типа

На представленных графиках видно, что особи одного вида (рис. 1.15) демонстрируют практически идентичную форму функции глубины $f_G$; виды разных классов (рис. 1.16) – сильно отличающуюся форму данной функции.

## 1.4. Характеристики зависимости однородных цепей

В данном подразделе на основе наглядных графических образов (диаграмм) представлены ***характеристики пространственной зависимости*** пары однородных цепей ($j$-й и $l$-й), взятых из состава полной неоднородной цепи [2, 3]. Это позволило соотнести субъективное суждение о степени «пространственной зависимости» элементов сопоставляемых цепей с формальной числовой характеристикой, представляющей такую зависимость.

На приведенной ниже диаграмме выделена в полной неоднородной цепи «бинарно-однородная» цепь (далее бинарная), в состав которой включены только те пары разных знаков $j$ и $l$ сравниваемых однородных цепей, которые определяются знаками $i$ «ближайшими справа» относительно знаков $j$.




Предполагается, что наличие бинарной цепи свидетельствует о такой причинно-следственной связи цепей, при которой хотя бы некоторые компоненты (следствия) $l$-й цепи расположены вслед за компонентами (причинами) $j$-й цепи на определенных интервалах.

Пространственная связь такого типа названа **«*причинной зависимостью*»** событий одной $l$-й цепи от другой $j$-й цепи.

Обозначим: $\Delta(l/j)_i$ – интервал между $i$-ми ближайшими справа знаками $l$ по отношению к знакам $j$; $n(l/j)$ – число пар знаков $j$ и $l$, связанных интервалами $\Delta(l/j)_i$. В примере: $\Delta(l/j)_1 = \Delta(l/j)_3 = 1;\ \Delta(l/j)_2 = 2;\quad \Delta(l/j)_4 = 3.$

Среднее геометрическое значение всех $n(l/j)$ установленных интервалов (бинарной цепи) $\Delta(l/j)_i$ между «смежными» элементами двух однородных цепей определено в виде

$$\Delta(l/j)_{cp} = \sqrt[n(l/j)]{\prod_{i=1}^{n(l/j)} \Delta(l/j)_i}. \qquad (1.30)$$

Интервалы только между теми знаками выделенной однородной цепи $l$, которые являются «ближайшими справа» относительно знаков цепи $j$, обозначены $\Delta(l_j)_i$. В примере: $\Delta(l_j)_1 = 9;\ \Delta(l_j)_2 = 2;\ \Delta(l_j)_3 = 4;\ \Delta(l_j)_4 = 1$. Среднее

геометрическое значение интервалов $\Delta(l_j)_i$ между выделенными элементами $l_j$ в данной однородной цепи определено в виде

$$\Delta(l_j)_{cp} = \sqrt[n(l/j)]{\prod_{i=1}^{n(l/j)} \Delta(l_j)_i}. \tag{1.31}$$

При условии $\Delta(l_i)_{cp} > \Delta(l/j)_{cp}$, разность вида

$$v(l_j) = \left(1 - \frac{\Delta(l/j)_{cp}}{\Delta(l_j)_{cp}}\right) \tag{1.32}$$

представляет ***«избыточность» l-й цепи, зависимой от j-й***. Некоторые элементы такой цепи «связаны справа» с элементами $j$-ой цепи. В противном случае при $\Delta(l_i)_{cp} \leq \Delta(l/j)_{cp}$, данная разность свидетельствует об отсутствии избыточности (диаграммы 1 и 5). Исходя из этого введем ***коэффициент частичной зависимости l-й однородной цепи (от j-й цепи)*** в виде

$$K_1(l/j) = (n(l/j)/n_l) \cdot v(l_j). \tag{1.33}$$

Отметим, что соотношение $n(l/j)/n_l$ представляет собой ***условную вероятность события***, состоящего в появления пары знаков $l$ и $j$, связанных причинной зависимостью, от появления знаков цепи $l$.

Ниже приведены пять диаграмм (1–5) бинарно-однородных цепей, представляющие наглядно-очевидные случаи пространственной зависимости и соответствующие им коэффициенты зависимости.

На диаграммах элементы $j$-й цепи обозначены черными квадратами, а элементы $l$-й – серыми.

Диаграмма 1

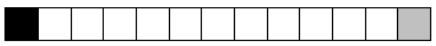


Диаграмма 2

Диаграмма 3

Диаграмма 4

Диаграмма 5

| Завис-ть | $K_1$ | $K_2$ | $K_3$ |
|---|---|---|---|
| От | -11 | -11 | 0 |
| От | 0 | 0 | |
| от | 0,733 | 0,44 | 0,282 |
| от | 0,129 | 0,181 | |
| от | 0,939 | 0,939 | 0,712 |
| от | 0,539 | 0,539 | |
| от | 0,5 | 0,5 | 0,353 |
| от | 0,25 | 0,25 | |
| от | 0,996 | 0,996 | 0 |
| от | -0,6 | -0,6 | |

Если все элементы $l$-й цепи связаны интервалами $\Delta(l/j)_i$ с элементами $j$-й однородной цепи $n(l/j) = n_l$ (диаграммы 2–5), то оценка частичной зависимости становится оценкой ***полной зависимости*** и определяется в виде $K_1(l/j) = \upsilon(l_i)$. Если, кроме отмеченного, числа выделенных элементов сравниваемых однородных цепей будут $n_l = n_j$, то такие цепи считаются полностью взаимозависимыми (диаграммы 3–5).

Причинная зависимость названа ***установленной, определенной*** или ***закономерной***, если размеры всех интервалов $\Delta(l/j)_i$ бинарной цепи задаются определенным операционным преобразованием. Соответственно ***неустановленная*** или ***неопределенная причинная зависимость*** отмечается в такой бинарной цепи, размеры интервалов $\Delta(l/j)_i$ для которой не удалось установить в форме определенного преобразования. Частный случай определенной зависимости назван ***фиксированной причинной зависимостью***. При этом «следование за» в бинарной цепи представлено равными интервалами $\Delta(l/j)_i = \Delta(l/j)_{cp} = const, \forall i = 1, 2, \dots, n(l/j)$ (диаграммы 4 и 5). Предельный случай определенной зависимости назван ***непосредственной причинной зависимостью***; при этом бинарная цепь задается единичными интервалами $\Delta(l/j)_i = 1, \forall i = 1, 2, \dots, n(l/j)$ (диаграмма 5).

Степень зависимости одной цепи от другой с учетом «полноты её участия» в составе обеих однородных цепей, определена в виде

$$K_2(l/j) = \frac{2n(l/j)}{n_j + n_l} \cdot \upsilon(l_j). \tag{1.34}$$

Для примера построена диаграмма 2, в которой числовые характеристики зависимостей $K_1$ и $K_2$ не равны.

Если не требуется учитывать индивидуальную зависимость $l$-й цепи от $j$-й, то вычисляется (средний) ***коэффициент взаимной зависимости*** в виде

$$K_3(j, l) = \sqrt{K_2(l/j) \cdot K_2(j/l)}. \tag{1.35}$$

В случае если подкоренное выражение отрицательно (один из $K_2 < 0$), то цепи считаются взаимно независимыми, и коэффициент $K_3$ искусственно приравнивается к 0 (диаграмма 5).

В результате просмотра около 300 диаграмм подобного вида субъективные суждения о степени зависимости цепей в основном совпадали с формальными оценками.

Субъективная оценка зависимостей в длинных бинарно-однородных цепях практически невозможна. В то же время для реальных текстов формальные оцен-

ки зависимостей пар слов, составляющих бинарно-однородные цепи, оказываются правдоподобными. В качестве иллюстрации приведены фрагменты двух ***матриц зависимостей компонентов*** бинарно-однородных цепей для наиболее зависимых слов известных литературных произведений (табл. 1.1, 1.2).

*Таблица 1.1*

Значения причинной зависимости ($K_1$) для некоторых пар слов из текста повести А.С. Пушкина «Пиковая дама» (в первом столбце слова-причины)

| | *n*(*l*/*j*) / *nl* | *v*(*lj*) | *K*1 | *n*(*l*/*j*) / *nl* | *v*(*lj*) | *K*1 | *n*(*l*/*j*) / *nl* | *v*(*lj*) | *K*1 | *n*(*l*/*j*) / *nl* | *v*(*lj*) | *K*1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **выигрыш** | | | **долг** | | | **игроки** | | | **деньги** | | |
| **карты** | 0,900 | 0,957 | 0,862 | 0,858 | 0,820 | 0,703 | 1,000 | 0,681 | 0,681 | 0,243 | 0,837 | 0,204 |
| | **вечер** | | | **дом** | | | **года** | | | **день** | | |
| **время** | 0,750 | 0,854 | 0,640 | 0,800 | 0,849 | 0,679 | 0,715 | 0,841 | 0,601 | 0,750 | 0,797 | 0,598 |
| | **игроки** | | | **лица** | | | **люди** | | | **долг** | | |
| **гости** | 0,625 | 0,847 | 0,530 | 0,556 | 0,923 | 0,513 | 0,715 | 0,628 | 0,449 | 0,715 | 0,644 | 0,460 |
| | **люди** | | | **ужас** | | | **смотрит** | | | **слово** | | |
| **старуха** | 1,000 | 0,903 | 0,903 | 0,910 | 0,906 | 0,824 | 0,875 | 0,932 | 0,816 | 0,924 | 0,876 | 0,809 |

*Таблица 1.2*

Значения причинной зависимости ($K_1$) для некоторых пар слов из текста повести И.С. Тургенева «Муму» (в первом столбце слова-причины)

| | *n*(*l*/*j*) / *nl* | *v*(*lj*) | *K*1 | *n*(*l*/*j*) / *nl* | *v*(*lj*) | *K*1 | *n*(*l*/*j*) / *nl* | *v*(*lj*) | *K*1 | *n*(*l*/*j*) / *nl* | *v*(*lj*) | *K*1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **старая** | | | **руки** | | | **знаки** | | | **деревня** | | |
| **Герасим** | 1 | 0,969 | 0,969 | 1 | 0,956 | 0,956 | 1 | 0,952 | 0,952 | 0,9 | 0,926 | 0,833 |
| | **знаки** | | | **вместе** | | | **хвост** | | | **каморка** | | |
| **Муму** | 0,75 | 0,852 | 0,639 | 0,728 | 0,842 | 0,612 | 0,556 | 0,87 | 0,483 | 0,5 | 0,867 | 0,434 |
| | **говорила** | | | **лицо** | | | **старая** | | | **быт** | | |
| **барыня** | 1 | 0,903 | 0,903 | 1 | 0,893 | 0,893 | 0,875 | 0,941 | 0,824 | 0,917 | 0,838 | 0,768 |
| | **работать** | | | **Москва** | | | **дело** | | | **сторона** | | |
| **деревня** | 0,667 | 0,778 | 0,519 | 0,6 | 0,751 | 0,451 | 0,4 | 0,887 | 0,355 | 0,445 | 0,748 | 0,333 |

Наконец отметим, что наличие «сильной» причинной зависимости (диаграммы 2, 3 и 5), может быть использовано для «сжатия» информации при описании взаимного расположения элементов массива данных.

## 1.5. Характеристики соответствия однородных цепей

В данном подразделе на основе наглядных графических образов (диаграмм) представлены ***характеристики соответствия*** пар однородных цепей ($j$-й и $l$-й) [2]. Это позволяет соотнести субъективное суждение о степени «пространственного соответствия» элементов сопоставляемых цепей с формальной числовой характеристикой, представляющей такое соответствие.

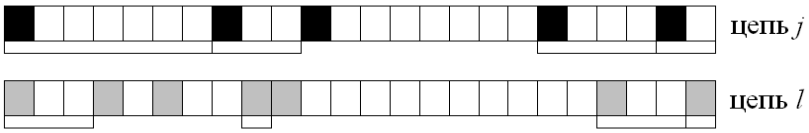
Если сопоставляемые цепи берутся из состава разных полных неоднородных цепей (различных знаковых последовательностей), то их необходимо пред-

варительно расположить относительно общего «конца» или «начала»; другими словами, сравниваемые цепи устанавливаются относительно одного «начала координат» (диаграмма 6). Обоснование для выбора начала координат здесь не рассматривается.

В случае, когда сравниваемые однородные цепи не пересекаются, можно считать их не соответствующими друг другу.

Выделяются, как это показано на диаграмме 6, пары сопоставляемых интервалов $\langle \Delta(j_l)_i, \Delta(l/j)_i \rangle$ для отмеченных элементов $j$ и $l; \forall i = 1, 2, \dots, n_{j,l}$, где $n_{j,l}$ – число пар сопоставляемых интервалов у сравниваемых однородных цепей.

Диаграмма 6




В примере:

$$\langle \Delta(j_l)_1 = 7, \Delta(l/j)_1 = 3 \rangle; \langle \Delta(j_l)_2 = 3, \Delta(l/j)_2 = 1 \rangle; \langle \Delta(j_l)_3 = 4, \Delta(l/j)_3 = 3 \rangle; \langle \Delta(j_l)_4 = 2, \Delta(l/j)_4 = 1 \rangle.$$

***Степень локального соответствия*** пары сопоставляемых интервалов $l$-й цепи $j$-й цепи определена в виде

$$a(l/j)_i = \frac{2\sqrt{\Delta(j_l)_i \cdot \Delta(l/j)_i}}{\Delta(j_l)_i + \Delta(l/j)_i}. \quad (1.36)$$

***Степень частичного соответствия*** $l$-й цепи определена в виде

$$a_1(l/j) = \frac{2n(l/j)}{n_j + n_l} \cdot \left( \prod_{i=1}^{n(l/j)} a(l/j)_i \right)^{\frac{1}{n(l/j)}}, \quad (1.37)$$

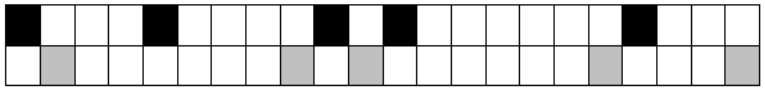
где $n(l/j)$ – число сопоставляемых пар интервалов (способом определенным выше). ***Полное соответствие*** между цепями имеет место при равенстве чисел $n_j = n_l = n(l/j)$ (диаграмма 9). При этом чем меньше различаются сопоставляемые интервалы, тем больше степень соответствия цепей (диаграмма 8). Если же все сопоставляемые интервалы равны, то степень соответствия максимальна и равна 1 (диаграмма 9).

Если неважно индивидуальное соответствие одной цепи по отношению к другой, то вычисляется (средний) ***коэффициент взаимного соответствия*** в виде

$$a_2(l,j) = \sqrt{a_1(l/j) \cdot a_1(j/l)}. \qquad (1.38)$$

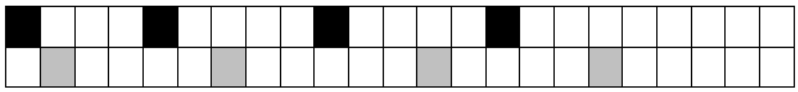
Характеристика соответствия названа ***установленной***, ***определенной*** или ***закономерной***, если размеры всех сопоставляемых интервалов задаются определенным операционным преобразованием. ***Неустановленная или неопределенная характеристика соответствия*** отмечается для такого случая, когда размеры сопоставляемых интервалов не задаются в форме определенного преобразования (диаграммы 7, 10–13).

Диаграмма 7



| Соответствие | | $a_1$ | $a_2$ |
|---|---|---|---|
| от | | 0,889 | 0,842 |
| от | | 0,797 | |

Диаграмма 8

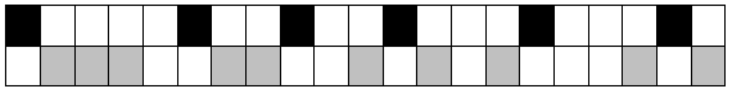


| | | | |
|---|---|---|---|
| от | | 0,992 | 0,856 |
| от | | 0,738 | |

Диаграмма 9

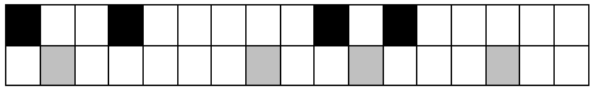
| | | | |
|---|---|---|---|
| от | | 1 | 1 |
| от | | 1 | |

Диаграмма 10



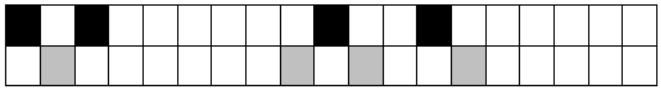
| | | | |
|---|---|---|---|
| от | | 0,674 | 0,645 |
| от | | 0,617 | |

Диаграмма 11



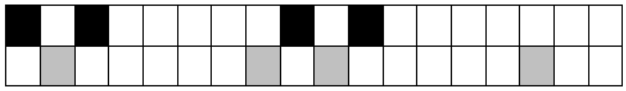
| | | | |
|---|---|---|---|
| от | | 0,942 | 0,835 |
| от | | 0,739 | |

Диаграмма 12



| | | | |
|---|---|---|---|
| от | | 0,911 | 0,812 |
| от | | 0,723 | |

Диаграмма 13



| | | | |
|---|---|---|---|
| от | | 0,872 | 0,807 |
| от | | 0,747 | |

Частный случай определенной характеристики соответствия, когда локальные соответствия одинаковы $a(l/j)_i = const, \forall\, i = 1,2,\dots,n_{j,l}$, называется ***фиксированной характеристикой соответствия***.

Семь диаграмм (7–13) наглядно представляют пары однородных цепей и их коэффициенты соответствия. В каждом примере приведены диаграммы пар

таких однородных цепей, которые входят в состав одной и той же полной неоднородной цепи и рассматриваются на предмет пространственного соответствия друг другу. На диаграммах элементы *j*-й цепи обозначены черными квадратами, а элементы *l*-й – серыми.

Аналогично матрицам зависимостей компонентов в бинарно-однородных цепях одной и той же знаковой последовательности возможно построение ***матриц соответствия пар однородных цепей*** для отдельной знаковой последовательности или при сравнении двух знаковых последовательностей.

## 1.6. Аналоги характеристик строя числовым характеристикам случайных величин

Определим числовые характеристики строя цепи, которые по форме аналогичны разного рода моментам случайных величин [2]. В формулах (1.15–1.18), (1.20) и (1.21) определены представленные ниже числовые характеристики строя цепи. «Удаленностью» очередного $(i+1)$-го вхождения элемента относительно его *i*-го вхождения в *j*-й однородной цепи названа величина, полученная логарифмированием интервала между ними $\log \Delta_{ij}$.

Средняя удаленность элементов *j*-й однородной цепи (1.16) определена в виде

$$g_j = \log_2 \Delta_{gj} = G_j/n_j = 1/n_j \sum_{j_{i=1}}^{n} \log_2 \Delta_{ij},$$

где средний геометрический интервал j-й цепи определён в виде (1.2):

$$\Delta_{gj} = \sqrt[n_j]{V}.$$

Средняя удаленность элементов данной полной неоднородной цепи определена в виде (1.18):

$$g = \log_2 \Delta_g = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta_{ij},$$

где средний геометрический интервал полной неоднородной цепи определён в виде (1.5):

$$\Delta_g = \sqrt[n]{V}.$$

Назовем отношение величин вида

$$\delta\Delta_{gi} = \Delta_{gi}/\Delta_g \tag{1.39}$$

***нормированным средним геометрическим интервалом j-й цепи.***

Логарифм отношения средних интервалов, определенный разностью

$$\log\Delta_{gi} - \log\Delta_g, \tag{1.40}$$

назовем ***центрированной средней удаленностью элементов j-й однородной цепи.***

Аналогично математическому ожиданию центрированной случайной величины имеем

$$\sum_{j=1}^{m} \frac{n_j}{n}(\log\Delta_{gi} - \log\Delta_g) = 0\,. \tag{1.41}$$

Однако ***среднее арифметическое отклонение средних удаленностей однородных цепей***

$$v_g = \sum_{j=1}^{m} \frac{n_j}{n}|\log\Delta_{gj} - \log\Delta_g| \tag{1.42}$$

не равно нулю и является ***числовой характеристикой «разброса» средних удаленностей однородных цепей***.

Второй центральный момент или ***дисперсию средних удаленностей*** однородных цепей, а также ***среднее квадратичное отклонение*** (СКО) средних удаленностей определим соответственно в виде

$$D_g = \sum_{j=1}^{m} \frac{n_j}{n}\left(\log\Delta_{gi} - \log\Delta_g\right)^2; \tag{1.43}$$

$$\sigma_g = \sqrt[2]{D_g}. \tag{1.44}$$

Как известно, СКО является более удобной числовой характеристикой разброса, так как имеет совпадающую со средней удаленностью размерность.

Третий центральный момент или ***асимметрию средних удаленностей*** однородных цепей определим в виде

$$\mu_{3g} = \sum_{j=1}^{n} \frac{n_j}{n} (\log\Delta_{gi} - \log\Delta_g)^3 = \sum_{j=1}^{n} \frac{n_j}{n} (g_j - g)^3. \quad (1.45)$$

Нормированием этой величины в виде

$$As_g = \mu_{3g}/\sigma_g^3 \quad (1.46)$$

получаем безразмерную характеристику – ***коэффициент асимметрии (скошенность) распределения средних удаленностей однородных цепей***.

Наконец, определение четвертого центрального момента и его нормирование дают ***характеристику эксцесса (крутизна или плосковершинность) – для распределения средних удаленностей*** однородных цепей в виде

$$\varepsilon x_g = \frac{\mu_{4g}}{\sigma_g^4}. \quad (1.47)$$

Заменим в формуле (1.18) суммирование по однородным цепям сложением удаленностей интервалов $\log\Delta_k$ (независимо от их принадлежности к однородным цепям) в виде

$$\log\Delta_g = \sum_{j=1}^{m} \frac{n_j}{n} \log\Delta_{gi} = \sum_{k=1}^{M} \frac{n_k}{n} \log\Delta_k, \quad (1.48)$$

где $n_k/n$ – частота вхождений любого интервала длиной $\Delta_k$ в строй полной неоднородной цепи;

$M$– число разных интервалов $\Delta_k$ в полной неоднородной цепи.

На основе отношения величин вида

$$\sigma\Delta_k = \Delta_k/\Delta_g, \quad (1.49)$$

получим ***нормированный интервал строя полной неоднородной цепи***. Логарифмирование этого отношения дает разность в виде

$$\log\Delta_k - \log\Delta_g, \quad (1.50)$$

представляющую ***центрированную удаленность произвольных одинаковых элементов в строе полной неоднородной цепи.***

Это позволяет получить аналогичные моменты и характеристики, сформулированные на множестве разных интервалов.

Введем для таких характеристик строя множество обозначений, аналогичное рассмотренным выше:

$$\{\upsilon_\Delta, D_\Delta, \sigma_\Delta, \mu_{3\Delta}, As_\Delta, \mu_{4\Delta}, \varepsilon x_\Delta\}. \tag{1.51}$$

В качестве примера представим выражение для вычисления ***среднего квадратичного отклонения удаленностей одинаковых интервалов в полной неоднородной цепи*** в виде

$$\sigma_\Delta = \sqrt{\sum_{k=1}^{M} \frac{n_k}{n} (\log \Delta_k - \log \Delta_g)^2}. \tag{1.52}$$

Рассмотренные числовые характеристики позволяют более подробно описывать оригинальный строй цепи.

Для «полного» описания строя всей неоднородной цепи следует использовать два аналога статистических распределений: частотное распределение средних удаленностей всех $m$ ($j$-ых) однородных цепей и частотное распределение всех $M$ разных ($k$-ых) удаленностей, которые представимы двумя множествами пар вида

$$\{\langle n_j/n, \log \Delta_{gi}\rangle\} \text{ и } \{\langle n_k/n, \log \Delta_k\rangle\}. \tag{1.53}$$

Существуют аналогичные по форме рассмотренным выше характеристики строя разнородных цепей (см. п. 1.2.1).

## 1.7. Меры расхождения построений информационных цепей

Определим ***меры расхождения*** оригинальных построений разных неоднородных последовательностей $A$ и $B$. Для приближённого сравнения на основе глубин $G$ и средних удаленностей $g$, определим выражения в виде

$$\Delta G_1 = |G_A - G_B|; \tag{1.54}$$

$$\Delta g_1 = |g_A - g_B|. \tag{1.55}$$

Вторая формула позволяет сравнивать последовательности без учёта их длины.

Для более точного сравнения можно использовать расхождение по однородным цепям:

$$\Delta G_2 = \sum_{j=1}^{m} \left| G_{Aj} - G_{Bj} \right|; \tag{1.56}$$

$$\Delta g_2 = \sum_{j=1}^{m} \left| g_{Aj} - g_{Bj} \right|. \tag{1.57}$$

И, наконец, для наиболее точного сравнения можно использовать поинтервальную меру расхождения.

$$\Delta g_3 = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left| g_{Aji} - g_{Bji} \right|. \tag{1.58}$$

## 1.8. Об использовании интервалов и числовых характеристик строя для кодирования

В рамках формального анализа строя полной неоднородной знаковой цепи представляется очевидным кодирование интервалов (при декомпозиции такой цепи на неполные однородные (неоднородные) цепи), а также использование числовых характеристик строя, для следующих целей:

– компактного представления массивов данных (для сжатия информации);

– защиты массивов данных от помех (для помехоустойчивого кодирования);

– защиты от несанкционированного доступа к массивам данных (криптография).

Однако в данном разделе не рассматриваются эти традиционные темы, актуальные для систем передачи данных, в пользу анализа особенностей и обнаружения закономерностей построения цепей событий разной природы. К настоящему времени известны работы Браиловского И.В. [18], посвященные оптимальному (эффективному) кодированию интервалов, в которых удачно осуществляется так называемое «интервальное преобразование текста». В этом преобразовании при выделении интервалов очередной однородной цепи ис-

пользуется «текст», в котором вычеркнуты однородные цепи, рассмотренные на предыдущих шагах. Такая процедура позволяет сильно «сжать» исходный текст. В этих работах утверждается, что до недавнего времени было затруднено кодирование интервалов в знаковых последовательностях, которые не удавалось интерпретировать как композицию однородных цепей. Данное утверждение спорно, так как в теории массового обслуживания давно введены понятия потоков однородных событий (заявок), которые при композиции могут составлять неоднородный поток событий. ***В качестве случайной величины в потоке однородных событий учитывается (временной) интервал между соседними (одинаковыми) заявками.*** Несмотря на отмеченное, авторами опубликована одна работа, посвященная вопросам эффективного кодирования интервалов [19].

## 1.9. Замечания по разделу

Не отрицая традиционные средства для исследования детерминированных сигналов и других массивов данных, следует провести границу в области применения вероятностно-статистического и энтропийного подходов, с одной стороны, и непосредственным формальным анализом строя цепи сообщений – с другой стороны.

В тех ситуациях, когда нет необходимости или нет инструментальных возможностей учитывать порядок следования событий, применимы средства теории вероятностей и математической статистики. В ситуациях, когда необходимо учитывать порядок следования событий и имеются инструментальные средства для этого, более адекватными, по мнению авторов, являются средства формального анализа строя цепи.

## 1.10. Контрольные вопросы и задания

1. Чем обоснована необходимость формального анализа расположения компонентов в массивах данных?

2. Определение понятия «строй цепи»; примеры прямого преобразования знаковых последовательностей в соответствующие им строи.

3. Определение понятия «вектор строя»; примеры векторов строя.

4. Определение декомпозиции знаковой последовательности на однородные цепи; примеры декомпозиции короткой последовательности (строки текста) на однородные последовательности.

5. Отображение знаковой цепи матрицей интервалов; примеры отображения знаковой последовательности матрицей интервалов.

6. Формулы основных числовых характеристик строя цепи; примеры вычисления характеристик строя для коротких последовательностей.

7. Определение распределения числовых характеристик строя. Связь распределений характеристик строя с распределениями вероятностей (частот), частотно-ранговыми распределениями.

8. Определение функций характеристик строя.

9. Определение бинарно-однородной цепи в составе полной неоднородной последовательности; примеры выделения таких цепей из коротких последовательностей.

10. Определение числовых характеристик пространственной зависимости компонентов бинарно-однородной цепи; структура матрицы зависимостей; примеры матрицы зависимостей для короткой последовательности.

11. Определение числовых характеристик соответствия однородных цепей; структура матрицы соответствия пар однородных цепей; пример вычисления матрицы соответствия одной или двух коротких последовательностей.

12. Определение характеристик строя, аналогичных числовым характеристикам случайных величин; пример вычисления таких характеристик для короткой цепи.

13. Назовите задачи информатики, кроме задач анализа расположения компонентов, в которых может использоваться модель строя и числовые характеристики строя.

# 2. ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ ТЕОРИИ СВЯЗИ К. ШЕННОНА

Материал, изложенный в данном разделе, традиционно принято относить к теории информации, которая основана на понятии энтропии [4, 5, 6, 7]. Однако теория, разработанная К. Шенноном, не определяет понятия информации и не является теорией информации по существу, так как рассматривает вопросы передачи и кодирования отдельных сообщений так называемым ***статистическим источником***. При этом целью ***приёмника*** является выбор и поиск, то есть идентификация места расположения эталона сообщения в его памяти. Явления, формализованные в данном разделе соответствуют фундаментальному понятию кода, описанному в учебном пособии «Прикладная теория информации».

## 2.1. Информация и дихотомический поиск сообщений

Под ***информацией*** Р. Хартли понимал логическую инструкцию для выбора и поиска (идентификации) полученного сообщения или состояния из некоторого множества.

При этом ***количество информации*** (по Хартли, 1928 г.) определяется в виде

$$I = \log_2 N, \tag{2.1}$$

где $N$ – число различных состояний наблюдаемого объекта или мощность алфавита сообщений источника; $I$– количество информации в битах. Если $N = 2$, то $I_{min} = 1$ бит.

При такой процедуре (программе) идентификации, еще до передачи сообщений, множество эталонов всех сообщений расположено определённым образом в памяти приемника: разбито на пары подмножеств (вплоть до отдельных элементов), образуя бинарное дерево. Если двигаться от вершины дерева к месту расположения в памяти эталона искомого сообщения, то можно наметить ***путь движения*** к данному сообщению, в котором повороты направо «→» можно обозначить «1», а повороты налево «←» – «0». Этот путь и есть ***информация для идентификации*** места расположения эталона некоторого сообщения в памяти приемника. Таким образом, информация (по Хартли) – это процедура (программа) ***дихотомического*** (***бинарного***) поиска. Она представлена последовательностью команд или инструкций для приёмника вида: «направо», «налево» или с учетом обозначений «1», «0». Заметим, что ***дихотомия*** – это последовательное деление целого на две части. По существу, дихотомическая проце-

дура идентификации представлена адресом эталона сообщения в памяти приёмника.

Заметим, что поиск объектов можно осуществлять различными процедурами, в зависимости от наличия априорной (заранее известной приемнику) информации о них:

– простым перебором или сканированием эталонов;
– случайным выбором эталона;
– оптимальным поиском;
– комбинированным поиском.

***Оптимальная процедура поиска*** – это процедура, обеспечивающая минимальное среднее количество команд для поиска данного объекта.

Пусть имеется восемь различных сообщений, которые в памяти приемника упорядочены, как это указано на рис. 2.1. Оптимальная программа поиска объекта-сообщения «∅» имеет вид «101». Из рисунка видно, что это кратчайший путь поиска.

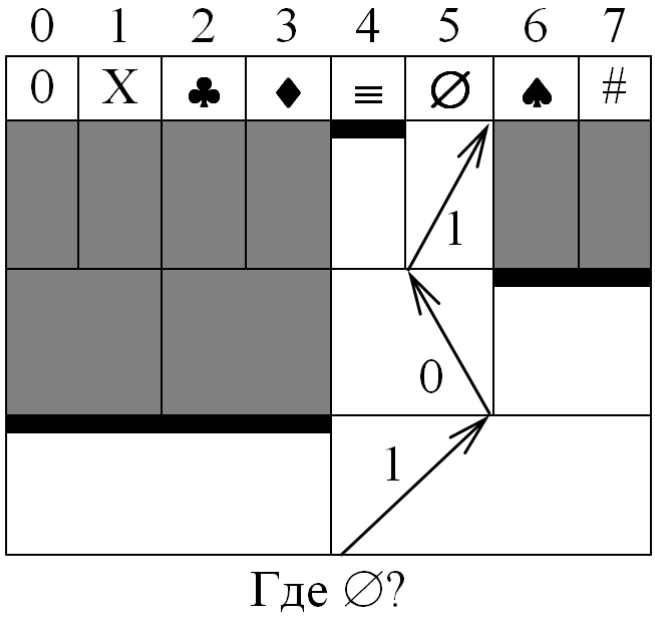



*Рис. 2.1.* Дихотомическая процедура поиска – информация по Р. Хартли

Таким же образом можно обозначить все восемь сообщений; программы их поиска окажутся закодированы трёхразрядными двоичными числами. Количество информации в каждой программе, определенное по формуле Р. Хартли, равно трём битам: $I = \log_2 8 = 3$ бит.

Также из рисунка видно, что приёмник не опознаёт собственно сообщения, а лишь находит путь к их эталонам. Однако в информатике, как правило, сообщением является сама двоичная последовательность (адрес).

Длина программы для выбора при таком равном разбиении совпадает с количеством информации. При данной схеме моделью источника является ***источник случайных сообщений***. Кроме того, при использовании меры информации Хартли, неявно предполагается, что эти сообщения равновероятны. В данной схеме приемник в своей памяти имеет только алфавит состояний (эталонов); он не способен запоминать даже один предыдущий элемент. Приемник воспринимает сообщения как статистически независимые и равновероятные состояния.

Рассмотрим систему передачи сообщений «***вероятностный источник – приемник***», в которой источник формирует случайные события (сообщения). При этом приёмник, также как в схеме Хартли, должен идентифицировать принимаемое сообщение, т.е. найти место его эталона в памяти. В реальности массивы данных измерений или тексты, передаваемые в системе, обычно представляют собой закономерно упорядоченные последовательности и не являются случайными событиями. Однако такой приёмник не может воспринимать всю последовательность целиком, а только отдельные сообщения (буквы, числа, слова, блоки текста или данных), которые воспринимаются им как случайные события.

В теории Р. Хартли неявно допускалось, что выбор искомого состояния (сообщения) осуществляется из множества всех разных, т.е. равновозможных состояний. Однако на практике во многих ситуациях вероятности состояний источника неодинаковы и могут быть известны приёмнику сообщений. Очевидно, что наличие априорной информации, имеющейся у приемника в виде вероятностей состояний (сообщений) источника позволяет оптимизировать условия выбора, поиска, идентификации определённого состояния, сообщения или объекта. В этом случае целесообразнее просматривать в первую очередь эталоны, представляющие более вероятные (чаще встречающиеся) сообщения. Для этого все отличающиеся эталоны сообщений рассматриваемого множества должны быть предварительно ***ранжированы*** (упорядочены) в памяти приёмника по убыванию частот их появления. Сама программа поиска может также осуществляться методом последовательного деления выбранного множества на два подмножества с последующим выбором одного из них. Однако каждое из двух подмножеств необходимо формировать таким набором сообщений (состояний), чтобы суммарные частоты их появления по возможности были одинаковы или хотя бы близки по значениям. В примере, показанном на рис. 2.2, вероятности смежных подмножеств искусственно подобраны равными. Заметим, что если бы все подмножества на каждом уровне были бы равновероятны, то в каждом подмножестве оказалось бы равное число сообщений, как это было в простейшем случае для множества, состоящего только из разных сообщений и сама поисковая процедура, учитывающая вероятности, оказалась бы идентична по оптимальности процедуре, не учитывающей вероятности.

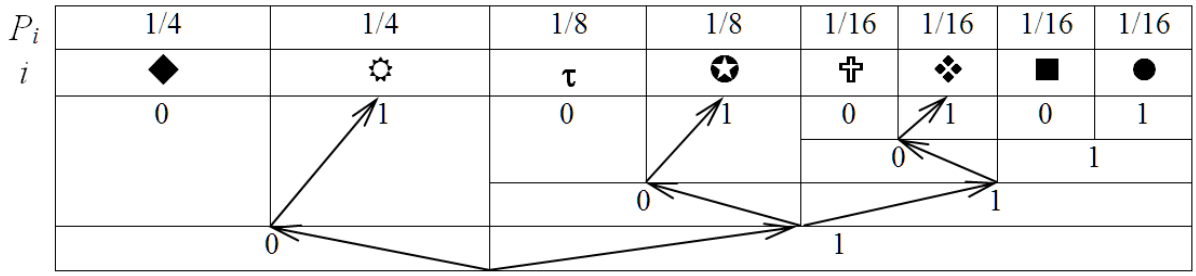



*Рис. 2.2.* Оптимальный дихотомический поиск сообщений
с разными вероятностями появления:
($i$– сообщения;$P_i$ – вероятность $i$-го сообщения)

Как видно из рис. 2.2, длина инструкции для дихотомического поиска в этом случае различна. Для поиска в памяти приемника самых частых сообщений с вероятностью 1/4 требуются всего две команды. Длина программы идентификации самых редких сообщений с вероятностью 1/16 состоит из четырех команд. Средняя длина программы поиска подсчитывается следующим образом:

$L = 2 \cdot \frac{1}{4} \cdot 2 + 2 \cdot \frac{1}{8} \cdot 3 + 4 \cdot \frac{1}{16} \cdot 4 = 2{,}75$ элемента.

Это согласуется со здравым смыслом в тех случаях, когда мы говорим, что менее вероятные, т.е. более неожиданные события несут больше информации и наоборот – более вероятные, т.е. менее неожиданные события содержат меньше информации. В данном примере с «идеальным» распределением вероятностей величина средней длины программ идентификации совпадает с количеством информации (2,75 бит), вычисляемым по формуле К. Шеннона:

$$I = H = -\sum_{j=1}^{m} \log_2 P_j ,$$

где $I$ – количество информации, $H$– энтропия источника сообщений.

Для случая равновероятных сообщений, когда $P_j = \frac{1}{m}$, эта формула принимает вид формулы Р. Хартли: $I = \log_2 m$.

## 2.2. Энтропия вероятностного источника сообщений

Пусть для идентификации некоторого знака $j$, появляющегося с вероятностью $P_j$ среди множества разных знаков $m$, требуется $I_j$ последовательных операций деления на ***равновероятные*** подмножества. Последняя операция деления, которая привела к определению $j$, состояла в делении множества с суммарной вероятностью $2P_j$ на две равновероятные части. Предыдущая операция

делила диапазон $2^2 P_j$, а ещё раньше делился диапазон $2^3 P_j$ и т.д. Рассуждая подобным образом, мы, наконец, достигаем первой операции деления пополам всего множества разных знаков, вероятность которого $\sum_{j=1}^{m} P_j = 1$.

Следовательно:

$$2^{I_j} \cdot P_j = 1 \text{ или } 2^{I_j} = 1/P_j. \tag{2.2}$$

После логарифмирования получим

$$I_j = L_j = log_2 \frac{1}{P_j} = - log_2 P_j. \tag{2.3}$$

Таким образом определяется длина программы $L_j$ или количество информации $I_j$ для идентификации отдельного $j$-го сообщения (при вероятностях сообщений кратных 2). Эта же величина представляет и ***неожиданность*** сообщения $H_j$, которая уменьшается при увеличении вероятности появления сообщения, соответственно уменьшается количество информации, переносимое сообщением для его идентификации. При $P_j = 1$ событие достоверно, поэтому оно не является неожиданным ($\log_2 1 = 0$) и не несёт информации. Зависимость

$$H_j = - \log_2 P_j, \tag{2.4}$$

будем называть также функцией неожиданности сообщения, как это предложил Альфред Реньи [20].

Зная неожиданности, размеры программ или числа информаций для идентификации каждого из $m$ сообщений данного множества можно, определить их средние значения (точнее математические ожидания) в виде

$$H = I = L = - \sum_{j=1}^{m} P_j \log_2 P_j = \sum_{j=1}^{m} P_j H_j = \sum_{j=1}^{m} P_j I_j = \sum_{j=1}^{m} P_j L_j, \quad \text{(2.5), (2.6), (2.7)}$$

где $H$ – средняя неожиданность сообщения, степень неопределённости состояния, энтропия источника сообщений;

$L, I$ – средняя длина программы и среднее количество информации для идентификации отдельного сообщения соответственно.

Заметим, что длина программы $L_j$ и количество информаций $I_j$ для идентификации отдельного сообщения в отличие от его неожиданности могут быть

только целыми числами. Средние значения величин $L$ и $I$ могут быть и дробными, как и значение $H$.

При идентификации не всегда удаётся на каждом шаге разбить исходное или выбранное на предыдущем шаге множество сообщений на два подмножества с равными или близкими суммарными вероятностями. При этом установленная процедура дихотомического выбора осуществляется не из равновероятных множеств, и соответствующие программы для идентификации содержат такие команды или инструкции, которые не всегда осуществляют лучшие, наиболее полные, выборы. Поэтому такие программы уже не могут считаться информациями. Легко убедиться, что в подобных случаях средняя длина программы для идентификации сообщений $L$ будет превышать энтропию источника сообщений $H$. Это свидетельствует о том, что совокупность всех программ обеспечивает выбор нужного сообщения в среднем медленнее, чем в случае последовательного деления на равновероятные подмножества.

Последовательные разбиения на близкие по суммарным вероятностям подмножества можно осуществить, если заменить данное множество соответствующим ему множеством, алфавит которого представлен большим числом разных сообщений. Причём частоты появления сообщений из нового алфавита должны отличаться от частот сообщений исходного алфавита, так чтобы приблизиться к ситуации, которая имела место при последовательном разбиении множества, состоящего только из разных сообщений (см. рис. 2.1). Подобные условия можно выполнить, если приёмник будет одновременно воспринимать не одно сообщение исходного алфавита, а сразу их целую большую группу, которая представляет отдельное макросообщение нового алфавита. Так как повторные появления таких длинных «слов» практически невозможны, то вероятности сообщений нового алфавита можно считать равными. Программа выбора такого слова длиннее, чем в случае идентификаций отдельного сообщения, поэтому для сравнительной оценки удобнее пользоваться средней длиной программы для выбора отдельной буквы.

Пусть число длинных слов равно $N$; каждое из таких слов формируется из $m$ разных сообщений и состоит из $n$ статистически независимых букв; числа вхождений каждой из $m$ букв в длинное слово и соответствующие им вероятности будут [6, 7]:

$$n_1, n_2, \cdots, n_j, \cdots, n_m, \qquad \sum_{j=1}^{m} n_j = n;$$

$$P_1, P_2, \cdots, P_j, \cdots, P_m, \qquad \sum_{j=1}^{m} P_j = 1.$$

Так как повторные появления одинаковых длинных слов практически невозможны, то, как уже отмечалось ранее, вероятности таких событий можно считать равными ($P = 1/N$). В соответствии с положениями теории вероятностей формирование длинного слова из букв следует рассматривать как сложное случайное событие типа «произведение событий» [17]. Вероятность появления такого слова $P$ определяется через вероятности составляющих его независимых букв $P_j$ в виде

$$P = \underbrace{P_1 \cdot P_1 \cdot \ldots \cdot P_1}_{n_1} \cdot \underbrace{P_2 \cdot P_2 \cdot \ldots \cdot P_2}_{n_2} \cdot \underbrace{P_m \cdot P_m \cdot \ldots \cdot P_m}_{n_m} \tag{2.8}$$

или компактнее в виде

$$P = P_1^{n1} \cdot P_2^{n2} \cdot \ldots \cdot P_m^{nm} = \prod_{j=1}^{m} P_j^{nj}. \tag{2.9}$$

Если допустить, что $n \to \infty$, то в соответствии с законом больших чисел $n_j = n \cdot P_j$ и

$$P = \prod_{j=1}^{m} P_j^{n \cdot P_j}. \tag{2.10}$$

В соответствии с мерой Хартли количество информации для идентификации длинного слова будет

$$I_n = \log_2 N = \log_2 \frac{1}{P} = -\log_2 P \tag{2.11}$$

Подставляя (2.10) в (2.11), получим

$$I_n = -\log_2 \prod_{j=1}^{m} P_j^{n \cdot P_J} = -n \cdot \log_2 \prod_{j=1}^{m} P_j^{P_j} = -n \cdot \sum_{j=1}^{m} P_j \cdot \log P_j \tag{2.12}$$

Среднее количество информации для идентификации отдельной буквы $I = I_n/n$, т.е.

$$H = I = -\sum_{j=1}^{m} P_j \cdot log\ P_j. \tag{2.13}$$

Определяемую таким образом величину К. Шеннон в своей книге «Математическая теория связи», вышедшей в 1948 г., назвал ***энтропией источника*** *H* (по аналогии с термодинамической энтропией) или ***количеством информации*** *I*. Так же, как и в термодинамике, где энтропия характеризует неопределённость теплового состояния вещества, энтропия в теории информации служит мерой неопределённости сообщения (состояния источника). В соответствии с определением А. Реньи, эта мера представляет среднюю неожиданность сообщений.

Если сообщения статистически зависимы, то при определении энтропии необходимо учитывать не только безусловные, но также и их условные вероятности.

При получении того или иного сообщения, т.е. в процессе его идентификации, неопределённость сообщения или состояния источника уменьшается и, наконец, снимается полностью. При этом получаемое количество информации, идентифицирующей сообщение, равно уменьшению энтропии. Таким образом, ***энтропия и количество информации – величины взаимно обратные***. В начальный момент времени энтропия (неопределенность состояния) источника сообщений максимальна, так как сообщения еще не получены приемником. После завершения передачи сообщений энтропия источника равна нулю, а количество принятой информации на приемной стороне равно максимальной энтропии источника сообщений в начальный момент времени. Образно говоря, энтропия источника сообщений «перетекает» в информацию для приемника. ***Этим объясняется равенство количества информации I и энтропии H*** в формуле К. Шеннона (несмотря на разную суть этих понятий). Подчеркнём, что ***энтропия является характеристикой именно источника сообщений*** и представляет его информационную способность, степень неопределённости состояния, степени неупорядоченности, степень хаоса, среднюю неожиданность сообщений. При этом предполагается, что более неожиданные (редкие) сообщения – более информативны и наоборот более частые сообщения – менее информативны.

Необходимо помнить, что «количество информации по Шеннону» и энтропия являются усреднёнными характеристиками сообщений и состояний источника; идентификация конкретного сообщения требует определённого количества информаций.

## 2.3. Функция неожиданности и понятие «энтропия» по А. Реньи

Замечательная интерпретация (информационной) энтропии, дана выдающимся венгерским математиком Альфредом Реньи (1976 г.) в научно-популярной книге «Трилогия о математике» [20].

Этот формализм обычно с трудом воспринимается даже специалистами. Понятие «энтропия» поясняется как «степень неопределенности состояния», «неопределенность», «информационная способность» вероятностного источника сообщений, как «степень неупорядоченности», как «характеристика хаоса» наблюдаемой системы. Однако из самой формулы К. Шеннона эти интерпретации не следуют; отсюда возникают определённые затруднения в понимании формализма энтропии. Известная формула обычно представлена в виде (2.13)

А. Реньи ввел «***функцию неожиданности***» случайного сообщения, которая имеет интуитивно-приемлемую по смыслу простую зависимость вида

$$H_j = -\log P_j = \log \frac{1}{P_j}. \qquad (2.14)$$

Ее график представлен на рис. 2.3, из которого видно, что редкие сообщения (для которых $0 < P_j \ll 1$) очень неожиданны для приемника, а частые сообщения (когда $P_j \to 1$) имеют малую неожиданность. Заменив в формуле (2.13) $(-\log_2 P_j)$ на $H_j$ из (2.14), А. Реньи получает простую для интерпретации формулу энтропии в виде

$$H = \sum_{j=1}^{m} P_j \cdot H_j. \qquad (2.15)$$

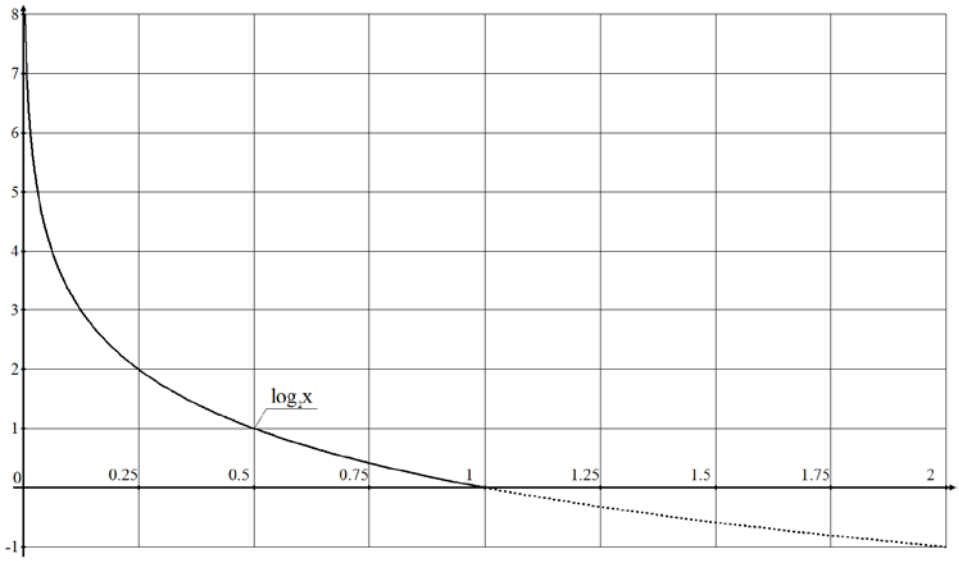



*Рис. 2.3.* Функция неожиданности

Из теории вероятностей известно, что это формула математического ожидания дискретной случайной величины, для которой различается $m$ значений. Таким образом, по А. Реньи, ***энтропия – это математическое ожидание неожиданностей разных сообщений*;** или короче: ***энтропия – это средняя неожиданность сообщений*.** А. Реньи утверждает, что событие (сообщение) может быть неожиданно, а величина (энтропия) – неопределенна.

## 2.4. Свойства энтропии

Рассмотрим свойства энтропии [6, 7].

1. По определению, представленному выражением (2.13) следует, что энтропия не бывает отрицательной, т.е. $H \geq 0$.

2. Если одно из сообщений достоверно $(P_j = 1)$, а остальные невозможны $(P_{j\neq i} = 0)$, то никакой неопределённости в состоянии источника нет, т.е. $H = 0$.

3. Энтропия ***элементарного источника***, принимающего всего два состояния, т.е. генерирующего два разных сообщения (например, «1» и «0»), вероятности которых соответственно равны $P$ и $(1 - P)$, описывается выражением

$$H = -\sum_{j=1}^{2} P_j \cdot \log_2 P_j = H(P) = -[P \cdot \log_2 P + (1-P) \cdot \log_2(1-P)] = \\ = P \cdot H_1(P) + (1-P) \cdot H_2(1-P). \tag{2.16}$$

Графики функций энтропии $H(P)$ и неожиданности первого сообщения $H_1(P)$ представлены на рис. 2.4, из которых видно, что функция $H(P)$ принимает максимальное значение, равное одному биту, когда вероятности первого и второго сообщений равны, т.е. если $P = (1 - P) = 0{,}5$.

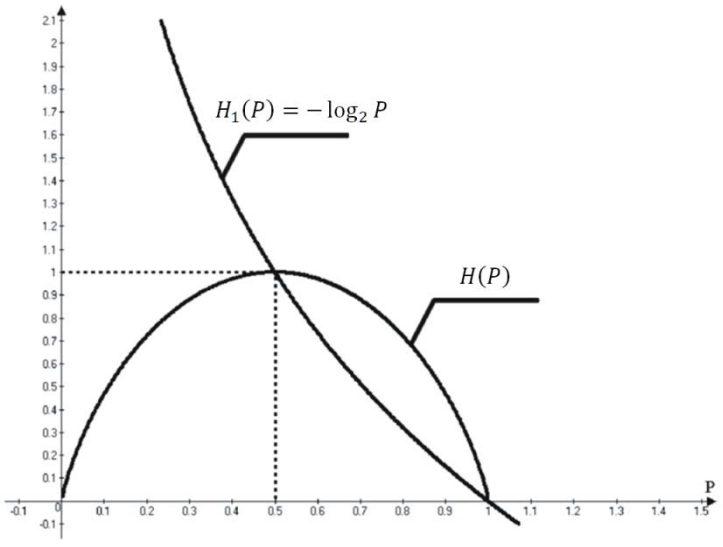



*Рис. 2.4.* Графики функций энтропии $H(P)$ и неожиданности первого сообщения $H_1(P)$

4. Степень неопределённости состояния системы – энтропия – максимальна, когда вероятности состояний одинаковы, т.е. если

$$P_1 = P_2 = \cdots = P_j = \cdots P_m = \frac{1}{m};$$

В этом случае энтропия источника сообщений определяется выражением

$$H_{max} = -\sum_{j=1}^{m} \frac{1}{m} \log_2 \frac{1}{m} = \log_2 m. \qquad (2.17)$$

Заметим, что это выражение совпадает с выражением (2.1) для определения количества информации аддитивной мерой Хартли (если $m = N$).

5. Энтропия сложной системы $H_\Sigma$ (являющейся объединением нескольких систем) или многоканального источника складывается из безусловных энтропий составляющих её систем $H_e$ , если поведение этих систем статистически независимо, т.е.

$$H_\Sigma = H_1 + H_2 + \cdots + H_e + \cdots + H_k = \sum_{e=1}^{k} H_e, \qquad (2.18)$$

где $k$ – количество систем или каналов источника.

Также определяется и количество информаций, идентифицирующих состояние сложной системы (многоканального источника).

Для описания сложной системы, состоящей из статистически зависимых систем, необходимо, кроме безусловных энтропий, учитывать также условные, совместные и взаимные энтропии.

## 2.5. Энтропия объединения

Рассмотрим систему «вероятностный источник – приемник», в которой приемник обладает следующими «способностями»: различает сообщения из конечного набора, воспринимает их как ***статистически зависимые в парах случайные события*** (запоминает и может делать суждения о следующем сообщении на основе предыдущего), знает безусловные и условные вероятности всех сообщений. При этом источник может быть вероятностным, то есть формирующим статистически зависимые сообщения в парах сообщений или в кортежах большей длины, хотя на практике источник, как правило, является не ве-

роятностным и порождает закономерно упорядоченные последовательности сообщений в виде массивов данных или текстов.

Определим энтропию вероятностного источника, формирующего статистически зависимые в парах сообщения (двухканальный источник) [6, 7]. Обозначим множество сообщений, формируемых первым каналом данного источника $X = \{x_1, x_2, \dots, x_i, \dots, x_m\}$, множество сообщений второго канала – $Y = \{y_1, y_2, \dots, y_j, \dots, y_l\}$, безусловные вероятности – $P(x_i), P(y_j)$ и условные вероятности $P(x_i/y_j), P(y_j/x_i)$, энтропию двухканального источника (энтропию объединения) – $H(X, Y)$.

Воспользуемся формулой (2.13) для энтропии источника, формирующего статистически независимые отдельные сообщения, и запишем энтропию объединения в виде

$$H(X,Y) = -\sum_{i=1}^{m}\sum_{j=1}^{l} P(x_i, y_j)\log_2 P(x_i, y_j), \qquad (2.19)$$

где $P(x_i, y_j)$ – вероятность сложного события (типа «произведение событий»), состоящего в совместном появлении сообщений $x_i$ на выходе первого канала и сообщения $y_j$ на выходе второго канала.

Из теории вероятностей известно [17], что вероятность сложного события типа «произведение событий» для статистически зависимых событий определяется в виде

$$P(x_i, y_j) = P(x_i)\cdot P(y_j/x_i);$$
$$P(x_i, y_j) = P(y_j)\cdot P(x_i/y_j),$$

Подставим в (2.19) первое определение $P(x_i, y_j)$:

$$H(X,Y) = -\sum_{i=1}^{m}\sum_{j=1}^{l} P(x_i)\cdot P(y_j/x_i)\cdot \log_2\left(P(x_i)\cdot P(y_j/x_i)\right). \qquad (2.20)$$

Заменим логарифм произведения суммой логарифмов:

$$H(X,Y) = -\sum_{i=1}^{m}\sum_{j=1}^{l} P(x_i)\cdot P(y_j/x_i)\cdot \left(\log_2 P(x_i) + \log_2 P(y_j/x_i)\right). \qquad (2.21)$$

Заменим двойную сумму на два слагаемых:

$$H(X,Y) = -\sum_{i=1}^{m} P(x_i) \cdot \log_2 P(x_i) \cdot \sum_{j=1}^{l} P\left(y_j/x_i\right) - \\ -\sum_{i=1}^{m} P(x_i) \sum_{j=1}^{l} P\left(y_j/x_i\right) \cdot \log_2 P\left(y_j/x_i\right), \tag{2.22}$$

Так как сумма $\sum_{j=1}^{l} P\left(y_j/x_i\right) = 1$, то первое слагаемое в формуле (2.22) запишем компактнее.

Тогда

$$H(X,Y) = -\sum_{i=1}^{m} P(x_i) \cdot \log_2 P(x_i) - \sum_{i=1}^{m} P(x_i) \sum_{j=1}^{l} P\left(y_j/x_i\right) \cdot \log_2 P\left(y_j/x_i\right). \tag{2.23}$$

Первое слагаемое, очевидно, представляет безусловную энтропию первого источника (канала) $X$, поэтому последнее выражение (2.21) запишем в виде

$$H(X,Y) = H(X) - \sum_{i=1}^{m} P(x_i) \sum_{j=1}^{l} P\left(y_j/x_i\right) \cdot \log_2 P\left(y_j/x_i\right). \tag{2.24}$$

Обозначим

$$-\sum_{j=1}^{l} P\left(y_j/x_i\right) \cdot \log_2 P\left(y_j/x_i\right) = H(Y/x_i), \tag{2.25}$$

Назовем $H(Y/x_i)$ ***частной условной энтропией*** канала $Y$ при условии наблюдения сообщения $x_i$ в первом канале. С учётом последнего обозначения перепишем формулу взаимной энтропии (2.24) в виде

$$H(X,Y) = H(X) + \sum_{i=1}^{m} P(x_i) \cdot H(Y/x_i). \tag{2.26}$$

Очевидно, что правое слагаемое представляет собой математическое ожидание (среднее значение) величин $H(Y/x_i)$.

Обозначим

$$\sum_{i=1}^{m} P(x_i) \cdot H(Y/x_i) = H\left(\frac{Y}{X}\right) \tag{2.27}$$

и назовём $H(Y/X)$ ***условной энтропией одного*** канала (источника) $Y$ относительно другого статистически зависимого канала $X$.

С учетом введенных понятий и обозначений запишем выражение (2.26) в виде

$$H(X,Y) = H(X) + H(Y/X). \tag{2.28}$$

Так как статистическая зависимость источников случайных событий взаимна (симметрична), то подставляя в формулу (2.19) второе определение $P(x_i, y_j)$, получим в результате аналогичных преобразований и подстановок выражение вида

$$H(X,Y) = H(Y) + H(X/Y). \tag{2.29}$$

Таким образом, из выражений (2.28) и (2.29) следует, что ***для двух статистически зависимых источников энтропия их объединения определяется суммой безусловной энтропии одного из каналов и условной энтропией второго канала относительно первого.***

Сложный источник, генерирующий пары статистически зависимых сообщений, формирует последовательность событий, называемую ***марковской цепью*** (***цепь Маркова первого порядка***).

Так как для ***статистически независимых источников*** $P(x_i, y_j) = P(x_i) \cdot P(y_j)$, то энтропия их объединения определяется по формуле

$$H(X,Y) = H(X) + H(Y), \tag{2.30}$$

то есть энтропия объединения в таком случае определяется суммой безусловных энтропий отдельных каналов.

***При наличии функциональной зависимости между каналами*** очевидно совпадение их энтропий с энтропией двухканальной системы

$$H(X,Y) = H(X) = H(Y). \tag{2.31}$$

Если рассматривается трёхканальный источник (объединение трёх статистически зависимых источников $X, Y, Z$), то формула энтропии по аналогии с формулой (2.29) принимает вид

$$H(X,Y,Z) = H(X,Y) + H(Z/XY), \tag{2.32}$$

или подставляя вместо $H(X,Y)$ её определение, получаем

$$H(X,Y,Z) = H(X) + H(Y/X) + H(Z/XY), \tag{2.33}$$

где $H(Z/XY)$ – это условная энтропия третьего источника относительно двух других.

Сложный источник, генерирующий тройки статистически зависимых сообщений, формирует последовательность событий, называемую ***марковской цепью второго порядка***. Очевидно понятие многоканального источника (состоящего из $n$ статистически зависимых источников). Сложный источник, генерирующий $(n+1)$ статистически зависимых сообщений, формирует последовательность событий, называемую ***марковской цепью n-го порядка***. Энтропия объединения такого источника определяется аналогично. При этом приёмник, запоминающий $n$ предыдущих сообщений, потребует $\sum_{i=1}^{n} m^i$ ячеек памяти, т.е. количество требуемой памяти будет расти экспоненциально при линейном увеличении количества запоминаемых сообщений.

Рассмотрим в качестве примера (табл. 2.1) энтропии буквенного текста для русского и английского языков. Символом $H$ обозначена энтропия, приходящаяся на одну букву в том случае, когда все буквы считаются равновероятными (вероятности букв неизвестны или не учитываются); $H(X)$ – безусловная энтропия на одну букву (учитываются вероятности только отдельных букв); $H(X/Y)$ – условная энтропия при учёте взаимозависимости двух букв, следующих друг за другом (учитываются, кроме отмеченного, условные вероятности букв, следующих друг за другом).

*Таблица 2.1*

| Языки | Энтропия | | |
|---|---|---|---|
| | $H$ | $H(X)$ | $H(X/Y)$ |
| Русский | 5,00 | 4,35 | 3,52 |
| Английский | 4,76 | 4,03 | 3,32 |

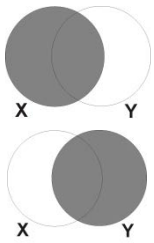
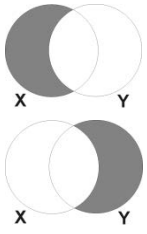
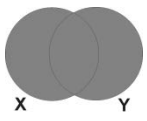
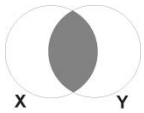
Рассмотренные соотношения между энтропиями и данный пример показывают, что энтропия источника уменьшается, если учитывается всё больше данных о статистических свойствах источника. Кроме того, количество информации убывает сильнее всего при переходе от учёта отдельных букв к учёту их пар (количество учитываемой информации увеличивается в 2 раза), в то же время при учёте трех букв количество информации по сравнению с учётом двух букв увеличивается лишь на треть.

Заметим, что величины $H$, представленные в табл. 2.1, рассчитываются просто – по формуле Хартли, т.е. для русского языка это $\log_2 33 \approx 5$ бит, а для английского языка $\log_2 27 \approx 4{,}76$.

## 2.6. Виды энтропий и соотношения между ними

В предыдущем параграфе даны понятия (безусловной) энтропии источника, условной энтропии, энтропии объединения. Эти виды энтропий, соотношения между ними, а также диаграммы Эйлера – Венна, изображающие соответствующие источники, представлены в табл. 2.2.

*Таблица 2.2*

| Обозначение | Название | Соотношения энтропий | Диаграмма |
|---|---|---|---|
| $H(X)$<br>$H(Y)$ | Безусловная энтропия | $H(X) \geq H(X/Y)$<br>$H(X) = H(X/Y) + H(X \cdot Y)$<br>$H(Y) \geq H(Y/X)$<br>$H(Y) = H(Y/X) + H(X \cdot Y)$ |  |
| $H(X/Y)$<br>$H(Y/X)$ | Условная энтропия | $H(X/Y) = H(X) - H(X \cdot Y)$<br>$H(Y/X) = H(Y) - H(X \cdot Y)$ |  |
| $H(X,Y) = H(Y,X)$ | Энтропия объединения | $H(X,Y) = H(X) + H(Y/X)$<br>$H(X,Y) = H(Y) + H(X/Y)$<br>$H(X,Y) = H(X) + H(Y) - H(X \cdot Y)$ |  |
| $H(X \cdot Y) = H(Y \cdot X)$ | Взаимная энтропия | $H(X \cdot Y) = H(X) - H(X/Y)$<br>$H(X \cdot Y) = H(Y) - H(Y/X)$<br>$H(X \cdot Y) = H(X,Y) - H(X/Y) - H(Y/X)$ |  |

В данной таблице только взаимная энтропия $H(X \cdot Y)$ не пояснялась ранее. Она характеризует энтропию (информацию), одинаковую для источников $X$ и $Y$.

Условную энтропию также называют зависимой энтропией.

Энтропию объединения также называют совместной или коррелированной.

## 2.7. Контрольные вопросы и задания

1. Схема <источник – приёмник> сообщений.

2. Понятие статистического (вероятностного) источника сообщений.

3. Функция приёмника сообщений в схеме <вероятностный источник–приёмник>.

4. Понятие и определение информации по Р. Хартли.

5. Количество информации по Р. Хартли.

6. Перечислите и поясните процедуры поиска элемента (объекта) в массиве.

7. Дихотомический (бинарный) поиск.

8. Организация памяти приёмника для дихотомического поиска.

9. Вывод формулы энтропии источника с идеальным набором частот сообщений.

10. Формулы количества информации и средней длины программы при дихотомической идентификации.

11. Вывод (с использованием модели длинного слова) формулы энтропии источника с реальным набором частот (вероятностей сообщений).

12. Объясните причину равенства $H = I$ (энтропии и количества информации).

13. Функция неожиданности сообщения (по А. Реньи).

14. Понятие и определение энтропии по А. Реньи.

15. Свойство энтропии источника.

16. Понятие и определение энтропии объединения двух источников.

17. Вывод формулы энтропии объединения двух статистически-зависимых источников.

18. Понятия и определения частной условной и условной энтропии.

19. Энтропия объединения двух независимых и зависимых источников.

20. Энтропия объединения нескольких статистически-зависимых источников.

21. Понятие и определение марковских цепей первого, второго и $n$-го порядков; описание источников марковских цепей разного порядка.

22. Понятия и определения других видов энтропий.

23. Какое формальное понятие характеризует вероятностный источник, а какое приёмник сообщений?

# 3. ЭЛЕМЕНТЫ ТЕОРИИ КОДИРОВАНИЯ

## 3.1. Основные положения о кодировании сообщений

В данном разделе рассматриваются положения, соответствующие фундаментальному понятию кода из первого раздела и представляющие ***особенности кодирования в технических системах*** [6, 7].

***Цель и суть любого кодирования в широком смысле*** – это представление сообщений в форме, удобной для разнообразной последующей обработки, в том числе для передачи, хранения, выполнения арифметических и логических операций. С учётом этого даже простое усиление аналогового сигнала по амплитуде и разнообразные модуляции сигнала – все это является ***кодированием***. Такое определение также совпадает с понятием кода в теории М. Мазура.

В узком смысле слова (в информатике) кодированием принято называть отображение сообщений кодовыми словами. При этом исходные сообщения сами могут быть в форме кодовых слов, тогда речь идёт о представлении одних кодовых слов другими.

***В технических системах кодирование используется для следующих конкретных целей:***

1) обеспечения построения простой и надежной аппаратуры, предназначенной для обработки закодированных сообщений;

2) защиты сообщений от помех (при их обработке, передаче по каналам связи, хранении); для этого используется помехоустойчивое кодирование;

3) компрессии или сжатия информации, т.е. для компактного представления данных; в этом случае применяется эффективное (оптимальное) кодирование;

4) сжатия информации с последующей защитой ее от помех; при этом используется двойное последовательное кодирование;

5) обнаружения и исправления ошибок при выполнении арифметико-логических операций; в этих случаях применяются арифметические коды.

Заметим, что ***для построения именно простой надежной аппаратуры кодовые слова и числа в цифровых машинах являются двоичными, а не троичными или, например, десятичными.***

Способы шифрования, изучаемые криптографией в рамках данного раздела и теории кодирования не рассматриваются, т.к. выделяются в самостоятельную дисциплину как один из способов обеспечения защиты информации то несанкционированного доступа. В то же время все вопросы шифрования полностью соответствуют теории Мазура.

## 3.2. Эффективное кодирование

### 3.2.1. Цель и идея эффективного кодирования

***Целью эффективного кодирования*** является представление массивов сообщений с размером алфавита $M$ компактными текстами, которые записаны кодовыми словами, составленными из символов алфавита меньшей мощности: $m < M$.

Так как сравнивать эффективность разных кодов с помощью длины закодированных последовательностей затруднительно из-за отсутствия представительных выборок массивов данных или эталонного массива (текста), то для оценки эффективности используется нормированная величина, инвариантная к размеру текста.

Таким образом, показателем качества эффективного кода является средняя длина (точнее математическое ожидание) кодового слова, определяемая в виде:

$$L = \sum_{j=1}^{M} P_j \cdot L_j, \qquad (3.1)$$

где $P_j$ – вероятность получения данного кодового слова;

$L_j$ – длина кодового слова $i$-го сообщения;

$M$ – количество разных кодовых слов (мощность алфавита сообщений).

Идея «сжатия» текста состоит в том, что наиболее частые сообщения кодируются короткими словами, а более редкие – длинными. При этом средняя длина кодового слова будет минимальна. Заметим, что и в естественных языках наиболее частые слова – короткие, а редкие – длинные.

При сжатии информации желательно обойтись без разделителей между кодовыми словами, так как они сильно удлиняют закодированную последовательность. Для решения этой задачи кодовые слова необходимо строить так, чтобы более длинные из них не начинались с символов более коротких. Множество построенных таким способом кодовых слов, в которых все приставки (префиксы) различны, называется ***префиксным кодом***. При этом возможна однозначная дешифрация закодированного массива, если известно его начало, и в нем нет искажений отдельных «разрядов» помехами. Иначе ошибочное чтение «первого» или «очередного» слова распространяется обычно на несколько последующих слов. Такое неоднократно возобновляющееся ошибочное чтение называется ***треком ошибки***.

Леон Георг Крафт (1949 г.) установил ограничение для существования множества слов разделимого префиксного кода мощностью $M$, записанных в алфавите $m$, в виде:

$$\sum_{k=1}^{M} m^{-n_k} = \sum_{k=1}^{M} \frac{1}{m^{n_k}} \leq 1. \qquad (3.2)$$

### 3.2.2. Теорема Крафта–Макмиллана

Пусть $M$ – мощность множества кодируемых сообщений и соответствующих им кодовых слов; $m$ – мощность алфавита символов, используемых для построения кодовых слов; $n_k$ – длина кодового слова, представляющего $k$-е сообщение; $l_j$ – количество разных слов длины $j$; $n$ – максимальная длина слова.

Пусть для построения кодовых слов используется алфавит $\{a, b, c, d, e, f, g, h\}$; $m = 8$.

Сначала для построения ***однобуквенных слов*** используем следующие символы $\{\langle a\rangle, \langle b\rangle\}$.

При этом очевидно, что число однобуквенных слов $l_1 = 2$ и в общем случае

$$l_1 \leq m.$$

Далее, для построения ***двухбуквенных слов*** используем следующие пары, не начинающиеся с выбранных ранее однобуквенных слов

$$\{\langle da\rangle, \langle ea\rangle, \langle db\rangle, \langle cc\rangle\};\ l_2 = 4$$

и в общем случае

$$l_2 \leq (m - l_1) \cdot m = m^2 - m \cdot l_1,$$

где $(m - l_1)$ – число неиспользованных однобуквенных слов, а второй компонент соответствует любому второму символу. Таким образом, правая часть неравенства представляет мощность множества всех двухбуквенных слов. Это множество строится как прямое произведение множества символов, неиспользованных для построения однобуквенных слов, на множество всех символов.

Для построения ***трёхбуквенных слов*** используем следующие тройки, не начинающиеся с выбранных ранее двух- и однобуквенных слов

$$\{\langle dca\rangle, \langle cbc\rangle, \langle ddd\rangle, \langle hag\rangle, \langle faa\rangle\}, l_3 = 5$$

и в общем случае

$$l_3 \leq ((m - l_1) \cdot m - l_2) \cdot m = m^3 - m^2 \cdot l_1 - m \cdot l_2.$$

При этом правая часть неравенства представляет мощность множества всех трёхбуквенных слов, которое строится как прямое произведение множества неиспользованных двухбуквенных слов на множество всех букв.

Наконец, используя такую же процедуру отбора и построения всё более длинных кодовых слов и обобщая вышеприведённое неравенство, запишем, что в общем случае число слов длиной *n* ограничено следующим неравенством

$$l_n \leq m^n - m^{n-1} \cdot l_1 - m^{n-2} \cdot l_2 - \ldots - m \cdot l_{n-1}.$$

Разделим последнее неравенство на $m^n$ и перепишем его в виде

$$\frac{1}{m^n} \cdot l_n + \frac{1}{m^{n-1}} \cdot l_{n-1} + \ldots + \frac{1}{m^2} \cdot l_2 + \frac{1}{m} \cdot l_1 \leq 1.$$

В компактной форме это неравенство имеет вид

$$\sum_{j=1}^{n} \frac{1}{m^j} \cdot l_j \leq 1, \tag{3.3}$$

где $m^j$ – мощность множества всех кортежей $(A_j)$ длиной $j$ в алфавите $m$; $l_j$ – мощность множества кодовых слов $(B_j)$, выбранных из множества всех кортежей $(A_j)$, при этом $B_j \subseteq A_j$; $\frac{1}{m^j} \cdot l_j$ – ***относительная мощность множества кодовых слов*** на множестве всех кортежей $A_j$.

С учётом введенных обозначений, неравенство (3.3) можно интерпретировать следующим образом: суммарное относительное число кодовых слов на множестве множеств всех кортежей разной длины $\{A_1, A_2, \ldots, A_j, \ldots, A_n\}$ для префиксного кода должно быть не больше единицы.

Заменим произведение $\frac{1}{m^j} \cdot l_j$ суммой компонентов в количестве $l_j$ и перепишем формулу (3.3) в виде

$$\sum_{j=1}^{n}(\underbrace{\frac{1}{m^j}+\frac{1}{m^j}+\cdots+\frac{1}{m^j}}_{\left|\leftarrow \quad l_j \quad \rightarrow\right|}) \leq 1.$$

Заметим, что в последнем неравенстве слагаемые компоненты представляют разные кодовые слова одинаковой длины $j$. С учетом этого заменим суммирование по разным словам одинаковой длины суммированием по всем разным словам в виде

$$\sum_{k=1}^{M}\frac{1}{m^{n_k}} = \sum_{k=1}^{M} m^{-n_k} \leq 1, \qquad (3.4)$$

где $\frac{1}{m^{n_k}}$ – относительная мощность отдельного кодового слова (представляющего $k$-$е$ сообщение) на множестве всех кортежей $\left(A_{n_k}\right)$ длиной $n_k$.

Из неравенств (3.2) и (3.4) видно, что $A_{n_k}$ – это одно и то же множество для кортежей одинаковой длины, имеющее лишь разное обозначение.

Таким образом, доказано, что множество слов разделимого префиксного кода мощностью $M$ существует, если выполняется неравенство Крафта.

Интерпретируем неравенство Крафта с учётом введенных пояснений следующим образом: ***суммарное относительное число отдельных кодовых слов на множестве всех кортежей A должно быть не больше единицы***.

Рассмотрим с другой точки зрения $k$-е слагаемое $1/m^{n_k}$ суммы (3.4). Знаменатель этой суммы $m^{n_k}$ представляет собой мощность множества кортежей (слов) длиной $n_k$ всего множества последовательностей, составленных из элементов алфавита $m$, которые потенциально могут представлять $k$-е сообщение. Только одна из них выбирается на роль кодового слова.

Пусть $m^{n_1} \leq m^{n_2} \leq \cdots \leq m^{n_M}$, и $A$ – полное $m$-арное дерево глубины $n_M$ (рис. 3.1). Каждое слово длиной $n_k \leq n_M$ $m$-арного алфавита соответствует узлу дерева на уровне $n_k$.

При выборе очередного слова длиной $n_k$ от полного дерева отсекается поддерево с корнем в узле, соответствующем данному слову, и имеющее мощность (количество узлов) $m^{n_k}$. Таким образом, $1/m^{n_k}$ – это доля мощности отсекаемого поддерева в мощности всего дерева $A$.

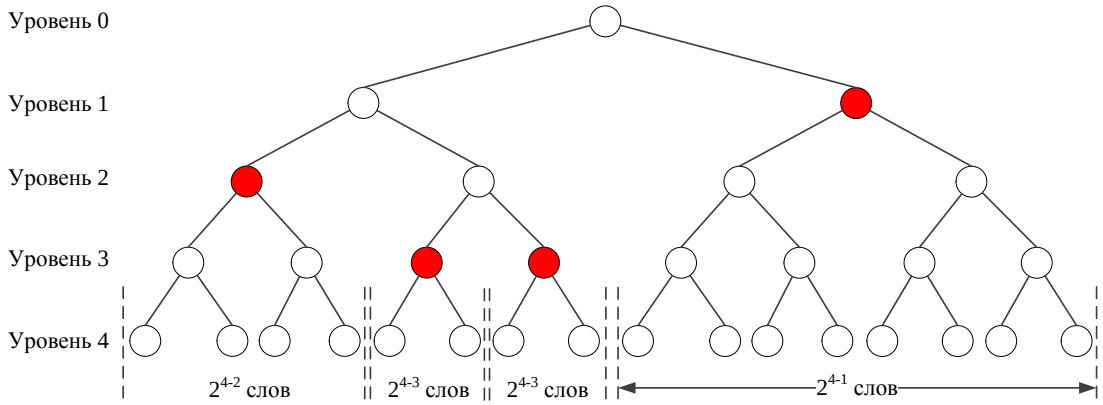



*Рис. 3.1.* Дерево всех возможных слов длиной до 5 двоичного алфавита

Таким образом, неравенство Крафта связывает мощности алфавитов исходной и закодированной последовательностей $(M, m)$ с длиной кодовых слов $n_k$ результирующей последовательности.

Для конкретного кода возможны три варианта:

1. Неравенство Крафта строго выполняется: $\sum_{k=1}^{M} m^{-n_k} < 1$; в этом случае код является избыточным и закодированный текст может быть прочитан однозначно.

2. Неравенство Крафта выполняется в виде равенства: $\sum_{k=1}^{M} m^{-n_k} = 1$; тогда код безызбыточен (оптимален) и закодированный текст может быть прочитан однозначно.

3. Неравенство Крафта не выполняется: $\sum_{k=1}^{M} m^{-n_k} > 1$; закодированный текст не может быть прочитан однозначно без разделительных знаков.

Учитывая статистические свойства источника сообщений, можно минимизировать среднее число двоичных символов, требующихся для отображения одного сообщения, что при отсутствии помех (шума) позволяет уменьшить время передачи или необходимую емкость массива данных.

### 3.2.3. Теорема Шеннона об эффективном кодировании для дискретного канала без помех

Эффективное кодирование сообщений с произвольными вероятностями основано на теореме Клода Элвуда Шеннона для дискретного канала без шума. Шеннон доказал, что мультимножество сообщений (совокупность сообщений, в которой некоторые из них неоднократно повторяются) некоторого алфавита можно закодировать так, что средняя длина двоичного слова (кодирующего отдельное сообщение) будет сколь угодно близка к энтропии источника этих сообщений, но не меньше этой величины, т.е.

$$L \geq H. \tag{3.5}$$

### *Утверждение теоремы К. Шеннона об эффективном кодировании для дискретного канала без шумов*

При кодировании множества сообщений $X$ с энтропией $H(X)$ в алфавите мощностью $m$, при условии отсутствия помех на всех этапах (передачи и хранения), средняя длина кодового слова не может быть меньше частного от деления энтропии на количество информации в одном элементе, то есть

$$L \geq \frac{H(X)}{\log_2 m}. \tag{3.6}$$

Если вероятности сообщений не являются целочисленными отрицательными степенями числа $m$, то есть $P(x_k) \neq m^{-n_k}$, то точное достижение указанной нижней границы невозможно, и $L > H(X)/\log_2 m$; но если кодируются не отдельные сообщения, а достаточно длинные их блоки, то к этой границе можно приблизиться сколь угодно близко и

$$L \approx H(X)/\log_2 m. \tag{3.7}$$

Рассмотрим ***формальное доказательство теоремы***.

Из теории вероятностей известно, что для любых двух распределений вероятностей $\{P(x_k)\}$ и $\{Q(x_k)\}$ на одном и том же множестве сообщений выполняется неравенство

$$\sum_{k=1}^{M} P(x_k) \cdot \log_2 \frac{P(x_k)}{Q(x_k)} \geq 0, \tag{3.8}$$

запишем логарифм частного разностью логарифмов:

$$\sum_{k=1}^{M} P(x_k) \cdot (\log_2 P(x_k) - \log_2 Q(x_k)) \geq 0. \tag{3.9}$$

Распишем левую часть неравенства (3.9) в виде разности двух сумм:

$$\sum_{k=1}^{M} P(x_k) \cdot \log_2 P(x_k) - \sum_{k=1}^{M} P(x_k) \cdot \log_2 Q(x_k) \geq 0. \tag{3.10}$$

Перенесём первое слагаемое в правую часть неравенства

$$-\sum_{k=1}^{M} P(x_k) \cdot \log_2 Q(x_k) \geq -\sum_{k=1}^{M} P(x_k) \cdot \log_2 P(x_k). \tag{3.11}$$

Очевидно, правую часть неравенства представляет выражение (2.13) для энтропии $H(X)$. Перепишем неравенство (3.11) в виде

$$-\sum_{k=1}^{M} P(x_k) \cdot \log_2 Q(x_k) \geq H(X). \tag{3.12}$$

Так как по допущению теоремы распределение вероятностей $\{Q(x_k)\}$ может быть любым, зададим вероятности сообщений этого распределения следующим образом:

$$Q(x_k) = \frac{m^{-n_k}}{\sum_{r=1}^{M} m^{-n_r}}. \tag{3.13}$$

Подставляя это выражение в левую часть неравенства (3.12), получим

$$-\sum_{k=1}^{M} P(x_k) \cdot \log_2 \frac{m^{-n_k}}{\sum_{r=1}^{M} m^{-n_r}} \geq H(X), \tag{3.14}$$

запишем логарифм частного разностью логарифмов

$$-\sum_{k=1}^{M} P(x_k) \cdot \left( \log_2 m^{-n_k} - \log_2 \sum_{r=1}^{M} m^{-n_r} \right) \geq H(X). \tag{3.15}$$

Представим сумму (3.15) в виде двух слагаемых:

$$-\sum_{k=1}^{M} P(x_k) \cdot \log_2 m^{-n_k} + \sum_{k=1}^{M} P(x_k) \cdot \log_2 \sum_{r=1}^{M} m^{-n_r} \geq H(X). \tag{3.16}$$

Из неравенства Крафта

$$\sum_{r=1}^{M} m^{-n_r} \leq 1$$

следует, что

$$\log_2 \sum_{r=1}^{M} m^{-n_r} \leq \log_2 1 = 0, \tag{3.17}$$

то есть $\log_2 \sum_{r=1}^{M} m^{-n_r} \leq 0$ и поэтому второе слагаемое неравенства (3.16) только уменьшает его левую часть. Значит, исключение данного слагаемого лишь усиливает неравенство (3.16), которое теперь можно записать в виде

$$-\sum_{k=1}^{M} P(x_k) \cdot \log_2 m^{-n_k} \geq H(X). \tag{3.18}$$

Перепишем логарифм степени в виде произведения

$$\sum_{k=1}^{M} P(x_k) \cdot n_k \cdot \log_2 m \geq H(X). \tag{3.19}$$

Обозначим

$$\sum_{k=1}^{M} P(x_k) \cdot n_k = L. \tag{3.20}$$

Из теории вероятностей известно, что сумма произведений вероятностей (частот) слов на их длины есть математическое ожидание длины слова или другими словами средняя длина слова. Подставим (3.20) в (3.19) и запишем его в виде

$$L \cdot \log_2 m \geq H(X). \tag{3.21}$$

Окончательно получим

$$L \geq \frac{H(X)}{\log_2 m}. \tag{3.22}$$

Так как в современной цифровой вычислительной технике кодовые слова представлены в алфавите $\{0,1\}$ мощностью $m = 2$, то утверждение теоремы Шеннона выглядит проще, а именно $L \geq H$.

Покажем, что утверждение (3.22) может быть получено проще.

Обозначим и определим количество информаций, необходимых для идентификации конкретного сообщения $i(x_k) = -\log_2 P(x_k)$. Максимальное количество информации, необходимое для идентификации любой буквы алфавита мощностью $m$, определяется в виде

$$i_{max} = \log_2 m. \tag{3.23}$$

При кодировании конкретного сообщения $x_k$ информацию терять нельзя, поэтому для определения длины $n_k$ кодового слова, представляющего это сообщение, справедливо неравенство

$$n_k \geq \frac{i(x_k)}{i_{max}} = \frac{-\log_2 P(x_k)}{\log_2 m}. \tag{3.24}$$

Так как длина кодового слова – натуральное число, то очевидно, что эта величина должна определяться в диапазоне

$$\frac{-\log_2 P(x_k)}{\log_2 m} \leq n_k \leq \frac{-\log_2 P(x_k)}{\log_2 m} + 1, \tag{3.25}$$

то есть нужно брать «ближайшее целое сверху» к величине $-\log_2 P(x_k)/\log_2 m$.

Усредним неравенства (3.25) по множеству кодируемых сообщений:

$$\frac{-\sum_{k=1}^{M} P(x_k) \cdot \log_2 P(x_k)}{\log_2 m} \leq \sum_{k=1}^{M} P(x_k) \cdot n_k \leq \frac{-\sum_{k=1}^{M} P(x_k) \cdot \log_2 P(x_k)}{\log_2 m} + 1, \tag{3.26}$$

с учётом введенных выше обозначений запишем оба неравенства в виде

$$\frac{H(X)}{\log_2 m} \leq L \leq \frac{H(X)}{\log_2 m} + 1. \tag{3.27}$$

Как видно, левое неравенство соответствует утверждению первой части теоремы Шеннона.

Если вероятности реального и идеального распределений $\{P(x_k)\}$ и $\{Q(x_k)\}$ совпадают и имеют следующие значения

$$P(x_k) = Q(x_k) = m^{-n_k},$$

то вместо неравенства (3.27) имеем равенство, задающее ***границу Шеннона для эффективных кодов*** в виде

$$L = H \,/\, \log_2 m. \tag{3.28}$$

При $m = 2$ такие идеальные вероятности кратны 2, а $L = H$. Примеры с такими вероятностями сообщений рассмотрены выше.

Однако в реальной ситуации, обычно, вероятности сообщений не кратны 2. При этом не удается разбивать упорядоченные по убыванию вероятностей сообщения на буквально равные по суммарным вероятностям подмножества и на каждом шаге идентификации сообщения (когда нужно выбрать одно из двух подмножеств) выбор будет уже не оптимален, и средняя длина кодового слова будет больше энтропии (в отличие от случая, где вероятности сообщений кратны 2). Однако, если оказывается, что ***избыточность*** построенного кода невелика, то он принимается для эксплуатации, иначе следует построить другой код.

Докажем ***утверждение второй части теоремы Шеннона***. Если статистические вероятности сообщений сильно отличаются от идеальных ($P(x_k) = m^{-n_k}$), то средняя длина кодовых слов может значительно превышать границу, указанную Шенноном ($H(X)/\log_2 m$), и соответствующий код будет иметь большую ***избыточность***. При этом код будет не оптимальным, а лишь эффективным. Добиться построения оптимального кода можно путём кодирования целых последовательностей исходных сообщений (блоков) одним кодовым словом.

Обозначим длину блока исходных сообщений буквой $K$. Число всех возможных разных блоков (мощность декартова гиперкуба) равно $M^K$. Допустим, что исходные сообщения $y_j$, формирующие отдельные блоки, статистически независимы между собой. Аналогично тому, как это сделано для отдельных исходных сообщений, можно определить границы для длины кодового слова $n_j$, представляющего отдельный блок следующими неравенствами

$$\frac{-\log_2 P(y_j)}{\log_2 m} \le n_j \le \frac{-\log_2 P(y_j)}{\log_2 m} + 1, \tag{3.29}$$

где $P(y_j)$ – вероятность формирования $j$-го блока из множества возможных блоков мощностью $M^K$. При этом для кодирования отдельных блоков кодовыми словами префиксного кода должно выполняться неравенство Крафта:

$$\sum_{j=1}^{M^K} m^{-n_j} \le \sum_{j=1}^{M^K} P(y_j) = 1. \tag{3.30}$$

Усредняя неравенство (3.29) по множеству всех блоков, получим

$$\frac{-\sum_{k=1}^{M^K} P(y_j)\log_2 P(y_j)}{\log_2 m} \le \sum_{k=1}^{M^K} P(y_j)n_j \le \frac{-\sum_{k=1}^{M^K} P(y_j)\log_2 P(y_j)}{\log_2 m} + 1. \quad (3.31)$$

Используя обозначения для энтропии блока $H(Y)$ и для средней длины кодового слова, представляющего блок $L_{\text{блока}}$, перепишем неравенства в виде

$$\frac{H(Y)}{\log_2 m} \le L_{\text{блока}} \le \frac{H(Y)}{\log_2 m} + 1. \quad (3.32)$$

В соответствии с допущением теоремы, сообщения, формирующие блок, статистически независимы, и поэтому энтропия блока (энтропия объединения) равна сумме одинаковых энтропий $H(X)$ отдельных сообщений, т.е.

$$H(Y) = K \cdot H(X). \quad (3.33)$$

После подстановки формулы (3.33) в (3.32) получим

$$\frac{K \cdot H(X)}{\log_2 m} \le L_{\text{блока}} \le \frac{K \cdot H(X)}{\log_2 m} + 1. \quad (3.34)$$

Разделим оба неравенства на длину блока К. С учётом сокращений получим

$$\frac{H(X)}{\log_2 m} \le \frac{L_{\text{блока}}}{K} \le \frac{H(X)}{\log_2 m} + \frac{1}{K}. \quad (3.35)$$

Обозначим $L_{\text{блока}}/K = L'$, где $L'$ – приведенная средняя длина кодового слова, приходящаяся на одно сообщение блока. Очевидно, что при увеличении длины блока, когда $K \gg 1$, величина $1/K \approx 0$, а оба неравенства вырождаются в одно равенство:

$$L' = \frac{H(X)}{\log_2 m}. \quad (3.36)$$

Таким образом, доказано утверждение второй части теоремы Шеннона.

Следует отметить, что ***теорема Шеннона*** не указывает конкретного способа построения кода, так как ***является теоремой «существования»***, которая

только доказывает, что оптимальные коды существуют, и указывает для них границу средней длины слова.

Однако из теоремы следует, что при построении эффективного кода необходимо стремиться к тому, чтобы каждый символ кодовой комбинации нес максимальное количество информации (равное $\log_2 m$). Для этого каждый символ должен быть по возможности равновероятным и статистически независимым от других символов.

### 3.2.4. О законе сохранения числа идентифицирующих информаций

Дадим интерпретацию неравенству (3.21), которое, по нашему мнению, представляет собой ***закон сохранения количества идентифицирующих информаций***. Из этого закона следует, что любое перекодирование множества исходных сообщений (мощности $M$) словами в алфавите меньшей мощности ($m < M$) не должно приводить к потере информации.

Кроме того, можно дать и физическую интерпретацию полученному неравенству, которое можно рассматривать как соотношение неопределённостей.

Введём соответствия для двух ситуаций:

1) если кодовые слова передаются последовательно, то

$n_k \sim \Delta t$ и $\log_2 m \sim \Delta E$, а $H(X) \sim \hbar$;

2) если кодовые слова передаются параллельно, то

$n_k \sim \Delta x$ и $\log_2 m \sim \Delta p$, а $H(X) \sim \hbar$.

Запишем соответствующие неравенства

$$L \cdot \log_2 m \geq H(X);$$

$$\Delta t \cdot \Delta E \geq \hbar;$$

$$\Delta x \cdot \Delta p \geq \hbar.$$

Второе неравенство можно интерпретировать так: неопределённость момента появления слова, помноженная на неопределённость количества энергии, переносимой этим словом, не меньше некоторой константы, аналогичной энтропии, размерностью «время, умноженное на энергию».

Третье неравенство можно интерпретировать так: неопределённость позиции слова, помноженная на неопределённость импульса, не меньше некото-

рой константы, аналогичной энтропии, размерностью «время, умноженное на энергию».

### 3.2.5. О факторах и средствах, обеспечивающих сжатие данных

***Абсолютная избыточность эффективного кода*** определяется разностью средней длины кодового слова и энтропии, т.е. в виде

$$\Delta L = L - H. \tag{3.37}$$

По существу, Шеннон доказал ***теорему существования границы оптимальных кодов***, которая задана в виде

$$L_{min} = H/\log_2 m \tag{3.38}$$

для $m = 2, L_{min} = H$.

Если при кодировании отдельных сообщений исходного текста избыточность кода оказывается велика, то следует кодировать в исходной последовательности не отдельные сообщения, а целые их группы – блоки. При этом можно приблизиться к границе, указанной Шенноном, когда $L' \approx H$. ***Чем длиннее блоки, тем эффективнее код***. Заметим, что ***дополнительный эффект сжатия информации*** получается здесь не за счет того, что учитываются более дальние статистические связи. Этот эффект достигается тем, что алфавит исходных сообщений в количестве $M$ заменяется большим набором макросообщений или блоков в количестве $M^K$, где $K$ – длина блока. Такой набор удается точнее разбить на близкие по суммарным вероятностям подмножества. В пределе, при $K \to \infty$, вероятности появления этих блоков становятся одинаковыми, так как они встречаются в последовательности по одному разу, и мы приходим к ситуации, описанной Р. Хартли. В этом случае для идентификации любого блока требуются программы одинаковой длины (кодовые слова одинаковой длины). Для сравнения эффективности разных кодов, кодирующих отдельные сообщения или блоки из $K$ сообщений, следует использовать приведенную меру

$$L' = L_{\text{блок}}/K,$$

так как для кодирования блока из $K$ сообщений требуется, очевидно, более длинное слово, чем для отдельного сообщения.

***Основной эффект сжатия*** массива данных, записанного кодовыми словами, получается за счет того, что более частые сообщения кодируются корот-

кими словами, а редкие – длинными, так что средняя длина кодового слова получается минимальной.

***Дополнительный эффект сжатия*** получается за счет устранения разделительных знаков между словами оптимального кода. При этом более длинные кодовые слова строятся таким образом, что не начинаются с символов коротких.

Однако очевидны ситуации, когда на вход приемника поступают ***статистически зависимые сообщения***. Реальные тексты, файлы – это последовательности вовсе не случайных сообщений. Однако ***оптимальный код предназначен для кодирования только статистически независимых сообщений***. В отмеченных ситуациях требуется предварительно осуществлять ***декорреляцию последовательности сообщений***. Для этого есть два способа:

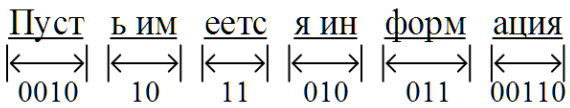
– декорреляция блоками;

– декорреляция $L$-граммами (более эффективна).

При первом способе не удается учесть статистические связи на границах между блоками. Это устраняется при кодировании $L$-граммами.

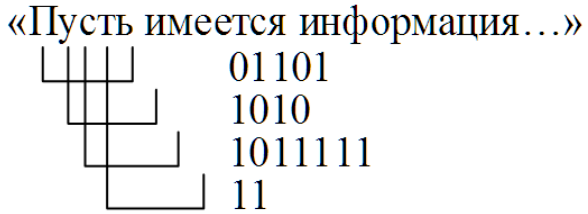
При декорреляции любым способом исходный текст заменяется другим текстом, который записан макросимволами – блоками. Этот новый текст состоит из сообщений, которые статистически почти не зависимы, и кодируется обычным способом.

Рассмотрим декорреляцию (сегментацию) и кодирование блоками по четыре сообщения следующей фразы: «Пусть имеется информация...» →




Если $n$ – длина исходной последовательности, $K$– длина блока, то число заменяющих ее блоков $N_{\text{блок}} = n/K$.

Декорреляция и кодирование $L$-граммами по четыре сообщения имеют вид:




Число $L$-грамм (по $K$ сообщений), заменяющих исходную последовательность длиной $n$ сообщений, определяется в виде $N_{\text{блок}} = n - (K - 1)$.

Специфика технической реализации оптимального кодирования обусловлена тем, что кодовые слова имеют разную длину. Если кодовые слова передаются по каналу связи через равные интервалы времени $\Delta t$ или размещаются в памяти машины, то будут появляться интервалы времени $\Delta\tau_i$ между словами, когда канал не занят сообщениями, а в ячейках памяти присутствуют незаполненные разряды (как это показано в табл. 3.1 и на рис. 3.2).

*Таблица 3.1*

$n = 8$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | 0 | 0 | 1 |
| 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| | | | | | 1 | 1 | 0 |
| | | | | | | 0 | 1 |
| | | | | | 1 | 0 | 0 |
| | | | | 1 | 0 | 1 | 0 |

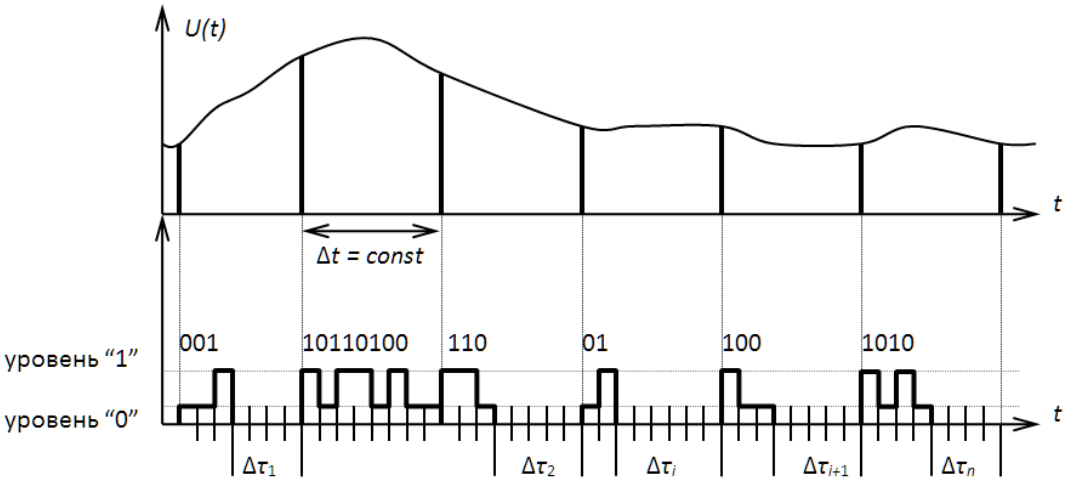



*Рис. 3.2.* Неэффективная передача сообщений с интервалами $\Delta\tau_i$

Несколько отвлекаясь от темы данного раздела, заметим, что на рис. 3.2 представлены два поперечных множества сообщений, в каждом из которых по восемь сообщений. Первичные сообщения представляют амплитуды сигналов, зафиксированные в определённые моменты времени, а вторичные сообщения являются кодовыми словами. Будучи упорядоченными оба множества являются информационными цепями, связи между сообщениями этих поперечных множеств являются кодами. Число таких кодов – 8.

Для эффективной передачи сообщений по каналу и при заполнении массива данных в памяти нужно установить буферное запоминающее устройство – БЗУ (накопитель) с числом разрядов не меньше удвоенного числа разрядов самого длинного кодового слова. Из этого накопителя в канал связи выдаются не отдельные кодовые слова, а непрерывная последовательность «1» и «0» с по-

стоянной скоростью и без пробелов. При записи в ячейки памяти на выходе накопителя формируются слова одинаковой длины по числу разрядов в них. Соответственно при считывании данных из массива или для выделения кодовых слов нужен еще один такой же накопитель. Если кодирование осуществляется блоками или *L*-граммами, то на стороне источника сигналов необходим формирователь блоков или *L*-грамм, а на стороне приемника сообщений – их распознаватель (расформирователь).

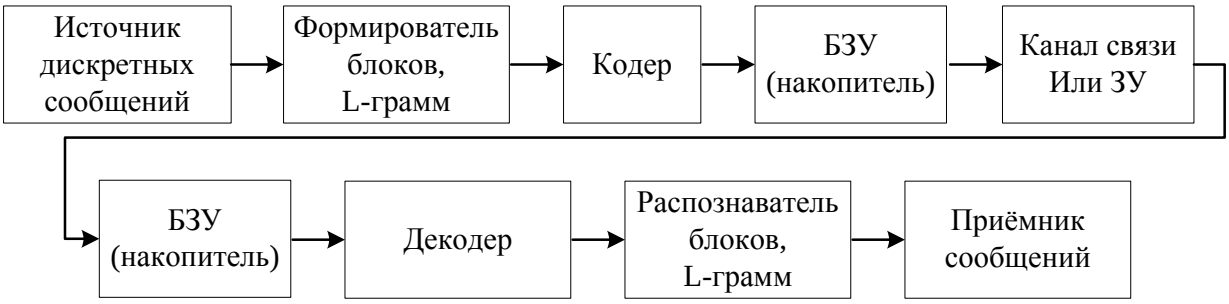



*Рис. 3.3.* Система передачи данных эффективным кодом

Полная структурная схема системы эффективного кодирования, передачи и декодирования сообщений представлена на рис. 3.3.

### 3.2.6. Построение эффективного кода Шеннона – Фэно

Методы построения эффективных кодов были впервые разработаны Шенноном и Фэно и существенно не различаются, поэтому соответствующий код получил название кода Шеннона – Фэно. Исходные данные для построения кода – это алфавит сообщений и их вероятности. Код строится следующим образом: все множество сообщений с их вероятностями выписывается в таблицу в порядке убывания вероятностей. Затем все множество разбивается на два подмножества так, чтобы суммы вероятностей в каждом из них были по возможности одинаковы. Всем сообщениям подмножества с наиболее вероятными сообщениями в качестве первого символа их кодовых слов приписывается «0», а всем сообщениям второго подмножества назначается «1». Каждое из подмножеств, в свою очередь, аналогично разбивается (без перемешивания!) на два меньших подмножества, обозначается собственно «0» или «1» и т.д. Такой процесс повторяется до тех пор, пока в каждом подмножестве не останется по одному сообщению.

***Примечание:*** При формировании блоков или *L*-грамм кодирование осуществляется точно так же; только в этом случае исходными данными для построения кода являются частоты (в пределе – вероятности) блоков или *L*-грамм.

Пример построения эффективного кода для восьми упорядоченных по убыванию вероятностей сообщений приведен в табл. 3.2.

*Таблица 3.2*

| Сообщение | Вероятности | Разбиения | | | | | Кодовые слова |
|---|---|---|---|---|---|---|---|
| | | I | II | III | IV | V | |
| $U_1$ | 0,4 | } | | | | | 0 |
| $U_2$ | 0,3 | ⎫ | } | | | | 10 |
| $U_3$ | 0,1 | ⎪ | ⎫ | } | | | 1100 |
| $U_4$ | 0,06 | ⎪ | ⎪ | ⎭ | } | | 1101 |
| $U_5$ | 0,04 | ⎪ | ⎪ | ⎫ | ⎫ | } | 11100 |
| $U_6$ | 0,04 | ⎪ | ⎪ | ⎪ | ⎭ | } | 11101 |
| $U_7$ | 0,03 | ⎪ | ⎪ | ⎭ | ⎫ | } | 11110 |
| $U_8$ | 0,03 | ⎭ | ⎭ | | ⎭ | } | 11111 |

При построении кода Шеннона – Фэно оказываются возможными неоднозначные разбиения и при этом можно построить несколько разных кодов. Лучше использовать тот код, который имеет минимальную среднюю длину из всех других вариантов кодов. Для этого нужно построить множество различных вариантов и выбрать лучший.

### 3.2.7. Построение оптимального кода

Дэвид Хаффман (1952 г.) предложил такую процедуру построения эффективного кода, которая сразу дает наилучший вариант. Этапы этой процедуры следующие:

1. Строится таблица, число столбцов в которой на один больше, чем размер алфавита сообщений; во второй столбец вписываются по убыванию вероятности, а в первый – соответствующие сообщения.

2. Вероятности двух самых редких сообщений суммируются, образуя вероятность дополнительного псевдосообщения.

3. Вероятности сообщений из 2-го столбца, кроме двух самых редких, вместе с псевдосообщением переписываются в порядке убывания частот в 3-й столбец справа.

4. Действия (2) и (3) выполняются до тех пор, пока не образуется в последнем столбце псевдосообщение с вероятностью «1»; при этом переупорядочиваются сообщения и псевдосообщения из 3-го, 4-го и т.д. столбцов.

5. Далее для построения кодовых слов необходимо проследить все переходы данного сообщения по строкам и столбцам полученной диагональной матрицы и соответственно каждый переход закодировать символами «0» или «1».

Пример построения диагональной таблицы приведен в табл. 3.3.

*Таблица 3.3*

| $U_i$ | $P_i$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $U_1$ | 0,4 | 0,4 | 0,4 | 0,4 | 0,4 | 0,4 | 0,6 | 1 |
| $U_2$ | 0,16 | 0,16 | 0,16 | 0,18 | 0,26 | 0,34 | 0,4 | |
| $U_3$ | 0,12 | 0,12 | 0,14 | 0,16 | 0,18 | 0,26 | | |
| $U_4$ | 0,1 | 0,1 | 0,12 | 0,14 | 0,16 | | | |
| $U_5$ | 0,08 | 0,08 | 0,1 | 0,12 | | | | |
| $U_6$ | 0,08 | 0,08 | 0,08 | | | | | |
| $U_7$ | 0,04 | 0,06 | | | | | | |
| $U_8$ | 0,02 | | | | | | | |
| | $\sum P_i = 1$ | | | | | | | |

Для наглядности сопоставления отдельным сообщениям соответствующих им кодовых слов строится бинарное ***кодовое дерево,*** в вершине которого размещается псевдосообщение с вероятностью 1. От вершины этого дерева опускаются две ветви, которые оканчиваются псевдосообщениями, полученными на предпоследнем шаге построения таблицы. Ветвь сообщения с большей вероятностью кодируется символом «1», а с меньшей – «0».

Такое дихотомическое ветвление дерева продолжается до тех пор, пока не дойдем до вероятности каждого сообщения. Теперь, двигаясь по кодовому дереву сверху вниз, можно записать для каждого сообщения соответствующее ему кодовое слово. Кодовое дерево и эффективный код в виде кодирующей таблицы (табл. 3.4) для восьми сообщений представлено на рис. 3.4.

*Таблица 3.4*

| $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | $U_6$ | $U_7$ | $U_8$ |
|---|---|---|---|---|---|---|---|
| 0 | 110 | 100 | 1111 | 1110 | 1011 | 10101 | 10100 |

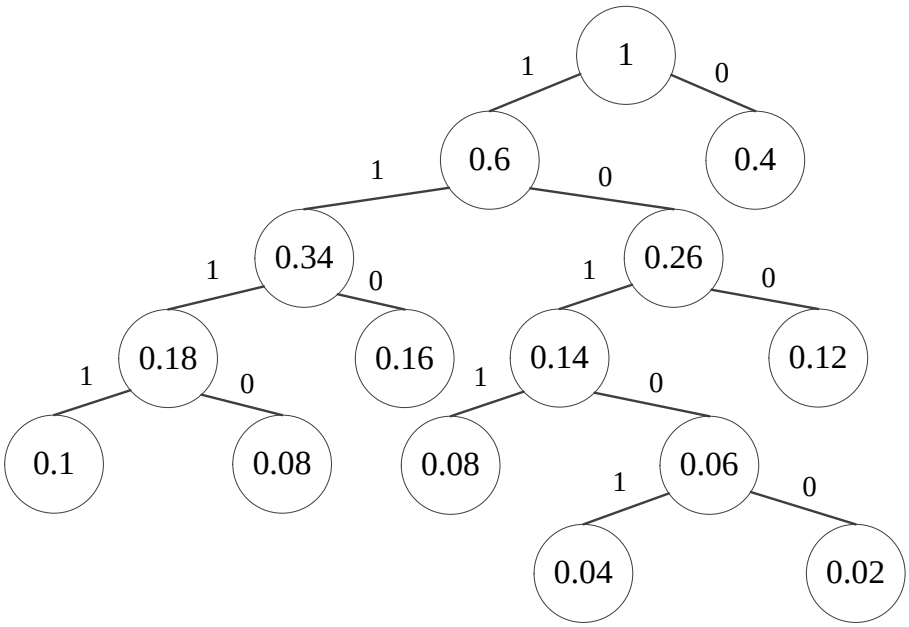



*Рис. 3.4.* Кодовое дерево

### 3.2.8. Контрольные вопросы и задания

1. Цель и суть любого кодирования в технических системах.
2. Определение кодирования в информатике.
3. Конкретные цели кодирования в технических системах.
4. Цель эффективного (оптимального) кодирования.
5. Задача эффективного кодирования.
6. Основная идея эффективного кодирования.
7. Понятие префиксного кода.
8. Показатель эффективности оптимального кода; избыточность такого кода.
9. Понятие трека ошибки.
10. Утверждение теоремы Крафта – Макмиллана.
11. Доказательство теоремы Крафта – Макмиллана.
12. Утверждение теоремы К. Шеннона об оптимальном кодировании.
13. Доказательство теоремы Шеннона об оптимальном кодировании.
14. Граница существования оптимальных кодов.
15. Способы декорреляции последовательности сообщений.
16. Особенности реализации систем оптимального кодирования.
17. Процедуры построения оптимальных кодов.
18. Особенности строения кодовых слов эффективного кода.
19. Перечислите все факторы, за счет которых достигается эффект «сжатия» текста.
20. Процедура построения эффективного кода Шеннона – Фено.
21. Процедура построения оптимального кода Хаффмана.

## 3.3. Помехоустойчивое кодирование

### 3.3.1. Цель и идея помехоустойчивого кодирования

«Теория помехоустойчивого кодирования базируется на результатах исследований, проведённых Шенноном и сформулированных им в виде основной теоремы для дискретного канала с помехами (шумом):

*при любой скорости передачи двоичных символов, меньшей, чем пропускная способность канала, существует такой код, при котором вероятность безошибочного декодирования будет сколь угодно мала; вероятность ошибки не может быть сделана сколь угодно малой, если скорость передачи больше пропускной способности канала*» [7].

Рассмотрим идею помехоустойчивого кодирования на примере кодирования сообщений словами равной длины, в которых фиксированы позиции ин-

формационных и дополнительных разрядов. Для этих целей используются так называемые равномерные разделимые блоковые коды.

Кодирующее устройство (шифратор) осуществляет следующее преобразование над входным безызбыточным *k*-разрядным кодовым словом, которое несет только полезную информацию. Кодер наращивает длину слова, увеличивая число разрядов кодового слова до $n > k$, при этом появляются ***дополнительные (избыточные, проверочные или контрольные) разряды***, кроме так называемых ***информационных (k) разрядов***, несущих полезную информацию. Таким образом, кодовое слово на выходе кодера содержит $n - k = m$ избыточных разрядов. Содержимое дополнительных (избыточных) разрядов кодер определяет в соответствии с алгоритмом кодирования на основе содержимого информационных разрядов. Избыточная информация в помехоустойчивом кодовом слове представлена содержимым определенных информационных и дополнительных разрядов. Сама же избыточная информация – это, по существу, алгоритм формирования избыточных разрядов, т.е. алгоритм кодирования, который известен дешифратору (декодеру). То есть для дешифратора данный алгоритм кодирования является избыточной информацией – это то постоянное преобразование, которое сохраняется независимо от того, какие кодовые слова передаются от источника к приемнику. Используя эту избыточную информацию, дешифратор принимает очередное слово и проверяет содержимое всех его разрядов на соответствие данному алгоритму кодирования. Если данное слово не удовлетворяет используемому алгоритму кодирования, то дешифратор делает вывод об обнаружении ошибки и в зависимости от того, «в какой степени» это соответствие не выполняется, может опознавать и исправлять некоторые ошибки.

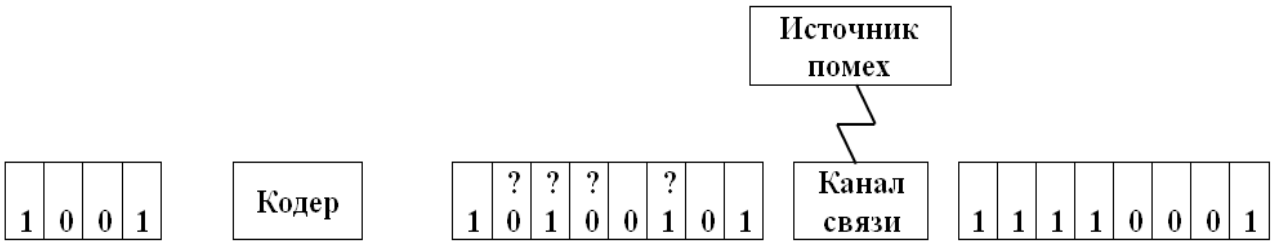



*Рис. 3.5.* Помехоустойчивое кодирование
и искажение слова помехой

Кратко это можно выразить следующим образом: ***идея помехоустойчивого кодирования*** состоит во внесении кодером избыточной информации в виде алгоритма (правил) кодирования с помощью дополнительных разрядов помехоустойчивого кодового слова, с последующей проверкой декодером этого слова на соответствие принятому алгоритму кодирования (рис. 3.5).

### 3.3.2. Элементы теории помехоустойчивого кодирования

Все множество помехоустойчивых кодов различается разнообразными алгоритмами кодирования, каждый из которых разработан для защиты кодовых слов от помех определенного характера и их кратности. Обычно различают помехи ***некоррелированные*** (при которых искажение слова в данном разряде не связано с искажением в других разрядах) и ***коррелированные*** или ***пакеты ошибок*** (при которых искажение в данном разряде приводит к искажению в других разрядах). Помехи различной интенсивности порождают ***ошибки разной кратности: однократные, 2-кратные и R-кратные***, когда в кодовом слове меняется на противоположное (искажается) содержимое 1, 2,…, $R$ разрядов одновременно (1 меняется на 0, а 0 меняется на 1).

Никакие коды не могут не только исправлять, но и обнаруживать все ошибки. Это объясняется следующим. Пусть требуется закодировать двоичным кодом $Q = 2^k$ разных сообщений ($Q$ – объем кода, $k$ – число двоичных разрядов). Для каждого безызбыточного входного слова, отображающего конкретное сообщение, декодер должен сформировать на выходе только одно помехоустойчивое слово, поэтому ***разрешенных помехоустойчивых кодовых слов*** тоже $2^k$. Все множество $n$-разрядных кодовых слов имеет мощность $2^n$, из которых в подмножестве разрешенных кодовых слов только $2^k$.

Помехоустойчивое слово на входе дешифратора может иметь искаженные разряды, поэтому все множество $n$-разрядных разных кодовых слов, которые могут иметь место на входе дешифратора, имеет мощность $2^n$. На множестве $n$-разрядных кодовых слов можно выделить, кроме подмножества разрешенных кодовых слов, подмножество ***запрещенных кодовых слов*** мощностью $(2^n - 2^k)$.

При построении помехоустойчивых кодов важно правильно использовать подмножество запрещенных кодовых слов. Это значит, что запрещенных кодовых слов должно быть как можно меньше, но их наличие должно позволять дешифратору исправлять как можно большее число разнообразных ошибок. При этом не пришлось бы сильно удлинять помехоустойчивое кодовое слово, в котором дополнительные и информационные разряды равно подвержены воздействию помех.

***Абсолютная избыточность кода*** определяется в виде $m = n - k$.

Относительная избыточность $R_n = (n - k)/n$ или $\mathrm{R_k} = (n - k)/k$.

С учетом понятия избыточности кода легко определить понятие «***оптимальный помехоустойчивый код***». Это такой код, который обеспечивает заданную корректирующую способность (обнаружение или исправление ошибок определенной кратности) минимальным числом дополнительных разрядов. Из-

быточные разряды могут быть искажены так же, как и информационные, поэтому, удлиняя слово дополнительными разрядами, мы снижаем его «помехоустойчивость».

Кодовое слово может передаваться от шифратора к дешифратору с ошибкой и без нее. Таким образом, возможны ***два варианта передачи кодового слова: правильная и неправильная***. Число вариантов правильной передачи, когда разрешенное кодовое слово, проходя путь от кодера к декодеру, трансформируется само в себя, равно $2^k$.

Следует различать ***два вида неправильной передачи***:

1) разрешенное кодовое слово на пути от кодера к декодеру трансформируется в иное разрешенное слово. В этих случаях декодер, проверяя структуру и содержимое принятого кодового слова на соответствие данному алгоритму кодирования, вынужден принять решение, что кодовое слово правильно. При этом дешифратор не только не исправит эту ошибку, но даже и не обнаружит ее. Так как каждое разрешенное слово может трансформироваться в любое другое разрешенное слово, то число вариантов такой передачи $2^k \cdot (2^k - 1)$;

2) разрешенное кодовое слово трансформируется в запрещенное. В таких случаях дешифратор способен обнаружить ошибку, а в некоторых – и исправить. Так как каждое разрешенное слово может трансформироваться в любое запрещенное слово (число которых $2^n - 2^k$), то число вариантов такой ошибочной передачи $2^k \cdot (2^n - 2^k)$.

Суммируя числа разных вариантов передачи, получим общее число вариантов передачи:

$$2^k \cdot 2^n = 2^k + 2^k \cdot (2^k - 1) + 2^k \cdot (2^n - 2^k).$$

Разные варианты передачи кодовых слов графически представлены на рис. 3.6.

Если в кодовых словах позиции избыточных и информационных разрядов фиксированы, то такой ***код называется разделимым***; если позиции не фиксированы, то такой ***код называется неразделимым***.

При построении помехоустойчивого кода следует разумно использовать подмножество запрещенных кодовых слов. Есть ***два способа разбиения подмножества запрещенных кодовых слов на непересекающиеся подмножества***. В зависимости от принятого способа разбиения возможно построение разных по сути помехоустойчивых кодов и соответственно два разных подмножества помехоустойчивых корректирующих кодов. И те, и другие коды могут обнаруживать и исправлять ошибки, но делают это по-разному.

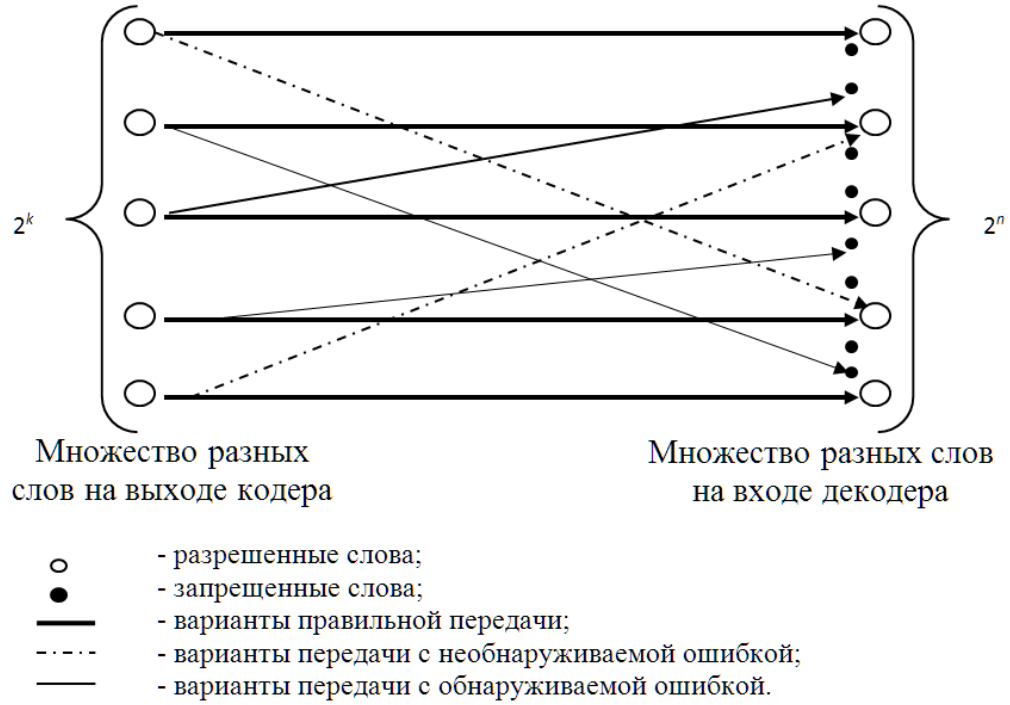



*Рис. 3.6.* Варианты передачи кодовых слов от кодера к декодеру

***1-й способ*:** разбиение всех запрещенных слов на непересекающиеся подмножества по принципу принадлежности (близости) запрещенного слова к разрешенному кодовому слову. При этом «вокруг» каждого разрешенного кодового слова группируются такие запрещенные слова, которые «ближе» к нему, чем к другим разрешенным словам (рис. 3.7). В этом случае в качестве разрешенных кодовых слов следует выбирать такие, которые составляют множество элементов, удаленных друг от друга на расстояние не меньше некоторой величины (называемой ***минимальным хэмминговым расстоянием***).

При таком способе разбиения дешифратор выносит решение в пользу того разрешенного слова, расстояние от которого до принятого слова меньше, чем до других разрешенных слов. Количество непересекающихся подмножеств запрещенных кодовых слов при этом равно числу разрешенных слов $2^k$.

***2-й способ*:** разбиение по принципу принадлежности запрещенного кодового слова к вектору ошибки или к ***классу смежности***. При таком разбиении декодер опознает не переданное ему слово, а вектор ошибки, которой оно оказалось поражено. Для этого декодер, учитывая содержимое избыточных и информационных разрядов, проверяет принятое слово на соответствие данному алгоритму кодирования и в результате вычисляет ***опознаватель (синдром) ошибки***, который указывает на принадлежность принятого слова к одному из непересекающихся подмножеств запрещенных слов (классов смежности), «порожденных» определенным вектором ошибки. В такой системе кодер должен по определенным ***правилам кодирования*** определять содержимое избыточных

разрядов на основе известного содержимого информационных разрядов. Эти правила или ***алгоритм кодирования*** представляют собой систему уравнений, в которых данными (известными величинами) являются значения информационных разрядов. Для определения содержимого каждого избыточного разряда применяется свое уравнение. Дешифратор проверяет на истинность каждое из этих уравнений, проверка дает либо «0», либо «1». Проверки всех уравнений дают множество нулей и единиц, называемое ***опознавателем ошибки***. Если опознаватель состоит только из одних нулей, декодер делает вывод об отсутствии ошибки, иначе, по виду ненулевого опознавателя, декодер может определить тип ошибки, так как опознаватель указывает на принадлежность принятого слова к подмножеству запрещенных слов, порожденных данным вектором ошибки.

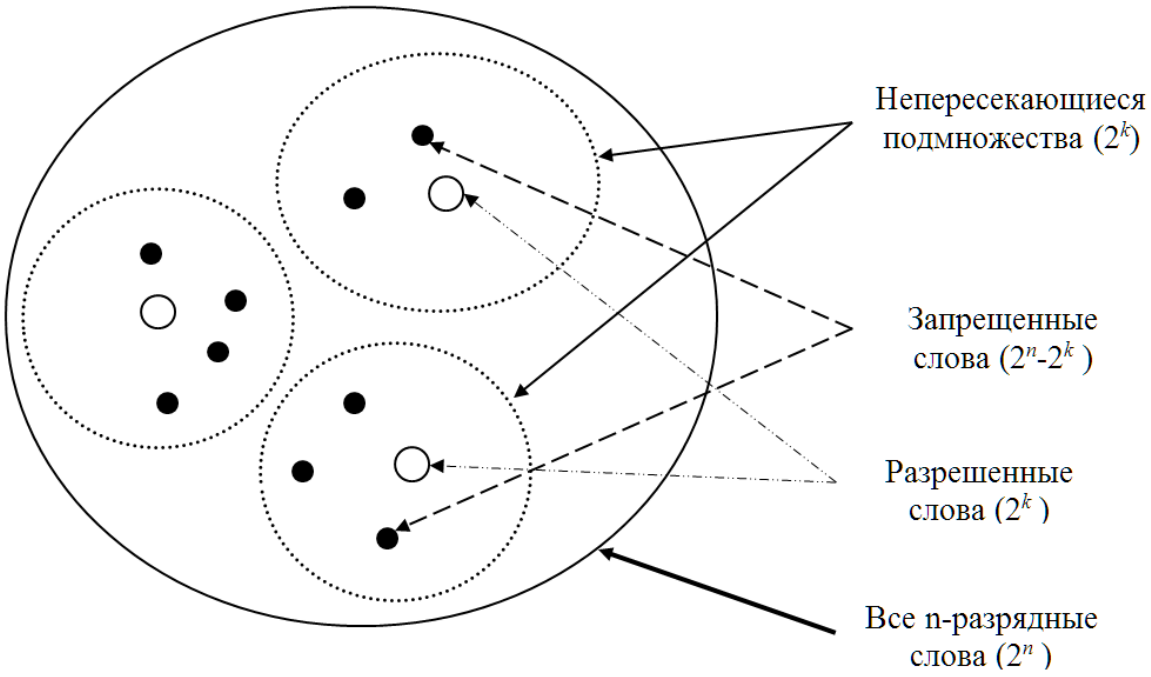


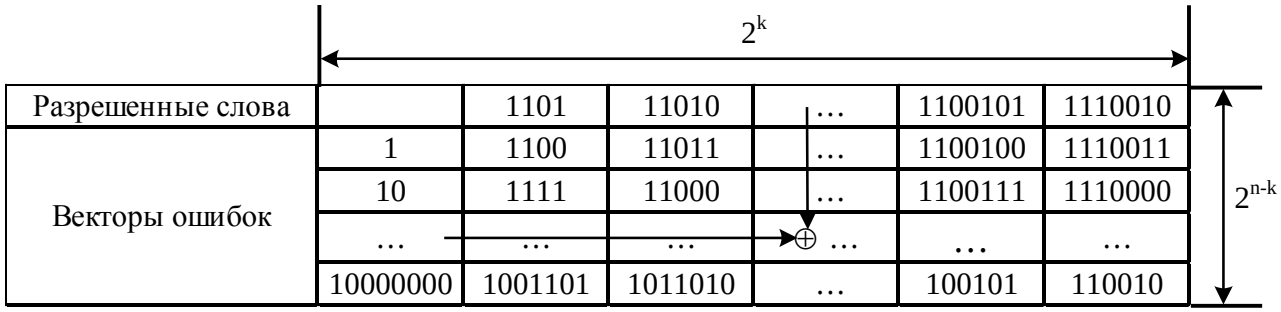

*Рис. 3.7.* Разбиение всех запрещенных слов
по принципу близости к разрешенным словам

Два способа разбиения запрещенных слов на непересекающиеся подмножества хорошо интерпретируются с помощью табл. 3.5, в ячейках которой размещены все $n$-разрядные запрещенные и разрешенные слова в количестве $2^n$. Следовательно, «площадь» этой таблицы, измеренная числом ее элементов, также равна $2^n$. В первой строке таблицы размещаются только разрешенные слова, поэтому ее ширина (число столбцов) равна $2^k$, а число строк – $2^{n-k}$. В каждой другой строке размещаются запрещенные слова, образованные из разрешенных слов и соответствующего подлежащего исправлению вектора

ошибок. Все строки, кроме первой, представляют ***непересекающиеся (по векторам ошибок) подмножества запрещенных кодовых слов, называемые классами смежности***; их число равно $2^{n-k}-1$. ***Класс смежности*** – это подмножество запрещенных слов (в количестве $2^k$), порожденных одним вектором ошибки. В каждом столбце, начиная со второго элемента, размещается непересекающееся подмножество запрещенных слов (в количестве $2^{n-k}-1$), порожденное одним разрешенным словом.

Таким образом, разбиение таблицы по столбцам демонстрирует разбиение всего множества запрещенных кодовых слов по ***принципу близости к разрешенным кодовым словам***, а разбиение по строкам – ***по принципу принадлежности к вектору ошибки*** (классу смежности).

*Таблица 3.5*




Из табл. 3.5 видно, что число классов смежности равно $2^{n-k}-1$; а общее число всех $n$-разрядных слов $2^n = 2^k \cdot 2^{n-k}$.

Коды, использующие или первое, или второе разбиение, способны обнаруживать и исправлять ошибки, но возникает вопрос: как выбирать корректирующие коды? На практике проблемы выбора нет. В тех случаях, когда разработчик знает векторы ошибок, которые могут нанести непоправимый ущерб системе, приходится строить корректирующий код на основе разбиения по принципу принадлежности к заданному вектору ошибки или классу смежности. Если такой опасности нет, то разработчик должен разрабатывать код, исправляющий более вероятные ошибки и обеспечивающий простую и надежную реализацию.

Для определения степени различия между кодовыми словами вводится специальная метрика – ***кодовое*** или ***хэммингово расстояние***. Расстояние между двумя кодовыми словами определяется числом разрядов с различным содержимым. Формально кодовое расстояние можно определить, подсчитав число единиц в кодовом слове, полученном поразрядным суммированием по модулю 2 сравниваемых кодовых слов.

**Пример:**
10011101
⊕
01011101
11000000

Кодовое расстояние $d = 1 + 1 = 2$.

***Минимальное хэммингово расстояние*** задает такое множество разрешенных слов, для любой пары в котором простое кодовое расстояние не меньше заданной величины.

Хэмминг установил ***границу существования оптимальных корректирующих кодов***. Пусть имеется возможность кодировать словами длиною *n*-разрядов. При этом все множество разных двоичных слов (включая запрещенные) составляет величину $2^n$. Требуется узнать, какое количество слов из этого множества можно использовать в качестве разрешенных, если необходимо исправлять все ошибки вплоть до *S*-кратных. Для того чтобы все ошибки названной кратности были исправляемы, в каждом из непересекающихся подмножеств данного разбиения число всех запрещенных слов должно быть не меньше числа ошибок, порождающих эти слова.

Число возможных ошибок в *n*-разрядном кодовом слове:

– однократных: $C_n^1 = n;$

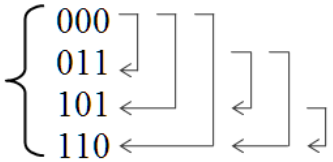
– двукратных: $C_n^2 = \frac{n!}{2!(n-2)!};$

– *S*-кратных: $C_n^S = \frac{n!}{S!(n-S)!};$

– общее число, включая S-кратные:

$$C_n^1 + C_n^2 + \cdots + C_n^S = \sum_{i=1}^{S} C_n^i .$$

**Пример.** Для множества трехразрядных слов
$\{000, 001, 011, 100, 101, 110, 111\}$ $d_{min} = 1$
выберем те слова, для которых $d_{min} = 2$.



Это разрешенные слова для кода, обнаруживающего однократные ошибки. Если дешифратор принимает кодовое слово с нечетным числом единиц, то это значит, что произошла однократная ошибка.

Выберем те слова, для которых $d_{min} = 3$.

$$\begin{cases} 000 \\ 111 \end{cases}$$

Это разрешенные слова для кода, исправляющего однократные ошибки. Если дешифратор принимает слова:

100 или 010 или 001 или 000, то это значит, что передано 000;

011 или 101 или 110 или 111, то это значит, что передано 111.

Поскольку все ошибки исправить невозможно, то разработчик строит код в расчете на те ошибки, которые исправлять необходимо. В дальнейшем ***вектором ошибки*** будем называть двоичное ***псевдослово***, содержащее «1» в тех разрядах, содержимое которых искажено помехами в данном помехоустойчивом кодовом слове.

**Пример**

10111011 – помехоустойчивое разрешенное кодовое слово

$\oplus$

00101000 – вектор ошибки

10010011 – запрещенное слово.

Множеству подлежащих исправлению векторов ошибок должно соответствовать множество запрещенных слов, порожденных из данного разрешенного слова этими ошибками. Таким образом, в каждом подмножестве столько запрещенных слов, сколько разных ошибок мы хотим исправить. Если требуется исправлять все ошибки кратности не более $S$, то количество запрещенных слов в каждом подмножестве, вместе с соответствующим разрешенным словом, равно

$$\sum_{i=1}^{S} C_n^i + 1 .$$

Отношение числа всех $n$-разрядных слов к числу слов в каждом подмножестве определяет предельное количество непересекающихся подмножеств, а значит – максимальное число разрешенных слов, т.е.

$$\frac{2^n}{\sum_{i=1}^{S} C_n^i + 1} \geq Q; \quad \text{если} \quad C_n^0 = 1, \quad \text{то} \quad \frac{2^n}{\sum_{i=0}^{S} C_n^i} \geq Q. \qquad (3.39)$$

Таким образом, определяется ***граница Хэмминга для существования оптимальных корректирующих кодов***.

***Пример.*** Пусть для построения кода, корректирующего все однократные ошибки, используются однобайтовые слова. Требуется определить максимальное число разрешенных слов, выбираемых из всего множества однобайтовых слов.

Число однократных ошибок $C_n^1 = n$, т.е. $C_n^1 = 8$.

Максимальное число разрешенных слов $Q \leq \frac{2^8}{1+8} = \frac{256}{9}$, т.е. $Q = 28$.

Кодирование и декодирование на основе определенного алгоритма (правила) кодирования можно заменить ***табличным кодированием и декодированием***. В этом случае необходимо заранее, до создания кодера и декодера, «вычислить по алгоритму» все множество разрешенных и соответствующих им запрещенных слов. При этом в памяти кодера размещается кодирующая таблица, которая представляет взаимнооднозначное соответствие входных безызбыточных $k$-разрядных и помехоустойчивых $n$-разрядных слов. В памяти декодера размещается декодирующая таблица со списком соответствующих друг другу запрещенных и разрешенных слов (в терминологии баз данных соотношение между запрещёнными и разрешёнными словами M:1). Алгоритм табличного кодирования (декодирования) сводится к поиску в первой колонке таблицы слова, тождественного входному. Идентификация входного слова завершается считыванием из соответствующей строки второй колонки выходного слова. Кодирование на основе кодирующих таблиц – самый распространенный способ кодирования. Рассмотрим для примера построение группового кода, корректирующего однократные ошибки.

### 3.3.3. Процедура построения группового кода

Рассмотрим процесс построения группового кода на примере.

***Исходные данные для построения кода*:**

– тип исправляемых ошибок – некоррелированные;

– кратность исправляемых ошибок $S = 1$;

– объем кода (число кодируемых сообщений) $Q = 15$.

Процедура состоит из четырех этапов.

*1. Расчёт числа информационных разрядов*

$$2^k - 1 \geq Q. \tag{3.40}$$

В левой части неравенства записано число двоичных $k$-разрядных слов без нулевой комбинации (слова состоящего только из нулей).

*2. Расчёт числа избыточных разрядов.*

Если необходимо исправлять все ошибки вплоть до S-кратных, то число информационных и избыточных разрядов должно определяться из соотношений

$$2^{n-k} - 1 \geq \sum_{i=1}^{S} C_n^i. \tag{3.41}$$

Здесь в левой части неравенства записано число комбинаций, составляемых из дополнительных разрядов, и равное этому числу количество классов смежности. Таким образом, данное неравенство утверждает, что число классов смежности должно быть не меньше числа исправляемых ошибок.

Для кода исправляющего только однократные ошибки эта же формула имеет вид

$$2^{n-k} \geq n; \tag{3.42}$$

$$k = 4; n = 7; n - k = 3,$$

где $k$ – число информационных разрядов; $n$ – число разрядов помехоустойчивого слова; $(n - k)$ – число избыточных разрядов.

Дополнительных разрядов в кодовом слове должно быть столько, чтобы породить нужное число запрещенных слов или классов смежности, а именно $2^{n-k} - 1$ (см. выше прямоугольную таблицу разбиений). Число классов смежности должно быть не меньше числа исправляемых ошибок, поэтому здесь $2^{n-k} - 1 \geq n$.

*3. Построение таблицы опознавателей ошибок* (табл. 3.6).

*Таблица 3.6*

| Векторы ошибок | | | | | | | Опознаватели (данный код) | Опознаватели (код Хэмминга) |
|---|---|---|---|---|---|---|---|---|
| a7 | a6 | a5 | a4 | a3 | a2 | a1 | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 101 | 001 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 110 | 010 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 001 | 011 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 100 | 100 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 010 | 101 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 011 | 110 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 111 | 111 |
| $n = 7$ | | | | | | | $n - k = 3$ | |

Каждой ошибке соответствует собственный опознаватель. Если для кодов, исправляющих однократные ошибки, в качестве опознавателей однократных ошибок используют соответствующие номерам искаженных разрядов двоичные числа, то такой код называется кодом Хэмминга (1-я и 3-я колонки табл. 3.6).

*4. Определение проверочных равенств*

Проверочное равенство представляет собой одно из преобразованных уравнений алгоритма кодирования. Пусть уравнение, с помощью которого формируется содержимое некоторого избыточного разряда, включает в себя в качестве известных величин содержимое определенных информационных разрядов, которые суммируются по модулю 2. В таком случае проверка дешифратором каждого из этих уравнений сводится к проверке на четность.

Если проверка дешифратором уравнения, определяющего данный избыточный разряд, показала нарушение принятого условия (четности), то соответствующий разряд опознавателя окажется равным 1. Просмотр во второй колонке таблицы содержимого данного разряда всех опознавателей ошибок позволяет выделить номера разрядов помехоустойчивого кодового слова, в которых возможно искажение. Для разрабатываемого кода разряды опознавателя определяются следующими группами разрядов помехоустойчивого кодового слова:

1-й (младший) – $a1 \cup a3 \cup a6 \cup a7$;
2-й – $a2 \cup a5 \cup a6 \cup a7$;
3-й – $a1 \cup a2 \cup a4 \cup a7$

(здесь знак $\cup$ соответствует логическому «или»).

При отсутствии однократных ошибок в слове дешифратор вычислит «нулевой опознаватель» (состоящий из одних нулей – 000). Поэтому можно записать «проверочные равенства» дешифратора в виде следующей системы уравнений:

$$\begin{cases} a1 \oplus a3 \oplus a6 \oplus a7 = 0 \\ a2 \oplus a5 \oplus a6 \oplus a7 = 0 \\ a1 \oplus a2 \oplus a4 \oplus a7 = 0 \end{cases}$$

– уравнения, формирующие 1-й, 2-й и 3-й разряды опознавателя.

При построении группового кода исправляющего, кроме однократных, ошибки большей кратности, необходимо учитывать следующее:

– опознаватели всех подлежащих исправлению векторов ошибок должны быть разными;

– число разрядов опознавателя ошибки должно быть по возможности меньшим, чтобы уменьшить избыточность кода;

– опознаватели векторов ошибок с единицами в нескольких разрядах устанавливаются как суммы по модулю 2 опознавателей однократных ошибок в этих разрядах, поэтому для определения правил построения кода и составления проверочных равенств достаточно правильно подобрать только опознаватели однократных ошибок в каждом из разрядов;

– выбирая в качестве опознавателей однократных ошибок в $i$-ом разряде двоичную комбинацию с числом разрядов меньшим $i$, необходимо убедиться в том, что для всех остальных подлежащих исправлению векторов ошибок, имеющих единицы в $i$-ом и более младших разрядах, получаются опознаватели, отличные от уже использованных.

В табл. 3.7 приведены опознаватели одиночных ошибок для кодов, исправляющих однократные и двукратные независимые ошибки вплоть до 15 разряда [7].

*Таблица 3.7*

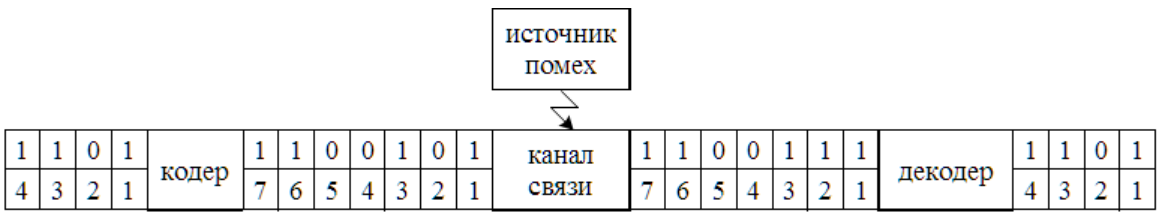
| N разряда | Опознаватель | N разряда | Опознаватель | N разряда | Опознаватель |
|---|---|---|---|---|---|
| 1 | 00000001 | 6 | 00010000 | 11 | 01101010 |
| 2 | 00000010 | 7 | 00100000 | 12 | 10000000 |
| 3 | 00000100 | 8 | 00110011 | 13 | 10010110 |
| 4 | 00001000 | 9 | 01000000 | 14 | 10110101 |
| 5 | 00001111 | 10 | 01010101 | 15 | 11011011 |

*5. Построение алгоритма кодирования*

Имея данную систему уравнений, на роль избыточных разрядов следует выбирать те, которые встречаются в проверочных равенствах по одному разу, т.е. $a3, a4, a5$. Выделение избыточных разрядов сопровождается определением информационных разрядов помехоустойчивого кодового слова. При этом для данного кода будут получены правила кодирования

$$\begin{cases} a3 = a1 \oplus a6 \oplus a7 \\ a4 = a1 \oplus a2 \oplus a7 \\ a5 = a2 \oplus a6 \oplus a7 \end{cases}$$

### 3.3.4. Процесс кодирования и декодирования



*Кодирование:*

| 1 | 1 | ? | ? | ? | 0 | 1 |
|---|---|---|---|---|---|---|
| *a7* | *a6* | *a5* | *a4* | *a3* | *a2* | *a1* |

*a3 = 1 ⊕ 1 ⊕ 1 = 1*

*a4 = 1 ⊕ 0 ⊕ 1 = 0*

*a5 = 0 ⊕ 1 ⊕ 1 = 0*

*Искажение помехой:*

1100101

⊕

0000010 – вектор ошибки

1100111 – запрещенное кодовое слово

*Декодирование*

– вычисление опознавателя ошибки:

1 ⊕ 1 ⊕ 1 ⊕ 1 = 0 (первый разряд опознавателя);

1 ⊕ 0 ⊕ 1 ⊕ 1 = 1 (второй разряд опознавателя);

1 ⊕ 1 ⊕ 0 ⊕ 1 = 1 (третий разряд опознавателя).

По таблице опознавателей отыскивается соответствующий опознавателю 110 вектор ошибки 0000010;

– исправление ошибки (восстановление кодового слова)

1100111 (запрещенное слово)

⊕

0000010 (вектор ошибки)

1100101 (разрешенное слово);

– выделение содержимого информационных разрядов ($a_1, a_2, a_6, a_7$).

На выходе декодера – 1101.

### 3.3.5. Контрольные вопросы и задания

1. Цель и суть любого кодирования.
2. Цели кодирования в технических системах.
3. Классификация помехоустойчивых кодов.
4. Цель и идея помехоустойчивого кодирования.
5. Понятие дополнительных, избыточных, проверочных, контрольных и информационных разрядов.
6. Типы (модели) помех.
7. Виды ошибок.
8. Что такое корректирующая способность кода?

9. Утверждение теоремы Шеннона для дискретного канала с помехами.

10. Варианты передачи кодового слова от шифратора к дешифратору и числа разных вариантов передач.

11. Что такое избыточная информация любого помехоустойчивого кода, для «кого» она таковой является и «кем» она формируется?

12. Какой по сути является избыточная информация у любого разделимого кода?

13. Какой по сути является избыточная информация у любого группового кода?

14. Какой по сути является избыточная информация у конкретного группового кода?

15. Чем представлена избыточная информация в кодовом слове любого помехоустойчивого кода?

16. Чем представлена избыточная информация в кодовом слове любого разделимого кода?

17. Чем представлена избыточная информация в кодовом слове любого группового кода?

18. Чем представлена избыточная информация в любом кодовом слове данного группового кода?

19. Чем представлена избыточная информация в конкретном кодовом слове данного группового кода?

20. Понятие хэммингового (кодового) расстояния (пояснить примером).

21. Понятие минимального хэммингового расстояния.

22. Минимальное хэммингово расстояние для кодов, которые только обнаруживают ошибки кратности $r$ (пояснить геометрически, примером разрешенных слов, взятых из множества трехразрядных кодовых слов).

23. Минимальное хэммингово расстояние для кодов, которые только исправляют ошибки кратности $S$ (пояснить геометрически, примером разрешенных слов, взятых из множества трехразрядных кодовых слов).

24. Минимальное хэммингово расстояние для кодов, которые только обнаруживают ошибки кратности $r$ и исправляют ошибки кратности $S$.

25. Построить множество $n$-разрядных кодовых слов с минимальным хэмминговым расстоянием, равным $d_{min}$.

26. Способы разбиения всего множества запрещенных кодовых слов на непересекающиеся подмножества; числа всех кодовых слов, запрещенных и разрешенных слов, непересекающихся подмножеств.

27. Понятие класса смежности. Как соотносятся между собой число классов смежности и количество исправляемых ошибок?

28. Что такое опознаватель или синдром ошибки?

29. Как определяется число информационных, избыточных разрядов и разрядов опознавателей ошибок?

30. Как определить абсолютную и относительную избыточность кода?

31. Понятие оптимального помехоустойчивого кода.

32. Граница Хэмминга для существования оптимального корректирующего кода (понятие и формула).

33. Ограничение при выборе способа разбиения множества запрещенных слов на непересекающиеся подмножества, используемого для построения кода.

34. Алгоритм работы кодера группового кода и его устройство.

35. Алгоритм работы дешифратора группового кода и его устройство.

36. Как организована работа кодера без выполнения вычислительных операций алгоритма кодирования?

37. Как организована работа дешифратора, реализующего процедуру «максимум правдоподобия»?

# 4. ИЗМЕРЕНИЕ КОЛИЧЕСТВА ИНФОРМАЦИИ

Важнейшим свойством информации является её измеримость, то есть возможность точного количественного подсчёта или учёта информации. В настоящее время существует несколько направлений или подходов, в которых поразному определяется количество информации различных типов. Рассмотрим основные из них.

## 4.1. Три подхода при измерении информации

При измерении информации можно различать три подхода: структурный, статистический и семантический [7].

При ***структурном подходе*** измерение информации осуществляется простым подсчетом ***числа сообщений*** (информационных элементов), составляющих информационный массив, или ***количеством преобразований*** сообщений (связей между информационными элементами), или ***числом комбинаций***, которые можно реализовать из этих сообщений (элементов). Этот подход применяется для оценки ёмкости информационных систем без учёта условий их применения. К таким системам относят средства формирования, передачи, обработки, преобразования и хранения сообщений.

При ***статистическом подходе*** количество информации ставится в зависимость от вероятности появления сообщений. Такой подход позволяет оценить информационные возможности систем с учетом конкретных статистических свойств сообщений и помех. Подчеркнём, что математическая теория связи К. Шеннона, называемая обычно теорией информации, является в настоящее время наиболее развитым направлением в общей теории информации. Однако этот подход описывает далеко не все информационные явления даже в технических системах. Сфера его применения ограничена статистическим (***случайным, неупорядоченным***) характером анализируемых информационных явлений, в то время как большая часть явлений, объектов природы и техники не только не случайны, но либо закономерно реализуются, либо существуют как факты в виде упорядоченных и организованных структур. Статистический подход получил широкое распространение для информационных оценок во многих областях человеческой деятельности, но это только лишний раз указывает на необходимость и возможность более общей теории для анализа информационных явлений.

***Семантический подход*** предназначен для учёта целесообразности, ценности, полезности, существенности и содержательности информации, для оценки эффективности логического опыта (например, более ценной можно считать ту

информацию, которая содержится в более короткой программе из возможных программ, ведущих к одному и тому же состоянию цели из заданного исходного состояния). Так как это направление в настоящее время недостаточно разработано, то здесь не рассматриваются связанные с ним вопросы.

## 4.2. Контрольные вопросы и задания

1. Какие подходы принято различать при измерении информации?

2. Определите область применения структурного подхода при измерении информации.

3. Каковы преимущества и недостатки подхода при измерении информации в котором учитываются статистические свойства сообщений и помех.

4. Для чего используется семантический подход?

## 4.3. Структурные меры информации

При использовании структурных мер информации учитывается только дискретное строение информационного массива, а именно – число сообщений (элементов) в массиве, количество связей между ними или число комбинаций из данных сообщений.

К структурным мерам информации относятся геометрическая, комбинаторная и аддитивная.

***Геометрическая мера*** информации употребляется в измерении «длины линии», «площади» или «объёма» данного информационного массива (комплекса) в единицах дискретных элементов (сообщений) этого массива. Этой мерой обычно измеряют ***информационную емкость*** массива, комплекса и т.п.

Пусть требуется измерить информационную ёмкость черно-белой фотографии, ширина и длина которой соответственно равны $y_{max}$ и $x_{max}$, а разность между минимальной и максимальной яркостью участков изображения равна $z_{max}$. Будем считать, что резкость изображения и чувствительность фотобумаги или разрешающая способность приборов, используемых для оценки яркости каждого элемента изображения, не позволяют различать участки размером менее $\Delta y$ и $\Delta x$ по смежным сторонам фотографии и перепады яркости менее $\Delta z$. Тогда, подсчитав возможное число отсчётов по каждой из сторон и число уровней яркости в виде

$$n_x = x_{max}/\Delta x\,;$$

$$n_y = y_{max}/\Delta y\,;$$

$$n_z = z_{max}/\Delta z.$$

определим информационную ёмкость фотографии как «объём» информационного массива в виде

$$N_{\text{г}_3} = n_x \cdot n_y \cdot n_z.$$

Аналогично этому можно определить информационную ёмкость процесса изменения напряжения как «***площадь***» массива данных, представляющих этот процесс, т.е. в виде $N_{\text{г}_2} = n_u \cdot n_t$, где $n_u$ – число различимых вольтметром значений напряжения, $n_t$ – число отсчётов напряжения во времени.

В общем случае, когда минимальное значение измеряемой величины не равно нулю, число отсчётов в диапазоне соответствующей величины вычисляется в виде

$$n_j = \frac{j_{max} - j_{min}}{\Delta j}. \tag{4.1}$$

При измерении геометрической мерой массива данных размерности, большей трёх используем понятие «***гиперобъема***», определяемого в виде

$$N_{\text{г}_j} = \prod_{j=1}^{m} n_j, \tag{4.2}$$

где $m$ – размерность массива.

Ёмкость запоминающего устройства ЦВМ часто измеряют числом ячеек памяти (элементов массива данных).

***Комбинаторная мера*** информации употребляется для оценки возможностей систем, в которых передача и хранение информации осуществляются при помощи различных комбинаций из набора (алфавита) сообщений мощностью $m$. Заметим, что сопоставление сообщениям множества большой мощности $N$ комбинаций из сообщений множества меньшей мощности $m < N$ является одним из способов кодирования, а сами комбинации (группы сообщений, символов) обычно называются ***кодовыми комбинациями*** или ***кодовыми словами***.

***Глубина m кодового слова*** или числа – это количество ***различных*** сообщений (элементов, состояний системы, знаков, символов), содержащихся в принятом алфавите. Глубина числа соответствует основанию позиционной системы счисления.

***Длина n кодового слова*** или числа – это количество повторений символов алфавита для образования данного кодового слова или числа. Длина числа соответствует принятой разрядности системы счисления.

Комбинирование возможно в комплексах с неодинаковыми сообщениями, переменными связями или разнообразными позициями.

Одинаковые сообщения могут стать различными, если учесть их положение, позицию, как это делается, например, в позиционных системах счисления.

В комбинаторике рассматривают различные виды соединений из элементов.

***Сочетания*** из $m$ элементов по $n$ различаются составом элементов, их число определяется в виде

$$N_{к_1} = C_m^n = \frac{m!}{n! \cdot (m-n)!}. \quad (4.3)$$

***Перестановки*** из $m$ элементов различаются только их порядком следования, их число определяется в виде

$$N_{к_2} = P_m = m! = n! \quad (4.4)$$

Возможны ***перестановки с неоднократными повторениями*** некоторых элементов, их число определяется в виде

$$N_{к_3} = P_m^{пов} = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_m!}. \quad (4.5)$$

Заметим, что отдельные перестановки с повторениями могут отображать массивы данных, представленные символами, знаковые последовательности, тексты.

***Размещения*** из $m$ элементов по $n$ элементов различаются составом элементов и их порядком; их число определяется в виде

$$N_{к_4} = A_m^n = \frac{m!}{(m-n)!}. \quad (4.6)$$

Из сопоставления $N_{к_2}$ и $N_{к_1}$ видно, что

$$A_m^n = P_m \cdot C_m^n = \frac{m! \cdot n!}{n! \cdot (m-n)!}. \quad (4.7)$$

Возможны размещения с повторениями одинаковых элементов; число таких размещений

$$N_{к_5} = A_m^{n_{пов}} = m^n. \quad (4.8)$$

Комбинаторная мера определяет количество информации числом возможных или реализуемых комбинаций из элементов, т.е. оценивает ***структурное разнообразие*** информационных комплексов. Так как число комбинаций значительно больше количества комбинируемых элементов, то комбинационная мера информации компактнее геометрической.

***Аддитивная мера*** информации или «***мера Хартли***» находит широкое применение. Для рассмотрения этой меры используем понятия глубины и длины кодового слова или числа.

При глубине $m$ и равной длине $n$ количество разных кодовых слов или чисел определяется в виде $N_{\text{к}_5} = m^n$, т.е. экспоненциально (нелинейно) зависит от длины числа $n$. Так при $m = 2$ и $n = 2, 3, 4, 5$ соответственно $N_{\text{к}_5} = 4, 8, 16, 32$. Вследствие нелинейной зависимости от $n$ число $N_{\text{к}_5}$ не является удобной мерой для оценки информационной ёмкости систем. Очевидно, было бы удобнее иметь такую меру информации, при которой увеличение длины кодового слова, числа сообщений или состояний, представляющих наблюдаемую систему, приводило бы к пропорциональному увеличению содержащейся в них информации. А при наличии нескольких кодовых слов (или систем) количество содержащейся в них информации определялось бы простой суммой количества информации в каждом из них. Поэтому в 1928 г. Р. Хартли ввёл ***аддитивную двоичную логарифмическую меру***, представляющую количество информации в двоичных единицах – битах (bits – binary digits). При определении количества информации этой мерой используется двоичный логарифм от числа разных сообщений, кодовых слов или состояний системы $N$, т.е.

$$I = \log_2 N. \tag{4.9}$$

Если $N = m^n$, то выражение (4.9) можно записать в виде

$$I = \log_2 m^n = n \cdot \log_2 m. \tag{4.10}$$

Обозначим

$$\log_2 m = K, \tag{4.11}$$

где $K$ – постоянная величина.

Подставляя формулу (4.11) в (4.10), получим, что количество информации в битах пропорционально длине кодового слова, так как

$$I = K \cdot n. \tag{4.12}$$

Из выражения (4.12) видно, что минимальное количество информации – один бит – содержит система (***элементарный источник***) с двумя (равновероятными) состояниями, т.е. когда $n = 1, m = 2$. Такой ***элементарный приёмник*** соответственно получает один бит информации при реализации одного из этих состояний или при выборе одного из двух равновозможных сообщений.

При наличии нескольких систем, числа состояний которых представлены значениями $N_1, N_2, \cdots, N_e, \cdots, N_к$, количество информации во всех системах определяется суммой вида

$$I(N_1, N_2, \cdots, N_к) = I(N_1) + I(N_2) + \cdots + I(N_к) = \sum_{e=1}^{k} \log_2 N_e, \qquad (4.13)$$

где к – количество наблюдаемых систем, кодовых слов и т.п.

Таким образом, мера Хартли удобна благодаря свойству ***аддитивности***, которое обеспечивает возможность ***сложения*** и ***пропорциональности*** количества информации к длине кодового слова. Заметим, что свойством аддитивности обладает и геометрическая мера информации, но она менее компактна, чем мера Хартли.

Как отмечалось в п. 2.1 (раздел 2) по определению Р. Хартли, ***информация*** – это особого рода ***логическая инструкция***, набор указаний или программа ***для выбора***, поиска (идентификации) определённого сообщения (состояния) из некоторого их множества. Заметим, что слово «идентичный» означает «тождественный», «одинаковый». Соответственно идентификация – это установление тождества, одинаковости, т.е. отождествление, приравнивание, уподобление. Далее будет показано, что «информация по Хартли» – это лишь один из двух видов информаций, так называемая «идентифицирующая информация», в отличие от «описательной». Однако здесь важно заметить, что ***Хартли определил информацию как определённого рода преобразование***, переводящее приёмник из одного состояния в другое, что в значительной степени совпадает с определением информации, данном в п. 1.10 учебного пособия [1].

Наконец, подчеркнём, что в отличие от широко распространённого представления, «информация по Хартли» не описывает объект, а предназначена только для его выбора, поиска, идентификации. Далее в п. 4.7 будет показано, что означает описание объекта.

Аддитивная мера Хартли используется для количественной оценки данных, полученных при измерениях, для оценки информационной емкости технических и других систем, а также различных документов, графической информации и в других случаях, где можно различать конечное число состояний. На-

пример, если вольтметр различает $n_u$ значений напряжения и во время измерений выполнено $n_t$ замеров, то полученный массив данных содержит

$$I = n_t \cdot \log_2 n_u.$$

Заметим, что идентификация имеет место при любых измерениях, которые, однако, не всегда осуществляются методом последовательного деления на два.

Ёмкость оперативного запоминающего устройства (ОЗУ) ЦВМ, состоящего из 3000 восьмиразрядных ячеек для хранения двоичных кодовых слов, составляет

$$I = 3000 \cdot 8 \text{ бит} = 24000 \text{ бит}.$$

В этом случае в каждой ячейке может храниться одно из $2^8 = 256$ возможных кодовых слов.

В чёрно-белой фотографии, содержащей $n_y \cdot n_x$ элементарных участков, в каждом из которых различимо $n_z$ уровней яркости, содержится количество информации

$$I = n_y \cdot n_x \cdot \log_2 n_z.$$

Более крупная единица информации – один ***байт*** (byte) равен восьми битам (1Б = 8 бит).

Для измерения количества информации в больших массивах данных используются еще более крупные единицы:

$2^{10} = 1024$ байта составляют один *килобайт* (Кбайт),

1024 килобайта образуют один *мегабайт* (Мбайт),

1024 мегабайта равняются одному *гигабайту* (Гбайт),

1024 гигабайта составляют один *терабайт* (Тбайт).

## 4.4. Контрольные вопросы и задания

1. Какие виды структурных мер информации следует различать?

2. Приведите примеры измерения ёмкости информационных массивов, представленных «объёмом», «площадью» и «длиной».

3. Какие существуют виды комбинаций из элементов?

4. На чём основано кодирование, использующее комбинации из элементов? Что представляет собой кодовое слово?

5. Как измеряется ёмкость информационного массива комбинаторной мерой?

6. Что понимается под глубиной и длиной кодового слова?

7. Определите меру Хартли и ее свойства. В чём проявляется свойство аддитивности?

8. Что такое бит?

9. Приведите определение информации, данное Хартли.

10. Запишите (закодируйте) все «информации Хартли» для идентификации каждого из шестнадцати различных сообщений.

11. Является ли «информация Хартли» описанием сообщения?

12. О каких программах поиска можно сказать, что их длина равняется количеству информации, учтённому аддитивной мерой?

13. Приведите примеры измерения количества информации аддитивной мерой Хартли.

14. Что такое байт?

## 4.5. Статистические меры информации

В теории Р. Хартли неявно допускалось, что выбор искомого состояния (сообщения) осуществляется из множества всех разных, т.е. равновозможных состояний. Однако на практике во многих ситуациях вероятности состояний источника неодинаковы и бывают известны приёмнику сообщений. Очевидно, что наличие априорной информации, имеющейся у приемника в виде вероятностей состояний (сообщений) источника, позволяет изменить условия выбора, поиска, идентификации определённого состояния, сообщения или объекта. В этом случае целесообразнее просматривать в первую очередь группы или подмножества, содержащие эталоны сообщений (состояний), появление которых более вероятно. Для этого все отличающиеся эталоны рассматриваемого множества должны быть предварительно ***ранжированы*** (упорядочены) в порядке убывания частот появления сообщений (состояний). Сама программа поиска может также осуществляться методом последовательного деления выбранного множества на два подмножества с последующим выбором одного из них. Как было показано в п. 2.1, 2.2 (раздел 2), при известных вероятностях $P_j$ появления сообщений на входе приёмника и дихотомическом поиске их эталонов среднее число информаций, необходимых для идентификации отдельного сообщения, определяется формулой (2.13) вида

$$H = I = -\sum_{j=1}^{m} P_j \cdot \log P_j.$$

Данную величину К. Шеннон в своей книге «Математическая теория связи», вышедшей в 1948 г., назвал аналогично термодинамическому понятию ***энтропией H*** источника или ***количеством информации I***. Если сообщения статистически зависимы, то при определении энтропии необходимо учитывать не только безусловные, но также и их условные вероятности. Так же как в термодинамике, где энтропия характеризует неопределённость теплового состояния вещества, энтропия в теории К. Шеннона служит мерой неопределённости состояния источника (сообщения), характеристикой информационной способности источника, мерой неупорядоченности сообщений, а мерой хаоса. В соответствии с определением А. Реньи эта мера представляет среднюю неожиданность сообщений (см. п. 2.2 и 2.3 в разделе 2).

При получении того или иного сообщения, т.е. в процессе его идентификации неопределённость сообщения или состояния источника либо уменьшается, либо снимается полностью. При этом количество информации, используемое для идентификации сообщения, равно уменьшению энтропии, т.е. энтропия и количество информации – величины взаимно обратные. Таким образом, энтропия также характеризует информационную способность источника сообщений. Необходимо помнить, что «количество информации по Шеннону» и энтропия являются усреднёнными характеристиками сообщений и состояний источника; для идентификации конкретного сообщения используется определённое число информаций.

## 4.6. Контрольные вопросы и задания

1. Возможна ли идентификация сообщений с неравными вероятностями их появления с помощью информации Хартли?

2. Какова зависимость между вероятностью отдельного сообщения, его неожиданностью, длиной программы и информацией для его идентификации? Какие из этих величин могут быть только целыми числами?

3. Напишите выражения для энтропии, количества информации и средней длины программы, необходимой для идентификации сообщения; дайте качественное толкование этих величин.

4. Возможна ли с помощью «информации Хартли» идентификация сообщений с такими вероятностям их появления, которые не допускают разбиений на равновероятные подмножества? Ответьте на этот вопрос в том числе путём вывода выражения для энтропии.

5. Почему энтропия и количество информации, идентифицирующей сообщение, величины взаимно обратные?

6. Какова разница между «количеством информации по Шеннону» и информацией для идентификации отдельного сообщения?

7. Опишите свойства энтропии.

## 4.7. Количество информации, определяемое числом преобразований сообщений в информационной цепи

### 4.7.1. Полезная, избыточная и паразитная информации в процессе управления

В п. 1.10 учебного пособия [1] определено понятие «информация» и её различные виды как преобразования, связывающие сообщения в поперечных множествах сообщений цепи управления. Число этих информаций-преобразований, содержащихся в информационных цепях, т.е. в цепях оригиналов, образов и промежуточных сообщений, можно подсчитывать [16]. Рассмотрим, что может дать такой подсчёт и как будут согласовываться полученные выражения для учёта числа информации с общепринятыми формулами (2.13) и (4.9) статистической теории информации. Для ответа на эти вопросы необходимо рассмотреть следующие понятия.

***Полезная информация*** – это информация, относящаяся к группе, состоящей из наименьшего количества информаций данной информационной цепи, необходимых для данного процесса управления.

***Избыточная информация*** – это такая информация, которая получена из других информаций данной информационной цепи.

***Паразитная информация*** – это информация, возникающая вне данного процесса управления.

К избыточной информации относится повторная и обратная информации, дополнительные операционные информации, помимо той, которая необходима для данного процесса управления, а также симуляционная псевдоинформация.

К паразитной информации относятся симуляционная дезинформация и симуляционная парадезинформация.

В дальнейшем при подсчете информации будем учитывать только полезную информацию.

Среди полезных информаций будем различать описательные информации и идентифицирующие информации.

### 4.7.2. Описательные информации и подсчет их числа в информационной цепи

***Описательная информация*** – это такая информация, которая относится к наименьшему возможному числу информаций, необходимых для описания некоторого ***определённого*** сообщения в информационной цепи.

***Исходная*** (реперная) ***информация*** – это описательная информация, необходимая для определения первого сообщения в информационной цепи.

***Исходное сообщение*** – это сообщение, которое следует преобразовать с помощью исходной информации для получения первого сообщения в информационной цепи.

Для описания последнего сообщения в информационной цепи необходимо знать информацию, которая преобразует предпоследнее сообщение в последнее, а для описания предпоследнего сообщения, в свою очередь, необходима информация, которая преобразует предыдущее сообщение, и т.д. – любое последующее сообщение описывается на основе предыдущего. Для определения же самого первого сообщения информационной цепи необходима исходная информация, которой должно быть подвергнуто некоторое исходное сообщение, не принадлежащее к данной информационной цепи.

Так как исходное сообщение не играет роли в процессе управления, то его можно выбрать произвольным образом, но в соответствии с ним должна быть выбрана исходная информация. Для определения, например, значений температуры в качестве исходной температуры берётся температура таяния льда, температура абсолютного нуля. В качестве исходной информации можно взять операционную информацию вида $t_0 + \Delta t_{01} = t_1$, где $\Delta t_{01} = (t_1 - t_0)$ – параметр операции. За исходную высоту местности принимается уровень моря; за исходный электрический потенциал берётся потенциал земли и т.п.

Наиболее удобно выбирать исходное сообщение, совпадающим с первым сообщением информационной цепи, так как исходная информация при этом будет тривиальной. К такому случаю относятся формулировки: «дана точка на плоскости, относительно которой…», «дано начальное состояние системы, относительного которого…».

Выясним, от чего зависит число описательных информаций.

**Теорема 4.1.** Число информаций $D$, описывающих одно сообщение в информационной цепи, состоящей из $n$ различных сообщений, равно числу этих сообщений.

Доказательство этой теоремы основано на том обстоятельстве, что для полного описания любого $j$-го сообщения информационной цепи, состоящей из $n$ сообщений, необходимо связать это сообщение со всеми $(n-1)$ остальными сообщениями этой цепи с помощью такого же числа описательных информаций и ещё одна – исходная информация – требуется для связи данного сообщения с исходным сообщением. Таким образом, $D_j = n$. Но те же утверждения справедливы для любого сообщения информационной цепи, поэтому

$$D_1 = D_2 = \cdots = D_j = \cdots = D_n = D = n. \qquad (4.14)$$

Из теоремы 4.1 следует, что при полном описании одного из сообщений информационной цепи одновременно оказываются описанными и все остальные сообщения данной цепи. Так бывает, например, когда имеется система из $n$ уравнений с $n$ переменными, $(n-1)$ из которых заданы, и требуется определить оставшиеся.

Здесь важно уяснить, что ***описать сообщение*** – это значит ***установить его связи со всеми другими сообщениями данной информационной цепи***, а также с исходным сообщением с помощью преобразований (описательных информаций). Невозможно, например, описать красный цвет не соотнося его с другими цветами (желтым, зелёным, синим и т.д.) спектра видимого излучения.

**Теорема 4.2.** Для описания одного сообщения информационной цепи, содержащей основную информацию и состоящей из произвольного числа $n$ сообщений достаточно двух описательных информаций, т.е.

$$D_1 = D_2 = \cdots = D_j = \cdots = D_n = D = 2. \qquad (4.15)$$

Данная теорема доказывается тем, что для определения любого сообщения из $n$ сообщений требуется $(n-1)$ – кратное применение одной и той же основной информации $I$ и применения одной исходной информации $I_{01}$ (см. п. 1.10 в учебном пособии [1].

Для дальнейших рассуждений необходимо определить следующие понятия.

***Полная описательная информация*** – это информация, которая представлена наименьшим ***множеством описательных информаций***, необходимых для описания определённого сообщения в информационной цепи, т.е. полная информация содержится в наименьшем множестве преобразований, связывающих данное сообщение с исходным и всеми другими сообщениями информационной цепи.

***Описательная информация множества сообщений*** информационной цепи – это информация, которая представлена наименьшим множеством полных описательных информаций, необходимых для описания каждого из сообщений информационной цепи.

***Редкость*** сообщения или описательной информации – это величина, обратная частоте (в пределе – вероятности) встречаемости сообщения данного класса или описательной информации данного типа среди множества всех рассматриваемых сообщений или информации. Следовательно, чем меньше вероятность появления некоторого объекта, тем реже он встречается и наоборот.

Редкость будем обозначать в виде $n/n_j$ или $1/P_j$, где $n_j$ – число одинаковых обобщений, составляющих $j$-й класс, $n$ – общее число сообщений; $P_j/n$ и $P_j$ – соответственно частота (относительное число) и вероятность одинаковых сообщений $j$-го класса.

**Теорема 4.3.** Если в информационной цепи, состоящей из $n$ сообщений, имеется $m$ классов, состоящих соответственно из $n_a, n_b, \dots, n_m$ одинаковых сообщений, то среднее число описательных информаций можно определить следующим выражением:

$$D = \left(\frac{n}{n_a}\right)^{\frac{n_a}{n}} \cdot \left(\frac{n}{n_b}\right)^{\frac{n_b}{n}} \cdot \dots \cdot \left(\frac{n}{n_m}\right)^{\frac{n_m}{n}}. \tag{4.16}$$

Заметим, что возможными реализациями такой цепи могут быть обычные тексты, знаковые последовательности, массивы данных измерений (в которых некоторые значения повторяются).

Докажем теорему. Дано, что в информационной цепи имеется:

$n_a$ сообщений $a_1 = a_2 = a_3 = \dots = a$,
$n_b$ сообщений $b_1 = b_2 = b_3 = \dots = b$,
……..
$n_m$ сообщений $m_1 = m_2 = m_3 = \dots = m$,
причём $n_a + n_b + \dots + n_m = n$.

В соответствии с теоремой 4.1 число информаций, описывающих сообщение $a_1$, будет $D_{a_1} = n$. Точно также для числа информаций, описывающих остальные сообщения этого же класса, имеем $D_{a_2} = n, D_{a_3} = n$ и т.д. Вследствие одинаковости сообщений $a_1 = a_2 = a_3 = \dots = a$ *полные информации*, описывающие каждое из этих сообщений, также одинаковы, поэтому для полной информации, описывающей одно из этих сообщений, информации, описывающие остальные сообщения этого класса, являются избыточными и должны быть исключены. При этом среднее число описательных информаций, приходящееся на каждое из этих сообщений, будет меньше общего числа сообщений данной информационной цепи во столько раз, сколько имеется одинаковых сообщений в данном классе, т.е. $D_a = n/n_a$. Также определяются средние числа описательных информаций, приходящихся на каждый из остальных классов, т.е.

$$\mathrm{D_a} = \frac{n}{n_a}, D_b = \frac{n}{n_b}, \dots, D_m = \frac{n}{n_m}. \tag{4.17}$$

Среднее число описательных информаций одного сообщения информационной цепи определим как среднее геометрическое средних чисел информаций, описывающих сообщения для всех классов:

$$D = \sqrt[n]{D_a^{n_a} \cdot D_b^{n_b} \cdot D_m^{n_m}}, \tag{4.18}$$

или с учетом (4.17), получим

$$D = \left(\frac{n}{n_a}\right)^{\frac{n_a}{n}} \cdot \left(\frac{n}{n_b}\right)^{\frac{n_b}{n}} \cdot \left(\frac{n}{n_m}\right)^{\frac{n_m}{n}}, \tag{4.19}$$

или то же в компактном виде

$$D = \prod_{j=1}^{m} \left(\frac{n}{n_j}\right)^{\frac{n_j}{n}}, \tag{4.20}$$

где $j$ – номер класса, $m$ – число классов.

Преобразуем уравнение (4.20) к виду

$$D = n \Big/ \sqrt[n]{\prod_{j=1}^{m} n_j^{n_j}}. \tag{4.21}$$

Обозначив

$$\sqrt[n]{\prod_{j=1}^{m} n_j^{n_j}} = n_q, \tag{4.22}$$

где $n_q$ – среднее (геометрическое) число сообщений одного класса, и подставляя эту величину в знаменатель выражения (4.21), получим

$$D = {}^{n}\!/_{n_q}. \tag{4.23}$$

В частном случае, когда все сообщения информационной цепи различны, т.е. когда каждый класс представлен всего одним сообщением и $m = n$, из формулы (4.20) имеем $D = n$, что совпадает с утверждением теоремы 4.1.

Заметим, что по определению величина $D_a = \frac{n}{n_a}$ оценивает редкость встречаемости одинаковых сообщений класса «$a$» и среди всего множества $n$ сообщений, составляющих информационную цепь. Если рассмотреть любую из полных описательных информаций сообщений этого класса, то станет видно, что эта же величина, является редкостью тривиальных информаций среди всего множества $n$ информаций, описывающих отдельное сообщение класса «$a$», а также, что она представляет, кроме того, редкость одинаковых полных информаций, описывающих сообщения класса «$a$», среди множества $n$ полных информаций, описывающих все сообщения информационной цепи. Таким образом, величина $D_j = {}^{n}/_{n_j}$ представляет одновременно ***редкости встречаемости следующих объектов***: одинаковых сообщений $j$-го класса, тривиальных описательных информаций, входящих в состав полной описательной информации отдельного сообщения $j$-го класса, одинаковых полных описательных информаций сообщений $j$-го класса, образующих вместе с другими описательную информацию множества всех сообщений информационной цепи.

С учётом данных замечаний среднее число описательных информаций $D$, определённое в виде формул (4.18) – (4.21) и (4.23), позволяет также оценить ***средние редкости*** одинаковых сообщений или тривиальных описательных информаций, или одинаковых полных информаций, описывающих сообщения одного класса.

### 4.7.3. Идентифицирующие информации и подсчёт их числа в информационной цепи

***Идентифицирующая информация*** или преобразование для установления тождества – это такая информация, которая относится к наименьшему возможному числу информаций, необходимых для идентификации ***отдельного*** сообщения информационной цепи.

Такие информации следует отличать от описательных информаций, которые предназначены для описания определённого сообщения (см. п. 4.7.2).

В п. 2.1 (раздел 2) отмечалось, что ***идентификация как процесс установления тождества осуществляется путём выбора, отыскания*** некоторого заданного эталона (сообщения, предмета, объекта, явления) из множества других сообщений. При этом считается, что у системы, осуществляющей такой поиск, имеется эталон искомого сообщения, т.е. имеется полное его описание, позволяющее отличить данное сообщение от других сообщений. Найти или выбрать сообщение – это значит указать, ***где*** оно находится или ***когда*** появилось, или ***где*** и ***когда*** оно имеет место, или каков его ***номер*** в ряду других сообщений. Таким образом, ***критериями выделения сообщения могут быть пространство, время, пространство-время, а также порядок следования***. На-

личие эталона искомого сообщения обеспечивает возможность проверки выполнения условий, требуемых одним из установленных критериев.

***При выборе, выделении (идентификации) сообщений делается только различие между тривиальной и нетривиальной информацией*** независимо от того, с помощью каких преобразований получена нетривиальная информация.

Поясним это утверждение примером. Пусть информационная цепь состоит из четырёх сообщений: $< x_1, x_2, x_3, x_4 >$, из которых под определённым углом зрения необходимо выделять сообщение $x_2$.

В соответствии с теоремой 4.1 для описания произвольного сообщения этой информационной цепи требуются четыре описательные информации, которые связывают данные сообщения с исходным сообщением $x_0$:

$$I_{01}x_0 = x_1; \qquad I_{02}x_0 = x_2; \qquad I_{03}x_0 = x_3; \qquad I_{04}x_0 = x_4.$$

При идентификации в качестве исходного сообщения берётся именно искомое сообщение, точнее говоря, его эталон, т.е. в нашем примере $x_0 = x_2$. При этом одна из рассмотренных информаций будет тривиальной, а остальные нетривиальными:

$$I_{21}x_2 = x_1; \qquad I^{\circ}x_2 = x_2; \qquad I_{23}x_2 = x_3; \qquad I_{24}x_2 = x_4.$$

Таким образом, видно, что идентификация некоторого сообщения основана на утверждении, что среди описательных информаций есть тривиальная. Для этого необходимо найти ответы на следующие вопросы:

$$I_{21} = I^{\circ}? \qquad I_{22} = I^{\circ}? \qquad I_{23} = I^{\circ}? \qquad I_{24} = I^{\circ}?$$

В результате ответов, которые имеют вид «да» или «нет», будут получены следующие четыре описательные информации:

$$I_{21} \neq I^{\circ}; \qquad I_{22} = I^{\circ}; \qquad I_{23} \neq I^{\circ}; \qquad I_{24} \neq I$$

или

$$x_2 \neq x_1; \qquad x_2 \neq x_2; \qquad x_2 \neq x_3; \qquad x_2 \neq x_4.$$

Из примера видно, что число возможных описательных информаций выделенного сообщения зависит от порядка задания вопроса, так как тривиальная информация в наилучшем случае может быть найдена после первого же вопроса, в наихудшем – после третьего, а при наличии $n$ сообщений – после $(n-1)$-го во-

проса. В последнем случае $n$-й вопрос не нужен, так как очевидно, что это будет тривиальная информация.

Рассмотрим условия, при которых число описательных информаций, необходимых для идентификации сообщений, не зависит от порядка следования вопросов. Это и будут идентифицирующие информации. Число информаций, идентифицирующих одно сообщение, будем обозначать символом $H$.

**Теорема 4.4.** Число описательных информаций, необходимых для идентификации сообщения, однозначно определено только в информационной цепи из двух сообщений; при этом $H_2 = 1$.

Данное утверждение доказывается тем, что в такой информационной цепи мы избавлены от случайности в порядке выбора сообщений, так как любой ответ на первый же вопрос – либо о наличии тривиальной информации, либо о выявлении нетривиальной описательной информации (в этом случае очевидно, что тривиальной информацией является вторая) – позволяет однозначно выделить искомое сообщение. Таким образом, самая первая выявленная описательная информация является информацией, идентифицирующей выделенное сообщение, т.е. $H_2 = 1$. Число же описательных информаций здесь будет $D_2 = 2$. Одна из них – тривиальная информация, утверждающая, что описанное сообщение является выделенным сообщением, а вторая – нетривиальная – утверждает, что остальные сообщения не являются выделенными.

Однозначностью выбора объясняется рассмотренный в п. 2.1 (раздел 2) известный способ идентификации (дихотомическая процедура), основанный на последовательном делении выбранного множества на две равные части с выбором одной из них и т.д. на каждом шаге вплоть до идентификации пары сообщений и выделением в ней искомого сообщения.

**Теорема 4.5.** Число информаций, идентифицирующих одно сообщение в информационной цепи, содержащей $n$ различных сообщений, может быть определено как двоичный логарифм этого числа сообщений.

Докажем теорему. Дана информационная цепь из $n$ сообщений, являющаяся половиной информационной цепи из $2n$ сообщений. В соответствии с теоремой 4.1 числа информаций, описывающих одно сообщение в этих информационных цепях соответственно будут:

$$D_n = n;\ \ D_{2n} = 2n,$$

вследствие чего

$$D_{2n} = 2 \cdot D_n. \qquad (4.24)$$

Пусть число информаций, идентифицирующих одно сообщение в информационной цепи из $n$ сообщений, равно $H_n$. В соответствии с теоремой 4.4 для идентификации самой этой цепи, как одной из двух составляющих информационную цепь из $2n$ сообщений, необходима одна идентифицирующая информация. Поэтому общее число информаций, идентифицирующих данное сообщение в информационной цепи из $2n$ сообщений, будет

$$H_{2n} = H_n + 1. \tag{4.25}$$

Связь между $H$ и $D$ можно найти, прологарифмировав формулу (4.24) по основанию 2:

$$\log_2 D_{2n} = \log_2 2D_n = \log_2 D + \log_2 2,$$

или

$$\log_2 D_{2n} = \log_2 D_n + 1. \tag{4.26}$$

Из сравнения уравнений (4.25) и (4.26) следует, что на основании одной из них может быть найдена другая, если зависимость $H = f(D)$ имеет вид

$$H = \log_2 D. \tag{4.27}$$

Для $D_n = n$ уравнение (4.27) приобретает вид

$$H_n = \log_2 n, \tag{4.28}$$

т.е. идентично формуле (4.9) для определения количества информации аддитивной мерой Хартли.

Заметим, что формулой Хартли (4.9) мы пользовались безотносительно к числу описательных информаций $D$, однако, она справедлива лишь для частного случая, когда $D = n$; в более общих ситуациях, когда $D \leq n$, справедливо уравнение

$$H = \log_2 D.$$

Так как для идентификации одного из двух подмножеств множества сообщений требуется одна идентифицирующая информация, то число идентифицирующих информаций $H_n$ равно числу последовательных разбиений множества

на два равных подмножества, необходимых для идентификации сообщения из целой информационной цепи. Это совпадает с тем, что в п. 2.1 и 2.2 (раздел 2) называется длиной программы или количеством информации для выбора сообщения.

**Теорема 4.6.** Если в информационной цепи из $n$ сообщений имеется $m$ классов, каждый из которых состоит соответственно из $n_a, n_b, \dots, n_m$ одинаковых сообщений, то среднее число идентифицирующих информаций можно выразить соотношением:

$$H = \frac{n_a}{n}\log_2\frac{n}{n_a} + \frac{n_b}{n}\log_2\frac{n}{n_b} + \cdots + \frac{n_m}{n}\log_2\frac{n}{n_m}.$$

Заметим, что возможными реализациями такой цепи могут быть обычные тексты, знаковые последовательности, массивы данных измерений (в которых некоторые значения повторяются).

Докажем теорему. Дано, что в информационной цепи имеется:

$n_a$ сообщений $a_1 = a_2 = a_3 = \cdots = a$;
$n_b$ сообщений $b_1 = b_2 = b_3 = \cdots = b$;
……
$n_m$ сообщений $m_1 = m_2 = m_3 = \cdots = m$,
причём $n_a + n_b + \cdots + n_m = n$.

В соответствии с теоремой 4.5 число информаций, идентифицирующих сообщение $a_1$ в информационной цепи из $n$ сообщений, составляет $H_n = \log_2 n$, а число информаций, «идентифицирующих» это же сообщение внутри своего класса, состоящего из $n_a$ сообщений, составляет $H_{n_a} = \log_2 n_a$. Но выделить сообщение $a_1$ внутри своего класса невозможно, так как сообщения $a_1, a_2, a_3, \dots$ неразличимы из-за того, что они одинаковы. Таким образом, ***идентификация позволяет*** только установить принадлежность рассматриваемого сообщения к некоторому классу, но не позволяет указать, какое именно из этих сообщений идентифицируется. Поэтому для идентификации любого сообщения, принадлежащего классу $a$, потребуется меньше идентифицирующих информаций на величину $H_{n_a}$, т.е.

$$H_a = H_n - H_{n_a} = \log_2 n - \log_2 n_a. \qquad (4.29)$$

Аналогично находится число идентифицирующих информаций, приходящихся на каждое сообщение остальных классов:

$$H_b = \log_2 n - \log_2 n_b,\ \ldots,\ H_m = \log_2 n - \log_2 n_m, \tag{4.30}$$

т.е.

$$H_a = \log_2 \frac{n}{n_a},\ H_b = \log_2 \frac{n}{n_b}, \ldots,\ H_m = \log_2 \frac{n}{n_m}. \tag{4.31}$$

Среднее количество информаций, идентифицирующих отдельное сообщение, можно определить средним значением чисел идентифицирующих информаций сообщений всех классов:

$$H = \frac{1}{n}(n_a H_a + n_b H_b + \cdots + n_m H_m), \tag{4.32}$$

откуда с учётом (4.31) получаем

$$H = \frac{n_a}{n}\log_2 \frac{n}{n_a} + \frac{n_b}{n}\log_2 \frac{n}{n_b} + \cdots + \frac{n_m}{n}\log_2 \frac{n}{n_m} \tag{4.33}$$

или то же в компактном виде

$$H = \sum_{j=1}^{m} \frac{n_j}{n}\log_2 \frac{n}{n_j}, \tag{4.34}$$

где $j$ – номер класса сообщения, $m$ – число классов.

В частном случае, когда все сообщения, информационной цепи различны, т.е. когда каждый класс представлен всего одним сообщением и $m = n$, из (4.34) имеем $H = \log_2 n$, что совпадает с утверждением теоремы 4.5 и формулой (4.9) для определения количества информации мерой Хартли.

**Теорема 4.7.** Среднее число идентифицирующих информаций равно двоичному логарифму среднего числа описательных информаций.

Данное утверждение доказывается сравнением формул (4.20) и (4.34), из которого следует, что

$$H = \log_2 D \tag{4.35}$$

Для усвоения ***фундаментального характера полученной формулы*** полезно ещё раз обратить внимание на замечание к формулам (4.27) и (4.28).

Из уравнения (4.35) видно, что число идентифицирующих информаций $H$ меньше числа описательных информаций $D$. Это объясняется тем, что описание одного сообщения из множества сообщений является также описанием всех сообщений данного множества, в то время как идентификация одного сообщения лишь означает утверждение, что ни одно из остальных сообщений не является искомым сообщением, т.е. при идентификации утверждается только то, что кроме одной тривиальной информации, все остальные нетривиальны без указания какими именно являются нетривиальные информации.

При большом числе сообщений в информационной цепи, когда $n \to \infty$, частоты появления сообщений разных классов можно заменить вероятностями этих сообщений [16]:

$$\frac{n_a}{n} = P_a, \quad \frac{n_b}{n} = P_b, \ldots, \quad \frac{n_j}{n} = P_j, \ldots, \quad \frac{n_m}{n} = P_m. \tag{4.36}$$

После подстановки (4.36) в (4.20) выражение для среднего количества описательных информаций примет вид

$$D = \prod_{j=1}^{m} \left(\frac{1}{P_j}\right)^{P_j} \qquad \text{или} \qquad \mathrm{D} = 1 \Big/ \prod_{j=1}^{m} \left(P_j\right)^{P_j}, \tag{4.37}$$

где $j$ – номер класса сообщения, $m$ – число классов.

Выражение (4.34) для среднего количества идентифицирующих информаций после аналогичной подстановки преобразуется к виду

$$H = \sum_{j=1}^{m} P_j \log_2 \frac{1}{P_j} \qquad \text{или} \qquad H = -\sum_{j=1}^{m} P_j \log_2 P_j.$$

Данное выражение, кроме того, может быть получено логарифмированием по основанию 2 зависимости (4.37) и является идентичным формуле (2.13), предложенной К. Шенноном для измерения количества информации.

### 4.7.4. Использование мер описательной и идентифицирующей информаций

В п. 2.2 (раздел 2) отмечалось, что величина $H$, определённая формулой (2.13), также представляет энтропию источника или неопределенность его состояния, среднюю неожиданность или неопределённость сообщения. Рассмотрим подробнее сущность меры, которая оценивает среднее число описа-

тельных информаций $D$, определяемое в виде (4.20) и (4.37). Как видно из формул (4.17)–(4.20) при выводе выражения для $D$ исключались одинаковые, а значит – избыточные информации, описывающие сообщения одного и того же класса. Поэтому по аналогии с тем как средняя неожиданность сообщений представляет степень неопределённости или энтропию, будем полагать, что среднее число описательных информаций $D$ (или средняя редкость одинаковых сообщений информационной цепи или их одинаковых описательных информаций) представляет ***степень разнообразия неоднородности***, неповторяемости, нерегулярности сообщений или их описательных информаций.

В п. 4.3 отмечались недостатки статистической меры информации (Шеннона), которая является обобщением аддитивной меры (Хартли) на случай неравновероятных сообщений. Как показано в п. 2.1 и 2.2 (раздел 2), применение этих мер ограничено измерением числа информаций (типа «команд по выбору одного из двух равных множеств»), которые предназначены для идентификации случайного сообщения.

На практике возникает необходимость идентифицировать сообщения, имеющие различную физическую природу, например:

- выигрышный лотерейный билет на получение определённой суммы денег – среди всех типов выигрышных и проигрышных билетов;
- любимая мелодия – среди множества транслируемых по радио музыкальных произведений;
- символ буквы «$a$» – среди множества символов в тексте;
- значение измеряемой величины – среди множества всех возможных значений этой величины;
- кодовое слово «1 0 1 1 1» – среди некоторого множества двоичных комбинаций; например; хранящихся в запоминающем устройства ЦВМ;
- «лёгкие» или «трудные» билеты – среди множества экзаменационных билетов, отличающихся по степени сложности.

Сообщения разных классов в подобных множествах в общем случае представлены разными вероятностями, которые могут быть получены из опыта или других источников. В процессе идентификации такие сообщения могут восприниматься приёмником как случайные – независимые или в лучшем случае, статистически зависимые, т.е. как сообщения, которые *не составляют закономерно упорядоченное множество*.

В п. 1.1, 1.3, 1.10 учебного пособия [1] отмечается, что информации и явления информирования имеет смысл рассматривать только в процессах управления. В п. 4.7.2 установлено также, что в каждом процессе управления имеются сообщения и определяющие их описательные информации, но заметим, что не всегда требуется идентифицировать сообщения, а значит, подсчитывать число идентифицирующих информаций. В большинстве случаев в процессах управле-

ния имеют место множества неслучайных, упорядоченных сообщений, связанных между собой совершенно определённым образом. Подобные абстрактные объекты, называемые в математике упорядоченными множествами или кортежами, не фиксируют принципиально разной роли информационных и кодовых цепей в процессах информирования и управления. Упорядоченными множествами, называемыми здесь информационными цепями, являются, например: текст, состоящий из слов или букв (сообщений), график функции, осциллограмма, последовательность звуков, представляющая речь человека, музыку и т.д.; массив чисел, отображающий некоторый процесс и представленный в ЗУ ЦВМ последовательностью двоичных кодовых слов, а на бумаге печатающего устройства или экране дисплея – последовательностью символов, программа для ЦВМ, составленная из символов алгоритмического языка, сообщения или показания одного или нескольких датчиков или приборов, отображающие один и тот же или несколько взаимосвязанных процессов, математическое выражение, представляющее зависимости между его переменными (сообщениями).

Очевидно, что в этих случаях зачастую возникает необходимость в определении числа описательных информаций, которые связывают сообщения в некоторую конкретную информационную цепь. Это позволяет оценить возможности информационных систем, которые в процессе управления на разных этапах производят генерацию информационных цепей, их передачу, хранение и преобразование.

Ещё раз подчеркнём, что по известному числу $D$ описательных информаций всегда можно определить число $H$ идентифицирующих информаций, но не всегда имеет смысл это делать. Так, в исторических событиях, засвидетельствованных в книге, очевидно, что имеются описательные информации, число которых бывает необходимо подсчитывать. Однако осуществлять идентификацию какого-либо исторического события не имеет смысла, так как они не могут появляться с некоторыми вероятностями, а существуют как факты.

## 4.8. Контрольные вопросы и задания

1. Определите полезную, избыточную и паразитную информации и приведите их примеры.

2. Что такое описательная информация? Каково назначение исходного сообщения и исходной информации и какими они могут быть?

3. Определите понятие «редкость».

4. Определите полную описательную информацию множества сообщений информационной цепи; приведите примеры таких информаций.

5. Приведите соотношения для определения числа описательных информаций следующих информационных цепей: 1) состоящей только из разных сооб-

щений; 2) содержащей основную информацию; 3) содержащей группы одинаковых сообщений. Докажите справедливость приведённых соотношений.

6. Что такое идентификация? Что может быть критерием выделения сообщения? На каком утверждении основана идентификация сообщения? Каково число вопросов или выборов возможно при идентификации сообщения?

7. Что такое идентифицирующая информация и каково её отличие от описательной информации?

8. В какой информационной цепи возможна однозначная идентификация?

9. Приведите соотношения для определения числа идентифицирующих информаций в информационной цепи, состоящей только из разных сообщений, а также – содержащей группы одинаковых сообщений; докажите справедливость приведенных соотношений.

10. Какова зависимость между числом описательных и идентифицирующих информаций?

11. Почему идентифицирующих информаций меньше, чем описательных в данной информационной цепи?

12. Какова связь между соотношением для определения числа идентифицирующих информации и формулами Р. Хартли и К. Шеннона?

13. Какие количественные характеристики информационной цепи определяет среднее число описательных информаций?

14. Какова связь между энтропией и разнообразием описательных информации в различных информационных цепях?

15. Укажите области явлений, где целесообразен подсчёт описательных и идентифицирующих информации?

## 4.9. Связь числовых характеристик строя с формулами для подсчёта описательных и идентифицирующих информаций

В разделе 1 представлен ряд числовых характеристик, которые позволяют компактно или более детально описывать расположение компонентов в отдельной цепи. Среди них выделим и проанализируем следующие:

- средний геометрический интервал между одинаковыми компонентами данной цепи, представленный формулой (1.5);
- среднюю удалённость одинаковых компонентов в цепи, представленную формулами (1.17) и (1.18).

Покажем, что эти характеристики определяют числа информаций, необходимых для описания и идентификации интервалов (расположения компонентов) [2].

Подставим выражение для объема цепи (1.3) в формулу (1.5), а затем вместо $V_j$ – формулу (1.1). В результате получим два выражения вида

$$\Delta_g = \sqrt[n]{\prod_{j=1}^{m} V_j} = \sqrt[n]{\prod_{j=1}^{m}\prod_{i=1}^{n_j} \Delta_{ji}}. \qquad (4.38),(4.39)$$

Обозначим

$$\Delta_{gj} = \sqrt[n_j]{\prod_{i=1}^{n_j} \Delta_{ji}},$$

с учетом введенного обозначения $V_j = \Delta_{gj}^{n_j}$, подставляя это выражение в (4.39) получим

$$\Delta_g = \sqrt[n]{\prod_{j=1}^{m} \Delta_{gj}^{n_j}} = \prod_{j=1}^{m} \Delta_{gj}^{\frac{n_j}{n}}. \qquad (4.40)$$

Логарифмируя последнее выражение, получим формулу для средней удалённости в виде

$$g = \sum_{j=1}^{m} \log_2 \Delta_{gj}^{\frac{n_j}{n}} = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \log_2 \Delta_{gj}. \qquad (4.41)$$

Рассмотрим связь основных характеристик строя с формулами Мазура и Шеннона.

Назовём ***«регулярной» знаковой цепью*** такую, в которой все интервалы каждой однородной цепи будут: $\Delta_{ij} = \Delta_{aj} = n/n_j = const$ $(j = 1,2,\cdots,m)$. Выражения для числовых характеристик строя регулярной цепи (4.40) и (4.41) соответственно принимают вид:

$$\Delta_g = \Delta_{g_{max}} = \prod_{j=1}^{m} \Delta_{gj}^{\frac{n_j}{n}} = \prod_{j=1}^{m} \Delta_{aj}^{\frac{n_j}{n}}. \qquad (4.42)$$

Так как $\Delta_{aj} = n/n_j$, то выражение (4.42) принимает вид

$$\Delta_{g_{max}} = \prod_{j=1}^{m} \left(\frac{n}{n_j}\right)^{\frac{n_j}{n}}. \tag{4.43}$$

Из сравнения формул для числа описательных информаций (4.20) и максимального среднего геометрического интервала (4.43) видно, что они совпадают, т.е.

$$D = \Delta_{g_{max}}. \tag{4.44}$$

Для регулярной знаковой цепи выражение (4.41) для средней удалённости принимает вид

$$g = g_{max} = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \log_2 \Delta_{gj} = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \log_2 \Delta_{aj}. \tag{4.45}$$

Подставляя $\Delta_{aj} = \frac{n}{n_j}$, получим формулу вида

$$g_{max} = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \log_2 \frac{n}{n_j}. \tag{4.46}$$

Из сравнения формул для числа идентифицирующих информаций (4.34) и максимальной средней удалённости (4.46) видно, что они совпадают, т.е.

$$H = g_{max}. \tag{4.47}$$

Для бесконечной знаковой цепи ($n \to \infty$) формула Мазура (4.34), в которой $\Delta_{aj} = \left(n/n_j\right) \to \left(1/P_j\right)$, принимает вид формулы К. Шеннона (2.13) для энтропии или количества информации. Такая информация используется только для дихотомической идентификации (но не для описания) отдельных сообщений.

Для «***регулярно-периодичной***» цепи, в которой интервалы всех однородных цепей равны: $\Delta_{ij} = \Delta_{aj} = \Delta_a = n/m$ и $n \gg m$ (т.е. в случае, когда все интервалы в цепи равны), формула (4.43) принимает вид

$$\Delta_{g_{max}} = \prod_{j=1}^{m} \left(\frac{n}{m}\right)^{\frac{m}{n}} = \left(\frac{n}{m}\right)^{\frac{m^2}{n}}. \tag{4.48}$$

Соответственно формула (4.46) примет вид

$$g_{max} = \frac{m^2}{n} \cdot \log_2 \frac{n}{m}. \tag{4.49}$$

Пример регулярно-периодичной цепи: ACBDACBDACBDACBDACBD.

Для *текстов и других нерегулярных последовательностей* формулы Мазура и Шеннона дают оценку строя только сверху, так как в этих случаях

$$D > \Delta_g; \tag{4.50}$$

$$H > g = \log_2 \Delta_g. \tag{4.51}$$

Таким образом, *формулы для среднего геометрического интервала и средней удаленности знаковой цепи обобщают формулы Мазура и Шеннона*, так-как, в отличие от последних при описании строя данной цепи учитывают не только мощность состава, но и взаимное расположение ее компонентов (знаков, слов).

Соответственно числовые характеристики строя на основе однородных цепей принимают минимальные значения для *«сплошных» последовательностей*, в которых все одинаковые элементы расположены подряд.

Пусть компоненты информационной цепи расположены на её позиции сплошными сериями по $n_j$ элементов в каждой и при вычислении характеристик используется привязка к концу цепи. Пример цепи с сериями сплошных последовательностей: `fffff22dddddzzzzzzwjjjj`. Заметим, что при любом расположении серий компонентов на позиции цепи

$$\sum_{j=1}^{m} n_j = n.$$

В этом случае объем строя цепи определяется в виде

$$V = \prod_{k=1}^{m-1} \sum_{j=1}^{k} n_j, \tag{4.52}$$

(при помощи индексов $k$ ведется счёт серий одинаковых компонентов, начиная с конца цепи).

При расположении компонентов по убыванию числа вхождений, когда

$$n_j \geq n_{j+1}, j = \overline{1,m},$$

то есть первые места на позиции последовательности занимает самый частый компонент, а последние места заняты самыми редкими (возможно одноразовыми) компонентами, объем строя примет минимальное значение

$$V = V_{min}.$$

Пример такой последовательности: *55555555555ДДДДДДДДТТТВЕК*.

Пусть количество компонентов во всех сериях одинаково и определяется в виде

$$n_j = {}^{n}/_{m}, \qquad j = \overline{1,m}.$$

После замены в формуле (4.52) суммирование произведением длин разных серий, она примет вид

$$V = \prod_{k=1}^{m-1} k \cdot \frac{n}{m} = (m-1)! \cdot \left(\frac{n}{m}\right)^{m-1}. \tag{4.53}$$

Средний геометрический интервал определится в виде

$$\Delta_g = \sqrt[n]{(m-1)! \cdot \left(\frac{n}{m}\right)^{m-1}}. \tag{4.54}$$

Логарифмируя, получим

$$g = \frac{1}{n}\log_2\left((m-1)! \cdot \left(\frac{n}{m}\right)^{m-1}\right) = \frac{1}{n}\log_2(m-1)! + \frac{m-1}{n}\log_2\frac{n}{m}. \tag{4.55}$$

Важно отметить взаимосвязь характеристик строя цепи, полученных на основе однородных и разнородных цепей (см. п 1.2.1 раздела 1). Так, для рассмотренных выше моделей регулярной и сплошной последовательностей, характеристики строя на основе разнородных цепей принимают соответственно минимальное и максимальное значения.

## 4.10. Контрольные вопросы и задания

1. Определение понятий регулярной и регулярно-периодичных цепей.

2. Покажите связь характеристик строя с формулами для подсчёта описательных и идентифицирующих информаций.

3. Определите понятие сплошной последовательности.

4. Определите формулы характеристик строя для сплошных последовательностей.

## 4.11. Формулы для подсчёта количества информации на основе структурного и статистического подходов

### 4.11.1. Формулы на основе структурного подхода

*1. Геометрические меры*

| № | Формула | Пояснение |
|---|---|---|
| 1 | $N_{\text{г}_1} = n_x$ | Длина «линии»; размер одномерного массива |
| 2 | $N_{\text{г}_2} = n_x \cdot n_y$ | «Площадь»; размер двумерного массива |
| 3 | $N_{\text{г}_3} = n_x \cdot n_y \cdot n_z$ | «Объем»; размер трехмерного массива |
| 4 | $N_{\text{г}_j} = \prod_{j=1}^{m} n_j$ | «Гиперобъем»; размер $j$-мерного массива |
| 5 | $n_j = \frac{j_{max} - j_{min}}{\Delta j}$ | Количество отсчётных значений в диапазоне измеряемой величины |

*2. Комбинаторные меры*

| № | Формула | Пояснение |
|---|---|---|
| 6 | $N_{\text{к}_1} = C_m^n = \frac{m!}{n! \cdot (m-n)!}$ | Число сочетаний из $m$ элементов по $n$ |
| 7 | $N_{\text{к}_2} = P_m = m! = n!$ | Число перестановок (из $m$ элементов), различающихся только их порядком следования |
| 8 | $N_{\text{к}_3} = P_m^{\text{пов}} = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_m!}$ | Число перестановок с неоднократными повторениями |
| 9 | $N_{\text{к}_4} = A_m^n = \frac{m!}{(m-n)!}$ | Число размещений из $m$ элементов по $n$ элементов |
| 10 | $N_{\text{к}_5} = A_m^{n_{\text{пов}}} = m^n$ | Число размещений с повторениями одинаковых элементов |

*3. Аддитивная мера*

| № | Формула | Пояснение |
|---|---|---|
| 11 | $I = \log_2 N$ | Число информаций по Р. Хартли |
| 12 | $I = \log_2 m^n = n \cdot \log_2 m$ | Для $N = m^n$ |

*4. Меры М. Мазура*

| № | Формула | Пояснение |
|---|---|---|
| 13 | $D = \prod_{j=1}^{m} \left(\frac{n}{n_j}\right)^{\frac{n_j}{n}}$ | Число описательных информаций одного сообщения в цепи |
| 14 | $D = n = m$ | Для $n = m$ |
| 15 | $H = \sum_{j=1}^{m} \frac{n_j}{n} \log_2 \frac{n}{n_j}$ | Число идентифицирующих информаций отдельного сообщения |
| 16 | $H = \log_2 n = \log_2 m$ | Для $n = m$ |
| 17 | $H = \log_2 D$ | Связь чисел идентифицирующих и описательных информаций |

*5. Меры расположения компонентов (строя) цепи*

| № | Формула | Пояснение |
|---|---|---|
| 18 | $\Delta_g = \prod_{j=1}^{m} \Delta_{gj}^{\frac{n_j}{n}}$ | Средний геометрический интервал; число описательных информаций одного интервала в цепи |
| 19 | $g = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \log_2 \Delta_{gj}$ | Средняя удалённость; число идентифицирующих информаций отдельного интервала в цепи |
| 20 | $g = \log_2 \Delta_g$ | Связь средней удалённости со средним геометрическим интервалом |

### 4.11.2. Формулы на основе статистического подхода

| № | Формула | Пояснение |
|---|---|---|
| 21 | $I = -\sum_{j=1}^{m} P_j \log_2 P_j$ | Количество информации по К. Шеннону |
| 22 | $I = \log_2 n = \log_2 m$ | При $P_j = 1/m$ и $m = n$ |

### 4.11.3. Связь мер Мазура, Шеннона и строя цепи

| № | Формула | Пояснение |
|---|---|---|
| 23 | $\Delta_{g_{max}} = \prod_{j=1}^{m} \Delta_{aj}^{\frac{n_j}{n}} = \prod_{j=1}^{m} \left(\frac{n}{n_j}\right)^{\frac{n_j}{n}} = D$ | Для $n/n_j = \Delta_{aj} = \Delta_{gj}$ |
| 24 | $g = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \log_2 \Delta_{aj} = \sum_{j=1}^{m} \frac{n_j}{n} \log_2 \frac{n}{n_j} = H$ | Для $n/n_j = \Delta_{aj} = \Delta_{gj}$ |
| 25 | $g = H = -\sum_{j=1}^{m} P_j \log_2 P_j$ | Для $n \to \infty,\ n_j/n \to P_j$ |
| 26 | $\Delta_g < D$ | Для всех остальных случаев $(\Delta_{aj} > \Delta_{gj})$ |
| 27 | $g < I$ | Для всех остальных случаев $(\Delta_{aj} > \Delta_{gj})$ |

# БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Гуменюк, А.С. Прикладная теория информации [Электронный ресурс] : учеб. пособие / А.С. Гуменюк, Н.Н. Поздниченко. – Омск : Изд-во ОмГТУ, 2015. – 167 с.

2. Алгоритмы анализа структуры сигналов и данных : монография / А.С. Гуменюк, Ю.Н. Кликушин, В.Ю. Кобенко, В.Н. Цыганенко ; под научн. ред. д-ра техн. наук Ю.Н. Кликушина. – Омск : Изд-во ОмГТУ, 2010.

3. О средствах формального анализа строя нуклеотидных цепей / А.С. Гуменюк, Н.Н. Поздниченко, С.Н. Шпынов, И.Н. Родионов // Математическая биология и биоинформатика, 2013. Т. 8. № 1. С. 373–397. Электронный доступ: http://www.matbio.org/article.php?journ_id=15&id=158

4. Шеннон, К. Работы по теории информации и кибернетике : [пер. с англ.] / К. Шеннон. – М. : Изд-во иностранной лит-ры, 1963. – 829 с.

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

6. Темников, Ф.Е. Теоретические основы информационной техники / Ф. Е. Темников, В.А. Афонин, В.И. Дмитриев. − 2-е изд., испр. и доп. − М. : Энергия, 1979. − 512 с.

7. Дмитриев, В.И. Прикладная теория информации : учеб. для студентов / В.И. Дмитриев. – М. : Высш.шк., 1989. – 320 с.

8. Душин, В.К. Теоретические основы информационных процессов и систем : учебник. – 2-е изд. – Издательско-торговая корпорация «Дашков и К», 2006. – 348 с.

9. Гуменюк, А.С. Элементы классической и современной теории информации (Организация информирования) : учеб. пособие / А.С. Гуменюк. – Омск : Изд-во ОмГТУ, 1994. – 82 с.

10. Гуменюк, А.С. Элементы информатики и теории информации : конспект лекций / А.С. Гуменюк. – Омск : Изд-во ОмГТУ, 2006. – 76 с.

11. Гуменюк, А.С. Прикладная теория информации : конспект лекций / А.С. Гуменюк. – Омск : Изд-во ОмГТУ, 2006. – 76 с.

12. Гуменюк, А.С. Информатика : учеб. пособие / А.С. Гуменюк, И.В. Потапов. – Омск : ИПЦ ОмГМА, 2012. – 176 с.

13. Шрейдер, Ю.А. Системы и модели / Ю.А. Шрейдер, А.А. Шаров. – М. : Радио и связь, 1982. – 152 с.

14. URL: [https://ru.wikipedia.org/wiki/%D0%9E%D0%BA%D1%83%D0%BB%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F]

15. URL: [http://habrahabr.ru/company/alee/blog/118398/]

16. Мазур, М. Качественная теория информации : [пер. с польск.] / М. Мазур. − М. : Мир, 1974. − 240 с.

17. Вентцель, Е.С. Теория вероятностей / Е.С. Вентцель. − М. : Наука, 1969. − 576 с.

18. Брайловский, И.В. Новый метод частичных обобщенных интервальных преобразований кодирования источников без памяти / И.В. Брайловский // Чебышевский сборник. – М. : Изд-во МГУ. 2003. Т. 4. Вып. 4. – С. 36–47.

19. Гуменюк, А.С. О длине двоичных слов для кодирования интервалов строя знаковой цепи // Информационные технологии и математическое моделирование (ИТММ-2005) : Матер. IV Всеросс. науч.-техн. конф. – Томск : Изд-во Том. унт-та, 2005. Ч. 1. – С. 44–46.

20. Реньи, А. Трилогия о математике : [пер. с венгр.] / А. Реньи. − М. : Мир, 1980. – 376 с.